Dương Ngọc Sơn

Bài giảng

Xác suất và thống kê

25 tháng 1, 2026

Lời nói đầu

Tài liệu học tập này gồm 4 chương, được tổng hợp từ các bài giảng của tác giả cho học phần *Xác suất và thống kê* tại Đại học Phenikaa trong học kì 2 năm học 2023/24 và đã được chỉnh sửa, bổ sung qua các kì học sau đó đến nay. Mục tiêu chính của việc biên soạn tài liệu này là nhằm cung cấp cho sinh viên một nguồn học liệu phù hợp với nội dung của môn học, bám sát đề cương môn Xác suất và thống kê do Khoa Khoa học cơ bản xây dựng.

Trong quá trình biên soạn, tài liệu có tham khảo một số giáo trình quen thuộc như cuốn *Thống kê và Ứng dụng* của GS. Đặng Hùng Thắng [1], cuốn *Advanced Engineering Mathematics, Part G: Probability, Statistics* của nhà toán học người Đức Erwin Kreyszig [2], và một số tài liệu chuyên sâu khác như cuốn sách kinh điển *Probability* của nhà toán học Nga Albert Nikolaevich Shiryaev.

Bên cạnh các nội dung lý thuyết, tài liệu còn giới thiệu về công cụ tính toán quan trọng trong lĩnh vực thống kê: *Phần mềm thống kê R*. Đây là công cụ hỗ trợ đắc lực trong các bài thực hành sử dụng hệ đại số máy tính (gọi tắt là CAS). Các bài thực hành CAS được lồng ghép nhằm phục vụ hai mục tiêu: một là minh họa sinh động cho các kiến thức lý thuyết về xác suất và thống kê, là hai là giúp sinh viên bước đầu làm quen với một công cụ thiết yếu cho các ứng dụng thực tiễn của thống kê.

Tài liệu được biên soạn trong thời gian ngắn, song song với quá trình giảng dạy trên lớp trong năm học 2023/24. Dù đã có nhiều chỉnh sửa, bổ sung trong các năm sau đó, nhưng chắc chắn tài liệu vẫn còn những điểm chưa hoàn thiện. Tác giả hy vọng tài liệu này sẽ hỗ trợ hữu ích cho công tác giảng dạy và học tập môn *Xác suất và thống kê* ở bậc đại học trong các trường kĩ thuật không chuyên về toán học, đồng thời tác giả mong nhận được những ý kiến đóng góp từ bạn đọc để tài liệu ngày càng có ích hơn.

Hà Nội, tháng 1 năm 2026 Dương Ngọc Sơn

Mục lục

Chương 1

Phân tích dữ liệu. Lí thuyết xác suất

Mục tiêu của chúng ta trong chương này là tìm hiểu những vấn đề cơ bản của phân tích dữ liệu và lí thuyết xác suất, bao gồm:

- Phân tích dữ liệu: Cách biểu diễn dữ liệu (biểu đồ) các đặc trưng của dữ liệu như trung bình, phương sai, độ lệch chuẩn; trung vị, và tứ phân vị.
- Lí thuyết xác suất: Phép thử, xác suất, sự kiện, xác suất có điều kiện, công thức Bayes, công thức xác suất toàn phần, biến ngẫu nhiên và phân phối xác suất.

Tài liệu tham khảo: Kreyszig [1, Chapter 24], Đ. H. Thắng [2, Chương 1, 2], V. H. Nhự (chủ biên) [3], Akritas [4], Cramer [5], Shiryaev [6].

Nội dung

1.1 Biểu diễn dữ liệu

Dữ liệu được thu thập trong thực tế thường là dữ liệu thô. Chúng ta cần phân tích dữ liệu để giúp hiểu rõ thông tin, hỗ trợ việc ra quyết định, phát hiện mẫu và xu hướng. Một bước quan trọng trong phân tích dữ liệu là biểu diễn dữ liệu một cách hiệu quả. Biểu diễn dữ liệu giúp chúng ta truyền tải những thông tin về dữ liệu một cách tốt hơn.

1.1.1 Biểu diễn dữ liệu dạng thân-lá

Biểu đồ thân-lá (stem-and-leaf) là một công cụ thống kê được sử dụng để biểu diễn dữ liệu số dưới dạng biểu đồ, trong đó, con số hàng đầu của mỗi dòng được gọi là "thân" (stem), còn con số cuối cùng (hoặc con số cuối cùng và các chữ số đằng sau) của mỗi dòng được gọi là "lá" (leaf).

Ví dụ 1.1. Giả sử một dữ liệu cho bởi dãy số 23, 27, 27, 31, 32, 33, 33, 33, 36, 37, 38, 38, 41, 45, 45, 49. Hãy biểu diễn dữ liệu trên bằng biểu đồ thân lá và giải thích cách làm.

Lời giải

Dữ liệu đó có thể được biểu diễn dưới dạng biểu đồ thân-lá như sau:

thân	lá
2	3 7 7
3	1 2 3 3 3 6 7 8 8
4	1 5 5 9

Trong biểu đồ thân-lá như trên, các con số hàng đầu (thân) là 2, 3 và 4, và các con số cuối cùng (lá) là những con số còn lại. Ví dụ, số 6 được đánh dấu ở trên biểu thị giá trị 36. Cách biểu diễn này tuy đơn giản nhưng giúp tổ chức và hiển thị dữ liệu một cách trực quan và dễ hiểu hơn.

Trong biểu đồ thân-lá, các giá trị trong phần lá có thể được lặp lại một số lần nhất định, thể hiện "tần suất" của giá trị dữ liệu tương ứng. Như trong ví dụ trên, tần suất của các giá trị 27 và 45 là 2, trong khi đó, tần suất của giá trị 33 là 3.

Sử dụng phần mềm để biểu diễn dữ liệu

Trong suốt bài giảng này, chúng ta dùng phần mềm R để thực hành tính toán và trình bày kết quả. Phần mềm R là phần mềm hỗ trợ ngôn ngữ lập trình phân tích thống kê chuyên nghiệp với giao diện đơn giản, dễ sử dụng và hoàn toàn miễn phí với mã nguồn mở. Để biết thêm thông tin, tải và cài đặt, vào `https://www.r-project.org/`. Bạn đọc có thể tham khảo thêm Akritas [4].

CAS 1.1. Trong phần mềm thống kê R, chức năng `stem()` giúp tạo biểu đồ thân-lá, như dưới đây:

```
# Nạp dữ liệu
data = c(23, 27, 27, 31, 32, 33, 33, 33, 36, 37, 38, 38, 41,
         45, 45, 49)

# Hàm stem() tạo biểu đồ thân-lá
stem(data)

##
##   The decimal point is 1 digit(s) to the right of the |
##
```

```
##   2 | 377
##   3 | 123336788
##   4 | 1559
```

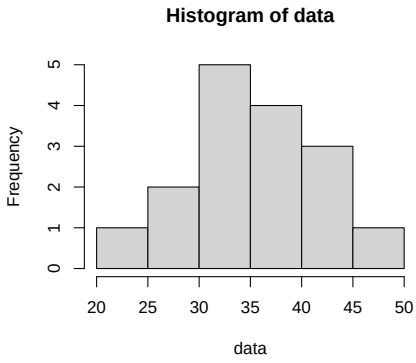
Với dữ liệu của Ví dụ 1.1 trong biến `data`, dòng lệnh `stem(data)` cho kết quả đúng như trên.

1.1.2 Biểu diễn dữ liệu dạng đồ thị

Đồ thị tần suất (histogram) hay *biểu đồ tần suất* (hoặc "tổ chức đồ") là một biểu đồ thống kê mà sử dụng các hình cột để biểu diễn phân phối của một tập hợp dữ liệu. Trục ngang của đồ thị tần suất thường là các khoảng giá trị của dữ liệu, trong khi trục đứng thể hiện tần suất tương đối (tỉ lệ phần trăm) của các quan sát trong mỗi khoảng đó.

Bước đầu tiên trong quy trình vẽ một đồ thị tần suất là xác định số lượng các khoảng giá trị của dữ liệu. Chúng ta cần quyết định số lượng khoảng dữ liệu muốn phân chia. Số lượng các khoảng này có thể ảnh hưởng đến cách mà dữ liệu được hiển thị, nên thường cần thử nghiệm và điều chỉnh để đạt được hiệu quả tốt nhất.

Tiếp theo, chúng ta cần tính toán biên của các khoảng, tức là chia phạm vi giá trị của dữ liệu thành các khoảng dựa trên số lượng khoảng đã chọn. Mỗi khoảng sẽ có một giới hạn trên và dưới.



Hình 1.1: Biểu đồ cột của dữ liệu của Ví dụ 1.1

Trong Ví dụ 1.1, chúng ta phân chia dữ liệu làm 6 khoảng: 21–25, 26–30, 31–35, 36–40, 41–45, 46–50, đếm số lượng quan sát trong mỗi khoảng, và tính *tần suất tuyệt đối* (absolute frequency) của mỗi điểm dữ liệu trong mỗi khoảng đó. Tần suất tương đối (relative frequency), kí hiệu $f_{\text{rel}}(x_i)$, bằng số các điểm dữ liệu nhận giá trị x_i chia cho tổng số điểm dữ liệu, là một số dương bé hơn 1, hoặc tính theo phần trăm thì bé hơn 100%.

Kết quả của việc tính tần suất của các điểm dữ liệu ghi trong Bảng 1.1 dưới đây.

Cuối cùng, hãy vẽ một cột cho mỗi khoảng, với chiều cao của cột thể hiện số lượng quan sát trong khoảng đó và kết quả thu được là Hình 1.1.

Đồ thị tần suất thường được sử dụng để hiểu phân phối của dữ liệu, cho phép nhận biết các đặc điểm như xu hướng trung tâm, sự phân tán

Lớp	21–25	26–30	31–35	36–40	41–45	46–50
Tần suất tuyệt đối	1	2	5	4	3	1
f_{rel} (%)	6,25	12,50	31,25	25,00	18,75	6,25

Bảng 1.1: Tần suất lớp tuyệt đối và tương đối

và hình dạng của dữ liệu. Nó là một công cụ quan trọng trong phân tích thống kê và trực quan hóa dữ liệu.

Trung vị và tứ phân vị

Trung vị là gì? Giả sử có một mẫu x. Xét các giá trị mẫu có tính chất sau, gọi là tính chất trung vị,

> *Ít nhất 50% các giá trị mẫu lớn hơn hoặc bằng giá trị đó và ít nhất 50% các giá trị mẫu nhỏ hơn hoặc bằng giá trị đó.*

Khi đó, có nhiều nhất hai giá trị trong mẫu có tính chất trung vị. *Trung vị mẫu* (sample median) bằng trung bình cộng các giá trị có tính chất trung vị nêu trên.

Nếu mẫu $x = \{x_1, x_2, \ldots, x_m\}$ được sắp xếp theo thứ tự tăng dần,

$$x_{\min} = x_1 \le x_2 \le \cdots \le x_n = x_{\max},$$

thì xét hai trường hợp sau:

- Trường hợp n chẵn, $n = 2k$. Khi đó, các giá trị x_k và x_{k+1} thoả mãn tính chất trung vị. Nếu $x_k < x_{k+1}$ thì đó là hai giá trị mẫu thoả mãn tính chất trung vị và không còn giá trị nào khác. Còn nếu $x_k = x_{k+1}$ thì giá trị chung đó là giá trị trung vị duy nhất. Trong cả hai trường hợp, trung vị của mẫu là
$$m = \frac{1}{2}(x_k + x_{k+1}).$$
- Trường hợp n lẻ, $n = 2k+1$. Khi đó, mọi giá trị mẫu có tính chất trung vị đều phải bằng x_{k+1} và do đó
$$m = x_{k+1}.$$

Hiệu của giá trị lớn nhất $x_{\max}$ và giá trị bé nhất $x_{\min}$ gọi là *khoảng biến thiên* (range) của dữ liệu x, tức là $R = x_{\max} - x_{\min}$.

Ví dụ 1.2. Giả sử điểm kiểm tra giữa kì môn Xác suất thống kê của 9 bạn sinh viên lớp A là 4, 5, 7, 7, 7, 8, 8, 9, 9. Khi đó, giá trị $x_3 = x_4 = x_5 = 7$ là giá trị duy nhất thoả mãn tính chất trung vị nêu trên. Do đó, trung vị của mẫu là $m = 7$.

Ví dụ 1.3. Giả sử điểm kiểm tra giữa kì của một lớp Xác suất thống kê của 10 bạn sinh viên lớp B là 4, 5, 6, 7, 7, 8, 8, 9, 9, 10. Khi đó, có $n = 10$ điểm mẫu. Hai giá trị $x_4 = x_5 = 7$ và $x_6 = x_7 = 8$ là hai giá trị mẫu thoả mãn tính chất trung vị nêu trên. Khi đó, trung vị mẫu bằng

$$m = \frac{1}{2}(7+8) = 7{,}5.$$

Tứ phân vị là gì? Khái niệm tứ phân vị được xác định tương tự như trung vị, nhưng thay mức phân chia 50% – 50% của trung vị bằng 25% – 75% và 75% – 25%.

CAS 1.2. Trong phần mềm thống kê R, chức năng `summary()` đưa ra tổng kết về giá trị bé nhất, lớn nhất (min, max), các tứ phân vị (bao gồm median), và trung bình.

```
# Khai báo dữ liệu vào biến data
data = c(23, 27, 27, 31, 32, 33, 34, 34, 36, 37, 38, 40)

# Thực hiện chức năng summary()
summary(data)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23,0    30,0    33,5    32,7    36,2    40,0
```

Đọc kết quả thu được thì trung vị của dữ liệu là $m = 33{,}5$, là trung bình cộng của hai giá trị 33 và 34 thoả mãn tính chất trung vị. Hai tứ phân vị còn lại là phân vị $Q1 = 30$ và $Q3 = 36{,}25$. Chúng ta có thể kiểm tra các tính toán này bằng tay như sau: Mỗi giá trị trong hai giá trị mẫu 27 và 31 thoả mãn tính chất: có ít nhất 25% điểm mẫu bé hơn hoặc bằng nó và ít nhất 75% lớn hơn hoặc bằng nó. Do đó, $Q1 = 0{,}25 \times 27 + 0{,}75 \times 31 = 30$. Tương tự như vậy, $Q3$ được xác định từ hai giá trị mẫu 36 và 37 như sau: $Q3 = 0{,}75 \times 36 + 0{,}25 \times 37 = 36{,}25$.

Biểu đồ hộp

Biểu đồ hộp (boxplot, hoặc box-and-whiskers plot) là một công cụ trực quan được sử dụng trong thống kê để mô tả sự phân phối của dữ liệu và nhận diện các giá trị bất thường (outliers). Chúng giúp so sánh giữa các nhóm dữ liệu một cách trực quan và cung cấp thông tin về sự biến thiên và xu hướng của dữ liệu.

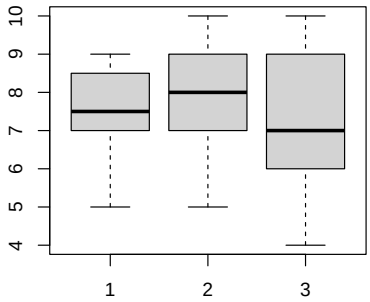
Biểu đồ hộp gồm những gì? Các thành phần chính của biểu đồ hộp gồm hộp (box), đường giữa (median line), và các râu (whiskers). Phần hộp thể hiện *khoảng tứ phân vị giữa* (interquartile range - IQR), tức là khoảng từ

phân vị thứ nhất (Q1, tức 25%) đến phân vị thứ ba (Q3, tức 75%) của dữ liệu. Đường giữa hộp là đường nằm ngang ở giữa hộp biểu diễn giá trị trung vị (median) của dữ liệu, tức là phân vị thứ hai (Q2, 50%). Các râu là các đường kéo dài từ hai đầu của hộp biểu diễn khoảng dữ liệu bên ngoài từ Q1 đến Q3. Thông thường, râu kéo dài đến 1,5 lần độ dài IQR (hiệu $Q3-Q1$). Cuối cùng, các giá trị ngoại lai (nếu có) là những điểm dữ liệu nằm ngoài râu, thường được biểu diễn bằng các dấu chấm nhỏ.

Ứng dụng của biểu đồ hộp là gì? Biểu đồ hộp giúp so sánh phân phối dữ liệu một cách dễ dàng, cho thấy một các trực quan sự phân phối giữa các nhóm dữ liệu khác nhau. Chúng cũng giúp nhận diện giá trị ngoại lai, là những giá trị "bất thường" trong tập dữ liệu. Biểu đồ hộp cũng thể hiện xu hướng trung tâm và mức độ biến động: Trung vị và các phân vị giúp thấy rõ mức độ tập trung và biến động của dữ liệu.

Chúng ta có thể vẻ biểu đồ hộp của dữ liệu rất dễ dàng bằng phần mềm, như trong CAS sau đây.

CAS 1.3. Trong R, để vẽ biểu đồ hộp của một tập dữ liệu, chỉ cần dùng `boxplot()`.



Hình 1.2: Biểu đồ hộp của các dữ liệu trong Ví dụ 1.2

```
x = c(5, 7, 7, 7, 8, 8, 9, 9)
y = c(5, 6, 7, 7, 8, 8, 9, 9, 10)
z = c(4, 6, 6, 7, 7, 7, 9, 9, 10)
boxplot(x, y, z)
```

Kết quả thu được là Hình 1.2 biểu diễn dữ liệu ở Ví dụ 1.2 và 1.3. Bạn đọc có thể quan sát và tự rút ra các kết luận về các dữ liệu và sự so sánh giữa chúng.

Bài tập 1.1. Tìm các tứ phân vị của dữ liệu trong các Ví dụ 1.2 và 1.3.

1.1.3 Trung bình, phương sai mẫu, và độ lệch chuẩn mẫu

Đối với các mẫu dữ liệu, bên cạnh tứ phân vị và trung vị, người ta thường quan tâm đến một số tham số khác, bao gồm trung bình (mean), phương sai (variance), và độ lệch chuẩn (standard deviation). Chúng là những đặc

trưng tuy đơn giản nhưng rất quan trọng, cho những thông tin ban đầu về mẫu.

Định nghĩa 1.1:

Cho một mẫu dữ liệu gồm các giá trị $x_1, x_2, \ldots, x_n$, gồm n điểm mẫu. *Trung bình* $\bar{x}$ của nó là

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.1}$$

Trong công thức (1.1), kí hiệu "sigma" $\sum$ có nghĩa là

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Nhận xét 1.1. Nếu $f_{\text{rel}}(x_i)$ là tần suất tương đối của điểm dữ liệu x_i, thì trung bình còn có thể được tính là

$$\bar{x} = \sum_{i} x_i f_{\text{rel}}(x_i). \tag{1.2}$$

Như vậy, trung bình $\overline{x}$ bằng tổng của tất cả các giá trị x_i xuất hiện trong mẫu nhân với tần suất tương đối của nó.

Tham số tiếp theo cần nói đến là *phương sai* (variance) của dữ liệu. Phương sai đo *mức độ phân tán* của dữ liệu quanh giá trị trung bình, tức là mức độ biến động của dữ liệu so với giá trị trung bình của dữ liệu. Cụ thể, phương sai được tính bằng cách lấy trung bình bình phương của khoảng cách giữa mỗi giá trị dữ liệu và giá trị trung bình:

Định nghĩa 1.2:

Cho một mẫu dữ liệu gồm các giá trị $x_1, x_2, \ldots, x_n$, gồm n điểm mẫu. *Phương sai* s^2 của mẫu được cho như sau:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.3}$$

Như trên, trong công thức (1.3), kí hiệu "sigma" $\sum$ có nghĩa là

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2.$$

Để tính toán thuận tiện hơn, người ta có thể sử dụng công thức (1.4) sau đây cho phương sai mẫu hiệu chỉnh.

Bài tập 1.2. Cho $x=(x_1,x_2,\ldots,x_n)$ là một mẫu và $x^2=(x_1^2,x_2^2,\ldots,x_n^2)$ là mẫu gồm bình phương của các giá trị mẫu trong x. Khi đó, phương sai mẫu hiệu chỉnh của x có thể tính là

$$s^2=\frac{n}{n-1}\left(\overline{x^2}-\bar{x}^2\right), \tag{1.4}$$

trong đó, $\overline{x^2}$ là trung bình của mẫu x^2. Hãy chứng minh công thức trên.

Ví dụ 1.4. Cho mẫu dữ liệu $x=(45,48,48,49,50)$. Để tính trung bình và phương sai, chúng ta tính các giá trị x_i^2 và ghi vào bảng như sau:

i	x_i	x_i^2
1	45	2025
2	48	2304
3	48	2304
4	49	2401
5	50	2500
Trung bình	48	2306,8

Giá trị ở dòng cuối cùng là trung bình cộng các giá trị phía trên. Từ đó, trung bình của x là 48, trong khi đó, phương sai của x tính theo công thức (1.4) là

$$s^2=\frac{5}{5-1}\left(2306{,}8-(48)^2\right)=3{,}5.$$

Lấy căn bậc hai của phương sai mẫu, chúng ta thu được độ lệch chuẩn (standard deviation) của mẫu.

Định nghĩa 1.3:

Độ lệch chuẩn s của mẫu là

$$s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{1.5}$$

Nhận xét 1.2. Độ lệch chuẩn mẫu có cùng đơn vị với các giá trị mẫu trong khi đó phương sai của mẫu có đơn vị là bình phương của đơn vị của mẫu.

Trung bình, phương sai, và độ lệch chuẩn của mẫu dữ liệu có thể tính tự động bằng máy tính khoa học, chẳng hạn như Casio 580VNX, và bằng các phần mềm máy tính, thường là các *hệ thống đại số máy tính* (computer algebra system – CAS). Mục CAS dưới đây minh họa cách tính toán bằng phần mềm R.

CAS 1.4. Trong R, một mẫu dữ liệu đơn giản có thể được đưa vào bằng tay qua lệnh concatenation, viết tắt là c. Tuy nhiên, lưu ý rằng R cũng có thể đọc dữ liệu từ các tệp bảng tính (như tệp excel hoặc csv). Trung bình, phương sai, và độ lệch chuẩn được tính, theo thứ tự, bằng các lệnh `mean()`, `variance()`, và `sd()`.

Đối với dữ liệu trong Ví dụ 1.1, ta tính được trung bình, phương sai, và độ lệch chuẩn của nó và thể hiện ở màn hình như sau:

```
data = c(23, 27, 27, 31, 32, 33, 33, 33, 36, 37, 38, 38, 41,
         45, 45, 49)
print(mean(data)) # trung bình mẫu

## [1] 35,5

print(var(data)) # phương sai mẫu

## [1] 50,67

print(sd(data)) # độ lệch chuẩn mẫu

## [1] 7,118
```

Bài tập mục 1.1

1. Cho dữ liệu $x = \{20, 21, 21, 21, 19, 19, 20, 20, 19\}$.

(a) Hãy sắp xếp dữ liệu theo thứ tự tăng dần.

(b) Biểu diễn dữ liệu dạng thân-lá.

(c) Tính trung bình, phương sai, và độ lệch chuẩn của mẫu dữ liệu đó.

(d) Tìm trung vị và các tứ phân vị của dữ liệu.

2. Cho một mẫu dữ liệu biểu diễn dưới dạng biểu đồ thân-lá như sau:

1	1 3 5
1	5 5 6 7 8 9
2	0 1 3 3 4 5 5
2	6 7 8 9 9
3	1 2 2 3 5
3	6 6 7

(a) Hãy vẽ biểu đồ cột (histogram) biểu diễn dữ liệu trên.

(b) Hãy tính trung bình, phương sai, và độ lệch chuẩn của mẫu.

3. Chứng minh rằng trung bình $\overline{x}$ của dữ liệu x tuỳ ý luôn nằm giữa giá trị bé nhất và giá trị lớn nhất: $x_{\min} \le \overline{x} \le x_{\max}$.

1.2 Phép thử, kết quả, sự kiện và các phép toán trên sự kiện

Mục tiêu của phần này là giúp bạn đọc hiểu rõ các khái niệm cơ bản của môn xác suất, gồm các khái niệm cơ bản về phép thử (trial), sự kiện (hay biến cố - event), các phép toán trên sự kiện. Tài liệu tham khảo: Kreyszig [1, §24.2].

1.2.1 Một số khái niệm cơ bản

Một phép thử (trial) là một lần thực hiện một thí nghiệm duy nhất. Kết quả của một phép thử gọi là một *kết quả* hay kết cục (outcome), và là một *điểm mẫu* (sample point). Một *không gian mẫu* (sample space) của một thí nghiệm là tập tất cả các kết quả có thể xảy ra, thường được kí hiệu là Ω.

Để hình dung rõ hơn, hãy xem mấy ví dụ sau:

- Kiểm tra một sản phẩm, kết quả có thể là sản phẩm đó có lỗi, hoặc không có lỗi. Như vậy, có thể viết $\Omega = \{\text{lỗi}, \text{không lỗi}\}$.
- Gieo một xúc xắc và ghi lại số chấm thu được. Không gian mẫu có thể mô tả là tập $\Omega = \{1,2,3,4,5,6\}$.
- Gieo một xúc xắc hai lần và ghi lại số chấm thu được cùng với thứ tự. Không gian mẫu có thể mô tả là tập
$$\Omega = \{1,2,3,4,5,6\}^2,$$
gồm 36 phần tử.
- Gieo hai đồng xu và ghi lại số mặt sấp. Không gian mẫu có thể mô tả là tập $\Omega = \{0,1,2\}$, gồm 3 phần tử.

Trong các ví dụ trên, chúng ta không thể biết được kết quả nào sẽ xảy ra trước khi phép thử được thực hiện. Các phép thử như vậy là đối tượng nghiên cứu chủ yếu của lí thuyết xác suất, và được gọi là *phép thử ngẫu nhiên* (random trial).

Trong một phép thử, chúng ta có thể quan tâm đến một kết quả duy nhất, hoặc một nhóm các kết quả có một tính chất chung nào đó. Ví dụ, trong phép thử gieo xúc xắc, xét sự kiện "*Số chấm xuất hiện là số lẻ*". Sự kiện này xảy ra khi chúng ta thu được một trong nhóm các kết quả 1, 3, hoặc 5, và được mô tả bởi tập hợp $A = \{1,3,5\}$. Nó là tập con của không gian mẫu, tức là $A \subset \Omega$. Điều này được tổng quát hoá thành khái niệm *sự kiện* (một số sách gọi là *biến cố*) như sau:

Định nghĩa 1.4:

Một tập con của không gian mẫu được gọi là một *sự kiện*.

Trong một số sách, *sự kiện* còn được gọi là *biến cố*.

Như vậy, nếu A là một sự kiện và Ω là không gian mẫu (cùng liên đến một thí nghiệm T nào đó) thì $A \subset \Omega$. Trong ví dụ trước, "Số chấm xuất hiện là số lẻ" cho sự kiện $A = \{1,3,5\}$, là một tập con của không gian mẫu $\Omega = \{1,2,3,4,5,6\}$.

Nếu A là một sự kiện liên quan đến một thí nghiệm T đã được tiến hành, kết quả a xảy ra với $a \in A$, thì chúng ta nói A *xảy ra*. Chúng ta cũng nói a là một *kết quả thuận lợi* cho sự kiện A. Trong ví dụ nói trên, $A = \{1,2,3\}$, nếu $a = 3$ xảy ra thì a là một kết quả thuận lợi cho A.

1.2.2 Các phép toán trên sự kiện

Các phép toán cơ bản trên sự kiện cũng là các phép toán cơ bản trong lí thuyết tập hợp "ngây thơ" (naive set theory), được biết đến trong toán học ở bậc trung học và cần thiết cho hầu hết các lĩnh vực của toán học.

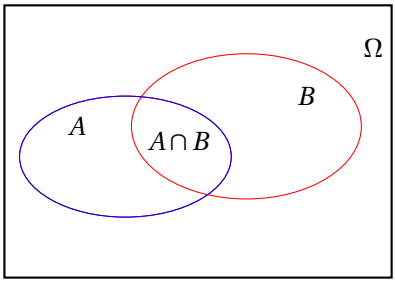
Hai phép toán cơ bản nhất đối với các sự kiện là phép hợp thành (tổng) và phép giao (tích), được định nghĩa như sau:

Định nghĩa 1.5:

Giả sử A và B là hai sự kiện liên quan đến cùng một phép thử T.

- *Tổng* $A \cup B$ là sự kiện gồm tất cả các điểm mẫu của A *hoặc* của B (hoặc của cả hai).
- *Tích* $A \cap B$ là sự kiện gồm tất cả các điểm mẫu thuộc vào cả A và B.

Tổng và tích của nhiều sự kiện được định nghĩa tương tự.



Hình 1.3: Biểu đồ Venn minh họa không gian mẫu Ω và hai sự kiện A và B với giao khác rỗng

Như vậy, hai phép toán trên hoàn toàn giống với hai phép *hợp thành* và *giao* trong lí thuyết tập hợp. Chúng cũng là tổng và tích trong đại số Boole (Boolean algebra), thường được minh hoạ bằng các biểu đồ Venn (Venn diagram, xem Hình 1.3).

Ví dụ 1.5. Giả sử T là phép thử gieo xúc xắc 1 lần, A là sự kiện "mặt xuất hiện là chẵn", và B là sự kiện "mặt xuất hiện lớn hơn 3". Khi đó,

$$A = \{2,4,6\}, \quad B = \{4,5,6\}.$$

Như vậy,

$$A \cup B = \{2,4,5,6\},$$

là sự kiện bao gồm các kết quả a thoả mãn một trong hai điều kiện a chẵn hoặc $a > 3$, hoặc cả hai. Trong khi đó,

$$A \cap B = \{4,6\}$$

là sự kiện bao gồm các kết quả a thoả mãn *đồng thời* hai điều kiện a chẵn và $a > 3$.

Hai sự kiện là xung khắc (hay "loại trừ nhau" – mutually exclusive) nếu chúng không thể xảy ra đồng thời. Nói cách khác, không có kết quả nào có thể thuận lợi cho cả hai sự kiện xung khắc.

Định nghĩa 1.6:

Nếu hai sự kiện A và B liên quan đến một phép thử T thoả mãn

$$A \cap B = \emptyset$$

thì A và B gọi là hai sự kiện *xung khắc*.

Chú ý 1.1. Trong trường hợp A và B là xung khắc, nếu A không xảy ra thì *không nhất thiết* B phải xảy ra. Ví dụ, trong phép thử gieo xúc xắc, nếu $A = \{1,2\}$ và $B = \{4,6\}$ thì A và B là xung khắc, nhưng nếu kết quả xuất hiện mặt 3 chấm thì rõ ràng B không xảy ra, và A cũng không xảy ra.

Định nghĩa 1.7:

Nếu A và B xung khắc và nếu $A \cup B = \Omega$ (không gian mẫu) thì B được gọi là *sự kiện đối* của A, kí hiệu $B = \overline{A}$.

Như vậy, nếu sự kiện đối $\overline{A}$ của A gồm các kết quả nằm trong phần bù $\overline{A} = \Omega \setminus A$ trong Ω của A. Ví dụ, trong thí nghiệm gieo xúc xắc, $\Omega = \{1,2,3,4,5,6\}$, giả sử $A = \{1,2\}$. Khi đó,

$$\overline{A} = \Omega \setminus A = \{3,4,5,6\}.$$

Bài tập §1.2

1. Xác định hoặc mô tả không gian mẫu của các thí nghiệm sau:

(a) Lấy ngẫu nhiên 3 trái cây trong hộp chứa cả táo và đào

(b) Tung hai đồng xu

(c) Tung hai xúc xắc

(d) Tung một xúc xắc liên tục nhiều lần cho đến khi mặt "Sáu" xuất hiện

(e) Chọn ra một nhóm gồm 2 người trong 5 người

(f) Lấy từng trái cam (không hoàn lại) trong một hộp có 10 trái, trong đó 1 trái bị hỏng

2. Tung 3 xúc xắc. Các sự kiện A: *Tổng số chấm chia hết cho 3* và B: *Tổng số chấm chia hết cho 5* có xung khắc hay không?

3. (a) Sự kiện đối của không gian mẫu là gì?

(b) Chứng minh rằng sự kiện đối của tổng hai sự kiện bằng tích của các sự kiện đối của chúng.

(c) Chứng minh rằng sự kiện đối của tích hai sự kiện bằng tổng của các sự kiện đối của chúng.

1.3 Hoán vị, chỉnh hợp, tổ hợp

Quy tắc đếm là một trong những công cụ cơ bản và quan trọng nhất trong toán học tổ hợp và xác suất. Những quy tắc này giúp chúng ta tính toán số lượng khả năng có thể xảy ra của một sự kiện hoặc kết hợp các đối tượng khác nhau mà không cần phải liệt kê tất cả các trường hợp. Trong xác suất, việc biết cách đếm đúng số trường hợp có thể xảy ra là bước đầu tiên để tính toán xác suất của một sự kiện. Các quy tắc đếm cơ bản mà chúng ta sẽ nói chi tiết sau đây là quy tắc cộng, quy tắc nhân, chỉnh hợp, tổ hợp, và số hoán vị.

Tài liệu tham khảo: Kreyszig [1, §24.3].

1.3.1 Quy tắc cộng và nhân

Nếu sự kiện A có n_1 cách xảy ra và sự kiện B có n_2 cách xảy ra, và hai sự kiện này không chồng chéo (tức là không xảy ra đồng thời), thì số cách để A hoặc B xảy ra là tổng của n_1 và n_2. Ví dụ, một nhóm có 5 nam và 7 nữ. Chọn một người từ nhóm này, vậy số cách chọn được là $5+7=12$ cách.

Mặt khác, nếu một sự kiện có thể được chia thành nhiều bước, và mỗi bước có số lượng khả năng khác nhau, thì toàn bộ cách thực hiện sự kiện đó là tích của số cách thực hiện từng bước. Đó là quy tắc nhân. Ví dụ, một quán cà phê có 3 loại cà phê và 4 loại sữa. Để chọn một loại đồ uống gồm cà phê và sữa, chúng ta có $3\times 4=12$ cách chọn.

1.3.2 Chỉnh hợp, tổ hợp, và số hoán vị

Chỉnh hợp là số cách sắp xếp k phần tử từ một tập hợp gồm n phần tử. Số các chỉnh hợp chập k của n phần tử là

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

Mặt khác, tổ hợp là số cách chọn k phần tử từ một tập hợp gồm n phần tử mà không quan tâm đến thứ tự. Số các tổ hợp chập (combination) k của n phần tử là

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

Cuối cùng, số các hoán vị của n phần tử là

$$P = n!.$$

Lưu ý, để thuận tiện về mặt kí hiệu, chúng ta quy ước $0! = 1$.

Ví dụ 1.6. Một hộp chứa 5 quả cầu màu xanh và 8 quả cầu màu vàng. Xét phép thử chọn ngẫu nhiên đồng thời 3 quả cầu từ hộp đó. Khi đó số phần tử của không gian mẫu là

$$C_{13}^3 = \frac{13!}{3!(13-3)!} = 286.$$

CAS 1.5. Trong R, số các cách chọn k phần tử trong một tổng thể gồm n phần tử được tính bằng hàm `choose()`. Ví dụ, để tính số các cách chọn một số lá bài trong một bộ bài, chúng ta tính như sau:

```
# Khai báo các tham số n và k
n = 52; k = 9

# Tính tổ hợp chập k = 9 của n = 52 phần tử
result = choose(n, k)

# In ra kết quả
print(result)

## [1] 3,679e+09
```

Kết quả thu được là số tổ hợp chập 9 của 52 phần tử, $C_{52}^9 = 3{,}6791 \times 10^9$. Đó chính là số cách chọn ra 9 lá bài trong bộ bài gồm 52 lá bài.

Bài tập mục 1.3

1. Liệt kê tất cả các hoán vị của các chữ cái A, B, C, và D. Có tất cả bao nhiêu hoán vị của chúng?
2. Có bao nhiêu cách chọn ra 5 chiếc bút trong hộp 12 chiếc?
3. Có bao nhiêu cách xếp 10 lái xe vào n xe bus?
4. Có bao nhiêu cách chọn 3 sinh viên trong lớp 40 sinh viên vào một hội đồng gồm các vị trí *chủ tịch, thư kí,* và *uỷ viên*?

1.4 Định nghĩa và các tính chất cơ bản của xác suất

Trong mục này, chúng ta nhắc lại những kiến thức cơ bản về xác suất, bao gồm các định nghĩa và những quy tắc cơ bản về xác suất, xác suất có điều kiện, công thức cộng xác suất, công thức nhân xác suất, công thức xác suất toàn phần và định lí Bayes. Tài liệu tham khảo: Kreyszig [1, §24.3].

1.4.1 Định nghĩa tổng quát của xác suất

Nếu một phép thử T được thực hiện n lần thì *tần suất tuyệt đối* (absolute frequency) của một sự kiện A, kí hiệu là $f(A)$ (f lấy từ chữ frequency), là số lần xuất hiện kết quả thuận lợi đối với A. *Tần suất tương đối* (relative frequency) của A trong n phép thử (đã thực hiện) là

$$f_{\text{rel}}(A) = \frac{f(A)}{n}. \tag{1.6}$$

Tần suất tương đối có một số tính chất sau:

$$0 \le f_{\text{rel}}(A) \le 1, \quad f_{\text{rel}}(\emptyset) = 0, \quad f_{\text{rel}}(\Omega) = 1.$$

Hơn nữa, với cùng n phép thử đó, và với một sự kiện B xung khắc với A, ta có

$$f_{\text{rel}}(A \cup B) = f_{\text{rel}}(A) + f_{\text{rel}}(B).$$

Hoặc tổng quát hơn, với A và B tuỳ ý

$$f_{\text{rel}}(A \cup B) = f_{\text{rel}}(A) + f_{\text{rel}}(B) - f_{\text{rel}}(A \cap B).$$

Một nhu cầu thực tế nảy sinh là ước tính các tần suất tương đối của một sự kiện trước khi n phép thử đó được thực hiện. Nhu cầu này liên quan đến khái niệm xác suất thực nghiệm (empirical probability) của một sự kiện.

Xác suất thực nghiệm của một sự kiện A liên quan đến một phép thử T được dùng để đánh giá khả năng xảy ra A khi phép thử T được thực hiện nhiều lần. Nếu thực nghiệm tung một đồng xu cân xứng, đồng chất nhiều

lần, khả năng xuất hiện mặt sấp (S) và ngửa (N) ngang nhau. Xác suất thực nghiệm của sự kiện xuất hiện mặt (S) cũng như sự kiện xuất hiện mặt (N) sẽ đều xấp xỉ bằng 0,5. Nếu một xúc xắc (dice) *cân xứng, đồng chất* được gieo nhiều lần, thì xác suất thực nghiệm của sự kiện xuất hiện các mặt 1, 2, đến 6 là ngang nhau và đều xấp xỉ bằng 1/6. Đó là hai ví dụ về những thực nghiệm mà không gian mẫu có hữu hạn điểm mẫu trong đó, với những lí do về tính đối xứng (tính cân xứng, đồng chất), mà khả năng xuất hiện mỗi kết quả được xem là như nhau. Khi đó, chúng ta có thể cho tương ứng mỗi sự kiện với một giá trị số mà chúng có thể đánh giá một cách hợp lí khả năng một sự kiện xảy ra. Mỗi giá trị như vậy gọi là xác suất (lí thuyết) của sự kiện A. Điều quan trọng ở đây là *luật số lớn* đảm bảo rằng các xác suất thực nghiệm của một sự kiện sẽ hội tụ, theo một nghĩa nào đó, tới xác suất lí thuyết của nó.

Định nghĩa 1.8:

Giả sử một phép thử T có không gian mẫu Ω chứa một số hữu hạn các điểm mẫu, khả năng các điểm mẫu xuất hiện là như nhau, thì *xác suất* cho một sự kiện A xảy ra là

$$P(A) = \frac{\text{Số các điểm mẫu trong A}}{\text{Số các điểm mẫu trong } \Omega} = \frac{|A|}{|\Omega|}, \tag{1.7}$$

ở đó $|A|$ là số các điểm mẫu trong một sự kiện A (và tất nhiên $|\Omega|$ là số các điểm mẫu trong không gian mẫu).

Từ định nghĩa, chúng ta có $P(\emptyset) = 0$, $P(\Omega) = 1$, trong khi đó

$$0 \le P(A) \le 1 \quad \forall A \subset \Omega. \tag{1.8}$$

Các tính chất này tương tự như tính chất của tần suất tương đối.

Ví dụ 1.7. Gieo một con xúc xắc đồng nhất. Tìm xác suất của sự kiện sau: *"Số chấm thu được là số lẻ"*.

Lời giải

Với giả thiết xúc xắc là đồng nhất, chúng ta coi khả năng xuất hiện bất kì một mặt nào đó trong 6 mặt là như nhau. Do đó, nếu A là sự kiện xảy ra khi xuất hiện mặt có số điểm là lẻ thì $A = \{1,3,5\}$, và vậy thì A có 3 điểm mẫu: $|A| = 3$. Mặt khác, không gian mẫu $\Omega = \{1,2,3,4,5,6\}$ có 6 điểm mẫu: $|\Omega| = 6$. Do đó,

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = 0{,}5.$$

Chúng ta cũng nói xác suất xuất hiện sự kiện A là 0,5, hoặc 50%.

Ví dụ 1.8. Trong cửa hàng còn 100 bóng đèn trong đó có 4 bóng đèn bị hỏng. Người bán hàng lấy ngẫu nhiên 2 bóng đèn giao cho khách hàng. Tìm xác suất người khách hàng đó mua ít nhất một bóng đèn hỏng.

Lời giải

Giả sử chúng ta đánh số tác bóng đèn từ 1 đến 100, trong đó bóng thứ nhất và thứ hai là các bóng bị hỏng. Phép thử ở đây là chọn ngẫu nhiên hai bóng đèn trong 100 bóng đèn. Kết quả của mỗi phép thử như vậy là một cặp số (a,b) với $a \neq b$, là thứ tự của các bóng đèn được chọn. Không gian mẫu Ω là tập tất cả các kết quả và như vậy: $|\Omega| = C_{100}^2 = 4950$.

Gọi $A \subset \Omega$ là tập tất cả các cặp (a,b) trong đó $a \neq b$ và ít nhất một trong chúng thuộc $\{1,2\}$. Người bán hàng chọn phải ít nhất một bóng hỏng nếu kết quả phép thử nằm trong A . Mặt khác, $|A| = 99+98 = 197$. Mỗi kết quả (a,b) có cùng khả năng xảy ra nên xác suất A xảy ra là

$$P(A) = \frac{|A|}{|\Omega|} = \frac{197}{4950} = 0{,}03979798.$$

Xác suất người hàng mua phải bóng hỏng đúng bằng xác suất của A, xấp xỉ 4%.

Trong trường hợp tổng quát ở đó giả thiết về khả năng xuất hiện các kết quả trong không gian mẫu như nhau không còn nữa, chúng ta có khái niệm xác suất và không gian xác suất (probability space) như sau:

Định nghĩa 1.9:

Cho một không gian mẫu Ω, mỗi sự kiện $A \subset \Omega$, tương ứng với một số $P(A)$, gọi là *xác suất* của A, sao cho các tiên đề sau đồng thời thỏa mãn:

1. Với mỗi A trong Ω,

$$0 \leq P(A) \leq 1. \tag{1.9}$$

2. Toàn bộ không gian mẫu có xác suất bằng đơn vị:

$$P(\Omega) = 1. \tag{1.10}$$

3. Với hai sự kiện A và B *xung khắc*, tức là $A \cap B = \emptyset$,

$$P(A \cup B) = P(A) + P(B). \tag{1.11}$$

Nếu Ω có vô hạn các phần tử, tiên đề này được thay bởi tiên đề:

3′. Với dãy các sự kiện $A_1, A_2, \ldots, A_n, \ldots$ xung khắc từng đôi (nghĩa là $A_i \cap A_j = \emptyset$ với mọi $i \neq j$)

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \tag{1.12}$$

trong đó, vế phải là một chuỗi số dương hội tụ.

Nhận xét 1.3. Trong trường hợp Ω có vô hạn phần tử, đếm được (như tập số tự nhiên), hay không đếm được (như tập các số thực trên $[0,1]$), thì người ta sẽ chỉ xét một lớp nhất định gọi là $\mathscr{F}$, các tập con của Ω thoả mãn các tiên đề của các "*σ-đại số*"[1]. Đó là một lớp các tập con mà người ta có thể gán cho mỗi tập con một xác suất một cách "hợp lí". Một sự kiện sẽ là một tập con của Ω nằm trong $\mathscr{F}$. Như vậy, không phải mọi tập con của Ω đều là một sự kiện. May mắn là hạn chế này không ảnh hưởng tới khả năng ứng dụng của lí thuyết xác suất trong thực tiễn, bởi vì các tập con của Ω không nằm trong σ-đại số $\mathscr{F}$ thường không xuất hiện trong thực tiễn, trong khi việc xây dựng chúng một cách trừu tượng cũng không phải là một việc đơn giản.

[1] Sigma đại số là một khái niệm quan trọng trong lí thuyết xác suất và giải tích, đặc biệt là trong lí thuyết tích phân Lebesgue.

1.4.2 Các định lí cơ bản của xác suất

Một số quy tắc tính xác suất tuy đơn giản nhưng rất hữu ích, chẳng hạn như quy tắc bù sau đây:

Mệnh đề 1.1: (Quy tắc bù)

Cho một sự kiện A và sự kiện đối lập của nó là $\overline{A}$ trong một không gian mẫu Ω, khi đó

$$P\left(\overline{A}\right) = 1 - P(A). \tag{1.13}$$

Chứng minh. Từ định nghĩa của sự kiện đối, chúng ta có $A \cap \overline{A} = \emptyset$ và $A \cup \overline{A} = \Omega$. Do đó

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

Từ đó dễ dàng thu được đẳng thức cần chứng minh. □

Ví dụ 1.9. Tung 10 đồng xu đồng nhất đồng thời. Tìm xác suất của biến cố: *"Ít nhất 2 mặt sấp xuất hiện"*.

Lời giải

Không gian mẫu Ω có thể được xem là tập các chuỗi gồm 10 kí tự S và N thể hiện các kết quả sấp (S) hoặc ngửa (N) của các đồng xu sau khi tung. Rõ ràng

$$|\Omega| = 2^{10} = 1\,024,$$

trong khi đó khả năng thu được một chuỗi nhất định là như nhau. Đây là một giả thiết về tính *đồng khả năng*, có được từ những yếu tố thực tế như các đồng xu là cân xứng và được tung một cách độc lập nhau.

Để tìm xác suất của A một cách trực tiếp, chúng ta cần đếm số các điểm mẫu trong A. Cách làm này tương đối khó khăn, do có 9 trường hợp tương ứng với số mặt sấp đúng bằng 2, 3, ..., 10. Tuy nhiên, nếu đếm các điểm mẫu trong sự kiện đối $\overline{A}$: "Xuất hiện nhiều nhất 1 mặt sấp," thì chúng ta dễ dàng thấy được

$$\overline{A} = B_0 \cup B_1, \quad B_1 \cap B_2 = \emptyset,$$

trong đó, B_i là sự kiện "Xuất hiện đúng i mặt sấp". Dễ thấy,

$$|B_0| = 1, \quad |B_1| = 10.$$

Áp dụng định nghĩa cổ điển của xác suất để thu được

$$P(\overline{A}) = P(B_0) + P(B_1) = \frac{11}{1024},$$

và vậy thì $P(A) = 1 - \frac{11}{1024} = \frac{1013}{1024}$.

Mệnh đề 1.2: (Quy tắc cộng)

Nếu A và B là hai sự kiện liên quan đến cùng một phép thử thì

$$P(A \cap B) = P(A) + P(B) - P(A \cap B). \tag{1.14}$$

Chứng minh. Bài tập! □

Ví dụ 1.10. Tung một xúc xắc cân xứng. Gọi A là sự kiện "Số chấm lẻ" và B là sự kiện "Số chấm lớn hơn ba". Khi đó,

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2},$$

bởi vì $A = \{1,3,5\}$ trong khi đó $B = \{4,5,6\}$. Hơn nữa,

$$A \cap B = \{5\},$$

nên $P(A \cap B) = \frac{1}{6}$. Vậy

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}.$$

Kết quả này cũng có thể suy ra từ nhận xét $A \cup B = \{1,3,4,5,6\}$ có 5 kết quả.

Bài tập §1.4

1. Trong hộp có 4 bút chì màu xanh và 2 bút chì mầu đỏ. Lần lượt lấy ngẫu nhiên hai chiếc bút chì trong hộp (không trả lại). Tìm xác suất để:

(a) Lần đầu tiên lấy được bút màu xanh và lần thứ hai lấy được bút màu đỏ.

(b) Lần đầu tiên lấy được bút màu đỏ và lần thứ hai lấy được bút màu xanh.

2. Trong một chiếc hộp có 20 chiếc bút màu đen và 12 chiếc bút màu trắng. Lấy ngẫu nhiên đồng thời 4 chiếc bút trong hộp đó, tính xác suất để lấy được hai bút màu đen và hai bút màu trắng.

3. Lấy ngẫu nhiên 4 lá bài từ một bộ bài 52 lá bài.

(a) Xác định không gian mẫu và số các điểm mẫu.

(b) Tìm xác suất của sự kiện "lấy được 4 lá bài chất cơ".

(c) Tìm xác suất của sự kiện "lấy được 3 lá bài chất rô và 1 lá bài chất nhép".

(d) Tìm xác suất của sự kiện "lấy được 2 lá bài chất bích và 2 lá bài chất rô".

4. Trong một hộp có 8 viên bi màu xanh, 4 viên bi màu đỏ, và 3 viên bi màu vàng. Lấy ngẫu nhiên đồng thời 2 viên bi từ đó. Tính xác suất để có ít nhất 1 viên bi màu đỏ.

1.5 Xác suất có điều kiện

Xác suất có điều kiện (conditional probability) là một khái niệm cơ bản trong lí thuyết xác suất, giúp xác định khả năng một sự kiện xảy ra, dựa trên thông tin rằng một sự kiện khác đã xảy ra. Được ký hiệu là $P(B \mid A)$, nó là xác suất xảy ra sự kiện B khi biết rằng sự kiện A đã xảy ra.

Xác suất có điều kiện được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau. Ví dụ, trong Y học, nó dùng để đánh giá khả năng mắc bệnh khi có các triệu chứng hoặc kết quả xét nghiệm. Ví dụ, nếu biết rằng một người có kết quả xét nghiệm dương tính cho một bệnh, xác suất họ thực sự mắc bệnh l à bao nhiêu dựa trên tỉ lệ mắc bệnh trong cộng đồng. Trong Kinh tế và tài chính, xác suất có điều kiện có thể dùng để đánh giá rủi ro trong đầu tư. Chẳng hạn, xác suất thị trường chứng khoán tăng giá khi có thông tin về thay đổi chính sách tiền tệ là bao nhiêu? Trong thương mại điện tử, xác suất có điều kiện được dùng để đề xuất sản phẩm, như xác suất khách hàng sẽ mua một sản phẩm khi họ đã xem qua sản phẩm tương tự. Trong bảo hiểm hoặc quản trị rủi ro, xác suất có điều kiện giúp đánh giá rủi ro dựa trên các sự kiện đã xảy ra như thiên tai, tai nạn hay hỏa hoạn, v.v.

1.5.1 Định nghĩa xác suất có điều kiện

Định nghĩa 1.10:

Giả sử A là một sự kiện với xác suất dương, $P(A) > 0$. Xác suất của một sự kiện B, biết rằng sự kiện A đã xảy ra, kí hiệu là $P(B \mid A)$, gọi là *"xác suất của A với điều kiện B xảy ra"*. Xác suất điều kiện này được cho bởi công thức sau:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}. \tag{1.15}$$

Lưu ý, nếu A là sự kiện có xác suất bằng 0 thì $P(B \mid A)$ không có nghĩa, do điều kiện A không có khả năng xảy ra.

Ví dụ 1.11. Tung một xúc xắc đồng nhất 2 lần và ghi lại kết quả thu được. Các kết quả có thể viết dưới dạng (a, b) với $a, b \in \{1, 2, \ldots, 6\}$. Ví dụ, nếu lần tung thứ nhất thu được 3 chấm và lần tung thứ hai thu được 2 chấm thì kết quả ghi $(3, 2)$. Không gian mẫu của thực nghiệm này là

$$\Omega = \{(1,1), (1,2), \ldots, (1,6), \ldots, (6,1), (6,2), \ldots, (6,6)\},$$

gồm 36 kết quả: $|\Omega| = 36$.

Xét sự kiện A: *Số chấm thu được ở lần tung thứ nhất là lẻ*". Có nghĩa là

$$A = \{(a,b) \in \Omega \mid a \text{ lẻ}\} = \{1,3,5\} \times \{1,2,3,4,5,6\}.$$

Khi đó, $|A| = 18$ và vậy thì

$$P(A) = \frac{18}{36} = 0{,}5.$$

Tiếp theo, xét sự kiện B: *"Tổng hai kết quả của hai lần tung bằng 5"*. Khi đó,

$$B = \{(1,4),(2,3),(3,2),(4,1)\},$$

với xác suất $P(B) = |B|/|\Omega| = 4/36 = 1/9$.

Bây giờ, lưu ý rằng

$$A \cap B = \{(1,4),(3,2)\},$$

và do đó

$$P(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{2}{36} = \frac{1}{18}.$$

Giả sử chúng ta đã biết A xảy ra, có nghĩa là kết quả lần tung thứ nhất là lẻ. Khi đó, xác suất của B với điều kiện A, xác suất $P(B \mid A)$, là

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{1}{18} \div 0{,}5 = \frac{1}{9}.$$

Nhận xét 1.4. Khi chúng ta tính xác suất của B với điều kiện A, chúng ta đã thu hẹp không gian mẫu Ω về một không gian mẫu mới $\Omega' = A$ và tìm xác suất của một sự kiện $B' = B \cap A$ trong không gian mẫu này. Trong ví dụ trên, với điều kiện A, chúng ta thu hẹp không gian mẫu để có được

$$\Omega' = \{(a,b) \mid a = 1,3,5 \text{ và } b = 1,\ldots,6\},$$

với $|\Omega'| = 18$ kết quả, trong khi đó

$$B \cap \Omega' = \{(1,4),(3,2)\},$$

gồm hai kết quả. Từ đây, dễ dàng thấy $P(B \mid A) = \frac{1}{9}$, đúng như kết quả đã tính.

Từ định nghĩa của xác suất điều kiện, chúng ta có quy tắc nhân như sau:

Mệnh đề 1.3:

Nếu $P(A) > 0$ thì

$$P(A \cap B) = P(B \mid A)P(A). \tag{1.16}$$

Công thức (1.16) gọi là công thức nhân hoặc công thức chuỗi (giống như công thức đạo hàm hàm hợp trong giải tích).

Ví dụ 1.12. Xác suất để một nhân viên nữ trong công ty được chọn ngẫu nhiên sinh vào ngày chủ nhật là 14,2%. Biết rằng tỉ lệ nhân viên nữ trong công ty là 55%. Chọn ngẫu nhiên một nhân viên trong công ty. Tìm xác suất để người được chọn là nữ sinh vào chủ nhật.

Lời giải

Nếu A là sự kiện "Người được chọn là nữ", thì $P(A) = 0{,}55$. Nếu B là sự kiện "Người được chọn sinh vào chủ nhật" thì, theo giả thiết, $P(B \mid A) = 0{,}142$. Xác suất cần tính là $P(A \cap B)$. Theo công thức trên,

$$P(A \cap B) = P(B \mid A)P(A) = 0{,}55 \times 0{,}142 = 0{,}0781.$$

1.5.2 Các sự kiện độc lập

Khái niệm các sự *kiện độc lập* dùng để mô tả các sự kiện mà việc sự kiện này xảy ra hay không không phụ thuộc vào sự kiện kia xảy ra hay không. Ví dụ, nếu gieo hai đồng xu, thì đồng xu thứ nhất là "sấp" hay "ngửa" không phụ thuộc vào trạng thái của đồng xu thứ hai. Trong lí thuyết xác suất, tính chất này được định nghĩa một cách chính xác như sau:

Định nghĩa 1.11:

Hai sự kiện A và B thỏa mãn

$$P(A \cap B) = P(A)P(B), \tag{1.17}$$

được gọi là hai sự kiện *độc lập*.

Ví dụ 1.13. Xét thí nghiệm gieo xúc xắc hai lần. Không gian mẫu Ω gồm các cặp (a, b), với $a, b \in \{1,2,\ldots,6\}$, gồm 36 điểm mẫu. Xét sự kiện A: "Kết quả lần gieo thứ nhất là chẵn", và B: "Kết quả lần gieo lần thứ hai là bé hơn 3". Khi đó,

$$A = \{2,4,6\} \times \{1,2,3,4,5,6\},$$

gồm $3 \times 6 = 18$ điểm mẫu: $|A| = 18$. Trong khi đó,

$$B = \{1,2,3,4,5,6\} \times \{1,2\}$$

gồm $2 \cdot 6 = 12$ điểm mẫu: $|B| = 12$. Như vậy,

$$P(A) = 18/36 = 1/2, \quad P(B) = 12/36 = 1/3.$$

Để kiểm tra xem A và B có độc lập hay không, chúng ta tìm $A \cap B$, đó là tập các kết quả với lần gieo thứ nhất được số chẵn chấm còn lần gieo thứ hai

có số chấm ít hơn 3. Rõ ràng,

$$A \cap B = \{(2,1),(2,2),(4,1),(4,2),(6,1),(6,2)\},$$

gồm 6 điểm mẫu: $|A \cap B| = 6$. Từ đó

$$P(A \cap B) = 6/36 = 1/6 = P(A)P(B).$$

Từ kết quả này, chúng ta kết luận A và B là hai sự kiện độc lập.

Ví dụ trên phản ánh một một quan sát thực tế là, trong mô hình trên, kết quả của lần gieo thứ nhất không ảnh hưởng của tới kết quả của lần gieo thứ hai, và "khả năng" cả hai sự kiện cùng xảy ra được đánh giá bằng tích của các khả năng của từng sự kiện độc lập xảy ra.

Ở một góc nhìn khác, từ quy tắc nhân, chúng ta có nếu A và B là hai sự kiện với xác suất dương thì chúng độc lập nếu và chỉ nếu

$$P(A \mid B) = P(A),$$

hoặc

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

Nói cách khác, xác suất để sự kiện này xảy ra không phụ thuộc vào điều kiện sự kiện kia có xảy ra hay không.

Ví dụ 1.14. Hai khẩu súng cùng bắn vào một đích với xác suất bắn trúng lần lượt là 0,6 và 0,8. Tìm xác suất

(a) ít nhất một khẩu súng bắn trúng đích;

(b) cả hai khẩu cùng trúng đích.

Lời giải

Nếu gọi A là sự kiện khẩu súng thứ nhất bắn trúng và B là sự kiện khẩu súng thứ hai bắn trúng đích thì $P(A) = 0{,}6$ và $P(B) = 0{,}8$. Xác suất cả hai khẩu bắn trúng đích là xác suất của sự kiện tích $A \cap B$. Do A và B độc lập nên

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ &= 0{,}6 \times 0{,}8 \\ &= 0{,}48. \end{aligned}$$

Cuối cùng, sử dụng công thức cộng, ta tính xác suất ít nhất một khẩu bắn trúng là

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0{,}6 + 0{,}8 - 0{,}48 = 0{,}92,$$

tức là 92%.

Lấy mẫu

Giả sử chúng ta có một nhóm đối tượng với các đặc tính nào đó. Xét thử nghiệm lấy ngẫu nhiên một đối tượng và quan sát đặc tính của nó. Khi đó, xác suất để thu được một đặc tính nhất định phụ thuộc vào số đối tượng có đặc tính đó trong nhóm. Một tình huống thú vị xảy ra khi chúng ta tiếp tục lấy mẫu trong nhóm đối tượng đó. Để tìm hiểu kĩ hơn, hãy xét ví dụ sau đây.

Ví dụ 1.15. Trong hộp có 100 sản phẩm trong đó có 15 sản phẩm lỗi và 85 sản phẩm không lỗi. Lấy ngẫu nhiên liên tiếp 2 sản phẩm từ trong hộp (không trả lại sản phẩm sau khi lấy). Tìm xác suất của sự kiện B: "*Sản phẩm lấy lần thứ hai bị lỗi*" với điều kiện A: "*Sản phẩm lấy lần thứ nhất bị lỗi.*"

Điều thú vị là khi xem hai lần lấy mẫu là hai phép thử liên tiếp, thì xác suất liên quan đến phép thử thứ hai phụ thuộc vào kết quả của phép thử thứ nhất.

Chúng ta có hai trường hợp:

- Nếu lần đầu tiên lấy được sản phẩm lỗi, thì xác suất để lấy sản phẩm lần 2 bị lỗi là $P_2 = 14/99$, vì lúc này có 99 sản phẩm trong hộp với 14 sản phẩm lỗi.
- Nếu lần lấy mẫu đầu tiên thu được sản phẩm không lỗi, thì xác suất để lần thứ hai có lỗi là $P_2 = 15/99$.

Trong thí nghiệm này, không gian mẫu Ω gồm tất cả các cặp *có thứ tự* gồm 2 sản phẩm. Khi đó,

$$|\Omega| = A^2_{100} = 9900.$$

Biến cố A: "*Sản phẩm lấy lần thứ nhất bị lỗi*" đồng nhất với tập con $A \subset \Omega$ gồm các cặp hai sản phẩm (a, b), a lỗi. Bởi quy tắc đếm trong mục trước, $|A| = 15 \cdot 99 = 1485$. Do đó,

$$P(A) = \frac{1485}{9900} = 0{,}15.$$

Kết quả này hoàn toàn phù hợp với quan sát khả năng lấy được sản phẩm lỗi hay không trong lần lấy thứ nhất không phụ thuộc vào kết quả của lần thứ hai. Trong khi đó, trước khi lấy sản phẩm lần thứ nhất, có 100 sản phẩm với 15 sản phẩm lỗi.

Mặt khác, nếu B là sự kiện "*Sản phẩm lấy lần thứ hai bị lỗi*" thì $A \cap B$ là sự kiện *"Cả hai sản phẩm đều bị lỗi"*, tức là tập các cặp sản phẩm (a, b) cùng bị lỗi. Có tất cả 15 sản phẩm lỗi nên số các cặp như vậy là số các chỉnh hợp chập 2 của 15 phần tử: $|A \cap B| = A^2_{15} = 210$. Từ đó

$$P(A \cap B) = \frac{2100}{9900} = \frac{7}{330} \approx 0{,}0212121\ldots$$

Như vậy, xác suất của sự kiện *"Sản phẩm thứ hai có lỗi"* với điều kiện *"Sản phẩm thứ nhất có lỗi"* là

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{14}{99} \approx 0{,}1414141\ldots$$

Điều thú vị là kết quả này hoàn toàn phù hợp với quan sát thực tế là nếu A xảy ra thì trước khi lấy sản phẩm lần thứ hai, có 99 sản phẩm trong hộp trong đó có 14 sản phẩm bị lỗi.

1.5.3 Xác suất toàn phần và công thức Bayes

Mục này nói về công thức Bayes[1] và công thức xác suất toàn phần liên quan đến khái niệm *hệ đầy đủ* các sự kiện.

Định nghĩa 1.12: (Hệ đầy đủ)

Các sự kiện $A_1, A_2, \ldots, A_n$ của một không gian mẫu Ω được gọi là một *hệ đầy đủ* nếu

1. Chúng xung khắc từng đôi: $A_i \cap A_j = \emptyset$, với mọi $i \neq j$,
2. $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$.

Rõ ràng nếu $A_1, A_2, \ldots, A_n$ là một hệ đầy đủ thì

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

Như vậy, xác suất của một sự kiện B có thể tính bằng tổng các xác suất với điều kiện A_i, $i = 1, 2, \ldots, n$. Cụ thể hơn, chung ta có định lí sau.

Định lí 1.1:

Giả sử $\{A_1, \ldots, A_n\}$ là một hệ đầy đủ các sự kiện trong không gian mẫu Ω. Khi đó, với mọi sự kiện $B \subset \Omega$, ta có

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i). \tag{1.18}$$

Chứng minh. Bài tập. □

[1] Thomas Bayes (khoảng 1701–1761) là một nhà toán học và thống kê học người Anh.

Công thức Bayes

Một trong những kết quả cơ bản nhưng đóng vai trò quan trọng trong xác suất và thống kê là công thức Bayes. Công thức này là nền tảng cho lí thuyết thống kê Bayes.

Nếu A và B là hai sự kiện trong một không gian Ω với các xác suất dương: $P(A), P(B) > 0$. Theo công thức xác suất điều kiện,

$$P(A \cap B) = P(B \mid A)P(A).$$

Từ đó suy ra

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}. \tag{1.19}$$

Nhưng A và $\overline{A}$ lập thành một hệ đầy đủ nên

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

Thay vào (1.19), chúng ta thu được

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})(1 - P(A))}. \tag{1.20}$$

Đó chính là trường hợp đơn giản nhất của công thức Bayes .

Ví dụ 1.16. Giả sử có một loại xét nghiệm để phát hiện một bệnh hiếm gặp với độ chính xác là 95%. Giả sử 1% dân số mắc bệnh này. Nếu một người được chọn ngẫu nhiên từ cộng đồng được xét nghiệm có kết quả dương tính, thì xác suất thực sự mắc bệnh là bao nhiêu?

Lời giải

Gọi D là biến cố người bị nhiễm bệnh và T là biến cố xét nghiệm cho kết quả dương tính. Độ chính xác của xét nghiệm là 95% có nghĩa là đối với người có bệnh, xác suất để họ dương tính với xét nghiệm là

$$P(T \mid D) = 0{,}95,$$

và đối với người không bị bệnh, xác suất họ xét nghiệm âm tính là

$$P(\overline{T} \mid \overline{D}) = 0{,}95.$$

Do đó,

$$P(T \mid \overline{D}) = 1 - P(\overline{T} \mid \overline{D}) = 0{,}05.$$

Chúng ta được yêu cầu tìm xác suất một người bị bệnh với điều kiện người đó có xét nghiệm dương tính, tức là xác suất có điều kiện $P(D \mid T)$.

Chúng ta có

$$\begin{aligned} P(T) &= P(T \mid D)P(D) + P(T \mid \overline{D})P(\overline{D}) \\ &= 0{,}95 \cdot 0{,}01 + 0{,}05 \cdot 0{,}99 = 0{,}059, \end{aligned}$$

và khi thay vào công thức Bayes, thì

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T)} = \frac{0{,}95 \times 0{,}01}{0{,}059} \approx 0{,}161.$$

Vậy xác suất thực sự mắc bệnh khi có kết quả dương tính từ xét nghiệm đó là khoảng 16,1%.

Nhận xét 1.5. Trong ví dụ trên kết quả $P(D \mid T) \approx 0{,}161$ tính được do xác suất bị bệnh trong cộng đồng, $P(D) = 0{,}01$, đã biết. Giá trị $P(D \mid T) \approx 0{,}161$ không biết được nếu chúng ta không biết được mức độ phổ biến của bệnh trong cộng đồng.

Trong thực tế, xác suất để một loại xét nghiệm cho kết quả chính xác hay không phụ thuộc vào người được xét nghiệm có bị bệnh hay không, có nghĩa là $P(T \mid D)$ không nhất thiết bằng $P(\overline{T} \mid \overline{D})$ như trong ví dụ trên. Giá trị $P(T \mid D)$, cho xác suất người có bệnh được chuẩn đoán đúng, được gọi là *độ nhạy* (sensitivity) của xét nghiệm, giá trị $P(\overline{T} \mid \overline{D})$, cho xác suất người không có bệnh được chuẩn đoán đúng, được gọi là *độ đặc hiệu* (specificity) của xét nghiệm. Giá trị $P(D \mid T)$ gọi là *giá trị dự báo dương tính* (positive predictive value – PPV), còn $P(D)$ là mức độ phổ biến của bệnh trong cộng đồng (prevalence of disease).

Trở lại Ví dụ 1.16, nếu α là độ nhạy của xét nghiệm, còn β là độ đặc hiệu của xét nghiệm, và p là mức độ phổ biến của bệnh trong cộng đồng thì,

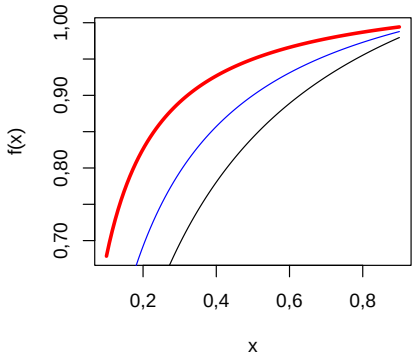
$$\begin{aligned} P(T \mid D) &= \alpha, \\ P(T \mid \overline{D}) &= 1 - P(\overline{T} \mid \overline{D}) \\ &= 1 - \beta, \end{aligned}$$

trong khi đó

$$P(D) = p.$$

Lập luận như trên, từ công thức Bayes chúng ta dễ dàng suy ra giá trị dự báo dương tính PPV(p) là

$$\begin{aligned} \text{PPV}(p) &= P(D \mid T) \\ &= \frac{\alpha p}{(\alpha + \beta - 1)p + 1 - \beta}. \end{aligned} \tag{1.21}$$



Hình 1.4: Đồ thị của PPV với những giá trị khác nhau của α và β

Nhận thấy khi $\beta = 1$ thì
$PPV(p) = 1$ với mọi p, như kì vọng.
Nhưng trường hợp này hiếm gặp trên thực tế. Thông thường chúng ta có $0 < \alpha < 1$ và $0 < \beta < 1$. Với điều kiện $\beta < 1$ thì $\text{PPV}(p)$ phụ thuộc một cách phi tuyến vào p (nó là một hàm hữu tỉ bậc 1 của p).

Từ (1.21) ta có $\text{PPV}(p) \to 0$ khi $p \to 0$. Như vậy, đối với các bệnh hiếm gặp thì giá trị dự báo PPV cũng rất bé, ít có ý nghĩa. Mặt khác, khai triển Taylor tại $p = 0$ có dạng

$$\text{PPV}(p) = \frac{\alpha p}{1-\beta} + O(p^2), \quad p \to 0.$$

Hệ số $\alpha/(1-\beta)$ cho thấy, đối với bệnh hiếm gặp, người được xét nghiệm có kết quả dương tính có khả năng mắc bệnh cao gấp bao nhiêu lần người không được xét nghiệm.

Từ (1.20) và công thức xác suất toàn phần (1.18), chúng ta có thể mở rộng công thức Bayes cho một hệ đầy đủ các sự kiện như sau:

Định lí 1.2:

Giả sử $\{A_1, A_2, \ldots, A_n\}$ là một hệ đầy đủ các sự kiện trong không gian mẫu Ω. Nếu $B \subset \Omega$ sao cho $P(B) > 0$, thì

$$P(A_i \mid B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)}.$$

Bài tập 1.3. Giả sử theo một thống kê, có 68% thư điện tử là thư rác. Một bộ lọc thư điện tử có thể phân loại đúng được 98% thư rác, và phân loại nhầm 3% thư thường là thư rác. Tính xác suất một thư điện tử là thư thường biết rằng nó bị bộ lọc này phân loại là thư rác.

1.5.4 Dãy phép thử Bernoulli

Giả sử chúng ta có một phép thử với hai kết quả, gọi là "thành công" và "thất bại". Gọi A là sự kiện thu được kết quả thành công, với xác suất $P(A) = p$ là một tham số cố định, $0 < p < 1$. Chúng gọi là *phép thử Bernoulli*.

Một dãy phép thử Bernoulli gồm các phép thử Bernoulli *độc lập* và có cùng xác suất thành công p. Dãy phép thử Bernoulli có hai thuộc tính cơ bản:

- Độc lập: Mỗi phép thử trong dãy là độc lập với các phép thử khác. Điều này có nghĩa là kết quả của mỗi phép thử không ảnh hưởng đến kết quả của các phép thử khác.

- Cùng xác suất thành công: Tất cả các phép thử trong dãy có cùng xác suất thành công p.

Dãy phép thử Bernoulli là mô hình xác suất của n phép thử độc lập trong đó không gian mẫu có 2^n điểm mẫu, $|\Omega| = 2^n$, với các điểm mẫu có thể được ghi dưới dạng dãy các số 0 hoặc 1, với số 1 (hoặc 0) ở vị trí thứ k biểu thị kết quả thành công (hoặc thất bại) ở phép thử thứ k. Không gian mẫu trong dãy phép thử Bernoulli không phải là đồng khả năng. Một điểm mẫu có k thành công và $n-k$ thất bại sẽ có xác suất xảy ra là $p^k(1-p)^{n-k}$. Số các điểm mẫu như vậy là tổ hợp chập k của n phần tử, tức là bằng $C_n^k = n!/(k!(n-k)!$. Như vậy, xác suất của sự kiện có đúng k kết quả thành công là

$$P_n(k;p) = C_n^k p^k (1-p)^{n-k}. \tag{1.22}$$

Nhận xét 1.6. Lưu ý rằng, từ công thức nhị thức Newton, chúng ta có

$$\sum_{k=0}^{n} P_n(k;p) = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = (p+1-p)^n = 1.$$

Đẳng thức này phù hợp với nhận xét rằng các sự kiện A_k: "có đúng k phép thử thành công", $k = 0,1,2,\ldots,n$, là một hệ đầy đủ.

Ví dụ 1.17. Một lô hàng có một số lượng lớn bưởi da xanh được người ta phân loại theo cách sau: Lấy ngẫu nhiên 25 quả bưởi làm mẫu và kiểm tra chất lượng xem chúng "đạt" hay "không đạt". Nếu không có quả nào không đạt chất lượng thì lô hàng được xếp loại A; Nếu có một hoặc hai quả không đạt chất lượng thì lô hàng xếp loại B; Còn nếu có từ 3 quả không đạt chất lượng thì lô hàng xếp loại C. Giả sử lô hàng đó thực tế có 2% quả bưởi không đạt chất lượng.

(a) Tìm xác suất để lô hàng đó được xếp loại A.

(b) Tìm xác suất để lô hàng đó được xếp loại B.

(c) Tìm xác suất để lô hàng đó được xếp loại C.

Lời giải

Gọi A, B, và C lần lượt là sự kiện lô hàng được xếp loại A, B, và C.

(a) Vì lô hàng có một số lượng lớn quả bưởi da xanh với 2% số bưởi không đạt, nên ta có thể xem việc lấy mỗi quả bưởi để kiểm tra đều có xác suất 2% lấy được quả không đạt chất lượng và có xác suất 98% lấy được quả đạt chất lượng. Như vậy, chúng ta có thể dùng mô hình dãy

25 phép thử Bernoulli để tính các xác suất như yêu cầu. Cụ thể, xác suất để lấy được 25 quả đạt chất lượng là

$$P(A) = (0{,}98)^{25} \approx 0{,}6034647.$$

(b) Tương tự như (a), xác suất để lô hàng được xếp hạng B là

$$\begin{aligned} P(B) &= C_{25}^{1} \times (0{,}02) \times (0{,}98)^{24} + C_{25}^{2} \times (0{,}02)^{2} \times (0{,}98)^{23} \\ &\approx 0{,}3832918. \end{aligned}$$

Ở trên, chúng ta đã sử dụng công thức cộng: Sự kiện lô hàng xếp hạng B là tổng của hai sự kiện "có 1 quả không đạt chất lượng" và "có 2 quả không đạt chất lượng".

(c) Vì A, B, và C là một hệ đầy đủ nên xác suất để lô hàng được xếp loại C là

$$P(C) = 1 - P(A) - P(B) \approx 0{,}01324343.$$

Bài tập §1.5

1. Cho A là một sự kiện. Khi nào thì A độc lập với chính nó?

2. Nếu A và B là hai sự kiện độc lập thì các sự kiện đối lập $\overline{A}$ và $\overline{B}$ có độc lập hay không? Tại sao?

3. Gieo đồng thời hai con xúc xắc cân đối. Tính xác suất để tổng số chấm xuất hiện lớn hơn 10, biết rằng ít nhất một mặt 5 chấm xuất hiện.

4. Một chùm chìa khoá gồm 9 chiếc trong đó có 2 chiếc mở được ổ khoá. Thử ngẫu nhiên từng chìa vào ổ khoá. Tìm xác suất để ổ khoá được mở ở lần thử thứ ba.

5. (Cassells–Schoenberger–Graboys[1]) Giả sử trong thành phố có 1/1000 người mắc một loại bệnh. Giả sử có một loại xét nghiệm mà ai mắc bệnh cũng cho kết quả "dương tính" trong khi đó có 5% người không mắc bệnh cũng cho kết quả dương tính. Xét nghiệm ngẫu nhiên một người. Hỏi xác suất người đó mắc bệnh là bao nhiêu nếu biết người đó cho kết quả xét nghiệm là dương tính?

1.6 Khái niệm về biến ngẫu nhiên

Nội dung quan trọng trong mục này là về khái niệm *biến ngẫu nhiên* (một số sách gọi là *đại lượng ngẫu nhiên*). Cũng giống như phép thử ngẫu nhiên, các biến ngẫu nhiên (BNN) phụ thuộc vào kết quả của phép thử và do đó, chúng ta không xác định được giá trị của nó trước khi thực hiện phép thử. Trong lí thuyết xác suất, điều chúng ta quan tâm không phải là khi nào một BNN nhận một giá trị cho trước nào đó, mà là khả năng xảy ra của sự kiện

[1] Ward Casscells, B.S., Arno Schoenberger, M.D., and Thomas B. Graboys, M.D., *Interpretation by Physicians of Clinical Laboratory Results*, N Engl J Med 1978;299:999-1001

một BNN nhận một giá trị nào đó hoặc giá trị đó nằm trong một khoảng hay một tập hợp nào đó.

Tài liệu tham khảo: Kreyszig [1, §24.5], Cramer [5, Chapters 13-20], Đ. H. Thắng [2].

1.6.1 Khái niệm chung về biến ngẫu nhiên

Định nghĩa 1.13:

Cho Ω là một không gian mẫu. Một *biến ngẫu nhiên* (BNN) là một hàm số $X : \Omega \to \mathbb{R}$ sao cho các xác suất $P(X = a)$ và $P(a < X < b)$ hoàn toàn xác định.

Chúng ta kí hiệu $P(X = a)$ là xác suất của sự kiện "$X = a$", tức là tập con $\{p \in \Omega \mid X(p) = a\}$ của không gian mẫu Ω. Tương tự, $P(a < X < b)$ là xác suất của sự kiện $\{p \in \Omega \mid a < X(p) < b\}$. Điều kiện các xác suất trên hoàn toàn xác định luôn thoả mãn nếu không gian mẫu là hữu hạn. Trong trường hợp Ω là vô hạn thì hai điều kiện này có nghĩa là các tập con $\{X = a\}$ và $\{a < X < b\}$ nằm trong sigma-đại số các tập hợp có xác suất.

Ví dụ 1.18. Lấy ngẫu nhiên hai lần liên tiếp, mỗi lần một quả bóng từ một hộp có 5 quả bóng màu trắng và 3 quả bóng màu vàng và quan sát màu của các quả bóng lấy được. Gọi X là biến ngẫu nhiên cho bởi số lượng bóng màu trắng lấy được. Xác định không gian mẫu, tập các giá trị của X, và các xác suất $P(X = a)$.

Chứng minh. Viết mỗi kết quả của phép thử là TT, TV, VT, và VV, tương ứng với các kết quả hai bóng trắng, bóng thứ nhất màu trắng và bóng thứ hai màu vàng, v.v. Không gian mẫu lúc này là $\Omega = \{TT, TV, VT, VV\}$. Khi đó, hàm số

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

xác định bởi bảng sau:

Mẫu	X
TT	2
TV	1
VT	1
VV	0

Lưu ý rằng, do số bóng màu trắng và màu vàng là khác nhau, nên xác suất lấy được bóng màu trắng và vàng khác nhau. Do đó, các điểm mẫu trong Ω không phải là đồng khả năng.

Để tính được xác suất, chẳng hạn,

$$P(X=2)=P(\{TT\}),$$

chúng ta cần xây dựng một không gian mẫu mà các kết quả là đồng khả năng, dựa trên giả thiết mỗi quả bóng đều có khả năng được lấy như nhau.

Đánh số các quả bóng từ 1 đến 8 sao cho các bóng màu trắng có số thứ tự 1, 2, và 3, trong khi đó các bóng màu vàng nhận các thứ tự còn lại. Không gian mẫu bây giờ là tập tất cả các cặp 2 quả bóng có thứ tự. Rõ ràng

$$|\Omega'|=A_8^2=56.$$

Bây giờ, chúng ta có thể tìm xác suất của các sự kiện TT: "hai bóng lấy ra màu trắng", TV: "bóng thứ nhất màu trắng và thứ hai màu vàng", VT: "bóng thứ nhất màu vàng và bóng thứ hai màu trắng", và VV: "cả hai bóng màu vàng".

Ví dụ, số các điểm mẫu trong sự kiện TT là $5\times 4=20$, do có 5 cách chọn bóng thứ nhất màu trắng và 4 cách chọn bóng thứ 2 màu trắng. Từ đó suy ra

$$P(X=2)=P(\{TT\})=\frac{20}{56}=\frac{5}{14},$$

Tương tự, chúng ta có thể tính

$$\begin{aligned}P(0<X\le 1)&=P(\{TV,VT\})\\&=P(\{TV\})+P(\{VT\})\\&=\frac{15}{56}+\frac{15}{56}=\frac{15}{28}.\end{aligned}$$

Các xác suất khác được xác định tương tự; chi tiết dành cho bạn đọc. □

Nhận xét 1.7. Trong ví dụ trên, để tính các xác suất chúng ta cũng có thể xét không gian mẫu gồm các cặp không sắp thứ tự của hai trong số 8 quả bóng. Khi ấy, số điểm mẫu là $C_8^2=28$, khác với cách giải trước (sự khác biệt này là do chúng ta đã ghi các kết quả của thí nghiệm theo hai cách khác nhau). Với cách này, chúng ta cũng tính được, chẳng hạn, xác suất để lấy được hai quả màu trắng là

$$P(X=2)=\frac{C_5^2}{C_8^2}=\frac{10}{28}=\frac{5}{14},$$

đúng như trên đã tính. Tương tự, xác suất để lấy được một trắng một vàng, không kể thứ tự) là

$$P(X=1)=\frac{C_5^1\times C_3^1}{C_8^2}=\frac{15}{28},$$

cũng đúng như kết quả đã tính. Cuối cùng,

$$P(X=0)=P(\{VV\})=\frac{C_3^2}{C_8^2}=\frac{3}{28}.$$

Lưu ý, X trong trường hợp này có đúng 3 giá trị và chúng ta có

$$P(X=0)+P(X=1)+P(X=2)=\frac{5}{14}+\frac{15}{28}+\frac{3}{28}=1.$$

1.6.2 Biến ngẫu nhiên rời rạc và liên tục

Các biến ngẫu nhiên xuất hiện trong thực tế thường rơi vào một trong hai loại: *rời rạc và liên tục.*

Biến ngẫu nhiên rời rạc

Một BNN được gọi là rời rạc nếu tập tất cả các giá trị có thể có của nó là một tập hữu hạn hoặc đếm được[1]. Như vậy, nếu $X:\Omega\to\mathbb{R}$ là một BNN rời rạc thì có thể viết

$$X(\Omega)=\{x_1,x_2,\ldots,x_n,\ldots\}.$$

Như vậy, nếu BNN X là rời rạc thì $P(X=x_i)>0$ với mọi $x_i\in X(\Omega)$.

Biến ngẫu nhiên rời rạc là một công cụ quan trọng trong lý thuyết xác suất để mô hình hóa các hiện tượng có tập giá trị rời rạc, chẳng hạn như số lần xảy ra một sự kiện trong các thí nghiệm.

Biến ngẫu nhiên liên tục

Một BNN liên tục X xác định trên không gian mẫu với vô hạn điểm mẫu trong đó các xác suất $P(X=x)=0$, tức là xác suất để X nhận bất kì một giá trị cụ thể nào đó đều bằng 0. Các BNN liên tục thường được dùng để mô hình hóa các đại lượng có thể thay đổi liên tục, chẳng hạn như thời gian, khoảng cách, nhiệt độ, v.v., của các không gian mẫu vô hạn, hoặc dùng để mô hình hoá một cách gần đúng các mô hình với một số "lớn" không xác định các điểm mẫu.

Trong thực tế các đại lượng liên tục như vậy thường chỉ được xác định (thông qua các phép đo đạc hoặc tính toán) một cách gần đúng. Do đó, đối với những BNN liên tục, chúng ta sẽ quan tâm đến các xác suất để giá trị của X rơi vào một khoảng nào đó, tức là các xác suất dạng $P(a<X<b)$, trong khi đó, xác suất để nó nhận một giá trị cụ thể nào đó đều bằng 0.

[1]Một tập B có vô hạn phần tử gọi là đếm được nếu nó có thể viết dưới dạng một dãy $B=\{x_1,x_2,\ldots,x_n,\ldots\}$. Ví dụ về các tập đếm được bao gồm tập các số tự nhiên, tập số nguyên, hay tập các số hữu tỉ. Tập các số vô tỉ là một ví dụ về các tập không đếm được.

Mặt khác, lưu ý rằng do $P(X=a)=P(X=b)=0$, nên

$$P(a<X<b)=P(a<X\le b)=P(a\le X<b)=P(a\le X\le b).$$

Điều này có nghĩa là việc có xét các giá trị tại hai đầu mút hay không, trong trường hợp X là BNN liên tục, là không quan trọng.

Bài tập §1.6

1. Biến ngẫu nhiên là gì? Sự khác biệt giữa biến ngẫu nhiên rời rạc và liên tục là gì?
2. Lấy ngẫu nhiên liên tiếp (không trả lại) hai quả bóng trong một thùng gồm 4 quả bóng đỏ và 3 quả bóng trắng. Gọi Y là số bóng đỏ lấy được. Xác định các kết quả có thể có và giá trị tương ứng của Y.

1.7 Phân bố của biến ngẫu nhiên

Trong mục này, chúng ta nói về hàm xác suất của BNN rời rạc, hàm mật độ xác suất của BNN liên tục, hàm phân phối tích luỹ của các BNN rời rạc và liên tục, v.v.

1.7.1 Hàm xác suất, hàm mật độ, hàm phân phối xác suất tích lũy

Định nghĩa 1.14: (Hàm xác suất)

Một hàm số $f: X(\Omega)\to\mathbb{R}$ được gọi là hàm xác suất (probability function) của một biến ngẫu nhiên rời rạc X nếu các điều kiện sau được thỏa mãn:

(i) $\sum\limits_x f(x)=1$,

(ii) $P(X=x)=f(x)$.

Hiển nhiên, hàm xác suất của một BNN rời rạc là duy nhất.

Theo định nghĩa, $P(X=x)>0$ nếu $x\in X(\Omega)$. Do đó, nếu $f(x)$ là hàm xác suất của một BNN rời rạc X thì $f(x)>0$ với mọi $x\in X(\Omega)$.

Ví dụ 1.19. Gọi X là bình phương của số chấm xuất hiện trong thử nghiệm gieo xúc xắc một lần. Khi đó, tập các giá trị của X là

$$X(\Omega)=\{1,4,9,16,25,36\},$$

Với giả thiết xúc xắc là cân xứng, hàm xác suất của X là

$$\{(1,1/6),(4,1/6),(9,1/6),(16,1/6),(25,1/6),(36,1/6)\}.$$

Chú ý 1.2. Hàm xác suất của BNN rời rạc nhận hữu hạn giá trị có thể được cho dưới dạng bảng mà ở đó, cột thứ nhất ghi các giá trị của x_i thuộc tập giá trị $X(\Omega)$ còn cột thứ hai ghi các giá trị $f(x_i) = P(X = x_i)$. Như vậy, trong Ví dụ 1.19, hàm xác suất được viết như sau:

i	x_i	$f(x_i) = P(X = x_i)$
1	1	1/6
2	4	1/6
3	9	1/6
4	16	1/6
5	25	1/6
6	36	1/6

Định nghĩa 1.15:

Giả sử X là một BNN với hàm xác suất $f(x)$. *Hàm phân phối tích luỹ* của X là

$$F(x) = \sum_{t \le x} f(t). \tag{1.23}$$

Ví dụ 1.20. Trong một kì thi, điểm của 10 sinh viên lần lượt là

$$6, 6, 7, 7, 7, 7, 8, 8, 9, 10.$$

Chọn ngẫu nhiên một sinh viên trong 10 sinh viên nói trên và gọi X là điểm số của sinh viên đó.

(a) Tìm tập giá trị của X.

(b) Tìm hàm xác suất của X.

(c) Tìm xác suất của các sự kiện A_k: *"Điểm số của sinh viên được chọn bé hơn hoặc bằng k"*, với $k = 5, 6, 7, 8, 9$.

Lời giải

(a) Không gian mẫu Ω là tập hợp gồm 10 sinh viên trong lớp, trong khi đó số điểm $X : \Omega \to \mathbb{R}$ là một BNN nhận các giá trị $6, 7, 8, 9, 10$:

$$X(\Omega) = \{6, 7, 8, 9, 10\}.$$

(b) Hàm xác suất là hàm số f chỉ nhận giá trị khác không tại các điểm thuộc tập giá trị của X. Tại các điểm trong tập giá trị này, $f(x)$ cho bởi bảng sau:

i	x_i	$f(x_i)=P(X=x_i)$
2	6	0,2
3	7	0,4
4	8	0,2
5	9	0,1
6	10	0,1

(c) Từ giả thiết, $A_k=\{p\in\Omega \mid X(p)\le k\}$. Do đó, $P(A_k)=P(X\le k)=F(k)$, trong đó $F(x)$ là hàm phân phối tích luỹ của X, cho bởi

$$F(x)=\begin{cases}0 & \text{nếu } x<6\\ 0{,}2 & \text{nếu } 6\le x<7\\ 0{,}6 & \text{nếu } 7\le x<8\\ 0{,}8 & \text{nếu } 8\le x<9\\ 0{,}9 & \text{nếu } 9\le x<10\\ 1{,}0 & \text{nếu } x\ge 10\end{cases}$$

Lưu ý, các hàm phân phối tích lũy của các BNN rời rạc là các hàm hằng số trên từng đoạn.

Đối với các BNN liên tục, chúng ta sẽ cần đến hàm mật độ xác suất. Các hàm mật độ xác suất này sẽ cho chúng ta các xác suất khi lấy tích phân của chúng trên khoảng cần tính.

Định nghĩa 1.16:

Một hàm số $f(x)\ge 0$ được gọi là *hàm mật độ xác suất* của một BNN X nếu

$$P(a<X<b)=\int_a^b f(x)\,\mathrm{d}x$$

xảy ra với mọi $a<b$.

Rõ ràng một điều kiện cần để $f(x)$ là một hàm mật độ của một BNN nào đó là

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x=P(-\infty<X<\infty)=P(\Omega)=1, \tag{1.24}$$

ở đó, vế trái là một *tích phân suy rộng* [1](improper integral) của một hàm không âm ($f(x)\ge 0$).

[1]Tích phân suy rộng $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$ được định nghĩa là giới hạn của các tích phân $\int_A^B f(x)\,\mathrm{d}x$ khi cho đồng thời $A\to-\infty$ và $B\to+\infty$ độc lập nhau.

Ví dụ 1.21. Giả sử

$$f(x)=\begin{cases}\alpha x^2, & 0\le x\le 3,\\ 0 & x\in(-\infty)\cup(3,+\infty),\end{cases}$$

là hàm mật độ xác suất của một BNN liên tục X.

(a) Hãy xác định tham số α.

(b) Tìm $P(0<X<1)$.

Lời giải

(a) Để tìm α, chúng ta dựa vào điều kiện (1.24). Cụ thể như sau:

$$1=\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x=\int_0^3 \alpha x^2\,\mathrm{d}x=\frac{\alpha x^3}{3}\Big|_0^3=9\alpha.$$

Từ đó suy ra $\alpha=1/9$.

(b) Với $\alpha=1/9$, ta có

$$P(0<X<1)=\int_0^1 \alpha x^2\,\mathrm{d}x=\frac{\alpha x^3}{3}\Big|_0^1=\frac{\alpha}{3}=\frac{1}{27}.$$

Định nghĩa 1.17:

Hàm phân phối xác suất tích luỹ của một BNN liên tục X là

$$F(x)=P(X\le x). \tag{1.25}$$

Như vậy, đối với một BNN liên tục, mối liên hệ giữa hàm mật độ $f(x)$ và hàm phân phối tích luỹ $F(x)$ là

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

và do đó $F'(x)=f(x)$. Nói cách khác, $F(x)$ khả vi và nó là một nguyên hàm của $f(x)$.

Chú ý 1.3. Lưu ý, trong nhiều sách, người ta dùng kí hiệu $f(x)$ cho hàm xác suất (probability mass function) của BNN rời rạc, và cũng dùng kí hiệu $f(x)$ cho hàm mật độ xác suất (probability density function). Cách kí hiệu này dễ gây ra sự nhầm lẫn giữa hai khái niệm xác suất và mật độ xác suất.

1.7.2 Phân phối đồng thời và các biến ngẫu nhiên độc lập

Trong nhiều tình huống thực tế, chúng ta cần ghi chép các kết quả của phép thử ngẫu nhiên như các véctơ biểu thị nhiều đại lượng (chẳng hạn, chiều cao và cân nặng của các cá thể). Lúc đó, chúng ta quan tâm đến *phân phối xác suất đồng thời* (joint probability distribution) của hai hoặc nhiều hơn các biến ngẫu nhiên.

Định nghĩa 1.18:

Giả sử X và Y là hai BNN rời rạc. Hàm xác suất đồng thời của X và Y là hàm số của hai biến x và y cho bởi

$$f(x,y) = P(X = x, Y = y). \tag{1.26}$$

Như vậy, $f(x,y)$ là xác suất để xảy ra đồng thời X nhận giá trị x và Y nhận giá trị y.

Ví dụ 1.22. Tung một đồng xu liên tiếp 3 lần. Gọi X là số lần mặt sấp (S) xuất hiện ở lần tung thứ nhất và Y là số lần mặt sấp xuất hiện trong hai lần tung tiếp theo. Không gian mẫu gồm các kết quả có thể thu được là

$$\Omega = \{NNN, SNN, NSN, NNS, SSN, SNS, NSS, SSS\},$$

với $|\Omega| = 8$ và các kết quả đồng khả năng.

Giá trị của các BNN X và Y cho bởi bảng sau:

Điểm mẫu	X	Y
NNN	0	0
SNN	1	0
NSN	0	1
NNS	0	1
SSN	1	1
SNS	1	1
NSS	0	2
SSS	1	2

Khi đó, hàm phân phối xác suất đồng thời là

$$\begin{aligned}
f(0,0) &= P(NNN) = 0{,}125,\\
f(0,1) &= P(\{NSN, NNS\}) = 0{,}25,\\
f(0,2) &= P(NSS) = 0{,}125,\\
f(1,0) &= P(SNN) = 0{,}125,\\
f(1,1) &= P(\{SSN, SNS\}) = 0{,}25,\\
f(1,2) &= P(SSS) = 0{,}125,
\end{aligned}$$

trong khi đó $f(x,y) = 0$ trong mọi trường hợp khác.

Lưu ý, nếu X nhận một số hữu hạn các giá trị $x_i, i = 1,2,\ldots,n$ và Y nhận một số hữu hạn các giá trị $y_j,\ j = 1,2,\ldots,m$ thì phân phối đồng thời của X và Y hoàn toàn được xác định bởi ma trận p_{ij} với

$$p_{ij} = f(x_i, y_j) = P(X = x_i, Y = y_j).$$

Như trong ví dụ trên, X nhận 2 giá trị 0 và 1, trong khi đó Y nhận 3 giá trị 0, 1 và 2. Ma trận p_{ij} sẽ là ma trận cỡ 2×3

$$P = \begin{bmatrix} 0{,}125 & 0{,}25 & 0{,}125 \\ 0{,}125 & 0{,}25 & 0{,}125 \end{bmatrix}.$$

Định nghĩa 1.19:

Giả sử X và Y là các BNN liên tục. Hàm hai biến $f(x,y)$ được gọi là hàm *hàm mật độ xác suất đồng thời* của X và Y nếu

(i) $f(x,y) \geq 0$, với mọi (x,y),

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$,

(iii) Với mọi sự kiện A,

$$P((X,Y) \in A) = \iint_A f\,\mathrm{d}x\,\mathrm{d}y.$$

Từ hàm xác suất đồng thời của các BNN, chúng ta có thể đưa ra khái niệm các BNN độc lập (independent random variables), chi tiết như trong định nghĩa sau:

Định nghĩa 1.20:

ai BNN X và Y với hàm xác suất tương ứng là $g(x)$ và $h(y)$ được gọi là *độc lập* nếu hàm xác suất đồng thời $f(x,y)$ của chúng thoả mãn

$$f(x,y) = g(x)h(y), \tag{1.27}$$

với mọi x, y.

Bạn đọc có thể kiểm tra các BNN trong Ví dụ 1.22 là các BNN độc lập.

Nhận xét 1.8. Mối quan hệ giữa khái niệm độc lập của BNN và của các sự kiện như sau: Giả sử X và Y là hai BNN liên tục độc lập trên cùng một không gian mẫu. Nếu I và J là hai khoảng trong tập số thực và A và B tương ứng là hai sự kiện $X \in I$, $Y \in J$. Khi đó, $A \cap B = \{(X,Y) \in I \times J\}$. Theo định lí Fubini (xem [7, §VII.3]),

$$\begin{aligned} P(A \cap B) &= \iint_{I\times J} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint_{I\times J} g(x)h(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\int_I g(x)\,\mathrm{d}x\right)\left(\int_J h(y)\,\mathrm{d}y\right) \\ &= P(A)P(B). \end{aligned} \tag{1.28}$$

Khi đó, A và B là hai sự kiện độc lập.

Bài tập §1.7

1. Trong kho có 6 bộ ti vi, trong đó có 2 bộ bị hỏng. Lấy ngẫu nhiên 3 bộ để gửi cho khách hàng. Gọi X là số bộ ti vi hỏng trong lô hàng gửi đi. Tìm phân bố xác suất của X.
2. Một BNN liên tục X nhận giá trị từ $x = 0$ đến $x = 2$ và có hàm mật độ là $f(x) = 1/2$ với $x \in [0,2]$. Tìm $P(1 < X < 1{,}5)$ và $P(X \leq 1{,}7)$.

1.8 Các số đặc trưng của biến ngẫu nhiên

Nội dung chính của phần này bao gồm: trung bình (mean), phương sai (variance), và độ lệch chuẩn (standard deviation) của các biến BNN. Phần cuối, chúng ta cũng nói về trung vị (median) và mode của các BNN.

1.8.1 Trung bình

Trung bình μ của một biến ngẫu nhiên X là tương tự với trung bình $\bar{x}$ của tần suất. Trung bình μ đặc trưng cho vị trí trung tâm của phân phối, còn được gọi là **kì vọng toán** (mathematical expectation), và còn được kí hiệu $\mathbb{E}[X]$, bởi vì nó cho giá trị trung bình của X được mong đợi khi thực hiện nhiều phép thử.

Định nghĩa 1.21: Trung bình/Kì vọng

Cho X là một BNN rời rạc hoặc liên tục. Trung bình μ_X hay kì vọng $\mathbb{E}[X]$ của X được định nghĩa như sau:

$$\mu_X = \mathbb{E}[X] = \sum_x xP(X=x) \qquad \text{(phân phối rời rạc)} \qquad (1.29)$$

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x \qquad \text{(phân phối liên tục),} \qquad (1.30)$$

trong đó, nếu X là BNN liên tục thì $f(x)$ là hàm mật độ của X.

Chú ý 1.4. Lưu ý, đối với một BNN rời rạc hay liên tục X, để kì vọng $\mathbb{E}[X]$ tồn tại thì các vế phải của (1.29) hoặc (1.30) phải tồn tại. Điều này tương đương với tính hội tụ không điều kiện của tổng trong vế phải của (1.29) hoặc tính hội tụ của tích phân suy rộng trong vế phải (1.30).

Ví dụ 1.23. Cho thực nghiệm tung một đồng xu đồng nhất cân xứng hai lần liên tiếp. Gọi X là số mặt sấp (S) xuất hiện. Tính trung bình của biến ngẫu nhiên này.

Lời giải

Nhận thấy X là BNN rời rạc xác định trên một không gian mẫu có 4 kết quả: $\Omega = \{SS, SN, NS, NN\}$. Nó có hàm xác suất cho bởi bảng sau đây:

x	**Sự kiện** $\{X=x\}$	$f(x)=P(X=x)$
0	$\{NN\}$	0,25
1	$\{SN, NS\}$	0,50
2	$\{SS\}$	0,25

Từ đó, trung bình (hay kì vọng) của X là

$$\mu_X = \mathbb{E}[X] = \sum_x xf(x) = 0\times 0{,}25 + 1\times 0{,}50 + 2\times 0{,}25 = 1.$$

Ví dụ 1.24. Tính trung bình của BNN có hàm mật độ $f(x)$ trong Ví dụ 1.21.

Lời giải

Hàm mật độ có dạng $f(x) = x^2/9$ với $0 \le x \le 3$ và bằng 0 bên ngoài $(0,3)$. Do đó, trung bình của X là

$$\mu_X = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_0^3 \frac{x^3}{9}\,\mathrm{d}x = \frac{x^4}{36}\Big|_0^3 = \frac{9}{4}.$$

Chú ý 1.5. Hai BNN có cùng phân phối thì có cùng trung bình. Do đó, chúng ta cũng coi trung bình của một BNN là trung bình của phân phối của nó.

1.8.2 Phương sai và độ lệch chuẩn

Phương sai (variance) σ^2 đặc trưng cho sự phân tán của BNN cũng như của phân phối xác suất của nó. Phương sai của X được định nghĩa như sau:

$$\sigma_X^2 = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]. \tag{1.31}$$

Từ công thức của kì vọng, ta có

$$\sigma^2 = \sum_x (x-\mu)^2 p(X=x) \qquad \text{(phân phối rời rạc)} \tag{1.32}$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \qquad \text{(phân phối liên tục).} \tag{1.33}$$

Lưu ý, vì tổng hoặc tích phân suy rộng trong công thức phương sai là không âm, nên nếu X có trung bình hữu hạn, thì phương sai của X luôn tồn tại, hữu hạn hoặc bằng $+\infty$.

Ví dụ 1.25. Tính phương sai của BNN X trong Ví dụ 1.23.

Lời giải

Từ tính toán trong ví dụ trước, ta có X nhận các giá trị $0,1,2$ với trung bình $\mu_X = 1$. Do đó, $Y = (X-\mu)^2$ nhận các giá trị 1 và 0 với các xác suất tương ứng cho bởi bảng sau:

y	**Sự kiện** $\{Y=y\}$	$g(y) = P(Y=y)$
0	$\{SN, NN\}$	0,5
1	$\{SS, NN\}$	0,5

Từ đó, phương sai của X là

$$\sigma_X^2 = E(X-\mu_X)^2 = E(Y) = \sum_y y\,P(Y=y) = 0\times 0{,}5 + 1\times 0{,}5 = 0{,}5.$$

Ví dụ 1.26. Tính phương sai của BNN có hàm mật độ $f(x)$ trong Ví dụ 1.21.

Lời giải

Hàm mật độ có dạng $f(x) = x^2/9$ với $0 \le x \le 3$ và bằng 0 bên ngoài $(0,3)$. Trung bình của X đã tính được trong Ví dụ 1.24 là

$$\mu_X = \frac{9}{4}.$$

Từ đó, chúng ta tính phương sai như sau:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)\,\mathrm{d}x = \int_0^3 \left(x-\frac{9}{4}\right)^2 \frac{x^2}{9}\,\mathrm{d}x = \frac{27}{80}.$$

Định lí 1.3:

Nếu X và X^2 có trung bình hữu hạn thì phương sai σ_X^2 của một BNN X có thể tính được bằng công thức

$$\sigma_X^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2. \tag{1.34}$$

Chứng minh. Ta sẽ chứng minh kết quả này cho trường hợp BNN rời rạc. Trường hợp BNN liên tục dành cho bạn đọc.

Theo định nghĩa,

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 f(x)\\ &= \sum_x (x^2 - 2\mu x + \mu^2) f(x)\\ &= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x).\end{aligned}$$

Thay $\sum_x x^2 f(x) = \mathbb{E}\left[X^2\right]$, $\mu = \sum_x x f(x)$, và $\sum_x f(x) = 1$ vào biểu thức cuối cùng, chúng ta dễ dàng thu được

$$\sigma^2 = \mathbb{E}\left[X^2\right] - 2\mu^2 + \mu^2 = \mathbb{E}\left[X^2\right] - \mu^2.$$

Chứng minh trong trường hợp BNN X là liên tục dành cho bạn đọc. □

Định nghĩa 1.22:

Giá trị $\sigma = \sqrt{\sigma^2}$ gọi là *độ lệch chuẩn* của BNN X.

Nếu X là một BNN và g là một hàm số, thì $g(X)$ cũng là một BNN. Mối quan hệ giữa hàm phân phối của X và của $g(X)$ hoặc mỗi liên hệ giữa trung bình và phương sai của X và của $g(X)$ thường không đơn giản. Tuy nhiên, trong trường hợp g là hàm bậc nhất thì mối liên hệ giữa trung bình và phương sai của X và của $g(X)$ như sau:

Định lí 1.4:

Cho X là một BNN với trung bình μ và độ lệch chuẩn $\sigma > 0$. Gọi $Y = c_1 + c_2 X$. Khi đó,

$$\mu_Y = \mathbb{E}[Y] = c_1 + c_2\mu, \quad \sigma_Y = c_2\sigma. \tag{1.35}$$

Nói riêng, nếu

$$Z = \frac{X-\mu}{\sigma}$$

thì

$$\mu_Z = 0, \quad \sigma_Z = 1.$$

Chú ý 1.6. Đối với một BNN X,

$$Z = \frac{X-\mu_X}{\sigma_X}$$

được gọi là BNN chuẩn hoá (standardized random variable) của X.

Ví dụ 1.27 (Phân phối đều rời rạc)**.** Một BNN rời rạc X có phân phối đều nếu X nhận hữu hạn giá trị $\{x_1, x_2, \ldots, x_n\}$ với xác suất như nhau[1]

$$P(X = x_i) = \frac{1}{n}, \quad i = 1, 2, \ldots, n.$$

Ví dụ, cho X là một BNN rời rạc nhận các giá trị $1, 2, \ldots, n$ và có phân phối đều. Hãy tính trung bình và phương sai của X.

Bắt đầu với kì vọng, ta có

$$\begin{aligned}\mu_X = \mathbb{E}[X] &= \sum_{i=1}^{n} i\,P(X=i) \\ &= \sum_{i=1}^{n} \frac{i}{n} = \frac{n+1}{2}.\end{aligned}$$

Như vậy, trung bình μ_X chính là trung bình số học của n giá trị của X.

[1] Một số sách yêu cầu các giá trị x_i của phân phối đều rời rạc có dạng $x_i = a + i - 1$ với $a \in \mathbb{R}$.

Để tính phương sai của X, hãy bắt đầu với BNN X^2 nhận các giá trị $1,4,9,\ldots,n^2$. Do đó,

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \sum_{i=1}^{n} i^2 P(X^2 = i^2) \\ &= \sum_{i=1}^{n} i^2 P(X = i) \\ &= \sum_{i=1}^{n} \frac{i^2}{n} \\ &= \frac{(n+1)(2n+1)}{6}.\end{aligned}$$

Từ đó, phương sai của X là

$$\begin{aligned}\sigma_X^2 = \mathrm{Var}(X) &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \\ &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n^2-1}{12}.\end{aligned}$$

Ví dụ 1.28 (Phân phối đều liên tục)**.** Chúng ta nói một BNN liên tục X có phân phối đều nếu X có hàm mật độ xác suất có dạng

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

Khi đó, chúng ta viết $X \sim U(a,b)$. Dễ dàng tính được

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b x f(x)\,\mathrm{d}x \\ &= \int_a^b \frac{x}{b-a}\,\mathrm{d}x \\ &= \left.\frac{x^2}{2(b-a)}\right|_a^b \\ &= \frac{b+a}{2},\end{aligned}$$

trong khi đó,

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \int_a^b x^2 f(x)\,\mathrm{d}x \\ &= \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x \\ &= \frac{x^3}{3(b-a)}\Big|_a^b \\ &= \frac{a^2+ab+b^2}{3}.\end{aligned}$$

Từ đó, phương sai của X là

$$\begin{aligned}\sigma_X^2 = D(X) &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \\ &= \frac{a^2+ab+b^2}{3} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{(a-b)^2}{12}.\end{aligned}$$

1.8.3 Trung vị và mode

Hai đặc trưng quan trọng khác của các BNN là mode (mốt) và trung vị (median).

Mốt của một BNN

Đối với một BNN rời rạc X, mode của nó là một hoặc nhiều giá trị x_i của X mà có xác suất $P(X = x_i)$ lớn nhất.

Nếu có nhiều giá trị có cùng xác suất lớn nhất, biến ngẫu nhiên sẽ có nhiều mốt và được gọi là *đa mốt* (multimodal). Một ví dụ đơn giản của BNN rời rạc có nhiều mode là các BNN có phân phối đều.

Mode của một BNN liên tục X là giá trị của biến x mà tại đó hàm mật độ xác suất $f(x)$ đạt giá trị cực đại. Nói cách khác, mode là điểm có mật độ xác suất cao nhất trên đường cong của hàm mật độ.

Nếu hàm mật độ có nhiều cực đại, thì X có nhiều mốt và được gọi là *đa mốt* (multimodal). Ngược lại, nếu hàm mật độ có một cực đại duy nhất, biến ngẫu nhiên đó là *đơn mốt* (unimodal).

Trung vị

Trung vị (median) của một BNN liên tục X là giá trị m sao cho diện tích dưới đường cong của hàm mật độ xác suất từ $-\infty$ đến m bằng $1/2$, tức là:

$$\int_{-\infty}^{m} f(x)\,\mathrm{d}x = \frac{1}{2}, \tag{1.36}$$

trong đó $f(x)$ là hàm mật độ xác suất của X. Như vậy, $P(X \le m) = \frac{1}{2}$.

Ví dụ 1.29. Xét một BNN X với hàm mật độ cho bởi

$$f(x) = \begin{cases} 2x & \text{nếu } x \in [0;\ 1], \\ 0 & \text{trong các trường hợp còn lại.} \end{cases}$$

Hãy tìm mode, trung vị, và trung bình của X.

Lời giải

Rõ ràng, mode của X là giá trị $x = 1$. Mặt khác, nếu m là trung vị thì

$$\int_{0}^{m} 2x\,\mathrm{d}x = 0{,}5.$$

Từ đó, tính tích phân trong vế trái của điều kiện trên, ta có

$$m^2 = 0{,}5,$$

và do đó, trung vị của X là $m = \sqrt{0{,}5}$.

Cuối cùng, trung bình của X là

$$\mu_X = \int_0^1 x f(x)\,\mathrm{d}x = \int_0^1 2x^2\,\mathrm{d}x = \frac{2}{3}.$$

Nhận xét 1.9. Thông thường, các BNN liên tục chỉ có 1 trung vị. Thật vậy, nếu có hai giá trị trung vị, hãy gọi là m_1 và m_2 (với $m_1 < m_2$), thì bắt buộc $f(x) = 0$ trên toàn bộ đoạn $[m_1, m_2]$. Nhưng những trường hợp như vậy hiếm xảy ra trong các ứng dụng của xác suất thống kê.

Đối với BNN rời rạc X thì tồn tại nhiều nhất hai giá trị m thuộc miền giá trị của X sao cho

$$P(X \le m) \ge \frac{1}{2}, \quad P(X \ge m) \ge \frac{1}{2}. \tag{1.37}$$

Trung vị của X được định nghĩa là trung bình cộng của các giá trị như vậy.

Trung vị là một chỉ số thường được sử dụng để đo lường xu hướng trung tâm của dữ liệu, đặc biệt khi dữ liệu có những giá trị ngoại lai (outliers) hoặc phân phối không đối xứng, vì trung vị ít bị ảnh hưởng bởi các giá trị cực đoan hơn so với trung bình.

1.8.4 Bất đẳng thức Markov và Chebyshev

Bất đẳng thức Markov[1] nói rằng đối với một biến ngẫu nhiên không âm X và với mọi $a > 0$, ta có:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{1.38}$$

Chú ý 1.7. Bất đẳng thức Markov chỉ hữu ích khi $a > \mathbb{E}[X]$.

Ví dụ 1.30. Khối lượng mỗi trái bưởi trong một lô hàng có trung bình $\mu = 560$ gram. Xác suất để một trái bưởi được chọn ngẫu nhiên có khối lượng lớn hơn 800 gram nhiều nhất là bao nhiêu?

Lời giải

Gọi X là khối lượng bưởi trong lô hàng đó. Theo giả thiết, $E(X) = 560$. Bất đẳng thức Markov cho chúng ta ước lượng

$$P(X \geq 800) \leq \frac{\mathbb{E}[X]}{800} = \frac{560}{800} = 0{,}7.$$

Vậy, xác suất để một trái bưởi được chọn ngẫu nghiên có khối lượng ≥ 800 gram nhiều nhất là 70%.

Chứng minh bất đẳng thức Markov trong trường hợp X có một hàm mật độ $f(x)$ như sau: Do X không âm

$$E(X) = \int_0^\infty x f(x)\,\mathrm{d}x \geq \int_a^\infty x f(x)\,\mathrm{d}x \geq a \int_a^\infty f(x)\,\mathrm{d}x = aP(X \geq a).$$

Chia hai vế cho a chúng ta thu được (1.38).

Chứng minh BĐT Markov cho trường hợp BNN rời rạc cũng khá đơn giản. Giả sử X là BNN rời rạc với hàm xác suất $f(x)$ xác định trên tập giá trị $X(\Omega) = \{x_1, x_2, \ldots\}$, với $x_i > 0$. Khi đó,

$$E(X) = \sum_i x_i f(x_i) \geq \sum_{i|x_i \geq a} x_i f(x_i) \geq a \sum_{x_i \geq a} f(x_i) = aP(X \geq a).$$

Từ đó chúng ta dễ dàng thu được BĐT (1.38).

Nhận xét 1.10. Bất đẳng thức Markov cũng đúng cho các BNN không âm tuỳ ý (bao gồm các BNN chứa phần "kì dị").

Một ứng dụng đơn giản của bất đẳng thức Markov là bất đẳng thức Chebyshev[2], là một kết quả quan trọng trong lí thuyết xác suất, cung cấp

[1] Andrey Andreyevich Markov (1856–922) là một nhà toán học người Nga.
[2] Pafnuty Lvovich Chebyshev (1821–1894) là một nhà toán học người Nga.

một giới hạn trên cho xác suất mà một biến ngẫu nhiên lệch khỏi giá trị kì vọng của nó.

Định lí 1.5: (Chebyshev)

Giả sử X là một biến ngẫu nhiên với kì vọng μ và phương sai σ^2. Khi đó, với mọi $k>1$:

$$P(|X-\mu| \le k\sigma) \ge 1-\frac{1}{k^2}. \tag{1.39}$$

Nói cách khác, xác suất mà giá trị của biến ngẫu nhiên X nhân giá trị sai khác với giá trị kì vọng một lượng bé hơn k lần độ lệch chuẩn ít nhất bằng $1-1/k^2$.

Chứng minh. Dễ thấy

$$P(|X-\mu| \ge k\sigma) = P((X-\mu)^2 \ge (k\sigma)^2).$$

Mặt khác, áp dụng bất đẳng thức Markov với $Y=(X-\mu)^2$ và $a=(k\sigma)^2$, chúng ta có:

$$\begin{aligned} P((X-\mu)^2 \ge (k\sigma)^2) &\le \frac{E((X-\mu)^2)}{(k\sigma)^2)} \\ &= \frac{\sigma^2}{(k\sigma)^2} \\ &= \frac{1}{k^2}, \end{aligned}$$

do $E((X-\mu)^2) = \mathrm{Var}(X) = \sigma^2$. Từ đó dễ dàng suy ra bất đẳng thức cần chứng minh. □

Ví dụ 1.31. Khối lượng mỗi trái cam trong một lô hàng là một BNN X có trung bình $\mu = 180$ g và phương sai $\sigma^2 = 12$ g^2. Xác suất để một trái cam được chọn ngẫu nhiên có khối lượng rơi vào khoảng 150 gram đến 210 gram ít nhất bằng bao nhiêu?

Lời giải

Áp dụng BĐT Chebyshev, chúng ta có

$$P(150 \le X \le 210) = P(|X-180| \le 30) \ge 1-\frac{\sigma^2}{30^2} = 1-\frac{12}{30} = 0{,}6.$$

Xác suất để một trái cam được chọn ngẫu nhiên có khối lượng từ 150 đến 210 gram ít nhất 60%.

Tất nhiên, nếu có giả thiết khối lượng của các trái cam tuân theo phân phối chuẩn $X \sim \mathcal{N}(180,12)$ thì chúng ta sẽ tính được xác suất để một trái cam được chọn ngẫu nhiên có khối lượng trong một khoảng cho trước tuỳ ý. Ở đây, mặc dù không biết dạng của phân phối của X, nhưng BĐT Chebysev vẫn cho chúng ta một ước lượng của xác suất đó.

Ví dụ 1.32. Giả sử X là một BNN với hàm mật độ

$$f(x)=\begin{cases}3x(2-x)/4, & 0\le x\le 2,\\ 0, & \text{trong các trường hợp khác.}\end{cases}$$

Tính xác suất $P(\mu-2\sigma\le X\le\mu+2\sigma)$, trong đó, μ là trung bình và σ là độ lệch chuẩn của X. So sánh kết quả thu được với ước lượng cho bởi bất đẳng thức Chebysev.

Lời giải

Tính toán cụ thể, chúng ta thu được trung bình của X là

$$\mu=\int_0^2 x f(x)\,\mathrm{d}x=\int_0^2 \frac{3}{4}x^2(2-x)\,\mathrm{d}x=1,$$

và do đó phương sai của X là

$$\sigma^2=\int_0^2 (x-1)^2 f(x)\,\mathrm{d}x=\frac{1}{5}.$$

Vậy

$$\begin{aligned}P(\mu-2\sigma\le X\le\mu+2\sigma)&=P\left(1-\frac{2}{\sqrt{5}}\le X\le 1+\frac{2}{\sqrt{5}}\right)\\&=\int_{1-\frac{2}{\sqrt{5}}}^{1+\frac{2}{\sqrt{5}}}\frac{3}{4}x^2(2-x)\,\mathrm{d}x\\&=\frac{11}{5\sqrt{5}}\approx 0{,}98387.\end{aligned}$$

Mặt khác, bất đẳng thức Chebyshev cho ước lượng ($k=2$)

$$P(\mu-2\sigma\le X\le\mu+2\sigma)\ge 1-\frac{1}{4}=0{,}75.$$

Bạn đọc hãy tự đưa ra nhận xét cho mình về hai kết quả này.

Bài tập mục 1.8

1. Hãy tính trung bình μ_X và phương sai σ_X^2 của BNN rời rạc X nhận các giá trị $x_i = 0,1,2,3$ với các xác suất tương ứng cho bởi bảng sau:

x_i	0	1	2	3
$P(X=x_i)$	0,2	0,3	0,1	0,4

2. Xét BNN X như trong Bài tập 1. Hãy tính trung bình của X^2 và dùng giá trị μ_X và μ_{X^2} để tính $\mu_{(X+1)^2}$.

3. Một BNN rời rạc X được cho bởi $P(X=x)=k(x-1)$ với $x=2,3,4,5$. Tìm k và kì vọng $E(X)$.

4. Một BNN rời rạc X với kì vọng (trung bình) $E(X)=2,3$ cho bởi bảng sau:

x_i	0	1	2	3	α
$P(X=x_i)$	0,1	0,2	0,2	0,3	p

Tìm p và α.

5. Một hàm mật độ xác suất $f(x)$ của một BNN X có dạng

$$f(x)=\begin{cases} c & \text{nếu } 0\le x\le 2,\\ 0 & \text{nếu } x\in(-\infty,0)\cup(2,+\infty)\end{cases}$$

trong đó c là một hằng số. Tìm $P(0\le X\le 1)$.

6. Cho X là một BNN với hàm xác suất như trong Bài tập 5. Tìm c sao cho $P(X<c)=95\%$.

7. Tuổi thọ của một loại bóng đèn là một BNN liên tục X với hàm mật độ xác suất như sau:

$$f(x)=\begin{cases} c\,x^2(5-x) & \text{nếu } 0<x<5,\\ 0 & \text{trong các trường hợp còn lại.}\end{cases}$$

Tìm xác suất để bóng đèn hỏng trước khi nó được 1 năm tuổi.

8. Tìm hàm phân phối tích luỹ của BNN X trong Bài tập 7.

9. Tính trung bình và phương sai của một BNN X có hàm mật độ xác suất $f(x)$ trong các trường hợp sau:

(a) $f(x)=mx$ trong khoảng $[1,2]$ và bằng 0 bên ngoài khoảng đó (m là một hằng số phù hợp).

(b) X là số chấm có được trong một lần gieo xúc xắc cân xứng đồng chất.

(c) $f(x)=c(x-0,9)(x-1,1)$; tìm k, μ, và σ^2.

10. Giả sử X là một BNN liên tục có hàm phân phối tích luỹ là

$$F(x)=\begin{cases} 0 & x\le 0\\ 1-e^{-kx} & x>0\end{cases},$$

trong đó k là một hằng số dương. Tìm σ_X^2 theo k.

11. Chứng minh rằng hàm số

$$f(x)=\begin{cases}\frac{1}{x^2} & \text{nếu } x\geq 1\\ 0 & \text{nếu } x<1\end{cases}$$

là hàm mật độ của một BNN X. Chứng minh rằng khi ấy X có kì vọng vô hạn.

12. Giả sử X là một BNN liên tục có hàm mật độ xác suất là

$$f(x)=\begin{cases}\alpha(1-x) & \text{khi } 2\leq x\leq 4\\ 0 & \text{khi } x<2 \text{ hoặc } x>4.\end{cases}.$$

Tìm α và trung vị của X.

1.9 Một số phân phối xác suất thường gặp

Như chúng ta đã biết ở mục trước, điều kiện để một hàm số là hàm mật độ của BNN nào đó là khá nhẹ nhàng. Do đó, có thể dễ dàng xây dựng được những ví dụ về các phân bố xác suất. Tuy nhiên, trong các ứng dụng thực tiễn, chỉ có một số phân bố thường xuyên xuất hiện và do đó chúng đóng vai trò rất quan trọng trong lí thuyết xác suất và thống kê. Mục tiêu của mục này là tìm hiểu sâu về một số phân phối xác suất như vậy, bao gồm phân phối nhị thức, phân phối Poisson, phân phối chuẩn, phân phối t (phân phối Student), phân phối nhị thức âm, và một số phân phối khác. Tài liệu tham khảo: Kreyszig [1, §24.7, 24.8].

1.9.1 Dãy phép thử Bernoulii và phân phối nhị thức

Phân phối nhị thức (binomial distribution), kí hiệu $B(n,p)$ là một mô hình xác suất để mô tả thí nghiệm trong đó người ta thực hiện n phép thử độc lập, không ảnh hưởng lẫn nhau, với xác suất của xuất hiện một sự kiện A đã định là p. Ở đây, n là số nguyên dương và $0<p<1$.

Xét một phép thử T với sự kiện A có xác suất xuất hiện p. Phép thử T được thực hiện n lần liên tiếp. Gọi X là số lần xuất hiện sự kiện A trong n phép thử. Khi đó, X là một BNN rời rạc nhận các giá trị $0,1,2,\ldots,n$, với các xác suất

$$P(X=k)=C_n^k p^k(1-p)^{n-k}, \tag{1.40}$$

trong đó, C_n^k là tổ hợp chập k của n phần tử:

$$C_n^k=\frac{n!}{k!(n-k)!}. \tag{1.41}$$

Như vậy, hàm xác suất của phân phối nhị thức là

$$f(k)=P(X=k)=C_n^k p^k(1-p)^{n-k}. \tag{1.42}$$

Ví dụ 1.33. Tung một xúc xắc cân xứng 10 lần. Gọi X là số lần mặt 6 chấm xuất hiện. Khi đó, X là một BNN có phân phối Bernoulli với $n = 10$ và $p = 1/6$. Xác suất để có đúng hai lần xuất hiện mặt 6 chấm là

$$P(X=2) = C_{10}^{2}\left(\frac{1}{6}\right)^{2}\left(\frac{5}{6}\right)^{8}.$$

Ví dụ 1.34. Tung một xúc xắc đồng chất và cân xứng bốn lần. Tính xác suất để nhận được ít nhất hai mặt sáu điểm.

Lời giải

Gọi X là BNN cho số lần mặt sáu điểm xuất hiện trong 4 lần tung. Khi đó, X có phân phối nhị thức với $n = 4$ và $p = \frac{1}{6}$. Ta có hàm xác suất tương ứng là

$$f(k) = C_{4}^{k}\left(\frac{1}{6}\right)^{k}\left(\frac{5}{6}\right)^{n-k} \qquad k = 0,1,2,3,4.$$

Từ đó, chúng ta có thể tính được xác suất nhận được ít nhất hai mặt sáu điểm là

$$P(X \geq 2) = f(2) + f(3) + f(4) = \frac{171}{1296} \approx 13{,}19444\%.$$

CAS 1.6. Trong R, các giá trị của hàm xác suất của phân phối nhị thức là `dbinom()`, với các tham số lần lượt là giá trị cần tính x, cỡ (size) n, và xác suất p. Với $n = 4$ và $p = 1/6$, tổng $\sum_{i=2}^{4} f(i)$ xuất hiện trong ví dụ trên có thể được tính bởi đoạn code sau:

```
# Khai báo các tham số
n = 4; p = 1/6; result = 0

# Tính tổng f(i) với i chạy từ 2 đến 4
for (i in 2:4) {
        result = result + dbinom(i, n, p)
}

# In ra kết quả
print(result)

## [1] 0,1319
```

Mệnh đề 1.4:

Nếu X là BNN có phân phối nhị thức $B(n,p)$ thì X có kì vọng

$$\mu_X = np, \tag{1.43}$$

và phương sai

$$\sigma_X^2 = np(1-p). \tag{1.44}$$

Chứng minh. Trước hết, chúng ta nhắc lại một đẳng thức đơn giản về tổ hợp mà chúng ta cần dùng, đó là

$$kC_n^k = nC_{n-1}^{k-1}, \quad k = 1,2,\ldots,n.$$

Từ đó,

$$\begin{aligned}
\mu_X &= \sum_{k=0}^{n} kf(k) = \sum_{k=1}^{n} kC_n^k p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} nC_{n-1}^{k-1} p^k (1-p)^{n-k} \\
&= np \sum_{j=0}^{n-1} C_{n-1}^{j} p^j (1-p)^{n-1-j} \qquad \text{(đổi chỉ số } j = k-1\text{)} \\
&= np(p+(1-p))^{n-1} \\
&= np.
\end{aligned}$$

Ta thu được công thức của trung bình, như mong muốn.

Chú ý rằng trong lập luận và tính toán phía trên, chúng ta đã chứng minh được đẳng thức

$$\sum_{i=0}^{m} iC_m^i p^i (1-p)^{m-i} = mp, \tag{1.45}$$

đúng với mọi $m \geq 1$. Bây giờ, áp dụng đẳng thức này hai lần, chúng ta dễ dàng tính được kì vọng của X^2. Cụ thể như sau:

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{k=0}^{n} k^2 f(k) \\
&= \sum_{k=1} k^2 C_n^k p^k (1-p)^{n-k} \\
&= n \sum_{k=1}^{n} C_{n-1}^{k-1} p^k (1-p)^{n-k} \\
&= np \sum_{j=0}^{n-1} (j+1) C_{n-1}^{j} p^j (1-p)^{n-1-j} \quad \text{(đổi chỉ số } j = k-1\text{)} \\
&= np \left(\sum_{j=0}^{n-1} C_{n-1}^{j} p^j (1-p)^{n-1-j} + \sum_{j=0}^{n-1} j C_{n-1}^{j} p^j (1-p)^{n-1-j} \right) \\
&= np\left(1+(n-1)p\right) \quad \text{(áp dụng (1.45)).}
\end{aligned}$$

Từ đó, áp dụng Định lí 1.3, ta có

$$\sigma_X^2 = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = np\left(1+(n-1)p\right) - (np)^2 = np(1-p).$$

Đó là điều phải chứng minh. □

Ví dụ 1.35. Trong một kỳ thi, xác suất để một thí sinh bất kỳ thi đỗ là 80%. Chọn ngẫu nhiên 5 thí sinh. Trung bình có bao nhiêu thí sinh trong số đó thi đỗ? Xác suất để xảy ra đúng 4 thí sinh thi đỗ?

Lời giải

Gọi X là số thí sinh đỗ trong 5 thí sinh được chọn ngẫu nhiên. Thế thì X có phân phối nhị thức $B(n,p)$ với $n=5$ và $p=0{,}8$ (80%). Do đó, trung bình có

$$\mu = \mathbb{E}[X] = np = 5 \times 0{,}8 = 4$$

thí sinh thi đỗ. Mặt khác, xác suất để có đúng 4 thí sinh thi đỗ là

$$P(X=4) = \binom{5}{4} (0{,}8)^4 (1-0{,}8)^1 = 0{,}4096,$$

tức là chỉ có khoảng 40,96%.

Kết thúc mục này, chúng ta chứng minh định lí sau:

Định lí 1.6: (Bernoulli)

Nếu f_n là tần suất xuất hiện sự kiện A trong n phép thử độc lập và p là xác suất xuất hiện sự kiện đó trong mỗi phép thử thì với mọi $\varepsilon > 0$

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

Chứng minh. Gọi X_n là số lần xuất hiện sự kiện A trong n phép thử độc lập thì X_n có phân phối nhị thức $B(n,p)$. Do đó

$$\mathbb{E}[X_n] = np, \quad \text{Var}(X_n) = np(1-p).$$

Từ tính chất của kì vọng và phương sai, ta suy ra

$$\mathbb{E}[X_n/n] = np/n = p,$$

trong khi đó,

$$\mathbb{E}[X_n/n] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Do đó, áp dụng bất đẳng thức Chebyshev (Định lí 1.5), ta có

$$1 \geq P(|X_n/n - 1| \geq \epsilon) \geq 1 - \frac{\text{Var}(X_n/n)}{\epsilon^2} = 1 - \frac{p(1-p)}{n\epsilon^2}.$$

Lấy giới hạn khi $n \to \infty$ và áp dụng định lí giới hạn bị kẹp, ta có

$$\lim_{n\to\infty} P(|X_n/n - 1| \geq \epsilon) = 1.$$

Chú ý rằng $f_n = X_n/n$, và ta có điều phải chứng minh. □

Định lí 1.6, còn gọi là "Luật số lớn dạng yếu", cho thấy tần suất thực nghiệm f_n sẽ *hội tụ theo xác suất* về xác suất lí thuyết p khi số phép thử n tăng lên. Nó là cơ sở lí thuyết cho định nghĩa thống kê của xác suất, và cho thấy rằng trong thực tiễn, nếu số quan sát đủ lớn, xác suất có thể được ước lượng gần đúng bằng tần suất.

Cần lưu ý rằng đây là sự hội tụ theo xác suất, không phải hội tụ tuyệt đối: với mỗi giá trị hữu hạn của n, sai lệch giữa f và p vẫn có thể đáng kể, nhưng xác suất để sai lệch này vượt quá một ngưỡng bất kỳ sẽ tiến về 0 khi $n \to \infty$.

1.9.2 Phân phối Poisson

Phân phối xác suất của một BNN rời rạc X với vô hạn đếm được (countable) giá trị với hàm xác suất

$$P(X=k)=f(k)=\frac{\mu^k}{k!}e^{-\mu},\quad k=0,1,2,\ldots \tag{1.46}$$

gọi là *phân phối Poisson*[1] với tham số μ, ở đây $\mu>0$. Lưu ý, theo khai triển Taylor của hàm e^x, chúng ta có

$$\sum_{k=0}^{\infty}f(k)=\sum_{k=0}^{\infty}\frac{\mu^k}{k!}e^{-\mu}=e^{-\mu}\left(\sum_{k=0}^{\infty}\frac{\mu^k}{k!}\right)=e^{-\mu}e^{\mu}=1.$$

Phân phối Poisson có thể xem là giới hạn của dãy các phân phối nhị thức $B(n,p)$ khi $p\to 0$, $n\to\infty$, và $np\to\mu$, và thường được sử dụng để mô hình hóa số lần các sự kiện xảy ra trong một khoảng thời gian hoặc trong một không gian nhất định, trong đó các sự kiện xảy ra độc lập với nhau và có cùng xác suất xảy ra trong mỗi đơn vị thời gian hoặc không gian.

Ví dụ 1.36. Giả sử chúng ta đang thực hiện một thử nghiệm lâm sàng để đánh giá hiệu quả của một loại thuốc mới. Trong quá trình thử nghiệm, chúng ta ghi nhận số lượng các phản ứng phụ xảy ra do thuốc trong một nhóm bệnh nhân trong một khoảng thời gian nhất định, ví dụ như trong vòng 1 tuần.

Thông thường, các phản ứng phụ có thể là những hiện tượng hiếm gặp và xảy ra độc lập với nhau. Giả sử chúng ta thu được dữ liệu và thấy rằng trung bình có khoảng 2 phản ứng phụ xảy ra mỗi tuần trong nhóm bệnh nhân này. Hỏi xác suất để xảy ra đúng 3 phản ứng phụ trong một tuần là bao nhiêu?

Lời giải

Gọi X là BNN cho số phản ứng phụ trong một tuần. Khi đó, X có phân phối Poisson với tham số $\mu=2$. Xác suất để xảy ra đúng 3 phản ứng phụ trong một tuần được được cho bởi $P(X=3)$. Do đó,

$$P(X=3)=\frac{2^3}{3!}e^{-2}\approx 0{,}1804.$$

Vậy, xác suất cần tính là xấp xỉ 18,04%.

CAS 1.7. Trong R, phân phối Poisson được tính bằng hàm `dpois()`, với các tham số x (giá trị mà xác suất BNN nhận giá trị đó cần tính) và μ là tham số chính của phân phối. Với $\mu=2$ và $x=3$, chúng ta tính như sau:

[1] Siméon Denis Poisson (Poát-sông, 1781–1840) là một nhà toán học người Pháp.

```
# Khai báo các tham số mu và x
mu = 2; x = 3

# In ra giá trị của hàm xác suất phân phối Poisson
print(dpois(x, mu))

## [1] 0,1804
```

Định lí 1.7:

Trung bình và phương sai của một biến ngẫu nhiên X với phân phối Poisson với tham số μ đều bằng μ.

Chứng minh. Trung bình của phân phối Poisson là

$$\mu = \sum_{k=0}^{\infty} k f(k) = e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} = e^{-\mu}\mu \sum_{k=1}^{\infty} \frac{\mu^{(k-1)}}{(k-1)!} = \mu,$$

như đã khẳng định.

Để tính phương sai của X, chúng ta tính $E(X^2)$. Ta có

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \sum_{k=0}^{\infty} k^2 f(k) \\
&= e^{-\mu} \sum_{k=1}^{\infty} k^2 \frac{\mu^k}{k!} \\
&= e^{-\mu} \sum_{k=1}^{\infty} (k(k-1)+k) \frac{\mu^k}{(k-1)!} \\
&= e^{-\mu} \left(\sum_{k=2}^{\infty} \frac{\mu^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} \right) \\
&= e^{-\mu}(\mu^2 e^{\mu} + \mu e^{\mu}) \\
&= \mu^2 + \mu.
\end{aligned}$$

Từ đó, áp dụng Định lí 1.3, ta có

$$\sigma_X^2 = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = (\mu^2 + \mu) - \mu^2 = \mu.$$

Đó là điều phải chứng minh. □

1.9.3 Phân phối chuẩn

Phân phối chuẩn (normal distribution), hay còn được gọi là phân phối Gauss[1], là một trong những phân phối xác suất quan trọng nhất trong thống kê và xác suất. Nó được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau như kinh tế học, khoa học xã hội, khoa học máy tính, và nhiều lĩnh vực khác.

Phân phối chuẩn được mô tả bởi hai tham số: trung bình và độ lệch chuẩn. Đặc điểm quan trọng của phân phối chuẩn là nó có hình dạng đặc trưng là hình chuông và được đối xứng qua giá trị trung bình. Điều này có nghĩa là phần lớn dữ liệu trong phân phối nằm gần giá trị trung bình, và phân phối có đuôi dài ở hai phía.

Phân phối chuẩn được cho bởi bằng hàm mật độ xác suất:

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

Trong đó: ,

- μ là trung bình của phân phối.
- σ là độ lệch chuẩn của phân phối.
- x là biến ngẫu nhiên.
- e là số Euler (2,71828 ...), và π là số pi (3,14159 ...).

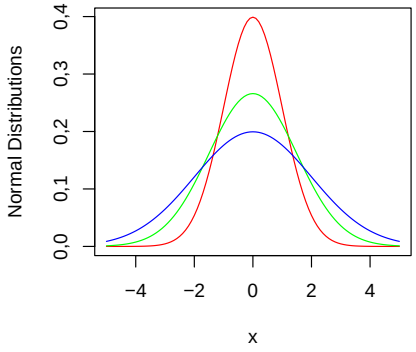
Phân phối chuẩn có nhiều ứng dụng trong đó phổ biến nhất là việc mô hình hóa dữ liệu thực tế và làm việc với các phép kiểm định thống kê như kiểm định giả thiết (test of hypothesis) và xây dựng các khoảng tin cậy (confidence interval).

Từ công thức hàm mật độ, chúng ta có hàm phân phối tích luỹ của phân phối chuẩn là

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, \mathrm{d}t \quad (1.47)$$

Đặc biệt, với $\mu = 0$ và $\sigma = 1$, chúng ta có

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t. \quad (1.48)$$



Hình 1.5: Hàm mật độ của các phân phối chuẩn có cùng trung bình

[1] Carl Friedrich Gauss (1777-1855) là một nhà toán học người Đức

Một BNN X có phân phối chuẩn với trung bình 0 và phương sai 1, $X \sim \mathcal{N}(0,1)$, gọi là BNN có *phân phối chuẩn tắc*.

Việc tính toán đối với hàm phân phối $F(x)$ thường được đưa về hàm Laplace $\Phi(x)$. Thật vậy, từ công thức đổi biến số trong tích phân, dễ dàng thu được định lí sau:

Định lí 1.8:

Giả sử $F(x)$ là hàm phân phối của một phân phối chuẩn với trung bình μ và độ lệch chuẩn σ. Khi đó

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{1.49}$$

Việc tính giá trị của Φ và hàm ngược Φ^{-1} thường khá phức tạp nên người ta tính sẵn các giá trị của nó và lập thành một bảng. Trong R, giá trị hàm $\Phi(z)$ và hàm ngược Φ^{-1} được tính bằng cách gọi hàm `pnorm()` và `qnorm()`.

Định lí 1.9:

Giả sử X là một BNN có phân phối chuẩn với trung bình μ và độ lệch chuẩn σ. Khi đó,

$$P(a < X \le b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{1.50}$$

Ví dụ 1.37. Giả sử BNN X có phân phối chuẩn với trung bình $\mu = 8$ và phương sai $\sigma^2 = 0{,}0625$. Biết rằng $\Phi(1{,}645) \approx 0{,}95$. Tìm c sao cho $P(X \le c) = 95\%$.

Lời giải

Gọi $F(x)$ là hàm phân phối xác suất của X, ta có $F(c) = P(X \le c) = 0{,}95$. Theo giả thiết, X có phân phối chuẩn với trung bình $\mu = 8$ và độ lệch chuẩn $\sigma = \sqrt{\sigma^2} = \sqrt{0{,}0625} = 0{,}25$ nên

$$F(x) = \Phi\left(\frac{x-8}{0{,}25}\right).$$

Từ đó, điều kiện đối với c là

$$\Phi\left(\frac{c-8}{0{,}25}\right) = F(c) = 0{,}.$$

Từ đó, ta tìm được

$$\frac{c-8}{0{,}25} \approx 1{,}645,$$

và giải được

$$c = 8 + 0{,}25 \cdot 1{,}645 \approx 8{,}41125.$$

Ví dụ 1.38. Trong một quy trình sản xuất, đường kính X của sản phẩm có phân phối chuẩn với trung bình là 5 cm và độ lệch chuẩn là 0,005 cm.

(a) Hỏi xác suất sản phẩm được sản xuất ra là phế phẩm là bao nhiêu nếu chúng ta cho phép sai số của đường kính là $5 \pm 0{,}01$ cm ?

(b) Tìm dung sai d sao cho tỉ lệ phế phẩm là 1%?

Lời giải

(a) Nếu $F(x)$ là hàm phân phối tích luỹ của X thì xác suất để sản phẩm sản xuất ra là chính phẩm là

$$P_{CP} = P(5-0{,}01 < X < 5+0{,}01) = F(5{,}01) - F(4{,}99).$$

Theo giả thiết, X có phân phối chuẩn với trung bình $\mu = 5$ và độ lệch chuẩn $\sigma = 0{,}01$ nên

$$F(x) = \Phi\left(\frac{x-5}{0{,}005}\right).$$

Từ đó

$$P_{CP} = \Phi(2) - \Phi(-2) \approx 0{,}9545 = 95{,}45\%.$$

Ta suy ra xác suất để sản phẩm là phế phẩm là

$$P_{PP} = 1 - P_{CP} \approx 1 - 0{,}9545 = 0{,}0455 = 4{,}55\%.$$

(b) Như phân tích ở trên, nếu dung sai cho phép là d thì xác suất để sản phẩm lấy ra là phế phẩm là

$$\begin{aligned} P_{\text{PP}} = 1 - P_{CP} &= 1 - P(5-d < X \le 5+d) \\ &= 1 - (F(5+d) - F(5-d)) \\ &= 1 - \Phi(d/\sigma) + \Phi(-d/\sigma) \\ &= 2\Phi(-d/\sigma) \end{aligned}$$

Hay

$$\Phi(-d/\sigma) = 0{,}005.$$

Từ đó, ta tìm được

$$-\frac{d}{\sigma} = \Phi^{-1}(0{,}005) \approx -2{,}5758,$$

có nghĩa là $d \approx 2{,}5758\sigma \approx 1{,}2879 \cdot 10^{-2}$ (cm).

Chú ý 1.8. Đối với BNN liên tục như phân phối chuẩn, giá trị của hàm mật độ xác suất $f(x)$ của phân phối chuẩn *không* phải là xác suất mà, đúng như tên gọi của nó, là *mật độ* xác suất: Giá trị $f(x)$ cho một xấp xỉ của xác suất X nhận giá trị rơi vào trong một khoảng I "rất nhỏ" xung quanh điểm x trên mỗi đơn vị độ dài:

$$f(x) = \lim_{x \in I, |I| \to 0} \frac{P(X \in I)}{|I|},$$

trong đó $|I|$ là độ dài của I (xem mô phỏng trong CAS 1.9).

Lưu ý, mặc dù đều là những khái niệm cơ bản và quan trọng trong lí thuyết biến ngẫu nhiên, hai khái niệm xác suất và mật độ xác suất vẫn thường bị nhầm lẫn với nhau. Cuốn *"Phân tích dữ liệu với R"*[8, trang 57] viết: "Chẳng hạn như xác suất một phụ nữ Việt Nam có chiều cao 160 cm bằng

$$P(X = 160 \mid \mu = 156, \sigma = 4.6) = [\text{ công thức }] = 0.0594."$$

Ở đây, vế trái biểu thị xác suất của sự kiện $X = 160$, và xác suất này bằng 0 do X là BNN liên tục, trong khi đó vế phải là mật độ xác suất tại $x = 160$, và luôn khác không. Ở đó, tác giả của [8] đã hoàn toàn nhầm lẫn giữa xác suất và mật độ của nó.

CAS 1.8. Trong R, để tính toán với các phân phối chuẩn, chúng ta dùng `pnorm()` cho hàm phân phối tích luỹ và `dnorm()`[1] cho hàm mật độ của phân phối liên tục.

Với trung bình $\mu = 5$ và độ lệch chuẩn $\sigma = 0{,}005$, hàm phân phối tích luỹ sẽ là `pnorm(5.01, mean = 5, sd = 0.005)`.

```
# Định nghĩa hàm xác suất tích luỹ F
F <- function(x) pnorm(x, mean = 5, sd = 0.005)

# Tính xác suất X rơi trong khoảng 4,99 và 5,01.
p <- F(5.01) - F(4.99)

# Hiển thị kết quả
print(p)

## [1] 0,9545
```

Kết quả tính được là P_{CP} =0,9545 (làm tròn tới 4 chữ số sau dấu phẩy thập phân).

[1] Chữ cái "d" trong "dnorm" có lẽ lấy từ chữ cái đầu của từ "density", trong Tiếng Anh có nghĩa là *mật độ*. Tuy nhiên, hàm xác suất của một số BNN rời rạc trong R cũng được đặt tên bắt đầu bằng chữ "d", như dpois() (phân phối Poisson) hay dbinom() (phân phối nhị thức). Trong trường hợp này, chữ "d" không còn mang nghĩa mật độ (density) nữa.

CAS 1.9. Sau đây, ta xét $X \sim \mathcal{N}(156,4,6^2)$, tính xác suất $P(160-\epsilon < X < 160+\epsilon)$, và chia các kết quả cho độ dài 2ϵ, với $\epsilon = 1, 1/2, 1/3, 1/4, 1/5$. Chương trình này minh hoạ sự khác biệt và mối liên hệ giữa `pnorm()` và `dnorm()`.

```
# Hàm mật độ
f <- function(x) dnorm(x, mean = 156, sd = 4.6)

# Hàm xác suất tích luỹ
F <- function(x) pnorm(x, mean = 156, sd = 4.6)

for (n in 1:5) {

    # Bán kính
    e = 1 / n

    # Tính các xác suất để X rơi vào khoảng 160 - e và 160 + e
    p <- F(160 + e) - F(160 - e)

    # Tính trung bình xác suất trên mỗi đơn vị độ dài
    d <- p / (2 * e)

    # Hiển thị kết quả
    print(d)
}
```

```
## [1] 0,0593082
## [1] 0,0593948
## [1] 0,0594107
## [1] 0,0594163
## [1] 0,0594189
```

Kết quả cho thấy 5 giá trị có xu hướng tiệm cận đến giá trị hàm mật độ xác suất tại $x = 160$.

Ví dụ 1.39. Chỉ số BMI[1] của nhóm các bệnh nhân trong một thử nghiệm là biến ngẫu nhiên có phân phối chuẩn với kì vọng là 29 ($\mathrm{kg\,m^{-2}}$) và phương sai là 36 $\mathrm{kg^2 m^{-4}}$. Một người có chỉ số BMI bình thường nếu chỉ số BMI của người đó nằm trong khoảng (18,5; 24,9). Kiểm tra ngẫu nhiên BMI của một người trong nhóm.

(a) Tính xác suất để người này có BMI bé hơn 35.

(b) Tính xác suất để người này có BMI ở mức bình thường.

Lời giải

[1]BMI: *Body Mass Index*, chỉ số về khối lượng cơ thể.

(a) Bởi vì $X \sim \mathcal{N}(\mu = 29, \sigma^2 = 36)$ nên hàm phân phối xác suất tích lũy của X là

$$F(x) = \Phi\left(\frac{x-29}{\sqrt{36}}\right),$$

trong đó $\Phi(z)$ là hàm phân phối tích lũy của phân phối chuẩn tắc $\mathcal{N}(0,1)$ (hàm Laplace). Do đó

$$P(X < 35) = F(35) = \Phi\left(\frac{35-29}{\sqrt{36}}\right) = \Phi(1).$$

Tra bảng phân phối chuẩn tắc, hoặc sử dụng máy tính, chúng ta có $\Phi(1) \approx 0{,}8413447$. Vậy, xác suất để người được chọn có BMI bé hơn 35 là xấp xỉ 84,13%.

(b) Lập luận như trường hợp 1 hoặc định lí 1.9, chúng ta có xác suất để người được chọn có BMI ở mức bình thường là

$$\begin{aligned} P(18{,}5 < X < 24{,}9) &= \Phi\left(\frac{24{,}9-29}{\sqrt{36}}\right) - \Phi\left(\frac{18{,}5-29}{\sqrt{36}}\right) \\ &\approx \Phi(-0{,}683) - \Phi(-1{,}75) \\ &\approx 0{,}2072. \end{aligned}$$

Trong bước tính giá trị của Φ, chúng ta cần tra bảng phân phối chuẩn, hoặc sử dụng máy tính. Kết luận, xác suất để người này có BMI ở mức bình thường là 20,72%.

1.9.4 Phân phối hình học và phân phối nhị thức âm

Trong tiểu mục này, ta nói về phân phối hình học (geometric distribution) và phân phối nhị thức âm (negative binomial distribution).

Phân phối hình học

Giả sử chúng ta thực hiện một dãy (vô hạn) các phép thử Bernoulli độc lập với xác suất thành công là p và xác suất thất bại là $1-p$. Gọi X là lần số thứ tự của lần thành công đầu tiên trong dãy. Khi đó, X là một BNN rời rạc với nhận các giá trị $1, 2, \ldots, n, \ldots$. Dễ dàng thấy rằng, $X = 1$ nếu kết quả đầu tiên là thành công. Như vậy, $P(X = 1) = p$. Tương tự, $X = 2$ nếu phép thử đầu tiên thất bại, và phép thử thứ hai thành công. Xác suất để hai sự kiện độc lập này cùng xảy ra là $(1-p)(p)$ và vậy thì

$$P(X = 2) = (1-p)p.$$

Tổng quát hơn,

$$P(X=n)=(1-p)^{n-1}p. \tag{1.51}$$

Định nghĩa 1.23:

Một BNN rời rạc X được gọi là có phân phối hình học với tham số p nếu hàm xác suất của nó cho bởi phương trình (1.51).

Biến ngẫu nhiên X với phân phối hình học nhận vô hạn giá trị và điều kiện tổng các xác suất bằng 1 có dạng

$$\begin{aligned}\sum_{n=1}^{\infty} f(n) &= \sum_{n=1}^{\infty}(1-p)^{n-1}p \\ &= p\sum_{m=0}^{\infty} q^m \quad \text{với } q=1-p \\ &= p\left(\frac{1}{1-q}\right)=1.\end{aligned}$$

Lưu ý rằng trong vế phải xuất hiện chuỗi số dương quen thuộc trong giải tích – chuỗi hình học (geometric series):

$$\sum_{m=0}^{\infty} q^m = \lim_{N\to\infty}\left(\sum_{m=0}^{N} q^m\right) = \lim_{N\to\infty}\left(\frac{1-q^{N+1}}{1-q}\right) = \frac{1}{1-q},$$

khi $|q|<1$.

Định lí 1.10:

Giả sử BNN X có phân phối hình học với tham số p. Khi đó, $\mathbb{E}[X]=1/p$ và $\sigma_X^2=(1-p)/p^2$.

Chứng minh. Theo công thức của kì vọng, chúng ta có

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=1}^{\infty} nP(X=n) \\ &= \sum_{n=1}^{\infty} npq^{n-1} \\ &= p\lim_{N\to\infty}\left(\sum_{n=1}^{N} nq^{n-1}\right), \quad q=1-p.\end{aligned}$$

Mặt khác, lưu ý rằng

$$\begin{aligned}\sum_{n=1}^{N} nq^{n-1} &= \frac{\mathrm{d}}{\mathrm{d}q}\left(1+q+\cdots q^N\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{1-q^{N+1}}{1-q}\right)\\ &= \frac{1-q^{N+1}-(N+1)(1-q)q^N}{(1-q)^2}.\end{aligned}$$

Từ đó, lấy giới hạn khi $N \to \infty$ ta có

$$\begin{aligned}\mathbb{E}[X] &= p\lim_{N\to\infty}\left(\frac{1-q^{N+1}-(N+1)(1-q)q^N}{(1-q)^2}\right)\\ &= p\frac{1}{(1-q)^2} = \frac{1}{p}.\end{aligned}$$

Đó là công thức cho kì vọng. □

Phân phối nhị thức âm

Phân phối nhị thức âm (negative binomial distribution) là một mở rộng của phân phối hình học. Một BNN rời rạc X gọi là có phân phối nhị thức âm với hai tham số $r \in \mathbb{Z}_{\geq 1}$ và $p \in (0,1)$ nếu X nhận các giá trị $0,1,2,\ldots$, với xác suất

$$P(X=0) = C_{k+r-1}^{k} p^r (1-p)^k, \quad k = 1,2,\ldots$$

Phân phối nhị thức âm dùng để mô tả chuỗi các phép thử Bernoulli không giới hạn số lần trên đó BNN X là số lần thử để xuất hiện lần thành công thứ r.

Ví dụ 1.40. Tỉ lệ sản phẩm bị lỗi của một dây chuyền sản xuất là 0,08. Bộ phận quản lí chất lượng lấy ngẫu nhiên và liên tiếp một số lượng lớn sản phẩm để kiểm tra. Tính xác suất để trong lần lấy sản phẩm thứ 9 họ lấy được sản phẩm lỗi thứ 3.

Lời giải

Nếu X là số lần lấy được sản phẩm tốt cho đến khi lấy được sản phẩm lỗi thứ 3, thì X là một BNN có phân phối nhị thức âm với $r = 3$ và $p = 0,08$. Sự kiện lấy được sản phẩm lỗi thứ 3 ở lần lấy thứ 9 là sự kiện lấy được đúng 6 sản phẩm tốt cho đến lần lấy thứ 9. Do đó, xác suất cần tính là

$$P(X=6) = C_8^6 (0,08)^3 (1-0,08)^6 = 0,008693.$$

Như vậy, xác suất để lấy được sản phẩm lỗi thứ 3 trong lần lấy thứ 9 là $\approx$ 0,87%.

1.9.5 Phân phối Gamma, phân phối mũ, và phân phối chi-bình phương

Hàm gamma là hàm số xác định bởi công thức

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, \mathrm{d}x, \tag{1.52}$$

với $\alpha > 0$.

Lưu ý, việc tính toán các giá trị hàm gamma tương đối khó khăn, do chúng ta phải tính các tích phân suy rộng. Tuy nhiên, với các giá trị $\alpha = n$ nguyên dương, có thể tính được

$$\Gamma(n) = (n-1)!.$$

Chẳng hạn, $\Gamma(1) = 0! = 1$, $\Gamma(3) = (3-1)! = 2$, v.v.

Định nghĩa 1.24:

Một BNN liên tục X có *phân phối gamma* với tham số $\alpha > 0$ và $\beta > 0$ nếu nó có hàm mật độ là

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0,$$

và $f(x) = 0$ với $x \le 0$.

Hai trường hợp đặc biệt và quan trọng của phân phối gamma là trường hợp $\alpha = 1$ và $\alpha = \nu/2$, ν nguyên dương, và $\beta = 2$.

Với $\alpha = 1$, chúng ta có

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0,$$

và nếu X có mật độ như trên thì chúng ta nói X có *phân phối mũ* (exponential distribution).

Với $\alpha = \nu/2$, với ν nguyên dương, và $\beta = 2$, hàm $f(x)$ trở thành hàm mật độ xác suất của phân phối *chi-bình phương* (kí hiệu χ^2). Phân phối chi-bình phương đóng một vai trò quan trọng trong một số bài toán kiểm định giả thiết thống kê, như bài toán kiểm định phương sai của phân phối chuẩn. Chúng ta sẽ trở lại với phân phối này trong §3.3.

Định lí 1.11:

Nếu BNN liên tục X có phân phối gamma với hai tham số α, β như trên thì trung bình μ_X và phương sai σ_X^2 của X lần lượt là

$$\mu_X = \mathbb{E}[X] = \alpha\beta, \tag{1.53}$$

và

$$\sigma_X^2 = \alpha\beta^2. \tag{1.54}$$

Chứng minh. Từ công thức của hàm mật độ, chúng ta có

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \\
&= \int_0^{\infty} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^\alpha e^{-x/\beta}\,\mathrm{d}x \\
&= \frac{\beta}{\Gamma(\alpha)} \int_0^{\infty} t^\alpha e^{-t}\,\mathrm{d}t \quad \text{(đổi biến } x = \beta t,\ \mathrm{d}x = \beta\,\mathrm{d}t) \\
&= \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \beta\alpha.
\end{aligned}$$

Ở bước cuối cùng, chúng ta đã sử dụng tính chất $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. (1.53) được chứng minh.

Để chứng minh (1.54), cũng bằng phép đổi biến $x = \beta t$

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_0^{\infty} x^2 f(x)\,\mathrm{d}x \\
&= \int_0^{\infty} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha+1} e^{-x/\beta}\,\mathrm{d}x \\
&= \frac{\beta^2}{\Gamma(\alpha)} \int_0^{\infty} t^{\alpha+1} e^{-t}\,\mathrm{d}t \\
&= \frac{\beta^2\Gamma(\alpha+2)}{\Gamma(\alpha)} \\
&= \beta^2\alpha(\alpha+1).
\end{aligned}$$

Từ đó,

$$\sigma_X^2 = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \beta^2\alpha(\alpha+1) - (\alpha\beta)^2 = \alpha\beta^2.$$

Đó là điều phải chứng minh. □

1.9.6 Phân phối Fisher

Phân phối Fisher[1] (hay còn gọi là *phân phối F*) là một phân phối xác suất liên quan đến thống kê suy luận, thường được sử dụng trong phân tích phương sai (ANOVA) và kiểm định giả thiết.

Phân phối F được nảy sinh khi tìm phân phối của tỉ lệ của hai biến ngẫu nhiên có phân phối chi-bình phương, và được dùng để so sánh hai phương sai của các nhóm mẫu. Nếu U_1 và U_2 là hai biến ngẫu nhiên có phân phối chi-bình phương với số bậc tự do tương ứng là d_1 và d_2, thì BNN F được định nghĩa như sau:

$$F = \frac{U_1/d_1}{U_2/d_2} \tag{1.55}$$

là một BNN tuân theo phân phối F với d_1 và d_2 là bậc tự do.

Hàm mật độ xác suất của phân phối Fisher được cho bởi:

$$f_{d1,d2}(x) = \frac{\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{(d_1/2-1)}}{\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)\left(1+\frac{d_1}{d_2}x\right)^{\frac{d_1+d_2}{2}}}, \tag{1.56}$$

với các điều kiện: $x \geq 0$, $d_1 > 0$ và $d_2 > 0$ (bậc tự do là số nguyên dương). Ở trên, B là hàm Beta, được định nghĩa như:

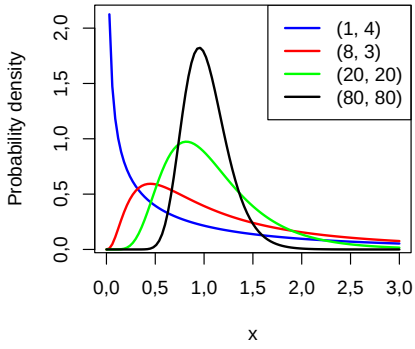
$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t. \tag{1.57}$$

Nhân tử chứa hàm Beta nhằm đảm bảo rằng tổng diện tích dưới đường cong của hàm mật độ là 1.

Một số đặc điểm của phân phối Fisher như sau:

- Phân phối Fisher không đối xứng và lệch phải, với đuôi dài hơn ở phía bên phải.
- Giá trị kì vọng của phân phối Fisher là gần 1 nếu số bậc tự do của tử số và mẫu số tương đối gần nhau.

Trong thực tiễn, phân phối F thường được tính toán thông qua các phần mềm thống kê khi thực hiện kiểm định F.



Hình 1.6: Đồ thị của hàm mật độ của phân phối F với các cặp bậc tự do khác nhau

[1] Ronald Fisher (1890 - 1962) là một nhà thống kê học người Anh.

Kì vọng và phương sai của một biến ngẫu nhiên với phân phối Fisher được cho bởi hai công thức trong định lí sau đây.

Định lí 1.12:

Giả sử X là một BNN có phân phối Fisher với song bậc tự do (d_1, d_2). Khi đó,

1. Nếu $d_2 > 2$ thì X có kì vọng
$$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}.$$
2. Nếu $d_2 > 4$ thì X có phương sai là
$$\mathrm{Var}(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)}.$$

1.9.7 Xấp xỉ phân phối nhị thức bằng phân phối chuẩn

Xấp xỉ phân phối nhị thức bằng phân phối chuẩn là một kĩ thuật thường được sử dụng trong thống kê khi số lượng mẫu n của phân phối nhị thức lớn. Phương pháp này được gọi là "phép xấp xỉ chuẩn" (normal approximation) và đặc biệt hữu ích khi cần tính toán nhanh chóng hoặc không có bảng giá trị phân phối nhị thức.

Cố định $p \in (0,1)$. Với n nguyên dương, hàm phân phối tích luỹ của phân phối nhị thức $B(n,p)$ là

$$B_n(x;p) = \sum_{0 \le k \le x} C_n^k p^k (1-p)^{n-k}.$$

Khi đó, định lí giới hạn của phân phối nhị thức của de Moivre (1733), còn gọi là định lí de Moivre–Laplace, có thể phát biểu như sau: *Với mọi z,*

$$\lim_{n\to\infty} B_n\left(np + z\sqrt{np(1-p)}; p\right) = \Phi(z), \tag{1.58}$$

với $\Phi(z)$ là hàm Laplace đã nói ở mục phân phối chuẩn.

Nếu đặt $x = np + z\sqrt{np(1-p)}$ thì $z = (x-np)/\sqrt{np(1-p)}$. Chúng ta có thể viết lại giới hạn trên như sau:

$$\lim_{n\to\infty} B_n(x;p) = \Phi(z). \tag{1.59}$$

Giới hạn (1.59) cho phép chúng ta, trong trường hợp n lớn, tính gần đúng xác suất trong phân phối nhị thức bằng phân phối chuẩn.

Giả sử $X \sim B(n,p)$ với n "lớn". Với mọi cặp số tự nhiên a và b, $a<b$,

$$P(a \le X \le b) = \sum_{k=a}^{b} C_n^k p^k (1-p)^{n-k} \sim \Phi(\beta) - \Phi(\alpha), \tag{1.60}$$

với

$$\alpha = \frac{a-\mu-0{,}5}{\sigma}, \quad \beta = \frac{b-\mu+0{,}5}{\sigma}, \tag{1.61}$$

trong đó

$$\mu = np \text{ và } \sigma = \sqrt{np(1-p)} \tag{1.62}$$

lần lượt là trung bình và độ lệch chuẩn của X.

Ví dụ 1.41. Trong một dây chuyền sản xuất một sản phẩm A, xác suất để một sản phẩm sản xuất ra bị lỗi là $p = 0{,}01$. Lấy ngẫu nhiên một mẫu gồm 1240 sản phẩm. Bằng cách xấp xỉ phân phối nhị thức bằng phân phối chuẩn, tính gần đúng xác suất để mẫu lấy được có số sản phẩm bị lỗi ít nhất 10 và nhiều nhất là 15.

Lời giải

BNN mà chúng ta quan tâm là số các sản phẩm lỗi trong mẫu 1240 sản phẩm, kí hiệu là X. Có thể xem X có phân phối nhị thức với $n = 1240$ và $p = 0{,}01$, tức là $X \sim B(1240;\ 0{,}01)$. Khi đó trung bình và phương sai của X là

$$\mu = np = 12{,}4,$$

và

$$\sigma = \sqrt{np(1-p)} = 12{,}276.$$

Với $a = 10$ và $b = 15$, từ công thức (1.61), chúng ta tính được

$$\alpha = \text{-}0{,}8277, \quad \beta = 0{,}8848.$$

Từ công thức xấp xỉ trên, ta có

$$P(10 \le X \le 15) \approx \Phi(0{,}8848) - \Phi(\text{-}0{,}8277) \approx 0{,}608.$$

Vậy xác suất để có ít nhất 10 và nhiều nhất 15 sản phẩm lỗi trong mẫu 1240 sản phẩm là $\approx 0{,}608$, dựa trên xấp xỉ phân phối nhị thức bằng phân phối chuẩn tắc.

Nhận xét 1.11. Nếu tính trực tiếp bằng phân phối nhị thức, chúng ta được kết quả $\approx 0{,}6072$. Bạn đọc hãy so sánh kết quả được tính trực tiếp và kết quả tính bằng xấp xỉ phân phối chuẩn và rút ra nhận xét cho mình.

Bài tập mục 1.9

1. Trong thành phố có 25% gia đình sở hữu xe điện. Chọn ngẫu nhiên 16 gia đình và gọi X là số gia đình sở hữu xe điện trong đó.

(a) X có phân bố gì?

(b) Tính xác suất để có đúng 4 gia đình có xe điện.

(c) Tính kì vọng $\mathbb{E}[X]$ và phương sai σ_X^2.

2. Giả sử chiều cao của phụ nữ hiện nay là phân phối chuẩn với trung bình $\mu = 156$ cm và độ lệch chuẩn $\sigma = 4{,}6$ cm.

(a) Một sinh viên đã sử dụng công thức cho hàm mật độ $f(x)$ của phân phối chuẩn với $\mu = 156$ và $\sigma = 4{,}6$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

để tính $f(160) \approx 0{,}0594$ và từ đó kết luận xác suất để một phụ nữ được chọn ngẫu nhiên có chiều cao 160 cm là khoảng 5,95%. Kết luận của bạn sinh viên đó đúng hay sai? Tại sao?

(b) Tìm xác suất của một phụ nữ được chọn ngẫu nhiên có chiều cao rơi vào khoảng từ 159,5 cm đến 160,5 cm.

3. Giả sử X có phân phối chuẩn với trung bình 5 và phương sai 4. Tính các xác suất $P(X < 6)$, $P(X > 2)$, và $P(3 < X < 5)$.

4. Giả sử X có phân phối chuẩn với trung bình 15 và độ lệch chuẩn 2. Tìm c sao cho $P(X < c) = 95\%$.

5. Giả sử X có phân phối chuẩn với trung bình 10 và độ lệch chuẩn 1,5. Tìm c sao cho $P(X > c) = 0{,}2$.

6. (Rutherford–Geiger) Năm 1910, Rutherford–Geiger bằng thực nghiệm phát hiện ra rằng số hạt α phát ra mỗi giây (s) của một chất phóng xạ là một BNN với phân phối Poisson. Giả sử X có kì vọng $\mu_X = 0{,}6$. Tính xác suất của sự kiện có 2 hay nhiều hơn hạt α phát ra trong khoảng thời gian một giây.

7. Trong một nhóm các bệnh nhân đang điều trị một loại thuốc mới, tỉ lệ gặp phải phản ứng phụ là p. Chọn ngẫu nhiên 167 bệnh nhân. Hỏi p lớn nhất có thể bằng bao nhiêu để xác suất không có bệnh nhân nào gặp phải phản ứng phụ không nhỏ hơn 5%?

1.10 Hàm của các biến ngẫu nhiên

Giả sử X và Y là hai biến ngẫu nhiên trên một không gian mẫu Ω. Nếu $\varphi(x, y)$ là một hàm số xác định với mọi $x \in X(\Omega)$ và $y \in Y(\Omega)$ thì $Z = \varphi(X, Y)$ cũng là một BNN trên Ω, nếu xác suất các sự kiện $\{Z < a\}$ xác định (chẳng hạn, khi φ là hàm liên tục). Khi đó, Z là hàm của hai BNN X và Y.

Sau đây, chúng ta tìm hiểu mối quan hệ giữa kì vọng và phương sai của Z với kì vọng và phương sai của X và Y. Tiếp theo, chúng ta nêu định nghĩa của khái niệm "hiệp phương sai" (covariance) của X và Y.

Trước hết, chúng ta xét trường hợp các BNN X và Y là rời rạc và $f(x,y)$ là hàm xác suất đồng thời của X và Y, tức là

$$f(x,y) = P(X = x, Y = y).$$

Khi đó, từ công thức cộng xác suất, giá trị $P(Z = z)$ là tổng tất cả các giá trị $f(x,y)$ khi x, y chạy qua các giá trị sao cho $z = \varphi(x,y)$. Cụ thể hơn, nếu $f(z)$ là hàm xác suất của Z thì

$$f(z) = P(g(X,Y) = z) = \sum_{\varphi(x,y)=z} f(x,y).$$

Từ đó, kì vọng của Z là

$$\begin{aligned}\mathbb{E}[Z] &= \sum_z z f(z) \\ &= \sum_z z \left(\sum_{\varphi(x,y)=z} f(x,y) \right) \\ &= \sum_x \sum_y \varphi(x,y) f(x,y), \end{aligned} \tag{1.63}$$

trong đó, tổng thứ nhất lấy theo mọi $z \in Z(\Omega)$ và tổng thứ hai lấy theo mọi cặp $(x,y) \in X(\Omega) \times Y(\Omega)$ sao cho $z = \varphi(x,y)$.

Nói riêng, nếu $g(x,y) = x + y$ thì

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_x \sum_y (x+y) f(x,y) \\ &= \sum_x \sum_y x f(x,y) + \sum_x \sum_y y f(x,y).\end{aligned}$$

Mặt khác, theo công thức xác suất toàn phần thì

$$\begin{aligned}\sum_x \sum_y x f(x,y) &= \sum_x x \sum_y P(X = x, Y = y) \\ &= \sum_x x P(X = x) \\ &= \mathbb{E}[X],\end{aligned}$$

và

$$\begin{aligned}\sum_x \sum_y y f(x,y) &= \sum_y y \sum_x P(X = x, Y = y) \\ &= \sum_y y P(Y = y) \\ &= \mathbb{E}[Y].\end{aligned}$$

Khi tổng hợp các đẳng thức trên, chúng ta thu được

$$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]. \tag{1.64}$$

Điều đó có nghĩa là kì vọng của tổng hai BNN bằng tổng các kì vọng của chúng.

Bằng quy nạp, chúng ta có kết quả sau tổng quát sau đây:

Định lí 1.13:

Kì vọng của một tổng các biến ngẫu nhiên bằng tổng các kì vọng của chúng:

$$\mathbb{E}[X_1+X_2+\cdots+X_n]=\mathbb{E}[X_1]+\mathbb{E}[X_2]+\cdots+\mathbb{E}[X_n] \tag{1.65}$$

Tiếp theo, chúng ta phân tích trường hợp $\varphi(x,y)=xy$, giả sử thêm rằng X và Y là độc lập với các hàm xác suất $g(x)$ và $h(y)$. Khi đó, kì vọng của X và Y lần lượt là

$$\mathbb{E}[X]=\sum_x xg(x), \quad \mathbb{E}[Y]=\sum_y yh(y),$$

trong khi đó, hàm xác suất đồng thời của X và Y là

$$f(x,y)=g(x)h(y).$$

Do đó, áp dụng (1.63) với $\varphi(xy)=xy$ và $f(x,y)=g(x)h(y)$, chúng ta thu được

$$\begin{aligned}\mathbb{E}[X]=\mathbb{E}[XY]&=\sum_x\sum_y xyg(x)h(y)\\&=\left(\sum_x xg(x)\right)\left(\sum_y yh(y)\right)\\&=\mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

Như vậy, kì vọng của tích của hai BNN *độc lập* bằng tích của các kì vọng của chúng.

Cũng như trường hợp tổng của hai BNN, bằng quy nạp, chúng ta thu được định lí sau:

Định lí 1.14:

Kì vọng của một tích các biến ngẫu nhiên độc lập bằng tích các kì vọng của chúng. Có nghĩa là với các BNN độc lập $X_1, X_2, \ldots, X_n$,

$$\mathbb{E}[X_1X_2\cdots X_n]=\mathbb{E}[X_1]\mathbb{E}[X_2]\cdots\mathbb{E}[X_n]. \tag{1.66}$$

Nhận xét 1.12. Hai định lí trên cũng đúng cho các BNN liên tục. Tổng quát hơn, các kết quả đó vẫn đúng cho các BNN bất kì, rời rạc, liên tục tuyệt đối, hay liên tục kì dị, v.v. Chúng ta không đi sâu vào các BNN như vậy.

Tiếp theo, chúng ta nói về hiệp phương sai (covariance) của hai BNN.

Định nghĩa 1.25:

Hiệp phương sai của X và Y là

$$\sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{1.67}$$

Nhận xét 1.13. Nói riêng, phương sai của một BNN X bằng hiệp phương sai của X với chính nó: $\sigma_X^2 = \sigma_{XX}$. Mặt khác, hiệp phương sai của hai BNN độc lập bằng 0. Như vậy, hiệp phương sai đo tính không độc lập của hai BNN.

Dễ thấy hiệp phương sai có tính đối xứng:

$$\sigma_{XY} = \sigma_{YX}.$$

Định lí 1.15:

Cho X và Y là hai BNN và $Z = aX + bY + c$, với $a, b, c \in \mathbb{R}$. Khi đó

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \tag{1.68}$$

Nói riêng, với $a = b = 1$ và $c = 0$ thì

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY},$$

và vậy thì nếu X và Y độc lập thì phương sai của tổng $X + Y$ bằng tổng các phương sai của X và Y.

Chứng minh. Bài tập. □

Ví dụ 1.42. Xét một phép thử Bernoulli với xác suất thành công $p \in (0,1)$. Phép thử được thực hiện n lần liên tiếp, độc lập. Gọi I_i là BNN cho giá trị 1 nếu lần thử thứ i là thành công và nhận giá trị 0 nếu lần thử thứ i thất bại. Khi đó, các BNN I_i là độc lập, có cùng phân phối, với trung bình và phương sai cho bởi

$$\mu_i = \mu_{I_i} = p, \quad \sigma_i^2 = \sigma_{I_i}^2 = p(1-p).$$

Gọi X là số lần thử thành công trong n lần thử. Khi đó, theo cách xác định I_i,

$$X = I_1 + I_2 + \cdots + I_n.$$

Áp dụng các kết quả về trung bình và phương sai của tổng các BNN độc lập, chúng ta có

$$\mu_X = \sum_{i=1}^{n} \mu_i = np,$$

trong khi đó, do các BNN I_i là độc lập,

$$\sigma_X^2 = \sum_{i=1}^{n} \sigma_i^2 = np(1-p).$$

Chúng ta thu được một chứng minh khác của định lí về trung bình và phương sai của phân phối Bernoulli.

Bài tập §1.10

1. Giả sử X là một BNN với hàm mật độ xác suất $f(x) = e^{-x}$ với $x > 0$ và $f(x) = 0$ với $x \leq 0$. Tìm kì vọng của $Y = e^{2X/3}$.

2. Gọi X là số chấm thu được khi tung một xúc xắc cân xứng. Tìm kì vọng của $Y = 2X^2 - 5$.

3. Vận tốc V của một phân tử khí trong trạng thái cân bằng là một BNN với hàm mật độ xác suất cho bởi

$$f(v) = \begin{cases} \alpha v^2 e^{-bv^2} & \text{khi } v > 0, \\ 0 & \text{khi } v \leq 0. \end{cases}$$

Tìm phân phối xác suất của động năng $W = mV^2/2$ theo m, α, và b.

Tài liệu

[1] E. Kreyszig, *Advanced Engineering Mathematics, Part G: Probability, Statistics*, 10 **edition**. New York: John Wiley & Sons, Inc., 2011.

[2] Đ. H. Thắng, *Thống kê và Ứng dụng: Giáo trình dành cho các trường Đại học và Cao đẳng*, 6 **edition**. NXB Giáo dục Việt Nam, 2022.

[3] V. H. Nhự, C. H. Nam, L. Đ. Ninh **and** P. Q. Sáng, *Bài giảng Xác suất và Thống kê*. Tài liệu lưu hành nội bộ của ĐH Phenikaa, 2025.

[4] M. Akritas, *Probability and Statistics with R: for engineers and scientists*. New York: Pearson, 2015.

[5] H. Cramer, *Phương pháp toán học trong thống kê (dịch)*. NXB Khoa học và Kĩ thuật, 1970.

[6] A. N. Shiryaev, *Probability* (Graduate Text in Mathematics 95), 2 **edition**. Springer New York, 1996.

[7] N. V. Khuê, P. N. Thao, L. M. Hải **and** N. Đ. Sang, *Toán cao cấp, tập II (A2)*. NXB Giáo dục, 2008.

[8] N. V. Tuấn, *Phân tích dữ liệu với R*, 2 **edition**. NXB Tổng hợp TP. Hồ Chí Minh, 2014.

Chương 2

Ước lượng tham số

Mục tiêu của chúng ta trong chương này bao gồm:

- Các ước lượng điểm của tham số tổng thể như ước lượng của trung bình và phương sai, sơ lược về phương pháp hợp lí cực đại để tìm ước lượng điểm của tham số.
- Các ước lượng khoảng của tham số: Các phương pháp xây dựng khoảng tin cậy cho các tham số tổng thể, bao gồm khoảng tin cậy cho trung bình và phương sai của phân phối chuẩn, khoảng tin cậy cho xác suất p của phân phối $B(1,p)$.

Tài liệu tham khảo: Kreyszig [1, §25.1 – 25.3], Đ. H. Thắng [2, Chương 3].

Nội dung

2.1 Mẫu ngẫu nhiên và đặc trưng mẫu

Trong mục này, chúng ta nói về mẫu ngẫu nhiên và đặc trưng mẫu. Một số khái niệm trong đây đã biết ở §1.1.

2.1.1 Sơ lược về mẫu ngẫu nhiên

Giả sử chúng ta có một quần thể Ω nào đó, như tập tất cả các trái dưa hấu trong một vùng nguyên liệu, tập các sản phẩm của một dây chuyền sản xuất, hay tập tất cả các cử tri trong một cuộc bầu cử, v.v. Chọn ngẫu nhiên một đối tượng trong đó và quan sát một thuộc tính nào đó. Khi các thông tin quan sát được được biểu diễn dưới dạng số, chúng ta thu được một biến ngẫu nhiên giá trị thực, hãy gọi là $X : \Omega \to \mathbb{R}$. Ví dụ, gọi X là khối lượng (tính theo đơn vị nào đó) của một trái dưa hấu trên một vùng nguyên liệu. Khi thực hiện chọn một quả dưa hấu cụ thể, khối lượng của quả dưa được chọn sẽ cho giá trị của X.

Khi muốn tìm hiểu về tổng thể Ω, trong thực tế chúng ta không quan sát toàn bộ các đối tượng trong tổng thể. Thay vào đó, chúng ta quan sát một số lượng nhất định các đối tượng trong tổng thể. Khi đó, chúng ta thu được một mẫu. Số lượng các quan sát gọi là cỡ của mẫu n. Thông thường, quá trình lựa chọn các mẫu quan sát là ngẫu nhiên và độc lập, đảm bảo cho giá trị quan sát được có cùng phân phối xác suất và độc lập. Quá trình thu thập n mẫu ngẫu nhiên và độc lập như vậy được mô hình hóa bởi n biến ngẫu nhiên $X_1, X_2, \ldots, X_n$, gọi là một vectơ ngẫu nhiên n chiều.

Từ các BNN X_i, $i = 1,2,\ldots,n$, chúng ta xây dựng hai BNN mới có vai trò quan trọng trong những nội dung tiếp theo của môn học: Đó là "trung bình"

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{2.1}$$

và "phương sai"

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad n \geq 2. \tag{2.2}$$

Chúng thường xuyên xuất hiện trong các bài toán về ước lượng tham số và kiểm định giả thiết (trong Chương 3).

Bộ giá trị $x = (x_1, x_2, \ldots, x_n)$ của các BNN trên gọi là một mẫu ngẫu nhiên cỡ n. Các mẫu ngẫu nhiên này là nguyên liệu quan trọng trong các công cụ

thống kê. Trong những phần tiếp theo, chúng ta sẽ xây dựng các phương pháp thống kê học để đưa ra những quan sát và ước lượng đối với tổng thể dựa trên những mẫu ngẫu nhiên như vậy.

2.1.2 Đặc trưng mẫu

Nhắc lại rằng hai số đặc trưng quan trọng của một mẫu là *trung bình mẫu* (sample mean) và *phương sai mẫu* (sample variance). Trung bình mẫu được định nghĩa là

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), \tag{2.3}$$

trong khi đó, phương sai mẫu s^2 được xác định là

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right). \tag{2.4}$$

Như vậy, nếu xem một mẫu ngẫu nhiên là một dữ liệu, thì trung bình mẫu và phương sai mẫu chính là trung bình và phương sai của các dữ liệu đã nói trong Mục 1.1

Ví dụ 2.1. Quan sát giá gạo ở 5 cửa hàng được chọn ngẫu nhiên tại Hà Nội, người ta thấy rằng giá gạo tháng này đã tăng so với tháng trước lần lượt là 10, 15, 17, 19, và 22 đồng mỗi kg. Tìm phương sai của mẫu của dữ liệu này.

Lời giải

Để tìm phương sai, trước hết ta tìm trung bình mẫu:

$$\bar{x} = \frac{10+15+17+19+22}{5} = 16{,}6.$$

Do đó, phương sai mẫu là

$$s^2 = \frac{1}{3}\left((10-16{,}6)^2 + (15-16{,}6)^2 + (17-16{,}6)^2 + (19-16{,}6)^2 + (22-16{,}6)^2\right).$$

Kết quả thu được là $s^2 = 20{,}3$.

CAS 2.1. Như đã biết trong Chương 1, trong phần mềm thống kê R, trung bình và phương sai có thể được tính bằng `mean` và `var`. Trong ví dụ sau, chúng ta đưa vào các dữ liệu của Ví dụ 2.1 và tính trung bình và phương sai:

```
x = c(10, 15, 17, 19, 22)
print(mean(x))
```

```
## [1] 16,6

print(var(x))

## [1] 20,3
```

Kết quả thu được cho thấy trung bình của mẫu là $\overline{x} = 16{,}6$, trong khi đó phương sai là $s^2 = 20{,}3$.

Bên cạnh các phần mềm máy tính, các máy tính khoa học (Scientific Calculator) cũng có chức năng tính trung bình và phương sai của một tập dữ liệu một cách nhanh chóng và hiệu quả.

2.2 Ước lượng điểm của các tham số

Ước lượng điểm (point estimation) của tham số là một phương pháp trong thống kê để ước tính giá trị cụ thể của một tham số chưa biết của tổng thể, dựa trên dữ liệu mẫu. Giá trị ước lượng điểm là một con số duy nhất được tính toán từ mẫu để đại diện cho tham số tổng thể.

Để nói cụ thể hơn, chúng ta bắt đầu với hai khái niệm cơ bản về vấn đề này, đó là

- *Tham số tổng thể:* Là giá trị cần ước lượng, như trung bình tổng thể (μ), phương sai tổng thể (σ^2), tỉ lệ (p) hay bất kỳ đặc trưng nào khác của tổng thể mà chúng ta quan tâm.
- *Ước lượng điểm:* Là một con số duy nhất được tính toán từ mẫu để đại diện cho tham số tổng thể.

Một cách cụ thể hơn, một ước lượng điểm của một tham số θ tính từ mẫu cỡ n là một hàm số $\hat{\theta} = T_n(x_1, x_2, \ldots, x_n)$, trong đó $(x_1, x_2, \ldots, x_n)$ là giá trị quan sát được từ một véc tơ ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ với X_j là các BNN độc lập và có phân phối của tổng thể. Như vậy, T_n sẽ là một BNN với giá trị phụ thuộc vào mẫu quan sát được.

Để có thể đánh giá các ước lượng điểm của tham số, người ta đưa ra các khái niệm *ước lượng không chệch* (unbias estimator), *ước lượng vững* (weakly consistent estimator), và *ước lượng hiệu quả* (effective estimator).

Định nghĩa 2.1: (Ước lượng không chệch)

Ước lượng $\hat{\theta} = T_n(x_1, \ldots, x_n)$ được gọi là ước lượng không chệch của θ nếu kì vọng của T_n đúng bằng θ, có nghĩa là

$$\mathbb{E}[T_n] = \theta. \tag{2.5}$$

Trong thực tế, việc ước lượng thường được tiến hành dựa trên các mẫu có kích thước khác nhau. Ta kỳ vọng rằng khi kích thước mẫu tăng thì chất lượng của các ước lượng cũng được cải thiện. Do đó, đối với mỗi kích thước mẫu n, ta xét một *ước lượng* T_n, trong đó T_n là một hàm của n biến ngẫu nhiên quan sát được. Nói cách khác, một ước lượng không chỉ là một hàm đơn lẻ, mà được hiểu như một *dãy các hàm số*

$$\{T_n\}_{n\geq 1},$$

trong đó mỗi T_n nhận n giá trị quan sát từ mẫu và cho ra một giá trị số dùng để ước lượng tham số quan tâm.

Định nghĩa 2.2: (Ước lượng vững)

Một ước lượng T_n, $n \geq 1$, được gọi là một ước lượng vững nếu với mọi $\epsilon > 0$,

$$\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) = 1. \tag{2.6}$$

Trong mục nhỏ tiếp theo, chúng ta sẽ có những ví dụ về ước lượng không chệch và ước lượng vững.

2.2.1 Ước lượng điểm của trung bình, phương sai, và tỉ lệ

Hai tham số quan trọng của tổng thể là trung bình và phương sai. Các ước lượng điểm cho trung bình và phương sai được cho như sau.

- *Ước lượng trung bình tổng thể:* Nếu chúng ta muốn ước lượng trung bình tổng thể (μ), một cách ước lượng điểm tự nhiên là sử dụng *trung bình mẫu* ($\bar{x}$):

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

  trong đó x_i là các quan sát trong mẫu, và n là kích thước mẫu.

- *Ước lượng tỉ lệ tổng thể:* Nếu quan tâm đến tỉ lệ thành công trong tổng thể (p), ước lượng điểm của tỉ lệ tổng thể là tỉ lệ mẫu:

$$\hat{p} = \frac{x}{n},$$

  trong đó x là số quan sát thành công trong mẫu và n là tổng số quan sát.

- *Ước lượng phương sai tổng thể:* Nếu cần ước lượng phương sai tổng thể (σ^2), ước lượng điểm của phương sai tổng thể là phương sai mẫu

hiệu chỉnh:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

trong đó $\bar{x}$ là trung bình mẫu.

Các ước lượng trên đều là các ước lượng không chệch cho tham số tổng thể.

Định lí 2.1:

Trung bình mẫu là ước lượng không chệch cho trung bình tổng thể. Hơn nữa, nếu tổng thể có phương sai hữu hạn thì trung bình mẫu là một ước lượng vững của trung bình tổng thể.

Chứng minh. Ta có trung bình mẫu $\overline{x}$ là giá trị quan sát được của BNN $\overline{X}$, với

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{2.7}$$

ở đó, $X_1, X_2, \ldots, X_n$ là các BNN độc lập và có cùng phân phối với tổng thể X. Từ đó, có thể tính được trung bình và phương sai của $\overline{X}$, đó là

$$\mathbb{E}\left[\overline{X}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu$$

do mỗi X_i có cùng trung bình μ của tổng thể.

Nếu X có phương sai $\sigma^2 < \infty$ thì $\overline{X}$ có phương sai $\sigma^2_{\overline{X}} = \sigma^2/n$. Do đó, theo bất đẳng thức Chebyshev, với $\epsilon > 0$ tuỳ ý,

$$P(|\overline{X} - \mu| > \epsilon) \leq \frac{\sigma^2_{\overline{X}}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \quad \text{khi } n \to \infty.$$

Vậy, $\lim_{n\to\infty} P(|\overline{X} - \mu| > \epsilon) = 0$, có nghĩa là (dãy) các trung bình mẫu là một ước lượng vững của trung bình tổng thể. □

Định lí 2.2:

Phương sai mẫu là ước lượng không chệch cho phương sai tổng thể.

Chứng minh. Bài tập! □

Một ưu điểm rõ ràng của ước lượng điểm là nó cung cấp một con số cụ thể, dễ hiểu và trực tiếp để ước lượng tham số tổng thể. Do đó, phương pháp này thường là bước đầu trong quá trình phân tích thống kê. Tuy nhiên, sự đơn giản của các ước lượng điểm cũng đi liền với nhược điểm: Do chỉ đưa ra một giá trị cụ thể, ước lượng điểm không phản ánh được sự không chắc chắn hoặc sai số của quá trình ước lượng: Nó không cho biết mức độ tin cậy của kết quả ước lượng. Để khắc phục nhược điểm này, người ta thường sử dụng ước lượng khoảng, đi kèm với khoảng tin cậy. Đó là những khái niệm sẽ nói trong §2.3.

2.2.2 Sơ lược về phương pháp hợp lí cực đại

Mục này chúng ta tìm hiểu một phương pháp xây dựng ước lượng điểm của tham số, khởi xướng bởi Ronald Fisher (1890 – 1962), một nhà di truyền học và thống kê học người Anh vào đầu thế kỷ 20, gọi là *phương pháp hợp lí cực đại* (tiếng Anh gọi là "maximum likelihood estimation" – MLE).

Giả sử trong một tập nền gồm các sản phẩm có một tỉ trọng p sản phẩm là chính phẩm ($0 \le p \le 1$). Để ước tính p, chúng ta chọn ngẫu nhiên một mẫu có n sản phẩm và quan sát xem chúng là phế phẩm hay chính phẩm. Như vậy, với một mẫu nhất định, tỉ lệ p được ước tính như thế nào?

Hãy bắt đầu với $n = 1$. Giả sử chúng ta chọn ngẫu nhiên một sản phẩm và quan sát xem nó là phế phẩm hay chính phẩm. Kết quả thu được sẽ rơi vào một trong hai trường hợp:

- Sản phẩm đã chọn là chính phẩm. Xác suất để có kết quả là này p, lớn nhất khi $p = 1$. Lúc này, chúng ta ước lượng $p = 1$.
- Sản phẩm đã chọn là phế phẩm. Xác suất có kết quả này là $1-p$, lớn nhất khi $p = 0$. Lúc này chúng ta ước lượng $p = 0$.

Như vậy, với một mẫu cỡ $n = 1$, chỉ có thể ước lượng tỉ lệ chính phẩm là 0 hoặc 1.

Tiếp theo, hãy xem vấn đề như thế nào nếu $n = 2$. Lúc này, chúng ta có 3 trường hợp:

- Cả hai sản phẩm đã chọn là chính phẩm. Xác suất để có kết quả là này p^2, lớn nhất khi $p = 1$. Lúc này, chúng ta ước lượng $p = 1$.
- Có đúng một trong hai sản phẩm là chính phẩm. Xác suất để có kết quả này là $p(1-p)$, lớn nhất $p = 1/2$. Lúc này, chúng ta ước lượng $p = 1/2$.
- Cả hai sản phẩm đã chọn là phế phẩm. Xác suất để có kết quả này là $(1-p)^2$, lớn nhất khi $p = 0$. Lúc này chúng ta ước lượng $p = 0$.

Đó chính là nguyên lí của phương pháp hợp lí cực đại trong đó, người ta ước tính giá trị của tham số (trong ví dụ trên là p) là giá trị sao cho xác suất để thu được một mẫu đã có từ thực nghiệm là lớn nhất.

Tổng quát hơn, xét tình huống mẫu cỡ n tuỳ ý. Gọi X là BNN cho giá trị 1 nếu kết quả của phép thử là một sản phẩm tốt, và 0 nếu kết quả là một phế phẩm. Khi đó, X nhận hai giá trị 0 và 1 và hàm xác suất của nó là

$$f(x)=\begin{cases} p & \text{nếu } x=1 \\ 1-p & \text{nếu } x=0 \end{cases}$$

Rõ ràng, ta có thể viết $f(x)=p^x(1-p)^{1-x}$.

Lấy ngẫu nhiên từ tập nền n lần, ta thu được một mẫu cỡ n, kí hiệu $m=\{x_1,x_2,\ldots,x_n\}$. Khi đó, xác suất để thu được một mẫu như vậy là

$$P(m)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

Với các x_i đã biết, vế phải của đẳng thức trên là một hàm số của p, kí hiệu là $\ell(x_i;p)$:

$$\ell(x_i;p)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}. \tag{2.8}$$

Trong phương pháp hợp lí cực đại, người ta tìm giá trị của $\hat{p}$ sao cho ℓ đạt cực đại khi $p=\hat{p}$. Cụ thể là

$$\hat{p}=\arg\max \ell(x_i;p). \tag{2.9}$$

Lúc này, ℓ là một tích của nhiều nhân tử. Để đơn giản, chúng ta lấy logarith (để đưa tích về tổng) và xét bài toán tìm p sao cho $\ln(\ell(p))$ đạt cực đại. Từ đây, theo định lí Fermat (xem [3, Chương II, §2]), chúng ta cần tìm $\hat{p}$ sao cho

$$\left.\frac{\partial \ln \ell(x_i;p)}{\partial p}\right|_{p=\hat{p}}=0. \tag{2.10}$$

Ta có

$$\ln \ell(x_i;p)=\sum_{i=1}^{n}\Big(x_i\ln(p)+(1-x_i)\ln(1-p)\Big).$$

Do đó,

$$\begin{aligned}\left.\frac{\partial \ln \ell(x_i;p)}{\partial p}\right|_{p=\hat{p}} &= \sum_{i=1}^{n}\left(\frac{x_i}{p}-\frac{1-x_i}{1-p}\right)\Bigg|_{p=\hat{p}} \\ &= n\left(\frac{\bar{x}}{p}-\frac{1-\bar{x}}{1-p}\right)\Bigg|_{p=\hat{p}} \\ &= \frac{n(\bar{x}-\hat{p})}{\hat{p}(1-\hat{p})}.\end{aligned}$$

Từ đó, phương pháp này cho ta ước lượng

$$\hat{p} = \bar{x}.$$

Ví dụ 2.2. Trong một cuộc khảo sát với 520 người trong thành phố, người ta thấy 314 người trong số họ ủng hộ chính sách mới. Dựa trên mẫu này, ước lượng tỉ lệ ủng hộ chính sách trong toàn bộ dân số thành phố.

Lời giải

Trong ví dụ này, chúng ta lấy ngẫu nhiên một người trong thành phố và kết quả sẽ là "thành 'công" nếu người được chọn ủng hộ chính sách mới và kết quả sẽ là "thất bại" nếu người được chọn không ủng hộ nó. Với mỗi thành công, ta lấy $x_i = 1$ và với mỗi thất bại ta lấy $x_i = 0$. Khi đó, ta thu được một mẫu $x = (x_1, x_2, \ldots, x_{520})$, trong đó, có 314 chỉ số i với $x_i = 1$, còn 206 chỉ số còn lại $x_i = 0$. Ước lượng điểm của tỉ lệ tổng thể p (tỉ lệ người ủng hộ trong toàn bộ dân số) là

$$\hat{p} = \bar{x} = \frac{1}{520}\sum_{i=1}^{520} x_i = \frac{314}{520} = 0{,}6038.$$

Từ đó, chúng ta đưa ra kết luận sau: Có 60,38 % người dân trong toàn bộ thành phố ủng hộ chính sách mới, dựa trên mẫu khảo sát.

Trường hợp BNN liên tục

Bây giờ, xét một BNN liên tục X với hàm mật độ xác suất $f(x;\theta)$, với θ là một tham số (hoặc một vectơ các tham số). Giả sử $x_1, x_2, \ldots, x_n$ là một mẫu. Chúng ta biết rằng, vì X là BNN liên tục, nên xác suất để thu được một mẫu như vậy luôn bằng 0. Phương pháp hợp lí cực đại, trong trường hợp này, không tìm cách xác định tham số để xác suất thu được mẫu đó là cực đại mà tìm tham số sao cho *hàm mật độ đồng thời* của n biến ngẫu nhiên độc lập cùng phân phối $f(x,\theta)$ đạt cực đại tại x.

Ví dụ 2.3. Giả sử có một tập dữ liệu gồm chiều cao của 5 người với các giá trị sau (theo đơn vị cm):

$$X = \{160, 170, 165, 175, 180\}.$$

Hơn nữa, dữ liệu này được lấy ra từ một tổng thể tuân theo luật phân phối chuẩn với hàm mật độ xác suất được cho bởi:

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

Đó là hàm mật độ của $\mathcal{N}(\mu, \sigma^2)$ với trung bình μ và phương sai σ^2. Mục tiêu của chúng ta là ước lượng các giá trị của μ và σ.

Các bước thực hiện như sau:

- Thiết lập hàm hợp lí: Hàm hợp lí cho tập dữ liệu này là tích của các mật độ xác suất riêng lẻ của mỗi quan sát:

$$\ell(\mu,\sigma^2 \mid X) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

  Trong đó, $x_1, x_2, \ldots, x_5$ là các giá trị quan sát được.

- Hàm log-hợp lí. Để đơn giản hóa quá trình tính toán, chúng ta tính log của hàm hợp lí (log-likelihood), nhằm đưa tích về tổng:

$$\ln \ell(\mu,\sigma^2 \mid X) = \sum_{i=1}^{5} \left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \right),$$

  và tách các thành phần:

$$\ln \ell(\mu,\sigma^2|X) = -\frac{5}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{5}(x_i-\mu)^2.$$

- Ước lượng giá trị trung bình μ: Để tối ưu hoá hàm log-hợp lí, chúng ta lấy đạo hàm của nó theo μ:

$$\frac{\partial}{\partial \mu} \ln \ell(\mu,\sigma^2|X) = \frac{1}{\sigma^2}\sum_{i=1}^{5}(x_i-\mu).$$

- Đặt đạo hàm bằng 0 và giải phương trình:

$$\sum_{i=1}^{5}(x_i-\mu) = 0,$$

  chúng ta thu được

$$\mu = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{160+170+165+175+180}{5} = 170.$$

  Do đó, ước lượng của μ là 170 cm.

- Ước lượng phương sai σ^2: Tương tự, lấy đạo hàm của log-hợp lí theo σ^2 và đặt nó bằng 0:

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu,\sigma^2|X) = -\frac{5}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{5}(x_i-\mu)^2 = 0,$$

và do đó,

$$\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \mu)^2.$$

Tính toán cụ thể hơn,

$$\sum_{i=1}^{5}(x_i - \mu)^2 = (160-170)^2+(170-170)^2+(165-170)^2+(175-170)^2+(180-170)^2.$$

$$= 100+0+25+25+100 = 250.$$

Như vậy,

$$\sigma^2 = \frac{250}{5} = 50.$$

- Thông qua phương pháp hợp lí cực đại, ước lượng giá trị trung bình μ là 170 cm và phương sai σ^2 là 50 cm^2.

Lưu ý, các tính toán trên cho thấy $\mu = 170$ và $\sigma^2 = 50$ là điểm tới hạn duy nhất của hàm log-hợp lí và nó cũng là cực tiểu toàn cục duy nhất.

Nhận xét 2.1. Có thể tiếp tục phân tích ví dụ trên một cách tổng quát để thấy rằng nếu X tuân theo phân phối chuẩn, thì ước lượng hợp lí cực đại của trung bình và phương sai của X là

$$\hat{\mu}_X = \overline{x}, \quad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

trong đó $x = (x_1, x_2, \ldots, x_n)$ là mẫu dùng để ước lượng. Đáng chú ý là ước lượng MLE của phương sai không phải là ước lượng không chệch.

Ví dụ 2.4. Một BNN liên tục X có hàm mật độ xác suất có dạng $f(x) = ve^{-vx}$ nếu $x \geq 0$ và $f(x) = 0$ nếu $x < 0$, trong đó v là một tham số. Sử dụng phương pháp hợp lí cực đại, hãy ước lượng μ_X từ một mẫu ngẫu nhiên $x = (0{,}4,\ 0{,}8,\ 0{,}9,\ 1{,}1)$.

Lời giải

Trước hết, nhận xét rằng hàm $f(x)$ là một hàm mật độ xác suất với mọi giá trị của v. Thật vậy, $f(x) \geq 0$ và

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{0}^{\infty} ve^{-vx}\,\mathrm{d}x = -e^{-vx}\Big|_{x=0}^{\infty} = 1.$$

Tiếp theo, chúng ta tính trung bình của X theo ν. Ta có

$$\begin{aligned}\mu_X = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x &= \int_0^{\infty} \nu x e^{-\nu x}\,\mathrm{d}x \\ &= -\frac{e^{-\nu x}(\nu x+1)}{\nu}\Big|_{x=0}^{\infty} \\ &= \frac{1}{\nu}.\end{aligned}$$

Như vậy, nếu biết ước lượng của ν thì có ước lượng cho μ_X.

Bây giờ, xét 4 BNN liên tục độc lập X_i, $i = 1,\ldots,4$ có cùng hàm mật độ $f(x)$. Khi đó, hàm mật độ xác suất đồng thời cho bởi

$$g(x_1, x_2, x_3, x_4; \nu) = f(x_1)f(x_2)f(x_3)f(x_4).$$

Chúng ta tìm tham số ν sao cho hàm g đạt cực đại tại giá trị mẫu đã cho. Từ đó, theo định lí Fermat (như trên)

$$\frac{\partial \log(g(x_1, x_2, x_3, x_4))}{\partial \nu} = 0 \tag{2.11}$$

tại mẫu đã cho.

Vế trái của (2.11) có dạng

$$\sum_{i=1}^{4} \frac{\partial \log f(x_i)}{\partial \nu} = \sum_{i=1}^{4} \frac{\partial \log(\nu) - x_i \nu}{\partial \nu} = \sum_{i=1}^{4}\left(\frac{1}{\nu} - x_i\right) = \frac{4}{\nu} - \sum_{i=1}^{4} x_i.$$

Thay vào phương trình (2.11), chúng ta thu được

$$\nu = \frac{4}{\sum_{i=1}^{4} x_i}.$$

Với mẫu x_i như đã cho, chúng ta thu được điểm tới hạn của hàm g, xem như hàm của một biến ν, là $\nu = 1{,}25$. Giá trị này thực sự là cực đại của g vì đạo hàm cấp hai $\partial^2 g/\partial \nu^2(\nu) = -4/\nu^2 < 0$.

Kết luận: $\hat{\mu} = 1/\nu = 0{,}8$ là ước lượng điểm của trung bình theo phương pháp hợp lí cực đại.

Bài tập §2.2

1. Ước lượng điểm cho các tham số tổng thể là gì? Nêu ưu và nhược điểm của ước lượng điểm.
2. Nêu công thức của ước lượng điểm cho trung bình và phương sai của một tổng thể dựa trên một mẫu dữ liệu.
3. Bằng phương pháp hợp lí cực đại, hãy ước lượng trung bình của một BNN X có phân phối chuẩn dựa vào một mẫu ngẫu nhiên

x	0–2	2–4	4–6	6–8	8–10
Tần số	4	11	16	18	7

2.3 Khoảng tin cậy của tham số

Ước lượng khoảng (interval estimation) là một phương pháp trong thống kê dùng để ước lượng một tham số chưa biết của tổng thể bằng cách xác định một khoảng giá trị thay vì chỉ một con số cụ thể như trong bài toán ước lượng điểm. Khoảng này được gọi là *khoảng tin cậy* (confidence interval) và chứa giá trị thực của tham số đó với một "mức độ tin cậy" nhất định.

Mục tiêu của chúng ta sau đây là một số phương pháp xây dựng các khoảng tin cậy cho các tham số của tổng thể từ các mẫu ngẫu nhiên, bao gồm khoảng tin cậy cho trung bình của phân phối chuẩn trong hai tình huống phương sai của tổng thể đã biết và chưa biết.

2.3.1 Khoảng tin cậy của μ của phân phối chuẩn khi đã biết σ^2

Trong tiểu mục này, ta xây dựng phương pháp ước lượng khoảng cho trung bình của một tổng thể tuân theo phân phối chuẩn với phương sai đã biết. Lưu ý rằng mặc dù tình huống này ít gặp trong thực tế nhưng chúng ta vẫn đề cập đến nó ở đây nhằm minh họa một quy trình xây dựng khoảng tin cậy.

Cơ sở lí thuyết của quy trình xây dựng khoảng tin cậy nêu sau đây là định lí về tổng của các biến số ngẫu nhiên độc lập có phân phối chuẩn.

Định lí 2.3:

Giả sử $X_1, \ldots, X_n$ là các BNN độc lập có phân phối chuẩn với trung bình μ và phương sai σ^2. Khi đó,

(i) Tổng $X_1 + \cdots + X_n$ là BNN có phân phối chuẩn với trung bình $n\mu$ và phương sai $n\sigma^2$.

(ii) Biến ngẫu nhiên
$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n),$$
có phân phối chuẩn với trung bình μ và phương sai σ^2/n.

(iii) BNN được định nghĩa như sau:
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$
có phân phối chuẩn với trung bình 0 và phương sai 1.

Chú ý 2.1. Giá trị trung bình và phương sai của BNN $\overline{X}$ có thể được suy ra

từ các kết quả của Mục §1.10 về trung bình và phương sai của tổ hợp tuyến tính của các BNN độc lập. Tuy nhiên, lưu ý rằng định lí trên còn nói rằng BNN $\overline{X}$ có phân phối chuẩn, giống như các BNN X_i.

Trước khi nói về vấn đề chính ở đây là bài toán xây dựng các khoảng tin cậy, ta hãy minh họa một ứng dụng khá thú vị của định lí trên trong hai ví dụ sau đây.

Ví dụ 2.5. Trứng gà thu hoạch từ một trang trại A có khối lượng tuân theo phân phối chuẩn với trung bình $\mu = 45$ gram và độ lệch chuẩn $\sigma = 6{,}5$ gram[1]. Chọn ngẫu nhiên 12 quả trứng để đóng hộp. Tìm xác suất để hộp 12 quả trứng có tổng khối lượng bé hơn 520 gram.

Lời giải

Nếu X_i là khối lượng của quả trứng được chọn trong lần thứ i, với $i = 1,2,\ldots,12$, thì tổng khối lượng của 12 quả trứng là

$$X = X_1 + X_2 + \cdots + X_{12}.$$

Khi đó, X là tổng của 12 biến ngẫu nhiên độc lập có cùng một phân phối chuẩn $\mathcal{N}(\mu = 45, \sigma = 6{,}5)$. Do đó, từ Phần (i) của Định lí 2.3, ta suy ra X là một BNN với trung bình $\mu_X = 12 \times 45 = 540$ và phương sai là $\sigma_X^2 = 12 \times 6{,}5^2 \approx 507$. Từ đó, xác suất để hộp 12 quả trứng gà có tổng khối lượng bé hơn 520 gram là

$$\begin{aligned} P(X < 520) = F(520) &= \Phi\left(\frac{520-540}{\sqrt{507}}\right) \\ &\approx \Phi(-0{,}8882) \\ &\approx 0{,}1872, \end{aligned}$$

trong đó F là hàm xác suất tích luỹ của X và Φ, như thường lệ, là hàm Laplace.

Ví dụ 2.6. Kết quả phân tích một hormone trên một nhóm bệnh nhân cho thấy chúng có giá trị trung bình là 52 và độ lệch chuẩn là 12. Giả sử các giá trị này tuân theo phân phối chuẩn. Chọn ngẫu nhiên 10 người và tính giá trị trung bình của mức hormone của họ. Xác định xác suất để thu được kết quả nằm trong khoảng 45 đến 55.

Lời giải

[1] Với giả thiết khối lượng X của trứng gà tuân theo luật phân phối chuẩn, $X \sim \mathcal{N}(\mu = 45, \sigma = 6{,}5)$, thì mô hình của chúng ta sẽ cho một xác suất rất nhỏ cho những sự kiện không xảy ra trong thực tế (như sự kiện $(X < 0)$). Trong mô hình, chúng ta tính được $P(X < 0) \approx 2{,}2 \times 10^{-12}$. Thậm chí, sẽ không thể xảy ra sự kiện một quả trứng có khối lượng quá bé. Ví dụ, xác suất một quả trứng có khối lượng ít hơn 10 gram tính được theo mô hình là $P(X < 10) \approx 3{,}63 \times 10^{-8}$ cũng là một số rất bé.

Theo Định lí 2.3, nếu $\overline{X}$ là giá trị trung bình của mức hormone của 10 bệnh nhân được chọn ngẫu nhiên thì $\overline{X}$ là BNN có phân phối chuẩn với trung bình 52 và độ lệch chuẩn $12/\sqrt{10}$. Từ đó, xác suất để giá trị trung bình này nằm trong khoảng 45 đến 55 là

$$\begin{aligned} P(45 < X \leq 55) &= F(55) - F(45) \\ &= \Phi\left(\frac{55-52}{12/\sqrt{10}}\right) - \Phi\left(\frac{45-52}{12/\sqrt{10}}\right) \\ &= \Phi(0{,}791) - \Phi(-1{,}844) \\ &\approx 0{,}753, \end{aligned}$$

trong đó F là hàm xác suất tích luỹ của $\overline{X}$ và Φ là hàm Laplace. Như vậy, nếu chúng ta lấy mẫu rất nhiều lần, mỗi lần 10 bệnh nhân, thì khoảng 75,3% các lần lấy mẫu đó cho giá trị trung bình nằm trong khoảng từ 45 đến 55.

Phân vị mức của phân phối chuẩn

Trở lại với vấn đề xây dựng khoảng tin cậy, chúng ta sẽ sử dụng Định lí 2.3 để xác định khoảng tin cậy cho trung bình μ của phân phối chuẩn khi biết phương sai σ^2. Trước tiên, chúng ta đưa ra định nghĩa về *giá trị tới hạn mức α (phân vị mức α)* của phân phối chuẩn tắc.

Định nghĩa 2.3: (Phân vị mức)

Giả sử $X \sim \mathcal{N}(0,1)$ và $\alpha \in (0,1)$. Giá trị U_α được gọi là *phân vị mức α* hay *giá trị tới hạn mức α* của X nếu

$$P(X \leq U_\alpha) = 1 - \alpha. \tag{2.12}$$

Chú ý 2.2. Do $X \sim \mathcal{N}(0,1)$, nên điều kiện (2.12) tương đương với

$$\Phi(U_\alpha) = 1 - \alpha. \tag{2.13}$$

ở đó Φ là hàm Laplace:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du.$$

Như đã biết trong Chương 1, các giá trị hàm Laplace và hàm ngược của nó có thể được tính bằng phần mềm máy vi tính, máy tính khoa học cầm tay (như máy tính Casio fx 580VNX), hoặc tra bảng tính sẵn (Bảng A7 trong Kreyszig [1]).

CAS 2.2. Trong R, giá trị hàm Laplace được cho bởi hàm `pnorm()` trong khi đó hàm ngược của nó là `qnorm()`. Ví dụ, đoạn code R dưới đây cho mức phân vị 5% của phân phối chuẩn.

```
qnorm(0.95)

## [1] 1,645
```

Đối với máy tính khoa học cầm tay Casio fx580 VNX, tính toán như trong hướng dẫn của nhà sản xuất.

Các bước xác định khoảng tin cậy cho μ

Các bước xác định khoảng tin cậy cho trung bình của một tổng thể theo phân phối chuẩn với độ lệch chuẩn $\sigma > 0$ đã biết, dựa trên một mẫu $x_1, \ldots, x_n$ thu thập được từ một tổng thể có phân phối chuẩn như sau:

1. Xác định độ tin cậy γ đối với ước lượng cần tính.
2. Xác định giá trị tới hạn $U_{\alpha/2}$ tương ứng với mức $\alpha/2 = (1-\gamma)/2$, tức là phân vị mức $\alpha/2$ của phân phối chuẩn tắc. Trong bước này cần tính các giá trị của hàm Laplace ngược. (Bảng A7 trong Kreyszig [1]).
3. Tính trung bình mẫu $\bar{x}$ từ mẫu x_1, x_2, …, x_n đã có.
4. Tính "bán kính" của khoảng ước lượng,
$$k = \frac{U_{\alpha/2}\sigma}{\sqrt{n}}. \tag{2.14}$$
5. Kết luận khoảng tin cậy cho μ với độ tin cậy γ là
$$\bar{x} - k \le \mu \le \bar{x} + k. \tag{2.15}$$

Khi đó, nếu lấy mẫu lặp đi lặp lại nhiều lần từ tổng thể thì $100\gamma\%$ của tất cả các khoảng ước lượng xây dựng theo các bước trên sẽ chứa trung bình μ của tổng thể.

CAS 2.3. Việc tính phân vị mức $U_{\alpha/2}$ có thể được thực hiện bằng máy tính. Đoạn mã R sau đây tính phân vị mức đối với một số giá trị của độ tin cậy γ, sử dụng hàm `qnorm()`. Kết quả thu được sẽ là dãy phân vị mức của phân phối chuẩn của giá trị đó.

```r
# Các giá trị độ tin cậy
gamma <- c(0.90, 0.95, 0.99)

# Tính alpha
alpha <- 1 - gamma

#Tính phân vị mức alpha/2
result <- qnorm(1 - 0.5 * alpha)

# In ra kết quả
print(result)

## [1] 1,645 1,960 2,576
```

Kết quả thu được là các phân vị mức $U_{\alpha/2}$ tương ứng với các giá trị của độ tin cậy γ như bảng sau:

γ	90%	95%	99%
$\alpha/2$	5%	2,5%	0,5%
$U_{\alpha/2}$	1,645	1,960	2,576

Tiếp theo, chúng ta chứng minh khẳng định trên.

Chứng minh. Cho $X_1, \cdots, X_n$ là n biến ngẫu nhiên độc lập có phân phối chuẩn và có chung kì vọng μ và phương sai σ^2. Khi đó, chúng ta đã biết rằng

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

là một BNN có phân phối chuẩn với trung bình μ và phương sai σ^2/n, trong khi đó, BNN chuẩn hoá

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{2.16}$$

có phân phối chuẩn tắc: $Z \sim \mathcal{N}(0,1)$.

Chúng ta cần chỉ ra có xác suất γ xảy ra sự kiện khoảng $(\overline{x} - k, \overline{x} + k)$ chứa giá trị trung bình μ, có nghĩa là

$$P(\overline{X} - k \leq \mu \leq \overline{X} + k) = \gamma.$$

Thật vậy, ta có

$$\begin{aligned} P(\overline{X}-k\le\mu\le\overline{X}+k) &= P\left(-\frac{k\sqrt{n}}{\sigma}\le\frac{(\overline{X}-\mu)\sqrt{n}}{\sigma}\le\frac{k\sqrt{n}}{\sigma}\right)\\ &= P(-U_{\alpha/2}\le Z\le U_{\alpha/2})\\ &= \Phi(U_{\alpha/2})-\Phi(-U_{\alpha/2})\\ &= 2\Phi(U_{\alpha/2})-1 \quad (\text{bởi vì } \Phi(-U_{\alpha/2})=1-\Phi(U_{\alpha/2}))\\ &= 2(1-\alpha/2)-1\\ &= \gamma. \end{aligned}$$

Đó là điều phải chứng minh. □

Ví dụ 2.7. Một nhà khoa học vì muốn ước lượng trung bình của một loại enzyme trong máu của một nhóm tình nguyện viên nên lấy một mẫu của 24 người và tính được trung bình của mẫu là $\bar{x}=25$. Giả thiết rằng mức enzyme trong máu là một biến ngẫu nhiên tuân theo luật phân phối chuẩn với độ lệch chuẩn $\sigma=4$. Với độ tin cậy $\gamma=95\%$, hãy ước lượng trung bình mức enzyme của nhóm các tình nguyện viên.

Lời giải

Với $\gamma=95\%$, chúng ta xác định được $\alpha=0{,}05$. Giá trị $U_{\alpha/2}=1{,}96$ được xác định từ điều kiện

$$\Phi(U_{\alpha/2})=1-\frac{\alpha}{2}=0{,}975$$

(Bảng A7 trong Kreyszig [1]).

Đã biết trung bình mẫu là $\bar{x}=25$ với số mẫu $n=24$, và do đó

$$k=\frac{U_{\alpha/2}\sigma}{\sqrt{24}}=\frac{1{,}96\times 4}{\sqrt{24}}\approx 1{,}6.$$

Từ đó, kết luận được, với độ tin cậy 95%, trung bình của mức enzyme trên nằm trong khoảng $23{,}4<\mu<26{,}6$.

Ví dụ 2.8. Tìm khoảng tin cậy 90% của trung bình của một phân phối chuẩn dựa trên dữ liệu sau: 39, 51, 49, 43, 57, 59, biết độ lệch chuẩn là $\sigma=4$.

Lời giải

Với $\gamma=90\%$, chúng ta xác định được $\alpha/2=(1-0{,}9)/2=0{,}05$ và $U_{\alpha/2}=1{,}64$.

Trung bình mẫu $\overline{x}\approx 49{,}67$, trong khi đó số mẫu dữ liệu là $n=6$. Khoảng tin cậy 90% là $(\overline{x}-k,\overline{x}+k)$, với

$$k=\frac{U_{\alpha/2}\sigma}{\sqrt{n}}=\frac{1{,}645\times 4}{\sqrt{6}}=2{,}69.$$

Tóm lại, khoảng cần tính là $(49{,}67-2{,}68;\ 49{,}67+2{,}68)=(46{,}98;\ 52{,}36)$.

CAS 2.4. Trong phần mềm R, chúng ta có thể thu được khoảng tin cậy từ hàm `z.test()` trong gói *TeachingDemos*[1]. Đối với ví dụ trên, ta bắt đầu nhập dữ liệu vào một vectơ, ở đây kí hiệu là `data` và gọi hàm `z.test(data, sd = 4)` (tham số độ lệch chuẩn (sd) bằng 4).

```
# Tuỳ chọn hiển thị số chữ số có nghĩa = 6
options(digits = 6)

# Khai báo gói TeachingDemos
library(TeachingDemos)

# Khai báo dữ liệu mẫu
data = c(39, 51, 49, 43, 57, 59)

# Thực hiện hàm z.test() cho data với độ tin cậy 90%
result = z.test(data, sd = 4, conf.level = 0.9)

# In ra kết quả
print(result$conf.int)

## [1] 46,9806 52,3527
## attr(,"conf.level")
## [1] 0,9
```

Lưu ý, kết quả từ máy tính và kết quả tính bằng tay có một sự khác biệt nhỏ sinh ra từ sai số làm tròn.

CAS 2.5. Trong CAS này, chúng ta sẽ tính khoảng tin cậy cho trung bình từ các dữ liệu tạo ra một cách ngẫu nhiên. Chúng ta sẽ tạo ra 1000 dữ liệu mô phỏng mẫu ngẫu nhiên lấy từ một phân phối chuẩn tắc. Trong R, việc này được thực hiện bằng `rnorm()`. Sau đó, chúng ta kiểm tra xem có bao nhiêu khoảng tin cậy được xây dựng thực sự chứa trung bình $\mu=0$ của tổng thể.

```
library(TeachingDemos)
# Số các mẫu được tạo
N = 1000; k = 0

for (i in 1:N)  {

    # Tạo một mẫu ngẫu nhiên cỡ n = 30
    sample = rnorm(30)
```

[1] TeachingDemos: Demonstrations for Teaching and Learning, DOI: `https://doi.org/10.32614/CRAN.package.TeachingDemos`

```
    # Xây dựng KTC từ mẫu ngẫu nhiên đã được tạo
    test = z.test(sample, sd = 1, conf.level = 0.90)

    # Đếm số các KTC thực sự chứa trung bình tổng thể
    if (test$conf.int[1] * test$conf.int[2] < 0) {
        k = k + 1
    }
}

# Hiển thị tỉ lệ KTC thực sự chứa trung bình tổng thể
print(k/N)

## [1] 0,921
```

Kết quả mô phỏng ghi nhận có 92,1% các khoảng tin cậy xây dựng bằng `z.test()` thực sự chứa trung bình tổng thể $\mu = 0$, xấp xỉ bằng độ tin cậy 90% (cho bằng tham số `conf.level = 0.90`).

Chú ý 2.3. Với một mẫu đã có, độ dài các khoảng tin cậy tăng khi độ tin cậy tăng, có nghĩa là việc tăng độ tin cậy dẫn đến việc giảm độ chính xác. Mặt khác, do độ lệch chuẩn σ của tổng thể đã biết, việc tăng cỡ mẫu sẽ thu hẹp khoảng tin cậy.

2.3.2 *Khoảng tin cậy của μ của phân phối chuẩn khi chưa biết σ^2*

Trong phần trước, chúng ta biết ước lượng μ khi phương sai σ^2 đã biết. Tuy nhiên trong thực tế, phương sai σ^2 của một tổng thể thường không biết được, vì thế sẽ cần tìm hiểu phương pháp ước lượng trung bình μ cho trường hợp phương sai σ^2 *chưa biết*. Cơ sở toán học của phương pháp này là định lí về *phân phối* t[1]

[1]Phân phối t còn được gọi là phân phối Student, theo bút danh của William Sealy Gosset (13/6/1876–16/10/1937), một nhà thống kê và hoá học người Anh. Phân phối t xuất hiện lần đầu trên một bài báo của ông, được đăng trên tạp chí Biometrica vào năm 1907.

Định nghĩa 2.4: (Phân phối t)

Phân phối xác suất có hàm mật độ cho bởi

$$h(t)=K_{n-1}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}, \tag{2.17}$$

với hằng số K_{n-1} xác định như sau: Với tham số m nguyên thì

$$K_m=\frac{\Gamma(m/2+1/2)}{\sqrt{m\pi}\Gamma(m/2)}, \tag{2.18}$$

trong đó $\Gamma(\alpha)$ là hàm Gamma, gọi là phân phối t với $(n-1)$ bậc tự do.

Nhận xét 2.2. Vai trò của hằng số K_m là để tích phân hàm mật độ $h(t)$ trên toàn đường thẳng thực bằng đơn vị, tức là $\int_{-\infty}^{\infty} h(t)dt=1$.

Công thức của K_m khá phức tạp, nhưng điều này không quá quan trọng trong ứng dụng, vì các giá trị của hàm $h(t)$ cũng như hàm xác suất tích luỹ tương ứng có thể được tính bằng máy tính hoặc tra bảng tính sẵn (Kreyszig [1, Bảng A9] hoặc Đ. H. Thắng [2, Bảng 2]).

Một số quan sát ban đầu về phân phối t như sau:

- Đồ thị hàm mật độ của phân phối t với $(n-1)$ bậc tự do cũng có dạng "hình chuông" (bell-shaped), tương tự phân phối chuẩn.
- Hàm phân phối xác suất $F(z)$ được cho bởi công thức (2.17) có tính chất đối xứng:
$$F(-z)=1-F(z)$$
với mọi $z\in\mathbb{R}$.
- Giá trị t_α được gọi là *giá trị tới hạn mức* α $(0<\alpha<1)$ của phân phối t với $(n-1)$ bậc tự do nếu
$$F(t_\alpha)=1-\alpha, \tag{2.19}$$
với F là hàm phân phối xác suất tích luỹ, tức là
$$F(z)=\int_{-\infty}^{z} h(t)\,\mathrm{d}t, \tag{2.20}$$
trong đó $h(t)$ là hàm mật độ cho bởi (2.17). Điều này tất nhiên có nghĩa là nếu X là một BNN có phân phối t với $m=(n-1)$ bậc tự do thì
$$P(X<t_\alpha)=1-\alpha.$$

Tại sao phân phối t lại quan trọng trong bài toán khoảng tin cậy. Đó là vì Định lí 2.4 nêu ngay dưới đây. Trước hết, nhắc lại rằng, nếu $X_1, X_2, \ldots, X_n$ là một dãy các BNN thì $\overline{X}$ và S^2 được xác định lần lượt bởi

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n), \tag{2.21}$$

và

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{2.22}$$

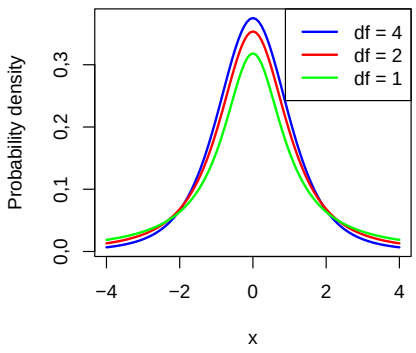
Khi đó, chúng ta có kết quả sau:

Định lí 2.4:

Giả sử $X_1, \cdots, X_n$, $n \geq 2$, là các biến ngẫu nhiên độc lập có phân phối chuẩn, có cùng kì vọng μ và phương sai σ^2. Khi đó, biến ngẫu nhiên

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}, \tag{2.23}$$

có phân phối t với $n-1$ bậc tự do.



Hình 2.1: Đồ thị của hàm mật độ của phân phối t với 4, 2, và 1 bậc tự do

Lưu ý rằng, khác với công thức của BNN Z trong (2.16), công thức xác định T (công thức (2.23)) là một công thức không chứa độ lệch chuẩn σ của tổng thể. Do đó, người ta dùng T để xây dựng khoảng tin cậy của trung bình μ khi σ *chưa biết*. Tương tự như bài toán ước lượng khoảng trung bình phân phối chuẩn khi đã biết phương sai của tổng thể, chúng ta thường phải tính t_α khi biết α trong bài toán ước lượng nêu dưới đây.

Khoảng tin cậy

Nhờ định lí 2.4, người ta xây dựng phương pháp xác định khoảng tin cậy cho μ của phân phối chuẩn khi chưa biết phương sai σ^2. Các bước xây dựng khoảng tin cậy trong trường hợp này tương tự như trường hợp σ đã biết, mô tả trong Mục 2.3.1. Điểm khác biệt ở đây là độ lệch chuẩn σ

của tổng thể được thay bằng độ lệch chuẩn s của mẫu và phân vị mức $U_{\alpha/2}$ trong phân phối chuẩn bởi phân vị mức $t_{\alpha/2}$ của phân phối t với $m = n-1$ bậc tự do, trong đó, n là cỡ mẫu.

Chi tiết các bước như sau:

1. Xác định một độ tin cậy γ cần tính.

2. Xác định giá trị tới hạn $t_{\alpha/2}$ ứng với mức $\alpha/2 = (1-\gamma)/2$ từ phương trình sau:
$$F(t_{\alpha/2}) = 1 - \frac{\alpha}{2} = \frac{1}{2}(1+\gamma), \tag{2.24}$$
với $F(z)$ là hàm phân phối tích luỹ của phân phối t với $m = n-1$ bậc tự do, cho bởi (2.20). Thông thường, để xác định được giá trị tới hạn này, chúng ta cần dùng bảng tính sẵn (Bảng A9 của Kreyszig [1]) hoặc máy tính.

3. Tính trung bình mẫu $\bar{x}$ và phương sai mẫu hiệu chỉnh s^2 từ mẫu đã cho.

4. Tính "bán kính" của khoảng ước lượng
$$k = \frac{t_{\alpha/2}\, s}{\sqrt{n}}.$$

5. Kết luận: Khoảng ước lượng cho μ với độ tin cậy γ được xác định như sau
$$\bar{x} - k \le \mu \le \bar{x} + k. \tag{2.25}$$

Khi đó, nếu lấy mẫu lặp đi lặp lại nhiều lần từ tổng thể, có khoảng $100\gamma\%$ của tất cả các khoảng ước lượng xây dựng theo các bước trên sẽ chứa trung bình μ của tổng thể.

Chứng minh. Theo định lí 2.4, biến ngẫu nhiên
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$
có phân phối t với $(n-1)$ bậc tự do.

Chúng ta cần chỉ ra có xác suất γ xảy ra sự kiện khoảng $(\overline{x} - k, \overline{x} + k)$ chứa giá trị trung bình μ, có nghĩa là
$$P(\overline{X} - k \le \mu \le \overline{X} + k) = \gamma.$$

Thật vậy, chúng ta có

$$\begin{aligned}
P(\overline{X}-k\le\mu\le\overline{X}+k) &= P\left(-\frac{k\sqrt{n}}{S}\le\frac{(\overline{X}-\mu)\sqrt{n}}{S}\le\frac{k\sqrt{n}}{S}\right)\\
&= P(-t_{\alpha/2}\le T\le t_{\alpha/2})\\
&= F(t_{\alpha/2})-F(-t_{\alpha/2})\\
&= 2F(t_{\alpha/2})-1\\
&= 2\left(1-\frac{\alpha}{2}\right)-1\\
&= 1-\alpha\\
&= \gamma.
\end{aligned}$$

Chúng ta thu được khẳng định cần chứng minh. □

Ví dụ 2.9. Chiều cao của một loại thông ở Đà Lạt là một BNN X có phân phối chuẩn. Đo một số cây, người ta thu được mẫu sau:

x_i (m)	3	3,5	4	4,5	5	5,5
Số cây	2	8	23	32	23	12

Tìm khoảng tin cậy 95% của trung bình của X.

Lời giải

Từ dữ liệu, chúng ta có được tổng số cây là 100 cây, trong khi đó, trung bình mẫu là

$$\overline{x}=\frac{1}{100}(2\times3+8\times3{,}5+23\times4+32\times4{,}5+23\times5+12\times5{,}5)=4{,}51.$$

Vì X có phân phối chuẩn với phương sai chưa biết nên chúng ta tìm khoảng tin cậy dựa trên phân phối t với $m=99$ bậc tự do ($m=n-1$ với $n=100$).

Với $\gamma=95\%$, chúng ta có $\alpha/2=(1-\gamma)/2=0{,}025$. Tra bảng t-Student, ta thu được giá trị tới hạn

$$t_{\alpha/2}=1{,}984.$$

Tiếp theo, phương sai mẫu chúng ta tính được là

$$s^2=\frac{1}{99}\left(2(3-4{,}51)^2+\cdots\right)\approx0{,}3635,$$

và do đó,

$$s=\sqrt{s^2}=\sqrt{0{,}3635}\approx0{,}6029.$$

Từ đó, chúng ta có khoảng tin cậy là $(\overline{x}-k,\ \overline{x}+k)$, với

$$k=t_{\alpha/2}s/\sqrt{100}=1{,}984\times0{,}6039/10=0{,}12.$$

Chú ý 2.4. Quy trình xây dựng KTC trong trường hợp này rất giống với quy trình xây dựng KTC khi đã biết phương sai. Do đó, cần tránh sự nhầm lẫn khi áp dụng. Hai nhầm lẫn hay thấy là tính giá trị thống kê Z nhưng lại sử dụng phân vị mức của phân phối t, và tính giá trị thống kê T nhưng lại sử dụng phân vị mức của phân phối chuẩn.

Một lưu ý khác là trong trường hợp cỡ mẫu lớn, chẳng hạn $n > 30$, kết quả thu được bằng một trong hai cách tính sai lầm như trên và kết quả đúng thường có giá trị xấp xỉ nhau.

CAS 2.6. Trong phần mềm thống kê R, để tính khoảng tin cậy cho trung bình của một tổng thể khi chưa biết phương sai từ một mẫu ngẫu nhiên, chúng ta dùng chức năng `t.test()`.

Trước tiên, cần nhập dữ liệu vào một biến trong R. Trong Ví dụ 2.9, các giá trị của x_i xuất hiện với tần suất khác nhau nên cần biến x ghi các giá trị x_i và biến w ghi các tần suất tương ứng.

```
# Giá trị mẫu
x = c(3, 3.5, 4, 4.5, 5, 5.5)

# Tần suất
w = c(2, 8, 23, 32, 23, 12)

# Mẫu
data = rep(x, w)

# Kết quả của t.test()
result = t.test(data)

# Hiển thị trường conf.int (khoảng tin cậy)
print(result$conf.int)

## [1] 4,39036 4,62964
## attr(,"conf.level")
## [1] 0,95
```

Kết quả của `t.test()` cho thấy khoảng tin cậy 95% (95 percent confidence interval) của trung bình của tổng thể là từ 4,39 đến 4,63.

Lưu ý, `t.test()` tính khoảng tin cậy với mức tin cậy mặc định 95%. Để tính với mức tin cậy khác, chẳng hạn 90%, cần đặt tham số"conf.level = 0.9".

Ví dụ 2.10. Giả sử chúng ta có mẫu dữ liệu

x_i	6,5 – 7	7 – 7,5	7,5 – 8	8 – 8,5	8,5 – 9	9 – 9,5
Số mẫu	3	5	11	12	6	3

từ một BNN X tuân theo quy luật phân phối chuẩn. Tìm khoảng tin cậy của chiều cao trung bình với độ tin cậy $\gamma = 90\%$.

Lời giải

Do BNN X có phân phối chuẩn với phương sai chưa biết nên chúng ta sẽ dùng ước lượng dựa trên phân phối t. Trong phương pháp này, chúng ta cần tìm trung bình và phương sai của mẫu. Tuy nhiên, do cách ghi dữ liệu không cụ thể. Ví dụ, trong lớp thứ nhất, giá trị x_i nằm trong khoảng 6,5 – 7 tương ứng với 3 điểm mẫu, nên chúng ta thay 3 điểm mẫu này bởi 3 điểm mẫu cùng nhận giá trị trung bình 6,75 của lớp thứ nhất, ... Từ đó, trung bình mẫu là

$$\overline{x} \approx 8{,}025,$$

và phương sai mẫu hiệu chỉnh là

$$s^2 = 0{,}435.$$

(Tính toán chi tiết dành cho bạn đọc).

Với mức độ tin cậy $\gamma = 90\%$, chúng ta có $\alpha/2 = (1-\gamma)/2 = 0{,}05$. Phân vị mức $\alpha/2$ tương ứng với phân phối t với $m = n-1 = 39$ bậc tự do (Bảng A9 (t-Distribution) trong Kreyszig [1]) là

$$t_{\alpha/2} = 1{,}685.$$

Từ đó, bán kính của khoảng ước lượng là k với

$$k = \frac{t_{\alpha/2}\, s}{\sqrt{n}} = \frac{1{,}685 \times \sqrt{0{,}435}}{\sqrt{40}} \approx 0{,}1757.$$

Kết luận: Với độ tin cậy 90%, chiều cao trung bình tối thiểu $\overline{x} - k = 8{,}025 - 0{,}176 = 7{,}849$, trong khi đó trung bình tối đa là $\overline{x} + k = 8{,}025 + 0{,}1757 = 8{,}201$.

2.3.3 Khoảng tin cậy cho tỉ lệ p

Trong mục này, chúng ta tìm hiểu một cách xây dựng khoảng tin cậy cho xác suất p, được đưa ra bởi Abraham Wald[1] Mặc dù phương pháp này có nhiều nhược điểm và ngày nay ít được sử dụng trong thực tế, nhưng vì nó khá đơn giản nên nó được giảng dạy rộng rãi trong thống kê học.

Định lí giới hạn trung tâm

Khoảng tin cậy Wald cho tỉ lệ được xây dựng dựa trên một định lí cơ bản của xác suất và thống kê, đó là định lí giới hạn trung tâm. Cho $X_1, X_2, \ldots, X_n, \ldots$

[1] Abraham Wald (1902–1950) là một nhà toán học và thống kê học người Mỹ gốc Hungary.

là các biến ngẫu nhiên độc lập có cùng hàm phân phối với trung bình μ và kì vọng σ^2. Đặt

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n). \tag{2.26}$$

Đối với trường hợp hàm phân phối của các X_i không nhất thiết là phân phối chuẩn, thì chúng ta sử dụng kết quả sau, gọi là Định lí giới hạn trung tâm. Đây là một định lí trụ cột của lí thuyết xác suất và thống kê và có ứng dụng rất nhiều trong những vấn đề khác nhau.

Định lí 2.5: (Lindeberg-Lévy)

Giả sử $X_1, X_2, \ldots, X_n$ là các BNN độc lập có cùng phân phối với trung bình μ và phương sai σ^2. Khi đó, BNN

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}, \tag{2.27}$$

với $\overline{X}_n$ xác định bởi (2.26), được xấp xỉ tiệm cận bởi một phân phối chuẩn tắc, có nghĩa là

$$\lim_{n\to\infty} F_n(x) = \Phi(x), \tag{2.28}$$

với $F_n(x)$ là hàm phân phối của Z_n và $\Phi(x)$ là hàm Laplace.

Chứng minh của định lí này có thể tìm thấy ở [4, Theorem 15.38].

Trong thực tế, khi phải làm việc với các giá trị xấp xỉ của trung bình và phương sai, và trong trường hợp n "lớn", chúng ta có thể giả thiết rằng Z_n có phân phối chuẩn với trung bình μ và phương sai σ^2/n.

Khoảng tin cậy cho tỉ lệ p của tổng thể

Giả sử p là xác suất của một tổng thể để xuất hiện sự kiện A. Gọi X là BNN cho giá trị 1 nếu X xảy ra và 0 trong trường hợp còn lại. Khi đó, X có phân phối Bernoulli với $n = 1$, có nghĩa là $X \sim A(p) = B(1, p)$. Do đó,

$$\mu = \mathbb{E}[X] = p, \quad \sigma^2 = \mathrm{Var}[X] = p(1-p).$$

Xét một mẫu $x = (x_1, x_2, ..., x_n)$ có kích thước n. Khi đó, theo cách xác định X, *tần suất* xuất hiện sự kiện A trong mẫu bằng trung bình mẫu $\bar{x}$ và là một ước lượng (điểm) của xác suất tổng thể p:

$$f = \overline{x} \approx p.$$

Để xây dựng một ước lượng khoảng cho p, chúng ta lập luận như sau: Bởi định lí giới hạn trung tâm, chúng ta có thể xấp xỉ

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

bởi phân phối chuẩn tắc $\mathcal{N}(0,1)$ khi n đủ lớn. Do đó, theo công thức (2.15) khoảng tin cậy cho trung bình $\mu = p$ với độ tin cậy γ là

$$\bar{x} - \frac{U_{\alpha/2}\sigma}{\sqrt{n}} \le \mu = p \le \bar{x} + \frac{U_{\alpha/2}\sigma}{\sqrt{n}}, \tag{2.29}$$

ở đó giá trị tới hạn $U_{\alpha/2}$ ứng với mức $\alpha/2$ của phân phối chuẩn tắc $\mathcal{N}(0,1)$.

Tuy nhiên vì p chưa biết, nên thay p bởi ước lượng điểm $\bar{p} = f$ và thay phương sai σ^2 bởi phương sai mẫu $\bar{f}(1-\bar{f})$ vào công thức (2.29), chúng ta thu được khoảng tin cậy cho xác suất p với độ tin cậy γ (với n đủ lớn) là

$$\hat{p} - U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} \le p \le \hat{p} + U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}. \tag{2.30}$$

Lưu ý, do chúng ta cần định lí giới hạn trung tâm, nên trong các ứng dụng, người ta thường yêu cầu n khá lớn, thông thường là $n \ge 30$.

Ví dụ 2.11. Để ước lượng tỉ lệ phế phẩm của một lô sản phẩm, bộ phận quản lí chất lượng kiểm tra ngẫu nhiên 400 sản phẩm và nhận thấy có 12 phế phẩm. Với mức tin cậy 95%, ước lượng tỉ lệ phế phẩm tối thiểu và tối đa.

Lời giải

Gọi p là tỉ lệ phế phẩm của tổng thể, là giá trị cần ước lượng. Bài toán yêu cầu ước lượng tỉ lệ (hay xác suất) p của trung bình các BNN độc lập dựa trên định lí giới hạn trung tâm. Khoảng ước lượng trong trường hợp này có dạng

$$\overline{x} - k \le p \le \overline{x} + k,$$

trong đó, $\overline{x}$ là trung bình mẫu và k được tính dựa vào cỡ mẫu $n = 400$ và phân vị mức $\alpha/2$ của phân phối chuẩn tắc.

Mẫu ngẫu nhiên của chúng ta có 12 phế phẩm trên 400 sản phẩm kiểm tra, nên trung bình mẫu (bằng tần suất tương đối của phế phẩm) bằng

$$\overline{x} = f = \frac{12}{400} = 0{,}03.$$

Với mức tin cậy $\gamma = 95\%$, chúng ta có $\alpha/2 = (1-\gamma)/2 = 0{,}025$. phân vị mức $\alpha/2$ của phân phối chuẩn tắc là

$$U_{\alpha/2} = 1{,}960.$$

Từ đó, chúng ta tính được

$$k = U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} = 1{,}960\sqrt{\frac{0{,}03(1-0{,}03)}{400}} = 0{,}0167.$$

Như vậy, với độ tin cậy 95%, tỉ lệ phế phẩm rơi vào khoảng từ 1,33% đến 4,67%.

CAS 2.7. Trong R, hàm `binom.confint()` trong thư viện `binom`[1] cho phép tính khoảng tin cậy cho tỉ lệ p với nhiều phương pháp khác nhau. Phương pháp vừa trình bày ở trên còn được gọi là phương pháp tiệm cận. Để sử dụng phương pháp này trong R, chúng ta dùng chỉ dẫn `methods = "asymptotic"` cho `binom.conf()`. Dưới đây là một ví dụ chi tiết.

```
# Dùng goi binom
library(binom)

# Hàm binom.confint với phương thức "tiệm cận"
binom.confint(12, 400, methods = "asymptotic")

##       method  x   n mean     lower     upper
## 1 asymptotic 12 400 0,03 0,0132828 0,0467172
```

Ví dụ 2.12. Để xác định tỉ lệ thành công trong một xét nghiệm một loại bệnh, nhà nghiên cứu làm thử 500 xét nghiệm và quan sát thấy 468 xét nghiệm thành công. Với độ tin cậy $\gamma = 95\%$, hãy xác định tỉ lệ thành công trong xét nghiệm này rơi vào khoảng nào.

Lời giải

Chúng ta có cỡ mẫu $n = 500$ với tần suất thành công là

$$f = 468/500 = 0{,}936.$$

Đó là một ước lượng điểm của tỉ lệ thành công.

Với $\gamma = 95\%$, chúng ta có $U_{\alpha/2} = 1{,}960$. Từ đó, chúng ta tính được

$$U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} = 1{,}960 \times \sqrt{\frac{0{,}936(1-0{,}936)}{500}} \approx 0{,}0131.$$

Từ đó, chúng ta có ước lượng

$$0{,}9360 - 0{,}0131 < p < 0{,}9360 + 0{,}0131.$$

Có nghĩa là, với độ tin cậy 95%, tỉ lệ thành công thấp nhất là 92,129% và cao nhất là 94,91%.

[1] binom: Binomial Confidence Intervals for Several Parameterizations, `https://doi.org/10.32614/CRAN.package.binom`

Ví dụ 2.13. Một nhà khoa học vì muốn ước lượng được số thỏ trên một thung lũng nên đã bắt 400 con, đánh dấu chúng, và sau đó thả chúng trở lại thung lũng. Sau một thời gian, nhà khoa học bắt lại 500 con và thấy có 72 con có dấu. Hãy ước lượng số thỏ trên cánh đồng với độ tin cậy là 0,95.

Lời giải

Đầu tiên, chúng ta ước lượng tỉ lệ thỏ có dấu trong thung lũng dựa trên mẫu 500 con thỏ với 72 thỏ được đánh dấu. Tần suất bắt được cá thể được đánh dấu là $f = 72/500 = 0{,}144$. Với độ tin cậy $\gamma = 95\%$, ta có $\alpha/2 = 0{,}0275$ và do đó $U_{\alpha/2} \approx 1{,}960$ (theo bảng phân phối chuẩn). Từ đó, chúng ta thu được

$$k = U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} = 1{,}960 \times \sqrt{\frac{0{,}14(1-0{,}14)}{500}} \approx 0{,}0308.$$

Từ đó, với độ tin cậy 95%, chúng ta có ước lượng tỉ lệ thỏ được đánh dấu là

$$0{,}1132 \le p \le 0{,}1748.$$

Mặt khác, với 400 thỏ được đánh dấu, nên chúng ta coi $p = 400/N$ và do đó có ước lượng

$$2\,288 \le N \le 3\,593.$$

2.3.4 Khoảng tin cậy cho phương sai của phân phối chuẩn

Trong mục này, chúng ta nói về phương pháp xây dựng khoảng tin cậy cho phương sai của một phân phối chuẩn. Phương pháp này dựa trên định lí về phân phối chi-bình phương.

Phân phối chi-bình phương (phân phối χ^2) là một trường hợp riêng của phân phối Gamma, đã nói trong Chương 1. Để cho tiện cho bạn đọc, chúng ta nêu lại định nghĩa như sau.

Định nghĩa 2.5: (Phân phối χ^2)

Biến ngẫu nhiên X được gọi là có phân phối *chi-bình phương* với m bậc tự do nếu hàm phân phối tích luỹ $F(x)$ của nó có dạng

$$F(x) = \frac{1}{2^{m/2}\Gamma(m/2)}\int_0^x e^{-t/2}t^{(m-2)/2}\,\mathrm{d}t \quad \text{nếu } x \ge 0, \tag{2.31}$$

và $F(x) = 0$ nếu $x < 0$, trong đó $\Gamma(t)$ là hàm Gamma.

Phân phối χ^2 có công thức tường minh nên các tính toán liên quan đến nó có thể dễ dàng được thực hiện bằng máy tính.

Dáng điệu của hàm mật độ của phân phối χ^2 được minh hoạ bởi Hình 2.2. Lưu ý, trong trường hợp $m = 2$, dáng điệu của hàm mật độ thể hiện trong đồ thị với đường màu xanh, rất khác so với các trường hợp $m \geq 3$.

Định lí 2.6: (Lindeberg-Lévy)

Giả sử $X_1, \ldots, X_n$ là các biến ngẫu nhiên độc lập có phân phối chuẩn với trung bình μ và phương sai σ^2. Đặt

$$S^2 = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right]. \tag{2.32}$$

Khi đó, BNN

$$T = (n-1)\frac{S^2}{\sigma^2}$$

có phân phối χ^2 (chi-bình phương) với $(n-1)$ bậc tự do.

Chứng minh cho Định lí 2.6 bạn đọc có thể tìm thấy trong Walpole–Myers–Myers–Ye [5, Theorem 8.4].

Còn bây giờ, chúng ta hãy sử dụng định lí trên để xây dựng khoảng tin cậy cho phương sai σ^2 của phân phối chuẩn.

Các bước tìm khoảng tin cậy cho σ^2

1. Chọn độ tin cậy γ.
2. Xác định các phân vị mức $\chi^2_{\alpha/2}$ và $\chi_{1-\alpha/2}$ trong các điều kiện

   $$F(\chi^2_{\alpha/2}) = 1 - \alpha/2 \tag{2.33}$$

   và

   $$F(\chi^2_{1-\alpha/2}) = \alpha/2 \tag{2.34}$$

   bằng cách tra bảng phân phối (Bảng A10 trong Kreyszig [1]) hoặc bằng CAS, hoặc bằng máy tính khoa học cầm tay.
3. Tính giá trị $(n-1)s^2$, trong đó n là cỡ mẫu và s^2 là phương sai mẫu.
4. Tính các giá trị

   $$k_1 = \frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \quad k_2 = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}. \tag{2.35}$$

5. Kết luận khoảng tin cậy $100\gamma\%$ của σ^2 là

   $$k_1 \leq \sigma^2 \leq k_2. \tag{2.36}$$

   Đó là khoảng tin cậy cần tìm.

Ví dụ 2.14. Giả sử chúng ta có một mẫu gồm $n = 30$ quan sát, và phương sai mẫu $s^2 = 10$. Hãy xây dựng khoảng tin cậy 95% cho phương sai σ^2.

Lời giải

Các bước tìm KTC cho phương sai được thực hiện như sau:

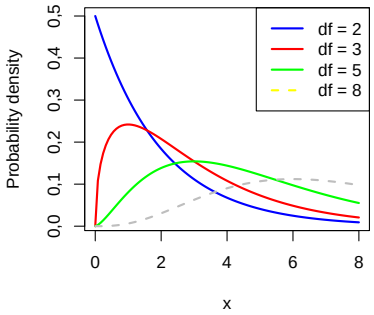
1. Xác định các thông số:
   - Số mẫu: $n = 30$.
   - Phương sai mẫu: $s^2 = 10$.
   - Độ tin cậy: $\gamma = 95\%$, do đó $\alpha = 0{,}05$, và $\alpha/2 = 0{,}025$.
   - Số bậc tự do: $m = n - 1 = 29$.
2. Tra bảng phân phối χ^2: Chúng ta cần tra các phân vị mức $\alpha/2$ của phân phối chi-bình phương với $n - 1 = 29$ bậc tự do.

   $$\chi^2_{0{,}025} \approx 16{,}047, \quad \chi^2_{0{,}975} \approx 45{,}722.$$

3. Tính khoảng tin cậy cho phương sai σ^2: Áp dụng các giá trị trên vào công thức:

   $$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = \left(\frac{(30-1)\times 10}{45{,}722}, \frac{(30-1)\times 10}{16{,}047}\right)$$

   Từ đó, bằng tính toán cụ thể, chúng ta thu được khoảng tin cậy cần tìm là $(6{,}34; 18{,}07)$.



Hình 2.2: Hàm mật độ của phân phối χ^2

Từ khoảng tin cậy cho σ^2 chúng ta suy ra khoảng tin cậy cho độ lệch chuẩn σ bằng cách lấy căn bậc hai của hai đầu mút, đó là

$$\left(\sqrt{6{,}34}, \sqrt{18{,}07}\right) = (2{,}52;\ 4{,}25).$$

Vậy khoảng tin cậy 95% cho độ lệch chuẩn σ là $(2{,}52;\ 4{,}25)$.

Chú ý 2.5. Đồ thị hàm mật độ của phân phối chi-bình phương không đối xứng (xem Hình 2.2). Do đó, không có mối liên hệ đơn giản giữa χ^2_α và $\chi^2_{1-\alpha}$ như đối với phân phối chuẩn và phân phối t. Như vậy, chúng ta phải tính lần lượt các giá trị này một cách riêng biệt.

CAS 2.8. Trong R, chúng ta có thể dùng hàm `varTest()` (One-Sample Chi-Squared Test on Variance) trong gói `EnvStats`[1] để tính toán khoảng tin cậy cho phương sai của phân phối chuẩn. Sau đây là một ví dụ:

```
# Dùng gói EnvStats
library(EnvStats)

# Nạp dữ liệu
x = c(89, 85, 87, 82, 89, 86, 91, 90, 78, 89, 87, 99, 83, 81)

# Thực hiện varTest
result = varTest(x)

# Hiển thị kết quả
print(result$conf.int)

##      LCL      UCL
## 13,8955 68,6228
## attr(,"conf.level")
## [1] 0,95
```

Cận dưới của KTC cho bởi LCL (lower confident limit) và cận trên cho bởi UCL (upper confident limit).

2.3.5 Những hiểu sai về khoảng tin cậy

Giả sử X là một BNN với trung bình μ. Nếu một nghiên cứu nói $[\alpha,\beta]$ *là một khoảng tin cậy 95% của* μ thì điều này có nghĩa là gì?

Đây mặc dù là câu hỏi cơ bản nhưng lại khá tế nhị mà nhiều người làm thống kê ứng dụng có kinh nghiệm cũng dễ cho câu trả lời sai.

Nhắc lại rằng, để hiểu ý nghĩa của một khoảng tin cậy ta phải quay lại quy trình xây dựng khoảng tin cậy từ một mẫu ngẫu nhiên. Quy trình này có tính chất sau: *Có khoảng* $100\gamma\%$*, trong đó* γ *là độ tin cậy, các khoảng xây dựng theo nó từ các mẫu ngẫu nhiên và độc lập sẽ chứa tham số tổng thể (ở trong ví dụ trên là trung bình* μ*).*

Các diễn giải sai lầm về khoảng tin cậy thường bắt từ việc quên đi quy trình xây dựng mà chỉ chú ý đến kết quả của quy trình xây dựng được, tức là chỉ chú ý đến $[\alpha,\beta]$.

Có phải nếu $[\alpha,\beta]$ là khoảng tin cậy 95% của μ thì có 95% khả năng μ thuộc vào $[\alpha,\beta]$?

Câu trả lời ở đây là "không". Vì sao?

[1] EnvStats: An R Package for Environmental Statistics, https://doi.org/10.1007/978-1-4614-8456-1

Vì chỉ có một BNN X, là một mô hình cho một tổng thể duy nhất. Do đó trung bình μ đã hoàn toàn xác định. Sự kiện μ có thuộc một khoảng nào đó hay không là sự kiện tất định, không phải là sự kiện ngẫu nhiên. Đánh giá khả năng xảy ra sự kiện tất định hay không là vô nghĩa, vì không có một không gian xác suất hợp lí cho sự kiện đó.

Nếu lấy mẫu x nhiều lần từ BNN X và tính các trung bình mẫu $\overline{x}$, thì khoảng 95% kết quả thu được nằm trong $[\alpha, \beta]$.

Mặc dù cách hiểu này mới nghe có vẻ hợp lí, nhưng nó cũng là một cách hiểu sai rất nghiêm trọng! Điều bất ngờ là cách hiểu sai này cũng xuất hiện khá phổ biến, trong cả các sách đã xuất bản. Ví dụ, tác giả N. V. Tuấn [6][1], sau khi tính được một khoảng tin cậy 95% cho một ví dụ là 0,68 và 0,71, đưa ra diễn giải như sau: "Nói cách khác, nếu nghiên cứu được lặp lại nhiều lần, chúng ta kỳ vọng rằng giá trị trung bình của mẫu (sample means) của mật độ xương sẽ dao động trong khoảng 0.68 đến 0.71 g/cm^2 (trung bình là 0.695)."

Sai lầm nằm ở đâu trong diễn giải trên?

Do các tham số tổng thể không biết được, nên không thể tính được kì vọng các trung bình mẫu rơi trong khoảng bất kì nào: Chúng ta không bao giờ đánh giá được xác suất mà các trung bình mẫu sẽ thu được rơi vào khoảng $[\alpha, \beta]$ vừa tìm là bao nhiêu.

Vậy cách hiểu đúng là gì? Như đã nói ở trên, là nếu nghiên cứu được lặp lại nhiều lần, thì ta kì vọng rằng có xấp xỉ 95% trong các khoảng tin cậy được tính trong các nghiên cứu được lặp lại đó thực sự chứa trung bình của tổng thể.

Trong ví dụ trên, cặp giá trị vừa tính được, 0,68 và 0,71 chỉ là các giá trị ngẫu nhiên. Đứng một mình, các giá trị của BNN, hay các kết quả của phép thử ngẫu nhiên (giống như một kết quả đồng xu sấp hay ngửa) không có ý nghĩa đối với các phép thử độc lập tiếp theo.

2.3.6 Một số mô phỏng trong R

Trong tiểu mục này, chúng ta mô phỏng một số tình huống giúp bạn đọc hiểu sâu hơn cũng như kiểm chứng các kết quả về mặt lí thuyết thống kê.

CAS 2.9. Nhắc lại rằng có khoảng 100γ % các khoảng xây dựng từ các mẫu ngẫu nhiên khác nhau từ một phân phối chuẩn theo quy tắc trên sẽ thực sự chứa phương sai của tổng thể. Trong phần mô phỏng bằng R dưới đây, chúng ta sẽ tạo ra 1000 mẫu ngẫu nhiên từ phân phối chuẩn tắc

[1] N. V. Tuấn (2019), *Phân tích dữ liệu với R (hỏi và đáp)*, NXB Tổng hợp TP. Hồ Chí Minh: Đây là một cuốn sách khá hữu ích cho người làm ứng dụng thống kê. Rất tiếc, cũng như cuốn "Phân tích dữ liệu với R" của cùng tác giả, cuốn sách này chứa nhiều lỗi sai nghiêm trọng về các vấn đề cơ sở lí thuyết xác suất và thống kê.

$\mathcal{N}(0,1)$, tính các khoảng tin cậy từ các mẫu ngẫu nhiên đó, và tính tỉ lệ % các khoảng tin cậy thực sự chứa phương sai $\sigma^2 = 1$.

```
library(EnvStats)
N = 1000; k = 0

for (i in 1:N)  {

      # Tạo các mẫu của phân phối chuẩn tắc.
      sample = rnorm(30, mean = 0, sd = 1)

      # Tính khoảng tin cậy của mẫu với độ tin cậy 99%.
      ci = varTest(sample, conf.level = 0.99)

      # Đếm các khoảng thực sự chứa phương sai của tổng thể.
      if ((ci$conf.int[1] - 1) * (ci$conf.int[2] - 1) < 0) {
            k = k + 1
      }
}

# In ra tỉ lệ % các KTC thực sự chứa phương sai
print(100 * k / N)

## [1] 99
```

Kết quả thu được có 99% các khoảng tin cậy thực sự chứa phương sai của phân phối chuẩn tắc $\sigma^2 = 1$.

Một câu hỏi đặt ra là nếu tổng thể không tuân theo phân phối chuẩn thì quy tắc xây dựng khoảng tin cậy như trên còn đúng không? Câu trả lời là "không". Đó là vì khi tổng thể X không có phân phối chuẩn, biến ngẫu nhiên $T = (n-1)S^2/\sigma^2$ không nhất thiết có phân phối χ^2 với $n-1$ bậc tự do. Do đó, việc tính KTC dựa trên phân phối χ^2 là không có cơ sở.

CAS dưới đây cho thấy nếu áp dụng quy tắc KTC trên cho trường hợp phân phối mũ thì kết quả thu được không đúng.

CAS 2.10. Sửa đổi đoạn mã R trong CAS 2.9, chúng ta tạo các mẫu với phân phối mũ với tham số $\lambda = 0{,}2$, phương sai $\sigma^2 = \lambda^{-2} = 25$, và xây dựng các khoảng theo công thức trên, đếm các khoảng thực sự chứa σ^2. Cụ thể như sau:

```
library(EnvStats)
# Tham số N = 1000 cho số mẫu sẽ tạo
N = 1000; k = 0

# Tạo 1000 mẫu từ phân phối mũ, áp dụng công thức KTC,
```

```
# và đếm số KTC thực sự chứa tham số tổng thể
for (i in 1:N)  {

      # Sinh ra mẫu cỡ 30 từ phân phối mũ tham số 0,2
      sample = rexp(30, 0.2)

      # Thực hiện varTest đối với mẫu với độ tin cậy 90%
      ci = varTest(sample, conf.level = 0.9)

      # Kiểm tra và đếm số mẫu chứa phương sai sigma^2 = 25
      if ((ci$conf.int[1] - 25) * (ci$conf.int[2] - 25) < 0) {
            k = k + 1
      }
}

# In ra tỉ lệ (theo %) các khoảng chứa sigma^2 = 25
print(100 * k / N)

## [1] 63,8
```

Kết quả cho thấy chỉ có 63,8 %, nhỏ hơn nhiều so với mức mong muốn 90%, các khoảng được xây dựng như vậy thực sự chứa phương sai $\sigma^2 = 25$. Thực nghiệm này chỉ ra rằng công thức khoảng tin cậy cho phương sai của phân phối chuẩn *không thể* áp dụng cho các phân phối khác.

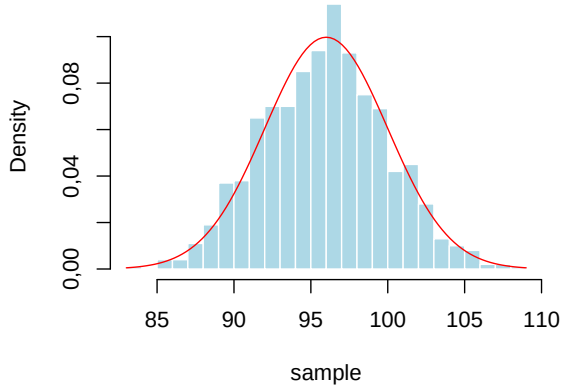
CAS 2.11. Giả sử X_n là một dãy các BNN độc lập có cùng phân phối đều $U(0,1)$. Dễ dàng tính được trung bình $\mu_{X_n} = 1/2$ và phương sai $\sigma^2_{X_n} = 1/12$. Mặt khác, nếu đặt $X = X_1 + \cdots + X_n$ và $\overline{X} = \frac{1}{n}$, thì $\mu_X = n/2$ và $\sigma_X = n/12$. Hơn nữa, định lí giới hạn trung tâm cho biết

$$X_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

có phân phối xấp xỉ $\mathcal{N}(0,1)$ khi n lớn.

Chúng ta hãy xem xét thí nghiệm sau trong R: Tạo ra $N = 1000$ mẫu cỡ $n = 192$ của phân phối đều $U(0,1)$ và tính tổng của chúng. Như vậy, ta thu được một mẫu cỡ 1000, là 1000 giá trị quan sát của BNN X với phân phối xấp xỉ phân phối chuẩn $\mathcal{N}(96,16)$. Tiếp theo, vẽ biểu đồ cột (histogram) của mẫu thu được cùng với đồ thị của hàm mật độ của phân phối chuẩn $\mathcal{N}(\mu = 96, \sigma^2 = 16)$. Bạn đọc hãy quan sát Hình 2.3 thu được và đưa ra nhận xét cho riêng mình.

```
N = 1000
n = 192
```



Hình 2.3: Xấp xỉ phân phối chuẩn của tổng các BNN với phân phối đều $U(0,1)$

```
sample = numeric(N)

# Tạo ra 1000 mẫu cỡ n = 192 từ phân phối đều và tính tổng
for (i in 1:N)  {
    x = runif(n, min = 0, max = 1)
    sample[i] = sum(x)
}

# Vẽ biểu đồ cột biểu thị các tổng thu được
hist(sample, breaks = 20, probability = TRUE,
    col = "lightblue", border = "white", main = "")

# Vẽ đường con biểu thị mật độ của phân phối chuẩn
curve(dnorm(x, mean = 96, sd = 4), col = "red", add = TRUE)
```

Bài tập mục 2.3

1. Phân phối t là gì? Nêu những điểm giống và khác nhau giữa phân phối t và phân phối chuẩn. Nêu một ứng dụng quan trọng của phân phối t.
2. Tính xấp xỉ các phân vị mức 90%, 95% và 97,5% của phân phối t với 24 và 30 bậc tự do (xem Bảng A9 trong Kreyszig [1]).
3. Thu thập ngẫu nhiên một mẫu dữ liệu gồm 100 mẫu từ một tổng thể tuân theo luật phân phối chuẩn, người ta tính được trung bình mẫu $\overline{x} = 0{,}1$ trong khi đó độ lệch

trung bình của mẫu là $s = 0{,}014$. Hãy ước lượng giá trị trung bình tổng thể, với độ tin cậy 95%, từ mẫu trên.

4. Điều tra ngẫu nhiên điểm trung bình môn toán của 100 thí sinh dự thi vào một trường đại học A, người ta thu được trung bình mẫu là 7,3 với phương sai mẫu $s^2 = 6{,}25$. Tìm khoảng tin cậy 95% cho điểm trung bình môn Toán của toàn bộ thí sinh dự thi dựa vào mẫu này.

5. Một nghiên cứu nhằm tìm hiểu các giao dịch mua sắm trực tuyến cho thấy thời gian trung bình để hoàn tất một giao dịch mua sắm trực tuyến là 5,8 phút với độ lệch chuẩn 2,4 phút. Nghiên cứu này dựa trên 105 mẫu. Giả sử thời gian để hoàn tất một giao dịch như vậy là một BNN tuân theo phân phối chuẩn. Hãy xây dựng khoảng tin cậy 95% cho thời gian trung bình để hoàn tất một giao dịch mua sắm.

6. Giả sử chiều cao của một nam 10 tuổi là một biến ngẫu nhiên X có phân phối chuẩn. Đo ngẫu nhiên 16 học sinh và thu được mẫu số liệu sau: 129, 130, 131, 131, 131, 132, 132, 132, 133, 133, 134, 134, 135, 135, 136, 137 (đơn vị: cm).

(a) Tính trung bình và phương sai của mẫu dữ liệu đó.

(b) Với độ tin cậy 95%, hãy ước lượng trung bình của X.

7. Tìm ước lượng điểm và khoảng tin cậy 95% cho trung bình của một BNN X với phân phối chuẩn với phương sai $\sigma^2 = 4$ dựa trên mẫu quan sát sau: $23, 24, 24, 25, 25, 25, 26, 27$.

8. Giả sử chiều dài của một quả dưa chuột được thu hoạch ở một nông trường là một BNN có phân phối chuẩn. Đo chiều dài của một số quả dưa và ghi lại số liệu như trong bảng sau:

Chiều dài	4,2 – 4,5	4,5 – 4,8	4,8 – 4,1	5,1 – 5,4	5,4 – 5,7
Số trái	4	7	11	7	5

Hãy xây dựng khoảng tin cậy 95% của trung bình của chiều dài loại dưa chuột nói trên.

9. Một mẫu ngẫu nhiên gồm $n = 89$ khách hàng cho thấy rằng 48 khách hàng có kế hoạch mua một loại ti vi mới. Tìm khoảng tin cậy 95% của tỉ lệ p của các khách hàng trong tổng thể tất cả các khách hàng mà có kế hoạch mua ti vi mới.

10. (K. Pearson) Tìm khoảng tin cậy 99% cho xác suất p của một phân phối nhị thức với một mẫu kết quả của 24 000 phép thử trong đó có 12 012 kết quả thành công.

2.4 Hàm đặc trưng và định lí giới hạn trung tâm

Trong mục này, chúng ta tìm hiểu về hàm đặc trưng của một phân phối xác suất, với mục tiêu chính là trình bày sơ lược về một cách chứng minh khá ngắn gọn của định lí giới hạn trung tâm (Định lí 2.5).

Tài liệu tham khảo: Cramer [7].

2.4.1 Hàm đặc trưng

Hàm đặc trưng là một công cụ mạnh trong xác suất và thống kê, thường dùng để chứng minh các định lí giới hạn như định lí giới hạn trung tâm.

Nó có liên hệ mật thiết với biến ngẫu nhiên, phân phối xác suất, và các biến đổi Fourier. Hàm đặc trưng tồn tại và xác định, kể cả khi một số đặc trưng như kì vọng hay phương sai không tồn tại.

Định nghĩa 2.6:

Hàm đặc trưng của biến ngẫu nhiên giá trị thực X là hàm cho tương ứng mỗi $t \in \mathbb{R}$ với kì vọng toán của biến ngẫu nhiên giá trị phức e^{itX}, và được kí hiệu là $\varphi_X(t)$. Như vậy,

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right], \quad t \in \mathbb{R}. \tag{2.37}$$

Vì X có giá trị thực, ta có thể viết $e^{itX} = \cos(tX) + i\sin(tX)$ và thu được

$$\varphi_X(t) = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)].$$

Nếu X là biến ngẫu nhiên rời rạc thì:

$$\varphi_X(t) = \sum_x e^{itx} P(X = x).$$

Chẳng hạn, nếu $X \sim A(p)$ thì

$$\varphi_X(t) = P(X = 0) + e^{it} P(X = 1) = 1 - p + pe^{it}.$$

Lưu ý, hàm trưng của hai BNN cùng phân phối là như nhau. Do đó, ta cũng nói $\varphi_X(t)$ là hàm đặc trưng của phân phối của X. Như vậy, hàm số $t \mapsto 1 - p + pe^{it}$ cũng là hàm đặc trưng của phân phối $A(p)$.

Nếu X là biến ngẫu nhiên liên tục thì

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,\mathrm{d}x.$$

Như vậy, hàm đặc trưng của một phân phối chính là *biến đổi Fourier* của hàm mật độ xác suất của nó.

Bảng sau đây cho biết các hàm đặc trưng của một số phân phối xác suất cơ bản.

Phân phối xác suất	**Hàm đặc trưng**
$B(n,p)$	$\left(1 - p + pe^{it}\right)^n$
$\mathrm{Pois}(\lambda)$	$\exp\left(\lambda\left(e^{it} - 1\right)\right)$
$\mathcal{N}(\mu,\sigma^2)$	$\exp\left(i\mu t - \dfrac{\sigma^2 t^2}{2}\right)$
$U(a,b)$	$\dfrac{e^{itb} - e^{ita}}{it(b-a)}$

Một số tính chất của hàm đặc trưng như sau :

1. Nếu X là một BNN giá trị thực thì $|\varphi_X(t)| \le 1$. Thật vậy,

$$\begin{aligned}\left|\int e^{itx} f(x)\,\mathrm{d}x\right| &\le \int \left|e^{itx}\right| |f(x)|\,\mathrm{d}x \\ &= \int |f(x)|\,\mathrm{d}x \\ &= 1 \quad (\text{vì } f(x) \ge 0).\end{aligned}$$

2. Nếu $Y = aX + b$ với a, b là các hằng số thực thì $\varphi_Y(t) = e^{itb}\varphi_X(at)$. Thật vậy, theo định nghĩa:

$$\begin{aligned}\varphi_Y(t) &= \mathbb{E}\left[e^{itY}\right] \\ &= \mathbb{E}\left[e^{it(aX+b)}\right] \\ &= e^{itb}\mathbb{E}\left[e^{iatX}\right] \\ &= e^{itb}\varphi_X(at).\end{aligned}$$

3. Hàm phân phối $F(x)$ xác định một cách duy nhất hàm đặc trưng $\varphi_X(t)$. Đây là một tính chất suy từ biến đổi Fourier và biến đổi Fourier ngược.

4. Nếu $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên độc lập và

$$X = X_1 + X_2 + \cdots + X_n$$

thì

$$\varphi_X(t) = \prod_{k=1}^{n} \varphi_{X_k}(t).$$

Thật vậy, dễ thấy rằng vì các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập nên:

$$\begin{aligned}\varphi_X(t) &= \mathbb{E}\left[e^{it(X_1+\cdots+X_n)}\right] \\ &= \mathbb{E}\left[e^{itX_1}\cdots e^{itX_n}\right] \\ &= \prod_{k=1}^{n} \mathbb{E}\left[e^{itX_k}\right] \\ &= \prod_{k=1}^{n} \varphi_{X_k}(t).\end{aligned}$$

5. Nếu tồn tại $\mathbb{E}\left[|X|^k\right]$ thì hàm đặc trưng $\varphi_X(t)$ cũng tồn tại đạo hàm đến bậc k tại mọi điểm t (khả vi đến bậc k tại mọi điểm t).

2.4.2 Định lí giới hạn trung tâm

Sử dụng các hàm đặc trưng, có thể chứng minh định lí giới hạn trung tâm, một định lí rất quan trọng trong những ứng dụng đã biết.

Định lí 2.7: (Lindeberg–Lévy)

Cho $(X_i)_{i\geq 1}$ là dãy các biến ngẫu nhiên độc lập và cùng phân phối thỏa mãn

$$\mathbb{E}[X_1]=\mu, \quad 0<\mathrm{Var}(X_1)=\sigma^2<\infty.$$

Gọi $S_n=\sum_{i=1}^n X_i$. Khi đó,

$$\frac{S_n-n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad (\text{khi } n\to\infty),$$

tức là phân phối chuẩn chuẩn hoá chuẩn tắc là giới hạn phân phối của tổng được chuẩn hoá.

Chứng minh định lí 2.7. Với n cố định, ta đặt $Y_i=(X_i-\mu)/(\sigma\sqrt{n})$ và đặt

$$Z_n := \frac{S_n-n\mu}{\sigma\sqrt{n}} = \sum_{i=1}^n Y_i.$$

Ta sẽ chứng minh rằng hàm đặc trưng $\varphi_{Z_n}(t)$ hội tụ điểm tới $\varphi(t)=e^{-t^2/2}$ (hàm đặc trưng của $\mathcal{N}(0,1)$). Thật vậy, theo tính độc lập và tính đồng phân phối của các BNN X_i ta suy ra tính độc lập và cùng phân phối của Y_i. Do đó

$$\varphi_{Z_n}(t)=\mathbb{E}\left[\exp\left(it\sum_{i=1}^n Y_i\right)\right]=\left(\varphi_{Y_1}(t)\right)^n=\left(\varphi_{Z_1}\left(t/\sqrt{n}\right)\right)^n.$$

(Lưu ý, ở trên ta đã dùng $Z_1=(S_1-\mu)/\sigma)=\sqrt{n}\,Y_1$.)

Vì $\mathbb{E}\left[Z_1^2\right]<\infty$, ta có thể lấy đạo hàm dưới dấu tích phân:

$$\varphi'_{Z_1}(u)=\mathbb{E}\left[iZ_1e^{iuZ_1}\right], \quad \varphi''_{Z_1}(u)=\mathbb{E}\left[-Z_1^2e^{iuZ_1}\right],$$

và do đó $\varphi'_{Z_1}(0)=i\mathbb{E}[Z_1]=0$, $\varphi''_{Z_1}(0)=-\mathbb{E}\left[Z_1^2\right]=-1/n$. Từ đó (do sự tồn tại và liên tục của đạo hàm bậc hai tại 0) suy ra khai triển Taylor

$$\varphi_{Z_1}(u)=1-\frac{1}{2}u^2+o(u^2) \quad (u\to 0).$$

Với t cố định, ta có

$$\varphi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)=1-\frac{1}{2n}t^2+o(1/n), \quad n\to\infty.$$

Cùng với trên, ta có

$$\begin{aligned}\log\varphi_{Z_n}(t) &= n\log\varphi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\\ &= n\log\left(1-\frac{t^2}{2n}+o\left(\frac{1}{n}\right)\right)\\ &= -\frac{1}{2}t^2+o(1), \quad n\to\infty.\end{aligned}$$

Do đó

$$\varphi_{Z_n}(t)=\exp\left(\log\varphi_{Z_n}(t)\right)\to e^{-t^2/2} \quad (n\to\infty),$$

với hội tụ điểm theo $t\in\mathbb{R}$.

Theo định lí liên tục của Lévy (Lévy's continuity theorem), nếu các hàm đặc trưng $\varphi_{Z_n}(t)$ hội tụ điểm tới một hàm liên tục $\varphi(t)$ tại mọi t, thì phân phối của Z_n hội tụ tới phân phối có hàm đặc trưng φ. Ở đây $\varphi(t)=e^{-t^2/2}$ là hàm đặc trưng của $\mathcal{N}(0,1)$ và liên tục, nên

$$Z_n=\frac{S_n-n\mu}{\sigma\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,1).$$

Điều này hoàn tất chứng minh. □

Tài liệu

[1] E. Kreyszig, *Advanced Engineering Mathematics, Part G: Probability, Statistics*, 10 **edition**. New York: John Wiley & Sons, Inc., 2011.

[2] Đ. H. Thắng, *Thống kê và Ứng dụng: Giáo trình dành cho các trường Đại học và Cao đẳng*, 6 **edition**. NXB Giáo dục Việt Nam, 2022.

[3] N. V. Khuê, P. N. Thao, L. M. Hải **and** N. Đ. Sang, *Toán cao cấp, tập I (A1)*. NXB Giáo dục, 2008.

[4] A. Klenke, *Probability theory: a comprehensive course*, 3 **edition**. Springer Science & Business Media, 2020.

[5] R. E. Walpole, R. H. Myers, S. L. Myers **and** K. Ye, *Probability & Statistics for Engineers & Scientists*, 9 **edition**. Prentice Hall, Pearson, 2002.

[6] N. V. Tuấn, *Phân tích dữ liệu với R (Hỏi và đáp)*, 2 **edition**. NXB Tổng hợp TP. Hồ Chí Minh, 2019.

[7] H. Cramer, *Phương pháp toán học trong thống kê (dịch)*. NXB Khoa học và Kĩ thuật, 1970.

Chương 3

Kiểm định giả thiết thống kê

Mục tiêu của chương này là tìm hiểu một số khái niệm và phương pháp tiêu chuẩn để kiểm định các giả thiết thống kê, bao gồm các giả thiết về trung bình (kiểm định z, kiểm định t), giả thiết về phương sai (kiểm định χ^2), so sánh hai trung bình (kiểm định t hai mẫu), và so sánh hai phương sai.

Tài liệu tham khảo chính: Kreyszig [1, Chapter 25], Đ. H. Thắng [2, Chương 4, 5].

Nội dung

3.1 Khái niệm, quy tắc chung, các dạng kiểm định

Kiểm định giả thiết thống kê (statistical hypothesis testing) là một phương pháp thống kê được sử dụng để kiểm tra giả thiết về một hoặc nhiều thông số của quần thể dựa trên mẫu dữ liệu thu thập được. Mục tiêu của kiểm định thường là để xác định xem liệu có bằng chứng thống kê đủ mạnh để bác bỏ hoặc không bác bỏ giả thiết đã được đặt ra.

Quy trình kiểm định thường bắt đầu với việc thiết lập hai giả thiết:

- *Giả thiết chính*, kí hiệu H_0: Đây là giả thiết mà bạn muốn kiểm tra, thường là giả thiết về sự không khác biệt hoặc một mức độ cụ thể của sự khác biệt.
- *Đối thiết*, kí hiệu H_1: Đây là giả thiết mà bạn muốn chứng minh là đúng, thường là giả thiết về sự khác biệt.

Người ta chọn một quy trình kiểm định và thực hiện trên dữ liệu mẫu. Kết quả của phép kiểm định được so sánh với một ngưỡng quyết định suy ra từ *mức ý nghĩa* (significance level) được chọn trước. Nếu kết quả đạt được vượt qua ngưỡng này, có đủ bằng chứng để bác bỏ giả thiết chính H_0 và chấp nhận giả thiết thay thế H_1; ngược lại, nếu kết quả không vượt qua ngưỡng ý nghĩa, thì không có đủ bằng chứng để bác bỏ giả thiết chính.

Có nhiều loại kiểm định thống kê khác nhau, bao gồm kiểm định z, kiểm định t cho trung bình, kiểm định chi-bình phương cho phương sai, kiểm định ANOVA (Analysis of Variance) cho dữ liệu có thể được coi là liên tục, kiểm định chi bình phương (χ^2) cho dữ liệu phân loại. Mỗi loại kiểm định có ứng dụng và giả định riêng biệt tùy thuộc vào bối cảnh nghiên cứu và loại dữ liệu được sử dụng.

3.1.1 Quy trình kiểm định giả thiết thống kê

Các bước cơ bản để thực hiện một phép kiểm định thống kê như sau:

1. Xác định cặp giả thiết:
   - Phát biểu giả thiết chính H_0, chẳng hạn $\theta = \theta_0$ với θ là một tham số nào đó (như trung bình hay phương sai) và θ_0 là một giá trị cho trước.
   - Phát biểu đối thiết H_1.
2. Chọn mức ý nghĩa α (chẳng hạn $\alpha = 5\%$, 1%, hoặc thậm chí là 0,1%).
3. Chọn *tiêu chuẩn kiểm định* $T = g(X_1, X_2, \ldots, X_n)$ phù hợp với từng tình huống và xác định quy luật phân phối của T với điều kiện giả thiết H_0 là đúng.

4. Tìm *miền bác bỏ* W_α sao cho

$$P(T \in W_\alpha) = \alpha, \tag{3.1}$$

tức là *trong trường hợp giả thiết* H_0 *là đúng thì xác suất xảy ra* $T \in W_\alpha$ *bằng* α. Lưu ý, miền bác bỏ W_α được chọn không chỉ phụ thuộc vào giả thiết chính mà phụ thuộc cả vào đối thiết.

5. Từ mẫu cụ thể $(x_1, x_2, \ldots, x_n)$ tính *giá trị quan sát* của tiêu chuẩn kiểm định

$$T_* = g(x_1, x_2, \ldots, x_n).$$

6. Chấp nhận hoặc bác bỏ giả thiết H_0 dựa trên mối quan hệ giữa giá trị quan sát T_* và miền bác bỏ W_α. Cụ thể như sau:

   - Nếu $T_* \in W_\alpha$ thì *bác bỏ* H_0.
   - Nếu $T_* \notin W_\alpha$, thì chưa có đủ cơ sở để bác bỏ H_0.

Các tiêu chuẩn kiểm định chúng ta sẽ nói trong chương này gồm kiểm định z và kiểm định t cho trung bình của phân phối chuẩn, kiểm định χ^2 cho phương sai của phân phối chuẩn, và kiểm định t hai mẫu cho vấn đề so sánh hai trung bình.

3.1.2 Sai lầm của kiểm định giả thiết thống kê

Kiểm định giả thiết thống kê là một công cụ quan trọng trong việc đưa ra kết luận dựa trên dữ liệu. Tuy nhiên, nó không hoàn hảo và có những sai lầm phổ biến mà người sử dụng cần nhận biết. Dưới đây là một số sai lầm lớn khi thực hiện kiểm định giả thiết:

Sai lầm loại I

Sai lầm loại I xảy ra khi bác bỏ giả thuyết gốc H_0 trong khi nó đúng. Điều này có nghĩa là chúng ta đưa ra kết luận sai rằng có bằng chứng để tin vào giả thuyết thay thế H_1 khi thực tế không phải vậy.

Ví dụ, giả sử chúng ta đang kiểm tra giả thuyết về một loại thuốc mới với H_0: "Thuốc không có tác dụng", và H_1: "Thuốc có tác dụng." Nếu chúng ta bác bỏ H_0 và kết luận rằng thuốc có tác dụng, nhưng trên thực tế thuốc không có tác dụng, thì chúng ta đã mắc sai lầm loại I.

Xác suất mắc sai lầm loại I là mức ý nghĩa (α) của kiểm định. Đây là xác suất cho thấy một sự khác biệt đáng kể trong dữ liệu (theo H_1) khi thực tế sự khác biệt đó không tồn tại (theo H_0). Thông thường, mức α được đặt ở 0,05, nghĩa là có 5% khả năng chúng ta quan sát được dữ liệu cho sự bác bỏ H_0.

Sai lầm loại I thường dẫn đến kết luận sai rằng một yếu tố nào đó có tác động hoặc có hiệu quả khi thực tế nó không có. Điều này đặc biệt nguy hiểm trong các lĩnh vực như y tế, khi bạn có thể khuyến cáo sử dụng một loại thuốc không hiệu quả hoặc thậm chí có thể gây hại.

Sai lầm loại II

Sai lầm loại II xảy ra khi không bác bỏ giả thuyết gốc H_0 trong khi nó sai. Xác suất mắc sai lầm này được kí hiệu là β, và *độ mạnh* của kiểm định (power of test) là $1-\beta$, biểu thị khả năng phát hiện sự khác biệt khi nó thực sự tồn tại.

Ví dụ, giả sử chúng ta đang kiểm tra loại thuốc như trên, với H_0: "Thuốc không có tác dụng", và H_1: "Thuốc có tác dụng". Nếu không bác bỏ H_0 và kết luận rằng thuốc không có tác dụng, nhưng thực tế thuốc có tác dụng, thì chúng ta đã mắc sai lầm loại II.

Xác suất mắc sai lầm loại II được ký hiệu là β. Giá trị này thường không được biết trước và phụ thuộc vào nhiều yếu tố như kích thước mẫu và hiệu ứng thật sự trong dữ liệu. Độ mạnh của kiểm định được xác định bởi $1-\beta$, nghĩa là khả năng phát hiện ra sự khác biệt khi sự khác biệt thực sự tồn tại. Thường một kiểm định mạnh có độ mạnh ít nhất là 0,8, tương ứng với 20% khả năng mắc sai lầm loại II.

Sai lầm loại II dẫn đến việc bỏ qua một phát hiện quan trọng. Ví dụ, trong nghiên cứu y tế, có thể bỏ qua tác dụng có lợi của một loại thuốc. Điều này làm giảm khả năng phát hiện ra các yếu tố hoặc tác nhân có giá trị, khiến nghiên cứu hoặc thử nghiệm không mang lại hiệu quả cao.

Tóm tắt về hai loại sai lầm nêu trên được ghi trong bảng sau.

Yếu tố	**Sai lầm loại I**	**Sai lầm loại II**
Kết luận sai	Bác bỏ H_0 khi H_0 đúng	Không bác bỏ H_0 khi H_0 sai
Xác suất	α (thường là 0,05 hoặc 0,01)	β (không được biết trước)
Hậu quả	Kết luận sai rằng có hiệu ứng	Bỏ sót một hiệu ứng thực sự
Ưu tiên giảm thiểu	Đặt ra mức ý nghĩa thấp	Tăng cỡ mẫu, tăng độ mạnh

Trong kiểm định giả thiết thống kê, ngoài hai sai lầm nêu trên còn có một số sai lầm khác với các lí do đến từ thực tiễn mà chúng ta không nói đến ở đây.

3.2 Kiểm định trung bình của phân phối chuẩn

Trong mục này chúng ta tìm hiểu chi tiết bài toán kiểm định trung bình của một phân phối chuẩn trong hai tình huống: (1) phương sai đã biết, và

(2) phương sai chưa biết. Bạn đọc có thể tham khảo thêm Kreyszig [1, tr. 1081–1086] hoặc Đ. H. Thắng [2, Chương 4, §2].

Giả sử X là một BNN có phân phối chuẩn với trung bình μ và phương sai σ^2, nghĩa là $X \sim \mathcal{N}(\mu, \sigma^2)$. Chúng ta muốn kiểm tra một giả thiết phát biểu rằng trung bình μ của X bằng một giá trị μ_0 nào đó. Lúc này, giả thiết $\mu = \mu_0$ là *giả thiết chính*, kí hiệu là H_0.

Tuỳ theo những yêu cầu cụ thể, chúng ta sẽ kiểm định giả thiết H_0 đối với một trong ba *đối thiết* sau:

Loại kiểm định	Một phía trái	Một phía phải	Hai phía
Giả thiết chính H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
Đối thiết H_1	$\mu < \mu_0$	$\mu > \mu_0$	$\mu \neq \mu_0$

Chúng ta sẽ lần lượt tìm hiểu mỗi trường hợp trên trong các tiểu mục ngay sau đây.

3.2.1 Kiểm định trung bình μ khi phương sai σ^2 đã biết

Bài toán kiểm định giả thiết về trung bình μ của một phân phối chuẩn khi giá trị σ^2 đã biết dựa trên kết quả sau: Giả sử $X_1, X_2, \ldots, X_n$ là các BNN độc lập có cùng phân phối với X. Khi đó,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

có cùng trung bình μ và có phương sai σ^2/n. Hơn nữa, phân phối chuẩn hoá

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{3.2}$$

có phân phối chuẩn tắc $\mathcal{N}(0,1)$.

Chúng ta sẽ chia làm 3 trường hợp sau đây tương ứng với 3 giả thiết thay thế H_1 như sau:

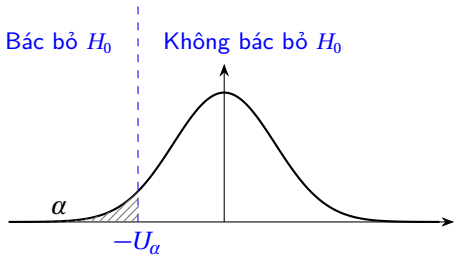
Trường hợp 1: Kiểm định một phía trái

Trong trường hợp này, giả thiết thay thế H_1 là $\mu < \mu_0$. Nếu H_0 đúng, tức là $\mu = \mu_0$, thì theo định lí vừa nêu phía trên, biến ngẫu nhiên Z cho bởi

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{3.3}$$

có phân phối chuẩn tắc $\mathcal{N}(0,1)$. Do đó, với mọi $c \in \mathbb{R}$, chúng ta có thể tính được xác suất

$$P(Z \leq c) = \Phi(c), \tag{3.4}$$



Hình 3.1: Miền bác bỏ cho kiểm định một phía trái: $W_\alpha = \{Z < -U_\alpha\}$

trong đó Φ, như thường lệ, là hàm Laplace (xem Bảng A7 trong Kreyszig [1]).

Từ nhận xét trên, chúng ta thấy rằng với *mức ý nghĩa* α, miền bác bỏ đối với bài toán kiểm định có thể chọn là

$$W_\alpha = \{Z \le -U_\alpha\},$$

trong đó Z cho bởi công thức (3.3) và U_α là phân vị mức α của phân phối chuẩn tắc $\mathcal{N}(0,1)$, đã nói đến ở Chương 2, có nghĩa là

$$\Phi(U_\alpha) = 1-\alpha.$$

Lưu ý, miền bác bỏ W_α có tính chất: *Nếu* H_0 *đúng, thì*

$$\begin{aligned} P(Z \in W_\alpha) &= P(Z \le -U_\alpha) \\ &= \Phi(-U_\alpha) \\ &= 1-\Phi(U_\alpha) \\ &= 1-(1-\alpha) \\ &= \alpha, \end{aligned}$$

như mong muốn. Xem minh họa trong Hình 3.1.

Giả sử bằng thực nghiệm, ta thu được một mẫu dữ liệu x cỡ n với trung bình $\overline{x}$ và tính được giá trị quan sát Z_*, thu được bằng cách thay $\overline{x}$ vào $\overline{X}$ trong (3.3), tức là

$$Z_* = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}.$$

Xét các trường hợp sau:

- Nếu $Z_* \in W_\alpha$ thì chúng ta bác bỏ giả thiết H_0, vì nếu H_0 đúng thì chỉ có xác suất α quan sát được kết quả thuộc vào W_α.

- Nếu $Z_* \notin W_\alpha$ thì chưa có đủ cơ sở để bác bỏ H_0.

Ví dụ 3.1. Giả sử một BNN X có phân phối chuẩn với trung bình μ và phương sai $\sigma^2 = 4$. Hãy kiểm định giả thiết $\mu = 50$ với đối thiết $\mu < 50$, sử dụng mẫu có kích thước $n = 25$ với trung bình mẫu $\overline{x} = 48{,}5$ và mức ý nghĩa $\alpha = 5\%$.

Lời giải

Trong ví dụ này, giả thiết chính H_0 là $\mu = 50$, trong khi đó, giả thiết thay thế H_1 là $\mu < 50$ và kiểm định là *một phía trái*.

Với mức ý nghĩa $\alpha = 5\%$, chúng ta tính được

$$U_\alpha = 1{,}645,$$

dựa trên điều kiện (Bảng A7)

$$\Phi(U_\alpha) = 1 - \alpha = 0{,}95.$$

Như vậy, miền bác bỏ là

$$W_\alpha = \{Z < -1{,}645\}.$$

Giá trị quan sát của $\overline{X}$ là $\overline{x} = 48{,}5$ nên giá trị quan sát của Z là

$$Z_* = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{48{,}5 - 50}{2/\sqrt{25}} = -3{,}75.$$

Do $Z_* \in W_\alpha$ nên chúng ta bác bỏ H_0 và chấp nhận H_1.

Trường hợp 2: Kiểm định một phía phải

Trong trường hợp này, H_1 là $\mu > \mu_0$. Ta vẫn sử dụng công thức (3.3) và lập luận tương tự như trường hợp kiểm định một phía trái, miền bác bỏ lúc này là

$$W_\alpha = \{Z \geq U_\alpha\},$$

thoả mãn $P(Z \in W_\alpha) = \alpha$.

Ví dụ 3.2. Từ một tổng thể có phân phối chuẩn với trung bình μ chưa biết và phương sai $\sigma^2 = 240{,}25$, người ta lấy được một mẫu ngẫu nhiên x gồm 100 quan sát và tính được trung bình mẫu $\overline{x} = 128{,}5$. Với mức ý nghĩa $\alpha = 1\%$, kiểm định cặp giả thiết

$$\begin{cases} H_0 : \mu = 125, \\ H_1 : \mu > 125. \end{cases}$$

Lời giải

Đây là bài toán kiểm định một phía phải cho trung bình μ khi phương sai đã biết. Với mức ý nghĩa $\alpha = 0{,}01$, chúng ta tính được từ điều kiện

$$\Phi(U_\alpha) = 0{,}99$$

giá trị phân vị mức 1% của phân phối chuẩn là $U_\alpha \approx 2{,}33$ và do đó miền bác bỏ là

$$W_\alpha = \{Z > 2{,}33\}.$$

Giá trị quan sát của Z là

$$Z = \frac{128{,}5 - 125}{\sqrt{240{,}25}/\sqrt{100}} \approx 2{,}26.$$

Giá trị quan sát của Z không thuộc W_α nên chúng ta không có đủ cơ sở thống kê để bác bỏ H_0.

Trường hợp 3: Kiểm định hai phía

Bằng những lập luận tương tự như Trường hợp 1 & 2, trong trường hợp kiểm định hai phía, H_1 là $\mu \neq \mu_0$, với Z cho bởi (3.3), miền bác bỏ là

$$W_\alpha = \{|Z| \geq U_{\alpha/2}\}.$$

Như vậy, nếu H_0 đúng thì có $\alpha/2$ khả năng quan sát được $Z_* > U_{\alpha/2}$ và cũng có $\alpha/2$ khả năng quan sát được $Z_* < -U_{\alpha/2}$.

Ví dụ 3.3. Một công ty sản xuất bóng đèn cho biết tuổi thọ trung bình của sản phẩm là 1200 giờ, với độ lệch chuẩn tổng thể $\sigma = 100$ giờ. Một nhóm kiểm định viên nghi ngờ giá trị này không chính xác, nên lấy mẫu ngẫu nhiên gồm $n = 64$ bóng đèn và thu được trung bình mẫu $\bar{x} = 1170$ giờ.

Với mức ý nghĩa $\alpha = 0{,}05$, hãy kiểm định xem tuổi thọ trung bình thực sự có khác 1200 giờ hay không.

Lời giải

Trong tình huống này, cặp giả thuyết kiểm định sẽ là

$$\begin{cases} H_0 : \mu = 1200, \\ H_1 : \mu \neq 1200. \end{cases}$$

Đây là kiểm định hai phía.

Vì phương sai tổng thể đã biết, nên chúng ta sử dụng kiểm định Z:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Với mức ý nghĩa $\alpha = 0{,}05$, ta có:

$$U_{\alpha/2} = U_{0{,}025} = 1{,}96.$$

Miền bác bỏ cho kiểm định hai phía lúc này sẽ là

$$W_\alpha = \{|Z| > 1{,}96\}.$$

Giá trị quan sát tính được bằng cách thay số:

$$Z = \frac{1170 - 1200}{100/\sqrt{64}} = \frac{-30}{12{,}5} = -2{,}4.$$

Kết luận: Vì $|Z| = 2{,}4 > 1{,}96$ nên Z_* thuộc vào miền bác bỏ, ta *bác bỏ giả thuyết* H_0: ở mức ý nghĩa 5%, có đủ bằng chứng để kết luận rằng tuổi thọ trung bình của bóng đèn *khác 1200 giờ*.

Tóm tắt quy trình kiểm định μ khi σ^2 đã biết

Với giả thiết chính $H_0 : \mu = \mu_0$, tiêu chuẩn kiểm định

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}},$$

việc tìm miền bác bỏ trong 3 trường hợp nêu trên được tóm tắt trong bảng sau:

Loại kiểm định	Một phía trái	Một phía phải	Hai phía
Giả thiết chính H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
Đối thiết H_1	$\mu < \mu_0$	$\mu > \mu_0$	$\mu \neq \mu_0$
Miền bác bỏ W_α	$\{Z \leq -U_\alpha\}$	$\{Z \geq U_\alpha\}$	$\{\|Z\| \geq U_{\alpha/2}\}$

Bảng 3.1: Kiểm định z: Miền bác bỏ trong các trường hợp của H_0 và H_1

3.2.2 *Kiểm định trung bình μ khi phương sai σ^2 chưa biết*

Trong tiểu mục này, chúng ta tìm hiểu bài toán kiểm định trung bình của phân phối chuẩn với phương sai chưa biết, gọi là *kiểm định t*. Bài toán này xuất hiện nhiều trong thực tế hơn so với bài toán đã biết trong tiểu mục trước, bởi vì phương sai của tổng thể nhìn chung không biết trước được.

Cơ sở lí thuyết của phương pháp kiểm định t là Định lí 2.4, phát biểu lại như sau: *Nếu $X_1, X_2, \ldots, X_n$ là các BNN độc lập có phân phối chuẩn và có trung bình $\mu = \mu_0$ và phương sai σ^2 thì BNN*

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \tag{3.5}$$

có phân phối t[1] *với* $m = n-1$ *bậc tự do, trong đó*

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n). \tag{3.6}$$

Trong bài toán này, chúng ta thực hiện quy trình kiểm định tương tự như trong §3.2, với tiêu chuẩn T cho bởi (3.5), sử dụng phương sai mẫu và phân phối t với $n-1$ bậc tự do (ở đây, n là cỡ mẫu). Miền bác bỏ lúc này được xác định qua Bảng 3.2.

Loại kiểm định	Một phía trái	Một phía phải	Hai phía
Giả thiết chính H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
Đối thiết H_1	$\mu < \mu_0$	$\mu > \mu_0$	$\mu \neq \mu_0$
Miền bác bỏ W_α	$\{T \leq -t_\alpha\}$	$\{T \geq t_\alpha\}$	$\{\lvert T\rvert \geq t_{\alpha/2}\}$

Bảng 3.2: Kiểm định t: Miền bác bỏ trong các trường hợp của H_0 và H_1

Ví dụ 3.4. Theo báo cáo của nhà sản xuất, mỗi tủ lạnh nhãn hiệu A tiêu thụ khoảng 46 (kWh/năm). Từ một mẫu gồm 25 người dùng được nghiên cứu, cho thấy tủ lạnh nhãn hiệu A tiêu thụ trung bình 42,5 (kWh/năm) với độ lệch chuẩn 11,9 kWh. Với mức ý nghĩa 5%, hãy kiểm định báo cáo của nhà sản xuất, với giả sử tổng thể đang xét có phân phối chuẩn.

Lời giải

Giả thiết H_0 là $\mu = 46$, trong khi đó, đối thiết H_1 là $\mu < 46$, kiểm định là một phía trái.

Mẫu quan sát gồm $n = 25$ điểm mẫu với trung bình mẫu $\overline{x} = 42{,}5$ và độ lệch chuẩn $s = 11{,}9$. Do phương sai của tổng thể chưa biết, nên chúng ta dùng kiểm định t,

$$T = \frac{\overline{X} - 46}{S/\sqrt{25}}.$$

Với mức ý nghĩa $\alpha = 5\%$, chúng ta tính được giá trị tới hạn từ phân phối t với $n-1 = 24$ bậc tự do, cụ thể là

$$t_\alpha = 1{,}71.$$

Miền bác bỏ, theo Bảng 3.2, là

$$W_\alpha = \{T < -t_\alpha\} = \{T < -1{,}71\}.$$

Với những thông số đã biết, ta tính được $T_* = (\overline{x} - 46)/(11{,}9/\sqrt{25}) = -1{,}47$. Nhận thấy $T_* \notin W_\alpha$, và do đó chưa đủ cơ sở để bác bỏ H_0.

[1] Còn gọi là phân phối Student.

Ví dụ 3.5. Một công ty dược phẩm tiến hành kiểm định hiệu quả của một loại thuốc trị cúm mới. Thuốc được thử nghiệm trên 120 bệnh nhân được chọn ngẫu nhiên. Kết quả cho thấy thời gian trung bình mà các bệnh nhân còn xuất hiện triệu chứng cúm là 5,5 ngày, với độ lệch chuẩn 1,2 ngày. Được biết rằng, nếu không dùng thuốc, thời gian trung bình người bị cúm có triệu chứng là 7 ngày. Hãy kiểm định ở mức ý nghĩa $\alpha = 0{,}05$ xem loại thuốc mới này có thực sự làm giảm thời gian bị cúm hay không.

Lời giải

Phân tích tình huống giả định trong bài ra, chúng ta thấy yêu cầu là kiểm định giả thiết $H_0 : \mu = 7$ so với đối thiết $H_1 : \mu < 7$, trong đó μ là thời gian trung bình một người bị cúm và dùng thuốc đang muốn kiểm nghiệm. Nếu kết quả của kiểm định là bác bỏ H_0 thì ta chấp nhận giả thiết *thuốc đang kiểm định là có tác dụng* (vì nó làm ngắn thời gian có triệu chứng trung bình của người bị mắc cúm).

Chúng ta sẽ dùng kiểm định t với miền bác bỏ là

$$W_\alpha = \left\{ T = \frac{\overline{X} - 7}{S/\sqrt{120}} \;\middle|\; T < -t_\alpha \right\}.$$

Giá trị quan sát của thống kê T có được từ thử nghiệm trên nhận được bằng cách thay $\overline{X} = 5{,}5$ và $S = 1{,}2$:

$$T_* = \frac{5{,}5 - 7}{1{,}2\sqrt{120}} = -13{,}6931.$$

Mặt khác, giá trị t_α, với $\alpha = 0{,}05$, tính từ phân phối t với 119 bậc tự do là

$$t_\alpha = 1{,}6578.$$

Nhận thấy $T_* \in W_\alpha$ nên bác bỏ H_0 và chấp nhận H_1, nghĩa là chấp nhận giả thiết thuốc có tác dụng.

Ví dụ 3.6. Theo một báo cáo, chiều cao trung bình của cây lúa giống ST25 là 98 cm. Để kiểm chứng tính chính xác của báo cáo này, một nhóm nghiên cứu đã tiến hành đo ngẫu nhiên $n = 50$ cây lúa cùng giống. Kết quả cho thấy mẫu thu được có trung bình $\bar{x} = 100$ cm và độ lệch chuẩn mẫu $s = 4{,}5$ cm.

Với mức ý nghĩa $\alpha = 0{,}05$, hãy kiểm định giả thuyết rằng chiều cao trung bình thực sự của giống lúa ST25 *lớn hơn 100 cm* so với giá trị báo cáo.

Lời giải

Ta bắt đầu với việc thiết lập cặp giả thuyết

$$\begin{cases} H_0 : \mu = 98, \\ H_1 : \mu > 100. \end{cases}$$

Đây là kiểm định một phía (phía phải).

Tiếp theo, để xác định thống kê kiểm định, lưu ý rằng vì phương sai tổng thể chưa biết, nên ta sử dụng thống kê kiểm định theo phân phối t

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

trong đó: $\bar{x} = 100$, $\mu_0 = 98$, $s = 4{,}5$, $n = 50$. Để xác định miền bác bỏ, lưu ý rằng bậc tự do là $\text{df} = n - 1 = 49$. Với $\alpha = 0{,}05$ và $\text{df} = 49$, giá trị tới hạn là:

$$t_{0,95} \approx 1{,}676.$$

Vì kiểm định phía phải nên miền bác bỏ sẽ là

$$W_\alpha = \{T > 1{,}676\}.$$

Từ mẫu quan sát được, ta tính giá trị thống kê

$$T = \frac{100 - 98}{4{,}5/\sqrt{50}} = \frac{2}{0{,}6364} \approx 3{,}142.$$

Vì $T_* = 3{,}142 > 1{,}676$, ta *bác bỏ* H_0: Ở mức ý nghĩa 5%, có đủ bằng chứng để kết luận rằng *chiều cao trung bình của cây lúa giống ST25 lớn hơn 98 cm.*

3.2.3 Giá trị p

Tiểu mục này, ta tìm hiểu về giá trị p. Khái niệm này khá quan trọng, nhất là khi các phần mềm thống kê thường đưa ra kết quả của bài toán kiểm định bằng cách báo cáo giá trị p.

Trong các bài toán kiểm định cặp giả thiết (H_0, H_1) bằng thống kê T và một mẫu x, người ta có thể trình bày kết quả bằng **giá trị** p (p-value). Ở đây, thay vì xây dựng một miền bác bỏ W_α, chúng ta tính xác suất, với giả thiết H_0 đúng, của sự kiện thu được một mẫu có giá trị quan sát T_* bằng hoặc "cực đoan" hơn giá trị quan sát T_* của mẫu đã thu thập được. Vấn đề này có vẻ khá đơn giản, nhưng lại khá tế nhị! Lưu ý bạn đọc rằng một số tác giả giải thích một cách đại khái rằng trong bước này ta tính toán xác suất dữ liệu xảy ra, nhưng không nói đến thống kê T. Điều này có vẻ khó hiểu vì thật ra nó sai: Đối với các mô hình liên tục, xác suất một dữ liệu đơn lẻ xảy ra luôn bằng 0!

Giá trị p chỉ có ý nghĩa khi gắn với một thống kê T. Thống kê T đóng vai trò là một thước đo giúp chúng ta xác định được sự kiện gồm các mẫu "cực đoan hơn" mẫu đã thu thập được. Chính sự kiện này là sự kiện mà ta cần tính xác suất để thu được giá trị p.

Đối với phương pháp giá trị p, bạn đọc chỉ cần nhớ tính chất sau: *Giá trị quan sát T_* nằm trong miền bác bỏ W_α, tức là $T_* \in W_\alpha$, nếu và chỉ nếu giá trị p thoả mãn $p \le \alpha$.*

Kết quả của bài toán kiểm định được báo cáo một cách cụ thể hơn thông qua độ lớn của giá trị p.

Chúng ta hãy xem ví dụ sau trong R.

CAS 3.1. Đo nhiệt độ nóng chảy T của kẽm, người ta thu được các mẫu đo như sau: 419,9; 420,4; 418,9; 419,8; 419,7; 419; 418,5; 417,8; 421,6; 421,9. Với mức ý nghĩa 5%, kiểm định giả thiết nhiệt độ nóng chảy trung bình của kẽm bằng 419.

Đoạn mã R sau đây thực hiện kiểm định t hai phía và trình bày kết quả theo dạng giá trị p.

```
# Đưa dữ liệu vào biến x
x = c(419.9, 420.4, 418.9, 419.8, 419.7, 419.0, 418.5, 417.8,
     421.6, 421.9)

# Chạy t.test() với tham số mu = 419
t.test(x, mu = 419)

##
##  One Sample t-test
##
## data:  x
## t = 1,828, df = 9, p-value = 0,101
## alternative hypothesis: true mean is not equal to 419
## 95 percent confidence interval:
##  418,822 420,678
## sample estimates:
## mean of x
##    419,75
```

Kết quả cho thấy giá trị p của kiểm định bằng 0,101 (`p-value = 0.101`), lớn hơn mức ý nghĩa $\alpha = 0{,}05$. Điều này có nghĩa là không đủ cơ sở để bác bỏ giả thiết H_0.

Bài tập §3.2

1. Một nông trại cho biết khối lượng trung bình của mỗi quả bưởi là 450 g. Để kiểm tra thông tin này, người ta lấy ngẫu nhiên 51 quả bưởi và đo được trung bình mẫu là 445 g với độ lệch chuẩn mẫu 30 g. Hãy kiểm định ở mức ý nghĩa $\alpha = 0{,}01$ xem khối lượng trung bình thực sự của bưởi có nhỏ hơn 450 g (tức là bưởi nhẹ hơn báo cáo) hay không.

2. Khảo sát về chiều cao X của trẻ nam 6 tuổi, người ta thu được mẫu dữ liệu sau:

x_i (cm)	$104-108$	$108-112$	$112-116$	$116-120$	$120-124$
Số mẫu	8	11	14	11	6

Giả sử là chiều cao X là một BNN tuân theo phân phối chuẩn. Với mức ý nghĩa 5%, hãy kiểm định giả thiết H_0 phát biểu rằng chiều cao trung bình của các học sinh đó là $\mu_0 = 110$ (cm) với đối thiết H_1 phát biểu rằng chiều cao trung bình lớn hơn 110 (cm).

3. Một dây chuyền chế tạo một chi tiết máy với độ dài 1 cm. Nghi ngờ dây chuyền hoạt động không bình thường, người ta chọn ra một mẫu ngẫu nhiên gồm 100 sản phẩm thì thấy như sau:

x_i (cm)	0,95	0,97	0,99	1,01	1,03	1,05
Số mẫu	8	30	38	19	3	2

Với mức ý nghĩa 5%, có thể nói gì về độ dài trung bình của chi tiết máy chế tạo bởi dây chuyền đó?

3.3 Kiểm định cho phương sai của phân phối chuẩn

Mục này nói về bài toán kiểm định phương sai cho phân phối chuẩn. Tài liệu tham khảo: Đ. H. Thắng [2, Chương 4, §6.] , Kreyszig [1, page 1084].

Giả sử BNN X có phân phối chuẩn $\mathcal{N}(\mu,\sigma^2)$. Nếu chúng ta có cơ sở để giả thiết rằng *giá trị của phương sai* σ^2 *bằng* σ_0^2, tức chúng ta có giả thiết thống kê

$$H_0 : \sigma^2 = \sigma_0^2. \tag{3.7}$$

Chúng ta tìm hiểu bài toán kiểm định giả thiết H_0 thông qua một mẫu ngẫu nhiên. Cơ sở thống kê toán học cho quy trình kiểm định này là Định lí 2.6 trong mục trước nói rằng: *Giả sử* $X_1,\ldots,X_n$ *là dãy các biến ngẫu nhiên độc lập và có cùng phân phối với* X. *Đặt*

$$S^2 = \frac{1}{n-1}\left[(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2\right]. \tag{3.8}$$

Khi đó, nếu X *có phương sai* σ_0^2, *thì BNN*

$$T = (n-1)\frac{S^2}{\sigma_0^2}, \tag{3.9}$$

có phân phối χ^2 *(chi-bình phương) với* $(n-1)$ *bậc tự do.*

Chú ý 3.1. Lưu ý, phân phối xác suất chi-bình phương (χ^2) đã được nói đến trong §1.9.5.

Từ định lí vừa nói, người ta xây dựng quy trình kiểm định giả thiết $H_0 : \sigma = \sigma_0$ đối với một trong 3 đối thiết H_1 dựa trên phân phối χ^2. Quy trình kiểm định này tương tự như quy trình kiểm định trung bình trong đó thống kê T cho bởi (3.9) và cách xác định miền bác bỏ W_α được tóm tắt như trong Bảng 3.3.

Loại kiểm định	Một phía trái	Một phía phải	Hai phía
Giả thiết H_0	$\sigma = \sigma_0$	$\sigma = \sigma_0$	$\sigma = \sigma_0$
Đối thiết H_1	$\sigma < \sigma_0$	$\sigma > \sigma_0$	$\sigma \neq \sigma_0$
Miền bác bỏ W_α	$\{T \leq \chi^2_{1-\alpha}\}$	$\{T \geq \chi^2_{1\alpha}\}$	$\{T \leq \chi^2_{1-\alpha/2}$ hoặc $T > \chi^2_{\alpha/2}\}$

Bảng 3.3: Miền bác bỏ W_α trong kiểm định phương sai

Ở đây, tiêu chuẩn kiểm định T cho bởi công thức (3.9), còn giá trị χ^2_β là giá trị tới hạn tính từ phân phối χ^2 với $n-1$ bậc tự do (phân vị mức β của phân phối chi-bình phương với $n-1$ bậc tự do), có nghĩa là

$$F(\chi^2_\beta) = 1-\beta, \quad 0 < \beta < 1. \tag{3.10}$$

Giá trị tới hạn của phân phối χ^2 có thể tìm thấy trong Kreyszig [1, Bảng A10] hoặc bằng phần mềm R, xem CAS 3.2.

Bây giờ, với một mẫu ngẫu nhiên $\overline{x}$, chúng ta tính giá trị quan sát

$$T_* = (n-1)\frac{s^2}{\sigma_0^2},$$

và tuỳ vào việc giá trị quan sát này có thuộc vào miền bác bỏ hay không mà đưa ra quyết định bác bỏ hay chấp nhận H_0.

Ví dụ 3.7. Giả sử chúng ta có một mẫu ngẫu nhiên kích thước $n = 25$ và phương sai mẫu $s^2 = 15$ từ một tổng thể có phân phối chuẩn. Hãy kiểm định giả thiết $\sigma^2 = \sigma_0^2 = 12$ so với đối thiết $\sigma^2 > 12$, với mức ý nghĩa $\alpha = 5\%$.

Lời giải

Giả thiết H_0: $\sigma^2 = 12$ và $H_1 : \sigma^2 > 12$, kiểm định là *một phía phải*.

Với mức ý nghĩa $\alpha = 0{,}05$, giá trị tới hạn χ^2_α của phân phối χ^2 với 24 bậc tự do được tính từ điều kiện

$$F(\chi^2_\alpha) = 1-\alpha = 0{,}95.$$

Tra Bảng A10 trong Kreyszig [1] với $(n-1) = 24$ bậc tự do, hoặc sử dụng phần mềm R, chúng ta thu được $\chi^2_\alpha \approx 36{,}4$. Từ Bảng 3.3, miền bác bỏ sẽ là

$$W_\alpha = \left\{ T = (n-1)\frac{S^2}{\sigma_0^2} \,\middle|\, T > 36{,}4 \right\}.$$

Bây giờ, giá trị quan sát từ mẫu là

$$T_* = (25-1) \times \frac{15}{12} = 30.$$

Kết luận: $T_* \notin W_\alpha$ và vậy thì chấp nhận H_0.

CAS 3.2. Trong R, chúng ta có thể tìm giá trị tới hạn χ^2_α bằng chức năng `qchisq()`. Ví dụ, với $\alpha = 0{,}05$ và $1-\alpha = 0{,}95$, bậc tự do df $= 24$, ta thực hiện như sau:

```
# Yêu cầu tính chính xác đến 6 chữ số
options(digits = 8)

# Tính phân vị mức của phân phối chi-bình phương
qchisq(0.95, df = 24)

## [1] 36,415029
```

Kết quả thu được, với $\alpha = 0{,}95$, là $\chi^2_\alpha \approx 36{,}415028$, với 8 chữ số có nghĩa.

CAS 3.3. Cho mẫu dữ liệu $x = (14{,}5, 16{,}2, 14{,}5, 19{,}4, 14{,}8, 17{,}4, 15{,}1, 19{,}8, 16{,}8)$ lấy từ một phân phối chuẩn. Sử dụng `varTest()` (đã dùng trong CAS 2.9) trong gói EnvStats, chúng ta sẽ kiểm định giả thiết $H_0 : \sigma^2 = 12{,}5$ so với đối thiết $H_1 : \sigma^2 < 12{,}5$. Tham số cho hàm `varTest()` sẽ bao gồm `sigma.squared = 12.5` (cho H_0) và `alternative = "less"` (cho H_1).

```
# Sử dụng gói EnvStats
library(EnvStats)

# Mẫu dữ liệu
x = c(14.5, 16.2, 14.5, 19.4, 14.8, 17.4, 15.1, 19.8, 16.8)

# Hàm varTest()
varTest(x, sigma.squared = 12.5, alternative = "less")

## $statistic
## Chi-Squared
##      2,6512
##
## $parameters
## df
##  8
##
## $p.value
## [1] 0,045692869
##
## $estimate
## variance
##   4,1425
##
```

```
## $null.value
## variance
##     12,5
##
## $alternative
## [1] "less"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "x"
##
## $conf.int
##       LCL       UCL
##  0,000000 12,127481
## attr(,"conf.level")
## [1] 0,95
##
## attr(,"class")
## [1] "htestEnvStats"
```

Kết quả cho thấy giá trị p là 0,0457; với mức ý nghĩa 5% chúng ta bác bỏ H_0 và chấp nhận giả thiết $\sigma^2 < 12.5$.

Bài tập §3.3

1. Một kĩ sư kiểm tra độ chính xác của đường kính bi thép dùng trong ổ trục. Dữ liệu thu được từ một mẫu gồm 6 viên bi (đơn vị: mm) như sau:

$$10{,}2 \quad 10{,}1 \quad 10{,}3 \quad 10{,}0 \quad 10{,}4 \quad 10{,}2$$

Giả sử đường kính bi thép tuân theo phân phối chuẩn. Trên thông tin kĩ thuật, nhà sản xuất ghi rằng độ lệch chuẩn của đường kính là $\sigma = 0{,}15$ (mm). Hãy kiểm định thông tin này đối với đối thiết nói rằng độ lệch chuẩn $\sigma > 0{,}15$.

2. Cho biết cân nặng của dê con mới sinh tuân theo phân bố chuẩn. Mẫu theo các lớp (đơn vị: g):

Cân nặng	2500-2700	2700-2900	2900-3100	3100-3300	3300-3500
Số dê con	12	28	40	30	10

Lấy điểm giữa của mỗi lớp làm giá trị đại diện: 2600, 2800, 3000, 3200, 3400. Với mức ý nghĩa $\alpha = 0{,}05$, kiểm định giả thiết

$$H_0 : \sigma = 225 \text{ g} \qquad \text{với đối thiết} \qquad H_1 : \sigma < 225 \text{ g}.$$

3. Cùng dữ kiện như Bài 2, kiểm định giả thiết $H_0 : \sigma = 220$ so với đối thiết $H_1 : \sigma \neq 220$.

3.4 So sánh trung bình và phương sai hai phân phối chuẩn

Trong mục này, chúng ta nói về bài toán so sánh các tham số của hai phân phối chuẩn. Tài liệu tham khảo: Kreyszig [1, Chapter 25.4], Đ. H. Thắng [2, tr. 133–136].

3.4.1 So sánh trung bình hai phân phối chuẩn

Giả sử chúng ta có hai BNN với phân phối chuẩn X và Y: $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ và $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Hơn nữa, giả sử rằng $\sigma_X = \sigma_Y$ (chúng có thể chưa biết). Nếu có cơ sở để giả thiết rằng hai kì vọng μ_X và μ_Y là bằng nhau, thì nảy sinh bài toán kiểm định giả thiết thống kê

$$H_0 : \mu_X = \mu_Y$$

đối với các đối thiết

- $H_1 : \mu_X < \mu_Y$, hoặc
- $H_1 : \mu_X > \mu_Y$, hoặc
- $H_1 : \mu_X \neq \mu_Y$.

Chúng ta sử dụng hai mẫu tương ứng: mẫu $(x_1, \ldots, x_m)$ có kích thước m ứng với các biến ngẫu nhiên $(X_1, \ldots, X_m)$ độc lập và có cùng phân phối với X, và mẫu $(y_1, \ldots, y_n)$ có kích thước n tương ứng với các BNN $(Y_1, \ldots, Y_n)$ độc lập và có cùng phân phối với Y.

Như trước, chúng ta đặt

$$\overline{X} = \frac{1}{m}(X_1 + \cdots + X_m), \quad S_X^2 = \frac{1}{m-1}\left[(X_1 - \bar{X})^2 + \cdots + (X_m - \bar{X})^2\right].$$

Tương tự như vậy, ta cũng đặt

$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n), \quad S_Y^2 = \frac{1}{n-1}\left[(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2\right].$$

Hơn nữa, đặt

$$S = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}. \tag{3.11}$$

(S^2 gọi là *phương sai gộp*).

Chúng ta sẽ sử dụng tiêu chuẩn kiểm định

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}. \tag{3.12}$$

Cơ sở thống kê toán học cho bài toán kiểm định sẽ nói sau đây là định lí nói rằng nếu giả thiết H_0 đúng thì T là biến ngẫu nhiên có phân phối t với df $= m+n-2$ bậc tự do. Từ đó, chúng ta xây dựng quy trình kiểm định tương tự như trên, với miền bác bỏ W_α được cho bởi Bảng 3.4 trong đó thống kê T cho bởi (3.12) và t_α là giá trị tới hạn trong phân phối t.

Loại kiểm định	Một phía trái	Một phía phải	Hai phía
Giả thiết H_0	$\mu_X = \mu_Y$	$\mu_X = \mu_Y$	$\mu_X = \mu_Y$
Đối thiết H_1	$\mu_X < \mu_Y$	$\mu_X > \mu_Y$	$\mu_X \neq \mu_Y$
Miền bác bỏ W_α	$\{T \leq -t_\alpha\}$	$\{T \geq t_\alpha\}$	$\{\|T\| \geq t_{\alpha/2}\}$

Bảng 3.4: Miền bác bỏ W_α trong bài toán so sánh trung bình

Ví dụ 3.8. Xét hai mẫu ngẫu nhiên từ hai tổng thể tuân theo phân phối chuẩn với phương sai bằng nhau:

x:	108	95	105	102	103	99	98	104	97	109
y:	96	92	91	102	105	94	100	90	95	96

Với mức ý nghĩa 5%, có thể nói hai tổng thể trên có trung bình như nhau hay không?

Lời giải

Hai mẫu dữ liệu đã cho gồm $m = n = 10$ mẫu. Để tính giá trị quan sát T_* của thống kê T định nghĩa bởi (3.12), trước hết chúng ta tính trung bình và phương sai của x và y. Cụ thể, chúng ta có

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{7} x_i = 102, \quad s_x^2 = 22,$$

trong khi đó,

$$\overline{y} = 96{,}1, \quad s_Y^2 \approx 23{,}88$$

Tiếp theo, s được tính theo công thức (3.11) là

$$s = \sqrt{\frac{9\times 22{,}0 + 9\times 23{,}88}{18}} \approx 4{,}79.$$

Cuối cùng, giá trị T_* được tính bởi

$$T_* = \frac{102{,}0 - 96{,}1}{4{,}79\times\sqrt{\frac{1}{10}+\frac{1}{10}}} \approx 2{,}75.$$

Bước tiếp theo, với mức ý nghĩa $\alpha = 5\%$ chúng ta tính được $t_{\alpha/2}$ từ bảng phân phối t dựa trên điều kiện

$$F(t_{\alpha/2}) = 1 - \alpha/2 = 0{,}975,$$

trong đó F là hàm phân phối t với $m + n - 2 = 18$ bậc tự do. Cụ thể là $t_{\alpha/2} \approx 2{,}10$.

Kiểm định là hai phía nên miền bác bỏ là

$$W_\alpha = \{|T| > 2{,}10\}$$

(xem dòng cuối cùng của Bảng 3.4). Nhận thấy $T_* \in W_\alpha$ nên chúng ta bác bỏ giả thiết H_0.

CAS 3.4. Phép kiểm định so sánh trung bình hai phân phối chuẩn khi chưa biết phương sai gọi là *phép kiểm định t hai mẫu* (Two Sample t-test). Trong R, phép kiểm định này thực hiện như sau.

```
# Đưa các mẫu vào hai biến x và y
x = c(108, 95, 105, 102, 103, 99, 98, 104, 97, 109)
y = c(96, 92, 91, 102, 105, 94, 100, 90, 95, 96)

# Áp dụng chức năng t.test cho hai mẫu
t.test(x, y, paired = FALSE, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  x and y
## t = 2,75455, df = 18, p-value = 0,013045
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1,4000144 10,3999856
## sample estimates:
## mean of x mean of y
##     102,0      96,1
```

Kết quả tính toán, $T_* = 2{,}75$. Giá trị `p-value = 0,01305` nhỏ hơn α và do đó bác bỏ giả thiết H_0.

Nhận xét 3.1. Trong trường hợp cỡ mẫu bằng nhau, $m = n$, công thức của T viết lại một cách đơn giản hơn là

$$T = \sqrt{n}\frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2 + S_Y^2}}.$$

3.4.2 So sánh phương sai hai phân phối chuẩn

Vấn đề so sánh phương sai của hai phân phối chuẩn được giải quyết dựa trên phân phối xác suất Fisher, thực hiện như trong ví dụ sau.

Ví dụ 3.9. Chúng ta hãy kiểm định giả thiết $\sigma_X^2 = \sigma_Y^2$ của hai phân phối chuẩn X và Y, so với đối thiết $\sigma_X^2 < \sigma_Y^2$, sử dụng dữ liệu trong Ví dụ 3.8.

Lời giải

Với dữ liệu này, tính được $s_x^2 = 22$ và $s_y^2 \approx 23{,}88$. Do đó, $v_0 = s_y^2/s_x^2 \approx 1{,}09$. Chọn mức ý nghĩa $\alpha = 5\%$, sử dụng $P(V \le F_\alpha) = 1-\alpha = 0{,}95$, từ bảng phân phối F (phân phối Fisher) với $(9,9)$ bậc tự do, chúng ta có $F_\alpha \approx 3{,}18$. Do $v_0 < F_\alpha$, chấp nhận giả thiết H_0: "$\sigma_x^2 = \sigma_y^2$".

CAS 3.5. Bài toán so sánh hai phương sai được gọi là *kiểm định F*[1] Trong R, chúng ta thực hiện như sau:

```
# Đưa các dữ liệu mẫu vào x và y
x = c(108, 95, 105, 102, 103, 99, 98, 104, 97, 109)
y = c(96, 92,91, 102, 105, 94, 100, 90, 95, 96)

# Áp dụng var.test() cho hai mẫu
var.test(x, y)

##
##  F test to compare two variances
##
## data:  x and y
## F = 0,921359, num df = 9, denom df = 9, p-value = 0,90489
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0,22885249 3,70938503
## sample estimates:
## ratio of variances
##         0,92135877
```

Kết quả thu được *giá trị p* rất lớn, xấp xỉ 0,9049, nên chúng ta sẽ chấp nhận H_0.

Bài tập §3.4

1. Nếu sự giống nhau và khác nhau của bài toán so sánh trung bình ghép cặp và không ghép cặp.
2. Kiểm tra đường kính các viên thuốc do hai máy dập viên thuốc cho kết quả như sau:

[1] F-test với chữ F là chữ cái đầu của tên họ nhà thống kê Ronald Fisher.

	Máy 1	**Máy 2**
Trung bình	$\overline{x}_1 = 5{,}44$	$\overline{x}_2 = 5{,}52$
Phương sai	$s_1^2 = 0{,}0095$	$s_2^2 = 0{,}0105$
Số mẫu	$n_1 = 28$	$n_2 = 30$

Với mức ý nghĩa 5%, có thể nói đường kính trung bình của các viên thuốc do hai máy dập là như nhau hay không?

3. Tìm hiểu về thời gian đi học (học phổ thông hoặc cao hơn) trung bình của người trưởng thành người ta thu được dữ kiện sau:

	Nam	**Nữ**
Trung bình	$\overline{x}_1 = 13{,}7$ (năm)	$\overline{x}_2 = 14{,}1$ (năm)
Độ lệch chuẩn	$s_1 = 3{,}04$	$s_2 = 2{,}72$
Số mẫu	$n_1 = 565$	$n_2 = 425$

Với mức ý nghĩa 1%, có thể nói nữ giới có thời gian được đi học dài hơn nam giới hay không?

4. Tìm hiểu về điểm trung bình của học sinh trong một kì thi người ta thu được hai mẫu từ hai nhóm học sinh như sau:

Nhóm 1:	70	95	72	88	79	80	68	55	73	82
Nhóm 2:	96	90	71	82	70	85	98	77	52	65

Giả định rằng các mẫu được lấy ra từ hai nhóm học sinh có điểm thi tuân theo phân phối chuẩn. Với mức ý nghĩa 1%, có thể nói hai nhóm học sinh có điểm thi trung bình giống nhau hay không?

3.5 Kiểm định mức độ phù hợp của mô hình

Kiểm định mức độ phù hợp với phân phối xác suất (Goodness-of-fit test) là phương pháp thống kê để xác định xem một tập dữ liệu có phù hợp với một phân phối xác suất nhất định hay không. Mục tiêu là kiểm tra xem dữ liệu có thể được mô tả bằng một phân phối xác suất cụ thể (ví dụ: phân phối chuẩn, phân phối Poisson, v.v.) hay không.

3.5.1 Kiểm định chi-bình phương

Mục này, chúng ta tìm hiểu phương pháp chi-bình phương (χ^2 test) nhằm kiểm tra xem dữ liệu phân loại (categorical data) có phù hợp với một phân phối lí thuyết hay không.

Phân phối đều

Sau đây là một ví dụ về kiểm định chi-bình phương cho một dữ liệu xem có phù hợp với mô hình phân phối đều rời rạc hay không.

Ví dụ 3.10. Giả sử bạn có một tập dữ liệu về số lần xuất hiện của 6 mặt xúc xắc trong 60 lần gieo xúc xắc, và bạn muốn kiểm định xem dữ liệu này có phù hợp với phân phối đồng đều (uniform distribution), tức là mỗi mặt xúc xắc có khả năng xuất hiện như nhau.

Mặt xúc xắc	Số lần xuất hiện
1	10
2	8
3	9
4	12
5	11
6	10

Thiết lập giả thiết vô hiệu và giả thiết thay thế:

– giả thiết H_0: Dữ liệu phù hợp với phân phối đồng đều.

– giả thiết H_1: Dữ liệu không phù hợp với phân phối đồng đều.

Ta tính số lần xuất hiện kì vọng cho mỗi mặt xúc xắc: $E = \frac{60}{6} = 10$.

Sau đó, sử dụng công thức chi-bình phương để tính thống kê kiểm định:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

Trong đó O_i là tần số quan sát và E_i là tần số kì vọng.

Sau khi tính ra giá trị chi-bình phương, so sánh với giá trị tới hạn (critical value) từ bảng phân phối chi-bình phương để kết luận xem có bác bỏ giả thiết H_0 hay không. Tính toán chi tiết như sau:

$$\begin{aligned}\chi^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(10-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(10-10)^2}{10} \\ &= \frac{0}{10} + \frac{4}{10} + \frac{1}{10} + \frac{4}{10} + \frac{1}{10} + \frac{0}{10} = 1.\end{aligned}$$

Bậc tự do $\text{df} = k-1 = 6-1 = 5$. Với mức ý nghĩa $\alpha = 0{,}05$ và bậc tự do $\text{df} = 5$, từ bảng phân phối chi-bình phương, giá trị tới hạn là 11,07. Vì giá trị $\chi^2 = 1{,}0$ nhỏ hơn giá trị tới hạn 11,07 chúng ta không bác bỏ giả thiếtH_0. Điều này có nghĩa là không có bằng chứng đủ mạnh để cho rằng viên xúc xắc này không công bằng.

Phân phối chuẩn

Dưới đây, chúng ta sử dụng sự khác biệt giữa số tần suất thực nghiệm (quan sát) và số tần suất kì vọng (tính toán dựa trên phân phối chuẩn) để kiểm tra giả thiết rằng dữ liệu tuân theo phân phối chuẩn. Cụ thể hơn, chúng ta kiểm định giả thiết H_0: "Dữ liệu tuân theo phân phối chuẩn" so với đối thiết H_1: "Dữ liệu không tuân theo phân phối chuẩn".

Các bước thực hiện:

- Phân nhóm dữ liệu: Dữ liệu được phân chia thành các nhóm (bins), và số lượng phần tử trong mỗi nhóm sẽ được tính toán dựa trên dữ liệu thực tế (số liệu quan sát). Bước này cần thiết khi kiểm định các mô hình phân phối liên tục.

- Tính số liệu kì vọng: Trong mỗi nhóm, số liệu kì vọng được tính toán dựa trên phân phối chuẩn lí thuyết với các tham số như trung bình và độ lệch chuẩn của dữ liệu.

- Tính giá trị chi-bình phương: Giá trị kiểm định chi-bình phương được tính bằng công thức:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

  Trong đó O_i là số liệu quan sát được trong nhóm thứ i và E_i là số liệu kì vọng trong nhóm thứ i.

- So sánh với giá trị tới hạn: So sánh giá trị chi-bình phương tính được với giá trị tới hạn từ bảng phân phối chi-bình phương tương ứng với mức ý nghĩa α và bậc tự do được tính bằng công thức:

$$\text{df} = k - p - 1,$$

  trong đó k là số nhóm và p là số tham số được ước lượng từ dữ liệu (đối với phân phối chuẩn là 2: trung bình và độ lệch chuẩn).

- Quyết định: Nếu giá trị chi-bình phương tính toán lớn hơn giá trị tới hạn, ta bác bỏ giả thuyết H_0, tức là tổng thể không tuân theo phân phối chuẩn. Trong trường hợp ngược lại, giá trị chi-bình phương nhỏ hơn hoặc bằng giá trị tới hạn, thì không có đủ bằng chứng để bác bỏ H_0.

Ví dụ 3.11. Giả sử có dữ liệu chiều cao của 100 người (cm), và chúng ta muốn kiểm tra xem dữ liệu này có tuân theo phân phối chuẩn hay không. Dữ liệu thô có giá trị bé nhất là 162 và lớn nhất là 183 và đã được phân nhóm như sau:

Nhóm (cm)	161–165	166–170	171–175	176–180	181–185
Tần suất	10	25	30	22	13

Bằng tính toán quen thuộc, chúng ta thu được trung bình mẫu là $\overline{x} = 173{,}15$ và phương sai mẫu là $s^2 \approx 35{,}0783$. Chúng ta sẽ kiểm định mức độ phù hợp của dữ liệu này với mô hình phân phối chuẩn $\mathcal{N}(173{,}15; \sigma^2 = 31{,}0783)$.

Chúng ta tính các xác suất kì vọng cho mỗi nhóm. Ví dụ, với nhóm thứ nhất, 161–165, chúng ta tính được số liệu kì vọng bằng xác suất kì vọng nhân với tổng số mẫu. Cụ thể là

$$P(155{,}5 \le X \le 160{,}5) \times 100 = 1{,}49.$$

Với các nhóm khác, chúng ta cũng tính toán tương tự như vậy. Các kết quả thu được ghi trong bảng sau:

Nhóm (cm)	156–160	161–165	166–170	171–175	176–180	181–185
Quan sát	2	8	25	30	22	13
Kì vọng	1,49	8,19	22,9	32,7	23,85	8,88
$(O_i - E_i)^2/E_i$	0,175	0,004	0,193	0,223	0,144	1,912

Từ bảng, chúng ta tính được tổng giá trị chi-bình phương

$$\chi^2 = 0{,}4 + 0{,}193 + 0{,}223 + 0{,}144 + 1{,}912 = 2{,}651.$$

Với mức ý nghĩa 0,05, từ bảng tới hạn của phân phối chi-bình phương với

$$m = \text{số nhóm} - \text{số tham số} - 1 = 6 - 2 - 1 = 3$$

bậc tự do, chúng ta có $\chi^2_{0{,}05} \approx 7{,}8147$. Do giá trị tính được nhỏ hơn giá trị tới hạn, chúng ta chấp nhận giả thiết H_0: "dữ liệu phù hợp với phân phối chuẩn".

3.5.2 Một số kiểm định khác

Kiểm định chi-bình phương nêu trong mục trước thường được sử dụng khi chúng ta có một tập dữ liệu đủ lớn và muốn kiểm tra xem dữ liệu có phù hợp với phân phối chuẩn hay không. Nó đặc biệt hữu ích khi dữ liệu được chia thành các nhóm hoặc lớp, và chúng ta cần kiểm tra sự phù hợp giữa số liệu quan sát được và số liệu kì vọng từ mô hình chuẩn.

Tuy nhiên, với dữ liệu có số lượng quan sát nhỏ hoặc bị ảnh hưởng bởi các ngoại lệ, chúng ta nên cân nhắc các kiểm định khác như Shapiro-Wilk

hoặc Kolmogorov-Smirnov[1] (K–S test) hoặc kiểm định Anderson-Darling vì chúng có thể cho kết quả chính xác hơn trong những trường hợp này. Rất tiếc, các bài toán kiểm định này nằm ngoài khuôn khổ của môn học hiện tại của chúng ta. Bạn đọc có thể tham khảo Walpole–Myers–Myers –Ye [3].

3.6 Kiểm định phi tham số

Các phương pháp kiểm định nói trong các mục trước (kiểm định tham số) dựa trên một điều kiện rất quan trọng, đó là các dữ liệu được lấy từ các tổng thể có phân phối chuẩn. Trong thực tế khi các điều kiện này bị vi phạm,thì các phương pháp kiểm định tham số sẽ đưa ra các kết luận sai lệch. Trong những tình huống này, kiểm định phi tham số (nonparametric hypothesis test) là một phương pháp thay thế.

Sau đây, chúng ta tìm hiểu kĩ một ví dụ về *kiểm định dấu cho trung vị* (Sign test for median).

Ví dụ 3.12. Trong một nghiên cứu về dinh dưỡng có 12 người tham gia. Sau một giai đoạn thử nghiệm chế độ dinh dưỡng đặc biệt, người ta thu được dữ liệu về tăng/giảm cân nặng như sau: $-1,-3,-3,-3,4,-5,-6,-3,-7,-10,-2,4$, trong đó có 10 người giảm cân và 3 người tăng cân. Sử dụng dữ liệu này, chúng ta kiểm định giả thiết trung vị của sự tăng/giảm cân: $m = 0$ (giả thiết H_0, chế độ dinh dưỡng không ảnh hưởng tới cân nặng) đối với giả thiết $m < 0$, (giả thiết thay thế H_1, chế độ dinh dưỡng có tác dụng giảm cân).

Gọi X là số người tham gia giảm cân. Nếu H_0 đúng, thì X có phân phối nhị thức với $p = 0{,}5$. Mẫu của chúng ta có 12 người, trong đó 2 người không thay đổi cân nặng. Xác suất để quan sát được một mẫu 12 người trong đó có 10 hoặc nhiều hơn số người giảm cân là

$$P(X = 10) + P(X = 11) + P(X = 12) = 0{,}01928.$$

Như vậy, nếu chế độ dinh dưỡng không có tác dụng giảm cân, thì chỉ có 1,9% xác suất chúng ta quan sát được kết quả như trên. Do đó, chúng ta chấp nhận giả thiết H_1.

CAS 3.6. Trong R, các bài toán kiểm định như trong Ví dụ 3.12 có thể được thực hiện bởi `binom.test()`; cụ thể như sau:

```
# Đưa dữ liệu vào biến x
x = c(-1, -3, -3, -3, 4, -5, -6, -3, -7, -10, -2, 4)
```

[1] Andrey Nikolaevich Kolmogorov (1903–1987) và là một nhà toán học Xô-Viết, người đóng vai trò trung tâm trong xây dựng lí thuyết xác suất hiện đại. Nikolai Vasilyevich Smirnov (1900–1966) cũng là một nhà toán học người Xô-viết, được biết đến với các cống hiến trong xác suất và thống kê, bao gồm kiểm định Kolmogorov–Smirnov.

```r
# Áp dụng chức năng binom.test() với tham số "greater" (lớn hơn)
binom.test(sum(x < 0), length(x), p = 0.5, alternative = "greater")

##
##  Exact binomial test
##
## data:  sum(x < 0) and length(x)
## number of successes = 10, number of trials = 12, p-value = 0,019287
## alternative hypothesis: true probability of success is greater than 0,5
## 95 percent confidence interval:
##  0,56189456 1,00000000
## sample estimates:
## probability of success
##             0,83333333
```

Lưu ý, giá trị p cho bởi `p-value = 0.019` chính là xác suất chúng ta đã tính trong Ví dụ 3.12.

Một phát triển của kiểm định dấu cho trung vị là kiểm định hạng Wilcoxon, bạn đọc có thể xem Walpole et al [3].

Tài liệu

[1] E. Kreyszig, *Advanced Engineering Mathematics, Part G: Probability, Statistics*, 10 **edition**. New York: John Wiley & Sons, Inc., 2011.

[2] Đ. H. Thắng, *Thống kê và Ứng dụng: Giáo trình dành cho các trường Đại học và Cao đẳng*, 6 **edition**. NXB Giáo dục Việt Nam, 2022.

[3] R. E. Walpole, R. H. Myers, S. L. Myers **and** K. Ye, *Probability & Statistics for Engineers & Scientists*, 9 **edition**. Prentice Hall, Pearson, 2002.

Chương 4

Tương quan và hồi quy

Mục tiêu chính của chương này là tìm hiểu hai nội dung chính: *tương quan* và *hồi quy*. Các khái niệm và phương pháp cần hiểu và nhớ bao gồm *hiệp phương sai, hệ số tương quan*, khái niệm *đường hồi quy*, và phương pháp tìm đường *hồi quy tuyến tính*.

Tài liệu tham khảo: Kreyszig [1, §25.9], Đ. H. Thắng [2, Chương 7], V. H. Như [3, Chương 4].

Nội dung

4.1 Tương quan: hiệp phương sai, hệ số tương quan

Phân tích tương quan (analysis of variance) là quá trình đánh giá mối quan hệ giữa hai hoặc nhiều biến số hoặc thuộc tính để xác định mức độ tương đồng hoặc tương phản giữa chúng. Quá trình này thường được sử dụng để đo lường và đánh giá sự liên quan giữa các biến trong dữ liệu số liệu. Kết quả của phân tích tương quan thường được biểu thị dưới dạng *hệ số tương quan* (correlation coefficient), cho biết độ mạnh và hướng của mối

quan hệ giữa các biến số. Trong mục này, chúng ta nói về hệ số tương quan Pearson và mối quan hệ tuyến tính giữa các biến số.

4.1.1 Hệ số tương quan

Giả sử chúng ta có mẫu với n cặp giá trị (x_1, y_1), ..., (x_n, y_n). *Hệ số tương quan* (Pearson) của mẫu, kí hiệu là r, được xác định bởi

$$r = \frac{s_{xy}}{s_x s_y}, \tag{4.1}$$

trong đó s_{xy} là *hiệp phương sai mẫu* (sample covariance), cho bởi

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right). \end{aligned} \tag{4.2}$$

Các đại lượng s_x^2 và s_y^2 là phương sai mẫu của x và y, đã biết đến ở §1.1 trong Chương 1, đó là

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right), \tag{4.3}$$

và tương tự cho s_y^2.

Mệnh đề 4.1:

Hệ số tương quan có các tính chất sau:

(i) $-1 \le r \le 1$;

(ii) $r = \pm 1$ nếu và chỉ nếu tất cả các điểm mẫu nằm trên một đường thẳng.

Chứng minh. Dữ liệu $u = (x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x})$ và $v = (y_1 - \overline{y}, y_2 - \overline{y}, \ldots, y_n - \overline{y})$ có thể xem là hai véctơ trong $\mathbb{R}^n$. Khi đó, từ công thức hiệp phương sai,

$$s_{xy} = \frac{u \cdot v}{n-1},$$

trong đó biểu thức $u \cdot v$ ở vế phải là tích vô hướng thông thường của u và v. Cũng vậy, từ công thức của phương sai,

$$s_x^2 = \frac{\|u\|^2}{n-1}, \quad s_y^2 = \frac{\|v\|^2}{n-1},$$

với chuẩn $\|u\|^2 = u \cdot u$ là "chuẩn Euclid" trong $\mathbb{R}^n$. Từ đó, áp dụng BĐT Buniakovsky–Cauchy–Schwarz quen thuộc,

$$|u \cdot v| \leq \|u\| \|v\|,$$

chúng ta dễ dàng thu được

$$|r| = \left| \frac{s_{xy}}{s_x s_y} \right| = \frac{|u \cdot v|}{\|u\| \|v\|} \leq 1,$$

và (i) đã được chứng minh.

Trong trường hợp dấu bằng xảy ra, u và v là phụ thuộc tuyến tính và khi đó dễ thấy các điểm dữ liệu phải nằm trên cùng một đường thẳng. □

Nhận xét 4.1. Các vectơ u và v trong chứng minh trên thể hiện các độ lệch của điểm mẫu của x và y so với giá trị trung bình của chúng. Từ công thức (4.1), hệ số tương quan r là $cos(\alpha)$ của góc α giữa u và v trong không gian Euclid $\mathbb{R}^n$. Khi $r \approx 1$ thì $\alpha \approx 0$, biểu thị xu hướng thay đổi của x và y so với trung bình là gần giống nhau. Khi $r \approx -1$, thì $\alpha \approx 180°$, cho thấy xu hướng thay đổi này ngược nhau. Khi $r \approx 0$ thì góc $\alpha \approx 90°$, cho thấy xu hướng thay đổi của x và y không có mối liên hệ nào.

Ý nghĩa hình học của hệ số tương quan Pearson gợi ý rằng nó chỉ phù hợp cho các mối quan hệ tuyến tính.

Ví dụ 4.1. Tính hệ số tương quan và hiệp phương sai của mẫu sau:

$$\{(0;\ 0{,}2), (2;\ 2{,}2), (4;\ 3{,}1), (6;\ 3{,}8), (8;\ 5{,}5), (10;\ 7{,}2)\}.$$

Lời giải

Ta có hai mẫu $x = (0, 2, 4, 6, 8, 10)$ và $y = (0{,}2;\ 2{,}2;\ 3{,}1;\ 3{,}8;\ 5{,}5;\ 7{,}2)$. Từ công thức của phương sai mẫu,

$$s_x^2 = 14, \quad s_y^2 = 6{,}070667.$$

Trong khi đó, hiệp phương sai mẫu tính bởi công thức trên là

$$s_{xy}^2 = 9{,}12.$$

Đưa các tính toán trên vào công thức hệ số tương quan, chúng ta thu được

$$r_{xy} = \frac{9{,}12}{\sqrt{14} \times \sqrt{6{,}070667}} \approx 0{,}989265.$$

Chú ý 4.1. Đối với dữ liệu trong ví dụ trên, hệ số tương quan khác 1 nên các điểm mẫu không nằm trên một đường thẳng. Tuy nhiên, giá trị của nó khá gần 1 (bằng 0,983999) nên các điểm mẫu nằm khá gần một đường thẳng; xem Hình 4.1.

CAS 4.1. Trong phần mềm R, hiệp phương sai mẫu được tính bởi `cov()`, trong khi đó, hệ số tương quan mẫu được tính bởi `cor()`.

Trong đoạn mã R sau đây, ta tính các kết quả trong ví dụ trên.

```
# Đưa vào các dữ liệu x và y
x = c(0, 2, 4, 6, 8, 10)
y = c(0.2, 2.2, 3.1, 3.8, 5.5, 7.2)

# Tính hiệp phương sai theo phương pháp Pearson
cov(x, y, method = "pearson")

## [1] 9,12

# Tính hệ số tương quan
cor(x, y)

## [1] 0,98926496
```

Nhận xét 4.2. Như đã thấy, một ưu điểm của hệ số tương quan Pearson r là dễ tính toán, dễ hiểu và dễ sử dụng trong phân tích dữ liệu. Trong khi đó, nhược điểm của nó là chỉ phù hợp với dữ liệu có phân phối chuẩn và mối quan hệ tuyến tính.

Nếu dữ liệu không tuân theo phân phối chuẩn hoặc có mối quan hệ phi tuyến tính, người ta có thể cần sử dụng các phương pháp khác như hệ số tương quan Spearman hoặc Kendall. Bạn đọc có thể tham khảo [4, Chapter 16.7].

4.1.2 Hiệp phương sai

Như đã biết trong §1.10 của Chương 1, đối với hai BNN X và Y, đại lượng

$$\sigma_{XY} = \mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big], \tag{4.4}$$

trong đó $\mu_X = \mathbb{E}[X]$ và $\mu_Y = \mathbb{E}[Y]$, lần lượt là trung bình của X và Y, được gọi là *hiệp phương sai* của hai BNN X và Y.

Lưu ý rằng hiệp phương sai còn có thể tính qua công thức

$$\sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Từ hiệp phương sai của hai BNN, chúng ta có thể định nghĩa hệ số tương quan (lí thuyết) cho hai BNN X và Y như sau:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{4.5}$$

trong đó

$$\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}\big[(X-\mathbb{E}[X])^2\big],$$

là phương sai của X, trong khi đó σ_Y^2 với công thức tương tự là phương sai của Y.

Ví dụ 4.2. Cho hai BNN liên tục X và Y với hàm mật độ đồng thời

$$f(x,y)=\begin{cases}2 & \text{nếu } 0<x<y<1,\\ 0 & \text{trong các trường hợp còn lại.}\end{cases}$$

Tính hiệp phương sai và hệ số tương quan.

Lời giải

Ta bắt đầu tính các kì vọng của X, Y, và XY. Từ hàm mật độ đồng thời, ta tính được, qua định lí Fubini (xem ví dụ [5, §VII.3, §3]),

$$\begin{aligned}\mathbb{E}[X]&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\,\mathrm{d}x\,\mathrm{d}y\\&=\int_0^1\mathrm{d}y\int_0^y 2x\,\mathrm{d}x\\&=\int_0^1 y^2\,\mathrm{d}y\\&=\frac{1}{3}.\end{aligned}$$

Bằng những tính toán tương tự, có thể tính được

$$\mathbb{E}[Y]=\frac{2}{3},$$

trong khi đó

$$\mathbb{E}[XY]=\int_0^1\mathrm{d}y\int_0^y 2xy\,\mathrm{d}x=\frac{1}{4}$$

Từ đó, hiệp phương sai của X và Y là

$$\begin{aligned}\sigma_{XY}&=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\\&=\frac{1}{4}-\frac{1}{3}\cdot\frac{2}{3}\\&=\frac{1}{36}.\end{aligned}$$

Để tính hệ số tương quan, ta còn phải tính các phương sai của X và Y: Ta có

$$\mathbb{E}\left[X^2\right]=\int_0^1\mathrm{d}y\int_0^y 2x^2\,\mathrm{d}x=\frac{1}{6},$$

và từ đó suy ra

$$\sigma_X^2 = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{18}.$$

Tính toán tương tự, ta cũng thu được

$$\sigma_Y^2 = \frac{1}{18}.$$

Tổng hợp các tính toán, ta tính được hệ số tương quan ρ của X và Y là

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{1}{2}.$$

Mối liên hệ giữa hệ số tương quan của hai BNN và sự phụ thuộc tuyến tính của chúng như sau:

Định lí 4.1:

Hệ số tương quan ρ của hai BNN X và Y có các tính chất sau:

(i) $-1 \le \rho \le 1$;

(ii) $\rho = \pm 1$ nếu và chỉ nếu X và Y là phụ thuộc tuyến tính, có nghĩa là tồn tại các hệ số γ và δ, hoặc γ^* và δ^* sao cho

$$Y = \gamma X + \delta,$$

hoặc

$$X = \gamma^* Y + \delta^*.$$

Chứng minh Định lí 4.1 cũng tương tự như chứng minh Mệnh đề 4.1, sử dụng một mở rộng của BĐT Buniakovski–Cauchy–Schwarz cho không gian véctơ tuỳ ý với một tích vô hướng (không nhất thiết hữu hạn chiều). Chúng ta bỏ qua chi tiết.

Từ hệ số tương quan, ta đưa ra khái niệm *không tương quan* (uncorrelated) cho hai BNN như sau.

Định nghĩa 4.1:

Hai BNN X và Y được gọi là *không tương quan* nếu $\rho = 0$.

Nếu X và Y là hai BNN độc lập thì hiệp phương sai của chúng bằng 0, tức là $\sigma_{XY} = 0$, và do đó hệ số tương quan cũng bằng 0, tức là $\rho = 0$. Điều ngược lại nói chung không đúng: tồn tại các BNN không độc lập, nhưng chúng có hệ số tương quan bằng 0.

Ví dụ 4.3. Giả sử $X \sim \mathcal{N}(0,1)$ và $Y = X^2$. Rõ ràng $\mathbb{E}[X] = 0$, $XY = X^3$, và do đó

$$\begin{aligned}\sigma_{XY} &= \mathbb{E}[XY] - \underbrace{\mathbb{E}[X]}_{=0}\mathbb{E}[Y] \\ &= \mathbb{E}\left[X^3\right] \\ &= \int_{-\infty}^{+\infty} x^3 f(x)\,\mathrm{d}x,\end{aligned}$$

trong đó $f(x)$ là hàm mật độ của phân phối chuẩn tắc: $f(x) = e^{-x^2/2}/\sqrt{2\pi}$.

Do tính đối xứng của $f(x)$, ta dễ thấy tích phân trên bằng 0, và do đó, $\sigma_{XY} = 0$, có nghĩa là X và Y không tương quan.

Tuy nhiên, X và Y cũng không độc lập. Thật vậy, xét hai sự kiện $A = \{X \geq 1\}$ và $B = \{Y \geq 1\}$. Do điều kiện $X \geq 1$ kéo theo điều kiện $Y \geq 1$ nên $A \subset B$ và vậy thì $A \cap B = \{X \geq 1\} = A$. Vậy,

$$P(A \cap B) = P(A) > P(A)P(B).$$

Như vậy, A và B không độc lập và X và Y cũng không độc lập.

Định nghĩa 4.2:

Véctơ ngẫu nhiên 2 chiều (X, Y) có *phân phối chuẩn hai chiều* nếu hàm mật độ đồng thời của nó có dạng

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}h(x,y)\right),$$

với

$$h(x,y) = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right].$$

Đối với phân phối chuẩn hai chiều, chúng ta có mối quan hệ sau:

Định lí 4.2:

Nếu (X, Y) có phân phối chuẩn, X, Y không tương quan, thì X và Y là độc lập.

Chứng minh của định lí này bạn đọc có thể xem Shiryaev [6] hoặc Klenke [7]. Lưu ý, trong Định lí 4.2, điều kiện (X, Y) có phân phối chuẩn là cần thiết; xem Ví dụ 4.3.

4.1.3 Kiểm định cho hệ số tương quan

Kiểm định hệ số tương quan ρ của hai biến giúp xác định xem có mối quan hệ tuyến tính giữa hai biến và liệu mối quan hệ đó có ý nghĩa thống kê hay không. Chi tiết hơn, chúng ta kiểm định giả thiết H_0: $\rho = 0$, so với đối thiết H_1: $\rho > 0$[1], sử dụng mẫu gồm n cặp $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Quy trình kiểm định hệ số tương quan này được nêu ra bởi Fisher[2]

Các bước thực hiện kiểm định:

1. Chọn mức ý nghĩa α, thường là 5% hoặc 1%.

2. Tính giá trị tới hạn t_α của phân phối t với $n-2$ bậc tự do, trong đó n là kích thước mẫu (số cặp quan sát):

$$F(t_\alpha) = 1 - \alpha.$$

   (Sử dụng Bảng t-Distribution (bảng A9) trong Kreyszig [1]).

3. Sử dụng kiểm định t với công thức:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

   trong đó r là hệ số tương quan Pearson của mẫu.

4. Kết luận:

   - Nếu $t > t_\alpha$ thì bác bỏ giả thuyết H_0, tức là hệ số tương quan khác 0 có ý nghĩa thống kê.
   - Nếu $t \leq t_\alpha$ thì không đủ bằng chứng để bác bỏ H_0, nghĩa là không có mối liên hệ tuyến tính rõ ràng.

Ví dụ 4.4. Sử dụng dữ liệu trong Ví dụ 4.1, kiểm định giả thiết H_0: $\rho = 0$, so với đối thiết H_1: $\rho > 0$ với mức ý nghĩa 1%.

Lời giải

Chúng ta có các thông số sau:

- Mức ý nghĩa $\alpha = 0{,}05$.
- Số điểm mẫu $n = 6$.

[1] Lưu ý rằng trong mọi tình huống thì $\rho \geq 0$.

[2] Fisher, R.A. (1915) *Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population.* Biometrika, 10, 507-521.

- Mức tới hạn t_α trong phân phối t với $n-2=4$ bậc tự do:

$$t_\alpha \approx 3{,}75.$$

  (Tra bảng t-Distribution (Bảng A9) trong Kreyszig [1] hoặc sử dụng máy vi tính).

- Giá trị của kiểm định t: Hệ số tương quan Pearson $r = 0{,}989265$ và do đó

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}989265 \times \sqrt{4}}{\sqrt{1-(0{,}989265)^2}} = 13{,}53927.$$

- Kết luận: Giá trị t lớn hơn giá trị tới hạn (với mức ý nghĩa 1%) nên bác bỏ H_0.

Kết luận của quy trình kiểm định phù hợp với nhận định ban đầu khi quan sát thấy hệ số tương quan mẫu rất gần 1.

CAS 4.2. Trong R, kiểm định cho hệ số tương quan bằng phương pháp Pearson cho dữ liệu trong ví dụ trên thực hiện như sau:

```
# Đưa dữ liệu vào các biến x và y
x <- c(0, 2, 4, 6, 8, 10)
y <- c(0.2, 2.2, 3.1, 3.8, 5.5, 7.2)

# Chức năng cor.test() cho (x, y) theo phương pháp Pearson
cor.test(x, y, method="pearson")

##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 13,5392, df = 4, p-value = 0,00017224
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0,90135889 0,99887794
## sample estimates:
##       cor
## 0,98926496
```

Kết quả cho thấy giá trị p rất nhỏ, nhỏ hơn mức ý nghĩa 5% rất nhiều, nên ta bác bỏ H_0. Hơn nữa, chúng ta cũng thu được giá trị của thống kê t, đúng như đã tính bằng tay.

Bài tập §4.1

1. Tính hiệp phương sai và hệ số tương quan của mẫu x và y cho bởi bảng sau:

x	2	4	6	8
y	4,15	7,85	12,20	15,86

2. Cũng câu hỏi như Bài 1 với mẫu sau:

x	1,2	2,3	3,2	4,4	5,1
y	10,4	14,1	18,9	22,5	26,1

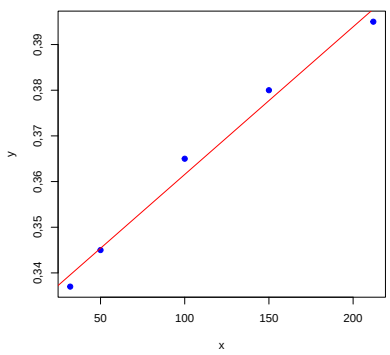
3. Giả sử X nhận các giá trị $-1, 0, 1$ với các xác suất là $\frac{1}{3}$ và $Y = X^2$. Hãy tính hệ số tương quan ρ của X và Y. Hai BNN X và Y có độc lập hay không? Tại sao?

4. Gieo hai xúc xắc đồng chất cân xứng. Gọi X là số chấm của xúc xắc thứ nhất và Y là tổng số chấm của cả hai xúc xắc. Tính hệ số tương quan ρ của X và Y.

4.2 Hồi quy tuyến tính đơn

Mục này nói về trường hợp đơn giản nhất của hồi quy, đó là hồi quy tuyến tính đơn biến (simple linear regression). Tuy đơn giản, nhưng đây lại là trường hợp hay gặp nhất trong các ứng dụng của thống kê. Hơn nữa, khi hiểu rõ ý tưởng và phương pháp hồi quy tuyến tính đơn, việc tìm hiểu hồi quy phi tuyến và hồi quy đa biến trở nên dễ dàng hơn. Tài liệu tham khảo: Kreyszig [1, Chapter 25.9], Đ. H. Thắng [2, Chương 7].

4.2.1 *Đường hồi quy tuyến tính mẫu, hệ số xác định*



Hình 4.1: Biểu đồ phân tán của dữ liệu và đường hồi quy tuyến tính

Trong phân tích hồi quy (regression analysis), chúng ta tìm hiểu sự phụ thuộc của một đại lượng ngẫu nhiên Y vào một đại lượng x (nếu có) thông qua sự phụ thuộc của trung bình $\mu_{Y(x)}$ của Y vào x, tức là tìm hiểu $\mu_{Y(x)} = \mu(x)$ như một hàm số theo nghĩa thông thường: Mục tiêu cụ thể của chúng ta ở đây là ước tính hàm số $\mu(x)$ từ một mẫu gồm các cặp giá trị (x_i, y_i) quan sát được.

Đồ thị của hàm $y = \mu(x)$ trên mặt phẳng Oxy được gọi là *đường hồi quy* (regression curve) của Y. Khi $\mu(x)$ là một hàm bậc nhất:

$$\mu(x) = \kappa_0 + \kappa_1 x, \qquad (4.6)$$

thì đường hồi quy là đường thẳng, được gọi là *đường hồi quy tuyến tính*.

Bây giờ, giả sử chúng ta có một mẫu ngẫu nhiên gồm n cặp giá trị (x_1, y_1), (x_2, y_2), …, (x_n, y_n), biểu thị n điểm trên mặt phẳng Oxy. Chúng

ta muốn xây dựng một đường thẳng để từ đó ta có thể xấp xỉ giá trị của $\mu(x)$ tại một giá trị cụ thể của x. Trong tình huống này, người ta sử dụng một phương pháp, gọi là phương pháp *bình phương tối thiểu* (least square principle–OLS), được khởi xướng bởi Gauss[1] và Legrendre[2].

Trong phương pháp bình phương tối thiểu, đường hồi quy tuyến tính được tìm dựa vào điều kiện sau: "*Đường hồi quy tuyến tính có tổng của bình phương 'khoảng cách' (theo phương song song với trục O y) từ các điểm dữ liệu tới đường hồi quy là nhỏ nhất.*"

Để phương pháp này xác định duy nhất nghiệm thì phải có điều kiện sau đây đối với các mẫu.

Giả thiết 4.1 (A1). Tồn tại ít nhất hai giá trị phân biệt trong các giá trị x_i của các điểm mẫu.

Dễ thấy Giả thiết A1 tương đương với điều kiện $s_x^2 > 0$.

Do khoảng cách thẳng đứng từ điểm (x_i, y_i) đến đường hồi quy mẫu là

$$d_i = |y_i - k_0 - k_1 x_i|,$$

nên nguyên lí bình phương tối thiểu cho hai hệ số k_0 và k_1 sao cho của đường hồi quy sao cho

$$\begin{aligned} q = q(k_0, k_1) &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} |y_i - k_0 - k_1 x_i|^2 \end{aligned}$$

là bé nhất.

Để tìm k_1 và k_0 sao cho q bé nhất, ta có thể áp dụng phương pháp tìm cực trị của hàm hai biến số: Các giá trị tới hạn của k_0 và k_1 thỏa mãn hệ phương trình $\nabla q(k_0, k_1) = 0$ (xem, ví dụ [5, §VI.6]), tức là

$$\begin{cases} \dfrac{\partial q}{\partial k_0} = 0, \\ \dfrac{\partial q}{\partial k_1} = 0, \end{cases} \tag{4.7}$$

[1]Johann Carl Friedrich Gauß (1777–1855) là một nhà toán học người Đức.
[2]Adrien-Marie Legendre (1752–1833) là một nhà toán học người Pháp.

trong đó, $\frac{\partial q}{\partial k_j}$ là đạo hàm riêng của q theo k_j $j=0,1$. Bằng tính toán cụ thể,

$$\begin{aligned}\frac{\partial q}{\partial k_0} &= \sum_{i=1}^{n} \frac{\partial d_i}{\partial k_0} \\ &= \sum_{i=1}^{n} 2\left(y_i - k_0 - k_1 x_i\right) \\ &= 2\left\{\sum_{i=1}^{n} y_i - n k_0 - k_1 \sum_{i=1}^{n} x_i\right\}.\end{aligned}$$

Vậy, phương trình $\partial q/\partial k_0 = 0$ có thể viết lại là

$$n k_0 + k_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \tag{4.8}$$

Tương tự, tính toán trực tiếp phương trình $\partial q/\partial k_1$ và kết hợp với phương trình vừa rồi, chúng ta thu được hệ phương trình tương đương với (4.7),

$$\begin{cases} n k_0 + k_1 \sum\limits_{i=1}^{n} x_i & = \sum\limits_{i=1}^{n} y_i \\ k_0 \sum\limits_{i=1}^{n} x_i + k_1 \sum\limits_{i=1}^{n} x_i^2 & = \sum\limits_{i=1}^{n} x_i y_i, \end{cases} \tag{4.9}$$

được gọi là *phương trình chuẩn* (normal equation) của bài toán hồi quy. Đó là một hệ phương trình tuyến tính gồm hai phương trình của hai ẩn k_0 và k_1.

Bởi Giả thiết A1, phương trình chuẩn là một hệ phương trình Cramer[1], do định thức của ma trận các hệ số là khác không:

$$\begin{aligned} D := \begin{vmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{vmatrix} &= n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \\ &= n(n-1)s_x^2 > 0. \end{aligned}$$

Bây giờ, để áp dụng công thức Cramer quen thuộc (xem Kreyszig [1, §7.7]),

[1] Hệ phương trình Cramer là một khái niệm cơ bản trong đại số tuyến tính. Đó là các hệ phương trình tuyến tính có số ẩn bằng số phương trình và ma trận hệ số là ma trận khả nghịch. Các hệ phương trình chuẩn nảy sinh từ bài toán bình phương bé nhất thường là các hệ phương trình tuyến tính với ma trận hệ số là ma trận đối xứng. Do đó, ngoài phương pháp Cramer, những hệ này còn có thể giải bằng các phương pháp số khác như phương pháp phân rã Cholesky, v.v.

ta tính định thức tương ứng với k_1:

$$D_1 := \begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i y_i \end{vmatrix} = n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$
$$= n(n-1)s_{xy}.$$

Từ đó, nghiệm của phương trình chuẩn (4.9) là

$$k_1 = \frac{D_1}{D} = \frac{s_{xy}}{s_x^2}. \tag{4.10}$$

Hệ số k_1 như trên gọi là *hệ số hồi quy* của mẫu.

Từ k_1, chúng ta thay vào phương trình đầu tiên trong (4.9) và giải ra được

$$k_0 = \overline{y} - k_1\overline{x}. \tag{4.11}$$

Lưu ý, trung bình mẫu $(\overline{x},\ \overline{y})$ nằm trên đường hồi quy $y = k_0 + k_1 x$.

Ví dụ 4.5. Tính phương trình đường hồi quy tuyến tính của mẫu (x_i, y_i), $i = 1,2,3,4$, cho bởi bảng sau:

x_i	$-1{,}2$	$-0{,}1$	$0{,}2$	$1{,}4$
y_i	$0{,}105$	$1{,}101$	$0{,}890$	$1{,}925$

Từ đó, tính xấp xỉ giá trị của Y khi $x = 0{,}5$.

Lời giải

Để tính s_x^2 và s_{xy}, chúng ta sẽ tính các giá trị x_i^2 và $x_i y_i$ với $i = 1,2,\ldots,n = 4$ và ghi các kết quả thu được trong bảng sau:

i	x_i	y_i	x_i^2	$x_i y_i$	$y = 0{,}6813494\,x + 0{,}9541488$
1	$-1{,}2$	$0{,}105$	$1{,}44$	$-0{,}126$	$0{,}13653$
2	$-0{,}1$	$1{,}101$	$0{,}01$	$-0{,}1101$	$0{,}88602$
3	$0{,}2$	$0{,}890$	$0{,}04$	$0{,}178$	$1{,}09042$
4	$1{,}4$	$1{,}925$	$1{,}96$	$2{,}695$	$1{,}90804$
$\sum_i$	$0{,}3$	$4{,}021$	$3{,}45$	$2{,}6369$	$q(k_0,k_1) \approx 0{,}087669$

(Dòng cuối cùng là tổng các giá trị của các dòng phía trên.) Từ đó, ta tính được từ cột thứ hai và thứ tư (với $n = 4$)

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{3}\left(3{,}45 - \frac{1}{4}(0{,}3)^2\right) = 1{,}1425.$$

Với công thức tương tự, ta có (cũng với $n=4$) hệ số tương quan mẫu là

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) \\ &= \frac{1}{3}\left(2{,}6369 - \frac{1}{4}(0{,}3)(4{,}021)\right) \\ &= 0{,}7784417. \end{aligned}$$

Từ đó, ta tính được hệ số hồi quy theo công thức 4.10 là

$$k_1 = \frac{s_{xy}}{s_x^2} = \frac{1{,}425}{0{,}7784417} = 0{,}6813494.$$

Đường hồi quy lúc này có dạng

$$y - \overline{y} = k_1(x - \overline{x}),$$

trong đó, $\overline{y} = 4{,}021/4 = 1{,}00525$ còn $\overline{x} = 0{,}075$. Ta thu được hàm bậc nhất $y = 0{,}6813494x + 0{,}9541488$ với tổng bình phương các sai số xấp xỉ 0,087669.

Khi $x = 0{,}5$, thay vào phương trình đường hồi quy, ta tính được

$$0{,}6813494 \cdot 0{,}5 + 0{,}9541488 \approx 1{,}2948.$$

CAS 4.3. Phần mềm R có chức năng `lm()` (viết tắt của **l**inear **m**odel) để tính hệ số hồi quy và đường tuyến tính mẫu như trong ví dụ sau:

```
# Đưa vào dữ liệu x và y
x = c(-1.2, -0.1, 0.2, 1.4)
y = c(0.105, 1.101, 0.890, 1.925)

# Gọi chức năng lm() để tính mô hình tuyến tính
model = lm(y ~ x)

# Hiển thị các hệ số hồi quy thông qua $coefficients
print(model$coefficients)

## (Intercept)           x
##  0,95414880  0,68134938
```

Từ đó, chúng ta thu được hệ số hồi quy và đoạn chắn (intercept) của đường hồi quy tuyến tính của dữ liệu trong Ví dụ 4.5.

CAS 4.4. Sau đây, chúng ta sử dụng `plot()` của phần mềm R để vẽ đồ thị phân tán (scatter plot) cho dữ liệu. Để làm ví dụ, xét dữ liệu sau (lấy từ [1, Problem Set 25.9]).

x	32	50	100	150	212
y	0,337	0,345	0,365	0,380	0,395

Đoạn code R sau đây cho bức tranh minh hoạ của các điểm dữ liệu trên mặt phẳng toạ độ và đường hồi quy tuyến tính của nó. Kết quả thu được thể hiện trong Hình 4.1 trong §4.2.

```
# Đưa vào các dữ liệu x và y
x = c(32, 50, 100, 150, 212)
y = c(0.337, 0.345, 0.365, 0.380, 0.395)

# Vẽ biểu đồ x và y
plot(y ~ x, pch = 19, col = "blue")

# Tính các mô hình tuyến tính
model = lm(y ~ x)

# Vẽ đường thẳng hồi qui dựa trên các hệ số tính được
curve(model$coefficients[2] * x + model$coefficients[1],
    from = 20, to = 220, add = "TRUE", col = "RED")
```

4.2.2 Khoảng tin cậy trong phân tích hồi quy

Trong tiểu mục trước, ta tìm hệ số hồi quy và kết quả tìm được là một ước lượng điểm của hệ số đó. Trong tiểu mục này, ta sẽ nói về khoảng tin cậy của hệ số hồi quy đó trong trường hợp hai giả thiết sau đây được thoả mãn:

Giả thiết 4.2. Với mỗi x, biến ngẫu nhiên Y có phân phối chuẩn với phương sai σ^2 không phụ thuộc vào x, trong khi đó, trung bình của Y là $\mu(x)$ thoả mãn

$$\mu(x) = \kappa_0 + \kappa_1 x. \tag{4.12}$$

Giả thiết 4.3. Các thực nghiệm lấy mẫu $(x_1, y_1), (x_2, y_2), \ldots,$ là độc lập.

Hệ số κ_1 được gọi là *hệ số hồi quy* của tổng thể. Mặt khác, có thể chứng minh được (trong tình huống hai giả thiết trên được thoả mãn) ước lượng hợp lí cực đại của κ_1 chính là hệ số hồi quy mẫu k_1.

Khoảng tin cậy của κ_1

Với các giả thiết (4.1)–(4.3), chúng ta có thể xây dựng khoảng tin cậy của κ_1 như sau:

1. Chọn mức tin cậy γ (chẳng hạn 95%).

2. Tìm giá trị tới hạn $t_{\alpha/2}$, trong đó $\alpha/2 = (1-\gamma)/2$, của phân phối Student với $n-2$ bậc tự do (n là cỡ mẫu). Như vậy,

$$F(t_{\alpha/2}) = 1-\alpha/2 = 1-(1-\gamma)/2 = (1+\gamma)/2.$$

3. Tính các giá trị phương sai mẫu s_x^2, s_{xy}, và s_y^2, cùng với hệ số hồi quy mẫu k_1. Đặt

$$q_0 = (n-1)(s_y^2 - k_1^2 s_x^2) \tag{4.13}$$

4. Tính

$$K = t_{\alpha/2}\sqrt{\frac{q_0}{(n-2)(n-1)s_x^2}}. \tag{4.14}$$

Khoảng tin cậy của κ_1 là

$$\{k_1 - K \le \kappa_1 \le k_1 + K\}. \tag{4.15}$$

CAS 4.5. Xét một số mẫu dữ liệu [1, Problem Set 25.9] về Revolutions = số vòng quay/phút và Power = công suất động cơ Diesel như sau:

Revolutions	400	500	600	700	750
Power	5800	10300	14200	18800	21000

Đoạn code R sau đây cho kết quả r là một giá trị rất gần 1, cho thấy mối quan hệ tuyến tính "mạnh" giữa x và y.

```
# Đưa vào các dữ liệu
Revolutions = c(400, 500, 600, 700, 750)
Power = c(5800, 10300, 14200, 18800, 21000)

# Tính hệ số tương quan Pearson r bằng chức năng cor()
r = cor(Revolutions, Power, method = "pearson")

# Hiển thị hệ số tương quan r
print(r)

## [1] 0,99968214
```

Để tính khoảng tin cậy, chúng ta hãy tạo một khung dữ liệu (`data.frame`) chứa cả hai giá trị vòng quay và công suất và sử dụng `lm()`, và ghi kết quả vào biến my_model.

```
# Tạo một khung dữ liệu
PR = data.frame(Power, Revolutions)
```

```
# Tính hệ số hồi quy của data frame PR
my_model = lm(PR)
coeffs = my_model$coefficients

# Hiển thị các hệ số hồi quy
print(coeffs[2])

## Revolutions
##   43,182927
```

Kết quả thu được cho thấy hệ số hồi quy $k_1 \approx 43{,}1829$. Cuối cùng, khoảng tin cậy 95% của κ_1 được tính dự bằng `confint()` (viết tắt của **conf**ident **int**erval). Cụ thể, đoạn code R tiếp theo cho các khoảng tin cậy 95% cho κ_0 (Intercept) và κ_1 (Revolutions).

```
# Tính khoảng tin cậy trong mô hình tuyến tính
confint(my_model)

##                      2,5 %        97,5 %
## (Intercept) -12666,020801 -10249,832857
## Revolutions     41,181904     45,183949
```

Bài tập §4.2

1. Tìm phương trình đường hồi quy tuyến tính của các mẫu sau [1, Problem Set 25.9]:

(a) x = biến dạng của một loại thép (mm), y = độ cứng Brinell $\mathrm{kg\,mm^{-2}}$.

x	6	9	11	13	22	26	28	33	35
y	68	67	65	53	44	40	37	34	32

(b) (Định luật Ohm) x = điện áp (V), y = cường độ dòng điện (A). Tìm biến trở R ohm)

x	40	40	80	80	110	110
y	5,1	4,8	0,0	10,3	13,0	12,7

(c) Khoảng cách dừng của ô tô. x = tốc độ (mph), y = khoảng cách dừng (feet). Tìm y tại tốc độ 35 (mph).

x	30	40	50	60
y	160	240	330	435

2. Tìm khoảng tin cậy 95% của hệ số hồi quy, với các giả thiết phù hợp, sử dụng các mẫu trong Bài tập 1.

3. Sử dụng các mẫu trong Bài tập 1, kiểm định giả thiết $\rho = 0$ so với đối thiết $\rho > 0$, sử dụng mức ý nghĩa $\alpha = 0{,}05$.

4. (CAS) Vẽ các đồ thị phân tán (scatter plot) và đường hồi quy tuyến tính cho các dữ liệu trong Bài tập 1.

Tài liệu

[1] E. Kreyszig, *Advanced Engineering Mathematics, Part G: Probability, Statistics*, 10 **edition**. New York: John Wiley & Sons, Inc., 2011.

[2] Đ. H. Thắng, *Thống kê và Ứng dụng: Giáo trình dành cho các trường Đại học và Cao đẳng*, 6 **edition**. NXB Giáo dục Việt Nam, 2022.

[3] V. H. Nhự, C. H. Nam, L. Đ. Ninh **and** P. Q. Sáng, *Bài giảng Xác suất và Thống kê*. Tài liệu lưu hành nội bộ của ĐH Phenikaa, 2025.

[4] R. E. Walpole, R. H. Myers, S. L. Myers **and** K. Ye, *Probability & Statistics for Engineers & Scientists*, 9 **edition**. Prentice Hall, Pearson, 2002.

[5] N. V. Khuê, P. N. Thao, L. M. Hải **and** N. Đ. Sang, *Toán cao cấp, tập II (A2)*. NXB Giáo dục, 2008.

[6] A. N. Shiryaev, *Probability* (Graduate Text in Mathematics 95), 2 **edition**. Springer New York, 1996.

[7] A. Klenke, *Probability theory: a comprehensive course*, 3 **edition**. Springer Science & Business Media, 2020.

Chỉ mục