ĐÀO HỮU HỒ

XÁC SUẤT THỐNG KÊ

(In lần thứ mười)

NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA HÀ NỘI

LỜI NÓI ĐẦU

Bộ môn Xác suất - Thống kê, khoa Toán - Cơ - Tin học, trường Đại học Tổng hợp trước kia và nay là trường Đại học Khoa học Tự nhiên - Đại học Quốc gia Hà Nội đã nhiều năm tham gia giảng dạy giáo trình Lý thuyết Xác suất và Thống kê ứng dụng cho các đối tượng không thuộc chuyên ngành toán học, trong đó có sinh viên các hệ của khoa Kinh tế, khoa Sinh học, khoa Tâm lý - Xã hội học..., đặc biệt là các học viên cao học và nghiên cứu sinh của các ngành đó.

Do quy mô đào tạo ngày càng mở rộng, để có một giáo trình thống nhất, chúng tôi đã biên soạn cuốn sách "**Xác suất Thống kê**" nhằm góp phần nâng cao hiệu quả của việc học tập và giảng dạy.

Nội dung của giáo trình được biên soạn vừa theo chương trình của ĐHQG vừa được mở rộng, nâng cao cho phù hợp với chương trình thi tuyển sau đại học môn Toán cao cấp của Bộ Giáo dục và Đào tạo cho các ngành Kinh tế - Nông - Lâm - Sinh - Y. Đồng thời giáo trình cũng tham khảo chương trình Xác suất thống kê của trường Đại học Nông nghiệp Wageningen - Hà Lan.

Với thời gian từ 60 đến 75 tiết được phân bố cho cả phần xác suất và cả phần thống kê (25 - 30 tiết cho phần Lý thuyết Xác suất; 35 - 45 tiết cho phần Thống kê ứng dụng), giáo trình chỉ có thể đề cập đến các khái niệm cơ bản và các kết luận phổ quát của Lý thuyết Xác suất và Thống kê ứng dụng. Chúng tôi đã cố gắng diễn đạt các kết luận và các khái niệm dưới ngôn ngữ giản dị, thích hợp với độc giả chỉ được trang bị những kiến thức tối thiểu về Toán cao cấp. Mặt khác, các thí dụ và bài tập minh hoạ đã được lựa chọn ít nhiều liên quan đến các bài toán thường gặp trong thực tế của các lĩnh vực Kinh tế, Nông nghiệp, Lâm nghiệp, Sinh học và Y học...

Giáo trình có thể sử dụng rộng rãi cho các đối tượng là sinh viên thuộc các hệ của các ngành kể trên. Giáo trình cũng là tài liệu tham khảo tốt cho giáo viên giảng dạy môn lý thuyết xác suất và thống kê của các trường đại học và cao đẳng.

Vì thời gian và khả năng có hạn, chắc chắn giáo trình khó tránh khỏi thiếu sót. Bộ môn Xác suất - Thống kê và Tác giả rất mong nhận được sự góp ý và lượng thứ của bạn đọc.

Hà Nội, ngày 27 tháng 9 năm 1996

PGS. PTS. Đào Hữu Hồ

Phần I

CƠ SỞ LÝ THUYẾT XÁC SUẤT

Chương I

KHÁI NIỆM CƠ BẢN VỀ XÁC SUẤT

§ 1. GIẢI TÍCH TỔ HỢP

1. Hoán vị:

Giả sử có n phần tử được xếp ở n vị trí. Ta đổi chỗ các phần tử cho nhau. Số cách đổi chỗ của n phần tử cho nhau được gọi là số hoán vị của n phần tử. Số cách ấy được chứng minh bằng

$$n! = n(n-1) \ldots 2.1$$

(tích của n thừa số nguyên bắt đầu từ n đến 1)

hay diễn đạt cách khác tương đương là: Ta có n phần tử và n vị trí Xếp n phần tử vào n vị trí đã cho sao cho mỗi chỗ có một và chỉ một phần tử. Số cách xếp như vậy cũng chính là n!

Ta có:

$n! = n(n-1)! = n(n-1)(n-2) \ldots (n-k+1)(n-k)!$

Quy ước:

$$0! = 1$$

Ví dụ 1: Ta có 3 người A, B, C xếp vào 3 chỗ ngồi. Rõ ràng ta có: 3! = 3.2 = 6 cách xếp như sau: ABC, ACB, BAC, BCA, CAB, CBA.

2. Tổ hợp:

Ta lấy ngẫu nhiên ra k phần tử từ tập gồm n phần tử ($k \leq n$), sao cho 2 cách lấy được gọi là khác nhau nếu giữa chúng có ít nhất một phần tử khác nhau. Số cách lấy ra k phần tử như vậy được gọi là tổ hợp chập k của n, ký hiệu là: C_n^k và được chứng minh bằng:

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

Ta có: $C_n^k = C_n^{n-k}$

bởi vì: $$C_n^{n-k} = \frac{n!}{(n-k)!(n-(n-k))!} = \frac{n!}{k!(n-k)!} = C_n^k$$

hoặc nói cách khác: *Một cách lấy ra k phần tử thì cũng chính là một cách lấy ra n − k phần tử còn lại.*

Rõ ràng: $C_n^0 = C_n^n = 1 \qquad C_n^1 = C_n^{n-1} = n$

Trong các tài liệu có thể có các cách ký hiệu khác nhau sau về C_n^k chẳng hạn ${}_nC_k$, $\binom{k}{n}$.

Ví dụ 2: Chọn ngẫu nhiên ra 2 người từ một nhóm 3 người A, B, C.

Ta có : $C_3^2 = \frac{3!}{2!1!} = 3$ cách chọn như sau: AB, AC, BC.

3. Chỉnh hợp:

Ta lấy ngẫu nhiên ra k phần tử từ 1 tập gồm n phần tử sao cho 2 cách lấy được gọi là khác nhau nếu giữa chúng hoặc có ít nhất một phần tử khác nhau, hoặc thứ tự lấy ra của các phần tử là khác nhau. Số cách lấy ra k phần tử như vậy được gọi là chỉnh hợp chập k của n, ký hiệu là A_n^k và được xác định bởi công thức sau:

$$A_n^k = n(n-1)....(n-k+1)$$

(tích của k thừa số nguyên liên tiếp, số lớn nhất là n).

Ví dụ 3: Chọn ngẫu nhiên 2 người từ một nhóm 3 người A, B, C để đi làm một nhiệm vụ nào đó. Ai được chọn đầu tiên sẽ được làm nhóm trưởng của nhóm ấy.

Từ ví dụ 2 ta sẽ có 3 cách chọn: AB, AC, BC. Bây giờ ta lại thêm được 3 cách chọn nữa: BA, CA, CB. Như vậy ta có 6 cách chọn có thể.

Theo công thức:

$$A_3^2 = 3.2 = 6$$

Nhận xét: Với một cách chọn theo nghĩa tổ hợp, ta có k! cách đổi chỗ cho nhau của k phần tử được chọn. Tuy nhiên k! cách này đối với nghĩa tổ hợp thì vẫn chỉ là 1 cách chọn, nhưng theo nghĩa của chỉnh hợp thì lại có k! cách. Vậy

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

Cũng có các cách ký hiệu khác của chỉnh hợp, chẳng hạn như ${}_nP_k$.

Ngoài 3 khái niệm vừa giới thiệu trên còn có một số cách chọn khác, nhưng đơn giản, chẳng hạn lấy ra từng phần tử một k lần, lấy có hoàn trả lại và không hoàn trả lại.

Đối với các bạn đọc lần đầu tiếp xúc với môn học này thường hay lúng túng trong việc xác định xem trường hợp đang xét thuộc vào cách chọn nào. Để giúp bạn đọc nhanh chóng nắm được vấn đề chúng tôi xin dừng lại làm một số nhận xét sau:

a. Khái niệm hoán vị tương đối khác biệt với khái niệm lấy ra k phần tử từ n phần tử, do đó ta gạt khái niệm hoán vị sang một bên và chỉ xét các cách lấy ra k phần tử từ tập n phần tử.

b. Có 4 cách lấy ra k phần tử từ tập gồm n phần tử. Đó là:

1- Cách lấy theo nghĩa tổ hợp

2- Cách lấy theo nghĩa chỉnh hợp

3- Cách lấy ra từng phần tử một không hoàn lại k lần

4- Cách lấy ra từng phần tử một có hoàn lại k lần.

- Trong 4 cách này, 2 cách đầu được lấy ra cùng một lúc, lấy 1 lần, 2 cách sau lấy từng phần tử một, lấy k lần.

- k phần tử được lấy ra theo 3 cách đầu là không có phần tử nào của tập hợp n phần tử có mặt lớn hơn 1 lần, có nghĩa là k phần tử này đều khác nhau (nếu tập gồm n phần tử khác nhau). Nói cách khác k phần tử lấy ra theo 3 cách đầu, mỗi phần tử của tập chung gồm n phần tử chỉ có mặt không quá 1 lần.

Trong khi đó trong k phần tử lấy ra theo cách 4 có thể có những phần tử được lấy trùng lại.

Như vậy nếu yêu cầu k phần tử chọn ra có thể trùng nhau cũng được thì đó là lấy từng phần tử và có hoàn lại. Nếu yêu cầu k phần tử chọn ra phải khác nhau thì thuộc một trong ba cách đầu.

c. Cách 3 phân biệt với hai cách đầu ở chỗ lấy k lần hay 1 lần.

d. Cách 1 và cách 2 phân biệt với nhau ở chỗ thứ tự các phần tử lấy ra có kể đến hay không kể đến, hoặc thứ tự lấy ra của các phần tử có ý nghĩa gì không?

e. Về định lượng giữa 4 cách lấy, ta có so sánh sau:

Cách 1		Cách 2		Cách 3		Cách 4
C_n^k	$<$	A_n^k	$=$	$n(n-1)...(n-k+1)$	$<$	n^k

f. Luật tích: Việc lấy các phần tử ra từ 1 tập hợp chung tuân theo luật tích: Nếu ta có 2 việc A_1 và A_2 khác nhau sao cho có k_1 cách thực hiện A_1, k_2 cách thực hiện A_2 thì số cách thực hiện hai việc A_1 và A_2 liên tiếp là $k_1.k_2$ (là tích chứ không phải là tổng như một số bạn đọc lầm tưởng !).

Ví dụ 4: Có bao nhiêu cách lấy ra 5 con bài từ cỗ bài tú lơ khơ 52 con sao cho trong 5 con bài lấy ra có 3 con At và 2 con 10.

Số cách lấy ra 3 con At: $C_4^3 = 4$

Số cách lấy ra 2 con 10: $C_4^2 = 6$

Số cách lấy ra 3 con At và 2 con 10 là: $C_4^3 C_4^2 = 4.6 = 24$

Luật tích trên mở rộng cho m việc $A_1, A_2, \ldots, A_m$, mỗi việc A_i có k_i cách có thể. Khi đó số cách có thể để $A_1, A_2, \ldots, A_m$ đều được thực hiện là: $k_1 \times k_2 \times \ldots \times k_m$.

Trường hợp đặc biệt: Lấy từng phần tử một k lần có hoàn lại từ tập n phần tử, A_i là lấy một phần tử ở lần thứ i. Vì có hoàn lại nên số cách thực hiện A_i là như nhau và bằng $C_n^1 = n$. Do đó lấy k lần có hoàn lại như đã nhận xét trên, số cách có thể là n^k. Còn nếu lấy k lần không hoàn lại thì số cách thực hiện A_1 là $C_n^1 = n$, số cách thực hiện A_2 là $C_{n-1}^1 = n-1, \ldots,$ số cách thực hiện A_k là $C_{n-k+1}^1 = n - k + 1$. Ta nhận được số cách lấy theo cách 3 là $n(n-1) \ldots (n-k+1)$.

§ 2. ĐỊNH NGHĨA XÁC SUẤT

Trước hết ta phải làm quen với khái niệm phép thử và biến cố.

– Gieo một đồng tiền trên một mặt phẳng: Đó là một phép thử. Kết quả có thể xảy ra khi gieo đồng tiền: "Xuất hiện mặt sấp" (mặt quốc huy) hoặc "xuất hiện mặt ngửa" (mặt chữ số).

"Xuất hiện mặt sấp" – Đó là một biến cố.

"Xuất hiện mặt ngửa" – Đó là một biến cố.

Hai biến cố này gọi là biến cố sơ cấp của phép thử trên.

– Đo nhiệt độ ngoài trời: Đó là một phép thử.

"Nhiệt độ ngoài trời là t^oC" – Đó là một biến cố.

– Gieo 1 con xúc xắc: Đó là một phép thử.

"Xuất hiện mặt k chấm ở mặt trên của xúc xắc" – Đó là một biến cố. Tương ứng k = 1, 2, 3, 4, 5, 6 – là 6 biến cố sơ cấp ứng với phép thử đã cho.

"Xuất hiện một mặt có 7 chấm" – Đó cũng là một biến cố, nhưng biến cố này không thể xảy ra khi phép thử được thực hiện. Ta gọi là biến cố không thể (hay biến cố trống, biến cố rỗng) ký hiệu là $\varnothing$.

"Xuất hiện mặt có số chấm ≤ 6 và ≥ 1" – Đây cũng là một biến cố. Biến cố này luôn luôn xảy ra khi ta gieo một con xúc xắc trên nền mặt phẳng. Biến cố này được gọi là biến cố chắc chắn – ký hiệu là Ω.

"Xuất hiện mặt có số chấm chẵn" – Đây cũng là một biến cố. Khi ta gieo con xúc xắc, biến cố này có thể xảy ra mà cũng có thể không xảy ra. Loại biến cố như thế ta gọi là biến cố ngẫu nhiên. Ta sẽ ký hiệu các biến cố ngẫu nhiên bởi các chữ A, B, C,...

Bây giờ ta nêu định nghĩa xác suất dạng cổ điển

Định nghĩa xác suất (dạng cổ điển)

Xuất phát từ giả thiết về tính đồng khả năng của các biến cố sơ cấp (hay tính đồng khả năng của các trường hợp có thể xảy ra) ta có định nghĩa sau:

Xác suất của biến cố A là một số không âm, ký hiệu P(A) (P viết tắt từ chữ Probability), biểu thị khả năng xảy ra biến cố A và được xác định như sau:

$$P(A) = \frac{m}{n} = \frac{\text{Số trường hợp thuận lợi cho A}}{\text{Số trường hợp có thể có khi phép thử thực hiện}}$$

(Những khả năng hoặc các biến cố sơ cấp - nếu chúng xảy ra thì suy ra A xảy ra - gọi là những trường hợp thuận lợi cho A).

Theo định nghĩa của biến cố không thể và biến cố chắc chắn thì số trường hợp thuận lợi cho biến cố không thể là 0 và mọi trường hợp có thể của phép thử đều là thuận lợi cho biến cố chắc chắn nên

$$P(\varnothing) = \frac{0}{n} = 0 \qquad , \qquad P(\Omega) = \frac{n}{n} = 1 \qquad 0 \leq P(A) \leq 1$$

Ví dụ 5: Rút ngẫu nhiên từ cỗ bài tú lơ khơ gồm 52 con bài ra 8 con bài. Tìm xác suất sao cho trong 8 con bài rút ra có:

a. 3 con At, 2 con 10, 1 con 2, 1 con K, 1 con J.

b. 2 con Cơ, 1 con Rô, 2 con Pic, 3 con Nhép.

c. 5 con màu đỏ, 3 con màu đen.

d. 3 con chủ bài (3 con đồng chất nào đó, chất đó đã được xác định trước, thí dụ 3 con Cơ).

Phép thử của ta là rút ngẫu nhiên ra 8 con bài. Số có thể là C_{52}^{8}.

Gọi A = {trong 8 con bài rút ra có 3 con At, 2 con 10, 1 con 2, 1 con K, 1 con J}

Tương tự B, C, D là các biến cố tương ứng với các câu b; c; d.

- Số trường hợp thuận lợi cho A theo luật tích sẽ là:

$$C_4^3 \times C_4^2 \times C_4^1 \times C_4^1 \times C_4^1 = 4 \,.\, 6 \,.\, 4 \,.\, 4 \,.\, 4$$

Do đó: $\quad P(A) = \dfrac{4^4 \,.\, 6}{C_{52}^{8}}$

- Số trường hợp thuận lợi cho B là:

$$C_{13}^2 \,.\, C_{13}^1 \,.\, C_{13}^2 \,.\, C_{13}^3 \rightarrow P(B) = \frac{C_{13}^2 \,.\, C_{13}^1 \,.\, C_{13}^2 \,.\, C_{13}^3}{C_{52}^{8}}$$

– Số trường hợp thuận lợi cho C là:

$$C_{26}^5 . C_{26}^3 \rightarrow P(C) = \frac{C_{26}^5 . C_{26}^3}{C_{52}^8}$$

– Số trường hợp thuận lợi cho D là:

$$C_{13}^3 . C_{39}^5 \rightarrow P(D) = \frac{C_{13}^3 . C_{39}^5}{C_{52}^8}$$

Ví dụ 6: Một số điện thoại ở Hà Nội gồm 6 chữ số. Giả sử ta chọn số điện thoại một cách ngẫu nhiên. Tìm khả năng để chọn được một số điện thoại gồm:

a. Chữ số 5 đầu tiên và 6 chữ số khác nhau.

b. Chữ số 5 đầu tiên và số điện thoại là số chẵn

c. Chữ số 5 đầu tiên, 5 chữ số còn lại khác nhau, chữ số cuối cùng chẵn

d. Chữ số 5 đầu tiên và 5 chữ số còn lại là một số đối xứng

e. Chữ số 5 đầu tiên, chữ số 0 cuối cùng và 4 chữ số giữa trùng với năm sinh của chủ hộ.

Ta thấy mọi số điện thoại đều lập nên từ tập hợp gồm 10 chữ số: 0, 1, 2,...,9. Mà 1 số điện thoại có thể có các chữ số trùng nhau. Như vậy để lập nên các số điện thoại gồm 6 chữ số thì phép thử của ta chính là chọn ngẫu nhiên từng số một có hoàn lại 6 lần. Do đó số trường hợp có thể là 10^6.

Đặt A, B, C, D, E là các biến cố tương ứng với 5 trường hợp cần tìm xác suất.

a. A = {Chữ số 5 đầu và 6 chữ số khác nhau}. Số thuận lợi cho A là 1. 9. 8. 7. 6. 5 hoặc tương đương $1.A_9^5$

$$P(A) = \frac{A_9^5}{10^6} = \frac{9.8.7.6.5}{10^6} = \frac{15120}{10^6} = 0{,}01512$$

b. Số thuận lợi cho B là: $1.10.10.10.10.5 = 5.10^4$ (có 5 chữ số chẵn: 0, 2, 4, 6, 8)

$$P(B) = \frac{5 \cdot 10^4}{10^6} = 0{,}05$$

c. Số thuận lợi cho C là: 1. 9. 8. 7. 6. 5 hoặc tương đương $1.5.A_9^4$ (chữ số đầu tiên có 1 khả năng, chữ số cuối cùng có 5 khả năng, 4 chữ số còn lại khác nhau và khác chữ số cuối cùng)

$$P(C) = \frac{5A_9^4}{10^6} = \frac{5.9.8.7.6}{10^6} = 0{,}01512$$

d. Số thuận lợi cho D là:

$$1.10.10.10.1.1 = 10^3 \rightarrow P(D) = \frac{10^3}{10^6} = 0{,}001$$

e. Số thuận lợi cho E là:

$$1.1.1.1.1.1 = 1 \rightarrow P(E) = \frac{1}{10^6} = 10^{-6}$$

Ví dụ 7: (xem [3]) Có 5 tấm bìa vuông như nhau. Trên mỗi tấm bìa có ghi một chữ cái H, O, N, A, I. Ta sắp xếp ngẫu nhiên 5 tấm bìa đó thành một hàng ngang. Tìm xác suất để được chữ HANOI

Số có thể: 5!

Số thuận lợi: 1

Xác suất cần tìm là 1/5!

Ví dụ 8: Một tổ gồm 10 người tổ chức buổi liên hoan ngồi quanh bàn tròn. Mọi người ngồi vào chỗ một cách ngẫu nhiên. Tìm khả năng để cho A và B ngồi cạnh nhau.

Số có thể: 10!

Số thuận lợi: 10.2.8!.

(A có thể ngồi 1 trong 10 chỗ, 2 chỗ bên cạnh dành cho B, 8 chỗ còn lại dành cho 8 người kia).

Vậy xác suất cần tìm là: $\dfrac{10.2.8!}{10!} = \dfrac{2}{9}$

Định nghĩa xác suất theo phương pháp thống kê:

Làm đi làm lại một phép thử nào đó n lần mà có m lần biến cố A xuất hiện thì tỷ số m/n gọi là tần suất của biến cố A.

Khi n thay đổi, tần suất m/n cũng thay đổi nhưng nó luôn dao động quanh một số cố định nào đó, n càng lớn thì m/n càng gần số cố định đó. Số cố định ấy được gọi là xác suất của biến cố A theo nghĩa thống kê. Trên thực tế khi n đủ lớn ta xấp xỉ P(A) bởi m/n.

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

Ví dụ 9: Buffon đã gieo một đồng tiền cân đối, đồng chất 4040 lần thấy có 2048 lần xuất hiện mặt sấp.

$$\frac{m}{n} = 0{,}5069$$

Pearson đã gieo 12000 lần thấy 6019 lần sấp.

$$\frac{m}{n} = 0{,}5016$$

Pearson đã gieo 24000 lần thấy 12012 lần sấp.

$$\frac{m}{n} = 0{,}5005$$

Số cố định cần tìm trong trường hợp này là 0,5. Tức là xác suất xuất hiện mặt sấp khi ta gieo đồng tiền cân đối và đồng chất bằng 0,5.

Một ví dụ cổ điển nữa là vấn đề tính xác suất sinh con trai hay con gái. Vấn đề này đã được nhiều nhà sinh lý học, nhân chủng học nghiên cứu từ lâu. Người cổ Trung Hoa từ năm 2228 trước Công nguyên đã qua thống kê kinh nghiệm đưa ra tỷ số sinh con gái là 0,5. Laplace nghiên cứu sinh đẻ ở London, Petersburg và Berlin trong 10 năm và đưa ra tỷ số sinh cháu gái là 21/43.

Đacnon nghiên cứu sinh đẻ ở Pháp và cho các số liệu sau:

Năm	1806	1816	1836	1856	1903	1920
Tần suất sinh con gái	0,485	0,484	0,485	0,487	0,488	0,489

Cramer nghiên cứu ở Thụy Điển trong năm 1935: Số cháu sinh là 88273. Số cháu gái là 42591. Vậy tần suất là 0,4825.

Nhận xét: Định nghĩa xác suất dạng thống kê hay định nghĩa xác suất theo tần suất chỉ cho ta giá trị xấp xỉ và mức độ chính xác của việc xấp xỉ tùy thuộc vào số lần thực hiện phép thử.

§ 3. QUAN HỆ GIỮA CÁC BIẾN CỐ

– Ta thực hiện một phép thử. Các kết quả có thể có khi phép thử được thực hiện gọi là các biến cố sơ cấp (hoặc các biến cố cơ bản).

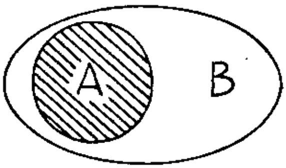
– Quan hệ kéo theo: Biến cố A gọi là kéo theo biến cố B và ký hiệu là $A \subset B$ nếu và chỉ nếu A xảy ra thì suy ra B xảy ra.

Mô tả hình học của quan hệ này có thể hình dung A là tập con của B, tập A được chứa trong B.

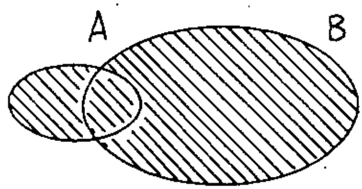
– Quan hệ tương đương: 2 biến cố A và B gọi là tương đương với nhau, ký hiệu $A = B$, khi và chỉ khi $A \subset B$ và $B \subset A$.

– Tổng của hai biến cố: Tổng của hai biến cố A và B là một biến cố được ký hiệu là $A \cup B$, sao cho biến cố tổng $A \cup B$ xảy ra khi và chỉ khi hoặc A xảy ra hoặc B xảy ra (nói cách khác khi và chỉ khi có ít nhất một trong các biến cố A và B xảy ra).

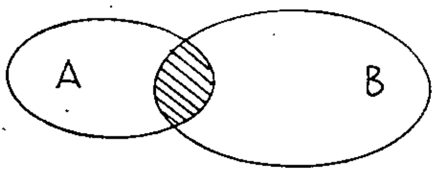
– Tích của hai biến cố: Tích của hai biến cố A và B là một biến cố được ký hiệu là $A \cap B$ hoặc AB, sao cho biến cố tích AB xảy ra khi và chỉ khi A xảy ra và B xảy ra.



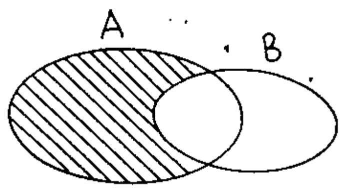
$A \subset B$



$A \cup B$



AB



$A \setminus B$

– Hai biến cố xung khắc: A và B được gọi là xung khắc với nhau nếu $AB = \varnothing$

Nói cách khác A và B gọi là xung khắc nếu xảy ra biến cố này thì không xảy ra biến cố kia.

– Hiệu của biến cố A và biến cố B: là một biến cố được ký hiệu là $A \setminus B$ sao cho biến cố hiệu $A \setminus B$ xảy ra khi và chỉ khi A xảy ra nhưng B không xảy ra.

– Biến cố đối lập: $\overline{A}$ được gọi là biến cố đối lập của biến cố A khi và chỉ khi $\overline{A}$ xảy ra thì A không xảy ra và ngược lại, tức là:

$$\overline{A} = \Omega \setminus A$$

Nhận xét:

a. Ta có thể mở rộng các quan hệ biến cố cho 3, 4 biến cố hoặc nhiều hơn nữa.

b. Khi xét quan hệ giữa các biến cố bạn đọc không nên dùng minh hoạ hình học để thay thế cho định nghĩa mà phải bám chặt vào định nghĩa để xét, bởi vì minh hoạ hình học chỉ nhằm mục dích giúp bạn dọc hình dung ra một cách cơ bản các mối quan hệ, biểu diễn hình học không thể phản ánh chính xác trong mọi trường hợp.

Chẳng hạn ta xét trường hợp sau:

A = {Anh thứ 1 bắn trúng bia}

B = {Anh thứ 2 bắn trúng bia}

Khi đó khó có thể mô tả biến cố tích AB, và nếu ta biểu diễn biến cố tích AB là phần chung giữa A và B thì trong trường hợp này phần chung đó biểu diễn bằng hình học thế nào. Nhiều bạn đọc sẽ nhầm rằng phần chung giữa A và B không có. Vì thế A và B xung khắc,... Nhưng ta bám sát định nghĩa thì biến cố tích AB là anh thứ 1 bắn trúng và anh thứ 2 bắn trúng, nghĩa là cả 2 anh đều bắn trúng. Hai biến cố này không xung khắc..

Sau khi mở rộng quan hệ biến cố cho một số lớn hơn 2 biến cố, ta xét nhóm đầy đủ các biến cố:

– n biến cố $A_1, A_2, ..., A_n$ gọi là một nhóm đầy đủ các biến cố nếu thỏa mãn hai điều kiện sau:

1. Chúng xung khắc với nhau từng đôi một: $A_iA_j = \varnothing$, $i \neq j$.

2. Tổng của n biến cố tương đương với biến cố chắc chắn:

$$A_1 \cup A_2 \cup ... \cup A_n = \Omega$$

Ví dụ đơn giản của nhóm đầy đủ là: $\overline{A}$, A (ở đây n = 2)

– Quy tắc đối ngẫu De Morgan (đọc là Đờ–moóc–gan):

$$\overline{A \cup B \cup C} = \overline{A}\,\overline{B}\,\overline{C}$$

$$\overline{ABC} = \overline{A} \cup \overline{B} \cup \overline{C}$$

Quy tắc trên mở rộng cho một số hữu hạn bất kỳ các biến cố.

– Phép tổng và tích biến cố có tính chất giao hoán, kết hợp à phân phối.

$$A(B \cup C) = AB \cup AC$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Ví dụ 10: Một ban nhạc nhẹ đang muốn tuyển thêm thành iên mới. Người được chọn sẽ là người đáp ứng một số trong ác điều kiện sau:

Điều kiện A = {tự do vào tối thứ 7}

Điều kiện B = {phải yêu thích âm nhạc}

Điều kiện C = {biết chơi một nhạc cụ nào đấy}

Điều kiện D = {phải biết và sử dụng được nhạc lý}

Điều kiện E = {có giọng tốt}

Người được chọn sẽ là người đáp ứng các điều kiện hỗn hợp sau:

$$A \cap B \cap \left[(C \cap D) \cup E\right]$$

Vậy để được chọn vào ban nhạc thì phải đáp ứng hoặc 3 iều kiện A, B và E, hoặc 4 điều kiện A, B, C, D. Thoả mãn ả 4 điều kiện A, C, D và E nhưng không thỏa mãn B cũng ẫn không được chấp nhận.

Ví dụ 11: Hai người cùng bắn, mỗi người bắn một viên vào ia. Gọi $A_i \triangleq$ {Người thứ i bắn trúng bia}, $i = 1, 2$.

Hãy viết biến cố sau qua A_1, A_2

a. Chỉ có người thứ 1 bắn trúng: $A_1\overline{A}_2$

b. Có một người bắn trúng: $A_1\overline{A}_2 \cup \overline{A}_1A_2$

c. Có ít nhất một người bắn trúng: $A_1 \cup A_2$

d. Cả hai cùng bắn trúng: A_1A_2

e. Không có ai bắn trúng: $\overline{A}_1\overline{A}_2$ hoặc $\overline{A_1 \cup A_2}$

f. Nhóm đầy đủ biến cố: $A_1, \overline{A}_1$ hoặc $\overline{A}_1 \overline{A}_2$, $\overline{A}_1 A_2$, $A_1\overline{A}_2$, A_1A_2

§ 4. CÔNG THỨC CỘNG XÁC SUẤT

$$P(A \cup B) = P(A) + P(B) - P(AB) \qquad (1)$$

Ta chứng minh (1) theo định nghĩa cổ điển.

Giả sử số trường hợp có thể có của phép thử là n. Số trường hợp thuận lợi cho A, B, AB tương ứng là n_A, n_B, n_{AB}. Số trường hợp thuận lợi cho $A \cup B$ là $n_A + n_B - n_{AB}$.

Do đó:

$$P(A \cup B) = \frac{n_A + n_B - n_{AB}}{n} = \frac{n_A}{n} + \frac{n_B}{n} - \frac{n_{AB}}{n}$$

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

Nếu A, B xung khắc:

$AB = \varnothing$, $P(AB) = P(\varnothing) = 0$ nên

$$P(A \cup B) = P(A) + P(B) \qquad (2)$$

Từ (2) suy ra

$$P(A) = 1 - P(\overline{A}) \qquad (3)$$

(Vì A, $\overline{A}$ xung khắc, $A \cup \overline{A} = \Omega$ nên $P(A \cup \overline{A}) = P(\Omega) = 1$. Mặt khác $P(A \cup \overline{A}) = P(A) + P(\overline{A})$, tức là $P(A) + P(\overline{A}) = 1$).

Mở rộng (1) và (2) cho trường hợp 3 biến cố:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) \qquad (4)$$

Nếu các biến cố A, B, C xung khắc với nhau từng cặp thì:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \qquad (5)$$

§ 5. CÔNG THỨC NHÂN XÁC SUẤT

1. Xác suất có điều kiện:

Xác suất có điều kiện của biến cố A với điều kiện biến cố đã xảy ra là một con số không âm, được ký hiệu P(A/B), nó ểu thị khả năng xảy ra biến cố A trong tình huống biến cố B ĩ xảy ra. Giả sử :

- Số kết quả có thể có khi phép thử thực hiện là n
- Số trường hợp thuận lợi cho biến cố B là n_B.
- Số trường hợp thuận lợi cho cả biến cố A và biến cố B là n_{AB}.

Khi đó theo định nghĩa

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

Vậy
$$P(A/B) = \frac{P(AB)}{P(B)} \qquad (6)$$

Nhìn chung $\quad P(A/B) \neq P(A)$

Ví dụ 12: Giả sử một lớp chia làm 3 nhóm thực tập. Nhóm có 30 sinh viên trong đó có 10 nữ. Nhóm II có 25 sinh viên :ong đó có 10 nữ. Nhóm III có 25 sinh viên trong đó có 8 nữ. họn ngẫu nhiên trong lớp ra một sinh viên.

Gọi $\quad$ A = {sinh viên được chọn ra là nữ}

$\qquad\quad$ B = {sinh viên được chọn ra thuộc nhóm II}

Ta có:

$$P(A) = \frac{28}{80} = 0{,}35; \quad P(A/B) = \frac{10}{25} = 0{,}4; \quad P(A) \neq P(A/B)$$

Tính chất của xác suất có điều kiện:

a. $0 \leqslant P(A/B) \leqslant 1$

b. $P(B/B) = 1$

c. Nếu $AC = \varnothing$ thì $P(A \cup C/B) = P(A/B) + P(C/B)$

d. $P(\overline{A}/B) = 1 - P(A/B)$

2. Hai biến cố độc lập:

Hai biến cố A và B gọi là độc lập với nhau nếu:

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

Xét tính độc lập của 2 biến cố bằng cách kiểm chứng sự thỏa mãn 1 trong 2 đẳng thức trên (theo định lượng) thường không phải là dễ dàng và thuận tiện. *Về định tính, thực chất đẳng thức trên có nghĩa là việc xảy ra biến cố này hay không đều không ảnh hưởng đến khả năng xảy ra biến cố kia.* Do đó trong thực tế ta có thể dùng tính chất này để xét tính độc lập của hai biến cố.

Ví dụ 13: Hai chị A và B cùng đến nhà hộ sinh để sinh con.

Gọi A = {chị A sinh con trai}

B = {chị B sinh con trai}

Rõ ràng dù A xảy ra hay $\overline{A}$ xảy ra, tức là dù chị A sinh con trai hay con gái đều không ảnh hưởng đến khả năng sinh con trai hay gái của chị B. Hai biến cố A và B độc lập.

Nhận xét: Nếu A, B độc lập với nhau thì

$A, \overline{B}$ độc lập với nhau

$\overline{A}, B$ độc lập với nhau

$\overline{A}, \overline{B}$ độc lập với nhau

(chẳng hạn $P(\overline{A}/B) = 1 - P(A/B) = 1 - P(A) = P(\overline{A})$ tức là $\overline{A}$, B độc lập với nhau)

3. Công thức nhân xác suất:

$$P(AB) = P(A) \,.\, P(B/A) = P(A/B) \,.\, P(B) \qquad (6')$$

Nếu A, B độc lập: $P(AB) = P(A).\ P(B)$ (7)

Chứng minh (6') được trực tiếp suy từ (6).

Nhận xét: Hai xác suất P(A/B) và P(B/A) trong (6') có vai trò tương đương. Xác suất P(AB) có thể tính qua 1 trong 2 biểu thức ở vế phải của (6') đều được. Song trong những trường hợp cụ thể, ta có thể dùng được biểu thức này mà có thể không nên dùng biểu thức kia, chẳng hạn trường hợp phép thử liên quan đến B xảy ra trước phép thử liên quan đến A, khi đó P(A/B) dễ dàng tìm được nhưng P(B/A) có khi khó xác định được hoặc không có nghĩa.

Ta xét trường hợp sau:

Một hộp có 5 bi đỏ và 3 bi xanh. Lấy ngẫu nhiên ra 1 viên bi ta được bi đỏ. Sau đó lấy tiếp 1 bi nữa trong số bi còn lại.

Gọi B = {viên bi lấy ra lần đầu là đỏ}

A = {viên bi lấy ra lần sau là đỏ}

Dễ dàng ta có $P(A/B) = \frac{4}{7}$

Trong khi đó P(B/A) khó tìm hơn nhiều và chắc không ít bạn đọc sẽ lúng túng khi phải tính nó.

Do đó trường hợp này ta nên dùng:

$$P(AB) = P(B).P(A/B) = \frac{5}{8}.\frac{4}{7} = \frac{5}{14}$$

Ví dụ 14: Một người có 3 con gà mái, 2 con gà trống nhốt chung trong một lồng. Một người đến mua, người bán gà bắt ngẫu nhiên ra một con. Người mua chấp nhận mua con đó.

a. Tìm xác suất để người đó mua được con gà mái.

Người thứ hai đến mua, người bán gà lại bắt ngẫu nhiên ra một con.

b. Tìm xác suất người thứ hai mua được gà trống.

c. Xác suất này sẽ bằng bao nhiêu nếu người bán gà quên mất rằng con gà bán cho người thứ nhất là gà trống hay mái?

Gọi B_i = {người thứ i mua được gà mái}, i = 1, 2. Ta có:

a. $$P(B_1) = \frac{3}{5} = 0,6$$

b. Vì người thứ hai mua sau khi người thứ nhất đã mua xong, cho nên:

$$P(\overline{B}_2/B_1) = \frac{2}{4} = 0,50$$

c. Ta có $\overline{B}_2 = \Omega . \overline{B}_2 = (B_1 \cup \overline{B}_1)\overline{B}_2 = B_1\overline{B}_2 \cup \overline{B}_1\overline{B}_2$

Vì $B_1\overline{B}_1 = \varnothing$ nên $(B_1\overline{B}_2) \cap (\overline{B}_1\overline{B}_2) = \varnothing$

Theo (2):

$$P(\overline{B}_2) = P(B_1\overline{B}_2) + P(\overline{B}_1\overline{B}_2)$$

Do B_1, B_2 không độc lập nên $B_1, \overline{B}_2$ và $\overline{B}_1, \overline{B}_2$ không độc lập, theo (6'):

$$P(\overline{B}_2) = P(B_1) . P(\overline{B}_2/B_1) + P(\overline{B}_1) . P(\overline{B}_2/\overline{B}_1)$$

$$= \frac{3}{5} . \frac{2}{4} + \frac{2}{5} . \frac{1}{4} = \frac{8}{20} = 0,40$$

Ví dụ 15: Hai người cùng bắn vào mục tiêu một cách độc lập với nhau. Xác suất trúng đích của chiến sỹ A là 0,8 còn của chiến sỹ B là 0,7. Tìm khả năng xảy ra các tình huống sau:

a. Chiến sỹ A bắn trúng đích ngay trong 3 phát đầu

b. Chiến sỹ B bắn trúng đích ngay từ phát thứ 3

c. Hai người cùng bắn trúng đích khi mỗi người bắn một phát.

d. Ít nhất có 1 người bắn trúng đích khi mỗi người bắn một phát.

Gọi A_i = {chiến sỹ A bắn trúng đích ở phát thứ i}, i = 1, 2, 3

B_i = {chiến sỹ B bắn trúng đích ở phát thứ i}, i = 1, 2, 3

D_1, D_2, D_3, D_4 là 4 biến cố tương ứng cần tìm xác suất trong 4 câu a, b, c, d ở trên. Ta có:

$$D_1 = A_1 \cup A_2 \cup A_3$$

$$D_2 = \overline{B}_1 \overline{B}_2 B_3$$

$$D_3 = A_1 B_1$$

$$D_4 = A_1 \cup B_1$$

A_i, B_i độc lập với nhau, A_1, A_2, A_3 độc lập, B_1, B_2, B_3 độc lập. Nhưng A_i, B_i không xung khắc.

Vậy:

$$P(D_1) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3)$$
$$-P(A_2A_3) + P(A_1A_2A_3)$$

$$P(D_1) = 0{,}8 \,.\, 3 - 3 \,.\, 0{,}8 \,.\, 0{,}8 + 0{,}8 \,.\, 0{,}8 \,.\, 0{,}8 = 0{,}992$$

$$P(D_2) = P(\overline{B}_1) \,.\, P(\overline{B}_2) \,.\, P(B_3) = (1 - 0{,}7) \,.\, (1 - 0{,}7) \,.\, 0{,}7 = 0{,}063$$

$$P(D_3) = P(A_1) \,.\, P(B_1) = 0{,}8 \,.\, 0{,}7 = 0{,}56$$

$$P(D_4) = P(A_1) + P(B_1) - P(A_1B_1) = 0{,}8 + 0{,}7 - 0{,}56 = 0{,}94$$

Ví dụ 16: Bắn liên tiếp vào một mục tiêu cho đến khi có một viên đạn đầu tiên trúng mục tiêu thì ngừng bắn. Tìm xác suất sao cho phải bắn đến viên thứ 4, biết rằng xác suất trúng mục tiêu của mỗi lần bắn là như nhau và bằng 0,3.

Gọi A_i = {viên thứ i trúng mục tiêu}, i = 1, 2...

A = {bắn đến viên thứ tư mới ngừng}

$$A = \overline{A}_1 \overline{A}_2 \overline{A}_3 A_4$$

Các biến cố A_1, A_2, A_3, A_4 không độc lập vì việc xảy ra biến cố A_i sẽ ảnh hưởng đến khả năng xảy ra A_{i+1}

$$P(A_{i+1}/A_i) = 0\,,\; P(A_{i+1}/\overline{A}_i) = 0{,}3$$

Do đó:

$$P(A) = P(\overline{A}_1)\,.\,P(\overline{A}_2/\overline{A}_1)\,.\,P(\overline{A}_3/\overline{A}_1\overline{A}_2)\,.\,P(A_4/\overline{A}_1\overline{A}_2\overline{A}_3)$$

$$= \left[1 - P(A_1)\right].\left[(1 - P(A_2/\overline{A}_1)\right].\left[(1 - P(A_3/\overline{A}_1\overline{A}_2)\right].$$

$$.P(A_4/\overline{A}_1\,\overline{A}_2\,\overline{A}_3)$$

Hơn nữa: $\overline{A}_2 \subset \overline{A}_1$, $\overline{A}_3 \subset \overline{A}_2 \subset \overline{A}_1$

nên $\quad \overline{A}_1\overline{A}_2 = \overline{A}_2$, $\overline{A}_1\,\overline{A}_2\,\overline{A}_3 = \overline{A}_3$

$$P(A) = (1 - 0{,}3)\,(1 - 0{,}3)(1 - 0{,}3)\,.\,0{,}3 = 0{,}1029$$

Ví dụ 17: Để dập tắt nạn sâu bệnh hại lúa, Đội bảo vệ thực vật của Hợp tác xã đã tiến hành phun thuốc 3 lần liên tiếp trong 1 tuần. Xác suất sâu bị chết sau lần phun thứ I là 0,5. Nếu sâu sống sót thì khả năng bị chết sau lần phun thứ II là 0,7. Tương tự sau lần phun thứ III là 0,9. Tìm xác suất sâu bị chết sau đợt phun thuốc.

Gọi $\quad A$ = {sâu bị chết sau đợt phun thuốc}

A_i = {sâu bị chết sau lần phun thứ i}, $i = 1, 2, 3$

A_1, A_2, A_3 không độc lập.

Ta có: $A = A_1 \cup \overline{A}_1A_2 \cup \overline{A}_1\,\overline{A}_2A_3$. Áp dụng (5) và (6') ta tính được P(A).

Hoặc ta làm cách sau:

$\overline{A} = \overline{A}_1\,\overline{A}_2\,\overline{A}_3$ = {sâu sống sót sau đợt phun thuốc}

$$P(\overline{A}) = P(\overline{A}_1)\,.\,P(\overline{A}_2/\overline{A}_1)\,.\,P(\overline{A}_3/\overline{A}_1\,\overline{A}_2)$$

$$= \left[1 - P(A_1)\right].\left[1 - P(A_2/\overline{A}_1)\right].\left[1 - P(A_3/\overline{A}_2)\right]$$

$$= (1 - 0{,}5)\,.\,(1 - 0{,}7)\,.\,(1 - 0{,}9) = 0{,}015$$

$$P(A) = 1 - P(\overline{A}) = 1 - 0{,}015 = 0{,}985.$$

Ví dụ 18: Một hộp chứa 3 bi trắng, 7 bi đỏ và 15 bi xanh. Một hộp khác chứa 10 bi trắng, 6 bi đỏ và 9 bi xanh. Ta lấy ngẫu nhiên từ mỗi hộp ra một bi. Tìm xác suất để hai viên bi lấy ra là cùng màu (Xem [1]).

Gọi A = {hai viên bi lấy ra là cùng màu}

1_T = {viên bi rút ra từ hộp 1 là bi trắng}. Tương tự 1_D, 1_X.

2_T = {viên bi rút ra từ hộp 2 là bi trắng}. Tương tự 2_D, 2_X.

Ta có: $A = 1_T 2_T \cup 1_D 2_D \cup 1_X 2_X$

Mỗi viên bi chỉ có một màu nên tính xung khắc thỏa mãn. Việc rút bi từ hộp 1 và từ hộp 2 độc lập với nhau. Vậy:

$$P(A) = P(1_T).P(2_T) + P(1_D)P(2_D) + P(1_X)P(2_X)$$

$$= \frac{3}{25}\frac{10}{25} + \frac{7}{25}\frac{6}{25} + \frac{15}{25}\frac{9}{25} = \frac{207}{625}$$

Ví dụ 19: Ở một nông trường chăn nuôi người ta phát hiện một trại lợn bị bệnh. Lợn có thể bị bệnh A với xác suất 0,7, bị bệnh B với xác suất 0,5, có thể bị cả hai bệnh A và B. Người ta dùng cả hai loại thuốc T_1 và T_2 để điều trị cho lợn.

Điều trị loại thuốc T_1, xác suất khỏi bệnh của lợn bệnh A là 0,8, của lợn bệnh B là 0,6 và của lợn bị cả hai bệnh A và B là 0,3.

Điều trị loại thuốc T_2, khả năng khỏi của lợn bệnh A, bệnh B, bệnh AB tương ứng là 0,6 ; 0,7 và 0,4.

a. Tìm xác suất để lợn bị cả bệnh A và B

b. Khả năng chữa khỏi cho lợn bằng T_1, bằng T_2. Loại thuốc nào cho khả năng chữa khỏi cao hơn.

Gọi A = {lợn bị bệnh A}

B = {lợn bị bệnh B}

AB = {lợn bị 2 bệnh A và B}

3 biến cố này không xung khắc từng đôi một.

D = {lợn bị bệnh}

$D = A \cup B \cup AB$

Gọi E_i = {lợn khỏi bệnh do điều trị thuốc T_i}, $i = 1, 2$

a. Vì lợn bị bệnh A hay B là độc lập nên:

$$P(AB) = P(A).\ P(B) = 0,7\ .\ 0,5 = 0,35$$

b. Biến cố "lợn bị bệnh và được chữa khỏi bằng thuốc T_1" là:

$$DE_1 = (A \cup B \cup AB)E_1 = AE_1 \cup BE_1 \cup (AB)E_1$$

Theo công thức tính xác suất (4) ta có

$$P(DE_1) = P(AE_1) + P(BE_1) + P[(AB)E_1] - P(AE_1BE_1)$$
$$- P(AE_1(AB)E_1) - P(BE_1(AB)E_1) + P(AE_1.BE_1.(AB)E_1)$$
$$= P(A).P(E_1/A) + P(B).P(E_1/B) + P(AB).P(E_1/AB)$$
$$- P(AE_1)P(BE_1) - P(AE_1)P((AB)E_1) - P(BE_1)P((AB)E_1)$$
$$+ P(AE_1)P(BE_1)P((AB)E_1)$$
$$= 0,7\ .\ 0,8 + 0,5\ .\ 0,6 + 0,35\ .\ 0,3 - 0,56\ .\ 0,30$$
$$- 0,56\ .\ 0,105 - 0,30\ .\ 0,105 + 0,56\ .\ 0,30\ .\ 0,105$$
$$= 0,72434$$

Biến cố "lợn bị bệnh và được chữa khỏi bằng thuốc T_2 là:

$$DE_2 = (A \cup B \cup AB)E_2 = AE_2 \cup BE_2 \cup (AB)E_2$$
$$P(DE_2) = P(AE_2) + P(BE_2) + P[(AB)E_2] - P(AE_2BE_2)$$
$$- P(AE_2(AB)E_2) - P(BE_2(AB)E_2) + P(AE_2.BE_2.(AB)E_2)$$
$$= P(A).P(E_2/A) + P(B)\ .\ P(E_2/B) + P(AB)\ .\ P(E_2/AB)$$
$$- P(AE_2\ BE_2) - P(AE_2(AB)E_2) - P(BE_2(AB)E_2)$$
$$+ P(AE_2.BE_2.ABE_2)$$
$$= 0,7.0,6 + 0,5.0,7 + 0,35.0,4 - 0,42.0,35$$
$$- 0,42.0,14 - 0,35.0,14 + 0,42.0,35.0,14$$
$$= 0,67578$$

Như vậy khả năng khỏi bệnh do điều trị thuốc T_1 cao hơn thuốc T_2.

§ 6. CÔNG THỨC XÁC SUẤT ĐẦY ĐỦ VÀ BAYES

Giả sử $B_1, B_2, ..., B_n$ là một nhóm đầy đủ các biến cố. Xét biến cố A sao cho A xảy ra chỉ khi một trong các biến cố $B_1, B_2, ..., B_n$ xảy ra. Nói cách khác A xảy ra thì một biến cố B_i nào đó xảy ra. Khi đó:

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A/B_i) \tag{8}$$

Công thức (8) được gọi là công thức xác suất đầy đủ.

$$P(B_k/A) = \frac{P(B_k) \cdot P(A/B_k)}{\sum_{i=1}^{n} P(B_i) \cdot P(A/B_i)} \tag{9}$$

Công thức (9) được gọi là công thức Bayes.

Ta chứng minh công thức (8)

$$A = A\Omega = (B_1 \cup B_2 \cup ... \cup B_n)A = B_1A \cup ... \cup B_nA$$

Các biến cố B_iA, $i = 1, 2, ..., n$ xung khắc nên:

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

Công thức (9) suy ra từ đẳng thức sau:

$$P(B_k)P(A/B_k) = P(A) \cdot P(B_k/A) = \left[P(B_kA)\right]$$

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

Ví dụ 20: Một lô hạt giống được phân làm 3 loại. Loại I chiếm 2/3 số hạt cả lô. Loại II chiếm 1/4, còn lại là loại III.

Loại I có tỷ lệ nảy mầm 80%, loại II có tỷ lệ nảy mầm 60% và loại III có tỷ lệ nảy mầm 40%. Hỏi tỷ lệ nảy mầm chung của lô hạt giống là bao nhiêu? (Nói cách khác: Ta lấy ngẫu nhiên từ lô ra một hạt. Tìm xác suất để được hạt nảy mầm) (Xem [2]).

Gọi B_i = {hạt giống lấy ra thuộc loại i}, i = 1, 2, 3

A = {hạt giống lấy ra thuộc loại hạt nảy mầm}

Dễ dàng thấy B_1, B_2, B_3 là nhóm đầy đủ các biến cố. A xảy ra thì hạt đó phải thuộc 1 trong 3 loại, tức là một trong ba biến cố B_1, B_2, B_3 phải xảy ra.

Ta có:

$P(B_1) = 2/3$ $\quad P(B_2) = 1/4$ $\quad P(B_3) = 1/12$

$P(A/B_1) = 0,80$ $\quad P(A/B_2) = 0,60$ $\quad P(A/B_3) = 0,40$

Vậy:

$$P(A) = 2/3.0,80 + 1/4.0,60 + 1/12.0,40 = 0,70$$

Ví dụ 21: Trong một trạm cấp cứu bỏng: 80% bệnh nhân bỏng do nóng, 20% bệnh nhân bỏng do hóa chất. Loại bỏng do nóng có 30% bị biến chứng, loại bỏng do hóa chất có 50% bị biến chứng.

a. Từ tập hồ sơ bệnh nhân người ta chọn ngẫu nhiên ra một bệnh án. Tìm xác suất để gặp một bệnh án của bệnh nhân bị biến chứng.

b. Rút ngẫu nhiên được 1 bệnh án của bệnh nhân bị biến chứng.

Tìm xác suất để bệnh án đó là của bệnh nhân bị biến chứng do nóng gây ra? do hóa chất gây ra?

Gọi A = {bệnh án rút ra là của bệnh nhân bị biến chứng}

B = {bệnh nhân bị bỏng do nóng}

C = {bệnh nhân bị bỏng do hóa chất}

Dễ dàng thấy B, C là nhóm đầy đủ biến cố.

Biến cố A xảy ra thì hoặc B hoặc C xảy ra.

Ta có: $P(B) = 0,80$ $\quad P(C) = 0,20$

$P(A/B) = 0,30$ $\quad P(A/C) = 0,50$

a. Theo công thức xác suất đầy đủ ta có:

$$P(A) = 0{,}8.0{,}3 + 0{,}2.0{,}50 = 0{,}34$$

b. Theo công thức Bayes ta có:

$$P(B/A) = \frac{0{,}8 \,.\, 0{,}3}{0{,}34} \approx 0{,}706$$

$$P(C/A) = \frac{0{,}2 \,.\, 0{,}5}{0{,}34} \approx 0{,}294$$

Ví dụ 22: Tỷ lệ người dân nghiện thuốc lá là 30%, biết rằng tỷ lệ người viêm họng trong số người nghiện thuốc lá là 60%. Còn tỷ lệ người viêm họng trong số người không hút thuốc lá là 40%.

a. Chọn ngẫu nhiên một người, biết rằng người đó viêm họng. Tính xác suất người đó nghiện thuốc.

b. Nếu người đó không bị viêm họng, tính xác suất để người đó nghiện thuốc (Xem [3]).

Gọi A = {chọn ra một người bị viêm họng}

B = {người được chọn ra là người nghiện thuốc}

Nhóm đầy đủ ở đây là B và $\overline{B}$

Ta có:

$P(B) = 0{,}30 \qquad P(\overline{B}) = 0{,}70$

$P(A/B) = 0{,}60 \qquad P(A/\overline{B}) = 0{,}40$

Theo (8) ta có:

$$P(A) = 0{,}3.0{,}6 + 0{,}7.0{,}4 = 0{,}46$$

a. $$P(B/A) = \frac{0{,}3 \,.\, 0{,}6}{0{,}46} \approx 0{,}39$$

b. $$P(\overline{A}) = 1 - P(A) = 0{,}54$$

$$P(B/\overline{A}) = \frac{P(B)\,.\,P(\overline{A}/B)}{P(\overline{A})} = \frac{0{,}3 \,.\, 0{,}4}{0{,}54} \approx 0{,}222$$

Nhận xét:

Bạn đọc thường hay lúng túng trong việc chỉ ra một nhóm đầy đủ các biến cố. Xin nêu một vài nhận xét mang tính kinh nghiệm, để hy vọng giúp bạn đọc nhanh chóng tìm ra nhóm đầy đủ biến cố.

Trước hết nhóm đầy đủ không duy nhất. Để tính xác suất của biến cố A có thể dựa vào nhóm đầy đủ này hay nhóm đầy đủ khác, miễn là quan hệ giữa A với nhóm đầy đủ đó phải phù hợp, tức là A xảy ra thì một trong các biến cố của nhóm đầy đủ phải xảy ra.

. *Nếu phép thử gồm 2 giai đoạn, biến cố A liên quan đến giai đoạn sau, thì các kết quả có thể có của giai đoạn đầu chính là một nhóm đầy đủ cần tìm* (chẳng hạn trong bài tập 26: giai đoạn đầu của phép thử là lấy từ mỗi lô ra một sản phẩm, giai đoạn sau là lấy ra một sản phẩm từ hai sản phẩm được lấy ra).

Hoặc nếu hiện tượng đang xét có thể phân chia làm 2 bước, khi đó các kết quả có thể có của bước I chính là nhóm đầy đủ cần tìm (chẳng hạn trường hợp các bệnh nhân bỏng: bước 1 là phân loại bỏng, bước 2 là bị biến chứng; hoặc trường hợp viêm họng: bước 1 là phân loại nghiện và không nghiện, bước 2 là xét sự viêm họng trên mỗi loại của bước 1).

Bạn đọc tự đối chiếu nhận xét trên với các ví dụ cũng như với các bài tập cuối chương, sẽ thấy sáng tỏ vấn đề hơn nhiều.

§ 7. DÃY PHÉP THỬ BERNOULLI

1. Định nghĩa:

Tiến hành n phép thử độc lập (tức là các kết quả của phép thử này không ảnh hưởng gì đến kết quả phép thử kia) được gọi là n phép thử Bernoulli (hoặc một lược đồ Bernoulli) nếu thỏa mãn 2 điều kiện sau:

1. Mỗi phép thử có hai kết quả: A và $\overline{A}$

2. P(A) = p; P(A) như nhau đối với mọi phép thử.

Ví dụ 23:

+ Gieo một đồng tiền 10 lần, đó là 10 phép thử Bernoulli.

+ Một người bắn 5 viên đạn, bắn từng viên một vào một mục tiêu. Đó là 5 phép thử Bernoulli. (Nhưng nếu 5 người bắn, mỗi người bắn một viên thì nói chung đó lại không phải là 5 phép thử Bernoulli).

+ Gieo 1 con xúc xắc 100 lần, A = {xuất hiện mặt lục}. Đó là 100 phép thử Bernoulli.

2. Tần số xuất hiện biến cố A:

Ta tìm xác suất sao cho trong n phép thử Bernoulli biến cố A xuất hiện m lần. Ký hiệu xác suất này là P_n (m, p). Ta có :

$$P_n(m, p) = C_n^m \times p^m \times (1-p)^{n-m} \qquad (10)$$

trong đó m = 0, 1, 2,...,n

Thật vậy các kết quả có thể của n phép thử Bernoulli sẽ là một dãy gồm n chữ A và $\overline{A}$ (ở phép thử thứ i A xuất hiện ta ghi A, $\overline{A}$ xuất hiện ta ghi $\overline{A}$)

$$\underbrace{A\ \overline{A}\ A\ A\ \ldots\ldots\ldots\ \overline{A}\ A}_{\text{n chữ}}$$

Để cho trong n phép thử này biến cố A xuất hiện m lần thì trong dãy có m chữ A, (n−m) chữ $\overline{A}$.

Với 1 cách sắp xếp cố định m chữ A, (n−m) chữ $\overline{A}$, do tính độc lập nên xác suất tương ứng là $p^m \,.\, (1-p)^{n-m}$.

Nhưng ta lại có C_n^m cách xếp m chữ A trong n vị trí.

Vậy xác suất cần tìm là:

$$P_n(m, p) = C_n^m \,.\, p^m \,.\, (1-p)^{n-m}$$

3. Số có khả năng nhất:

Ta gieo một đồng tiền cân đối, đồng chất 5 lần. A = {xuất hiện mặt sấp}. $P(A) = \frac{1}{2}$. Số mặt sấp xuất hiện có thể từ 0 đến 5 tương ứng với các xác suất.

$$P_5(m, \frac{1}{2}) = C_5^m . (\frac{1}{2})^m . (\frac{1}{2})^{5-m} \quad m = 0, 1, 2, ..., 5$$

Trong 6 con số trên sẽ tồn tại số lớn nhất. Số m tương ứng với xác suất lớn nhất sẽ là số hay xảy ra nhất. Trong trường hợp trên m = 2 và 3, tức là trong 5 lần gieo đồng tiền mặt sấp có thể xuất hiện 0 lần, 1 lần, ..., 5 lần, nhưng xuất hiện 2 lần và 3 lần là có khả năng nhất.

Số m_o mà ứng với nó $P_n(m_o, p)$ lớn nhất, được gọi là số có khả năng nhất.

$$P_n(m_o, p) = \max_{0 \le m \le n} P_n(m, p)$$

Quy tắc tìm số có khả năng nhất như sau:

- Nếu $np + p - 1$ là một số nguyên thì m_o chính là $np + p - 1$ và $np + p$.

- Nếu $np + p - 1$ là một số thập phân thì m_o chính là số nguyên bé nhất nhưng lớn hơn $np + p - 1$, tức là

$$m_o = [np + p - 1] + 1 \quad ([x] \text{ là hàm phần nguyên}).$$

Ví dụ 24: Tỷ lệ mắc bệnh basedow ở một vùng nào đó là 10%. Trong đợt khám tuyển nghĩa vụ quân sự người ta đã khám cho 100 người. Tìm xác suất:

a. Trong 100 người có 6 người bị basedow

b. Trong 100 người có 95 người không bị basedow

c. Trong 100 người có ít nhất một người bị basedow

d. Tìm số người bị basedow có khả năng nhất. Tính xác suất tương ứng.

Ở đây ta có 100 phép thử Bernoulli với A = {bị bệnh basedow} và P(A) = p = 0,10. Do đó ta có:

a. $P_{100}(6\ ;\ 0{,}10) = C_{100}^{6} \cdot 0{,}1^{6} \cdot 0{,}9^{94}$.

b. $P_{100}(95\ ;\ 0{,}90) = C_{100}^{95} \cdot 0{,}9^{95} \cdot 0{,}1^{5}$.

c. $P_{100}(m \geqslant 1\ ;\ 0{,}1) = 1 - P_{100}(m = 0\ ;\ 0{,}1)$

$$= 1 - C_{100}^{0} \cdot 0{,}1^{0} \cdot 0{,}9^{100}$$

$$= 1 - 0{,}9^{100}$$

d. $np + p - 1 = 100.0{,}1 + 0{,}1 - 1 = 9{,}1$

Số người bị basedow có khả năng nhất khi khám 100 người là 10 người

$$P_{100}(10\ ;\ 0{,}1) = C_{100}^{10} \cdot 0{,}1^{10} \cdot 0{,}9^{90}.$$

Ví dụ 25: Một lô hạt giống với tỷ lệ hạt lép là 5%. Cần phải lấy một mẫu cỡ bao nhiêu sao cho xác suất để bị ít nhất một hạt lép không bé hơn 0,95.

Ta có:

$$P_n(m \geqslant 1\ ;\ 0{,}05) = 1 - P_n(m = 0\ ;\ 0{,}05)$$

$$= 1 - C_n^0 \cdot 0{,}05^0 \cdot 0{,}95^n$$

$$= 1 - 0{,}95^n$$

Theo đầu bài:

$$1 - 0{,}95^n \geqslant 0{,}95$$

$$0{,}05 \geqslant 0{,}95^n \Rightarrow n \geqslant \frac{\ln 0{,}05}{\ln 0{,}95}$$

BÀI TẬP CHƯƠNG I

1. Một lô hàng gồm 100 sản phẩm, trong đó có 10 phế phẩm. Lấy ngẫu nhiên ra 20 sản phẩm. Tìm xác suất để cho trong 20 sản phẩm lấy ra:

a. Có 5 phế phẩm

b. Bị cả 10 phế phẩm

c. Có đúng 5 chính phẩm

2. Lớp học môn xác suất gồm 70 sinh viên trong đó có 25 nữ. Chọn ngẫu nhiên ra một nhóm gồm 10 sinh viên. Tìm xác suất để trong nhóm chọn ra có 4 sinh viên nữ.

3. Đoàn tàu điện gồm 3 toa tiến vào một sân ga, ở đó đang có 12 hành khách chờ lên tàu. Giả sử các hành khách lên tàu một cách ngẫu nhiên và độc lập với nhau và mỗi toa còn ít nhất 12 chỗ trống. Tìm khả năng xảy ra các tình huống sau:

a. Toa I có 4 người, toa II có 5 người, còn lại là toa III.

b. Mỗi toa có 4 người

c. Hai hành khách A và B cùng lên một toa.

4. Thang máy của một khách sạn 10 tầng xuất phát từ tầng 1 với 5 khách. Coi như mọi người chọn tầng một cách ngẫu nhiên và độc lập. Tìm khả năng xảy ra các tình huống sau:

a. Tất cả cùng ra ở tầng 5

b. Tất cả cùng ra ở một tầng

c. Mỗi người ra ở một tầng khác nhau

d. Hai người cùng ra một tầng, 3 người kia ra 3 tầng khác nhau, tức là 5 người ra 4 tầng khác nhau

5. Một em bé có 5 bìa với các chữ N, N, A, H, H. Tìm xác suất để em bé trong khi sắp ngẫu nhiên thu được chữ NHANH (xem [1]).

6. Gieo đồng thời 2 con xúc xắc cân đối và đồng chất. Tìm xác suất sao cho :

a. Tổng số nốt ở mặt trên hai con xúc xắc bằng 8.

b. Hiệu số nốt ở mặt trên hai con xúc xắc có trị số tuyệt đối bằng 2.

c. Số nốt ở mặt trên hai con xúc xắc bằng nhau (xem [1]).

7. Xét ví dụ 8 khi mọi người ngồi theo hàng ngang.

8. Viết 5 con số 1, 2, 3, 4, 5 lên 5 quả cầu như nhau. Chọn hú hoạ liên tiếp ra 3 quả cầu và xếp theo thứ tự từ trái sang phải. Tìm xác suất để nhận được số chẵn.

9. Rút ngẫu nhiên ra 5 con bài từ bộ bài tam cúc gồm 32 con. Tìm xác suất sao cho trong 5 con bài rút ra có:

a. 1 con tướng, 1 con sỹ, 2 con xe và 1 con tốt.

b. Lập được "tứ tử".

c. Lập được "ngũ tử"

d. Lập được 1 bộ ba "xe – pháo – mã".

e. Có 3 con màu đỏ, 2 con màu đen.

10. Biển đăng ký xe máy loại $50cm^3$ ở Hà Nội gồm 3 phần. Phần đầu là số chỉ vùng Hà Nội: số 29. Phần giữa là 3 chữ số. Phần cuối gồm 2 chữ cái.

a. Tính xem có thể lập được bao nhiêu biển đăng ký xe máy $50cm^3$ ở Hà Nội.

b. Giả sử ta chọn ngẫu nhiên một biển đăng ký. Tìm xác suất để nhận được biển gồm 3 số 468?

c. Tìm xác suất để nhận được một biển có tổng 3 số phần giữa lớn hơn 24.

d. Tìm xác suất nhận được biển có 3 số phần giữa lập thành 1 số chẵn và phần cuối cùng là HK.

11. Một em bé có một hộp chứa 2 bi trắng và 4 bi đỏ. Em rút hú hoạ từng viên bi một cho đến viên cuối cùng. Tìm xác suất để viên bi cuối cùng là bi đỏ (xem [1]).

12. Ở Hạ nghị viện của một quốc gia nào đó có 20 nghị sĩ thuộc Đảng Cộng hòa, có 10 nghị sĩ thuộc Đảng Dân chủ. Cần lập một tiểu ban gồm 5 nghị sĩ. Tìm xác suất để tiểu ban được chọn ngẫu nhiên có:

a. 3 Nghị sĩ thuộc Đảng Cộng hòa và 2 nghị sĩ thuộc Đảng Dân chủ.

b. Cả 5 nghị sĩ thuộc vào một đảng.

13. Một người mua buôn 15 tivi. Anh ta sẽ đồng ý cho xếp lô tivi 15 chiếc này lên xe nếu anh ta kiểm tra ngẫu nhiên 4 chiếc, không có chiếc nào khuyết tật. Vậy xác suất để anh ta chấp nhận lô hàng 15 chiếc này là bao nhiêu nếu trong lô này có 3 chiếc bị khuyết tật.

14. Ta kiểm tra lần lượt 10 sản phẩm. Mỗi sản phẩm thuộc 1 trong 2 loại: Chính phẩm hoặc phế phẩm. Ký hiệu A_k = {sản phẩm kiểm tra thứ k là chính phẩm}, $k = \overline{1,10}$.

Hãy biểu diễn qua A_k các biến cố sau :

a. Cả 10 sản phẩm đều là chính phẩm

b. Có ít nhất một sản phẩm là phế phẩm

c. Các sản phẩm kiểm tra theo thứ tự chẵn là chính phẩm, còn các sản phẩm kiểm tra theo thứ tự lẻ là phế phẩm.

d. Có 1 phế phẩm và 9 chính phẩm

e. Có 2 phế phẩm và 8 chính phẩm (chỉ ra 1 biến cố đại diện và số các biến cố dạng như thế)

15. Ba người cùng bắn vào bia, mỗi người bắn 1 viên. A_i = {Người thứ i bắn trúng bia}. Hãy biểu diễn các biến cố sau qua A_1, A_2, A_3:

a. Chỉ có người thứ nhất bắn trúng

b. Có ít nhất một người bắn trúng

c. Cả ba người cùng bắn trúng

d. Người đầu bắn trúng, người thứ ba bắn trượt

e. Có đúng một người bắn trúng

g. Có đúng hai người bắn trúng

h. Có ít nhất 2 người bắn trúng

i. Không có ai bắn trúng

k. Có không quá 2 người bắn trúng

16. Gieo liên tiếp 1 đồng tiền cho đến khi nào xuất hiện mặt sấp đầu tiên thì dừng lại.

a. Hãy mô tả các biến cố sơ cấp

b. Tính xác suất sao cho phải gieo đến lần thứ ba

17. Một công nhân trong công việc lắp ráp cần sử dụng đến ốc và vít. Anh ta có 2 hộp, 1 hộp ốc với tỷ lệ chính phẩm là 85%, 1 hộp vít với tỷ lệ chính phẩm là 90%.

Chọn ngẫu nhiên một ốc và một vít, tìm khả năng anh ta nhận được 1 bộ ốc vít tốt.

18. Ở một cơ quan nọ có 3 chiếc xe ôtô. Khả năng có sự cố của mỗi ôtô tương ứng là 0,15 ; 0,20 và 0,10.

Tìm khả năng cả 3 ôtô cùng bị hỏng

Tìm khả năng có ít nhất 1 cái hoạt động được.

Tìm khả năng cả 3 ôtô cùng hoạt động được

Tìm xác suất có không quá 2 ôtô bị hỏng

19. Một chi tiết được gia công qua 3 công đoạn nối tiếp với nhau và chất lượng chi tiết chỉ được kiểm tra sau khi đã được gia công xong. Xác suất gây ra khuyết tật cho chi tiết ở các công đoạn tương ứng là 0,2; 0,15; 0,10. Tìm xác suất để sau khi gia công chi tiết có khuyết tật. Bị ít nhất 2 khuyết tật. Bị cả 3 khuyết tật. Không bị khuyết tật nào. Bị không quá một khuyết tật.

20. Tín hiệu thông tin được phát 3 lần với xác suất thu được mỗi lần là 0,4.

a. Tìm xác suất để nguồn thu nhận được thông tin đó

b. Nếu muốn xác suất thu được thông tin lên 0,9 thì phải phát bao nhiêu lần (Xem [3]).

21. Ba người, mỗi người bắn 1 viên vào mục tiêu với xác suất trúng của mỗi người tương ứng là 0,6; 0,8 và 0,7.

Tìm xác suất:

a. Chỉ có anh thứ 2 bắn trúng.

b. Có đúng một người bắn trúng

c. Có ít nhất một người bắn trúng

d. Cả ba người cùng bắn trúng

e. Có đúng 2 người bắn trúng

f. Có ít nhất 2 người bắn trúng

h. Có không quá 2 người bắn trúng

22. Trong điều trị bệnh lao có hiện tượng kháng thuốc. Gọi A là hiện tượng "kháng INH của vi khuẩn lao", B là hiện tượng "kháng PAS của vi khuẩn lao", C là hiện tượng "kháng streptomycin của vi khuẩn lao"

Qua theo dõi ta biết khả năng kháng INH của vi khuẩn lao là 20%, nghĩa là P(A) = 0,20. Tương tự P(B) = 0,4, P(C) = 0,30. Việc kháng các loại thuốc khác nhau là độc lập với nhau.

Nếu phối hợp cả 3 loại thuốc thì khả năng khỏi bệnh là bao nhiêu.

23. Một lô thỏ gồm 3/4 thỏ có gen dị hợp tử Xt, 1/4 thỏ có gen dị hợp tử Xx, X là gen màu xám (gen trội), t là gen màu trắng (gen lặn). Bắt ngẫu nhiên từng con một ra 2 thỏ.

a. Tìm xác suất để 2 thỏ cùng gen

b. Giả sử 2 thỏ bắt được có 1 thỏ đực và 1 thỏ cái. Cặp thỏ này sinh được 4 con thỏ xám.

Tìm xác suất để cặp bố mẹ cùng gen Xt, cùng gen XX.

24. Một trận không chiến giữa máy bay ta và địch. Máy bay ta bắn trước với xác suất trúng là 0,5. Nếu bị trượt, máy bay địch bắn trả lại với xác suất trúng là 0,4. Nếu không bị trúng đạn, máy bay ta lại bắn trả với xác suất trúng 0,3. Tìm xác suất:

a. Để máy bay địch bị rơi trong cuộc không chiến trên.

b. Để máy bay ta bị rơi trong cuộc không chiến trên.

25. Ta có 10 hộp bi trong đó 4 hộp loại I, mỗi hộp có 3 bi trắng và 5 bi đỏ; 3 hộp loại II, mỗi hộp có 4 bi trắng, 6 bi đỏ; 3 hộp loại III, mỗi hộp 2 bi trắng, 5 bi đỏ.

a. Rút hú hoạ một hộp và từ đó lấy ngẫu nhiên một bi. Tìm xác suất để được bi đỏ.

b. Rút hú hoạ một hộp và từ đó lấy ngẫu nhiên một bi thì được bi trắng. Tìm xác suất viên bi đó rút ra từ hộp loại II (xem [1]).

26. Có hai lô sản phẩm. Lô 1 có 10 sản phẩm loại I và 2 sản phẩm loại II. Lô 2 có 16 sản phẩm loại I và 4 sản phẩm loại II. Từ mỗi lô ta lấy ngẫu nhiên ra một sản phẩm. Sau đó trong 2 sản phẩm thu được ta lại lấy hú hoạ ra một sản phẩm.

Tìm xác suất để sản phẩm lấy ra sau cùng là sản phẩm loại I.

27. Có hai lô gà giống. Lô I gồm 15 con trong đó có 3 con trống. Lô II gồm 20 con trong đó có 4 con trống. Một con từ lô II nhảy sang lô I. Từ lô I ta bắt ngẫu nhiên ra một con. Tìm xác suất để con gà bắt ra là gà trống.

28. Ta biết rằng các trẻ sinh đôi có thể là sinh đôi thật (do 1 trứng sinh ra), trong trường hợp đó chúng cùng giới hoặc giả sinh đôi (do 2 trứng sinh ra), trong trường hợp đó xác suất để chúng cùng giới là 1/2. Ta giả thiết rằng đã biết xác suất p sao cho trong một họ đã cho hai trẻ sinh đôi là sinh đôi thật.

a. Tìm xác suất để cho hai trẻ sinh đôi là sinh đôi thật biết rằng chúng cùng giới.

b. Tìm xác suất để cho hai trẻ sinh đôi là giả sinh đôi biết rằng chúng khác giới (Xem [4]).

29. Biết rằng tỷ lệ người mắc bệnh nào đó ở địa phương nào đó là 2%. Người ta sử dụng một phản ứng mà nếu người bị bệnh thì phản ứng luôn luôn dương tính, nếu không bị bệnh thì phản ứng có thể dương tính với xác suất 0,20.

a. Tìm xác suất phản ứng dương tính

b. Tìm xác suất bị bệnh, không bị bệnh trong nhóm người có phản ứng dương tính

c. Qua phương pháp thử này ta có thể ước lượng tỷ lệ mắc bệnh là bao nhiêu.

30. Một người có 3 chỗ ưa thích như nhau để câu cá. Xác suất câu được cá ở những chỗ đó tương ứng là 0,6 ; 0,7 và 0,8. Biết rằng ở mỗi một chỗ người đó đã thả câu 3 lần và chỉ câu được 1 con cá. Tìm xác suất để cá câu được ở chỗ thứ nhất. (Xem [3]).

31. Theo kết quả điều tra về bệnh lao, tỷ lệ người bị lao ở vùng nọ là 0,001. Tìm xác suất để khi khám cho 10 người:

a. Không ai bị lao

b. 5 người bị lao

c. Ít nhất 1 người bị lao

d. Số người không bị lao có khả năng nhất

32. Trong một cuộc thi bắn quốc tế, mỗi xạ thủ bắn 60 viên vào bia. Xạ thủ của Việt Nam bắn trúng tâm với xác suất 0,92. Tìm xác suất để:

a. Xạ thủ này bắn trúng tâm cả 60 viên

b. Xạ thủ này bị trượt ngoài tâm 2 viên

c. Xạ thủ này bị trượt ngoài tâm ít nhất 1 viên

d. Tìm số viên trúng tâm có khả năng nhất. Tính xác suất tương ứng.

33. Một bác sỹ có tiếng về chữa 1 bệnh nào đó. Xác suất chữa khỏi bệnh là 8/10. Có người nói rằng cứ 10 người đến chữa thì chắc chắn có 8 người khỏi. Điều khẳng định đó có đúng không?

Tìm xác suất sao cho bác sỹ đó chữa cho 10 người thì có 8 người khỏi.

34. Một chiến sỹ tự vệ tập bắn súng, xác suất để chiến sỹ này bắn trúng tâm là 0,3. Hỏi chiến sỹ này phải bắn ít nhất bao nhiêu viên để với xác suất không bé hơn 0,80 chiến sỹ này bắn trúng tâm ít nhất 1 viên.

35. Hai đấu thủ chơi cờ ngang tài ngang sức thi đấu với nhau. Hỏi rằng khả năng nào cao hơn giữa hai khả năng:

- Thắng 2 ván trong 4 ván
- Thắng 3 ván trong 6 ván

36. Một lô hàng có tỷ lệ chính phẩm là 95%. Lấy liên tiếp ra 2 sản phẩm. Tìm xác suất để nhận được

a. Cả 2 chính phẩm

b. Ít nhất 1 chính phẩm

c. Chỉ có cái thứ 2 là chính phẩm

d. Có đúng 1 chính phẩm

37. Có hai lô hàng cũ. Lô I có 10 cái tốt, 2 cái hỏng. Lô II có 12 cái tốt, 3 cái hỏng. Từ mỗi lô lấy ngẫu nhiên ra một cái. Tìm xác suất để :

a. Nhận được 2 cái tốt

b. Nhận được 2 cái cùng chất lượng

c. Nếu lấy từ cùng 1 lô ra 2 cái thì nên lấy từ lô nào để được 2 cái tốt với khả năng cao hơn.

38. Tỷ lệ cha mắt đen, con mắt đen là 0,05; cha mắt đen, con mắt xanh là 0,079. Cha mắt xanh, con mắt đen là 0,089; cha mắt xanh, con mắt xanh là 0,782.

a. Tìm khả năng con mắt xanh biết rằng cha mắt xanh?

b. Tìm khả năng con mắt không đen biết rằng cha mắt đen?

39. Trẻ sinh đôi cùng giới gấp đôi trẻ sinh đôi khác giới. Xác suất sinh đôi khác giới trong các trường hợp là như nhau. Tìm xác suất để đứa thứ 2 là trai với điều kiện đứa thứ nhất trong cặp sinh đôi là trai. Biết rằng khả năng sinh con trai trong mỗi lần sinh là 0,49.

40. Trên một bảng quảng cáo người ta mắc 2 hệ thống bóng đèn. Hệ thống I gồm 2 bóng mắc nối tiếp. Hệ thống II gồm 2 bóng mắc song song. Khả năng bị hỏng của mỗi bóng đèn sau 6 giờ thắp sáng liên tục là 15% và việc hỏng của các bóng coi như độc lập. Tìm xác suất.

a. Hệ thống I bị hỏng

b. Hệ thống II không bị hỏng

c. Cả hai hệ thống bị hỏng

d. Chỉ có hệ thống I bị hỏng

41. Trong một làng tỷ lệ nam / nữ là 12 :13. Khả năng mắc bệnh bạch tạng ở nam là 0,6%, ở nữ là 0,35%.

a. Khả năng gặp một người trong làng bị mắc bệnh bạch tạng là bao nhiêu.

b. Gặp trong làng một người không mắc bệnh. Khả năng người gặp đó là nữ cao hơn hay là nam cao hơn.

42. Có một bệnh nhân mà bác sỹ chẩn đoán mắc bệnh A với xác suất 70%, mắc bệnh B với xác suất 30%. Để có thêm thông tin chẩn đoán bác sỹ đã cho xét nghiệm sinh hóa. Sau 3 lần thử thấy có 1 lần dương tính, biết rằng khả năng dương tính của mỗi lần xét nghiệm đối với bệnh A và B tương ứng là 10% và 30%. Hãy cho biết nên chẩn đoán bệnh nhân mắc bệnh nào?

43. Có 2 lô sản phẩm. Lô I gồm 90% chính phẩm. Lô II có tỷ lệ phế phẩm/chính phẩm là 1/4. Lấy ngẫu nhiên ra một lô, rồi từ đó lấy ngẫu nhiên một sản phẩm ta được chính phẩm. Trả sản phẩm này trở lại lô của nó. Từ lô này ta lại lấy ra một sản phẩm. Tìm xác suất để lấy phải phế phẩm.

Chương II

ĐẠI LƯỢNG NGẪU NHIÊN VÀ HÀM PHÂN PHỐI

§ 1. ĐẠI LƯỢNG NGẪU NHIÊN

– Một đại lượng (hay một biến) nhận các giá trị của nó với xác suất tương ứng nào đấy gọi là đại lượng ngẫu nhiên hay biến ngẫu nhiên.

Ta thường ký hiệu các biến ngẫu nhiên bởi các chữ X, Y, Z... hoặc ξ, η, ζ,... (đọc là "si", "ê–ta", "giê–ta"). Các giá trị mà biến ngẫu nhiên nhận thường viết bằng chữ nhỏ: x, y, z...

– Phân loại các đại lượng ngẫu nhiên. Căn cứ vào giá trị mà biến ngẫu nhiên nhận ta phân các đại lượng ngẫu nhiên ra làm hai loại chính: biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục.

1. Biến ngẫu nhiên rời rạc:

Nếu tập các giá trị mà biến ngẫu nhiên nhận là một tập gồm một số hữu hạn điểm hoặc vô hạn nhưng đếm được, khi đó biến ngẫu nhiên gọi là biến ngẫu nhiên rời rạc.

Giả sử biến ngẫu nhiên ξ nhận các giá trị x_1, x_2,... x_n... và $P\{\xi = x_i\} = p_i$, $i = 1, 2,...$

Để mô tả (hoặc xác định) biến ngẫu nhiên rời rạc ξ ta dùng bảng sau:

ξ	x_1	x_2	...	x_n	...
$P\{\xi = x_i\}$	p_1	p_2	...	p_n	...

Trong đó: $\sum_i p_i = 1, \quad p_i > 0 \ \forall i = 1, 2, \ldots$

Bảng với hai thông tin cho trên xác định biến ngẫu nhiên ξ được gọi là bảng phân phối xác suất.

Ví dụ 1: Gieo đồng thời 2 đồng tiền cân đối và đồng chất. Gọi ξ là biến ngẫu nhiên chỉ số lần xuất hiện mặt sấp. Ta có bảng phân phối xác suất sau:

ξ	0	1	2
$P\{\xi = x_i\}$	1/4	2/4	1/4

Gọi A_i = {đồng tiền thứ i xuất hiện mặt sấp}, i = 1, 2...
A_1, A_2 độc lập

Biến cố $\{\xi = 0\} = \overline{A}_1 \overline{A}_2$

$$\rightarrow P\{\xi = 0\} = P(\overline{A}_1)\, P(\overline{A}_2) = 1/2 \times 1/2 = 1/4$$

$$\{\xi = 1\} = A_1 \overline{A}_2 \cup \overline{A}_1 A_2$$

$$\rightarrow P\{\xi = 1\} = P(A_1)\, P(\overline{A}_2) + P(\overline{A}_1)\, P(A_2)$$

$$= 1/2 \times 1/2 + 1/2 \times 1/2 = 2/4$$

$$\{\xi = 2\} = A_1 A_2$$

$$\rightarrow P\{\xi = 2\} = 1/4$$

Trong trường hợp các giá trị x_i, p_i có tính quy luật, thay cho việc lập bảng trên ta có thể mô tả bởi đẳng thức dạng sau:

$$P\{\xi = x_i\} = p_i\,, i = 1, 2, \ldots$$

Ví dụ 2: Gieo một con xúc xắc cân đối đồng chất 100 lần. Gọi X là biến ngẫu nhiên chỉ số lần xuất hiện mặt lục trong 100 lần gieo trên. Khi đó theo (10) (chương I) ta có phân phối xác suất của X là:

$$P\{X = m\} = C_{100}^{m} \times (1/6)^{m} \times (5/6)^{100-m}; \ m = 0, 1, 2, \ldots, 100$$

trong đó p = P{xuất hiện mặt lục} = 1/6

Ví dụ 3: Một xạ thủ đem theo 5 viên đạn đến trường bắn để chỉnh súng trước ngày thi đấu. Anh ta bắn từng viên một vào bia với xác suất trúng tâm là 0,9. Anh ta thử súng theo quy tắc sau:

a. Nếu có 3 viên liên tiếp trúng tâm thì thôi không bắn nữa.

b. Nếu có 3 viên trúng tâm thì thôi không bắn nữa.

Gọi X và Y là số đạn mà anh ta đã dùng để thử súng tương ứng theo 2 nguyên tắc trên.

Lập bảng phân phối xác suất của X và Y.

X	3	4	5
P{X = k}	$0{,}9^3$ (= 0,729)	$0{,}9^3 \times 0{,}1$ (= 0,0729)	1 - (0,729 + 0,0729) (= 0,1981)

Chẳng hạn $\{X = 4\} = \overline{T}\ T\ T\ T$ (T : trúng, $\overline{T}$: trượt).

Song không ít bạn đọc sẽ xác định như sau :

$$\{X = 5\} = \overline{T}\ \overline{T}\ T\ T\ T \cup T\ \overline{T}\ T\ T\ T$$

Khi đó P(X = 5) = 0,0729 < 0,1981, do đó tổng các xác suất vẫn nhỏ hơn 1.

Sở dĩ như vậy là vì biến cố {X = 5} bạn đọc mới chỉ tính đến 2 trường hợp dừng lại theo đúng quy tắc đặt ra, chứ chưa kể đến các trường hợp không thỏa mãn quy tắc nhưng vì hết đạn nên không thể bắn tiếp được. Đó là các trường hợp bắn hết 5 viên nhưng không trúng viên nào hoặc chỉ trúng 1 viên, 2 viên hoặc 3 viên nhưng không liên tiếp, trúng 4 viên nhưng không liên tiếp.

Trong trường hợp này làm như trên là nhanh nhất

Y	3	4	5
P{Y = k}	$0{,}9^3$ (= 0,729)	$C_3^1 \times 0{,}9^3 \times 0{,}1$ (= 0,2187)	$1 - 0{,}9^3 - 0{,}3 \times 0{,}9^3$ (= 0,0523)

Ví dụ 4: Cho 2 biến ngẫu nhiên X, Y độc lập với bảng phân phối xác suất như sau:

X	0	1	2
$P\{X = i\}$	0,3	0,4	0,3

Y	-1	1
$P\{Y = j\}$	0,4	0,6

Hãy lập bảng phân phối xác suất của X^2, $X+Y$, XY

X và Y gọi là độc lập với nhau nếu mọi biến cố liên quan đến X độc lập với biến cố bất kỳ liên quan đến Y.

Ta có $P\{X^2 = k^2\} = P(X = k)$, tức là khả năng X nhận giá trị k cũng chính là khả năng X^2 nhận giá trị k^2.

X^2	0	1	4
$P\{X^2 = i^2\}$	0,3	0,4	0,3

$X + Y$	-1	0	1	2	3
$P\{X + Y = k\}$	0,12	0,16	0,30	0,24	0,18

Do X, Y độc lập nên

$$P(X + Y = -1) = P(X = 0) \times P(Y = -1) = 0,12$$

$$(X + Y = 0) = (X = 1) \cap (Y = -1)$$

$$(X + Y = 1) = (X = 0) \cap (Y = 1) \cup (X=2) \cap (Y = -1)$$

$$(X + Y = 2) = (X = 1) \cap (Y = 1)$$

$$(X + Y = 3) = (X = 2) \cap (Y = 1)$$

XY	-2	-1	0	1	2
$P\{XY = k\}$	0,12	0,16	0,3	0,24	0,18

$$(XY = -2) = (X = +2) \cdot (Y = -1)$$

$$(XY = -1) = (X = 1)\ (Y = -1)$$

$$(XY = 0) = (X = 0)\ [(Y = -1) \cup (Y = 1)]$$

$$= (X = 0).\ \Omega = (X = 0)$$

$$(XY = 1) = (X = 1)\ (Y = 1)$$

$$(XY = 2) = (X = 2)\ (Y = 1)$$

Ví dụ 5: Biến ngẫu nhiên ξ có phân phối xác suất

$$P\{\xi = k\} = e^{-\lambda}\frac{\lambda^k}{k!}, \quad (\lambda > 0)$$

$k = 0, 1, 2, \ldots$ được gọi là biến ngẫu nhiên Poisson.

2. Biến ngẫu nhiên liên tục:

Nếu tập các giá trị biến ngẫu nhiên nhận lấp đầy một khoảng nào đó, khi đó biến ngẫu nhiên được gọi là biến ngẫu nhiên liên tục.

- Để mô tả (hoặc xác định) biến ngẫu nhiên liên tục ta dùng khái niệm hàm mật độ.

Hàm p(x) được gọi là hàm mật độ của biến ngẫu nhiên nào đấy nếu thỏa mãn hai điều kiện sau:

1. $p(x) \geq 0 \quad \forall x$ thuộc $(-\infty, +\infty)$

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

Trong trường hợp này, xác suất để ξ thuộc vào khoảng (x_0, x_1) được tính như sau:

$$P\{x_o < \xi < x_1\} = \int_{x_o}^{x_1} p(x)dx$$

Ví dụ 6:

$$p(x) = \begin{cases} 0 & \text{nếu } x < a \text{ hoặc } x > b \\ \dfrac{1}{b-a} & \text{nếu } a \leq x \leq b \end{cases}$$

p(x) là hàm mật độ của biến ngẫu nhiên nhận mọi giá trị trên [a, b] với khả năng đều như nhau, gọi tắt là mật độ đều trên [a, b].

Ví dụ 7:

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

p(x) là hàm mật độ của biến ngẫu nhiên nhận giá trị trên toàn trục số. Hàm mật độ này được gọi là mật độ chuẩn.

Ví dụ 8: Cho hàm $p(x) = a\sin 2x$. Xác định hằng số a để p(x) trở thành hàm mật độ của biến ngẫu nhiên nhận giá trị tập trung trong khoảng $[0, \pi/2]$

Như vậy

$$p(x) = \begin{cases} 0 & \text{nếu } x < 0 \text{ hoặc } x > \pi/2 \\ a\sin 2x & \text{nếu } 0 \leqslant x \leqslant \pi/2 \end{cases}$$

Trong $[0,\pi/2]$ thì $\sin 2x \geqslant 0$ nên $a \geqslant 0$

Ta có:

$$\int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{\pi/2} a\sin 2xdx + \int_{\pi/2}^{+\infty} 0dx$$

$$= a\int_{0}^{\pi/2} a\sin 2xdx = \frac{-a}{2}\cos 2x\Big|_{0}^{\pi/2} = \frac{-a}{2}(-2)$$

$$= a = 1$$

Vậy $a = 1$ và $p(x) = \sin 2x$ là hàm mật độ của biến ngẫu nhiên nhận giá trị tập trung trong $[0,\pi/2]$.

§ 2. HÀM PHÂN PHỐI

1. Định nghĩa:

Cho biến ngẫu nhiên ξ, ta xác định hàm phân phối của ξ như sau:

$$F_\xi(x) = P\{\xi < x\} \tag{1}$$

Trong định nghĩa t ên x là biến của hàm F, x nhận giá trị thực, x thuộc $(-\infty, +\infty)$. Tại một điểm x bất kỳ hàm F(x) chính là xác suất để biến ngẫu nhiên nhận giá trị nhỏ hơn x hoặc để biến ngẫu nhiên nhận giá trị bên trái x.

Chỉ số của hàm $F_\xi(x)$ để chỉ hàm phân phối của biến ngẫu nhiên ξ. Trường hợp không cần thiết có thể bỏ qua không cần viết chỉ số đó.

2. Tính chất:

Hàm phân phối của biến ngẫu nhiên có một số tính chất cơ bản sau:

1^o. Hàm phân phối xác định $\forall x \in (-\infty, +\infty)$

2^o. $0 \leqslant F(x) \leqslant 1, \forall x$; $F(-\infty) = 0, F(+\infty) = 1$

3^o. Hàm phân phối là hàm không giảm: Nếu $x_1 < x_2$ thì $F(x_1) \leqslant F(x_2)$

4^o. $P\{a \leqslant \xi < b\} = F(b) - F(a)$ (2)

Ngoài ra còn một số tính chất khác nữa, nhưng ở đây ta bỏ qua.

Chứng minh: Tính chất 1^o và 2^o trực tiếp suy từ định nghĩa (1).

Để chứng minh chặt chẽ $F(-\infty) = 0$, $F(+\infty) = 1$, ở giáo trình này ta không thể đề cập đến được. Bỏ qua tính chặt chẽ của toán học ta có thể hình dung biến cố $\{\xi < -\infty\}$ và $\{\xi < +\infty\}$ tương đương với biến cố $\varnothing$ và Ω.

Chứng minh tính chất 3^o và 4^o ta có:

Nếu $x_1 < x_2$, biến cố $\{\xi < x_2\} = \{\xi < x_1\} \cup \{x_1 \leqslant \xi < x_2\}$

Do xung khắc nên $P\{\xi < x_2\} = P\{\xi < x_1\} + P\{x_1 \leqslant \xi < x_2\}$

$$F(x_2) = F(x_1) + P\{x_1 \leqslant \xi < x_2\}$$

Từ đẳng thức cuối ta suy ra 3^o và 4^o.

Bây giờ ta quay trở lại định nghĩa $F_\xi(x)$

Nếu ξ rời rạc $P\{\xi = x_i\} = p_i$

$$F_\xi(x) = \sum_{i:\, x_i < x} p_i \qquad (3)$$

Nếu ξ liên tục với hàm mật độ p(x)

$$F_\xi(x) = \int_{-\infty}^{x} p(t)dt \qquad (4)$$

Nhận xét:

- Nếu hàm mật độ liên tục tại x thì tại đó ta có $F'(x) = p(x)$

- Nếu hàm phân phối của biến ngẫu nhiên ξ liên tục tại x_o thì $P\{\xi = x_o\} = 0$

Thật vậy:

$$P\{\xi = x_o\} = \lim_{\Delta x \to 0} P\{x_o \leqslant \xi < x_o + \Delta x\}$$

$$= \lim_{\Delta x \to 0} \left[F(x_o + \Delta x) - F(x_o)\right] = F(x_o) - F(x_o) = 0$$

Từ đó suy ra nếu biến ngẫu nhiên ξ là biến ngẫu nhiên liên tục với điều kiện các giá trị của ξ lấp đầy 1 khoảng liên tục, thì xác suất để ξ nhận giá trị tại một điểm nào đó là bằng 0.

– Để cho tiện, từ nay ta dùng ký hiệu $\xi \simeq F(x)$ nghĩa là biến ngẫu nhiên ξ có hàm phân phối là F(x).

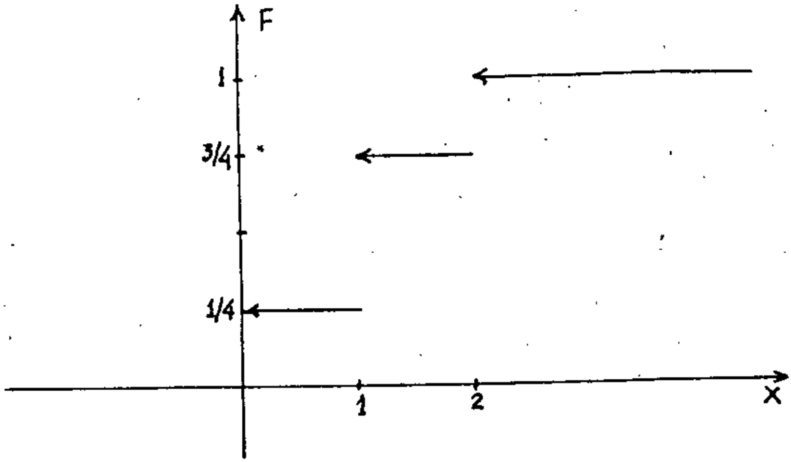
3. Các ví dụ:

Tìm hàm phân phối của các biến ngẫu nhiên trong các ví dụ 1 - 8 ở trên.

– Đối với biến ngẫu nhiên ξ trong ví dụ 1 ta có hàm phân phối như sau:

$$F_\xi(x) = \begin{cases} 0 & \text{nếu } x \leqslant 0 \\ 1/4 & \text{nếu } 0 < x \leqslant 1 \\ 1/4 + 2/4 = 3/4 & \text{nếu } 1 < x \leqslant 2 \\ 1/4 + 2/4 + 1/4 = 1 & \text{nếu } 2 < x \end{cases}$$

Đồ thị của $F_\xi(x)$



Tính $P\{0 < \xi \leqslant 2\}$ và $P\{1 \leqslant \xi < 5\}$

Để tính $P\{0 < \xi \leqslant 2\}$ ta có thể tính trực tiếp:

$$P\{0 < \xi \leqslant 2\} = P\{(\xi = 1) \cup (\xi = 2)\}$$
$$= P\{\xi = 1\} + P\{\xi = 2\} = 2/4 + 1/4 = 3/4$$

Hoặc ta có thể tính thông qua hàm phân phối bằng cách áp dụng (2):

$$P\{0 < \xi \leqslant 2\} = P\{0 \leqslant \xi < 2\} - P(\xi = 0) + P(\xi = 2)$$
$$= F(2) - F(0) - P(\xi = 0) + P(\xi = 2) = 3/4 - 0 - 1/4 + 1/4 = 3/4$$

Tương tự ta tính

$$P\{1 \leqslant \xi \leqslant 5\} = P(\xi = 1) + P(\xi = 2) = 2/4 + 1/4 = 3/4$$

Hoặc

$$P\{1 \leqslant \xi \leqslant 5\} = P\{1 \leqslant \xi < 5\} + P\{\xi = 5\}$$
$$= F(5) - F(1) + P\{\xi = 5\} = 1 - 1/4 + 0 = 3/4$$

– Đối với biến ngẫu nhiên Y trong ví dụ 3 ta có:

$$F_Y(x) = \begin{cases} 0 & \text{nếu } x \le 3 \\ 0{,}729 & \text{nếu } 3 < x \le 4 \\ 0{,}9477 & \text{nếu } 4 < x \le 5 \\ 1 & \text{nếu } 5 < x \end{cases}$$

- Đối với biến ngẫu nhiên trong ví dụ 5

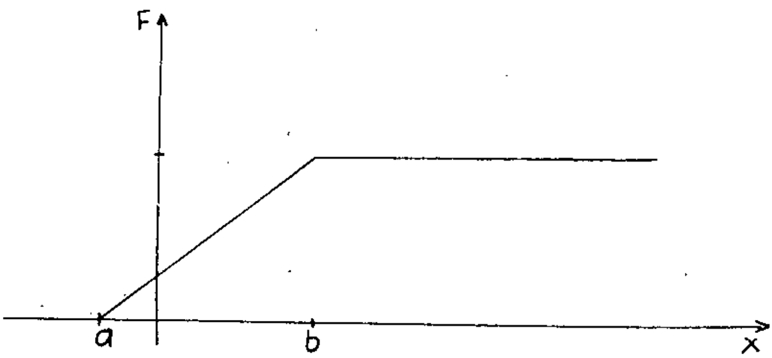
$$F(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ e^{-\lambda} \sum_{s=0}^{n} \frac{\lambda^s}{s!} & \text{nếu } n < x \le n+1 \end{cases}$$

Ta ký hiệu phân phối Poisson với tham số λ là $P_\lambda . \xi \doteq P_\lambda$ tức là biến ngẫu nhiên ξ có phân phối Poisson.

- Đối với ví dụ 6: Biến ngẫu nhiên X có phân phối đều trên [a,b]

$$F(x) = \begin{cases} 0 & \text{nếu } x \le a \\ \int_{-\infty}^{x} p(t)dt & \text{nếu } a < x \le b \\ \int_{-\infty}^{a} 0dt + \int_{a}^{b} \frac{dt}{b-a} + \int_{b}^{x} 0dt = \int_{a}^{b} \frac{dt}{b-a} = 1 & \text{nếu } a < x \le b \end{cases}$$

$$= \begin{cases} 0 & \text{nếu } x \le 3 \\ \frac{x-a}{b-a} & \text{nếu } a < x \le b \\ 1 & \text{nếu } x > b \end{cases}$$



Nếu X có phân phối đều trên [0, 1], ta có:

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

– Đối với hàm mật độ ở ví dụ 7, ta có

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

Biến ngẫu nhiên ứng với hàm mật độ ở ví dụ 7 hoặc hàm phân phối trên được gọi là biến ngẫu nhiên có phân phối chuẩn. Ta ký hiệu phân phối chuẩn là $N(a, \sigma^2)$.

4. Một số phân phối một chiều quen thuộc và ứng dụng thực tế:

a. Phân phối nhị thức:

Xét n phép thử Bernoulli với xác suất thành công P(A) = p. Gọi ξ là số lần xuất hiện biến cố A trong n phép thử trên. Phân phối của ξ được gọi là phân phối nhị thức và ký hiệu $\xi = B(n, p)$.

$$P(\xi = m) = C_n^m p^m (1-p)^{n-m} \qquad m = 0, 1, 2, \ldots, n$$

Dãy phép thử Bernoulli thường gặp nhiều trong thực tế do đó biến ngẫu nhiên có phân phối nhị thức cũng thường gặp trong các ứng dụng.

b. Phân phối Poisson

Phân phối này đã được giới thiệu ở các ví dụ trước. Người đầu tiên mô tả phân phối này là Simeon Denis Poisson vào năm 1837. Phân phối này đã có nhiều ứng dụng đối với nhiều quá trình có liên quan đến số quan sát đối với một đơn vị thời gian hoặc không gian. Chẳng hạn số cuộc điện thoại nhận được ở một trạm điện thoại trong một phút, số khách hàng đến nhà băng đối với mỗi một chu kỳ 30 phút. Số máy bị hỏng trong một ngày... Nói chung là dòng vào của một hệ phục vụ (quán bia, hiệu cắt tóc, hiệu chữa xe, trạm điện thoại, một cửa hàng nào đó...) là các biến ngẫu nhiên tuân theo luật Poisson. Hoặc giả sử có n điểm phân phối đều trên [a, b]. Khi đó số các điểm rơi vào một đoạn có độ dài đơn vị là biến ngẫu nhiên Poisson v.v...

c. Phân phối đều

Hàm mật độ và hàm phân phối đã được đưa ra ở trên. Từ biến ngẫu nhiên phân phối đều người ta nhận được bảng các số ngẫu nhiên.

d. Phân phối chuẩn $N(\mu, \sigma^2)$

Hàm mật độ chuẩn tổng quát $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$ với $-\infty < x < +\infty$.

Đường cong mật độ này đối xứng qua đường $x = \mu$, nhận trục Ox làm tiệm cận ngang và có giá trị cực đại tại $x = \mu$ với tung độ cực đại là $\frac{1}{\sigma\sqrt{2\pi}}$

Trường hợp đặc biệt: $\xi \approx N(0, 1)$. Khi đó hàm mật độ được ký hiệu là $\varphi(x)$:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \text{với } -\infty < x < +\infty$$

là hàm đối xứng qua trục tung, đồ thị có dạng hình chuông. Hàm phân phối N(0, 1) được ký hiệu $\Phi(x)$ (xem bảng I - Phụ lục).

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Phân phối chuẩn chiếm vị trí quan trọng trong lý thuyết xác suất, là vị trí trung tâm trong các kết luận thống kê sau này.

Trong thực tế nhiều biến ngẫu nhiên, nhiều quy luật tuân theo luật chuẩn hoặc gần chuẩn, chẳng hạn trọng lượng và chiều cao của người lớn, mức độ thông minh của trẻ em, điểm thi của các thí sinh, lực chịu đựng của một thanh sắt, các sai số đo đạc, sai số quan sát, độ bền dẻo của máy móc, trung bình cộng của một số lớn các đại lượng ngẫu nhiên độc lập...

Trong buôn bán, kinh tế và trong khoa học xã hội, nhiều phân phối không giống phân phối chuẩn, nhưng phân phối của trung bình cộng đối với mỗi trường hợp lại có thể xem như phân phối chuẩn miễn là n lớn v.v...

Trong nghiên cứu địa chất, phân phối chuẩn cũng được ứng dụng để mô tả nhiều hiện tượng địa chất, chẳng hạn như hàm lượng và khoáng vật trong đá, hàm lượng nước trong đá trầm tích, hàm lượng một số nguyên tố hóa học, độ lỗ hổng của cát kết, mực nước trong lỗ khoan... Bên cạnh đó phân phối loga chuẩn cũng đóng vai trò rất quan trọng trong nghiên cứu địa chất. Một số nhà địa chất, chẳng hạn Ahrens, đã đề nghị coi phân phối loga chuẩn như một quy luật cơ bản của địa hóa học. Biến ngẫu nhiên X gọi là có phân phối loga chuẩn nếu $\lg_a X$ tuân theo luật phân phối chuẩn. Chẳng hạn là đường kính của hạt trầm tích, hoặc X là bề dày các lớp trầm tích, hoặc hàm lượng các nguyên tố hiếm trong đá, hệ số thẩm thấu của đá trầm tích v.v... (xem [5]).

Do vai trò và vị trí đặc biệt của phân phối chuẩn cho nên người ta đã lập bảng tính giá trị của hàm mật độ $\varphi(x)$ và giá trị của hàm phân phối $\Phi(x)$.

Tuy về lý thuyết biến ngẫu nhiên chuẩn N(0, 1) nhận giá trị trên toàn đường thẳng, song trên thực tế biến ngẫu nhiên N(0, 1) nhận giá trị trong khoảng (–3; 3) với xác suất 0,9973, nhận giá trị trong khoảng (–3,5; 3,5) với xác suất 0,9996. Thêm vào đó hàm $\varphi(x)$ là hàm đối xứng, cho nên chúng ta chỉ cần 1 trang sách để lập bảng tính hàm $\varphi(x)$ và $\Phi(x)$ với sự sai khác của các giá trị x là 0,01.

Mặt khác nếu $X = N(\mu, \sigma^2)$ thì ta có thể đưa về chuẩn $N(0, 1)$ bằng phép biến đổi sau: $Y = \dfrac{X - \mu}{\sigma}$

Phép biến đổi này gọi là phép chuẩn hóa biến ngẫu nhiên. Trong tiết sau ta sẽ thấy được rằng biến ngẫu nhiên Y có phân phối chuẩn N(0, 1). Ta có:

$$P\{a \leq X < b\} = P\left\{\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right\} = \text{(theo (2))}$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad (5)$$

e. Phân phối mũ

Biến ngẫu nhiên ξ có phân phối mũ nếu hàm mật độ của nó được xác định bởi

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

Hàm phân phối có dạng:

$$F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 1 - e^{-\lambda x} & \text{nếu } x > 0 \end{cases}$$

Trong thực tế, nhiều đại lượng ngẫu nhiên phù hợp với phân phối mũ, chẳng hạn thời gian phục vụ của hệ phục vụ đám đông là các biến ngẫu nhiên có phân phối mũ (thời gian nói chuyện của các cuộc đàm thoại, thời gian một khách hàng cần phục vụ ở quán bia, ở hiệu uốn tóc...). Thời gian sống của một bóng đèn điện, thời gian phục vụ của một dụng cụ điện tử, v.v... cũng tuân theo luật mũ.

Phân phối mũ là trường hợp riêng của họ phân phối Weibull (xem bài tập 10 cuối chương), nằm trong họ các phân phối giảm, mà họ các phân phối này có nhiều ứng dụng trong lâm nghiệp. Chẳng hạn phân phối số cây theo kích thước đường kính là phân phối giảm cho tất cả loài cây thuộc một lâm phần hoặc cho từng loại cây. Số cây thông nhựa vào thời kỳ còn non tính theo đường kính của nó khá phù hợp với phân phối Weibull. Trái lại phân phối của số cây theo đường kính ở những rừng tư nhiên khác tuổi nhiều tầng thường tuân theo những phân phối giảm (xem [2]) v.v...

Ngoài các phân phối trên, bây giờ ta làm quen với phân phối χ^2 (đọc là *khi bình phương*), phân phối Student và phân phối F. Ba dạng phân phối này ít được dùng để mô tả các biến ngẫu nhiên trong thực tế, song chúng lại có vai trò rất quan trọng trong các bài toán thống kê sau này.

f. Phân phối Student hay phân phối t

Phân phối này do William S. Gosset đưa ra năm 1908. Nhiều bạn đọc thắc mắc tại sao lại gọi là phân phối Student ("phân phối sinh viên" vì "student" nghĩa là "sinh viên"), hay Student là tên riêng của người tìm ra phân phối này. Theo Lincohn L. Chao thì phân phối này do William S. Gosset đưa ra. Gosset khi đó đang làm thuê cho Guinness Brewery ở Dublin và hãng này không cho phép Gosset dùng tên thật của mình để công bố các kết quả lý thuyết của ông. Vì thế ông đã viết dưới bút danh "Student" và tên phân phối Student bắt nguồn như vậy. Nhiều khi ta còn gọi phân phối Student là phân phối t trong đó t là một biến ngẫu nhiên, t là một thống kê tiêu chuẩn xác định bởi:

$$t = \frac{\bar{X} - \mu}{s}$$

(ở đây chỉ đơn thuần giới thiệu biểu thức của t, để hiểu rõ các ký hiệu $\bar{X}$ và s, xin mời bạn đọc tìm hiểu ở các tài liệu về thống kê toán học).

Hàm mật độ của t xác định bởi

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

Trong đó

$$\Gamma(u) = \int_0^{\infty} e^{-x} x^{u-1}\,dx\,,\ \Gamma(u+1) = u\Gamma(u)\,,\ \Gamma(1) = 1$$

$$\Gamma(1/2) = \sqrt{\pi}$$

Hàm mật độ của phân phối t cũng là hàm đối xứng qua trục tung, dạng đồ thị của nó cũng có dạng hình chuông rất giống hàm mật độ chuẩn $\varphi(x)$.

Số nguyên n được gọi là số bậc tự do của phân phối t.

Ta có kết quả sau: Nếu $X_1, \ldots X_n$ độc lập, cùng phân phối N(0, 1) thì $\dfrac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$ có phân phối Student.

g. Phân phối χ^2

Hàm mật độ của phân phối χ^2 có dạng

$$p(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ \dfrac{x^{n/2-1}}{2^{n/2}\,\Gamma(n/2)} e^{-\frac{x}{2}} & \text{nếu } x > 0 \end{cases}$$

n gọi là bậc tự do của phân phối χ^2.

Thực chất của phân phối χ^2 với n bậc tự do chính là phân phối của biến ngẫu nhiên $\sum_{i=1}^{n} X_i^2$ trong đó $X_1, X_2, \ldots X_n$ độc lập, cùng phân phối N(0, 1).

Phân phối χ^2 do Karl Pearson đưa ra vào năm 1900.

h. Phân phối F.

Phân phối F do R. A. Fisher đưa ra.

Phân phối F là phân phối của tỷ số hai biến ngẫu nhiên độc lập có phân phối χ^2 với n_1 và n_2 bậc tự do.

Biến ngẫu nhiên

$$F = \frac{\chi^2_{n_1}}{n_1} : \frac{\chi^2_{n_2}}{n_2} = \frac{n_2}{n_1} \times \frac{\chi^2_{n_1}}{\chi^2_{n_2}}$$

Hàm mật độ của phân phối F có dạng:

$$p(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ C(n_1, n_2)\dfrac{x^{\frac{n_1}{2}-1}}{(n_2 + n_1x)^{\frac{(n_1+n_2)}{2}}} & \text{nếu } x > 0 \end{cases}$$

Trong đó $C(n_1, n_2)$ là hằng số thích hợp.

Ví dụ 9: Tiến hành 10 quan sát độc lập về biến ngẫu nhiên X có phân phối N(5; 0,16)

a. Tìm xác suất $P\{4 \leq X \leq 5{,}5\}$

b. Tìm xác suất sao cho trong 10 quan sát độc lập về X có 6 lần X nhận giá trị trong [4; 5,5]

a. Theo (5) ta có (ở đây $\mu = 5$, $\sigma = \sqrt{0{,}16} = 0{,}4$):

$$P\{4 \leq X \leq 5{,}5\} = P\left\{\frac{4-5}{0{,}4} \leq \frac{X-5}{0{,}4} \leq \frac{5{,}5-5}{0{,}4}\right\}$$

$$= P\left\{-2{,}5 \leq \frac{X-5}{0{,}4} \leq 1{,}25\right\} = \Phi(1{,}25) - \Phi(-2{,}5)$$

$$= 0{,}8944 - 0{,}0062 = 0{,}8882$$

(Do X là biến ngẫu nhiên liên tục nên $P(X = x_0) = 0$ nghĩa là $P(\frac{X-5}{0{,}4} = 1{,}25) = 0$).

b. 10 quan sát độc lập chính là 10 phép thử Bernoulli với $A = \{4 \leq X \leq 5{,}5\}$, $p = 0{,}8882$. Vì vậy xác suất để trong 10 quan sát có 6 lần $X \in [4; 5{,}5]$ sẽ là: $C_{10}^6 \times 0{,}8882^6 \cdot 0{,}1118^4$.

§ 3. CÁC SỐ ĐẶC TRƯNG CỦA ĐẠI LƯỢNG NGẪU NHIÊN

1. Kỳ vọng (giá trị trung bình):

Định nghĩa: Kỳ vọng của biến ngẫu nhiên ξ là một con số được ký hiệu là $E\xi$ và được xác định như sau:

$$E\xi = \begin{cases} \sum_i x_i p_i & \text{nếu } P(\xi = x_i) = p_i \\ \int_{-\infty}^{\infty} xp(x)dx & \text{nếu } \xi \text{ có mật độ } p(x) \end{cases}$$

Ý nghĩa: Kỳ vọng của biến ngẫu nhiên là giá trị trung bình mà biến ngẫu nhiên nhận, hoặc kỳ vọng của biến ngẫu nhiên là trọng tâm của phân phối xác suất với khối lượng 1. Chính vì vậy mà người ta dùng kỳ vọng để xác định vị trí của phân phối.

Ta đã quen biết với trung bình cộng hoặc trung bình số học. Trung bình cộng sẽ tốt trong trường hợp các số trị đang xét có cùng khả năng, nhưng trung bình cộng phản ánh không tốt trong trường hợp các số trị được nhận không đồng khả năng, khi đó dùng trung bình có trọng lượng hoặc trung bình theo nghĩa xác suất trên sẽ tốt hơn nhiều. Trường hợp các p_i bằng nhau ta lại nhận được trung bình số học.

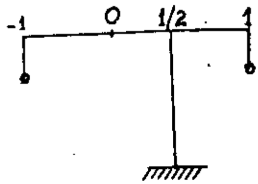
Để thấy rõ ý nghĩa nêu trên ta minh hoạ bởi ví dụ đơn giản sau:

X	-1	1
P	1/4	3/4

Khi đó: $EX = (-1).1/4 + 1.3/4 = 1/2$

(trong khi đó nếu trung bình số học sẽ là $1/2(-1+1) = 0$).

Ta hình dung một thanh có độ dài như bên, tại 2 đầu mút ta treo 2 trọng lượng 1/4 kg và 3/4kg. Khi đó trọng tâm của thanh chính là điểm 1/2, tức là điểm EX, điểm đó lệch về đầu mút có trọng lượng lớn.



Tính chất:

a. $EC = C$ (C là hằng số)

b. $ECX = CEX$

c. $E(X \pm Y) = EX \pm EY$

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

e. $Ef(X) = \sum_i f(x_i)p_i$ nếu $P(X = x_i) = p_i$

$Ef(X) = \int_{-\infty}^{\infty} f(x)p(x)dx$ nếu X có mật độ p(x)

Chứng minh: Ta chứng minh các tính chất trên trong trường hợp biến ngẫu nhiên rời rạc

a. Vì $P(X = C) = 1$ nên $EX = EC = C.1 = C$

b. Ta có thể đưa một hằng số ra ngoài dấu tổng nên tính chất b được chứng minh:

$$P(X = x_i) = p_i,\ P(C.X = C.x_i) = p_i$$
$$EC.X = \sum_i Cx_ip_i = C\sum_i x_ip_i = CEX$$

c. Ta xác định phân phối xác suất của X + Y:

Giả sử: $P(X = x_i) = p_i$

$P(Y = y_j) = q_j$

$P\{X = x_i, Y = y_j\} = P\{(X = x_i) \cap (Y = y_j)\}$

(nếu X, Y độc lập) $= P(X = x_i) \times P(Y = y_j) = p_iq_j$

$$E(X+Y) = \sum_{i,j}(x_i + y_j)p_iq_j = \sum_{i,j} x_ip_iq_j + \sum_{i,j} y_jp_iq_j$$

$$= \sum_{i} x_ip_i \sum_{j} q_j + \sum_{j} y_jq_j \sum_{i} p_i = EX\,.\,1 + EY\,.\,1$$

$$= EX + EY$$

Tính chất c. không đòi hỏi tính độc lập, nhưng để đơn giản trong chứng minh ta xét trường hợp độc lập.

d. Tương tự ta chứng minh tính chất d)

$$P(XY = x_iy_j) = P(X = x_i) \times P(Y = y_j) = p_iq_j$$

$$E(XY) = \sum_{i,j} x_iy_jp_iq_j = \sum_{i} x_ip_i \sum_{j} y_j\,q_j = EX\,.\,EY$$

e. Vì $P(X = x_i) = p_i$ nên $P(f(X) = f(x_i)) = p_i$

Ta suy ra tính chất e).

2. Median (Trung vị):

Median của biến ngẫu nhiên ξ là một số được ký hiệu μ_ξ và được xác định như sau:

$$P(\xi < \mu_\xi) = F(\mu_\xi) \leqslant 1/2$$

$$P(\xi \leqslant \mu_\xi) = F(\mu_\xi + 0) \geqslant 1/2$$

Trong đó F là hàm phân phối của ξ. Nếu hàm phân phối F liên tục thì hai hệ thức trên tương đương với:

$$F(\mu_\xi) = 1/2 \tag{8}$$

Nếu có nhiều nghiệm, chẳng hạn m_o và m_1 là nghiệm thì mọi điểm thuộc $[m_o, m_1]$ cũng đều là nghiệm.

$m_o \equiv m_1$: có một trung vị

$m_o \neq m_1$: có nhiều trung vị

Như vậy trung vị là điểm phân đôi khối lượng xác suất thành hai phần bằng nhau.

3. Mode:

Nếu ξ rời rạc thì Mode là giá trị của ξ mà tại đó xác suất tương ứng lớn nhất.

Nếu ξ liên tục có mật độ $p(x)$ thì Mode là giá trị x_0 mà tại đó $p(x)$ đạt cực đại.

Nhận xét: Nếu phân phối của biến ngẫu nhiên ξ đối xứng và có một Mode thì cả 3 đặc trưng: Kỳ vọng, Median và Mode trùng nhau.

Nếu phân phối của ξ đối xứng hoặc gần đối xứng thì dùng kỳ vọng định vị là tốt nhất.

Nếu phân phối của ξ quá lệch thì dùng trung vị và Mode để định vị sẽ tốt hơn.

4. Phương sai:

Định nghĩa: Phương sai của biến ngẫu nhiên ξ là một số không âm, ký hiệu là $D\xi$, được xác định bởi:

$$D\xi = E(\xi - E\xi)^2 \tag{9}$$

$$= E\xi^2 - (E\xi)^2 \tag{10}$$

(Từ (9): $E(\xi - E\xi)^2 = E\{\xi^2 + (E\xi)^2 - 2\xi E\xi\}$

$= E(\xi^2) + (E\xi)^2 - 2E\xi \,.\, E\xi = E\xi^2 - (E\xi)^2$)

Trong đó theo tính chất e) của kỳ vọng:

$$E\xi^2 = \sum_i p_i x_i^2 \qquad \text{nếu } P(\xi = x_i) = p_i$$

$$= \int_{-\infty}^{\infty} x^2 p(x)dx \qquad \text{nếu } \xi \text{ có mật độ } p(x)$$

Ý nghĩa: Phương sai của biến ngẫu nhiên là một số không âm dùng để đo mức độ phân tán (mức độ tản mát) của các giá trị của biến ngẫu nhiên ξ xung quanh tâm ($E\xi$) của nó. $D\xi$ nhỏ thì mức độ phân tán nhỏ, độ tập trung lớn. $D\xi$ càng lớn thì độ phân tán càng cao.

$E\xi$: Tâm của phân phối

$\xi - E\xi$: Khoảng cách từ giá trị của biến ngẫu nhiên ξ đến tâm

$(\xi - E\xi)^2$: Bình phương khoảng cách trên

$E(\xi - E\xi)^2$: Trung bình của bình phương khoảng cách trên.

Tính chất

a. $Dc = 0$, $c = \text{const.}$

b. $Dc\xi = c^2 D\xi$

c. Nếu ξ và η độc lập thì $D(\xi \pm \eta) = D\xi + D\eta$

Chứng minh các tính chất này đều dựa vào định nghĩa (9), chẳng hạn:

$$
\begin{aligned}
Dc\xi &= E\{c\xi - Ec\xi\}^2 = Ec^2(\xi - E\xi)^2 = c^2 D\xi \\
D(\xi + \eta) &= E\{(\xi + \eta) - (E\xi + E\eta)\}^2 \\
&= E\{(\xi - E\xi) + (\eta - E\eta)\}^2 \\
&= E(\xi - E\xi)^2 + E(\eta - E\eta)^2 + 2E(\xi - E\xi)(\eta - E\eta) \\
&= D\xi + D\eta + 2E(\xi - E\xi)\, E(\eta - E\eta) \\
&= D\xi + D\eta + 0
\end{aligned}
$$

Bạn đọc cần lưu ý: Từ tính chất b. ta thấy $D(-\xi) = D\xi$

Ký hiệu $\sigma = \sqrt{D\xi}$ gọi là độ lệch tiêu chuẩn (σ đọc là xích ma). Còn $\frac{\sigma}{\mu}.100\%$ gọi là hệ số biến thiên.

5. Phân vị cấp p:

x_p được gọi là phân vị cấp p của phân phối $F(x)$ nếu:

$$F(x_p) \leqslant p$$

$$F(x_p + 0) \geqslant p$$

Nếu hàm phân phối liên tục: $F(x_p) = p$

– Trường hợp $p = 1/2$ ta có trung vị

– Ta có các tứ phân vị $x_{1/4}$, $x_{2/4}$, $x_{3/4}$

Khi đó $P\{\xi \in [x_{1/4}, x_{3/4}]\} = 0{,}50$ (nếu ξ liên tục). Khoảng $(x_{1/4}, x_{3/4})$ được gọi là khoảng tứ phân vị. Khoảng này cũng được dùng để đặc trưng độ tập trung, phân tán của biến ngẫu nhiên.

6. Kỳ vọng và phương sai của một số phân phối thường gặp:

Ví dụ 10: $\xi = B(n, p)$

Ta sẽ chỉ rõ: $E\xi = np$, $D\xi = np(1-p)$

Trước hết ta xây dựng các biến ngẫu nhiên $\xi_1, \dots, \xi_n$

$$\text{trong đó } \xi_i = \begin{cases} 1 & \text{nếu ở phép thử thứ i A xuất hiện} \\ 0 & \text{nếu ở phép thử thứ i } \bar{A} \text{ xuất hiện .} \end{cases}$$

$$\xi_i = \begin{cases} 1 & \text{với xác suất } P(A) = p \\ 0 & \text{với xác suất } P(\bar{A}) = 1 - p = q \end{cases}$$

Do n phép thử độc lập nên $\xi_1, \xi_2, \dots, \xi_n$ độc lập.

Phân phối ξ_i xác định như trên (i = 1, 2,..., n) tức là chúng cùng phân phối.

Ta có: $E\xi_i = 1p + 0q = p$

$$E\xi_i^2 = 1^2p + 0q = p$$

$$D\xi_i = E\xi_i^2 - (E\xi_i)^2 = 1p - p^2 = p(1-p) = pq$$

Mặt khác: $\xi = \sum_i^n \xi_i$ = số lần xuất hiện A trong n phép thử. Vậy theo tính chất kỳ vọng và phương sai ta có:

$$E\xi = \sum_i^n E\xi_i = nE\xi_i = np$$

$$D\xi = \sum_i^n D\xi_i = nD\xi_i = npq$$

Ví dụ 11: $\xi \simeq P_\lambda$, ξ có phân phối Poisson với tham ẩn $\lambda > 0$

Khi đó $\quad E\xi = \lambda, \qquad D\xi = \lambda$

Thật vậy: $$E\xi = \sum_{k=0}^{\infty} k \,.\, e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k k}{k(k-1)!}$$

$$= e^{-\lambda} \sum_{s=0}^{\infty} \frac{\lambda^{s+1}}{s!}$$

$$= e^{-\lambda} \lambda \sum_{s=0}^{\infty} \frac{\lambda^s}{s!} = e^{-\lambda} e^{\lambda} \lambda = \lambda$$

$$E\xi^2 = \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k k}{(k-1)!} = e^{-\lambda} \sum_{s=0}^{\infty} \frac{(s+1)\lambda^{s+1}}{s!}$$

$$= e^{-\lambda} \lambda \sum_{s=0}^{\infty} \frac{s\lambda^s}{s!} + \lambda e^{-\lambda} \sum_{s=0}^{\infty} \frac{\lambda^s}{s!}$$

$$= \lambda E\xi + \lambda e^{-\lambda} e^{\lambda} = \lambda^2 + \lambda$$

$$D\xi = E\xi^2 - (E\xi)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Ví dụ 12: ξ có phân phối đều trên [a, b]

$$E\xi = \int_{-\infty}^{+\infty} xp(x)dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2}\Big|_a^b = \frac{a+b}{2}$$

$$E\xi^2 = \int_{-\infty}^{+\infty} x^2 p(x)dx = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a} \frac{x^3}{3}\Big|_a^b = \frac{a^2+ab+b^2}{3}$$

$$D\xi = E\xi^2 - (E\xi)^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

Ví dụ 13: $\xi \simeq N(\mu, \sigma^2)$

Ta có: $\quad E\xi = \mu, \qquad D\xi = \sigma^2$

Thật vậy: $$E\xi = \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Đặt $\dfrac{x-\mu}{\sigma} = t$ ta có

$$E\xi = \int_{-\infty}^{\infty} \frac{\mu+\sigma t}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}} \sigma dt$$

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

$$= \mu \,.\, 1 + 0 = \mu$$

$$E\xi^2 = \int_{-\infty}^{\infty} \frac{x^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{(\mu+\sigma t)^2}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2}} \sigma dt$$

$$= \frac{\mu^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \sigma dt + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \times e^{\frac{-t^2}{2}} dt + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \times e^{\frac{-t^2}{2}} dt$$

$$= \mu^2 \,.\, 1 + \frac{\sigma^2}{\sqrt{2\pi}} \left\{ -t \times e^{\frac{-t^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{\infty} e^{\frac{-t^2}{2}} dt \right\} + 0$$

$$= \mu^2 + \frac{\sigma^2}{\sqrt{2\pi}} \left\{ 0 + \int_{-\infty}^{\infty} e^{\frac{-t^2}{2}} dt \right\} = \mu^2 + \sigma^2$$

$$D\xi = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

Từ tính chất của kỳ vọng và phương sai ta thấy ngay rằng: Nếu $\xi = N(\mu, \sigma^2)$ thì $\dfrac{\xi-\mu}{\sigma}$ có kỳ vọng 0 và phương sai 1. Từ biến ngẫu nhiên chuẩn tùy ý ta có thể đưa về chuẩn N(0, 1).

Ví dụ 14: ξ có phân phối mũ với tham ẩn $\lambda > 0$

Khi đó: $E\xi = \frac{1}{\lambda}$, $D\xi = \frac{1}{\lambda^2}$

Thật vậy: $E\xi = \int_0^{+\infty} x\lambda e^{-\lambda x}dx = -xe^{-\lambda x}\Big|_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x}dx$

$$= 0 - \frac{1}{\lambda}e^{-\lambda x}\Big|_0^{+\infty} = \frac{1}{\lambda}$$

$$E\xi^2 = \int_0^{+\infty} x^2\lambda e^{-\lambda x}dx = -x^2e^{-\lambda x}\Big|_0^{+\infty} + 2\int_0^{+\infty} xe^{-\lambda x}dx$$

$$= 0 + \frac{2}{\lambda}E\xi = \frac{2}{\lambda^2}$$

$$D\xi = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Ví dụ 15: Giả sử $\xi \simeq N(3 ; 0,25)$, $\eta \simeq P_4$ và ξ, η độc lập. Tìm

a. $E(-2\xi + 3\eta + 5)$

b. $D(-2\xi + 3\eta + 5)$

c. $E(\xi^2 - 2\eta^2 + 3\xi\eta - 2\eta + 1)$

Theo các kết quả trên ta có:

$$E\xi = 3 ;\ D\xi = 0,25 ;\ E\eta = 4 ,\ D\eta = 4$$

$$E\xi^2 = D\xi + (E\xi)^2 = 0,25 + 3^2 = 9,25$$

$$E\eta^2 = D\eta + (E\eta)^2 = 4 + 4^2 = 20$$

a. $E(-2\xi + 3\eta + 5) = -2E\xi + 3E\eta + 5$

$= -2.3 + 3.4 + 5 = 11$

b. $D(-2\xi + 3\eta + 5) = (-2)^2D\xi + 3^2D\eta$

$= 4.0,25 + 9.4 = 37$

c. $E(\xi^2-2\eta^2 + 3\xi\eta-2\eta+1) = E\xi^2-2E\eta^2 + 3E\xi E\eta-2E\eta+1$

$= 9,25 - 2.20 + 3.3.4 - 2.4 + 1 = -1,75$

Bây giờ ta xét một số ví dụ mang tính chất tổng hợp, ứng dụng các kết quả đã đề cập ở các phần trên.

Ví dụ 16: Trong một lô hàng (chẳng hạn lô các chai thuốc sâu, lô hàng mỳ chính, lô các chai bia,...) có tỷ lệ hàng giả là 25%. Đội quản lý thị trường đã lấy ngẫu nhiên ra từng chiếc một cho đến khi nào gặp chiếc giả thì thôi không lấy nữa. Gọi ξ là số chiếc thật (số sản phẩm chính phẩm) đã lấy ra.

a. Lập bảng phân phối xác suất của ξ

b. Về trung bình phải lấy bao nhiêu chiếc mới gặp chiếc giả.

c. Nếu gọi η là số chiếc đã lấy ra (kể cả chiếc giả), khi đó phân phối của η và $E\eta$ sẽ như thế nào?

a. Ta có bảng phân phối xác suất như sau:

$$P(\xi = m) = 0{,}75^m . 0{,}25 \text{ với } m = 0, 1, 2, \ldots$$

(vì m chiếc đầu là hàng thật, còn chiếc thứ (m+1) là hàng giả).

Phân phối dạng trên được gọi là *phân phối hình học*. Đó là phân phối của số phép thử không thành công được thực hiện cho đến khi thành công (có kết quả). Nó là trường hợp riêng của *phân phối Pascal và phân phối nhị thức âm.*

b. Để trả lời câu hỏi này ta dùng $E\xi$, vì $E\xi$ chính là trung bình số các sản phẩm giả đã lấy ra.

$$E\xi = \sum_{m=o}^{\infty} mP(\xi = m)$$

$$= \sum_{m=o}^{\infty} m0{,}75^m . 0{,}25 = 0{,}25 . \sum_{m=o}^{\infty} m\, 0{,}75^m$$

$$= 0{,}25 . \frac{0{,}75}{(1-0{,}75)^2} = \frac{0{,}75}{0{,}25} = 3$$

(ta có $\sum\limits_{m=o}^{\infty} mp^m = p\sum\limits_{m=1}^{\infty} mp^{m-1} = p\left(\sum\limits_{m=1}^{\infty} p^m\right)' = p\left(\frac{1}{1-p}\right)' = \frac{p}{(1-p)^2}$ với $|p| < 1$). Như vậy $E\xi = \dfrac{1-p}{p}$

c. Ta dễ dàng nhận được

$$P(\eta = m) = 0{,}75^{m-1} . 0{,}25 \text{ với } m = 1, 2, 3, \ldots$$

$$E\eta = E\xi + 1 = \frac{1-p}{p} + 1 = \frac{1}{p} = \frac{1}{0,25} = 4$$

(hoặc $= 3 + 1 = 4$)

Ví dụ 17: Khả năng xuất hiện một loại vi trùng (mà ta đang quan tâm) ở một thí nghiệm là 10%. Một cán bộ nghiên cứu đã làm từng thí nghiệm một cho đến khi nào thành công (nhận được loại vi trùng trên) thì thôi. Nhưng anh ta chỉ được cấp kinh phí để làm tối đa 20 thí nghiệm. Gọi ξ là số thí nghiệm không thành công mà cán bộ trên đã làm. Hãy lập bảng phân phối xác suất của ξ (Phân phối dạng này được gọi là *phân phối hình học bị chặt).*

Ta có $\quad P(\xi = m) = 0,9^m . 0,1$ với $m = 0, 1, 2,..., 19$

$P(\xi = 20) = 0,9^{20}$

Dễ dàng kiểm tra lại tổng các xác suất bằng 1:

$$\sum_{m=0}^{n-1} p^m . q + p^n = q . \frac{1-p^n}{1-p} + p^n = 1; \ (1 - p = q)$$

Ví dụ 18: Trở lại ví dụ 16 về lô hàng có chứa hàng giả. Đội quản lý thị trường lấy ngẫu nhiên từng chiếc một ra 10 chiếc thì dừng lại. Gọi η là số chiếc giả trong 10 chiếc lấy ra.

a. Lập bảng phân phối xác suất của η

b. Về trung bình trong 10 chiếc lấy ra có mấy chiếc giả.

c. Hãy đoán xem trong 10 chiếc lấy ra có mấy chiếc giả là xảy ra nhiều hơn cả. Tính xác suất xảy ra trường hợp đó.

d. Tính $E(5\eta - 3)$, $D(3 - 5\eta)$.

a. Cho biết tỷ lệ hàng giả là 25%, việc lấy ra 10 chiếc (lấy 10 lần) chính là thực hiện 10 phép thử Bernoulli với $p = 0,25$. Do đó η có phân phối nhị thức $B(10; 0,25)$:

$$P(\eta = m) = C_{10}^m \ 0,25^m \ 0,75^{10-m}; \ m = 0, 1, ..., 10$$

b. Về trung bình sẽ có $E\eta = 10.0,25 = 2,5$ chiếc giả trong 10 chiếc lấy ra

c. Để trả lời câu hỏi này ta dùng Mode hoặc chính là số có khả năng nhất

$$np + p - 1 = 10.0,25 + 0,25 - 1 = 1,75$$

suy ra $Mod(\eta) = 2$

Xác suất xảy ra trường hợp này là:

$$P(\eta = 2) = C_{10}^2\, 0,25^2 \,.\, 0,75^8 \approx 0,2816$$

d. $$E(5\eta - 3) = 5 \,.\, E\eta - 3 = 5 \,.\, 10 \,.\, 0,25 - 3 = 9,5$$

$$D(3 - 5\eta) = (-5)^2 D\eta = 25 \,.\, 10 \,.\, 0,25 \,.\, 0,75 = 46,875$$

Ví dụ 19: Giả sử trong 10 chiếc lấy ra có 3 chiếc giả. Lấy ngẫu nhiên ra 4 chiếc (lấy cùng lúc) từ 10 chiếc này. Gọi ξ là số chiếc giả trong 4 chiếc lấy ra.

a. Mô tả phân phối của ξ

b. Dự đoán xem trong 4 chiếc lấy ra sẽ gặp mấy chiếc giả.

c. Viết biểu thức hàm phân phối của ξ

d. Tính $E\xi$, $D\xi$, $P\{0 < \xi \leq 2,2\}$

a. Phép thử ở đây là lấy cùng lúc và chỉ có một phép thử (do đó ξ sẽ không phải là biến ngẫu nhiên nhị thức như trong ví dụ 18), cho nên ta có bảng phân phối của ξ như sau (Phân phối dạng này được gọi là *phân phối siêu hình học*)

ξ	0	1	2	3
$P(\xi = k)$	$\frac{C_3^0 C_7^4}{C_{10}^4}$	$\frac{C_3^1 C_7^3}{C_{10}^4}$	$\frac{C_3^2 C_7^2}{C_{10}^4}$	$\frac{C_3^3 C_7^1}{C_{10}^4}$
	0,17	0,50	0,3	0,03

b. Vì $Mod(\xi) = 1$ nên ta dự đoán trong 4 chiếc lấy ra sẽ có 1 chiếc giả (xác suất tương ứng = 0,5 là lớn nhất)

c. Hàm phân phối

$$F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 0,17 & 0 < x \leq 1 \\ 0,67 & 1 < x \leq 2 \\ 0,97 & 2 < x \leq 3 \\ 1 & 3 < x \end{cases}$$

d. $E\xi = 0 + 0{,}50 + 2\,.\,0{,}3 + 3\,.\,0{,}03 = 1{,}19$

$E\xi^2 = 0 + 0{,}50 + 2^2.\,0{,}3 + 3^2\,.\,0{,}03 = 1{,}97$

$D\xi = 1{,}97 - 1{,}19^2 = 0{,}5539$

$P\{0 < \xi \leqslant 2{,}2\} = P\{(\xi = 1) \cup (\xi = 2)\}$

$= P(\xi = 1) + P(\xi = 2) = 0{,}80$

Phân phối siêu hình học (cũng còn gọi là phân phối siêu bội) có thể xây dựng tổng quát như sau: Một tập gồm N phần tử trong đó có M phần tử mang dấu hiệu I, do đó (N–M) phần tử mang dấu hiệu II, ta lấy ngẫu nhiên đồng thời ra n phần tử. Gọi ξ là số phần tử mang dấu hiệu I trong n phần tử được lấy ra. ξ nhận các giá trị m = 0, 1, 2,... min(n,M). Khi đó phân phối xác suất của ξ được xác định như sau:

$$P(\xi = m) = \frac{C_M^m\, C_{N-M}^{n-m}}{C_N^n};\ m = 0,\ 1,\ ...,\ \min\ (n,M)$$

Bỏ qua tính toán ta có kết quả sau:

$$E\xi = n\,.\,\frac{M}{N}\ ;\ D\xi = n\,.\,\frac{M}{N}\,.\,\frac{N-M}{N}\,\frac{1-\frac{n}{N}}{1-\frac{1}{N}}$$

$$= \frac{nM(N-M)}{N^2}\left(1 - \frac{n-1}{N-1}\right)$$

Áp dụng kết quả này vào ví dụ 19 ta có

$$E\xi = \frac{4.3}{10} = 1{,}2$$

$$D\xi = \frac{4.3.7}{10^2}\left(1 - \frac{3}{9}\right) = 0{,}56$$

(Kết quả nhận được ở đây và ở trong ví dụ 19 đều là các kết quả đúng. Sở dĩ có sự sai khác này là do trong ví dụ 19 các xác suất đã được tính xấp xỉ).

Ví dụ 20: Gọi X là số nhu cầu gọi đến trung tâm điều phối xe taxi trong khoảng thời gian 1 giờ. Giả sử X có phân phối Poisson với $\lambda = 6{,}5$. Cho biết $e^{-6,5} \approx 0{,}0016$.

a. Về trung bình trong 1 giờ có bao nhiêu nhu cầu dùng xe taxi.

b. Khả năng vắng khách trong thời gian 1 giờ là bao nhiêu?

c. Tìm Mod(X)

d. Tính $P\{EX - \sqrt{DX} \leqslant X \leqslant EX + \sqrt{DX}\}$. Diễn đạt kết quả nhận được bằng lời dưới dạng dễ hiểu.

a. Theo giả thiết ta thấy về trung bình trong 1 giờ có 6,5 nhu cầu dùng xe taxi.

b. $P(X = 0) = e^{-6,5} \approx 0{,}0016$

c. Để tìm Mod(X) trong phân phối Poisson ta thấy:

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!} = p_{k-1}\frac{\lambda}{k}$$

Suy ra $\quad p_k / p_{k-1} = \dfrac{\lambda}{k} = \begin{cases} \geqslant 1 & \text{nếu } \lambda \geqslant k \\ < 1 & \text{nếu } \lambda < k \end{cases}$

Các giá trị p_k với $k = 0, 1, 2...$ sẽ tăng dần đến max rồi lại giảm dần đến 0. Vậy nếu λ nguyên thì Mod(X) là $\lambda-1$ và λ. Nếu λ không nguyên thì Mod(X) là $[\lambda]$, tức là số nguyên lớn nhất nhưng bé hơn λ.

Áp dụng vào ví dụ này ta thấy Mod(X) là 6.

d. $$P\{EX - \sqrt{DX} \leqslant X \leqslant EX + \sqrt{DX}\}$$
$$= P\{6{,}5 - \sqrt{6{,}5} \leqslant X \leqslant 6{,}5 + \sqrt{6{,}5}\}$$
$$= P\{3{,}95 \leqslant X \leqslant 9{,}05\} = P\{(X = 4) \cup (X = 5) \cup ... \cup (X = 9)\}$$
$$= e^{-6,5}\left(\frac{6{,}5^4}{4!} + ... + \frac{6{,}5^9}{9!}\right) \approx 0{,}0016 \,.\, 509{,}21 \approx 0{,}8147$$

Điều đó có nghĩa khả năng để trong 1 giờ có từ 4 đến 9 nhu cầu dùng taxi là 81,47%.

Ví dụ 21: Gọi X là trọng lượng của con gà công nghiệp 4 tháng tuổi. Giả sử $X \simeq N(3{,}5;\ (0{,}25)^2)$, (X^{kg}).

a. Con số 3,5 nói lên điều gì

b. Tính tỷ lệ đàn gà có trọng lượng trên 3 kg.

c. Xuất chuồng ngẫu nhiên 100 con. Có bao nhiêu con trọng lượng trên 3 kg là có khả năng nhất. Viết công thức tính xác suất xảy ra trường hợp đó.

d. Gọi Y là số con gà nặng trên 3kg trong 100 con bắt ra. Mô tả phân phối xác suất của Y.

a. Con số 3,5 trong phân phối chuẩn cho ta biết trọng lượng trung bình của một con gà là 3,5kg.

b. Tỷ lệ đàn gà có trọng lượng trên 3kg chính là $P(X > 3)$.

$$P(3 < X) = P(3 < X < 10) = \Phi\left(\frac{10-3{,}5}{0{,}25}\right) - \Phi\left(\frac{3-3{,}5}{0{,}25}\right)$$

$$= \Phi(26) - \Phi(-2) = 1 - \Phi(-2) = \Phi(2) = 0{,}98$$

Vậy tỷ lệ đàn gà có trọng lượng trên 3kg là 98%.

c. Đây là 100 phép thử Bernoulli, $p = 0{,}98$, nên ta có $np + p - 1 = 100.0{,}98 + 0{,}98 - 1 = 97{,}98$.

Vậy có 98 con. Xác suất cần tìm là $C_{100}^{98}\ (0{,}98)^{98}\ (0{,}02)^2$

d. $Y \simeq B(100;\ 0{,}98)$

$P(Y = m) = C_{100}^{m}\,(0{,}98)^m\,(0{,}02)^{100-m}\ ;\ m = 0,1,\ldots,100$

Ví dụ 22: Một mảng mạch điện tử có thời gian sống tuân theo luật mũ với $\lambda = 0{,}0001$ (đơn vị đo là giờ).

a. Cho biết thời gian sống trung bình của mảng mạch điện tử.

b. Nếu thời gian bảo hành là 6 tháng (tức 4380 giờ) thì tỷ lệ các mảng mạch điện tử loại trên cần đưa đến trạm bảo hành là bao nhiêu %.

a. Thời gian sống trung bình là $\frac{1}{\lambda} = 10\,000$ giờ

b. Tỷ lệ các mảng mạch điện tử cần đưa đến trạm bảo hành chính là xác suất để mảng mạch bị hỏng trước 4380 giờ.

P{mảng mạch điện tử làm việc $\leq$ 4380 giờ}

$$= 1 - e^{-\lambda.4380} = 1 - e^{-0,438} \approx 1 - 0,65 = 0,35$$

Vậy tỷ lệ cần tìm là 35%.

Qua ví dụ 18, 19 và 21 (câu c) bạn đọc cần tự lý giải và phân biệt xem khi lấy ra k phần tử từ một tập các phần tử thì khi nào là lấy từng phần tử một (như ví dụ 18, 21 (c)), khi nào lấy cùng lúc (như ví dụ 19) (nếu như đề bài không chỉ rõ). Đây cũng là một điểm nút của nhiều bài toán mà không ít bạn đọc lúng túng hoặc nhầm lẫn.

§ 4. VECTƠ NGẪU NHIÊN

1. Phân phối đồng thời:

Giả sử $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ trong đó ξ_i, i = 1, 2, ..., n là các biến ngẫu nhiên một chiều. ξ được gọi là vectơ ngẫu nhiên n chiều.

Ta định nghĩa hàm phân phối của vectơ ngẫu nhiên ξ như sau:

$$F_\xi(x_1, \dots, x_n) = P\left\{\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n\right\}.$$

Nếu các thành phần ξ_i, i = 1, 2, ..., n của vectơ ngẫu nhiên ξ độc lập với nhau thì

$$F_\xi(x_1, x_2, \dots, x_n) = P\left\{\xi_1 < x_1\right\}.P\left\{\xi_2 < x_2\right\} \dots P\left\{\xi_n < x_n\right\}$$
$$= F_{\xi_1}(x_1)\, F_{\xi_2}(x_2) \dots F_{\xi_n}(x_n)$$

Trong trường hợp này, nếu ξ_i có hàm mật độ thì

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n} = f_{\xi_1}(x_1) . f_{\xi_2}(x_2) \dots f_{\xi_n}(x_n)$$

trong đó f, f_{ξ_i} ký hiệu là hàm mật độ của vectơ ξ và thành phần ξ_i.

Bây giờ ta xét kỹ hơn trong trường hợp 2 chiều (X, Y).

Nếu X, Y đều là các đại lượng ngẫu nhiên rời rạc với $P(X = x_i) = p_i$, $P(Y = y_j) = q_j$, ta có bảng phân phối xác suất của vectơ ngẫu nhiên (X, Y) như sau:

X \ Y	y_1	.....	y_j	....
x_1	p_{11}	.....	p_{1j}	....
$\vdots$	$\vdots$		$\vdots$	
x_i	p_{i1}	.....	p_{ij}	....
$\vdots$	$\vdots$		$\vdots$	

trong đó $p_{ij} = P\left\{ (X = x_i) \quad (Y = y_j)\right\}$

Nếu X, Y độc lập thì $p_{ij} = p_i q_j$

Rõ ràng ta có:

$$(X = x_i) = (X = x_i)\Omega = (X = x_i)\left\{ \bigcup_j (Y = y_j)\right\} = \bigcup_j (X = x_i)(Y = y_j)$$

Do các biến cố $(Y = y_j)$, $j = 1, 2, \ldots$ xung khắc với nhau nên

$$p_i = P(X = x_i) = \sum_j P\left\{ (X = x_i)(Y = y_j)\right\} = \sum_j p_{ij} \quad ; \ i = 1, 2, \ldots$$

$$q_j = P(Y = y_j) = \sum_i p_{ij} \quad ; \ j = 1, 2, \ldots$$

Ví dụ 23: Gieo 2 đồng tiền cân đối đồng chất. Gọi X, Y là 2 biến ngẫu nhiên chỉ kết quả nhận được của 2 đồng tiền tương ứng. Lập bảng phân phối xác suất của (X, Y).

Y \ X	S	N	p_i
S	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
N	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
q_j	$\frac{1}{2}$	$\frac{1}{2}$	1

Ví dụ 24: Có hai cái hộp, mỗi hộp đựng, 6 bi; trong hộp I có: 1 bi mang số 1, 2 bi mang số 2, 3 bi mang số 3; trong hộp II có: 2 bi mang số 1, 3 bi mang sô 2, 1 bi mang số 3. X là số ghi trên bi rút ra từ hộp I, Y là số ghi trên bi rút ra từ hộp II. Rút từ mỗi hộp một bi. Hãy lập bảng phân phối xác suất của vectơ ngẫu nhiên (X, Y).

Mỗi hộp có 6 bi cho nên số khả năng có thể của phép thử là 6.6 = 36. 36 trường hợp này là có cùng khả năng xảy ra. Khi đó ta có:

Y \ X	1	2	3	p_i
1	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{9}$	$\frac{1}{6}$	$\frac{1}{18}$	$\frac{1}{3}$
3	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{2}$
q_j	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$	1

(Trong 36 trường hợp có thể thì có 2 trường hợp (1, 1), 3 trường hợp (1, 2), 4 trường hợp (2, 1), v.v...)

Ví dụ 25: Hàm mật độ chuẩn 2 chiều

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}}$$

trong đó $EX = \mu_1$, $DX = \sigma_1^2$, $EY = \mu_2$, $DY = \sigma_2^2$, ρ là hệ số tương quan (xem định nghĩa ở mục dưới).

2. Các đặc trưng của vectơ ngẫu nhiên:

Gọi $\xi = (\xi_1, \xi_2, ..., \xi_n)$

- Vectơ kỳ vọng $E\xi = (E\xi_1, E\xi_2, ..., E\xi_n)$

- Covariance hay là hiệp phương sai

$$\operatorname{cov}(\xi_i, \xi_j) = E\{(\xi_i - E\xi_i)(\xi_j - E\xi_j)\}$$

Đơn giản ta ký hiệu $\operatorname{cov}(\xi_i, \xi_j) = \lambda_{ij}$. Rõ ràng

$$\lambda_{ij} = \lambda_{ji}$$

$$\lambda_{ii} = D\xi_i$$

- Ma trận mômen, ma trận Covariance của vectơ ngẫu nhiên ξ:

$$\Lambda = (\lambda_{ij})_{n\ n} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nn} \end{pmatrix}$$

Ta có ma trận mômen Λ là ma trận đối xứng, hơn nữa ma trận mômen xác định không âm hoặc các định thức con chính không âm, trong đó $\det\Lambda \geqslant 0$.

Bây giờ ta xét trường hợp hai chiều $\xi = (\xi_1, \xi_2)$.

- Hệ số tương quan: Hệ số tương quan giữa 2 biến ngẫu nhiên ξ_1 và ξ_2 được xác định như sau:

$$\rho = \frac{E\{(\xi_1 - E\xi_1)(\xi_2 - E\xi_2)\}}{\sqrt{D\xi_1}\sqrt{D\xi_2}} = \frac{\lambda_{12}}{\sqrt{\lambda_{11} \times \lambda_{22}}} = \frac{\lambda_{12}}{\sigma_1 \sigma_2} \quad (11)$$

– Trong trường hợp 2 chiều ma trận mômen Λ có dạng:

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix}$$

Do $\det\Lambda = \sigma_1^2\sigma_2^2(1 - \rho^2) \geq 0$,

từ đó suy ra $-1 \leq \rho \leq 1$.

Ở giáo trình này ta công nhận các kết luận sau:

Hệ số tương quan là số đo mức độ phụ thuộc tuyến tính giữa 2 biến ngẫu nhiên. Nếu $|\rho|$ càng gần 1 thì mức độ phụ thuộc tuyến tính giữa chúng càng chặt. Khi đó:

Nếu $\rho > 0$ thì sự phụ thuộc giữa 2 biến là đồng biến.

Nếu $\rho < 0$ thì sự phụ thuộc là nghịch biến.

Nếu $|\rho| = 1$ thì ta có $P\{\xi_2 = a\xi_1 + b\} = 1$ với a, b là các hằng số nào đó.

Nếu $|\rho|$ càng gần 0 thì mức độ phụ thuộc tuyến tính giữa chúng càng yếu.

Nếu $\rho = 0$ ta nói 2 biến ngẫu nhiên không tương quan.

Rõ ràng: Nếu 2 biến ngẫu nhiên độc lập thì chúng không tương quan, nhưng điều ngược lại không đúng.

– Ma trận tương quan: ký hiệu ρ_{ij} là hệ số tương quan giữa ξ_i và ξ_j, ta có ma trận tương quan sau:

$$P = (\rho_{ij})_{n\times n} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{pmatrix}$$

Trở lại trường hợp 2 chiều: $P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

Ta dễ dàng rút ra hệ thức sau: $\Lambda = \sum P \sum$ (12)

trong đó $\sum = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$

Hệ thức (12) còn đúng cho vectơ ngẫu nhiên n chiều.

Ví dụ 26: Trở lại ví dụ 24. Hãy

a. Tìm EX. EY.

b. Lập ma trận covariance

c. Tính hệ số tương quan ρ.

a. Từ bảng phân phối xác suất ở ví dụ 24 ta có

X	1	2	3
$P(X = x_i)$	1/6	1/3	1/2

Y	1	2	3
$P(Y = y_j)$	1/3	1/2	1/6

Do đó $EX = \frac{7}{3}$, $EY = \frac{11}{6}$

(Cũng có thể tính theo công thức $EX = \sum_i \sum_j x_i p_{ij}$

$\left(= \sum_i x_i \sum_j p_{ij} = \sum_i x_i p_i\right)$).

b. Ta cần tìm $\lambda_{11} = DX = EX^2 - (EX)^2 = 6 - \frac{49}{9} = \frac{5}{9}$

Ta cũng có thể tính

$$DX = \sum_i \sum_j p_{ij}(x_i - EX)^2 = \sum_i (x_i - EX)^2 \cdot \sum_j p_{ij}$$

$$= \left(-\frac{4}{3}\right)^2 \cdot \frac{1}{18} + \left(-\frac{1}{3}\right)^2 \cdot \frac{1}{9} + \left(\frac{2}{3}\right)^2 \cdot \frac{1}{6} + \left(-\frac{4}{3}\right)^2 \cdot \frac{1}{12} + \left(-\frac{1}{3}\right) \cdot \frac{1}{6}$$

$$+ \left(\frac{2}{3}\right)^2 \cdot \frac{1}{4} + \left(-\frac{4}{3}\right)^2 \cdot \frac{1}{36} + \left(-\frac{1}{3}\right)^2 \cdot \frac{1}{18} + \left(\frac{2}{3}\right)^2 \cdot \frac{1}{12}$$

$$= \frac{16}{9}\left(\frac{1}{18} + \frac{1}{12} + \frac{1}{36}\right) + \frac{1}{9}\left(\frac{1}{9} + \frac{1}{6} + \frac{1}{18}\right) + \frac{1}{9}\left(\frac{1}{6} + \frac{1}{4} + \frac{1}{12}\right)$$

$$= \frac{16}{9}\cdot\frac{1}{6} + \frac{1}{9}\cdot\frac{1}{3} + \frac{1}{9}\cdot\frac{1}{2} = \frac{5}{9}$$

Tương tự $\lambda_{22} = DY = \dfrac{17}{36}$

Ta tính

$$\lambda_{12} = E\{(X - EX)(Y - EY)\} = \sum_i \sum_j (x_i - EX)(y_j - EY)p_{ij}$$

$$= \left(-\frac{4}{3}\right)\left(-\frac{5}{6}\right)\frac{1}{18} + \left(-\frac{4}{3}\right)\cdot\frac{1}{6}\cdot\frac{1}{12} + \left(-\frac{4}{3}\right)\cdot\frac{7}{6}\cdot\frac{1}{36} + \ldots$$

$$+ \frac{2}{3}\left(-\frac{5}{6}\right)\frac{1}{6} + \frac{2}{3}\cdot\frac{1}{6}\cdot\frac{1}{4} + \frac{2}{3}\cdot\frac{7}{6}\cdot\frac{1}{12} = 0$$

Để tiện tính λ_{ij}, thay cho bảng phân phối xác suất của (X, Y) ta lập bảng phân phối của vectơ (X - EX, Y - EY):

$X - \frac{7}{3}$ \ $Y - \frac{11}{6}$	$-\frac{5}{6}$	$\frac{1}{6}$	$\frac{7}{6}$
$-\frac{4}{3}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{36}$
$-\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{6}$	$\frac{1}{18}$
$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$

Ma trận covariance của vectơ (X, Y) sẽ là

$$\Lambda = \begin{pmatrix} \frac{5}{9} & 0 \\ 0 & \frac{17}{36} \end{pmatrix}$$

c. Hệ số tương quan ρ cần tìm là: $\rho = \dfrac{0}{\sqrt{\dfrac{5}{9}\cdot\dfrac{17}{36}}} = 0$

§ 5. LUẬT SỐ LỚN

Để trình bày luật số lớn chúng ta cần biết được hai dạng hội tụ cơ bản trong lý thuyết xác suất: hội tụ theo xác suất và hội tụ hầu chắc chắn. Ở giáo trình này chúng ta không đề cập đến các dạng hội tụ của dãy đại lượng ngẫu nhiên, do đó không thể đi sâu vào luật số lớn. Dưới đây chúng ta nêu vài kết quả đơn giản của luật số lớn hay được dùng đến ngay ở trong phần sau của giáo trình này.

1. Luật số lớn dạng Tchebychev:

Nếu dãy các đại lượng ngẫu nhiên $\xi_1, \dots, \xi_n, \dots$ độc lập, có phương sai giới nội đều ($D\xi_n \leq C$, $\forall n$) thì ta có

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - \frac{1}{n}\sum_{k=1}^{n}E\xi_k\right| > \varepsilon\right\} = 0 \ , \ \forall\varepsilon > 0$$

2. Luật số lớn dạng Khinchin:

Nếu dãy các biến ngẫu nhiên $\xi_1, \dots, \xi_n, \dots$ độc lập cùng phân phối có kỳ vọng hữu hạn thì ta có

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - E\xi_k\right| > \varepsilon\right\} = 0 \ , \ \forall\varepsilon > 0$$

Nói cách khác $\frac{1}{n}\sum_{k=1}^{n}\xi_k$ hội tụ tới $E\xi_k$ (theo xác suất).

Áp dụng 2 kết quả trên cho dãy ξ_k là dãy các phép thử Bernoulli ta nhận được định lý Bernoulli:

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k = \frac{m}{n} \text{ hội tụ tới } p \text{ (theo xác suất)}$$

khi n tăng ra vô hạn (nghĩa là tần suất hội tụ tới xác suất khi n tăng ra vô hạn).

3. Luật số lớn dạng Borel:

Xét n phép thử Bernoulli, trong đó $P(\xi_k = 1) = p$; $P(\xi_k = 0) = 1 - p$. Khi đó ta có:

$$P\left\{ \left(\frac{1}{n} \sum_{k=1}^{n} \xi_k - p\right) \to 0 \right\} = 1 \text{ khi } n \to \infty.$$

Tức là $\frac{1}{n} \sum_{k=1}^{n} \xi_k = \frac{m}{n}$ hội tụ tới p với xác suất 1.

(Luật số lớn xét theo sự hội tụ với xác suất 1 gọi là luật mạnh số lớn. Vì vậy luật số lớn dạng Borel gọi là luật mạnh số lớn).

§ 6. ỨNG DỤNG CỦA ĐỊNH LÝ GIỚI HẠN

Trong mục này chúng ta đề cập đến một vài kết quả của định lý giới hạn mà trong giáo trình thống kê toán học sẽ cần dùng đến, đồng thời dùng để tính xấp xỉ xác suất (10) chương I. Xác suất $P_n(m; p)$ đã được chỉ ra, nhưng với m, n lớn việc tính giá trị cụ thể của $P_n(m, p)$ lại là vấn đề khó khăn. Song lý thuyết đã chỉ ra rằng $P_n(m, p)$ sẽ dao động quanh giá trị giới hạn nào đó.

1. Định lý giới hạn địa phương Moivres - Laplace:

Xét n phép thử Bernoulli với $P(A) = p$, $(0 < p < 1)$, khi đó ta có:

$$\left| \frac{P_n(m, p)}{\varphi_0(x_m)} - 1 \right| < \frac{C_1}{n} + \frac{C_2 |x_m|^3}{\sqrt{n}}$$

trong đó: $x_m = \dfrac{m - np}{\sqrt{npq}}$

$$\varphi_0(x_m) = \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x_m^2}{2}} = \frac{1}{\sqrt{npq}} \varphi(x_m)$$

$q = 1 - p$, C_1, C_2 là các hằng số dương,

$\varphi(x)$ mật độ $N(0, 1)$ (Xem bảng VIII, phụ lục).

$$\frac{x_m^3}{\sqrt{n}} \to 0 \text{ khi } n \to \infty$$

Từ định lý ta có:

$$P_n(m, p) \approx \varphi_0(x_m) = \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x_m^2}{2}} = \frac{1}{\sqrt{npq}} \varphi(x_m) \qquad (13)$$

2. Định lý giới hạn trung tâm:

Nếu các biến ngẫu nhiên $\xi_1, \xi_2, \ldots, \xi_n \ldots$ độc lập cùng phân phối với phương sai hữu hạn khác 0 thì

$$\lim_{x \to \infty} P\left\{ \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (\xi_k - \mu) < x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt$$

đều $\forall x$, trong đó $E\xi_k = \mu$, $D\xi_k = \sigma^2$

Nói cách khác biến ngẫu nhiên $\dfrac{\sum_{k=1}^{n} \xi_k - E(\sum_{k=1}^{n} \xi_k)}{\sqrt{D(\sum_{k=1}^{n} \xi_k)}}$

có phân phối giới hạn là phân phối $N(0, 1)$.

Áp dụng kết quả trên cho dãy phép thử Bernoulli:

$$\xi_k = \begin{cases} 1 & \text{với xác suất } p \\ 0 & \text{với xác suất } 1-p \end{cases}$$

$\sum_{k=1}^{n} \xi_k$ chính là số lần xuất hiện biến cố A trong n phép thử Bernoulli. Thay cho $\sum_{k=1}^{n} \xi_k$ ta viết m, ta có:

$$\lim_{x \to \infty} P\left\{\frac{m - np}{\sqrt{npq}} < x\right\} = \Phi(x) \text{ đều với mọi } x.$$

Nghĩa là biến ngẫu nhiên nhị thức có phân phối giới hạn là phân phối chuẩn.

Từ kết quả này ta lại có:

$$P\left\{a < \frac{m - np}{\sqrt{npq}} < b\right\} \approx \Phi(b) - \Phi(a)$$

hoặc $P\{m_0 < m < m_1\} = P\left\{\frac{m_0 - np}{\sqrt{npq}} < \frac{m - np}{\sqrt{npq}} < \frac{m_1 - np}{\sqrt{npq}}\right\}$

$$\approx \Phi\left(\frac{m_1 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{m_0 - np}{\sqrt{npq}}\right) \qquad (14)$$

Ví dụ 27: Gieo 3200 lần một đồng tiền cân đối và đồng chất. Gọi ξ là số lần xuất hiện mặt sấp trong 3200 lần gieo đó.

a. Tìm số lần xuất hiện mặt sấp có khả năng nhất. Tính xác suất tương ứng.

b. Tìm xác suất sao cho giá trị của ξ nằm trong khoảng $(1600 + 5\sqrt{2}\ ;\ 1600 + 10\sqrt{2})$ (xem [1]).

Ta có: $n = 3200,\ p = 0,5$

$np + p - 1 = 3200.0,5 + 0,5 - 1 = 1599,5$

Vậy số lần xuất hiện mặt sấp có khả năng nhất là 1600. Theo (10) chương I ta có:

$$P_{3200}(1600\,;\ 0{,}5) = C_{3200}^{1600} \cdot 0{,}5^{3200}$$

Theo (13) ta có: $x_m = \dfrac{1600 - 3200 \cdot 0{,}5}{\sqrt{3200 \cdot 0{,}5 \cdot 0{,}5}} = 0$

$$P_{3200}(1600\,;\ 0{,}5) = \frac{1}{\sqrt{3200 \cdot 0{,}5 \cdot 0{,}5}}\,\varphi(0) \approx \frac{0{,}39894}{28{,}28} \approx 0{,}014$$

Theo (4) ta có:

$$P\{1600 + 5\sqrt{2} < \xi < 1600 + 10\sqrt{2}\}$$

$$\approx \Phi\left(\frac{1600 + 10\sqrt{2} - 1600}{20\sqrt{2}}\right) - \Phi\left(\frac{1600 + 5\sqrt{2} - 1600}{20\sqrt{2}}\right)$$

$$\approx \Phi(0{,}5) - \Phi(0{,}25) \approx 0{,}691462 - 0{,}598706 \approx 0{,}092756$$

3. Định lý Poisson:

Nếu npq nhỏ thì việc xấp xỉ phân phối nhị thức bởi phân phối chuẩn như trên không được tốt, do đó khi npq nhỏ người ta thường xấp xỉ phân phối nhị thức bởi phân phối Poisson thể hiện qua định lý Poisson dưới đây:

Giả sử với mỗi n = 1, 2, ... dãy $\left\{\xi_{nk}, k = 1, 2, \ldots, n\right\}$ là dãy đại lượng ngẫu nhiên độc lập, phân phối như nhau với

$$P(\xi_{nk} = 1) = p_n\,,\ P(\xi_{nk} = 0) = 1 - p_n\ ,\ \forall k = 1, 2, \ldots n$$

Ký hiệu $\mu_n = \sum\limits_{k=1}^{n} \xi_{nk}$. Rõ ràng $\mu_n \simeq B(n, p_n)$.

Giả sử $\lim np_n = \lambda$. Khi đó ta có

$$\lim_{n\to\infty} P\left\{\mu_n = m\right\} = e^{-\lambda}\frac{\lambda^m}{m!}\ ,\ \forall m = 0, 1, 2, \ldots$$

Nghĩa là với n đủ lớn phân phối nhị thức $B(n, p_n)$ được xấp xỉ bởi phân phối Poisson với $\lambda_n = np_n$.

$$P\left\{\mu_n = m\right\} \approx e^{-\lambda_n} . \frac{\lambda_n^m}{m!}, \quad \forall m = 0, 1, 2, \ldots \tag{15}$$

Ví dụ 28: Trong số 500 trang của một cuốn sách có 10 lỗi in. Tìm xác suất sao cho khi lấy hú họa (lấy ngẫu nhiên) một trang sách thì có:

a. Đúng 2 lỗi in

b. Không ít hơn 2 lỗi in (xem [1])

a. Gọi X là số lỗi trên một trang sách in. Vì xác suất p_n để một chữ bị lỗi là rất nhỏ và số chữ n trong một trang sách là lớn cho nên ta xấp xỉ phân phối của X bởi phân phối Poisson với tham số $\lambda_n = np_n$ = số lỗi trung bình trên một trang sách,

$\lambda_n = \frac{10}{500} = 0{,}02.$

$$P(X = 2) = e^{-0{,}02} . \frac{(0{,}02)^2}{2!} = e^{-0{,}02} \frac{0{,}0004}{2} = 0{,}0002 . e^{-0{,}02}$$

b. $$P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - e^{-0{,}02} - e^{-0{,}02} . 0{,}02 = 1 - 1{,}02\, e^{-0{,}02} \approx 0$$

BÀI TẬP CHƯƠNG II

1. Gieo một con xúc xắc cân đối và đồng chất. Gọi X là số chấm ở mặt trên của con xúc xắc.

a. Lập bảng phân phối xác suất của X.

b. Viết biểu thức hàm phân phối. Vẽ đồ thị của nó.

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

a. Lập bảng phân phối xác suất của Y.

b. Viết biểu thức hàm phân phối của Y.

3. Trong một cái bát có để 5 hạt đậu trong đó có 2 hạt đỏ. Lấy ngẫu nhiên ra 2 hạt. Gọi X là số hạt đậu đỏ được lấy ra.

a. Lập bảng phân phối xác suất của X.

b. Viết biểu thức hàm phân phối của X.

c. Tính $P\{0 < X < 2\}$ bằng cách tính trực tiếp và bằng cách thông qua hàm phân phối.

4. 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. Tìm phân phối xác suất của X.

b. Viết biểu thức hàm phân phối của X.

c. Tính $P\{0 < X \leq 4\}$ theo 2 cách.

5. Một xạ thủ dùng 5 viên đạn để thử súng. Anh ta bắn từng viên vào bia với xác suất trúng tâm là 0,95. Nếu có 2 viên liên tiếp trúng tâm thì thôi không bắn nữa. Gọi X là số đạn còn thừa ra.

a. Lập bảng phân phối xác suất của X.

b. Viết biểu thức hàm phân phối của X.

6. Một xạ thủ đem 6 viên đạn để bắn kiểm tra trước ngày thi bắn. Anh ta bắn từng viên vào bia với xác suất trúng vòng 10 là 0,85. Nếu bắn được 3 viên liên tiếp trúng vòng 10 thì thôi không bắn nữa. Gọi X là số đạn anh ta đã bắn.

a. Lập bảng phân phối xác suất của X

b. Viết biểu thức hàm phân phối của X

c. Xét trường hợp anh ta bắn được 3 viên trúng vòng 10 thì ngừng bắn. Gọi Y là số đạn còn lại. Tìm quy luật phân phối của Y.

7. Cho 2 biến ngẫu nhiên X và Y độc lập với các phân phối xác suất như sau:

X	-1	0	1	2
P	0.2	0.3	0.3	0.2

Y	-1	0	1
P	0.3	0.4	0.3

Lập bảng phân phối xác suất của X^2, $X + Y$, $2Y$, $X - 2Y$ và XY.

8. Ta có 2 hộp bi, hộp I có 3 bi trắng và 1 bi đỏ. Lấy ngẫu nhiên từ hộp I ra 2 viên bi bỏ vào hộp II đã có sẵn 2 bi trắng và 2 bi đỏ. Sau đó lại lấy ngẫu nhiên từ hộp II ra 2 viên bỏ vào hộp I. Gọi X và Y là số bi trắng ở hộp I và hộp II sau 2 lần chuyển bi như trên. Lập bảng phân phối xác suất của X và Y.

9. Cho biến ngẫu nhiên Cauchy với hàm phân phối

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\,\text{arctg}x$$

a. Tìm xác suất của biến cố: $\{0 < X < 1\}$

b. Tìm hàm mật độ của X (xem [1])

10. Biến ngẫu nhiên Weibull có hàm phân phối $F(x) = 1 - e^{\frac{-x^m}{x_0}}$ nếu $x > 0$ (x_0 là hằng số dương).

a. Tìm hàm mật độ tương ứng

b. Tìm phân vị cấp p.

c. Tìm Mod (xem [1])

11. Xác định hằng số a để hàm $p(x) = ae^{-x^2}$ là hàm mật độ của biến ngẫu nhiên nào đó. Tìm xác suất để cho giá trị của biến ngẫu nhiên tương ứng trên thuộc khoảng $(-\infty, 0)$ (xem [1]).

23. Tỷ lệ mắc bệnh viêm gan truyền nhiễm ở một vùng dân cư (X‰) là biến ngẫu nhiên có hàm mật độ:

$$p(x) = \frac{1}{20} \text{ nếu } 15‰ < x < 35‰$$

a. Tính tỷ lệ mắc bệnh trung bình

b. Tính $P\{|X - 20| > 5\}$

24. Mỗi người góp vào x nghìn đồng, tham gia trò chơi như sau: Gieo đồng thời 2 con xúc xắc cân đối, đồng chất, nếu được 2 mặt lục thì được nhận 14 nghìn; một mặt lục thì được nhận 4 nghìn.

a. Hỏi x là bao nhiêu để về trung bình trò chơi là vô thưởng vô phạt (không lỗ, không lãi).

b. Cần tối thiểu bao nhiêu người tham gia chơi để tổng số tiền góp vào ≥ 14 nghìn đồng?

25. Tiêm một loại vacxin chống toi gà. Khả năng miễn dịch là 80%. Một tổ kiểm tra bắt ngẫu nhiên ra từng con cho đến khi nào gặp con không miễn dịch thì thôi.

a. Mô tả phân phối xác suất của biến ngẫu nhiên chỉ số gà mà tổ kiểm tra đã bắt ra.

b. Về trung bình tổ phải bắt ra bao nhiêu con gà.

c. Tính xác suất phải bắt không quá 3 con gà.

26. Gọi X là trọng lượng một bao phân bón được đóng gói tự động. $X \simeq N(10; 0{,}05^2)$ (Xkg)

a. Có thể nói những gì về trọng lượng của một bao từ giả thiết đã cho.

b. Tìm tỷ lệ các bao phân bón có trọng lượng sai lệch so với trọng lượng quy định (10kg) không quá 100gam.

c. Tìm xác suất để khi chọn ngẫu nhiên một bao gặp bao có trọng lượng trên 10,1kg.

d. Nếu một máy khác đóng gói tự động cho trọng lượng Y của bao là biến ngẫu nhiên $N(10; 0{,}1^2)$, thì ta có thể nói gì về 2 máy đóng gói? Nói gì về 2 lô hàng do 2 máy đóng gói.

e. Nếu chọn ngẫu nhiên ra 100 bao từ lô hàng do máy đầu đóng gói thì có thể nói gì về tần số xuất hiện của các bao.

f. Mô tả phân phối xác suất của số bao có trọng lượng X thỏa mãn: $|X - 10| < 0{,}1$ trong 100 bao lấy ra.

g. Về trung bình trong 100 bao lấy ra có bao nhiêu bao có trọng lượng từ 9,9kg đến 10,1kg?

Con số đó có phải là số có khả năng nhất không?

27. Gọi X là số người tới trạm điện thoại trung tâm của Quận trong thời gian 5 phút. Giả sử X tuân theo luật phân phối Poisson với $\lambda = 3$. Cho biết $e^{-3} \approx 0{,}05$.

a. Nói rằng trong 5 phút có 3 người tới trạm điện thoại có đúng không?

b. Tính xác suất để trong 5 phút có không quá 3 người, có từ 2 đến 5 người, có trên 5 người tới trạm điện thoại?

28. Gọi Y là thời gian nói chuyện điện thoại của khách (Y phút). Giả sử Y tuân theo luật phân phối mũ với $\lambda = \frac{1}{3}$.

a. Cho biết thời gian trung bình khách nói chuyện điện thoại là bao nhiêu phút.

b. Theo luật trên hãy tính tỷ lệ khách hàng nói chuyện điện thoại dưới 3 phút, từ 3 phút đến 10 phút, trên 10 phút.

(Cho biết $e^{-1} \approx 0{,}37$; $e^{-\frac{10}{3}} \approx 0{,}037$)

29. Một công ty xây dựng cần mua 10000 gạch men của xưởng gạch men Thanh Tân. Theo hợp đồng, khi nhập hàng công ty kiểm tra ngẫu nhiên 100 viên, nếu thấy bị không quá 3 viên thứ phẩm thì chấp nhận lô hàng, nếu thấy có trên 3 viên thứ phẩm thì trả lô hàng. Xưởng gạch đem giao lô hàng với tỷ lệ thứ phẩm là 3%.

a. Tìm khả năng xưởng gạch gặp may trong việc giao lô hàng này.

b. Nếu kiểm tra 100 viên thấy không có viên nào thứ phẩm, công ty sẽ thanh toán mỗi viên 1500đ (lô hàng được xếp loại A); nếu thấy 1 viên thứ phẩm, lô hàng xếp loại B với 3% số viên chỉ được thanh toán 1000đ/1viên; nếu thấy 2 hoặc 3 viên thứ phẩm lô hàng xếp loại C với 10% số viên chỉ được thanh toán 1000đ/1viên. Chi phí vận chuyển 1 lượt là 500.000đ.

Nếu bạn là người giao lô hàng, bạn hãy đoán xem khả năng nào xảy ra đối với lô hàng của bạn. Bạn có chấp nhận kết luận đó không?

30. Cho bảng phân phối xác suất của vectơ ngẫu nhiên (X, Y) như sau:

X \ Y	1	2	3
1	0,17	0,13	0,25
2	0,10	0,30	0,05

a. Lập ma trận covariance

b. Tìm hệ số tương quan ρ

31. Cho bảng xác định luật phân phối của vectơ ngẫu nhiên (X, Y) như sau:

X \ Y	20	40	60
10	3λ	λ	0
20	2λ	4λ	2λ
30	λ	2λ	5λ

Xác định:

a. Tham số λ

b. EX, EY

c. Ma trận covariance

d. Hệ số tương quan ρ.

32. Một máy tính điện tử gồm 10000 bóng bán dẫn, trong đó có 1000 bóng loại I với khả năng bị hỏng của mỗi bóng là 0,0005; 3000 bóng loại II với khả năng bị hỏng của mỗi bóng là 0,0003; 6000 bóng loại III với khả năng bị hỏng của mỗi bóng là 0,0001. Máy tính sẽ ngừng làm việc nếu có ít nhất một bóng bị hỏng.

Giả sử rằng các bóng bị hỏng là độc lập với nhau.

a. Tìm xác suất để có ít nhất một bóng loại i (i = 1, 2, 3) bị hỏng.

b. Tìm khả năng máy tính ngừng làm việc.

HƯỚNG DẪN VÀ ĐÁP SỐ

Chương I

1. a. $\dfrac{C_{10}^{5}\ C_{90}^{15}}{C_{100}^{20}}$ b. $\dfrac{C_{10}^{10}\ C_{90}^{10}}{C_{100}^{20}}$ c. 0

2. $\dfrac{C_{25}^{4}\ C_{45}^{6}}{C_{70}^{10}}$

3. a. $\dfrac{12!}{4!\,5!\,3!\,3^{12}}$ b. $\dfrac{12!}{(4!)^3\,3^{12}}$ c. $\dfrac{1}{3}$

4. a. $\dfrac{1}{9^5}$ b. $\dfrac{1}{9^4}$ c. $\dfrac{A_9^5}{9^5}$ d. $\dfrac{C_5^2\ A_9^4}{9^5}$

5. $\dfrac{4}{5!}$

6. a. $\dfrac{5}{36}$ b. $\dfrac{2}{9}$ c. $\dfrac{1}{6}$

7. $\dfrac{1}{5}$

8. $\dfrac{2}{5}$

9. a. $\dfrac{C_2^1\,C_4^1\,C_4^2\,C_{10}^1}{C_{32}^5}$ b. $\dfrac{C_5^4\,C_{27}^1}{C_{32}^5}$ c. $\dfrac{2C_5^5}{C_{32}^5}$

d. $\dfrac{2(C_2^1)^3 C_{29}^2}{C_{32}^5}$ e. $\dfrac{C_{16}^3 C_{16}^2}{C_{32}^5}$

10. a. $10^3 \times 24^2$ b. $\dfrac{1}{10^3}$ c. $\dfrac{1}{10^2}$

d. $\dfrac{5}{10 \times 24^2}$

11. $\dfrac{2}{3}$

12. a. $\dfrac{C_{20}^3 C_{10}^2}{C_{30}^5}$ b. $\dfrac{C_{20}^5 + C_{10}^5}{C_{30}^5}$

13. $\dfrac{C_{12}^4}{C_{15}^4}$

14. a. $A_1 A_2 A_3 \dots A_{10}$

b. $\overline{A_1 A_2 \dots A_{10}}$

c. $\overline{A}_1 A_2 \overline{A}_3 A_4 \overline{A}_5 A_6 \overline{A}_7 A_8 \overline{A}_9 A_{10}$

d. $\overline{A}_1 A_2 \dots A_{10} \cup A_1 \overline{A}_2 A_3 \dots A_{10} \cup A_1 \dots A_9 \overline{A}_{10}$

e. $A_1 \dots \overline{A}_i \dots \overline{A}_j \dots A_{10}$ có C_{10}^2 biến cố dạng trên.

15. a. $A_1 \overline{A}_2 \overline{A}_3$

b. $A_1 \cup A_2 \cup A_3$

c. $A_1 A_2 A_3$

d. $A_1 A_2 \overline{A}_3$

e. $A_1 \overline{A}_2 \overline{A}_3 \cup \overline{A}_1 A_2 \overline{A}_3 \cup \overline{A}_1 \overline{A}_2 A_3$

g. $A_1 A_2 \overline{A}_3 \cup A_1 \overline{A}_2 A_3 \cup \overline{A}_1 A_2 A_3$

h. $A_1 A_2 \cup A_1 A_3 \cup A_2 A_3$

i. $\overline{A}_1 \overline{A}_2 \overline{A}_3$

k. $\overline{A}_1 \cup \overline{A}_2 \cup \overline{A}_3$

16. a. NN ... NS (Gieo k+1 lần, k lần là N)

b. $\frac{1}{2^3} = \frac{1}{8}$

17. 0,765

18. 0,003; 0,997; 0,612; 0,997

19. 0,388

20. a. 0,784 b. Phát 5 lần

21. Dùng các kết quả của bài 15 và công thức cộng, nhân xác suất.

Lưu ý rằng 3 biến cố A_i = {người thứ i bắn trúng mục tiêu} (i = 1, 2, 3) là độc lập và không xung khắc.

22. 0,976

23. a. $\frac{5}{8}$ b. P{Bố mẹ cùng gen XT} = 0

P{Bố mẹ cùng gen XX} = $\frac{1}{3}$

(cố 4 khả năng của gen bố mẹ: (XT, XT), (XT, XX), (XX, XT) và (XX, XX). Ba trường hợp sau thỏ con đều là thỏ xám: XX hoặc XT. Trường hợp đầu thỏ con có thể là XX, XT, TT).

24. a. 0,59 b. 0,20

25. B_i = {Bi được rút từ hộp loại i}, i = 1, 2, 3 là nhóm đầy đủ.

a. Dùng công thức xác suất đầy đủ (= 0,644).

b. Dùng công thức Bayes (= 0,337).

26. A_i = {Sản phẩm từ lô i là loại I}, i = 1, 2

$B_1 = A_1 A_2$ $B_2 = A_1 \overline{A}_2$

$B_3 = \overline{A}_1 A_2$ $B_4 = \overline{A}_1 \overline{A}_2$

B_1, B_2, B_3, B_4 là nhóm đầy đủ.

Sau đó dùng công thức xác suất đầy đủ.

27. Gọi B = {Con gà từ lô I lạc sang lô II là gà mái}

B, $\overline{B}$ là nhóm đầy đủ.

Dùng công thức xác suất đầy đủ.

28. A = {Trẻ sinh đôi cùng giới}

B = {Trẻ sinh đôi là sinh đôi thật}

$\overline{B}$ = {Trẻ sinh đôi là sinh đôi giả}

B, $\overline{B}$ là nhóm đầy đủ $P(A) = \frac{1}{2}(1 + p)$.

a. $P(B/A) = 2p/p + 1$.

b. $P(\overline{B}/\overline{A}) = 1$.

29. A = {Có phản ứng dương tính}

B = {Người bị bệnh}

B, $\overline{B}$ là nhóm đầy đủ.

a. Dùng công thức xác suất đầy đủ

b. Dùng công thức Bayes $P(B/A)$, $P(\overline{B}/A)$

c. $P(B/A)$

30. A = {Câu được cá sau 3 lần thả câu}

B_i = {Câu cá ở địa điểm thứ i}, i = 1, 2, 3

$$P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$$

$$P(A/B_1) = C_3^1 \times 0{,}6 \times 0{,}4^2$$

$$P(A/B_2) = C_3^1 \times 0{,}7 \times 0{,}3^2$$

$$P(A/B_3) = C_3^1 \times 0{,}8 \times 0{,}2^2$$

Dùng công thức Bayes

31. Đây là 10 phép thử Bernoulli, p = P(người bị lao) = 0,001

c. $P_{10}(m \geq 1; 0{,}001) = 1 - P_{10}(0; 0{,}001)$

d. $10 \times 0{,}999 + 0{,}999 - 1 = 9{,}989$

Số người không bị lao có khả năng nhất là 10.

32. Đây là phép thử Bernoulli với $n = 60$, $p = 0,92$

a. $P_{60}(60;\ 0,92)$

b. $P_{60}(2;\ 0,08)$

c. $1 - P_{60}(0;\ 0,08)$

d. 56 và $P_{60}(56;\ 0,92)$

33. Điều khẳng định đó không đúng

$$P_{10}(8\,;0,8) = C_{10}^{8} \times 0,8^{0,8} \times 0,2^{2}$$

34. $1-P_n(m = 0;\ 0,3) = 1 - 0,7^n \geqslant 0,8 \Rightarrow n = 5$

35. $P_4(2;\ 0,5) > P_6(3;\ 0,5)$

36. Gọi A_i = {Sản phẩm lấy ra lần thứ i là chính phẩm}

A_1, A_2 độc lập, không xung khắc

a. $A = A_1A_2 \quad \Rightarrow P(A) = (0,95)^2 = 0,9025$

b. $B = A_1 \cup A_2 \quad \Rightarrow P(B) = 0,9975$

c. $C = \overline{A}_1 A_2 \quad \Rightarrow P(C) = 0,0475$

d. $D = \overline{A}_1 A_2 \cup A_1 \overline{A}_2 \quad \Rightarrow P(D) = 0,0950$

37. a. 0,67 b. 0,70 c. Nên lấy từ lô I

38. Dùng công thức xác suất có điều kiện để tính

a. $P(\text{conMX}/\text{chaMX}) = \dfrac{P(\text{conMX}, \text{chaMX})}{P(\text{chaMX})} = \dfrac{0,782}{0,871} = 0,8978$

Vì

$P(\text{chaMX}) = P(\text{chaMX,conMĐ}) + P(\text{chaMX,conMX}) = 0,089 + 0,782 = 0,871$

b. 0,612

39. P (cùng giới) $= \dfrac{2}{3}$; $P(TG) = P(GT) = \dfrac{1}{2}$

(T: trai; G: gái; T_i = {Đứa thứ i là trai}

$$P\left({}^{T_2}/_{T_1}\right) = {}^{P(T_1T_2)}/_{P(T_1)} = \frac{2}{3}\cdot{}^{0,49}/_{0,4933} \approx 0,66$$

(Vì $P(T_1) = P(T_1T_2) + P(T_1G) = \frac{2}{3} . 0{,}49 + \frac{1}{3} . \frac{1}{2} = 0{,}4933$)

40. a. 0,2775 b. 0,9775 c. 0,006 d. 0,2712

41. a. Xác suất đầy đủ với B = {Gặp được nam}

B, $\overline{B}$: nhóm đầy đủ; xác suất là 0,47%

b. Khả năng người gặp là nữ cao hơn (nữ: 52%, nam 48%)

42. Công thức Bayes. Nên chẩn đoán bệnh A.

43. 2 lần áp dụng công thức xác suất đầy đủ, Bayes. Lần I sẽ cho ta xác suất để chính phẩm thuộc lô I và lô II tương ứng là 9/17 và 8/17. Lần II áp dụng công thức xác suất đầy đủ với xác suất lấy từ lô I là 9/17, xác suất lấy từ lô II là 8/17. Khi đó xác suất để lấy ra một phế phẩm sẽ bằng:

$$\frac{9}{17} . 0{,}1 + \frac{8}{17} . \frac{1}{5} = 0{,}147$$

Chương II

2. Ta lập bảng

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Từ bảng đó ta dễ dàng lập được bảng phân phối xác suất của Y.

Từ đó viết được biểu thức hàm phân phối.

3. a.

X	0	1	2
P	$\frac{C_3^2}{C_5^2}$	$\frac{C_3^1 C_2^1}{C_5^2}$	$\frac{C_2^2}{C_5^2}$

X	0	1	2
P	0,3	0,6	0,1

b. $$F(x) = \begin{cases} 0 & x \leqslant 0 \\ 0,3 & 0 < x \leqslant 1 \\ 0,9 & 1 < x \leqslant 2 \\ 1 & 2 < x \end{cases}$$

c. $$P\{0 < x < 2\} = F(2) - F(0) - P(x = 0)$$
$$= 0,9 - 0 - 0,3 = 0,6 = P(x = 1) = 0,6$$

4. a.

X	0	1	2	3
P(X = k)	0,42	0,425	0,14	0,015

c. 0,58

5. a.

X	0	1	2	3
P(X=k)	0,00725	0,045125	0,045125	0,9025

6. a.

X	3	4	5	6
P(X=k)	0,614125	0,0921187	0,013818	0,279938

c.

Y	0	1	2	3
P(Y=k)	0,02661	0,082908	0,276357	0,614125

7.

X^2	0	1	4
$P(X^2=k)$	0,3	0,5	0,2

X+Y	-2	-1	0	1	2	3
P	0,06	0,17	0,27	0,27	0,17	0,06

2Y	-2	0	2
P	0,3	0,4	0,3

X-2Y	-3	-2	-1	0	1	2	3	4
P	0,06	0,09	0,17	0,18	0,18	0,17	0,09	0,06

XY	-2	-1	0	1	2
P	0,06	0,15	0,58	0,15	0,06

8.

X	1	2	3	4
P(X=k)	$\frac{1}{30}$	$\frac{11}{30}$	$\frac{15}{30}$	$\frac{3}{30}$

Y	1	2	3	4
P(Y=k)	$\frac{3}{30}$	$\frac{15}{30}$	$\frac{11}{30}$	$\frac{1}{30}$

9. a. 0,25 b. $p(x) = \dfrac{1}{\pi(1 + x^2)}$

10. a. $p(x) = \dfrac{m}{x_o} x^{m-1} e^{\frac{-x^m}{x_o}}$ nếu $x \geqslant 0$

b. $F(x_p) = p$ tức là $1 - e^{\frac{-x_p^m}{x_o}} = p \Rightarrow x_p = \sqrt[m]{-x_o \ln(1 - p)}$

c. $p'(x) = 0 \Rightarrow x = \sqrt[m]{\frac{m-1}{m} x_o} = \text{Mode}$.

11. $a = \frac{1}{\sqrt{\pi}}$, $\frac{1}{2}$

12. a. $C = \frac{1}{\pi}$

b. $F(x) = \begin{cases} 0 & x \leqslant -2 \\ \frac{1}{\pi}\arcsin\frac{x}{2} + \frac{1}{2} & -2 < x \leqslant 2 \\ 1 & 2 < x \end{cases}$

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

13. a. $F(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{a}^{\frac{\lg x - \lg x_o}{\sigma}} e^{\frac{-x^2}{2}} dx$

b. Mod $= 10^{-5C}$

14. $P_{100}(60 ;\ 0{,}5) = C_{100}^{60} \frac{1}{2^{100}}$

15. a. $C = \frac{1}{2}$ b. $F(x) = \begin{cases} 0 & x \leqslant -\frac{\pi}{2} \\ \frac{1}{2} + \frac{1}{2}\sin x & \frac{-\pi}{2} < x \leqslant \frac{\pi}{2} \\ 1 & \frac{\pi}{2} < x \end{cases}$

c. $\frac{\sqrt{2}}{4}$ d. 2 ; $P_5(2 , \frac{\sqrt{2}}{4}) = C_5^2 \frac{2}{16}(1 - \frac{\sqrt{2}}{4})^3$

16. 3 trường hợp đầu áp dụng công thức trường hợp rời rạc. Trường hợp bài 9: Không tồn tại kỳ vọng, phương sai.

2 trường hợp cuối áp dụng công thức trong trường hợp liên tục.

17. a. $p(x) = \dfrac{\alpha x_0^\alpha}{x^{\alpha+1}}$ nếu $x \geqslant x_0$ và $= 0$ nếu $x < x_0$

b. Nếu $\alpha > 1$ $E\xi = \dfrac{\alpha}{\alpha - 1} x_0'$; $\alpha \leqslant 1$ không tồn tại $E\xi$.

18. a. X - Số phế phẩm trong ca. Đó là biến ngẫu nhiên nhị thức với $n = 100$, $p = 0,02$.

b. $EX = np = 2$ $\quad$ $P_{100}(2 ; 0,02) = C_{100}^2 . 0,02^2 . 0,98^{98}$

19. $P\{158 < X < 162\} = \Phi(\frac{1}{3}) - \Phi(-\frac{1}{3}) \approx 0,26$

$$P_4(m \geqslant 1; 0,26) = 1 - 0,74^4$$

20. a. $EX = 2$; $EY = 5$. Áp dụng tính chất của kỳ vọng

b. $9DX + 4DY$, trong đó $DX = 0,09$; $DY = 25$

c. $EX^2 = DX + (EX)^2$, $EY^2 = DY + (EY)^2$

21. Dùng công thức xấp xỉ (13)

22. Dùng công thức xấp xỉ (14)

23. a. $EX = \dfrac{0,035 + 0,015}{2} = 0,020 = 20‰$

b. $P\{|X - 20| > 5\} = P\{(X < 15) \cup (X > 25)\}$

$$= P\{X < 15\} + P\{X > 25\} = 0 + \frac{35 - 25}{35 - 15} = 0,50$$

24. a. Để trò chơi là vô thưởng vô phạt cần thỏa mãn $EX = x$ trong đó X là biến ngẫu nhiên chỉ phần thắng thu được.

$$EX = 14 . \frac{1}{6^2} + 4 . \frac{2.5}{6^2} + 0 . \frac{25}{6^2} = 1,5 \text{ nghìn đồng}$$

Vậy mỗi người cần góp x = 1500 đồng.

b. Số người tham gia chơi cần $\geq \frac{14}{1,5} \approx 9,33$, tức là tối thiểu cần 10 người tham gia.

25. a. $P(X = m) = 0,8^{m-1}.0,2 \qquad m = 1, 2, 3...$

b. $EX = \frac{1}{0,2} = 5$

c. $P(X \leq 3) = \sum_{m=1}^{3} P(X = m) = 0,2 + 0,8.0,2 + 0,8^2.0,2$

$= 0,488$

26. a. Không thể nói trọng lượng của mỗi bao đều là 10kg, mà chỉ có thể nói trọng lượng trung bình của các bao là 10kg, còn trọng lượng của một bao bất kỳ dao động quanh giá trị 10 với độ lệch chuẩn là 50gam.

b. Tỷ lệ cần tìm chính là

$P\{|X - 10| < 0,1\} = 2\Phi(2) - 1 \approx 0,9545$

c. $P\{X > 10,1\} = 1 - \Phi(2) \approx 0,0227$

d. Máy đầu cân đo chính xác hơn máy sau. Lô hàng do máy đầu đóng gói là đồng đều hơn, chênh lệch giữa bao nặng và bao nhẹ ít hơn so với lô hàng do máy sau đóng gói.

e. Tần số xuất hiện có tính tập trung với tính đối xứng (một cách tương đối) so với giá trị 10kg.

f. Đó là phân phối B(100; 0,9545).

g. Về trung bình có 100.0,9545 = 95,45 bao.

Số trên không phải là số có khả năng nhất (m_o nguyên và = 96)

27. a. Nói vậy không đúng. Trong 5 phút có thể không có ai, có thể có 1 người, 3 người, 5 người, 10 người... đến trạm điện thoại. Nhưng về trung bình có 3 người đến.

b. 0,65; 0,714; 0,08

28. a. 3 phút

b. 0,63; 0,333; 0,037

29. a. 0,598

b. Khả năng lô hàng xếp loại C là cao nhất (với xác suất 0,4318 và bị thiệt 350.000đ do bị tính oan 7% thứ phẩm). Chấp nhận giao lô hàng với sự xếp loại C, vì nếu đem về sẽ mất thêm 500.000đ vận chuyển nữa.

30. a. $\Lambda = \begin{pmatrix} 0,25 & -0,06 \\ -0,06 & 0,57 \end{pmatrix}$

b. $\rho = -0,159$

31. a. $\lambda = 1/20$ b. $EX = 22$, $EY = 41$

c. $\Lambda = \begin{pmatrix} 56 & 68 \\ 68 & 259 \end{pmatrix}$ d. $\rho = 0,565$

32. a. Gọi X_i là số bóng loại i (i = 1, 2, 3) bị hỏng. X_i xấp xỉ Poisson với tham số 0,5; 0,9 và 0,6. X_i độc lập.

$$P(X_1 \geqslant 1) = 1 - P(X_1 = 0) = 1 - e^{-0,5} \approx 0,394$$

Tương tự $P(X_2 \geqslant 1) \approx 0,594$ $P(X_3 \geqslant 1) = 0,452$

b. $P(X_1 + X_2 + X_3 \geqslant 1) = 1 - P(X_1 + X_2 + X_3 = 0)$

$$= 1 - P\{(X_1 = 0)(X_2 = 0)(X_3 = 0)\} = 1 - 0,606 . 0,406 . 0,548 \approx 0,865$$

TÀI LIỆU THAM KHẢO

1. Hoàng Hữu Như, Nguyễn Văn Hữu. *Bài tập Lý thuyết xác suất và Thống kê toán học*. NXB Đại học và THCN, Hà Nội 1976.

2. Nguyễn Hải Tuất. *Thống kê toán học trong lâm nghiệp*. NXB Nông nghiệp, Hà Nội 1982

3. Nguyễn Đình Cử, Trương Giêu. *Bài tập xác suất và Thống kê toán học*. Đại học Kinh tế quốc dân, 1992.

4. G. Lefort. *Bài tập Giải tích, Đại số và Xác suất*. Tập III. NXB Đại học và THCN, Hà Nội 1985.

5. Đặng Mai. *Các phương pháp toán trong địa chất*. Hà Nội 1990.

Phần II

THỐNG KÊ ỨNG DỤNG

Chương I

LÝ THUYẾT MẪU

§ 1. CÁC PHƯƠNG PHÁP LẤY MẪU ĐƠN GIẢN

Thông tin đầu tiên và nhiều khi cũng là thông tin duy nhất mà chúng ta dựa vào để nghiên cứu, phân tích chính là các kết quả quan sát có được, vì vậy các kết quả này phải đảm bảo tính chính xác, tính ngẫu nhiên của nó, phải là các đại diện một cách trung thực cho hiện tượng hoặc cho đại lượng mà chúng ta đang nghiên cứu.

Xuất phát từ thông tin sai lệch thì các kết luận nhận được sẽ phản ánh không đúng hiện tượng nghiên cứu, thậm chí còn làm cho ta nghi ngờ ngay cả tính hiệu quả của phương pháp ta sử dụng. Do vậy trước tiên ta quan tâm đến việc thu thập thông tin ban đầu.

Các quan sát độc lập hay các phép thử độc lập: Các quan sát (phép thử) được tiến hành một cách độc lập với nhau, kết quả của quan sát (phép thử) này không phụ thuộc vào kết quả của quan sát (phép thử) khác và cũng không ảnh hưởng đến khả năng xảy ra kết quả của quan sát (phép thử) khác.

Các phép thử lặp: Các phép thử được tiến hành trong các điều kiện hoàn toàn như nhau.

Lấy mẫu ngẫu nhiên có hoàn lại: Rút ngẫu nhiên từ một tập nào đó ra một phần tử. Ghi lại các số đặc trưng cần thiết từ phần tử đó, sau đó trả nó trở lại tập ban đầu trước khi rút tiếp ngẫu nhiên lần sau.

Lấy mẫu ngẫu nhiên không hoàn lại: Tương tự như trên, chỉ khác ở chỗ các phần tử được rút ra sẽ không được trả lại tập ban đầu.

Ví dụ 1: Lấy mẫu máu của 100 người, ta tiến hành xét nghiệm một số thông số (chẳng hạn lượng hồng cầu, bạch cầu, Glucoza. Cholestérol...) trong mỗi mẫu máu. Đó là 100 phép thử độc lập.

Một y tá dùng một loại thước đo chiều cao cho 50 thanh niên trong đợt khám tuyển nghĩa vụ quân sự. Đó là 50 quan sát độc lập.

Để xác định tỷ lệ người bị mắc bệnh A ở một vùng nào đó ta khám ngẫu nhiên cho một số người. Những người đã khám rồi sẽ loại khỏi danh sách những người đang chờ gọi khám. Đó là lấy mẫu không hoàn lại.

Thực ra vấn đề lấy mẫu là một vấn đề rất quan trọng và cũng rất phong phú trong nghiên cứu thống kê. W.G.Corchan đã trình bày vấn đề này trong cuốn sách Sampling Techniques dày hơn 400 trang. Ngoài các phương pháp lấy mẫu chung dùng trong khoa học thống kê, đối với mỗi lĩnh vực riêng biệt như nông nghiệp, lâm nghiệp, sinh học, địa chất, xã hội học... lại có những phương pháp lấy mẫu mang tính đặc thù của nó.

Để minh hoạ chúng ta xét phương pháp lấy mẫu trong nghiên cứu xã hội học mà có thể khái quát các phương pháp thu thập thông tin trong xã hội bằng sơ đồ sau:

b. Chọn mẫu với xác suất không đều

Trên các lĩnh vực xã hội sự không đồng đều về quy mô của hiện tượng là phổ biến, do đó việc thu thập thông tin dựa vào các mẫu đồng loạt sẽ hạn chế tính đại diện, chẳng hạn các hiện tượng xã hội, văn hóa, tinh thần không xuất hiện đồng đều theo thời gian mà thường rộ lên ở một số thời điểm nhất định. Chọn mẫu theo thời gian với xác suất không đều tùy theo quy mô xuất hiện các hiện tượng cho các mẫu đại diện tốt hơn các mẫu chọn theo xác suất đều.

(Hiện tượng mê tín dị đoan tập trung vào dịp lễ hội, vào nơi có lễ hội. Những người nhiễm HIV tập trung ở những gái mại dâm, người tiêm chích...).

Chẳng hạn việc chọn hộ gia đình ở thành phố có thể thực hiện theo hai cách:

- Chọn mẫu theo hệ thống quản lý hành chính gồm các cấp: quận - phường - tổ dân phố - hộ gia đình. Nếu việc phân chia các đơn vị hành chính tương đối đều về số dân thì mẫu nhiều cấp với xác suất đều là hợp lý.

- Chọn mẫu theo kết cấu tự nhiên của dân cư thành phố, gồm 3 cấp: các đường phố - các số nhà - các hộ gia đình (khu tập thể coi là một phố). Do quy mô của đơn vị mẫu ở mỗi cấp không đều nhau: số lượng số nhà của một đường phố, số lượng hộ gia đình của mỗi số nhà, cho nên hợp lý nhất là chọn mẫu với xác suất không đều tỷ lệ với các số lượng đó; chỉ có chọn đường phố là theo xác suất đều.

c. Điều tra nhóm trội

Điều tra nhóm trội là một dạng điều tra trọng điểm, nhưng với những điều kiện nhất định có khả năng suy rộng cho toàn bộ tổng thể. Chẳng hạn, nghiên cứu quỹ thời gian rỗi của nhóm người có điều kiện sống cao, có giao lưu văn hóa mạnh, có trình độ văn hóa nhất định... ta có thể suy rộng (có điều chỉnh) ra cho toàn bộ dân cư với khoảng thời gian đủ dài để nâng cao mức sống vật chất và tinh thần của toàn thể lên mức hiện tại của nhóm trội.

§ 2. MẪU NGẪU NHIÊN

Tiến hành n quan sát độc lập về biến ngẫu nhiên X nào đó. Ta gọi X_i là việc quan sát lần thứ i về biến ngẫu nhiên X. Khi đó $(X_1, X_2, \ldots, X_n)$ được gọi là mẫu ngẫu nhiên, n được gọi là cỡ mẫu hay số lần quan sát. Như vậy mẫu ngẫu nhiên cỡ n thực chất là n biến ngẫu nhiên độc lập, cùng phân phối như biến ngẫu nhiên X.

Ta gọi x_i là kết quả quan sát được ở lần thứ i. Khi đó $(x_1, x_2, \ldots, x_n)$ là n giá trị cụ thể ta quan sát được. Đó là một giá trị cụ thể mà mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ nhận.

Từ nay về sau khi nói rằng ta có một mẫu ngẫu nhiên cỡ n được rút ra từ biến ngẫu nhiên X, ta sẽ hiểu đó là n biến ngẫu nhiên độc lập cùng phân phối nếu ta không quan tâm đến kết quả cụ thể quan sát được mà ta muốn nghiên cứu các tính chất chung của mẫu, của các đặc trưng mẫu, còn ta sẽ hiểu đó là n giá trị cụ thể quan sát được nếu ta quan tâm đến kết quả cụ thể và ta cần những tính toán cụ thể.

Như vậy chúng ta có trong tay một mẫu ngẫu nhiên, dựa trên đó ta sẽ dùng các phương pháp và kết quả của thống kê để phân tích và rút ra những kết luận cần thiết.

§ 3. PHÂN PHỐI THỰC NGHIỆM

Giả sử ta có mẫu ngẫu nhiên $(x_1, x_2, \ldots, x_n)$. Xuất phát từ n giá trị cụ thể mà biến ngẫu nhiên nhận ta xây dựng hàm số:

$$F_n(x) = \frac{\#\{x_i < x\}}{n}$$

trong đó $\# \{x_i < x\}$ là số các giá trị mẫu x_i mà nhỏ hơn x. Khi x thay đổi, ta nhận được hàm $F_n(x)$ theo biến số thực x. Hàm số này được gọi là hàm phân phối thực nghiệm.

Xuất phát từ các mẫu cụ thể khác nhau ta nhận được các hàm phân phối thực nghiệm khác nhau. Đồ thị của chúng đều là các đường bậc thang. Các đường bậc thang khác nhau đều có chung một tính chất là: Khi cỡ mẫu tăng vô hạn các hàm phân phối thực nghiệm tiến đến hàm phân phôi lý thuyết cần tìm. Điều đó được thể hiện qua định lý sau:

Định lý Glivencô: Giả sử F(x) là hàm phân phối của biến ngẫu nhiên X mà ta đang cần tìm. $F_n(x)$ là hàm phân phối thực nghiệm nhận được từ mẫu ngẫu nhiên cỡ n. Khi đó

$$P\{ \sup_{-\infty < x < \infty} |F_n(x) - F(x)| \to 0 \text{ khi } n \to \infty \} = 1$$

Như vậy hàm phân phối thực nghiệm là một xấp xỉ của hàm phân phối lý thuyết. Xấp xỉ đó càng tốt khi cỡ mẫu n càng lớn. Với n cố định hàm phân phối thực nghiệm cho ta hình ảnh hình học về phân phối lý thuyết cần tìm.

§ 4. ĐA GIÁC TẦN SUẤT VÀ TỔ CHỨC ĐỒ

Giả sử biến ngẫu nhiên X rời rạc với phân phối xác suất $P(X = c_i) = p_i$. Trên hệ trục toạ độ ta đặt các điểm (c_i, p_i). Ta được tập các điểm trên mặt phẳng toạ độ. Đó là biểu đồ xác suất của biến ngẫu nhiên X mà ta cần tìm.

Ta có một mẫu ngẫu nhiên $(x_1, x_2, \ldots, x_n)$. Với n giá trị cụ thể đó ta có thể thu gọn bằng cách gộp các giá trị giống nhau lại và tính đến số lần nhận giá trị đó trong mẫu đã cho ta được:

$$\begin{cases} x_{(1)} < x_{(2)} < \dots < x_{(k)} \\ m_1 \quad m_2 \quad \dots \quad m_k \end{cases} \qquad \sum m_i = n$$

(m_i lần trong mẫu nhận giá trị $x_{(i)}$). Ở đây để cho tiện ta sắp xếp các giá trị $x_{(i)}$ theo thứ tự tăng dần. Nói chung tập $(x_{(1)}, x_{(2)}, \dots, x_{(k)})$ là tập con (có thể trùng) của tập giá trị $(c_1, c_2, \dots, c_i, \dots)$.

Trên hệ toạ độ ta đặt các điểm $(x_{(i)}; \frac{m_i}{n})$, $i = 1, \dots, k$. Tập các điểm đó được gọi là biểu đồ tần suất hoặc đa giác tần suất (nối các điểm lại ta được một đa giác).

Gọi $$p_i = P\{X = x_{(i)}\}$$

Theo luật số lớn ta có:

$$P\left\{\frac{m_i}{n} \to p_i\right\} = 1$$

nghĩa là khi n lớn tung độ của biểu đồ tần suất xấp xỉ tung độ của biểu đồ xác suất cần tìm.

Ví dụ 2: Kết quả của 25 phép thử độc lập về biến ngẫu nhiên X nào đấy như sau:

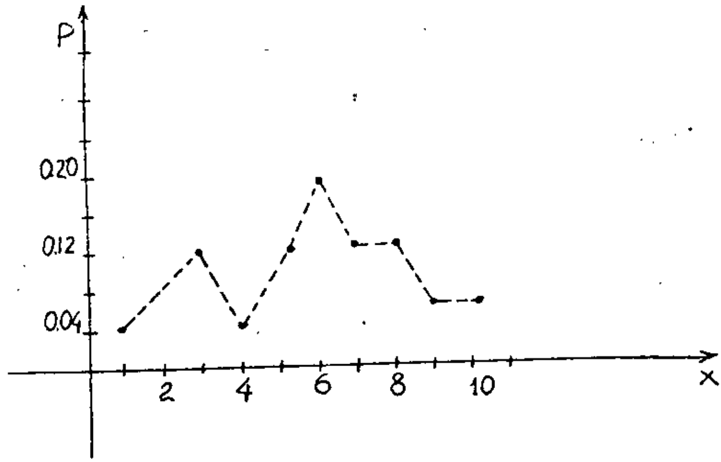
$x_{(i)}$	1	2	3	4	5	6	7	8	9	10
m_i	1	2	3	1	3	5	3	3	2	2

Hãy xây dựng biểu đồ tần suất.

Ta có

$x_{(i)}$	1	2	3	4	5	6	7	8	9	10
$\frac{m_i}{n}$	0,04	0,08	0,12	0,04	0,12	0,2	0,12	0,12	0,08	0,08

Biểu đồ tần suất nhận được ở hình 1



Hình 1: Biểu đồ tần suất ứng với ví dụ 2

Giả sử X là biến ngẫu nhiên liên tục với hàm mật độ p(x) chưa biết.

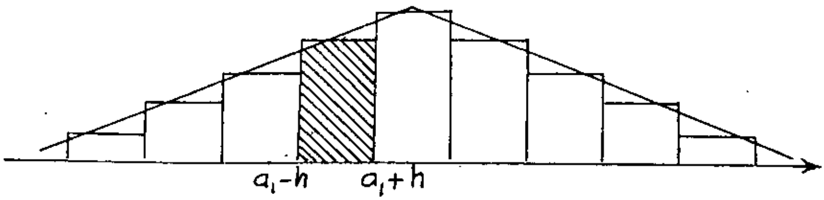
Ta chia khoảng biến thiên của mẫu thành các khoảng con bằng nhau với độ dài 2h bởi các điểm chia $a_i - h$; $a_i + h$, $i = 1, 2, \ldots$ Trên mỗi khoảng con ta lập một hình chữ nhật cạnh là khoảng $(a_i - h; a_i + h)$ có độ dài 2h và cạnh kia có độ dài $\dfrac{m_i}{2h \cdot n}$ trong đó m_i là số giá trị mẫu rơi vào khoảng $(a_i - h; a_i + h)$.

Khi đó diện tích hình chữ nhật con: $\dfrac{m_i}{2h \cdot n} \cdot 2h = \dfrac{m_i}{n}$.

Tổng diện tích các hình chữ nhật con:

$$\sum_i 2h \times \frac{m_i}{2h \cdot n} = \sum_i \frac{m_i}{n} = 1$$

Đường gấp khúc nhận được gọi là tổ chức đồ (hình 2)



Hình 2: Tổ chức đồ

Xấp xỉ đường gấp khúc bởi một đường cong nào đó. Đường cong này cho ta một hình ảnh hình học ban đầu về đường cong mật độ cần tìm.

Để làm sáng tỏ điều nói trên ta xét trường hợp đường cong mật độ $p(x)$ liên tục,

$$p_i = P\{X \in (a_i - h,\ a_i + h)\} = \int_{a_i-h}^{a_i+h} p(x)dx = p(v_i) \times 2h$$

Đẳng thức cuối là do ta áp dụng định lý trung bình tích phân, v_i là điểm nào đó trong khoảng $(a_i-h,\ a_i+h)$.

Theo luật số lớn, với xác suất 1 khi $n \to \infty$, ta có:

$$\frac{m_i}{n} \to p_i \text{ hay } \frac{m_i}{n} \to p(v_i) \,.\, 2h$$

$$\frac{m_i}{n \,.\, 2h} \to p(v_i)$$

tức là trong mỗi khoảng tung độ của tổ chức đồ: $\dfrac{m_i}{n \,.\, 2h}$ sẽ xấp xỉ giá trị của đường cong mật độ $p(x)$.

Ví dụ 3: (xem [2])

Trong cuộc điều tra Glucoza trong máu ở 100 người ta thu được kết quả như sau (mg%):

Khoảng Glucoza	Số người	Khoảng	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		

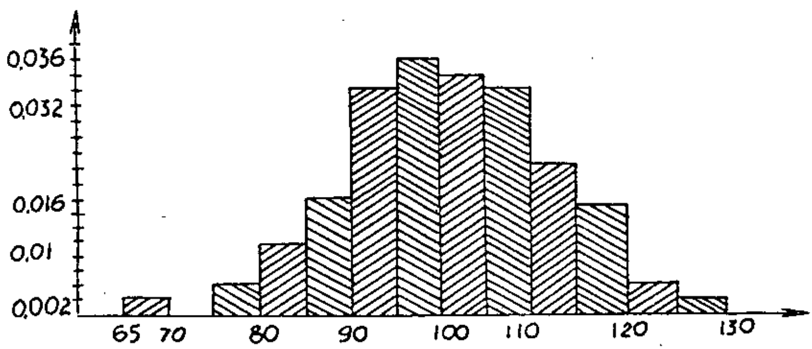
Hãy xây dựng tổ chức đồ ứng với mẫu trên.

Ở đây các số liệu đã được ghép thành các khoảng bằng nhau có độ dài là 5, do đó ta chọn luôn các khoảng đã cho là các khoảng (a_i-h, a_i+h).

Ta tính $\dfrac{m_i}{n\,.\,2h} = \dfrac{m_i}{100\,.\,5}$ tương ứng với các khoảng:

$\dfrac{m_i}{500}$	0,002	0	0,004	0,01	0,016	0,032	0,036
	0,034		0,032	0,018	0,014	0.004	và 0,002

Nhìn vào tổ chức đồ ta thấy có tính đối xứng, có thể xấp xỉ bởi đường cong mật độ chuẩn.



Hình 3: Tổ chức đồ ứng với ví dụ 3

§ 5. CÁC ĐẶC TRƯNG MẪU

Giả sử ta nghiên cứu biến ngẫu nhiên X

Ký hiệu $\mu = EX$; $\sigma^2 = DX$

Các số đặc trưng của X gọi là các số đặc trưng lý thuyết (các số này ta chưa biết và đang cần tìm chúng).

$(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên được rút ra từ X.

Ta xây dựng biến ngẫu nhiên rời rạc X' nhận n giá trị mẫu với xác suất đều và bằng 1/n: $P\{X' = X_i\} = 1/n$

Dễ dàng thấy ngay rằng: $F_n(x) = P\{X' < x\}$.

Hàm phân phối thực nghiệm $F_n(x)$ còn được gọi là hàm phân phối mẫu.

Các số đặc trưng của biến ngẫu nhiên X' gọi là các số đặc trưng mẫu: Kỳ vọng mẫu, phương sai mẫu, mômen mẫu cấp k,...

1. Kỳ vọng mẫu:

$$\overline{X} = EX' = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Do $X_1, X_2, ... X_n$ là các biến ngẫu nhiên độc lập cùng phân phối như X, nên kỳ vọng mẫu là một biến ngẫu nhiên. Do đó ta lại tìm kỳ vọng và phương sai của $\overline{X}$:

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n} \times n \times EX = \mu$$

$$D\overline{X} = \frac{1}{n^2}\sum_{1}^{n} DX_i = \frac{1}{n^2} \times n \times DX = \frac{DX}{n} = \frac{\sigma^2}{n}$$

2. Phương sai mẫu:

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

Ta có:

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

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

$$= \frac{1}{n}n\,.\,DX - D\overline{X} = DX - \frac{DX}{n} = \frac{n-1}{n} \times \sigma^2$$

Vì giá trị trung bình của s^2 không đúng bằng σ^2, do đó nhiều khi thay cho s^2 ta dùng:

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

Khi đó:

$$E\hat{s} = \frac{n}{n-1} \times Es^2 = \frac{n}{n-1} \times \frac{n-1}{n} \times \sigma^2 = \sigma^2$$

Ta có thể tính được các mômen cấp cao hơn nữa của $\overline{X}$, s^2. Nhưng trong giáo trình này ta dừng lại ở đây.

3. Phân phối của $\overline{X}$ và s^2:

– Nếu $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên được rút ra từ biến ngẫu nhiên X có phân phối nhị thức B(m, p) thì $n\overline{X} = \sum_{i=1}^{n} X_i$ cũng có phân phối nhị thức B(mn, p).

– Nếu X có phân phối Poisson với tham số λ thì $n\overline{X} = \sum_{i=1}^{n} X_i$ cũng có phân phối Poisson với tham số $n\lambda$.

– Nếu X có phân phối χ^2_m thì $n\overline{X}$ cũng có phân phối χ^2_{nm}.

– Nếu X có phân phối chuẩn $N(\mu, \sigma^2)$ thì:

a. $n\overline{X}$ cũng có phân phối chuẩn $N(n\mu, n\sigma^2)$

$$\text{và } \overline{X} \simeq N(\mu, \frac{\sigma^2}{n})$$

b. $\dfrac{ns^2}{\sigma^2} \simeq \chi^2_{n-1}$

c. $\overline{X}$ độc lập với s^2 (và ngược lại).

d. $t = \dfrac{\overline{X} - \mu}{s}\sqrt{n-1}$ hoặc $t = \dfrac{\overline{X} - \mu}{\hat{s}}\sqrt{n}$ có phân phối Student với n – 1 bậc tự do.

– Nếu $(X_1, X_2, ..., X_n)$ là mẫu từ biến ngẫu nhiên chuẩn $N(\mu_1, \sigma_1^2)$ còn $(Y_1, Y_2, ..., Y_m)$ là mẫu từ biến ngẫu nhiên chuẩn độc lập với mẫu trên, khi đó:

$$\overline{X} \pm \overline{Y} \simeq N(\mu_1 \pm \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$

$$t = \frac{\{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)\} \times \sqrt{(n+m-2) \cdot m \cdot n}}{\sqrt{(n+m) \cdot (ns_x^2 + ms_y^2)}}$$

có phân phối Student với n + m –2 bậc tự do, trong đó s_x^2, s_y^2 là hai phương sai mẫu tương ứng với mẫu X và mẫu Y.

Các kết quả trên được chứng minh dễ dàng khi ta dùng công cụ hàm đặc trưng trong lý thuyết xác suất. Song ở giáo trình này ta phải công nhận các kết quả trên để sử dụng sau này.

4. Cách tính $\overline{X}$ và s^2:

Cho mẫu ngẫu nhiên $(x_1, x_2, ..., x_n)$ hoặc mẫu thu gọn:

$$\begin{cases} x_{(1)} & x_{(2)} \ \cdots \ x_{(k)} \\ m_1 & m_2 \ \ldots \ m_k \end{cases} \qquad \sum_{i=1}^{k} m_i = n$$

Để đơn giản cách viết, từ nay về sau thay cho viết $x_{(i)}$ ta viết x_i, chỉ cần nhớ rằng x_i trong mẫu thu gọn không nhất thiết là x_i trong mẫu ban đầu:

khi đó $\overline{X} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i$ $\qquad s^2 = \frac{1}{n}\sum_{i=1}^{k} m_i x_i^2 - \overline{X}^2$

Nếu các giá trị mẫu x_i không gọn, lại cách đều nhau thì ta có thể thu gọn số liệu bằng phép biến đổi tuyến tính.

$$u_i = \frac{x_i - x_o}{h}$$

(có nghĩa là ta thay đổi gốc tính và đơn vị tính). Chọn hằng số x_o và h ta căn cứ vào dãy số liệu cụ thể, thông thường ta lấy x_o là giá trị x_i ở quãng giữa, còn h là độ dài khoảng cách đều giữa các x_i. Ta lập bảng tính như sau:

x_i	m_i	u_i	$m_i u_i$	$m_i u_i^2$
$\sum$	n		(*)	(*)

Từ bảng tính trên ta tính được $\overline{u} = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot u_i$

$$s_u^2 = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot u_i^2 - \overline{u}^2$$

Trở lại số liệu ban đầu ta có:

$$\overline{X} = x_o + h \cdot \overline{u}$$

$$s_x^2 = h^2 \cdot s_u^2$$

Trong trường hợp tính toán bằng các phương tiện thô sơ thì việc lập bảng tính và thu gọn số liệu vẫn rất cần thiết và có lợi.

Nếu mẫu được cho dưới dạng các khoảng, ta chọn mỗi khoảng một điểm đại diện, thông thường là điểm giữa khoảng lúc đó ta lại có mẫu thu gọn.

Ví dụ 4: Trở lại với số liệu được cho trong ví dụ 3.

Gọi X là lượng Glucoza trong máu (mg%). Tính glucoza trung bình $\overline{X}$ và bình phương độ lệch mẫu s^2.

Theo số liệu đã cho, chọn x_i là điểm giữa của mỗi khoảng.

Chọn: $x_0 = x_7 = 97{,}5$: $h = 5$

$$u_i = \frac{x_i - 97{,}5}{5}$$

Ta có bảng tính sau:

Các khoảng Glucoza	m_i	x_i	u_i	$m_i u_i$	$m_i u_i^2$
1	2	3	4	5	6
65 - 70	1	67,5	-6	-6	36
70 - 75	0	72,5	-5	0	0
75 - 80	2	77,5	-4	-8	32
80 - 85	5	82,5	-3	-15	45
85 - 90	8	87,5	-2	-16	32
90 - 95	16	92,5	-1	-16	16
95 - 100	18	97,5	0	0	0
100 - 105	17	102,5	1	17	17
105 - 110	16	107,5	2	32	64
110 - 115	9	112,5	3	27	81
115 - 120	5	117,5	4	20	80
120 - 125	2	122,5	5	10	50
125 - 130	1	127,5	6	6	36
$\sum$	100			51	489

$$\overline{u} = \frac{51}{100} = 0{,}51$$

$$s_u^2 = \frac{489}{100} - 0{,}51^2 \approx 4{,}63$$

$$\overline{X} = 97{,}5 + 5 \cdot 0{,}51 \approx 100{,}05$$

$$s_X^2 = 5^2 \cdot 4{,}63 = 115{,}75$$

$$s \approx 10{,}76$$

Ví dụ 5: Một tổ gồm 50 công nhân xây dựng, trong một tuần thực hiện lương khoán theo sản phẩm, thu nhập của từng người trong tuần đó như sau (đơn vị nghìn đồng)

49	62	68	73	76	80	87	89	92	94
51	63	68	73	77	83	87	90	92	95
54	65	69	74	79	85	87	91	93	96
56	65	70	74	79	86	88	91	93	96
59	67	72	74	80	86	89	92	94	97

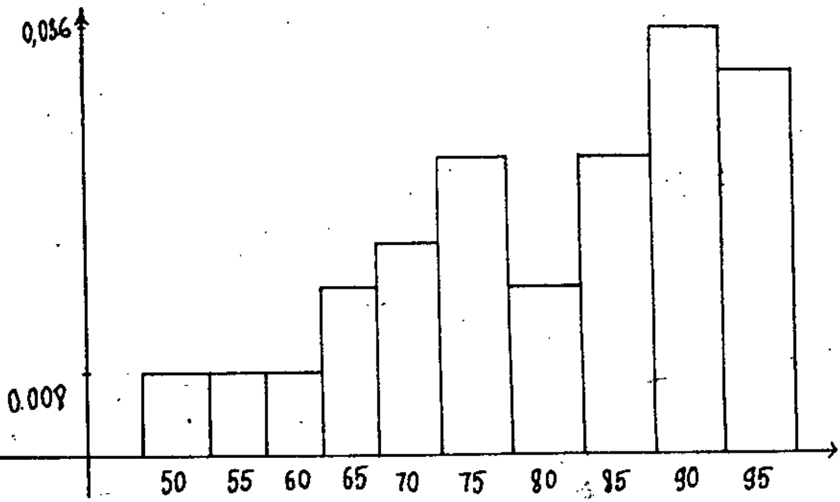
a. Hãy thu gọn số liệu trên bằng cách nhóm lại thành các khoảng.

b. Xây dựng tổ chức đồ.

c. Tìm $\overline{X}$ và s^2.

Để không gặp giá trị mẫu nào trùng với điểm phân khoảng ta chọn các điểm nằm giữa 2 số nguyên, khi đó việc thu gọn số liệu thể hiện ở 2 cột đầu, còn để xây dựng tổ chức đồ ta cần thêm cột 3 và để tính $\overline{X}$, s^2 ta cần 4 cột cuối.

Khoảng	m_i	$m_i/2h.n$	Điểm đại diện x_i	$u_i = x_i$ -70/5	$m_i u_i$	$m_i u_i^2$
1	2	3	4	5	6	7
47,5 - 52,5	2	0,008	50	-4	-6	32
52,5 - 57,5	2	0,008	55	-3	-6	18
57,5 - 62,5	2	0,008	60	-2	-4	8
62,5 - 67,5	4	0,016	65	-1	-4	4
67,5 - 72,5	5	0,020	70	0	0	0
72,5 - 77,5	7	0,028	75	1	7	7
77,5 - 82,5	4	0,016	80	2	8	16
82,5 - 87,5	7	0,028	85	3	21	63
87,5 - 92,5	9	0,036	90	4	36	144
92,5 - 97,5	8	0,032	95	5	40	200
$\sum$	50				90	492



Từ bảng tính trên ta tính được:

$\bar{u} = \frac{90}{50} = 1,8; s_u^2 = \frac{492}{50} - 1,8^2 = 6,6$

$\overline{X} = 70 + 5.\bar{u} = 70 + 5.1,8 = 79$

$s_x^2 = 5^2 \times 6,6 = 165$

$s \approx 12,845$

§6. SAI SỐ QUAN TRẮC

Trong việc lấy mẫu, do nhiều nguyên nhân khác nhau, sẽ không tránh khỏi các sai số trong số liệu mẫu. Vì vậy trước khi dùng các phương pháp thống kê để phân tích, xử lý ta cần loại bỏ các sai số không đáng có ở trong mẫu đã cho.

Giả sử x là kết quả quan sát được, a là giá trị chân thực (giá trị đúng) của đại lượng ta quan sát, Z là sai số.

Ta có

Sai số = Kết quả quan trắc - Giá trị chân thực.

Z = x - a

Nói chung vì a chưa biết nên sai số Z cũng chưa biết.

Để cho tiện việc xử lý ta phân loại các sai số như sau:

1. *Sai số thô:* Là sai số sinh ra do vi phạm các điều kiện cơ bản của việc lấy mẫu hoặc do sơ suất của người thực hiện, chẳng hạn người kiểm tra cố ý chọn ra các sản phẩm tốt để kiểm tra khi đánh giá chất lượng, hoặc người kỹ thuật viên ghi nhầm kết quả thu được...

2. *Sai số hệ thống:* Là sai số do không điều chỉnh chính xác dụng cụ hoặc không thống nhất giữa những kỹ thuật viên về

cách xác định một đại lượng nào đó..., do vậy dẫn đến một loạt kết quả quan sát được bị lệch đi một tỷ lệ nhất định nào đó.

3. *Sai số ngẫu nhiên*: Là sai số sinh ra do một số lớn các nguyên nhân mà tác dụng của chúng bé đến mức không thể tách riêng và tính riêng biệt cho từng nguyên nhân được.

Trong ba loại sai số trên, sai số thô và sai số hệ thống cần phát hiện sớm và khử bỏ ngay, còn sai số ngẫu nhiên không thể khử bỏ được trong mỗi lần quan sát. Do đó từ nay về sau khi các kết quả quan sát được đưa vào xử lý bằng các phương pháp toán học ta sẽ giả thiết rằng chúng chỉ chứa các sai số ngẫu nhiên.

4. ***Phân phối của sai số ngẫu nhiên:***

Sau khi loại bỏ sai số thô và hệ thống, sai số quan trắc chỉ còn lại sai số ngẫu nhiên: $Z = x - a$. Thông thường ta lấy luật phân phối chuẩn làm luật phân phối xác suất của sai số ngẫu nhiên. Điều đó thường khá phù hợp với thực nghiệm bởi lẽ luật phân phối chuẩn phản ánh tính đối xứng của các sai số ngẫu nhiên: Các sai số ngẫu nhiên có dấu khác nhau thường được gặp gần như nhau, hơn nữa luật chuẩn còn phản ánh tính tập trung của các sai số ngẫu nhiên: Sai số ngẫu nhiên có trị số tuyệt đối bé thường gặp hơn các sai số có trị số tuyệt đối lớn. Hơn nữa về phương diện lý thuyết, sai số ngẫu nhiên gây ra bởi các yếu tố rất nhỏ, độc lập với nhau nên theo định lý giới hạn trung tâm nó phải tuân theo luật chuẩn tức là $Z \simeq N(0, \sigma^2)$, trong đó σ là độ chính xác của các quan trắc.

Ta có:

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

$$P\{|Z| < k\sigma\} = \Phi(k) - \Phi(-k)$$

Trong đó $\Phi(x)$ là hàm phân phối của đại lượng ngẫu nhiên chuẩn $N(0, 1)$.

Với $k = 2$ ta có $P\{-2\sigma < Z < 2\sigma\} \approx 95\%$

$k = 3$ ta có $P\{|Z| > 3\sigma\} \approx 0{,}0027$

Xác suất 0,0027 quá bé cho nên ta xem như trong thực tế sai số ngẫu nhiên không thể vượt quá giới hạn $\pm 3\sigma$. Đó là nội dung của quy tắc 3σ.

5. Phương pháp khử sai số thô:

Để phát hiện được sai số thô ta thường căn cứ vào sự khác nhau rõ rệt về giá trị của một kết quả quan sát với các kết quả khác. Đó là cảm tính ban đầu. Để khẳng định chắc chắn ta làm như sau (cơ sở lý luận của thuật toán này ta sẽ hiểu kỹ ở chương III trong giáo trình này).

Giả sử x^* là giá trị bị nghi ngờ có chứa sai số thô. $x_1, x_2, \ldots, x_n$ là các giá trị còn lại.

a. Khi σ đã biết.

– Tính $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $u = \dfrac{|x^* - \overline{X}|}{\sigma\sqrt{\dfrac{n+1}{n}}}$

– Tra bảng phân phối chuẩn $N(0, 1)$ hàm $\Phi(x)$ tìm $2(1 - \Phi(u))$ (Bảng I phụ lục).

– Cho trước xác suất α khá bé thông thường $\alpha = 0{,}10; 0{,}05; 0{,}01, \ldots$

– Nếu $2(1 - \Phi(u)) < \alpha$ thì ta sẽ kết luận x^* chứa sai số thô và loại bỏ x^* khỏi mẫu.

Kết luận như vậy sẽ được đảm bảo với xác suất $\geq 1-\alpha$ và có thể sai với xác suất $\leq \alpha$.

Ví dụ 5: Trong 41 quan sát độc lập với sai số bình phương trung bình $\sigma = 0{,}133$ ta thấy có một giá trị đột xuất $x^* = 6{,}866$, đồng thời giá trị trung bình của 40 kết quả còn lại là $\overline{X} = 6{,}5$. Vậy với độ tin cậy 95% ta có thể xem x^* chứa sai số thô được không?

Ta tính $u = \dfrac{|x^* - \overline{X}|}{\sigma.\sqrt{\dfrac{n+1}{n}}} = \dfrac{6{,}866 - 6{,}500}{0{,}133.\sqrt{\dfrac{41}{40}}} \approx 2{,}72$

Tra bảng I: $2 \times (1 - \Phi(2{,}72)) = 2(1 - 0{,}9967) = 0{,}0067 < 0{,}05$.

Vậy với độ tin cậy 95% ta coi x^* chứa sai số thô. (Với kết quả tính được thậm chí ta coi x^* chứa sai số thô ngay cả với độ tin cậy 99,3%)

b. Khi σ chưa biết.

Tính $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{s}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$.

$$t = \frac{|x^* - \overline{X}|}{\hat{s}}$$

- Với α đã cho tra bảng II (phụ lục) ta tìm được giá trị tới hạn $t_n(1 - \alpha)$.

- Nếu $t > t_n(1 - \alpha)$ thì ta kết luận x^* chứa sai số thô và loại bỏ x^* khỏi tập các giá trị quan sát. Độ tin cậy của kết luận trên $\geq 1 - \alpha$.

Ví dụ 6: Xét lại ví dụ 5 nhưng σ chưa biết. Từ 40 kết quả quan sát ta tính được:

$\overline{X} = 6{,}5$; $\hat{s} = 0{,}13$ còn $x^* = 6{,}866$ với $\alpha = 0{,}05$ có thể coi x^* chứa sai số thô được không?

Vì $\overline{X}$, $\hat{s}$ đã được tính nên ta chỉ còn tính t.

$$t = \frac{6{,}866 - 6{,}500}{0{,}13} \approx 2{,}82$$

Tra bảng II ta được $t_{40}(0{,}95) = 2{,}048$.

Vậy với độ tin cậy 95% ta coi x^* chứa sai số thô.

Thực ra ta có thể kết luận với độ tin cậy 99% vì $t_{40}(0,99) = 2,742$.

BÀI TẬP CHƯƠNG I

1. Điều tra doanh số hàng tháng của 100 hộ kinh doanh một ngành nào đó ta thu được số liệu sau:

Doanh số X (triệu đồng)	10,1	10,2	10,4	10,5	10,7	10,8	10,9	11	11,3	11,4
Số hộ	2	3	8	13	25	20	12	10	6	1

a. Xây dựng đa giác tần suất

b. Tính $\overline{X}, s^2, \hat{s}^2, s, \hat{s}$.

2. Điều tra 365 điểm trồng lúa của một huyện ta được các số liệu sau (Xem [5]):

Năng suất X (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm trồng lúa	6	13	38	74	106	85	30	10	3

a. Xây dựng đa giác tần suất

b. Tính năng suất lúa trung bình và độ phân tán của năng suất lúa của 365 điểm nói trên.

3. Đo chiều cao của 50 cây ta được kết quả sau (X mét) (Số liệu trích từ [3]):

11.7	14,2	10,3	9,5	11,1	11.5	13.6	14.5	8.7	10
7,1	7,0	12.8	10,8	10,5	9.7	5.9	9,1	10,2	7,2
11,1	11,5	8.7	12.2	7,0	5.1	11.3	8.4	12.5	13
10,0	12.6	11,6	6.6	9.5	8.5	11.2	8,5	9.0	10.5
9,3	10,2	10,2	7,2	9,4	8.0	8,2	10	4.2	10.3

a. Rút gọn số liệu bằng cách ghép khoảng.

b. Lập tổ chức đồ.

c. Tính $\overline{X}$, s^2 và $\hat{s}$.

4. Kết quả thi môn xác suất cơ sở của lớp gồm 30 sinh viên được cho như sau:

(X: Điểm thi – cho theo thang điểm 10)

8	9	7	6	8	5	10	9	6	8	10	10
7	8	4	5	7	4	9	8	6	3	5	3
10	9	4	7	6	8						

a. Vẽ đa giác tần suất.

b. Tính số điểm trung bình mà lớp đạt được, độ phân tán tiêu chuẩn của điểm thi của các sinh viên.

Chương II

VỀ BÀI TOÁN ƯỚC LƯỢNG THAM SỐ

Nghiên cứu biến ngẫu nhiên X và giả sử ta biết được phân phối của X thuộc họ phân phối phụ thuộc vào tham số θ nào đó: $X \simeq F(x, \theta)$.

Khi đó để xác định hoàn toàn phân phối của X ta phải xác định được các giá trị tham số θ mà phân phối đó nhận. Chẳng hạn ta biết X có phân phối Poisson, nhưng tham số λ bằng bao nhiêu (θ chính là λ), hoặc ta biết X có phân phối chuẩn, nhưng μ và σ^2 nhận giá trị nào thì lại chưa rõ (θ chính là cặp (μ, σ^2)),...

Ngay trong trường hợp ta chưa biết gì về phân phối của X, khi đó biết được các số đặc trưng của X là rất có giá trị.

Do đó bài toán đi tìm các ước lượng cho các tham ẩn của phân phối hoặc ước lượng các số đặc trưng của biến ngẫu nhiên là bài toán rất cần thiết.

§ 1. ƯỚC LƯỢNG ĐIỂM CHO KỲ VỌNG, MEDIAN, PHƯƠNG SAI VÀ XÁC SUẤT

1. Ước lượng điểm:

Giả sử θ là tham số cần ước lượng. Trong tay chúng ta chỉ có một mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$. Vì vậy để ước lượng cho θ không thể không dựa vào mẫu ngẫu nhiên.

Ta sẽ dùng một hàm nào đó của mẫu, tức là một hàm nào đó của n biến $X_1, X_2, ..., X_n$ để làm ước lượng cho θ. Ký hiệu hàm đó là $\theta^*(X_1, X_2, ..., X_n)$. Như vậy $\theta^*(X_1, X_2, ..., X_n)$ là một biến ngẫu nhiên vì $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên độc lập, cùng phân phối. $\theta^*(X_1, X_2, ..., X_n)$ nhận một giá trị cụ thể (một điểm) khi mẫu $(X_1, X_2, ..., X_n)$ nhận một giá trị cụ thể $(x_1, x_2, ..., x_n)$ nên θ^* gọi là ước lượng điểm.

2. Ước lượng không chệch:

Vì $\theta^*(X_1, X_2, ..., X_n)$ là biến ngẫu nhiên nên ta không thể đòi hỏi $\theta^*(X_1, X_2, ..., X_n)$ đúng bằng giá trị θ cần tìm được.

Ước lượng $\theta^*(X_1, X_2, ..., X_n)$ được gọi là ước lượng không chệch của θ nếu thoả mãn:

$$E\theta^*(X_1, X_2, ..., X_n) = \theta$$

Nếu $E\theta^*(X_1, X_2, ..., X_n) = \theta + C$ thì $\theta^*(X_1, X_2, ..., X_n)$ gọi là ước lượng chệch đối với độ chệch C.

Như vậy

$\overline{X}$ là ước lượng không chệch cho EX

s^2 là ước lượng không chệch cho DX

$\hat{s}^2$ là ước lượng chệch của DX đối với độ chệch $-\dfrac{DX}{n}$

3. Ước lượng điểm cho kỳ vọng của biến ngẫu nhiên:

$\overline{X}$ là ước lượng không chệch của kỳ vọng của biến ngẫu nhiên bất kỳ. Cụ thể $\overline{X}$ là ước lượng không chệch cho tham ẩn μ phân phối chuẩn, cho tham ẩn λ của phân phối Poisson và cho tham ẩn $\frac{1}{\lambda}$ phân phối mũ...

4. Ước lượng điểm cho Median (trung vị):

Ước lượng điểm cho Median của biến ngẫu nhiên là Median mẫu, ký hiệu là Med.

Med là giá trị mà nó chia dãy số liệu đã được sắp theo thứ tự (tăng hoặc giảm) thành 2 phần sao cho số các phần tử mẫu ở trên Med bằng số các phần tử mẫu ở dưới Med.

a. Tính Med đối với các số liệu không nhóm lại.

Giả sử $x_{(1)}, x_{(2)}, ..., x_{(n)}$ là dãy các giá trị mẫu được xếp lại theo thứ tự tăng dần (hoặc giảm dần).

Nếu n lẻ thì $\quad \text{Med} = x_{(n+1)/2}$

Nếu n chẵn thì $\quad \text{Med} = \frac{1}{2}(x_{n/2} + x_{n/2+1})$

b. Tính Med đối với các số liệu được gộp lại thành các khoảng.

Khoảng thứ i: (x_i, x_{i+1}) có m_i quan sát nằm trong khoảng đó, $\sum_i m_i = n$

Giả sử Med nằm trong khoảng thứ k: (x_k, x_{k+1}).

Khi đó Med được xác định như sau:

$$\text{Med} = \frac{\frac{n}{2} - \sum_{i}^{k-1} m_i}{m_k}(x_{k+1} - x_k) + x_k$$

Ví dụ 1: Tiến hành đo chiều cao cho 100 em học sinh lớp 3 (8 tuổi) ở một trường phổ thông cơ sở, ta có kết quả sau:

Khoảng chiều cao (cm)	Số em	$m_1 + ... + m_i$
(110 - 112]	5	5
(112 - 114]	8	13
(114 - 116]	14	27
(116 - 118]	17	44
(118 - 120]	20	64
(120 - 122]	16	80
(122 - 124]	10	90
(124 - 126]	6	96
(126 - 128]	4	100
$\sum$	100	

Từ số liệu ta thấy Med sẽ nằm trong khoảng thứ 5 (k = 5),

$$\sum_{i=1}^{4} m_i = 44 \ , \ m_5 = 20 \, , \ x_5 = 118 \, , \ x_6 = 120.$$

$$\text{Vậy Med} = \frac{\frac{100}{2} - 44}{20} \times (120 - 118) + 118 \approx 118{,}6$$

5. Ước lượng điểm cho phương sai:

$\hat{s}^2$ là ước lượng điểm không chệch cho $DX = \sigma^2$

s^2 là ước lượng điểm của DX với độ chệch : $\dfrac{-DX}{n}$

6. Ước lượng điểm cho xác suất:

Ước lượng cho xác suất p của biến cố A nào đó (hay ước lượng cho tỷ lệ nào đó) là

$p^* = \dfrac{m}{n}$, trong đó n là số lần quan sát, m là số lần xảy ra biến cố A.

$p^* = \frac{m}{n}$ cũng là ước lượng không chệch cho xác suất p.

(vì $Ep^* = E\left(\frac{m}{n}\right) = \frac{1}{n}.Em = \frac{np}{n} = p$)

Ví dụ 2: Trở lại ví dụ 1 về chiều cao của 100 học sinh lớp 3.

– Gọi X là chiều cao của em học sinh (cm). Hãy chỉ ra ước lượng điểm cho EX, DX và p = P{116 < X < 124}.

Ta lập bảng tính $\overline{X}$, s^2.

Gọi x_i là điểm đại diện cho mỗi khoảng.

Ta thực hiện phép thu gọn số liệu $u_i = \frac{x_i - 119}{2}$

Khoảng chiều cao	m_i	x_i	u_i	$m_i u_i$	$m_i u_i^2$
(110 - 112]	5	111	-4	-20	80
(112 - 114]	8	113	-3	-24	72
(114 - 116]	14	115	-2	-28	56
(116 - 118]	17	117	-1	-17	17
(118 - 120]	20	119	0	0	0
(120 - 122]	16	121	1	16	16
(122 - 124]	10	123	2	20	40
(124 - 126]	6	125	3	18	54
(126 - 128]	4	127	4	16	64
$\sum$	100			-19	399

$$\overline{u} = \frac{-19}{100} = -0{,}19 \Rightarrow \overline{X} = 119 + 2\overline{u} = 118{,}62$$

$$s_u^2 = \frac{399}{100} - (-0{,}19)^2 = 3{,}9529$$

$$s_x^2 = 4 \times 3{,}9539 = 15{,}8156$$

$$s \approx 3{,}977$$

$$p^* = \frac{63}{100} = 0{,}63 \quad (63 = 17 + 20 + 16 + 10)$$

Vậy: ước lượng điểm cho EX là 118,62.

ước lượng điểm cho DX là 15,8156.

ước lượng điểm cho p là 0,63.

Ví dụ 3: Để đánh giá tỷ lệ người mắc bệnh bướu cổ ở một vùng cao, ta chọn ngẫu nhiên vài bản làng và điều tra số người mắc bệnh ở bản này. Kết quả thấy trong số 264 người có 156 người bị mắc bệnh bướu cổ. Hỏi tỷ lệ mắc bệnh bướu cổ ở vùng cao này là bao nhiêu. (Ta coi như tình hình mắc bệnh ở các bản khác nhau trong vùng là như nhau).

Ta có $p^* = \dfrac{156}{246} \approx 0,59 = 59\%$

Ta ước lượng tỷ lệ mắc bệnh thực sự của cả vùng là 59%.

§ 2. ƯỚC LƯỢNG KHOẢNG

1. Định nghĩa ước lượng khoảng:

Một khoảng với hai đầu mút ngẫu nhiên $(\theta_1^*(X_1, ..., X_n)$, $\theta_2^*(X_1, ..., X_n))$ được gọi là ước lượng khoảng cho tham ẩn θ với độ tin cậy $1 - \alpha$ nếu tham ẩn θ thuộc vào khoảng nói trên với xác suất $1 - \alpha$ tức là:

$$P\{\theta_1^*(X_1, ..., X_n) < \theta < \theta_2^*(X_1, ..., X_n)\} = 1 - \alpha$$

(Ở đây $(X_1, ..., X_n)$ vẫn ký hiệu là mẫu ngẫu nhiên). Như vậy xác suất để tham số không nằm trong khoảng $(\theta_1^*(X_1, ..., X_n), \theta_2^*(X_1, ..., X_n))$ chỉ bằng α.

Khoảng $(\theta_1^*(X_1,..., X_n), \theta_2^*(X_1, ..., X_n))$ được gọi là khoảng tin cậy. Giá trị $1 - \alpha$ gọi là độ tin cậy. Còn $\theta_2^* - \theta_1^*$ gọi là độ chính xác của ước lượng.

Rõ ràng với cùng một độ tin cậy thì khoảng tin cậy càng hẹp càng giúp ta xác định chính xác được tham số cần tìm.

2. Ước lượng khoảng đối với kỳ vọng của biến ngẫu nhiên:

Nếu $(X_1, \ldots, X_n)$ được rút ra từ họ chuẩn thì $\overline{X} = N(EX, \frac{DX}{n})$ (xem §5 chương I).

Nếu mẫu ngẫu nhiên được rút ra từ biến ngẫu nhiên X không phải chuẩn, khi đó $\overline{X}$ cũng có phân phối xấp xỉ chuẩn với kỳ vọng EX và phương sai $\frac{DX}{n}$ nếu n đủ lớn. (Kết quả này nhận được từ định lý giới hạn trung tâm trong lý thuyết xác suất). Vấn đề cần nói đến ở đây là: Thế nào gọi là đủ lớn? Tốt nhất là $n \geq 100$, song trong nhiều ứng dụng $n \geq 30$ cũng được coi là đủ lớn để sử dụng xấp xỉ chuẩn của $\overline{X}$ (xem [2]).

a. Nếu DX đã biết ($DX = \sigma^2$):

Khi đó ước lượng khoảng cho kỳ vọng $\mu = EX$ (nếu X không chuẩn thì đòi hỏi $n \geq 30$) với độ tin cậy $1 - \alpha$ là:

$$\left\{\overline{X} - u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} ; \overline{X} + u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right\} \tag{1}$$

Trong đó $u\left(\frac{\alpha}{2}\right)$ xác định từ $\Phi(u(\frac{\alpha}{2})) = 1 - \frac{\alpha}{2}$

Cơ sở lý luận của kết luận trên như sau:

Với giả thiết đặt ra (về tính chuẩn hoặc $n \geq 30$) ta có:

$$\overline{X} \simeq N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \simeq N(0, 1)$$

Do đó $P\left\{\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} \leq u_{(\alpha/2)}\right\} = 1 - \alpha$

Hay $$P\left\{\overline{X} - u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

Từ đó rút ra khoảng tin cậy cần tìm.

Ví dụ 4: Để xác định chiều cao trung bình của các cây bạch đàn trong khu rừng rộng trồng bạch đàn ta không có điều kiện đo chiều cao của mọi cây trong khu rừng, do đó ta tiến hành đo ngẫu nhiên 35 cây. Kết quả cho được như sau:

Khoảng chiều cao (m)	(6,50 - 7,0]	(7,0 - 7,5]	(7,5 - 8]	(8 - 8,5]	(8,5 - 9]	(9 - 9,5]
Số cây	2	4	10	11	5	3

Với xác suất 95% ta có thể nói chiều cao trung bình của cây bạch đàn thuộc khu rừng trên nằm trong khoảng nào. Giả sử sai số tiêu chuẩn σ của phân phối chiều cao bạch đàn đã biết và bằng 0,64 (giả thiết này có thể nhận được nếu ta đã có các số liệu về chiều cao bạch đàn cùng lứa tuổi và cùng điều kiện sinh trưởng. Dùng phương sai mẫu nhận được từ các số liệu trước làm phương sai cho chiều cao cây bạch đàn trong ví dụ này).

Từ mẫu trên ta tính được $\overline{X} = 8{,}06$

$$\overline{X} \pm u\left(\frac{0{,}05}{2}\right)\frac{\sigma}{\sqrt{n}} = 8{,}06 \pm 1{,}96 \times \frac{0{,}64}{\sqrt{35}} = 8{,}06 \pm 0{,}21$$

Vậy chiều cao trung bình của cây bạch đàn nằm trong khoảng (7,85; 8,27). Khẳng định đó đúng với xác suất 95%.

b. Nếu DX chưa biết:

Giả sử $(X_1, \ldots, X_n)$ là mẫu ngẫu nhiên rút ra từ biến ngẫu nhiên chuẩn, $X_i \simeq N(\mu, \sigma^2)$. Để đưa ra ước lượng khoảng cho $\mu = EX$ ta xuất phát từ khẳng định:

Đại lượng $\frac{\overline{X}-\mu}{s}\sqrt{n-1}$ có phân phối Student với n - 1 bậc tự do (xem §5 chương I).

Từ đó $P\left\{\frac{|\overline{X}-\mu|}{s}\sqrt{n-1} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha$

hay

$$P\left\{\overline{X} - t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n-1}} < \mu < \overline{X} + t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n-1}}\right\} = 1-\alpha$$

Với giá trị $t_{n-1}\left(\frac{\alpha}{2}\right)$ được tìm từ bảng phân phối Student ứng với n - 1 bậc tự do (xem bảng III, phụ lục):

$$P\left\{|T_{n-1}| \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha$$

Do vậy với độ tin cậy 1 - α, ước lượng khoảng của μ là

$$\left\{\overline{X} - t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n-1}},\ \overline{X} + t_{n-1}\left(\frac{\alpha}{2}\right) \times \frac{s}{\sqrt{n-1}}\right\} \quad (2)$$

Ví dụ 5: Trở lại ví dụ 4 và coi như σ^2 chưa biết, giả thiết X chuẩn, ta tính được $s^2 = 0{,}40$ ($s \approx 0{,}63$; $\alpha = 0{,}05$), tra bảng III ta có:

$$t_{34}\left(\frac{0{,}05}{2}\right) = 2{,}03$$

Khoảng tin cậy (2) sẽ là:

$$\left(8{,}06 - 2{,}03 \times \frac{0{,}63}{\sqrt{34}};\ 8{,}06 + 2{,}03 \times \frac{0{,}63}{\sqrt{34}}\right) = (7{,}84\ ;\ 8{,}28)$$

Vậy với độ tin cậy 95% ta khẳng định chiều cao trung bình của cây bạch đàn ở khu rừng trên nằm trong khoảng (7,84; 8,28). Khẳng định như vậy có khả năng đúng là 95%, khả năng khẳng định bị sai là 5%.

Nhận xét: Trong trường hợp tính chuẩn không thoả mãn, phương sai cũng chưa biết, nhưng nếu n lớn ta có thể thay phương sai chưa biết σ^2 bởi $\hat{s}^2$ và đưa về áp dụng trường hợp phương sai đã biết. Khi đó ta có khoảng tin cậy xấp xỉ cho μ với độ tin cậy $(1 - \alpha)$ là

$$\left\{ \overline{X} - u(\frac{\alpha}{2}) \frac{\hat{s}}{\sqrt{n}} ; \overline{X} + u(\frac{\alpha}{2}) \frac{\hat{s}}{\sqrt{n}} \right\} \tag{3}$$

3. Ước lượng khoảng đối với phương sai của biến ngẫu nhiên chuẩn:

Vì $(X_1, ..., X_n)$ là mẫu chuẩn nên $\frac{ns^2}{\sigma^2}$ hoặc $\frac{(n-1)\hat{s}^2}{\sigma^2}$ có phân phối χ^2_{n-1} (xem §5 chương I). Từ đó ta có khoảng tin cậy đối với phương sai $DX = \sigma^2$ như sau:

$$\left(\frac{(n-1)\hat{s}^2}{\chi^2_{n-1}(\alpha/2)} ; \frac{(n-1)\hat{s}^2}{\chi^2_{n-1}(1-\alpha/2)} \right) \tag{4}$$

Trong đó $\chi^2_{n-1}(1-\alpha/2)$ và $\chi^2_{n-1}(\alpha/2)$ tra từ bảng phân phối χ^2_{n-1} (bảng IV - phụ lục) thoả mãn hệ thức:

$$P\left\{ \chi^2_{n-1} > \chi^2_{n-1}(\alpha/2) \right\} = \alpha/2$$

$$P\left\{ \chi^2_{n-1} > \chi^2_{n-1}(1-\alpha/2) \right\} = 1 - \alpha/2$$

Ví dụ 6: Chỉ ra ước lượng khoảng của σ^2 trong ví dụ 4 với độ tin cậy 95%.

Theo ví dụ 5 ta tính được $s^2 = 0{,}40$;

suy ra $\hat{s}^2 = \frac{35}{34} \times s^2 \approx 0{,}4118$

Tra bảng IV ta được $\chi^2_{34}(0{,}975) = 19{,}8$ $\chi^2_{34}(0{,}025) = 52$

Khoảng tin cậy (4) là: (0,269; 0,707).

Ví dụ 7: Để tham khảo độ chính xác của một dụng cụ đo độ dài người ta đo trên cùng một mục tiêu 30 lần bằng dụng cụ ấy - kết quả nhận được $\hat{s}^2 = 0{,}05$. Hãy tìm ước lượng khoảng cho độ chính xác của dụng cụ đo với độ tin cậy 95%.

Theo §6 chương I ta coi sai số ngẫu nhiên Z tuân theo luật chuẩn $N(0, \sigma^2)$. Vì vậy ta có thể dùng (4) để ước lượng cho độ chính xác của dụng cụ đo.

Ta có $\chi^2_{29}(0{,}975) = 16$; $\chi^2_{29}(0{,}025) = 45{,}7$

Khoảng tin cậy (4) sẽ là: (0,032; 0,09).

4. Ước lượng khoảng đối với tỷ lệ hay xác suất:

Ước lượng đối với tỷ lệ (xác suất) nào đó là bài toán rất hay gặp ở trong các lĩnh vực. Xác suất p chính là tham ẩn của phân phối nhị thức. Trong §1 ta đã chỉ ra ước lượng điểm của p là:

$$p^* = \frac{m}{n}$$

Ta có thể biểu diễn cách khác của p^* như sau:

Giả sử $(X_1, \ldots, X_n)$ là mẫu ngẫu nhiên

trong đó $X_i = \begin{cases} 1 & \text{nếu ở phép thử thứ i biến cố A xuất hiện} \\ 0 & \text{nếu ở phép thử thứ i biến cố } \overline{A} \text{ xuất hiện} \end{cases}$

hoặc $X_i = \begin{cases} 1 & \text{với xác suất p} \\ 0 & \text{với xác suất } q = 1 - p \end{cases}$

Khi đó $$p^* = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

(Viết p^* theo tần suất $\frac{m}{n}$ hoặc theo kỳ vọng mẫu $\overline{X}$ đều như nhau).

Nhưng theo cách thứ 2 ta cần lưu ý mẫu ngẫu nhiên ở đây là mẫu của dãy phép thử Bernoulli, nếu không dễ bị nhầm lẫn!).

Ta dễ dàng nhận được

$$Ep^* = p, \quad Dp^* = \frac{p(1-p)}{n}$$

$$m = \sum_{i=1}^{n} X_i \approx B(n, p)$$

a. Nếu n lớn và p không quá gần 0 và 1

Vì n lớn ta có thể xấp xỉ phân phối của p^* bằng phân phối chuẩn (theo định lý giới hạn trung tâm trong lý thuyết xác suất). Do đó ước lượng khoảng của p với độ tin cậy $1 - \alpha$ là

$$\left(\frac{m}{n} - u\left(\frac{\alpha}{2}\right)\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\ ;\ \frac{m}{n} + u\left(\frac{\alpha}{2}\right)\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\right) \quad (5)$$

Ví dụ 8: Để xác định tỷ lệ nảy mầm của lô hạt giống, người ta gieo thử 300 hạt, thấy có 276 hạt nảy mầm. Với độ tin cậy 95% ta có thể nói tỷ lệ nảy mầm của lô hạt giống tối đa là bao nhiêu?

Theo (5), với độ tin cậy 95% ta khẳng định tỷ lệ nảy mầm của lô hạt giống sẽ thuộc vào khoảng:

$$\left(\frac{276}{300} - 1{,}96\sqrt{\frac{0{,}92 \times (1-0{,}92)}{300}}\ ;\ 0{,}92 + 1{,}96\sqrt{\frac{0{,}92 \times (1-0{,}92)}{300}}\right)$$

$$= (0{,}89;\ 0{,}95).$$

Như vậy với độ tin cậy 95% ta kết luận tỷ lệ nảy mầm tối đa của lô là 95%, tỷ lệ tối thiểu là 89%.

b. Nếu n lớn và p gần 0 hoặc gần 1

Vì p gần 0 hoặc gần 1 nên ta không xấp xỉ phân phối nhị thức của p bằng phân phối chuẩn được mà ta xấp xỉ bằng phân phối Poisson với tham ẩn $\lambda = np$. Trên cơ sở đó ta lập được các giá trị trong bảng V (phụ lục). Tương ứng với giá trị m. tra bảng ta tìm được giá trị np_1 và np_2 với độ tin cậy 95%. Từ đó khoảng tin cậy 95% của xác suất p là (p_1, p_2).

Ví dụ 9: Làm xét nghiệm AIDS cho 1000 người, thấy có 15 người cho kết quả dương tính. Xét nghiệm tiếp đối với những người này thấy có 3 người mắc bệnh AIDS. Hãy chỉ ra ước lượng khoảng cho tỷ lệ người mắc bệnh AIDS ở thành phố trên với độ tin cậy 95%.

Ta có ước lượng điểm $p^* = \dfrac{3}{1000} = 0{,}003$ khá bé nên tra bảng V với $m = 3$ ta được $np_1 = 0{,}62$; $np_2 = 8{,}76$. Do đó

$$p_1 = 0{,}00062, \; p_2 = 0{,}00876.$$

Vậy với độ tin cậy 95% tỷ lệ người mắc bệnh AIDS ở thành phố nọ vào khoảng (0,00062; 0,00876).

c. Nếu n bé nhưng p không quá gần 0 hoặc 1

Giả sử $4 \leq n \leq 10$, m là số lần xuất hiện biến cố A ($p = p(A)$).

Tương ứng với m đã cho, tra bảng VI (phụ lục) ta tìm được giới hạn dưới p_1 và giới hạn trên p_2 với độ tin cậy 95%.

Ví dụ 10: (xem [5]) Trong đợt vận động bầu cử tổng thông người ta phỏng vấn ngẫu nhiên 1600 cử tri, được biết 960 người trong số đó sẽ bỏ phiếu cho ứng cử viên A.

a. Hãy chỉ ra ước lượng điểm cho tỷ lệ thực (p) các cử tri bỏ phiếu cho ứng cử viên A.

b. Với độ tin cậy 95% tối thiểu ứng cử viên A sẽ chiếm được bao nhiêu phần trăm số phiếu.

Ta có ước lượng điểm cho p là:

$$p^* = \frac{960}{1600} = 0{,}6$$

Theo (5), ta có:

$$\left[\frac{960}{1600} - u(\frac{0{,}05}{2})\sqrt{\frac{0{,}6 \times 0{,}4}{1600}}\ ;\ 0{,}6 + 1{,}96\sqrt{\frac{0{,}6 \times 0{,}4}{1600}}\ \right]$$

$$= (0{,}6 - 0{,}024;\ 0{,}6 + 0{,}024) = (0{,}576;\ 0{,}624)$$

Vậy với độ tin cậy 95%, tối thiểu ứng cử viên A chiếm được 57% số phiếu bầu của cử tri.

5. Ước lượng khoảng của sự khác nhau giữa hai giá trị trung bình:

Giả sử X và Y là 2 biến ngẫu nhiên với các giá trị trung bình là μ_1 và μ_2, còn phương sai là σ_1^2 và σ_2^2 đã biết.

Gọi $D = \mu_1 - \mu_2$ là sự khác nhau giữa hai kỳ vọng.

Giả sử hai mẫu ngẫu nhiên cỡ n_1 và n_2 được rút ra độc lập với nhau từ 2 biến X và Y. Từ đó ta nhận được các trung bình mẫu $\overline{X}$, $\overline{Y}$ và $\overline{D} = \overline{X} - \overline{Y}$.

Nếu X, Y phân phối chuẩn thì $\overline{D}$ cũng phân phối chuẩn.

Nếu X, Y không phải chuẩn thì $\overline{D}$ được xấp xỉ chuẩn nếu n_1, n_2 khá lớn ($n_1, n_2 \geqslant 30$).

Do đó, $\overline{D} \simeq N(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2})$

Từ đây ta nhận được ước lượng khoảng đối với hiệu $\mu_1 - \mu_2$ với độ tin cậy $1 - \alpha$ như sau:

$$\left(\overline{X} - \overline{Y} - u(\frac{\alpha}{2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ \overline{X} - \overline{Y} + u(\frac{\alpha}{2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \quad (6)$$

Ví dụ 11: Lấy 100 quả trứng từ lô trứng do nhóm gà A đẻ ra, xác định được trọng lượng trứng trung bình là 40gam.

Lấy 120 quả từ lô trứng do nhóm gà B đẻ ra, xác định được trọng lượng trứng trung bình là 44gam.

Với $\alpha = 5\%$, sự sai khác giữa hai loại trứng gà nằm trong khoảng nào? Giả thiết trọng lượng quả trứng gà cả hai loại đều tuân theo quy luật chuẩn $\sigma_1^2 = \sigma_2^2 = 15$

Theo (6) ta có:

$$\left\{ 44 - 40 - 1{,}96\sqrt{15 \times \left(\frac{1}{100} + \frac{1}{120}\right)}\ ;\right.$$

$$\left. 44 - 40 + 1{,}96\sqrt{15 \times \left(\frac{1}{100} + \frac{1}{120}\right)} \right\} = (2{,}97\,;\,5{,}03).$$

§ 3. ĐỘ CHÍNH XÁC CỦA ƯỚC LƯỢNG VÀ SỐ QUAN SÁT CẦN THIẾT

Trong §2 chúng ta đã giải quyết bài toán xây dựng ước lượng khoảng. Khi cho mẫu ngẫu nhiên, có nghĩa là cho cỡ mẫu n, cho độ tin cậy $1 - \alpha$, ta xây dựng được ước lượng khoảng (θ_1^*, θ_2^*). Như vậy độ chính xác của ước lượng $\varepsilon = \theta_2^* - \theta_1^*$ cũng sẽ chỉ ra. Trong trường hợp ước lượng khoảng đối xứng dạng $(\theta^* - b, \theta^* + b)$, trong đó b có thể là hằng số, có thể là ngẫu nhiên, thì $\varepsilon = 2b$, có lúc để đơn giản ta lấy $\varepsilon = b$.

Trong các trường hợp trình bày ở §2.2, §2.4 các ước lượng khoảng đều đối xứng và đều phụ thuộc cỡ mẫu n.

Bây giờ bài toán đặt ngược lại: cho độ tin cậy, cho độ chính xác ε, hãy tìm số quan sát n cần thiết để nhận được ước lượng với độ tin cậy và độ chính xác đã cho.

Bài toán được đưa về tìm n là số nguyên bé nhất thoả mãn: $b(n) < \varepsilon$.

Chẳng hạn ta có $$b(n) = u(\frac{\alpha}{2}) \times \frac{\sigma}{\sqrt{n}} \leqslant \varepsilon \quad \text{(xem (1))}$$

hoặc $$b(n) = t_{n-1}(\frac{\alpha}{2}) \times \frac{s}{\sqrt{n-1}} < \varepsilon \quad \text{(xem (2), (3))}$$

hoặc $$b(n) = u(\frac{\alpha}{2})\sqrt{\frac{p^* \times (1-p^*)}{n}} < \varepsilon \quad \text{(xem (5))}$$

Nhìn chung b(n) phụ thuộc vào n ở dạng $n^{-1/2}$ do đó muốn tăng độ chính xác lên 2 lần thì phải tăng n lên 4 lần.

Ví dụ 12: Trở lại ví dụ 5, với $n = 35$, $\alpha = 0,05$ ta đã nhận được ước lượng khoảng cho chiều cao trung bình của cây bạch đàn là (7,84; 8,28). Như thế độ chính xác của ước lượng nhận được là $\varepsilon \approx 0,22$.

Bây giờ ta muốn có ước lượng khoảng với độ chính xác $\varepsilon = 0,10$ thì cần đo bao nhiêu cây?

Theo (2) ta có:

$$t_{n-1}\left(\frac{0,05}{2}\right) \times \frac{0,63}{\sqrt{n-1}} = \frac{2,03 \times 0,63}{\sqrt{n-1}} < 0.10$$

Suy ra

$$n > 1 + \left\{\frac{2,03 \times 0,63}{0,1}\right\}^2 = 164,56 \Rightarrow n \geqslant 165$$

Hoặc trở lại ví dụ 8, $n = 300$, $\alpha = 0,05$ ta nhận được ước lượng khoảng (0,89; 0,95), với độ chính xác $\varepsilon = 0,03$

Bây giờ ta muốn nâng độ chính xác lên 2 lần, tức là $\varepsilon = 0,015$ thì cần bao nhiêu quan sát?

Ta có

$$u\left(\frac{0,05}{2}\right) \times \sqrt{\frac{p^* \times (1 - p^*)}{n}} = 1,96 \times \sqrt{\frac{0,92 \times (1 - 0,92)}{n}} < 0,015.$$

Suy ra

$$n > \left(\frac{1,96}{0,015}\right)^2 \times 0,92 \times 0,08 = 1256,63. \text{ Vậy } n \geq 1257$$

Ví dụ 13: Cần nghiên cứu sản lượng sữa của một đàn bò của một nông trường nhằm xem xét:

a. Sản lượng sữa trung bình một ngày của một con bò là bao nhiêu?

b. Bao nhiêu % đàn bò cho sản lượng sữa trên 11kg trong ngày?

Để làm việc đó ta tiến hành điều tra một cách ngẫu nhiên trên 100 con bò, kết quả như sau:

X (sản lượng sữa/ngày (kg))	< 9kg	9-11	11-13	13-15	>15
Số con bò	10	24	42	16	8

Giả sử X tuân theo luật chuẩn.

c. Muốn độ tin cậy của kết luận là 99,73%, sai số khi nghiên cứu sản lượng không vượt quá 0,5kg, sai số khi nghiên cứu tỷ lệ bò cho sữa > 11kg/ngày không vượt quá 12% thì cần phải điều tra bao nhiêu con bò?

Giải

a. Ước lượng điểm cho sản lượng sữa trung bình của một con bò/ngày chính là $\overline{X}$. Chọn điểm đại diện cho mỗi khoảng là điểm giữa ta có:

$$\overline{X} = 8 \times \frac{10}{100} + 10 \times \frac{24}{100} + 12 \times \frac{42}{100} + 14 \times \frac{16}{100} + 16 \times \frac{8}{100}$$

$$= 11,76 \text{ kg}$$

b. Tỷ lệ bò cho sữa > 11kg/ngày là

$$p = \frac{42 + 16 + 8}{100} \approx 0,66$$

c. Ở đây $1 - \alpha = 99,73$

$$u\left(\frac{0,0027}{2}\right) = 3\ ,\ t_{99}\left(\frac{0,0027}{2}\right) = 3,1$$

$$t_{99}\left(\frac{0,0027}{2}\right) \times \frac{s}{\sqrt{n - 1}} \leq 0,5$$

$$u\left(\frac{0,0027}{2}\right) \times \sqrt{\frac{0,66 \times 0,34}{n}} \leq 0,12$$

Từ số liệu ta tính được

$$EX^2 = \frac{14194}{100} = 141,94$$

$$s^2 = 142,72 - 11,76^2 = 4,4226 \Rightarrow s = 2,103$$

Vậy $$n \geq 1 + \frac{3,1^2 \times 4,4226}{(0,5)^2} = 171,005$$

$$n \geq \frac{3^2}{0,12^2} \times 0,66 \times 0,34 = 140,25$$

Do đó cần phải điều tra 172 con bò, nghĩa là điều tra thêm 72 con bò nữa.

BÀI TẬP CHƯƠNG II

1. Với số liệu đã cho trong bài tập 1 chương I, hãy:

   a. Chỉ ra Median mẫu.

   b. Với độ tin cậy 95%, có thể nói doanh số trung bình/tháng của các hộ nằm trong khoảng nào? Giả thiết X tuân theo luật chuẩn.

c. Ước lượng tỷ lệ % các hộ có doanh số tháng ≥ 11 triệu Với độ tin cậy 99% tỷ lệ này thấp nhất là bao nhiêu?

2. Với số liệu đã cho trong bài tập 2 chương I, hãy:

a. Chỉ ra Median mẫu, giá trị trung bình mẫu?

b. Với độ tin cậy 95% năng suất lúa trung bình của huyện thấp nhất và cao nhất là bao nhiêu? (giả thiết năng suất lúa là biến ngẫu nhiên chuẩn).

c. Tỷ lệ % số điểm trồng lúa có năng suất cao > 35tạ/ha? Tỷ lệ này thấp nhất là bao nhiêu với độ tin cậy 99%.

3. Có một khu rừng có diện tích rất lớn. Căn cứ vào kết quả điều tra ngẫu nhiên trên 31 ô, mỗi ô có diện tích trên 0,1 ha được giá trị trung bình mẫu (thể tích gỗ trung bình trên mỗi ô) và sai số tiêu chuẩn trên mỗi ô là $\overline{X} = 10,2m^3$, $s = 1,45m^3$. Hãy ước lượng số quan sát cần thiết để sai số ước lượng không vượt quá $0,4m^3$ với độ tin cậy 95%?

- Nếu muốn sai số không vượt quá 0,5 thì cần điều tra bổ sung hay không ?

Nếu muốn sai số không vượt quá 0,5 và số quan sát không vượt quá cỡ mẫu đã điều tra thì độ tin cậy của ước lượng là bao nhiêu (Số liệu trích từ [3]).

Hãy chỉ ra ước lượng khoảng đó.

Giả thiết X tuân theo luật chuẩn.

4. Với độ tin cậy 95% hãy ước lượng mức xăng tiêu hao trung bình cho một loại ôtô chạy từ A đến B, nếu chạy thử 30 lần trên đoạn đường này người ta ghi nhận được lượng xăng tiêu hao như sau:

Lượng xăng (lít)	(9,6 - 9,8]	(9,8 - 10]	(10 - 10,2]	(10,2 - 10,4]	(10,4 - 10,6]
Số lần tương ứng	3	5	10	8	4

Biết rằng lượng xăng hao phí là đại lượng ngẫu nhiên chuẩn (xem[5]).

5. Để ước lượng số cá trong hồ, người ta đánh bắt 2000 con, đánh dấu rồi thả xuống. Vài ngày sau, ta lại đánh bắt 400 con thì thấy 80 con cá được đánh dấu. Với độ tin cậy 95% số cá trong hồ có bao nhiêu con? (xem [5]).

6. Một xí nghiệp đưa ra thị trường một sản phẩm mới. Để xem đánh giá của người tiêu dùng đối với loại sản phẩm mới này như thế nào, người ta phát cho mỗi người mua hàng đó một phiếu thăm dò và yêu cầu gửi lại cho xí nghiệp chậm nhất là sau 3 tháng. Vì điều kiện thời gian nên xí nghiệp không thể hỏi ý kiến của tất cả khách hàng trong cả nước, cho nên họ chỉ gửi phiếu thăm dò cho khách hàng ở Hà Nội. Kết quả, sau 3 tháng, xí nghiệp nhận được 300 phiếu thăm dò, trong đó có 90 phiếu tỏ ra là thích loại sản phẩm này (cả về chức năng và giá cả). Hãy ước lượng tỷ lệ thực khách hàng thích loại sản phẩm này?

Với độ tin cậy 95% tỷ lệ đó cao nhất là bao nhiêu?

Muốn nhận được ước lượng khoảng cho tỷ lệ thực với độ chính xác là 0,03 thì cần thăm dò thêm bao nhiêu phiếu nữa.

Với mẫu $n = 300$, ước lượng khoảng có độ chính xác là 0,0436 thì độ tin cậy của kết luận về ước lượng khoảng là bao nhiêu?

7. Để nghiên cứu tuổi thọ của một loại bóng đèn người ta thắp thử 100 bóng và có số liệu sau:

Tuổi thọ X (giờ)	Số bóng tương ứng	Tuổi thọ X (giờ)	Số bóng tương ứng
1010 - 1030	2	1110 - 1130	20
1030 - 1050	3	1130 - 1150	12
1050 - 1070	8	1150 - 1170	10
1070 - 1090	13	1170 - 1190	6
1090 - 1110	25	1190 - 1210	1

Sau khi cải tiến kỹ thuật người ta lại thắp thử 100 bóng. Kết quả như sau:

Tuổi thọ Y (giờ)	1150	1160	1170	1180	1190	1200
Số bóng	10	15	20	30	15	10

a. Hãy chỉ ra ước lượng điểm và ước lượng khoảng (α = 0,05) cho tuổi thọ trung bình (EX, EY) và bình phương độ lệch của tuổi thọ bóng đèn (DX, DY) trước và sau khi cải tiến.

b. Với độ tin cậy 95% có thể nói việc cải tiến kỹ thuật đã làm tăng tuổi thọ trung bình của bóng đèn lên ít nhất bao nhiêu giờ.

c. Nếu ước lượng khoảng cho EX có độ chính xác là 6,05 thì độ tin cậy tương ứng là bao nhiêu.

d. Nếu muốn ước lượng khoảng cho EX với độ tin cậy 95%, độ chính xác là 5 thì cần quan sát thêm bao nhiêu bóng đèn nữa.

Giả sử X và Y đều tuân theo luật chuẩn.

(Số liệu trên được trích từ [6]).

Chương III

VỀ BÀI TOÁN KIỂM ĐỊNH GIẢ THIẾT

Trong chương này chúng ta sẽ giải quyết các bài toán dạng như sau:

Giả sử ta có hai giả thiết (hai khả năng) về một vấn đề nào đó chẳng hạn:

– Có ý kiến cho rằng tham ẩn θ của phân phối nào đó nhận giá trị θ_0, nhưng lại có ý kiến cho rằng θ nhận giá trị θ_1.

– Có ý kiến cho rằng năng suất trung bình của cây trồng theo 2 phương pháp là như nhau, nhưng lại có ý kiến cho rằng năng suất của phương pháp này cao hơn phương pháp kia.

– Có ý kiến cho rằng biến ngẫu nhiên đang xét tuân theo luật chuẩn, lại có ý kiến không tán thành.

– Có ý kiến cho rằng việc điều trị lần trước có ảnh hưởng đến kết quả điều trị lần này, nhưng lại có ý kiến cho rằng không ảnh hưởng, v.v...

Vấn đề đặt ra là ta phải chọn một trong hai giả thiết đặt ra. Nói cách khác là ta chọn giả thiết nào để khả năng đúng cao hơn, khả năng sai thấp hơn. Để cho tiện ta gọi một trong hai giả thiết đặt ra là giả thiết H, còn giả thiết kia là đối thiết K.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên. Vận dụng các kết quả của lý thuyết xác suất ta sẽ tìm một miền S, sao cho khi mẫu $(X_1, ..., X_n) \in S$ thì ta bác bỏ giả thiết H, còn khi $(X_1, ..., X_n) \notin S$ thì ta chấp nhận H cho đến khi có thông tin mới.

Miền S nói trên được gọi là miền tiêu chuẩn.

Khi bác bỏ hay chấp nhận giả thiết H chúng ta có thể mắc phải 2 loại sai lầm.

Ta bác bỏ H nhưng thực tế H đúng: Sai lầm loại I.

Ta chấp nhận H nhưng thực tế H sai: Sai lầm loại II.

Ta mong muốn chọn S sao cho cực tiểu cả 2 khả năng phạm sai lầm. Nhưng khi cỡ mẫu n cố định thì mong muốn trên không thể thực hiện được, do đó thông thường ta cho trước giới hạn trên của xác suất phạm sai lầm loại I, ký hiệu là α, α thường nhỏ (α = 0,10; 0,05; 0,01). Ta sẽ tìm miền S sao cho khả năng phạm sai lầm loại II đạt cực tiểu.

α được gọi là mức ý nghĩa của tiêu chuẩn.

Sau đây sẽ xét một số bài toán đơn giản.

§ 1. KIỂM ĐỊNH GIẢ THIẾT VỀ GIÁ TRỊ TRUNG BÌNH

Giả sử $(X_1, \ldots, X_n)$ là mẫu ngẫu nhiên rút từ biến ngẫu nhiên X có $EX = \mu$, $DX = \sigma^2$ (Nếu X không chuẩn thì $n \geq 30$).

$H : \mu = \mu_0$ với đối thiết $K : \mu \neq \mu_0$; α cho trước.

1. Nếu σ^2 đã biết:

Các bước tiến hành như sau:

– Xuất phát từ mẫu đã cho tính $\overline{X}$ và $u = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

– Tìm giá trị $u_{(\alpha/2)}$ từ bảng (I) sao cho $\Phi(u(\frac{\alpha}{2})) = 1 - \frac{\alpha}{2}$

So sánh u và $u(\frac{\alpha}{2})$

Nếu $|u| \geq u(\frac{\alpha}{2})$ ta bác bỏ H

Nếu $|u| < u(\frac{\alpha}{2})$ ta chấp nhận H.

Cơ sở lý luận của quy tắc trên như sau:

Nếu $X = N(\mu, \sigma^2)$ hoặc X không phải chuẩn nhưng $n \geq 30$ thì với giả thiết H đúng ($\mu = \mu_o$) ta có:

$$\overline{X} = N(\mu, \frac{\sigma^2}{n}). \text{ Suy ra } \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} = N(0, 1)$$

Do đó $$P\left\{ \frac{|\overline{X} - \mu_o|}{\sigma/\sqrt{n}} \geq u(\frac{\alpha}{2}) \right\} = \alpha$$

Như vậy giả thiết đúng, xác suất ta bác bỏ giả thiết $\leq \alpha$.

Tức là xác suất phạm sai lầm loại I $\leq \alpha$. (Việc chứng minh xác suất phạm sai lầm loại II cực tiểu bạn đọc xem trong [1]). Ở đây miền tiêu chuẩn

$$S = \left\{ (X_1, ..., X_n) : \frac{|\overline{X} - \mu_o|}{\sigma/\sqrt{n}} \geq u(\frac{\alpha}{2}) \right\}$$

Hoặc $$S = \left\{ (X_1, .., X_n) : \overline{X} \leq \mu_o - u(\frac{\alpha}{2}) \frac{\sigma}{\sqrt{n}} \right.$$

$$\text{hoặc } \left. \overline{X} \geq \mu_o + u(\frac{\alpha}{2}) \frac{\sigma}{\sqrt{n}} \right\}$$

Nếu ta xét H : $\mu = \mu_o$ với K : $\mu > \mu_o$ hoặc H : $\mu = \mu_o$ với K : $\mu < \mu_o$

Cũng vẫn với cơ sở lý luận như trên ta có ngay 2 miền tiêu chuẩn tương ứng là:

$$S = \left\{ (X_1, ..., X_n) : \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} \geq u(\alpha) \right\} \quad (2)$$

và $$S = \left\{ (X_1, \ldots, X_n) : \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq -u(\alpha) \right\} \quad (3)$$

trong đó $\Phi(u(\alpha)) = 1 - \alpha$, $\Phi(-u(\alpha)) = \Phi(u(1 - \alpha)) = \alpha$

2. Nếu σ^2 chưa biết:

Trong trường hợp này ta phải giả thiết về tính chuẩn của biến ngẫu nhiên X.

– Xuất phát từ mẫu đã cho ta tính $\overline{X}$, s^2 và $t = \frac{\overline{X} - \mu_0}{s}\sqrt{n-1}$

– Với α đã cho, tra bảng phân phối Student với n – 1 bậc tự do ta tìm được $t_{n-1}(\alpha/2)$ từ hệ thức $P\left\{ |T_{n-1}| \geq t_{n-1}(\alpha/2) \right\} = \alpha$ trong đó T_{n-1} là biến ngẫu nhiên với n – 1 bậc tự do.

– So sánh t và $t_{n-1}(\alpha/2)$

Nếu $|t| \geq t_{n-1}(\alpha/2)$ ta bác bỏ giả thiết H.

Nếu $|t| < t_{n-1}(\alpha/2)$ ta chấp nhận giả thiết H.

Như vậy miền tiêu chuẩn là:

$$S = \left\{ (X_1, \ldots, X_n) : \frac{|\overline{X} - \mu_0|}{s}\sqrt{n-1} \geq t_{n-1}(\alpha/2) \right\} \quad (4)$$

Có cơ sở lý luận của cách làm trên như sau:

Giả sử giả thiết H đúng, tức $X \simeq N(\mu_0, \sigma^2)$

Khi đó $\overline{X} \simeq N(\mu_0, \frac{\sigma^2}{n})$ và $\frac{\overline{X} - \mu_0}{s}\sqrt{n-1}$ có phân phối Student với n – 1 bậc tự do (xem §5 chương I), do đó

$$P\left\{ \frac{|\overline{X} - \mu_0|}{s}\sqrt{n-1} \geq t_{n-1}(\alpha/2) \right\} = \alpha$$

Tương tự nếu H : $\mu = \mu_0$, K : $\mu > \mu_0$

$$S = \left\{ (X_1, \ldots, X_n) : \frac{\overline{X} - \mu_0}{s}\sqrt{n-1} \geq t_{n-1}(\alpha) \right\} \quad (5)$$

Nếu H : $\mu = \mu_0$, K : $\mu < \mu_0$

$$S = \left\{ (X_1, \dots, X_n) : \frac{\overline{X} - \mu_0}{s} \sqrt{n-1} \leq -t_{n-1}(\alpha) \right\} \qquad (6)$$

trong đó:

$$P\left\{ T_{n-1} \geq t_{n-1}(\alpha) \right\} = \alpha$$

$$P\left\{ T_{n-1} \leq -t_{n-1}(\alpha) \right\} = P\left\{ T_{n-1} \leq t_{n-1}(1-\alpha) \right\} = \alpha$$

Ví dụ 1: (xem [3]) Một vườn cây con phi lao có chiều cao trung bình chưa xác định. Theo hợp đồng đã ký giữa người sản xuất cây con và lâm trường trồng cây thì chỉ khi nào chiều cao của cây đạt trên 1m mới đem ra trồng để đảm bảo tỷ lệ sống cao.

Người ta điều tra ngẫu nhiên 50 cây trong vườn và tính được chiều cao trung bình $\overline{X} = 1{,}1$m. Hỏi vườn cây phi lao nói trên đã đưa ra trồng được chưa? Cho biết sự biến động về chiều cao của cây phi lao trong giai đoạn vườn ươm ở trong những điều kiện tương tự là $\sigma = 0{,}1$.

Vì σ đã biết ta dùng (1) hoặc (2) hoặc (3). Ở đây giả thiết

H : $\mu = 1{,}0$ chọn K : $\mu > 1{,}0$; $\alpha = 0{,}05$

Ta có:

$$u = \frac{\overline{X} - 1}{0{,}1}\sqrt{50} = \frac{1{,}1 - 1}{0{,}1} \times 7{,}1 \approx 7{,}1$$

$$u(0{,}05) = 1{,}65$$

Theo (2) ta bác bỏ giả thiết $\mu = 1$ và chấp nhận đối thiết $\mu > 1{,}0$ nghĩa là có thể đưa vườn cây phi lao trên ra trồng được rồi. Bài toán đã được giải quyết.

Mang tính chất minh họa ta xét thêm bài toán H : $\mu = 1$ với K : $\mu < 1$, $\alpha = 0{,}05$.

Theo (3) ta có $\{(X_1, \dots, X_n): u = 7{,}1 > -1{,}65\}$ nghĩa là ta chấp nhận H : $\mu = 1$, bác bỏ K : $\mu < 1$.

Điều này không có gì mâu thuẫn với kết luận trước, bởi vì thực tế $\mu > 1$, cho nên giữa 2 khả năng $\mu = 1$ và $\mu < 1$ thì khả năng $\mu = 1$ ít sai hơn so với khả năng $\mu < 1$.

Nếu trong ví dụ trên ta chưa biết σ. Nhưng từ 50 quan sát ta tính được $s \approx 0{,}12$.

Khi đó để giải quyết bài toán $H : \mu = 1$ với $K : \mu > 1$ ta dùng (5)

$$\alpha = 0{,}05 \Rightarrow t_{49} = 1{,}68$$

còn
$$\frac{\overline{X} - \mu_0}{s}\sqrt{n-1} = \frac{1{,}1 - 1}{0{,}12}\sqrt{49} \approx 5{,}833$$

$$5{,}833 > 1{,}68$$

tức mẫu đã cho thuộc S(5) nên ta bác bỏ giả thiết $\mu = 1$, chấp nhận $\mu > 1$.

Nhận xét: Nếu giả thiết chuẩn không có, σ^2 lại chưa biết, nhưng cỡ mẫu n lớn ta thay σ^2 bởi $\hat{s}^2$ và áp dụng như trường hợp σ^2 đã biết.

§ 2. KIỂM ĐỊNH GIẢ THIẾT VỀ XÁC SUẤT (HOẶC TỶ LỆ)

Giả sử $(X_1,...,X_n)$ là mẫu ngẫu nhiên trong đó

$$X_i = \begin{cases} 1 & \text{với xác suất } p \\ 0 & \text{với xác suất } q = 1 - p \end{cases} \qquad p = P(A)$$

hay nói cách khác trong n quan sát có m lần xảy ra biến cố A,

$$m = \sum_{i=1}^{n} X_i \;;\;\; m \approx B(n, p)$$

Nếu n lớn ($n \geqslant 30$) thì $\frac{m}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ có phân phối xấp xỉ chuẩn. Lý luận tương tự như trường hợp §1 (tiêu chuẩn (1), (2), (3)) ta nhận được các miền tiêu chuẩn để kiểm định giả thiết H : $p = p_o$ với các đối thiết $p \neq p_o$, $p > p_o$ và $p < p_o$ mức α tương ứng sau:

$$p = p_o \text{ với } p \neq p_o \quad S = \left\{ \frac{\left|\frac{m}{n} - p_o\right|}{\sqrt{p_o(1-p_o)}}\sqrt{n} \geqslant u\left(\frac{\alpha}{2}\right) \right\} \tag{7}$$

$$p = p_o \text{ với } p > p_o \quad S = \left\{ \frac{\frac{m}{n} - p_o}{\sqrt{p_o(1-p_o)}}\sqrt{n} \geqslant u(\alpha) \right\} \tag{8}$$

$$p = p_o \text{ với } p < p_o \quad S = \left\{ \frac{\frac{m}{n} - p_o}{\sqrt{p_o(1-p_o)}}\sqrt{n} \leqslant -u(\alpha) \right\} \tag{9}$$

Ví dụ 2: (xem [3]) Một kho hạt giống có tỷ lệ nảy mầm xác định $p_o = 0{,}90$. Ngẫu nhiên có một thiết bị bị hỏng làm thay đổi điều kiện bên trong của kho. Hỏi tỷ lệ nảy mầm của kho hạt giống có còn giữ nguyên hay không (với $\alpha = 0{,}05$)?

Để có thông tin về tỷ lệ nảy mầm mới của kho ta làm thí nghiệm 200 hạt thấy có 140 hạt nảy mầm.

Ở đây giả thiết H : $p = 0{,}90$ với K: $p < 0{,}90$. Theo (9) ta tính:

$$\frac{\frac{m}{n} - p_o}{\sqrt{p_o(1-p_o)}}\sqrt{n} = \frac{\frac{140}{200} - 0{,}9}{\sqrt{0{,}9 \times 0{,}1}}\sqrt{200} = -9{,}5$$

Khoảng trọng lượng (gam)	Điểm đại diện x_i	Số trẻ con so (m_i)	Số trẻ con dạ (m'_i)	$m_i x_i$	$m'_i x_i$
1700 - 1900	1800	3	1	5400	1800
1900 - 2100	2000	5	0	10000	0
2100 - 2300	2200	7	2	15400	4400
2300 - 2500	2400	5	4	12000	9600
2500 - 2700	2600	13	7	33800	18200
2700 - 2900	2800	21	13	58800	36400
2900 - 3100	3000	20	22	60000	66000
3100 - 3300	3200	11	18	35200	57600
3300 - 3500	3400	6	15	20400	51000
3500 - 3700	3600	3	10	10800	36000
3700 - 3900	3800	0	7	0	26600
3900 - 4100	4000	1	3	4000	12000
4100 - 4300	4200		2		8400
4300 - 4500	4400		1		4400
$\sum$		$n_1 = 95$	$n_2 = 105$	265800	332400

Giả sử độ lệch bình phương trung bình đối với con so $\sigma_1^2 = 190000$, đối với con dạ $\sigma_2^2 = 200704$, với $\alpha = 0,05$, hãy kiểm định giả thiết H : $\mu_1 = \mu_2$ với đối thiết K : $\mu_1 \neq \mu_2$ và $\mu_1 < \mu_2$

Ở đây $n_1 = 95$, $n_2 = 105$

Từ hai cột cuối ta có:

$$\overline{X} = \frac{265800}{95} = 2797,8947 \approx 2798\text{gam}$$

$$\overline{Y} = \frac{332400}{105} = 3165,7142 \approx 3166 \text{ gam}$$

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{2798 - 3166}{\sqrt{\dfrac{190000}{95} + \dfrac{200704}{105}}} \approx -5,88$$

Theo (10) : $|-5{,}88| > 1{,}96 \Rightarrow$ Ta bác bỏ $\mu_1 = \mu_2$ và chấp nhận $\mu_1 \neq \mu_2$ là đúng

Theo (12) : $-5{,}88 < -1{,}65 \Rightarrow$ Ta bác bỏ $\mu_1 = \mu_2$, chấp nhận $\mu_2 > \mu_1$

2. Nếu giả thiết σ_1^2 , σ_2^2 chưa biết:

Trong trường hợp này ta phải giả thiết:

$$X = N(\mu_1 ; \sigma_1^2), \quad Y = N(\mu_2 , \sigma_2^2) \text{ và } \sigma_1^2 = \sigma_2^2$$

Khi đó các bước làm như sau:

+ Xuất phát từ hai mẫu đã cho tính $\overline{X}$, $\overline{Y}$, s_x^2, s_y^2

+ Tính

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}} \sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}}$$

+ Với α đã cho tra bảng phân phối Student tìm

$$t_{n_1 + n_2 - 2}\left(\frac{\alpha}{2}\right) \text{ hoặc } t_{n_1 + n_2 - 2}(\alpha)$$

Các miền tiêu chuẩn như sau:

Nếu $\mu_1 = \mu_2$ và $\mu_1 \neq \mu_2$: $S = \left\{ |t| \geq t_{n_1 + n_2 - 2}\left(\frac{\alpha}{2}\right) \right\}$ (13)

Nếu $\mu_1 = \mu_2$ và $\mu_1 > \mu_2$: $S = \left\{ t \geq t_{n_1 + n_2 - 2}(\alpha) \right\}$

Ví dụ 4: (xem [5]) Người ta thí nghiệm hai phương pháp chăn nuôi gà khác nhau, sau một tháng, kết quả tăng trọng như sau:

Phương pháp I: $n_1 = 100$ con , $\overline{X} = 1{,}1$ kg , $s_1^2 = 0{,}04$

Phương pháp II: $n_2 = 150$ con , $\overline{Y} = 1{,}2$ kg , $s_2^2 = 0{,}09$

Với mức ý nghĩa $\alpha = 0{,}05$ có thể kết luận phương pháp II hiệu quả hơn phương pháp I được không? Giả thiết mức tăng trọng của gà tuân theo luật chuẩn và $\sigma_1^2 = \sigma_2^2$

Ta có

$$t = \frac{1{,}2 - 1{,}1}{\sqrt{\dfrac{100 \times 0{,}04 + 150 \times 0{,}09}{100 + 150 - 2}}\ \sqrt{\dfrac{100 + 150}{100 \times 150}}} \approx 2{,}914$$

$$H : \mu_1 = \mu_2 \ ; \ K : \mu_2 > \mu_1$$

$$\alpha = 0{,}05 \Rightarrow t_{248}(0{,}05) = 1{,}645$$

Theo (14) $t = 2{,}914 > t_{248}(0{,}05) = 1{,}645$

Ta bác bỏ giả thiết $\mu_1 = \mu_2$ và chấp nhận $\mu_2 > \mu_1$, tức là phương pháp II hiệu quả hơn phương pháp I.

Nhận xét: Trong trường hợp σ_1, σ_2 chưa biết, ta đã 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 thoả mãn ta cũng có thể xấp xỉ σ_1^2 bởi $\hat{s}_1^2$, σ_2^2 bởi $\hat{s}_2^2$. Đó là các ước lượng không chệch và nếu cỡ mẫu lớn thì sự xấp xỉ khá tốt, do đó ta vẫn có thể dùng quy tắc như trường hợp σ_1, σ_2 đã biết.

Trong ví dụ trên, nếu áp dụng (11) ta có

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\hat{s}_1^2}{n_1} + \dfrac{\hat{s}_2^2}{n_2}}} = \frac{1{,}2 - 1{,}1}{\sqrt{\dfrac{s_1^2}{n_1 - 1} + \dfrac{s_2^2}{n_2 - 1}}} = \frac{1{,}2 - 1{,}1}{\sqrt{\dfrac{0{,}04}{99} + \dfrac{0{,}09}{149}}} \approx 3{,}147$$

$3{,}147 > 1{,}65$, do đó ta cũng chấp nhận đối thiết $\mu_2 > \mu_1$

Ví dụ 5: Để khảo sát tình hình đọc báo của sinh viên người ta tiến hành điều tra ở 3 trường đại học:

* Trường A: Điều tra 50 người thấy trung bình mỗi sinh viên bỏ ra 10 giờ một tháng để đọc báo, độ lệch bình phương mẫu là 10.

* Trường B: Điều tra 40 người, thấy trung bình là 12 giờ, phương sai mẫu là 5,6.

*Trường C: Điều tra 30 người, thấy trung bình là 11,5 giờ, phương sai mẫu là 5,4.

Với xác suất 95%, hãy kiểm tra xem sinh viên các trường trên dành thời gian đọc báo có như nhau không?

Trong thống kê đã có lời giải cho bài toán so sánh nhiều giá trị trung bình, song ở giáo trình này ta không đề cập đến, do vậy ta áp dụng bài toán so sánh hai giá trị trung bình.

Ở đây $n_i \geq 30$ ta dùng (10).

Trước hết so sánh hai trường B và C, khi đó theo (10):

$$\frac{|12 - 11{,}5|}{\sqrt{\dfrac{5{,}6}{40-1} + \dfrac{5{,}4}{30-1}}} = \frac{0{,}5}{0{,}57} \approx 0{,}88 < 1{,}96$$

Vậy chấp nhận ở hai trường B và C sinh viên có số giờ đọc báo như nhau.

Bây giờ so sánh 2 trường A và B:

$$\frac{|12 - 10|}{\sqrt{\dfrac{5{,}6}{40-1} + \dfrac{10}{50-1}}} = \frac{2}{0{,}59} \approx 3{,}39 > 1{,}96$$

Ta bác bỏ giả thiết sinh viên hai trường A và B có số giờ đọc báo như nhau.

Kết luận: Sinh viên trường A đọc báo khác hẳn trường B và C

Sinh viên 2 trường B và C đọc báo như nhau.

3. Dùng tiêu chuẩn phi tham số:

Để kiểm định sự khác nhau giữa trung bình của hai biến ngẫu nhiên theo các phương pháp trình bày ở trên ta phải giả thiết đến tính chuẩn, hơn nữa còn đòi hỏi hai phương sai bằng nhau. Bây giờ ta sẽ trình bày một vài tiêu chuẩn trong khá nhiều tiêu chuẩn mà giả thiết tính chuẩn không cần đặt ra. Lớp các tiêu chuẩn này được gọi là phi tham số.

Trước hết ta định nghĩa hạng (rank) của một phần tử.

Giả sử ta có một dãy các số thực được xếp theo thứ tự tăng dần, trong dãy này không có giá trị nào bằng nhau:

$$x_1 < x_2 < \dots < x_k < \dots < x_n.$$

Khi đó rank $(x_1) = 1$, rank $(x_2) = 2$, ..., rank $(x_k) = k$,...

Nếu trong dãy có những giá trị trùng nhau, chẳng hạn $x_k = x_{k+1}$ hoặc $x_k = x_{k+1} = x_{k+2}$,... Khi đó hạng của các phần tử trùng nhau được tính như sau :

Nếu $x_{k-1} < x_k = x_{k+1} < x_{k+2}$ thì

$$\text{rank}\ (x_k) = \text{rank}(x_{k+1}) = k + \frac{1}{2}$$

Nếu $x_{k-1} < x_k = x_{k+1} = x_{k+2} < x_{k+3}$ thì

$$\text{rank}\ (x_k) = \text{rank}(x_{k+1}) = \text{rank}(x_{k+2})$$

$$= \frac{1}{3}(k + k + 1 + k + 2) = k + 1$$

a. Tiêu chuẩn Mann – Whitney

Giả sử $(x_1, x_2, \dots, x_{n_1})$ là mẫu ngẫu nhiên rút ra từ X; $(y_1, y_2, \dots, y_{n_2})$ là mẫu ngẫu nhiên rút ra từ Y. Giả thiết H : X và Y thuần nhất (do đó EX = EY); đối thiết K : X và Y không thuần nhất (tức là 2 mẫu trên không phải được rút ra từ một biến ngẫu nhiên). Giả thiết X và Y đều có hàm phân phối liên tục.

Ta sắp xếp $n_1 + n_2$ giá trị từ hai mẫu trên theo thứ tự tăng dần, sau đó tính hạng của từng phần tử x_i, y_j.

Gọi R_i là tổng số hạng của các phần tử trong mẫu i.

Ta tính:
$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

Gọi U là U_1 hoặc U_2

Ta có thể chứng minh được rằng với giả thiết H đúng thì khi n_1 và n_2 tăng lên đại lượng U có phân phối tiến rất nhanh đến phân phối chuẩn với tham số EU, DU xác định như sau:

$$EU = \frac{n_1 n_2}{2} \qquad DU = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Do đó đại lượng $u = \dfrac{U - EU}{\sqrt{DU}} \simeq N(0, 1)$

Từ đó ta có các tiêu chuẩn sau:

– Tính $u = \dfrac{U - EU}{\sqrt{DU}}$

– Với α đã cho tra bảng I phụ lục tìm $u(\frac{\alpha}{2})$ từ phương trình

$$\Phi(u(\frac{\alpha}{2})) = 1 - \frac{\alpha}{2}$$

Miền tiêu chuẩn của bài toán là

$$S = \{|u| \geq u(\frac{\alpha}{2})\} \qquad (17)$$

Nghĩa là nếu S xảy ra thì ta chưa có cơ sở khẳng định EX = EY, còn nếu $|u| \leq u(\frac{\alpha}{2})$ thì ta chấp nhận EX = EY.

Ví dụ 6: Tại một trại chăn nuôi lợn người ta áp dụng thử một loại thuốc tăng trọng bổ sung vào khẩu phần thức ăn của 10 con lợn. Sau ba tháng thu được kết quả sau (X - kg):

62	63	64	60	62	68	65	64
63	61 (kg)						

Trong khi đó 15 con lợn khác không dùng thuốc tăng trọng có trọng lượng sau ba tháng như sau (Y - kg):

56	58	60	56	57	59	60	58	61
62	60	58	59	61	57 (kg)			

Phải chăng EX > EY (với α = 0,05), tức là thuốc tăng trọng đã tăng trọng lượng của lợn.

x_i	rank(x_i)	y_j	rank(y_j)	y_j	rank(y_j)
60	11,5	56	1,5	60	11,5
61	15	56	1,5	60	11,5
62	18	57	3,5	61	15
62	18	57	3,5	61	15
63	20,5	58	6	62	18
63	20,5	58	6		
68	25	58	6	R_2	127,5
64	22,5	59	8,5		
64	22,5	59	8,5		
65	24	60	11,5		
R_1	197,5				

$H: EX = EY$ với $K: EX \neq EY$

$$U_1 = 10.15 + \frac{10.11}{2} - 197{,}5 = 7{,}5$$

$$U_2 = 10.15 + \frac{15.16}{2} - 127{,}5 = 142{,}5$$

$$U = 7{,}5.$$

$$EU = \frac{10.15}{2} = 75$$

$$DU = \frac{10.15.26}{12} = 325,\ \sqrt{325} \approx 18{,}03$$

$$u = \frac{7{,}5-75}{17{,}32} \approx -3{,}744$$

Theo (17) ta có $|u| = 3{,}744 > 1{,}96 = u(\frac{0{,}05}{2})$

Vậy $EX \neq EY$ và rõ ràng $EX > EY$.

b. Tiêu chuẩn hạng được đánh dấu của Wilcoxon

Tiêu chuẩn này cũng dùng để kiểm tra giả thiết như đối với tiêu chuẩn Mann – Whitney trong trường hợp hai mẫu được cho theo từng cặp tương ứng (mẫu phụ thuộc): $(x_1, x_2, ..., x_n)$ và $(y_1, y_2, ..., y_n)$

Xuất phát từ hai mẫu ta tính $d_i = x_i - y_i$, $i = 1, 2, ..., n$.

Tính rank $(|d_i|)$. Bỏ qua các giá trị bằng 0.

Gọi n^+ là số các $d_i \neq 0$

Tính
$$T^- = \sum_{i:d_i<0} \text{rank}(|d_i|)$$

$$T^+ = \sum_{i:d_i>0} \text{rank}(|d_i|)$$

Gọi T là T^- hoặc T^+

Ta có kết quả sau: Nếu giả thiết H đúng ($\mu_1 = \mu_2$) và nếu $n^+ \geqslant 25$ thì T có phân phối xấp xỉ chuẩn với

$$ET = \frac{n^+ \times (n^+ + 1)}{4}, \quad DT = \frac{n^+ \times (n^+ + 1) \times (2n^+ + 1)}{24}.$$

(Thực tế T xấp xỉ chuẩn N(ET, DT) với $n^+ \geqslant 8$ (xem [4])).

Từ kết quả này ta nhận được miền tiêu chuẩn sau:

$$S = \left\{ \frac{|T - ET|}{\sqrt{DT}} \right\} \geqslant u(\frac{\alpha}{2}) \quad (18)$$

Ví dụ 7: Có ý kiến cho rằng loại dầu ăn mới làm giảm lượng cholesterol trong máu của người dùng. Một mẫu ngẫu nhiên đã thu được từ 25 sinh viên. Lượng cholesterol (mg/100g) đã được đo trước khi dùng loại dầu ăn trên (X) và sau ba tháng dùng loại dầu ăn này (Y) đối với 25 sinh viên trên. Kết quả thu được như sau:

x_i:	160	159	157	158	156	155	155	154	154
y_i:	165	162	157	156	153	156	154	157	153
x_i:	155	156	156	159	158	158	156	158	158
y_i:	157	158	154	160	162	161	160	157	155
x_i:	159	159	154	160	159	159	156		
y_i:	156	161	153	158	156	157	156		

Với $\alpha = 0{,}05$ hãy kiểm tra giả thiết EX = EY với đối thiết EX $\neq$ EY.

Từ số liệu trên ta tính $d_i = x_i - y_i$ và rank ($|d_i|$)

x_i	y_i	$d_i = x_i - y_i$	rank $\|d_i\|$	Hạng với dấu âm	Hạng với dấu dương
1	2	3	4	5	6
160	165	-5	23	23	
159	162	-3	17	17	
157	157	0			
158	156	+2	10		10
156	153	+3	17		17
155	156	-1	3,5	3,5	
155	154	+1	3.5		3,5
154	157	-3	17	17	
154	153	+1	3.5		3,5
155	157	-2	10	10	
156	158	-2	10	10	
156	154	+2	10		10
159	160	-1	3,5	3,5	
158	162	-4	21,5	21,5	
158	161	-3	17	17	
156	160	-4	21,5	21,5	
158	157	+1	3,5		3,5
158	155	+3	17		17
159	156	+3	17		17
159	161	-2	10	10	
154	153	+1	3,5		3,5
160	158	+2	10		10
159	156	+3	17		17
159	157	+2	10		10
156	156	0			
				$T^- = 154$	$T^+ = 122$

$$T = T^+ = 122, \ n^+ = 23$$

$$ET = \frac{23 \times 24}{4} = 138 \ , \ DT = \frac{23 \times 24 \times 47}{24} = 1081$$

$$\frac{T - ET}{\sqrt{DT}} = \frac{122-138}{32,88} = -0,4866$$

Vậy ta chấp nhận lượng cholesterol trước và sau khi dùng dầu ăn là như nhau. Ta bác bỏ ý kiến cho là dùng dầu ăn mới này làm giảm lượng cholesterol cho người dùng.

Ví dụ 8: Có ý kiến cho rằng ở một lớp chuyên nọ học sinh đạt điểm thi về khoa học tự nhiên cao hơn điểm thi về khoa học xã hội 5 điểm. Lấy ngẫu nhiên ra 20 học sinh và kết quả tổng điểm thi về khoa học tự nhiên (X) và tổng điểm thi về khoa học xã hội (Y) của họ như sau:

X	Y	Z=X−5	d=Y−Z	rank. $\|d_i\|$	Hạng với dấu âm	Hạng với dấu dương
1	2	3	4	5	6	7
49	44	44	0			
38	33	33	0			
47	40	42	-2	6	6	
43	38	38	0			
35	31	30	+1	3		3
40	38	35	+3	7,5		7,5
45	39	40	-1	3	3	
38	33	33	0			
33	25	28	-3	7,5	7,5	
30	30	25	+5	9		9
40	35	35	0			
43	38	38	0			
35	30	30	0			
45	39	40	-1	3	3	
33	28	28	0			
39	40	34	+6	10		10
38	33	33	0			
48	44	43	+1	3		3
34	30	29	+1	3		3
35	30	30	0			
				$\sum$	T^- = 19,5	T^+ = 35,5

Với α = 0,05 hãy kiểm tra giả thiết EX −5 = EY với đối thiết

$EX - 5 \neq EY$. Ở đây $n^+ = 10$, $T = 19,5$.

$$ET = \frac{10 \times (10 + 1)}{4} = 27,5\ ;\ DT = \frac{10 \times 11 \times 21}{24} = 96,25$$

Theo (18) ta có:

$$\frac{|T - ET|}{\sqrt{DT}} = \frac{|19,5 - 27,5|}{\sqrt{96,25}} = \frac{8}{9,81} < 1,96$$

Vậy chấp nhận giả thiết $EX - 5 = EY$, tức ý kiến đánh giá là đúng.

§ 4. SO SÁNH HAI XÁC SUẤT (HAY HAI TỶ LỆ)

Giả sử ta có hai mẫu ngẫu nhiên: $(X_1, X_2, ..., X_{n_1})$,

trong đó $P(X_i = 1) = p_1$, $P(X_i = 0) = q_1 = 1 - p_1$, $m_1 = \sum_{i=1}^{n_1} X_i$

và mẫu $(Y_1, ..., Y_{n_2})$, trong đó

$$P(Y_j = 1) = p_2,\ P(Y_j = 0) = q_2 = 1 - p_2,\ m_2 = \sum_{j=1}^{n_2} Y_j$$

Giả thiết $H : p_1 = p_2$, đối thiết $K : p_1 \neq p_2$ hoặc $p_1 > p_2$.

Vì $EX_i = p_1$, $EY_j = p_2$, $DX_i = p_1(1-p_1)$, $DY_j = p_2(1-p_2)$ cho nên so sánh hai xác suất p_1, p_2 chính là so sánh hai giá trị trung bình với phương sai chưa biết. Nếu giả thiết H đúng thì $DX_i = DY_j = \sigma^2$. Khi đó

$$D(\overline{X} - \overline{Y}) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

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:

$\dfrac{m_1 + m_2}{n_1 + n_2} \times \left(1 - \dfrac{m_1 + m_2}{n_1 + n_2}\right)$ 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).

Từ đó ta nhận được các miền tiêu chuẩn mức α tương ứng như sau:

$$u = \frac{\dfrac{m_1}{n_1} - \dfrac{m_2}{n_2}}{\sqrt{\dfrac{m_1 + m_2}{n_1 + n_2} \times \dfrac{n_1 + n_2 - m_1 - m_2}{n_1 + n_2}} \times \sqrt{\dfrac{n_1 + n_2}{n_1 . n_2}}} \tag{19}$$

Nếu H: $p_1 = p_2$, $K: p_1 \neq p_2 \Rightarrow S = \{|u| \geqslant u(\frac{\alpha}{2})\}$ (20)

Nếu H: $p_1 = p_2$, $K: p_1 > p_2 \Rightarrow S = \{u \geqslant u(\alpha)\}$ (21)

Ví dụ 9: Có hai loại thuốc A và B cùng điều trị một bệnh nào đó. Qua theo dõi ta thấy trong số 160 người dùng thuốc A có 120 người khỏi bệnh, còn trong số 56 người dùng thuốc B có 40 người khỏi bệnh. Hỏi tác dụng của hai loại thuốc trên trong việc chữa bệnh có như nhau không ($\alpha = 0,05$)?

Ở đây H : $p_1 = p_2$, K : $p_1 \neq p_2$

Theo (19) ta có $u = \dfrac{\dfrac{120}{160} - \dfrac{40}{56}}{\sqrt{\dfrac{160}{216} \times \dfrac{216-160}{216} \times \dfrac{216}{160 \times 56}}} \approx 0,529$

Theo (20): $u = 0{,}529 < 1{,}96 = u(\frac{0{,}05}{2})$

Ta chấp nhận công dụng của hai loại thuốc A và B đối với việc điều trị một bệnh nào đó là như nhau.

§ 5. SO SÁNH HAI PHƯƠNG SAI

Đối với phương sai chúng ta cũng có các bài toán kiểm định giả thiết dạng H : $\sigma^2 = \sigma_o^2$ với các đối thiết $\sigma^2 \neq \sigma_o^2$ hoặc $\sigma^2 > \sigma_o^2$ hoặc $\sigma^2 < \sigma_o^2$ (σ_o^2 đã biết):

Hoặc H : $\sigma_1^2 = \sigma_2^2$ với đối thiết $\sigma_1^2 \neq \sigma_2^2$

Trong kỹ thuật đo đạc, phương sai thể hiện sai số đo đạc, phương sai càng lớn thì kỹ thuật đo đạc càng kém chính xác. Việc so sánh độ chính xác của hai dụng cụ đo thường được đặt ra, do vậy ta sẽ trình bày về bài toán so sánh hai phương sai của hai biến ngẫu nhiên chuẩn.

Giả sử $(X_1, \dots, X_{n_1})$ trong đó $X_i = N(\mu_1, \sigma_1^2)$

$(Y_1, \dots, Y_{n_2})$ trong đó $Y_j = N(\mu_2, \sigma_2^2)$

Để kiểm định giả thiết H : $\sigma_1^2 = \sigma_2^2$ với K : $\sigma_1^2 > \sigma_2^2$ ta tiến hành các bước như sau:

- Xuất phát từ hai mẫu đã cho ta tính $\overline{X}, \hat{s}_1^2, \overline{Y}, \hat{s}_2^2$

- Lập tỷ số:

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} \quad \text{nếu } \hat{s}_1^2 > \hat{s}_2^2$$

hoặc $$F = \frac{\hat{s}_2^2}{\hat{s}_1^2} \quad \text{nếu } \hat{s}_1^2 < \hat{s}_2^2$$

– Với α đã cho, tra bảng VII của phân phối F với $n_1 - 1$ và $n_2 - 1$ bậc tự do (hoặc $n_2 - 1$ và $n_1 - 1$ bậc tự do), tìm được $F_{n_1-1,n_2-1}(\alpha)$ (hoặc $F_{n_2-1,n_1-1}(\alpha)$) (Bảng VII – phụ lục).

(Thông thường bậc tự do ứng với tử số được viết trước và được đọc trên hàng ngang của bảng, còn bậc tự do ứng với mẫu số được đọc ở cột dọc).

– Nếu $F < F_{n_1-1,n_2-1}(\alpha)$ ta chấp nhận giả thiết

$F \geq F_{n_1-1,n_2-1}(\alpha)$ ta bác bỏ giả thiết.

Ví dụ 10: (xem [2]) So sánh hai phương pháp định lượng nitơ cùng tiến hành trên một mẫu. Kết quả được cho như sau:

Theo phương pháp I:

x_i(g)	37	38	39	40	41	42	43	44	45	46	47
m_i:	1	0	2	2	4	7	4	2	0	2	1

Theo phương pháp II:

y_j(g)	39	40	41	42	43	44	45
m_j	2	1	6	9	8	3	1

Hãy so sánh độ chính xác của hai phương pháp trên (α=0,05)

Ta có $n_1 = 25$, $n_2 = 30$. Qua các tính toán quen thuộc ta nhận được:

$$\overline{X} = 42{,}08\ ;\ \hat{s}_1^2 = 5{,}16\ ;\ \overline{Y} = 42{,}10\ ;\ \hat{s}_2^2 = 1{,}96$$

$$F = \frac{5{,}16}{1{,}96} = 2{,}63$$

Tra bảng VII: $F_{24.29}(0,05) = 1,9$

Vậy ta bác bỏ giả thiết và kết luận độ chính xác của hai phương pháp là khác nhau, hơn nữa có thể kết luận phương pháp II chính xác hơn phương pháp I.

§ 6. TIÊU CHUẨN PHÙ HỢP χ^2

Giả sử ta có mẫu ngẫu nhiên $(x_1, ..., x_n)$ rút ra từ biến ngẫu nhiên X. Ta giả sử F(x) chính là hàm phân phối của X, F(x) đã hoàn toàn xác định.

Bài toán đặt ra là:

Giả thiết $X \simeq F(x)$ có đúng không? Nói cách khác số liệu thực nghiệm $(x_1, ..., x_n)$ có phù hợp với giả thuyết lý thuyết $X \simeq F(x)$ hay không? Để giải quyết bài toán trên ta tiến hành các bước sau:

– Chia không gian giá trị của biến X thành k khoảng rời nhau $S_1, S_2, ..., S_k$.

– Đếm m_i số các giá trị quan sát rơi vào khoảng S_i, $\sum_{i=1}^{k} m_i = n$

– Vì F(x) đã hoàn toàn biết (kể cả các tham số cũng đã nhận giá trị cụ thể) ta tính được

$$p_i = P\{X \in S_i\}, \ i = 1, ..., k$$

– Tính tổng $$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} \qquad (22)$$

– Với α đã cho tra bảng phân phối χ – bình phương với k-1 bậc tự do ta tìm được $\chi^2_{k-1}(\alpha)$ sao cho $P\{\chi^2_{k-1} \geq \chi^2_{k-1}(\alpha)\} = \alpha$

- Nếu $\chi^2 \geq \chi^2_{k-1}(\alpha)$ ta bác bỏ giả thiết cho là $X \simeq F(x)$

- Nếu $\chi^2 < \chi^2_{k-1}(\alpha)$ ta coi như $X \simeq F(x)$, tức số liệu thực nghiệm phù hợp với giả thuyết lý thuyết.

Nhận xét: Tiêu chuẩn χ^2 được sử dụng tốt khi n đủ lớn và $m_i \geq 10$, $\forall$ i (có sách ghi $m_i \geq 5$ $\forall$i).

Do vậy từ mẫu đã cho được ghép khoảng, nếu có khoảng nào m_i nhỏ thì ta gộp khoảng đó vào khoảng trước hoặc sau nó. Điều kiện này cho ta biết nên chọn k là bao nhiêu. Dĩ nhiên k > 1 (khoảng đầu và cuối cần điều chỉnh cho phù hợp với miền giá trị của biến ngẫu nhiên). Tương tự đối với mẫu dạng thu gọn.

Cơ sở lý luận của lời giải trên là do Pearson đã chứng minh được khi giả thiết đúng: $X \simeq F(x)$ thì tổng χ^2 có phân phối giới hạn là χ - bình phương với k−1 bậc tự do.

Ví dụ 11: (xem [2]) Người ta cho lai chéo hai giống cây khác nhau bởi hai cặp đặc tính A với a và B với b. Ở thế hệ đầu kết quả thu được khá thuần nhất. Ở thế hệ thứ hai thấy xuất hiện bốn kiểu cây, mà kiểu hình được đánh dấu bằng A−B−; A−bb; aaB−; aabb.

Nếu các đặc tính được di truyền theo luật Menđen thì tỷ số lý thuyết của bốn kiểu hình là 9/16, 3/16, 3/16, 1/16.

Quan sát cụ thể trên 160 cây ta thấy:

Kiểu A−B− có 100 cây, A−bb có 18 cây

Kiểu aaB− có 24 cây và kiểu aabb có 18 cây

Hỏi kết quả quan sát này có phù hợp với luật phân bố Menđen không? Giả thiết H: Kết quả quan sát phù hợp với phân bố Menđen.

Theo phân bố Menđen thì

$P(A\text{-}B\text{-}) = 9/16$, $P(A\text{-}bb) = 3/16$, $P(aaB\text{-}) = 3/16$, $P(aabb) = 1/16$.

Ở đây $k = 4$, $m_1 = 100$, $m_2 = 18$, $m_3 = 24$, $m_4 = 18$ do đó theo (22):

$$\chi^2 = \frac{(100 - 160 \cdot 9/16)^2}{160 \cdot 9/16} + \frac{(18 - 160 \cdot 3/16)^2}{160 \cdot 3/16}$$

$$+ \frac{(24 - 160 \cdot 3/16)^2}{160 \cdot 3/16} + \frac{(18 - 160 \cdot 1/16)^2}{160 \cdot 1/16} \approx 13{,}51$$

Nếu với $\alpha = 0{,}05$ thì $\chi_3^2(0{,}05) \approx 7{,}815$

$$\chi^2 > \chi_3^2(0{,}05)$$

Do đó ta bác bỏ giả thiết, có nghĩa là với độ tin cậy 95% thì các kết quả quan sát trên không phù hợp với phân bố Menđen.

Bây giờ ta mở rộng tiêu chuẩn phù hợp χ^2 cho trường hợp hàm phân phối $F(x, \theta)$, có chứa tham số θ chưa xác định. Giả sử θ là tham ẩn với số chiều r : $\theta = (\theta_1, \theta_2, \dots, \theta_r)$.

Trong trường hợp này các bước tiến hành vẫn như trường hợp trước, song bây giờ xác suất $p_i = P(X \in S_i)$ sẽ phụ thuộc vào θ : $p_i(\theta)$. Vì thế ta phải thay θ bởi ước lượng θ^* nào đó. Fisher đã đưa ra phương pháp minimun χ^2 để tìm θ^*. Trong giáo trình này để đơn giản ta sẽ thay θ bởi các ước lượng điểm mà ta đã biết trong chương II, chẳng hạn $X = N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, $r = 2$, μ thay bởi $\overline{X}$, σ^2 thay bởi s^2 hoặc $\hat{s}^2$.

Nếu $X =$ Poisson với tham ẩn λ, $\theta = \lambda$, $r = 1$ ta thay λ bởi $\overline{X}$

Nếu $X = B(n, p)$, $\theta = p$, $r = 1$ ta thay p bởi $\overline{X}$ hoặc $\frac{m}{n}$.

Nếu $X =$ mũ với tham ẩn λ, $\theta = \lambda$, $r = 1$, ta thay λ bởi $\frac{1}{\overline{X}}$

Nếu X phân phối đều trên $[a, b]$, $\theta = (a, b)$, $r = 2$ ta ước lượng a, b từ hệ thức sau:

$$EX = \frac{a+b}{2} = \overline{X} \qquad a = \overline{X} - \sqrt{3}\,s$$

$$\sqrt{DX} = \frac{b-a}{2\sqrt{3}} = s \qquad b = \overline{X} + \sqrt{3}\,s$$

Khi đó tổng

$$\chi^2(\theta^*) = \sum_{i=1}^{k} \frac{(m_i - np_i(\theta^*))^2}{np_i(\theta^*)}$$

có phân phối giới hạn là χ^2 với k−r−1 bậc tự do. Vậy để kiểm định H: X có phân phối $F(x, \theta)$ và đối thiết K: Phân phối của X không phải là $F(x, \theta)$, ta có miền tiêu chuẩn sau:

$$S = \{\chi^2(\theta^*) \geq \chi^2_{k-r-1}(\alpha)\}$$

Ví dụ 12: Tiến hành đo ngẫu nhiên chiều cao của 100 cây bạch đàn trong khu rừng trồng bạch đàn của một lâm trường ta được các kết quả sau:

Khoảng chiều cao (mét)	Số cây	Khoảng chiều cao	Số cây
8,275 - 8,325	1	8,625 - 8,675	17
8,325 - 8,375	2	8,675 - 8,725	12
8,375 - 8,425	4	8,725 - 8,775	9
8,425 - 8,475	5	8,775 - 8,825	7
8,475 - 8,525	8	8,825 - 8,875	6
8,525 - 8,575	10	8,875 - 8,925	0
8,575 - 8,625	18	8,925 - 8,975	1
		$\sum$	100

Hãy kiểm tra giả thiết cho rằng chiều cao cây bạch đàn ở khu rừng trên tuân theo luật phân phối chuẩn (α = 0,05)?

Từ số liệu trên, chọn điểm đại diện là điểm giữa mỗi khoảng. Thu gọn số liệu bởi phép biến đổi

$$u_i = \frac{x_i - 8,6}{0,05}$$

Qua các tính toán quen thuộc ta nhận được $\overline{X} = 8,63$, $s^2 = 0,01645 \approx (0,128)^2$. Bây giờ ta lập bảng tính χ^2. Thay μ bởi 8,63 ; σ^2 bởi 0,01645

$$H : X = N(8,63;\ 0,01645).$$

Từ các khoảng số liệu đã cho ta giữ nguyên các khoảng có $m_i \geq 5$, gộp 3 khoảng đầu lại, 3 khoảng cuối lại, ta có k = 10 và bảng tính tổng $\chi^2(\theta^*)$ như sau:

Các khoảng $S_i(a_i, a_{i+1})$	m_i	$t_i = \frac{a_i - 8,63}{0,128}$	$\Phi(t_i)$	p_i	np_i	$m_i - np_i$	$\frac{(m_i - np_i)^2}{np_i}$
		$-\infty$					
$(-\infty;\ 8,425)$	7	-1,602	0,054799	0,0548	5,48	1.52	0,4216
(8,425; 8,475)	5	-1,211	0,113139	0,0583	5,83	-0,83	0,1182
(8,475; 8,525)	8	-0,820	0,206108	0,0930	9,3	-1,30	0,1817
(8,525; 8,575)	10	-0,430	0,335598	0.1295	12,95	-2,95	0,6720
(8,575; 8,625)	8	-0,039	0,484047	0,1484	14,84	3,16	0,6729
(8,625; 8,675)	17	0,352	0,636831	0,1528	15,28	1,72	0,1936
(8,675; 8,725)	12	0,742	0,770350	0,1735	17,35	-1,35	0,1365
(8,725; 8,775)	9	1,133	0,870763	0,1004	10,04	-1,04	0,1077
(8,775; 8,825)	7	1,524	0,935744	0,0650	6,5	0,50	0,0385
$(8,825;\ +\infty)$	7	$+\infty$	1	0,0643	6,43	0,57	0,0505
$\sum$	100					$\chi^2(\theta^*)$	=2,5923

Để tính các xác suất p_i ta làm như sau:

$$p_i = P\{a_i < X < a_{i+1}\}$$

$$= P\left\{\frac{a_i - 8,63}{0,128} < \frac{X - 8,63}{0,128} < \frac{a_{i+1} - 8,63}{0,128}\right\}$$

$$= P\left\{t_i < \frac{X - 8,63}{0,128} < t_{i+1}\right\} = \Phi(t_{i+1}) - \Phi(t_i)$$

Với $\alpha = 0,05$, tra bảng IV ta được $\chi^2_7(0,05) = 14,0671$

$$\chi^2(\theta^*) = 2,5932 < 14,0671$$

Vậy ta chấp nhận giả thiết, tức là biến ngẫu nhiên chiều cao cây bạch đàn ở rừng nọ tuân theo luật phân phối chuẩn với độ tin cậy 95%.

§ 7. KIỂM TRA TÍNH ĐỘC LẬP

Giả sử ta có mẫu ngẫu nhiên cỡ n các quan sát đồng thời về hai biến ngẫu nhiên X và Y : (x_1,y_1), (x_2,y_2), ..., (x_n,y_n).

Giả thiết H : X và Y độc lập với nhau

Đối thiết K : X và Y không độc lập

– Ta ghép các giá trị mẫu $(x_1, ..., x_n)$ thành các khoảng, chẳng hạn r khoảng, ghép các giá trị mẫu $(y_1,..., y_n)$ thành s khoảng. Khi đó ta nhận được bảng hai lối vào gồm r.s ô chữ nhật con. Gọi ô (i,j) là ô ở hàng i cột j.

– Đếm số các quan sát từ mẫu đã cho rơi vào ô (i,j). Ký hiệu số đó là n_{ij}, $i = \overline{1..r}$; $j = \overline{1..s}$. Nói cách khác n_{ij} là số các giá trị mẫu mà có giá trị mẫu theo X rơi vào khoảng thứ i và giá trị mẫu theo Y rơi vào khoảng thứ j.

Cần lưu ý rằng các khoảng theo X và các khoảng theo Y không nhất thiết được phân chia theo định lượng mà có thể theo định tính, chẳng hạn tốt, trung bình, xấu hoặc giỏi, khá, trung bình, kém hoặc mầu xanh, đỏ, trắng, vàng...

– Tính $n_{i.} = \sum_{j=1}^{s} n_{ij}$ (lấy tổng theo hàng)

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ (lấy tổng theo cột)

$$n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

– Đối với mỗi ô (i,j) ở trong bảng ta tính $\dfrac{n_{i.} \times n_{.j}}{n}$

Để tiện tính toán ta đặt số này trong ô (i,j) cạnh số n_{ij} nhưng ta đặt trong ngoặc.

- Tính $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \dfrac{(n_{ij} - \dfrac{n_{i.}n_{.j}}{n})^2}{\dfrac{n_{i.}n_{.j}}{n}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \dfrac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right)$ (23)

- Với α đã cho tra bảng IV với (r−1) (s−1) bậc tự do tương ứng tìm được $\chi^2_{(r-1)(s-1)}(\alpha)$.

- Nếu $\chi^2 \geqslant \chi^2_{(r-1)(s-1)}(\alpha)$ ta bác bỏ tính độc lập của X và Y.

(Thực chất của tiêu chuẩn này là ứng dụng tiêu chuẩn phù hợp χ^2 (tìm hiểu chi tiết ở [1])).

X \ Y	1	2	...	j	...	s	$\sum$
1	n_{11}	n_{12}		n_{1j}		n_{1s}	$n_{1.}$
2	n_{21}	n_{22} $\left(\dfrac{n_{2.} \times n_{.2}}{n}\right)$		n_{2j}		n_{2s}	$n_{2.}$
...							
i	n_{i1}	n_{i2}		n_{ij} $\dfrac{(n_{i.} \times n_{.j})}{n}$		n_{is}	$n_{i.}$
...							
r	n_{r1}	n_{r2}		n_{rj} $\dfrac{(n_{r.} \times n_{.j})}{n}$		n_{rs}	$n_{r.}$
$\sum$	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.s}$	n

Ví dụ 13: Nghiên cứu ảnh hưởng của thành phần thức ăn của bố mẹ (X) đối với giới tính (Y) của con cái ta có kết quả sau (số liệu từ [5]):

Y – Giới tính con cái \ X – Thành phần thức ăn	Thiếu vitamin	Đủ vitamin
Trai	123	145
Gái	153	150

Với $\alpha = 0,05$, có thể xem X và Y độc lập với nhau hay không?

Ở đây $r = s = 2$, các n_{ij} đã được cho.

Các bước tính toán thể hiện trong bảng sau:

Y \ X	Thiếu vitamin	Đủ vitamin	$\sum$
Trai	123 (129,54)	145 (138,46)	268
Gái	153 (146,46)	150 (156,54)	303
$\sum$	276	295	571

$$\chi^2 = \frac{(123 - 129,54)^2}{129,54} + \frac{(145 - 138,46)^2}{138,46}$$

$$+ \frac{(153 - 146,46)^2}{146,46} + \frac{(150 - 156,54)^2}{156,54}$$

$$= 0,33 + 0,31 + 0,29 + 0,27 = 1,20$$

Tra bảng IV $\quad \chi_1^2(0,05) = 3,841$

$$\chi^2 < \chi_1^2(0,05)$$

Vậy ta chấp nhận giả thuyết cho rằng thành phần thức ăn của bố mẹ độc lập với giới tính của con cái.

§8. SO SÁNH NHIỀU TỶ LỆ

Trong §4 ta đã so sánh hai tỷ lệ. Bây giờ giả sử ta có s đối tượng (s tỷ lệ), mỗi đối tượng được phân chia theo hai dấu hiệu (hai lớp) chẳng hạn tốt, xấu hoặc khỏi, không khỏi hoặc có tác dụng, không có tác dụng, v.v... Khi đó ta có bảng hai lối vào như trường hợp ở §7 với 2s ô (2 hàng s cột hoặc s hàng 2 cột)

H: s tỷ lệ như nhau : $p_1 = p_2 = ... = p_s$

K : s tỷ lệ khác nhau.

Áp dụng bài toán kiểm tra tính độc lập của hai dấu hiệu đối với s đối tượng. Nếu hai dấu hiệu độc lập với các đối tượng thì điều đó có nghĩa là tỷ lệ đang xét ở s đối tượng là như nhau.

Nếu hai dấu hiệu phụ thuộc vào các đối tượng tức là tỷ lệ đang xét sẽ thay đổi theo các đối tượng.

Vậy lời giải đối với bài toán này hoàn toàn tương tự như đối với bài toán kiểm tra tính độc lập.

$$\text{Tính } \chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{s}\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{.i}n_{.j}}{n}} \qquad (23)$$

Với α đã cho tra bảng tìm $\chi^2_{(2-1)(s-1)}(\alpha) = \chi^2_{s-1}(\alpha)$

Nếu $\chi^2 \geq \chi^2_{s-1}(\alpha)$ ta bác bỏ giả thiết các tỷ lệ như nhau.

Ví dụ 14: So sánh tác dụng của 6 mẫu thuốc thử nghiệm (cùng điều trị một bệnh). Sau một thời gian dài theo dõi tác dụng trên 6 lô chuột song song với một lô đối chứng (không dùng thuốc) ta được kết quả sau:

Kết quả quan sát \ Mẫu thuốc	Mẫu 1	M2	M3	M4	M5	M6	Lô đối chứng
Có tác dụng	40	35	43	28	44	42	0
Không có tác dụng	39	47	34	55	32	39	80
Số bị chết trong thời gian thí nghiệm	21	18	23	17	24	19	20

Hỏi tác dụng dược lý của 6 mẫu thuốc nói trên có khác nhau không ($\alpha = 0{,}05$).

Để giải bài toán này ta phân tích thành các bài toán nhỏ mà đối với nó ta có thể áp dụng tiêu chuẩn χ^2 để so sánh nhiều tỷ lệ.

a. So sánh tỷ lệ chết giữa 6 lô thử thuốc xem có khác nhau không. Nếu có sự khác nhau thì bài toán được dừng lại và câu trả lời đã có. Nếu không khác nhau thì ta tiếp tục bài toán b. sau:

b. So sánh tỷ lệ chết giữa 6 lô thử thuốc và lô đối chứng. Nếu tỷ lệ này khác nhau thì ta dừng lại và bài toán đã được giải, còn nếu ngược lại, tức là tỷ lệ chết hoàn toàn do các nguyên nhân ngẫu nhiên, khi đó ta có thể loại trừ thông tin về số chuột chết trong thời gian thí nghiệm và loại bỏ cột đối chứng, chỉ cần quan tâm đến thông tin về số các tác dụng và không có tác dụng trong 6 lô thử thuốc và ta đi tiếp bài toán c.:

c. Tỷ lệ tác dụng của 6 mẫu thuốc có như nhau không?

Bây giờ ta lần lượt giải các bài toán a. ; b. ; c.

a. H : Tỷ lệ chết trong 6 lô thử thuốc là như nhau.

K : Tỷ lệ chết trong 6 lô thử thuốc là khác nhau.

Mẫu thuốc / Kết quả quan sát	M.1	M.2	M.3	M.4	M.5	M.6	$\sum$
Số sống	79 (79,7)	82	77	83	76	81 (79,7)	478
Số chết	21 (20,3)	18	23	17	24	19 (20,3)	122
$\sum$	100	100	100	100	100	100	600

(các số $\frac{n_{i.}n_{.j}}{n}$ trong cùng một hàng đều bằng nhau).

$$\chi^2 = \frac{1}{79,7}\left\{(79 - 79,7)^2 + (82 - 79,7)^2 + \ldots + (81 - 79,7)^2\right\}$$

$$+ \frac{1}{20,3}\left\{(21 - 20,3)^2 + (18 - 20,3)^2 + \ldots + (19 - 20,3)^2\right\}$$

$$\approx 2,432$$

$$\chi_5^2\,(0,05) = 11,0705 > 2,432$$

Vậy tỷ lệ chết trong 6 lô dùng thuốc là như nhau

b. H : Tỷ lệ chết giữa các lô thử thuốc và lô đối chứng là như nhau.

Kết quả	Lô dùng thuốc	Lô đối chứng	$\sum$
Số sống	478 (477,9)	80 (79,7)	558
Số chết	122 (122,1)	20 (20,3)	142
$\sum$	600	100	700

$$\chi^2 = \frac{(478-477,9)^2}{477,9} + \frac{(80-79,7)^2}{79,7}$$

$$+ \frac{(122-122,1)^2}{122,1} + \frac{(20-20,3)^2}{20,3} \approx 0,006$$

$$\chi_1^2(0,05) = 3,84146 > 0,006 = \chi^2$$

Vậy tỷ lệ chết giữa các lô thử thuốc và lô đối chứng là như nhau.

c. H : Tỷ lệ tác dụng ở 6 mẫu thuốc là như nhau.

Mẫu / Kết quả	M1	M2	M3	M4	M5	M6	$\sum$
Có tác dụng	40 (38,34)	35 (39,8)	43 (37,37)	28 (40,28)	44 (36,8)	42 (39,39)	232
Không có tác dụng	39 (40,66)	47 (42,2)	34 (39,63)	55 (42,72)	32 (39,2)	39 (41,69)	246
$\sum$	79	82	77	83	76	81	478

$$\chi^2 = \frac{(40-38,34)^2}{38,34} + \frac{(35-39,8)^2}{39,8} + \ldots + \frac{(39-41,69)^2}{41,69} \approx 13,21$$

$$\chi_5^2(0,05) = 11,0705$$

$$\chi^2 > \chi_5^2(0,05)$$

Vậy ta bác bỏ giả thiết cho rằng tác dụng của 6 mẫu thuốc là như nhau.

BÀI TẬP CHƯƠNG III

1. Với số liệu đã cho trong bài tập 7 chương II, hãy kiểm định giả thiết EX = EY với đối thiết EX < EY, $\alpha = 0{,}05$.

2. Trọng lượng các bao gạo là biến ngẫu nhiên chuẩn N(50; 0,01). Có nhiều ý kiến khách hàng phản ánh là trọng lượng bị thiếu. Một nhóm thanh tra đã cân ngẫu nhiên 25 bao gạo trong kho, kết quả như sau: (xem [5])

Trọng lượng bao gạo (kg)	48-48,5	48,5-49	49-49,5	49,5-50	50-50,5
Số bao	2	5	10	6	2

Hãy xem ý kiến khách hàng có đúng không bằng cách kiểm tra giả thiết: $\mu = 50$ và đối thiết $\mu < 50$, $\alpha = 0{,}05$.

3. Trong điều kiện chăn nuôi bình thường, lượng sữa trung bình của một con bò là 14kg/1ngày. Nghi ngờ điều kiện chăn nuôi kém đi làm cho lượng sữa giảm xuống, người ta điều tra ngẫu nhiên 25 con và tính được lượng sữa trung bình của một con trong một ngày là 12,5 và độ lệch tiêu chuẩn s = 2,5. Với mức ý nghĩa $\alpha = 0{,}05$ hãy kết luận điều nghi ngờ nói trên. Giả thiết lượng sữa bò là một biến ngẫu nhiên chuẩn (xem [5]).

4. Một máy sản xuất tự động với tỷ lệ chính phẩm 98%. Sau một thời gian hoạt động, người ta nghi ngờ tỷ lệ trên đã bị giảm. Kiểm tra ngẫu nhiên 500 sản phẩm thấy có 28 phế phẩm, với $\alpha = 0{,}05$ hãy kiểm tra xem chất lượng làm việc của máy có còn được như trước hay không?

5. Để so sánh trọng lượng trung bình của trẻ sơ sinh ở thành thị và nông thôn người ta cân thử trọng lượng của 10000 cháu và thu được kết quả sau đây:

Vùng	Số cháu được cân	Trọng lượng trung bình	Độ lệch tiêu chuẩn mẫu
Nông thôn Thành thị	8000 2000	3,0kg 3,2kg	0,3kg 0,2kg

Với mức ý nghĩa $\alpha = 0{,}05$ có thể coi trọng lượng trung bình của trẻ sơ sinh ở thành thị cao hơn ở nông thôn hay không? (Giả thiết trọng lượng trẻ sơ sinh là biến ngẫu nhiên chuẩn).

6. Thống kê số tai nạn lao động tại hai xí nghiệp có các số liệu sau:

Xí nghiệp	Số công nhân	Số tai nạn lao động
I II	200 800	20 120

Với mức ý nghĩa $\alpha = 0{,}05$ hãy kết luận xem chất lượng công tác bảo vệ an toàn lao động tại hai xí nghiệp trên có khác nhau không? (xem [5]).

7. Trồng cùng một giống lúa trên hai thửa ruộng như nhau và bón hai loại phân khác nhau. Đến ngày thu hoạch ta có kết quả như sau: Thửa thứ nhất lấy mẫu 1 000 bông lúa thấy số hạt trung bình mỗi bông $\overline{X} = 70$ hạt và $s_x = 10$.

Thửa thứ hai lấy mẫu 500 bông thấy số hạt trung bình mỗi bông $\overline{Y} = 72$ hạt và $s_y = 20$

Hỏi sự khác nhau giữa $\overline{X}$ và $\overline{Y}$ là ngẫu nhiên hay bản chất ($\alpha = 0{,}05$)? (xem [5])

8. Với số liệu đã cho trong bài tập 1 chương I, hãy kiểm tra giả thiết cho rằng số liệu đã cho phù hợp với phân phối chuẩn $\alpha = 0;05$.

9. Với số liệu đã cho trong bài tập 2 chương II hãy kiểm tra giả thiết cho rằng số liệu cho rút ra từ biến ngẫu nhiên chuẩn ($\alpha = 0,05$).

10. Số con của 2000 phụ nữ thủ đô dưới 25 tuổi cho ở bảng sau:

Số con X	0	1	2	3	4
Số phụ nữ	1090	650	220	30	10

Với mức ý nghĩa $\alpha = 0,05$ có thể xem X tuân theo luật Poisson hay không? (xem [5]).

11. Cùng một loại hạt giống đem xử lý theo 2 phương án khác nhau. Kết quả quan sát chiều cao cây con của mỗi phương án được cho dưới đây (số liệu được trích ra từ [3]).

Phương án I	39,2 37,3	29 27,7	28,5 23,4	33,5 33,4	41,7 29,2	37,2 35,6
Phương án II	20,8 31,0 20,7	33,8 27,4 17,6	28,6 19,5 29,4	23,4 29,6 27,7	22,7 23,2 25,5	30,9 18,7 14,5

Hãy dùng tiêu chuẩn phi tham số để kiểm tra xem 2 phương án xử lý có ảnh hưởng đến sinh trưởng chiều cao cây con hay không ($\alpha = 0,05$).

12. Giả sử ta muốn xác định xem hiệu quả của chế độ ăn kiêng đối với việc giảm trọng lượng như thế nào. 20 người quá béo đã thực hiện chế độ ăn kiêng. Trọng lượng của từng người trước khi ăn kiêng (Xkg) và sau khi ăn kiêng (Ykg) được cho như sau:

X	80	78	85	70	90	78	92	88	75		
Y	75	77	80	70	84	74	85	82	80		
X	75	63	72	89	76	77	71	83	78	82	90
Y	65	62	71	83	72	82	71	79	76	83	81

Dùng tiêu chuẩn phi tham số kiểm tra xem chế độ ăn kiêng có tác dụng làm giảm trọng lượng hay không ($\alpha = 0{,}05$)?

13. Dùng 3 phương án xử lý hạt giống, kết quả cho như sau:

Kết quả	Phương án I	Phương án II	Phương án III
Số hạt mọc	360	603	490
Số hạt không mọc	40	97	180

a. Các phương án xử lý có tác dụng như nhau đối với tỷ lệ nảy mầm hay không ($\alpha = 0{,}05$)?

b. Tìm phương án xử lý tốt nhất (xem [3]) ($\alpha = 0{,}05$)

14. Nghiên cứu sự ảnh hưởng của hoàn cảnh gia đình đối với tình hình phạm tội của trẻ em ở tuổi vị thành niên qua 148 em nhỏ người ta thu được kết quả sau:

Hoàn cảnh gia đình / Tình trạng phạm tội	Bố hoặc mẹ đã chết	Bố mẹ ly hôn	Còn cả bố mẹ
Không phạm tội	20	25	13
Phạm tội	29	43	18

Với mức ý nghĩa $\alpha = 0{,}05$ có thể kết luận là hoàn cảnh gia đình của trẻ em độc lập với phạm tội hay không (xem [5])

15. Theo dõi sự phụ thuộc giữa mầu mắt và mầu tóc ở 124 phụ nữ ở một nước châu Âu ta có kết quả sau:

Mầu mắt \ Mầu tóc	Vàng nâu	Nâu	Đen	Vàng hoe
Xanh	25	9	3	7
Xám	13	17	10	7
Nâu mực	7	13	8	5

Với $\alpha = 0{,}05$ hãy kiểm tra giả thiết cho rằng mầu của tóc và mầu của mắt độc lập với nhau (xem [2]).

16. Để xác định thời vụ phun thuốc diệt sâu có lợi nhất, tổ bảo vệ cây trồng đã theo dõi các lứa sâu trong từng thời kỳ và đếm số sâu non mới nở bắt được. Kết quả ghi ở bảng sau:

Thời kỳ theo dõi	Tháng 1	Tháng 2	Tháng 3	Tháng 4	Tháng 5
Số sâu non mới nở bắt được	62	28	70	75	15
Tổng số sâu bắt được	488	392	280	515	185

a. Tỷ lệ sâu non mới nở trong các thời kỳ quan sát khác nhau có ý nghĩa hay không ($\alpha = 0{,}05$)?

b. Xác định thời kỳ có tỷ lệ sâu mới nở cao nhất. Ước lượng tỷ lệ này. Để ước lượng có sai số không vượt quá 0,02 cần bắt bao nhiêu sâu? ($\alpha = 0{,}05$)

17. Đối với người Việt Nam lượng huyết sắc tố trung bình là 138,3g/l. Khám cho 80 công nhân ở nhà máy có tiếp xúc hoá chất thấy huyết sắc tố trung bình là 120g/l; s = 15g/l. Từ kết quả trên có thể kết luận lượng huyết sắc tố trung bình của công nhân nhà máy này thấp hơn mức chung hay không ($\alpha = 0{,}05$).

18. Đo huyết sắc tố cho 50 công nhân nông trường thấy có 60% ở mức dưới 110g/l. Số liệu chung của khu vực này là 30%

ở mức dưới 110g/l. Với mức ý nghĩa $\alpha = 0,10$ có thể kết luận công nhân nông trường có tỷ lệ huyết sắc tố dưới 110g/l cao hơn mức chung hay không?

19. Hàm lượng đường trong máu của công nhân sau 3 giờ làm việc với máy siêu cao tần đã được đo ở 2 thời điểm trước và sau 3 giờ làm việc. Ta có kết quả sau:

Trước: $n_1 = 50$; $\overline{X} = 60mg\%$; $s_1 = 7$

Sau: $n_2 = 40$; $\overline{Y} = 52mg\%$; $s_2 = 9,2$

Với mức ý nghĩa $\alpha = 0,05$ có thể khẳng định hàm lượng đường trong máu sau 3 giờ làm việc đã giảm đi hay không?

20. Đánh giá tác dụng của một chế độ ăn bồi dưỡng mà dấu hiệu quan sát là số hồng cầu. Người ta đếm số hồng cầu của 20 người trước và sau khi ăn bồi dưỡng:

x_i:	32	40	38	42	41	35	36	47	50
y_i:	40	45	42	50	52	43	48	45	55

x_i:	30	38	45	43	36	50	38	42	41	45	44
y_i:	34	32	54	58	30	60	35	50	48	40	50

Với $\alpha = 0,05$ có thể kết luận 2 dãy số liệu trên là thuần nhất (tức là 2 mẫu đều rút ra từ một biến ngẫu nhiên) hay không?

21. Gọi X là số người tới một trạm điện thoại trong thời gian 3 phút. Theo dõi 50 khoảng thời gian như vậy ta có các số liệu sau:

Số người đến (X)	0	1	2	3	4	5	6
Số khoảng xảy ra	8	15	12	9	4	1	1

Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận X tuân theo luật phân phối Poisson hay không?

22. Giao đồng thời 2 đồng tiền 50 lần. Tần số xuất hiện số mặt sấp được cho như sau:

Số mặt sấp	0	1	2
Tần số xuất hiện	10	28	12

Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận 2 đồng tiền là cân đối và đồng chất hay không?

23. Tiến hành 50 quan sát độc lập về thời gian ngồi uống bia của khách, ta được các số liệu sau:

Khoảng thời gian (T phút)	(0; 5)	[5; 10)	[10; 15)	[15; 20)	≥ 20
Số người	10	20	8	7	5

Với $\alpha = 0,10$ thử xem T có tuân theo luật mũ hay không?

24. Trong đợt thi đua, phân xưởng I báo cáo chất lượng sản phẩm làm ra như sau: có 85% loại I, 10% loại II và 5% loại III. Ban thi đua đã lấy ngẫu nhiên từ lô sản phẩm chưa phân loại của phân xưởng I ra 100 sản phẩm, thấy có 80 loại I, 13 loại II, 7 loại III. Với $\alpha = 0,10$ có thể kết luận gì về báo cáo của phân xưởng I.

25. Một nhà máy có 3 phân xưởng cùng sản xuất một loại sản phẩm. Chất lượng sản phẩm được chia thành 3 loại. Kiểm tra, phân loại ngẫu nhiên một số sản phẩm từ lô sản phẩm của 3 phân xưởng ta có số liệu sau:

Chất lượng \ Phân xưởng	PX I	PX II	PX III
Loại I	70	80	60
Loại II	25	20	15
Loại III	5	10	5

Với α = 0,05 có thể kết luận chất lượng sản phẩm phụ thuộc vào nơi làm ra chúng hay không?

Chương IV

VỀ BÀI TOÁN TƯƠNG QUAN VÀ HỒI QUY

Trong chương này ta sẽ tìm hiểu mức độ phụ thuộc giữa hai biến ngẫu nhiên và tìm biểu thức biểu diễn sự liên hệ giữa chúng.

§1. HỆ SỐ TƯƠNG QUAN MẪU

Trong phần cơ sở xác suất ta đã biết hệ số tương quan giữa hai biến ngẫu nhiên X và Y:

$$\rho = \frac{E(XY)-(EX)(EY)}{\sqrt{EX^2-(EX)^2}\sqrt{EY^2-(EY)^2}}$$

Đó là số đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên X và Y. Nhưng chưa biết phân phối của (X, Y) thì hệ số tương quan lý thuyết ρ cũng chưa tìm được. Do đó ta tìm cách ước lượng ρ theo mẫu quan sát được:

Giả sử ta có mẫu ngẫu nhiên cỡ n về véctơ ngẫu nhiên (X,Y) này: (x_i, y_i), i = 1, 2, ..., n hoặc mẫu thu gọn:

$$\begin{cases} (x_i, y_i) & i=1,2,\ldots,k \\ m_i & \sum_{i=1}^{k} m_i = n \end{cases}$$

Khi đó hệ số tương quan mẫu r được tính theo công thức:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{k} m_i x_i y_i - \overline{X} \times \overline{Y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{k} m_i x_i^2 - \overline{X}^2}\ \sqrt{\frac{1}{n}\sum_{i=1}^{k} m_i y_i^2 - \overline{Y}^2}} \quad (1)$$

hoặc

$$r = \frac{\sum_{i=1}^{k} m_i x_i y_i - n\overline{X} \times \overline{Y}}{\sqrt{\sum_{i=1}^{k} m_i x_i^2 - n\overline{X}^2}\ \sqrt{\sum_{i=1}^{k} m_i y_i^2 - n\overline{Y}^2}} \quad (2)$$

Nếu ta làm phép thu gọn số liệu:

$$u_i = \frac{x_i - x_o}{h}, \quad v_i = \frac{y_i - y_o}{l}$$

thì hệ số tương quan không thay đổi, khi đó

$$r = \frac{\sum_{i=1}^{k} m_i u_i v_i - n\overline{u} \times \overline{v}}{\sqrt{\sum_{i=1}^{k} m_i u_i^2 - n\overline{u}^2}\ \sqrt{\sum_{i=1}^{k} m_i v_i^2 - n\overline{v}^2}} \quad (3)$$

Trong trường hợp phải tính bằng máy tính thô sơ, để tránh nhầm lẫn ta lập bảng tính như sau:

x_i	y_i	m_i	u_i	v_i	m_iu_i	$m_iu_i^2$	m_iv_i	$m_iv_i^2$	$m_iu_iv_i$
x_1 x_2 x_k	y_1 y_2 y_k	m_1 m_2 m_k	u_1 u_k	v_1 v_k					
	$\sum$	n			*	*	*	*	*

Nhìn vào công thức tính r(3) ta thấy cần các số tổng ở 5 cột cuối cùng (chỗ đánh dấu *).

Ví dụ 1: Theo dõi sự phụ thuộc giữa mức suy giảm hàm lượng đường X(%) và thời gian chờ chế biến (t) ta có kết quả sau:

x:	30	30	35	35	40	40	40	45	45	45	50	50
t:	2	4	4	6	4	6	8	6	8	10	8	10
m:	1	1	3	1	1	2	2	2	3	1	1	2

Hãy tìm hệ số tương quan mẫu giữa X và t (Số liệu trích từ [5])

Ta thu gọn số liệu: $u_i = \dfrac{x_i - 40}{5}$; $v_i = \dfrac{t_i - 6}{2}$

Các bước tính toán được thể hiện trong bảng dưới đây:

x_i	t_i	m_i	u_i	v_i	m_iu_i	$m_iu_i^2$	m_iv_i	$m_iv_i^2$	$m_iu_iv_i$
30	2	1	-2	-2	-2	4	-2	4	4
30	4	1	-2	-1	-2	4	-1	1	2
35	4	3	-1	-1	-3	3	-3	3	3
35	6	1	-1	0	-1	1	0	0	0
40	4	1	0	-1	0	0	-1	1	0
40	6	2	0	0	0	0	0	0	0
40	8	2	0	1	0	0	2	2	0
45	6	2	1	0	2	2	0	0	0
45	8	3	1	1	3	3	3	3	3
45	10	1	1	2	1	1	2	4	2
50	8	1	2	1	2	4	1	1	2
50	10	2	2	2	4	8	4	8	8
	$\sum$	20			4	30	5	27	24

Từ (3) ta có

$$r = \frac{24 - 20 \times \frac{4}{20} \times \frac{5}{20}}{\sqrt{30 - 20 \times (0{,}2)^2}\sqrt{27 - 20 \times (0{,}25)^2}}$$

$$\approx \frac{23}{27{,}42} \approx 0{,}8398$$

Với kết quả $r \approx 0{,}84$ ta thấy sự phụ thuộc tuyến tính giữa hàm lượng đường suy giảm và thời gian chờ chế biến là chặt và đồng biến.

§2. ĐƯỜNG HỒI QUY BÌNH PHƯƠNG TRUNG BÌNH TUYẾN TÍNH THỰC NGHIỆM

Khi có sự phụ thuộc tuyến tính giữa hai biến ngẫu nhiên tương đối chặt chẽ ta có thể hy vọng xấp xỉ biến này bởi một hàm tuyến tính của biến kia. Nghĩa là cần tìm biểu thức $aX + b$ sao cho xấp xỉ Y tốt nhất theo nghĩa cực tiểu sai số bình phương trung bình $E(Y - aX - b)^2$.

Ta có:

$E(Y - aX - b)^2 = E\{(Y - EY)$

$-a(X - EX) + EY - aEX - b\}^2$

$= E(Y - EY)^2 + a^2E(X - EX)^2$

$+ (EY - aEX - b)^2$

$- 2aE\{(Y - EY)(X - EX)\} + 0 + 0$

Vế phải sẽ đạt cực tiểu nếu và chỉ nếu:

Tam thức bậc hai theo a: $a^2DX - 2a\rho\sqrt{DX}\sqrt{DY} + DY$ đạt cực tiểu và số hạng $(EY - aEX - b)^2 = 0$.

Do đó ta chọn b = EY - aEX còn a là toạ độ đỉnh của tam thức bậc hai:

$$a = -\frac{-2\rho\sqrt{DX}\sqrt{DY}}{2DX} = \rho\sqrt{\frac{DY}{DX}}$$

Khi đó giá trị nhỏ nhất của vế phải chính là giá trị của tam thức bậc hai theo a tại đỉnh của nó:

$$\min E(Y - aX - b)^2 = \rho^2\frac{DY}{DX}.DX - 2\rho\sqrt{\frac{DY}{DX}}\rho\sqrt{DX}\sqrt{DY}$$

$$+ DY = DY(1 - \rho^2)$$

Vậy biểu thức aX + b cần tìm chính là

$$\rho\sqrt{\frac{DY}{DX}}X + EY - \rho\sqrt{\frac{DY}{DX}}EX$$

Phương trình đường hồi quy bình phương trung bình tuyến tính của Y theo X là:

$$\rho\sqrt{\frac{DY}{DX}}X + EY - \rho\sqrt{\frac{DY}{DX}}EX = Y \qquad (4)$$

hoặc $$Y - EY = \rho\sqrt{\frac{DY}{DX}}(X - EX) \qquad (4')$$

Sai số bình phương trung bình khi dùng đường hồi quy trung bình tuyến tính để xấp xỉ Y là:

$$\sigma^2_{Y/X} = DY(1 - \rho^2) \qquad (5)$$

Sai số này càng nhỏ khi $|\rho|$ càng gần 1 tức là mức độ phụ thuộc tuyến tính giữa hai biến càng chặt.

Tương tự phương trình đường hồi quy bình phương trung bình tuyến tính của X theo Y là

$$X - EX = \rho\sqrt{\frac{DX}{DY}}(Y - EY) \qquad (6)$$

Sai số $$\sigma^2_{x/y} = DX(1-\rho^2) \tag{7}$$

Nhận thấy rằng đường hồi quy trung bình tuyến tính (4) hoặc (6) luôn luôn lập được, miễn là hai biến X và Y tồn tại phương sai dương hữu hạn. Song việc dùng đường hồi quy ấy để xấp xỉ biến này qua biến kia lại là vấn đề khác. Vấn đề đó phụ thuộc vào $|\rho|$ có đủ lớn hay không. Nếu $|\rho|$ bé thì việc xấp xỉ trên sẽ không tốt, khi đó ta không nên dùng.

Xuất phát từ mẫu ngẫu nhiên (x_i,y_i), $i = \overline{1, n}$ ta xây dựng đường hồi quy trung bình tuyến tính thực nghiệm bằng cách thay trong (4') EY bởi $\overline{Y}$, EX bởi $\overline{X}$ và $\rho\sqrt{\dfrac{DY}{DX}}$ bởi $r \times \dfrac{s_y}{s_x}$

$$y - \bar{y} = r \times \frac{s_y}{s_x}(x - \bar{x}) \tag{8}$$

hoặc $$x - \bar{x} = r \times \frac{s_x}{s_y}(y - \bar{y}) \tag{9}$$

và ước lượng sai số $\sigma^2_{y/x}$; $\sigma^2_{x/y}$ bởi

$$s^2_{y/x} = s^2_y(1 - r^2)$$

$$s^2_{x/y} = s^2_x(1 - r^2)$$

Như vậy bằng các kết quả tính toán ở bảng tính hệ số tương quan ta nhận được phương trình đường hồi quy trung bình tuyến tính thực nghiệm.

Trở lại ví dụ 1 ở chương này ta có:

$$\bar{x} = 40 + 5\,.\,\bar{u} = 40 + 5\,.\,4/20 = 41$$

$$\bar{t} = 6 + 2\,.\,\bar{v} = 6 + 2\,.\,5/20 = 6{,}5$$

$$s_x^2 = 5^2 s_u^2 = 25(30/20 - (0,2)^2) = 36,5$$

$$s_t^2 = 2^2 s_v^2 = 4(27/20 - (0,25)^2) = 5,15$$

Vậy theo (4') ta có:

$$x - 41 = 0,84\sqrt{\frac{36,5}{5,15}}\,(t - 6,5)$$

$$x = 2,236t + 26,5$$

$$s_{X/t}^2 = 36,5(1 - 0,84^2) \approx 10,73$$

Dùng đường hồi quy trung bình tuyến tính thực nghiệm nhận được ta có thể tìm được giá trị xấp xỉ của X khi biết t, chẳng hạn $t = 4 \Rightarrow x(4) = 35,444$, số liệu quan sát được là 35, $t = 5$ ta có $x(5) = 37,68$ hoặc $t = 9$ ta có $x(9) = 46,62$.

Ví dụ 2: Ở một vùng có nghề phụ thủ công, quan sát 10 gia đình về 2 tiêu thức: số trẻ em dưới 16 tuổi (X) và thu nhập thêm bằng nghề phụ (Y đơn vị nghìn đồng) thu được số liệu sau:

Gia đình	A	B	C	D	E	G	H	I	K	L
Số trẻ em dưới 16 tuổi	3	5	2	4	4	4	6	1	3	3
Thu nhập Y (nghìn đồng)	58	89	72	71	68	64	98	49	59	62

a. Hãy khảo sát mối tương quan giữa hai tiêu thức trên

b. Xây dựng đường hồi quy bình phương trung bình tuyến tính thực nghiệm của thu nhập theo số trẻ em.

c. Ước lượng sai số bình phương trung bình.

Ở đây $m_i = 1\ \forall i$, ta có bảng tính sau:

x_i	y_i	u_i	u_i^2	y_i^2	u_iy_i
3	58	-1	1	3364	-58
5	89	1	1	7921	89
2	72	-2	4	5184	-144
4	71	0	0	5041	0
4	68	0	0	4624	0
4	64	0	0	4096	0
6	98	2	4	9604	196
1	49	-3	9	2401	-147
3	59	-1	1	3481	-59
3	62	-1	1	3844	-62
$\sum$	690	-5	21	49560	-185

$$u_i = x_i - 4$$

Từ bảng tính ta có:

$$\bar{u} = -5/10 = -0{,}5 \Rightarrow \bar{X} = 4 + \bar{u} = 3{,}5$$

$$\bar{Y} = 690/10 = 69$$

$$s_u^2 = 21/10 - (0{,}5)^2 = 1{,}85 = s_x^2$$

$$s_y^2 = 49560/10 - 69^2 = 195$$

a. $$r = \frac{-185 - 10 \times (-0{,}5) \times 69}{\sqrt{21 - 10 \times (0{,}5)^2}\ \sqrt{49560 - 10 \times 69^2}} = 0{,}842$$

Điều đó chứng tỏ có sự phụ thuộc tuyến tính chặt giữa thu nhập và số trẻ em – sự phụ thuộc là đồng biến.

b. Theo (8):

$$y - 69 = 0{,}842 \times \sqrt{\frac{195}{1{,}85}} \times (x - 3{,}5)$$

$$y = 8{,}6446.x + 39{,}256.$$

c. Tiếp theo ta có ước lượng sai số bình phương trung bình:

$$s_{y/x}^2 = 195 \times (1 - 0{,}842^2) = 56{,}745.$$

§ 3. TỶ SỐ TƯƠNG QUAN MẪU

Hệ số tương quan dùng để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên. Như thế mức độ phụ thuộc phi tuyến chưa được đề cập đến. Mà điều đó càng cần phải biết đến khi nghiên cứu mức độ phụ thuộc giữa 2 biến ngẫu nhiên, nhất là khi mức độ phụ thuộc tuyến tính yếu ($|\rho|$ bé). Vì vậy người ta đã xây dựng một đại lượng khác để đo mức độ phụ thuộc hàm tính bất kỳ. Đại lượng đó được gọi là tỷ số tương quan. Với mức độ của giáo trình này ta không có điều kiện để tìm hiểu về tỷ số tương quan lý thuyết, ý nghĩa và tính chất của nó (vì để làm việc đó ta cần phải biết về kỳ vọng có điều kiện). Ta công nhận một số kết quả sau:

a. Tỷ số tương quan Y theo X được xác định bởi:

$$\Gamma^2_{Y/X} = 1 - \frac{E(Y - E(Y/X))^2}{DY}$$

trong đó E(Y/X) là kỳ vọng có điều kiện của Y với điều kiện X đã cho.

b. $0 \leqslant \Gamma^2_{Y/X} \leqslant 1$, $\Gamma^2_{aY+b/X} = \Gamma^2_{Y/X}$

c. $\Gamma^2_{Y/X}$ dùng để đo mức độ phụ thuộc giữa biến Y và biến X.

$\Gamma^2_{Y/X}$ càng lớn thì sự phụ thuộc càng chặt.

Rõ ràng $\Gamma^2_{Y/X} \geqslant \rho^2$

Bây giờ ta quan tâm đến việc tính tỷ số tương quan mẫu:

Xuất phát từ mẫu ngẫu nhiên (x_i, y_i), $i = 1, 2, \ldots n$, để tính tỷ số tương quan mẫu $R^2_{Y/X}$ ta cần tiến hành các bước sau:

1. Sắp xếp x_i, y_i thành dãy tăng dần:

$$x_{(1)} < x_{(2)} < \ldots < x_{(i)} < x_{(k)}$$
$$y_{(1)} < y_{(2)} < \ldots < y_{(j)} < y_{(l)}$$

2. Đếm n_{ij} là số phần tử mẫu (x_r, y_s) mà có $x_r = x_{(i)}$, $y_s = y_{(j)}$

Khi đó $\displaystyle\sum_{i,j}^{s} n_{ij} = n$

Kí hiệu $\displaystyle n_{i.} = \sum_{j=1}^{l} n_{ij}, \quad n_{.j} = \sum_{i=1}^{k} n_{ij}$

3. Tính $\displaystyle\sum_{j=1}^{l} n_{.j} y_{(j)} \qquad \sum_{j=1}^{l} n_{.j}(y_{(j)})^2$

4. Tính $\displaystyle\sum_{j=1}^{l} n_{ij} y_{(j)}$ với $i = 1, 2, ..., k$

$\displaystyle\frac{1}{n_{i.}}(\sum_{j=1}^{l} n_{ij} y_{(j)})^2$ với $i = 1, 2, ..., k$

$\displaystyle\sum_{i=1}^{k} \frac{1}{n_{i.}}(\sum_{j=1}^{l} n_{ij}\, y_{(j)})^2$

5. Tính:

$$R^2_{Y/X} = \frac{\displaystyle\sum_{i=1}^{k} \frac{1}{n_{i.}}(\sum_{j=1}^{l} n_{ij} y_{(j)})^2 - \frac{1}{n}(\sum_{j=1}^{l} n_{.j} y_{(j)})^2}{\displaystyle\sum_{j=1}^{l} n_{.j}(y_{(j)})^2 - \frac{1}{n}(\sum_{j=1}^{l} n_{.j} y_{(j)})^2}$$

Ta công nhận công thức tính $R^2_{Y/X}$ nêu trên.

Nhận xét: Do tính chất $\Gamma^2_{aY+b/X} = \Gamma^2_{Y/X}$ nên ta có thể thực hiện phép thu gọn số liệu đối với Y: $Z = \dfrac{Y - y_o}{h}$ khi tính $R^2_{Y/X}$

Từ công thức của tỷ số tương quan ta thấy $\Gamma^2_{Y/X} \neq \Gamma^2_{X/Y}$, do đó $R^2_{Y/X} \neq R^2_{X/Y}$, cho nên khi tính $R^2_{X/Y}$ ta cần thay đổi cho thích hợp.

Ta lập bảng tính tỷ số tương quan mẫu $R^2_{Y/X}$ như sau:

$Y_{(j)}$ \ $X_{(i)}$	$X_{(1)}$	$X_{(2)}$	$X_{(i)}$	$X_{(k)}$	$n_{.j}$	$n_{.j}y_{(j)}$	$n_{.j}(y_{(j)})^2$
$Y_{(1)}$	n_{11}	n_{21}	n_{i1}	n_{k1}	$n_{.1}$		
$Y_{(2)}$	n_{12}	n_{22}	n_{i2}	n_{k2}	$n_{.2}$		
$Y_{(j)}$	n_{1j}	n_{2j}	n_{ij}	n_{kj}	$n_{.j}$		
$Y_{(l)}$	n_{1l}	n_{2l}	n_{il}	n_{kl}	$n_{.l}$		
$n_{i.}$	$n_{1.}$	$n_{2.}$	$n_{i.}$	$n_{k.}$	n	*	*
$\sum_{j=1}^{l} n_{ij}y_{(j)}$	?	?	?	?		Tổng theo cột	
$\left\{\sum_{j=1}^{l} n_{ij}y_{(j)}\right\}^2$	?	?	?	?			
$\frac{1}{n_{i.}}\left\{\sum_{j=1}^{l} n_{ij}y_{(j)}\right\}^2$	?	?	?	?	*	Tổng theo hàng	

Ví dụ 3: (xem [2]) Kết quả thu hoạch (Y) theo lượng phân bón (X) của một loại hoa màu trên 100 thửa ruộng được đo như sau (số liệu có tính chất ví dụ minh hoạ) Y, X (kg).

Y \ X	20	25	30	35	40
140	10	8			
150		12	7		
160			28	6	
170				8	9
180					12

Hãy tính tỷ số tương quan mẫu R^2_{XY}?

Ta thực hiện rút gọn số liệu $u_i = \dfrac{x_i - 30}{5}$

Các bước tiến hành nêu ở phần lý thuyết sẽ được thực hiện lần lượt và thể hiện ở bảng tính dưới đây:

$x_{(i)}$	$u_{(i)}$ \ $y_{(j)}$	140	150	160	170	180	$n_{i.}$	$n_{i.}u_{(i)}$	$n_{i.}(u_{(i)})^2$
20	-2	10	0	0	0	0	10	-20	40
25	-1	8	12	0	0	0	20	-20	20
30	0	0	7	28	0	0	35	0	0
35	1	0	0	6	8	0	14	14	14
40	2	0	0	0	9	12	21	42	84
$n_{.j}$		18	19	34	17	12	100	16	158
$\sum_{i=1}^{k} n_{ij}u_{(i)}$		-28	-12	6	26	24			
$\left\{\sum_{i=1}^{k} n_{ij}u_{(i)}\right\}_j^2$		784	144	36	676	576			
$\frac{1}{n_{.j}}\left\{\sum_{i=1}^{k} n_{ij}u_{(i)}\right\}^2$		43,5	7,58	1,06	39,76	48	139,9		

Theo (10) ta có:

$$R^2_{X/Y} = \frac{139{,}9 - \frac{1}{100} \times 16^2}{158 - \frac{1}{100} \times 16^2} = \frac{137{,}34}{155{,}44} \approx 0{,}8836$$

§ 4. MỘT SỐ DẠNG TUYẾN TÍNH HÓA ĐƯỢC

Trong thực tế nếu ta gặp phải những trường hợp $|r|$ nhỏ nhưng $R^2_{Y/X}$ lớn, nghĩa là mức độ phụ thuộc giữa 2 biến là khá chặt nhưng chủ yếu là phi tuyến, thì khi đó việc dùng hồi quy trung bình tuyến tính để xấp xỉ sẽ khó chấp nhận được, song việc đi tìm quan hệ hàm phi tuyến cho phù hợp để xấp xỉ lại là vấn đề khó khăn. Trong phạm vi giáo trình này ta đề cập đến các trường hợp phụ thuộc phi tuyến, nhưng ta có thể đưa về tuyến tính được.

1. *Hồi quy dạng lũy thừa:* $Y = a + bX^m$

Đặt $Z = X^m$, ta có $Y = a + bZ$

Bây giờ thay cho việc ước lượng $\rho(Y, X)$ (mà có thể khá bé) ta ước lượng:

$$\rho(Y, X^m) \text{ hoặc } \rho(Y, Z)$$

Nếu $Y = a + bX^m$ thì $\rho(Y, X^m)$ sẽ khá lớn.

Với $m = -1$ ta có hồi quy Hypecbol: $Y = a + \frac{b}{X}$

Với $m = 2$ ta có hồi quy Parabol: $Y = a + bX^2$

Với $m = 3$ ta có hồi quy lũy thừa 3: $Y = a + bX^3$

Với $m = \frac{1}{2}$ ta có hồi quy căn bậc 2: $Y = a + b\sqrt{X}$

Nếu $a = 0$ ta có hồi quy dạng: $Y = bX^m$

$$\text{Logarit hai vế: } \ln|Y| = m\ln|X| + \ln|b|$$

Do đó nếu $\rho(\ln|Y|, \ln|X|)$ đủ lớn thì ta có thể dùng hồi quy dạng $Y = bX^m$

2. *Hồi quy dạng mũ*: $Y = ae^{bX}$

Ta có thể tuyến tính hóa bằng cách logarit 2 vế:

$$\ln|Y| = \ln|a| + bX$$

3. *Hồi quy dạng Logarit*: $Y = a + b\ln X$

$$\text{Đặt } Z = \ln X \text{ ta có } Y = a + bZ$$

Nhận xét: Thực ra qui luật phụ thuộc giữa các biến rất phức tạp, nhất là trong sinh học. Chọn mô hình toán để mô tả một cách chính xác luôn là bài toán khó giải. Để dự đoán được dạng phụ thuộc sau khi tính hệ số tương quan, tỷ số tương quan, phân tích thông tin nhận được, ta đem chấm các kết quả quan sát được lên trên hệ toạ độ và căn cứ vào chiều hướng của biểu đồ mà ta thêm lượng thông tin về dạng cần tìm.

Ví dụ 4: Để nghiên cứu sự phụ thuộc giữa hai biến ngẫu nhiên nào đó ta tiến hành 30 quan sát độc lập và được kết quả sau:

x_i	1	1,5	2	2	2,5	2,5	3	3	3
y_i	7	9,4	12,8	13	17,6	17,5	23	22,5	22,8
m_i	4	4	2	4	3	5	4	2	2

Phải chăng có sự phụ thuộc dạng $y = a + bx^2$?

Hãy xây dựng đường hồi quy thực nghiệm dạng trên.

Ta đặt $Z = x^2$. Trước hết ta tìm hệ số tương quan mẫu giữa y và x^2. Ta lập bảng tính như sau:

x_i	y_i	m_i	$Z_i = x_i^2$	m_iZ_i	$m_iZ_i^2$	m_iy_i	$m_iy_i^2$	$m_iy_iZ_i$
1	7	4	1	4	4	28	196,0	28,0
1,5	9,4	4	2,25	9	20,25	37,6	353,44	84,6
2	12,8	2	4	8	32	25,6	327,68	102,40
2	13	4	4	16	64	52,0	676,0	208
2,5	17,6	3	6,25	18,75	117,19	52,8	929,28	330
2,5	17,5	5	6,25	31,25	195,31	87,5	1531,25	546,87
3	23	4	9	36	324	92,0	2116,0	828
3	22,5	2	9	18	162	45	1012,5	405
3	22,8	2	9	18	162	45,6	1039,68	410,4
	$\sum$	30		159,0	1080,75	466,1	8181,83	2941,27

Theo (2) ta có:

$$r(Y,Z) = \frac{2941,27 - 30 \times \dfrac{159}{30} \times \dfrac{466,1}{30}}{\sqrt{1080,75 - 30 \times \left(\dfrac{159}{30}\right)^2}\ \sqrt{8181,83 - 30 \times \left(\dfrac{466,1}{30}\right)^2}}$$

$$\approx 0,996$$

Điều này chứng tỏ giữa y và x^2 có sự phụ thuộc tuyến tính rất chặt, tức $y = a + bx^2$

Để xây dựng đường hồi quy trên, theo (8) ta cần tính:

$$\bar{Z} = \overline{X}^2 = \frac{159}{30} = 5,3$$

$$\bar{Y} = \frac{466,1}{30} \approx 15,54$$

$$s_Z^2 = \frac{1080,75}{30} - (5,3)^2 = 7,935$$

$$s_Y^2 = \frac{8181,83}{30} - (15,54)^2 = 31,236$$

$$y - 15,54 = 0,996 \times \sqrt{\frac{31,236}{7,935}} \times (z - 5,3)$$

$$y = 1,976z + 5.07$$

Suy ra: $$y = 1,976x^2 + 5.07$$

Ví dụ 5: Giả sử hàm hồi quy của Y là "giá thành sản phẩm" đối với X là "khối lượng sản phẩm xuất ra" có dạng Hypecbol. Hãy ước lượng hàm hồi quy đó, nếu điều tra ở 30 xí nghiệp cùng loại sản phẩm ta có kết quả sau (số liệu trích từ [5])

Y \ X	50	100	150	200	250
100				4	4
110		2	6	1	1
120	1	4	2		
130	3		1		1
$\sum$	4	6	9	5	6

Theo giả thiết ta có: $Y = a + \frac{b}{X}$. Đặt $\frac{1}{X} = Z$

Ta lập bảng tính sau:

x_i	y_i	m_i	Z_i	v_i	m_iZ_i	$m_iZ_i^2$	m_iv_i	$m_iv_i^2$	$m_iv_iZ_i$
200	100	4	0,005	-1	0,020	0.000010	-4	4	-0.020
250	100	4	0,004	-1	0,016	0.000064	-4	4	-0.016
100	110	2	0.01	0	0,020	0,000200	0	0	0
150	110	6	0,007	0	0,042	0.000294	0	0	0
200	110	1	0,005	0	0,005	0.000025	0	0	0
250	110	1	0,004	0	0,004	0,000016	0	0	0
50	120	1	0,02	1	0,020	0,000400	1	1	0,020
100	120	4	0,01	1	0,040	0,000400	4	4	0.040
150	120	2	0,007	1	0,014	0,000098	2	2	0,014
50	130	3	0,02	2	0,060	0,001200	6	12	0,120
150	130	1	0,007	2	0,007	0,000049	2	4	0,014
250	130	1	0,004	2	0,004	0,000016	2	4	0,008
$\sum$		30			0,252	0.002862	9	35	0,180

$$\text{Đặt } z_i = \frac{y_i - 110}{10}$$

Theo (3) ta có

$$r(Y, Z) = \frac{0{,}18 - \dfrac{0{,}252 \times 9}{30}}{\sqrt{0{,}002862 - \dfrac{0{,}252^2}{30}}\ \sqrt{35 - \dfrac{9^2}{30}}} \approx 0{,}7176$$

Điều đó chứng tỏ có sự phụ thuộc tuyến tính tương đối chặt giữa Y và $\frac{1}{X}$. Quan hệ đồng biến, tức là nếu X giảm thì $\frac{1}{X}$ tăng do đó Y tăng. Quan hệ giữa Y và X lại là nghịch biến.

Để lập đường hồi quy thực nghiệm, theo (8) ta cần tính:

$$\overline{Z} = \frac{0{,}252}{30};$$

$$s_Z^2 = \frac{0{,}002772}{30} - \left(\frac{0{,}252}{30}\right)^2 = 0{,}0000218$$

$$\overline{Y} = 110 + 10 \times \frac{9}{30} = 113;$$

$$s_Y^2 = 10^2 \cdot \left(\frac{35}{30} - \left(\frac{9}{30}\right)^2\right) = 108{,}333$$

$$y - 113 = 0{,}7176\sqrt{\frac{108{,}33}{0{,}0000218}} \times (z - 0{,}0084)$$

$$y = 1599{,}337\,z + 99{,}566$$

$$\text{Vậy } y = 99{,}566 + \frac{1599{,}337}{x}$$

§ 5. HỒI QUY BÌNH PHƯƠNG TRUNG BÌNH TUYẾN TÍNH NHIỀU CHIỀU

1. Hồi quy bình phương trung bình tuyến tính nhiều chiều lý thuyết:

Giả sử $X_1, X_2, ..., X_m$ là các biến ngẫu nhiên có mômen cấp 2 hữu hạn. Ta muốn tìm một hàm tuyến tính của $X_2,...X_m$ dạng $a_2X_2 + ... + a_mX_m + C$ sao cho nó xấp xỉ X_1 tốt nhất theo nghĩa cực tiểu sai số bình phương trung bình, tức là

$E\{X_1 - (a_2X_2 + ... + a_mX_m + C)\}^2$ đạt cực tiểu. Ta sẽ tìm các hằng số $a_2,...,a_m$, C bằng phương pháp bình phương bé nhất.

Ta có $E\{X_1 - (a_2X_2 + ... + a_mX_m + C)\}^2$

$$= E\left\{X_1 - EX_1 - \sum_{k=2}^{m} a_k . (X_k - EX_k) - C + EX_1 - \sum_{k=2}^{m} a_k . EX_k\right\}^2$$

$$=E(X_1 - EX_1)^2 + E\left\{\sum_{k=2}^{m} a_k . (X_k - EX_k)\right\}^2 + (-C + EX_1 - \sum_{k=2}^{m} a_k .EX_k)^2$$

$$-2 \times \sum_{k=2}^{m} a_k E(X_1 - EX_1) \times (X_k - EX_k) + 2E(X_1 - EX_1)$$

$$\times (-C + EX_1 - \sum_{k=2}^{m} a_k . EX_k)^2$$

$$-2 \times \sum_{k=2}^{m} a_k . E(X_k - EX_k) \times (-C + EX_1 - \sum_{k=2}^{m} a_k EX_k)$$

Vì $(-C + EX_1 - \sum_{k=2}^{m} a_k . EX_k)^2 \geq 0$ nó sẽ nhỏ nhất nếu ta chọn

$$C = EX_1 - \sum_{k=2}^{m} a_k . EX_k$$

Ký hiệu: $\lambda_{rs} = E(X_r - EX_r) \times (X_s - EX_s)$

$$E(X_1 - \sum_{k=2}^{m} a_k X_k - C)^2 = \lambda_{11} + \sum_{r=2}^{m}\sum_{s=2}^{m} a_r a_s \lambda_{rs} - 2\sum_{k=2}^{m} a_k \lambda_{1k}$$

Để biểu thức trên đạt cực tiểu các hệ số $a_2, ..., a_k$ phải thỏa mãn hệ phương trình sau (do lấy đạo hàm theo các a_k):

$$\begin{cases} \lambda_{22}a_2 + \lambda_{23}a_3 + ... + \lambda_{2m}a_m = \lambda_{21} \\ \lambda_{32}a_2 + \lambda_{33}a_3 + ... + \lambda_{3m}a_m = \lambda_{31} \\ \quad ..... \\ \lambda_{m2}a_2 + \lambda_{m3}a_3 + ... + \lambda_{mm}a_m = \lambda_{m1} \end{cases} \qquad (11)$$

Đây là hệ m − 1 phương trình tuyến tính với m − 1 ẩn $a_2, ..., a_m$. Giả thiết rằng phân phối đồng thời của $X_1, X_2, ..., X_m$ rải ra trong không gian m chiều và không nằm trong bất kỳ siêu phẳng nào có số chiều ít hơn m, khi đó định thức của hệ trên sẽ dương, do đó hệ có nghiệm duy nhất.

Ký hiệu $\Lambda = (\lambda_{ij})_{m\times m}$ là ma trận mômen của $X_1, X_2, ..., X_m$

Λ^*_{1k} là phần phụ đại số của λ_{1k} trong ma trận mômen Λ.

Ký hiệu: β_{1k} (k = 2, 3, ..., m) là nghiệm của hệ phương trình trên. Theo quy tắc Cramer ta dễ dàng tìm được:

$$\beta_{1k} = -\frac{\Lambda^*_{1k}}{\Lambda^*_{11}}, \quad k = 2, 3, ..., m \qquad (12)$$

$$C = \overline{X}_1 + \sum_{k=2}^{m} \frac{\Lambda^*_{1k}}{\Lambda^*_{11}} \overline{X}_k$$

β_{1k}, k = 2, 3,... m, gọi là các hệ số hồi quy của biến X_1 theo biến X_k tương ứng.

Vậy biểu thức tuyến tính của $X_2, ..., X_m$ xấp xỉ tốt nhất X_1 là:

$$-\sum_{k=2}^{m} \frac{\Lambda^*_{1k}}{\Lambda^*_{11}} \times (X_k - EX_k) + EX_1$$

Để đơn giản khi viết ta ký hiệu biểu thức này là X_1^*. Phương trình hồi quy bình phương trung bình tuyến tính của X_1 theo m - 1 biến còn lại $X_2, ..., X_m$ là:

$$X_1 = -\sum_{k=2}^{m} \frac{\Lambda_{1k}^*}{\Lambda_{11}^*} \times (X_k - EX_k) + EX_1 \qquad (13)$$

Đây chính là một siêu phẳng trong không gian m chiều, nên phương trình (13) còn gọi là siêu phẳng hồi quy bình phương trung bình.

2. Phương sai phần dư:

Sai số bình phượng trung bình nhỏ nhất mắc phải khi xấp xỉ X_1 bởi siêu phẳng hồi quy bình phương trung bình chính là

$$E(X_1 - X_1^*)^2 = E\left\{ X_1 - (-\sum_{k=2}^{m} \frac{\Lambda_{1k}^*}{\Lambda_{11}^*} \times (X_k - EX_k) + EX_1)\right\}^2$$

Ta ký hiệu đại lượng này là $\sigma_{1.2...m}^2$

Ký hiệu $Y_{1.2.3...m} = X_1 - X_1^*$ là phần dư khi xấp xỉ X_1 bởi biểu thức X_1^*. Ta chứng minh được:

$$E(Y_{1.2.3...m} \cdot X_i) = \begin{cases} 0 & \text{nếu } i = 2, 3, ..., m \\ \dfrac{\det\Lambda}{\Lambda_{11}^*} & \text{nếu } i = 1 \end{cases}$$

Tức là phần dư $Y_{1.2.3....m}$ không tương quan với các biến $X_2, ..., X_m$. Để đơn giản ta giả thiết $EX_i = 0$, $\forall i = 1, 2, ... m$. Khi đó:

$$X_1^* = -\sum_{k=2}^{m} \frac{\Lambda_{1k}^*}{\Lambda_{11}^*} \times X_k$$

$$\sigma_{1.2...m}^2 = E(X_1 - X_1^*)^2 = EY_{1.2...m}^2 = DY_{1.2...m}$$

$$= EY_{1.2...m}(X_1 + \sum_{k=2}^{m} \frac{\Lambda_{1k}^*}{\Lambda_{11}^*} \times X_k) = EY_{1.2...m} \times X_1 = \frac{\det\Lambda}{\Lambda_{11}^*}$$

Vậy: $$\sigma_{1.2...m}^2 = \frac{\det\Lambda}{\Lambda_{11}^*} \qquad (14)$$

3. Hệ số tương quan bội:

Bây giờ ta xét mối liên hệ tuyến tính giữa X_1 và X_1^*. Ký hiệu đại lượng đó là $\rho_{1.2...m}$. Ta gọi $\rho_{1.2...m}$ là hệ số tương quan bội.

Theo định nghĩa:

$$\rho_{1.2...m} = \rho(X_1, X_1^*) = \frac{E(X_1X_1^*)}{EX_1^2 \times EX_1^{*2}}$$

(ta cũng giả thiết $EX_i = 0$, $i = 1, 2, ..., m$)

Ta có: $$E(X_1X_1^*) = E\{X_1(X_1 - Y_{1.2...m})\}$$

$$= EX_1^2 + E(X_1Y_{1.2...m}) = \lambda_{11} - \frac{\det\Lambda}{\Lambda_{11}^*}$$

$$E(X_1^*)^2 = E(X_1 - Y_{1.2...m})^2$$

$$= EX_1^2 + EY_{1.2...m}^2 - 2EX_1Y_{1.2...m}$$

$$= EX_1^2 - EY_{1.2...m}^2$$

$$= \lambda_{11} - \frac{\det\Lambda}{\Lambda_{11}^*} \geqslant 0$$

(vì $EY_{1.2...m}^2 = EX_1.Y_{1.2...m}$ (xem (14))

vậy $$\rho_{1.2...m} = \frac{\lambda_{11} - \frac{\det\Lambda}{\Lambda_{11}^*}}{\sqrt{\lambda_{11}}\ \sqrt{\lambda_{11} - \frac{\det\Lambda}{\Lambda_{11}^*}}} = \sqrt{1 - \frac{\det\Lambda}{\lambda_{11} \times \Lambda_{11}^*}} \qquad (15)$$

hoặc $$\rho_{1.2...m} = \sqrt{1 - \frac{\sigma^2_{1.2...m}}{\sigma^2_1}} \tag{16}$$

Nhận xét:

a. Từ biểu thức của nó ta thấy ngay $\rho_{1.2...m} \geq 0$

Do $\lambda_{11} > 0$, $\frac{\det\Lambda}{\Lambda^*_{11}} > 0$ nên $\rho_{1.2...m} \leq 1$

Vậy $0 \leq \rho_{1.2...m} \leq 1$

b. $\rho_{1.2...m} = 1$ khi và chỉ khi $X_1 = aX^*_1 + b$ với xác suất 1, tức X_1 biểu diễn chính xác qua tổ hợp tuyến tính của $X_2,...,X_m$.

c. $\rho_{1.2...m} = 0 \Leftrightarrow \lambda_{11} - \frac{\det\Lambda}{\Lambda^*_{11}} = 0 \Leftrightarrow E(X^*_1)^2 = 0 \Leftrightarrow X^*_1 = 0$ với xác suất 1

Nhưng $\Lambda^*_{11} > 0$ tức phân phối đồng thời của $X_2; X_3,...,X_m$ không thể nằm trong bất kỳ một siêu phẳng nào có số chiều ít hơn m − 1 được, cho nên $X^*_1 = 0$ hay

$\beta_{12}X_2 + ... + \beta_{1m}X_m = 0$ với xác suất 1 khi và chỉ khi

$\beta_{12} = \beta_{13} = ... = \beta_{1m} = 0$

Nghiệm duy nhất của hệ phương trình tuyến tính lại bằng 0 thì vế phải của hệ đó (xem (11)) phải bằng 0, tức là $\lambda_{1k} = 0$, k = 2, 3,...,m. Điều đó tương đương với việc X_1 không tương quan với các biến còn lại.

$$\rho_{1.2...m} = 0 \Leftrightarrow \rho(X_1, X_2) = \rho(X_1, X_3) = ... = \rho(X_1, X_m) = 0$$

4. Hồi quy bình phương trung bình tuyến tính nhiều chiều thực nghiệm:

Giả sử ta có mẫu ngẫu nhiên cỡ n: $(x_{1i}, x_{2i},..., x_{mi})$, i = 1,...,n

$$\text{hoặc mẫu thu gọn}\begin{cases}(x_{1i}, x_{2i}, \dots, x_{mi}) & i = 1, \dots, n \\ m_i & \sum_{i=1}^{k} m_i = n\end{cases}$$

Vì các hệ số hồi quy, phương sai phần dư và hệ số tương quan bội đều biểu diễn qua ma trận mômen Λ, cho nên trước tiên ta tìm các mômen mẫu $\bar{\lambda}_{ij}$ và các kỳ vọng mẫu $\bar{x}_i$, trong đó

$$\bar{\lambda}_{ij} = \frac{1}{n} \times \sum_{k=1}^{n} (x_{ik} - \bar{x}_i) \cdot (x_{jk} - \bar{x}_j)$$

$$= \frac{1}{n} \times \sum_{l=1}^{n} x_{il} \times x_{jl} - \bar{x}_i \cdot \bar{x}_j, \quad 1 \leq i \leq j \leq n$$

hoặc $\bar{\lambda}_{ij} = \frac{1}{n}\sum_{l=1}^{k} m_l \; x_{il} \cdot x_{jl} - \bar{x}_i \cdot \bar{x}_j$

Thay λ_{ij} bằng ước lượng $\bar{\lambda}_{ij}$ của nó ta lập được ma trận mômen mẫu $\bar{\Lambda}$. Từ đó ta nhận được ước lượng điểm cho các hệ số hồi quy β_{1k}, hệ số tương quan bội thực nghiệm, ước lượng điểm cho phương sai phần dư và phương trình siêu phẳng hồi quy bình phương trung bình tuyến tính nhiều chiều thực nghiệm.

Ví dụ 6: (xem [3]) Số liệu điều tra về thể tích ($V(m^3)$), chiều cao (h (mét)) và đường kính (d (mét)) ở độ cao 1,3m của 15 cây thông mã vĩ tuổi 30 ở Cẩm Phả được cho trong 3 cột đầu của bảng dưới đây (số liệu điều tra của bộ môn điều tra quy hoạch trường Đại học Lâm nghiệp).

Sự phụ thuộc giữa V và h, d được đưa ra ở dạng sau:

$V = a.h + b.h.d^2 + c$, a, b, c = các hệ số cần tìm

Đặt $X_1 = h$, $X_2 = h.d^2$

$V = a.X_1 + b.X_2 + c$

X_1	X_2	V	X_1^2	X_2^2	V^2	X_1V	X_2V	X_1X_2
15,5	1,06398	0,30761	240,2	1,13205	0,09462	4,76803	0,32779	16,49172
15,8	1,74153	0,01417	249,6	3,03295	0.37720	9,70391	1,06960	27,51634
13,0	0,57541	0,19722	176,8	0.33109	0.03889	2.62314	0,11348	7,65296
15,5	0,67059	0,27117	240,2	0.44969	0.07353	4.20313	0.18184	10,39417
13,4	0,56323	0,24969	180.9	0.31949	0.06234	3.35834	0.14113	7,62242
11,6	0,42762	0,17823	134,5	0,18286	0,03176	2,06751	0,07621	4.96041
15,0	1,02966	0,34115	225,0	1.06019	0,11638	5,11734	0.35127	15,44490
14,9	0,60197	0,16309	222,0	0.36337	0,02659	2,43011	0,09817	8,96942
12,6	0,33198	0,14771	160,0	0,11021	0,02182	1,86860	0,04903	4,19962
13,3	0,51323	0,18691	178,4	0.26341	0,03493	2,49714	0.09593	6,85686
13,1	0,50325	0,18764	171,6	0.25326	0,03520	2,45807	0,09443	6,59257
14,7	0,80765	0,33352	217,5	0.65230	0,11124	4,91953	0,26937	11,91285
14,4	1,08109	0,42422	207,3	1,16876	0,17996	6,10882	0,45862	16,56775
15,7	0,68579	0,28644	246,4	0.47031	0,08205	4,49718	0,17644	10,70693
13,2	0,65054	0,25641	174,2	0,42320	0,06574	3,38469	0,16681	8,58719
212,2	11,2495	4,14524	3025	10,2121	1,35233	60,0056	3.69018	163,5162

Trong trường hợp này $m_i = 1$, $\forall i$. Căn cứ vào bảng tính trên ta có:

$$\overline{X}_1 = \frac{212,21}{15} = 14,14$$

$$\overline{X}_2 = \frac{11,249527}{15} = 0,749971$$

$$\overline{V} = \frac{4,145244}{15} = 0,276349$$

$$\lambda_{11} = s_v^2 = \frac{1,352355}{15} - \overline{V}^2 = 0,0137866$$

$$\lambda_{22} = s_{x1}^2 = \frac{3025,2771}{15} - \overline{X}_1^2 = 1,5380996$$

$$\lambda_{33} = s_{x2}^2 = \frac{10,212182}{15} - \overline{X}_2^2 = 0,1183545$$

$$\lambda_{12} = \frac{60{,}005629}{15} - \overline{X}_1 . \overline{V} = 0{,}0907654$$

$$\lambda_{13} = \lambda_{vx2} = \frac{3{,}690185}{15} - \overline{X}_2 . \overline{V} = 0{,}0320914$$

$$\lambda_{23} = \lambda_{x1x2} = \frac{163{,}51621}{15} - \overline{X}_1 . \overline{X}_2 = 0{,}2909784$$

$$\overline{\Lambda} = \begin{pmatrix} 0{,}0137866 & 0{,}0907654 & 0{,}0320914 \\ 0{,}0907654 & 1{,}5380996 & 0{,}2909784 \\ 0{,}0320914 & 0{,}2909784 & 0{,}1183545 \end{pmatrix}$$

Từ đó $\det\overline{\Lambda} = +0{,}0004672$

$$\overline{\Lambda}^*_{11} = 0{,}096803 \,;\; \overline{\Lambda}^*_{12} = -0{,}001426 \,;\; \overline{\Lambda}^*_{13} = -0{,}022735$$

Ước lượng cho các hệ số hồi quy β_{1k} sẽ là (xem (12))

$$a = -\frac{\overline{\Lambda}^*_{12}}{\overline{\Lambda}^*_{11}} = -\frac{-0{,}001426}{0{,}096803} = 0{,}0147314$$

$$b = -\frac{\overline{\Lambda}^*_{13}}{\overline{\Lambda}^*_{11}} = -\frac{-0{,}022735}{0{,}096803} = 0{,}2348657$$

$$c = \overline{V} - (a\overline{X}_1 + b\overline{X}_2)$$

$$= 0{,}276349 - (0{,}0147314.14{,}14 + 0{,}2348657.0{,}749971)$$

$$= -\,0{,}108095$$

Vậy phương trình siêu phẳng hồi quy cần tìm là:

$$V = 0{,}0147314.X_1 + 0{,}2348657.X_2 - 0{,}108095$$

hoặc $\quad V = 0{,}0147314.h + 0{,}2348657.h.d^2 - 0{,}108095$

Theo (14) ước lượng cho $\sigma^2_{1.2.3}$ là:

$$s^2_{1.2.3} = \frac{\det\overline{\Lambda}}{\overline{\Lambda}^*_{11}} = \frac{0{,}0004672}{0{,}0968} = 0{,}0048264$$

Theo (16) ước lượng cho $\rho_{1.2.3}$ là:

$$r_{1.2.3} = \sqrt{1 - \frac{0{,}0048264}{0{,}0137866}} \approx 0{,}806$$

5. Một số dạng phi tuyến có thể đưa về tuyến tính:

Cũng như trong trường hợp giữa hai biến, trường hợp nhiều biến cũng có thể tuyến tính hóa được trong một số trường hợp, chẳng hạn:

a. $y = a + b.x + c.x^2$

Đặt $X_1 = y, \quad X_2 = x, \quad X_3 = x^2$

$X_1 = a + b.X_2 + c.X_3$

b. $y = a.x^b.Z^c$

Logarit hai vế, đặt $X_1 = \ln y$, $X_2 = \ln x$, $X_3 = \ln Z$ ta có

$$\ln y = b.\ln x + c\ln Z + \ln a$$

$$X_1 = bX_2 + c.X_3 + \ln a$$

Dạng tổng quát hơn: $y = a.x^b.Z^c.u^d.v^h$ cũng được đưa về một cách tương tự.

Trong lâm nghiệp quan hệ giữa thể tích thân cây với chiều cao và đường kính tuân theo dạng trên. Ví dụ 4 là một minh hoạ cho điều nhận xét này.

c. $y = a + \frac{b_1}{Z_1} + ... + \frac{b_m}{Z_m}$. Đặt $X_i = \frac{1}{Z_i}$, $i = 1, ..., m$

$$y = a + b_1 . X_1 + ... + b_m . X_m$$

d. $y = a + b_1 . Z_1 + b_2 . Z_2 + b_3 . Z_1^2 + b_4 . Z_2^2 + b_5 . Z_1 . Z_2$

(Hồi quy dạng toàn phương)

Đặt $Z_1 = X_1$, $Z_2 = X_2$, $Z_1^2 = X_3$, $Z_2^2 = X_4$, $Z_1 . Z_2 = X_5$

Ta có $y = a + b_1.X_1 + b_2.X_2 + b_3.X_3 + b_4.X_4 + b_5.X_5$

e. $y = e^{a + b_1 . Z_1 + ... + b_m . Z_m}$ (Hồi quy mũ)

Ta có $\ln y = a + b_1.Z_1 + ... + b_m.Z_m$

Thay cho việc xét y ta xét lny.

g. $y = \dfrac{1}{a + b_1 . Z_1 + ... + b_m Z_m}$

Suy ra $\frac{1}{y} = a + b_1 . Z_1 + ... + b_m Z_m$. Ta xét $\frac{1}{y}$ thay cho y

BÀI TẬP CHƯƠNG IV

1. Nghiên cứu mối liên hệ giữa X là số tiền đầu tư cho việc phòng bệnh tính trên đầu người và Y là tỷ lệ người mắc bệnh ở 50 địa phương thu được bảng tương quan thực nghiệm sau:

X đồng \ Y(%)	2	2,5	3	3,5	4
100				2	3
200			3	6	2
300		4	6	3	
400	1	6	4	1	
500	6	3			

(Số liệu trích từ [5])

a. Tìm hệ số tương quan mẫu

b. Xây dựng đường hồi quy tuyến tính thực nghiệm của Y theo X.

c. Nếu năm sau đầu tư cho phòng bệnh là 600đ/ người thì tỷ lệ mắc bệnh khoảng bao nhiêu %?

d. Ước lượng phương sai phần dư.

2. Qua nhiều tác giả nghiên cứu cho thấy giữa lượng đạm (N) và cácbon (C) trong mùn có liên hệ với nhau theo dạng tuyến tính. Hãy xác nhận lại nhận định trên qua ví dụ về đất lâm nghiệp ở Quảng Ninh sau đây (số liệu trích từ bộ môn Đất trường Đại học Lâm nghiệp) (xem [3]):

Hàm lượng C:	1,79	4,39	3,07	4,40	3,10	5,60
Hàm lượng N:	0,06	0,42	0,18	0,30	0,22	0,38
C:	7,81	3,95	4,71			
N:	0,46	0,23	0,42			

Tìm hệ số tương quan mẫu và xây dựng đường hồi quy tuyến tính thực nghiệm của N theo C.

3. Quan sát 2 biến ngẫu nhiên X và Y ta có số liệu sau (trích từ [5]):

Y \ X	7	8	9
200	41	7	
300	1	52	1
400		8	40

a. Tìm hệ số tương quan mẫu

b. Tìm tỷ số tương quan mẫu của Y theo X

c. Xây dựng đường hồi quy thực nghiệm bậc 2 của Y đối với X?

4. Giả sử X : Mức giá (nghìn đồng)

Y : Tổng thu nhập về loại hàng đó (triệu đồng)

Z : Tổng cung cấp (triệu đồng).

Ta có các số liệu sau: (mỗi bộ ba số liệu được cho tương ứng với nhau)

X:	10	10,5	11,0	11,5	12,0	12,5	13,0
	13,5	14,0	14,5				
Y:	101,5	55,2	40,8	28,5	21,1	16,2	14,1
	13,0	12,1	11,5				
Z	3,9	5,0	8,2	8,6	15,1	30,5	48,7
	70,3	100,5	120,4				

a. Giả sử sự phụ thuộc giữa nhu cầu (Y) với giá cả (X) có dạng Hypecbol, còn sự phụ thuộc của cung cấp (Z) vào giá cả (X) có dạng tuyến tính.

- Hãy ước lượng $R^2_{Y/X}$ và r(Z, X)

- Hãy xây dựng các phương trình hồi quy bình phương trung bình thực nghiệm giữa Y với X và giữa Z với X.

b. Dựa vào các phương trình nhận được trên hãy dự đoán mức giá ổn định của thị trường đối với loại hàng hóa trên và minh hoạ bằng đồ thị (xem [5])

HƯỚNG DẪN VÀ ĐÁP SỐ

Chương II

1.b. Đây là ước lượng khoảng đối với giá trị trung bình trường hợp phương sai chưa biết, hoặc dùng khoảng xấp xỉ như nhận xét (thử dùng cả 2 khoảng rồi so sánh để rút ra nên dùng khoảng nào)

c. Ước lượng điểm và mút bên trái của ước lượng khoảng đối với tỷ lệ

2.b. Mút trái và mút phải của khoảng tin cậy đối với giá trị trung bình (trường hợp phương sai chưa biết)

3. Nếu có giả thiết chuẩn ta dùng ước lượng khoảng đối với giá trị trung bình trường hợp phương sai chưa biết. Nếu không có tính chuẩn ta dùng khoảng xấp xỉ.

Với giả thiết chuẩn a. $n \geq 56$. b. $n \geq 36$

(Cần điều tra bổ sung 5 ô nữa).

c. $$\frac{t_{30}(\frac{\alpha}{2}) \cdot 1{,}45}{\sqrt{30}} < 0{,}5 \Rightarrow t_{30}(\frac{\alpha}{2}) < 1{,}888$$

Cho nên độ tin cậy > 90%. Khoảng tin cậy là ($9{,}7m^3$ - $10{,}7m^3$)

4. (10,046l - 10,222l) ($\overline{X} = 10{,}134$; $s^2 = 0{,}053644$)

5. Gọi N là số cá trong hồ. $p = \frac{2000}{N}$ là tỷ lệ cá đánh dấu. $p^* = \frac{80}{400} = 0,2$. Ta tìm được khoảng tin cậy cho p, từ đó nhận được khoảng tin cậy cho N (8369; 12423)

6. $p^* = 0,3$; p_{max} là 0,352; $n \geq 897$; Độ tin cậy là 90%

7. a. Áp dụng trường hợp giả thiết chuẩn thỏa mãn, phương sai chưa biết $\overline{X} = 1111,4$; $s_X^2 = 1402,4 = (37,44)^2$; $\overline{Y} = 1175,5$; $s_Y^2 = 204,75$

b. Mút dưới của ước lượng khoảng theo công thức (6) 56,3

c. Độ tin cậy là 90%

d. $n \geq 209 \Rightarrow$ cần quan sát thêm 105 bóng nữa.

Chương III

1. Áp dụng nhận xét

3. H : $\mu = 14$/K : $\mu < 14$; $\alpha = 0,05$ theo (6)

$t = -2,94$ nên chấp nhận điều nghi ngờ nêu ra

4. $p = 0,98$/ $p < 0,98$; $\alpha = 0,05$; $u = -5,8$ nên $p < 98\%$

5. $\mu_1 = \mu_2 / \mu_1 < \mu_2$; $\alpha = 0,05$. Áp dụng nhận xét $(\mu_1 < \mu_2)$

6. $p_1 = p_2$/ $p_1 \neq p_2$; $\alpha = 0,05$ $u = 0,576 \Rightarrow p_1 = p_2$

7. Sự khác nhau giữa $\overline{X}$ và $\overline{Y}$ là bản chất (Dùng nhận xét ta tính được $2,106 > 1,96$)

8. Gộp 2 số liệu đầu, 2 số cuối $\Rightarrow$ số khoảng $k = 8$

9. Gộp 2 số liệu cuối, số khoảng $k = 8$

10. $\chi^2 = 9,74$ nên X không tuân theo luật Poisson

11. Dùng tiêu chuẩn Mann-Whitney ($R_1 = 260$, $U_1 = 34$, $u = -3,13$)

12. Dùng Wilcoxon. $T^+ = 147,5$; $T^- = 23,5$. $|u| = 2,744$. nên chưa có cơ sở để kết luận 2 mẫu thuần nhất, tức là chưa có cơ sở kết luận EX = EY. Phần lớn các giá trị d_i là dương và $\overline{X} > \overline{Y}$ nên EX > EY, tức là có giảm đi.

13. a. $\chi^2 \approx 61$ nên tỷ lệ hạt mọc của ba phương án là khác nhau

b. Ta có $p_1^* > p_2^* > p_3^*$

Do đó trước hết ta kiểm tra $p_1 = p_2$ / $p_1 > p_2$; $\alpha = 0,05$ ta nhận được $p_1 > p_2$. Suy ra p_1 là max - phương án I là tốt nhất (giá trị của u = 1,9)

14. So sánh 3 tỷ lệ phạm tội hoặc kiểm tra sự độc lập giữa hoàn cảnh gia đình và tình trạng phạm tội. $\chi^2 = 0,2573$ nên chấp nhận sự độc lập

16. a. So sánh 5 tỷ lệ sâu non. $\chi^2 \approx 50,6$ nên kết luận các tỷ lệ khác nhau (Lưu ý tìm số sâu không non bắt được)

b. Ta có $p_3^* > p_4^* > p_1^* > p_5^* > p_2^*$

Do đó cần kiểm tra $p_3 = p_4$/ $p_3 > p_4$; $\alpha = 0,05$. Tính được u = 3,64 nên kết luận $p_3 > p_4$. Suy ra p_3 là max

17. $\mu = 138 / \mu < 138$; $\alpha = 0,05$. Áp dụng nhận xét (u = -8,54)

18. p = 0,3/ p > 0,3; $\alpha = 0,10$. (u = 10 > 1,28)

19. So sánh 2 trung bình (áp dụng nhận xét)

20. Dùng tiêu chuẩn Wilcoxon T = T' = 27; u = 2,913

21. Cần gộp khoảng cuối là {X ≥ 4}, suy ra k = 5 khoảng

22. Giả thiết 2 đồng tiền cân đối đồng chất tương đương với phân phối xác suất $P(X = 0) = P(X = 2) = \frac{1}{4}$

$P(X = 1) = 0,5$; trong đó X là biến ngẫu nhiên chỉ số mặt sấp xuất hiện. Nói cách khác là kiểm tra sự phù hợp giữa số liệu quan sát với giả thiết X có phân phối xác định như trên (cũng chính là phân phối nhị thức B(2; 0,5))

23. Kiểm tra sự phù hợp giữa số liệu quan sát với luật phân phối mũ

24. Kiểm tra sự phù hợp giữa số liệu quan sát với một phân phối hoàn toàn xác định về chất lượng sản phẩm ($\chi^2 \approx 2,0$)

25. Kiểm tra tính độc lập giữa chất lượng sản phẩm và nơi sản xuất ra chúng ($\chi^2 = 12,63$).

Chương IV

Các dạng bài và dạng câu hỏi đã rõ ràng. Cách làm hoàn toàn tương tự như ví dụ đã xét.

TÀI LIỆU THAM KHẢO

1. Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như. *Thống kê Toán học*. NXB Đại học và THCN, Hà Nội 1984.

2. Lê Khánh Trai, Hoàng Hữu Như. *Ứng dụng xác suất Thống kê trong Y, Sinh học*. NXB Khoa học và Kỹ thuật, Hà Nội 1979.

3. Nguyễn Hải Tuất. *Thống kê Toán học trong lâm nghiệp*. NXB Nông nghiệp, Hà Nội 1982.

4. Lincoln L. Chao. *Statistics: Methods and Analyses*. International Student Edition, 1969.

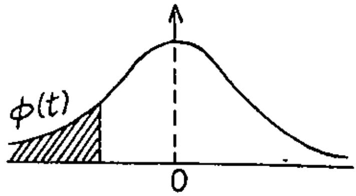
5. Nguyễn Đình Cử, Trương Giêu. *Bài tập xác suất và Thống kê Toán học*. Đại học Kinh tế Quốc dân - 1992.

PHỤ LỤC

Bảng I. Hàm phân bố chuẩn

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}x^2} dx$$

(t từ − 3.9 đến 0)



t	0	1	2	3	4	5	6	7	8	9
-0.0	0.5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
4	3446	3409	3372	3336	5300	3264	3228	3192	3156	3121
-0.5	3.3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
7	2420	2389	2358	2327	2297	2266	2236	2206	2177	2148
8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
-1.0	0.1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1	1357	1335	1314	1292	1291	1251	1230	1210	1190	1170
2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
3	0968	0951	0934	0918	9001	0885	0869	0853	0838	0823
4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
-1.5	0.0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
8	0359	0351	0344	0336	0329	0322	0314	0317	0301	0294
9	0288	0281	0274	0268	0262	0256	0250	0244	0239	0233
-2.0	0.0227	0222	0217	0212	0207	0202	0197	0192	0188	0183
1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
-2.5	0.0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
t	-3.0	-3.1	-3.2	-3.3	-3.4	-3.5	-3.6	-3.7	-3.8	-3.9
Φ(t)	0.0013	0010	0007	0005	0003	0002	0002	0001	0001	000

Bảng I. Hàm phân bố chuẩn

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}x^2} dx$$

(t từ 0 đến +3.9)

t	0	1	2	3	4	5	6	7	8	9
0.0	0,5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
7	7580	7611	7642	7673	7703	7734	7764	7794	7823	7852
8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	0,8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1	8643	8665	8686	8708	8709	8749	8770	8790	8810	8830
2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1,5	0,9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
8	9641	9649	9656	9664	9671	9678	9686	9683	9699	9706
9	9712	9719	9726	9732	9738	9744	9750	9756	9761	9767
2,0	0,9773	9778	9783	9788	9793	9798	9803	9808	9812	9817
1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
4	9918	9920	9922	9905	9927	9929	9931	9932	9934	9936
2.5	0.09938	9940	9941	9943	9945	9946	9948	9949	9951	9952
6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986

t	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
Φ(t)	0.9987	9990	9993	9995	9996	9997	9998	9999	9999	9999

Bảng II. Giá trị tới hạn tn(P) để bác bỏ giá trị "đột xuất" x^* (n là số các kết quả nhận được, P là độ tin cậy của kết luận)

n	P			
	0,95	0,98	0,99	0,999
5	3.04	4.11	5.04	9.43
6	2.78	3.64	4.36	7.41
7	2.62	3.36	3.96	6.37
8	2.51	3.18	3.71	5.73
9	2.43	3.05	3.54	5.31
10	2.37	2.96	3.41	5.01
11	2.33	2.89	3.31	4.79
12	2.29	2.83	3.23	4.62
13	2.26	2.78	3.17	4.48
14	2.24	2.74	3.12	4.17
15	2.22	2.71	3.08	4.28
16	2.20	2.68	3.04	4.20
17	2.18	2.66	3.01	4.13
18	2.17	2.64	2.98	4.07
20	1.145	2.602	2.932	3.979
25	2.105	2.541	2.852	3.819
30	2.079	2.403	2.802	3.719
35	2.061	2.476	2.768	3.652
40	2.048	2.456	2.742	3.602
45	2.038	2.441	2.722	3.565
50	2.030	2.429	2.707	3.532
60	2.018	2.411	2.683	3.492
70	2.009	2.399	2.667	3.462
80	2.003	2.389	2.665	3.439
90	1.998	2.382	2.646	3.423
100	1.994	2.377	2.639	3.409
∞	1.960	2.326	2.576	3.291

Bảng III. Giá trị $t_k(\alpha/2)$ và $t_k(\alpha)$ của phân phối Student

Bậc tự do K	Mức ý nghĩa α (trường hợp 2 phía)					
	0,10	0.05	0,02	0,01	0,002	0,001
1	6,31	12,71	31,82	63,66	318.31	636,62
2	2,92	4,30	6,97	9,92	22,33	31.60
3	2,35	3.18	4,54	5,84	10.21	12,92
4	2,13	2,78	3,75	4,60	7,17	8.61
5	2,01	2,57	3,37	4,03	5,89	6,87
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3.50	4,79	5,41
8	1,86	2,31	2,90	3,36	4,50	5,01
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,50
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,69	2,85	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,71	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,70	2.05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,47	2,76	3,41	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,34	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
VC	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Mức ý nghĩa α (trường hợp 1 phía)					

Bảng IV. *Giá trị $\chi^2_k(\alpha)$ của phân phối χ^2 với k bậc tự do (α tính theo %)*

K	99,0	97,5	95	90	70	50
1	0.000157	0.000982	0.00393	0.0158	0.048	0.455
2	0.0201	0.0506	0.103	0.211	0.713	1.39
3	0.0115	0.216	0.352	0.584	1.42	2.37
4	0.297	0.481	0.711	1.06	2.19	3.36
5	0.554	0.381	1.15	1.61	3.00	4.35
6	0.872	1.24	1.61	2.20	3.83	5.35
7	1.24	1.69	2.17	2.83	4.67	6.35
8	1.65	2.18	2.73	3.49	5.53	7.34
9	2.09	2.70	3.33	4.17	6.39	8.34
10	2.53	3.25	3.91	4.87	7.27	9.34
11	3.05	3.82	4.57	5.58	8.15	10.3
12	3.57	4.10	5.12	6.30	9.03	11.3
13	4.11	5.01	5.89	7.04	9.93	12.3
14	4.66	5.63	6.57	7.79	10.8	13.3
15	5.23	6.26	7.26	8.55	11.7	14.3
16	5.81	9.91	7.93	9.31	12.6	15.3
17	6.11	7.56	8.67	10.1	13.5	16.3
18	7.01	8.23	9.30	10.9	14.4	17.3
19	7.63	8.91	10.1	11.7	15.4	18.3
20	8.26	9.59	10.9	12.4	16.3	19.3
21	8.90	10.3	11.6	13.2	17.2	20.3
22	9.54	11.0	12.3	14.0	18.1	21.3
23	10.2	11.7	13.1	14.8	19.0	22.3
24	10.9	12.4	13.8	15.7	19.9	23.3
25	44.5	13.1	14.6	16.5	20.9	24.3
26	12.2	13.8	15.4	17.3	21.8	25.3
27	12.9	14.6	16.2	18.1	22.7	26.3
28	13.6	15.3	16.9	18.9	23.6	27.3
29	14.3	16.0	17.7	19.8	24.6	28.3
30	15.0	16.8	18.5	20.6	25.5	29.3
40	22.2	24.4	26.5	29.1	34.9	39.3
50	29.7	32.4	34.8	37.7	44.3	49.3
60	35.5	40.5	43.2	46.5	53.8	59.3
70	45.1	48.8	61.7	55.3	63.3	69.3
80	53.5	57.2	60.4	64.3	72.9	79.3
90	61.8	65.6	69.1	73.3	82.5	89.3
100	70.1	74.2	77.9	82.4	92.1	99.3

Bảng IV. *Giá trị $\chi^2_k(\alpha)$ của phân phối χ^2 với k bậc tự do (α tính theo %) (tiếp theo)*

	30	10	5	2,5	1	0.1
1	1.07	2,71	3.81	5.02	6.62	10.8
2	2.41	4,61	5.99	7.38	9.21	13.8
3	3.67	6,25	7.81	9.15	11.3	16.3
4	4.88	7,78	9.49	11.1	13.3	18.5
5	6.06	9.21	11.1	12.3	15.1	20.5
6.	7.23	10.6	12.6	14.1	16.8	22.5
7	8.38	12.0	14.1	16.0	18.5	24.3
8	9.52	13.4	15.5	17.5	20.1	26.1
9	10.7	14.7	16.9	19.0	21.7	27.9
10	11.8	16.0	18.3	20.5	23.2	29.6
11	12.9	17.3	19.7	21.9	24.7	31.3
12	14.0	18.5	21.0	23.3	26.2	32.9
13	15.1	19.8	22.4	24.7	27.7	34.5
14	16.2	21.1	23.7	26.1	29.1	36.1
15	17.3	22.3	25.0	27.5	30.6	37.7
16	18.4	23.5	26.3	28.8	32.0	39.3
17	19.5	24.8	27.6	30.2	33.4	40.8
18	20.6	26.0	28.9	31.5	34.8	42.3
19	21.7	27.2	30.1	32.9	36.2	43.8
20	22.8	28.4	31.4	34.2	37.6	45.3
21	23.9	29.6	32.7	35.5	38.9	46.8
22	24.9	30.8	33.9	36.8	40.3	48.3
23	26.0	32.0	35.2	38.1	41.6	49.7
24	27.1	33.2	36.4	39.4	43.0	51.2
25	28.2	34.4	37.7	40.6	44.3	52.6
26	29.2	35.6	38.9	41.9	45.6	54.1
27	30.3	36.7	40.1	43.2	47.0	55.5
28	31.4	37.9	41.3	44.5	48.3	56.9
29	32.5	39.1	42.6	45.7	49.6	58.3
30	33.5	40.3	43.8	47.0	50.9	59.7
40	44.2	51.8	55.8	59.3	63.7	73.4
50	54.7	63.2	67.5	71.4	76.2	86.7
60	65.2	74.4	69.1	83.3	88.4	92.6
70	75.1	85.5	90.5	95.0	100.4	121.3
80	86.1	96.6	101.9	106.6	112.3	124.8
90	96.5	107.6	113.1	118.1	124.1	137.2
100	106.9	118.5	124.3	129.6	135.8	149.1

Bảng V. Khoảng tin cậy của tỷ lệ $p \leqslant 0,1$ hoặc $p \geqslant 0,9$ (với $\alpha = 5\%$) (các giá trị của np_1 *và* np_2 *khi* $p \leqslant 0,1$*)*

m	0	1	2	3	4	5	6	7	8	9	10
00		0.025 5.572	0.24 7.22	0.62 8.76	1.09 10.24	1.62 11.67	2.20 13.06	2.31 14.42	3.45 15.76	4.12 17.08	
10	4.8 18,4	5.5 19.7	6.2 21.0	6.2 21.0	7.6 23.6	8.3 24.9	9.0 26.1	9.8 27.3	10.6 28.5	11.4 29.7	
20	12,2 30,9	13.0 32.1	13.8 33.3	13.8 33.3	15.4 35.7	16.2 36.9	17.0 38.1	17.8 39.3	18.6 40.5	19.9 41.7	
30	20,2 42,8	21.0 44.0	21.8 45.2	21.8 45.2	23.5 47.5	24.4 48.6	25.2 49.8	26.1 50.9	26.9 52.1	57.8 55.3	
40	28,6 54,4	29.5 55.5	30.3 56.7	30.3 56.7	32.0 59.0	32.9 60.1	33.7 61.3	34.6 62.5	35.4 63.6	36.3 64.7	
50	37,1 65,9	38.0 67.0	38.8 68.2	38.8 68.2	40.5 70.5	41.4 71.6	42.3 72.7	43.1 73.9	44.0 75.0	44.9 76.1	
60	45,8 77,2	46.6 78.4	47.5 79.5	47.5 79.5	49.3 81.8	40.2 82.9	51.0 84.0	51.9 85.1	52.8 86.2	53.7 87.3	
70	54,6 88,4	55.5 89.5	53.3 90.7	53.3 90.7	58.1 92.9	58.0 94.0	59.9 95.1	60.8 96.2	61.7 97.3	62.6 98.4	
80	63,4 99,6	64.3 100.7	65.5 101.8	66.1 102.9	67.0 104.0	67.9 107.3	68.8 106.2	69.7 107.3	70.8 108.4	71.5 109.5	
90	72,4 110,6	73.3 111.7	74.2 112.8	75.1 113.9	76.0 155.0	76.9 116.1	77.8 117.2	78.7 118.3	79.6 119.4	80.5 120.5	81.4 121.6

Bảng VI. Khoảng tin cậy của tỷ lệ $p = \frac{m}{n}$ *của mẫu bé.*

($3 < n \leq 10$ với $\alpha = 5\%$)

m n	0	1	2	3	4	5	6	7	8	9	10
4	0 60,2	0.6 80.6	6.8 93.2	19.4 99.4	39.9 100.0						
5	0 52,5	0.5 71.1	5.3 85.3	14.7 94.7	28.4 99.5	47.8 100.0					
6	0 45,9	0.4 64.1	4.3 77.7	11.8 88.2	22.3 95.7	35.9 99.6	54.1 100.0				
7	0 41,0	0.4 57.9	3.7 71.0	9.9 80.6	18.4 90.1	29.0 96.3	42.1 99.6	59.0 100.0			
8	0 36,9	0.3 52.7	3.2 65.2	8.5 75.5	15.7 84.3	24.5 91.5	34.8 96.8	47.3 99.7	63.1 100.0		
9	0 33,6	0.3 48.3	2.8 60.0	7.5 70.1	13.7 78.8	21.2 86.3	29.9 92.5	40.0 97.2	51.7 99.7	66.4 100.0	
10	0 30,8	0.3 44.5	2.5 55.6	6.7 65.2	12.2 73,8	18.7 81.3	26.2 87.8	34.8 93.3	44.4 97.5	55.5 99.7	69.2 100.0

Bảng VII. *Tìm giá trị* F_{05} *(hàng trên) và* F_{01} *(hàng dưới)*

	1	2	3	4	5	6	7	8	9	10	11	12
1	161	250	216	225	230	234	237	239	241	242	243	214
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106
2	18.51	19.00	19.17	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42
3	40.13	9.55	9.28	5.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.37	27.34	27.23	27.13	27.05
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	16.66	14.54	14.45	14.37
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	6.68	3.63	3.60	3.57
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.74
8	5.32	4.46	4.07	3.84	3.69	3.85	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	8.65	2.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67
9	5,12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.97
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	7.56	5.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.60	2.65	2.60	2.56	2.53
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.45
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	4.89	3.80	3.73	3.67
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45

Bảng VII. Tìm giá trị F_{05} *(hàng trên) và* F_{01} *(hàng dưới)*

	14	16	20	24	30	40	50	75	100	200	500	vc
1	245	246	248	249	250	251	252	253	253	254	254	254
	6142	6169	6208	6234	6258	6286	6302	6323	6334	6352	6361	6366
2	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50
3	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	26.92	26.84	26.69	26.60	26.50	26.41	26.35	6.27	26.23	26.18	26.14	26.12
4	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	14.24	14.15	14.02	13.93	13.85	13.74	13.69	13.61	13.57	13.52	13.88	13.46
5	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	9.71	9.68	9.55	9.47	9.38	9.29	9.21	9.17	9.13	9.07	9.04	9,02
6	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.64	6.90	6,88
7	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3,23
	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5,65
8	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86
9	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31
10	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	4.60	4.52	4.41	4.33	5.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91
11	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	4.20	4.20	4.10	4.02	3.94	3.86	3.80	3.74	3.79	3.66	3.62	3.60
12	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36
13	3.85	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	2.55	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16
14	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
15	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
16	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.89	2.86	2.80	2.77	2.75
17	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65

Bảng VII. Tìm giá trị F_{05} *(hàng trên) và* F_{01} *(hàng dưới)*

	1	2	3	4	5	6	7	8	9	10	11	12
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34
	8,28	6,01	5,09	4,58	4,25	4,01	3,85	3,71	3,60	3,51	3,44	3,37
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,34	2,31
	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2.45	2,40	2,35	2,31	2,28
	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,30	3,23
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25
	8,02	5,78	4,87	4,37	4,04	3,81	3,65	3,51	3,40	3,31	3,24	3,17
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,26	2,23
	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,36	2,32	2,28	2,24	2,20
	6,88	5,66	4,67	4,26	3,94	3,71	3,54	3,41	3,30	5,21	9,14	3,03
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18
	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3.36	3,25	3,17	3,09	3,03
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24	2.20	2,16
	7,78	3,57	4,68	4,18	3,86	3,63	3,46	3,32	3,21	3,13	3,05	2,99
26	4,22	3,37	2,94	1,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15
	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,17	3,09	3,02	2,96
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	3,20	2,16	2,13
	7,68	5,49	4,60	4,11	3,79	3,56	3,39	3,26	3,14	3,06	2,98	2,93
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,15	2,12
	7,64	5,45	4,57	4,07	3,76	3,53	3,36	3,23	3,11	3,03	2,95	2,90
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18	2,14	2,10
	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,08	3,00	2,92	2,87
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12	2,00
	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,90	2,84
32	4,15	3,30	2,90	2,67	2,51	2,40	2,32	2,25	2,19	2,14	2,10	2,07
	7,50	5,34	4,46	3,97	3,66	3,42	3,25	3,12	3,01	2,94	2,86	2,80
34	4,13	3,28	2,88	2,65	2,49	2,38	2.30	2,23	2,17	2,12	2,08	2,05
	7,44	5,29	4,42	3,13	3,61	3,38	3,21	3,08	2,97	2,89	2,82	2,76
36	4,11	3,26	2,86	2.63	2,48	2,36	2,28	2,21	2,15	2,10	2,06	2,03
	7,39	5,25	4,38	3,89	3,58	3,35	3,18	3,04	2,94	2,86	2,78	2,72
38	4,10	3,25	2,85	2.62	2,46	2,35	2,26	2,19	2,14	2,09	2,05	2,02
	7,35	5,21	4,34	3,86	3,54	3,32	3,15	3,02	2,91	2,82	2,75	2,69

Bảng VII. Tìm giá trị F_{05} (hàng trên) và F_{01} (hàng dưới)

	14	16	20	24	30	40	50	75	100	200	500	vc
18	2,29	2,25	2,79	2,15	2,11	2,07	2,04	2,00	1,98	1,95	1,93	1,92
	3,27	3,19	3,07	3,00	2.91	2,83	2.78	2,71	2.68	2.62	2,59	2,57
19	2,26	2,21	2,15	2,11	2,07	2,02	2,00	1,96	1,94	1,91	1,90	1,88
	3,19	3,12	3,00	2,92	2,84	2,76	2,70	2,63	2,60	2,54	2,51	2,49
20	2,23	2,18	2,12	2,08	2,04	1.99	1,96	1,92	1,90	1,87	1,85	1,84
	3,13	3,05	2,94	2,86	2,77	2,69	2.63	2,56	2.53	2,47	2,44	2,42
21	2,20	2,15	2,09	2,05	2,00	1,96	1.93	1,89	1m87	1,84	1,82	1,81
	3,07	2,99	2,88	2,80	2,72	2,63	2,58	2,51	2,47	2,42	2,38	2,36
22	2,18	2,13	2,07	2,03	1,98	1,93	1,91	1,87	1,84	1,81	1,80	1,87
	3,02	2,94	2,83	2,675	2,67	2,58	2,53	2,46	2,42	2,37	2,33	2,31
23	2,14	2.10	2,05	2,00	1,96	1,91	1,88	1,84	1,82	1,79	1,77	1,76
	2,97	2,89	2,78	2,70	2,62	2,53	2,48	2,41	2,37	2,32	2,28	2,26
24	2,13	2,09	2,02	1,98	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
	2,93	2,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21
25	2,11	2,06	2,00	1,96	1,92	1,87	1,84	1,80	1,77	1,74	1,72	1,71
	2,89	2,81	2,70	2,62	2,54	2,45	2,40	2,32	2,29	2,23	2,19	2,17
26	2,10	2,05	2,99	1,95	1,90	1,85	1,82	1,78	1,76	1,72	1,70	1,69
	2,86	2,77	2,66	2,58	2,50	2,41	2,36	2,28	2,25	2,19	2,15	2,13
27	2,08	2,03	1,97	1,93	1,88	1,84	1,80	1,76	1,74	1,71	1,68	1,67
	2,83	2,74	2,63	2,55	2,47	2,38	2,33	2,25	2,21	2,46	2,12	2,10
28	2,06	2,02	1,96	1,91	1,87	1,81	1,78	1,75	1,72	1,69	1,67	1,65
	2,80	2,71	2,60	2,52	2,44	2,35	2,30	2,22	2,18	2,13	2,09	2,06
29	2,05	2,00	1,94	1,90	1,85	1,80	1,77	1,73	1,71	1,68	1,65	1,64
	2,77	2.68	2,57	2,49	2,41	2,32	2,27	2,19	2.15	2,10	2,06	2,03
30	2,04	1,99	1,93	1,89	1,84	1,79	1,76	1,72	1,69	1,67	1,64	1,62
	2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01
32	2,02	1,97	1,91	1,86	1,82	1,76	1,74	1,69	1,67	1,64	1,61	1,59
	3,70	2,62	2,61	2,42	2,34	2,25	2.20	2,12	2,08	2,02	1,98	1,96
34	2,00	1,95	1,89	1,84	1,80	1,74	1,71	1,67	1,64	1,61	1,59	1,57
	2,66	2,58	2,47	2,38	2,30	2,21	2,15	2,08	2,04	1,98	1,94	1,91
36	1,98	1,93	1,87	1,82	1,78	1,72	1,69	1,65	1,62	1,59	1,56	1,55
	2,62	2,54	2,43	2,35	2,26	2,17	2,12	2,04	2,00	1,94	1,90	1,88
38	1,96	1,92	1,85	1,80	1,76	1,71	1,67	1,63	1,60	1,57	1,54	1,53
	2,59	2,51	2,40	2,32	1,22	1,14	2,08	2,00	1,97	1,90	1,86	1,84

Bảng VII. Tìm giá trị F_{05} *(hàng trên) và* F_{01} *(hàng dưới)*

·	1	2	3	4	5	6	7	8	9	10	11	12
40	4,08	3,23	2.84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04	2,00
	7,31	5,18	4,31	3,83	3,51	3,29	3.12	2,99	2,88	2.80	2,73	2,66
42	4,07	3,22	2,83	2,59	2,44	2,32	2,24	2,17	2,11	2.06	2,02	1,99
	7,27	5,15	4,29	3,80	3,49	3,26	3,10	2,96	2.86	2,77	2,70	2,64
44	4.06	3,21	2.82	2,58	2,43	2,31	2.23	2.10	2,10	2,05	2,01	1,98
	7,24	5,12	4,26	3,78	3,46	3,24	3,07	2,94	2,84	2,75	2,68	2,62
46	4,05	3,20	2,81	2,57	2,42	2,30	2,22	2,14	2,09	2,04	2,00	1,97
	7,21	5,10	4,24	2,76	3,44	3,22	3,05	2,92	2,82	2,73	2,66	2.60
48	4,04	3,19	2.80	2,56	2,44	2,30	2,21	2,14	2,08	2,03	1,99	2,96
	7,19	5,08	4,22	3,74	3,42	3,20	3.04	2,90	2,80	2,71	2,64	1,58
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98	1,95
	7,17	5,06	4,20	3.72	3,41	2,18	3,02	2.88	2,78	2,70	2,62	2,56
55	4,02	3,17	2,78	2.54	2,38	2,27	2,18	2,11	2,05	2,00	1,97	1,93
	7,08	5,01	4,16	3,68	3,37	3,15	2,98	2,85	2,75	2,66	2,59	2,53
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,01	1,99	1,95	1,92
	7,07	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50
65	3,98	3,14	2,75	2,51	2,36	2,24	2,15	2,08	2,02	1,98	1,94	1,90
	7.04	4,95	4,10	3,62	3,31	3,09	2,93	2,79	2,70	2,61	2,54	2,47
70	3.96	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	2,93	1,89
	7,04	4,92	4.08	3,60	3,29	3,07	2,91	2,77	2,67	2,59	2,51	2,45
80	3,96	3,11	2,72	2,48	2,33	2,21	2,12	2,05	1,99	1,95	1,91	1,88
	6,96	4,88	4,04	3,56	3,25	3,04	2,87	2,74	2,64	2,55	2,48	2,44
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	,92	1,88	1,85
	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,43	2,36
125	3,92	3,07	2,68	2,44	2,29	2,17	2,08	2,01	1,95	1,90	1,86	1,83
	6.84	4,78	3,94	3,47	3,17	2,95	2,79	2,65	2,56	2,47	2,40	2,33
150	3,91	3,06	2,67	3,43	2,27	2,16	2,97	2,00	1,94	1,89	1,85	1,82
	6,81	4,75	2,91	3,41	3,14	2,92	2,76	2,62	2,53	2,44	2,37	2,30
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	21,92	1,87	1,83	2,80
	6,76	4,71	3,88	3,41	3,11	2,90	2,73	2,60	2,50	2,40	2,34	1,28
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81	2,78
	6,70	4,66	3,83	3,36	3,06	2,85	2,69	2,55	2,46	2,37	2,29	1,23
1000	3,85	3,00	2,61	2,38	2,22	2,10	2,02	1,95	1,89	1,84	1,80	1,76
	6,66	4,62	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,26	2,20
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79	2,75
	6.64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24	7,28

Bảng VII. Tìm giá trị F_{05} (hàng trên) và F_{01} (hàng dưới)

	14	16	20	24	30	40	50	75	100	200	500	∞
40	1,95	1,90	1,84	1,79	1,74	1,69	1,69	1,61	1,59	1,55	1,53	1,51
	2,56	2,49	2,37	2,29	2,20	2,11	2,05	1,97	1,94	1,88	1,84	1m,81
42	1,94	1,89	1,82	1,78	1,73	1,68	1,64	1,60	1,57	1,54	1,51	1,49
	2,54	2,46	2,35	2,26	2,17	2,08	2,02	1,94	1,91	1,85	1,80	1,78
44	1,92	1,88	1,81	1,76	1,72	1,66	1,63	1,58	1,56	1,52	1,50	1,48
	2,32	2,44	2,32	2,24	2,15	206	2,00	1,92	1,88	1,82	1,78	1,75
46	1,91	1,87	1,80	2,75	1,71	1,65	1,62	1,57	1,54	1,51	1,48	1,46
	2,50	2,42	2,30	2,22	2,13	2,04	1,98	1,90	1,96	1,80	1,76	1,72
48	1,90	1,86	1,79	1,74	1,70	1,64	1,61	1,56	1,53	1,50	1,47	1,45
	2,48	2,40	2,28	2,20	2,11	2,02	1,96	1,58	1,84	1,78	1,73	1,70
50	1,90	1,85	1,78	1,74	1,69	1,63	1,60	1,55	1,52	1,48	1,46	1,44
	2,46	2,39	2,26	2,78	2,10	2,00	1,94	1,86	1,82	1,76	1,71	1,68
55	1,88	1,83	1,76	1,72	1,67	1,61	1,58	1,52	1,50	1,46	1,43	1,41
	2,43	2,35	2,23	2,15	2,06	1,96	1,90	1,82	1,78	1,71	1,66	1,64
60	1,86	1,81	1,75	1,70	1,65	1,59	1,56	1,50	1,48	1,44	1,41	1,39
	2,40	2,32	2,20	2,12	2,03	1,93	1,87	1,79	1,74	1,68	1,63	1,60
65	1,85	1,80	1,73	1,68	1,63	1,57	1,54	1,49	1,46	1,42	1,39	1,37
	2,37	2,30	2,78	2,09	2,06	1,90	1,84	1,76	1,71	1,64	1,60	1,56
70	1,84	1,79	1,72	1,67	1,62	1,56	1,53	1,47	1,45	1,10	1,37	1,35
	2,35	2,28	2,15	3,07	1,98	1,88	1,82	1,74	1,69	1,62	1,56	1,53
80	1,32	1,77	1,70	1,65	1,60	1,54	1,51	1,45	1,42	1,38	1,35	1,32
	2,32	2,24	2,11	2,03	1,94	1,84	1,78	1,70	1,65	1,57	1,52	1,49
100	1,79	1,75	1,68	1,63	1,57	1,51	1,43	1,42	1,39	1,34	1,30	1,28
	2,26	2,19	2.06	1,98	1,89	1,79	1,73	1,64	1,59	1,51	1,46	1,43
125	1,77	1,72	1,65	1,60	1,55	1,49	1,45	1,39	1,36	1,31	1,27	1,25
	2,23	2,15	2,03	1,94	1,85	1,75	1,68	1,59	1,54	1,46	1,40	1,37
150	1,76	1,71	1,64	159	1,54	1,47	1,44	1,37	1,34	1,29	1,25	1,22
	6,20	2,12	2,00	1,91	1,83	1,721,45	1,66	1,56	1,51	1,43	1,37	1,33
200	1,74	1,69	1,62	1,57	1,52		1,42	1,35	1,32	1,26	1,22	1,19
	2,178	2,09	1,97	1,88	1,79	1,69	1,62	1,53	1,48	1,39	1,33	1,28
400	1,72	1,67	1,60	1,4	1,49	1,42	1,38	1,32	1,28	1,22	1,16	1,13
	2,12	2,04	1,92	1,84	1,74	1,64	1,57	1,47	1,42	1,32	1,24	1,19
1000	1,70	1,65	1,58	1,53	1,47	1,41	1,36	1,30	1,26	1,19	1,13	1,08
	2,09	2,01	1,89	1,81	1,71	1,61	1,54	1,44	1,38	1,28	1,19	1,11
∞	1,69	1,64	1,57	1,52	1,46	1,40	1,35	1,28	1,24	1,17	1,11	1,00
	2,07	1,99	1,87	1,79	1,69	1,59	1,52	1,41	1,36	1,25	1,15	1,00

Bảng VIII. Tung độ của đường cong phân bố chuẩn tiêu chuẩn

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \text{ đối với } 0 \leqslant u \leqslant 3{,}9$$

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	,39894	39892	39886	39876	39862	39811	39822	39797	39767	39733
0,1	,39695	39651	39608	39559	39505	39118	39387	39322	39253	30181
0,2	,39101	39021	38910	38853	38762	38667	38568	38466	38361	38251
0,3	,38139	38023	37903	37780	37651	37524	37391	37255	37115	36973
0,4	,36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0,5	,33207	35029	34819	31667	31482	34294	31105	33912	33718	33521
0,6	,33322	33121	32918	32713	32506	33297	32086	31874	31659	31443
0,7	,31225	31006	30785	30563	30339	30111	29887	29659	29431	29200
0,8	,28969	28737	28501	28269	28034	27798	27562	27324	27086	26848
0,9	,26609	26369	26129	25888	25647	25406	25164	24923	21681	211139
1,0	,21197	23955	23713	23471	23230	22988	22741	22506	22265	22025
1,1	,17137	21546	21307	21069	20831	20594	20357	20121	19886	19652
1,2	,19119	19186	18951	18721	18491	18265	18037	17810	17585	17360
1,3	,17137	16915	16691	16474	16256	16038	15822	15608	15396	15183
1,4	,14937	11761	14556	14350	14146	13943	13742	13542	13344	13147
1,5	,12952	12758	12566	12376	12188	12051	11816	11632	11150	11270
1,6	,11092	10915	10711	10567	10396	10226	10059	09893	09728	09566
1,7	,09105	19246	09089	08933	08780	08628	08478	08329	08183	03038
1,8	,07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1,9	,06562	06138	06316	06195	06077	05959	05844	05730	05618	05508
2,0	,05399	05292	05186	05082	01980	04879	01780	04682	01586	01191
2,1	,01398	01307	04217	01428	01011	03955	03871	03788	03706	03626
2,2	,03517	03170	03391	03319	03246	03174	03103	03034	02965	02898
2,3	,02833	02768	02705	02613	02582	02522	02463	02406	02349	02291
2,4	,02239	02186	02131	02083	02033	01984	01936	01889	01812	01797
2,5	,01753	01709	01667	01625	04585	01545	01506	01468	01131	01394
2,6	,01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2,7	,01012	01014	00987	00961	00935	00909	00885	00861	00837	00811
2,8	,00792	0070	00748	00727	00707	00687	00668	00619	00631	00613
2,9	,00595	00578	00562	00545	00530	00514	00499	00485	00171	00457
	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3,0	10013	00327	0024	00172	00123	00087	00061	00012	00029	00020

Bảng IX. Giá trị hàm $e^{-\lambda}$

λ	$e^{-\lambda}$	λ	$e^{-\lambda}$
0.0	1.00000	3.1	0.04505
0.1	0.90484	3.2	0.04076
0.2	0.81873	3.3	0.03688
0.3	0.74082	3.4	0.03337
0.4	0.67302	3.5	0.03020
0.5	0.60653	3.6	0.02732
0.6	0.54881	3.7	0.02472
0.7	0.49659	3.8	0.02237
0.8	0.44933	3.9	0.02024
0.9	0.40657	4.0	0.01832
1.0	0.36788	4.1	0.01657
1.1	0.33287	4.2	0.01500
1.2	0.30119	4.3	0.01357
1.3	0.27253	4.4	0.01228
1.4	0.24660	4.5	0.01111
1.5	0.22313	4.6	0.01005
1.6	0.20190	4.7	0.00910
1.7	0.18268	4.8	0.00823
1.8	0.16530	4.9	0.00745
1.9	0.14957	5.0	0.00674
2.0	0.13534	6.0	0.00248
2.1	0.12246	7.0	0.00091
2.2	0.11080	8.0	0.00034
2.3	0.10026	9.0	0.00012
2.4	0.09072	10.0	0.00005
2.5	0.08209	11.0	0.00002
2.6	0.07427	12.0	0.00001
2.7	0.06721	13.0	0.000000
2.8	0.06081	14.0	0.000000
2.9	0.05502	15.0	0.000000
3.0	0.04979		

MỤC LỤC

NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA HÀ NỘI

16 Hàng Chuối - Hai Bà Trưng - Hà Nội

Điện thoại: (04) 9724852; (04) 9724770. Fax: (04) 9714899

Chịu trách nhiệm xuất bản:

Giám đốc: PHÙNG QUỐC BẢO

Tổng biên tập: NGUYỄN BÁ THÀNH

Chịu trách nhiệm nội dung:

Hội đồng nghiệm thu giáo trình
Trường Đại học Khoa học Tự nhiên

Người nhận xét: GS. TS NGUYỄN VĂN HỮU
PGS. TS PHẠM VĂN KIỀU

Biên tập: NGUYỄN HỮU DƯ
LAN HƯƠNG

Sửa bài: THU HƯƠNG

Trình bày bìa: NGỌC ANH

XÁC SUẤT THỐNG KÊ

Mã số: 1K-53 ĐH2007

In 3000 cuốn, khổ 14.5 x 20.5 cm tại Nhà in Đại học Quốc gia Hà Nội

Số xuất bản: 381 - 2007/CXB/09 - 64/ĐHQGHN, ngày 25/05/2007

Quyết định xuất bản số: 258 KH/XB

In xong và nộp lưu chiểu quý II năm 2007.