Lecture Notes: Xác Suất & Thống kê

• Chương 1: Cao Hoàng Nam (biên soạn lần đầu) & Vũ Hữu Nhự, Lê Đức Ninh (chỉnh sửa, bổ sung)
• Chương 2: Cao Hoàng Nam (biên soạn lần đầu) & Vũ Hữu Nhự, Lê Đức Ninh (chỉnh sửa, bổ sung)
• Chương 3,4: Vũ Hữu Nhự (biên soạn và chỉnh sửa)

Ngày 16 tháng 8 năm 2022

Mục lục

Chương 1

Xác suất

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

1.1.1 Một số khái niệm cơ bản trong lý thuyết xác suất

- *Thực nghiệm* (experiment) là một quá trình đo lường hay quan sát trong phòng thí nghiệm, trong một nhà máy, trên đường, trong tự nhiên, hay bất cứ đâu. Thực nghiệm mà chúng ta quan tâm trong lý thuyết xác suất là các quá trình ngẫu nhiên và thay đổi, vì thế chúng ta không thể tiên đoán một kết quả chính xác.

  Ví dụ 1.1. *Một số các thực nghiệm ngẫu nhiên có thể như sau:*

  1. *Tung một đồng xu.*
  2. *Gieo một con súc sắc.*
  3. *Đếm số tai nạn giao thông trong một ngày ở Hà Nội.*
  4. *Đo hiệu điện thế hai đầu một mạch điện.*

- Trong các thực nghiệm ngẫu nhiên, chúng ta thường thực hiện lặp đi lặp lại một nhóm các hành động để quan sát hiện tượng. Mỗi lần thực hiện một nhóm hành động như thế này được gọi là một phép thử (trial).

  Ví dụ 1.2. *Mỗi lần tung đồng xu hay mỗi lần gieo con súc sắc là một phép thử.*

- Kết quả của mỗi phép thử được gọi là một *kết cục* (outcome).

  Ví dụ 1.3. *1. Trong thực nghiệm tung một đồng xu, kết cục có thể là mặt sấp hoặc mặt ngửa.*

2. *Trong thực nghiệm gieo một con súc sắc, kết cục là số điểm xuất hiện trên mặt con súc sắc. Số điểm này có thể là* $1, 2, 3, 4, 5, 6$.
3. *Trong thực nghiệm đo hiệu điện thế hai đầu một mạch điện, kết cục là giá trị hiệu điện thế hai đầu mạch và nhận giá trị trong* $[0, +\infty)$ *(V).*

- Tập tất cả các kết cục có thể của thực nghiệm ngẫu nhiên được gọi là *không gian mẫu* (sample space) hay *không gian xác suất*. Không gian mẫu được ký hiệu là S. Mỗi điểm trong không gian mẫu được gọi là *điểm mẫu* (sample point) tương ứng với một kết cục của một phép thử.

  Ví dụ 1.4. 1. *Trong thực nghiệm tung một đồng xu, không gian mẫu là* $S = \{\textit{sấp}, \textit{ngửa}\}$.
  2. *Trong thực nghiệm gieo một con súc sắc, không gian mẫu là* $S = \{1, 2, 3, 4, 5, 6\}$.
  3. *Trong thực nghiệm đo hiệu điện thế hai đầu một mạch điên, không gian mẫu là* $S = [0, +\infty)$ *(V).*

- *Sự kiện* (hay *biến cố*) là một tập con của không gian mẫu. Nếu trong một phép thử, một kết cục a xuất hiện và $a \in A$ (a là một phần tử của A), chúng ta nói rằng sự kiện A xuất hiện (xảy ra). Kết cục có thể được xem là một *sự kiện đơn* (simple event) hay *sự kiện cơ sở*. Vì vậy, không gian mẫu có thể được gọi là không gian các sự kiện cơ sở.

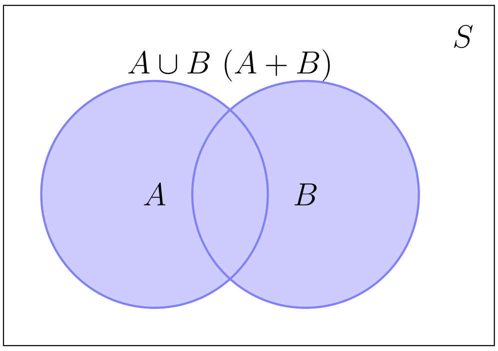
- *Sự kiện chắc chắn* là sự kiện luôn luôn xảy ra, biến cố này trùng với không gian mẫu S.

- *Sự kiện không thể* là sự kiện nhất định không xảy ra khi thực hiện phép thử. Biến cố không thể được kí hiệu là $\emptyset$.

  Ví dụ 1.5. *Trong thực nghiệm gieo một con súc sắc*

  1. $A = \{1, 3, 5\}$ *(số điểm xuất hiện trên mặt con súc sắc là lẻ),* $B = \{2, 4, 6\}$ *(số điểm xuất hiện trên mặt con súc sắc là chẵn),* $C = \{5, 6\}$*,... là các sự kiện.*
  2. $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$ *là các sự kiện cơ sở.*
  3. $D \equiv$ *"Xuất hiện mặt 7 chấm hoặc 0 chấm" là sự kiện không thể.*
  4. $S = \{1, 2, 3, 4, 5, 6\}$ *là sự kiện chắc chắn.*

  Khi thực hiện phép thử gieo một con súc sắc, nếu mặt 3 điểm xuất hiện, chúng ta nói rằng sự kiện A xuất hiện. Nếu mặt 6 điểm xuất hiện, chúng ta nói rằng sự kiện B (C) xuất hiện.



Hình 1.1: Biểu đồ Venn thể hiện hai sự kiện A, B trong không gian mẫu S và hợp của hai sự kiện $A \cup B$ (được tô màu)

1.1.2 Các phép toán trên sự kiện

- *Hợp (tổng)* của hai sự kiện A và B, được ký hiệu là $A \cup B$ (hoặc $A + B$), là một tập chứa tất cả các điểm mẫu trong A hoặc B hoặc cả hai. Hay nói cách khác, hợp của hai sự kiện A và B chỉ sự kiện khi có xuất hiện ít nhất một trong hai sự kiện trên. Tổng quát hợp của nhiều sự kiện như sau

$$\bigcup_{i=1}^{m} A_i = A_1 \cup A_2 \cup \cdots \cup A_m.$$

Ta có

$$\bigcup_{i=1}^{m} A_i \text{ xảy ra} \quad \Leftrightarrow \quad \left[\begin{array}{l} A_1 \text{ xảy ra} \\ A_2 \text{ xảy ra} \\ \ldots \\ A_m \text{ xảy ra.} \end{array}\right. \tag{1.1}$$

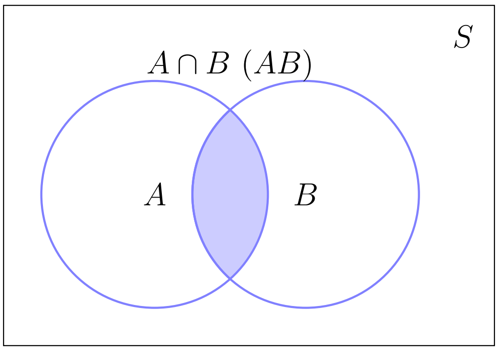
Biểu đồ Venn thể hiện hợp của hai sự kiện được thể hiện trong hình 1.1.

Ví dụ 1.6. *Trong thực nghiệm gieo một con súc sắc, cho các sự kiện sau*

$A \equiv$ *Số điểm lớn hơn 3*, $\quad B \equiv$ *Số điểm nhỏ hơn 6*, $\quad C \equiv$ *Số chẵn.*

Hợp của hai sự kiện A và B là $A \cup B = S$. Hợp của hai sự kiện A và C là $A \cup C = \{2, 4, 5, 6\}$.

- *Giao (tích)* của hai sự kiện A và B, được ký hiệu là $A \cap B$ (hoặc AB), là một tập chứa tất cả các điểm mẫu trong cả A và B. Hay nói cách



Hình 1.2: Biểu đồ Venn thể hiện hai sự kiện A, B trong không gian mẫu S và giao của hai sự kiện $A \cap B$ (được tô màu)

khác, giao của hai sự kiện A và B chỉ sự kiện khi có xuất hiện đồng thời hai sự kiện trên. Tổng quát giao của nhiều sự kiện như sau

$$\left(\prod_{i=1}^{m} A_i =\right) \quad \bigcap_{i=1}^{m} A_i = A_1 \cap A_2 \cap \cdots \cap A_m.$$

Ta có

$$\bigcap_{i=1}^{m} A_i \,\text{xảy ra} \quad \Leftrightarrow \quad \begin{cases} A_1 \,\text{xảy ra} \\ A_2 \,\text{xảy ra} \\ \ldots \\ A_m \,\text{xảy ra.} \end{cases} \tag{1.2}$$

Biểu đồ Venn thể hiện giao của hai sự kiện được thể hiện trong hình 1.2.

Ví dụ 1.7. *Xét Ví dụ 1.6 ở trên. Giao của A và B là $A \cap B = \{4, 5\}$. Giao của A và C là $A \cap C = \{4, 6\}$. Giao của A, B, và C là $A \cap B \cap C = \{4\}$.*

- *Sự kiện đối lập (biến cố đối lập)* của sự kiện A được ký hiệu là A^c hoặc $\bar{A}$, chỉ sự kiện không xuất hiện A. Đây là một tập con của S chứa tất cả các điểm mẫu không nằm trong A. Điều này nghĩa rằng

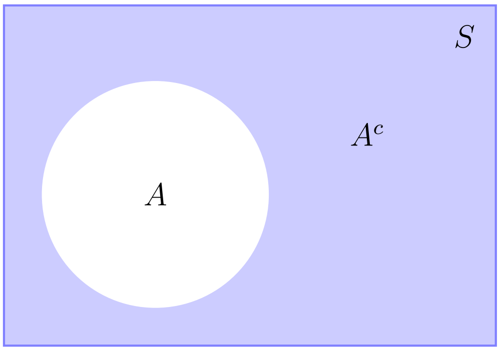
$$A \cap A^c = \emptyset, \quad A \cup A^c = S,$$

ở đây $\emptyset$ là sự kiện không thể. Ta có

$$A^c \,\text{xảy ra} \quad \Leftrightarrow \quad A \,\text{không xảy ra.} \tag{1.3}$$

Sự kiện A và sự kiện đối lập A^c được minh họa trong hình 1.3.

Ví dụ 1.8. *Xét Ví dụ 1.6 ở trên. Ta có $A^c = \{1, 2, 3\}$, $B^c = \{6\}$ và $C^c = \{1, 3, 5\}$.*



Hình 1.3: Biểu đồ Venn thể hiện sự kiện A và sự kiện đối lập A^c trong không gian mẫu S

1.2 Nhắc lại các quy tắc đếm

1.2.1 Qui tắc cộng

Để hoàn thành công việc người ta có thể chọn một trong k phương án.

- Phương án 1 có n_1 cách thực hiện
- Phương án 2 có n_2 cách thực hiện
  ....
- Phương án k có n_k cách thực hiện

Khi đó, có $n_1 + n_2 + \cdots + n_k$ cách thực hiện công việc đã cho.

1.2.2 Qui tắc nhân

Để hoàn thành công việc người ta chia công việc đó thành k giai đoạn.

- Giai đoạn 1 có n_1 cách thực hiện
- Giai đoạn 2 có n_2 cách thực hiện
  ....
- Giai đoạn k có n_k cách thực hiện

Khi đó, có $n_1 n_2 \ldots n_k$ cách thực hiện công việc đã cho.

1.2.3 Hoán vị

Một hoán vị của n phần tử là một cách sắp xếp có thứ tự n phần tử đó.

- Nếu n phần tử đôi một khác nhau, thì số các hoán vị n phần tử khác nhau là
$$P_n = n! = 1 \times 2 \times 3 \times \cdots \times n.$$
- Nếu có n phần tử được chia thành c lớp khác nhau, ở đây mỗi lớp chứa các vật giống nhau. Thì số hoán vị n phần tử là
$$\frac{n!}{n_1!n_2!\cdots n_c!} \quad (n_1 + n_2 + \cdots + n_c = n).$$

Nhận xét 1.1. *Chúng ta có một số lưu ý sau:*

(i) *Bởi định nghĩa ta có* $0! = 1$.

(ii) *Với mọi* $n \in \mathbb{N}$, *ta có* $(n+1)! = (n+1)n!$.

(iii) *Khi* n *lớn, chúng ta có gần đúng sau*
$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$
(công thức Stirling).

Ví dụ 1.9. *Sắp xếp 5 người vào một băng ghế. Hỏi có bao nhiêu cách?*

Giải*: Mỗi sắp xếp 5 người vào một băng ghế là một hoán vị của 5 phần tử. Vì thế số cách sắp xếp là* $5! = 120$.

Ví dụ 1.10. *Từ các chữ số* $1, 2$ *và* 3 *có thể lập được bao nhiêu số có 7 chữ số, trong đó chữ số 1 xuất hiện hai lần, chữ số 2 xuất hiện ba lần và chữ số 3 xuất hiện hai lần.*

Giải. $\frac{7!}{2!3!2!} = 201$ *số.*

1.2.4 Chỉnh hợp

- *Chỉnh hợp (không lặp)*: Một chỉnh hợp chập k của n phần tử là một nhóm có thứ tự gồm k phần tử lấy từ n phần tử đã cho, với $k \le n$. Số các chỉnh hợp chập k của n phần tử là
$$A_n^k = \frac{n!}{(n-k)!}.$$

- *Chỉnh hợp lặp*: Chỉnh hợp lặp chập k của n phần tử là một nhóm có thứ tự gồm k phần tử có thể giống nhau lấy từ n phần tử đã cho. Số các chỉnh hợp lặp chập k của n phần tử là

$$\tilde{A}_n^k = n^k.$$

Ví dụ 1.11. *Có một tập hợp gồm ba phần tử $\{a, b, c\}$. Hỏi có thể tạo ra bao nhiêu ký tự gồm hai phần tử từ tập hợp trên trong hai trường hợp sau:*

(i) Chú ý đến thứ tự và mỗi phần tử chỉ được chọn một lần.

(ii) Chú ý đến thứ tự nhưng mỗi phần tử có thể được chọn nhiều lần.

Giải: *(i) Số các ký tự thu được là thu được là $A_3^2 = 3.2 = 6$. Đó là ab, ba, ac, ca, bc, và cb.*

(ii) Số các ký tự thu được là $3^2 = 9$. Đó là aa, ab, ba, ac, ca, bb, bc, cb, và cc.

1.2.5 Tổ hợp

Một tổ hợp chập k của n phần tử là một nhóm gồm k phần tử khác nhau không phân biệt thứ tự lấy từ n phần tử đã cho, với $k \leq n$. Số các tổ hợp như vậy là

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!} \quad = \binom{n}{k}. \tag{1.4}$$

Ví dụ 1.12. *Hỏi có bao nhiêu cách lập một hội đồng gồm 3 thành viên được chọn từ 8 người?*

Giải*: Hội đồng là một nhóm gồm 3 người không phân biệt thứ tự được chọn từ 8 người, vì vậy ta có $C_8^3 = 56$ cách lập.*

1.3 Định nghĩa xác suất, các tính chất cơ bản của xác suất

1.3.1 Định nghĩa cổ điển của xác suất

Định nghĩa 1.1 (Xác suất cổ điển)**.** Cho không gian mẫu S của một thực nghiệm ngẫu nhiên, chứa một số hữu hạn các điểm mẫu (các kết cục). Giả sử khả năng cho các điểm mẫu xuất hiện là như nhau, thì xác suất cho một sự kiện A xuất hiện là

$$P(A) = \frac{\text{Số các điểm mẫu trong A}}{\text{Số các điểm mẫu trong S}} = \frac{\text{card}(A)}{\text{card}(S)}, \tag{1.5}$$

ở đó card(M) là số các điểm mẫu trong sự kiện M.

Ví dụ 1.13. *Cho thực nghiệm gieo một con súc sắc đồng nhất với xác suất xuất hiện các mặt là bằng nhau. Tìm xác suất để nhận các sự kiện sau*

(i) $A = \{1, 3, 5\}$ *(số điểm xuất hiện trên mặt con súc sắc là lẻ).*

(ii) $B = \{5, 6\}$.

Giải*: Không gian mẫu* $S = \{1, 2, 3, 4, 5, 6\}$.

(i) Số điểm trong A là 3 điểm, số điểm trong S là 6 điểm, vì thế $P(A) = \frac{3}{6} = \frac{1}{2}$.

(ii) Số điểm trong B là 2 điểm, vì thế $P(B) = \frac{2}{6} = \frac{1}{3}$.

Ngoài ra xác suất có thể được định nghĩa thông qua tần suất như sau

$$P_{\text{rel}}(A) = \frac{\text{Số lần sự kiện } A \text{ xuất hiện}}{\text{Số lần phép thử}}. \tag{1.6}$$

Xác suất được định nghĩa theo cách này là một giá trị gần đúng. Tuy nhiên trong nhiều ngành khoa học thực nghiệm, xác suất được xác định theo công thức (1.6) đạt độ chính xác khá lớn và rất phù hợp với tính toán lý thuyết. Khi số lần phép thử được thực hiện rất lớn (tiến ra vô cùng), xác suất được định nghĩa bằng tần suất sẽ tiến về giá trị thật sự.

1.3.2 Định nghĩa tổng quát

Định nghĩa 1.2 (Định nghĩa tổng quát về xác suất)**.** Cho một không gian mẫu S, với mỗi sự kiện A (một tập con của S) sẽ có một số tương ứng là $P(A)$ được gọi là *xác suất* để sự kiện A xuất hiện, sao cho các tiên đề sau đồng thỏa mãn:

1. Cho mỗi A trong S,
$$0 \le P(A) \le 1. \tag{1.7}$$

2. Không gian mẫu có xác suất

$$P(S) = 1. \tag{1.8}$$

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

$$P(A \cup B) = P(A) + P(B). \tag{1.9}$$

Nếu S có vô hạn các phần tử, tiền đề này được thay như sau

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots , \tag{1.10}$$

ở đây các sự kiện A_1, A_2,..., A_n,... xung khắc từng đôi một, nghĩa là $A_i \cap A_j = \emptyset \ \forall i \neq j$.

1.3.3 Các định lý cơ bản của xác suất

Định lý 1.1 (Qui tắc phần bù)**.** *Cho một sự kiện A và sự kiện đối lập của nó là A^c trong một không gian mẫu S, khi đó*

$$P(A^c) = 1 - P(A). \tag{1.11}$$

Ví dụ 1.14. *Tung năm đồng xu đồng nhất đồng thời. Tìm xác suất để sự kiện*

$$A \equiv \text{"Mặt sấp xuất hiện ít nhất một lần"}$$

xuất hiện.

Giải*: Không gian mẫu chứa $2^5 = 32$ kết cục (điểm mẫu). Xác suất cho mỗi kết cục là như nhau và bằng $1/32$. Sự kiện phủ định của A là*

$$A^c \equiv \text{"Không có bất kỳ mặt sấp xuất hiện"}.$$

Sự kiện A^c chỉ có một kết cục, vì thế xác suất để sự kiện A^c xuất hiện là $P(A^c) = 1/32$. Áp dụng Định lý 1.1, ta tìm được xác suất để sự kiện A xuất hiện là

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{32} = \frac{31}{32}.$$

Định lý 1.2 (Qui tắc cộng cho cặp sự kiện xung khắc)**.** *Cho các sự kiện A_1, A_2,..., A_n xung khắc nhau từng đôi một trong một không gian mẫu S, khi đó*

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \tag{1.12}$$

Ví dụ 1.15. *Biết trong một ngày một gara có số* $10-20$ *(10 đến 20),* $21-30$, $31-40$, *trên* 40 *xe đến sửa chữa tương ứng với khả năng (xác suất) là:* $0.20, 0.35, 0.25, 0.12$. *Hỏi xác suất để gara này nhận nhiều hơn* 20 *xe để phục vụ trong một ngày là bao nhiêu?*

Giải: *Vì các sự kiện được cho ban đầu là hoàn toàn xung khắc nhau, áp dụng Định lý 1.2 ta có: xác suất để gara này nhận nhiều hơn* 20 *xe đến sửa chữa trong một ngày là*

$$P = 0.35 + 0.25 + 0.12 = 0.72.$$

Định lý 1.3 (Qui tắc cộng)**.** *Cho hai sự kiện* A *và* B *trong không gian mẫu* S*, khi đó*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.13}$$

Ví dụ 1.16. *Trong thực nghiệm tung một con súc sắc đồng nhất, xét các sự kiện sau*

$A \equiv$ *"Số điểm xuất hiện trên mặt con súc sắc là lẻ",*

$B \equiv$ *"Số điểm xuất hiện trên mặt con súc sắc nhỏ hơn 4".*

Tìm xác suất để sự kiện $A \cup B$ *xuất hiện.*

Giải: Áp dụng Định lý 1.3, ta có

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3}.$$

1.4 Xác suất có điều kiện, công thức nhân xác suất, sự kiện độc lập

1.4.1 Xác suất có điều kiện

Xác suất của một sự kiện có thể phụ thuộc vào nhiều yếu tố, điều kiện khác nhau. Để chỉ ra một cách cụ thể hơn về việc xác suất của một sự kiện A nào đó phụ thuộc vào một điều kiện B nào đó ra sao, người ta đưa ra khái niệm xác suất có điều kiện. Điều kiện B cũng có thể hiểu là một sự kiện, tức là có sự kiện B.

Định nghĩa 1.3 (Xác suất có điều kiện-Conditional probability)**.** Xác suất của một sự kiện A nào đó, biết rằng sự kiện B khác đã xảy ra, được gọi là *xác suất có điều kiện của* A *khi* B *đã xảy ra.* Ký hiệu là $P(A|B)$ và đọc là "xác suất của A khi biết B xảy ra".

Giả sử trong một không gian mẫu, sự kiện B xuất hiện có xác suất $P(B) > 0$, thì xác suất của sự kiện A khi biết B xảy ra, được cho bởi công thức sau

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.14}$$

Tương tự, ta cũng có thể xét xác suất của sự kiện B khi biết A xảy ra với $P(A) > 0$, được ký hiệu là $P(B|A)$. Khi đó ta có

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

Xét hai sự kiện A và B là hai tập con trong không gian mẫu S ban đầu. Chúng ta có thể giải thích xác suất của sự kiện A dưới điều kiện B như sau. Khi xét điều kiện B xảy ra, thì chúng ta đã hạn chế không gian mẫu S xuống còn B, và hạn chế sự kiện A xuống còn $A \cap B$. Khi đó xác suất có điều kiện $P(A|B)$ chính là xác suất sự kiện $A \cap B$ trong không gian mẫu mới B.

Ví dụ 1.17. *Một lớp học có 30 bạn, trong đó có 17 bạn nữ và 13 bạn nam. Có ba bạn tên là Thanh, trong đó có 1 bạn nữ và 2 bạn nam. Thầy giáo gọi ngẫu nhiên 1 bạn lên bảng. Tìm xác suất cho các trường hợp sau:*

(i) để bạn đó có tên là Thanh?

(ii) để bạn đó có tên là Thanh với điều kiện đó là bạn nữ?

(iii) để bạn đó là nữ và biết rằng bạn được gọi tên là Thanh?

Giải: *Gọi các sự kiện sau $A \equiv$ "bạn được gọi có tên là Thanh" và $B \equiv$ "bạn được gọi là nữ". Không gian mẫu có 30 phần tử, sự kiện A có 3 phần tử, sự kiện B có 17 phần tử, và sự kiện $A \cap B$ có một phần tử.*

(i) $P(A) = \frac{3}{30} = \frac{1}{10}$.

(ii) $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/30}{17/30} = \frac{1}{17}$.

(iii) $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/30}{3/30} = \frac{1}{3}$.

1.4.2 Công thức nhân xác suất. Các sự kiện độc lập

Định lý 1.4 (Công thức nhân xác suất)**.** *Cho A và B là các sự kiện trong không gian mẫu S với $P(A), P(B) \neq 0$. Khi đó, ta có công thức*

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

Định nghĩa 1.4 (Các sự kiện độc lập)**.** Hai sự kiện A và B thỏa mãn

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

được gọi là các sự kiện *độc lập*.

Từ các phương trình (1.14)–(1.15) và Định lý 1.4, ta có

$$A, B \text{ độc lập} \implies P(A|B) = P(A), \quad P(B|A) = P(B).$$

Cho 3 sự kiện A, B, và C. Chúng được gọi là độc lập nếu

$$\begin{cases} P(A \cap B) = P(A)P(B), \\ P(B \cap C) = P(B)P(C), \\ P(C \cap A) = P(C)P(A), \\ P(A \cap B \cap C) = P(A)P(B)P(C). \end{cases}$$

Tương tự, n sự kiện được gọi là độc lập nếu chúng thỏa mãn các hệ thức sau

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k}), \tag{1.18}$$

với mọi $1 \leq j_1, j_2, \ldots j_k \leq n$ và với mọi $2 \leq k \leq n$.

Ví dụ 1.18. *Một hộp có 10 cái đinh ốc, trong đó có 3 cái bị lỗi. Lấy ngẫu nhiên lần lượt 2 cái đinh ốc ra khỏi hộp. Tìm xác suất để 2 đinh ốc lấy ra không bị lỗi trong hai trường hợp sau:*

(i) Đưa đinh ốc thứ nhất trở lại hộp sau khi lấy ra.

(ii) Không đưa đinh ốc thứ nhất trở lại hộp sau khi lấy ra.

Giải: *Gọi các sự kiện sau*

$$A \equiv \text{"Đinh ốc thứ nhất lấy ra không bị lỗi"},$$
$$B \equiv \text{"Đinh ốc thứ hai lấy ra không bị lỗi"}.$$

Chúng ta dễ dàng tìm thấy $P(A) = \frac{7}{10}$.

(i) Nếu đưa đinh ốc thứ nhất trở lại hộp sau khi lấy ra, thì $P(B|A) = P(B) = \frac{7}{10}$. *Theo Định lý 1.4, ta có*

$$P(A \cap B) = P(A)P(B|A) = \frac{7}{10}.\frac{7}{10} = 0.49$$

(ii) Nếu không đưa đinh ốc thứ nhất trở lại hộp sau khi lấy ra, thì $P(B|A) = \frac{6}{9}$. *Theo Định lý 1.4, ta có*

$$P(A \cap B) = P(A)P(B|A) = \frac{7}{10}.\frac{6}{9} \approx 0.47.$$

1.5 Công thức xác suất toàn phần, công thức Bayes

1.5.1 Công thức xác suất toàn phần

Định nghĩa 1.5 (Hệ đầy đủ các sự kiện)**.** Các sự kiện A_1, A_2,. . ., A_n ($n \geq 2$) của một không gian mẫu S được gọi là (hay tạo thành) một *hệ đầy đủ* nếu

(i) Chúng xung khắc từng đôi một: $A_i \cap A_j = \emptyset$, $\forall i \neq j$;

(ii) $A_1 \cup A_2 \cup \cdots \cup A_n = S$.

Dễ thấy $\{A, A^c\}$ là một hệ đầy đủ với bất kỳ sự kiện A.

Định lý 1.5 (Công thức xác suất toàn phần (đầy đủ))**.** *Cho $\{A_1, ..., A_n\}$ là một hệ đầy đủ các sự kiện trong không gian mẫu S. Khi đó với mọi sự kiện B (trong không gian S), ta có*

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i). \tag{1.19}$$

Ví dụ 1.19. *Một phân xưởng có 3 nhà máy I, II, III sản xuất cùng loại sản phẩm với tỷ lệ phế phẩm tương ứng là 1%, 0.5%, 0.2%. Biết rằng các nhà máy I, II, và III sản xuất tương ứng 35%, 45%, và 20% sản phẩm. Chọn ngẫu nhiên một sản phẩm, tìm xác suất đó là phế phẩm.*

Giải*: Gọi A_1, A_2, và A_3 tương ứng là các sự kiện chọn ra sản phẩm do nhà máy I, II, và III sản xuất. Chúng ta có thể chứng minh rằng tập $\{A_1, A_2, A_3\}$ tạo nên một hệ đầy đủ và có các xác suất*

$$P(A_1) = 0.35, \quad P(A_2) = 0.45, \quad P(A_3) = 0.2.$$

Gọi B là sự kiện rút được phế phẩm. Ta có thể tìm thấy các xác suất có điều kiện sau

$$P(B|A_1) = 0.01, \quad P(B|A_2) = 0.005, \quad P(B|A_3) = 0.002.$$

Áp dụng công thức xác suất toàn phần, ta có

$$\begin{aligned} P(B) = \sum_{i=1}^{3} P(B|A_i)P(A_i) &= 0.35 \times 0.01 + 0.45 \times 0.005 + 0.2 \times 0.002 \\ &= 0.00615. \end{aligned}$$

1.5.2 Công thức Bayes (Công thức xác suất hậu kiểm)

Định lý 1.6 (Công thức Bayes)**.** *Với mọi sự kiện B sao cho $P(B) > 0$, ta có*

$$P(C|B) = \frac{P(B|C)P(C)}{P(B)}.$$

Nếu biết $\{A_1, \ldots, A_n\}$ là một hệ đầy đủ các sự kiện, thì ta có thể tính $P(B)$ như sau:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i). \tag{1.20}$$

Nhận xét 1.2. *Trong Định lý 1.6 các xác suất $P(C)$ được gọi là* xác suất tiên nghiệm (tiền kiểm)*. Sau khi quan sát được sự kiện B, các xác suất $P(C|B)$ là* xác suất hậu kiểm*. Do đó công thức xác suất Bayes còn gọi là công thức xác suất hậu kiểm.*

Ví dụ 1.20. *Một mạch điện gồm hai bộ phận mắc nối tiếp, với xác suất làm việc bình thường trong một khoảng thời gian nào đó là* 0.95 *và* 0.98*. Ở một thời điểm trong khoảng thời gian trên người ta thấy mạch điện ngưng làm việc (do bộ phận nào đó hỏng). Tìm xác suất để chỉ đúng bộ phận thứ hai hỏng.*

Giải*: Do hai bộ phận mắc nối tiếp nên chỉ cần một bộ phận hỏng là mạch ngừng làm việc. Chúng ta có bốn khả năng sau:*

$A_1 \equiv$ *"Cả hai bộ phận đều làm việc bình thường",*

$A_2 \equiv$ *"Bộ phận I làm việc bình thường, bộ phận II hỏng",*

$A_3 \equiv$ *"Bộ phận I hỏng, bộ phận II làm việc bình thường",*

$A_4 \equiv$ *"Cả hai bộ phận đều hỏng".*

Dễ thấy $\{A_1, A_2, A_3, A_4\}$ là một hệ đầy đủ các sự kiện. Các xác suất tương ứng là

$$\begin{aligned} P(A_1) &= 0.95 \times 0.98 = 0.931, \\ P(A_2) &= 0.95 \times (1 - 0.98) = 0.019, \\ P(A_3) &= (1 - 0.95) \times 0.98 = 0.049, \\ P(A_4) &= (1 - 0.95)(1 - 0.98) = 0.001. \end{aligned}$$

Gọi B là sự kiện mạch điện không làm việc, ta có

$$P(B|A_1) = 0, \quad P(B|A_2) = P(B|A_3) = P(B|A_4) = 1.$$

Sử dụng công thức Bayes, ta tìm được

$$P(A_2|B) = \frac{1 \times 0.019}{0 \times 0.931 + 1 \times 0.019 + 1 \times 0.049 + 1 \times 0.001} = \frac{19}{69}.$$

Ví dụ 1.21. *Một trạm phát tín hiệu chỉ phát hai loại tín hiệu loại I và loại II với xác suất lần lượt là* 0.85 *và* 0.15*. Do có nhiễu trên đường truyền nên* $\frac{1}{7}$ *tín hiệu loại I bị méo và thu được như tín hiệu loại II, và* $\frac{1}{8}$ *tín hiệu loại II bị méo và thu được như tín hiệu loại I.*

(i) Tính xác suất để thu được tín hiệu loại I?

(ii) Giả sử đã thu được tín hiệu loại I. Tìm xác suất để thu được đúng loại tín hiệu đã phát?

Giải*: Gọi*

$$A_1 \equiv \text{"Tín hiệu loại I được phát"},$$
$$A_2 \equiv \text{"Tín hiệu loại II được phát"},$$
$$B \equiv \text{"Thu được tín hiệu loại I"}.$$

Dễ thấy hệ $\{A_1, A_2\}$ *là hệ đầy đủ. Hơn nữa, ta có*

$$P(A_1) = 0.85, \quad P(A_2) = 0.15, \quad P(B|A_1) = \frac{6}{7}, \quad P(B|A_2) = \frac{1}{8}.$$

(i) Theo công thức xác suất toàn phần, ta có

$$P(B) = P(B|A_1)P(A_1)+P(B|A_2)P(A_2) = \frac{6}{7}\times 0.85+\frac{1}{8}\times 0.15 \approx 0.747.$$

(ii) Theo công thức Bayes, xác suất để thu được đúng loại tín hiệu khi nhận được tín hiệu loại I là

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{0.85 \times 6/7}{0.747} \approx 0.975.$$

1.6 Khái niệm về biến ngẫu nhiên

Các bài toán thực tế đưa đến một khái niệm mới: *biến ngẫu nhiên* (random variable; stochastic variable), được viết tắt trong giáo trình này là BNN và được kí hiệu bởi các chữ cái in hoa $X, Y, Z, ...$ Vậy BNN là gì? Chúng ta hãy bắt đầu bằng một số ví dụ sau.

Ví dụ 1.22. *Gieo một con súc sắc đồng nhất. Không gian xác suất là* $S = \{1, 2, 3, 4, 5, 6\}$. *Chúng ta xét hai kịch bản sau:*

(i) *Kịch bản 1: Gieo con súc sắc đơn thuần. Trong trường hợp này, nếu chúng ta gọi X là số điểm trên mặt con súc sắc xuất hiện sau khi gieo, thì X có thể nhận các giá trị 1, 2, 3, 4, 5, 6. Ta nói rằng X là một biến ngẫu nhiên.*

(ii) *Kịch bản 2: Gieo con súc sắc nhận thưởng. Nếu mặt một điểm xuất hiện, bạn nhận được 10000 đồng tiền thưởng. Nếu mặt hai điểm xuất hiện, bạn nhận được 20000 đồng tiền thưởng. Tương tự cho mặt 3, 4, 5, 6 điểm xuất hiện, bạn sẽ nhận được 30000, 40000, 50000, 60000 đồng tiền thưởng tương ứng. Trong trường hợp này, nếu chúng ta gọi X là số tiền thưởng người chơi nhận sau khi gieo súc sắc, thì X có thể nhận các giá trị 10000, 20000, 30000, 40000, 50000, 60000 (đồng). Ta nói rằng X là một biến ngẫu nhiên.*

Ví dụ 1.23. *Bắn một viên đạn vào bia có bán kính 20 cm. Chúng ta xét hai kịch bản sau:*

(i) *Kịch bản 1: Bắn đạn một cách đơn thuần. Trong trường hợp này, chúng ta gọi X là khoảng cách d từ tâm bia đến điểm đạn dính trên bia. Khi đó X có thể nhận các giá trị nằm trong khoảng từ 0 (cm) đến 20 (cm). Ta nói rằng X là một biến ngẫu nhiên.*

(ii) *Kịch bản 2: Bắn đạn trong một cuộc thi. Nếu* $d \in [0, 2)$ *cm, người bắn được 10 điểm. Nếu* $d \in [2, 4)$ *cm, người chơi được 9 điểm. Tương tự, người chơi có thể nhận 8, 7, 6, 5, 4, 3, 2, 1 điểm. Người chơi nhận không điểm nếu* $d > 20$ *cm (đạn không dính bia). Trong trường hợp này, chúng ta gọi X là số điểm người tham gia thi bắn đạn nhận được. Khi đó X có thể nhận các giá trị 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Ta nói rằng X là một biến ngẫu nhiên.*

Định nghĩa 1.6. Một BNN X là một hàm được định nghĩa trên không gian xác suất, giá trị của nó là các số thực. Điều này được thể hiện thông qua ánh xạ sau:

$$X : S \to \mathbb{R}.$$

Với mỗi $a \in \mathbb{R}$ và bất kì khoảng $I \subset \mathbb{R}$, các xác suất $P(X = a)$ và $P(X \in I)$ đều được xác định.

BNN có thể được phân thành hai loại: BNN rời rạc và BNN liên tục.

- BNN rời rạc xuất hiện trong các thực nghiệm ở đó chúng ta thực hiện việc đếm: đếm số xe trên đường phố, đếm số sản phẩm bị hỏng trong quá trình sản xuất,...
- BNN liên tục xuất hiện trong các thực nghiệm ở đó chúng ta thực hiện phép đo lường: đo cường độ dòng điện, đo trọng lượng, đo độ dài,...

1.7 Phân bố của biến ngẫu nhiên

Định nghĩa 1.7 (Hàm phân bố xác suất)**.** *Hàm phân bố xác suất* (distribution function; cumulative distribution function–viết tắt là CDF) của BNN X là hàm $F(x)$ xác định với mọi $x \in \mathbb{R}$ bởi công thức sau

$$F(x) = P(X \leq x). \tag{1.21}$$

Ta thấy rằng $F(x)$ là xác suất của các biến ngẫu nhiên X không vượt quá một giá trị x bất kì nào đó. Hàm này có đặc điểm là một hàm không giảm, tức là nếu $a < b$ thì $F(a) \leq F(b)$.

1.7.1 Phân bố của biến ngẫu nhiên rời rạc

- Cho X là một biến ngẫu nhiên rời rạc. X có thể nhận các giá trị sau x_1, x_2, x_3, $\cdots$, với các xác suất (dương) như sau

  $$p_1 = P(X = x_1), \quad p_2 = P(X = x_2), \quad p_3 = P(X = x_3), \ldots,$$

  với $P(X \in I) = 0$ cho bất kỳ khoảng I không chứa các giá trị có thể xuất hiện trong thực nghiệm và $\sum_i p_i = 1$.

- Chúng ta định nghĩa *hàm xác suất* như sau

  $$f(x) = \begin{cases} p_i, & x = x_i \\ 0, & \text{các trường hợp khác,} \end{cases} \qquad (i = 1, 2, 3, \cdots). \tag{1.22}$$

  Bảng phân bố xác suất của X có dạng

X	x_1	x_2	$\cdots$
P	p_1	p_2	$\cdots$

- *Hàm phân bố* được định nhĩa như sau

  $$P(X \leq x) \equiv F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} p_i \tag{1.23}$$

- Xác suất tương ứng với một khoảng

$$P(a < X \leq b) = F(b) - F(a) = \sum_{a < x_i \leq b} p_i. \tag{1.24}$$

Ví dụ 1.24. *Tìm hàm xác suất và phân bố xác suất của BNN X trong đó X là số chấm xuất hiện khi thực nghiệm gieo con súc sắc đồng nhất.*

Giải: *Ta có*

X	1	2	3	4	5	6
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$F(x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1

Từ bảng trên ta có, hàm xác suất cần tìm là

$$f(x) = \begin{cases} \frac{1}{6} & \textit{với } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \textit{với } x \notin \{1, 2, 3, 4, 5, 6\} \end{cases} \tag{1.25}$$

và hàm phân bố xác suất là

$$F(x) = \begin{cases} 0 & \textit{với } x < 1, \\ \dfrac{1}{6} & \textit{với } 1 \leq x < 2, \\ \dfrac{1}{3} & \textit{với } 2 \leq x < 3, \\ \dfrac{1}{2} & \textit{với } 3 \leq x < 4, \\ \dfrac{2}{3} & \textit{với } 4 \leq x < 5, \\ \dfrac{5}{6} & \textit{với } 5 \leq x < 6, \\ 1 & \textit{với } x \geq 6. \end{cases}$$

Đồ thị của hàm xác suất $f(x)$ và hàm phân bố $F(x)$ được minh họa trong Hình 1.4

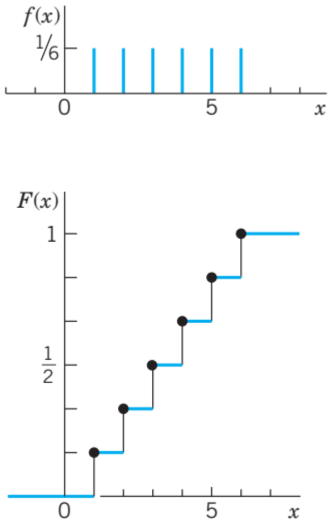
1.7.2 Phân bố của biến ngẫu nhiên liên tục

- *Hàm phân bố* được định nhĩa như sau

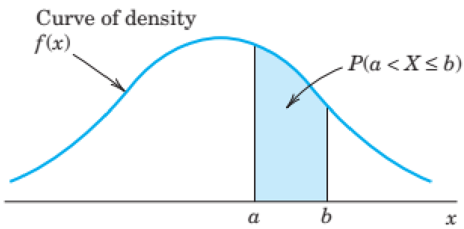
$$P(X \leq x) \equiv F(x) = \int_{-\infty}^{x} f(v)dv. \tag{1.26}$$

Trong đó, $f(x)$ được gọi là *hàm mật độ xác suất* và ta có

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



Hình 1.4: Hàm mật độ xác suất $f(x)$ và hàm phân bố $F(x)$



Hình 1.5: Hình minh họa công thức (1.27).

- Xác suất tương ứng với một khoảng

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(v)dv. \tag{1.27}$$

  Công thức này được minh họa trong Hình 1.5.

- Các xác suất cho các khoảng sau

$$a < X \leq b, \quad a \leq X < b, \quad a \leq X \leq b, \quad a < X < b,$$

  bằng nhau, tức là

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

Ví dụ 1.25. *Cho X là một BNN với hàm mật độ xác suất sau*

$$f(x) = \begin{cases} 0.75(1 - x^2), & -1 \leq x \leq 1 \\ 0, & \textit{các trường hợp khác.} \end{cases}$$

(i) Tìm hàm phân bố xác suất.

(ii) Tìm xác suất $P(-\frac{1}{2} \leq X \leq \frac{1}{2})$.

Giải*:*

(i) Hàm phân bố xác suất tương ứng với mật độ xác suất trên là

$$F(x) = \begin{cases} 0, & x < -1 \\ 0.75 \int_{-1}^{x} (1 - v^2) dv = 0.5 + 0.75x - 0.25x^3, & -1 \leq x \leq 1 \\ 1, & x > 1. \end{cases}$$

(ii) Xác suất $P(-\frac{1}{2} \leq X \leq \frac{1}{2})$ *được tính như sau*

$$P(-\frac{1}{2} \leq X \leq \frac{1}{2}) = F(\frac{1}{2}) - F(-\frac{1}{2}) = 68.75\%$$

1.8 Các số đặc trưng của biến ngẫu nhiên

Trung bình μ và *phương sai* σ^2 của một biến ngẫu nhiên X và phân bố của nó là tương tự với trung bình $\bar{x}$ và phương sai s^2 của phân bố tần suất được thảo luận trong mục 2.1.2. Trung bình μ đặc trưng cho vị trí trung tâm của phân bố. Ngoài ra trung bình μ còn được gọi là *kỳ vọng* của X, bởi nó đưa ra giá trị trung bình của X được mong muốn trong nhiều phép thử. Trung bình μ được định nghĩa như sau

$$\mu = \sum_i x_i p_i \qquad \text{(Phân bố rời rạc)} \tag{1.28a}$$

$$\mu = \int_{-\infty}^{+\infty} x f(x) dx \qquad \text{(Phân bố liên tục).} \tag{1.28b}$$

Trong khi đó phương sai σ^2 đặc trưng cho sự phân tán của phân bố và được định nghĩa như sau

$$\sigma^2 = \sum_i (x_i - \mu)^2 p_i \qquad \text{(Phân bố rời rạc)} \tag{1.29a}$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \qquad \text{(Phân bố liên tục).} \tag{1.29b}$$

Từ định nghĩa chúng ta có thể thấy rằng $\sigma^2 \geq 0$. Dấu bằng xảy ra trong trường hợp như biến ngẫu nhiên X là hằng số khắp mọi nơi (đối với phân bố liên tục) hay X chỉ nhận một giá trị và bằng μ (đối với phân bố rời rạc). Giá trị dương của căn bậc hai của phương sai σ^2 được gọi là *độ lệch chuẩn* của biến ngẫu nhiên X và phân bố của nó, và được ký hiệu là σ.

Ví dụ 1.26. *Cho thực nghiệm tung một đồng xu đồng nhất. Gọi biến ngẫu nhiên X là số mặt sấp (S) xuất hiện. Tính trung bình, phương sai của biến ngẫu nhiên này.*

Giải*: Các giá trị có thể của X là 0 (nếu mặt ngửa (N) xuất hiện) và 1 (nếu mặt sấp xuất hiện). Do đồng xu đồng nhất nên ta có các xác suất $P(X=0)=\frac{1}{2}$ và $P(X=1)=\frac{1}{2}$. Bằng cách áp dụng các công thức trên cho phân bố gián đoạn, chúng ta có thể tính trung bình và phương sai như sau*

$$\mu = 0\times\frac{1}{2}+1\times\frac{1}{2}=\frac{1}{2},$$
$$\sigma^2=(0-\frac{1}{2})^2\times\frac{1}{2}+(1-\frac{1}{2})^2\times\frac{1}{2}=\frac{1}{4}.$$

Ví dụ 1.27. *Cho hàm mật độ xác suất của một biến ngẫu nhiên X như sau*

$$f(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b\\ 0, & \textit{khác}.\end{cases}$$

Tính trung bình và phương sai của phân bố này.

Giải*: Bằng cách áp dụng các công thức trên cho phân bố liên tục, chúng ta có thể tính trung bình và phương sai như sau*

$$\mu=\int_a^b \frac{x}{b-a}dx=\frac{a+b}{2},$$
$$\sigma^2=\int_a^b\left(x-\frac{a+b}{2}\right)^2\frac{1}{b-a}dx=\frac{(b-a)^2}{12}.$$

Định lý 1.7 (Trung bình của phân bố đối xứng)**.** *Nếu một phân bố X đối xứng tương ứng với $x=c$, có nghĩa là $f(x-c)=f(x+c)$ với mọi x, thì $\mu=c$.*

Định lý 1.8 (Phép biến đổi của trung bình và phương sai)**.** *Cho biến ngẫu nhiên X có trung bình μ và phương sai σ^2. Khi đó*

(a) Biến ngẫu nhiên $X^*=a_1+a_2X$ $(a_2>0)$ có trung bình μ^* và phương sai σ^{*2} được tính bởi

$$\mu^*=a_1+a_2\mu \quad \sigma^{*2}=a_2^2\sigma^2.$$

(b) Đặc biệt, BNN chuẩn hóa (standardized random variable) Z tương ứng với X, có nghĩa,

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

có trung bình $\mu_Z=0$ và phương sai $\sigma_Z^2=1$.

1.9 Một số phân bố xác suất thường gặp

1.9.1 Phân bố nhị thức $X \sim B(n, p)$

Chúng ta quan tâm đến số lần sự kiện A xuất hiện trong n phép thử. Ở đây, chúng ta giả sử rằng các phép thử là độc lập và không ảnh hưởng lẫn nhau. Trong mỗi phép thử, sự kiện A có cùng một xác suất $P(A) = p$. Điều ngày gợi ý rằng, sự kiện A không xuất hiện trong mỗi phép thử với xác suất $P(A^c) = 1 - p$. Trong thực nghiệm này, biến số ngẫu nhiên chúng ta quan tâm là

$X \equiv$ "số lần sự kiện A xuất hiện trong n phép thử".

Các giá trị có thể của X là $0, 1, \ldots, n$. Chúng ta cần xác định hàm xác suất và phân bố xác suất tương ứng với biến ngẫu nhiên này. Hàm xác suất tương ứng được cho như sau

$$f(x) = P(X = x) = C_n^x p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x} \quad (x = 0, 1, 2, \cdots, n) \tag{1.30}$$

và được minh họa trong Hình 1.6. Phân bố xác suất tương ứng với hàm xác suất này được gọi là *phân bố nhị thức* (binomial distribution) hay *phân bố Bernoulli* (Bernoulli distribution).

Chúng ta có thể chứng minh giá trị trung bình và phương sai của phân bố nhị thức như sau

$$\mu = np, \tag{1.31}$$

$$\sigma^2 = np(1-p). \tag{1.32}$$

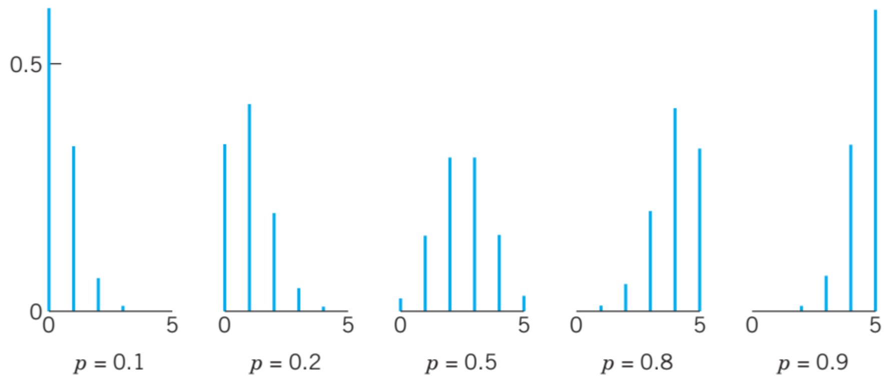
Trong trường hợp *đối xứng* của khả năng xuất hiện và không xuất hiện của sự kiện A ($p = \frac{1}{2}$), trung bình $\mu = \frac{n}{2}$, phương sai $\sigma^2 = \frac{n}{4}$ và hàm xác suất là

$$f(x) = P(X = x) = \binom{n}{x} \left(\frac{1}{2}\right)^n \quad (x = 0, 1, \ldots, n).$$

Ví dụ 1.28. *Tính xác suất để nhận được ít nhất hai mặt sáu điểm (mặt sáu điểm xuất hiện ít nhất hai lần) trong thực nghiệm tung con súc sắc đồng nhất bốn lần.*

Giải*: Ta có hàm xác suất tương ứng với thực nghiệm này như sau*

$$f(x) = C_4^x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{n-x} \quad (x = 0, 1, 2, 3, 4).$$



Hình 1.6: Hàm xác suất (1.30) của phân bố nhị thức với $n = 5$ và p nhận một số giá trị khác nhau

Sử dụng hàm xác suất này, chúng ta có thể tính được xác suất nhận được ít nhất hai mặt sáu điểm như sau

$$P[X \geq 2] = f(2) + f(3) + f(4) = \frac{171}{1296} = 13.2\%.$$

1.9.2 Phân bố Poisson $X \sim P(\mu)$

Phân bố rời rạc của biến ngẫu nhiên nhận vô hạn các giá trị có hàm xác suất như sau

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad (x = 0, 1, 2, 3, \cdots), \tag{1.33}$$

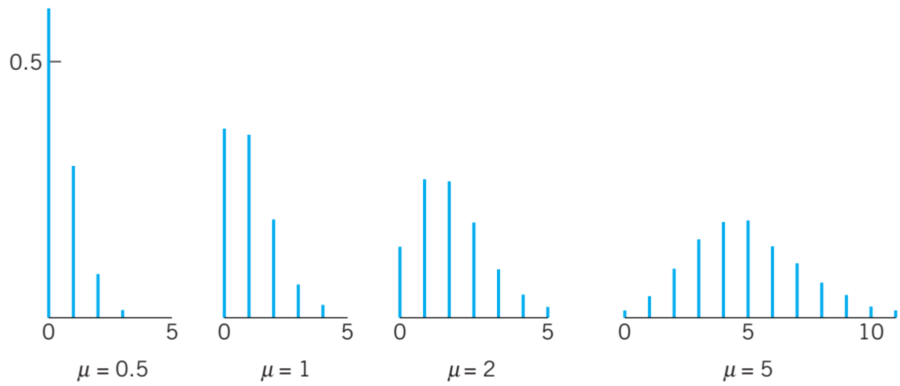
được gọi là *phân bố Poisson* với μ là giá trị trung bình của phân bố. Phân bố này có thể được xem như là giới hạn của phân bố nhị thức khi $p \to 0$ và $n \to \infty$ nhưng giá trị trung bình $\mu = np$ tiến đến một giá trị hữu hạn. Phương sai của phân bố Poisson là

$$\sigma^2 = \mu. \tag{1.34}$$

Hàm xác suất của phân bố Poisson được minh họa trong Hình 1.7.

Ví dụ 1.29. *Biết xác suất một sản phẩm bị lỗi xuất hiện trong quá trình sản xuất là $p = 0.01$. Hỏi xác suất số sản phẩm bị lỗi xuất hiện không nhiều hơn 2 trong 100 sản phẩm được tạo ra là bao nhiêu?*

Giải: *Gọi X là BNN nhận giá trị là số sản phẩm bị lỗi trong quá trình sản xuất. Từ phân bố nhị thức, chúng ta tìm được trung bình $\mu = np = 1$.*



Hình 1.7: Hàm xác suất (1.33) của phân bố Poisson với μ nhận một số giá trị khác nhau

Bởi p rất nhỏ và n lớn, một cách gần đúng chúng ta có có thể sử dụng phân bố Poisson để tính như sau.

$$P[X \leq 2] \approx f(0) + f(1) + f(2) = (1 + 1 + \frac{1}{2})e^{-1} = 91.97\%.$$

Từ phân bố nhị thức, chúng ta nhận được xác suất của sự kiện trên là $P[X \leq 2] = 92.06\%$. *Vì vậy, phân bố Poisson là một xấp xỉ khá tốt của phân bố nhị thức khi p nhỏ và n lớn.*

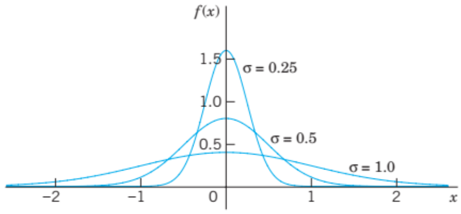
1.9.3 Phân bố chuẩn $X \sim \mathcal{N}(\mu, \sigma^2)$

Phân bố chuẩn (normal distribution) hay *phân bố Gauss* (Gauss distribution) là phân bố liên tục. Đây là phân bố liên tục quan trọng nhất vì trong nhiều ứng dụng, các biến ngẫu nhiên hoặc có phân bố chuẩn hoặc được xấp xỉ bởi phân phối chuẩn hoặc có thể biến đổi về phân phối chuẩn. Hàm mật độ xác suất của phân bố chuẩn được cho như sau

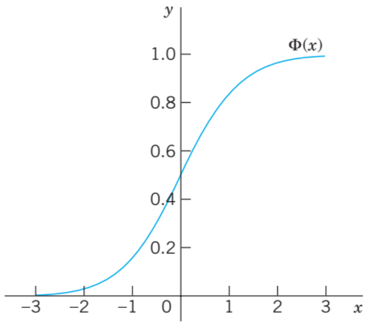
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{1.35}$$

với $\mu \in \mathbb{R}$ và $\sigma > 0$. Hàm xác suất của phân bố chuẩn được minh họa bằng đồ thị trong hình 1.8. Dưới đây là một số tính chất của phân phối chuẩn.

1. μ là trung bình và σ là độ lệch chuẩn.
2. Hệ số $\frac{1}{\sigma\sqrt{2\pi}}$ đảm bảo tích phân $\int_{-\infty}^{+\infty} f(x)dx = 1$ hay tổng xác suất bằng 1.



Hình 1.8: Hàm xác suất của phân bố chuẩn trong trường hợp $\mu = 0$ với các giá trị khác nhau của σ.



Hình 1.9: Hàm phân bố chuẩn tắc $\Phi(z)$ với trung bình $\mu = 0$ và phương sai $\sigma = 1$.

3. Đồ thị của hàm $f(x)$ đối xứng qua đường thẳng $x = \mu$.

Hàm phân bố tương ứng được cho như sau

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2} dv. \tag{1.36}$$

Trong trường hợp trung bình $\mu = 0$ và độ lệch chuẩn $\sigma = 1$, phân bố chuẩn được gọi là *phân bố chuẩn tắc*, được ký hiệu là $\Phi(z)$ và được cho như sau

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du. \tag{1.37}$$

Hàm phân bố (1.37) được minh họa Hình 1.9. Dưới đây là một số tính chất quan trọng của hàm Φ.

(i) $\Phi(+\infty) = 1$ hay tương đương với $\int_{-\infty}^{\infty} e^{-u^2/2} du = \sqrt{2\pi}$.

(ii) Với mọi $z \in \mathbb{R}$,
$$\Phi(-z) = 1 - \Phi(z).$$

Mối liên hệ giữa phân bố chuẩn và phân bố chuẩn tắc được thể hiện thông qua định lí sau.

Định lý 1.9. *Giả sử $F(x)$ là hàm phân bố của một phân bố chuẩn với trung bình μ và độ lệch chuẩn σ. Khi đó*

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{1.38}$$

Dưới đây là một hệ quả của Định lí 1.9.

Định lý 1.10 (Xác suất chuẩn cho các khoảng)**.** *X là BNN phân bố chuẩn với trung bình μ và độ lệch chuẩn σ. Khi đó*

$$P(a < X \leq b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{1.39}$$

Ví dụ 1.30. *Cho BNN X có phân bố chuẩn với trung bình $\mu = 5$ và phương sai $\sigma^2 = 0.04$. Tìm c và k sao cho*

(a) $P(X \leq c) = 95\%$.

(b) $P(5-k \leq X \leq 5+k) = 90\%$.

(c) $P(X \geq c) = 1\%$.

Giải. (a) Ta có $\sigma = 0.2$ và

$$P(X \leq c) = 95\% \Leftrightarrow \Phi\left(\frac{c-5}{0.2}\right) = 95\% \Leftrightarrow \frac{c-5}{0.2} = 1.654 \quad (\textit{Xem bảng A8}).$$

Do đó $c = 5.329$.

(b) Áp dụng Định lý 1.10, ta có

$$\begin{aligned} P(5-k \leq X \leq 5+k) &= \Phi\left(\frac{(5+k)-5}{0.2}\right) - \Phi\left(\frac{(5-k)-5}{0.2}\right) \\ &= \Phi(k/0.2) - \Phi(-k/0.2) = 2\Phi(k/0.2) - 1. \end{aligned}$$

Theo giả thiết, ta có $\Phi(k/0.2) = 95\%$. Từ bảng A8, ta nhận được $k = 1.645 \times 0.2 = 0.329$.

(c) Ta có

$$P(X \leq c) = 1 - P(X > c) = 1 - P(X \geq c) = 99\%.$$

Tương tự như câu (a), ta có $(c-5)/0.2 = 2.326$ (xem bảng A8). Vậy, $c = 5.465$.

Ví dụ 1.31. *Trong sản xuất các thanh sắt hình trụ, đường kính X của thanh sắt có phân bố chuẩn với trung bình là 2 in và độ lệch chuẩn là 0.008 in.*

(a) Hỏi xác suất để thanh sắt được sản xuất ra là phế phẩm là bao nhiêu nếu chúng ta cho phép sai số đường kính như sau 2 ± 0.02 in?

(b) Tìm dung sai c sao cho tỷ lệ phế phẩm là 4%?

Giải*: Ta có $\mu = 2$ (in) và $\sigma = 0.008$ (in).*

(a) Gọi A là sự kiện sản xuất các thanh sắt là phế phẩm với sai số đường kính được phép là 2 ± 0.02. Xác suất của sự kiện đối lập A^c là

$$P(A^c) = P(1.98 \leq X \leq 2.02) = \Phi\left(\frac{2.02-2}{0.008}\right) - \Phi\left(\frac{1.98-2}{0.008}\right) = 98\frac{3}{4}\%$$

(xem bảng Table A7.) Vì vậy, xác suất của sự kiện A là $P(A) = 1\frac{1}{4}\%$.

(b) Ta có

$$\begin{aligned} &P(2-c \leq X \leq 2+c) = 1 - 4\% = 96\% \\ \Leftrightarrow \quad &\Phi\left(\frac{(2+c)-2}{0.008}\right) - \Phi\left(\frac{(2-c)-2}{0.008}\right) = 96\%. \end{aligned}$$

Hay $\Phi(c/0.008) = 0.98$ và do đó $c = 2.054 \times 0.008 = 0.0164$ (xem bảng A8).

1.9.4 Xấp xỉ phân bố chuẩn của phân bố nhị thức

Khi n lớn, phân bố nhị thức $B(n, p)$ có thể được xấp xỉ bởi phân bố chuẩn, ở đây hàm xác suất $f(x)$ của phân bố nhị thức được xấp xỉ như sau

$$f(x) = \binom{n}{x} p^x (1-p)^{1-x} \simeq \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(x-np)^2}{2np(1-p)}}. \tag{1.40}$$

Vế phải của phương trình (1.40) là hàm mật độ xác suất của phân bố chuẩn với trung bình $\mu = np$ và phương sai $\sigma^2 = np(1-p)$. Hơn nữa, cho các số nguyên không âm a và b với $a < b$, ta có xác suất $P(a \leq X \leq b)$ được tính như sau

$$P(a \leq X \leq b) = \sum_{x=a}^{b} \binom{n}{x} p^x (1-p)^{n-x} \approx \Phi(\beta) - \Phi(\alpha), \tag{1.41}$$

ở đây

$$\alpha = \frac{a - \mu - 0.5}{\sigma}, \qquad \beta = \frac{b - \mu + 0.5}{\sigma}.$$

Trong thực tế, người ta thường dùng công thức xấp xỉ (1.41) khi

$$\left[\begin{array}{l} np > 5 \quad \text{và} \quad n(1-p) > 5, \\ np(1-p) > 20. \end{array}\right.$$

Ví dụ 1.32. *Xác suất mỗi cây sống sau thời gian trồng T là $p = 0.8$. Trồng 400 cây. Tìm xác suất để có từ khoảng 300 đến 360 cây sống sau thời gian trồng T.*

Giải. *Ta có $n = 400$, $p = 0.8$ và $X \equiv$"Số cây sống sau một thời gian trồng T" là biến ngẫu nhiên có phân bố nhị thức. Vì $np = 320 > 5$ và $n(1-p) = 80 > 5$, nên theo công thức (1.41) ta có*

$$P(300 \leq X \leq 360) \approx \Phi(\beta) - \Phi(\alpha)$$

với

$$\alpha = \frac{a - np - 0.5}{\sqrt{np(1-p)}} = \frac{300 - 320 - 0.5}{\sqrt{400 \times 0.8 \times (1-0.8)}} = -2.5625,$$
$$\beta = \frac{b - np + 0.5}{\sqrt{np(1-p)}} = \frac{360 - 320 + 0.5}{\sqrt{400 \times 0.8 \times (1-0.8)}} = 7.5625.$$

Từ bảng A7, ta có $\Phi(\beta) \approx 1$ và

$$\Phi(\alpha) = \Phi(-2.5625) = 1 - \Phi(2.5625) \approx 1 - \Phi(2.56) = 1 - 0.9948 = 0.0052.$$

Vậy xác suất để có từ khoảng 300 đến 360 cây sống sau thời gian trồng T là

$$P(300 \leq X \leq 360) \approx 1 - 0.0052 = 99.48\%.$$

Bài tập Chương 1

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

Bài 1.1. Xác định không gian mẫu (không gian sự kiện cơ sở) trong thực nghiệm gieo hai con súc sắc.

Bài 1.2. Xác định các sự kiện sau trong **Bài 1.1.**:

(a) Các mặt bằng nhau;

(b) Tổng số chấm của hai con súc sắc lớn hơn 9;

(c) Tổng chấm của hai con súc sắc bằng 7.

Bài 1.3. Xác định không gian mẫu khi tung một đồng xu cho đến khi mặt ngửa xuất hiện.

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

Bài 1.4. Một hộp có 100 con ốc vít, trong đó 10 con ốc bị lỗi (phế phẩm). Rút (không hoàn trả) ngẫu nhiên 3 con ốc vít từ hộp này, hỏi xác suất để có ít nhất một con ốc vít bị lỗi là bao nhiêu?

Bài 1.5. Gieo đồng thời bốn đồng xu đồng nhất, tìm các xác suất sau:

(a) Cả bốn mặt giống nhau xuất hiện;

(b) Có đúng 2 mặt sấp;

(c) Có ít nhất hai mặt ngửa.

Bài 1.6. Cho một thí nghiệm gieo 2 con súc sắc đồng nhất. Tìm các xác suất:

(a) để tổng số chấm của hai con súc sắc lớn hơn 4 và nhỏ hơn hoặc bằng 7;

(b) để số chấm bằng nhau hoặc có tổng là một số chẵn.

Bài 1.7. Một lớp học có 30 học sinh trong đó có 4 giỏi, 8 khá, 10 trung bình, và số còn lại là yếu. Chọn ngẫu nhiên ra 3 người, tìm các xác suất để:

(a) cả ba đều là học sinh yếu;

(b) có ít nhất một học sinh giỏi;

(c) có đúng một học sinh giỏi.

Xác suất có điều kiện, công thức nhân xác suất, và sự kiện độc lập:

Bài 1.8. Gieo hai con súc sắc giống nhau. Biết rằng tổng số chấm xuất hiện trên hai con súc sắc là một số chẵn, tìm xác suất để tổng đó bằng 6.

Bài 1.9. Cho một mạch điện với 3 thiết bị điều khiển đóng ngắt mắc nối tiếp với nhau. Xác suất để mạch điện làm việc trong một khoảng thời gian T là 95%. Hỏi xác suất bị lỗi của mỗi thiết bị điều khiển đóng ngắt trong khoảng thời gian T là bao nhiêu? Giả sử 3 thiết bị điều khiển đóng ngắt là đồng nhất.

Bài 1.10. Một thiết bị điều khiển áp suất chứa 4 van đồng nhất. Thiết bị này sẽ không làm việc nếu tất cả các van không hoạt động. Xác suất bị lỗi của mỗi van trong một khoảng thời gian đã cho là 3%. Hỏi xác suất để thiết bị điều khiển làm việc là bao nhiêu?

Bài 1.11. Một hộp chứa 100 đinh ốc, trong đó có 10 đinh ốc bị lỗi. Lấy ra 3 đinh ốc. Tìm xác suất để 3 đinh ốc bị đều lỗi trong hai trường hợp sau:

(a) đưa đinh ốc trở lại hộp sau mỗi lần ra;

(b) không đưa đinh ốc trở lại hộp sau mỗi lần ra.

Công thức xác suất toàn phần, công thức Bayes:

Bài 1.12. Theo một số liệu thống kê, năm 2004 ở Canada có 65.0% đàn ông thừa cân và 53.4% đàn bà thừa cân. Giả sử rằng số đàn ông và đàn bà ở Canada là bằng nhau. Hỏi rằng, trong năm 2004, xác suất để một người Canada được chọn ngẫu nhiên là người thừa cân là bao nhiêu?

Bài 1.13. Được biết có 5% đàn ông bị mù màu và 0.25% đàn bà bị mù màu. Giả sử số đàn ông bằng số đàn bà. Chọn ngẫu nhiên một người bị mù màu. Hỏi rằng xác suất để người đó là đàn ông là bao nhiêu?

Bài 1.14. Có hai hộp áo, hộp I có 10 chiếc áo trong đó có một chiếc áo bị lỗi, hộp II có 8 chiếc áo trong đó có hai chiếc áo bị lỗi. Lấy ngẫu nhiên 1 chiếc áo từ hộp I bỏ sang hộp II, sau đó từ hộp này chọn ngẫu nhiên ra 2 chiếc áo. Tìm xác suất để cả hai chiếc áo đó đều bị lỗi.

Bài 1.15. Tại một phòng khám chuyên khoa, tỷ lệ người đến khám có bệnh là 83%. Theo thống kê biết rằng nếu chẩn đoán có bệnh thì đúng tới 90%, còn nếu chẩn đoán không bệnh thì đúng 80%.

(a) Tính xác suất chẩn đoán đúng.

(b) Biết có một trường hợp chẩn đoán đúng, tìm xác suất người được chẩn đoán đúng có bệnh.

Bài 1.16. Một nhà máy ô tô A có ba phân xưởng I, II và III cùng sản xuất một loại pít-tông. Phân xưởng I, II, III sản xuất tương ứng 40%, 35%, 25% sản lượng của nhà máy, với tỉ lệ phế phẩm tương ứng là 5%, 8%, 4%. Chọn ngẫu nhiên một pít-tông do nhà máy A sản xuất.

(a) Tìm xác suất để pít-tông được chọn là phế phẩm.

(b) Biết pít-tông được chọn là phế phẩm. Tìm xác suất để sản phẩm đó do phân xưởng II sản xuất.

Biến ngẫu nhiên và phân bố của biến ngẫu nhiên:

Bài 1.17. Một biến ngẫu nhiên X có hàm xác suất $f(x) = \frac{k}{2^x}$ (với $x = 0$, 1, 2, 3, ...). Hỏi giá trị của tham số k và $P(X \geq 4)$?

Bài 1.18. Cho hàm xác suất $f(x) = kx^2$ (k phù hợp) nếu $0 \leq x \leq 5$ và $f(x) = 0$ trong miền còn lại của x ($x < 0$ và $x > 5$). Xác định hàm phân bố $F(x)$.

Bài 1.19. Giả sử độ dài L của các bu-lông ốc vít được mô tả bởi hàm sau $L = 200 + X$ (mm) ở đây X là một biến số ngẫu nhiên với hàm xác suất $f(x) = \frac{3}{4}(1 - x^2)$ nếu $-1 \leq x \leq 1$ và $f(x) = 0$ trong miền còn lại của x. Xác định c sao cho 95% bu-lông ốc vít có chiều dài nằm trong khoảng $200 - c$ và $200 + c$.

Bài 1.20. Một biến ngẫu nhiên có hàm phân bố $F(x) = 1 - e^{-3x}$ nếu $x > 0$ và $F(x) = 0$ nếu $x \leq 0$. Tìm hàm xác suất $f(x)$ và tìm x sao cho $F(x) = 0.9$.

Bài 1.21. Gọi X là tỷ số giữa doanh thu và lợi nhuận của một số công ty. Giả sử X có hàm phân bố

$$F(x) = \begin{cases} 0 & \text{nếu } x < 2, \\ (x^2 - 4)/5 & \text{nếu } 2 \leq x < 3, \\ 1 & \text{nếu } x \geq 3. \end{cases}$$

Tìm và vẽ đồ thị của hàm xác suất $f(x)$. Tìm xác suất để X nhận giá trị từ 2.5 (40% lợi nhuận) và 5 (20% lợi nhuận)?

Các số đặc trưng của biến ngẫu nhiên:

Bài 1.22. Tìm trung bình và phương sai của biến ngẫu nhiên X với hàm xác suất $f(x)$ được cho trong các trường hợp sau:

(a) X là số điểm xuất hiện khi gieo một con súc sắc đồng nhất.

(b) $f(0) = 0.512$, $f(1) = 0.384$, $f(2) = 0.096$, $f(3) = 0.008$.

(c) $f(x) = 2x$ nếu $0 \leq x \leq 1$ và $f(x) = 0$ trong miền còn lại.

Bài 1.23. Một trạm xăng nhỏ được cung cấp đầy xăng vào mỗi chiều thứ bảy. Thể tích xăng bán ra được mô tả bởi một biến ngẫu nhiên X và được cho theo đơn vị 10000 galon. Hàm xác suất của X được cho bởi $f(x) = 6x(1-x)$ nếu $0 \leq x \leq 1$ và $f(x) = 0$ trong miền còn lại. Xác định:

(a) trung bình, phương sai và độ lệch chuẩn của X;

(b) thể tích (chứa xăng) của trạm xăng trên là bao nhiểu để xác suất bán hết xăng trong một tuần là 95%.

Một số phân bố thường gặp:

Bài 1.24. Bốn đồng xu đồng nhất được tung đồng thời. Gọi X là biến ngẫu nhiên xác định số mặt ngửa của thực nghiêm trên.

(a) Xác định hàm xác suất của biến ngẫu nhiên X.

(b) Tìm xác suất để nhận được ít nhất một mặt ngửa.

Bài 1.25. Giả sử 4% các thanh thép được sản xuất bởi một nhà máy là bị lỗi, biết rằng các lỗi xảy ra ngẫu nhiên trong quá trình sản xuất. Người ta đóng gói các thanh sắt thành các bó, mỗi bó gồm 100 thanh sắt. Lấy ngẫu nhiên một bó. Bằng cách sử dụng tính chất xấp xỉ phân bố Poisson của phân bố nhị thức, tính gần đúng xác suất để bó sắt được chọn có chứa $x = 0, 1, 2, \ldots, 5$ thanh thép lỗi?

Bài 1.26. Gọi X là số xe oto đi qua điểm A trong một phút trong thời gian từ 8 đến 10 giờ sáng ngày Chủ nhật. Giả sử X có phân bố Poisson với trung bình 5. Tính xác suất để quan sát thấy nhiều nhất 4 xe oto đi qua điểm A trong thời gian một phút?

Bài 1.27. Cho X là một biến ngẫu nhiên có phân bố chuẩn với kỳ vọng 10 và phương sai 4. Tìm các xác suất sau: (i) $P(X < 10)$; (ii) $P(X > 12)$; (iii) $P(9 < X < 13)$.

Bài 1.28. Cho X là biến ngẫu nhiên có phân bố chuẩn với kỳ vọng 50 và phương sai 9. Hãy xác định số c trong các trường hợp sau: (i) $P(X < c) = 5\%$; (ii) $P(X > c) = 1\%$; (iii) $P(-c < X - 50 < c) = 50\%$.

Bài 1.29. Tuổi thọ của một loại ắc-qui ôtô có phân bố chuẩn với tuổi thọ trung bình là bốn năm và độ lệch chuẩn là một năm. Nhà sản xuất mong muốn đảm bảo pin làm việc không dưới bốn năm. Để đảm bảo yêu cầu này, hỏi bao nhiêu phần trăm pin cần sản xuất để thay thế?

Bài 1.30. Chi phí bảo trì và sữa chữa máy móc hàng tháng ở một nhà máy **A** là biến ngẫu nhiên có phân phối chuẩn với trung bình \$12000 và độ lệch chuẩn \$2000. Hỏi xác suất cho chi phí bảo trì và sữa chữa máy móc ở nhà máy này trong tháng tới vượt \$15000 là bao nhiêu?

Bài 1.31. Biết điện trở của các dây điện loại **B** là biến ngẫu nhiên có phân bố chuẩn với kỳ vọng 0.01 Ω và độ lệch chuẩn 0.001 Ω. Hỏi trong 1000 dây điện loại **B** có trung bình bao nhiêu dây có điện trở nằm giữa 0.009 Ω và 0.011 Ω?

Bài 1.32. Điện trở do một nhà máy sản xuất có kháng trở là một biến ngẫu nhiên có phân bố chuẩn với kỳ vọng $\mu = 150\Omega$ và độ lệch chuẩn $\sigma = 5\Omega$. Tính xác suất để một điện trở có:

(a) kháng trở từ 148Ω đến 152Ω;

(b) kháng trở từ 140Ω đến 160Ω.

Chương 2

Ước lượng tham số

Từ chương này chúng ta bắt đầu nghiên cứu thống kê (toán học). Thống kê có thể coi là tổng thể các phương pháp toán học, dựa trên lý thuyết xác suất và các công cụ khác, nhằm đưa ra được những thông tin mới, kết luận mới, có giá trị từ những bảng số liệu thô ban đầu nhằm giải quyết những vấn đề nào đó nảy sinh từ thực tế. Có thể kể một số mục đính chính của thống kê như sau: mô tả số liệu; ước lượng và dự đoán các đại lượng; tìm ra các mối quan hệ giữa các đại lượng; kiểm định các giả thuyết. Các phương pháp quan trọng của thống kê là ước lượng điểm của các tham số đặc trưng, xác định khoảng tin cậy và kiểm định các giả thuyết.

2.1 Mẫu ngẫu nhiên và các đặc trưng mẫu

2.1.1 Mẫu ngẫu nhiên

Trong công việc hằng ngày ta phải làm việc với các dãy số liệu. Chúng có thể là kết quả của việc đếm khi quan sát, đo đạc thông qua các thiết bị đo,... và cần được thu nhập, lưu trữ và phân tích. Để làm được điều đó ta cần sắp xếp lại các số, tổng hợp và xử lý bước đầu nhằm tìm kiếm các thông tin quan trọng của tập số liệu. Phần công việc này và vấn đề thu thập số liệu được gọi là thống kê mô tả.

Dãy số liệu thống kê thường được gọi là *mẫu*. Nó có nguồn gốc từ một tập lớn hơn mà ta sẽ gọi là *tập nền* (hay còn gọi là *tập hợp chính*). Chính vì thế mẫu sẽ mang thông tin nào đó từ tập nền, mặc dù các thông tin đó có thể khác nhau ở những mẫu khác nhau.

Muốn có đầy đủ thông tin về đối tượng nào đó ta phải làm việc với tập nền. Tuy nhiên việc nghiên cứu tập nền sẽ vô cùng khó khăn vì:

- do tập nền quá lớn dẫn đến đòi hỏi quá nhiều chi phí về vật chất và thời gian;

- do trình độ tổ chức và nghiên cứu hạn chế khi làm việc với quy mô lớn, không nắm bắt và kiểm soát được quá trình nghiên cứu;

- do nhiều khi không thể làm được nếu tập nền biến động nhanh, các phần tử thay đổi thường xuyên,...

Như vậy việc nghiên cứu trên tập nền, trừ các tập đủ bé, thường không thể thực hiện được. Từ đó đặt ra vấn đề chọn mẫu và nghiên cứu trên tập mẫu. Nếu mẫu được chọn ngẫu nhiên và với số lượng đủ, chúng ta hy vọng rằng việc xử lý chúng sẽ cho ta kết quả vừa nhanh vừa đỡ tốn kém mà vẫn đạt được độ chính xác và tin cậy cần thiết.

Một mẫu ngẫu nhiên kích thước n được lấy từ tập nền có biến ngẫu nhiên X là một tập giá trị $(x_1, x_2, \cdots, x_n)$, ở đây các số x_i được gọi là các *giá trị thực nghiệm* của X.

Ví dụ 2.1. *Một nhà sản xuất dưa chuột muối đóng hộp muốn biết phân bố chiều dài các quả dưa chuột (chiều dài trung bình, độ lệch chuẩn,...), để làm vỏ hộp với kích thước phù hợp. Nhà sản xuất này sẽ không đi đo hết chiều dài của hàng triệu quả dưa chuột (tập nền) sẽ được đóng hộp. Họ chỉ đo chiều dài của một số n quả dưa chuột (tập mẫu) được chọn ngẫu nhiên, rồi từ đó ước lượng ra phân bố chiều dài. Số n ở đây có thể sẽ là một con số khá lớn, chẳng hạn 100 quả hay 1000 quả, nhưng nó là một số rất nhỏ của tổng số các quả dưa chuột.*

Mẫu thực nghiệm $(x_1, x_2, \cdots, x_n)$ cho ta một phân bố xác suất $\hat{P}_n$, gọi là phân bố xác suất thực nghiệm, như sau: nó là phân bố xác suất rời rạc tập trung tại các điểm x_1, $x_2, \cdots$, x_n, sao cho mỗi điểm x_i có xác suất là $1/n$.

2.1.2 Biểu diễn dữ liệu

Biểu diễn dữ liệu dạng số và đồ thị

Dữ liệu được tập hợp thông qua quan sát hoặc đo lường. Dữ liệu có thể được biểu diễn bằng dạng số hoặc đồ thị theo nhiều cách khác nhau. Ví dụ, các báo hằng ngày có thể chứa các bảng giá cổ phiếu, tỉ giá hối đoái; các đường cong hay biểu đồ tần suất thể hiện sự phát triển kinh tế của một đất nước nào đó. Trong mục này, chúng ta sẽ tìm hiểu một số biểu diễn dữ liệu điển hình trong thống kê.

Chúng ta đưa ra hai cách biểu diễn dữ liệu điển hình (*Stem-and-Leaf* và *Histogram*) thông qua ví dụ sau.

Key: $7|8 = 78$

CAF	Stem	Leaf
1	7	8
4	8	1 3 4
11	8	6 7 7 9 9 9 9
13	9	0 1
14	9	9

Hình 2.1: Biểu diễn dữ liệu dạng **Stem-and-Leaf**, CAF = Tần suất tuyệt đối tích lũy

Ví dụ 2.2. *Đo lường sức căng của các tấm thép, người ta thu được dữ liệu chứa đựng 14 giá trị mẫu theo đơn vị kg/mm^2 như sau*

$$89 \quad 84 \quad 87 \quad 81 \quad 89 \quad 86 \quad 91 \quad 90 \quad 78 \quad 89 \quad 87 \quad 99 \quad 83 \quad 89. \tag{2.1}$$

Đầu tiên chúng ta hãy sắp xếp dữ liệu trên như sau

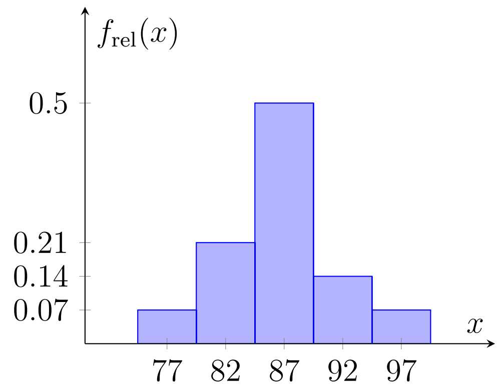
$$78 \quad 81 \quad 83 \quad 84 \quad 86 \quad 87 \quad 87 \quad 89 \quad 89 \quad 89 \quad 89 \quad 90 \quad 91 \quad 99. \tag{2.2}$$

- **Biểu diễn dữ liệu dạng *Stem-and-Leaf***: Đây là cách biểu diễn dữ liệu đơn giản nhất, ở đó dữ liệu được tách thành một “leaf" (con số cuối hoặc các con số sau dấu phẩy cho trường hợp số thập phân) và một “stem" (các con số còn lại hoặc các con số trước dấu phẩy cho trường hợp số thập phân). Từ (2.2), chúng ta nhìn thấy rằng các số trong dữ liệu ban đầu (2.1) có giá trị từ 78 đến 99. Chúng ta chia các số này thành năm nhóm: $75-79$, $80-84$, $85-89$, $90-94$, $95-99$. Các chữ số trong hàng chục của các nhóm này tương ứng là: 7, 8, 8, 9, 9. Chúng hình thành *stem* (xem hình 2.1). *Leaf* đầu tiên là 8, thể hiện 78. *Leaf* thứ hai là 134, thể hiện 81, 83, 84. Tương tự cho các *Leaf* tiếp theo.

  Số lần một giá trị xuất hiện trong một tập các điểm mẫu được gọi là *tần suất tuyệt đối* (absolute frequency). Giá trị 78 có tần suất là 1, giá trị 89 có tần suất là 4. Cột tận cùng bên trái của đồ thị **Stem-and-Leaf** thể hiện *tần suất tuyệt đối tích lũy* (cumulative absolute frequency), nó bằng tổng tất cả các tần suất tuyệt đối của các giá trị đến dòng *Leaf* tương ứng. Số 4 ở cột tận cùng bên trái của ô bên trái trong hình 2.1 thể hiện có bốn giá trị tính từ dòng *Leaf* đầu tiên đến dòng *Leaf* thứ hai. Trong khi đó, số 11 thể hiện có mười một giá trị tính từ dòng *Leaf* đầu tiên đến dòng *Leaf* thứ ba.

Lớp	$74.5 - 79.5$	$79.5 - 84.5$	$84.5 - 89.5$	$89.5 - 94.5$	$94.5 - 99.5$
$f_{\text{rel}}(x)$	0.07	0.21	0.5	0.14	0.07

Bảng 2.1: Tần suất lớp tương đối



Hình 2.2: Biểu diễn dữ liệu theo tần suất (Histogram)

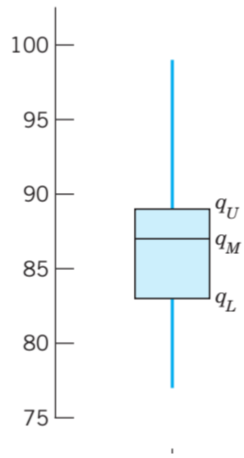
- **Biểu diễn dữ liệu dạng tần suất (Histogram)**: Trong trường hợp dữ liệu là một tập lớn, biểu diễn dữ liệu ở dạng tần suất sẽ hiện thị phân bố dữ liệu tốt hơn so với biểu diễn dữ liệu dạng **Stem-and-Leaf**. Chúng ta chia dữ liệu (2.2) thành các lớp (class interval) có độ rộng bằng nhau như sau: $74.5 - 79.5$, $79.5 - 84.5$, $84.5 - 89.5$, $89.5 - 94.5$, $94.5 - 99.5$. Điểm chính giữa của các lớp này lần lượt là: $x = 77, 82, 87, 92, 97$. *Tần suất lớp tương đối* (relative class frequency) $f_{\text{rel}}(x)$ của một lớp nhận x làm điểm chính giữa được tính như sau

$$f_{\text{rel}}(x) = \frac{\text{Số các phần tử của lớp có điểm chính giữa là } x}{\text{Tổng số phần tử của dữ liệu}}.$$

  Tần suất lớp tương đối của 5 khoảng trong dữ liệu (2.2) được cho bởi bảng 2.1 và được minh họa bằng biểu đồ *histogram* như trong hình 2.2.

- **Biểu đồ hộp (boxplot). Trung vị (median). Khoảng biến thiên tứ phân vị (Interquartile Range). Dữ liệu ngoại lai (Outlier)**.

  Biểu đồ hộp của một tập dữ liệu (đã được sắp thứ tự) $\{x_1, x_2, ..., x_n\}$ diễn tả 5 vị trí phân bố của dữ liệu đó gồm: *giá trị nhỏ nhất (min), tứ vị phân thứ nhất q_L (lower quartile), trung vị q_M (median, middle quartile), tứ vị phân thứ ba q_U (upper quartile)* và *giá trị lớn nhất (max)*.



Hình 2.3: Biểu đồ hộp của tập dữ liệu (2.2)

Trong đó, q_M, q_L, q_U được tính như sau:

$$q_M = \begin{cases} x_{(n+1)/2} & \text{nếu } n \text{ lẻ} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{nếu } n \text{ chẵn} \end{cases}$$

$$= \begin{cases} \text{giá trị của phần tử nằm chính giữa bộ dữ liệu} & \text{nếu } n \text{ lẻ} \\ \text{trung bình cộng của hai phần tử nằm chính giữa} & \text{nếu } n \text{ chẵn} \end{cases}$$

$q_L =$ **trung vị** của các phần tử của bộ dữ liệu **nằm dưới** giá trị q_M,

$q_U =$ **trung vị** của các phần tử của bộ dữ liệu **nằm trên** giá trị q_M.

Độ rộng của tập dữ liệu là $R = x_{\max} - x_{\min}$. Tuy nhiên, một thông tin quan trọng hơn về độ rộng của bộ dữ liệu là *khoảng biến thiên tứ phân vị (Interquartile Range)*: $IQR = q_U - q_L$.

Trong bộ dữ liệu (2.2), ta có

$$q_M = \frac{x_7 + x_8}{2} = \frac{87 + 98}{2} = 88, \quad q_U = x_{11} = 98, \quad q_L = x_4 = 84,$$
$$IQR = 5, \quad R = 99 - 78 = 21$$

Biểu đồ hộp của tập dữ liệu (2.2) được minh họa trong Hình 2.3.

Như trong Hình 2.3 ta thấy đường thẳng phía trên hộp khá dài, do đó trong bộ dữ liệu (2.2) có thể chứa những *dữ liệu ngoại lai (outlier)*. Một dữ liệu được gọi là *dữ liệu ngoại lai* nếu khoảng cách từ nó tới biên gần nhất của hộp lớn hơn 1.5 lần độ dài khoảng biến thiên tứ phân vị. Trong dữ liệu (2.2), ta có

$$q_U + 1.5IQR = 89 + 1.5 \times 5 = 96.5 < 99 = x_{14}.$$

Vậy $x_{14} = 99$ là một dữ liệu ngoại lai.

2.1.3 Trung bình, phương sai, và độ lệch chuẩn (của mẫu)

Định nghĩa 2.1. *Trung bình* (mean) $\bar{x}$ đo kích cỡ trung bình của các giá trị dữ liệu và được cho như sau

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n). \tag{2.3}$$

Ví dụ 2.2 (*tiếp*). Giá trị trung bình tương ứng với dữ liệu (2.1) là

$$\bar{x} = \frac{1}{14}(89 + 84 + \cdots + 89) = \frac{611}{7} \approx 87.3.$$

Biến thiên của dữ liệu được đo bởi *phương sai* s^2 hoặc *độ lệch chuẩn* s (căn bậc hai của s^2) với

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (2.4)\\ &= \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]. \end{aligned}$$

Lưu ý rằng, độ lệch chuẩn có cùng đơn vị với các giá trị dữ liệu.

Từ phương trình (2.4) ta có

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i)^2 - 2\bar{x}x_i + \bar{x}^2\right] \\ &= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2\right] \\ &= \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right), \qquad (2.5) \end{aligned}$$

ở đây $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2$. Dòng cuối của phương trình (2.5) cho ta một công thức tính nhanh phương sai s^2.

Ví dụ 2.2 (*tiếp*). Phương sai tương ứng với dữ liệu (2.1) là

$$\begin{aligned} s^2 &= \frac{1}{13}\left[\left(89 - \frac{611}{7}\right)^2 + \left(84 - \frac{611}{7}\right)^2 + \cdots + \left(89 - \frac{611}{7}\right)^2\right] \\ &= \frac{176}{7} \approx 25.14. \end{aligned}$$

Độ lệch chuẩn tương ứng là $s = \sqrt{176/7} \approx 5.014$.

Ví dụ 2.3. *Tính trung bình mẫu và phương sai mẫu của một mẫu gồm chiều cao của 400 cây con được cho bởi trong bảng phân bố ghép lớp sau*

Chiều cao (cm)	*Tần số r_i*	*Chiều cao (cm)*	*Tần số r_i*
$[4.5, 9.5]$	*18*	$(16.5, 19.5]$	*57*
$(9.5, 11.5]$	*58*	$(19.5, 22.5]$	*42*
$(11.5, 13.5]$	*62*	$(22.5, 26.5]$	*36*
$(13.5, 16.5]$	*72*	$(26.5, 29.5]$	*55*

Giải. Ta có kích thước mẫu $n = 400$ và có $k = 8$ lớp. Các điểm đại diện của các lớp lần lượt là

$$x_1 = 7, \quad x_2 = 10.5, \quad x_3 = 12.5, \quad x_4 = 15,$$
$$x_5 = 18, \quad x_6 = 21, \quad x_7 = 24.5, \quad x_8 = 28\,(cm)$$

và các tần số xuất hiện của các lớp là

$$r_1 = 18, \quad r_2 = 58, \quad r_3 = 62, \quad r_4 = 72,$$
$$r_5 = 57, \quad r_6 = 42, \quad r_7 = 36, \quad r_8 = 55.$$

Dễ thấy $\sum_{i=1}^{k} r_i = 400 = n$. Trung bình của mẫu là

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i r_i = 17.3\,(cm).$$

và phương sai của mẫu là

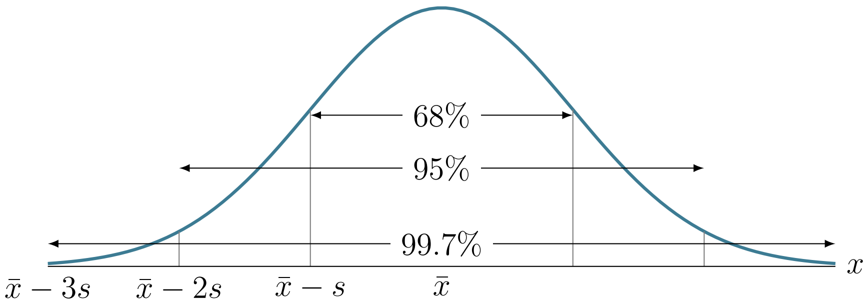
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} r_i(x_i - \bar{x})^2 = 38.01\,(cm^2).$$

Nhận xét 2.1 (Quy tắc đã kiểm chứng (Empirical rule))**.** *Qui tắc đã kiểm chứng chỉ ra rằng: Đối với dữ liệu có hình "quả chuông" (xem hình 2.4), thì*

(i) 68% *các điểm của dữ liệu nằm trong khoảng* $(\bar{x} - s, \bar{x} + s)$*;*

(ii) 95% *các điểm của dữ liệu nằm trong khoảng* $(\bar{x} - 2s, \bar{x} + 2s)$*;*

(iii) 99.7% *các điểm của dữ liệu nằm trong khoảng* $(\bar{x} - 3s, \bar{x} + 3s)$*.*



Hình 2.4: Qui tắc đã kiểm chứng

2.2 Ước lượng điểm của các tham số

Một ước lượng điểm của một tham số đặc trưng θ (của tập nền) là một con số $\hat{\theta}$ (một điểm trên trục thực) được tính từ một tập mẫu $(x_1, x_2, \cdots, x_n)$ đã cho và được hiểu như là một xấp xỉ của một giá trị chính xác chưa biết của tham số θ.

Tham số đặc trưng θ có thể một trong các đặc trưng sau của mẫu: trung bình, phương sai, trung vị (median), mode,... Cách đơn giản nhất để ước lượng đó là sử dụng các đặc trưng mẫu.

2.2.1 Ước lượng giá trị của kỳ vọng (trung bình)

Cho $\hat{\mu}$ là một ước lượng của trung bình (kỳ vọng) μ của một tập nền, thì $\hat{\mu}$ có thể được tính như sau

$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

2.2.2 Ước lượng giá trị của phương sai

Tương tự, một ước lượng phương sai $\hat{\sigma}^2$ cho phương sai σ^2 của một tập nền được tính như sau

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

2.2.3 Ước lượng giá trị của xác suất

Gọi p là xác suất tổng thể của một tập nền để xuất hiện sự kiện A. Cho biến ngẫu nhiên X (của tập nền) được xác định như sau

$$X = \begin{cases} 0 & \text{nếu không xảy ra sự kiện } A \text{ trong phép thử} \\ 1 & \text{nếu xảy ra sự kiện } A \text{ trong phép thử.} \end{cases}$$

Ta có

$$P(X = 0) = 1 - p \quad \text{và} \quad P(X = 1) = p.$$

Xét một mẫu $(x_1, x_2, ..., x_n)$ có kích thước n. Khi đó $(x_1 + x_2 + \cdots + x_n)$ bằng số lần xuất hiện sự kiện A. *Tần suất* xuất hiện sự kiện A trong mẫu là

$$f = \frac{(x_1 + x_2 + \cdots + x_n)}{n} = \bar{x}.$$

Do đó $f = \bar{x}$ là một ước lượng của xác suất tổng thể p.

2.2.4 Phương pháp hợp lý cực đại (Maximum likelihood method)

Một phương pháp khác để ước lượng tham số là *phương pháp hợp lý cực đại (Maximum likelihood method)*. Chúng ta xét một biến số ngẫu nhiên (rời rạc hoặc liên tục) X với hàm xác suất (hàm mật độ) $f(x;\theta)$ phụ thuộc vào tham số θ. Chúng ta lấy một mẫu (tương ứng với biến ngẫu nhiên này) có n giá trị độc lập như sau x_1, x_2, $\ldots$, x_n. Vì các phép lấy mẫu là độc lập (tương ứng với các biến ngẫu nhiên độc lập: $X_1, X_2, ..., X_n$), hàm xác suất đồng thời của các biến ngẫu nhiên $X_1, X_2, ..., X_n$ có dạng

$$l(x_1, ..., x_n; \theta) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta). \tag{2.6}$$

Hàm l trong công thức (2.6) được gọi là *hàm hợp lý cực* của tham số θ.

Ý tưởng cơ bản của phương pháp hợp lý cực đại là: *tìm ước lượng $\hat{\theta}$ của tham số θ sao cho xác suất l của mẫu quan sát là lớn nhất khi $\theta = \hat{\theta}$.*

Điều kiện cần vì vậy được xác định như sau

$$\frac{\partial l}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0 \quad \Leftrightarrow \quad \frac{\partial \ln l}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0. \tag{2.7}$$

Nghiệm $\hat{\theta}$ của phương trình (2.7) được gọi là *ước lượng hợp lý cực đại* của tham số θ.

Ví dụ 2.4. *Tìm ước lượng hợp lý cực đại của xác suất (tần suất) tổng thể p, biết hàm xác suất có dạng $f(x,p) = p^x(1-p)^{1-x}$ với $x \in \{0,1\}$.*

Giải: *Hàm hợp lý của xác suất p có dạng*

$$\begin{aligned} l(p) &= f_1(x_1,p)f_2(x_1,p)\cdots f_n(x_1,p) \\ &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n}. \end{aligned}$$

Do đó

$$\ln l(p) = \sum_{i=1}^{n}(x_i \ln p + (1-x_i)\ln(1-p))$$

và

$$\frac{\partial \ln l(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-x_i).$$

Giải phương trình

$$\frac{\partial \ln l(p)}{\partial p} = 0$$

ta được nghiệm là $p = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}$.

Vậy $\hat{p} = \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ là ước lượng hợp lý cực đại của p.

Ví dụ 2.5. *Sử dụng phương pháp hợp lý cực đại để ước lượng kỳ vọng và độ lệch chuẩn của phân bố chuẩn từ một tập mẫu có kích thước n, trung bình mẫu $\bar{x}$ và phương sai mẫu s^2.*

Giải: *Ta có hàm l trong trường hợp này như sau*

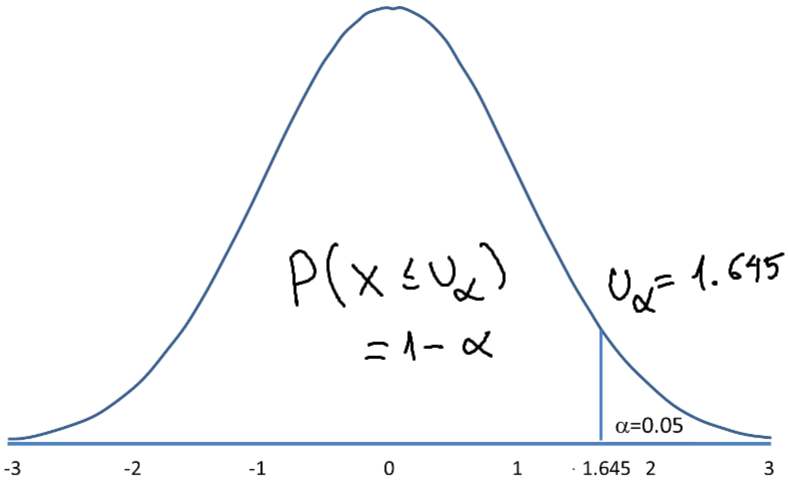
$$l = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-h}, \qquad \text{với} \qquad h = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Sự triệt tiêu của các đạo hàm theo $\theta_1 = \mu$ và $\theta_2 = \sigma$ được cho như sau:

$$\begin{aligned} \frac{\partial \ln l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0, \\ \frac{\partial \ln l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0. \end{aligned}$$

Giải các phương trình này, ta thu được các ước lượng hợp lý cực đại cho μ và σ như sau

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = s\sqrt{\frac{n-1}{n}}.$$



Hình 2.5: Giá trị tới hạn U_α mức $\alpha = 5\%$

2.3 Ước lượng khoảng của tham số

Ước lượng điểm có một nhược điểm cơ bản là khi mẫu có kích thước nhỏ thì ước lượng điểm có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng. Hơn nữa phương pháp ước lượng điểm cũng không thể biết được độ chính xác của các giá trị ước lượng $\hat{\theta}$ so với giá trị của đại lượng cần ước lượng θ. Để khắc phục các hạn chế đó, người ta đưa ra khái niệm ước lượng bằng một khoảng giá trị. Tất nhiên một khoảng ước lượng vẫn có thể sai giống như mọi ước lượng khác. Nhưng khác với ước lượng điểm, xác suất sai lầm có thể biết và trong chừng mực nào đó có thể hy vọng kiểm soát được.

Khoảng tin cậy cho một tham số θ của một phân bố nào đó (chẳng hạn θ là kỳ vọng của phân bố chuẩn) là một khoảng $\theta_1 \leq \theta \leq \theta_2$ với một *xác suất cao (độ tin cậy)* γ. Ở đây θ_1 và θ_2 được gọi là độ tin cậy dưới và trên tương ứng, trong khi γ được gọi là *độ tin cậy*. Như vậy bài toán ở đây là xác định θ_1 và θ_2 sao cho xác suất để tìm thấy θ trong miền $[\theta_1, \theta_2]$ bằng γ, nghĩa là

$$P(\theta_1 \leq \theta \leq \theta_2) = \gamma. \tag{2.8}$$

2.3.1 Khoảng tin cậy cho μ của phân bố chuẩn khi biết phương sai σ^2

Trước tiên ta trình bày định lý về tổng của các biến số ngẫu nhiên độc lập có phân bố chuẩn.

Định lý 2.1 (Tổng của các biến ngẫu nhiên độc lập có phân bố chuẩn)**.** *Cho $X_1, \cdots, X_n$ là các biến số ngẫu nhiên độc lập có phân bố chuẩn và có chung kỳ vọng μ và phương sai σ^2. Khi đó, các khẳng định sau là đúng:*

(a) Tổng $X_1 + \cdots + X_n$ là biến số ngẫu nhiên có phân bố chuẩn với kỳ vọng $n\mu$ và phương sai $n\sigma^2$.

(b) Biến ngẫu nhiên được định nghĩa như sau

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n),$$

là biến số ngẫu nhiên có phân bố chuẩn với kỳ vọng μ và phương sai $\frac{\sigma^2}{n}$.

(c) Biến số ngẫu nhiên được định nghĩa như sau

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

có phân bố chuẩn với kỳ vọng 0 và phương sai 1.

Sử dụng định lý này, chúng ta có thể xác định khoảng tin cậy cho trung bình (kỳ vọng) μ của phân bố chuẩn khi biết phương sai σ^2. Trước tiên ta đưa ra định nghĩa về *giá trị tới hạn mức α (phân vị mức α)* của phân bố chuẩn tắc.

Định nghĩa 2.2. Cho biến ngẫu nhiên $X \sim N(0,1)$ có phân bố chuẩn tắc và số $\alpha \in (0,1)$. Giá trị U_α được gọi là *phân vị mức α* hay *giá trị tới hạn mức α* của biến ngẫu nhiên X nếu

$$P(X \le U_\alpha) = 1 - \alpha \tag{2.9}$$

(xem Hình 2.5) hay

$$\Phi(U_\alpha) = 1 - \alpha.$$

ở đó $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$ là hàm Laplace.

Giá trị của U_α được cho bởi bảng A8. Chẳng hạn, với $\alpha = 5\%$, thì $\Phi(U_\alpha) = 95\%$ và $U_\alpha = 1.645$.

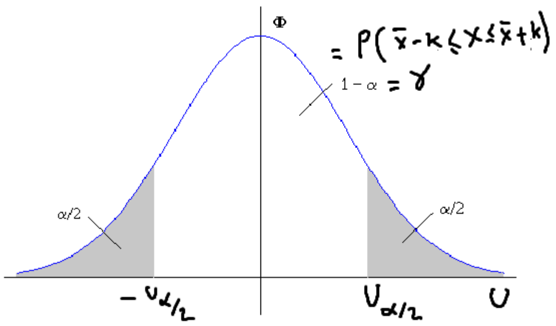
Các bước xác định khoảng tin cậy cho μ trong trường hợp này như sau:

Bước 1: Chọn độ tin cậy γ (95%, 99% hoặc khác).

Bước 2: Xác định giá trị tới hạn $U_{\alpha/2}$ tương ứng với mức $\frac{\alpha}{2} = \frac{1-\gamma}{2}$ từ Bảng 2.2 sau (hoặc xem bảng A8)

γ	90%	95%	99%	99.9%
$\alpha/2 = (1-\gamma)/2$	5%	2.5%	0.5%	0.05%
$U_{\alpha/2}$	1.645	1.960	2.576	3.291

Bảng 2.2: Một số giá trị tới hạn $U_{\alpha/2}$



Hình 2.6: Khoảng tin cậy cho μ của phân bố chuẩn khi biết phương sai σ^2

Bước 3: Tính trung bình mẫu $\bar{x}$ từ tập mẫu $x_1, x_2, \cdots, x_n$ đã cho.

Bước 4: Tính $k = \frac{U_{\alpha/2}\sigma}{\sqrt{n}}$. Khoảng tin cậy cho μ với độ tin cậy γ là

$$\bar{x} - k \leq \mu \leq \bar{x} + k. \tag{2.10}$$

(Xem Hình 2.6.)

Chứng minh công thức (2.10) ***là khoảng tin cậy cần tìm***. Cho $X_1, \cdots, X_n$ là các biến số ngẫu nhiên độc lập có phân bố chuẩn và có chung kỳ vọng μ và phương sai σ^2. Khi đó theo Định lý 2.1,

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

là biến số ngẫu nhiên có phân bố chuẩn với kỳ vọng μ và phương sai $\frac{\sigma^2}{n}$ và biến số ngẫu nhiên

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

có phân bố chuẩn tắc. Khi đó, ta có

$$\begin{aligned}
P(-k \leq \overline{X} - \mu \leq k) &= P\left(-\frac{k\sqrt{n}}{\sigma} \leq \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \leq \frac{k\sqrt{n}}{\sigma}\right) \\
&= P(-U_{\alpha/2} \leq Z \leq U_{\alpha/2}) \\
&= \Phi(U_{\alpha/2}) - \Phi(-U_{\alpha/2}) \\
&= 2\Phi(U_{\alpha/2}) - 1 \\
&= 2(1 - \alpha/2) - 1 \\
&= \gamma.
\end{aligned}$$

□

Ví dụ 2.6. *Tìm khoảng ước lượng với độ tin cậy 95% cho kỳ vọng μ của phân bố chuẩn với phương sai $\sigma^2 = 9$ sử dụng tập mẫu có kích thước $n = 100$ và trung bình mẫu $\bar{x} = 5$.*

Giải: *Cấp độ tin cậy được yêu cầu là $\gamma = 95\%$, vì thế $\alpha/2 = (1-0.95)/2 = 2.5\%$ và $U_{\alpha/2} = 1.960$. Trung bình mẫu $\bar{x}$ đã được xác định là 5. Độ lệch chuẩn $\sigma = \sqrt{9} = 3$. Chúng ta tính $k = \frac{U_0\sigma}{\sqrt{n}} = \frac{1.960\times 3}{\sqrt{100}} = 0.588$. Khoảng ước lượng cho μ với cấp độ tin cậy 95% là*

$$5 - 0.588 \leq \mu \leq 5 + 0.588 \quad \textit{hay} \quad 4.412 \leq \mu \leq 5.588.$$

Ví dụ 2.7. *Tìm kích thước mẫu n nhỏ nhất trong Ví dụ 2.6 sao cho khoảng tin cậy có độ dài $L = 0.4$ với độ tin cậy $\gamma = 0.95$?*

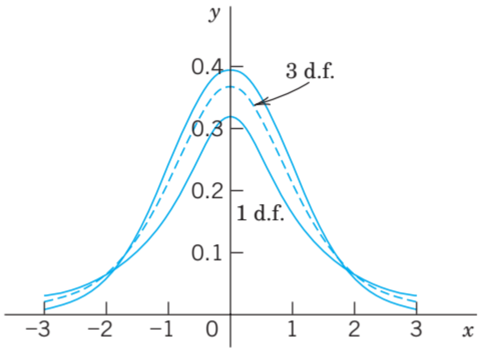
Giải. *Độ dài khoảng tin cậy là $L = 2k = 2U_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Vì $\gamma = 0.95$ nên $U_{\alpha/2} = 1.960$ và do đó*

$$n = \left(\frac{2U_{\alpha/2}\sigma}{L}\right)^2 = \left(\frac{2\times 1.960\times 3}{0.4}\right)^2 \approx 865.$$

2.3.2 Khoảng tin cậy cho μ của phân bố chuẩn khi chưa biết σ^2

Trong thực tế, phương sai σ^2 thường không biết, vì thế phương pháp ước lượng trên không còn hữu ích. Trong mục này chúng ta sẽ đưa ra phương pháp khác để ước lượng kỳ vọng μ cho trường hợp phương sai σ^2 không biết. Trước tiên ta có định lý về phân bố Student [1].

[1]"Student" là bút danh của nhà bác học người Anh tên thật là William Sealy Gosset.



Hình 2.7: Hàm mật độ của phân bố Student với 1 và 3 bậc tự do và hàm mật độ của phân bố chuẩn tắc

Định lý 2.2 (Phân bố Student)**.** *Cho $X_1, \cdots, X_n$ là các biến số ngẫu nhiên độc lập có phân bố chuẩn, có cùng kỳ vọng μ và phương sai σ^2. Biến số ngẫu nhiên được định nghĩa như sau*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}, \tag{2.11}$$

có phân bố Student $(n-1)$ bậc tự do và với hàm phân bố xác suất được cho như sau

$$F(z) = K_{n-1} \int_{-\infty}^{z} \left(1 + \frac{u^2}{n-1}\right)^{-\frac{n}{2}} du, \tag{2.12}$$

với

$$K_m = \frac{\Gamma(m/2 + 1/2)}{\sqrt{m\pi}\Gamma(m/2)},$$

trong đó $\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt$ là hàm gamma. Ở đây

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n), \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Nhận xét 2.2. *(i) Đồ thị hàm mật độ của phân bố Student $(n-1)$ bậc tự do cũng có dạng "hình chuông" tương tự phân bố chuẩn tắc (xem Hình 2.7.)*

(ii) Hàm phân bố xác suất $F(z)$ được cho bởi công thức (2.12) *có các tính chất sau:*

- $F(+\infty) = 1$*;*

- $F(-z) = 1 - F(z)$ *với mọi* $z \in \mathbb{R}$.

(iii) Giá trị t_α *được gọi là* giá trị tới hạn mức α *của phân bố Student* $(n-1)$ *bậc tự do nếu*

$$F(t_\alpha) = 1 - \alpha \tag{2.13}$$

với F *là hàm được cho bởi công thức* (2.12). *Giá trị của* t_α *được cho bởi bảng A9. Chẳng hạn, xét* $\alpha = 0.005$ *và* $n - 1 = 4$, *ta có* $F(t_\alpha) = 0.995$ *và* $t_\alpha = 4.60$.

(iv) Tại giới hạn số bậc tự do là ∞ *thì phân bố Student trùng với phân bố chuẩn.*

Sử dụng Định lý 2.2, chúng ta có thể xác định khoảng tin cậy cho μ của phân bố chuẩn khi chưa biết phương sai σ^2. Các bước xác định khoảng tin cậy cho μ trong trường hợp này như sau:

Bước 1: Chọn một độ tin cậy γ.

Bước 2: Xác định giá trị tới hạn $t_{\alpha/2}$ ứng với mức $\frac{\alpha}{2} = \frac{1-\gamma}{2}$ từ phương trình sau

$$F(t_{\alpha/2}) = 1 - \frac{\alpha}{2} = \frac{1}{2}(1 + \gamma), \tag{2.14}$$

nhờ bảng $A9$.

Bước 3: Tính trung bình mẫu $\bar{x}$ và phương sai mẫu s^2 từ tập mẫu kích thước n đã cho.

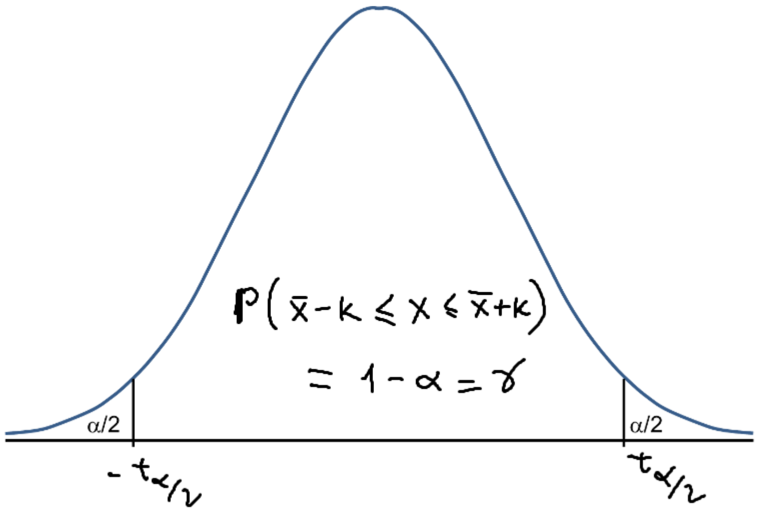
Bước 4: Tính $k = \frac{t_{\alpha/2}s}{\sqrt{n}}$. Khoảng tin cậy cho μ với kỳ vọng γ được xác định như sau

$$\bar{x} - k \leq \mu \leq \bar{x} + k. \tag{2.15}$$

(Xem Hình 2.8.)

Chứng minh công thức (2.15) ***là khoảng tin cậy cần tìm***. Cho $X_1, \cdots, X_n$ là các biến số ngẫu nhiên độc lập có phân bố chuẩn và có chung kỳ vọng μ và phương sai σ^2. Khi đó theo Định lý 2.2, biến ngẫu nhiên

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$



Hình 2.8: Khoảng tin cậy cho μ của phân bố chuẩn khi chưa biết phương sai σ^2

có phân bố Student $(n-1)$ bậc tự do. Khi đó, ta có

$$
\begin{aligned}
P(-k \le \overline{X} - \mu \le k) &= P\left(-\frac{k\sqrt{n}}{S} \le \frac{(\overline{X}-\mu)\sqrt{n}}{S} \le \frac{k\sqrt{n}}{S}\right) \\
&= P(-t_{\alpha/2} \le T \le t_{\alpha/2}) \\
&= F(t_{\alpha/2}) - F(-t_{\alpha/2}) \\
&= 2F(t_{\alpha/2}) - 1 \\
&= 2(1-\alpha/2) - 1 = 1 - \alpha \\
&= \gamma.
\end{aligned}
$$

□

Ví dụ 2.8. *Nhiệt độ (^oF) ở điểm đốt cháy của dầu Diesel được đo năm lần với các giá trị lần lượt như sau:* 144 147 146 142 144. *Giả sử tập nền có phân bố chuẩn, hãy tìm khoảng ước lượng cho kỳ vọng với độ tin cậy* 99%.

Giải: *Độ tin cậy được yêu cầu là* $\gamma = 0.99$ *và do đó* $\alpha = 0.01$. *Phương trình cho* $t_{\alpha/2}$ *là,* $F(t_{\alpha/2}) = 1 - \alpha/2 = 0.995$*, sử dụng bảng A9, ta tìm được* $t_{\alpha/2} = 4.60$. *Từ tập mẫu ta tính được trung bình mẫu và phương sai mẫu như sau,* $\bar{x} = 144.6$ *và* $s^2 = 3.8$. $k = \frac{4.60\sqrt{3.8}}{\sqrt{5}} = 4.01$. *Khoảng ước lượng cho* μ *với cấp độ tin cậy* 99% *là*

$$144.6 - 4.01 \le \mu \le 144.6 + 4.01 \quad \text{hay} \quad 140.59 \le \mu \le 148.61.$$

2.3.3 Khoảng tin cậy cho xác suất p của biến ngẫu nhiên có phân bố tuân theo quy luật không–một $A(p) \sim B(1,p)$

Các mục 2.3.1 và 2.3.2 được xây dựng cho biến ngẫu nhiên có phân bố chuẩn. Trong mục này chúng ta chỉ ra rằng chúng cũng được áp dụng để tìm khoảng tin cậy đối với các biến ngẫu nhiên bất kỳ nếu **kích thước** n **của mẫu đủ lớn**.

Cho $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên độc lập có cùng trung bình μ và phương sai σ^2. Khi đó $\overline{X} = (X_1 + \cdots + X_n)/n$ có tính chất sau:

(i) $\overline{X}$ có trung bình là μ và phương sai là σ^2/n;

(ii) Nếu $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên có phân bố chuẩn, thì $\overline{X}$ cũng có phân bố chuẩn.

Khi $X_1, X_2, \ldots, X_n$ không phải là các biến ngẫu nhiên có phân bố chuẩn, thì (ii) không áp dụng được. Tuy nhiên ta vẫn có thể xấp xỉ $\overline{X}$ bởi một biến ngẫu nhiên có phân bố chuẩn nhờ *Định lý Giới hạn trung tâm* dưới đây.

Định lý 2.3 (Định lý Giới hạn trung tâm (Central Limit Theorem))**.** *Cho $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên độc lập có cùng trung bình μ và phương sai σ^2. Đặt $\overline{X} = (X_1 + \cdots + X_n)/n$. Khi đó*

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{2.16}$$

được xấp xỉ tiệm cận bởi một phân bố chuẩn có trung bình 0 và phương sai 1 (phân bố chuẩn tắc), có nghĩa là

$$\lim_{n\to+\infty} F_n(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du \tag{2.17}$$

với $F_n(x)$ là hàm phân bố của Z_n.

Ta sẽ áp dụng Định lý Giới hạn trung tâm để xây dựng khoảng tin cậy cho xác suất tổng thể p. Gọi p là xác suất tổng thể của một tập nền để xuất hiện sự kiện A. Cho biến ngẫu nhiên X (của tập nền) được xác định như sau

$$X = \begin{cases} 0 & \text{nếu không xảy ra sự kiện } A \text{ trong phép thử} \\ 1 & \text{nếu xảy ra sự kiện } A \text{ trong phép thử.} \end{cases}$$

Ta có $X \sim A(p) = B(1,p)$ và $\mu = EX = p$, $\sigma^2 = DX = p(1-p)$. Xét một mẫu $(x_1, x_2, ..., x_n)$ có kích thước n. Khi đó $(x_1 + x_2 + \cdots + x_n)$ bằng số lần xuất hiện sự kiện A. *Tần suất* xuất hiện sự kiện A trong mẫu là

$$f = \frac{(x_1 + x_2 + \cdots + x_n)}{n} = \bar{x}.$$

Ta đã biết tần suất mẫu f là một ước lượng (điểm) của xác suất tổng thể p. Theo Định lý Giới hạn trung tâm, ta có thể xấp xỉ $Z_n = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ bởi phân bố chuẩn tắc khi n đủ lớn. Do đó, theo công thức (2.10) khoảng tin cậy cho kỳ vọng $\mu = p$ với độ tin cậy γ là

$$\bar{x} - \frac{U_{\alpha/2}\sigma}{\sqrt{n}} \le \mu = p \le \bar{x} + \frac{U_{\alpha/2}\sigma}{\sqrt{n}}, \tag{2.18}$$

ở đó giá trị tới hạn $U_{\alpha/2}$ ứng với mức $\alpha/2$ của phân bố chuẩn tắc, được tính từ bảng A8 (hoặc bảng A9 với bậc tự do bằng ∞). Tuy nhiên vì p chưa biết, nên thay p bởi f và đo đó thay σ bởi $\sqrt{f(1-f)}$ vào công thức (2.18), ta thu được khoảng tin cậy cho xác suất p với độ tin cậy γ (với n đủ lớn) là

$$f - U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} \le p \le f + U_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}, \tag{2.19}$$

Ví dụ 2.9. *Trong một cuộc bầu cử tổng thống ở nước Mỹ, người ta phỏng vấn ngẫu nhiên 1600 cử tri và được biết có 960 người bầu cử cho ứng viên A. Với độ tin cậy $\gamma = 95\%$, tối thiểu ứng cử viên A sẽ chiếm được bao nhiêu phần trăm số phiếu bầu?*

Giải. *Gọi p là tỷ lệ phiếu bầu cho ứng cử viên A. Gọi X là biến ngẫu nhiên được cho bởi*

$$X = \begin{cases} 0 & \text{nếu ứng cử viên A không được bầu} \\ 1 & \text{nếu ứng cử viên A được bầu.} \end{cases}$$

Khi đó X là biến ngẫu nhiên có phân bố không-một $A(p) \sim B(1,p)$. Ta có, tần suất mẫu là

$$f = \frac{960}{1600} = 0.6.$$

Vì $\gamma = 0.95$ nên $\alpha/2 = 2.5\%$ và $U_{\alpha/2} = 1.960$ (sử dụng bảng A8) và đó đó khoảng tin cậy của p là

$$0.6 - 1.960 \times \sqrt{\frac{0.6(1-0.6)}{1600}} \le p \le 0.6 + 1.960 \times \sqrt{\frac{0.6(1-0.6)}{1600}}$$

hay

$$0.576 \leq p \leq 0.624.$$

Vậy với độ tin cậy 95%, *thì tối thiểu có* 57.6% *và tối đa* 62.4% *cử tri sẽ bỏ phiếu cho ứng viên A.*

Bài tập Chương 2

Bài 2.1. Biểu diễn dữ liệu bằng biểu đồ stem-and-lead và histogra của các dữ liệu sau:

(a) Độ dài của móng tay (mm):

20 21 21 19 20 19 21 19

(b) Huyết áp của 15 bệnh nhân nữ lứa tuổi 20–22

156 158 154 133 141 130 144 137
151 146 156 138 138 149 139

(c) Release time [sec] of a relay (rơ-le)

1.3 1.2 1.4 1.5 1.3 1.3 1.4 1.1 1.5 1.4
1.6 1.3 1.5 1.1 1.4 1.2 1.3 1.5 1.4 1.4

Bài 2.2. Vẽ biểu đồ hộp và tìm các dữ liệu ngoại lai outlier (nếu có) của các dữ liệu **Bài 2.1.**

Bài 2.3. Tính trung bình, phương sai, độ lệch chuẩn của dữ liệu **Bài 2.1.**

Bài 2.4. Hãy tính trung bình mẫu và độ lệch chuẩn s của mẫu sau:

x_i	21	24	26	27	29	30	32	33
n_i	10	11	13	17	20	18	15	12

Bài 2.5. Thống kê tuổi thọ trung bình X của một loại bóng đèn, người ta thu được bảng dữ liệu sau:

Tuổi thọ X (giờ)	Số bóng đèn	Tuổi thọ X (giờ)	Số bóng đèn
1000–1020	3	1020–1040	4
1040–1060	7	1060–1080	20
1080-1100	40	1100–1120	39
1120–1140	25	1140–1160	18
1160–1180	9	1180-1200	2

(a) Hãy tính trung bình mẫu và độ lệch chuẩn s của mẫu.

(b) Biết tuổi thọ trung bình X của loại bóng đèn trên tuân theo quy luật phân bố chuẩn với độ lệch chuẩn là $\sigma = 130$ giờ. Hãy tìm khoảng tin cậy của tuổi thọ trung bình với độ tin cậy $\gamma = 95\%$.

Bài 2.6. Cho biến ngẫu nhiên X nhận giá trị là số các phép thử độc lập cho đến khi một sự kiện **A** xuất hiện. Biết X có hàm xác suất (hàm mật độ) được xác định bởi $f(x) = pq^{x-1}$ với $x = 1, 2, \cdots$, p là xác suất sự kiện **A** xuất hiện trong một phép thử và $q = 1 - p$. Cho một mẫu kích thước n $(x_1, x_2, \cdots, x_n)$ tương ứng với các giá trị quan sát của X. Sử dụng phương pháp hợp lý cực đại, hãy ước lượng p.

Bài 2.7. Một biến ngẫu nhiên X có hàm xác suất được cho bởi $f(x) = \theta e^{-\theta x}$ nếu $x \geq 0$ và $f(x) = 0$ nếu $x < 0$. Cho một mẫu ngẫu nhiên tương ứng với X như sau: 0.5, 0.7, 0.1, 1.1, 0.1. Sử dụng phương pháp hợp lý cực đại, hãy ước lượng θ.

Bài 2.8. Tìm khoảng ước lượng với độ tin cậy 95% cho kỳ vọng μ của phân bố chuẩn với độ lệch chuẩn $\sigma = 4.00$ từ mẫu thực nghiệm sau: 30 42 40 34 48 50.

Bài 2.9. Tìm khoảng ước lượng với độ tin cậy 90% cho kỳ vọng μ của phân bố chuẩn với phương sai $\sigma^2 = 0.25$ sử dụng mẫu có kích thước bằng 100 và trung bình mẫu bằng 212.3.

Bài 2.10. Hãy xác định kích thước mẫu để khoảng ước lượng với độ tin cậy 95% cho kỳ vọng của phân bố chuẩn có chiều dài 2σ.

Bài 2.11. Hãy xác định kích thước mẫu để khoảng ước lượng với độ tin cậy 99% cho kỳ vọng của phân bố chuẩn với phương sai $\sigma^2 = 25$ có chiều dài 2.0.

Bài 2.12. Đo chiều dài X của 20 con ốc vít, người ta thu được mẫu với trung bình mẫu 20.2 cm và phương sai mẫu 0.04 cm^2. Biết X có phân bố chuẩn, hãy tìm khoảng ước lượng cho chiều dài trung bình của con ốc vít với độ tin cậy 99%.

Bài 2.13. Đo độ cứng Knoop Y (đơn vị: N/m^2) của kim cương, người ta thu được mẫu thực nghiệm sau: 9500 9800 9750 9200 9400 9550. Biết Y có phân bố chuẩn, hãy tìm khoảng ước lượng cho độ cứng Knoop trung bình của kim cương với độ tin cậy 99%.

Bài 2.14. Đo nhiệt độ nóng chảy T (đơn vị: ^oC) của nhôm, người ta thu được mẫu thực nghiệm sau: 660 667 654 663 662. Biết T có phân bố chuẩn, hãy tìm khoảng ước lượng cho nhiệt độ nóng chảy trung bình của nhôm với độ tin cậy 99%.

Bài 2.15. Đo lường lượng phát thải CO (đơn vị: gram/mile) của một loại xe chở khách (di chuyển ở tốc độ 55 mph), người ta thu được mẫu thực nghiệm sau: 17.3 17.7 18.0 17.7 18.2 17.4 17.6 18.1. Biết sự phát thải CO trong trường hợp này có phân bố chuẩn, hãy tìm khoảng ước lượng cho lượng CO trung bình được phát thải từ loại xe này với độ tin cậy 99%.

Bài 2.16. Muốn ước lượng được số cá trong hồ, người ta bắt 2000 con cá trong hồ rồi đánh dấu chúng và thả chúng lại xuống hồ. Sau đó lại bắt lại 400 con và thấy có 53 con có dấu. Hãy ước lượng số cá trong hồ với độ tin cậy là 0.95.

Bài 2.17. Để ước lượng tỷ lệ phần trăm phế phẩm của một lô hàng người ta tiến hành kiểm tra ngẫu nhiên 400 sản phẩm và nhận thấy có 16 phế phẩm. Với mức tin cậy 95%, hãy ước lượng tỷ lệ phế phẩm tối thiểu và tối đa của lô hàng?

Bài 2.18. Hao phí nhiên liệu (gram) cho một đơn vị sản phẩm là một biến ngẫu nhiên X tuân theo quy luật phân bố chuẩn. Sản xuất thử một số sản phẩm và thu được bảng sau

X	19-19.2	19.2-19.4	19.4-19.6	19.9-19.8
Số sản phẩm	7	8	18	10

(a) Nếu biết $X \sim N(\mu, \sigma)$ với $\sigma = 0.02$ (gram), hãy ước lượng hao phí nhiên liệu trung bình cho một đơn vị sản phẩm với độ tin cậy $\gamma = 95\%$.

(b) Trong trường hợp không biết σ, hãy ước lượng hao phí nhiên liệu trung bình cho một đơn vị sản phẩm với độ tin cậy $\gamma = 99\%$.

Bài 2.19. Đo ngẫu nhiên chiều cao X của 40 cây con, ta thu được bảng dữ liệu sau

X (cm)	16.5-17	17-17.5	17.5-18	18-18.5	18.5-19	19-19.5
Số cây	3	5	11	12	6	3

Biết chiều cao X tuân theo quy luật phân bố chuẩn.

(a) Tìm khoảng tin cậy của chiều cao trung bình biết độ tin cậy $\gamma = 90\%$.

(b) Nếu muốn khoảng ước lượng có độ dài không quá $\epsilon = 0.1$ thì cần lấy ít nhất bao nhiêu mẫu?

Chương 3

Kiểm định giả thiết thống kê

Ý tưởng về khoảng tin cậy và về kiểm định giả thiết thống kê là hai ý tưởng quan trọng nhất trong thống kê hiện đại. Trong kiểm định giả thiết thống kê, chúng ta đưa ra suy luận từ mẫu đến tổng thể (tập nền) thông qua việc kiểm tra một giả thiết. Điều này được thực hiện dựa trên kinh nghiệm hoặc quan sát, dựa trên lý thuyết hoặc yêu cầu chất lượng,... Chẳng hạn, kết quả của sự kiểm định một giả thiết thông kê là cơ sở để quyết định mua (hay không mua) một mẫu xe ô tô nào đó, tùy thuộc vào việc kiểm định giả thiết về hiệu suất tiêu thụ nhiên liệu của mẫu xe đó; để sử dụng một loại thuốc, tùy thuộc vào việc thử nghiệm về tác dụng của nó; để tiến hành chiến lược tiếp thị một mặt hàng, tùy thuộc vào phản ứng của người tiêu dùng về mặt hàng đó.

3.1 Khái niệm, quy tắc chung và các dạng bài toán kiểm định

Ví dụ 3.1. *Chúng ta muốn mua* 100 *cuộn dây kim loại nào đó, miễn là chúng ta có thể xác minh được khẳng định của nhà sản xuất rằng dây có giới hạn đứt* $\mu = \mu_0 = 200$ *lb (pound; 1 lb* $\approx$ *0.45 kg). Đây là một bài kiểm định của* giả thuyết (hypothesis) $H_0 : \mu = \mu_0 = 200$. *Chúng ta sẽ không mua cuộn dây nếu kiểm định (thống kê) cho thấy rằng giới hạn đứt của cuộn dây thực sự là* $H_1 : \mu = \mu_1 < \mu_0$. $H_1 : \mu = \mu_1$ *được gọi là* đối thiết *(hay* giả thuyết thay thế*) của phép thử. Chúng ta sẽ* chấp nhận *giả thuyết nếu thử nghiệm cho thấy giả thiết đó đúng, ngoại trừ một xác suất lỗi nhỏ* α*–được gọi là* mức ý nghĩa *của kiểm định. Ngược lại chúng ta sẽ* bác bỏ giả thuyết. *Do đó,* α *là xác suất bác bỏ một giả thuyết mặc dù nó là đúng. Thông thường ta chọn mức ý nghĩa* $\alpha = 5\%$ *hay* $\alpha = 1\%$.

Để đưa ra được kết luận của kiểm định trong ví dụ trên chúng ta dựa trên nguyên lý sau.

- *Nếu một sự kiện có xác suất rất nhỏ thì trong một phép thử hay một vài phép thử sự kiện đó sẽ không xảy ra.*

Dưới đây là một thủ tục (quy trình) kiểm định của giả thiết thống kê.

3.1.1 Thủ tục kiểm định giả thiết thống kê

Bước 1: Phát biểu giả thiết $H_0 : \theta = \theta_0$ (chẳng hạn trong Ví dụ 3.1 ta có $H_0 : \mu = \mu_0 = 200$, ở đây $\theta = \mu$ và $\theta_0 = \mu_0$.)

Bước 2: Phát biểu đối thiết $H_1 : \theta < (\neq, >)\theta_0$ (chẳng hạn trong Ví dụ 3.1 ta có $H_1 : \mu < \mu_0 = 200$, ở đây $\theta = \mu$ và $\theta_0 = \mu_0$.)

Bước 3: Chọn mức ý nghĩa α (thông thường chọn $\alpha = 5\%$, 1%, 0.01%.)

Bước 4: Chọn *tiêu chuẩn kiểm định* $T = g(X_1, X_2, ..., X_n; \theta_0)$ và xác định quy luật phân bố của T với điều kiện giả thiết H_0 là đúng. Ở đó $(X_1, X_2, ..., X_n)$ là biến ngẫu nhiên được lập từ một mẫu có kích thước n.

Bước 5: Tìm *miền bác bỏ* W_α sao cho

$$P(T \in W_\alpha | H_0) = \alpha, \tag{3.1}$$

tức là *trong trường hợp giả thiết H_0 là đúng thì xác suất xảy ra lỗi $T \in W_\alpha$ bằng α.*

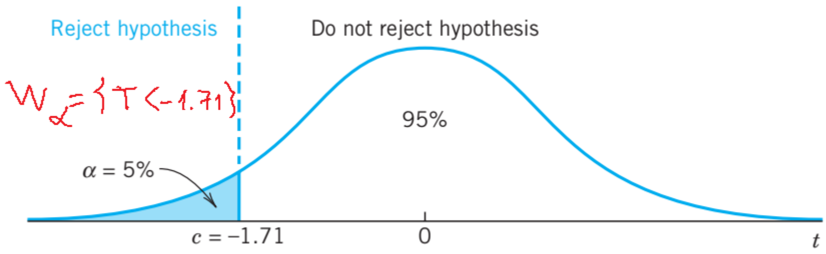
Bước 6: Từ mẫu cụ thể $(x_1, x_2, ..., x_n)$ tính *giá trị quan sát* của tiêu chuẩn kiểm định

$$T_{qs} = g(x_1, x_2, ..., x_n; \theta_0).$$

Bước 7: Chấp nhận hoặc bác bỏ giả thiết H_0 dựa trên mối quan hệ giữa giá trị quan sát T_{qs} và miền bác bỏ W_α:

- Nếu $T_{qs} \in W_\alpha$, theo nguyên tắc kiểm định thì H_0 sai, do đó **bác bỏ** H_0 và thừa nhận H_1.
- Nêu $T_{qs} \notin W_\alpha$, thì **chưa có đủ cơ sở để bác bỏ** H_0 (trên thực tế ta thừa nhận H_0.)

Ví dụ 3.1 *(tiếp). Để kiểm định được giới hạn đứt của cuộn dây kim loại, chúng ta thực nghiệm với mẫu ngẫu nhiên có kích thước $n = 25$ và thu được giới hạn đứt có giá trị trung bình mẫu $\bar{x} = 197$ lb và độ lệch chuẩn mẫu là $s = 6$ lb. Khi đó dựa trên thủ tục kiểm định giả thiết thống kê, ta có*



Hình 3.1: Miền bác bỏ W_α

- *Giả thiết* $H_0 : \mu = \mu_0 = 200$ *lb.*
- *Đối thiết* $H_1 : \mu < \mu_0$.
- *Mức ý nghĩa* $\alpha = 5\%$.
- *Chọn tiêu chuẩn kiểm định*

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

với $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ *và* $S = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right]$. *Nếu giả thiết* H_0 *đúng, thì* T *là phân bố Student* $(n-1)$ *bậc tự do (xem Định lý 2.2 trong Chương 2).*

- *Miền bác bỏ (xem Hình 3.1)*

$$W_\alpha = \{T < -t_\alpha\} = \{T < -1.71\}$$

với $t_\alpha = 1.71$ *là giá trị tới hạn thỏa mãn* $F(t_\alpha) = 1 - \alpha = 0.95$ *(xem bảng A9). Ở đó* $F(x)$ *là hàm phân bố của biến ngẫu nhiên* T.

- *Giá trị quan sát của tiêu chuẩn kiểm định*

$$T_{qs} = \frac{197 - 200}{6/\sqrt{25}} = -2.5$$

- *Vì giá trị quan sát* $T_{qs} = -2.5 < -1.71$, *tức là* $T_{qs} \in W_\alpha$, *do đó ta bác bỏ giả thiết* H_0. *Vậy giới hạn đứt của cuộn dây dường như nhỏ hơn lời khẳng định của nhà sản xuất.*

3.1.2 Sai lầm của kiểm định thống kê

Kiểm định (thống kê) có thể mắc hai loại sai lầm sau:

- **Sai lầm loại I:** Bác bỏ giả thiết H_0 trong khi nó đúng. Sai lầm loại này sinh ra khi kích thước mẫu quá nhỏ, do phương pháp lấy mẫu. Xác suất mắc sai lầm loại I đúng bằng mức ý nghĩa α.

- **Sai lầm loại II:** Thừa nhận H_0 trong khi nó sai. Sai lầm loại này xảy ra khi giá trị quan sát $T_{qs} \in W_\alpha$ trong khi H_1 đúng. Xác suất xảy ra sai lầm loại II được xác định như sau

$$P(T \notin W_\alpha | H_1) = \beta.$$

  Đại lượng $\eta = 1 - \beta = P(T \in W_\alpha | H_1)$ được gọi là *lực lượng kiểm định* (power of the test).

Ta muốn một quy tắc kiểm định mà xác suất xảy ra cả hai sai lầm loại I và II đều cực tiểu. Tuy nhiên không tồn tại một kiểm định nào lý tưởng như vậy, vì khi nếu xác suất xảy ra sai lầm loại I giảm thì xác suất xảy ra sai lầm loại II sẽ tăng lên và ngược lại. Trong thực tế, trước tiên chúng ta chọn α (5%, đôi khi chọn 1%), sau đó xác định miền bác bỏ W_α và từ đó tính được giá trị β. Nếu lực lượng kiểm định $\eta = 1 - \beta$ nhỏ, chúng ta lặp lại quá trình kiểm định bằng cách tăng kích thước mẫu.

3.1.3 Hai loại bài toán kiểm định

Trong chương này chúng ta xét hai loại bài toán kiểm định sau.

(i) *Kiểm định tham số* là kiểm định các tham số đặc trưng (như kỳ vọng, phương sai, tần suất,...) của biến ngẫu nhiên gốc trong tổng thể. Kiểm định về sự bằng nhau của hai kỳ vọng, của hai phương sai, hay của hai tần suất của tổng thể.

(ii) *Kiểm định phi tham số* bao gồm các bài toán về quy luật phân bố xác suất của biến ngẫu nhiên gốc của tổng thể, về tính độc lập của các dấu hiệu nghiên cứu định tính...

3.2 Kiểm định kỳ vọng của phân phối chuẩn

Giả sử biến ngẫu nhiên X trong tổng thể (tập nền) có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$. Nếu chúng ta có cơ sở để giả thiết rằng *giá trị của kỳ vọng μ bằng μ_0*, tức ta có giả thiết thống kê

$$H_0 : \mu = \mu_0. \tag{3.2}$$

Ta cần kiểm định giả thiết H_0.

3.2.1 Khi phương sai σ^2 đã biết

Xét mẫu ngẫu nhiên $(x_1, ..., x_n)$ tương ứng với biến ngẫu nhiên $W = (X_1, ..., X_n)$, ở đó $X_1, ..., X_n$ là các biến ngẫu nhiên có phân bố chuẩn và có cùng kỳ vọng và phương sai với X. Theo Định lý 2.1, kiểm định thống kê

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

là biến ngẫu nhiên phân bố chuẩn với kỳ vọng 0 và phương sai 1 (phân bố chuẩn tắc) nếu giả thiết H_0 đúng. Ta xét các bài toán kiểm định sau.

Bài toán 1

Bài toán *kiểm định hai phía*

$$\begin{cases} \text{Giả thiết } H_0 : \mu = \mu_0, \\ \text{Đối thiết } H_1 : \mu \neq \mu_0. \end{cases} \tag{3.3}$$

Miền bác bỏ (xem Hình 3.2)

$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, |T| > U_{\alpha/2} \right\} \tag{3.4}$$

ở đó giá trị tới hạn $U_{\alpha/2}$ được xác định bởi (xem bảng A8)

$$\Phi(U_{\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Bài toán 2

Bài toán *kiểm định một phía bên phải*

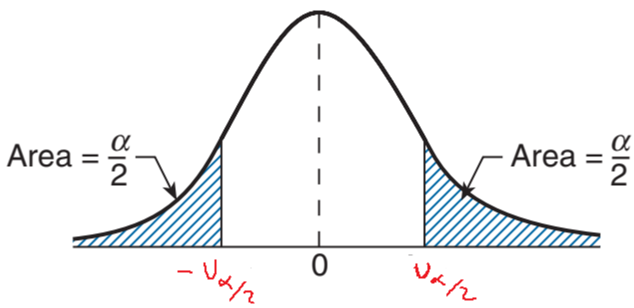
$$\begin{cases} \text{Giả thiết } H_0 : \mu = \mu_0, \\ \text{Đối thiết } H_1 : \mu > \mu_0. \end{cases} \tag{3.5}$$

Miền bác bỏ (xem Hình 3.3)

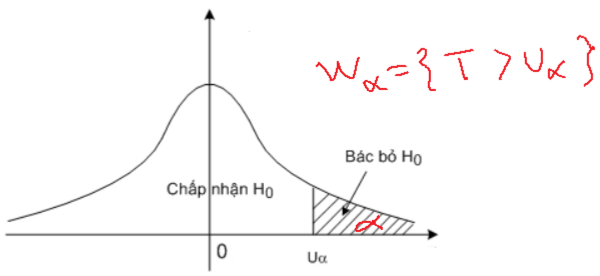
$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, T > U_\alpha \right\} \tag{3.6}$$

ở đó giá trị tới hạn U_α được xác định bởi (xem bảng A8)

$$\Phi(U_\alpha) = 1 - \alpha.$$



Hình 3.2: Miền bác bỏ W_α - phần gạch sọc



Hình 3.3: Miền bác bỏ W_α - phần gạch sọc

Bài toán 3

Bài toán *kiểm định một phía bên trái*

$$\begin{cases} \text{Giả thiết } H_0 : \mu = \mu_0, \\ \text{Đối thiết } H_1 : \mu < \mu_0. \end{cases} \tag{3.7}$$

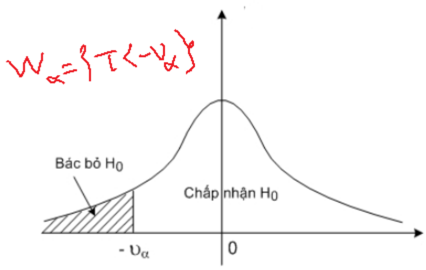
Miền bác bỏ (xem Hình 3.4)

$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \middle| T < -U_\alpha \right\} \tag{3.8}$$

ở đó giá trị tới hạn U_α được xác định bởi (xem bảng A8)

$$\Phi(U_\alpha) = 1 - \alpha.$$

Ví dụ 3.2. *Cho X là biến ngẫu nhiên có phân bố chuẩn với phương sai $\sigma^2 = 9$. Sử dụng một mẫu kích cỡ $n = 10$ với trung bình $\bar{x}$, kiểm định giả thiết $\mu = \mu_0 = 24$ với các đối thiết sau*



Hình 3.4: Miền bác bỏ W_α - phần gạch sọc

(i) $\mu \neq \mu_0$,

(ii) $\mu > \mu_0$,

(iii) $\mu < \mu_0$.

Biết mức ý nghĩa $\alpha = 5\%$.

Giải. *(i) Giá trị tới hạn* $U_{\alpha/2}$ *thỏa mãn*

$$\Phi(U_{\alpha/2}) = 1 - \frac{\alpha}{2} = 0.975.$$

Theo bảng A8, ta có $U_{\alpha/2} = 1.960$ *và vì vậy miền bác bỏ*

$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, |T| > 1.960 \right\}.$$

Giá trị quan sát của tiêu chuẩn kiểm định

$$T_{qs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 24}{3/\sqrt{10}}.$$

Vậy ta có

- *Nếu* $T_{qs} \in W_\alpha$ *hay* $x < 22.14$ *hoặc* $\bar{x} > 25.86$, *thì bác bỏ giả thiết* H_0, *chấp nhận* H_1.
- *Nếu* $T_{qs} \notin W_\alpha$ *hay* $22.14 \leq \bar{x} \leq 25.86$, *thì chấp nhận giả thiết* H_0.

(ii) Tương tự ta có $U_\alpha = 1.645$ *và miền bác bỏ*

$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, T > 1.645 \right\}.$$

Vậy ta có

- *Nếu $T_{qs} \in W_\alpha$ hay $\bar{x} > 25.56$, thì bác bỏ giả thiết H_0, chấp nhận H_1.*
- *Nếu $T_{qs} \notin W_\alpha$ hay $\bar{x} \leq 25.58$, thì chấp nhận giả thiết H_0.*

(iii) Tương tự ta có

- *Nếu $T_{qs} \in W_\alpha$ hay $\bar{x} < 22.44$, thì bác bỏ giả thiết H_0, chấp nhận H_1.*
- *Nếu $T_{qs} \notin W_\alpha$ hay $\bar{x} \geq 22.44$, thì chấp nhận giả thiết H_0.*

3.2.2 Khi phương sai σ^2 chưa biết

Xét mẫu ngẫu nhiên $(x_1, ..., x_n)$ tương ứng với biến ngẫu nhiên $W = (X_1, ..., X_n)$, ở đó $X_1, ..., X_n$ là các biến ngẫu nhiên có phân bố chuẩn và có cùng kỳ vọng và phương sai với X. Đặt $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ và $S = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right]$. Theo Định lý 2.2, kiểm định thống kê

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

là biến ngẫu nhiên phân bố Student $(n-1)$ bậc tự do nếu giả thiết H_0 đúng. Ta xét các bài toán kiểm định sau.

Bài toán 1

Bài toán *kiểm định hai phía*

$$\begin{cases} \text{Giả thiết } H_0 : \mu = \mu_0, \\ \text{Đối thiết } H_1 : \mu \neq \mu_0. \end{cases} \tag{3.9}$$

Miền bác bỏ (xem Hình 3.5)

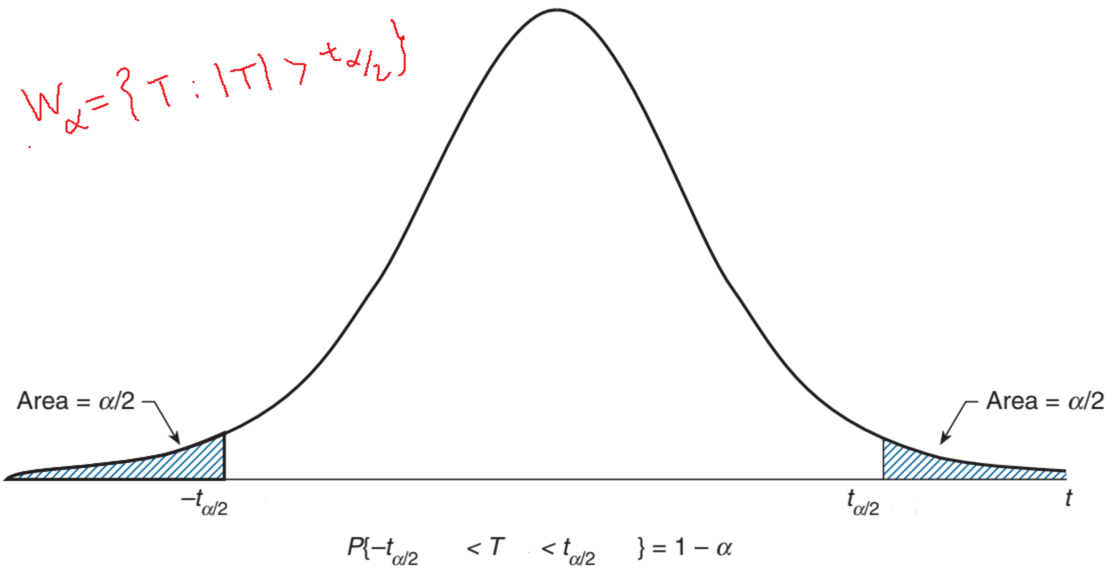
$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \,\middle|\, |T| > t_{\alpha/2} \right\} \tag{3.10}$$

ở đó giá trị tới hạn $t_{\alpha/2}$ được xác định bởi (xem bảng A9)

$$F(t_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

với $F(x)$ là hàm phân bố của biến ngẫu nhiên T.

$W_\alpha = \{T : |T| > t_{\alpha/2}\}$



Hình 3.5: Miền bác bỏ W_α - phần gạch sọc

Bài toán 2

Bài toán *kiểm định một phía bên phải*

$$\begin{cases} \text{Giả thiết } H_0 : \mu = \mu_0, \\ \text{Đối thiết } H_1 : \mu > \mu_0. \end{cases} \tag{3.11}$$

Miền bác bỏ (xem Hình 3.6)

$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \middle| T > t_\alpha \right\} \tag{3.12}$$

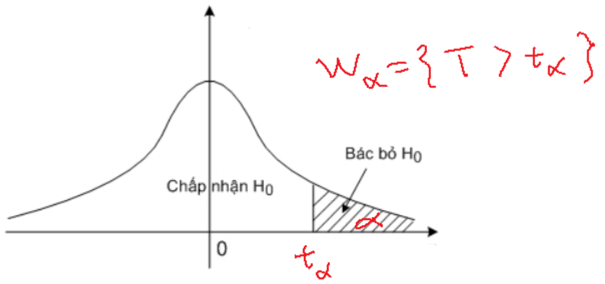
ở đó giá trị tới hạn t_α được xác định bởi (xem bảng A9)

$$F(t_\alpha) = 1 - \alpha.$$

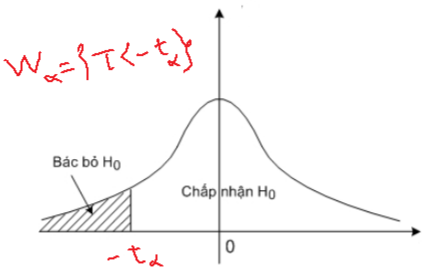
Bài toán 3

Bài toán *kiểm định một phía bên trái*

$$\begin{cases} \text{Giả thiết } H_0 : \mu = \mu_0, \\ \text{Đối thiết } H_1 : \mu < \mu_0. \end{cases} \tag{3.13}$$



Hình 3.6: Miền bác bỏ W_α - phần gạch sọc



Hình 3.7: Miền bác bỏ W_α - phần gạch sọc

Miền bác bỏ (xem Hình 3.7)

$$W_\alpha = \left\{ T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \,\middle|\, T < -t_\alpha \right\} \tag{3.14}$$

ở đó giá trị tới hạn t_α được xác định bởi (xem bảng A9)

$$F(t_\alpha) = 1 - \alpha.$$

Ví dụ 3.3. *Thử nghiệm sức căng của một mẫu gồm $n = 16$ sợi dây thừng (đường kính 3 in.) thu được trung bình mẫu $\bar{x} = 4482$ kg và độ lệch chuẩn mẫu $s = 115$ kg. Biết rằng sức căng của dây thừng là một biến ngẫu nhiên có phân bố chuẩn. Với mức ý nghĩa 5%, hãy kiểm định giả thiết sức căng trung bình của dây thừng là $\mu = \mu_0 = 4500$ kg so với đối thiết $\mu = \mu_1 = 4400$ kg. (Giá trị $\mu_0 = 4500$ kg được đưa ra bởi nhà sản xuất, trong khi đó giá trị $\mu_1 = 4400$ kg là kết quả được cho bởi thực nghiệm trước đó.)*

Giải. *Ta có*

$$\begin{cases} \textit{Giả thiết}\, H_0 : \mu = \mu_0 = 4500, \\ \textit{Đối thiết}\, H_1 : \mu = \mu_1 < \mu_0. \end{cases}$$

Giá trị tới hạn t_α thỏa mãn

$$F(t_\alpha) = 1 - \alpha = 0.95$$

Tra bảng A9 với $(n-1) = 15$ bậc tự do ta có $t_\alpha = 1.75$. Do đó miền bác bỏ

$$W_\alpha = \left\{ T = \frac{\bar{X} - 4500}{S/\sqrt{n}} \,\middle|\, T < -1.75 \right\}. \tag{3.15}$$

Giá trị quan sát

$$T_{qs} = \frac{4482 - 4500}{115/\sqrt{16}} = -0.626.$$

Vì $T_{qs} > -1.75$ nên $T_{qs} \notin W_\alpha$ và do đó chấp nhận H_0. (Chúng ta chấp nhận giá trị về sức căng trung bình của sợi dây thừng do nhà sản xuất đưa ra.)

3.3 Kiểm định phương sai của phân bố chuẩn

3.3.1 Phân bố χ^2 (khi-bình phương)

Định nghĩa 3.1 (Phân bố khi-bình phương)**.** Biến ngẫu nhiên X được gọi là có phân bố *khi-bình phương* với m bậc tự do và ký hiệu $X \sim \chi^2(m)$ nếu hàm phân bố $F(x)$ của nó có dạng

$$F(x) = \begin{cases} 0 & \text{nếu } x < 0, \\ C_m \displaystyle\int_0^x e^{-u/2} u^{(m-2)/2} du & \text{nếu } x \geq 0 \end{cases} \tag{3.16}$$

với $C_m = \frac{1}{2^{m/2}\Gamma(m/2)}$ và $\Gamma(t)$ là hàm Gamma.

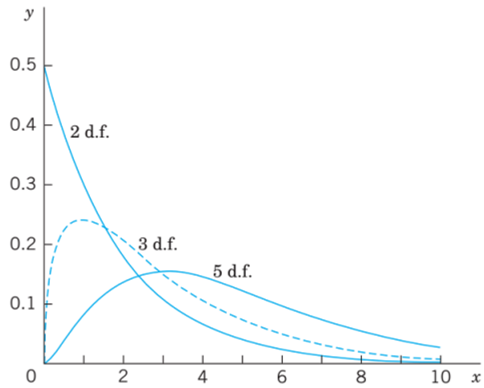
Hàm mật độ của phân bố khi-bình phương ứng với 2,3, và 5 bậc tự do được cho bởi Hình 3.8.

Định lý 3.1 (Phân bố khi-bình phương)**.** *Cho $X_1, \cdots, X_n$ là các biến số ngẫu nhiên độc lập có phân bố chuẩn, có cùng kỳ vọng μ và phương sai σ^2. Biến số ngẫu nhiên được định nghĩa như sau*

$$Y = (n-1)\frac{S^2}{\sigma^2}, \tag{3.17}$$

có phân bố khi-bình phương với $(n-1)$ bậc tự do. Ở đây

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



Hình 3.8: Hàm mật độ của phân bố χ^2 với 2,3, và 5 bậc tự do

3.3.2 Kiểm định cho phương sai của phân bố chuẩn

Giả sử biến ngẫu nhiên X trong tổng thể (tập nền) có phân bố 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 ta có giả thiết thống kê

$$H_0 : \sigma^2 = \sigma_0^2. \tag{3.18}$$

Ta cần kiểm định giả thiết H_0. Xét mẫu ngẫu nhiên $(x_1, ..., x_n)$ tương ứng với biến ngẫu nhiên $W = (X_1, ..., X_n)$, ở đó $X_1, ..., X_n$ là các biến ngẫu nhiên có phân bố chuẩn và có cùng kỳ vọng và phương sai với X. Đặt $S^2 = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right]$. Theo Định lý 3.1, kiểm định thống kê

$$T = (n-1)\frac{S^2}{\sigma_0^2}$$

là biến ngẫu nhiên có phân bố khi-bình phương $(n-1)$ bậc tự do nếu giả thiết H_0 đúng. Ta xét các bài toán kiểm định sau.

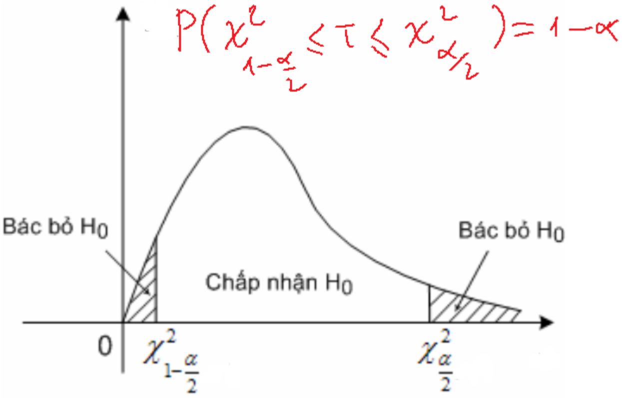
Bài toán 1

Bài toán *kiểm định hai phía*

$$\begin{cases} \text{Giả thiết } H_0 : \sigma^2 = \sigma_0^2, \\ \text{Đối thiết } H_1 : \sigma^2 \neq \sigma_0^2. \end{cases} \tag{3.19}$$

Miền bác bỏ (xem Hình 3.9)

$$W_\alpha = \left\{ T = (n-1)\frac{S^2}{\sigma_0^2} \,\middle|\, T < \chi^2_{1-\alpha/2} \text{ hoặc } T > \chi^2_{\alpha/2} \right\} \tag{3.20}$$



Hình 3.9: Miền bác bỏ W_α - phần gạch sọc

ở đó giá trị tới hạn χ^2_z được xác định bởi (xem bảng A10)

$$F(\chi^2_z) = 1 - z$$

với $F(x)$ là hàm phân bố của biến ngẫu nhiên T.

Bài toán 2

Bài toán *kiểm định một phía bên phải*

$$\begin{cases} \text{Giả thiết } H_0 : \sigma^2 = \sigma_0^2, \\ \text{Đối thiết } H_1 : \sigma^2 > \sigma_0^2. \end{cases} \tag{3.21}$$

Miền bác bỏ (xem Hình 3.10)

$$W_\alpha = \left\{ T = (n-1)\frac{S^2}{\sigma_0^2} \,\middle|\, T > \chi^2_\alpha \right\}. \tag{3.22}$$

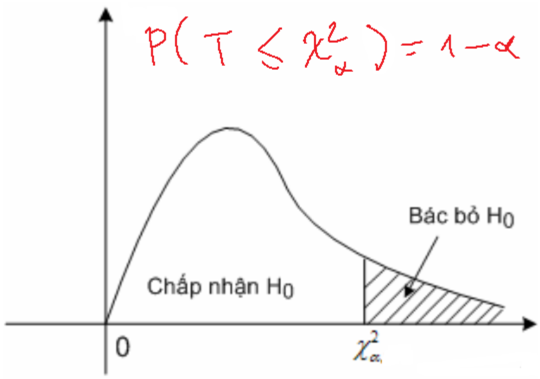
Bài toán 3

Bài toán *kiểm định một phía bên trái*

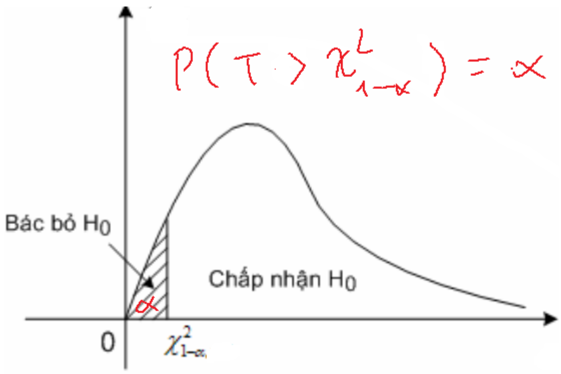
$$\begin{cases} \text{Giả thiết } H_0 : \sigma^2 = \sigma_0^2, \\ \text{Đối thiết } H_1 : \sigma^2 < \sigma_0^2. \end{cases} \tag{3.23}$$

Miền bác bỏ (xem Hình 3.11)

$$W_\alpha = \left\{ T = (n-1)\frac{S^2}{\sigma_0^2} \,\middle|\, T < \chi^2_{1-\alpha} \right\}. \tag{3.24}$$



Hình 3.10: Miền bác bỏ W_α - phần gạch sọc



Hình 3.11: Miền bác bỏ W_α - phần gạch sọc

Ví dụ 3.4. *Sử dụng mẫu ngẫu nhiên kích thước $n = 15$ và phương sai mẫu $s^2 = 13$ từ một tập nền có phân bố chuẩn, hãy kiểm định giả thiết $\sigma^2 = \sigma_0^2 = 10$ và đối thiết $\sigma^2 = \sigma_1^2 = 20$. Ở đây chọn mức ý nghĩa $\alpha = 5\%$.*

Giải. *Bài toán kiểm định bên phải*

$$\begin{cases} \textit{Giả thiết}\, H_0 : \sigma^2 = \sigma_0^2 = 10, \\ \textit{Đối thiết}\, H_1 : \sigma^2 = \sigma_1^2 > \sigma_0^2. \end{cases}$$

Giá trị tới hạn χ^2_α thỏa mãn

$$F(\chi^2_\alpha) = 1 - \alpha = 0.95.$$

Tra bảng A10 với $(n-1) = 14$ bậc tự do, ta có $\chi^2_\alpha = 23.68$ Miền bác bỏ

$$W_\alpha = \left\{ T = (n-1)\frac{S^2}{\sigma_0^2} \,\middle|\, T > 23.68 \right\}.$$

Giá trị quan sát

$$T_{qs} = (15-1) \times \frac{13}{10} = 18.2 \leq 23.68.$$

Do đó $T_{qs} \notin W_\alpha$ và vì vậy ta chấp nhận H_0.

3.4 So sánh kỳ vọng của hai phân phối chuẩn

Ta xét cùng lúc hai tổng thể. Tổng thể thứ nhất có dấu hiện nghiên cứu là biến ngẫu nhiên X với phân bố chuẩn $\mathcal{N}(\mu_X, \sigma_X^2)$. Tổng thể thứ hai có dấu hiện nghiên cứu là biến ngẫu nhiên Y với phân bố chuẩn $\mathcal{N}(\mu_Y, \sigma_Y^2)$. Giả sử rằng $\sigma_X = \sigma_Y$ (chúng có thể chưa biết.) Nếu có cơ sở để giả thiết rằng hai kỳ vọng μ_X và μ_Y là bằng nhau, ta sẽ kiểm định giả thiết thống kê

$$H_0 : \mu_X = \mu_Y.$$

Để kiểm định giả thiết trên, chúng ta sử dụng hai mẫu tương ứng với hai tổng thể: mẫu $(x_1, ..., x_{n_1})$ có kích thước n_1 ứng với biến ngẫu nhiên $(X_1, ..., X_{n_1})$ và mẫu $(y_1, ..., y_{n_2})$ có kích thước n_2 ứng với biến ngẫu nhiên $(Y_1, ..., Y_{n_2})$. Đặt

$$\bar{X} = \frac{1}{n_1}\left(X_1 + \cdots + X_{n_1}\right), \quad S_X^2 = \frac{1}{n_1 - 1}\left[(X_1 - \bar{X})^2 + \cdots + (X_{n_1} - \bar{X})^2\right]$$

$$\bar{Y} = \frac{1}{n_2}\left(Y_1 + \cdots + Y_{n_2}\right), \quad S_Y^2 = \frac{1}{n_2 - 1}\left[(Y_1 - \bar{Y})^2 + \cdots + (Y_{n_2} - \bar{Y})^2\right]$$

và

$$S = \sqrt{\frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1 + n_2 - 2}}.$$

Xét tiêu chuẩn kiểm định

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Người ta chứng minh được rằng, nếu giả thiết H_0 đúng thì T là biến ngẫu nhiên có phân bố Student $(n_1 + n_2 - 2)$ bậc tự do.

- Khi $n_1 = n_2 = n$, ta có

  $$T = \sqrt{n}\frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2 + S_Y^2}}$$

  và T là biến ngẫu nhiên có phân bố Student $(2n-2)$ bậc tự do khi giả thiết H_0 đúng.

Ta xét các bài toán kiểm định sau.

Bài toán 1

Bài toán *kiểm định hai phía*

$$\begin{cases} \text{Giả thiết } H_0: \mu_X = \mu_Y, \\ \text{Đối thiết } H_1: \mu_X \neq \mu_Y. \end{cases} \tag{3.25}$$

Miền bác bỏ

$$W_\alpha = \left\{ T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \,\middle|\, |T| > t_{\alpha/2} \right\}. \tag{3.26}$$

ở đó giá trị tới hạn $t_{\alpha/2}$ được xác định bởi (xem bảng A9)

$$F(t_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

với $F(x)$ là hàm phân bố của biến ngẫu nhiên T.

Bài toán 2

Bài toán *kiểm định một phía bên phải*

$$\begin{cases} \text{Giả thiết } H_0: \mu_X = \mu_Y, \\ \text{Đối thiết } H_1: \mu_X > \mu_Y. \end{cases} \tag{3.27}$$

Miền bác bỏ

$$W_\alpha = \left\{ T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \,\middle|\, T > t_\alpha \right\} \tag{3.28}$$

ở đó giá trị tới hạn t_α được xác định bởi (xem bảng A9)

$$F(t_\alpha) = 1 - \alpha.$$

Bài toán 3

Bài toán *kiểm định một phía bên trái*

$$\begin{cases} \text{Giả thiết } H_0: \mu_X = \mu_Y, \\ \text{Đối thiết } H_1: \mu_X < \mu_Y. \end{cases} \tag{3.29}$$

Miền bác bỏ

$$W_\alpha = \left\{ T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \,\middle|\, T < -t_\alpha \right\} \tag{3.30}$$

ở đó giá trị tới hạn t_α được xác định bởi (xem bảng A9)

$$F(t_\alpha) = 1 - \alpha.$$

Ví dụ 3.5. *Xét hai mẫu ngẫu nhiên thể hiện sản lượng của các công nhân sản xuất tấm thiếc trong hai môi trường làm việc khác nhau như sau*

X	*105*	*108*	*86*	*103*	*103*	*107*	*124*	*105*
Y	*89*	*92*	*84*	*97*	*103*	*107*	*111*	*97*

Giả thiết rằng hai tổng thể tương ứng có phân bố chuẩn có cùng phương sai. Với mức ý nghĩa 5%, *có thể nói sản lượng trung bình của các công nhân trong hai môi trường làm việc trên là như nhau hay không?*

Giải. *Bài toán kiểm định hai phía*

$$\begin{cases} \textit{Giả thiết}\, H_0 : \mu_X = \mu_Y, \\ \textit{Đối thiết}\, H_1 : \mu_X \neq \mu_Y. \end{cases}$$

Dựa vào giá trị của hai mẫu trên ta có

$$n_1 = n_2 = 8, \quad \bar{x} = 105.125, \quad \bar{y} = 97.500, \quad s_X^2 = 106.125, \quad s_Y^2 = 84.000.$$

Với mức ý nghĩa $\alpha = 5\%$, *từ bảng A9 ứng với* $(n_1 + n_2 - 2) = 14$ *bậc tự do và công thức*

$$F(t_{\alpha/2}) = 1 - \frac{\alpha}{2} = 0.975,$$

ta thu được $t_{\alpha/2} = 2.14$. *Do đó miền bác bỏ*

$$W_\alpha = \left\{ T = \sqrt{n}\frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2 + S_Y^2}} \,\middle|\, |T| > 2.14 \right\}$$

với $n = 8$. *Giá trị quan sát của tiêu chuẩn kiểm định* T *là*

$$T_{qs} = \sqrt{8}\frac{105.125 - 97.500}{\sqrt{106.125 + 84.000}} = 1.56.$$

Vì $|T_{qs}| \leq 2.14$ *nên* $T_{qs} \notin W_\alpha$ *và vì vậy ta chấp nhận* H_0. *Điều này có nghĩa rằng, trong hai môi trường làm việc trên, thì sản lượng trung bình tổng thể của hai tổng thể là bằng nhau.*

3.5 Kiểm định chất lượng: biểu đồ kiểm định cho trung bình (tùy chọn)

3.6 Kiểm định mức độ phù hợp của mô hình (kiểm định χ^2)

Kiểm định mức độ phù hợp của mô hình có nghĩa là thông qua một mẫu ngẫu nhiên $(x_1, ..., x_n)$ chúng ta kiểm định một hàm $F(x)$ có phải là hàm phân bố của biến ngẫu nhiên gốc X trong tổng thể hay không. Khi đó ta có giả thiết thống kê

$$H_0 : F(x) \text{ là hàm phân bố của } X$$

và đối thiết

$$H_1 : F(x) \text{ không là hàm phân bố của } X.$$

Để kiểm định mức độ phù hợp của mô hình trước tiên chúng ta định nghĩa *hàm phân bố mẫu* $\tilde{F}(x)$ như sau

$\tilde{F}(x) =$ Tổng của tần suất tương đối của tất cả các giá trị mẫu x_j với $x_j \leq x$.

Sau đó chúng ta kiểm tra liệu $\tilde{F}(x)$ có "đủ gần" (theo nghĩa nào đó) so với $F(x)$ hay không. Nếu $\tilde{F}(x)$ "đủ gần" so với $F(x)$, thì chúng ta chấp nhận giả thiết rằng $F(x)$ là hàm phân bố của tổng thể, ngược lại ta bác bỏ giả thiết đó.

Dưới đây là lược đồ kiểm định giả thiết rằng $F(x)$ là hàm phân bố của tổng thể thông qua một mẫu ngẫu nhiên $(x_1, ..., x_n)$.

Bước 1: Chia trục số thực thành K khoảng: $I_1, I_2, ..., I_K$ sao cho mỗi khoảng chứa ít nhất 5 giá trị của mẫu $(x_1, ..., x_n)$. Xác định *tần số thực nghiệm*

$$b_j = \text{Số các giá trị mẫu } x_1, x_2, ..., x_n \text{ nằm trong khoảng } I_j, \quad 1 \leq j \leq K.$$

Nếu một giá trị mẫu nào đó nằm trên phần giao của hai khoảng, thì cộng 0.5 cho hai số b_j tương ứng với hai khoảng đó.

Bước 2: Giả sử H_0 đúng, tức là $F(x)$ là hàm phân bố của X, ta tính được xác suất

$$p_j = P(X \in I_j), \quad 1 \leq j \leq K.$$

Từ đó tính *tần số lý thuyết*

$$e_j = np_j, \quad 1 \leq j \leq K.$$

Bước 3: Tính *độ lệch*

$$\chi_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}. \qquad (3.31)$$

Bước 4: Chọn mức ý nghĩa α (5%, 1%, hoặc giá trị khác).

Bước 5: Xác định giá trị tới hạn c từ phương trình

$$\hat{F}(c) = 1 - \alpha$$

và từ bảng A10. Ở đó $\hat{F}(x)$ là hàm phân bố của biến ngẫu nhiên χ^2 khi-bình phương với $K-1$ bậc tự do. Nếu r tham số của $F(x)$ chưa biết và phương pháp ước lượng hợp lý cực đại được sử dụng, thì $(K-r-1)$ bậc tự do sẽ thay thế cho $(K-1)$ bậc tự do.

Khi đó, ta có:

- Nếu $\chi_0^2 \leq c$, ta chấp nhận H_0.
- Nếu $\chi_0^2 > c$, ta bác bỏ H_0.

3.7 Kiểm định phi tham số (tự chọn)

Bài tập Chương 3

Bài 3.1. Kiểm định giả thiết $\mu = 0$ và đối thiết $\mu > 0$ sử dụng mẫu

$$0, \quad 1, \quad -1, \quad 3, \quad -8, \quad 6, \quad 1,$$

biết rằng biến ngẫu nhiên của tổng thể có phân bố chuẩn. Chọn mức ý nghĩa 5%.

Bài 3.2. Giả sử biến ngẫu nhiên của tổng thể có phân bố chuẩn với phương sai 9. Hãy kiểm định giả thiết $\mu = 60.0$ và đối thiết $\mu = 57.0$ sử dụng mẫu có kích thước 20 với trung bình mẫu $\bar{x} = 58.50$ và mức ý nghĩa $\alpha = 5\%$.

Bài 3.3. Kết luận của bài toán **Bài 3.2.** có thay đổi nếu chúng ta sử dụng một mẫu nhỏ có kích thước $n = 5$ với trung bình mẫu $\bar{x} = 58.05$ và mức ý nghĩa $\alpha = 5\%$.

Bài 3.4. Tìm miền bác bỏ trong bài toán **Bài 3.2.** trong trường hợp bài toán kiểm định hai phía với $\alpha = 5\%$.

Bài 3.5. Một công ty bán các can chứa dầu, theo quy định mỗi can chứa 5000 g dầu. Sử dụng mẫu ngẫu nhiên có kích thước 51 với trung bình 4990 g và độ lệch chuẩn 20 g để kiểm định sự sai khác của trọng lượng trung bình của lượng dầu do máy đong dầu đổ vào mỗi can so với khối lượng 5000 g theo quy định. Chọn mức ý nghĩa 5%. Biết lượng dầu do máy đong dầu đổ vào mỗi can là biến ngẫu nhiên có phân bố chuẩn.

Bài 3.6. Một mẫu ngẫu nhiên gồm 25 lốp xe có thời hạn sử dụng trung bình 37000 dặm (miles) và độ lệch chuẩn 5000 dặm. Nhà sản xuất khẳng định rằng thời hạn sử dụng trung bình của lốp xe là trên 35000 dặm. Với mức ý nghĩa 5%, liệu khẳng định của nhà sản xuất có được chấp nhận không? Biết thời hạn sử dụng của lốp xe là biến ngẫu nhiên có phân bố chuẩn.

Bài 3.7. Thực hiện nhiều lần đo đồng thời điện áp của một mạch điện bằng hai thiết bị vôn kế khác nhau, chúng ta thu được sự sai khác của hai bị vôn kế đó là

$$0.4, \quad -0.6, \quad 0.2 \quad 0.0, \quad 1.0, \quad 1.4 \quad ,0.4 \quad 1.6 \quad (V).$$

Với mức ý nghĩa 5%, liệu chúng ta có thể khẳng định rằng sự sai khác về kết quả đo điện áp của hai vôn kế đó là không đáng kể? Biết rằng sự sai khác về kết quả đo điệp áp của hai vôn kế tuân theo quy luật phân bố chuẩn.

Bài 3.8. Trọng lượng đóng bao của một loại sản phẩm A là biến ngẫu nhiên có phân bố chuẩn với trọng lượng trung bình theo quy định là 100 kg. Nghi ngờ sản phẩm bị đóng thiếu, người ta cân thử 29 bao loại này và thu được kết quả như sau

Trọng lượng (kg)	98.0 − 98.5	98.5 − 99.0	99.0 − 99.5	99.5 − 100.0	100.0 − 100.5
Số bao	2	6	10	7	4

Với mức ý nghĩa 5%, hãy kết luận về điều nghi ngờ trên.

Bài 3.9. Mức tiêu thụ xăng của một loại ô tô chạy trên một đoạn đường từ thành phố A tới thành phố B là một biến ngẫu nhiêu có phân bố chuẩn với kỳ vọng 50 lít. Người ta mới sửa lại đoạn đường từ A tới B và cho rằng mức tiêu thụ xăng trung bình đã giảm xuống. Quan sát 28 ô tô cùng loại thu được

Mức tiêu thụ (lít)	48.5 − 49.0	49.0 − 49.5	49.5 − 50.0	50.0 − 50.5	50.5 − 51.0
Số ô tô	4	10	9	3	2

Với mức ý nghĩa 2.5%, hãy kết luận về điều nghi ngờ trên.

Bài 3.10. Giả sử trong quá khứ độ lệch chuẩn của khối lượng các kiện hàng được đóng gói bởi các máy tự động là 0.8 oz (1 oz $\approx$ 0.028 kg). Hãy kiểm định giả thiết $H_0 : \sigma = 0.8$ oz và đối thiết $H_1 : \sigma > 0.8$ oz (trong trường hợp độ lệch chuẩn của khối lượng các kiện hàng tăng lên) sử dụng mẫu có kích thước 20 với độ lệch chuẩn 1.0 oz. Biết rằng khối lượng của kiện hàng là biến ngẫu nhiên có phân bố chuẩn. Chọn mức ý nghĩa $\alpha = 5\%$.

Bài 3.11. Có giả thiết cho rằng với thiết bị chạy bằng hệ thống pin nếu độ lệch chuẩn của tuổi thọ của pin nhỏ hơn một con số cố định, chẳng hạn 5 giờ thì chi phí sẽ tiết kiệm hơn khi chúng ta thay tất cả pin sau một thời gian cố định nào đó so với việc chúng ta thay riêng lẻ từng pin khi nó bị hỏng. Với mức ý nghĩa 5%, hãy kiểm định giả thiết trên thông qua mẫu có kích thước $n = 28$ với độ lệch chuẩn mẫu là $s = 3.5$ giờ. Biết tuổi thọ của pin là biến ngẫu nhiên có phân bố chuẩn.

Bài 3.12. Nếu máy móc hoạt động bình thường thì trọng lượng sản phẩm tuân theo quy luật chuẩn với độ lệch chuẩn là 1 kg. Có thể coi máy móc còn hoạt động tốt hay không nếu cân thử 30 sản phẩm ta thấy độ lệch chuẩn mẫu là 1.1 kg. Hãy kiểm định giả thiết trên với mức ý nghĩa 0.01.

Bài 3.13. Xăng nhãn hiệu A được sử dụng cho 16 xe ô tô cùng loại và thu được một mẫu gồm 16 giá trị về hiệu suất sử dụng của xăng nhẵn hiệu A có giá trị trung bình là 19.6 (số dặm đi được/1 gallon) và độ lệch chuẩn 0.4 (số dặm đi được/1 gallon). Trong cùng điều kiện tương tự, xăng nhãn hiệu B sử dụng cho 16 xe ô tô và thu được một mẫu có giá trị trung bình 20.2 (số dặm đi được/1 gallon) và độ lệch chuẩn 0.6 (số dặm đi được/1 gallon). Liệu hiệu suất sử dụng của xăng nhãn hiệu B tốt hơn xăng nhãn hiệu A? Hãy kiểm định điều đó. Biết rằng hiệu suất sử dụng (số dặm đi được/1 gallon) của hai loại xăng trên là biến ngẫu nhiên có phân bố chuẩn có cùng phương sai. Chọn mức ý nghĩa 5%.

Bài 3.14. Có hai loại bi bằng thép: loại I và loại II. Loại I có đường kính X là biến ngẫu nhiên có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma^2)$. Loại II có đường kính Y là biến ngẫu nhiên có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma^2)$. Người ta cho rằng đường kính trung bình của hai loại vòng bi này bằng nhau. Lấy ngẫu nhiên 10 viên bi loại I và 10 viên bi loại II. Ta thu được mẫu:

X	2.66	2.68	2.63	2.60	2.67	2.59	2.62	2.61
Y	2.63	2.60	2.57	2.56	2.58	2.58	2.61	2.62

Hãy kiểm định giả thiết trên với mức ý nghĩa 5%.

Bài 3.15. Để so sánh trọng lượng trung bình của trẻ sơ sinh thành thị và nông thôn, người ta theo dõi 1000 cháu và thu được kết quả

Kết quả	Số cháu	Trung bình mẫu	Độ lệch chuẩn mẫu
Nông thôn	500	3.0 kg	0.4 kg
Thành thị	500	3.2 kg	0.3 kg

Với mức ý nghĩa 5%, có thể coi trọng lượng trẻ sơ sinh ở thành thị cao hơn nông thôn hay không?

Chương 4

Tương quan và hồi quy

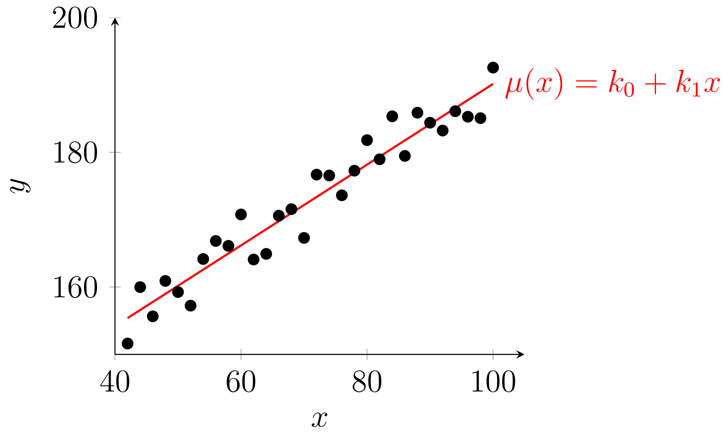
Trong chương này chúng ta thảo luận về các thực nghiệm liên quan tới việc quan sát hoặc đo lường đồng thời hai đại lượng X và Y. Do đó chúng ta làm việc với mẫu ngẫu nhiên mà các giá trị được cho bởi các cặp (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Nội dung chính của chương này như sau:

1. Trong **phân tích hồi quy (regression analysis)**, đại lượng X có thể được đo lường với độ chính xác cao và được coi như biến cơ sở. Ta gọi X là *biến độc lập (independent variable)* hay *biến điều khiển (control variable)*. Đại lượng Y còn lại là một biến ngẫu nhiên và chúng ta quan tâm tới sự phụ thuộc của Y vào X. Chẳng hạn chúng ta quan tâm tới: sự phụ thuộc của huyết áp Y của một người vào độ tuổi X của người đó; mối liên hệ giữa lượng cân tăng Y của một loài động vật nào đó phụ thuộc vào chế độ dinh dưỡng hàng ngày X;...

2. Trong **phân tích tương quan (correlation analysis)**, chúng ta quan tâm tới mối quan hệ (mối tương quan) giữa hai đại lượng ngẫu nhiên X và Y. Chẳng hạn, chúng ta nghiên cứu sự tương quan độ mòn X của lốp trước bên trái của một xe ô tô so với độ mòn Y của lốp trước bên phải; quan hệ giữa điểm số môn toán X và điểm số môn vật lý Y của một học sinh lớp 12;...

4.1 Hồi quy tuyến tính đơn: đường hồi quy tuyến tính mẫu, hệ số xác định

Trong phân tích hồi quy, sự phụ thuộc của đại lượng Y vào đại lượng $X = x$ giống như sự phụ thuộc của kỳ vọng μ của Y vào x, tức là $\mu = \mu(x)$. Đồ thị



Hình 4.1: Đường hồi quy tuyến tính tương ứng với đám mây điểm

của hàm $\mu(x)$ được gọi là *đường hồi quy (regression curve)* của Y. Khi

$$\mu(x) = \kappa_0 + \kappa_1 x, \tag{4.1}$$

thì đường hồi quy của Y là đường thẳng và được gọi là *đường hồi quy tuyến tính (regression line)*.

Cho mẫu ngẫu nhiên gồm n cặp giá trị: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) tương ứng với một đám mây điểm trên mặt phẳng OxY. Chúng ta muốn xây dựng một đường thẳng "phù hợp" đi xuyên qua đám mây điểm trên để từ đó ta có thể xấp xỉ giá trị của $\mu(x)$ tại một giá trị cụ thể của x. Đồ thị của đường hồi quy tuyến tính và đám mây điểm được thể hiện trong Hình 4.1.

Để tìm được phương trình của đường hồi quy tuyến tính, ta dùng *nguyên lý bình phương tối thiểu (least square principle)* sau:

- *Đường hồi quy tuyến tính đi xuyên qua đám mây điểm sao cho tổng của bình phương khoảng cách từ các điểm tới đường hồi quy là nhỏ nhất.*

Muốn đường hồi quy tuyến tính là duy nhất, ta cần giả thiết sau.

(A1) *Có ít nhất hai số có giá trị khác nhau trong các số* $x_1, x_2, \ldots, x_n$.

• **Xác định đường hồi quy tuyến tính dựa trên giả thiết (A1)**.

Giả sử ta có đám mây điểm $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ thỏa mãn giả thiết (A1). Ta gọi đường thẳng

$$y = k_0 + k_1 x \tag{4.2}$$

là **đường hồi quy (tuyến tính) mẫu**.

Khoảng cách đứng từ điểm (x_i, y_i) đến đường hồi quy mẫu là

$$\epsilon_i = |y_i - k_0 - k_1 x_i|.$$

Tổng bình phương các khoảng cách đó là

$$q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - k_0 - k_1 x_i)^2.$$

Để q bé nhất, ta áp dụng phương pháp cực trị của hàm hai biến số k_0 và k_1, có nghĩa k_0, k_1 thỏa mãn hệ phương trình

$$\begin{cases} \frac{\partial q}{\partial k_0} = 0 \\ \frac{\partial q}{\partial k_1} = 0 \end{cases}$$

hay

$$\begin{cases} nk_0 + k_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ k_0 \sum_{i=1}^n x_i + k_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases} \tag{4.3}$$

Phương trình (4.3) được gọi là *hệ phương trình chuẩn* với hai ẩn số k_0, k_1. Định thức của ma trận các hệ số là

$$\begin{aligned} \begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} &= n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \\ &= n(n-1)s_x^2 = n \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0 \end{aligned}$$

với

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

Theo Định lý Cramer, ta có

$$k_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n(n-1)s_x^2} = \frac{n(\overline{xy} - \bar{x}\bar{y})}{(n-1)s_x^2} \tag{4.4}$$

với

$$\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n), \quad \overline{xy} = \frac{1}{n}(x_1 y_1 + \cdots + x_n y_n).$$

Đại lượng s_{xy} được gọi là **hiệp phương sai mẫu (sample covariance)** và được cho bởi

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right) \tag{4.5}$$

hay

$$s_{xy} = \frac{n}{n-1}[\overline{xy} - \bar{x}\bar{y}]. \tag{4.6}$$

Từ (4.4), ta suy ra

$$k_1 = \frac{s_{xy}}{s_x^2}. \tag{4.7}$$

Từ phương trình thứ nhất của (4.3), hệ số k_0 được tính như sau

$$k_0 = \bar{y} - k_1\bar{x}.$$

Thay vào phương trình (4.2), ta có

$$y - \bar{y} = k_1(x - \bar{x}).$$

Do đó hệ số k_1 được cho bởi công thức (4.7) còn được gọi là **hệ số hồi quy (regression coefficient)** hay **hệ số xác định**.

Ví dụ 4.1. *Đo sự giảm thể tích $y[\%]$ của một tấm da dưới áp lực x [atmospheres] và nhận được giá trị sau*

x_i	*4000*	*6000*	*8000*	*10000*
y_i	*2.3*	*4.1*	*5.7*	*6.9*

Tìm đường hồi quy tuyến tính của y theo x.

Giải. *Ta có $n = 4$, $\bar{x} = 7000$, $\bar{y} = 4.75$ và*

$$s_x^2 = \frac{20000000}{3}, \quad s_{xy} = \frac{15400}{3}.$$

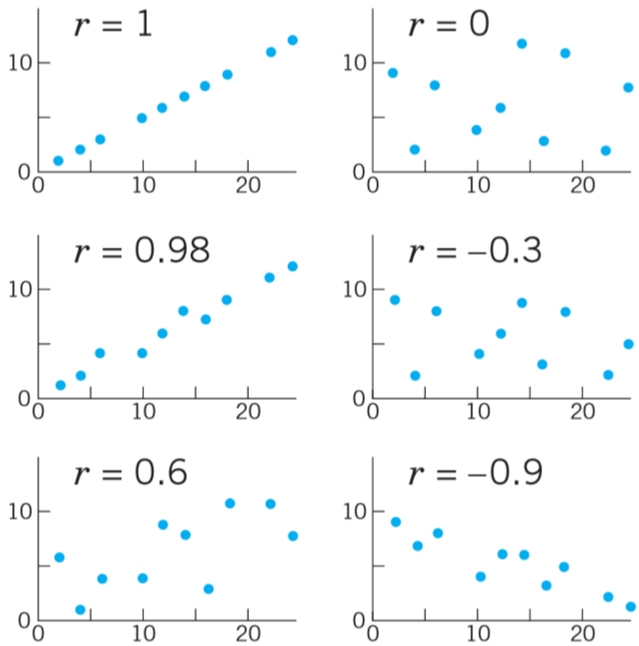
Do đó hệ số xác định $k_1 = \frac{s_{xy}}{s_x^2} = 0.00077$ và đường hồi quy cần tìm là

$$y - 4.74 = 0.00077(x - 7000) \quad hay \quad y = 0.00077x - 0.64.$$

4.2 Tương quan: hiệp phương sai, hệ số tương quan, hệ số tương quan mẫu

Phân tích tương quan liên quan tới mối liên hệ giữa hai biến ngẫu nhiên X và Y. Cho mẫu với n cặp giá trị (x_1, y_1), ..., (x_n, y_n). Mối quan hệ giữa các giá trị x_i và y_i $(1 \le i \le n)$ được thể hiện thông qua hiệp phương sai mẫu s_{xy} hay thông qua **hệ số tương quan mẫu (sample correlation coefficient)**

$$r = \frac{s_{xy}}{s_x s_y} \tag{4.8}$$



Hình 4.2: Mẫu với các giá trị của hệ số tương quan mẫu r

với s_{xy} được cho bởi (4.5) và

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n}{n-1}[\overline{x^2}-(\bar{x})^2],$$
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{x})^2 = \frac{n}{n-1}[\overline{y^2}-(\bar{y})^2].$$

Ta có định lý sau liên quan tới hệ số tương quan mẫu.

Định lý 4.1 (Hệ số tương quan mẫu)**.** *Hệ số tương quan mẫu r có các tính chất sau:*

(i) $-1 \le r \le 1$*;*

(ii) $r = \pm 1$ *nếu và chỉ nếu tất cả các điểm của mẫu nằm trên một đường thẳng.*

Một số giá trị của hệ số tương quan mẫu r được minh họa trong Hình 4.2.

Cho hai biến ngẫu nhiên X và Y.

- Đại lượng

$$\sigma_{XY} = E([X - \mu_X][Y - \mu_Y]) = E(XY) - E(X)E(Y) \tag{4.9}$$

được gọi là **hiệp phương sai (covariance)** của X và Y. Trong đó $\mu_X = E(X)$ và $\mu_Y = E(Y)$ tương ứng là kỳ vọng của X và Y.

- Đại lượng

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{4.10}$$

được gọi là **hệ số tương quan (theoretical correlation coefficient)** của X và Y. Trong đó $\sigma_X^2 = var(X) = E((X - EX)^2)$ và $\sigma_Y^2 = var(Y) = E((Y - EY)^2)$ tương ứng là phương sai của X và Y.

Tương tự Định lý 4.1, ta cũng có định lý sau về mối liên hệ giữa hệ số tương quan và độ phụ thuộc tuyến tính của X và Y.

Định lý 4.2 (Hệ số tương quan)**.** *Hệ số tương quan ρ của hai biến ngẫu nghiên X và Y có các tính chất sau:*

(i) $-1 \leq \rho \leq 1$*;*

(ii) $\rho = \pm 1$ *nếu và chỉ nếu X và Y là phụ thuộc tuyến tính, có nghĩa là*

$$\begin{cases} Y = \gamma X + \delta, \\ X = \gamma^* Y + \delta^*. \end{cases}$$

Bài tập Chương 4

Tính hệ số tương quan mẫu, hệ số xác định (hệ số hồi quy) và phương trình đường hồi quy tuyến tính của các mẫu sau.

Bài 4.1. $(0, 0.1), (2, 2.1), (4, 2.9), (6, 3.6), (8, 5.2)$.

Bài 4.2. $(-2, 3.5)$, $(1, 2.6)$, $(3, 3.1)$, $(5, 0.4)$.

Bài 4.3. $(1.1, 40)$, $(3.2, 65)$, $(3.4, 120)$, $(4.5, 150)$, $(5.6, 190)$.

Bài 4.4. x = số vòng quay/phút, y = công suất động cơ Diesel

x	400	500	600	700	750
y	5800	10300	14200	18800	21000

Bài 4.5. x = biến dạng của một loại thép (mm), y = độ cứng Brinell (kg/mm^2)

x	6	9	11	13	22	26	28	33	35
y	68	67	65	53	44	40	37	34	32

Bài 4.6. x = độ cứng Brinell (kg/mm^2), y = sức căng (đơn vị 1000 psi) của một loại thép

x	200	300	400	500
y	110	150	190	280

Bài 4.7. (Định luật Ohm) x = điện áp (V), y = cường độ dòng điện (A). Tìm biến trở R (Ω).

x	40	40	80	80	110	110
y	5.1	4.8	10.0	10.3	13.0	12.7

Bài 4.8. Sự dẫn nhiệt của nước. x = nhiệt độ (^oF), y = độ dẫn nhiệt ($Btu/(hr \cdot ft \cdot ^o F)$). Tìm y tại nhiệt độ 66^oF.

x	32	50	100	150	212
y	0.337	0.345	0.365	0.380	0.395

Bài 4.9. Khoảng cách dừng của ô tô. x = vận tốc (mph), y = khoảng cách dừng (ft). Tìm y tại vận tốc 35mph

x	30	40	50	60
y	160	240	330	435

Bài 4.10. x = độ ẩm của không khí (%), y = sự nở của gelatin (%) (gelatin = một chất gây đông trong thực phẩm). Tìm y tại $x = 25\%$.

x	10	20	30	40
y	0.8	1.6	2.3	2.8

APPENDIX 5

Tables

For Tables of Laplace Transforms see Secs. 6.8 and 6.9.
For Tables of Fourier Transforms see Sec. 11.10.

If you have a Computer Algebra System (CAS), you may not need the present tables, but you may still find them convenient from time to time.

Table A1 Bessel Functions

For more extensive tables see Ref. [GenRef1] in App. 1.

x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$
0.0	1.0000	0.0000	3.0	−0.2601	0.3391	6.0	0.1506	−0.2767
0.1	0.9975	0.0499	3.1	−0.2921	0.3009	6.1	0.1773	−0.2559
0.2	0.9900	0.0995	3.2	−0.3202	0.2613	6.2	0.2017	−0.2329
0.3	0.9776	0.1483	3.3	−0.3443	0.2207	6.3	0.2238	−0.2081
0.4	0.9604	0.1960	3.4	−0.3643	0.1792	6.4	0.2433	−0.1816
0.5	0.9385	0.2423	3.5	−0.3801	0.1374	6.5	0.2601	−0.1538
0.6	0.9120	0.2867	3.6	−0.3918	0.0955	6.6	0.2740	−0.1250
0.7	0.8812	0.3290	3.7	−0.3992	0.0538	6.7	0.2851	−0.0953
0.8	0.8463	0.3688	3.8	−0.4026	0.0128	6.8	0.2931	−0.0652
0.9	0.8075	0.4059	3.9	−0.4018	−0.0272	6.9	0.2981	−0.0349
1.0	0.7652	0.4401	4.0	−0.3971	−0.0660	7.0	0.3001	−0.0047
1.1	0.7196	0.4709	4.1	−0.3887	−0.1033	7.1	0.2991	0.0252
1.2	0.6711	0.4983	4.2	−0.3766	−0.1386	7.2	0.2951	0.0543
1.3	0.6201	0.5220	4.3	−0.3610	−0.1719	7.3	0.2882	0.0826
1.4	0.5669	0.5419	4.4	−0.3423	−0.2028	7.4	0.2786	0.1096
1.5	0.5118	0.5579	4.5	−0.3205	−0.2311	7.5	0.2663	0.1352
1.6	0.4554	0.5699	4.6	−0.2961	−0.2566	7.6	0.2516	0.1592
1.7	0.3980	0.5778	4.7	−0.2693	−0.2791	7.7	0.2346	0.1813
1.8	0.3400	0.5815	4.8	−0.2404	−0.2985	7.8	0.2154	0.2014
1.9	0.2818	0.5812	4.9	−0.2097	−0.3147	7.9	0.1944	0.2192
2.0	0.2239	0.5767	5.0	−0.1776	−0.3276	8.0	0.1717	0.2346
2.1	0.1666	0.5683	5.1	−0.1443	−0.3371	8.1	0.1475	0.2476
2.2	0.1104	0.5560	5.2	−0.1103	−0.3432	8.2	0.1222	0.2580
2.3	0.0555	0.5399	5.3	−0.0758	−0.3460	8.3	0.0960	0.2657
2.4	0.0025	0.5202	5.4	−0.0412	−0.3453	8.4	0.0692	0.2708
2.5	−0.0484	0.4971	5.5	−0.0068	−0.3414	8.5	0.0419	0.2731
2.6	−0.0968	0.4708	5.6	0.0270	−0.3343	8.6	0.0146	0.2728
2.7	−0.1424	0.4416	5.7	0.0599	−0.3241	8.7	−0.0125	0.2697
2.8	−0.1850	0.4097	5.8	0.0917	−0.3110	8.8	−0.0392	0.2641
2.9	−0.2243	0.3754	5.9	0.1220	−0.2951	8.9	−0.0653	0.2559

$J_0(x) = 0$ for $x = 2.40483, 5.52008, 8.65373, 11.7915, 14.9309, 18.0711, 21.2116, 24.3525, 27.4935, 30.6346$

$J_1(x) = 0$ for $x = 3.83171, 7.01559, 10.1735, 13.3237, 16.4706, 19.6159, 22.7601, 25.9037, 29.0468, 32.1897$

Table A1 ***(continued)***

x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$
0.0	$(-\infty)$	$(-\infty)$	2.5	0.498	0.146	5.0	−0.309	0.148
0.5	−0.445	−1.471	3.0	0.377	0.325	5.5	−0.339	−0.024
1.0	0.088	−0.781	3.5	0.189	0.410	6.0	−0.288	−0.175
1.5	0.382	−0.412	4.0	−0.017	0.398	6.5	−0.173	−0.274
2.0	0.510	−0.107	4.5	−0.195	0.301	7.0	−0.026	−0.303

Table A2 **Gamma Function** [see (24) in App. A3.1]

α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$
1.00	1.000 000	1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384
1.02	0.988 844	1.22	0.913 106	1.42	0.886 356	1.62	0.895 924	1.82	0.936 845
1.04	0.978 438	1.24	0.908 521	1.44	0.885 805	1.64	0.898 642	1.84	0.942 612
1.06	0.968 744	1.26	0.904 397	1.46	0.885 604	1.66	0.901 668	1.86	0.948 687
1.08	0.959 725	1.28	0.900 718	1.48	0.885 747	1.68	0.905 001	1.88	0.955 071
1.10	0.951 351	1.30	0.897 471	1.50	0.886 227	1.70	0.908 639	1.90	0.961 766
1.12	0.943 590	1.32	0.894 640	1.52	0.887 039	1.72	0.912 581	1.92	0.968 774
1.14	0.936 416	1.34	0.892 216	1.54	0.888 178	1.74	0.916 826	1.94	0.976 099
1.16	0.929 803	1.36	0.890 185	1.56	0.889 639	1.76	0.921 375	1.96	0.983 743
1.18	0.923 728	1.38	0.888 537	1.58	0.891 420	1.78	0.926 227	1.98	0.991 708
1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384	2.00	1.000 000

Table A3 **Factorial Function and Its Logarithm with Base 10**

n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$
1	1	0.000 000	6	720	2.857 332	11	39 916 800	7.601 156
2	2	0.301 030	7	5 040	3.702 431	12	479 001 600	8.680 337
3	6	0.778 151	8	40 320	4.605 521	13	6 227 020 800	9.794 280
4	24	1.380 211	9	362 880	5.559 763	14	87 178 291 200	10.940 408
5	120	2.079 181	10	3 628 800	6.559 763	15	1 307 674 368 000	12.116 500

Table A4 **Error Function, Sine and Cosine Integrals** [see (35), (40), (42) in App. A3.1]

x	erf x	Si(x)	ci(x)	x	erf x	Si(x)	ci(x)
0.0	0.0000	0.0000	∞	2.0	0.9953	1.6054	−0.4230
0.2	0.2227	0.1996	1.0422	2.2	0.9981	1.6876	−0.3751
0.4	0.4284	0.3965	0.3788	2.4	0.9993	1.7525	−0.3173
0.6	0.6039	0.5881	0.0223	2.6	0.9998	1.8004	−0.2533
0.8	0.7421	0.7721	−0.1983	2.8	0.9999	1.8321	−0.1865
1.0	0.8427	0.9461	−0.3374	3.0	1.0000	1.8487	−0.1196
1.2	0.9103	1.1080	−0.4205	3.2	1.0000	1.8514	−0.0553
1.4	0.9523	1.2562	−0.4620	3.4	1.0000	1.8419	0.0045
1.6	0.9763	1.3892	−0.4717	3.6	1.0000	1.8219	0.0580
1.8	0.9891	1.5058	−0.4568	3.8	1.0000	1.7934	0.1038
2.0	0.9953	1.6054	−0.4230	4.0	1.0000	1.7582	0.1410

Table A5 Binomial Distribution

Probability function $f(x)$ [see (2), Sec. 24.7] and distribution function $F(x)$

n	x	$p = 0.1$		$p = 0.2$		$p = 0.3$		$p = 0.4$		$p = 0.5$	
		$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
		0.		**0.**		**0.**		**0.**		**0.**	
1	0	9000	0.9000	8000	0.8000	7000	0.7000	6000	0.6000	5000	0.5000
	1	1000	1.0000	2000	1.0000	3000	1.0000	4000	1.0000	5000	1.0000
2	0	8100	0.8100	6400	0.6400	4900	0.4900	3600	0.3600	2500	0.2500
	1	1800	0.9900	3200	0.9600	4200	0.9100	4800	0.8400	5000	0.7500
	2	0100	1.0000	0400	1.0000	0900	1.0000	1600	1.0000	2500	1.0000
3	0	7290	0.7290	5120	0.5120	3430	0.3430	2160	0.2160	1250	0.1250
	1	2430	0.9720	3840	0.8960	4410	0.7840	4320	0.6480	3750	0.5000
	2	0270	0.9990	0960	0.9920	1890	0.9730	2880	0.9360	3750	0.8750
	3	0010	1.0000	0080	1.0000	0270	1.0000	0640	1.0000	1250	1.0000
4	0	6561	0.6561	4096	0.4096	2401	0.2401	1296	0.1296	0625	0.0625
	1	2916	0.9477	4096	0.8192	4116	0.6517	3456	0.4752	2500	0.3125
	2	0486	0.9963	1536	0.9728	2646	0.9163	3456	0.8208	3750	0.6875
	3	0036	0.9999	0256	0.9984	0756	0.9919	1536	0.9744	2500	0.9375
	4	0001	1.0000	0016	1.0000	0081	1.0000	0256	1.0000	0625	1.0000
5	0	5905	0.5905	3277	0.3277	1681	0.1681	0778	0.0778	0313	0.0313
	1	3281	0.9185	4096	0.7373	3602	0.5282	2592	0.3370	1563	0.1875
	2	0729	0.9914	2048	0.9421	3087	0.8369	3456	0.6826	3125	0.5000
	3	0081	0.9995	0512	0.9933	1323	0.9692	2304	0.9130	3125	0.8125
	4	0005	1.0000	0064	0.9997	0284	0.9976	0768	0.9898	1563	0.9688
	5	0000	1.0000	0003	1.0000	0024	1.0000	0102	1.0000	0313	1.0000
6	0	5314	0.5314	2621	0.2621	1176	0.1176	0467	0.0467	0156	0.0156
	1	3543	0.8857	3932	0.6554	3025	0.4202	1866	0.2333	0938	0.1094
	2	0984	0.9841	2458	0.9011	3241	0.7443	3110	0.5443	2344	0.3438
	3	0146	0.9987	0819	0.9830	1852	0.9295	2765	0.8208	3125	0.6563
	4	0012	0.9999	0154	0.9984	0595	0.9891	1382	0.9590	2344	0.8906
	5	0001	1.0000	0015	0.9999	0102	0.9993	0369	0.9959	0938	0.9844
	6	0000	1.0000	0001	1.0000	0007	1.0000	0041	1.0000	0156	1.0000
7	0	4783	0.4783	2097	0.2097	0824	0.0824	0280	0.0280	0078	0.0078
	1	3720	0.8503	3670	0.5767	2471	0.3294	1306	0.1586	0547	0.0625
	2	1240	0.9743	2753	0.8520	3177	0.6471	2613	0.4199	1641	0.2266
	3	0230	0.9973	1147	0.9667	2269	0.8740	2903	0.7102	2734	0.5000
	4	0026	0.9998	0287	0.9953	0972	0.9712	1935	0.9037	2734	0.7734
	5	0002	1.0000	0043	0.9996	0250	0.9962	0774	0.9812	1641	0.9375
	6	0000	1.0000	0004	1.0000	0036	0.9998	0172	0.9984	0547	0.9922
	7	0000	1.0000	0000	1.0000	0002	1.0000	0016	1.0000	0078	1.0000
8	0	4305	0.4305	1678	0.1678	0576	0.0576	0168	0.0168	0039	0.0039
	1	3826	0.8131	3355	0.5033	1977	0.2553	0896	0.1064	0313	0.0352
	2	1488	0.9619	2936	0.7969	2965	0.5518	2090	0.3154	1094	0.1445
	3	0331	0.9950	1468	0.9437	2541	0.8059	2787	0.5941	2188	0.3633
	4	0046	0.9996	0459	0.9896	1361	0.9420	2322	0.8263	2734	0.6367
	5	0004	1.0000	0092	0.9988	0467	0.9887	1239	0.9502	2188	0.8555
	6	0000	1.0000	0011	0.9999	0100	0.9987	0413	0.9915	1094	0.9648
	7	0000	1.0000	0001	1.0000	0012	0.9999	0079	0.9993	0313	0.9961
	8	0000	1.0000	0000	1.0000	0001	1.0000	0007	1.0000	0039	1.0000

Table A6 **Poisson Distribution**

Probability function $f(x)$ [see (5), Sec. 24.7] and distribution function $F(x)$

x	$\mu = 0.1$		$\mu = 0.2$		$\mu = 0.3$		$\mu = 0.4$		$\mu = 0.5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	9048	0.9048	8187	0.8187	7408	0.7408	6703	0.6703	6065	0.6065
1	0905	0.9953	1637	0.9825	2222	0.9631	2681	0.9384	3033	0.9098
2	0045	0.9998	0164	0.9989	0333	0.9964	0536	0.9921	0758	0.9856
3	0002	1.0000	0011	0.9999	0033	0.9997	0072	0.9992	0126	0.9982
4	0000	1.0000	0001	1.0000	0003	1.0000	0007	0.9999	0016	0.9998
5							0001	1.0000	0002	1.0000

x	$\mu = 0.6$		$\mu = 0.7$		$\mu = 0.8$		$\mu = 0.9$		$\mu = 1$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	5488	0.5488	4966	0.4966	4493	0.4493	4066	0.4066	3679	0.3679
1	3293	0.8781	3476	0.8442	3595	0.8088	3659	0.7725	3679	0.7358
2	0988	0.9769	1217	0.9659	1438	0.9526	1647	0.9371	1839	0.9197
3	0198	0.9966	0284	0.9942	0383	0.9909	0494	0.9865	0613	0.9810
4	0030	0.9996	0050	0.9992	0077	0.9986	0111	0.9977	0153	0.9963
5	0004	1.0000	0007	0.9999	0012	0.9998	0020	0.9997	0031	0.9994
6			0001	1.0000	0002	1.0000	0003	1.0000	0005	0.9999
7									0001	1.0000

x	$\mu = 1.5$		$\mu = 2$		$\mu = 3$		$\mu = 4$		$\mu = 5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	2231	0.2231	1353	0.1353	0498	0.0498	0183	0.0183	0067	0.0067
1	3347	0.5578	2707	0.4060	1494	0.1991	0733	0.0916	0337	0.0404
2	2510	0.8088	2707	0.6767	2240	0.4232	1465	0.2381	0842	0.1247
3	1255	0.9344	1804	0.8571	2240	0.6472	1954	0.4335	1404	0.2650
4	0471	0.9814	0902	0.9473	1680	0.8153	1954	0.6288	1755	0.4405
5	0141	0.9955	0361	0.9834	1008	0.9161	1563	0.7851	1755	0.6160
6	0035	0.9991	0120	0.9955	0504	0.9665	1042	0.8893	1462	0.7622
7	0008	0.9998	0034	0.9989	0216	0.9881	0595	0.9489	1044	0.8666
8	0001	1.0000	0009	0.9998	0081	0.9962	0298	0.9786	0653	0.9319
9			0002	1.0000	0027	0.9989	0132	0.9919	0363	0.9682
10					0008	0.9997	0053	0.9972	0181	0.9863
11					0002	0.9999	0019	0.9991	0082	0.9945
12					0001	1.0000	0006	0.9997	0034	0.9980
13							0002	0.9999	0013	0.9993
14							0001	1.0000	0005	0.9998
15									0002	0.9999
16									0000	1.0000

Table A7 Normal Distribution

Values of the distribution function $\Phi(z)$ [see (3), Sec. 24.8]. $\Phi(-z) = 1 - \Phi(z)$

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
	0.		**0.**		**0.**		**0.**		**0.**		**0.**
0.01	5040	0.51	6950	1.01	8438	1.51	9345	2.01	9778	2.51	9940
0.02	5080	0.52	6985	1.02	8461	1.52	9357	2.02	9783	2.52	9941
0.03	5120	0.53	7019	1.03	8485	1.53	9370	2.03	9788	2.53	9943
0.04	5160	0.54	7054	1.04	8508	1.54	9382	2.04	9793	2.54	9945
0.05	5199	0.55	7088	1.05	8531	1.55	9394	2.05	9798	2.55	9946
0.06	5239	0.56	7123	1.06	8554	1.56	9406	2.06	9803	2.56	9948
0.07	5279	0.57	7157	1.07	8577	1.57	9418	2.07	9808	2.57	9949
0.08	5319	0.58	7190	1.08	8599	1.58	9429	2.08	9812	2.58	9951
0.09	5359	0.59	7224	1.09	8621	1.59	9441	2.09	9817	2.59	9952
0.10	5398	0.60	7257	1.10	8643	1.60	9452	2.10	9821	2.60	9953
0.11	5438	0.61	7291	1.11	8665	1.61	9463	2.11	9826	2.61	9955
0.12	5478	0.62	7324	1.12	8686	1.62	9474	2.12	9830	2.62	9956
0.13	5517	0.63	7357	1.13	8708	1.63	9484	2.13	9834	2.63	9957
0.14	5557	0.64	7389	1.14	8729	1.64	9495	2.14	9838	2.64	9959
0.15	5596	0.65	7422	1.15	8749	1.65	9505	2.15	9842	2.65	9960
0.16	5636	0.66	7454	1.16	8770	1.66	9515	2.16	9846	2.66	9961
0.17	5675	0.67	7486	1.17	8790	1.67	9525	2.17	9850	2.67	9962
0.18	5714	0.68	7517	1.18	8810	1.68	9535	2.18	9854	2.68	9963
0.19	5753	0.69	7549	1.19	8830	1.69	9545	2.19	9857	2.69	9964
0.20	5793	0.70	7580	1.20	8849	1.70	9554	2.20	9861	2.70	9965
0.21	5832	0.71	7611	1.21	8869	1.71	9564	2.21	9864	2.71	9966
0.22	5871	0.72	7642	1.22	8888	1.72	9573	2.22	9868	2.72	9967
0.23	5910	0.73	7673	1.23	8907	1.73	9582	2.23	9871	2.73	9968
0.24	5948	0.74	7704	1.24	8925	1.74	9591	2.24	9875	2.74	9969
0.25	5987	0.75	7734	1.25	8944	1.75	9599	2.25	9878	2.75	9970
0.26	6026	0.76	7764	1.26	8962	1.76	9608	2.26	9881	2.76	9971
0.27	6064	0.77	7794	1.27	8980	1.77	9616	2.27	9884	2.77	9972
0.28	6103	0.78	7823	1.28	8997	1.78	9625	2.28	9887	2.78	9973
0.29	6141	0.79	7852	1.29	9015	1.79	9633	2.29	9890	2.79	9974
0.30	6179	0.80	7881	1.30	9032	1.80	9641	2.30	9893	2.80	9974
0.31	6217	0.81	7910	1.31	9049	1.81	9649	2.31	9896	2.81	9975
0.32	6255	0.82	7939	1.32	9066	1.82	9656	2.32	9898	2.82	9976
0.33	6293	0.83	7967	1.33	9082	1.83	9664	2.33	9901	2.83	9977
0.34	6331	0.84	7995	1.34	9099	1.84	9671	2.34	9904	2.84	9977
0.35	6368	0.85	8023	1.35	9115	1.85	9678	2.35	9906	2.85	9978
0.36	6406	0.86	8051	1.36	9131	1.86	9686	2.36	9909	2.86	9979
0.37	6443	0.87	8078	1.37	9147	1.87	9693	2.37	9911	2.87	9979
0.38	6480	0.88	8106	1.38	9162	1.88	9699	2.38	9913	2.88	9980
0.39	6517	0.89	8133	1.39	9177	1.89	9706	2.39	9916	2.89	9981
0.40	6554	0.90	8159	1.40	9192	1.90	9713	2.40	9918	2.90	9981
0.41	6591	0.91	8186	1.41	9207	1.91	9719	2.41	9920	2.91	9982
0.42	6628	0.92	8212	1.42	9222	1.92	9726	2.42	9922	2.92	9982
0.43	6664	0.93	8238	1.43	9236	1.93	9732	2.43	9925	2.93	9983
0.44	6700	0.94	8264	1.44	9251	1.94	9738	2.44	9927	2.94	9984
0.45	6736	0.95	8289	1.45	9265	1.95	9744	2.45	9929	2.95	9984
0.46	6772	0.96	8315	1.46	9279	1.96	9750	2.46	9931	2.96	9985
0.47	6808	0.97	8340	1.47	9292	1.97	9756	2.47	9932	2.97	9985
0.48	6844	0.98	8365	1.48	9306	1.98	9761	2.48	9934	2.98	9986
0.49	6879	0.99	8389	1.49	9319	1.99	9767	2.49	9936	2.99	9986
0.50	6915	1.00	8413	1.50	9332	2.00	9772	2.50	9938	3.00	9987

Table A8 Normal Distribution

Values of z for given values of $\Phi(z)$ [see (3), Sec. 24.8] and $D(z) = \Phi(z) - \Phi(-z)$
Example: $z = 0.279$ if $\Phi(z) = 61\%$; $z = 0.860$ if $D(z) = 61\%$.

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
1	−2.326	0.013	41	−0.228	0.539	81	0.878	1.311
2	−2.054	0.025	42	−0.202	0.553	82	0.915	1.341
3	−1.881	0.038	43	−0.176	0.568	83	0.954	1.372
4	−1.751	0.050	44	−0.151	0.583	84	0.994	1.405
5	−1.645	0.063	45	−0.126	0.598	85	1.036	1.440
6	−1.555	0.075	46	−0.100	0.613	86	1.080	1.476
7	−1.476	0.088	47	−0.075	0.628	87	1.126	1.514
8	−1.405	0.100	48	−0.050	0.643	88	1.175	1.555
9	−1.341	0.113	49	−0.025	0.659	89	1.227	1.598
10	−1.282	0.126	50	0.000	0.674	90	1.282	1.645
11	−1.227	0.138	51	0.025	0.690	91	1.341	1.695
12	−1.175	0.151	52	0.050	0.706	92	1.405	1.751
13	−1.126	0.164	53	0.075	0.722	93	1.476	1.812
14	−1.080	0.176	54	0.100	0.739	94	1.555	1.881
15	−1.036	0.189	55	0.126	0.755	95	1.645	1.960
16	−0.994	0.202	56	0.151	0.772	96	1.751	2.054
17	−0.954	0.215	57	0.176	0.789	97	1.881	2.170
18	−0.915	0.228	58	0.202	0.806	97.5	1.960	2.241
19	−0.878	0.240	59	0.228	0.824	98	2.054	2.326
20	−0.842	0.253	60	0.253	0.842	99	2.326	2.576
21	−0.806	0.266	61	0.279	0.860	99.1	2.366	2.612
22	−0.772	0.279	62	0.305	0.878	99.2	2.409	2.652
23	−0.739	0.292	63	0.332	0.896	99.3	2.457	2.697
24	−0.706	0.305	64	0.358	0.915	99.4	2.512	2.748
25	−0.674	0.319	65	0.385	0.935	99.5	2.576	2.807
26	−0.643	0.332	66	0.412	0.954	99.6	2.652	2.878
27	−0.613	0.345	67	0.440	0.974	99.7	2.748	2.968
28	−0.583	0.358	68	0.468	0.994	99.8	2.878	3.090
29	−0.553	0.372	69	0.496	1.015	99.9	3.090	3.291
30	−0.524	0.385	70	0.524	1.036			
31	−0.496	0.399	71	0.553	1.058	99.91	3.121	3.320
32	−0.468	0.412	72	0.583	1.080	99.92	3.156	3.353
33	−0.440	0.426	73	0.613	1.103	99.93	3.195	3.390
34	−0.412	0.440	74	0.643	1.126	99.94	3.239	3.432
35	−0.385	0.454	75	0.674	1.150	99.95	3.291	3.481
36	−0.358	0.468	76	0.706	1.175	99.96	3.353	3.540
37	−0.332	0.482	77	0.739	1.200	99.97	3.432	3.615
38	−0.305	0.496	78	0.772	1.227	99.98	3.540	3.719
39	−0.279	0.510	79	0.806	1.254	99.99	3.719	3.891
40	−0.253	0.524	80	0.842	1.282			

Table A9 ***t*-Distribution**

Values of z for given values of the distribution function $F(z)$ (see (8) in Sec. 25.3). Example: For 9 degrees of freedom, $z = 1.83$ when $F(z) = 0.95$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.32	0.29	0.28	0.27	0.27	0.26	0.26	0.26	0.26	0.26
0.7	0.73	0.62	0.58	0.57	0.56	0.55	0.55	0.55	0.54	0.54
0.8	1.38	1.06	0.98	0.94	0.92	0.91	0.90	0.89	0.88	0.88
0.9	3.08	1.89	1.64	1.53	1.48	1.44	1.41	1.40	1.38	1.37
0.95	6.31	2.92	2.35	2.13	2.02	1.94	1.89	1.86	1.83	1.81
0.975	12.7	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23
0.99	31.8	6.96	4.54	3.75	3.36	3.14	3.00	2.90	2.82	2.76
0.995	63.7	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17
0.999	318.3	22.3	10.2	7.17	5.89	5.21	4.79	4.50	4.30	4.14

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
0.7	0.54	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.53	0.53
0.8	0.88	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86
0.9	1.36	1.36	1.35	1.35	1.34	1.34	1.33	1.33	1.33	1.33
0.95	1.80	1.78	1.77	1.76	1.75	1.75	1.74	1.73	1.73	1.72
0.975	2.20	2.18	2.16	2.14	2.13	2.12	2.11	2.10	2.09	2.09
0.99	2.72	2.68	2.65	2.62	2.60	2.58	2.57	2.55	2.54	2.53
0.995	3.11	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.85
0.999	4.02	3.93	3.85	3.79	3.73	3.69	3.65	3.61	3.58	3.55

$F(z)$	Number of Degrees of Freedom									
	22	24	26	28	30	40	50	100	200	∞
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.25	0.25
0.7	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.52
0.8	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85	0.84	0.84
0.9	1.32	1.32	1.31	1.31	1.31	1.30	1.30	1.29	1.29	1.28
0.95	1.72	1.71	1.71	1.70	1.70	1.68	1.68	1.66	1.65	1.65
0.975	2.07	2.06	2.06	2.05	2.04	2.02	2.01	1.98	1.97	1.96
0.99	2.51	2.49	2.48	2.47	2.46	2.42	2.40	2.36	2.35	2.33
0.995	2.82	2.80	2.78	2.76	2.75	2.70	2.68	2.63	2.60	2.58
0.999	3.50	3.47	3.43	3.41	3.39	3.31	3.26	3.17	3.13	3.09

Table A10 Chi-square Distribution

Values of x for given values of the distribution function $F(z)$ (see Sec. 25.3 before (17)). Example: For 3 degrees of freedom, $z = 11.34$ when $F(z) = 0.99$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.005	0.00	0.01	0.07	0.21	0.41	0.68	0.99	1.34	1.73	2.16
0.01	0.00	0.02	0.11	0.30	0.55	0.87	1.24	1.65	2.09	2.56
0.025	0.00	0.05	0.22	0.48	0.83	1.24	1.69	2.18	2.70	3.25
0.05	0.00	0.10	0.35	0.71	1.15	1.64	2.17	2.73	3.33	3.94
0.95	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31
0.975	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48
0.99	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21
0.995	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.005	2.60	3.07	3.57	4.07	4.60	5.14	5.70	6.26	6.84	7.43
0.01	3.05	3.57	4.11	4.66	5.23	5.81	6.41	7.01	7.63	8.26
0.025	3.82	4.40	5.01	5.63	6.26	6.91	7.56	8.23	8.91	9.59
0.05	4.57	5.23	5.89	6.57	7.26	7.96	8.67	9.39	10.12	10.85
0.95	19.68	21.03	22.36	23.68	25.00	26.30	27.59	28.87	30.14	31.41
0.975	21.92	23.34	24.74	26.12	27.49	28.85	30.19	31.53	32.85	34.17
0.99	24.72	26.22	27.69	29.14	30.58	32.00	33.41	34.81	36.19	37.57
0.995	26.76	28.30	29.82	31.32	32.80	34.27	35.72	37.16	38.58	40.00

$F(z)$	Number of Degrees of Freedom									
	21	22	23	24	25	26	27	28	29	30
0.005	8.0	8.6	9.3	9.9	10.5	11.2	11.8	12.5	13.1	13.8
0.01	8.9	9.5	10.2	10.9	11.5	12.2	12.9	13.6	14.3	15.0
0.025	10.3	11.0	11.7	12.4	13.1	13.8	14.6	15.3	16.0	16.8
0.05	11.6	12.3	13.1	13.8	14.6	15.4	16.2	16.9	17.7	18.5
0.95	32.7	33.9	35.2	36.4	37.7	38.9	40.1	41.3	42.6	43.8
0.975	35.5	36.8	38.1	39.4	40.6	41.9	43.2	44.5	45.7	47.0
0.99	38.9	40.3	41.6	43.0	44.3	45.6	47.0	48.3	49.6	50.9
0.995	41.4	42.8	44.2	45.6	46.9	48.3	49.6	51.0	52.3	53.7

$F(z)$	Number of Degrees of Freedom							
	40	50	60	70	80	90	100	> 100 (Approximation)
0.005	20.7	28.0	35.5	43.3	51.2	59.2	67.3	$\frac{1}{2}(h - 2.58)^2$
0.01	22.2	29.7	37.5	45.4	53.5	61.8	70.1	$\frac{1}{2}(h - 2.33)^2$
0.025	24.4	32.4	40.5	48.8	57.2	65.6	74.2	$\frac{1}{2}(h - 1.96)^2$
0.05	26.5	34.8	43.2	51.7	60.4	69.1	77.9	$\frac{1}{2}(h - 1.64)^2$
0.95	55.8	67.5	79.1	90.5	101.9	113.1	124.3	$\frac{1}{2}(h + 1.64)^2$
0.975	59.3	71.4	83.3	95.0	106.6	118.1	129.6	$\frac{1}{2}(h + 1.96)^2$
0.99	63.7	76.2	88.4	100.4	112.3	124.1	135.8	$\frac{1}{2}(h + 2.33)^2$
0.995	66.8	79.5	92.0	104.2	116.3	128.3	140.2	$\frac{1}{2}(h + 2.58)^2$

In the last column, $h = \sqrt{2m - 1}$, where m is the number of degrees of freedom.

Table A11 *F*-Distribution with (*m*, *n*) Degrees of Freedom

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

Example: For (7, 4) d.f., $z = 6.09$ if $F(z) = 0.95$.

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Table A11 *F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)*

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	242	246	248	250	251	252	253	254
2	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
3	8.79	8.70	8.66	8.62	8.59	8.58	8.55	8.53
4	5.96	5.86	5.80	5.75	5.72	5.70	5.66	5.63
5	4.74	4.62	4.56	4.50	4.46	4.44	4.41	4.37
6	4.06	3.94	3.87	3.81	3.77	3.75	3.71	3.67
7	3.64	3.51	3.44	3.38	3.34	3.32	3.27	3.23
8	3.35	3.22	3.15	3.08	3.04	3.02	2.97	2.93
9	3.14	3.01	2.94	2.86	2.83	2.80	2.76	2.71
10	2.98	2.85	2.77	2.70	2.66	2.64	2.59	2.54
11	2.85	2.72	2.65	2.57	2.53	2.51	2.46	2.40
12	2.75	2.62	2.54	2.47	2.43	2.40	2.35	2.30
13	2.67	2.53	2.46	2.38	2.34	2.31	2.26	2.21
14	2.60	2.46	2.39	2.31	2.27	2.24	2.19	2.13
15	2.54	2.40	2.33	2.25	2.20	2.18	2.12	2.07
16	2.49	2.35	2.28	2.19	2.15	2.12	2.07	2.01
17	2.45	2.31	2.23	2.15	2.10	2.08	2.02	1.96
18	2.41	2.27	2.19	2.11	2.06	2.04	1.98	1.92
19	2.38	2.23	2.16	2.07	2.03	2.00	1.94	1.88
20	2.35	2.20	2.12	2.04	1.99	1.97	1.91	1.84
22	2.30	2.15	2.07	1.98	1.94	1.91	1.85	1.78
24	2.25	2.11	2.03	1.94	1.89	1.86	1.80	1.73
26	2.22	2.07	1.99	1.90	1.85	1.82	1.76	1.69
28	2.19	2.04	1.96	1.87	1.82	1.79	1.73	1.65
30	2.16	2.01	1.93	1.84	1.79	1.76	1.70	1.62
32	2.14	1.99	1.91	1.82	1.77	1.74	1.67	1.59
34	2.12	1.97	1.89	1.80	1.75	1.71	1.65	1.57
36	2.11	1.95	1.87	1.78	1.73	1.69	1.62	1.55
38	2.09	1.94	1.85	1.76	1.71	1.68	1.61	1.53
40	2.08	1.92	1.84	1.74	1.69	1.66	1.59	1.51
50	2.03	1.87	1.78	1.69	1.63	1.60	1.52	1.44
60	1.99	1.84	1.75	1.65	1.59	1.56	1.48	1.39
70	1.97	1.81	1.72	1.62	1.57	1.53	1.45	1.35
80	1.95	1.79	1.70	1.60	1.54	1.51	1.43	1.32
90	1.94	1.78	1.69	1.59	1.53	1.49	1.41	1.30
100	1.93	1.77	1.68	1.57	1.52	1.48	1.39	1.28
150	1.89	1.73	1.64	1.54	1.48	1.44	1.34	1.22
200	1.88	1.72	1.62	1.52	1.46	1.41	1.32	1.19
1000	1.84	1.68	1.58	1.47	1.41	1.36	1.26	1.08
∞	1.83	1.67	1.57	1.46	1.39	1.35	1.24	1.00

Table A11 *F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)*

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	4052	4999	5403	5625	5764	5859	5928	5981	6022
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	6056	6157	6209	6261	6287	6303	6334	6366
2	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	27.2	26.9	26.7	26.5	26.4	26.4	26.2	26.1
4	14.5	14.2	14.0	13.8	13.7	13.7	13.6	13.5
5	10.1	9.72	9.55	9.38	9.29	9.24	9.13	9.02
6	7.87	7.56	7.40	7.23	7.14	7.09	6.99	6.88
7	6.62	6.31	6.16	5.99	5.91	5.86	5.75	5.65
8	5.81	5.52	5.36	5.20	5.12	5.07	4.96	4.86
9	5.26	4.96	4.81	4.65	4.57	4.52	4.42	4.31
10	4.85	4.56	4.41	4.25	4.17	4.12	4.01	3.91
11	4.54	4.25	4.10	3.94	3.86	3.81	3.71	3.60
12	4.30	4.01	3.86	3.70	3.62	3.57	3.47	3.36
13	4.10	3.82	3.66	3.51	3.43	3.38	3.27	3.17
14	3.94	3.66	3.51	3.35	3.27	3.22	3.11	3.00
15	3.80	3.52	3.37	3.21	3.13	3.08	2.98	2.87
16	3.69	3.41	3.26	3.10	3.02	2.97	2.86	2.75
17	3.59	3.31	3.16	3.00	2.92	2.87	2.76	2.65
18	3.51	3.23	3.08	2.92	2.84	2.78	2.68	2.57
19	3.43	3.15	3.00	2.84	2.76	2.71	2.60	2.49
20	3.37	3.09	2.94	2.78	2.69	2.64	2.54	2.42
22	3.26	2.98	2.83	2.67	2.58	2.53	2.42	2.31
24	3.17	2.89	2.74	2.58	2.49	2.44	2.33	2.21
26	3.09	2.81	2.66	2.50	2.42	2.36	2.25	2.13
28	3.03	2.75	2.60	2.44	2.35	2.30	2.19	2.06
30	2.98	2.70	2.55	2.39	2.30	2.25	2.13	2.01
32	2.93	2.65	2.50	2.34	2.25	2.20	2.08	1.96
34	2.89	2.61	2.46	2.30	2.21	2.16	2.04	1.91
36	2.86	2.58	2.43	2.26	2.18	2.12	2.00	1.87
38	2.83	2.55	2.40	2.23	2.14	2.09	1.97	1.84
40	2.80	2.52	2.37	2.20	2.11	2.06	1.94	1.80
50	2.70	2.42	2.27	2.10	2.01	1.95	1.82	1.68
60	2.63	2.35	2.20	2.03	1.94	1.88	1.75	1.60
70	2.59	2.31	2.15	1.98	1.89	1.83	1.70	1.54
80	2.55	2.27	2.12	1.94	1.85	1.79	1.65	1.49
90	2.52	2.24	2.09	1.92	1.82	1.76	1.62	1.46
100	2.50	2.22	2.07	1.89	1.80	1.74	1.60	1.43
150	2.44	2.16	2.00	1.83	1.73	1.66	1.52	1.33
200	2.41	2.13	1.97	1.79	1.69	1.63	1.48	1.28
1000	2.34	2.06	1.90	1.72	1.61	1.54	1.38	1.11
∞	2.32	2.04	1.88	1.70	1.59	1.52	1.36	1.00

Table A12 Distribution Function $F(x) = P(T \leqq x)$ of the Random Variable T in Section 25.8

x	$n = 3$
	0.
0	167
1	500

x	$n = 4$
	0.
0	042
1	167
2	375

x	$n = 5$
	0.
0	008
1	042
2	117
3	242
4	408

x	$n = 6$
	0.
0	001
1	008
2	028
3	068
4	136
5	235
6	360
7	500

x	$n = 7$
	0.
1	001
2	005
3	015
4	035
5	068
6	119
7	191
8	281
9	386
10	500

x	$n = 8$
	0.
2	001
3	003
4	007
5	016
6	031
7	054
8	089
9	138
10	199
11	274
12	360
13	452

x	$n = 9$
	0.
4	001
5	003
6	006
7	012
8	022
9	038
10	060
11	090
12	130
13	179
14	238
15	306
16	381
17	460

x	$n = 10$
	0.
6	001
7	002
8	005
9	008
10	014
11	023
12	036
13	054
14	078
15	108
16	146
17	190
18	242
19	300
20	364
21	431
22	500

x	$n = 11$
	0.
8	001
9	002
10	003
11	005
12	008
13	013
14	020
15	030
16	043
17	060
18	082
19	109
20	141
21	179
22	223
23	271
24	324
25	381
26	440
27	500

x	$n = 12$
	0.
11	001
12	002
13	003
14	004
15	007
16	010
17	016
18	022
19	031
20	043
21	058
22	076
23	098
24	125
25	155
26	190
27	230
28	273
29	319
30	369
31	420
32	473

x	$n = 13$
	0.
14	001
15	001
16	002
17	003
18	005
19	007
20	011
21	015
22	021
23	029
24	038
25	050
26	064
27	082
28	102
29	126
30	153
31	184
32	218
33	255
34	295
35	338
36	383
37	429
38	476

x	$n = 14$
	0.
18	001
19	002
20	002
21	003
22	005
23	007
24	010
25	013
26	018
27	024
28	031
29	040
30	051
31	063
32	079
33	096
34	117
35	140
36	165
37	194
38	225
39	259
40	295
41	334
42	374
43	415
44	457
45	500

x	$n = 15$
	0.
23	001
24	002
25	003
26	004
27	006
28	008
29	010
30	014
31	018
32	023
33	029
34	037
35	046
36	057
37	070
38	084
39	101
40	120
41	141
42	164
43	190
44	218
45	248
46	279
47	313
48	349
49	385
50	423
51	461
52	500

x	$n = 16$
	0.
27	001
28	002
29	002
30	003
31	004
32	006
33	008
34	010
35	013
36	016
37	021
38	026
39	032
40	039
41	048
42	058
43	070
44	083
45	097
46	114
47	133
48	153
49	175
50	199
51	225
52	253
53	282
54	313
55	345
56	378
57	412
58	447
59	482

x	$n = 17$
	0.
32	001
33	002
34	002
35	003
36	004
37	005
38	007
39	009
40	011
41	014
42	017
43	021
44	026
45	032
46	038
47	046
48	054
49	064
50	076
51	088
52	102
53	118
54	135
55	154
56	174
57	196
58	220
59	245
60	271
61	299
62	328
63	358
64	388
65	420
66	452
67	484

x	$n = 18$
	0.
38	001
39	002
40	003
41	003
42	004
43	005
44	007
45	009
46	011
47	013
48	016
49	020
50	024
51	029
52	034
53	041
54	048
55	056
56	066
57	076
58	088
59	100
60	115
61	130
62	147
63	165
64	184
65	205
66	227
67	250
68	275
69	300
70	327
71	354
72	383
73	411
74	441
75	470
76	500

x	$n = 19$
	0.
43	001
44	002
45	002
46	003
47	003
48	004
49	005
50	006
51	008
52	010
53	012
54	014
55	017
56	021
57	025
58	029
59	034
60	040
61	047
62	054
63	062
64	072
65	082
66	093
67	105
68	119
69	133
70	149
71	166
72	184
73	203
74	223
75	245
76	267
77	290
78	314
79	339
80	365
81	391
82	418
83	445
84	473
85	500

x	$n = 20$
	0.
50	001
51	002
52	002
53	003
54	004
55	005
56	006
57	007
58	008
59	010
60	012
61	014
62	017
63	020
64	023
65	027
66	032
67	037
68	043
69	049
70	056
71	064
72	073
73	082
74	093
75	104
76	117
77	130
78	144
79	159
80	176
81	193
82	211
83	230
84	250
85	271
86	293
87	315
88	339
89	362
90	387
91	411
92	436
93	462
94	487