California Polytechnic State University, San Luis Obispo

JAY L. DEVORE

Tài liệu môn học

Probability and Statistics for Engineering and Sciences

XÁC SUẤT THỐNG KÊ ỨNG DỤNG

Người dịch:
Nguyễn Hồng Nhung
Hoàng Thị Minh Thảo
Lê Thị Mai Trang
Nguyễn Ngọc Tứ

Bộ môn Toán - ĐH SPKT, Tp. Hồ Chí Minh - Năm 2017

Chương 5

PHÂN PHỐI XÁC SUẤT ĐỒNG THỜI VÀ MẪU NGẪU NHIÊN

Giới thiệu

Trong chương 3 và 4, chúng ta đã phát triển các mô hình xác suất cho một biến ngẫu nhiên đơn biến. Nhiều vấn đề về xác suất thống kê liên quan đến các biến ngẫu nhiên đồng thời. Trong chương này, đầu tiên chúng ta thảo luận về các mô hình xác suất đồng thời của các biến ngẫu nhiên, đặc biệt nhấn mạnh vào trường hợp trong đó các biến số độc lập với nhau. Sau đó chúng ta nghiên cứu các giá trị kì vọng của các hàm số của các biến ngẫu nhiên, gồm hiệp phương sai và tương quan để đo mức độ liên kết giữa hai biến.

Ba phần cuối của chương xem xét các hàm của n biến ngẫu nhiên $X_1, X_2, ..., X_n$, và tập trung đặc biệt vào trung bình của chúng $(X_1 + ... + X_n)/n$. Một hàm như vậy không những là một biến ngẫu nhiên mà còn là một thống kê. Các phương pháp từ xác suất sẽ được sử dụng để lấy thông tin cho phân phối của một thống kê. Ta được kết quả đầu tiên là Định lý giới hạn trung tâm (Central Limit Theorem), cơ sở cho nhiều quy trình suy luận liên quan đến kích cỡ mẫu lớn.

5.1 Phân phối đồng thời của các biến ngẫu nhiên

Có rất nhiều tình huống thử nghiệm trong đó có nhiều hơn một biến ngẫu nhiên được người quan sát quan tâm. Trước tiên chúng ta xem xét phân phối xác suất đồng thời cho hai biến ngẫu nhiên rời rạc, sau đó cho hai biến liên tục và cuối

cùng cho nhiều hơn hai biến.

5.1.1 Hai biến ngẫu nhiên rời rạc

Hàm xác suất khối của một biến ngẫu nhiên rời rạc đơn X xác định xác suất khối được đặt trên mỗi giá trị có thể của X. Hàm xác suất khối đồng thời của hai biến rời rạc X và Y mô tả xác suất khối được đặt trên mỗi cặp giá trị (x, y).

Định nghĩa 5.1.1. Cho X và Y là hai biến rời rạc được định nghĩa trên không gian mẫu Ω của một phép thử. Hàm xác suất khối $p(x, y)$ cho mỗi cặp số (x, y) được định nghĩa bởi

$$p(x, y) = P(X = x \text{ and } Y = y)$$

thỏa $p(x, y) \geq 0$ và $\Sigma_x \Sigma_y p(x, y) = 1$.

Ví dụ 5.1 Cho biến ngẫu nhiên rời rạc X là chiều cao sinh viên (mét) với tập giá trị 1,5 ; 1,6 ; 1,7, biến ngẫu nhiên Y là cân nặng của sinh viên (kg) với tập giá trị là 50; 60; 65 và bảng phân phối xác suất đồng thời của X và Y như sau:

		Y		
	$p_{X,Y}$	50	60	65
X	1,5	0,1	0,2	0,1
	1,6	0,2	0,05	0,1
	1,7	0,05	0,1	0,1

Khi đó tính các xác suất
$p(1,5;60) = P(X = 1,5; Y = 60) = 0,2$
$P(Y \geq 60) = p(1,5;60)+p(1,6;60)+p(1,7;60)+p(1,5;65)+p(1,6;65)+p(1,7;65) = 0,2+0,05+0,1+0,1+0,1+0,1 = 0,65$
$P(X = 1,6) = 0,2+0,05+0,1 = 0,35$

Định nghĩa 5.1.2. Hàm xác suất khối biên duyên của X, được kí hiệu là p_X, được định nghĩa $p_X(x) = \Sigma_{y:p(x,y)>0} \quad p(x, y)$ với mỗi giá trị có thể của x.
Tương tự, **hàm xác suất khối biên duyên** của Y là $p_Y(y) = \Sigma_{x:p(x,y)>0} p(x, y)$ với mỗi giá trị có thể của y.

Việc sử dụng từ "biên duyên" ở đây là kết quả của thực tế rằng nếu hàm xác suất khối đồng thời được biểu diễn trong một bảng chữ nhật, thì các tổng xác suất các hàng (cột) chính là xác suất khối biên duyên của X và các tổng xác suất của các cột (hàng) chính là xác suất khối biên duyên của Y.

Ví dụ 5.2 (sử dụng dữ liệu của ví dụ 5.1)

		Y		
	$p_{X,Y}$	50	60	65
X	1,5	0,1	0,2	0,1
	1,6	0,2	0,05	0,1
	1,7	0,05	0,1	0,1

$p_X(1,5) = p(1,5;50) + p(1,5;60) + p(1,5;65) = 0,1 + 0,2 + 0,1 = 0,4$
$p_X(1,6) = p(1,6;50) + p(1,6;60) + p(1,56;65) = 0,2 + 0,05 + 0,1 = 0,35$
$p_X(1,7) = p(1,7;50) + p(1,7;60) + p(1,7;65) = 0,05 + 0,1 + 0,1 = 0,25$
Vậy hàm xác suất khối biên duyên của X là:

$$p_X(x) = \begin{cases} 0,4 & \text{nếu} X = 1,5 \\ 0,35 & \text{nếu} X = 1,6 \\ 0,25 & \text{nếu} X = 1,7 \end{cases}$$

Tương tự hãy tính hàm xác suất khối biên duyên cho Y?

5.1.2 Hai biến ngẫu nhiên liên tục

Xác suất mà giá trị quan sát thấy của một biến ngẫu nhiên liên tục X nằm trong một tập hợp một chiều A(ví dụ một khoảng)có được bằng cách lấy tích phân hàm mật độ $f(x)$ trên tập A. Tương tự như vậy, xác suất mà cặp (X, Y) của các biến ngẫu nhiên liên tục rơi vào tập hai chiều A (ví dụ hình chữ nhật) có được bằng cách lấy tích phân một hàm gọi tên là *hàm mật độ đồng thời* .

Định nghĩa 5.1.3. Cho X và Y là các biến ngẫu nhiên liên tục. Một **hàm mật độ đồng thời** $f(x, y)$ cho hai biến này là một hàm thỏa $f(x, y) \geq 0$ và $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dxdy = 1$. Khi đó với tập 2 chiều A ta có:

$$P[(X, Y) \in A] = \int \int_A f(x, y)dxdy$$

Đặc biệt, nếu A là hình chữ nhật 2 chiều $(x, y) : a \leq x \leq b, c \leq y \leq d$ thì

$$P[(X,Y) \in A] = P(a \le x \le b, c \le y \le d) = \int_a^b \int_c^d f(x,y)dxdy$$

Hàm mật độ biên duyên của mỗi biến có thể có được bằng cách tương tự như ta tìm cho trường hợp biến 2 chiều rời rạc. Hàm mật độ biên duyên của X tại giá trị x có được khi ta cố định x trong cặp (x,y) và lấy tích phân hàm mật độ đồng thời theo y. Còn nếu lấy tích phân hàm mật độ đồng thời theo x thì được hàm mật độ biên duyên theo Y.

Định nghĩa 5.1.4. Hàm mật độ xác suất biên duyên của X và Y , được kí hiệu $f_X(x)$ và $f_Y(y)$, được cho bởi công thức:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \text{ với } -\infty < x < \infty$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \text{ với } -\infty < y < \infty$$

Ví dụ 5.3 Một ngân hàng mở hai loại dịch vụ A và B, ngẫu nhiên chọn 1 ngày, gọi X là tỉ lệ thời gian mà dịch vụ A được sử dụng, và Y là tỉ lệ thời gian dịch vụ B được dùng. Thì tập hợp (X,Y) là một hình chữ nhật $D = \{(x,y) : 0 \le x \le 1, 0 \le y \le 1\}$. Giả sử hàm mật độ đồng thời của (X,Y) cho bởi

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & \text{với} \quad 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

Chú ý rằng điều kiện của hàm mật độ đồng thời là $\int_{\infty}^{\infty} \int_{\infty}^{\infty} f(x,y)dxdy = 1$ và $f(x,y) \ge 0$ đã thỏa.

a/ Xác suất không dịch vụ nào bận hơn $1/4$ thời gian là:

$P(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4}) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x+y^2)dxdy = \frac{6}{5} \int_0^{1/4} \int_0^{1/4} xdxdy + \frac{6}{5} \int_0^{1/4} \int_0^{1/4} y^2 dxdy = 0,0109$

b/ Tính $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 \frac{6}{5}(x+y^2) = \frac{6}{5}x + \frac{2}{5}$

Hàm mật độ biên duyên của X (chính là phân phối xác suất của thời gian bận của dịch vụ A mà không quan tâm đến dịch vụ B) là:

$$f_X(x) = \begin{cases} \frac{6}{5}x + \frac{2}{5} & \text{với} \quad 0 \le x \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

c/ Tìm hàm mật độ biên duyên của Y? Tính $P(\frac{1}{4} \le Y \le \frac{3}{4})$?

Ví dụ 5.4 Một công ty hạt có thị trường hộp các loại hạt pha trộn chứa hạnh nhân, hạt điều và đậu phộng. Giả sử trọng lượng tịnh của mỗi hộp có thể là 1 kg, nhưng trọng lượng của từng loại hạt là ngẫu nhiên. Bởi vì ba trọng số cộng lại bằng 1, một mô hình xác suất đồng thời cho hai loại hạt sẽ cung cấp thông tin cần thiết về trọng lượng cho loại hạt thứ ba. Đặt X là trọng lượng của hạnh nhân trong một chiếc hộp đã chọn và Y là trọng lượng hạt điều. Thì vùng để hàm mật độ dương là $D = (x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1$. Cho hàm mật độ đồng thời có dạng

$$f(x,y) = \begin{cases} kxy & \text{với} \quad 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1 \\ 0 & \text{nơi khác} \end{cases}$$

a/ Tính k ? (đs: 24)
b/ Tính xác suất mà hai loại hạt bất kì chiếm 50 phần trăm ? (Gợi ý: $A = \{(x, y) : 0 \leq x \leq 1, 0 \leq x \leq 1, x + y \leq 0, 5\}$) (đs: 0,0625)

5.1.3 Các biến ngẫu nhiên độc lập

Trong Chương 2, chúng ta đã chỉ ra một cách xác định tính độc lập của hai biến cố là thông qua điều kiện $P(A \cap B) = P(A).P(B)$. Đây là một định nghĩa tương tự cho sự độc lập của hai biến ngẫu nhiên.

Định nghĩa 5.1.5. Hai biến ngẫu nhiên X và Y là độc lập nếu mỗi cặp giá trị x và y thỏa

$$p(x, y) = p_X(x).p_Y(y) \text{ khi } X \text{ và } Y \text{ rời rạc}$$

hoặc

$$f(x, y) = f_X(x).f_Y(y) \text{ khi } X \text{ và } Y \text{ liên tục.}$$

Nếu hai công thức trên không thỏa cho tất cả (x, y) thì X và Y được gọi là phụ thuộc.

Ví dụ 5.5 Trong ví dụ 5.1 cho biến rời rạc, tồn tại $(p(1,6; 50) = 0,2) \neq (p(1,6) \times p(50) = 0.35 \times 0.35)$ nên X và Y không độc lập.
Ví dụ 5.6 Trong ví dụ 5.3 ; tồn tại $f_X(\frac{3}{4}) = f_Y(\frac{3}{4}) = \frac{9}{16}; f(\frac{3}{4}, \frac{3}{4}) = 0 \neq \frac{9}{16}.\frac{9}{16}$ nên hai biến trên không độc lập.

Sự độc lập của hai biến ngẫu nhiên khá quan trọng khi mô tả thí nghiệm nghiên cứu khi mà X và Y không ảnh hưởng lẫn nhau. Ngoài ra ta còn có kết quả sau: $P(a \leq X \leq b, c \leq X \leq d) = P(a \leq X \leq b).P(c \leq X \leq d)$

5.1.4 Nhiều hơn 2 biến

Để mô hình hoá các hàm của hơn hai biến ngẫu nhiên, chúng ta mở rộng khái niệm về phân phối đồng thời của hai biến.

Định nghĩa 5.1.6. Nếu $X_1, X_2, ..., X_n$ là tất cả biến ngẫu nhiên rời rạc, hàm xác suất đồng thời của các biến ngẫu nhiên là :

$$p(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

Nếu các biến liên tục, hàm mật độ đồng thời của $X_1, X_2, ..., X_n$ là hàm $p(x_1, x_2, ..., x_n)$ chẳng hạn với bất kì n khoảng $[a_1, b_1], ..., [a_n, b_n]$,

$$P(a_1 \leq X_1 \leq b_1, ..., a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} ... \int_{a_n}^{b_n} f(x_1, ..., x_n) dx_n ... dx_1$$

Trong một phép thử nhị thức, mỗi thí nghiệm chỉ cho một kết quả trong 2 kết quả có thể xảy ra. Bây giờ xét một phép thử gồm n thí nghiệm độc lập và giống nhau, mà trong đó mỗi phép thử chỉ cho một kết quả trong r kết quả có thể xảy ra. Cho $p_i = P$ (kết quả thứ i của bất kì phép thử cụ thể), và định nghĩa biến ngẫu nhiên $X_i =$ số lần thí nghiệm mà có kết quả $i (i = 1, ..., r)$. Một thí nghiệm như vậy được gọi là **phép thử đa thức - multinomial experiment** và hàm xác suất khối của $X_1, X_2, ..., X_n$ được gọi là **phân phối đa thức - multinomial distribution**. Bằng cách sử dụng lý luận tính tương tự với một biến trong trường hợp phân phối nhị thức, hàm xác suất đồng thời của $X_1, X_2, ..., X_n$ là

$$p(x_1, ..., x_r) = \begin{cases} \frac{n!}{(x_1!)(x_2!)...(x_r!)} . p_1^{x_1} ... p_r^{x_r}, & x_i = 0, 1, 2, ...; x_1 + ... + x_r = n \\ 0, \quad \text{nơi khác} \end{cases}$$

Trường hợp $r = 2$ sẽ cho phân phối nhị thức, với $X_1 =$số lần thành công và $X_2 = n = X_1 =$ số lần thất bại.

Định nghĩa 5.1.7. Biến ngẫu nhiên $X_1, X_2, ..., X_n$ được gọi là **độc lập** nếu với mỗi tập con $X_{i_1}, X_{i_2}, ..., X_{i_k}$ của các biến (mỗi cặp, mỗi bộ ba , ...)thì hàm xác suất đồng thời hay hàm mật độ đồng thời bằng với tích của các hàm xác suất biên duyên hay các hàm mật độ biên duyên..

Vì vậy, nếu các biến độc lập với $n = 4$, thì hàm xác suất đồng thời hay hàm mật độ đồng thời của 2 biến là tích của hai hàm biên duyên, tương tự cho trường hợp 3 và 4 biến. Quan trong nhất là khi ta đã có n biến độc lập thì hàm xác suất đồng thời hay hàm mật độ đồng thời là tích của n hàm biên duyên.

5.1.5 Phân phối điều kiện

Định nghĩa 5.1.8. Cho X và Y là 2 biến ngẫu nhiên liên tục với hàm mật độ biên duyên $f(x,y)$ và hàm mật độ biên duyên của X là $f_X(x)$. Thì với bất kì X có giá trị x mà $f_X(x) > 0$,**hàm mật độ điều kiện của Y cho bởi X=x** là

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \qquad -\infty < y < \infty$$

Nếu X và Y là biến rời rạc, thay hàm mật độ bằng hàm xác suất khối trong định nghĩa này thì được **hàm xác suất khối điều kiện của Y khi X=x.**

Ví dụ 5.7 Dựa vào dữ liệu của ví dụ 5.3.
a/ Hàm mật độ điều kiện của Y cho bởi $X = 0.8$ là
$f_{Y|X}(y|0,8) = \frac{f(0,8;y)}{f_X(0.8)} = \frac{1,2(0,8+y^2)}{1,2(0,8)+0,4} = \frac{1}{34}(24 + 30y^2)$ với $0 < y < 1$
b/ Xác suất để dịch vụ A bận trong nửa thời gian đầu khi biết $X = 0.8$
$P(Y \leq 0,5|X = 0,8) = \int_{-\infty}^{0,5} f_{Y|X}(y|0,8)dy = \int_0^{0,5} \frac{1}{34}(24 + 30y^2)dy = 0,39$
c/ Kì vọng có điều kiện $X = 0,8$ là:
$E(Y|X = 0,8) = \int_{-\infty}^{\infty} y.f_{Y|X}(y|0,8)dy = \frac{1}{34}\int_0^1 y(24 + 30y^2)dy = 0.574$

5.2 Giá trị Kỳ vọng, Phương sai và Hiệp phương sai

Bất kỳ hàm $h(X)$ của một biến ngẫu nhiên đơn thì chính nó cũng là một biến ngẫu nhiên. Tuy nhiên, để tính toán $E[h(X)]$ thì không cần phải có phân phối xác suất của $h(X)$; thay vào đó, $E[h(X)]$ được tính như một trung bình trọng số(a weighted average) của các giá trị $h(X)$, trong đó hàm trọng số (the weight function) là hàm xác suất khối $p(x)$ hoặc hàm mật độ $f(x)$ của X. Kết quả tương tự cho một hàm phân phối đồng thời 2 biến $h(x,y)$.

Mệnh đề 5.2.1. *Cho X và Y là biến ngẫu nhiên đồng thời với hàm xác suất khối $p(x,y)$ hoặc hàm mật độ $f(x,y)$ tương ứng với trường hợp biến rời rạc hoặc biến liên tục. Thì giá trị kì vọng của hàm $h(X,Y)$ được kí hiệu là $E[h(X,Y)]$ hoặc $\mu_{h(X,Y)}$ được tính như sau:*

$$E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y).p(x,y), & \text{nếu } X \text{ và } Y \text{ rời rạc} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y).f(x,y)dxdy, & \text{nếu } X \text{ và } Y \text{ liên tục} \end{cases}$$

Ví dụ 5.8 Năm người bạn đã mua vé cho một buổi hòa nhạc. Nếu vé là chỗ ngồi từ 1-5 trong một hàng nào đó và vé được phân phối ngẫu nhiên cho năm người, thì kì vọng cho số ghế khi chia cho 2 người là bao nhiêu? Đặt X và Y kí hiệu cho số ghế của người đầu tiên và người thứ 2. Cặp (X,Y) có thể có các giá trị (1, 2), (1, 3),. . . , (5, 4), và hàm xác suất khối đồng thời của (X, Y) là

$$p(x,y) = \begin{cases} \frac{1}{20} & x = 1,...,5; y = 1,...,5; x \neq y \\ 0 & \text{nơi khác} \end{cases}$$

Lúc đó số chỗ ngồi chia cho 2 người là hàm $h(X,Y) = |X-Y|-1$, ta có bảng tương ứng sau:

		x				
	$h_{X,Y}$	1	2	3	4	5
y	1	—	0	1	2	3
	2	0	—	0	1	2
	3	1	0	—	0	1
	4	2	1	0	—	0
	5	3	2	1	0	—

Vì vậy $E[h(X,Y)] = \sum_(x,y) \sum h(x,y).p(x,y) = \sum_{x=1}^{5} \sum_{y=1,x\neq y}^{5} (|x-y|-1).\frac{1}{20} = 1$

Ví dụ 5.9 Hàm mật độ đồng thời của X (lượng hạnh nhân) và Y (lượng hạt điều) trong 1kg hạt là:

$$f(x,y) = \begin{cases} 24xy & \text{với} \quad 0 \leq x \leq 1, 0 \leq y \leq 1, x+y \leq 1 \\ 0 & \text{nơi khác} \end{cases}$$

Nếu 1kg hạt hạnh nhân có giá 1\$; 1 kg hạt điều có giá 1,5\$; 1 kg hạt đậu phụng có giá 0.5\$; thì tổng giá trị của 1 hộp là:
$h(X,Y) = 1.X + 1,5.Y + 0,5.(1-X-Y) = 0,5 + 0,5.X + Y$
Kì vọng cho tổng giá trị là:
$E[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y).f(x,y)dxdy = \int_0^1 \int_0^{1-x} (0,5 + 0,5x + y).24xydydx = 1,1\$$

5.2.1 Hiệp phương sai

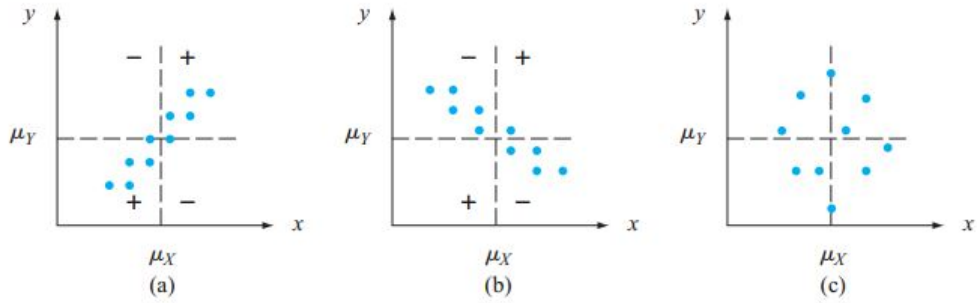
Định nghĩa 5.2.2. Hiệp phương sai giữa hai biến ngẫu nhiênX và Y là

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x,y) & , \quad X,Y \text{ rời rạc} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy, & X,Y \text{ liên tục} \end{cases}$$

Đó là, vì $X-\mu_X$ và $Y-\mu_Y$ là độ lệch của hai biến từ các giá trị trung bình tương ứng, hiệp phương sai là kì vọng của tích hai độ lệch. Chú ý rằng $Cov(X,X) = E[(X-\mu_X)^2] = V(X)$.

Giải thích định nghĩa: giả sử X và Y có một mối liên hệ dương với nhau, nghĩa là các giá trị lớn của X có khuynh xảy ra với các giá trị lớn của Y và các giá trị nhỏ của X với giá trị nhỏ của Y. Thì hầu như xác suất khối hay hàm mật độ sẽ được liên quan đến $x-\mu_X$ và $y-\mu_Y$, khi cả hai giá trị X,Y đều âm hay cả hai đều dương. Do đó với liên kết dương mạnh, $Cov(X,Y)$ sẽ hầu như dương (quite positive). Đối với một liên kết âm mạnh dấu của $X-\mu_X$ và $Y-\mu_Y$ sẽ có xu hướng ngược nhau, lúc đó thì $Cov(X,Y)$ hầu như âm. Nếu X,Y không có liên kết mạnh, tích âm và dương có xu hướng triệt tiêu lẫn nhau thì $Cov(X,Y) = 0$.

Hình dưới chỉ ra những khả năng khác nhau. Hiệp phương sai phụ thuộc vào các tập kết quả và xác suất. Hiệp phương sai có thể được thay đổi mà không cần điều chỉnh những tập kết quả, và có thể làm thay đổi nhiều đến giá trị của $Cov(X,Y)$.



p(x, y) = 1/10 for each of ten pairs corresponding to indicated points:
(a) positive covariance; (b) negative covariance; (c) covariance near zero

Ví dụ 5.10 Cho hàm xác suất khối và hàm phân phối biên duyên của hai biến X,Y như sau:

		y		
	$p_{X,Y}$	0	100	200
x	100	0,2	0,1	0,2
	250	0,05	0,15	0,3

x	100	250
$p_X(x)$	0,5	0,5

y	0	100	200
$p_Y(y)$	0,25	0,25	0,5

Áp dụng công thức tính Cov (X,Y) ? (Đs: 1875)

Mệnh đề 5.2.3. $Cov(X,Y) = E(X,Y) - \mu_X.\mu_Y$

Ví dụ 5.11

$$f(x,y) = \begin{cases} 24xy & \text{với} \quad 0 \le x \le 1, 0 \le y \le 1, x+y \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

$$f_X(x) = \begin{cases} 12x(1-x^2) & \text{với} \quad 0 \le x \le 1 \\ 0 & \text{nơi khác} \end{cases}$$

với $f_Y(y)$ cũng có được tương tự bằng cách thay x bởi y trong $f_X(x)$.
Hãy tính μ_X , μ_Y, $E(X,Y)$ từ đó suy ra $Cov(X,Y)$? (đs: -2/75)

5.2.2 Tương quan

Định nghĩa 5.2.4. Hệ số tương quan của X và Y, được kí hiệu là $Corr(X,Y), \rho_{X,Y}$ hay chỉ là ρ và được định nghĩa như sau

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$$

Ví dụ 5.12 Từ dữ kiện của ví dụ 5.10. Hãy tính hệ số tương quan của X và Y ? (Đs: 0,301)

Mệnh đề 5.2.5.

1. *Nếu a và c cùng âm hoặc cùng dương thì*

$$Corr(aX+b, cY+d) = Corr(X,Y)$$

2. *Với bất kì hai biến ngẫu nhiên X và Y thì $-1 \le Corr(X,Y) \le 1$*

Mệnh đề 5.2.6.

1. *Nếu X và Y độc lập, thì $\rho = 0$ nhưng $\rho = 0$ thì không suy ra được là có sự độc lập.*

2. *$\rho = 1$ hay -1 khi và chỉ khi $Y = aX + b$ với a, b là số nào đó và $a \neq 0$.*

Mệnh đề này chỉ ra rằng ρ là một thước đo về bậc của mối quan hệ **tuyến tính** giữa X và Y, và chỉ khi hai biến này quan hệ tuyến tính thì ρ có thể âm hoặc dương tùy ý. Giá trị tuyệt đối của $\rho < 1$ chỉ ra rằng mối quan hệ không tuyến tính hoàn toàn, nhưng vẫn có mối quan hệ phi tuyến tính trong đó. Cũng như vậy khi $\rho = 0$ thì không chỉ ra rằng X và Y độc lập, mà chỉ là X và Y không có quan hệ tuyến tính. Khi $\rho = 0$, X và Y được gọi là **không tương quan**. Hai biến có thể không tương quan nhưng chưa chắc phụ thuộc vì có quan hệ phi tuyến tính mạnh ở đây, vì vậy cần cẩn thận đưa ra kết luận khi gặp $\rho = 0$.

5.3 Các phân phối thống kê

Các quan sát trong một mẫu đơn được kí hiệu trong chương 1 là $x_1, x_2, ..., x_n$. Xem như chọn hai mẫu khác nhau có kích thước n từ cùng một phân phối tổng thể. x_i trong mẫu thứ hai sẽ hầu như luôn khác nhau một chút với các giá trị của mẫu đầu tiên. Ví dụ, một mẫu đầu tiên của $n = 3$ xe ô tô loại đặc biệt có thể cho hiệu quả nhiên liệu $x_1 = 30,7, x_2 = 29,4, x_3 = 31,1$, trong khi một mẫu thứ hai có thể cho $x_1 = 28,8, x_2 = 30.0$ và $x_3 = 32.5$. Trước khi chúng ta thu được dữ liệu, có sự không chắc chắn về giá trị của mỗi x_i. Vì điều này không chắc chắn, trước khi thu thập dữ liệu, chúng ta xem mỗi quan sát như là một biến ngẫu nhiên và kí hiệu mẫu bằng $X_1, X_2, ..., X_n$ (chữ in hoa cho biến ngẫu nhiên).
Sự thay đổi này trong các giá trị được quan sát giúp suy ra các giá trị của bất kỳ hàm nào của các quan sát mẫu - chẳng hạn như trung bình mẫu, độ lệch tiêu chuẩn mẫu. Trước tiên là để có $x_1, ..., x_n$, thì có một tất định cho giá trị $\overline{x}$, giá trị s...

Định nghĩa 5.3.1. Một **thống kê** là bất kì lượng nào mà giá trị có thể được tính từ dữ liệu mẫu. Trước khi thu thập dữ liệu, có một tất định về giá trị của bất kỳ thống kê đặc biệt sẽ cho kết quả. Do đó, thống kê là một biến ngẫu nhiên và sẽ được kí hiệu bởi một kí tự viết hoa; một kí tự viết thường được sử dụng để đại diện cho giá trị đã tính toán hoặc đã quan sát của thống kê.

Vì vậy trung bình mẫu trong thống kê được kí hiệu là $\overline{X}$, và giá trị tính được của thống kê này là $\overline{x}$. Tương tự S đại diện cho độ lệch chuẩn mẫu và giá trị được tín kí hiệu là s.
Bất kì thống kê nào (bản thân nó là một biến ngẫu nhiên) đều có một phân phối xác suất. Đáng chú ý là phân phối xác suất của trung bình mẫu $\overline{X}$.

Phân phối xác suất của một thống kê đôi khi gọi là phân phối mẫu để nhấn mạnh rằng nó mô tả cách mà thống kê khác nhau về giá trị trên tất cả các mẫu được lựa chọn.

5.3.1 Mẫu ngẫu nhiên

Phân bố xác suất của bất kỳ thống kê cụ thể phụ thuộc không chỉ vào phân phối tổng thể (chuẩn, đều,...) và kích thước mẫu n mà còn phụ thuộc phương pháp lấy mẫu. Xem xét lựa chọn một mẫu kích thước $n = 2$ từ tổng thể chỉ gồm ba giá trị 1, 5 và 10, và giả sử rằng thống kê được quan tâm là phương sai mẫu. Nếu lấy mẫu được thực hiện "có hoàn lại," thì $S^2 = 0$ khi $X_1 = X_2$. Tuy nhiên, S^2 không thể bằng 0 nếu mẫu "không hoàn lại". Vì vậy, $P(S^2 = 0) = 0$ với một phương pháp lấy mẫu, và xác suất này là dương với phương pháp lấy mẫu khác. Định nghĩa tiếp theo của chúng ta mô tả một phương pháp lấy mẫu thường gặp trong thực tế.

Định nghĩa 5.3.2. Các biến ngẫu nhiên $X_1, X_2, ..., X_n$ được gọi là mẫu ngẫu nhiên (đơn) cỡ mẫu n nếu:

1. Các X_i là biến ngẫu nhiên độc lập.

2. Các X_i có cùng phân phối mẫu.

Nếu lấy mẫu có hoàn lại từ tổng thể vô hạn, thì điều kiện 1 và 2 được thỏa mãn. Những điều kiện này cũng gần như thỏa mãn nếu lấy mẫu không hoàn lại, tuy nhiên mẫu cỡ n thì nhỏ hơn nhiều so với tổng thể cỡ N. Trong thực tế, nếu $n/N < 0.05$ (nhiều nhất 5% tổng thể được lấy mẫu), chúng ta có thể tiến hành khi X_i là biến ngẫu nhiên. Hiệu quả của phương pháp lấy mẫu này là phân phối xác suất của bất kỳ thống kê nào có thể dễ dàng thu được hơn so với bất kỳ phương pháp lấy mẫu nào khác.

Có hai phương pháp chung để thu thập thông tin về phân phối mẫu của một thống kê. Một phương pháp liên quan đến tính toán dựa trên các quy tắc xác suất, và thứ hai là thực hiện một thí nghiệm mô phỏng.

5.3.2 Tìm phân phối mẫu

Các quy tắc xác suất có thể được sử dụng để có được phân phối của một thống kê miễn là nó là một hàm "khá đơn giản" của các biến ngẫu nhiên X_i và hoặc là có vài giá trị X khác nhau trong tổng thể hoặc phân phối xác suất có dạng "tốt".

5.3.3 Những thí nghiệm mô phỏng

Phương pháp thứ hai để thu thập thông tin về phân phối mẫu của thống kê là thực hiện một thí nghiệm mô phỏng. Phương pháp này thường được sử dụng khi lấy đạo hàm thông qua các quy tắc xác suất là quá khó hoặc phức tạp để thực hiện. Một thí nghiệm như vậy luôn luôn được thực hiện với sự trợ giúp của máy tính. Các đặc điểm sau của thí nghiệm yêu cầu:

1. Thống kê của lãi suất ($\overline{X}, S$, một giá trị trung bình đã được cắt xén (trimmed),...)
2. Phân phối của tổng thể (chuẩn với $\mu = 100, \sigma = 15$; phân phối đều với giới hạn dưới $A = 5$ và giới hạn trên $B = 10$, ...)
3. Kích cỡ mẫu n (ví dụ, $n = 10$ hoặc $n = 50$)
4. Số lần lặp lại k (số mẫu cần lấy)

Sau đó sử dụng phần mềm thích hợp để lấy k mẫu ngẫu nhiên khác nhau, mỗi kích thước n, từ phân phối tổng thể đã được thiết kế. Đối với mỗi mẫu, tính toán giá trị của thống kê và xây dựng một biểu đồ tần số của các giá trị k . Biểu đồ này cung cấp một xấp xỉ phân phối mẫu của thống kê. Giá trị càng lớn hơn của k, thì xấp xỉ càng tốt hơn (phân phối mẫu thât sự sẽ xuất hiện khi $k \to \infty$). Trong thực hành, $k = 500$ hoặc 1000 thường là đủ nếu thống kê "khá đơn giản".

5.4 Phân phối của trung bình mẫu

Tầm quan trọng của trung bình mẫu $\overline{X}$ là để rút ra kết luận cho trung bình tổng thể μ. Một số phương pháp suy luận dựa trên tính chất của phân phối mẫu $\overline{X}$. Các tính chất tính toán và thí nghiệm mô phỏng đã nêu trong các phần trước khi ta đề cấp đến mối quan hệ $E(\overline{X})$ và μ hay $V(\overline{X}), \sigma^2$, và n

Mệnh đề 5.4.1. *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ một phân phối với trung bình μ và độ lệch chuẩn σ. Thì*

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$
2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n$ *và* $\sigma_{\overline{X}} = \sigma^2/\sqrt{n}$

Thêm nữa, với $T_o = X_1 + ... + X_n$ (tổng của mẫu), $E(T_o) = n\mu, V(T_o) = n\sigma^2, \sigma_{T_0} = \sqrt{n}\sigma$

Các chứng minh cho những kết quả trên sẽ được trình bày ở phần tiếp theo. Theo kết quả 1, phân phối mẫu của $\overline{X}$ (tức là xác suất) tập trung chính xác ở trung bình của tổng thể mà mẫu được chọn từ đó. Kết quả 2 cho thấy rằng phân phối $\overline{X}$ trở nên tập trung hơn về μ khi kích thước mẫu n tăng lên. Sự khác biệt quan trọng là phân phối của T_o trở nên lan rộng ra khi cỡ mẫu n tăng lên.

Ví dụ 5.13 Trong một thí nghiệm kiểm tra sức chịu mài mòn trên một mẫu titan, kì vọng số chu kỳ để phát ra âm thanh đầu tiên (dùng chỉ ra vết nứt đầu tiên) là $\mu = 28.000$ và độ lệch tiêu chuẩn của số chu kỳ là $\sigma = 5000$. Đặt $X_1, X_2, ..., X_{25}$ là một mẫu ngẫu nhiên có kích thước 25, trong đó mỗi X_i là số chu kỳ trên một mẫu titan khác nhau ngẫu nhiên được chọn. Thì giá trị kì vọng của số trung bình mẫu của chu kỳ cho đến khi phát âm thanh đầu tiên là $E(\overline{X}) = \mu = 28.000$ và tổng kì vọng cho số chu kỳ cho 25 mẫu tiatan là $E(T_0) = n\mu = 25(28000) = 700000$. Độ lệch chuẩn của $\overline{X}$ (sai số tiêu chuẩn của trung bình) và của T_0 là:

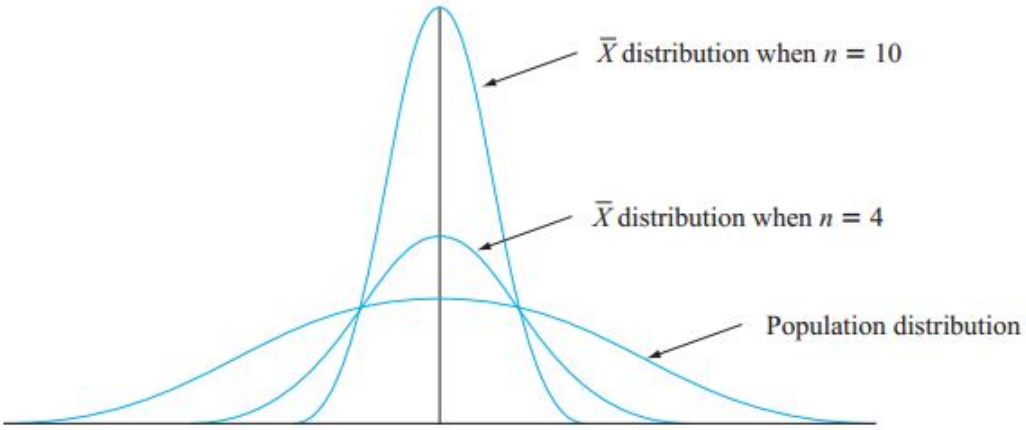
$$\sigma_{\overline{X}} = \sigma/\sqrt{n} = \frac{5000}{\sqrt{25}} = 1000$$
$$\sigma_{T_0} = \sqrt{n}\sigma = \sqrt{25}.5000 = 25000$$

Nếu cỡ mẫu tăng lên $n = 100$, $E(\overline{X})$ thì không đổi, nhưng $\sigma_{\overline{X}} = 500$ (bằng một nửa giá trị trước đó) (kích cỡ mẫu phải được tăng gấp bốn lần để giảm một nửa độ lệch tiêu chuẩn của $\overline{X}$).

5.4.1 Trường hợp phân phối chuẩn tổng thể

Mệnh đề 5.4.2. *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ phân phối chuẩn với trung bình μ và độ lệch chuẩn σ. Thì với mọi n, $\overline{X}$ được phân phối chuẩn (với trung bình μ và độ lệch chuẩn $\sigma/\sqrt{n}$), như là T_o (với trung bình $n\mu$ và độ lệch chuẩn $\sqrt{n}\sigma$) .*

Chúng ta có dữ kiện để biết phân phối $\overline{X}$ và T_o khi phân phối tổng thể là phân phối chuẩn. Đặc biệt, xác suất như $P(a \leq \overline{X} \leq b)$ và $P(c \leq T_o \leq d)$ có thể tính được bằng cách chuẩn hóa. Hình dưới giải thích cho mệnh đề



A normal population distribution and $\overline{X}$ sampling distributions

Ví dụ 5.14 Thời gian cho một con chuột của một phân loài nhất định tìm đường ra một mê cung là một phân phối chuẩn với $\mu = 1,5$ phút và $\sigma = 0,35$ phút. Giả sử năm con chuột được chọn để quan sát. Đặt $X_1, ..., X_5$ cho thời gian của chúng trong mê cung.

a/ Giả sử X_i là một mẫu ngẫu nhiên từ phân bố chuẩn này, tìm xác suất mà tổng thời gian $T_0 = X_1 + ... + X_5$ nằm trong khoảng từ 6 đến 8 phút?

Theo mệnh đề, T_0 có phân phối chuẩn với $\mu_{T_0} = n\mu = 5.(1,5) = 7,5$ và phương sai $\sigma^2_{T_0} = n\sigma^2 = 5(0,1225) = 0.6125$, vì vậy $\sigma_{T_0} = 0,783$.

$P(6 \leq T_0 \leq 8) = P(\frac{6-5,7}{0,783} \leq Z \leq \frac{8-7,5}{0,783}) = P(-1,92 \leq Z \leq 6.4) = \Phi(0,64) - \Phi(-1,92) = 0,7115$

b/ Xác định xác suất mà thời gian trung bình mẫu $\overline{X}$ nhiều nhất là 2 phút , với $\mu_{\overline{X}} = \mu = 1,5$ và $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 0,35/\sqrt{5} = 0,1565$. Thì

$P(\overline{X} \leq 2) = P(Z \leq \frac{2-1,5}{0,1565}) = P(Z \leq 3,19) = \Phi(3,19) = 0,9993$

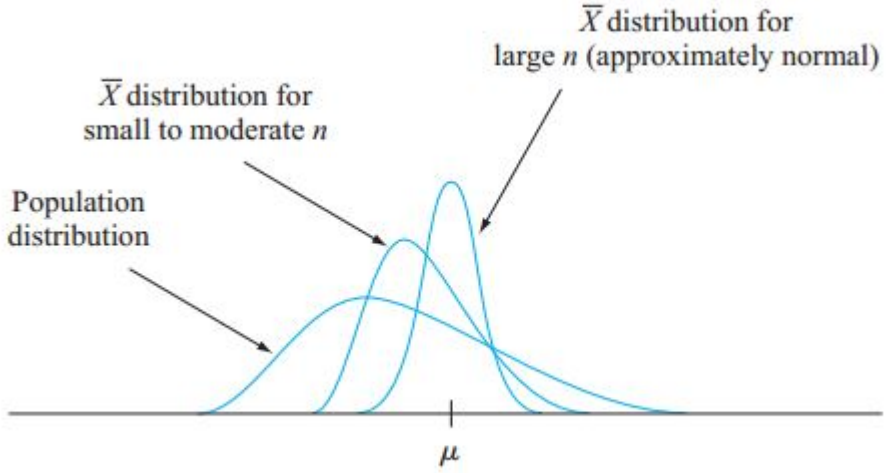
5.4.2 Định lý giới hạn trung tâm

Khi các X_i có phân phối chuẩn, thì $\overline{X}$ cũng có phân phối chuẩn cho mỗi mẫu cỡ n . Các kết quả từ ví dụ trên chỉ ra rằng khi phân phối tổng thể không chuẩn , lấy trung bình sẽ cho một phân phối tạo ra hình chuông hơn là mẫu được lấy. Nếu n lớn, một đường cong chuẩn thích hợp sẽ xấp xỉ gần đúng phân phối thật của $\overline{X}$.

Phát biểu cho ý này được trình bày trong định lý xác suất quan trọng sau đây.

Định lý 5.4.3. *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ phân phối với trung bình μ và phương sai σ^2. Thì nếu n đủ lớn, thì $\overline{X}$ có xấp xỉ gần đúng một phân phối chuẩn với $\mu_{\overline{X}} = \mu$ và $\sigma^2_{\overline{X}} = \sigma^2/n$ và T_o cũng có xấp xỉ gần đúng một phân phối chuẩn với $\mu_{T_o} = n\mu, \sigma^2_{T_o} = n\sigma^2$. Giá trị n càng lớn thì xấp xỉ càng tốt.*

Hình dưới minh họa Định lý Giới hạn Trung tâm. Theo ĐLGHTT, khi n lớn và ta muốn tính xác suất như $P(a \leq \overline{X} \leq b)$, ta chỉ cần "giả vờ" là $\overline{X}$ là chuẩn và chuẩn hóa nó, sử dụng bảng phân phối chuẩn. Kết quả sẽ được xấp xỉ đúng.



The Central Limit Theorem illustrated

Ví dụ 5.15 Lượng tạp chất cụ thể trong một lô sản phẩm hoá học nhất định là biến ngẫu nhiên có giá trị trung bình 4,0 g và độ lệch chuẩn 1,5 g. Nếu 50 lô được chuẩn bị độc lập, tính xác suất để trung bình mẫu tạp chất $\overline{X}$ nằm giữa 3,5 và 3,8g?
Ta có $n = 50$ đủ lớn để áp dụng ĐLGHTT. $\overline{X}$ xấp xỉ về phân phối chuẩn với trung bình $\mu_{\overline{X}} = 4$ và $\sigma_{\overline{X}} = 1,5/\sqrt{50} = 0,2121$ vì vậy
$P(3,5 \leq \overline{X} \leq 3,8) \approx P(\frac{3,5-4}{0,2121} \leq Z \leq \frac{3,8-4}{0,2121}) = \Phi(-0,94) - \Phi(-2,36) = 0,1645$

Ví dụ 5.16 Một tổ chức tiêu dùng nhất định thường báo cáo số lượng khiếm khuyết lớn của mỗi chiếc ô tô mới mà nó thử nghiệm. Giả sử số lỗi này là một biến ngẫu nhiên với giá trị trung bình là 3.2 và độ lệch chuẩn 2.4. Trong số 100 xe được lựa chọn ngẫu nhiên, tính xác suất trung bình mẫu của những lỗi này nhiều hơn 4? (đs: 0,0004)

Định lý giới hạn trung tâm cung cấp cái nhìn sâu sắc về lý do tại sao nhiều biến ngẫu nhiên có phân phối xác suất gần với phân phối chuẩn. Ví dụ, sai số đo lường trong một thí nghiệm khoa học có thể được coi là tổng của một số điểm nhiễu loạn đáng kể và sai số của độ đo nhỏ.
Một khó khăn thực tế trong việc áp dụng ĐLGHTT là khi nào n là đủ lớn. Vấn đề là độ chính xác của xấp xỉ cho một n cụ thể phụ thuộc vào hình dạng phân phối cơ bản ban đầu được lấy mẫu. Nếu phân phối cơ bản gần với một đường cong chuẩn, thì xấp xỉ tốt ngay cả đối với một n nhỏ, trong khi nếu nó xa phân phối chuẩn, thì cần n lớn.

Nguyên tắc chung: Nếu $n > 30$ định lý giới hạn trung tâm được sử dụng.

Có những phân phối tổng thể mà thậm chí n từ 40 hoặc 50 thì không đủ, nhưng những phân phối này hiếm gặp trong thực hành. Mặt khác, nguyên tắc chung thì bảo toàn đối với nhiều phân phối tổng thể, khi đó giá trị n ít hơn 30 thì đã đủ. Ví dụ, trong trường hợp phân phối đều của tổng thể, ĐLGHTT cho xấp xỉ tốt với $n \geq 12$.

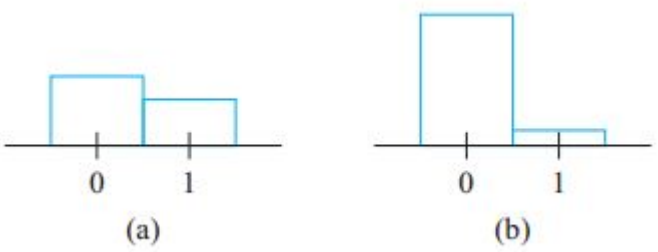
5.4.3 Những ứng dụng khác của định lý giới hạn trung tâm

ĐLGHTT có thể được sử dụng để làm rõ cho xấp xỉ chuẩn đối với phân phối nhị thức được thảo luận trong Chương 4. Nhớ lại rằng một biến nhị thức X là số lần thành công trong một thí nghiệm nhị thức bao gồm n thử nghiệm thành công / thất bại độc lập với $p = P(S)$ cho bất kỳ thử nghiệm cụ thể nào. Xác định một biến ngẫu nhiên mới X_1 bởi

$$X_1 = \begin{cases} 1 & , \quad \text{kết quả đầu thành công} \\ 0 & , \quad \text{kết quả đầu thất bại} \end{cases}$$

và định nghĩa $X_1, X_2, X_3, ..., X_n$ tương tự cho $n - 1$ thử nghiệm khác. Mỗi X_i cho biết một phép thử tương ứng có thành công hay không.
Bởi vì các thử nghiệm là độc lập và $P(S)$ là hằng số qua các thử nghiệm, các biến ngẫu nhiên X_i thì độc lập cùng phân phối (mẫu ngẫu nhiên được lấy từ phân phối Bernoulli). ĐLGTTT chỉ ra rằng nếu n là đủ lớn, cả tổng và trung bình của các X_i đều xấp xỉ phân phối chuẩn. Khi các X_i được cộng lại, tổng S được thêm 1 nếu biến đó xuất hiện và F được thêm 0 , vì vậy $X_1 + X_2 + ... + X_n = X$. Trung bình mẫu của các X_i là X/n chính là tỷ lệ thành công của mẫu. Nghĩa là, cả X và X/n đều

xấp xỉ chuẩn khi n lớn. Kích cỡ mẫu cần thiết cho phép xấp xỉ này phụ thuộc vào giá trị của p: Khi p gần 0,5, sự phân bố của mỗi X_i thì đối xứng (xem hình dưới), trong khi phân phối là khá lệch khi p gần 0 hoặc 1. Sử dụng xấp xỉ chỉ khi $np \geq 10$ và $n(1-p) \geq 10$ đảm bảo rằng n đủ lớn để vượt qua bất kỳ sự sai lệch nào trong phân phối Bernoulli bên dưới:



Nhắc lại: X có phân phối log chuẩn nếu $ln(X)$ có phân phối chuẩn.

Mệnh đề 5.4.4. *Cho $X_1, X_2, ..., X_n$ là một mẫu ngẫu nhiên từ một phân phối mà chỉ có giá trị dương thỏa $[P(X_i > 0) = 1]$. Thì với n đủ lớn, tích $Y = X_1X_2...X_n$ sẽ xấp xỉ phân phối log chuẩn*

5.5 Phân phối của tổ hợp tuyến tính

Trung bình mẫu $\overline{X}$ và tổng của mẫu T_o là những trường hợp đặc biệt của một loại biến ngẫu nhiên xuất hiện rất thường xuyên trong các ứng dụng thống kê.

Định nghĩa 5.5.1. Cho một tập n các biến ngẫu nhiên $X_1, ..., X_n$ và n hằng số $a_1, ..., a_n$ của biến ngẫu nhiên.

$$Y = a_1X_1 + ... + a_nX_n = \sum_{i=1}^{n} a_iX_i$$

được gọi là **tổ hợp tuyến tính** của biến ngẫu nhiên X_i

Mệnh đề 5.5.2. *Cho $X_1, X_2, ..., X_n$ có giá trị trung bình tương ứng $\mu_1, ..., \mu_n$ và phương sai tương ứng $\sigma_1^2, ..., \sigma_n^2$.*

1. *$X_i$có độc lập hay không đều có:*

$$E(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n) = a_1\mu_1 + ... + a_n\mu_n$$

2. *Nếu* $X_1, ..., X_n$ *độc lập*

$$V(a_1X_1 + a_2X_2 + ... + a_nX_n = a_1^2V(X_1) + a_2^2V(X_2) + ... + a_n^2V(X_n) = a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2$$

và

$$\sigma_{a_1X_1+a_2X_2+...+a_nX_n} = \sqrt{a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2}$$

3. *Với bất kì* $X_1, ...X_n$
$V(a_1X_1 + a_2X_2 + ... + a_nX_n) = \Sigma_{i=1}^n\Sigma_{j=1}^n a_ia_jCov(X_i, X_j)$

Ví dụ 5.17 Một trạm xăng bán ba loại xăng: regular, extra và super. Giá tương ứng các loại xăng này là 3 \$; 3,2\$ và 3,4\$ cho 1 gallon (4,4 lít). Gọi X_1, X_2, X_3 kí hiệu cho lượng gallon bán được trong 1 ngày nào đó. Giả sử mẫu X_i độc lập với $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300, \sigma_1 = 100, \sigma_2 = 80, \sigma_3 = 50$. Doanh thu kí hiệu là $Y = 3.X_1 + 3, 2X_2 + 3, 4X_3$. Hãy tính $E(Y), V(Y), \sigma(Y)$? (đs: 5620 ; 184,436 , 42,46)

5.5.1 Hiệu của hai biến ngẫu nhiên

Một trường hợp đặc biệt quan trọng của tổ hợp tuyến tính khi cho $n = 2, a_1 = 1, a_2 = -1$:

$$Y = a_1X_1 + a_2X_2 = X_1 - X_2$$

Từ đó ta có hệ quả sau:

Hệ quả 5.5.3. $E(X_1 - X_2) = E(X_1) - E(X_2)$ *với hai biến ngẫu nhiên bất kì* X_1 *và* X_2.
$V(X_1 - X_2) = V(X_1) + V(X_2)$ *nếu* X_1 *và* X_2 *độc lập.*

Ví dụ 5.18 Một nhà sản xuất ô tô trang bị một mô hình với một động cơ có 6 xy-lanh và động cơ 4 xy-lanh. Đặt X_1 và X_2 là lượng dầu cần tương ứng cho động cơ 6 xy-lanh và 4 xy-lanh. Với $\mu_1 = 22, \mu_2 = 26, \sigma_1 = 1, 2; \sigma_2 = 1, 5$ hãy tính $E(X_1 - X_2), V(X_1 - X_2), \sigma_{X_1-X_2}$? (đs: -4 ; 3,69 ; 1,92)

5.5.2 Trường hợp của biến ngẫu nhiên liên tục

Khi X_i là mẫu ngẫu nhiên lấy từ phân phối chuẩn, $\overline{X}$ và T_o thì cũng có phân phối chuẩn. Đây là kết quả tổng quát hơn liên quan đến tổ hợp tuyến tính.

Mệnh đề 5.5.4. *Nếu $X_1, X_2, ..., X_n$ độc lập, có phân phối chuẩn (có thể là trung bình khác nhau, phương sai khác nhau), thì bất kỳ tổ hợp tuyến tính của X_i cũng có phân phối chuẩn. Đặc biệt, hiệu của $X_1 - X_2$ cũng được phân phối chuẩn.*

Ví dụ 5.19 Theo ví dụ trên, cho tổng thu nhập khi bán 3 loại xăng ở 1 trạm là $Y = 3X_1 + 3,2X_2 + 3,4X_3$ và ta tính $\mu_Y = 5620$, $\sigma_Y = 429,46$. Nếu X_i có phân phối chuẩn thì xác suất để doanh thu hơn 4500 là:
$P(Y > 45000) = P(Z > \frac{4500-5620}{429,26}) = P(Z > -2,61) = 1 - \Phi(-2,61) = 0,9955$

CHÚ Ý: Trường hợp $n = 2$ thì ta có kết quả sau
$V(a_1X_1 + a_2X_2) = a_1^2V(X_1) + a_2^2V(X_2) + 2a_1a_2Cov(X_1, X_2)$

Chương 6

ƯỚC LƯỢNG ĐIỂM

6.1 Một số khái niệm tổng quan về ước lượng điểm

Các suy luận thống kê hầu hết xoay quanh việc rút ra các kết luận về một hay một vài tham số của tổng thể. Để thực hiện được điều này ta thường tiến hành một điều tra một mẫu của tổng thể về vấn đề đang quan tâm. Các kết luận được rút ra từ việc tính toán các giá trị của mẫu quan sát được.

Ví dụ ta ký hiệu μ là chiều cao trung bình của nữ sinh viên Trường Đại học Sư phạm Kỹ thuật. Quan sát một mẫu ngẫu nhiên gồm chiều cao của $n = 200$ sinh viên nữ của trường ta, thu được các giá trị $x_1, x_2, \ldots, x_{200}$. Trung bình mẫu $\overline{x}$ sẽ được sử dụng để đưa ra kết luận về μ.

Một cách tổng quát ta ký hiệu θ là tham số đang quan tâm, chưa biết và cần rút ra kết luận về θ. Giả sử ta cần ước lượng θ dựa trên mẫu ngẫu nhiên cỡ n là $X_1, X_2, \ldots, X_n$. Một thống kê trên mẫu ngẫu nhiên này là một hàm của các X_i, ví dụ như trung bình mẫu $\overline{X} = \dfrac{1}{n}\sum_{x=1}^{n} X_i$ là một biến ngẫu nhiên có thể sử dụng để đưa ra kết luận cho μ.

Điều này cũng đúng với trường hợp dữ liệu quan sát gồm nhiều mẫu. Ví dụ như có hai mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ và $Y_1, Y_2, \ldots, Y_m$; khi đó $\overline{X} - \overline{Y}$ là 1 hàm thống kê cho suy luận về sự khác biệt giữa 2 trung bình của 2 tổng thể là $\mu_1 - \mu_2$.

Định nghĩa 6.1. Một giá trị ước lượng điểm của tham số θ là một số mà số này được xem như là một giá trị hợp lý của θ. Một giá trị ước lượng điểm thu được bằng cách chọn một thống kê phù hợp và tính giá trị của thống kê này trên dữ liệu của mẫu. Khi đó hàm thống kê này được gọi là **ước lượng điểm** của θ.

Hàm thống kê là ước lượng điểm của θ tính toán trên mẫu thực nghiệm, kí hiệu là $\widehat{\theta}$, gọi là giá trị ước lượng điểm.

***Ví dụ* 6.1.** Quan sát thời gian sử dụng của máy điện thoại A (đơn vị: giờ) sau 10 lần sạc đầy pin ta có mẫu

$$x_1 = 5,5; x_2 = 5,33; x_3 = 6,2; x_4 = 6,5; x_5 = 7,0;$$

$$x_6 = 5,66; x_7 = 4,5; x_8 = 6,5; x_9 = 4,8; x_{10} = 5,0.$$

Gọi μ là thời gian sử dụng trung bình của máy điện thoại A sau sạc đầy pin.

$\widehat{\mu} = \overline{X} = \dfrac{1}{10}\sum_{x=1}^{10} X_i$ là một ước lượng điểm của μ.

$\overline{x} = \dfrac{1}{10}(5,5+5,33+6,2+6,5+7,0+5,66+4,5+6,5+4,8+5,0) = 5,699$ (giờ) là giá trị ước lượng điểm của μ ứng với ước lượng điểm $\overline{X}$.

Gọi Y là số lần máy điện thoại A có thời gian sử dụng từ 6 giờ trở lên sau n lần sạc pin đầy.

p là khả năng máy điện thoại A có thời gian sử dụng ít nhất là 6 giờ.

$\widehat{p} = \dfrac{Y}{n}$ là một ước lượng điểm của p.

$\dfrac{y}{n} = \dfrac{5}{10} = 0,5$ là một giá trị ước lượng điểm của p.

$\widetilde{X} = \dfrac{\min\limits_i X_i + \max\limits_i X_i}{2}$ cũng là một ước lượng điểm khác của trung bình μ.

$\widetilde{x} = \dfrac{4,5+7,0}{2} = 5,75$ cũng là một giá trị ước lượng điểm của trung bình μ tương ứng với ước lượng điểm $\widetilde{X}$.

$\widehat{\sigma}^2 = S^2 = \dfrac{\sum(X_i - \overline{X})^2}{n-1}$ là một ước lượng điểm của phương sai σ^2.

$s^2 = \dfrac{\sum(x_i - \overline{x})^2}{n-1} = 0,6820544444$ là một giá trị ước lượng điểm của phương sai σ^2 ứng với ước lượng điểm S^2.

$S'^2 = \dfrac{\sum(X_i - \overline{X})^2}{n}$ cũng là một ước lượng điểm khác của phương sai σ^2.

$s'^2 = \dfrac{\sum(x_i - \overline{x})^2}{n} = 0,613849$ là một giá trị ước lượng điểm của phương sai σ^2 ứng với ước lượng điểm S'^2.

Như vậy tương ứng với một tham số θ ta có thể có nhiều ước lượng điểm. Tối ưu nhất là ta cần tìm ước lượng điểm mà $\widehat{\theta} = \theta$. Tuy nhiên $\widehat{\theta}$ là một biến ngẫu nhiên nên giá trị ước lượng điểm có thể thay đổi trên các mẫu thực nghiệm khác nhau. Vậy làm thế nào lựa chọn một ước lượng điểm tốt nhất trong nhóm hữu hạn các ước lượng điểm, ta có một số tính chất mà một ước lượng điểm cần có như sau.

6.1.1 Ước lượng không chệch

Định nghĩa 6.2. Một ước lượng điểm $\widehat{\theta}$ được gọi là một ***ước lượng không chệch*** của θ nếu $E(\widehat{\theta}) = \theta$ với mọi giá trị có thể có của θ. Nếu $\widehat{\theta}$ là một ước lượng có chệch thì $E(\widehat{\theta}) - \theta$ được gọi là **sai số hệ thống của** θ.

Như vậy $\widehat{\theta}$ là một ước lượng không chệch của θ nếu phân phối xác suất của nó luôn có "trung tâm" tại giá trị đúng của tham số θ. "Trung tâm" ở đây là kỳ vọng của phân phối $\widehat{\theta}$.

Mệnh đề 6.3. *X là biến ngẫu nhiên có phân phối nhị thức với hai tham số n và p. Tỷ lệ mẫu $\widehat{p} = \dfrac{X}{n}$ là một ước lượng không chệch của p.*

Thật vậy, $E(\widehat{p}) = E\left(\dfrac{X}{n}\right) = \dfrac{1}{n}E(X) = \dfrac{1}{n}(np) = p.$

Mệnh đề 6.4. *Cho $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên có cùng phân phối với trung bình μ và phương sai σ^2. Khi đó ước lượng*

$$\widehat{\sigma}^2 = S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$$

là một ước lượng không chệch của σ^2.

Chứng minh

Với Y là một biến ngẫu nhiên, $V(Y) = E(Y^2) - [E(Y)]^2$.

Do đó ta có $E(Y^2) = V(Y) + [E(Y)]^2$.

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right] \\ E(S^2) &= \frac{1}{n-1}\left\{\sum E(X_i^2) - \frac{1}{n}E(\sum X_i)^2\right\} \\ &= \frac{1}{n-1}\left\{\sum(\sigma^2+\mu^2) - \frac{1}{n}\left\{V(\sum X_i) + [E(\sum X_i)]^2\right\}\right\} \\ &= \frac{1}{n-1}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}n\sigma^2 - \frac{1}{n}(n\mu)^2\right\} = \frac{1}{n-1}\left\{n\sigma^2 - \sigma^2\right\} = \sigma^2 \end{aligned}$$

Do đó S^2 là ước lượng không chệch của σ^2.

Mặt khác

$$E(S'^2) = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

nên

$$S'^2 = \frac{\sum(X_i - \overline{X})^2}{n}$$

là một ước lượng chệch của σ^2.

Khi chọn một ước lượng điểm trong nhiều ước lượng điểm của cùng một tham số ta chọn ước lượng không chệch.

$\widehat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ là một ước lượng điểm không chệch của μ. Thật vậy

$$\begin{aligned} E(\overline{X}) &= E\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) \\ &= \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\sum_{i=1}^n \mu = \frac{1}{n}n\mu = \mu \end{aligned}$$

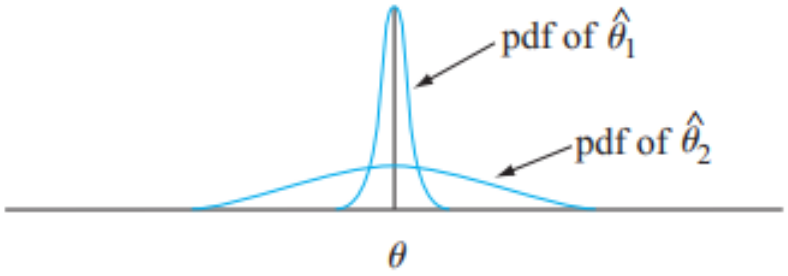
$\widehat{\mu}_2 = \dfrac{\min\limits_i X_i + \max\limits_i X_i}{2}$ cũng là một ước lượng điểm không chệch của μ. Vì

$$\begin{aligned} E(\widehat{\mu}_2) &= E\left(\frac{\min\limits_i X_i + \max\limits_i X_i}{2}\right) \\ &= \frac{1}{2}\Big(E(\min_i X_i) + E(\max_i X_i)\Big) \\ &= \frac{1}{2}\{\mu + \mu\} = \mu \end{aligned}$$

Vậy chúng ta cần chọn ước lượng điểm không chệch nào trong lớp các ước lượng điểm không chệch của θ? Để trả lời câu hỏi này ta xét mục tiếp theo.

6.1.2 Ước lượng hiệu quả

Giả sử $\widehat{\theta}_1$ và $\widehat{\theta}_2$ là hai ước lượng điểm không chệch của θ. Phân phối của hai ước lượng này đều có trung tâm tại giá trị đúng θ, tuy nhiên độ phân tán các giá trị có thể có của $\widehat{\theta}_1$, $\widehat{\theta}_2$ xung quanh giá trị θ có thể khác nhau. Nếu phương sai của $\widehat{\theta}_1$ nhỏ hơn phương sai của $\widehat{\theta}_2$ thì các giá trị có thể có của $\widehat{\theta}_1$ tập trung gần θ hơn các giá trị có thể có của $\widehat{\theta}_2$.



Hình 6.1: Đồ thị hai hàm phân phối của hai ước lượng điểm không chệch.

Định nghĩa 6.5. Trong lớp các ước lượng điểm không chệch của θ, ước lượng điểm $\widehat{\theta}$ có phương sai nhỏ nhất được gọi là ước lượng hiệu quả của θ so với các ước lượng không chệch khác.

Ước lượng hiệu quả có phương sai nhỏ nhất trong lớp tất cả các ước lượng không chệch của θ, hầu hết các giá trị của có thể có của ước lượng hiệu quả sẽ tập trung gần giá trị đúng θ.

Định lý 6.6. *Cho $X_1, X_2, \ldots, X_n$ là các biến ngẫu nhiên có phân phối chuẩn với hai tham số μ và σ^2. Khi đó ước lượng điểm $\widehat{\mu} = \overline{X}$ là một ước lượng hiệu quả của μ.*

6.1.3 Sai số chuẩn

Định nghĩa 6.7. ***Sai số chuẩn*** của ước lượng $\widehat{\theta}$ là độ lệch chuẩn của nó

$$\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}$$

Sai số chuẩn là tham số đo độ lệch chuẩn giữa một ước lượng điểm và giá trị đúng θ.

Nếu sai số chuẩn có chứa tham số chưa biết mà giá trị tham số đó có thể ước lượng thì ta sẽ ước lượng sai số chuẩn của một ước lượng. Ước lượng sai số chuẩn ký hiệu là $\widetilde{\sigma}_{\widehat{\theta}}$ hay $s_{\widehat{\theta}}$.

***Ví dụ* 6.2.** Cho biến ngẫu nhiên X có phân phối chuẩn với trung bình μ và phương sai σ^2.

Quan sát một mẫu ngẫu nhiên cỡ $n = 20$ ta thu được mẫu thực nghiệm:

1,65	1,71	1,66	1,69	1,72	1,68	1,64	1,66	1,71	1,74
1,62	1,64	1,61	1,69	1,73	1,76	1,70	1,69	1,67	1,68

$\widehat{\mu} = \overline{X}$ là một ước lượng điểm tốt nhất của μ với giá trị ước lượng là $\overline{x} = 1,6825$

Nếu ta biết $\sigma = 0,05$ thì sai số chuẩn của $\overline{X}$ là:

$\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0,05}{\sqrt{20}} = 0,0111803399.$

Nếu σ chưa biết, giá trị ước lượng của σ là $\widehat{\sigma} = s = 0,03971941061$ thì ta có ước lượng sai số chuẩn của ước lượng $\overline{X}$ là:

$\widehat{\sigma}_{\overline{x}} = s_{\overline{X}} = \dfrac{s}{\sqrt{n}} = 8,881530214.10^{-3}.$

Giả sử X là biến ngẫu nhiên có phân phối nhị thức với 2 tham số n và p. Tỉ lệ mẫu $\widehat{p} = \dfrac{X}{n}$ là ước lượng điểm tốt cho p với sai số chuẩn.

$$\sigma_{\widehat{p}} = \sqrt{V\left(\frac{X}{n}\right)} = \sqrt{\frac{V(X)}{n^2}} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}.$$

Khi một ước lượng điểm có phân phối xấp xỉ chuẩn, điều này thường xảy ra khi n lớn thì giá trị đúng của θ có nhiều khả năng thuộc vào khoảng với độ rộng là 2 sai số chuẩn của θ. Ví dụ như với mẫu có $n = 36$ có giá trị trung bình mẫu $\overline{x} = 29,35$ và độ lệch chuẩn mẫu $s = 0,36$ khi đó tính được ước lượng sai số chuẩn của $\widehat{\mu}$ là $\dfrac{s}{\sqrt{n}} = \dfrac{0,36}{\sqrt{36}} = 0,6$. Khi đó μ nhiều khả năng thuộc khoảng $29,35 \pm (2)(0,6) = (28,15; 30,55)$.

Nếu ước lượng không chệch $\widehat{\theta}$ không xấp xỉ chuẩn ta không thể sử dụng công thức trên để đưa ra khoảng giá trị để dự đoán θ thuộc vào. Để khắc phục tình huống này gần đây trong thống kê hiện đại có một phương pháp gọi là bootstrap. Phương pháp bootstrap là phương pháp coi mẫu gốc ban đầu đóng vai trò tổng thể mà từ đó nó được rút ra. Từ mẫu ban đầu lấy lại các mẫu ngẫu nhiên cùng cỡ với mẫu gốc bằng phương pháp lấy mẫu có hoàn lại, gọi là mẫu bootstrap. Với mỗi mẫu lấy lại

ta tính được giá trị tham số thống kê quan tâm gọi là tham số bootstrap. Sự phân bố của các tham số thống kê mẫu bootstrap là phân phối bootstrap.

Lấy mẫu có hoàn lại có nghĩa là sau khi chúng ta rút ra ngẫu nhiên một quan sát từ mẫu ban đầu, ta đặt nó trở lại trước khi lấy quan sát tiếp theo. Kết quả là, bất kỳ số có thể được rút ra một lần, nhiều hơn một lần, hoặc không được rút ra lần nào.

Với mẫu ban đầu $x = (x_1, x_2, \ldots, x_n)$, ta có mẫu bootstrap $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ với mỗi giá trị x_i^* được lấy ngẫu nhiên từ tập các giá trị $x_1, x_2, \ldots, x_n$ với xác suất $\frac{1}{n}$.

Tương ứng với mỗi mẫu bootstrap x^* ta có mô phỏng bootstrap của $\widehat{\theta}$ là $\widehat{\theta}^* = T(x^*)$, với hàm thống kê $T(x^*)$ tương tự với hàm thống kê $T(x)$ tác động lên mẫu x.

Mẫu lấy lại thứ nhất: $x_1^* = (x_{1_1}^*, x_{1_2}^*, \ldots, x_{1_n}^*)$ ta tính được giá trị $\widehat{\theta}_1^*$.

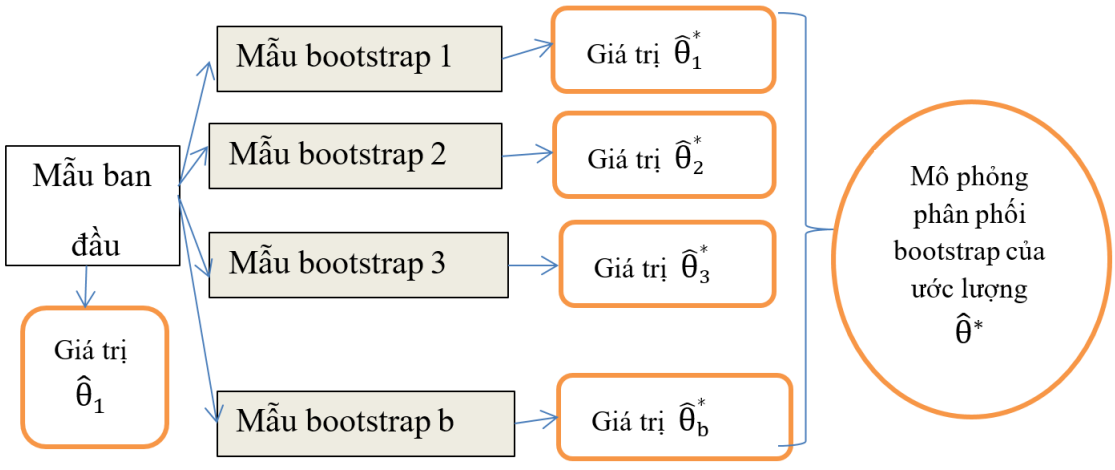
Mẫu lấy lại thứ hai: $x_2^* = (x_{2_1}^*, x_{2_2}^*, \ldots, x_{2_n}^*)$ ta tính được giá trị $\widehat{\theta}_2^*$.

...

Mẫu lấy lại thứ B: $x_B^* = (x_{B_1}^*, x_{B_2}^*, \ldots, x_{B_n}^*)$ ta tính được giá trị $\widehat{\theta}_B^*$.

Với $B = 1000$ hay 2000.

Với mẫu bootstrap ngẫu nhiên $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$, $\widehat{\theta}^* = T(X^*)$ là một thống kê trên mẫu bootstrap, khi đó $F^*(t) = P(\widehat{\theta}^* \leq t)$ là phân phối bootstrap của $\widehat{\theta}^*$.



Hình 6.2: Sơ đồ mô phỏng phân phối bootstrap.

Các bước dùng phương pháp bootstrap ước lượng sai số tiêu chuẩn của $\widehat{\theta}$ từ một mẫu gốc ban đầu:

Bước 1: Lấy theo phương pháp có hoàn lại từ mẫu gốc ban đầu được B mẫu bootstrap độc lập cùng cỡ với mẫu gốc $x_k^* = \left(x_{k_1}^*, x_{k_2}^*, \ldots, x_{k_n}^*\right), \quad k = 1, 2, \ldots, B.$

Bước 2: Với mỗi mẫu bootstrap có được ở bước 1 ta tính giá trị thống kê $\widehat{\theta}_k^* = T(x_k^*), \quad k = 1, 2, \ldots, B.$

Bước 3: Tính độ lệch chuẩn của B giá trị tính được ở bước 2.

$$S_{\widehat{\theta}} = \sqrt{\frac{1}{B-1}\sum(\widehat{\theta}_i^* - \overline{\theta}^*)^2} \qquad \text{với} \quad \overline{\theta}^* = \frac{1}{B}\sum \widehat{\theta}_i^*.$$

Độ lệch tiêu chuẩn này là ước lượng bootstrap của sai số tiêu chuẩn $\sigma_{\widehat{\theta}}$. Sai số chuẩn của ước lượng bootstrap là độ lệch chuẩn mẫu của giá trị $\widehat{\theta}_i^*$.

Ta có giá trị $S_{\widehat{\theta}}$ xấp xỉ $\sigma_{\widehat{\theta}}$ khi số lượng mẫu bootstrap B là lớn.

6.2 Các phương pháp ước lượng điểm

6.2.1 Phương pháp Moment

Định nghĩa 6.8. Cho $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên có cùng phân phối xác suất.

Moment thứ k của X (moment tổng thể hay moment lý thuyết) là $E(X^k)$.

Moment mẫu thứ k của X là $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$.

Cụ thể ta có:

- Moment thứ nhất của X là $E(X) = \mu$.
- Moment mẫu thứ nhất của X là $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.
- Moment thứ hai của X là $E(X^2)$.
- Moment mẫu thứ hai của X là $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$.

Định nghĩa 6.9. Cho $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên có cùng phân phối xác suất có hàm xác suất hay mật độ xác suất $f(x; \theta_1, \theta_2, \ldots, \theta_m)$ phụ thuộc vào m tham số chưa biết là $\theta_1, \theta_2, \ldots, \theta_m$. Ước lượng moment của m tham số này thu được bằng cách giải m phương trình m moment cấp k đầu tiên bằng m mẫu moment cùng bậc tương ứng.

***Ví dụ* 6.3.** Cho X là biến ngẫu nhiên có phân phối Bernoulli với tham số p chưa biết. X có hàm xác suất

$$p_X(x) = \begin{cases} p & \text{với } x = 1, \\ 1-p & \text{với } x = 0. \end{cases}$$

Quan sát một mẫu ngẫu nhiên có cỡ $n = 10$ ta có bộ số liệu $(1, 1, 0, 1, 0, 1, 1, 1, 0, 0)$. Hãy tìm ước lượng moment của p.

Giải:

Moment thứ nhất của X là $E(X) = p$.

Moment mẫu thứ nhất của X là $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = 0,6$.

Giải phương trình $E(X) = \overline{X}$ ta có $\widehat{p} = \overline{X}$ là ước lượng moment của p với giá trị ước lượng là $0,6$.

***Ví dụ* 6.4.** Thời gian hoạt động của loại thiết bị A là biến ngẫu nhiên X (đơn vị: giờ) có phân phối mũ với tham số λ. Quan sát thời gian hoạt động của một số thiết bị A ta có số liệu

$$(15,6; 16,0; 16,5; 15,8; 16,2; 17,5; 18,5; 17,8; 18,2; 16,8; 17,4; 17,9; 16,8; 16,0; 16,6)$$

Hãy tìm ước lượng của moment của λ.

Giải:

Moment thứ nhất của X là $E(X) = \dfrac{1}{\lambda}$.

Giá trị moment mẫu thứ nhất của X là $\overline{x} = 16,9$ (giờ).

$E(X) = \overline{X} \Rightarrow \widehat{\lambda} = \dfrac{1}{\overline{X}}$ là một ước lượng moment của λ.

Vậy giá trị ước lượng moment của λ là $\dfrac{1}{16,9}$.

***Ví dụ* 6.5.** Điểm thi xác suất của sinh viên trường Đại học M là biến ngẫu nhiên X có phân phối chuẩn $N(\mu, \sigma^2)$. Quan sát một mẫu ngẫu nhiên gồm $n = 1018$ bài thi của sinh viên trường M ta có bảng số liệu:

Điểm số	0	1	2	3	4	5	6	7	8	9	10
Số bài	2	24	50	98	151	193	189	162	95	46	8

Hãy tìm ước lượng moment của μ và σ^2.

Giải:

Moment cấp 1 của X là $E(X) = \mu$ với moment mẫu tương ứng là

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

Moment cấp 2 của X là $E(X^2) = \sigma^2 + \mu^2$ với moment mẫu tương ứng là

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Giá trị moment mẫu cấp 1 của X là $\overline{x}_n = 5,411591356$.

Giá trị moment mẫu cấp 2 của X là $\dfrac{1}{n}\sum_{i=1}^{n} x_i^2 = 33,09921415$.

Giải hệ phương trình $\begin{cases} \mu = \overline{X} \\ \sigma^2 + \mu^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}$

Ta có ước lượng moment của μ là $\widehat{\mu} = \overline{X}$ và giá trị ước lượng moment của μ bằng $5,411591356$.

Ước lượng moment của phương sai σ^2 là $\widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2$ và giá trị ước lượng moment của σ^2 bằng $3,813893141$.

6.2.2 Phương pháp ước lượng hợp lý cực đại

Định nghĩa 6.10. Cho $X_1, X_2, \ldots, X_n$ có cùng phân phối $f(x_1, x_2, \ldots, x_n, \theta_1, \theta_2, \ldots, \theta_m)$ với $\theta_1, \theta_2, \ldots, \theta_m$ là các tham số chưa biết. Hàm $f(x_1, x_2, \ldots, x_n, \theta_1, \theta_2, \ldots, \theta_m)$ là hàm của các giá trị quan sát được $x_1, x_2, \ldots, x_n$ và $\theta_1, \theta_2, \ldots, \theta_m$ được gọi là hàm hợp lý. Ước lượng $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_m$ làm cực đại hàm hợp lý gọi là ước lượng hợp lý cực đại của $\theta_1, \theta_2, \ldots, \theta_m$.

***Ví dụ* 6.6.** Tỷ lệ sản phẩm đạt chuẩn của nhà máy B là p chưa biết. Một người kiểm tra từng sản phẩm cho đến khi được 10 sản phẩm đạt chuẩn thì dừng. Biết người này dừng lại ở lần kiểm tra thứ 12. Hãy tìm ước lượng hợp lý cực đại của p.

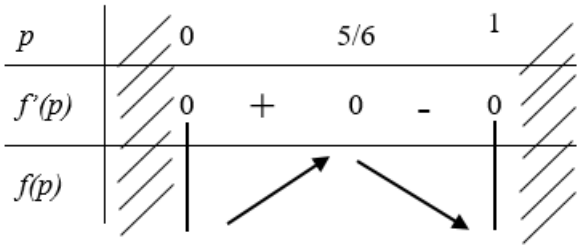
Giải:

Gọi X là số lần kiểm tra cho đến khi được 10 sản phẩm đạt chuẩn thì dừng $\Rightarrow X$ có phân phối nhị thức âm với hai hàm số là $r = 10$ và p.

$P(X=12) = C_{11}^{10} p^{10}(1-p)^2 = f(p)$ làm hàm hợp lý cực đại của p.

$$\begin{aligned} f'(p) &= 11[10p^9(1-p)^2 - 2p^{10}(1-p)] \\ &= 11p^9(1-p)[10(1-p) - 2p] \\ &= 11p^9(1-p)(10-12p) \end{aligned}$$

$f'(p) = 0$ tại $p = 0$ hoặc $p = 1$ hoặc $p = \dfrac{10}{12} = \dfrac{5}{6}$



$f(p)$ đạt cực đại tại 5/6. Vậy giá trị ước lượng hợp lý cực đại của p tương ứng với quan sát người kiểm tra dừng lại ở lần kiểm tra thứ 12 là 5/6.

***Ví dụ* 6.7.** Giả sử $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên có phân phối mũ với tham số λ

$X_1, X_2, \ldots, X_n$ độc lập nên ta có hàm hợp lý

$$\begin{aligned} f(x_1, x_2, \ldots, x_n; \lambda) &= (\lambda e^{-\lambda x_1}) \ldots (\lambda e^{-\lambda x_n}) \\ &= \lambda^n e^{-\lambda \sum x_i} \end{aligned}$$

$$ln[f(x_1, x_2, \ldots, x_n; \lambda)] = nln(\lambda) - \lambda \sum x_i$$

Đạo hàm $ln[f(x_1, x_2, \ldots, x_n; \lambda)]$ theo λ và cho bằng 0 ta có

$$\frac{n}{\lambda} - \sum x_i = 0 \Leftrightarrow \lambda = \frac{n}{\sum x_i} = \frac{1}{\overline{x}}$$

Suy ra ta có ước lượng hợp lý cực đại của λ là $\widehat{\lambda} = \dfrac{1}{\overline{X}}$

Kết quả này tương tự với kết quả thu được bằng phương pháp moment.
Tuy nhiên vì $E(\dfrac{1}{\overline{X}}) \neq \dfrac{1}{E(\overline{X})}$ nên đây là một ước lượng chệch của λ.

Bài tập 6.1

1. Quan sát điểm môn Toán của sinh viên trường Đại học A ta có bảng số liệu

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

   (a) Tìm một giá trị ước lượng điểm trung bình môn Toán của sinh viên trường Đại học A, chỉ ra ước lượng điểm tương ứng.

   (b) Tìm một giá trị của ước lượng điểm của tỷ lệ sinh viên trường Đại học A có điểm Toán từ 5 trở lên, chỉ ra ước lượng điểm tương ứng.

   (c) Tìm một giá trị ước lượng điểm của độ lệch chuẩn σ có điểm thi môn Toán sinh viên trường Đại học A, chỉ ra ước lượng điểm tương ứng.

   (d) Tìm một giá trị ước lượng điểm của tham số $\frac{\sigma}{\mu}$, chỉ ra ước lượng điểm tương ứng.

2. Quan sát số xe bán ra trong một ngày tại một số cửa hàng ta có bộ số liệu:

Số xe bán ra	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Số cửa hàng	3	5	7	10	11	15	18	21	19	17	15	13	12	10

   (a) Tìm một giá trị ước lượng điểm cho số xe trung bình bán ra trong một ngày tại các cửa hàng, chỉ ra ước lượng điểm tương ứng.

   (b) Tìm một giá trị ước lượng điểm cho tỉ lệ cửa hàng có số xe bán ra trong một ngày từ 10 xe trở lên, chỉ ra ước lượng điểm tương ứng

   (c) Tìm một giá trị của ước lượng điểm cho phương sai của số xe bán ra trong một ngày tại các cửa hàng, chỉ ra ước lượng điểm tương ứng.

   (d) Tìm một giá trị ước lượng điểm cho trung vị của số xe bán ra trong một ngày tại các cửa hàng, chỉ ra ước lượng điểm tương ứng.

   (e) Xác định sai số chuẩn của ước lượng trong câu (d).

3. (a) Quan sát số lượng ga (đơn vị: therm) sử dụng tại 10 ngôi nhà được chọn ngẫu nhiên tại vùng A trong tháng 1 , ta có bộ số liệu $103, 156, 118, 89, 125, 147, 122, 99, 138, 90$. Ký hiệu μ là lượng ga trung bình sử dụng tại mỗi nhà có sử dụng ga trong toàn bộ vùng A. Tính một giá trị ước lượng điểm của μ.

   (b) Giả sử có 10000 hộ sử dụng ga tại vùng A. Ký hiệu T là tổng lượng ga được sử dụng tại vùng A trong tháng 1. Ước lượng T với dữ liệu đã cho trong phần (a) và chỉ ra ước lượng điểm đã dùng.

   (c) Sử dụng dữ liệu trong phần (a), ước lượng tỉ lệ p là tỷ lệ hộ sử dụng ít nhất 100(therm)ga trong tháng 1.

4. Quan sát ngẫu nhiên n_1 nam giới hút thuốc, có X_1 người hút thuốc có đầu lọc. Quan sát ngẫu nhiên n_2 nữ giới hút thuốc, có X_2 người hút thuốc có đầu lọc. Ký hiệu p_1 và p_2 là tỷ lệ nam, nữ hút thuốc có đầu lọc trong số các nam giới, nữ giới có thuốc.

   (a) Chỉ ra rằng $\left(\frac{X_1}{n_1}\right) - \left(\frac{X_2}{n_2}\right)$ là một ước lượng không chệch của $p_1 - p_2$.

   (b) Xác định sai số chuẩn của ước lượng trong phần (a). Sử dụng các giá trị x_1, x_2 là các giá trị quan sát của các biến ngẫu nhiên X_1, X_2 để tính giá trị sai số chuẩn của ước lượng này.

   (c) Cho $n_1 = n_2 = 200, x_1 = 126, x_2 = 176$ sử dụng ước lượng trong phần a tính giá trị ước lượng của $p_1 - p_2$.

5. Cho mẫu ngẫu nhiên $X_1, \ldots, X_n$ có cùng phân phối xác suất, xác định bởi hàm mật độ xác suất

$$f(x, \theta) = 0,5(1 + \theta x) \qquad -1 \leq x \leq 1$$

với $-1 \leq \theta \leq 1$. Chỉ ra rằng $\widehat{\theta} = 3\overline{X}$ là một ước lượng không chệch của θ.

Bài tập 6.2

1. Một xét nghiệm được áp dụng cho n người chắc chắn không bệnh. Gọi X là số xét nghiệm dương tính (tức là có bệnh) trong n kết quả tức là X là số kết quả dương tính sai. Và p là tỉ lệ người không có bệnh bị kết luận là có bệnh sau xét nghiệm.

(a) Xác định ước lượng hợp lý cực đại của p. Nếu $n = 20$ và $x = 3$. Xác định giá trị của hợp lý cực đại của p.

(b) Ước lượng điểm trong phần (a) có là ước lượng không chệch không?

(c) Nếu $n = 20$ và $x = 3$ hãy xác định ước lượng hợp lý cực đại của xác suất $(1-p)^5$ là không có kết quả dương tính nào trong 5 xét nghiệm đối với 5 người không có bệnh.

2. Cho biến ngẫu nhiên có hàm mật độ xác suất

$$f(x,\theta) = \begin{cases} (\theta+1)x^\theta & 0 \le x \le 1, \\ 0 & \text{trường hợp ngược lại.} \end{cases}$$

với $\theta > -1$. Quan sát một mẫu ngẫu nhiên cỡ n của X ta thu được bộ số liệu

$$x_1 = 0,92; x_2 = 0,79; x_3 = 0,65; x_4 = 0,9; x_5 = 0,86;$$

$$x_6 = 0,47; x_7 = 0,73; x_8 = 0,97; x_9 = 0,76; x_{10} = 0,65.$$

(a) Sử dụng phương pháp moment tìm một ước lượng điểm của θ và tính giá trị của ước lượng này theo số liệu trên.

(b) Tìm ước lượng hợp lý cực đại của θ rồi tính giá trị của lượng này theo số liệu trên.

3. X_1, X_2 là các biến ngẫu nhiên có phân phối Poisson với tham số λ_1, λ_2. Xác định ước lượng hợp lý cực đại của λ_1, λ_2 và $\lambda_1 - \lambda_2$.

4. Cho một mẫu ngẫu nhiên $X_1, X_2, ..., X_n$ có hàm mật độ xác suất

$$f(x,\lambda,\theta) = \begin{cases} \lambda e^{-\lambda(x-\theta)} & x \ge 0, \\ 0 & \text{trường hợp ngược lại.} \end{cases}$$

(a) Xác định ước lượng hợp lý cực đại của θ và λ.

(b) Với mẫu thực nghiệm $3,11; 0,64; 2,55; 2,20; 5,44; 3,42; 10,39; 8,93; 17,82$ và $1,3$ tính giá trị ước lượng của θ và λ.

Chương 7

ƯỚC LƯỢNG KHOẢNG

Giới thiệu

Một ước lượng điểm, là 1 con số, tự bản thân nó không cung cấp bất kỳ thông tin nào về độ chính xác và độ tin cậy của ước lượng đó. Ví dụ như dùng thống kê $\overline{X}$ để tính ước lượng điểm cho giá trị trung bình đúng μ, giả sử dựa trên mẫu thực nghiệm thu được giá trị ước lượng điểm $\overline{x} = 9322,7$. Vì mẫu thay đổi nên hầu như không xảy ra trường hợp $\overline{x} = \mu$. Hơn nữa, ước lượng điểm không nói đến việc $\overline{x}$ gần hay xa μ.

Một giải pháp thay thế cho việc đưa ra một giá trị hợp lý duy nhất cho tham số là tính một khoảng giá trị hợp lý mà giá trị cần ước lượng thuộc vào, khoảng này gọi là khoảng ước lượng hay khoảng tin cậy. Để tính khoảng tin cậy cần xác định mức độ tin cậy của khoảng. Khoảng ước lượng với độ tin cậy 95% cho trung bình đúng có giới hạn dưới là 9162,5 và giới hạn trên 9482,9 thì μ có thể là bất kỳ giá trị nào thuộc khoảng (9162,5; 9482,9) với độ tin cậy 95%. Độ tin cậy 95% nghĩa là 95% của tất cả các mẫu sẽ cho ra khoảng ước lượng chứa μ hoặc bất kỳ tham số nào được ước lượng. Độ tin cậy thường dùng là 95%, 99% và 90%.

Khoảng ước lượng với độ tin cậy cao và độ rộng hẹp cho thông tin về tham số cần ước lượng khá chính xác. Ngược lại khoảng tin cậy rộng và độ tin cậy thấp cho thấy thông tin không chắc chắn về tham số cần ước lượng.

7.1 Các tính chất cơ bản của khoảng tin cậy

Giả sử tham số quan tâm là trung bình μ của một tổng thể và giả sử:

1. Tổng thể có phân phối chuẩn.

2. Giá trị độ lệch tiêu chuẩn của tổng thể là σ đã biết.

Thông thường giả thiết tổng thể có phân phối chuẩn là hợp lý, tuy nhiên nếu chưa biết giá trị của μ mà giá trị σ đã biết thường thì không đủ độ tin cậy.

Mẫu thực nghiệm $x_1, x_2, \ldots, x_n$ là kết quả quan sát của mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ có phân phối chuẩn với trung bình μ và độ lệch chuẩn σ. Khi đó biến ngẫu nhiên trung bình mẫu $\overline{X}$ có phân phối chuẩn với trung bình μ và độ lệch chuẩn $\sigma/\sqrt{n}$. Biến đổi chuẩn tắc hóa ta có

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{7.1}$$

Do đó các giá trị có thể có của Z thuộc khoảng -1,96 đến 1,96 với khả năng là 95%.

$$P\left(-1,96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1,96\right) = 0,95 \tag{7.2}$$

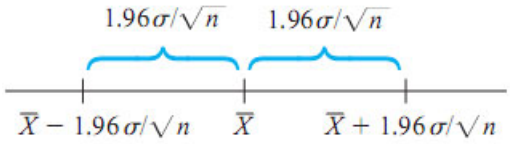
Biến đổi tương đương ta có

$$P\left(\overline{X} - 1,96.\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1,96.\frac{\sigma}{\sqrt{n}}\right) = 0,95 \tag{7.3}$$

Khoảng ngẫu nhiên

$$\left(\overline{X} - 1,96.\frac{\sigma}{\sqrt{n}}; \overline{X} + 1,96.\frac{\sigma}{\sqrt{n}}\right) \tag{7.4}$$

có trung tâm tại biến ngẫu nhiên $\overline{X}$.



Hình 7.1: Khoảng ngẫu nhiên $(\overline{X} - 1,96.\frac{\sigma}{\sqrt{n}}; \overline{X} + 1,96.\frac{\sigma}{\sqrt{n}})$ có trung tâm tại $\overline{X}$.

Định nghĩa 7.1. Từ mẫu thực nghiệm $x_1, x_2, \ldots, x_n$, tính được giá trị trung bình mẫu $\overline{x}$, thay $\overline{X}$ trong (7.4) bằng $\overline{x}$, khoảng cố định thu được gọi là khoảng tin cậy 95% cho μ

$$(\overline{x} - 1,96.\frac{\sigma}{\sqrt{n}}, \overline{x} + 1,96.\frac{\sigma}{\sqrt{n}})$$

Hay

$$\overline{x} - 1,96.\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1,96.\frac{\sigma}{\sqrt{n}}$$

với độ tin cậy 95% cho μ. Biểu diễn ngắn gọn là $\overline{x} \pm 1,96.\frac{\sigma}{\sqrt{n}}$, khi đó tương ứng với dấu - là điểm cuối bên trái (giới hạn dưới) và tương ứng với dấu + là điểm cuối bên phải (giới hạn trên).

Ví dụ 7.1 Sử dụng dữ liệu $\sigma = 3$, cỡ mẫu $n = 36$ và trung bình mẫu $\overline{x} = 80$ để tính khoảng tin cậy 95% cho chiều cao trung bình μ ta có kết quả là

$$\overline{x} \pm \frac{\sigma}{\sqrt{n}} = 80 \pm 1,96\frac{3}{\sqrt{36}} = 80 \pm 0,98 = (79,02; 80,98)$$

Tức là với độ tin cậy 95% thì $79,02 < \mu < 80,98$; khoảng ước lượng khá hẹp tức là độ chính xác của ước lượng của μ là khá cao.

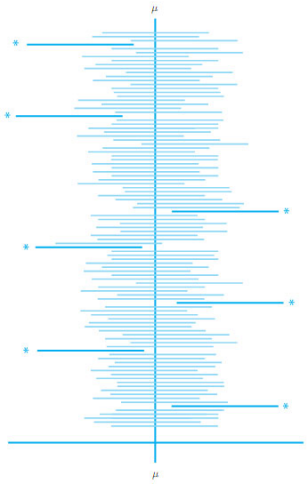
Khoảng giá trị $(79,02; 80,98)$ là một khoảng cố định không là khoảng ngẫu nhiên và μ là một hằng số (chưa biết) do đó khi viết $P(\mu \in (79,02; 80,98)) = 0,95$ là không chính xác.

Một giải thích chính xác của "độ tin cậy 95%" được dựa trên quan hệ tần số của xác suất, để nói một biến cố A có xác suất 0,95 tức là nếu thử nghiệm trên A được lặp đi lặp lại về lâu dài thì khả năng A sẽ xảy ra 95%. Giả sử lấy mẫu chiều cao của những người khác sẽ tính được một khoảng ước lượng với độ tin cậy 95% khác và lặp lại công việc này cho mẫu thứ 3, mẫu thứ 4, mẫu thứ 5,... Gọi A là biến cố $\overline{X} - 1,96.\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1,96.\frac{\sigma}{\sqrt{n}}$. Khi đó $P(A) = 0,95$ theo tính toán thì 95% của khoảng tin cậy sẽ chứa μ, điều này được minh họa trong hình sau khi đường thẳng đứng cắt trục đo tại giá trị μ (chưa biết) chú ý rằng 7 trong 100 khoảng sẽ không chứa μ tức là chỉ 5% của các khoảng được biểu diễn như vậy sẽ không chứa μ.

Dùng công thức khoảng tin cậy mặc dù đôi khi không thoả mãn yêu cầu bài toán. Khó khăn ở đây là ta cần giải thích về xác suất khi áp dụng trong một chuỗi các thí nghiệm chứ không phải là một khoảng duy nhất.

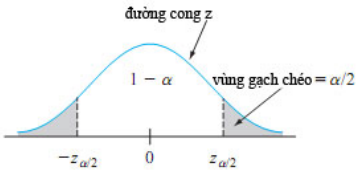
7.1.1 Các khoảng tin cậy khác

Độ tin cậy 95% được suy từ xác suất 0,95 trong bất đẳng thức (7.2), nếu muốn độ tin cậy 99% thì xác suất ban đầu phải thay bằng 0,99 khi đó cần phải thay giá trị phân vị mức của Z từ 1,96 thành 2,58 trong công thức tính độ tin cậy 95%. Trong



Hình 7.2: Khoảng tin cậy 95% (dấu * là khoảng không chứa μ)

thực tế muốn có độ tin cậy bao nhiêu thì thay 1,96 hay 2,58 bằng giá trị phân vị mức thích hợp của phân phối chuẩn.



Hình 7.3: $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

Định nghĩa 7.2. Khoảng tin cậy $100(1-\alpha)\%$ cho trung bình μ của một tổng thể có phân phối chuẩn với giá trị của σ cho trước là

$$\left(\overline{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) \tag{7.5}$$

hay $\overline{x} \pm z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$.

Công thức (7.5) cho khoảng tin cậy có thể phát biểu thành lời như sau

Ước lượng điểm của $\mu \pm$ (giá trị phân biệt của Z)(sai số chuẩn của kỳ vọng)

Ví dụ 7.2 Qui trình sản xuất một loại động cơ cụ thể gần đây đã được sửa đổi, lịch sử dữ liệu cho biết phân phối đường kính của lỗ ổ cắm trên vỏ động cơ có phân

phối chuẩn với độ lệch chuẩn 0,1 mm. Người ta tin rằng việc thay đổi này không ảnh hưởng đến phân phối cũng như độ lệch chuẩn, tuy nhiên giá trị trung bình của đường kính có thể thay đổi. Một mẫu gồm 100 động cơ được chọn, xác định được đường kính các lỗ ổ cắm trong mẫu là 5,426 mm. Yêu cầu đặt ra là tính khoảng tin cậy cho trung bình các đường kính lỗ với độ tin cậy 90%.

Giải

Độ tin cậy $100(1-\alpha)\%$ với $\alpha = 0,1$

suy ra giá trị phân vị mức $z_{\alpha/2} = z_{0,05} = 1,645$. Khoảng ước lượng cần tìm là

$$5,426 \pm (1,645).\frac{0,1}{\sqrt{100}} = 5,426 \pm 0,01645 = (5,40955; 5,44245)$$

7.1.2 Độ tin cậy, độ chính xác và cỡ mẫu

Độ tin cậy càng cao thì độ dài khoảng ước lượng càng rộng. Độ tin cậy 95% thì khoảng ước lượng kéo dài $1,96.\frac{\sigma}{\sqrt{n}}$ cho mỗi bên của $\overline{x}$, độ rộng của khoảng là $2(1,96).\frac{\sigma}{\sqrt{n}} = 3,92.\frac{\sigma}{\sqrt{n}}$. Tương tự độ rộng của khoảng tin cậy 99% là $2(2,58).\frac{\sigma}{\sqrt{n}} = 5,16.\frac{\sigma}{\sqrt{n}}$. Nghĩa là ta tin cậy nhiều hơn vì khoảng rộng hơn. **Độ rộng** của khoảng ước lượng với độ tin cậy $100(1-\alpha)\%$ là

$$w_0 = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Một nửa độ rộng của khoảng ước lượng $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ của độ tin cậy $100(1-\alpha)\%$ đôi khi được gọi là **sai số** hay **độ chính xác** của khoảng ước lượng. Sai số của khoảng tin cậy $100(1-\alpha\%)$ là

$$\varepsilon = \frac{w_0}{2} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Với một mẫu cố định từ độ tin cậy ta có thể xác định được độ chính xác của khoảng ước lượng và ngược lại biết độ rộng của khoảng ước lượng thì xác định được độ tin cậy của khoảng ước lượng. Còn với độ tin cậy cho trước cố định cỡ mẫu thay đổi sẽ làm độ rộng của khoảng ước lượng thay đổi.

Ví dụ 7.3 Quan sát thời gian chia sẻ lệnh của một hệ thống máy tính biết rằng thời gian cụ thể cho 1 lệnh chỉnh sửa có phân phối chuẩn với độ lệch tiêu chuẩn là 25 milisec. Một hệ điều hành mới được cài đặt, ta cần ước lượng thời gian trung bình thực sự là μ. Trong điều kiện mới giả sử thời gian vẫn có phân phối chuẩn với $\sigma = 25$. Xác định cỡ mẫu cần thiết để có độ tin cậy 95% và độ rộng (tối đa) là 10.

Giải

Cỡ mẫu cần thỏa mãn

$$10 = 2.(1,96)(25/\sqrt{n})$$

Suy ra

$$\sqrt{n} = 2.(1,96)(25)/10 = 9,80$$

Do đó

$$n = (9,80)^2 = 96,04$$

Vì n là số nguyên nên ta cần cỡ mẫu $n = 97$.

Cỡ mẫu cần cho khoảng tin cậy trong (7.5) với độ rộng w là $n = \left(2.z_{\alpha/2}.\frac{\sigma}{w}\right)^2$

7.1.3 Nguồn gốc của khoảng tin cậy

Cho mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ là mẫu mà trên đó tìm khoảng tin cậy của tham số θ. Giả sử tìm được biến ngẫu nhiên $h(X_1, X_2, \ldots, X_n; \theta)$ thỏa mãn 2 tính chất sau:

1. Là hàm phụ thuộc của cả $X_1, X_2, \ldots, X_n$ và θ.

2. Phân phối xác suất của biến này không phụ thuộc θ hay bất kỳ tham số nào.

Ví dụ như nếu phân phối của tổng thể là phân phối chuẩn khi biết σ và $\theta = \mu$ thì biến $h(X_1, X_2, \ldots, X_n; \mu) = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ thỏa cả hai tính chất là hàm phụ thuộc theo μ nhưng có phân phối chuẩn tắc nên không phụ thuộc μ.

Cho α bất kỳ thuộc 0 và 1 ta tìm được hằng số a và b thỏa mãn

$$P(a < h(X_1, X_2, \ldots, X_n; \theta) < b) = 1 - \alpha \tag{7.6}$$

Theo tính chất 2, a và b không phụ thuộc θ, theo ví dụ mẫu lấy từ tổng thể có phân phối chuẩn thì $a = -z_{\alpha/2}, b = z_{\alpha/2}$. Biến đổi (7.6) theo θ ta được xác suất

$$P(l(X_1, X_2, \ldots, X_n) < \theta < u(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$

Khi đó $l(x_1, x_2, \ldots, x_n)$ và $u(x_1, x2, \ldots, x_n)$ là giới hạn trên và giới hạn dưới của khoảng tin cậy với độ tin cậy $100(1-\alpha)\%$ trong ví dụ mẫu lấy từ tổng thể có phân phối chuẩn $l(X_1, X_2, \ldots, X_n) = \overline{X} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$ và $u(X_1, X_2, \ldots, X_n) = \overline{X} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$.

Ví dụ 7.4 Mô hình lý thuyết cho rằng thời gian để chất lỏng cách điện ở các điện cực bị phân hủy với một điện áp đặc biệt có phân phối mũ với tham số λ. Một mẫu ngẫu nhiên có kích thước $n = 10$ với chuỗi thời gian phân hủy có dữ liệu (theo phút) như sau $x_1 = 41,53; x_2 = 18,73; x_3 = 2,99; x_4 = 30,34; x_5 = 12,33; x_6 = 117,52; x_7 = 73,02; x_8 = 223,63; x_9 = 4; x_{10} = 26,78$. Với độ tin cậy 95% hãy tìm khoảng ước lượng cho λ từ đó tìm khoảng ước lượng cho thời gian phân hủy trung bình μ.

Giải

Đặt $h(X_1, X_2, \ldots, X_n; \lambda) = 2\lambda \sum X_j$, có $h(X_1, X_2, \ldots, X_n; \lambda)$ có phân phối Chi bình phương với bậc tự do $\nu = 2n$. Sử dụng bảng phụ lục các giá trị phân vị mức của phân phối Chi bình phương tương ứng với bậc tự do $\nu = 2(10) = 20$ và mức ý nghĩa 0,025 và 1-0,025 ta có

$$P\left(9,591 < 2\lambda \sum X_i < 34,17\right) = 0,95$$

hay

$$P\left(\frac{9,591}{2\sum X_i} < \lambda < \frac{34,170}{2\sum X_i}\right) = 0,95$$

Với dữ liệu đã cho $\sum x_i = 550,87$ có khoảng ước lượng cho λ là (0,00871; 0,03101).

Trung bình của phân phối mũ là $\mu = \frac{1}{\lambda}$ nên

$$P\left(\frac{2\sum X_i}{34,170} < \frac{1}{\lambda} < \frac{2\sum X_i}{9,591}\right)$$

do đó khoảng ước lượng cho trung bình μ với độ tin cậy 95% là

$$\left(\frac{2\sum x_i}{34,170}; \frac{2\sum x_i}{9,591}\right) = (32,24; 114,87)$$

7.1.4 Khoảng tin cậy bootstrap

Phương pháp bootstrap đã giới thiệu trong chương 6 như là cách để ước lượng sai số chuẩn $\sigma_{\hat{\theta}}$ của ước lượng điểm $\hat{\theta}$. Nó cũng được ứng dụng tìm khoảng tin cậy cho θ. Xét ước lượng của kỳ vọng μ của phân phối chuẩn với σ đã biết. Thay μ bởi θ và dùng $\hat{\theta} = \overline{X}$ như trong ước lượng điểm. Chú ý rằng $1,96\sigma\sqrt{n}$ là phân vị thứ 97,5 của phân phối $\hat{\theta} - \theta$[tức là, $P(\overline{X} - \mu < 1,96\sigma/\sqrt{n}) = P(Z < 1,96) = 0,9750$].

Tương tự $-1,96\sigma/\sqrt{n}$ là phân vị thứ 2,5; vì vậy

$$\begin{aligned} 0,95 &= P(\text{phân vị thứ} \quad 2,5 < \hat{\theta} - \theta < \text{phân vị thứ} \quad 97,5) \\ &= P(\hat{\theta} - \text{phân vị thứ} \quad 2,5 > \theta > \hat{\theta} - \text{phân vị thứ} \quad 97,5) \end{aligned}$$

Nghĩa là, với

$$l = \hat{\theta} - \text{phân vị thứ 97,5 của} \quad \hat{\theta} - \theta$$

$$u = \hat{\theta} - \text{phân vị thứ 2,5 của} \quad \hat{\theta} - \theta \tag{7.7}$$

Khoảng tin cậy cho θ là (l, u). Trong nhiều trường hợp, bách phân vị trong (7.7) không thể tính được nhưng ta có thể ước lượng được từ mẫu của chính nó. Giả sử lấy lại $B = 1000$ mẫu bootstrap và tính được

$$\hat{\theta}_1^*, ..., \hat{\theta}_{1000}^* \quad \text{và} \quad \overline{\theta}^*$$

được suy ra từ 1000 giá trị khác nhau

$$\hat{\theta}_1^* - \overline{\theta}^*, ..., \hat{\theta}_{1000}^* - \overline{\theta}^*.$$

Giá trị lớn nhất thứ 25 và giá trị nhỏ nhất thứ 25 của những sai khác này là ước lượng của những phân vị mức chưa biết trong (7.7).

7.2 Khoảng tin cậy của trung bình cho mẫu lớn

Khoảng tin cậy cho μ trong những phần trước đều giả sử tổng thể có phân phối chuẩn với σ đã biết. Giả sử khoảng tin cậy khi mẫu lớn không yêu cầu giả thiết này. Sau đây ta sẽ lập luận để đưa ra khoảng ước lượng cho trung bình khái quát trong trường hợp mẫu lớn và khoảng ước lượng cho tỉ lệ p của tổng thể.

7.2.1 Khoảng ước lượng của μ khi mẫu lớn

Mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ rút ra từ tổng thể có kỳ vọng μ và độ lệch chuẩn σ. Định lý giới hạn trung tâm chỉ ra rằng khi n lớn $\overline{X}$ dần tiến về phân phối chuẩn dù tổng thể có phân phối nào. Do đó có thể nói $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ có phân phối xấp xỉ chuẩn, vì thế

$$P\left(-Z_{\alpha/2} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\alpha/2}\right) \approx 1 - \alpha$$

Lập luận như trong (7.1) cho $\overline{x} \pm z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$ được khoảng tin cậy cho μ với độ tin cậy xấp xỉ $100(1-\alpha)\%$ khi mẫu lớn.

Khó khăn ở đây là việc tính toán khoảng tin cậy dựa trên giá trị σ mà giá trị này hiếm khi được biết. Xét biến chuẩn hóa $\frac{\overline{X}-\mu}{S/\sqrt{n}}$, trong biến chuẩn hóa này độ lệch chuẩn mẫu S thay cho σ. Trước đó chỉ có tử số là ngẫu nhiên theo Z bởi chứa $\overline{X}$. Trong biến chuẩn hóa mới này cả $\overline{X}$ và S đều có giá trị thay đổi từ mẫu này đến mẫu khác. Vì thế ta thấy rằng phân phối của biến mới này trải dài hơn đường cong z phản ánh sự thay đổi trong mẫu số. Điều này đúng khi n nhỏ. Tuy nhiên, khi n lớn chỉ có một chút thay đổi khi thay biến S cho σ, vì vậy biến này cũng có phân phối xấp xỉ chuẩn. Lập luận tương tự như trường hợp của σ đã biết, ta thu được một khoảng tin cậy cho μ khi mẫu lớn.

Mệnh đề 7.3. *Nếu n đủ lớn, biến chuẩn hóa $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ có phân phối xấp xỉ chuẩn, nghĩa là*

$$\overline{x} \pm z_{\alpha/2}.\frac{s}{\sqrt{n}} \tag{7.8}$$

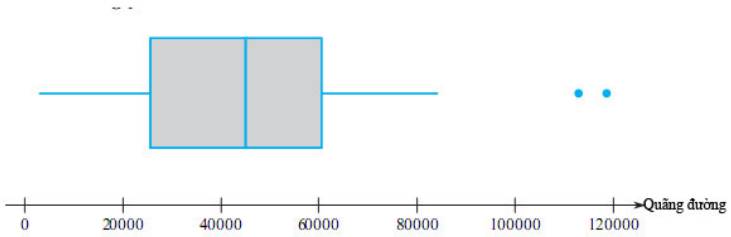
*là **khoảng tin cậy của trung bình** μ khi mẫu lớn với độ tin cậy xấp xỉ $100(1-\alpha)\%$. Công thức này đúng với mọi phân phối của tổng thể.*

Nói chung, $n > 40$ sẽ thỏa mãn để chứng minh các khoảng này. Điều này có phần hạn chế hơn công thức cho định lý giới hạn trung tâm CLT vì sử dụng S thay cho σ trong biến mới.

Ví dụ 7.5 Cho mẫu thực nghiệm

2948	2996	7197	8338	8500	8759	12710	12925
15767	20000	23247	24863	26000	26210	30552	30600
35700	36466	40316	40596	41021	41234	43000	44607
45000	45027	45442	46963	47978	49518	52000	53334
54208	56062	57000	57365	60020	60265	60803	62851
64404	72140	74594	79308	79500	80000	80000	84000
113000	118634						

Biểu đồ hộp của dữ liệu cho thấy, trừ hai ngoại lai ở phía cuối, phân phối của các giá trị tương đối đối xứng.



Xử lý số liệu được

- cỡ mẫu $n = 50$;
- giá trị trung bình mẫu $\overline{x} = 45.679,4$;
- giá trị trung vị mẫu $\tilde{x} = 45.013,5$
- giá trị độ lệch chuẩn mẫu $s = 26.641,675$,
- tứ phân vị lan truyền (khoảng cách từ tứ phân vị nhỏ đến tứ phân vị lớn) $f_s = 34265$.

Với độ tin cậy 95% giá trị $z_{0,025} = 1,96$ khoảng ước lượng cho trung bình là

$$\begin{aligned} 45.679,4 \pm (1,96)\left(\frac{26.641,675}{\sqrt{50}}\right) &= 45.679,4 \pm 7384,7 \\ &= (38.294,7; 53.064,1) \end{aligned}$$

Khoảng ước lượng này có độ rộng khá lớn do cỡ mẫu $n = 50$ chưa đủ lớn để vượt qua sự thay đổi đáng kể của mẫu.

7.2.2 Khoảng tin cậy tổng quát trong trường hợp mẫu lớn

Khoảng ước lượng $\overline{x} \pm z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$ và $\overline{x} \pm z_{\alpha/2}.\frac{s}{\sqrt{n}}$ là một trường hợp đặc biệt của khoảng tin cậy của tham số θ khi mẫu có cỡ lớn. Giả sử $\hat{\theta}$ là ước lượng hợp lý thỏa tính chất sau:

1. có phân phối xấp xỉ chuẩn;
2. là ước lượng không chệch;
3. một biểu diễn của độ lệch tiêu chuẩn của $\hat{\theta}$ là $\sigma_{\hat{\theta}}$ đã biết.

Ví dụ, trong trường hợp $\theta = \mu, \hat{\mu} = \overline{X}$ là một ước lượng không chệch của μ có phân phối xấp xỉ chuẩn khi n lớn và $\sigma_{\hat{\mu}} = \sigma_{\overline{X}} = \sigma/\sqrt{n}$. Chuẩn tắc hóa $\hat{\theta}$ thành biến ngẫu nhiên $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$, có phân phối xấp xỉ chuẩn.

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha \tag{7.9}$$

Đầu tiên giả sử $\sigma_{\hat{\theta}}$ không liên quan đến bất kỳ tham số nào (ví dụ, biết σ trong trường hợp $\theta = \mu$). Thay dấu $<$ trong (7.9) thành dấu $=$ ta có kết quả $\theta = \hat{\theta} \pm z_{\alpha/2}.\sigma_{\hat{\theta}}$, vì vậy giới hạn trên và giới hạn dưới của khoảng ước lượng là

$$\hat{\theta} - z_{\alpha/2}.\sigma_{\hat{\theta}} \quad \text{và} \quad \hat{\theta} + z_{\alpha/2}.\sigma_{\hat{\theta}}.$$

Giả sử $\sigma_{\hat{\theta}}$ không liên quan đến θ nhưng có liên quan đến một vài tham số. Đặt $s_{\hat{\theta}}$ là ước lượng của $\sigma_{\hat{\theta}}$ điều này có được bằng cách ước lượng tham số chưa biết. (ví dụ $s/\sqrt{n}$ là ước lượng của $\sigma/\sqrt{n}$). Dưới những điều kiện chung (cần để $s/\sqrt{n}$ dần về $\sigma_{\hat{\theta}}$ cho hầu hết mẫu), giá trị khoảng tin cậy là $\hat{\theta} \pm z_{\alpha/2}.s_{\hat{\theta}}$. Khoảng tin cậy của trung bình trong trường hợp mẫu lớn

$$\overline{x} \pm z_{\alpha/2}.s/\sqrt{n} \quad \text{là một ví dụ.}$$

Cuối cùng, giả sử $\sigma_{\hat{\theta}}$ liên quan đến θ chưa biết. Với trường hợp này, xét ví dụ khi $\theta = p$ là một tỉ lệ tổng thể thì $(\hat{\theta} - \theta)/\sigma_{\hat{\theta}} = z_{\alpha/2}$ có thể khó để giải. Một giải pháp xấp xỉ có thể dùng là thay θ trong $\sigma_{\hat{\theta}}$ bằng ước lượng $\hat{\theta}$. Kết quả được ước lượng cho độ lệch chuẩn $s_{\hat{\theta}}$ và khoảng ước lượng tương ứng là

$$\hat{\theta} \pm z_{\alpha/2}.s_{\hat{\theta}}.$$

7.2.3 Khoảng tin cậy cho tỉ lệ tổng thể

Đặt p là tỉ lệ tính chất quan tâm của tổng thể, đối tượng có tính chất cụ thể (ví dụ nhóm đối tượng tốt nghiệp từ một trường đại học, máy tính không cần dịch vụ bảo hành,... v.v). Mẫu ngẫu nhiên của n cá thể độc lập được chọn và X là số cá thể có tính chất quan tâm trong mẫu. Khi cỡ mẫu n nhỏ thì X có thể xem là một biến ngẫu nhiên nhị thức với $E(X) = np$ và $\sigma_X = \sqrt{np(1-p)}$. Nếu $np \geq 10, (q = 1 - p)$ thì X có phân phối xấp xỉ chuẩn.

Ước lượng hợp lý của p là $\hat{p} = X/n$, khi đó $\hat{p}$ có được bằng cách nhân X với hằng số $1/n$ cũng là phân phối xấp xỉ chuẩn. Như trong mục 6.1 $E(\hat{p}) = p$ ($\hat{p}$ là một ước

lượng không chệch của p) và sai số chuẩn $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. Độ lệch tiêu chuẩn $\sigma_{\hat{p}}$ liên quan đến tham số p chưa biết. Chuẩn hóa $\hat{p}$ suy ra

$$P(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}) \approx 1-\alpha$$

Biến đổi xác suất như trong mục 7.1, kết quả giới hạn khoảng tin cậy có được bằng cách thay $<$ bằng dấu $=$ và giải phương trình theo p. Ta được:

$$\begin{aligned} p &= \frac{\hat{p}+z^2_{\alpha/2}/2n}{1+z^2_{\alpha/2}/n} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\alpha/2}/4n^2}}{1+z^2_{\alpha/2}/n} \\ &= \tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\alpha/2}/4n^2}}{1+z^2_{\alpha/2}/n} \end{aligned}$$

Mệnh đề 7.4. *Đặt* $\tilde{p} = \dfrac{\hat{p}+z^2_{\alpha/2}/2n}{1+z^2_{\alpha/2}/n}$ *thì một khoảng tin cậy cho tỉ lệ của tổng thể* p *với độ tin cậy xấp xỉ* $100(1-\alpha)\%$ *là*

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}\hat{q}/n + z^2_{\alpha/2}/4n^2}}{1+z^2_{\alpha/2}/n} \tag{7.10}$$

với $\hat{q} = 1-\hat{p}$ *và khi đó dấu - trong (7.10) là giới hạn dưới và dấu + là giới hạn trên. Đây là* ***khoảng ước lượng điểm*** *cho* p.

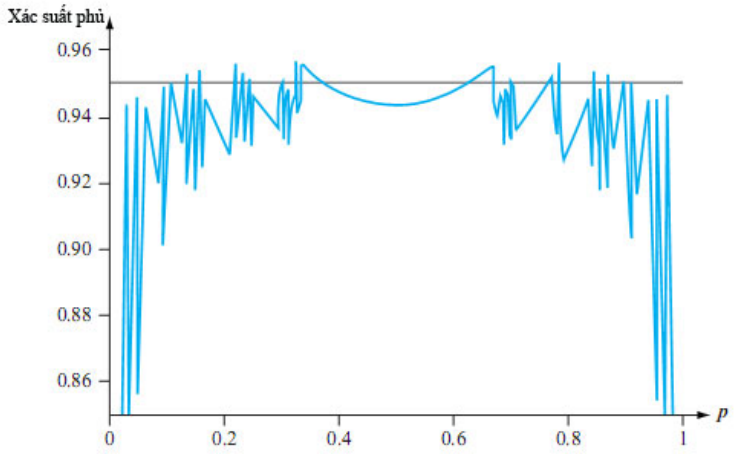
Khi cỡ mẫu n rất lớn thì $z^2/2n$ không đáng kể (nhỏ), z^2/n rất nhỏ khi so sánh với $\hat{p}$ và 1. Từ đó $\tilde{p} \approx \hat{p}$ trong trường hợp này so sánh giữa $z^2/4n^2$ và pq/n cũng không đáng kể (n^2 là số chia lớn hơn n), chiếm ưu thế trong biểu thức $\pm$ là $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$ khi đó tỉ lệ ước lượng của khoảng là xấp xỉ

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} \tag{7.11}$$

Đây là khoảng ước lượng có dạng là $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$ trong trường hợp mẫu lớn trong mục này. Khoảng tin cậy xấp xỉ (**??**) đã được giới thiệu trong nhiều sách về xác suất. Nó rõ ràng đơn giản và dễ hiểu hơn khoảng ước lượng điểm. Tại sao phải quan tâm đến khoảng ước lượng điểm?

Đầu tiên, giả sử dùng $z_{0,025} = 1,96$ trong công thức truyền thống (**??**), thì với độ tin cậy của khoảng ước lượng là xấp xỉ 95%, xác suất khoảng ngẫu nhiên bao gồm

cả giá trị thực của p (tức là xác suất phủ) nên là 0,95. Theo biểu diễn của hình sau cho trường hợp $n = 100$, xác suất phủ thực sự cho khoảng có thể khác nhau đáng kể so với xác suất chuẩn 0,95, đặc biệt khi p không gần với 0,5 (hình ảnh xác suất phủ p rất phức tạp vì phân phối xác suất nhị thức cơ bản là rời rạc chứ không liên tục). Đây là một nhược điểm của khoảng ước lượng truyền thống: mức độ tin cậy thực tế có thể khác với mức danh nghĩa ngay cả với cỡ mẫu lớn. Những nghiên cứu gần đây cũng chỉ ra rằng khoảng ước lượng điểm có thể khắc phục cho hầu như tất cả cỡ mẫu và tất cả giá trị của p, khoảng tin cậy thực sự sẽ ít khác biệt với khoảng ước lượng chuẩn do cách chọn $z_{\alpha/2}$. Điều này chủ yếu do khoảng ước lượng điểm được thay đổi một ít về 0,5 so với khoảng truyền thống. Đặc biệt điểm $\tilde{p}$ của khoảng ước lượng dấu gần về 0,5 so với điểm $\hat{p}$ của khoảng truyền thống. Đó là điều đặc biệt khi p nằm trong khoảng 0 đến 1.



Hình 7.4: Xác suất phủ thực sự cho khoảng (**??**) cho các giá trị khác nhau của p khi $n = 100$.

Thêm vào đó, khoảng ước lượng điểm có thể sử dụng cho hầu hết các mẫu và giá trị tham số. Không cần kiểm tra điều kiện $n\hat{p} \geq 10$ và $n(1 - \hat{p}) \geq 10$ thỏa yêu cầu của khoảng truyền thống.

Ví dụ 7.6 Một mẫu thực nghiệm cỡ $n = 48$ có 16 thí nghiệm có kết quả quan tâm. Gọi p là tỉ lệ có kết quả quan tâm trong mọi thí nghiệm. Một ước lượng điểm cho p là $\hat{p} = 16/48 = 1/3 = 0,333$. Khoảng tin cậy cho p với độ tin cậy xấp xỉ 95%

$$\frac{0,333 + (1,96)^2/96}{1 + (1,96)^2/48} \pm \frac{\sqrt{(0,333)(0,667)/48 + (1,96)^2/9216}}{1 + (1,96)^2/48}$$
$$= 0,345 \pm 0,129 = (0,216; 0,474)$$

Khoảng này khá rộng vì cỡ mẫu $n = 48$ chưa đủ lớn. Khoảng ước lượng truyền thống là

$$0,333 \pm 1,96\sqrt{\frac{(0,333)(0,667)}{48}} = 0,333 \pm 0,133 = (0,200; 0,466)$$

Hai khoảng này sẽ khá là gần nhau khi cỡ mẫu lớn.

Cỡ mẫu n cần thiết để khoảng ước lượng cho tỷ lệ p có độ rộng w là

$$n = \frac{2z^2\hat{p}\hat{q} - z^2w^2 \pm \sqrt{4z^4\hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^4z^4}}{w^2} \tag{7.12}$$

Hay

$$n \approx \frac{4z^2\hat{p}\hat{q}}{w^2}$$

Tuy nhiên kết quả này vẫn chứa $\hat{p}$ chưa biết. Cách tiếp cận hay nhất dùng giá trị lớn nhất của $\hat{p}\hat{q} = \hat{p}(1 - \hat{p})$ khi $\hat{p} = \hat{q} = 0,5$; do đó

$$n \approx \frac{z^2}{w^2}$$

Khi $w = 2\varepsilon$, ε gọi là sai số hoặc độ chính xác của khoảng ước lượng ta có công thức tương ứng là

$$n \approx \frac{z^2}{4\varepsilon^2}$$

7.2.4 Khoảng tin cậy bất đối xứng

Khoảng tin cậy được nói đến có cả giới hạn dưới của độ tin cậy và giới hạn trên của độ tin cậy cho ước lượng tham số. Trong một số trường hợp, điều tra viên chỉ mong muốn hai loại ràng buộc. Ví dụ nhà tâm lý học muốn tính khoảng tin cậy 95% chặn trên cho thời gian phản ứng trung bình với một kích thích đặc biệt, hay một kỹ sư mong muốn chỉ có chặn dưới cho thời gian hoạt động trung bình của một loại nhất định. Vì vùng diện tích tích lũy ở dưới đường cong chuẩn ở bên trái giá trị 1,645 là 0,95 nên

$$P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < 1,645\right) \approx 0,95$$

Dùng bất đẳng thức trong ngoặc đơn biến đổi theo μ và thay biến chuẩn hóa bằng cách tính toán các giá trị được bất đẳng thức $\mu > \overline{x} - 1,645.s/\sqrt{n}$. Vế phải biểu diễn chặn dưới của khoảng tin cậy. Với $P(-1,645 < Z) \approx 0,95$ và biến đổi bất đẳng thức ta được chặn trên của tin cậy. Lập luận tương tự cho ràng buộc của mỗi bên với một độ tin cậy khác nhau.

Mệnh đề 7.5. *Với một mẫu lớn, khoảng chặn trên của μ là*

$$\mu < \overline{x} + z_\alpha . \frac{s}{\sqrt{n}}$$

Với một mẫu lớn, khoảng chặn dưới của μ là

$$\mu > \overline{x} - z_\alpha . \frac{s}{\sqrt{n}}$$

Khoảng ước lượng một phía cho p là kết quả của việc thay $z_{\alpha/2}$ bằng z_α và $\pm$ bởi $+$ hay - trong công thức khoảng tin cậy cho p trong (7.10). Trong tất cả trường hợp độ tin cậy xấp xỉ là $100(1-\alpha)\%$.

Ví dụ 7.7 Một mẫu gồm 48 cá thể với giá trị trung bình mẫu $\overline{x} = 17,17$ và giá trị độ lệch chuẩn mẫu $s = 3,28$. Khoảng ước lượng chặn dưới của trung bình với độ tin cậy 95% là

$$17,17 - (1,645)\frac{(3,28)}{\sqrt{48}} = 17,17 - 0,78 = 16,39$$

Nghĩa là với độ tin cậy 95% ta có thể nói $\mu > 16,39$.

7.3 Các khoảng dựa trên phân phối chuẩn

Khoảng tin cậy cho μ dùng trong phần 7.2 có nghĩa khi n lớn. Kết quả là khoảng có thể sử dụng với bất kỳ phân phối nào của tổng thể. Tuy nhiên khi n nhỏ một cách để tìm khoảng ước lượng là đưa ra một giả thiết cụ thể về dạng phân phối của tổng thể, sau đó suy ra khoảng tin cậy phù hợp với giả thiết đó. Ví dụ, ta có thể đưa ra một khoảng tin cậy cho μ khi tổng thể được mô tả là có phân phối gamma, một khoảng khác cho trường hợp của phân phối Weibull, ... Các nhà thống kê đã thực hiện điều này cho một số họ phân phối khác nhau.

Giả Thiết 7.6. *Mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ lấy từ tổng thể có phân phối chuẩn với hai tham số μ và σ chưa biết.*

Kết quả chính của khoảng trong phần 7.2 là cho n lớn, biến ngẫu nhiên chuẩn hóa $Z = (\overline{X} - \mu)/(S/\sqrt{n})$ có phân phối xấp xỉ chuẩn. Khi n nhỏ, S có thể không dần về σ vì thế sự biến thiên của Z phát sinh trong cả tử số và mẫu số. Điều này suy ra rằng phân phối xác suất của $(\overline{X} - \mu)/(S/\sqrt{n})$ phân tán hơn phân phối chuẩn.

Kết quả suy luận dựa trên việc giới thiệu một họ mới của phân phối xác suất gọi là phân phối t.

Định lý 7.7. *Nếu $\overline{X}$ là trung bình của mẫu ngẫu nhiên có kích thước n từ một phân phối chuẩn với kỳ vọng μ, biến ngẫu nhiên*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{7.13}$$

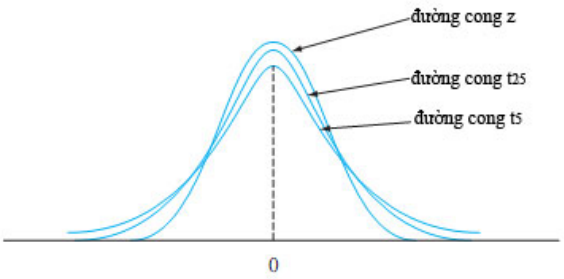
có một phân phối xác suất gọi là phân phối t với $n-1$ bậc tự do.

7.3.1 Tính chất của phân phối t

Ký hiệu ν là bậc tự do, viết tắt là df, là bậc tự do của phân phối t. Giá trị có thể có của ν là những số dương 1,2,3.... Cho một giá trị ν bất kỳ hàm mật độ xác định phân phối t tương ứng phức tạp hơn hàm mật độ chuẩn. May mắn là ta chỉ cần quan tâm đến một vài tính chất quan trọng của đường cong mật độ này. Đặt t_ν ký hiệu là phân phối t với bậc tự do ν.

Tính chất của phân phối t.

1. Đường cong t_ν có dạng chuông và trung vị tại 0.
2. Đường cong t_ν có độ trải rộng hơn đường cong của phân phối chuẩn tắc (z).
3. Khi ν tăng thì độ trải rộng của đường cong t_ν tương ứng giảm.
4. Khi $\nu \to \infty$, đường cong t_ν xấp xỉ với đường cong của phân phối chuẩn tắc (vì thế đường cong z thường được gọi là đường cong t với bậc tự do $df = \infty$.



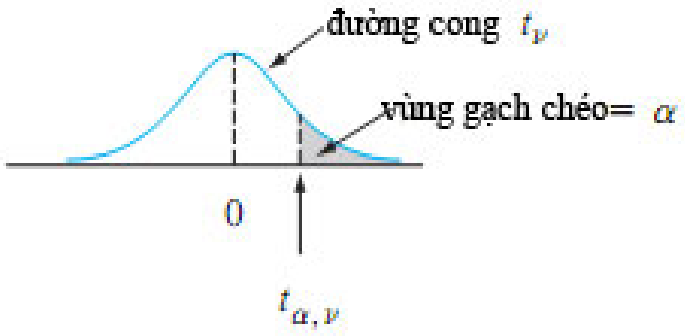
Hình 7.5: Đường cong t_ν và z.

Mặc dù S dựa trên n độ lệch $X_1 - \overline{X}, \ldots, X_n - \overline{X}$ nhưng $\sum(X_i - \overline{X}) = 0$ nên số bậc tự do (df) cho T trong (7.13) là $n-1$. Số bậc tự do cho biến t là căn cứ xác

định số độ lệch tự do của ước lượng độ lệch chuẩn trong mẫu số của T.

Chú ý: Đặt $t_{\alpha,\nu}$ là giá trị phân vị mức $100(1-\alpha)$ của phân phối t với ν bậc tự do (df); tức là diện tích miền giới hạn bởi trục hoành đường cong t nằm về phía bên phải của giá trị $t_{\alpha,\nu}$ là α.

Lấy ví dụ $t_{0,05;6}$ là giá trị phân vị mức 95 của phân phối t với 6 bậc tự do. Giá trị phân vị mức $t_{\alpha,\nu}$ được minh họa trong hình 7.6. Vì đường cong t đối xứng qua 0, $t_{1-\alpha,\nu} = -t_{\alpha,\nu}$. Giá trị phân vị mức $t_{\alpha,\nu}$ được biểu diễn trong bảng phụ lục A.5. Ví dụ để có được $t_{0,05;15} = 1,753$ ta đi từ cột $\alpha = 0,05$ chiếu xuống được dòng $\nu = 15$ và đọc được $t_{0,05;15} = 1,753$. Tương tự, $t_{0,05;22} = 1,717$ (cột 0,05 dòng $\nu = 22$), và $t_{0,01;22} = 2,508$. Với bậc tự do ν cố định, $t_{\alpha,\nu}$ tăng khi α giảm, khi đó ta phải di



Hình 7.6: Minh họa cho giá trị phân vị mức của phân phối t.

chuyển nhanh hơn về phía bên phải của 0 để xác định được khu vực α của phần đuôi. Cho α cố định, khi ν tăng (nghĩa là ta nhìn xuống bất kỳ cột nào của bảng t) giá trị của $t_{\alpha,\nu}$ giảm, vì thế không cần xa 0 để xác định miền đuôi có diện tích α. Thêm vào đó, $t_{\alpha,\nu}$ giảm chậm khi ν tăng. Do đó bảng giá trị phân vị mức của t biểu diễn sự tăng giữa 2 mốc bậc tự do 30 và 40 và nhảy tới $\nu = 50; 60; 120$ và cuối cùng là ∞. Vì t_∞ là đường cong của phân phối chuẩn. Họ các giá trị z_α xuất hiện vào dòng cuối của bảng. Công thức khoảng ước lượng CI sử dụng trước đó với mẫu lớn ($n > 40$) là ta xem phân phối t là xấp xỉ chuẩn.

7.3.2 Khoảng tin cậy khi cỡ mẫu nhỏ

Biến ngẫu nhiên chuẩn hóa T có phân phối t với $n-1$ bậc tự do (df) và diện tích miền giới hạn đường cong mật độ t và trục hoành từ $-t_{\alpha/2,n-1}$ và $t_{\alpha/2,n-1}$ là $1-\alpha$

(Diện tích mỗi bên đuôi là $\alpha/2$). Do đó

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha \tag{7.14}$$

Biểu thức 7.14 khác với biểu thức trong mục trước trong đó T và $t_{\alpha/2,n-1}$ thay cho Z và $z_{\alpha/2}$, nhưng nó được thao tác cùng một cách để có được khoảng tin cậy cho μ.

Mệnh đề 7.8. *Đặt $\overline{x}$ và s là trung bình mẫu và độ lệch tiêu chuẩn của mẫu của mẫu rút ngẫu nhiên từ tổng thể có phân phối chuẩn với kỳ vọng μ. Thì khoảng tin cậy $100(1-\alpha)\%$ cho μ là*

$$\left(\overline{x} - t_{\alpha/2,n-1}.\frac{s}{\sqrt{n}}; \overline{x} + t_{\alpha/2,n-1}.\frac{s}{\sqrt{n}}\right) \tag{7.15}$$

hay, gọn hơn , $\overline{x} \pm t_{\alpha/2,n-1}.s/\sqrt{n}$.
Khoảng tin cậy $100(1-\alpha)\%$ chặn trên cho μ là

$$\left(-\infty; \overline{x} + t_{\alpha,n-1}.\frac{s}{\sqrt{n}}\right)$$

Khoảng tin cậy $100(1-\alpha)\%$ chặn dưới cho μ là

$$\left(\overline{x} - t_{\alpha,n-1}.\frac{s}{\sqrt{n}}; \infty\right)$$

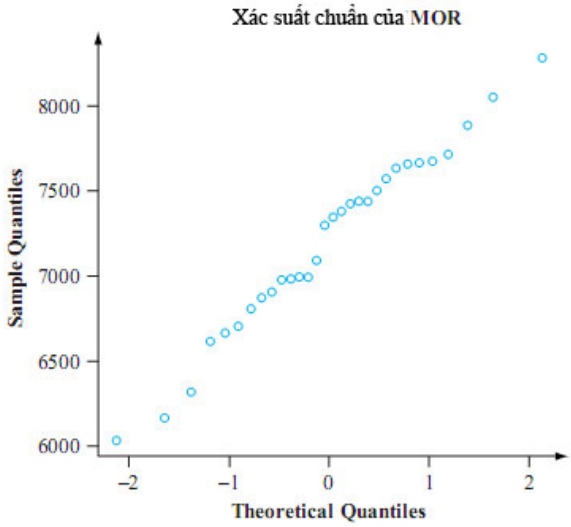
Ví dụ 7.8 Dữ liệu sau đây là giá trị MOR (đơn vị: MPa) trong bài viết "Phát triển của nền công nghiệp Laminated Planks từ gỗ xẻ Sweetgum" (Tạp chí Kỹ sư cầu đường, 2008:64-66)

6807,99	7637,06	6663,28	6165,03	6991,41	6992,23
6981,46	7569,75	7437,88	6872,39	7663,19	6032,28
6906,04	6617,17	6984,12	7093,71	7659,50	7378,61
7295,54	6702,76	7440,17	8053,26	8284,75	7347,95
7422,69	7886,87	6316,67	7713,65	7503,33	7674,99

Hình 7.7 biểu diễn biểu đồ xác suất chuẩn của dữ liệu bởi phần mềm R. Độ thẳng của các điểm trong biểu đồ cung cấp một giải thích mạnh mẽ cho phân phối của tổng thể MOR là tại đó ít xấp xỉ chuẩn nhất. Trung bình mẫu và độ lệch chuẩn mẫu lần lượt là $7203,191$ và $543,54$.

Bây giờ ta tìm khoảng tin cậy cho trung bình đúng của MOR với độ tin cậy 95%. Khoảng tin cậy này dựa vào $n-1 = 29$ bậc tự do, do đó giá trị tới hạn t cần tìm là $t_{0,025;29} = 2,045$. Khoảng ước lượng tìm được là:

$$\overline{x} \pm t_{0,025;29}.\frac{s}{\sqrt{n}} = 7230,191 \pm (2,045).\frac{543,54}{\sqrt{30}}$$



Hình 7.7: Biểu đồ xác suất chuẩn của dữ liệu MOR.

$$= 7230,191 \pm 202,938 = (7000,253; 7406,129)$$

Vậy ta ước lượng được $7000,253 < \mu < 7406,129$ với độ tin cậy 95%.
Chặn dưới của khoảng tin cậy 95% sẽ được tìm bằng cách thay giới hạn dưới của khoảng tin cậy vừa tìm với $t_{0,05;29} = 1,699$ thay cho $2,045$.

7.3.3 Dự đoán khoảng cho một giá trị tương lai đơn

Trong nhiều ứng dụng, vấn đề dự đoán giá trị đơn của một biến được quan sát tại một số thời điểm tương lai, được quan tâm hơn là ước lượng giá trị trung bình của biến đó.

Ví dụ 7.9 Xét mẫu sau về hàm lượng chất béo (theo phần trăm) của mẫu gồm $n = 10$ xúc xích được chọn ngẫu nhiên ("Đánh giá cảm quan và cơ học về chất lượng xúc xích Frankfurters", J. of Texture Studies , 1990:395-409):
25,2 21,3 22,8 17,0 29,8 21,0 25,5 16,0 20,9 19,5
Giả sử tổng thể có phân phối chuẩn. Một khoảng tin cậy 95% cho trung bình tổng thể của hàm lượng chất béo là:

$$\overline{x} \pm t_{0,025;9}.\frac{s}{\sqrt{n}} = 21,9 \pm 2,262\frac{4,134}{\sqrt{10}} = 21,9 \pm 2,96$$
$$= (18,94; 24,86)$$

Tuy nhiên, giả sử bạn chỉ định ăn đúng một cây xúc xích loại này và muốn dự đoán hàm lượng chất béo trong cây xúc xích sẽ ăn. Một dự đoán dạng điểm, tương tự ước

lượng điểm, chỉ là $\overline{x} = 21,9$. Dự đoán này không cho biết thông tin nào về độ tin cậy hay độ chính xác.

Thiết lập chung như sau: Ta có một mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ từ một tổng thể chuẩn, và muốn dự đoán giá trị đơn sẽ quan sát trong tương lai X_{n+1}. Một dự đoán dạng điểm là $\overline{X}$, và lỗi dự báo tương ứng là $\overline{X} - X_{n+1}$. Giá trị kỳ vọng của lỗi dự báo là:

$$E(\overline{X} - X_{n+1}) = E(\overline{X}) - E(X_{n+1}) = \mu - \mu = 0$$

Vì X_{n+1} độc lập với $X_1, \ldots, X_n$, nó cũng độc lập với $\overline{X}$, do đó phương sai của lỗi dự báo là:

$$V(\overline{X} - X_{n+1}) = V(\overline{X}) + V(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

Lỗi dự báo là một tổ hợp tuyến tính của các biến ngẫu nhiên có phân phối chuẩn và độc lập với nhau, vì vậy nó cũng có phân phối chuẩn. Do đó:

$$Z = \frac{(\overline{X} - X_{n+1}) - 0}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} = \frac{\overline{X} - X_{n+1}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}}$$

có phân phối chuẩn chính tắc. Nếu thay σ bởi độ lệch tiêu chuẩn mẫu S (của mẫu $X_1, \ldots, X_n$) thì ta có biến ngẫu nhiên:

$$T = \frac{\overline{X} - X_{n+1}}{S\sqrt{1 + \frac{1}{n}}}$$

có phân phối t với $(n-1)$ bậc tự do.
Biến đổi T thành $T = (\overline{X} - \mu)/(S/\sqrt{n})$ để xây dựng khoảng tin cậy cho kết quả sau đây.

Mệnh đề 7.9. *Một khoảng dự đoán (PI) cho một quan sát đơn từ tổng thể có phân phối chuẩn là*

$$\overline{x} \pm t_{\alpha/2,n-1}.s\sqrt{1 + \frac{1}{n}} \tag{7.16}$$

Mức dự đoán là $100(1-\alpha)\%$. *Chặn dưới của dự đoán có được từ việc thay* $t_{\alpha/2}$ *bởi* t_α *và bỏ dấu* $+$ *trong (7.16); làm tương tự để có chặn trên của dự đoán.*

Cách hiểu mức dự đoán 95% tương tự độ tin cậy 95%; nếu khoảng (7.16) được tính lần lượt sau nhiều lần lấy mẫu, về lâu dài 95% các khoảng này sẽ chứa giá trị tương lai phù hợp của X.

Ví dụ 7.10 (Tiếp ví dụ 7.9) Với $n = 10, \overline{x} = 21,9; s = 4,134$, và $t_{0,025;9} = 2,262$, một khoảng dự đoán 95% cho hàm lượng chất béo trong 1 cây xúc xích là:

$$21,9 \pm (2,262)(4,134)\sqrt{1 + \frac{1}{10}} = 21,9 \pm 9,81 = (12,09; 31,71)$$

Khoảng này khá rộng, cho thấy sự không chắc chắn đáng kể về hàm lượng chất béo. Chú ý là độ rộng của khoảng dự đoán gấp 3 lần so với khoảng tin cậy.

Lỗi dự đoán $\overline{X} - X_{n+1}$, là sai lệch giữa hai biến ngẫu nhiên, trong khi lỗi ước lượng $\overline{X} - \mu$, là sai lệch giữa một biến ngẫu nhiên và một giá trị cố định (tuy chưa biết). Khoảng dự đoán rộng hơn khoảng tin cậy vì có nhiều biến trong lỗi dự đoán hơn trong lỗi ước lượng.

7.3.4 Khoảng dung sai

Gọi k là một số nằm giữa 0 và 100. Một khoảng dung sai chứa ít nhất $k\%$ các giá trị của một tổng thể có phân phối chuẩn với độ tin cậy 95% có dạng:

$$\overline{x} \pm \text{ (giá trị tới hạn dung sai)}.s$$

Giá trị tới hạn dung sai cho $k = 90, 95$ và 99 kết hợp với các cỡ mẫu được cho trong Bảng A.6 phần Phụ lục. Bảng này cũng bao gồm các giá trị tới hạn cho độ tin cậy 99%. Thay $\pm$ bằng $+$ cho ta chặn trên dung sai, $-$ cho ta chặn dưới dung sai. Giá trị tới hạn để tính các chặn một phía cũng có trong Bảng A.6 phần phụ lục.

Ví dụ 7.11 Như một phần trong dự án lớn nghiên cứu về sự thay đổi của các tấm vỏ chịu lực, một thành phần cấu trúc đang được sử dụng rộng rãi ở Bắc Mỹ, bài viết "Tính uốn phụ thuộc thời gian của gỗ xẻ" (J.of Testing and Eval., 1996:187-193) báo cáo về các tính chất cơ học khác nhau của các mẫu gỗ thông Scotch. Xét những quan sát sau đây về mô-đun đàn hồi (MPa) thu được sau 1 phút tải một cấu hình nhất định:

10490 16620 17300 15480 12970 17260 13400 13900
13630 13260 14370 11700 15470 17840 14070 14760

Các lượng liên quan là $n = 16, \overline{x} = 14532.5, s = 2055.67$. Với độ tin cậy 95%, khoảng dung sai hai phía chứa ít nhất 95% giá trị mô-đun đàn hồi của các mẫu gỗ xẻ trong tổng thể với giá trị tới hạn dung sai 2,903 là:

$$14532,5 \pm (2,903)(2055,67) = 14532,5 \pm 5967,6 = (8564,9; 20500,1)$$

7.3.5 Khoảng dựa trên phân phối phi chuẩn của tổng thể

Khoảng tin cậy t trên một mẫu cho μ là robust tới nhỏ hay thậm chí lệch ra khỏi chuẩn trừ khi n khá nhỏ. Điều này có nghĩa là nếu giá trị phân vị mức cho khoảng tin cậy ví dụ như 95% thực sự gần với phân vị mức chuẩn với độ tin cậy là 95%. Nếu n nhỏ và phân phối tổng thể khá là phi chuẩn, thì độ tin cậy thực sự có sự khác biệt đáng kể so với mức giá trị phân vị có trong bảng t, nó sẽ gây ra phiền phức để tin rằng độ tin cậy của bạn là 95% trong khi thực ra là chỉ đạt được 88%. Phương pháp bootstrap, giới thiệu trong mục 7.1 là một cách tìm khoảng ước lượng tham số với trường hợp tổng thể phi chuẩn.

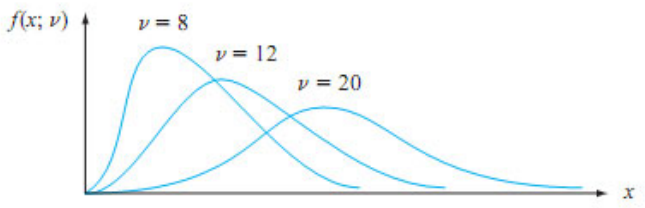
7.4 Khoảng tin cậy của phương sai và độ lệch chuẩn của phân phối chuẩn

Định lý 7.10. *Đặt $X_1, X_2, \ldots, X_n$ là mẫu ngẫu nhiên từ phân phối chuẩn với tham số μ và σ^2. Thì vecto ngẫu nhiên*

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

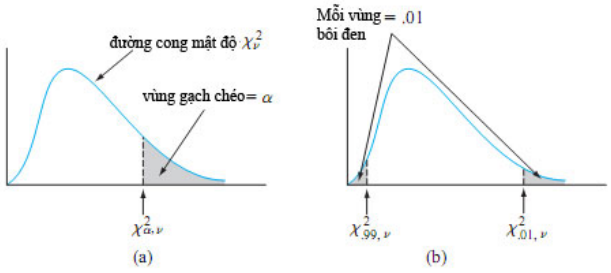
có phân phối chi bình phương χ^2 với $n-1$ bậc tự do (df).

Như đã bàn trong các mục 4.4 và 7.1, phân phối Chi-bình phương là một phân phối liên tục với tham số ν, gọi là bậc tự do, có thể nhận giá trị $1, 2, 3, \ldots$ Đồ thị một số hàm mật độ xác suất χ^2 được minh họa trong hình 7.8. Mỗi hàm mật độ xác suất $f(x;\nu)$ đều chỉ xác định dương khi $x > 0$, và có đồ thị nghiêng dương (phần đuôi trên kéo dài), tuy nhiên, phân phối di chuyển về bên phải và trở nên đối xứng hơn khi ν tăng.



Hình 7.8: Đồ thị các hàm mật độ Chi-bình phương.

Kí hiệu Gọi $\chi^2_{\alpha,\nu}$ là giá trị phân vị mức $100(1-\alpha)$ của phân phối Khi-bình phương ν bậc tự do, tức là phần diện tích dưới đường cong mật độ Khi-bình phương ν bậc tự do, trên trục hoành và ở bên phải $\chi^2_{\alpha,\nu}$ bằng α.



Hình 7.9: Minh họa cho kí hiệu $\chi^2_{\alpha,\nu}$.

Bảng A.7 phần Phụ lục chứa các giá trị phân vị mức của phân phối $\chi^2_{\alpha,\nu}$. Ví dụ, $\chi^2_{0,025;14} = 26,119$, và $\chi^2_{0,95;20} = 10,851$.

Biến ngẫu nhiên $(n-1)S^2/\sigma^2$ thỏa hai tính chất mà dựa vào đó phương pháp tổng quát để tìm khoảng tin cậy được xây dựng, đó là: biến này là một hàm theo tham số cần ước lượng σ^2; nhưng phân phối xác suất của nó (Chi-bình phương) lại không phụ thuộc vào tham số này. Phần diện tích dưới đường cong Chi-bình phương ν bậc tự do ở bên phải $\chi^2_{\alpha/2,\nu}$ là $\alpha/2$, bằng với phần diện tích ở bên trái $\chi^2_{1-(\alpha/2),\nu}$. Do đó, phần diện tích giữa hai giá trị tới hạn này là $(1-\alpha)$. Tức là:

$$P\left(\chi^2_{1-(\alpha/2),n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha \tag{7.17}$$

Bất đẳng thức trong (7.17) tương đương với

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2),n-1}}$$

Thay giá trị tính được s^2 vào các đầu mút cho ta khoảng tin cậy cho σ^2, và việc lấy tiếp căn bậc hai sẽ cho khoảng tin cậy của σ.

Một khoảng tin cậy $100(1-\alpha)\%$ cho phương sai σ^2 của một tổng thể chuẩn là

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}; \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2),n-1}}\right)$$

Một khoảng tin cậy cho σ có giới hạn dưới và giới hạn trên là căn bậc hai của giới hạn tương ứng của khoảng tin cậy cho σ^2. Một khoảng tin cậy chặn dưới hay chặn

trên được tính bằng cách thay $\alpha/2$ bởi α trong giới hạn tương ứng của khoảng tin cậy.

Ví dụ 7.12 Dữ liệu đi kèm về điện áp phóng điện của các mạch điện quá tải được đọc từ một biểu đồ xác suất chuẩn xuất hiện trong bài viết "Hư hỏng bảng mạch in linh hoạt liên quan đến sự tăng điện áp do tia sét" (IEEE, Transactions on Components, Hybrids, and Manuf. Tech., 1985:214-220). Tính tuyến tính của biểu đồ cung cấp cơ sở vững chắc cho giả thiết rằng điện áp phóng điện có phân phối xấp xỉ chuẩn.

1470 1510 1690 1740 1900 2000 2030 2100 2190
2200 2290 2380 2390 2480 2500 2580 2700

Đặt σ^2 là phương sai của phân phối điện áp phóng điện. Giá trị phương sai mẫu tính được là $s^2 = 137324.3$, cũng là ước lượng điểm của σ^2. Với bậc tự do là $n-1=16$, để tìm khoảng tin cậy 95% cho σ^2 ta cần các giá trị tới hạn Khi-bình phương gồm $\chi^2_{0.975,16} = 6,908$ và $\chi^2_{0,025;16} = 28,845$. Khoảng tìm được là:

$$\left(\frac{16(137324,3)}{28,845}, \frac{16(137324,3)}{6,908}\right) = (76172,3; 318064,4)$$

Lấy căn bậc hai mỗi đầu mút được $(276,0; 564,0)$ là khoảng tin cậy 95% cho σ.
Các khoảng trên khá rộng, do sự biến động đáng kể của điện áp phóng điện và cỡ mẫu khá nhỏ.

Khoảng tin cậy cho σ và σ^2 là khó tìm được trong trường hợp phân phối tổng thể là phi chuẩn. Các bạn có thể tìm hiểu trong các sách chuyên khảo về thống kê.

Bài tập Chương 7

Ước lượng trung bình tổng thể

Bài 1 Trọng lượng X (kg/con) của một số con heo ở thời kì xuất chuồng là:

X (kg/con)	65-85	85 – 95	95 – 105	105 – 115	115-135
Số con	8	40	60	42	10

Giả sử X có phân phối chuẩn. Hãy tìm khoảng tin cậy đối xứng 87% cho trọng lượng trung bình của loại heo trên. Biết trọng lượng 1 con heo được chọn ngẫu nhiên trong trại có phương sai là 225 kg^2.

Bài 2 Khảo sát chi tiêu X (triệu đồng/tháng) của một số người chọn ngẫu nhiên từ vùng A có thống kê sau:

X	3,2-3,7	3,7-4,2	4,2-4,7	4,7-5,2	5,2-5,7	5,7- 6,2	6,2-6,7
n_i	23	33	55	73	45	22	18

Tìm khoảng tin cậy đối xứng 98 % cho chi tiêu trung bình mỗi người dân vùng A.

Bài 3 Quan sát mức hao phí của 25 xe máy thuộc cùng một loại xe, chạy trên cùng một quãng đường, người ta thu được kết quả

Mức xăng (l)	1,9 – 2,1	2,1 – 2,3	2,3 – 2,5	2,5 – 2,7
Số xe	5	9	8	3

Hãy tìm ước lượng trung bình tối đa với độ tin cậy 99 % cho mức xăng hao phí trung bình của loại xe trên.

Bài 4 Quan sát 100 công nhân trong một xí nghiệp người ta tính được năng suất trung bình của một công nhân ở mẫu này là: $\overline{x} = 12$ sản phẩm/ngày và phương sai mẫu hiệu chỉnh là 25. Muốn ước lượng năng suất trung bình của một công nhân trong xí nghiệp với độ tin cậy 99% và độ chính xác $\varepsilon = 0,8$ thì cần quan sát năng suất của bao nhiêu công nhân nữa?

Bài 5 Mức tiêu thụ X của mỗi hộ gia đình vùng A trong mùa khô năm nay có phân phối chuẩn. Diều tra 1 số hộ gia đình vùng A :

X(kwh/t)	65-115	115-165	165-215	215-265	265-315	315-365	365-415
Số hộ	24	36	75	94	97	125	84

Nếu muốn ước lượng mức tiêu thụ điện trung bình các hộ vùng A trong mùa khô năm nay với độ chính xác 10 kwh/tháng thì độ tin cậy bằng bao nhiêu?

Ước lượng tỉ lệ tổng thể

Bài 6 Công ty M kiểm tra ngẫu nhiên 1200 sản phẩm do ca sáng sản xuất thấy có 45 sản phẩm không đạt chuẩn. Tính tỷ lệ sản phẩm đạt chuẩn tối đa do ca sáng sản xuất với độ tin cậy 99%.

Bài 7 Đo chiều dài 1 số sản phẩm do nhà máy A sản xuất:

X(cm)	53,8	53,81	53,82	53,83	53,84	53,85	53,86	53,87
Số sản phẩm	9	14	30	47	40	33	15	12

Biết X có phân phối chuẩn. Tìm khoảng ước lượng đối xứng 98 % cho tỉ lệ sản phẩm có chiều dài trên 53,84 cm.

Bài 8 Phỏng vấn 400 người ở một khu vực thấy 240 người ủng hộ dự luật A.

1. Với độ tin cậy 95%, hãy ước lượng tỷ lệ người ủng hộ dự luật A.
2. Nếu độ chính xác là 0,057 khi ước lượng tỉ lệ người ủng hộ dự luật A thì độ tin cậy là bao nhiêu?

Ước lượng phương sai tổng thể

Bài 9 Độ dày của bản kim loại tuân theo luật phân phối chuẩn. Đo 10 bản kim loại người ta tính được phương sai hiệu chỉnh của mẫu là 0,1367. Hãy xác định khoảng tin cậy 95% cho phương sai của độ dày đó.

Bài tập tổng hợp

Bài 10 Khảo sát chi tiêu X (triệu đồng/tháng) của một số người vùng A có thống kê sau (Biết X có phân phối chuẩn)

X	3,2-3,7	3,7-4,2	4,2-4,7	4,7-5,2	5,2-5,7	5,7- 6,2	6,2-6,7
Số người	23	33	55	73	45	22	18

1. Tìm khoảng tin cậy 95 % cho tỉ lệ người có chi tiêu trên 5,7 triệu đồng/ tháng vùng A.

2. Biết vùng A có 50 000 người. Tìm số người ở vùng A có chi tiêu trên 5,7 triệu đồng/tháng với độ tin cậy 95%.

3. Biết vùng A có 10 000 người chi tiêu trên 5,7 triệu đồng/ tháng . Tìm số người vùng A có với độ tin cậy 95 %

Bài 11 Khảo sát năng suất lúa thu được bảng số liệu sau:

Năng suất (tấn/ha)	5,1	5,4	5,5	5,6	5,8	6,2	6,4
Diện tích có năng suất tương ứng (ha)	10	20	30	15	10	10	5

Tìm ước lượng trung bình tối thiểu cho năng suất lúa trung bình ở vùng đó với độ tin cậy 95%

Bài 12 Do chiều dài 1 số sản phẩm do nhà máy A sản xuất, có thống kê sau: (X có phân phối chuẩn).

Chiều dài X(cm)	53,80	53,81	53,82	53,83	53,84	53,85	53,86	53,87
Số sản phẩm	9	14	30	47	40	33	15	12

1. Tìm khoảng tin cậy 95% cho chiều dài trung bình các sản phẩm do nhà máy A sản xuất.

2. Tìm khoảng tin cậy 98% cho tỉ lệ sản phẩm có chiều dài trên 53,84 cm.

Bài 13 Công ty M có 3000 đại lý, cho tiến hành điều tra ngẫu nhiên một số đại lý của mình và thu được bảng số liệu sau (X là doanh số, đơn vị: triệu đồng/tháng),

biết X có phân phối chuẩn.

Chiều dài X(cm)	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Số sản phẩm	7	12	18	27	22	17	13	4

1. Những đại lý có X>45 triệu đồng/tháng gọi là đại lý có doanh số cao. Hãy ước lượng số đại lý có doanh số cao với độ tin cậy 95%.

2. Hãy ước lượng doanh số trung bình/tháng của các đại lý với độ tin cậy 99%.

Bài 14 Khảo sát mức tiêu thụ điện X của một số hộ gia đình được chọn ngẫu nhiên ở vùng A ta được bảng số liệu sau:

X(kwh/tháng)	50-100	100-150	150-200	200-250	250-300	300-350	350-400
Số hộ	24	36	55	64	50	35	20

1. Ước lượng mức tiêu thụ điện trung bình của các hộ gia đình ở vùng A với độ tin cậy 99%.

2. Hộ có mức tiêu thụ điện dưới 100kwh/tháng gọi là hộ có mức tiêu thụ điện thấp. hãy ước lượng số hộ có mức tiêu thụ điện thấp ở vùng A với độ tin cậy 98% .Biết vùng A có 10.000 hộ dân.

Bài 15 Khảo sát thu nhập tại một doanh nghiệp có số liệu:

Thu nhập (triệu đồng/tháng)	2-4	4
Số lao động	24	1

1. Nếu dùng số liệu trên để ước lượng tỷ lệ người thu nhập thấp với sai số 1%. Hỏi độ tin cậy của ước lượng này khoảng bao nhiêu? Biết người thu nhập thấp có thu nhập từ 6 triệu đồng/tháng trở xuống.

2. Nếu dùng số liệu trên để ước lượng thu nhập trung bình của một người với sai số ko quá 0,5 triệu đồng/tháng thì điều tra ít nhất bao nhiêu người, với độ tin cậy 94 % và giả sử độ lệch chuẩn $\sigma = 1,85$.

Bài 16 Có số liệu thống kê về thu nhập (X: triệu đồng/tháng) của 100 người ở một công ty như sau:

Thu nhập (triệu đồng/tháng)	1-3	3-4	4-5	5-6	6-7	7-8	8-9	9-13
Số người	4	10	17	24	25	9	6	5

1. Nếu muốn độ chính xác khi ước lượng thu nhập trung bình là 0,25 (triệu đồng/ tháng) và thì độ chính xác là bao nhiêu?

2. Nếu muốn ước lượng thu nhập trung bình của các nhân viên ở công ty này có độ chính xác là 0,25 (triệu đồng/tháng) thì độ tin cậy đạt được bao nhiêu %?

Bài 17 Khảo sát chi tiêu X (triệu đồng/tháng) của một số hộ gia đình gồm 3 người (hai vợ chồng và một đứa con) chọn ngẫu nhiên từ vùng A, ta thu được số liệu sau :

X (triệu đồng)	7-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16	16-18
Số hộ gia đình	10	42	62	78	85	81	65	40	9

Giả sử X có phân phối chuẩn.

1. Tìm khoảng tin cậy 98% cho chi tiêu trung bình mỗi hộ gia đình gồm 3 người ở vùng A.

2. Tìm khoảng tin cậy 96% cho tỷ lệ hộ gia đình gồm 3 người ở vùng A có mức chi tiêu từ 10 triệu đồng trở lên trong một tháng.

3. Nếu độ rộng của khoảng ước lượng cho tỷ lệ hộ gia đình gồm 3 người ở vùng A có mức chi tiêu từ 10 triệu đồng trở lên trong một tháng là 0,02 thì độ tin cậy của khoảng này là bao nhiêu?

4. Nếu muốn tìm khoảng tin cậy 95% cho tỷ lệ hộ gia đình gồm 3 người ở vùng A có mức chi tiêu từ 10 triệu đồng trở lên trong một tháng với sai số không vượt quá 0,01 thì cần quan sát tối thiểu bao nhiêu hộ gia đình?

Bài 18 Quan sát trọng lượng X (kg/con) của một số con heo ở thời kì xuất chuồng ở trang trại M ta có số liệu

X (kg)	90-95	95-100	100-105	105-110	110-115	115-120	120-125
Số heo	8	40	60	83	88	82	66

1. Hãy ước lượng trọng lượng trung bình tối đa của 1 con heo thời kỳ xuất chuồng với độ tin cậy 96%.

2. Tỷ lệ heo xuất chuồng có trọng lượng ít nhất là 100kg tối đa là bao nhiêu? tối thiểu là bao nhiêu với độ tin cậy 98%. Biết trang trại có 1200 con heo đến thời kì xuất chuồng, hãy ước lượng số heo xuất chuồng có trọng lượng ít nhất là 100kg với độ tin cậy 98%.

3. Tìm khoảng tin cậy 97% cho trọng lượng các con heo thời kì xuất chuồng có trọng lượng từ 100kg trở lên.

Bài 19 Thống kê số sản phẩm loại S bán được tại 39 cửa hàng trong chuỗi các cửa hàng tiện lợi A trong một ngày ta thu được giá trị trung bình mẫu là 15,865 sản phẩm với độ lệch chuẩn mẫu là 2,38 sản phẩm. Giả sử số sản phẩm loại S bán được tại một cửa hàng trong chuỗi các hàng tiện lợi A trong một ngày có phân phối chuẩn.

1. Hãy tìm khoảng tin cậy 95% cho số sản phẩm loại S trung bình bán được trong một ngày tại một của hàng trong chuỗi các cửa hàng tiện lợi A.

2. Hãy tìm khoảng tin cậy 95% cho phương sai của số sản phẩm loại S bán được trong một ngày tại một của hàng trong chuỗi các cửa hàng tiện lợi A.

Bài 20 Để ước lượng số chim trong một vườn chim người ta bắt 400 con và đeo vòng vào chân chim, sau đó thả lại vào vườn chim. Sau 1 thời gian họ bắt 500 con chim trong vườn thấy có 46 con có đánh dấu. Tìm khoảng tin cậy 98% cho số chim trong vườn chim.

Chương 8

KIỂM ĐỊNH GIẢ THUYẾT DỰA TRÊN MỘT MẪU

8.1 Giả thuyết và thủ tục kiểm định

Một giả thuyết thống kê, hay giả thuyết, là một phát biểu hay sự khẳng định về

- giá trị của một tham số đơn (đặc điểm của tổng thể hay đặc điểm phân phối xác suất)
- giá trị của một số tham số
- hay về hình dạng của phân phối xác suất

Ví dụ như:

i. Giả thuyết phát biểu $\mu = 0,75$, khi μ là trung bình đúng của đường kính bên trong một loại ống PVC xác định. Ví dụ khác là phát biểu $p < 0,10$, khi p là tỉ lệ mạch bị lỗi trong tất cả các mạch được sản xuất từ một nhà máy.

ii. Nếu μ_1 và μ_2 ký hiệu cho trung bình đúng sức mạnh phá vỡ nút thắt của hai loại dây khác nhau. Một giả thuyết khẳng định $\mu_1 - \mu_2 = 0$, và một phát biểu khác là $\mu_1 - \mu_2 > 5$.

iii. Giả thuyết khẳng định rằng dưới một điều kiện xác định khoảng cách dừng có phân phối chuẩn.

Giả thuyết không về tham số sẽ được xét trong Chương 14. Trong chương này và một số chương, ta sẽ tập trung trên giả thuyết về tham số.

Trong bất kỳ vấn đề nào về kiểm định giả thuyết, cũng đều có hai giả thuyết mâu thuẫn được xét. Một giả thuyết có thể phát biểu $\mu = 0,75$ và giả thuyết khác là $\mu \neq 0,75$. Hay hai phát biểu mâu thuẫn có thể là $p \geq 10$ và $p < 10$. Dựa trên thông tin về mẫu sẽ quyết định giả thuyết nào sẽ đúng. Giống như trong một phiên tòa xét xử. Một tuyên bố khẳng định cá nhân bị buộc tội là vô tội. Trong hệ thống tư pháp U.S, khẳng định ban đầu này được cho là đúng, chỉ khi đối mặt với những bằng chứng mạnh mẽ ngược lại bị cáo thì bồi thẩm đoàn sẽ bác bỏ tuyên bố này để ủng hộ khẳng định khác là bị cáo phạm tội.

Tương tự, trong kiểm định giả thuyết thống kê, vấn đề là xây dựng phát biểu ban đầu được ủng hộ. Phát biểu ban đầu này sẽ không bị bác bỏ hay bị yêu cầu thay thế trừ khi có bằng chứng mạnh mẽ từ mẫu cung cấp cho khẳng định khác.

Định nghĩa 8.1. Giả thuyết không, ký hiệu H_0 là phát biểu ban đầu được giả định là đúng. Giả thuyết thay thế hay đối thuyết, ký hiệu là H_a là khẳng định mâu thuẫn với H_0.

Giả thuyết không sẽ được thay thế bởi giả thiết khác nếu bằng chứng từ mẫu cho thấy H_0 sai. Nếu không có mâu thuẫn mạnh với H_0, ta sẽ tiếp tục tin vào tính hợp lý của giả thuyết không. Hai kết luận có thể có của quá trình kiểm định giả thuyết là từ chối H_0 hay không từ chối H_0.

Một kiểm định của giả thuyết là phương pháp sử dụng thông tin từ dữ liệu mẫu để đưa ra quyết định bác bỏ hay chấp nhận giả thuyết không. Kiểm định giả thuyết $H_0 : \mu = 0,75$ với đối thuyết $H_a : \mu \neq 0,75$. Chỉ khi dữ liệu mẫu gợi ý mạnh rằng μ khác 0,75 thì sẽ bác bỏ giả thiết không. Trong trường hợp không có bằng chứng đó ta chấp nhận H_0.

Trong nhiều tình huống, H_a được gọi là "giả thuyết của nhà nghiên cứu" khi đó phát biểu của nghiên cứu viên được xác nhận. Ví dụ như 10% bảng mạch được sản xuất bởi cùng một nhà máy trong thời gian gần đây là có khiếm khuyết. Một kỹ sư đề nghị thay đổi dây chuyền sản xuất với niềm tin là nó sẽ có kết quả trong việc giảm tỉ lệ khiếm khuyết. Đặt p là tỉ lệ khiếm khuyết của mạch sau khi thay đổi dây chuyền sản xuất. Khi đó ta kiểm định giả thuyết $p \geq 0,10$ với đối thuyết là $H_a : p < 0,10$.

Tuy nhiên trong tài liệu này khi xét về kiểm định giả thuyết, H_0 được xem như một phát biểu bình đẳng. Nếu θ ký hiệu là tham số quan tâm, giả thuyết không sẽ có dạng $H_0 : \theta = \theta_0$, khi θ_0 là số đặc biệt gọi là giá trị không của tham số. Như ví dụ xét tình hình bảng mạch được nói như trên luận. Đối thuyết được đề nghị thay thế

cho giả thuyết $H_0 : p = 0,10$ là $H_a : p < 0,10$, tỉ lệ khiếm khuyết giảm khi sửa đổi quá trình sản xuất khi ta có bằng chứng cho việc thay đổi dây chuyền sản xuất là tốt. Nếu quá trình đổi mới xảy ra một trong hai tình huống hoặc tốt hơn hoặc xấu hơn quá trình cũ thì ta xét đối thuyết thay thế cho $H_0 : p = 0,10$ là $H_a : p \neq 0,10$.

Việc thay thế giả thuyết $H_0 : \theta = \theta_0$ giống như ba khẳng định sau:

1. $H_a : \theta > \theta_0$ (trong những trường hợp tiềm ẩn giả thuyết không là $\theta \leq \theta_0$)
2. $H_a : \theta < \theta_0$ (trong những trường hợp tiềm ẩn giả thuyết không là $\theta \geq \theta_0$)
3. $H_a : \theta \neq \theta_0$

8.1.1 Quá trình kiểm định

Quá trình kiểm định là một qui tắc, dựa trên dữ liệu mẫu, cho quyết định nên bác bỏ H_0 hay chấp nhận H_0.

Kiểm định giả thuyết $H_0 : p = 0,10$ với đối thuyết $H_a : p < 0,10$ trong ví dụ về bảng mạch có thể dựa trên kiểm định mẫu ngẫu nhiên của $n = 200$ bảng mạch. Ký hiệu X là số bảng mạch bị khiếm khuyết trong mẫu, X là một biến ngẫu nhiên nhị thức. x đại diện cho giá trị quan sát của X. Nếu H_0 đúng, $E(X) = np = 200.(0,10) = 20$ trong khi ta mong muốn có ít hơn 20 bảng mạch khiếm khuyết nếu H_a là đúng. Một giá trị x chỉ cần ít hơn 20 không mâu thuẫn mạnh với H_0 đó là lý do bác bỏ H_0 nếu $x \leq 15$ và không bác bỏ H_0 trong trường hợp ngược lại.

Quá trình kiểm định có hai phần:

1. Xác định một kiểm định thống kê, hay hàm của dữ liệu mẫu dùng để đưa ra quyết định.
2. Tìm vùng bác bỏ bao gồm cả giá trị x cho H_0 sẽ bị bác bỏ H_a.

Với ví dụ trên, giá trị bị bác bỏ bao gồm x=0, 1, 2,..., và 15. H_0 sẽ không bị bác bỏ nếu $x = 16, 17, \ldots, 199$, hay 200.

8.1.2 Sai lầm trong kiểm định giả thuyết

Định nghĩa 8.2. Sai lầm loại I là bác bỏ giả thuyết không H_0 khi nó đúng. Sai lầm loại II là việc chấp nhận H_0 khi H_0 sai.

Quá trình kiểm định tối ưu nhất là không xảy ra sai lầm. Tuy nhiên, ý tưởng này chỉ đạt được khi quyết định đưa ra dựa trên việc kiểm tra toàn bộ tổng thể ban đầu. Hơn nữa khó khăn của việc đưa ra quyết định của một quá trình là dựa trên dữ liệu mẫu. Một mẫu không đại diện cho tổng thể, ví dụ một giá trị của $\overline{X}$ khác μ hay giá trị của $\hat{p}$ khác giá trị ban đầu p sẽ là nguy cơ dẫn đến sai lầm của quá trình kiểm định.

Thay vì đòi hỏi một quá trình kiểm định không xảy ra sai lầm, ta phải tìm ra một quá trình hạn chế việc xảy ra cả hai loại sai lầm. Nghĩa là, một quá trình kiểm định tốt tức là xác suất xảy ra cả hai loại sai lầm đều nhỏ. Lựa chọn một vùng bác bỏ có giá trị xác suất phù hợp cho sai lầm loại I và loại II, xác suất xảy ra những sai lầm này được ký hiệu tương ứng bởi α và β. Vì H_0 có giá trị duy nhất là tham số, đó là giá trị đơn α. Tuy nhiên, với mỗi giá trị của tham số phù hợp với H_a có một giá trị khác của β.

Ví dụ 8.1 25% khách hàng hài lòng với chất lượng phục vụ tại cửa hàng A. Trong thời gian gần đây nhân viên cửa hàng nỗ lực tăng chất lượng phục vụ. Gọi p tỉ lệ khách hàng hài lòng. Hãy kiểm định giả thuyết cho p.

Giả thuyết $H_0 : p = 0,25$ và đối thuyết $H_a : p > 0,25$.

Khảo sát 20 khách hàng, giả thuyết H_0 sẽ bị bác bỏ khi số khách hàng hài lòng là lớn. Xét quá trình sau:

$$\text{Kiểm định thống kê:} \quad X = \text{số khách hàng hài lòng}$$
$$\text{Miền bác bỏ:} \quad R_8 = \{8, 9, \ldots, 19, 20\}; \quad \text{tức là, bác bỏ} \quad H_0 \quad \text{nếu} \quad x \geq 8$$

Khi H_0 đúng, X có phân phối nhị thức với $n = 20$ và $p = 0,25$.

$$\begin{aligned}\alpha &= P(\text{Sai lầm loại I}) = P(\text{bác bỏ} \quad H_0 \quad \text{khi nó đúng}) \\ &= P(X \geq 8 \quad \text{khi} \quad X \sim Bin(20; 0,25)) = 1 - B(7; 20; 0,25) \\ &= 1 - 0,989 = 0,102\end{aligned}$$

Ngược lại giá trị β có nhiều giá trị tương ứng với mỗi giá trị của p khác 0,25.

$$\begin{aligned}\beta(0,3) &= P(\text{Sai lầm loại II khi} \quad p = 0,3) \\ &= P(\text{Không bác bỏ} \quad H_0 \quad \text{khi nó sai vì} \quad p = 0,3) \\ &= P(X \leq 7 \quad \text{khi} \quad X \sim Bin(20; 0,3)) = 1 - B(7; 20; 0,3 = 0,772\end{aligned}$$

Tương tự $\beta(0,4) = 0,416; \quad \beta(0,5) = 0,132; \quad \beta(0,6) = 0,021; \quad \beta(0,7) = 0,001.$

8.2 Kiểm định về trung bình tổng thể

Những thảo luận chung trong chương 7 về Khoảng tin cậy cho một tổng thể có kỳ vọng μ tập trung ở ba trường hợp khác nhau. Ta sẽ xây dựng quá trình kiểm định cho những trường hợp này.

8.2.1 Trường hợp I: Một tổng thể chuẩn với σ đã biết

Mặc dù giả thuyết về giá trị σ đã biết rất hiếm gặp trong thực tế, tuy nhiên trường hợp này đơn giản và tính chất của nó có thể xây dựng được. Giả thuyết không trong cả ba trường hợp nói rằng μ có giá trị đặc biệt ký hiệu là μ_0.

Đặt $X_1, \ldots, X_n$ là một mẫu ngẫu nhiên kích thước n có phân phối chuẩn. Thì trung bình mẫu $\overline{X}$ cũng có phân phối chuẩn với giá trị kỳ vọng $\mu_{\overline{X}} = \mu$ và độ lệch chuẩn $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Khi H_0 đúng, $\mu_{\overline{X}} = \mu_0$. Xét thống kê Z chuẩn hóa bởi $\overline{X}$ với giả thiết H_0 đúng.

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Giả sử đối thuyết có dạng $H_a : \mu > \mu_0$. Thì một giá trị $\overline{x}$ ít hơn μ_0 chắc chắn không thể hỗ trợ cho việc chấp nhận H_a. Vì thế một $\overline{x}$ phù hợp để bác bỏ H_0 là một giá trị làm z âm (khi $\overline{x} - \mu_0$ là số âm nên khi chia cho $\sigma/\sqrt{n}$ là số âm). Tương tự, một giá trị $\overline{x}$ mà vượt quá μ_0 chỉ một số lượng nhỏ (tương ứng với z, là số dương nhưng nhỏ) cũng không gợi ý rằng nên bác bỏ H_0 và ủng hộ H_a. Bác bỏ H_0 phù hợp chỉ khi $\overline{x}$ vượt quá μ_0 đáng kể, nghĩa là, khi giá trị z dương và lớn. Vùng bác bỏ phù hợp, dựa trên kiểm định thống kê Z có dạng $z \geq c$.

Như trong lập luận của phần 8.1, giá trị c nên chọn để xác định xác suất của sai lầm loại I tại mức α mong muốn. Điều này dễ dàng làm được vì phân phối của kiểm định thống kê Z khi H_0 đúng là phân phối chuẩn tắc. Như một ví dụ, đặt $c = 1,645$, giá trị phân vị thứ 95 của phân phối chuẩn tắc ($z_{0,05} = 1,645$). Thì

$$\begin{aligned}\alpha &= P(\text{sai lầm loại I}) = P(\text{bác bỏ} \quad H_0 \quad \text{khi} \quad H_0 \quad \text{đúng}) \\ &= P(Z \geq 1,645 \quad \text{khi} \quad Z \sim N(0,1)) = 1 - \Phi(1,645) = 0,05\end{aligned}$$

Nói chung, vùng bác bỏ $z \geq z_\alpha$ có xác suất sai lầm loại I mức α.

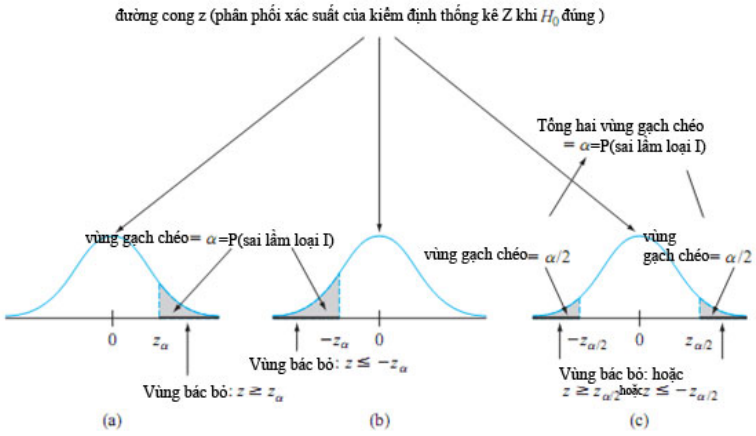
Lập luận tương tự cho đối thuyết $H_a : \mu < \mu_0$, vùng bác bỏ có dạng $z \leq c$, khi đó c là số âm được chọn thỏa mãn ($\overline{x}$ ở phía dưới μ_0 nếu và chỉ nếu z là số âm). Vì Z có phân phối chuẩn tắc khi H_0 đúng, lấy $c = -z_\alpha P(\text{ sai lầm loại I }) = \alpha$. Đây là

kiểm định bên trái. Lấy ví dụ , $z_{0,10} = 1,28$ suy ra vùng bác bỏ $z \leq -128$ với mức ý nghĩa 0,10.

Cuối cùng, khi đối thuyết là $H_a : \mu \neq \mu_0$, nên bác bỏ H_0 nếu $\overline{x}$ quá xa so với μ_0, điều này tương đương với bác bỏ H_0 hay nếu $z \geq c$ hoặc nếu $z \leq -c$. Giả sử $\alpha = 0,5$ thì:

$$\begin{aligned} 0,05 &= P(Z \geq c \quad \text{hay} \quad Z \leq -c) \\ &= \Phi(-c) + 1 - \Phi(c) = 2[1 - \Phi(c)] \end{aligned}$$

Giả thiết không: $H_0 : \mu = \mu_0$ Giá trị kiểm định thống kê: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	
Đối thuyết	Vùng bác bỏ cho kiểm định mức α
$H_a : \mu > \mu_0$	$z \geq z_\alpha$ (kiểm định phía bên phải)
$H_a : \mu < \mu_0$	$z \leq -z_\alpha$ (kiểm định phía bên trái)
$H_a : \mu \neq \mu_0$	$z \geq z_{\alpha/2}$ hoặc $z \leq -z_{\alpha/2}$ (kiểm định hai bên)



Hình 8.1: Vùng bác bỏ cho kiểm định: (a) kiểm định bên trái; (b) kiểm định bên phải; (c) kiểm định hai bên.

Thủ tục kiểm định giả thuyết về tham số.

1. Xác định tham số cần quan tâm và mô tả nó trong bối cảnh xảy ra vấn đề.

2. Xác định giá trị không và giả thuyết không.

3. Xác định đối thuyết.

4. Cho một công thức để tính giá trị của kiểm định thống kê.

5. Xác định vùng bác bỏ với mức ý nghĩa α.

6. Tính tham số mẫu cần thiết, thay thế vào công thức cho giá trị kiểm định thống kê để tính giá trị kiểm định thống kê.

7. Quyết định có nên bác bỏ H_0 hay không.

Công thức của bước 2 và 3 nên thực hiện trước khi kiểm tra dữ liệu.

Ví dụ 8.2 Nhà sản xuất hệ thống vòi phun nước dùng bảo vệ lửa trong cao ốc văn phòng phát biểu rằng trung bình đúng của hệ thống sẽ được kích hoạt khi nhiệt độ là 130^o. Một mẫu $n = 9$ hệ thống, khi kiểm tra có năng suất của trung bình mẫu kích hoạt với nhiệt độ $131,08^oF$. Nếu phân phối của thời gian hoạt động là chuẩn với độ lệch chuẩn $1,5^oF$ dữ liệu mâu thuẫn nào với phát biểu của nhà sản xuất có mức ý nghĩa $\alpha = 0,01$?

1. Tham số cần quan tâm μ là giá trị đúng của hệ thống.

2. Giả thuyết $H_0 : \mu = 130, \quad \mu_0 = 130.$

3. Đối thuyết $H_a : \mu \neq 130.$

4. Giá trị kiểm định
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{x} - 130}{1,5/\sqrt{n}}$$

5. Với mức ý nghĩa $\alpha = 0,01$ miền bác bỏ là $z \geq z_{0,005} = 2,58$ hoặc $z \leq -z_{0,005} = -2,58$.

6. Thay $n = 9$ và $\overline{x} = 131,08$
$$z = \frac{131,08 - 130}{1,5/\sqrt{9}} = \frac{1,08}{0,5} = 2,16$$

7. z không rơi vào miền bác bỏ giả thuyết. Vì thế không bác bỏ giả thuyết H_0 với mức ý nghĩa 0,01.

‘

8.2.2 Trường hợp II: Kiểm định cho mẫu lớn

Khi cỡ mẫu lớn, kiểm định z cho trường hợp I có thể dễ dàng xác định giá trị cho quá trình kiểm định mà không cần yêu cầu là phân phối chuẩn hay σ đã biết. Với cỡ mẫu n lớn, như lập luận trong chương 7 biến chuẩn tắc hóa là:

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

có phân phối xấp xỉ chuẩn tắc. Thay giá trị μ bằng giá trị không μ_0 ta được kiểm định thống kê

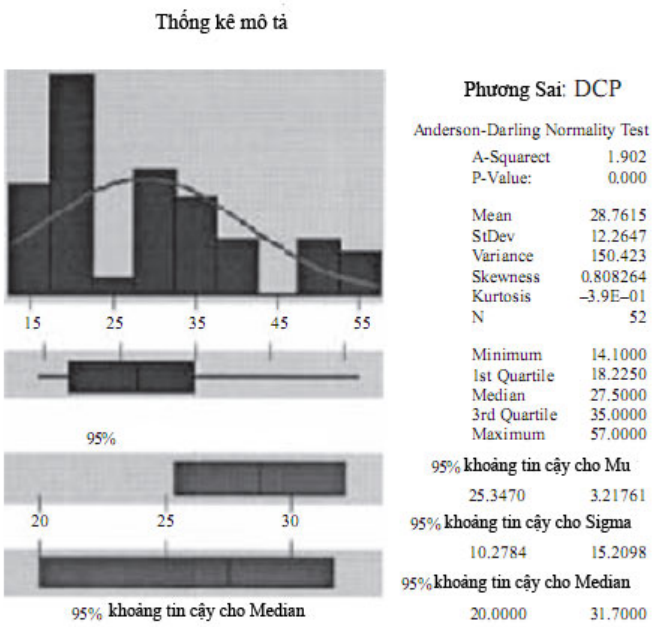
$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Có phân phối xấp xỉ chuẩn tắc khi H_0 đúng.

Ví dụ 8.3 Một máy đo trọng lực hình nón (DCP) dùng để đo độ thấm xuyên ($mm/blow$) có hình nón được đưa vào vỉa hè hay lề đường. Giả sử cho một ứng dụng đặc biệt mà có giá trị trung bình đúng DCP cho một loại hình nón đặc biệt ít hơn 30. Hình nón này sẽ không được ứng dụng nếu không có bằng chứng kết luận kết luận về đặc điểm kỹ thuật đã được đáp ứng. Đặt bài toán thích hợp về kiểm định giả thiết sử dụng dữ liệu sau ("Mô hình xác suất phân tích các giá trị kiểm tra sức cản thấm xuyên của nón trọng lực trong đánh giá cấu trúc vỉa hè" J.of Testing and Evaluation, 1999:7-14)

14.1	14.5	15.5	16.0	16.0	16.7	16.9	17.1	17.5	17.8
17.8	18.1	18.2	18.3	18.3	19.0	19.2	19.4	20.0	20.0
20.8	20.8	21.0	21.5	23.5	27.5	27.5	28.0	28.3	30.0
30.0	31.6	31.7	31.7	32.5	33.5	33.9	35.0	35.0	35.0
36.7	40.0	40.0	41.3	41.7	47.5	50.0	51.0	51.8	54.4
55.0	57.0								

1. μ là giá trị trung bình đúng của DCP.
2. Giả thuyết $H_0 : \mu = 30$.
3. Đối thuyết $H_a : \mu < 30$.
4. Giá trị kiểm định
$$z = \frac{\overline{x} - 30}{s/\sqrt{n}} = \frac{\overline{x} - 130}{1,5/\sqrt{n}}$$
5. Với mức ý nghĩa $\alpha = 0,05$ miền bác bỏ là $z \geq -1,6445$.



Hình 8.2: Biểu diễn một bảng tóm tắt từ phần mềm minitab.

6. Thay $n = 52$ và $\overline{x} = 28,76$ và $s = 12,2647$.

$$z = \frac{28,76 - 30}{12,2647/\sqrt{52}} = \frac{-1,24}{1,701} = -0,73$$

7. Vì $-0,7 > -1,645$ ta không bác bỏ giả thuyết H_0.

8.2.3 Trường hợp III: Phân phối chuẩn của tổng thể

Khi n nhỏ, không thể sử dụng định lý giới hạn trung tâm như trường hợp cỡ mẫu lớn. Ta sẽ giả định rằng phân phối của tổng thể ít nhất là xấp xỉ chuẩn và mô tả quá trình kiểm định có hiệu lực dựa trên giả định này. Nếu quan sát viên có lý do tốt để tin rằng phân phối của tổng thể khá là phi chuẩn thì một kiểm định cho họ phân phối khác với họ chuẩn sẽ được xét trong chương 15 hoặc có thể phát triển bằng quá trình bootstrap.

Nếu $X_1, X_2, \ldots, X_n$ là biến ngẫu nhiên từ phân phối chuẩn, biến ngẫu nhiên

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

có phân phối t với $n - 1$ bậc tự do.

Kiểm định giả thuyết $H_0 : \mu = \mu_0$ với đối thuyết $H_a : \mu > \mu_0$ bằng cách xét kiểm định thống kê $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$. Khi H_0 đúng, kiểm định thống kê T có phân phối t với $n-1$ bậc tự do. Khi H_0 đúng, vùng bác bỏ là vùng xảy ra sai lầm loại I. Sử dụng giá trị phân vị mức của phân phối t với $n-1$ bậc tự do $t_{\alpha,n-1}$ để xác định vùng bác bỏ $t \geq t_{\alpha,n-1}$ ta có

$$\begin{aligned} P(\text{ sai lầm loại I }) &= P(\text{Bác bỏ} \quad H_0 \quad \text{khi nó đúng}) \\ &= P(T \geq t_{\alpha,n-1} \quad \text{khi} \quad T \quad \text{có phân phối} \quad t \quad \text{với} \quad n-1 \quad \text{bậc tự do}) \\ &= \alpha \end{aligned}$$

Kiểm định thống kê này giống như trong trường hợp mẫu lớn nhưng đặt là T để nhấn mạnh rằng phân phối không của nó là một phân phối t với $n-1$ bậc tự đo chứ không phải là phân phối chuẩn z. Vùng bác bỏ cho kiểm định t khác kiểm định z là giá phân vị mức $t_{\alpha,n-1}$ thay cho giá trị phân vị mức z_α.

Kiểm định t cho một mẫu
Giả thuyết không: $H_0 : \mu = \mu_0$
Giá trị kiểm định: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

Đối thuyết	Vùng bác bỏ cho kiểm định mức α
$H_a : \mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (kiểm định bên phải)
$H_a : \mu < \mu_0$	$t \leq t_{\alpha,n-1}$ (kiểm định bên trái)
$H_a : \mu \neq \mu_0$	$t \geq t_{\alpha/2,n-1}$ hoặc $t \leq -t_{\alpha/2,n-1}$ (kiểm định hai bên)

8.3 Kiểm định về tỷ lệ

Ký hiệu p là tỉ lệ cá thể hoặc đối tượng trong một tổng thể có tính chất xác định (ví dụ tỷ lệ xe hơi truyền động bằng tay hay tỷ lệ người hút thuốc bằng bộ lọc điếu thuốc). Nếu một cá thể hay đối tượng với tính chất được cho là thành công (S) thì p là tỉ lệ thành công của tổng thể. Kiểm định liên quan đến p bằng cách dựa vào cỡ n của mẫu ngẫu nhiên lấy từ tổng thể. Gọi X là số cá thể thành công trong mẫu cỡ n, thì X có phân phối nhị thức với 2 tham số n, p. Khi n lớn ($np \geq 10$ và $n(1-p) \geq 10$), thì X có phân phối xấp xỉ chuẩn. Đầu tiên ta xét kiểm định với trường hợp mẫu lớn, sau đó với trường hợp mẫu nhỏ mà ta dùng trực tiếp phân phối nhị thức.

8.3.1 Kiểm định tỷ lệ p trường hợp mẫu lớn

Kiểm định p trường hợp mẫu là một trường hợp đặc biệt của quá trình tham số θ nói chung với cỡ mẫu lớn. Đặt $\hat{\theta}$ là ước lượng không chệch của θ và có xấp xỉ phân phối chuẩn với giả thuyết không có dạng $H_0 : \theta = \theta_0$ khi ký hiệu θ_0 là một số (giá trị không). Giả sử khi H_0 đúng, độ lệch chuẩn của $\hat{\theta}$, là $\sigma_{\hat{\theta}}$. Ví dụ như $\theta = \mu$ và $\hat{\theta} = \overline{X}, \sigma_{\hat{\theta}} = \sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Tiêu chuẩn kiểm định

$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

Nếu đối thuyết là $H_a : \theta > \theta_0$ một kiểm định bên phải với mức ý nghĩa xấp xỉ α được qui định bởi vùng bác bỏ $z \geq z_\alpha$.

Hai đối thuyết khác là $H_a : \theta < \theta_0$ và $H_a : \theta \neq \theta_0$ tương ứng với kiểm định bên trái và kiểm định hai bên cho z.

Trong trường hợp $\theta = p, \sigma_{\hat{\theta}}$ sẽ không liên quan đến bất kỳ tham số chưa biết nào khi H_0 đúng, đây không phải là trường hợp điển hình. Khi $\sigma_{\hat{\theta}}$ liên quan đến tham số chưa biết, thường thì có thể sử dụng ước lượng có độ lệch chuẩn $S_{\hat{\theta}}$ ở vị trí $\sigma_{\hat{\theta}}$ và Z vẫn có phân phối xấp xỉ chuẩn khi H_0 đúng (vì khi n lớn, $s_{\hat{\theta}} \approx \sigma_{\hat{\theta}}$ với hầu hết mẫu). Kiểm định mẫu lớn cho phần trước được trang bị ví dụ này: Vì σ thường không biết, ta dùng $s_{\hat{\theta}} = s_{\overline{X}} = s/\sqrt{n}$ thay cho $\sigma/\sqrt{n}$ trong mẫu của z.

Ước lượng $\hat{p}$ là không chệch ($E(\hat{p}) = p$), có phân phối xấp xỉ chuẩn, và độ lệch chuẩn của nó là $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. Thực tế đã được dùng trong Phần 7.2 để có khoảng tin cậy cho p. Khi H_0 đúng, $E(\hat{p}) = p_0$và $\sigma_{\hat{p}} = \sqrt{p_0(1-p_0)/n}$, vì thế $\sigma_{\hat{p}}$ không liên quan đến bất kỳ tham số chưa biết nào. Ta có kiểm định sau khi n lớn và H_0 đúng

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

có phân phối xấp xỉ chuẩn. Nếu giả thiết thay thế là $H_a : p > p_0$ và dùng vùng bác bỏ phía bên phải $z \geq z_\alpha$, thì

$$\begin{aligned} P(\text{sai lầm loại I}) &= P(H_0 \quad \text{bị bác bỏ khi nó đúng }) \\ &= P(Z \geq z_\alpha \text{ khi Z có phân phối xấp xỉ chuẩn }) \approx \alpha \end{aligned}$$

Vì thế mức ý nghĩa mong muốn của α đạt được bằng cách sử dụng giá trị tới hạn mà có trong vùng lấy trên đường cong phía bên phải của α. Tương tự có vùng bác bỏ cho hai đối thuyết bên trái $H_a : p < p_0$ và hai bên $H_a : p \neq p_0$.

Giả thiết không: $H_0 : p = p_0$		
Giá trị kiểm định thống kê: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$		
Đối thuyết	Vùng bác bỏ	
$H_a : p > p_0$	$z \geq z_\alpha$	(bên phải)
$H_a : p < p_0$	$z \leq z_\alpha$	(bên trái)
$H_a : p \neq p_0$	hoặc $z \geq z_\alpha$ hoặc $z \leq -z_{\alpha/2}$	(hai bên)
Quá trình kiểm định này được sử dụng khi $np_0 \geq 10$ và $n(1-p_0) \geq 10$		

8.3.2 Kiểm định tỷ lệ p trường hợp mẫu nhỏ

Quá trình kiểm định khi cỡ mẫu n nhỏ dựa trực tiếp vào phân phối nhị thức hơn là xấp xỉ chuẩn. Xét giả thiết thay thế $H_a : p > p_0$ và đặt X là số thành công trong mẫu. Thì là kiểm định thống kê, và vùng bác bỏ phía bên phải có dạng $x \geq c$ khi H_0 đúng. X có phân phối nhị thức với tham số n và p_0, vì thế

$$\begin{aligned} P(\text{sai lầm loại I}) &= P(H_0 \quad \text{bị bác bỏ khi nó đúng}) \\ &= P(X \geq c \quad \text{khi} \quad X \sim Bin(n, p_0)) \\ &= 1 - P(X \leq c-1 \quad \text{khi} \quad X \sim Bin(n, p_0)) \\ &= 1 - B(c-1; n, p_0) \end{aligned}$$

Khi giá trị tới hạn c giảm, nhiều giá trị X sẽ có trong vùng bác bỏ và P(sai lầm loại I) tăng vì X có phân phối xác suất rời rạc, thường không thể tìm được giá trị c chính xác cho P(sai lầm loại I) với mức ý nghĩa mong muốn là α (ví dụ 0,05 hay 0,01). Thêm vào đó vùng bác bỏ lớn nhất có dạng $\{c, c+1, \ldots, n\}$ thỏa $1 - B(c-1 : n, p_0) \leq \alpha$ được sử dụng. Quá trình kiểm định cho $H_a : p < p_0$ và cho $H_a : p \neq p_0$ tương tự như cách xây dựng trước, vùng bác bỏ xấp xỉ có dạng $x \leq c$ (một kiểm định bên trái). Giá trị tới hạn c là số lớn nhất thỏa $B(c; n, p_0) \leq \alpha$. Vùng bác bỏ khi giả thiết thay thế là $H_a : p \neq p_0$ bao gồm cả giá trị X lớn và nhỏ.

8.4 P giá trị

Thay vì sử dụng miền bác bỏ, ta sẽ xét cách khác để có kết luận trong kiểm định giả thuyết là P giá trị. Một cách tự nhiên, P giá trị cung cấp một độ đo về chiều dài của bằng chứng trong dữ liệu chống lại H_0.

Định nghĩa 8.3. Giả sử rằng thiết H_0 đúng, từ các giá trị của mẫu đã tính sẵn,

P giá trị là một xác suất biến cố một giá trị của kiểm định thống kê là mâu thuẫn với H_0.

Định nghĩa này khá ngắn gọn. Ta cần quan tâm những điểm chính sau

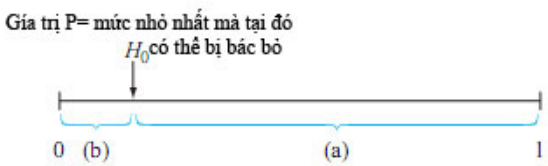
- P giá trị là một xác suất.
- Xác suất này tính được bằng cách giả sử H_0 đúng
- P giá trị không phải là xác suất để H_0 đúng, cũng không là xác suất sai lầm.
- Để xác định P giá trị, ta quyết định giá trị nào của kiểm định thống kê được tính từ mẫu có sẵn là mâu thuẫn với H_0.

Qui tắc quyết định trên P giá trị.
Chọn mức ý nghĩa α (xác suất mắc sai lầm loại I). Khi đó
Bác bỏ H_0 nếu giá trị P $\leq \alpha$
Không bác bỏ H_0 nếu giá trị P $> \alpha$

Do đó nếu giá trị P vượt quá mức ý nghĩa được lựa chọn, giả thiết không không thể bác bỏ tại mức đó, nhưng nếu giá trị P bằng hoặc nhỏ hơn α thì có đủ bằng chứng để chấp nhận H_0. Ví dụ ta tính giá trị $P = 0,0012$ thì phải sử dụng mức ý nghĩa 0,01. Ta sẽ bác bỏ giả thiết không và chấp nhận giả thiết thay thế vì $0,0012 < 0,01$. Tuy nhiên, giả sử ta chọn mức ý nghĩa là 0,001 thì khi đó phải có nhiều bằng chứng hơn từ dữ liệu trước khi H_0 có thể bị bác bỏ.

Hai quá trình của phương pháp vùng bác bỏ và phương pháp P giá trị trên thực tế là như nhau.

Mệnh đề 8.4. *P giá trị là mức ý nghĩa α nhỏ nhất mà tại đó giả thuyết không có thể bị bác bỏ.*



Hình 8.3: So sánh α và P giá trị P: (a) bác bỏ H_0; (b) không bác bỏ H_0.

8.4.1 P giá trị cho kiểm định z

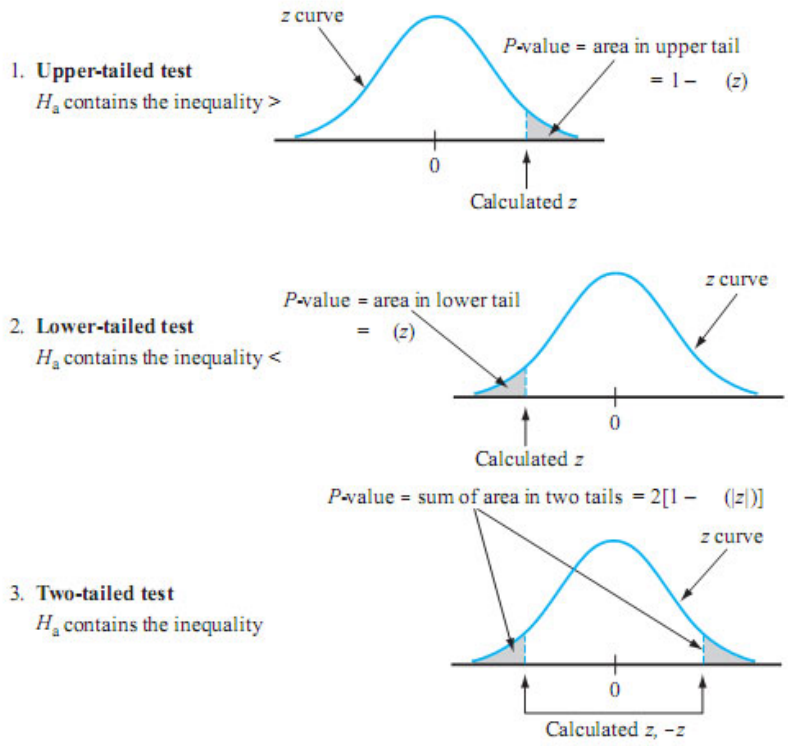
P giá trị cho kiểm định Z (ít nhất có phân phối xấp xỉ chuẩn). Xét một kiểm định bên phải và đặt z là giá trị tính được từ kiểm định thống kê Z. Giả thuyết không bị bác bỏ nếu $Z \geq z_\alpha$ và P giá trị là giá trị α nhỏ nhất trong trường hợp này. Khi z_α tăng thì α giảm, P giá trị là giá trị của α khi $z = z_\alpha$, tức là P giá trị $= 1 - \Phi(z)$. Lập luận tương tự cho kiểm định bên trái.

Với kiểm định hai bên, đầu tiên giả sử z là số dương thì P giá trị là giá trị α thỏa $z = z_{\alpha/2}$ (nghĩa là tính z bằng giá trị tới hạn bên phải). Điều này nói lên rằng xét vùng phía bên phải là một nửa của P giá trị, vì thế P giá trị $= 2[1 - \Phi(z)]$. Nếu z là số âm, P giá trị là α khi cho $z = -z_{\alpha/2}$, hay , tương đương với $-z = z_{\alpha/2}$, vì thế P giá trị $= 2[1 - \Phi(-z)]$. Khi đó $-z = |z|$ khi z là số âm, P giá trị $= 2[1 - \Phi(|z|)]$ cho cả z âm và z dương. Ba trường hợp này được minh họa trong hình 8.4.

$$P \text{ giá trị: } P = \begin{cases} 1 - \Phi(z) & \text{cho kiểm định bên phải của } z \\ \Phi(z) & \text{cho kiểm định bên trái của } z \\ 2[1 - \Phi(|z|)] & \text{cho kiểm định hai bên của } z \end{cases}$$

Ví dụ 8.4 Độ dày cho các tấm silicon được sử dụng trong một loại mạch tích hợp nhất định được kỳ vọng là 245 μm. Một mẫu gồm 50 tấm và mỗi tấm có một độ dày xác định, kết quả trung bình mẫu của độ dày là 246,18 μm và độ lệch chuẩn $3,60\mu m$. Dữ liệu này có cho thấy độ dày trung bình của tấm khác với giá trị kỳ vọng không?

1. Tham số quan tâm μ là độ dày trung bình của các tấm silicon.
2. Giả thuyết $H_0 : \mu = 245$.
3. Đối thuyết $H_a : \mu \neq 245$.
4. Giá trị kiểm định $z = \frac{\overline{x}-245}{s/\sqrt{n}}$
5. Tính giá trị kiểm đinh $\frac{246,18-245}{3,6/\sqrt{50}} = 2,32$
6. Với kiểm định hai phía, P giá trị $=2(1 - \Phi(2,32)) = 0,0204$
7. Với mức ý nghĩa $\alpha = 0,01$ ta không bác bỏ giả thuyết H_0 vì 0,0204>0,01. Với mức ý nghĩa này ta chưa đủ bằng chứng để bác bỏ độ dày của các tấm silicon khác độ dày kỳ vọng.



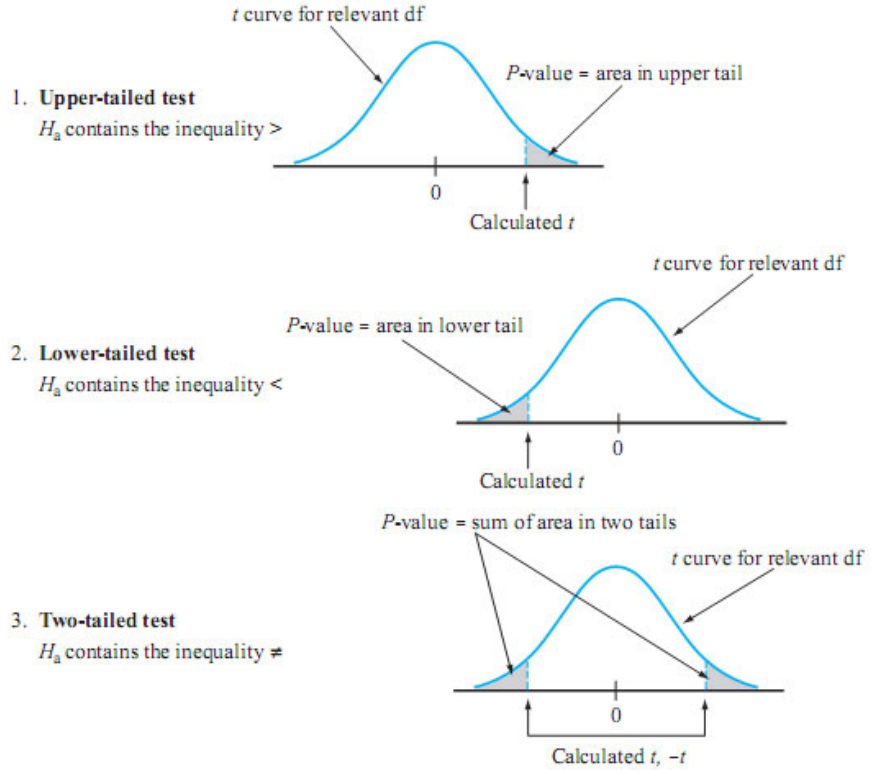
Hình 8.4: Xác định P giá trị cho kiểm định z.

8.4.2 P giá trị cho kiểm định t

Cũng giống như P giá trị cho kiểm định z là vùng đường cong z, P giá trị cho kiểm định t sẽ là vùng nằm dưới đường cong t với bậc tự do là $n-1$ được minh họa trong hình 8.5.

Bảng của giá trị tới hạn t dùng ở phần trước cho khoảng tin cậy và dự đoán sẽ không có đủ thông tin về bất kỳ phân phối t cụ thể nào để xác định chính xác vùng mong muốn trong đó có cả vùng bên phải của đường cong t, ta có bảng Phụ lục A.8. Mỗi cột khác nhau của của bảng cho một số bậc tự do khác nhau và những dòng này dùng để tính giá trị của kiểm định thống kê t được sắp xếp từ 0,0 đến 4,0 tăng dần đến 1. Cho ví dụ, số 0,074 xuất hiện tại giao của dòng tại 1,6 và cột tại bậc tự do (df) 8, vì thế vùng dưới đường cong t với 8 bậc tự do về phía bên phải của 1,6 (một vùng bên phải) là 0,074 vì đường cong t đối xứng nên 0,074 cũng có vùng dưới đường cong về phía bên trái của -1,6 (vùng dưới bên trái).

Ví dụ như, kiểm định giả thuyết $H_0 : \mu = 100$ với đối thuyết $H_a : \mu > 100$ dựa



Hình 8.5: P giá trị cho kiểm định t.

vào phân phối t với 8 bậc tự do. Nếu tính được giá trị thống kê t là 1,6 thì P giá trị cho kiểm định bên phải là 0,074 vì 0,074 vượt quá 0,05 nên ta không thể bác bỏ H_0 tại mức ý nghĩa 0,05. Nếu giả thiết thay thế là $H_a : \mu < 100$ và kiểm định dựa trên vùng 20 bậc tự do với $t = -3,2$, theo bảng phụ lục A.8 P giá trị vùng bên trái là 0,002. Giả thuyết không có thể bị bác bỏ tại mức 0,05 hoặc 0,01. Kiểm định giả thuyết $H_0 : \mu_1 - \mu_2 = 0$ với đối thuyết $H_a : \mu_1 - \mu_2 \neq 0$. Giả thuyết không nói rằng trung bình của 2 tổng thể được xác định trong khi giả thuyết thay thế nói rằng chúng khác nhau mà không thể xác định được H_0. Nếu một kiểm định t dựa trên 20 bậc tự do và $t = 3,2$ thì P giá trị cho kiểm định hai bên là $2(0,002) = 0,004$. Dây cũng là P giá trị với $t = -3,2$ vùng đuôi được gấp đôi lên vì giá trị của cả hai lớn hơn 3,2 và nhỏ hơn -3,2 điều này mâu thuẫn với H_0 đã tính (giá trị được sinh ra từ đuôi của đường cong t).

8.5 Một số chú ý về chọn thủ tục kiểm định

Một trong những thử nghiệm để trả lời cho câu hỏi cần quan tâm và phương pháp thu thập dữ liệu (thiết kế một thí nghiệm) để xây dựng kiểm định xấp xỉ bao gồm ba điểm chính sau:

1. Chỉ định một thử nghiệm thống kê (hàm của giá trị quan sát sẽ giúp cho việc đưa ra quyết định)
2. Quyết định dạng chung của vùng bác bỏ (hàm của các giá trị quan sát sử dụng cho việc đưa ra quyết định).
3. Chọn một số những giá trị quan trọng hoặc những giá trị mà sẽ không phụ thuộc vào vùng bác bỏ từ vùng chấp nhận được (bằng cách lấy phân phối của kiểm định thống kê khi H_0 đúng, và sau đó chọn ra một mức ý nghĩa).

Trong những ví dụ trước đó cả bước 1 và bước 2 đều được thực hiện theo một cách thức đặc biệt thông qua trực giác, ví dụ giả sử khi tổng thể về cơ bản có phân phối chuẩn với kỳ vọng μ và σ đã biết, từ $\overline{X}$ ta có biến chuẩn hóa cho kiểm định thống kê:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Kiểm định giả thuyết $H_0 : \mu = \mu_0$ với đối thuyết $H_a : \mu > \mu_0$, bằng trực giác ta đề nghị bác bỏ H_0 khi Z lớn, cuối cùng giá trị quan trọng được xác định bằng mức ý nghĩa của α và dùng giá trị thực Z có phân phối chuẩn, khi H_0 đúng. Độ tin cậy của kiểm định trong việc đưa ra một quyết định đúng có thể đánh giá được bằng những nghiên cứu về sai lầm loại II trong xác suất. Vấn đề cần xem xét ở đây là khi thực hiện các bước từ 1-3 có những câu hỏi cần giải quyết sau.

1. Ý nghĩa thực tế và hệ quả của chọn một mức ý nghĩa cụ thể khi đã xác định được thử nghiệm là gì?
2. Có tồn tại nguyên tắc chung không phụ thuộc vào trực giác mà có thể sử dụng để thu được một quá trình kiểm định tốt nhất hoặc tốt hay không?
3. Khi hai hay nhiều kiểm định là phù hợp trong cùng một bối cảnh phải so sánh những kiểm định đó như thế nào để đưa ra quyết định nên sử dụng cái nào?
4. Nếu một kiểm định xuất phát từ giả thiết có phân phối cụ thể hay từ mẫu tổng thể, kiểm định sẽ có dạng như thế nào khi vi phạm giả thuyết?

8.5.1 Ý nghĩa thống kê trong thực tế

Mặc dù quá trình đưa ra quyết định bằng cách sử dụng phương pháp luận cổ điển liên quan đến việc chọn mức ý nghĩa và sau đó bác bỏ hoặc không bác bỏ H_0 tại mức α đó, một báo cáo về việc sử dụng α để đưa ra quyết định lại truyền tải ít thông tin về dữ liệu mẫu. Đặc biệt khi kết quả của một thử nghiệm lại truyền đạt đến nhiều người, bác bỏ H_0 tại mức 0,05 sẽ thuyết phục hơn nếu giá trị quan sát của kiểm định thống kê sẽ vượt quá 5% giá trị điều kiện nếu nó vượt quá giá trị đó sẽ bác bỏ H_0. Điều này đúng với những yêu cầu dẫn đến khái niệm của P giá trị như cách có được mức ý nghĩa mà không cần áp đặt cho nó để có một giá trị α đặc biệt mà người khác muốn đưa ra quyết định cho riêng họ.

Thậm chí nếu P giá trị có trong bảng tóm tắt kết quả, có thể rất khó để giải thích giá trị này để đưa ra quyết định. Đó là vì, một P giá trị nhỏ sẽ chỉ ra một ý nghĩa thống kê, khi đó sẽ đề nghị bác bỏ H_0 và chấp nhận H_a, Khi kết quả là một mẫu lớn bắt đầu từ H_0 thì ít có ý nghĩa trong thực tế hơn. Trong nhiều tình huống, thử nghiệm chỉ bắt đầu từ H_0 lại ít có ý nghĩa trong thực tế trong khi với mức độ nhỏ từ H_0 lại có ý nghĩa thực tế hơn.

Xét ví dụ, Kiểm định $H_0 : \mu = 100$ ngược lại $H_a : \mu > 100$ khi μ là kỳ vọng của phân phối chuẩn với $\sigma = 10$, giả sử giá trị đúng của $\mu = 101$, sẽ không thấy được tầm quan trọng của H_0 bằng trực quan ta không bác bỏ H_0 khi $\mu = 101$ sẽ liên quan đến một sai lầm nghiêm trọng. Cho một mẫu lớn hợp lý kích thước n, với giá trị μ này sẽ cho một giá trị $\overline{x}$ gần 101 vì thế ta không muốn có bất kỳ bằng chứng gì về mẫu này để có sự tranh luận mạnh mẽ về việc bác bỏ H_0 khi $\overline{x} = 101$ là biến quan sát được. Khi có nhiều cỡ mẫu bảng 8.1 sẽ ghi lại tất cả giá trị P khi $\overline{x} = 101$ và xác suất của việc không bác bỏ H_0 tại mức .01 khi $\mu = 101$.

Cột thứ hai trong bảng 8.1 biểu diễn rằng ngay cả khi cỡ mẫu đủ lớn giá trị P của $\overline{x} = 101$ cũng gây ra tranh cãi mạnh mẽ cho việc bác bỏ H_0, khi $\overline{x}$ là biến quan sát và nó cũng gợi ý rằng giới hạn đúng trong thực tế ít khác với giá trị không $\mu_0 = 100$. Cột thứ ba chỉ ra rằng khi thực tế có sự khác biệt giữa giá trị đúng μ và giá trị không μ_0 cho mức ý nghĩa thích hợp và với một cỡ mẫu lớn sẽ luôn dẫn đến bác bỏ giả thiết không tại mức ý nghĩa đó. Tóm lại, ta phải đặc biệt cẩn trọng trong việc giải thích bằng chứng khi cỡ mẫu lớn, vì bất kỳ sự thay đổi nhỏ nào từ H_0 sẽ hầu chắc chắn được phát hiện bằng một kiểm định, nhưng một sự thay đổi như vậy có thể ít nghĩa thực tế.

Bảng 8.1 Minh họa của việc ảnh hưởng của cỡ mẫu lên P giá trị và β.

n	P giá trị khi $\overline{x} = 101$	β cho mức kiểm định 0,01
25	0,3085	0,9664
100	0,1587	0,9082
400	0,0228	0,6293
900	0,0013	0,2514
1600	0,0000335	0,0475
2500	0,000000297	0,0038
10,000	$7{,}69.10^{-24}$	0,0000

8.5.2 Nguyên tắc tỉ lệ hợp lý

Đặt $x_1, x_2, \ldots, x_n$ là những biến ngẫu nhiên được quan sát có kích thước n từ phân phối xác suất $f(x;\theta)$. Phân phối chung của những giá trị mẫu này là tích $f(x_1;\theta).f(x_2;\theta).\ldots.f(x_n;\theta)$. Điều này đã được nói đến về ước lượng của hàm hợp lý, hàm hợp lý là phân phối chung này, được xem như hàm theo θ, xét kiểm định H_0. θ thuộc Ω_0, ngược lại H_a thuộc Ω_a, khi Ω_0 và Ω_a rời nhau (ví dụ, $H_0 : \theta \leq 100$ ngược lại $H_a : \theta \geq 100$). Nguyên tắc tỉ lệ hợp lý cho kiểm định được xây dựng theo quá trình như sau:

1. Tìm một giá trị lớn nhất của hợp lý cho bất kỳ θ thuộc Ω_0 (bằng cách tìm ước lượng lớn nhất thuộc Ω_0 sau đó thay vào hàm hợp lý)

2. Tìm giá trị lớn nhất của hợp lý cho θ bất kỳ thuộc Ω_a

3. Tỉ lệ có dạng:

$$\lambda(x_1, \ldots, x_n) = \frac{\text{Giá trị hợp lý lớn nhất cho } \theta \in \Omega_0}{\text{Giá trị hợp lý lớn nhất cho } \theta \in \Omega_a}$$

Tỉ lệ $\lambda(x_1, \ldots, x_n)$ được gọi là giá trị tỉ lệ hợp lý thống kê, quá trình kiểm định bao gồm bác bỏ H_0 khi tỉ lệ này nhỏ, nghĩa là chọn một hằng số k sao cho H_0 bị bác bỏ nếu $\lambda(x_1, \ldots, x_n) \leq k$. Do đó, H_0 bị bác bỏ khi mẫu số của λ lớn hơn nhiều so với tử số, điều này chỉ ra rằng dữ liệu phù hợp với H_a hơn là với H_0.

Hằng số k được chọn sao cho có xác suất sai lầm loại I. Thường thì bất đẳng thức $\lambda \leq k$ có thể biến đổi để có điều kiện tương đương mà đơn giản hơn. Ví dụ, kiểm định giả thuyết $H_0 : \mu \leq \mu_0$, với đối thuyết $H_a : \mu > \mu_0$. Trong trường hợp

thông thường $\lambda \leq k$ tương đương với $t \geq c$. Nên với $c = t_{\alpha,n-1}$ thì kiểm định tỉ lệ hợp lý cũng là kiểm định một mẫu t.

Qui tắc về tỉ lệ hợp lý cũng có thể áp dụng khi X_i có những phân phối khác và thậm chí ngay cả khi nó độc lập, thông qua hàm hợp lý có thể làm cho phức tạp hơn trong trường hợp này. Một số quá trình kiểm định sẽ được trình bày trong chương sau có thể thu được từ qui tắc kiểm định hợp lý. Kiểm định đó thường xoay quanh việc "giảm thiểu β trong tất cả những kiểm định để có mức α mong muốn", vì thế đó là kiểm định đúng nhất. Để chi tiết hơn, ta có thể tham khảo một số ví dụ trong những tài liệu được liệt kê trong Chương 6.

Để có kiểm định t từ qui tắc tắc tỉ lệ hợp lý ta phải xây dựng một kiểm định thống kê hợp lý, có dạng của một phân phối xác suất, và mẫu phải được chỉ định lấy từ đó. Để lấy kiểm định t từ qui tắc tỉ lệ hợp lý, điều tra viên phải giả sử có một chuẩn pdf. Nếu điều tra viên cho rằng phân phối đó đối xứng nhưng không muốn có dạng chính xác (như chuẩn đồng dạng hay cauchy) thì sẽ thất bại vì không có cách nào để liên kết các giá trị cho tất cả các phân phối có tính đối xứng. Trong chương 15 ta sẽ trình bày một kiểm định thống kê phi tham số, vì thế, cái gọi là sai lầm xác suất loại I là sự kiểm soát tất cả những phân phối phụ thuộc khác nhau. Những quá trình này có ích khi điều tra viên có kiến thức hạn chế về phân phối phụ thuộc. Ta cũng sẽ thảo luận nhiều hơn về vấn đề 3 và 4 trong phần đầu.

Bài tập

Kiểm định trung bình tổng thể

Bài 1 Trọng lượng trung bình khi xuất chuồng ở một trại chăn nuôi gà công nghiệp năm trước là 2,8kg/con. Năm nay người ta sử dụng một loại thức ăn mới. Cân thử 25 con khi xuất chuồng người ta tính được trung bình mẫu 3,2kg và phương sai mẫu 0,25 . Với mức ý nghĩa 5%, hãy kết luận về tác dụng của loại thức ăn này có thực sự làm tăng trọng lượng trung bình của đàn gà lên hay không?

Bài 2 Cân thử 25 con khi xuất chuồng người ta tính được trung bình mẫu 3,2kg và phương sai mẫu 25 . Với mức ý nghĩa 5%, nếu trại chăn nuôi báo cáo trọng lượng trung bình khi xuất chuồng là 3,3 kg/con thì chấp nhận được không?

Bài 3 Theo báo cáo trước đây mức tiêu thụ điện trung bình trong một tháng ở phường X là 150kwh. Sau khi thực hiện chương trình tiết kiệm điện, kiểm tra ngẫu nhiên một số hộ ở phường này về mức tiêu dùng điện trong một tháng thì thu được

bảng số liệu:

X	100-110	110-120	120-130	130-140	140-150	150-160	160-170	170-180
n_i	13	44	56	69	57	45	23	12

Với mức ý nghĩa 5% hãy cho ý kiến về kết luận mức tiêu thụ điện trung bình trong một tháng ở phường X có giảm xuống hay không.

Kiểm định tỉ lệ tổng thể

Bài 4 Tỷ lệ khách hàng hài lòng với chất lượng dịch vụ tại chuỗi các salon chăm sóc sắc đẹp Y là 80%. Nhằm nâng cao hơn nữa chất lượng dịch vụ, các nhân viên được tập huấn đào tạo để nâng cao hơn nữa kỹ năng tay nghề. Sau đó, để khảo sát hiệu quả của việc tập huấn cho nhân viên, 1200 khách hàng ngẫu nhiên được hỏi về chất lượng dịch vụ của salon có 981 khách hàng hài lòng. Với mức ý nghĩa 3% có thể cho rằng việc tập huấn cho nhân viên là mang lại hiệu quả?

Bài 5 Người ta tiến hành điều tra ngẫu nhiên 400 người ở vùng A thì thấy có 22 người ở độ tuổi trưởng thành không biết chữ. Với mức ý nghĩa 2%, có thể cho rằng tỷ lệ dân số ở độ tuổi trưởng thành không biết chữ ở vùng này là 5% được hay không?

Bài 6 Tỉ lệ phế phẩm của một nhà máy trước đây là 8%. Năm nay nhà máy ứng dụng biện pháp kỹ thuật cải tiến hơn. Để xét hiệu quả của việc cải tiến kỹ thuật, người ta lấy một mẫu gồm 710 sản phẩm để kiểm tra và thấy có 30 phế phẩm. Với mức ý nghĩa 4% có thể cho rằng biện pháp kỹ thuật cải tiến có làm giảm tỉ lệ phế phẩm hay không ?

Kiểm định dùng P-value

Bài 7 Độ dày tiêu chuẩn của miếng Silicon wafers được sử dụng trong mạch là $245\mu m$. Một mẫu 50 wafers có độ dày trung bình là 246,18 μm và độ lệch chuẩn là $3,6\mu m$. Từ dữ liệu hãy kết luận liệu độ dày trung bình của loại Silicon wafer này có khác với độ dày tiêu chuẩn không?

Bài 8

1. Cho $H_0 : \mu = 100; H_a : \mu > 100$ với $n = 9$, $(\sigma)^2$ chưa biết. Tính được tiêu chuẩn kiểm định $t = 1,6$. Hãy tìm P giá trị? Và kết luận?

2. Cho $H_0 : \mu = 100; H_a : \mu < 100$ với $n = 21$, σ^2 chưa biết. Tính được tiêu chuẩn kiểm định $t = 1,6$ Hãy tìm P giá trị? Và kết luận?

2. Cho $H_0 : \mu = 100; H_a : \mu \neq 100$ với $n = 21$, σ^2 chưa biết. Tính được tiêu chuẩn kiểm định $t = 1,6$. Hãy tìm P giá trị? Và kết luận?

Bài tập tổng hợp

Bài 9 Năng suất lúa trung bình trong những vụ trước là 5,5 tấn/ha. Vụ lúa năm nay người ta áp dụng một biện pháp kĩ thuật mới. Điều tra 100 hecta lúa ta có bảng:

Năng suất (tạ/ha)	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-85
Diện tích (ha)	7	12	18	27	20	8	5	3

Với mức ý nghĩa 5%, kết luận xem biện pháp kĩ thuật mới có làm tăng năng suất lúa trung bình của vùng này lên không?

Bài 10 Trọng lượng sản phẩm A của nhà máy M là biến ngẫu nhiên X có phân phối chuẩn với khối lượng trung bình quy định 50kg và độ lệch chuẩn là 0,25 kg. Nghi ngờ dây chuyền sản xuất không bình thường nên tiến hành kiểm tra khối lượng một số sản phẩm được số liệu

X	49-49,25	49,25-49,5	49,5-49,75	49,75-50	50-50,25	50,25-50,5	50,5-50,75
n_i	5	8	10	16	14	11	9

Với mức ý nghĩa 3%, hãy cho nhận xét về nghi ngờ trên.

Bài 11 Tuổi thọ trung bình của một loại thiết bị điện tử B do nhà máy N sản xuất là 100 giờ. Sau khi cải tiến kỹ thuật quan sát thời gian sử dụng của 650 thiết bị B do nhà máy sản xuất ta thu được giá trị trung bình mẫu là 102,215 giờ và giá trị độ lệch chuẩn mẫu là 15,69 giờ. Đưa ra nhận xét về ý kiến việc cải tiến kỹ thuật có mang lại hiệu quả với mức ý nghĩa 3%.

Bài 12 Quan sát mức chi tiêu nhu yếu phẩm (triệu đồng/năm) của một hộ thì thu được bảng:

Chi tiêu	4	6	8	10	12
Số hộ	15	16	20	14	15

Những hộ chi tiêu dưới 7 triệu/tháng là chi tiêu thấp? Trước đây tỉ lệ chi tiêu thấp là 30%. Hãy kiểm định xem tỷ lệ hộ chi tiêu thấp bây giờ đã tăng lên chưa? Với $\alpha = 5\%$.

Bài 13 Trọng lượng của một loại sản phẩm do một nhà máy sản xuất là đại lượng ngẫu nhiên có phân phối chuẩn với trọng lượng trung bình là 500gr. Sau một thời gian sản xuất người ta nghi ngờ trọng lượng của loại sản phẩm này có xu hướng giảm sút nên tiến hành cân thử 25 sản phẩm và thu được kết quả cho ở bảng sau:

Trọng lượng (gr)	480	485	490	495	500	510
Số sản phẩm	2	3	8	5	3	4

Với mức ý nghĩa 5%, hãy kết luận điều nghi ngờ trên có đúng hay không?

Bài 14 Khảo sát thu nhập của một số người của một công ty, người ta thu được bảng sau:

Thu nhập (triệuđồng/năm)	26-32	32-36	36-40	40-44	44-48	48-54	54-60
Số người	8	12	20	25	20	10	5

Nếu công ty báo cáo mức thu nhập bình quân của một người là 3,6 triệu đ/tháng thì có chấp nhận được không? Kết luận với mức ý nghĩa $\alpha = 4\%$.

Bài 15 Một công ty lớn chuyên sản xuất phần mềm máy tính, cho rằng những người làm việc ở công ty này có thu nhập trung bình 5 triệu đồng/tháng. Lấy mẫu ở công ty được bảng:

Thu nhập (triệu đồng/tháng)	3	4	5	8	10
Số người	6	7	8	2	2

Giả sử thu nhập của những người làm việc ở công ty này có phân phối chuẩn. Với mức ý nghĩa 2%, hãy cho nhận xét về thông tin thu nhập trung bình ở trên có đáng tin hay không?

Bài 16 Trong 2115 trẻ sơ sinh chọn ngẫu nhiên ở vùng A có 1115 bé trai. Với mức ý nghĩa 5% có thể kết luận mất cân đối giới tính ở vùng A không?

Bài 17 Năm trước tỷ lệ đạt giải của đội tuyển Olympic tỉnh là 70%. Sau khi triển khai phương pháp học tập mới, người ta tiến hành khảo sát kết quả của đội tuyển thì 120 em chọn ngẫu nhiên thì thấy có 30 em bị trượt. Với $\alpha = 5\%$ hãy kiểm định xem phương pháp mới có mang lại hiệu quả hơn?

Bài 18 Khảo sát tuổi thọ X (đơn vị: tháng) của một số sản phẩm chọn ngẫu nhiên từ công ty A:

X	6-9	9-12	12-15	15-18	18-21	21-24	24-27
n_i	23	33	55	73	57	42	35

1. Tìm khoảng ước lượng với độ tin cậy 98% cho tuổi thọ trung bình của sản phẩm công ty A.

2. Dây chuyền sản xuất công ty A hoạt động bình thường nếu tuổi thọ trung bình của sản phẩm sản xuất ra là 18 tháng. Với mức ý nghĩa 1% hãy xem dây chuyền có hoạt động bình thường không?

Bài 19 Công ty M có 3000 đại lý, cho tiến hành điều tra ngẫu nhiên một số đại lý

của mình và thu được bảng số liệu sau (X là doanh số, đơn vị: triệu đồng/tháng), biết X có phân phối chuẩn.

X	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Số đại lý	7	12	18	27	22	17	13	4

1. Những đại lý có X>45 triệu đồng/tháng gọi là đại lý có doanh số cao. Hãy ước lượng số đại lý có doanh số cao với độ tin cậy 95%.

2. Có ý kiến cho rằng tỉ lệ đại lý có doanh số cao bằng 1/3 tỉ lệ đại lý có doanh thu còn lại. Hãy cho nhận xét về ý kiến này với mức ý nghĩa 1%.

3. Hãy ước lượng doanh số trung bình/tháng của các đại lý với độ tin cậy 99%.

Bài 20 Mức tiêu thụ X của mỗi hộ gia đình vùng A trong mùa khô năm nay có phân phối chuẩn. Điều tra 1 số hộ gia đình vùng A có thống kê sau

X(kwh/t)	65-115	115-165	165-215	215-265	265-315	315-365	365-415	415-465
Số hộ	24	36	75	94	97	125	84	75

1. Mức tiêu thụ điện trung bình các hộ gia đình vùng A trước là 280 kwh/tháng. Với mức ý nghĩa 2% hãy xét xem mức tiêu thụ điện trung bình các hộ gia đình vùng A năm nay có tăng lên không.

2.* Với mức ý nghĩa 5% so sánh tỉ lệ hộ gia đình có mức tiêu thụ $X > 315$kwh/t với tỉ lệ hộ gia đình có mức tiêu thụ $X \leq 315$kwh/t vùng A.

3. Hộ có $X > 315$kwh/t là hộ có mức tiêu thụ cao. Hãy ước lượng số hộ có mức tiêu thụ điện cao với độ tin cậy 95%, biết vùng này có 3000 hộ.

4. Nếu muốn ước lượng mức tiêu thụ điện trung bình các hộ vùng A trong mùa khô năm nay với độ chính xác 10 kwh/tháng thì độ tin cậy bằng ?

Bài 21 Tìm P giá trị trong các trường hợp sau:

1. Kiểm định phía phải với z =1,52

2. Kiểm định 2 phía với z = 2,2

3. Kiểm định phía phải với z = -1,2

4. Kiểm định 2 phía với z = -0,65

5. Kiểm định phải với bậc tự do (df) = 8; t=2
6. Kiểm định trái với n = 12 ; t = -2,5
7. Kiểm định 2 phía với bậc tự do (df) =15 ; t = -1,6

Bài 22 Quan sát dữ liệu về cường độ của bê tông (MPa)được mẫu sau:

112,3	97	92	80	101	99,2	95,8	103,5	89	86

Giả sử bê tông sẽ được sử dụng nếu cường độ trung bình của loại bê tông này lớn hơn 100MPa. Liệu bê tông này có được sử dụng không với mức ý nghĩa 5%? Sử dụng kiểm định theo phương pháp P giá trị.

Bài 23 Một mẫu 462 sinh viên trường X có 51 em sử dụng rượu bia thường xuyên. Có thể kết luận rằng tỉ lệ sinh viên sử dụng rượu bia thường xuyên của toàn trường lớn hơn 10% được không với mức ý nghĩa 10%? Dùng P giá trị để đưa ra kết luận.

Chương 9

CÁC KẾT LUẬN DỰA TRÊN HAI MẪU

Giới thiệu

Chương 7 và 8 đã trình bày về khoảng tin cậy và các bước kiểm định giả thuyết cho giá trị trung bình μ, tỷ lệ p, và phương sai σ^2. Ở đây chúng ta mở rộng các phương pháp cho trung bình, tỷ lệ, và phương sai của hai tổng thể khác nhau. Ví dụ, μ_1 kí hiệu cho thu nhập trung bình thực sự của dân cư quận Thủ Đức thuộc Thành phố Hồ Chí Minh năm 2018 và μ_2 thu nhập trung bình thực sự của dân cư quận Thủ Đức năm 2019. Quan sát viên có thể muốn sử dụng các mẫu quan sát về thu nhập trung bình của dân cư quận Thủ Đức mỗi năm làm cơ sở để ước lượng khoảng $\mu_1 - \mu_2$ (hiệu giữa hai thu nhập). Ví dụ khác, cho p_1 kí hiệu cho tỷ lệ sản phẩm lỗi thực sự của các điện thoại Samsung được sản xuất trong điều kiện hiện tại và p_2 kí hiệu cho tỷ lệ sản phẩm lỗi thực sự của các điện thoại Samsung được sản xuất trong điều kiện sản xuất đã thay đổi. Nếu nhờ vào điều kiện sản xuất sửa đổi mà làm giảm được tỷ lệ các sản phẩm lỗi, một kỹ sư quản lý chất lượng sẽ muốn sử dụng thông tin mẫu để kiểm định giả thuyết ban đầu $H_0 : p_1 - p_2 = 0$ (tức là $p_1 = p_2$) so với giả thuyết đối, $H_a : p_1 - p_2 > 0$ (tức là $p_1 > p_2$).

9.1 Kiểm định z và khoảng tin cậy cho hiệu giữa hai trung bình

Các kết luận được đề cập trong phần này liên quan đến $\mu_1 - \mu_2$ là hiệu của trung bình hai mẫu khác nhau.

Giả Thiết 9.1.

1. *$X_1, X_2, \ldots, X_m$ là một mẫu ngẫu nhiên từ một phân phối với giá trị trung bình là μ_1 và phương sai σ_1^2.*

2. *$Y_1, Y_2, \ldots, Y_m$ là một mẫu ngẫu nhiên từ một phân phối với giá trị trung bình là μ_2 và phương sai σ_2^2.*

3. *Mẫu X và Y độc lập với nhau.*

Việc sử dụng m cho số lần quan sát trong mẫu thứ nhất và n cho số lần quan sát trong mẫu thứ hai, cho phép hai cỡ mẫu khác nhau. Đôi khi điều này là vì việc lấy mẫu của một tổng thể này khó khăn hơn hoặc tốn chi phí hơn so với một tổng thể khác. Một số tình huống khác mà kích cỡ mẫu được lấy là bằng nhau ngay từ đầu, nhưng vì lý do vượt quá phạm vi thử nghiệm, mà kích thước mẫu thực tế có thể khác nhau.

$\overline{X} - \overline{Y}$ là ước lượng tự nhiên của $\mu_1 - \mu_2$, cũng là hiệu của trung bình các mẫu tương ứng. Các quy trình kết luận được dựa trên chuẩn hóa của ước lượng tự nhiên này, vì vậy chúng ta cần biểu thức cho giá trị kỳ vọng và độ lệch chuẩn của $\overline{X} - \overline{Y}$.

Mệnh đề 9.2. *Giá trị kỳ vọng của $\overline{X} - \overline{Y}$ là $\mu_1 - \mu_2$, cho nên $\overline{X} - \overline{Y}$ là một ước lượng không chệch của $\mu_1 - \mu_2$. Độ lệch chuẩn của $\overline{X} - \overline{Y}$ là:*

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Nếu chúng ta coi $\mu_1 - \mu_2$ là một tham số θ, thì ước lượng của nó là $\hat{\theta} = \overline{X} - \overline{Y}$ với độ lệch chuẩn $\sigma_{\hat{\theta}}$ được tính như mệnh đề trên. Với giá trị của σ_1^2 và σ_2^2 đều đã biết, thì giá trị của độ lệch chuẩn này có thể được tính như trên. Khi σ_1^2 và σ_2^2 đều không biết trước, phương sai mẫu được dùng để ước lượng $\sigma_{\hat{\theta}}$.

9.1.1 Các bước kiểm định tổng thể chuẩn với phương sai đã biết

Trong chương 7 và 8, khoảng tin cậy và các bước kiểm định đầu tiên cho trung bình tổng thể μ dựa trên giả định rằng tổng thể là phân phối chuẩn với giá trị của phương sai tổng thể σ^2 đã biết trước . Tương tự như vậy, trước tiên chúng ta giả sử rằng ở đây cả hai phân phối tổng thể là chuẩn và giá trị phương sai của cả hai là σ_1^2 và σ_2^2 đã biết. Các trường hợp trong đó một hoặc cả hai giả định này có thể được loại bỏ sẽ được trình bày ngắn gọn sau.

Bởi vì phân phối tổng thể là chuẩn, cả $\overline{X}$ và $\overline{Y}$ đều có phân phối chuẩn. Hơn nữa, sự độc lập của hai mẫu suy ra hai trung bình mẫu cũng độc lập với nhau. Do đó, hiệu $\overline{X} - \overline{Y}$ cũng là phân phối chuẩn, với giá trị kỳ vọng $\mu_1 - \mu_2$ và độ lệch chuẩn $\sigma_{\overline{X}-\overline{Y}}$ được suy ra từ mệnh đề nêu trên. Chuẩn hóa $\overline{X} - \overline{Y}$ cho biến chuẩn hóa.

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \tag{9.1}$$

Trong một bài toán kiểm định giả thuyết, giả thuyết ban đầu sẽ chỉ ra rằng $\mu_1 - \mu_2$ có một giá trị xác định. Gán giá trị giả thuyết này bằng Δ_0, chúng ta có $H_0 : \mu_1 - \mu_2 = \Delta_0$. Thông thường $\Delta_0 = 0$, trong trường hợp đó H_0 sẽ là $\mu_1 = \mu_2$. Một kết quả thống kê thử nghiệm từ việc thay thế $\mu_1 - \mu_2$ trong công thức (9.1) bằng giá trị Δ_0. Kiểm định thống kê Z thu được bằng cách chuẩn hóa $\overline{X} - \overline{Y}$ với giả định rằng H_0 là đúng, vì vậy trường hợp này nó có phân phối chuẩn chuẩn tắc. Kiểm định thống kê này có thể được viết là $(\hat{\theta} - \Delta_0)/\sigma_{\hat{\theta}}$, cùng dạng như một số kiểm định thống kê trong Chương 8.

Giả thuyết không $H_0 : \mu_1 - \mu_2 = \Delta_0$	
Giá trị kiểm định thống kê: $z = \frac{\overline{x}-\overline{y}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$	
Các giả thuyết đối:	Miền bác bỏ với kiểm định α
$H_a : \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (phía phải)
$H_a : \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (phía trái)
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$z \geq z_{\alpha/2}$ hoặc $z \leq -z_{\alpha/2}$ (2 phía)

Bởi vì đây là những kiểm định z, P giá trị đã được tính như cho kiểm định z ở chương 8 (ví dụ: P giá trị $= 1 - \Phi(z)$ cho bài kiểm định một phía phải).

Ví dụ 9.1 Phân tích một mẫu ngẫu nhiên gồm $m = 20$ mẫu pin AAA của hãng M cho một loại sản phẩm để xem tuổi thọ trung bình $\overline{x} = 35,5$ giờ. Một mẫu ngẫu nhiên thứ hai gồm $n = 25$ pin AAA của hãng N cho cùng loại sản phẩm đó có tuổi thọ trung bình là $\overline{y} = 34,7$ giờ . Giả sử rằng hai phân phối cho tuổi thọ trung bình của 2 mẫu trên có phân phối chuẩn với $\sigma_1 = 3,0$ và $\sigma_2 = 4,0$. Dữ liệu trên có chỉ ra được giá trị trung bình thật sự của hai tuổi thọ trên μ_1 và μ_2 là khác nhau không? Hãy thực hiện kiểm định với mức ý nghĩa $\alpha = 0,05$.

Giải

Để kiểm định hai tuổi thọ trung bình trên có khác nhau hay không, nghĩa là ta so sánh hiệu hai tuổi thọ trên với $\Delta_0 = 0$. Hơn nữa, đây là bài toán kiểm định hai phía.

Giả thuyết không: $H_0 : \mu_1 - \mu_2 = 0$; Giả thuyết đối: $H_a : \mu_1 - \mu_2 \neq 0$

Với $\alpha = 0,05$ thì $\phi(z_{\alpha/2}) = 1 - \alpha/2 = 0,975$ nên $z_{\alpha/2} = 1,96$

Tiêu chuẩn kiểm định: $z = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}} = 0,76626$

Kết luận: Theo tiêu chuẩn kiểm định 2 phía, do $z < z_{\alpha/2}$ nên tạm chấp nhận giả thuyết không $H_0 : \mu_1 - \mu_2 = 0 \to \mu_1 = \mu_2$. Vậy tuổi thọ trung bình pin AAA của hai hãng là bằng nhau.

9.1.2 Kiểm định với mẫu cỡ lớn

Các giả định về phân phối chuẩn của tổng thể và các giá trị σ_1 và σ_2 may mắn không cần thiết khi cả hai cỡ mẫu đều đủ lớn. Trong trường hợp này, Định lý Giới hạn Trung tâm (Central Limit Theorem - CLT) đảm bảo rằng $\overline{X} - \overline{Y}$ xấp xỉ phân phối chuẩn bất kể phân phối của tổng thể như thế nào. Hơn nữa, sử dụng S_1^2 và S_2^2 thay thế cho σ_1^2 và σ_2^2 và trong Biểu thức (9.1) đưa ra một biến mà phân phối của nó xấp xỉ phân phối chuẩn chuẩn tắc:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

Một kết quả từ kiểm định thống kê với mẫu lớn bằng cách thay thế $\mu_1 - \mu_2$ bởi Δ_0, giá trị kỳ vọng của $\overline{X} - \overline{Y}$ khi H_0 đúng. Thống kê Z này sẽ có xấp xỉ phân phối chuẩn chuẩn tắc khi H_0 là đúng. Kiểm định với mức ý nghĩa mong muốn mà thu được bằng cách sử dụng một giá trị tới hạn z vẫn chính xác như trước đây.

Sử dụng giá trị kiểm định thống kê:

$$z = \frac{\overline{x}-\overline{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$$

cùng với các trường hợp miền bác bỏ bên phải, trái, và 2 phía dựa trên các giá trị tới hạn z cho kiểm định mẫu lớn mức ý nghĩa xấp xỉ α.
Những kiểm định này thường thích hợp nếu cả $m > 40$ và $n > 40$.
P giá trị cũng được tính chính xác như các bài kiểm định z trước đây.

Ví dụ 9.2 Báo cáo dữ liệu tóm tắt dưới đây về lượng calo tiêu thụ khi hoạt động thể thao 30 phút đối với người 56kg. Lượng calo tiêu thụ (kcal) Từ dữ liệu này có

Môn thể thao	**Cỡ mẫu**	**Trung bình mẫu**	**Độ lệch chuẩn mẫu**
Bơi lội	92	320 kcal	80
Chạy 10km/h	110	290 kcal	60

kết luận được lượng calo tiêu thụ trung bình của người bơi lội cao hơn 5 kcal so với người chạy bộ không? Hãy kiểm định kết luận này với mức ý nghĩa 0,01.

Giải

Gọi μ_1, μ_2 lần lượt là lượng calo tiêu thụ trung bình của người bơi lội và người chạy bộ với vận tốc 10km/h trong 30 phút .

$n_1 = 92; s_1 = 80; \overline{x} = 320$

$n_2 = 110; s_2 = 60; \overline{y} = 290$

Giả thuyết không: $H_0 : \mu_1 - \mu_2 = 5$; $(\Delta_0 = 5)$

Giả thuyết đối: $H_a : \mu_1 - \mu_2 \neq 5$

Với $\alpha = 0,01$ thì $\phi(z_{\alpha/2}) = 1 - \alpha/2 = 0,995$ nên $z_{\alpha/2} = 2,576$.

Tiêu chuẩn kiểm định: $z = \frac{\overline{x}-\overline{y}-\Delta_0}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}} = 2,4718$

Kết luận: Theo tiêu chuẩn kiểm định 2 phía, do $-z_{\alpha/2} < z < z_{\alpha/2}$ nên chấp nhận giả thuyết $H_0 : \mu_1 - \mu_2 = 5$.

Vậy trong 30 phút, lượng calo tiêu thụ trung bình khi bơi lội sẽ hơn lượng calo trung bình khi chạy bộ 10km/h là 5kcal.

9.1.3 Khoảng tin cậy cho $\mu_1 - \mu_2$

Khi cả hai phân phối tổng thể là phân phối chuẩn, chuẩn hóa $\overline{X} - \overline{Y}$ cho một biến ngẫu nhiên Z với phân phối chuẩn chuẩn tắc. Vì phần diện tích dưới đường cong của z nằm giữa $-z_{\alpha/2}$ và $z_{\alpha/2}$ là $1 - \alpha$, nên theo đó .

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Biến đổi các bất đẳng thức nằm trong ngoặc đơn để tách biệt $\mu_1 - \mu_2$ nhận được đẳng thức xác suất sau

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha$$

Như vậy khoảng tin cậy $100(1-\alpha)$ % cho $\mu_1 - \mu_2$ có giới hạn dưới $\overline{x} - \overline{y} - z_{\alpha/2}.\sigma_{\overline{X}-\overline{Y}}$ và giới hạn trên $\overline{x} - \overline{y} + z_{\alpha/2}.\sigma_{\overline{X}-\overline{Y}}$. Khoảng này là một trường hợp đặc biệt của công thức tổng quát $\hat{\theta} \pm z_{\alpha/2}.\sigma_{\hat{\theta}}$.

Nếu cả m và n đều lớn, Định lý giới hạn trung tâm chỉ ra rằng khoảng này có giá trị ngay cả khi không có giả định về các tổng thể chuẩn; trong trường hợp này, độ tin cậy xấp xỉ $100(1-\alpha)\%$. Hơn nữa, việc sử dụng sai số mẫu S_1^2 và S_2^2 trong biến chuẩn hóa Z thu được một khoảng hợp lệ trong đó s_1^2 và s_2^2 thay thế cho σ_1^2 và σ_2^2.

Nếu m và n đủ lớn, khoảng tin cậy cho $\mu_1 - \mu_2$ với mức độ tin cậy xấp xỉ $100(1-\alpha)\%$ là

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

Trong đó dấu "–" cho giới hạn dưới và dấu "+" cho giới hạn trên của khoảng. Biên của khoảng tin cậy phía phải hoặc phía trái có thể được tính bằng cách giữ lại dấu hiệu (+ hoặc –) và thay thế $z_{\alpha/2}$ bởi z_α.

Ví dụ 9.3 Một thí nghiệm được thực hiện để nghiên cứu các đặc điểm khác nhau của bê tông theo cấp độ bền. Quan sát 70 mẫu về cường độ chịu nén của bê tông có cấp độ bền B15 (MPa), và 80 mẫu về cường độ chịu nén của bê tông có cấp độ bền B20. Tóm tắt các giá trị thể hiện ở bảng bên dưới. Hãy tìm khoảng tin cậy cho hiệu giữa cường độ chịu nén thực sự của hai loại cấp độ bền trên, với độ tin cậy 99%.

Giải

Cấp độ bền	**Cỡ mẫu**	**Trung bình mẫu**	**Độ lệch chuẩn mẫu**
B15	70	860 MPa	20
B20	80	1150 MPa	30

Gọi μ_1, μ_2 lần lượt là cường độ chịu nén của bê tông có cấp độ bền B15 và B20

Mẫu B15: $n_1 = 70; s_1 = 20; \overline{x} = 860$

Mẫu B20: $n_2 = 80; s_2 = 30; \overline{y} = 1150$

Với $\gamma = 0,99$ thì $\phi(z_{\alpha/2}) = \frac{1+\gamma}{2} = 0,995$ nên $z_{\alpha/2} = 2,58$

Sai số ước lượng: $\varepsilon = z_{\alpha/2}\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}} = 10,6264$

Khoảng tin cậy cho hiệu giữa cường độ chịu nén thực sự của hai loại cấp độ bền trên là :

$$(\overline{x} - \overline{y} - \varepsilon; \overline{x} - \overline{y} + \varepsilon) = (-300,6264; -279,3736).$$

Nếu các phương sai σ_1^2 và σ_2^2 đều đã biết (hoặc biết giá trị xấp xỉ) và điều tra viên sử dụng các cỡ mẫu giống nhau, thì kích thước mẫu chung n để thu được một khoảng tin cậy $100(1-\alpha)\%$ có chiều rộng w là

$$n = \frac{4z_{\alpha/2}^2(\sigma_1^2 + \sigma_2^2)}{w^2}$$

9.2 Kiểm định t cho 2 mẫu và khoảng tin cậy

Thông thường các giá trị của phương sai tổng thể sẽ không được biết trước. Trong phần trước, chúng ta đã minh họa cho các cỡ mẫu lớn khi sử dụng một bài kiểm định z và khoảng tin cậy, trong đó các phương sai mẫu đã được sử dụng thay cho phương sai tổng thể. Đối với các mẫu lớn, định lý giới hạn trung tâm cho phép chúng ta sử dụng các phương pháp này ngay cả khi hai tổng thể là không chuẩn.

Trong thực tế, mặc dù, thường xảy ra trường hợp ít nhất một mẫu có kích thước nhỏ và phương sai tổng thể là chưa biết. Nếu không sử dụng được định lý giới hạn trung tâm, chúng ta có thể đưa ra các giả định cụ thể cho phân phối tổng thể. Việc sử dụng các bước suy luận dựa theo những giả định này sẽ được giới hạn trong các tình huống mà các giả định ít nhất ở mức chấp nhận được. Ví dụ, chúng ta có thể giả định rằng cả hai phân phối tổng thể đều có phân phối Weibull hoặc cả hai đều là phân phối Poisson. Dù trong thực tế thì phân phối chuẩn thường là giả định hợp lý nhất.

Giả Thiết 9.3. *Cả 2 phân phối tổng thể là chuẩn, do đó $X_1, X_2, \ldots, X_m$ là mẫu ngẫu nhiên từ phân phối chuẩn và tương tự vậy cho $Y_1, Y_2, \ldots, Y_n$ (với mẫu X và mẫu Y là độc lập với nhau).*

Định lý 9.4. *Khi các phân phối tổng thể là chuẩn, biến được chuẩn hóa*

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \tag{9.2}$$

sẽ xấp xỉ một phân phối t với bậc tự do ν được ước lượng từ dữ liệu bởi công thức

$$v = \frac{(\frac{s_1^2}{m} + \frac{s_2^2}{n})^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = \frac{[(se_1)^2 + (se_2)^2]^2}{\frac{(se_1)^4}{m-1} + \frac{(se_2)^4}{n-1}}$$

Trong đó

$$se_1 = \frac{s_1}{\sqrt{m}}, se_2 = \frac{s_2}{\sqrt{n}}$$

(ν được làm tròn xuống số nguyên gần nhất)

Biến đổi T trong biểu thức xác suất để tách riêng $\mu_1 - \mu_2$ cho khoảng tin cậy, trong khi một kiểm định thống kê thu được từ thay thế $\mu_1 - \mu_2$ bằng giá trị không Δ_0.

Khoảng tin cậy của $\mu_1 - \mu_2$ với kiểm định t cho 2 mẫu và độ tin cậy $100(1-\alpha)\%$ là

$$\overline{x} - \overline{y} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

Khoảng tin cậy một phía có thể được tính như đã đề cập ở các phần trước.
Kiểm định t cho 2 mẫu để kiểm định giả thuyết $H_0 : \mu_1 - \mu_2 = \Delta_0$ như sau:

Giá trị kiểm định thống kê: $t = \frac{\overline{x}-\overline{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$

Các giả thuyết đối:	Vùng bác bỏ với kiểm định xấp xỉ mức ý nghĩa α
$H_a : \mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,\nu}$ một phía phải
$H_a : \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,\nu}$ một phía trái
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$t \geq t_{\alpha/2,\nu}$ hoặc $t \leq -t_{\alpha/2,\nu}$ hai phía.

Giá trị P có thể được tính như đã đề cập trong Chương 8 cho kiểm định t cho 1 mẫu.

Ví dụ 9.4 Quan sát doanh thu trung bình mỗi tháng (triệu đồng) của 15 cửa hàng tiện lợi Bách hoá xanh và 10 cửa hàng tiện lợi Satra năm 2018 thì được dữ liệu sau:

Cửa hàng	**Cỡ mẫu**	**Trung bình mẫu**	**Độ lệch chuẩn mẫu**
Bách hoá xanh	15	800	5
Satra	10	950	10

Giả sử doanh thu trung bình mỗi tháng cho hai loại cửa hàng này có phân phối chuẩn.

a/ Hãy tính khoảng tin cậy cho hiệu giữa doanh thu trung bình thực sự mỗi tháng của hai loại cửa hàng này với độ tin cậy 98%.

b/ Có thể kết luận là doanh thu trung bình thực sự mỗi tháng của Bách hoá xanh thấp hơn Satra được không? Kiểm định điều này với mức ý nghĩa 0,1.

Giải

a/ Gọi μ_1, μ_2 lần lượt là doanh thu trung bình mỗi tháng cho hai loại cửa hàng Bách hoá xanh và Satra

Mẫu Bách hóa xanh $m = 15; s_1 = 5; \overline{x} = 800$

Mẫu Satra $n = 10; s_2 = 10; \overline{y} = 950$

$$v = \frac{(\frac{s_1^2}{m} + \frac{s_2^2}{n})^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = 12,035 \approx 12$$

Với $\gamma = 0,98$ thì $\alpha = 1 - \gamma = 0,02$ nên $t_{\alpha/2;v} = t_{0,01;12} = 2,681$

Sai số ước lượng: $\varepsilon = t_{\alpha/2;v}\sqrt{\frac{{s_1}^2}{m} + \frac{{s_2}^2}{n}} = 9,1573$

Khoảng tin cậy cho hiệu giữa doanh thu trung bình thực sự mỗi tháng của Bách hoá xanh và Satra là :

$$(\overline{x} - \overline{y} - \varepsilon; \overline{x} - \overline{y} + \varepsilon) = (-159,1573; -140,8427).$$

b/ Giả thuyết không: $H_0 : \mu_1 - \mu_2 = 0$; $(\Delta_0 = 0)$

Giả thuyết đối: $H_a : \mu_1 - \mu_2 \neq 0$

Với $\alpha = 0,1$ thì $t_{\alpha/2;v} = t_{0,05;12} = 1,782$

Tiêu chuẩn kiểm định:

$$t = \frac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{\frac{{s_1}^2}{m} + \frac{{s_2}^2}{n}}} = -43,9155$$

Kết luận: Do $z < -z_{\alpha/2}$ nên bác bỏ H_0, chấp nhận $H_a : \mu_1 - \mu_2 \neq 0$. Mặt khác ta có $\overline{x} < \overline{y}$ nên ta chấp nhận kết luận doanh thu trung bình mỗi ngày của Bách hoá xanh trong năm 2018 thấp hơn Satra .

9.3 Phân tích trên số liệu ghép cặp

Trong Phần 9.1 và 9.2, chúng ta đã xem xét kết luận về hiệu giữa hai giá trị trung bình μ_1 và μ_2. Điều này được thực hiện bằng cách sử dụng các kết quả của mẫu ngẫu nhiên $X_1, X_2, \ldots, X_m$ từ phân phối với trung bình μ_1 và mẫu hoàn toàn độc lập (với mẫu X) $Y_1, Y_2, \ldots Y_n$ từ phân phối với giá trị trung bình μ_2. m đối tượng được chọn từ tổng thể 1 và n đối tượng khác từ tổng thể 2, và 2 tổng thể này là độc lập với nhau. Ngược lại, có một số tình huống thử nghiệm trong đó hai đặc tính được quan sát trên một bộ n cá thể hoặc đối tượng thí nghiệm.

Mẫu gồm n cặp được chọn độc lập $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$, với $E(X_i) = \mu_1$ và $E(Y_i) = \mu_2$. Đặt $D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \ldots, D_n = X_n - Y_n$ thì D_i là khác nhau từng cặp. D có trung bình $\mu_D = \mu_X - \mu_Y$ và $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$.

Vì mỗi cặp khác nhau là độc lập, nên mẫu D_i cũng độc lập với nhau. Đặt $D = X - Y$, trong đó X và Y tương ứng là quan sát đầu tiên và thứ hai trong một cặp tùy ý. Vậy hiệu kỳ vọng là:

$$\mu_D = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2$$

(quy tắc của các giá trị kỳ vọng được dùng ở đây là hợp lệ ngay cả khi X và Y phụ thuộc). Do đó bất kỳ giả thuyết nào về $\mu_1 - \mu_2$ có thể được diễn đạt như một giả thuyết về hiệu trung bình μ_D.

Khi mẫu X_i, Y_i có phân phối chuẩn thì mẫu D_i có phân phối chuẩn. Khoảng ước lượng cho $\mu_1 - \mu_2$ thu được khi tìm khoảng ước lượng cho μ_D dựa trên thông tin trên mẫu D_i. Kết luận về kiểm định giả thuyết đối với $\mu_1 - \mu_2 = \Delta_0$ tương ứng là kết luận cho giả thuyết $\mu_D = \Delta_0$ dựa trên mẫu D_i.

9.3.1 Kiểm định t cặp

Khi mẫu D_i tạo thành một mẫu ngẫu nhiên với trung bình μ_D, giả thuyết về μ_D có thể được kiểm định bằng cách sử dụng một kiểm định t một mẫu. Nghĩa là, để kiểm tra các giả thuyết về $\mu_1 - \mu_2$ khi dữ liệu là các cặp, ta tạo giá trị hiệu $D_1, D_2, \ldots, D_n$ và thực hiện một bài kiểm định t một mẫu (dựa trên bậc tự do $n-1$) trên các hiệu này. Khi số lượng các cặp n là lớn, giả định về phân phối chuẩn cho

Kiểm định t cặp.
Giả thuyết không: $H_0 : \mu_D = \Delta_0$ (trong đó $D = X - Y$ là hiệu giữa quan sát thứ nhất và quan sát thứ hai trong một cặp, và $\mu_d = \mu_1 - \mu_2$).
Giả trị thống kê: $t = \frac{\overline{d} - \Delta_0}{s_D/\sqrt{n}}$ (trong đó $\overline{d}$ và s_D là trung bình mẫu và độ lệch chuẩn tương ứng của mẫu d_i).

Các giả thuyết đối:	Miền bác bỏ với mức ý nghĩa α
$H_a : \mu_D > \Delta_0$	$t \geq t_{\alpha,n-1}$
$H_a : \mu_D < \Delta_0$	$t \leq -t_{\alpha,n-1}$
$H_a : \mu_D \neq \Delta_0$	$t \geq t_{\alpha/2,n-1}$ và $t \leq -t_{\alpha/2,n-1}$

Giá trị P được tính như các kiểm định t trước.

hiệu là không cần thiết. Theo định lý Giới hạn Trung tâm ta có kết quả của kiểm định z tương ứng.

Ví dụ 9.5 Một nghiên cứu để xác định môn học khác nhau có ảnh hưởng đến kết quả học tập không. Quan sát một mẫu gồm $n = 12$ sinh viên làm bài tập đại số và giải tích và được kết quả về điểm số như bảng sau.

Đối tượng sinh viên	1	2	3	4	5	6	7	8	9	10	11	12
Đại số	7.7	7.6	7.3	6.8	7.9	7.4	7.4	6.9	7.6	7.2	7.2	7.6
Giải tích	7	7.1	7.2	7.2	6.6	6.9	7.1	7.3	7.6	7.6	8	6.9
Hiệu	0.7	0.5	0.1	-0.4	1.3	0.5	0.3	-0.4	0	-0.4	-0.8	0.7

Dữ liệu có cho thấy điểm trung bình có khác nhau với môn học khác nhau không? Kiểm định điều này với mức ý nghĩa 0,05.

Giải

Gọi μ_D là hiệu của điểm trung bình thực sự của môn đại số với điểm trung bình thực sự của môn giải tích.

$n = 12; s = 0,6047; \overline{d} = 0,175$

Giả thuyết không: $H_0 : \mu_D = 0$; $(\Delta_0 = 0)$

Giả thuyết đối: $H_a : \mu_D \neq 0$

Với $\alpha = 0,05$ thì $t_{\alpha/2,n-1} = t_{0,025;11} = 2,201$

Tiêu chuẩn kiểm định: $t = \frac{\overline{d}-\Delta_0}{s_D/\sqrt{n}} = 1,002483$

Kết luận: do $-z_{\alpha/2} < z < z_{\alpha/2}$ nên chấp nhận $H_0 : \mu_D = 0$. Vậy không có sự khác nhau về điểm số trung bình của sinh viên về môn đại số và giải tích.

9.3.2 Khoảng tin cậy của cặp.

Cùng cách mà khoảng tin cậy t cho giá trị trung bình tổng thể đơn μ được dựa trên biến t là $T = (\overline{X} - \mu)/(S/\sqrt{n})$, khoảng tin cậy t cho $\mu_D(= \mu_1 - \mu_2)$ dựa trên thực tế là

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

có phân phối t với bậc tự do $n - 1$. Biến đổi biến t như các phần trước đã làm với khoảng tin cậy, thu được khoảng tin cậy $100(1 - \alpha)$ % như sau:

Khoảng tin cậy cặp t cho μ_D là

$$\overline{d} \pm t_{\alpha/2,n-1}.s_D/\sqrt{n}$$

Khoảng tin cậy một phía thu được từ công thức trên bằng cách thay thế $t_{\alpha/2}$ bởi t_α.

Khi n nhỏ, tính hợp lệ của khoảng này yêu cầu phân phối của các hiệu phải xấp xỉ chuẩn. Với n lớn, định lý Giới hạn Trung tâm đảm bảo cho kết quả của khoảng z là hợp lệ mà không có ràng buộc về phân phối của hiệu.

Ví dụ 9.6 Quan sát một mẫu gồm $n = 10$ điện thoại iphone 7 về thời lượng pin khi chỉ dùng nghe gọi (đặt là thời gian sử dụng 1) và khi có sử dụng các ứng dụng (nghe nhạc, chơi game, mạng xã hội,...) (đặt là thời gian sử dụng 2).

Đặt μ_D là hiệu trung bình thật sự của thời gian sử dụng 1 và thời gian sử dụng 2 (giờ). Sử dụng khoảng tin cậy t cặp để ước lượng khoảng tin cậy 99 % cho μ_D với

Điện thoại iphone 7	1	2	3	4	5	6	7	8	9	10
Thời gian sử dụng 1	14.3	13	12	13.8	13.7	14.1	14.7	14.3	14.3	14.4
Thời gian sử dụng 2	4.4	3.2	5	7.7	4.4	6	5.9	5.4	5.3	9
Hiệu	9.9	9.8	7	6.1	9.3	8.1	8.8	8.9	9	5.4

phân phối hiệu này xấp xỉ phân phối chuẩn.

Giải

$n = 10; s = 1,5578; \overline{d} = 8,23$

Với $\gamma = 0,99$ nên $\alpha = 0,01$thì $t_{\alpha/2,n-1} = t_{0,005;9} = 3,25$

Sai số ước lượng: $\varepsilon = t_{\alpha/2,n-1}.s_D/\sqrt{n} = 1,60101.$

Vậy khoảng ước lượng cho hiệu trung bình của thời gian sử dụng điện thoại khi chỉ nghe gọi với khi có xài các ứng dụng khác là:

$$(\overline{d} - \varepsilon; \overline{d} + \varepsilon) = (6,62899; 9,83101).$$

9.3.3 Dữ liệu cặp và quy trình kiểm định t cho hai mẫu

Xét kiểm định t cho hai mẫu trên dữ liệu cặp. Các tử số của hai kiểm định thống kê là đồng nhất, vì $\overline{d} = \Sigma d_i/n = [\Sigma(x_i - y_i)]/n = (\Sigma x_i)/n - (\Sigma y_i)/n = \overline{x} - \overline{y}$. Sự khác nhau giữa số liệu thống kê là hoàn toàn do mẫu số. Mỗi kiểm định thống kê thu được bằng cách chuẩn hóa $\overline{X} - \overline{Y}(= \overline{D})$. Nhưng với sự hiện diện của sự phụ thuộc, việc chuẩn hóa hai mẫu t là không chính xác. Để thấy điều này, nhớ lại Chương 5 rằng

$$V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X, Y)$$

Tương quan giữa X và Y là:

$$\rho = Corr(X, Y) = Cov(X, Y)/[\sqrt{V(X)}.\sqrt{V(Y)}]$$

Theo đó:

$$V(X - Y) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

Áp dụng vào $\overline{X} - \overline{Y}$ nhận được

$$V(\overline{X} - \overline{Y}) = V(\overline{D}) = V(\frac{1}{n}\Sigma D_i) = \frac{V(D_i)}{n} = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{n}$$

Kiểm định t cho hai mẫu dựa trên giả định về tính độc lập, trong trường hợp $\rho = 0$. Tuy nhiên, trong nhiều thí nghiệm cặp, sẽ có sự phụ thuộc dương mạnh mẽ giữa X và Y (X lớn liên quan đến Y lớn), do đó ρ sẽ dương và phương sai của $\overline{X} - \overline{Y}$

sẽ nhỏ hơn $\sigma_1^2/n + \sigma_2^2/n$. Do vậy, *bất cứ khi nào có sự phụ thuộc dương trong cặp, mẫu số cho thống kê t cặp phải nhỏ hơn so với t của kiểm định mẫu độc lập.* Thông thường t hai mẫu sẽ gần với số không hơn so với t cặp, giảm bớt đáng kể ý nghĩa của dữ liệu.

Tương tự như vậy, khi dữ liệu dạng cặp, khoảng tin cậy t cặp thường sẽ hẹp hơn khoảng tin cậy (không chính xác) của t cho hai mẫu. Điều này là vì thông thường sự biến thiên trong giá trị hiệu ít hơn so với biến thiên của giá trị x và y.

9.4 Các kết luận liên quan đến hiệu hai tỷ lệ

Sau khi trình bày các phương pháp so sánh trung bình của hai tổng thể khác nhau, chúng ta chuyển sự chú ý đến việc so sánh tỉ lệ của tổng thể. Cho rằng một cá nhân hoặc đối tượng được đánh giá là thành công "S" nếu anh/chị/đối tượng có được một số đặc trưng (ví dụ: người đã tốt nghiệp đại học, sở hữu chiếc tủ lạnh có chức năng tạo đá viên,...). Đặt

p_1= thành phần đạt tiêu chí S trong tổng thể 1
p_2= thành phần đạt tiêu chí S trong tổng thể 2

Ngoài ra, $p_1(p_2)$ có thể được coi là xác suất một cá nhân (hoặc đối tượng) ngẫu nhiên được chọn từ tổng thể đầu tiên (thứ hai) và thành công.

Giả sử một mẫu có kích thước m được chọn từ tổng thể thứ nhất và một mẫu độc lập kích thước n được chọn từ tổng thể thứ hai. Đặt X biểu thị số S trong mẫu đầu tiên và Y là số S trong mẫu thứ hai. Sự độc lập của hai mẫu suy ra rằng X và Y là độc lập. Nếu hai kích cỡ mẫu nhỏ hơn nhiều so với kích cỡ tổng thể tương ứng, X và Y có thể được coi là có phân phối nhị thức. Ước lượng tự nhiên cho $p_1 - p_2$ (hiệu về tỉ lệ tổng thể), là tương ứng với hiệu tỷ lệ mẫu $X/m - Y/n$.

Mệnh đề 9.5. *Đặt $\hat{p_1} = X/m$ và $\hat{p_2} = Y/n$, trong đó $X \sim Bin(m, p_1), Y \sim Bin(n, p_2)$, với X và Y là các biến độc lập. Vậy*

$$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2$$

vì $\hat{p_1} - \hat{p_2}$ là ước lượng không chệch của $p_1 - p_2$, và

$$V(\hat{p_1} - \hat{p_2}) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \tag{9.3}$$

với $q_i = 1 - p_i$

Đầu tiên chúng ta sẽ tập trung vào các tình huống trong đó cả m và n đều lớn. Sau đó vì mỗi $\hat{p_1}$ và $\hat{p_2}$ đều có phân phối xấp xỉ chuẩn, ước lượng của $\hat{p_1} - \hat{p_2}$ cũng có phân phối xấp xỉ chuẩn. Chuẩn hóa $\hat{p_1} - \hat{p_2}$ nhận được biến Z có phân phối xấp xỉ chuẩn:

$$Z = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}}}$$

9.4.1 Quy trình kiểm định mẫu lớn

Giả thuyết không mà một nhà điều tra hay gặp nhất sẽ có dạng $H_o : p_1 - p_2 = \Delta_0$. Mặc dù với trung bình tổng thể thì trường hợp $\Delta_0 \neq 0$ là không có gì khó khăn, thì với tỉ lệ của tổng thể $\Delta_0 = 0$ và $\Delta_0 \neq 0$ phải được xem xét riêng. Vì đa số các vấn đề thực tế của loại này thường liên quan $\Delta_0 = 0$ (như giả thuyết không $p_1 = p_2$), chúng ta sẽ tập trung vào trường hợp này. Khi $H_o : p_1 - p_2 = 0$ là đúng, đặt p kí hiệu giá trị chung của p_1 và p_2 (và tương tự cho q). Thì biến chuẩn hóa

$$Z = \frac{\hat{p_1} - \hat{p_2} - 0}{\sqrt{pq(\frac{1}{m} + \frac{1}{n})}} \tag{9.4}$$

sẽ có xấp xỉ phân phối chuẩn chuẩn tắc khi H_0 là đúng. Tuy nhiên, Z này không thể dùng như một kiểm định thống kê vì giá trị của p là chưa biết, H_0 chỉ khẳng định ở đây là có một giá trị chung của p, nhưng không nói giá trị đó là bao nhiêu. Một kiểm định thống kê là kết quả việc thay thế p và q trong (9.4) bằng các ước lượng thích hợp.

Giả định rằng $p_1 = p_2 = p$, thay vì các mẫu riêng biệt kích thước m và n từ hai tổng thể khác nhau (hai phân phối nhị thức khác nhau), chúng ta thực sự có một mẫu đơn cỡ $m + n$ từ một quần thể với tỷ lệ p. Tổng số cá thể trong mẫu kết hợp này có đặc trưng là $X + Y$. Ước lượng tự nhiên của p là:

$$\hat{p} = \frac{X + Y}{m + n} = \frac{m}{m + n}.\hat{p_1} + \frac{n}{m + n}.\hat{p_2} \tag{9.5}$$

Biểu thức thứ hai cho $\hat{p}$ thấy rằng nó thực sự là một trung bình trọng số của các ước lượng $\hat{p_1}$ và $\hat{p_2}$ thu được từ hai mẫu. Sử dụng $\hat{p}$ và $\hat{q} = 1 - \hat{p}$ thay cho của p và q trong (9.4) đưa ra một kiểm định thống kê có phân phối chuẩn chuẩn tắc khi H_0 là đúng.

Ví dụ 9.7 Nghiên cứu về tỉ lệ đạt 6.0 Ielts của sinh viên năm tư của một trường đại học có phụ thuộc vào phương pháp học hay không. Quan sát mẫu thứ nhất

Giả thuyết không: $H_0 : p_1 - p_2 = 0$	
Giá trị kiểm định thống kê (mẫu lớn) : $z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{m} + \frac{1}{n})}}$	
Các giả thuyết đối:	Miền bác bỏ với kiểm định xấp xỉ mức α
$H_a : p_1 - p_2 > 0$	$z \geq z_\alpha$
$H_a : p_1 - p_2 < 0$	$z \leq -z_\alpha$
$H_a : p_1 - p_2 \neq 0$	$z \geq z_{\alpha/2}$ hoặc $z \geq -z_{\alpha/2}$
Giá trị P được tính như các kiểm định z trước. Kiểm định có thể dùng được khi mà $m\hat{p_1}, m\hat{q_1}, n\hat{p_2}$ và $n\hat{q_2}$ đều tối thiểu là 10.	

gồm $n = 200$ sinh viên tự học tiếng anh qua các tài liệu và kênh học miễn phí trên Internet, thì có 110 sinh viên đạt 6.0 Ielts. Quan sát mẫu thứ hai gồm $n = 350$ sinh viên theo học một trung tâm tiếng anh thì có 180 sinh viên đạt 6.0 Ielts. Hãy kiểm định các giả thuyết trên với mức ý nghĩa là 0,05.

Giải

Gọi p_1, p_2 lần lượt là tỉ lệ đạt điểm Ielts 6.0 của sinh viên tự học tiếng anh qua các tài liệu, kênh học miễn phí trên Internet; và theo học một trung tâm tiếng anh.

Giả thuyết không: $H_0 : p_1 - p_2 = 0$

Giả thuyết đối: $H_a : p_1 - p_2 \neq 0$

Với $\alpha = 0,05$ thì $\phi(z_{\alpha/2}) = 1 - \alpha/2 = 0,975$ nên $z_{\alpha/2} = 1,96$

Tỷ lệ mẫu 1: $\hat{p_1} = \frac{X}{m} = \frac{110}{200}$;

Tỷ lệ mẫu 2: $\hat{p_2} = \frac{Y}{n} = \frac{180}{350}$

Tỷ lệ mẫu chung $\hat{p} = \frac{X+Y}{m+n} = \frac{110+180}{200+350} = \frac{290}{550}$, $\hat{q} = 1 - \hat{p} = \frac{260}{550}$.

Tiêu chuẩn kiểm định:

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{m} + \frac{1}{n})}} = 0,807$$

Do $-z_{\alpha/2} < z < z_{\alpha/2}$ nên chấp nhận $H_0 : p_1 - p_2 = 0$.

Vậy tỉ lệ đạt 6.0 Ielts của sinh viên năm tư không phụ thuộc vào phương pháp học.

9.4.2 Khoảng tin cậy cho mẫu lớn

Giống như trung bình, nhiều vấn đề hai mẫu liên quan đến vấn đề của so sánh thông qua kiểm định giả thuyết, nhưng đôi khi ước lượng khoảng cho $p_1 - p_2$ là thích hợp. Cả $\hat{p_1} = X/m$ và $\hat{p_2} = Y/n$ đều có phân phối xấp xỉ chuẩn khi m và n lớn. Nếu θ là $p_1 - p_2$, thì $\hat{\theta} = \hat{p_1} - \hat{p_2}$ đáp ứng các điều kiện cần thiết để có được một khoảng tin cậy của mẫu lớn. Cụ thể, ước lượng độ lệch tiêu chuẩn $\hat{\theta}$ là $\sqrt{(\hat{p_1}\hat{q_1}/m) + (\hat{p_2}\hat{q_2}/n)}$. Khoảng tin cậy tổng quát $100(1-\alpha)\%$ là $\hat{\theta} \pm z_{\alpha/2}.\hat{\sigma}_{\hat{\theta}}$ có dạng như sau. Chú ý rằng

Khoảng tin cậy $p_1 - p_2$ với độ tin cậy xấp xỉ $100(1-\alpha)\%$ là
$$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}.\hat{q_1}}{m} + \frac{\hat{p_2}\hat{q_2}}{n}}$$
Khoảng này có thể dùng được khi mà $m\hat{p_1}, m\hat{q_1}, n\hat{p_2}$ và $n\hat{q_2}$ đều tối thiểu là 10.

độ lệch chuẩn được ước lượng của $\hat{p_1} - \hat{p_2}$ (dạng căn thức) thì khác với khi kiểm định giả thuyết khi $\Delta_0 = 0$.

Các nghiên cứu gần đây cho thấy độ tin cậy thực sự của khoảng tin cậy truyền thống đôi khi chệch khỏi mức đa thức (mức mà ta nhận được khi sử dụng giá trị đặc trưng z, ví dụ 95% khi $z_{\alpha/2} = 1,96$). Cải tiến đề xuất là thêm một thành công và một thất bại cho hai mẫu và sau đó thay thế các giá trị mẫu $\hat{p}$ và mẫu $\hat{q}$ trong công thức nói trên bằng mẫu $\tilde{p}$ và mẫu $\tilde{q}$, trong đó $\tilde{p_1} = (x+1)/(m+2)$,v.v... Khoảng hiệu chỉnh này cũng có thể được sử dụng khi kích thước mẫu là khá nhỏ.

Ví dụ 9.8 Nghiên cứu về tỉ lệ đạt 6.0 Ielts của sinh viên năm tư của một trường đại học có phụ thuộc vào phương pháp học hay không. Quan sát mẫu thứ nhất gồm $n = 200$ sinh viên tự học tiếng anh qua các tài liệu và kênh học miễn phí trên Internet, thì có 110 sinh viên đạt 6.0 Ielts. Quan sát mẫu thứ hai gồm $n = 350$ sinh viên theo học một trung tâm tiếng anh thì có 180 sinh viên đạt 6.0 Ielts. Với độ tin cậy 99 % tìm khoảng tin cây cho hiệu giữa hai tỉ lệ trên?

Giải

Gọi p_1 , p_2 lần lượt là tỉ lệ đạt điểm Ielts 6.0 của sinh viên tự học tiếng anh qua các tài liệu, kênh học miễn phí trên Internet; và theo học một trung tâm tiếng anh.

Với $\gamma = 0,99$ thì $\phi(z_{\alpha/2}) = (1+\gamma)/2 = 0,995$ nên $z_{\alpha/2} = 2,58$

$\hat{p_1} = \frac{X}{m} = \frac{110}{200}$; $\hat{p_2} = \frac{Y}{n} = \frac{180}{350}$

Khoảng tin cậy 99 % cho hiệu giữa hai tỉ lệ trên là:

$$\frac{110}{200} - \frac{180}{350} \pm 2,58.\sqrt{\frac{\frac{110}{200}.\frac{90}{200}}{200} + \frac{\frac{180}{350}\frac{180}{350}}{350}} = (-0,0783; 0,6640)$$

9.4.3 Suy luận với mẫu nhỏ

Thỉnh thoảng suy luận liên quan $p_1 - p_2$ có thể phải dựa trên trường hợp có ít nhất một mẫu nhỏ. Các phương pháp thích hợp cho các tình huống như vậy không đơn giản như đối với các mẫu lớn và có nhiều tranh cãi giữa các nhà thống kê về các bước tiến hành. Một kiểm định thường xuyên được sử dụng, được gọi là kiểm định Fisher-Irwin, được dựa trên phân phối siêu bội.

9.5 Các kết luận liên quan đến hai phương sai tổng thể

Đôi khi cần đến phương pháp để so sánh hai phương sai tổng thể (hoặc độ lệch chuẩn), mặc dù các vấn đề này ít gặp hơn so với trung bình hoặc tỷ lệ. Đối với trường hợp trong đó các tổng thể đang được điều tra là chuẩn, các bước dựa trên một họ mới của phân phối xác suất mới.

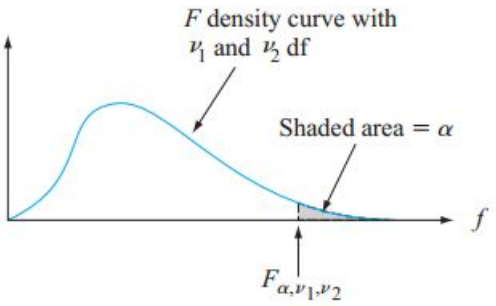
9.5.1 Phân phối F

Phân phối xác suất F có hai tham số, ký hiệu là ν_1 và ν_2. Tham số ν_1 được gọi là *số bậc tự do của tử* , và ν_2 là *số bậc tự do của mẫu* ; ở đây ν_1 và ν_2 là số nguyên dương. Một biến ngẫu nhiên có phân phối F không thể giả sử là một giá trị âm. Vì hàm mật độ là phức tạp và sẽ không được sử dụng rõ ràng, chúng ta bỏ qua công thức. Có một mối quan hệ quan trọng giữa một biến F và biến chi bình phương. Nếu X_1 và X_2 là các biến ngẫu nhiên có phân phối chi bình phương độc lập với bậc tự do ν_1 và ν_2 tương ứng, thì biến ngẫu nhiên

$$F = \frac{X_1/v_1}{X_2/v_2}$$

(tỉ số của 2 biến chi bình phương chia cho bậc tự do tương ứng), có thể được cho thấy để có phân phối F.

Hình dưới minh họa đồ thị của một hàm mật độ F điển hình. Tương tự với ký hiệu $t_{\alpha,v}$ và $\chi^2_{\alpha,v}$, chúng ta sử dụng F_{α,v_1,v_2} cho giá trị trên trục hoành mà α nằm trong vùng diện tích dưới đường cong mật độ F với bậc tự do ν_1 và ν_2 ở miền bên phải. Đường cong mật độ không đối xứng, vì vậy có vẻ như các giá trị tới hạn của phía trái và phải đều phải được lập bảng. Tuy nhiên điều này không cần thiết vì thực tế là $F_{1-\alpha,v_1,v_2} = 1/F_{\alpha,v_2,v_1}$.



Giả thuyết không: $H_0 : \sigma_1^2 = \sigma_2^2$	
Giá trị kiểm định thống kê: $f = s_2^1/s_2^2$	
Các giả thuyết đối:	Miền bác bỏ cho a với mức ý nghĩa α
$H_a : \sigma_1^2 > \sigma_2^2$	$f \geq F_{\alpha,m-1,n-1}$
$H_a : \sigma_1^2 < \sigma_2^2$	$f \leq F_{\alpha,m-1,n-1}$
$H_a : \sigma_1^2 \neq \sigma_2^2$	$f \geq F_{\alpha/2,m-1,n-1}$ hay $f \leq F_{\alpha/2,m-1,n-1}$

Bảng phụ lục cho F_{α,v_1,v_2}, và $\alpha = 0.10, 0.05, 0.01$ và 0.001, và các giá trị khác của ν_1 (trong các cột khác nhau của bảng) và ν_2 (trong các nhóm hàng khác nhau của bảng). Ví dụ, $F_{0.05,6,10} = 3.22$ và $F_{0.05,10,6} = 4.06$. Giá trị tới hạn $F_{0.95,6,10}$ mà chiếm 0.95 diện tích miền phải nó (và như vậy 0.05 ở bên trái) dưới đường cong F với $v_1 = 6$ và $v_2 = 10$, là $F_{0.95,6,10} = 1/F_{0.05,6,10} = 1/4.06 = 0.246$.

9.5.2 Kiểm định F cho các phương sai bằng nhau

Các bước kiểm định cho giả thuyết liên quan tỉ lệ σ_1^2/σ_2^2 dựa trên kết quả sau:

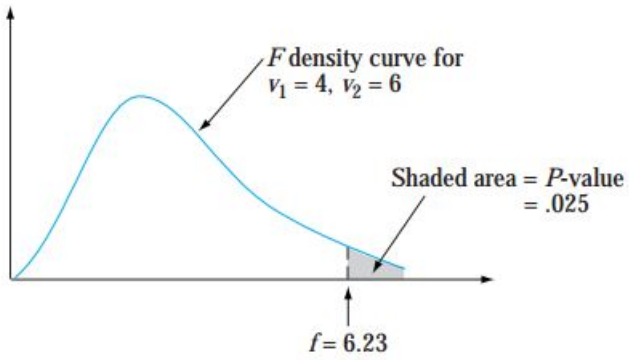
Định lý 9.6. *Đặt $X_1, \ldots, X_m$ là một mẫu ngẫu nhiên có phân phối chuẩn với phương sai σ_1^2, đặt $Y_1, \ldots, Y_n$ là một mẫu ngẫu nhiên khác (độc lập với mẫu X_i) có phân phối chuẩn với phương sai σ_2^2 , và đặt S_1^2 và S_2^2 biểu thị hai phương sai mẫu. Thì biến ngẫu nhiên*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

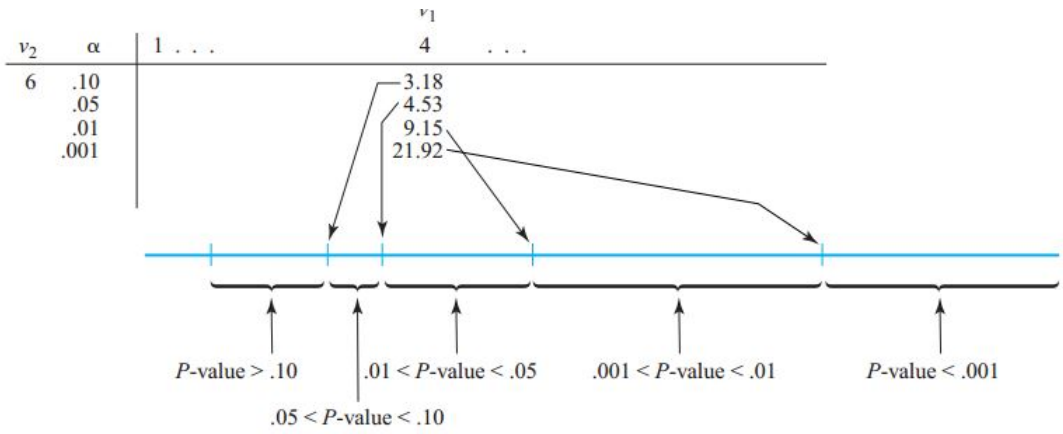
có phân phối F với $v_1 = m - 1$ và $v_2 = n - 1$.

Khi giá trị tới hạn được cho trong bảng chỉ cho $\alpha = 0.10, 0.05, 0.01$ và 0.001, kiểm định 2 phía chỉ có thể thực hiện ở mức $0.20, 0.10, 0.02, 0.002$. Các giá trị tới hạn F khác có thể thu được bằng phần mềm thống kê.

Nhắc lại về $P - value$ cho một kiểm định t bên phải trên là phần diện tích dưới đường cong t tương ứng (phần với bậc tự do phù hợp) đến bên phải của giá trị t được tính toán. Cùng một cách, giá trị P cho một kiểm định F miền phải là phần diện tích dưới đường cong F với bậc tự do của tử số và mẫu số thích hợp khi kiểm định bên phải của f đã được tính. Hình dưới minh họa điều này với một bài kiểm định dựa trên $v_1 = 4$ và $v_2 = 6$.



Lập bảng của phần diện tích miền phải đường cong F phức tạp hơn nhiều so với đường cong t bởi vì liên quan đến hai bậc tự do. Đối với mỗi sự kết hợp của v_1 và v_2, bảng F của chúng ta chỉ đưa ra bốn giá trị tới hạn mà diện tích là $0.10, 0.05, 0.01,$ và 0.001. Hình dưới cho thấy những gì có thể nói về P-value phụ thuộc vào nơi f rơi vào tương ứng với bốn giá trị tới hạn.



Ví dụ, kiểm định với $v_1 = 4$ và $v_2 = 6$,

$f = 5.70 \Rightarrow 0.01 <$P-value$< 0.05$

$f = 2.16 \Rightarrow$P-value> 0.10

$f = 25.03 \Rightarrow$P-value< 0.001

Chỉ khi f bằng một giá trị đã lập bảng thì chúng ta mới có được một giá trị P chính xác (ví dụ, nếu $f = 4.53$, thì $P - value = 0.05$). Một khi chúng ta biết rằng $0.01 < P - value < 0.05$, H_0 sẽ bị loại bỏ ở mức ý nghĩa 0.05 nhưng không ở mức 0.01. Khi $P - value < 0.001$, H_0 bị từ chối ở bất kỳ mức ý nghĩa hợp lý nào.

9.5.3 Khoảng tin cậy cho σ_1/σ_2

Khoảng tin cậy cho σ_1^2/σ_2^2 là dựa trên thay thế F trong biểu thức xác suất

$$P(F_{1-\alpha/2,v_1,v_2} < F < F_{\alpha/2,v_1,v_2}) = 1 - \alpha$$

bằng biến F ở công thức trên và biến đổi các bất đẳng thức để tách riêng σ_1^2/σ_2^2 . Một khoảng cho σ_1/σ_2 là kết quả từ lấy căn bậc hai của mỗi giới hạn.

9.6 Bài tập

9.6.1 Các kết luận trên hai mẫu độc lập

Bài 1 Báo cáo thống kê sức khỏe quốc gia A, bao gồm các thông tin sau đây về chiều cao (in.) cho hai nhóm phụ nữ da trắng không phải gốc Tây Ban Nha:

Nhóm tuổi	Cỡ mẫu	Trung bình mẫu	Độ lệch chuẩn mẫu
20-39	866	64,9	0,59
Từ 60 trở lên	934	63,1	0,79

1. Tìm khoảng tin cậy với độ tin cậy 95% cho hiệu giữa chiều cao trung bình tổng thể của phụ nữ trẻ và phụ nữ lớn tuổi.

2. Cho μ_1 là chiều cao trung bình tổng thể những phụ nữ độ tuổi 20–39 và μ_2 là chiều cao tổng thể những người độ tuổi từ 60 tuổi trở lên. Sử dụng miền bác bỏ giả thuyết hãy kiểm định giả thiết $H_0 : \mu_1 - \mu_2 = 1; H_1 : \mu_1 - \mu_2 > 1$ với mức ý nghĩa 0,001.

3. Giá trị P-value cho kiểm định thực hiện trong (2.) là bao nhiêu? Dựa trên giá trị P-value này, ta bác bỏ giả thuyết không ở mức ý nghĩa nào là hợp lý? Giải thích.

Bài 2 Kí hiệu μ_1 là tuổi thọ trung bình thực sự cho lốp của một thương hiệu R, và μ_2 kí hiệu tuổi thọ trung bình thực sự cho tuổi lốp của một thương hiệu khác có cùng kích thước. Kiểm định $H_0 : \mu_1 - \mu_2 = 5000; H_1 : \mu_1 - \mu_2 > 5000$ với mức ý nghĩa 0,01. Cho biết $m = 45; \overline{x} = 42500; s_1 = 2200; n = 45; \overline{y} = 36800; s_2 = 1500$.

Bài 3 Những người có hội chứng Reynaud có khuynh hướng bị suy giảm đột ngột tuần hoàn máu ở ngón tay và ngón chân. Trong một thí nghiệm để nghiên cứu mức độ suy giảm này, mỗi đối tượng nhúng một ngón trỏ trong nước rồi đo sản nhiệt (cal/cm2/min). Đối với $m = 10$ đối tượng có hội chứng, sản lượng nhiệt trung bình

là $\overline{x} = 0,64$, và cho $n = 10$ đối tượng không có hội chứng, sản lượng trung bình là 2,05. Gọi $\mu_1; \mu_2$ kí hiệu cho sản nhiệt trung bình thực sự cho hai đối tượng trên. Giả sử rằng phân phối sản nhiệt của hai loại đối tượng trên có phân phối chuẩn $\sigma_1 = 0,2; \sigma_2 = 0,4$.

1. Tìm khoảng tin cậy 98% cho hiệu $\mu_1 - \mu_2$.

2. Kiểm định giả thiết $H_0 : \mu_1 - \mu_2 = -1; H_1 : \mu_1 - \mu_2 < -1$ với mức ý nghĩa 0,1 theo cả hai cách miền bác bỏ và P giá trị.

Bài 4 Trong bài báo "Có sự khác biệt về giảng dạy toàn thời gian và bán thời gian?" (*College Teaching, 2009: 23–26*) báo cáo rằng cho một mẫu 125 các khóa học do giảng viên dạy toàn thời gian, điểm GPA trung bình là 2,7186 và độ lệch chuẩn là 0,63342, trong khi đối với mẫu của 88 khóa học được giảng dạy bán thời gian, trung bình và độ lệch chuẩn tương ứng là 2,8639 và 0,49241. Liệu có sự khác biệt điểm GPA trung bình thực sự giữa dạy bán thời gian với dạy toàn thời gian không? Hãy kiểm định với mức ý nghĩa 0,01.

Bài 5 Người quản lý công ty quan sát 75 buổi sáng đếm số sản phẩm sản xuất được trong mỗi buổi và tính được trung bình mẫu là 806 (sản phẩm/buổi) và độ lệch chuẩn mẫu là 185. Quan sát 100 buổi chiều và tính được trung bình mẫu là 723 (sản phẩm/ buổi) và độ lệch chuẩn mẫu là 164.

1. Có ý kiến cho rằng làm việc buổi sáng hiệu quả hơn buổi chiều. Hãy cho nhận xét với mức ý nghĩa 1 % ?

2. Hãy tìm khoảng ước lượng cho lượng sản phẩm chênh lệch của hai buổi với độ tin cậy 96 %?

Bài 6 Khi dân số già đi, có sự lo ngại ngày càng tăng về tai nạn chấn thương liên quan đến người già. Bài báo "Sự khác biệt về tuổi tác và giới tính trong việc lấy lại thăng bằng bằng 1 bước để khỏi bị ngã về phía trước" (J. của Gerontology, 1999: M44 – M50) được báo cáo trong một thí nghiệm trong đó góc nghiêng tối đa - xa nhất mà đối tượng có thể nghiêng và vẫn giữ lại thăng bằng bằng 1 bước được xác định từ mẫu phụ nữ trẻ 21-29 tuổi (young females) và mẫu phụ nữ lớn tuổi 67-81 tuổi (older females).

YF: 29, 34, 33, 27, 28, 32, 31, 34, 32, 27
OF: 18, 15, 23, 13, 12

Dữ liệu có cho thấy góc nghiêng tối đa trung bình thực sự của phụ nữ lớn tuổi thì lớn hơn phụ nữ nhỏ tuổi là 10 độ không? Hãy kiểm định làm rõ điều trên với mức ý nghĩa 0,1.

Bài 7 Bệnh thoái hóa khớp thường xuyên ảnh hưởng đến các khớp có trọng lượng như đầu gối. Bài viết "Dữ kiện về tái phân phối tải cơ học ở khớp đầu gối trong người cao tuổi khi tăng dần cầu thang và dốc " (Annals of Biomed. Engr., 2008: 467–476) trình bày dữ liệu tóm tắt sau về thời gian đứng (phút) cho các mẫu của cả người lớn tuổi và người trẻ tuổi.

Age	Sample Size	Sample Mean	Sample SD
Older	28	801	117
Younger	16	780	72

Giả sử rằng cả hai nhóm tuổi già và tuổi trẻ đều có phân phối chuẩn.

1. Tính khoảng ước lượng cho thời gian đứng trung bình của nhóm người già với độ tin cậy 99%.

2. Thực hiện kiểm định giả thuyết ở mức ý nghĩa 0,05 để quyết định xem thời gian đứng trung bình thực sự của người già có lớn hơn thời gian đứng trung bình thực sự của người trẻ không?

Bài 8 Để nghiên cứu tác dụng của việc bón phân đạm theo công thức A đối với sản lượng của bắp, người ta làm thí nghiệm trên các mảnh đất. Quan sát sản lượng thu được trên các mảnh đối chứng (không bón đạm) và các mảnh có bón phân đạm theo công thức A được bảng sau:

Sản lượng (tạ/ha)	55	53	30	37	49
Mảnh đối chứng	4	5	6	3	5

Sản lượng (tạ/ha)	60	58	29	39	47	35
Mảnh bón phân	4	7	3	5	8	3

Giả sử sản lượng của hai loại mảnh đất có phân phối chuẩn. Hãy cho kết luận về hiệu quả của việc bón phân đạm theo công thức A, với mức ý nghĩa 0,05.

Bài 9 Kiểm tra chất lượng nón bảo hiểm do 2 nhà máy A, B sản xuất được kết quả sau: trong số 500 nón bảo hiểm của nhà máy A, có 95 nón không đạt tiêu chuẩn.

Trong số 400 nón của nhà máy B có 95 nón không đạt tiêu chuẩn. Với mức ý nghĩa 3% hãy cho kết luận về chất lượng nón bảo hiểm của 2 nhà máy A, B.

Bài 10 Trong 500 sv nam có 45 sv đạt loại giỏi. Trong 400 sv nữ có 50 sv đạt loại giỏi. Với mức ý nghĩa 2 %, kết luận tỉ lệ giỏi của nam cao hơn nữ được không?

Bài 11 Cho một mẫu A có 28 phần tử với độ lệch chuẩn mẫu là 52,6. Một mẫu B có 26 phần tử, độ lệch chuẩn mẫu là 84,2. So sánh độ lệch chuẩn tổng thể của hai mẫu trên với mức ý nghĩa 2%?

9.6.2 Kết luận trên mẫu ghép cặp

Bài 12 Theo dõi thu nhập, chi tiêu (triệu đồng/tháng) của một số hộ gia đình trong vùng A có số liệu.

Thu nhập	15	18	19	21	23	27	29	19	17	24	22	28	35	38	40
Chi tiêu	12	15	15	17	21	25	22	18	17	21	18	21	30	25	26
Số dư	3	3	4	4	2	2	7	1	0	3	4	7	5	13	14

Giả sử thu nhập, chi tiêu trong 1 tháng của mỗi hộ gia đình là các biến ngẫu nhiên có phân phối chuẩn.

1. Tìm khoảng tin cậy 95 % cho số tiền dư trung bình trong một tháng của mỗi hộ gia đình.

2. Có ý kiến cho rằng số tiền dư trung bình của mỗi hộ trong 1 tháng là 4 triệu đồng. Hãy cho nhận xét về ý kiến này với mức ý nghĩa 5%.

9.6.3 Bài tập tổng hợp

Bài 13 Khảo sát chiều cao của sinh viên trường A, B ta có:

Chiều cao (m)	1,5-1,55	1,55-1,6	1,6-1,65	1,65-1,7	1,7-1,75	1,75-1,8	1,8-1,85	1,85-1,9
Số SV (A)	15	38	56	68	70	56	31	12
Số SV (B)	21	43	60	78	71	58	29	10

Có ý kiến cho rằng chiều cao trung bình của sinh viên trường A cao hơn trường B. Hãy cho nhận xét với mức ý nghĩa 5

Bài 14 Nghiên cứu khả năng chống cảm cúm của Vitamin C, có kết quả sau. Trong số 420 người không uống Vitamin C, có 93 người bị cảm cúm. Trong số 417 người, mỗi ngày uống 1g Vitamin C/mỗi người, có 51 người bị cảm cúm. Với mức ý nghĩa 1 % có thể cho rằng Vitamin C có khả năng chống cảm cúm hay không?

Bài 15 Giả thuyết rằng thời gian sử dụng điện thoại loại A, B có phân phối chuẩn. Quan sát thời gian sử dụng một số điện thoại A, B ta có số liệu:

Thời gian (h)	5-6	6-7	7-8	8-9	9-10	10-11	11-12
Số điện thoại A	1	2	5	7	6	3	1
Số điện thoại B	1	3	4	5	3	2	1

Hãy so sánh thời gian sử dụng trung bình của 2 loại điện thoại này với mức ý nghĩa 5

Bài 16 Mức tiêu thụ X, Y của mỗi hộ gia đình vùng A, B trong một tháng mùa khô trong 1 năm xác định có phân phối chuẩn. Diều tra ngẫu nhiên một số hộ gia đình ở vùng A, B ta có số liệu sau:

(kwh/tháng)	65-115	115-165	165-215	215-265	265-315	315-365	365-415	415-465
Số hộ vùng A	14	36	74	98	112	91	65	36
Số hộ vùng B	28	53	92	121	130	111	78	43

1. So sánh mức tiêu thu điện trung bình trong một tháng mùa khô của mỗi hộ gia đình vùng A và vùng B trong năm đang xét với mức ý nghĩa 2%.

2. Những hộ gia đình có mức tiêu thụ từ 315kwh/tháng là những hộ gia đình có mức tiêu thụ cao. Hãy so sánh tỷ lệ hộ có mức tiêu thu điện cao trong một tháng mùa khô ở hai vùng A, B với mức ý nghĩa 3%.

3. Với mức ý nghĩa 5% có thể cho rằng tỷ hộ có mức tiêu thụ điện cao bằng tỷ hộ có mức tiêu thụ điện không cao ở vùng A trong 1 tháng mùa khô.

4. Hãy ước lượng số hộ có mức tiêu thụ điện cao ở vùng B trong 1 tháng mùa khô với độ tin cậy 95%, biết vùng này có 3000 hộ.

5. Nếu muốn ước lượng mức tiêu thụ điện trung bình các hộ vùng A trong một tháng mùa khô năm xác định với độ chính xác 10 kwh/tháng thì độ tin cậy bằng bao nhiêu?

Bài 17 Để so sánh năng suất của hai giống lúa A, B người ta trồng thực nghiệm hai giống lúa trên các cặp mảnh ruộng, mỗi cặp mảnh ruộng có điều kiện chất đất, tưới tiêu, chăm sóc là giống nhau. Với đơn vị (tạ/ha), số liệu thu được như sau

Giống lúa A	55,0	51,3	52,8	55,2	57,3	58,2	59,1
Giống lúa B	56,0	51,8	53,1	55,4	58,1	58,1	59,3

Giống lúa A	55,9	51,3	52,8	55,2	57,3	58,2	59,1
Giống lúa B	56,0	51,8	53,1	55,4	58,1	58,1	59,3
Giống lúa A	54,5	52,3	52,6	56,2	57,4	58,3	59,2
Giống lúa B	54,6	52,8	53,1	56,4	58,0	58,3	59,3
Giống lúa A	53,5	53,3	53,8	56,2	59,3	58,2	59,1
Giống lúa B	53,6	53,8	54,1	56,4	59,4	58,4	59,5
Giống lúa A	52,5	54,3	55,8	57,2	55,3	53,2	53,1
Giống lúa B	52,6	54,8	56,1	57,4	55,1	53,2	53,3

Hãy so sánh năng suất của hai giống lúa này với mức ý nghĩa 5%.

Bài 18 Quan sát lượng điện tiêu thụ trong 1 tháng (kwh) của một số hộ gia đình ở vùng A trước và sau khi sử dụng bóng điện tiết kiệm điện ta có số liệu

Trước	29,5	42,0	45,5	56,5	28,0	43,5	57,0	65,0
Sau	28,0	41,5	43,0	52,0	27,5	42,0	54,5	63,0
Số hộ	15	12	24	15	17	23	14	15
Trước	28,0	45,0	45,0	56,0	28,5	44,0	55,0	66,5
Sau	26,5	43,0	42,5	52,5	27,0	42,0	54,5	63,0
Số hộ	12	23	14	15	14	23	24	13
Trước	24,5	41,5	35,0	36,5	38,0	33,0	35,0	36,0
Sau	22,0	41,0	33,0	32,0	36,0	38,0	32,5	32,0
Số hộ	11	12	15	14	12	13	15	9

Hãy đưa ra kết luận về hiệu quả của việc sử dụng bóng đèn tiết kiệm với mức ý nghĩa 3%.

Bài 19 Khảo sát chiều cao (đơn vị: cm) của một số học sinh tiểu học lớp 4 sau 3 tháng nghỉ hè ở vùng A ta thu được số liệu.

Trước hè	132,1	135,2	128,5	122,8	128,5	138,1	140,2	136,3
Sau hè	132,8	135,2	128,9	123,5	129,3	138,9	140,5	136,5
Trước hè	133,1	125,2	138,5	132,8	128,4	138,2	141,2	131,3
Sau hè	133,9	125,2	138,8	133,4	129,2	138,9	142,2	131,6
Trước hè	122,2	145,2	128,5	122,5	138,5	138,1	140,2	126,2
Sau hè	122,9	145,4	129,0	123,2	139,2	138,6	141,1	126,5

Hãy cho nhận xét về ý kiến sau 3 tháng nghỉ hè học sinh tiểu học lớp 4 tăng ít nhất

0,3 cm với mức ý nghĩa 5%.

9.6.4 Phân tích số liệu ghép đôi

Bài 14 Xem xét các dữ liệu về tải trọng phá vỡ (kg/25mm chiều rộng) cho các loại vải khác nhau trong cả điều kiện không mài mòn và điều kiện bị mài mòn ("Hiệu ứng của mài mòn ướt trên các đặc tính kéo của Cotton và Polyester-Cotton"- Tạp chí Kiểm tra và đánh giá, 1993: 84–93). Sử dụng kiểm định t cặp, như tác giả trong bài báo đã đề cập, để kiểm định $H_0 : \mu_D = 0; H_a : \mu_D > 0$ với mức ý nghĩa 0,01.

	Fabric							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
U	36.4	55.0	51.5	38.7	43.2	48.8	25.6	49.8
A	28.5	20.0	46.0	34.5	36.5	52.5	26.5	46.5

Chương 10

HỒI QUY TUYẾN TÍNH ĐƠN GIẢN VÀ TƯƠNG QUAN

Trong bài toán hai mẫu đã nghiên cứu trong chương 9 ta đi so sánh giá trị các tham số của hai phân phối x và y. Mặc dù mẫu đã cho ghép cặp trong chương 9 ta đã không sử dụng thông tin của một biến để nghiên cứu về biến còn lại. Dây chính xác là vấn đề mà hồi quy nghiên cứu: đó là khám phá ra mối quan hệ giữa hai hay nhiều biến, từ đó ta có thể thu được thông tin của một biến khi biết các biến còn lại.

Khi nói về x và y có mối quan hệ với nhau ta thường nghĩ đến mối quan hệ hàm $y = f(x)$. Ví dụ x là số sản phẩm A của cửa hàng bán lẻ B bán được trong 1 ngày. Mỗi sản phẩm A bán được cửa hàng lời 25 ngàn đồng, y số tiền lời của cửa hàng B thu được khi bán sản phẩm A trong 1 ngày. Khi đó $y = 25x$. Nếu một ngày cửa hàng B bán được $x = 8$ sản phẩm A thì cửa hàng thu được số tiền lời tương ứng là $y = 8.25 = 200$ ngàn đồng.

Tuy nhiên, trong thực tế có nhiều biến x, y có mối quan hệ với nhau nhưng x và y không có quan hệ hàm số $y = f(x)$. Ví dụ như điểm thi đại học môn Toán x và điểm môn Toán trung bình khi học đại học y của sinh viên trường đại học M có mối quan hệ với nhau. Tuy nhiên, trường hợp hai sinh viên có điểm Toán thi đại học giống nhau có điểm trung bình môn Toán khi học tại trường M là khác nhau thì có thể xảy ra.

Phân tích hồi quy là một phần của thống kê nghiên cứu mối quan hệ giữa hai hay nhiều biến không theo kiểu quan hệ hàm số. Trong chương 12, ta khái quát quan hệ hàm tuyến tính $y = \beta_0 + \beta_1 x$ cho một quan hệ xác suất tuyến tính và rút ra các kết luận cho mô hình. Trong chương 13, ta sẽ xét về kỹ thuật cho các mô

hình cụ thể và nghiên cứu các mối quan hệ phi tuyến tính cũng như mối quan hệ giữa nhiều hơn hai biến.

10.1 Mô hình hồi quy tuyến tính đơn giản

Quan hệ tuyến tính $y = \beta_0 + \beta_1 x$ giữa hai biến x và y là mối quan hệ toán học đơn giản giữa hai biến x và y. Tập hợp các điểm x và y xác định bởi $y = \beta_0 + \beta_1 x$ thuộc đường thẳng có hệ số dốc β_1 và hệ số tự do β_0 (tức là đường thẳng $y = \beta_0 + \beta_1 x$ cắt trục tung Oy tại điểm có tung độ bằng β_0).

Nếu hai biến x, y không có mối quan hệ hàm với nhau thì với giá trị x cố định ta không xác định được giá trị chắc chắn tương ứng của y. Ví dụ như quan sát mối quan hệ về số lượng từ vựng của trẻ em và độ tuổi của chúng thì tương ứng với một trường hợp cụ thể của tuổi trẻ em như $x = 5$ tuổi thì số lượng từ vựng của trẻ 5 tuổi là biến ngẫu nhiên y. Quan sát cụ thể đếm số lượng từ của trẻ A có tuổi bằng 5 thấy số lượng từ của em A khoảng 2000 từ. Khi đó ta nói một giá trị quan sát được của Y tương ứng $x = 5$ là $y = 2000$. Tổng quát, các biến có giá trị được cố định bởi người quan sát sẽ ký hiệu là x và gọi là **biến độc lập**, **biến để dự đoán** hay **biến giải thích** với x cố định, biến thứ hai sẽ là ngẫu nhiên. Ta thường ký hiệu biến ngẫu nhiên là Y và giá trị của nó là y và gọi biến này là **biến phụ thuộc** hay **biến đáp ứng, biến mong đợi**.

Thông thường các quan sát biến độc lập sẽ được thực hiện. Kí hiệu $x_1, x_2, \ldots, x_n$ là các giá trị quan sát được của biến độc lập; Y_i và y_i tương ứng là biến ngẫu nhiên và giá trị biến ngẫu nhiên tương ứng với giá trị quan sát x_i. Dữ liệu tồn tại dưới dạng ghép cặp $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Biểu diễn của dữ liệu này được gọi là một **biểu đồ chấm** sẽ cho những cảm nhận sơ bộ về mối quan hệ tự nhiên nào đó. Trong biểu đồ này, mỗi (x_i, y_i) được biểu diễn bởi một chấm trong hệ tọa độ hai chiều.

Mô hình tuyến tính xác suất

Đối với mô hình $y = \beta_0 + \beta_1 x$, giá trị quan sát của y là một hàm tuyến tính đối với x. Tổng quát hóa xấp xỉ này ta giả sử rằng giá trị mong đợi của Y là hàm tuyến tính của x cố định.

Định nghĩa 10.1. **Mô hình tuyến tính đơn giản**

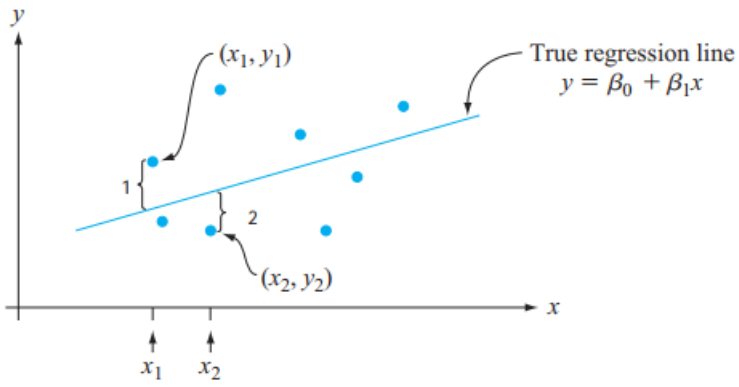
Các định tham số β_0, β_1 và σ^2 thỏa mãn với giá trị cố định x bất kỳ của biến

độc lập, biến phụ thuộc là biến ngẫu nhiên quan hệ với x qua phương trình

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{10.1}$$

Đại lượng ε trong mô hình là một biến ngẫu nhiên, được giả sử là có phân phối chuẩn với $E(\varepsilon) = 0$ và $V(\varepsilon) = \sigma^2$.

Biến ε thường được gọi **độ lệch ngẫu nhiên** hay **sai số ngẫu nhiên** trong mô hình. Nếu không có ε, cặp quan sát (x, y) bất kỳ sẽ tương ứng với một điểm chính thuộc đường $y = \beta_0 + \beta_1 x$, gọi là **đường hồi quy đúng (hay tổng thể)**. Suy luận về sai số ngẫu nhiên cho phép (x, y) nằm phía trên đường hồi quy đúng (khi $\epsilon > 0$) hay nằm dưới đường hồi quy đúng (khi $\varepsilon < 0$). Các điểm $(x_1, y_1), \ldots, (x_n, y_n)$ thu được từ n quan sát được biểu diễn thành các chấm về đường hồi quy đúng. Trong trường hợp các chấm xấp xỉ tuyến tính mô hình hồi quy tuyến tính đơn giản có thể được xem xét có dùng để biểu diễn cho số liệu này.



Hình 10.1: Các điểm tương ứng với các quan sát từ mô hình hồi quy tuyến tính đơn giản.

Ký hiệu x^* là một giá trị cụ thể của biến độc lập x và
$\mu_{Y.x^*}$ là giá trị trung bình của Y khi x có giá trị x^*
$\sigma^2_{Y.x^*}$ là phương sai của Y khi x có giá trị x^*

Thay cho các ký hiệu $E(Y|x^*)$ và $V(Y|x^*)$. Ví dụ như, x là tuổi của một đứa trẻ và y là số lượng từ của trẻ, thì $\mu_{Y.5}$ là lượng từ trung bình của tất cả các đứa trẻ 5 tuổi trong tổng thể biết và $\sigma^2_{Y.5}$ là đại lượng mô tả sự phân tán các giá trị của y (lượng từ vựng của những đứa trẻ 5 tuổi) trong tổng thể so với giá trị trung bình $\sigma^2_{Y.5}$.

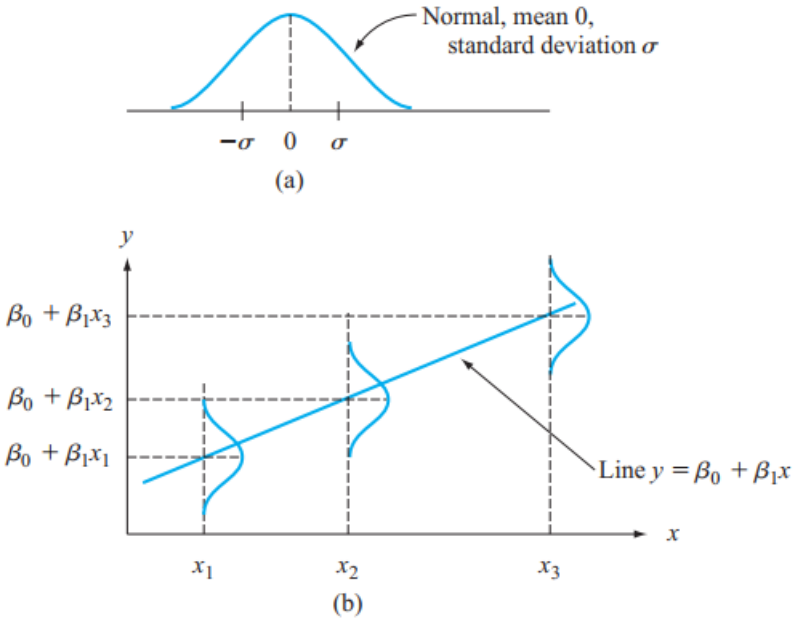
Với x cố định chỉ có duy nhất đại lượng ϵ trong vế phải của (12.1) là biến ngẫu nhiên.

$$\mu_{Y.x^*} = E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^*$$
$$\sigma^2_{Y.x^*} = V(\beta_0 + \beta_1 x^* + \epsilon) = V(\beta_0 + \beta_1 x^*) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Thay x^* bởi x ta có $\mu_{Y.x} = \beta_0 + \beta_1 x$ tức là trung bình của y là hàm tuyến tính của x. Đường hồi quy đúng $y = \beta_0 + \beta_1 x$ suy ra là đường của các giá trị trung bình của Y trên một giá trị cụ thể của x. Hệ số dốc β_1 được diễn tả như sự thay đổi mong đợi của Y tương ứng với sự tăng lên một đơn vị của x. $\sigma^2_{Y.x^*} = \sigma^2$ đẳng thức cho thấy phương sai của Y tại mỗi giá trị khác nhau của x là như nhau (phương sai không đổi).

Trong ví dụ tuổi của trẻ và lượng từ vựng, mô hình cho thấy lượng từ vựng trung bình của trẻ thay đổi tuyến tính với tuổi của trẻ và có phương sai không đổi tại mọi độ tuổi của trẻ.

Với mỗi x cố định Y là tổng của một hằng số $\beta_0 + \beta_1 x$ và sai số ngẫu nhiên có ϵ có phân phối chuẩn nên bản thân Y cũng có phân phối chuẩn.



Hình 10.2: (a) phân phối của ϵ, (b) phân phối của Y với các giá trị khác nhau của x.

Tham số σ^2 xác định khu vực phân tán của đường cong chuẩn so với giá trị

trung bình (cũng là cao độ của đường thẳng. Khi σ^2 nhỏ một điểm quan sát (x, y) sẽ hầu hết rơi vào vị trí gần đường hồi quy đúng và ngược lại khi σ^2 lớn.

***Ví dụ* 10.1.** Giả sử x và y có quan hệ hồi quy tuyến tính đơn giản với đường hồi quy đúng là $y = 65 - 1,2x$ và $\sigma = 8$. Khi đó với giá trị cố định x^* bất kỳ y có giá trị trung bình là $65 - 1,2x^*$ và độ lệch chuẩn là 8.

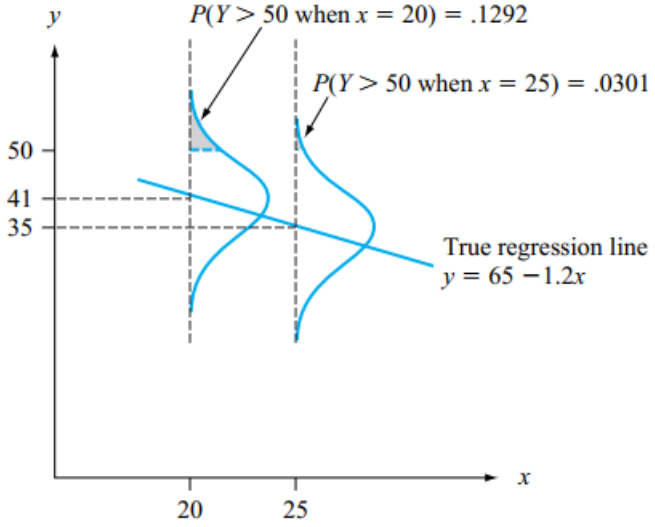
Với $x = 20$ biến ngẫu nhiên Y có giá trị trung bình $\mu_{Y.20} = 65 - 1,2(20) = 41$

$$P(Y > 50|x = 20) = P\left(Z = \frac{Y - 41}{8} > \frac{50 - 41}{8}\right) = 1 - \varnothing(1,13)$$
$$= 0,1292$$

Với $x = 25$ thì $\mu_{Y.25} = 65 - 1,2(25) = 35$

$$P(Y > 50|x = 25) = P\left(Z > \frac{50 - 35}{8}\right) = 1 - \varnothing(1,88)$$
$$= 0,0301$$

Các giá trị xác suất này được minh họa trong hình 12.5.



Hình 10.3: Xác suất dựa trên mô hình hồi quy tuyến tính đơn giản.

Giả sử Y_1 là quan sát được $x = 25$ và Y_2 là quan sát được thực hiện khi $x = 24$

Khi đó $Y_1 - Y_2$ là biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình $E(Y_1 - Y_2) = E(Y_1) - E(Y_2) = \beta_1 = -1,2$, giá trị phương sai $V(Y_1 - Y_2) =$

$V(Y_1)+V(Y_2)=\sigma^2+\sigma^2=128$, giá trị độ lệch chuẩn $\sigma=\sqrt{128}=11,314$. Xác suất Y_1 lớn hơn Y_2 là

$$P(Y_1-Y_2>0)=P(Z>\frac{0-(-1,2)}{11,314})=P(Z>0,11)=0,4562$$

Tức là mặc dù ta kỳ vọng Y giảm khi x tăng 1 đơn vị, quan sát của Y tại $x+1$ chưa chắc đã lớn hơn quan sát của Y tại x.

Bài tập 12.1

1. Tỷ số năng suất cho mẫu phép ngàm trong bể phốt phát được tính bằng cách lấy trọng lượng lớp phốt phát phủ ngoài chia cho lượng kim loại bị mất (cả hai cùng lấy đơn vị: mg/ft^2). Bài báo "Statistical Process Control of a Phosphate Coating Line" (Wire J. Intl., May 1997: 78–81) đưa ra bộ số liệu ghép cặp giữa nhiệt độ trong thùng (x) và tỷ số năng suất (y)

Temp.	170	172	173	174	174	175	176
Ratio	.84	1.31	1.42	1.03	1.07	1.08	1.04
Temp.	177	180	180	180	180	180	181
Ratio	1.80	1.45	1.60	1.61	2.13	2.15	.84
Temp.	181	182	182	182	182	184	184
Ratio	1.43	.90	1.81	1.94	2.68	1.49	2.52
Temp.	185	186	188				
Ratio	3.00	1.87	3.08				

   (a) Biểu diễn biểu đồ gốc lá của nhiệt độ (x) và tỷ số năng suất (y).

   (b) Giá trị của tỷ số hiệu suất được xác định đầy đủ và duy nhất ở nhiệt độ trong thùng hay không?

   (c) Vẽ biểu đồ chấm của dữ liệu. Biểu đồ có cho thấy ta có khả năng dự đoán tỷ số năng suất y bởi nhiệt độ trong thùng ngàm x hay không? Giải thích.

2. Cho bộ số liệu ghép cặp

x	17	32	35	40	40	48	65	70	84	88	94	97
y	38	62	54	68	85	80	93	105	116	117	127	114
x	99	100	110	111	120	123	134	168	172	178	182	191
y	132	136	134	139	142	170	149	164	188	195	200	215

Vẽ biểu đồ chấm của dữ liệu trên. Biểu đồ có cho thấy mối quan hệ x và y?

3. Giả sử rằng biến ngẫu nhiên Y có quan hệ hồi quy tuyến tính đơn giản với tính độc lập X với phương trình hồi quy đúng có dạng $y = 1800 + 1.3x$

   (a) Giá trị trung bình của Y là bao nhiêu khi $x = 2500$.

   (b) Giá trị Y thay đổi trung bình là bao nhiêu khi x tăng thêm 1 đơn vị.

   (c) Trả lời câu b khi x tăng thêm 100 đơn vị.

   (d) Trả lời câu b khi x giảm đi 100 đơn vị.

4. Tiếp tục với giả thuyết trong bài 3, giả sử thêm độ lệch chuẩn của sai số ngẫu nhiên ε là 350.

   (a) Tính xác suất giá trị của Y vượt quá 5000 khi giá trị của x là 2000.

   (b) Trả lời câu a khi giá trị của x là 2500.

   (c) Thực hiện hai quan sát độc lập tương ứng với $x = 2000$ và $x = 2500$. Tính xác suất quan sát thứ hai (tương ứng với $x = 2500$) lớn hơn quan sát thứ nhất (tương ứng với $x = 2000$) là 1000. Ký hiệu Y_1, Y_2 là các quan sát tương ứng với $x = x_1$ và $x = x_2$. Hỏi x_2 cần lớn hơn x_1 bao nhiêu để $P(Y_2 > Y_1) = 0,95$?

5. Tốc độ dòng chảy $y(m^3/\text{phút})$ trong một thiết bị sử dụng cho đo lường chất lượng khí phụ thuộc vào áp suất rơi x (đơn vị: inch) qua máy lọc thiết bị. Giả sử rằng giá trị của x trong khoảng 5 đến 20. Hai biến quan hệ qua mô hình hồi quy tuyến tính đơn giản với hàm hồi quy đúng $y = -0,12 + 0,95x$.

   (a) Tính kỳ vọng tốc độ dòng chảy thay đổi tương ứng khi áp suất rơi x tăng.

   (b) Tốc độ dòng chảy thay đổi trung bình bao nhiêu khi áp suất rơi giảm 5 inch.

   (c) Tốc độ dòng chảy trung bình là bao nhiêu khi áp suất rơi là 10 inch? 15 inch?

   (d) Giả sử $\sigma = 0,025$ và áp suất rơi là 10 inch. Tính xác suất giá trị tốc độ dòng chảy vượt quá 0,84?

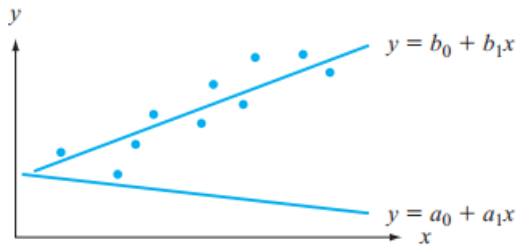
   (e) Tính xác suất tốc độ dòng chảy khi áp suất rơi và là 10 inch vượt quá tốc độ dòng chảy khi áp suất rơi là 11 inch.

6. Biến phụ thuộc Y quan hệ với biến độc lập x qua mô hình hồi quy tuyến tính đơn giản có hàm hồi quy đúng $y = 4000 + 10x$. Biết rằng $P(Y > 5500|x = 100) = 0,05$ và $P(Y > 6500|x = 200) = 0,10$. Tính độ lệch chuẩn σ của sai số ngẫu nhiên ϵ.

10.2 Ước lượng hệ số hồi quy

Ta sẽ giả sử trong mục này và các mục khác là các biến x và y có quan hệ hồi quy tuyến tính đơn giản. Các giá trị β_0, β_1 và σ^2 sẽ hầu hết không biết được trong một nghiên cứu. Thay vào đó có mẫu dữ liệu gồm n cặp giá trị $(x_1, y_1), \ldots, (x_n, y_n)$ từ dữ liệu này sẽ ước lượng được mô hình tham số và đường hồi quy đúng. Các quan sát được giả thuyết thực hiện một cách độc lập tức là y_i là giá trị quan sát của Y_i với $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ và n độ lệch $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ là độc lập. Dẫn đến $Y_1, Y_2, \ldots, Y_n$ độc lập.

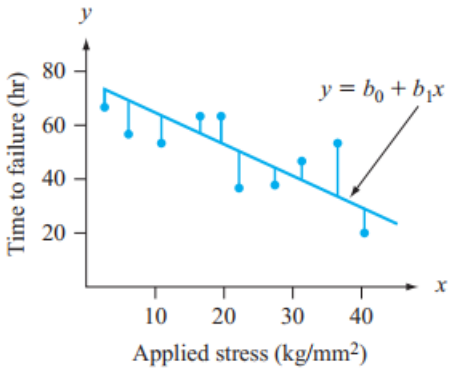
Qua mô hình các điểm quan sát sẽ bị phân phối bởi đường hồi quy đúng theo cách ngẫu nhiên. Hình 10.4 chỉ ra một kiểu các chấm của các quan sát của hai biến. Theo trực giác đường $y = a_0 + a_1 x$ không hợp lý để ước lượng đường hồi quy đúng $y = \beta_0 + \beta_1 x$ bởi vì nếu $y = a_0 + a_1 x$ là đường hồi quy đúng thì các điểm quan sát hầu hết phải nằm gần đường này. Đường thẳng $y = b_0 + b_1 x$ là ước lượng hợp lý bởi hầu hết các điểm nằm gần đường này hơn.



Hình 10.4: Hai ước lượng khác nhau của đường hồi quy đúng.

Theo hình 12.6 và thảo luận đã nói dẫn đến ước lượng của $y = \beta_0 + \beta_1 x$ nên là đường thẳng phù hợp nhất với các điểm quan sát. Theo nguyên lý bình phương tối thiểu được đưa ra bởi nhà toán học người Đức - Gauss (1777 - 1855) đường thẳng phù hợp với dữ liệu là đường có khoảng cách thẳng đứng giữa các điểm quan sát và đường này là nhỏ nhất (xem hình 10.5). Để đo sự phù hợp ta lấy tổng các bình

phương các độ lệch. Đường phù hợp nhất là đường có tổng các bình phương các độ lệch là nhỏ nhất.



Hình 10.5: Độ lệch của các dữ liệu quan sát với đường $y = b_0 + b_1x$.

Nguyên lý bình phương tối thiểu

Độ lệch thẳng đứng của điểm (x_i, y_i) với đường $y = b_0 + b_1x$ là *(cao độ của điểm)* - *(chiều cao của đường)* $= y_i - (b_0 + b_1x_i)$. Tổng các bình phương độ lệch thẳng đứng của các điểm $(x_1, y_1), \ldots, (x_n, y_n)$ tới đường thẳng là

$$f(b_0, b_1) = \sum_{x=1}^{n} [y_i - (b_0 + b_1x_i)]^2.$$

Ước lượng điểm của β_0 và β_1 ký hiệu là $\widehat{\beta}_0$ và $\widehat{\beta}_1$ và gọi là **ước lượng bình phương tối thiểu** là các giá trị làm cực tiểu $f(b_0, b_1)$. Tức là $\widehat{\beta}_0$ và $\widehat{\beta}_1$ thỏa mãn $f(\widehat{\beta}_1, \widehat{\beta}_1) \leq f(b_0, b_1)$ với mọi giá trị b_0 và b_1. **Đường hồi quy ước lượng** hay **đường bình phương tối thiểu** là đường có phương trình là $y = \widehat{\beta}_0 + \widehat{\beta}_1x$.

Đạo hàm riêng $f(b_0, b_1)$ theo b_0, b_1 và giải các phương trình các đạo hàm riêng bằng 0.

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1x_i)(-1) = 0$$
$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1x_i)(-x_i) = 0$$

Rút gọn nhân tử -2 và biến đổi tương đương ta có các phương trình sau gọi là **các phương trình chuẩn**

$$nb_0 + (\sum x_i)b_1 = \sum y_i$$
$$(\sum x_i)b_0 + (\sum x_i^2)b_1 = \sum x_i y_i$$

Các phương trình này là các phương trình tuyến tính theo hai ẩn b_0 và b_1. Cho rằng không phải tất cả các x_i là như nhau, ước lượng bình phương tối thiểu là nghiệm duy nhất của hệ phương trình này.

Ước lượng bình phương tối thiểu của hệ số dốc β_1 của đường hồi quy đúng là

$$b_1 = \widehat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{10.2}$$

Ước lượng bình phương tối thiểu của một hệ số tự do β_0 của đường hội quy đúng là

$$b_0 = \widehat{\beta}_0 = \frac{\sum y_i - \widehat{\beta}_1 \sum x_i}{n} = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{10.3}$$

Trong đó

$$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)$$
$$S_{xx} = \sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2.$$

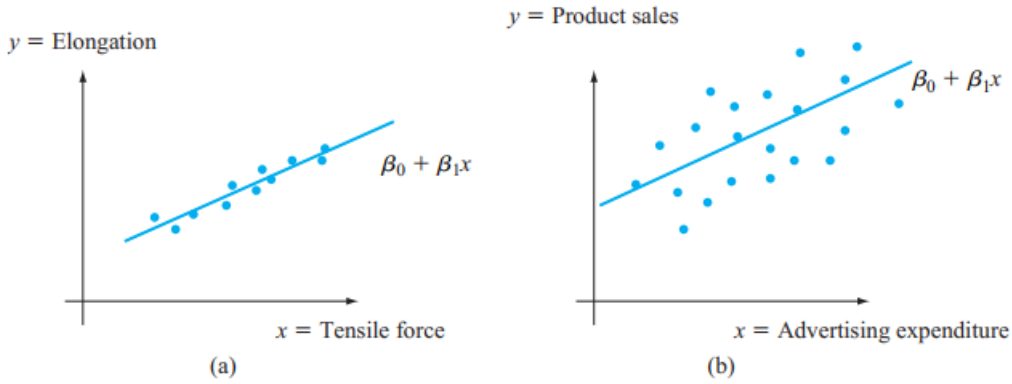
***Ví dụ* 10.2.** Cho bộ số liệu ghép cặp

x	132	128	120	115	114	102	105	98	95	98
y	46	48	51	52	54	55	53	57	56	58
x	92	91	89	78	75	73	68	65	71	59
y	61	60	61	63	65	66	67	72	65	70

$$\sum x_i = 1868 \qquad \sum x_i^2 = 183166$$
$$\sum y_i = 1180 \qquad \sum x_i y_i = 107326$$
$$n = 20 \qquad \sum y_i^2 = 70614$$
$$S_{xy} = 8694,8 \qquad b_1 = -0,3319225284$$
$$n = -2886 \qquad b_0 = 90,00156415$$

Ước lượng σ^2 và σ

Hệ số σ^2 xác định lượng thay đổi vốn có trong mô hình hồi quy. Khi σ^2 lớn, dẫn tới các quan sát (x_i, y_i) khá phân tán so với đường hồi quy thực, còn khi σ^2 nhỏ các điểm quan sát (x_i, y_i) tiến dần về đường hồi quy thực (xem hình 10.9). Ước lượng của σ^2 và quá trình kiểm định giả thuyết thống kê sẽ trình bày trong hai mục tiếp theo. Bởi vì phương trình đường thẳng là không biết, ước lượng được dựa trên phạm vi mẫu quan sát chệch so với đường ước lượng. Độ lệch lớn gợi ý giá trị σ^2 lớn còn các độ chệch có độ lớn là nhỏ thì gợi ý tới giá trị σ^2 là nhỏ.



Hình 10.6: Kiểu mẫu cho a) phương sai nhỏ, b) phương sai lớn.

Định nghĩa 10.2.

Các giá trị thích hợp (hay dự đoán) $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$ tính được bằng cách thế lần lượt các giá trị $x_1, x_2, \ldots, x_n$ vào phương trình hồi quy ước lượng: $\widehat{y}_1 = \widehat{\beta}_0 + \widehat{\beta}_1 x_1, \widehat{y}_2 = \widehat{\beta}_0 + \widehat{\beta}_2 x_2, \ldots, \widehat{y}_n = \widehat{\beta}_0 + \widehat{\beta}_n x_n$ các phần dư là hiệu giữa các giá trị quan sát và giá trị dự đoán $y_1 - \widehat{y}_1, y_2 - \widehat{y}_2, \ldots, y_n - \widehat{y}_n$.

Thông thường các độ lệch với trung bình trong một mẫu được sử dụng tính giá trị ước lượng $s^2 = \sum \dfrac{(x_i - \overline{x})^2}{n-1}$, ước lượng của σ^2 trong phân tích hồi quy được dựa trên tổng các bình phương độ lệch và tiếp tục sử dụng kí hiệu s^2 cho ước lượng của phương sai nên đừng bối rối đối với ký hiệu S^2 trước đó.

Định nghĩa 10.3. **Tổng bình phương các sai số** (hay tổng bình phương các độ lệch) kí hiệu là SSE

$$SSE = \sum (y_i - \widehat{y}_i)^2 = \sum \left[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right]^2$$

và ước lượng của σ^2 là

$$\widehat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}$$

Hiệu $n-2$ trong s^2 là bậc tự do (df) tương ứng của SSE và ước lượng s^2,
Công thức tương đương để tính giá trị SSE là

$$SSE = \sum y_i^2 - \widehat{\beta}_0 \sum y_i - \widehat{\beta}_1 \sum x_i y_i.$$

***Ví dụ* 10.3.** Cho bộ số liệu ghép cặp

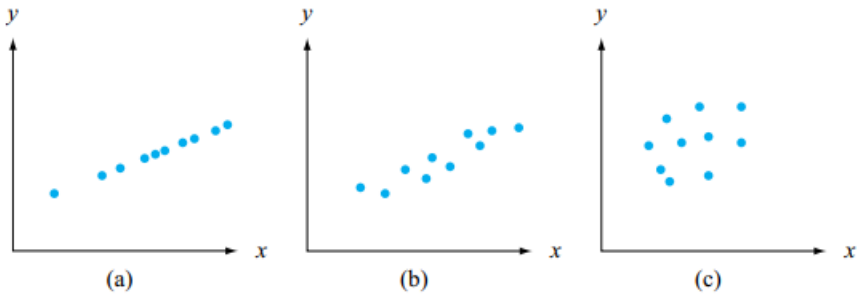
x	15	23	35	40	48	55	61	69	75	83
y	4,9	4,7	4,5	4,6	4,4	4,1	4,2	4,0	3,8	3,7
x	90	98	102	108	112	119	125	132	140	
y	3,6	3,4	3,5	3,3	3,0	2,8	2,6	2,3	2,0	

Cỡ mẫu $n = 19$
$\sum x_i = 1530;$ $\sum x_i^2 = 149030$
$\sum y_i = 69,4;$ $\sum y_i^2 = 266$
$\sum x_i y_i = 5032,5$
Từ đó ta tính được

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 25824,73684$$
$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = -556,0263158$$
$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -0,02153076406$$
$$\widehat{\beta}_0 = \frac{\sum y_i - \widehat{\beta}_1 \sum x_i}{n} = \overline{y} - \widehat{\beta}_1 \overline{x} = -5,386424685$$
$$SSE = 0,535696993.$$

Hệ số xác định (hệ số đo độ phù hợp của mô hình)

Hình 10.7 biểu diễn 3 biểu đồ chấm của mẫu ghép cặp. Trong cả 3 biểu đồ cao độ của các điểm khác nhau là biến thiên cho có sự thay đổi trong các giá trị quan sát y. Các điểm trong biểu đồ 1 chính xác thuộc một đường thẳng. Trong trường hợp này 100% mẫu quan sát được cho là x và y quan hệ tuyến tính. Biểu đồ 10.7(b) không rơi chính xác trên một đường, nhưng so sánh tất cả các độ chệch của y so với đường bình phương tối thiểu là nhỏ. Trong trường hợp này có lý do các giá trị khác nhau của y có thể tính được bằng cách xấp xỉ tuyến tính với biến được yêu cầu bởi mô hình hồi quy tuyến tính đơn giản. Khi biểu đồ giống hình 10.7(c) có mức độ



Hình 10.7: Sử dụng mô hình để giải thích sự biến thiên của y: a) dữ liệu với mọi sự biến thiên giải thích được, b) dữ liệu với hầu hết sự biến thiên giải được, c) dữ liệu với rất ít sự biến thiên giải thích được.

biến đổi giữa đường bình phương tối thiểu với các quan sát y nên mô hình hồi quy tuyến tính đơn giản không sử dụng để giải thích y bởi x.

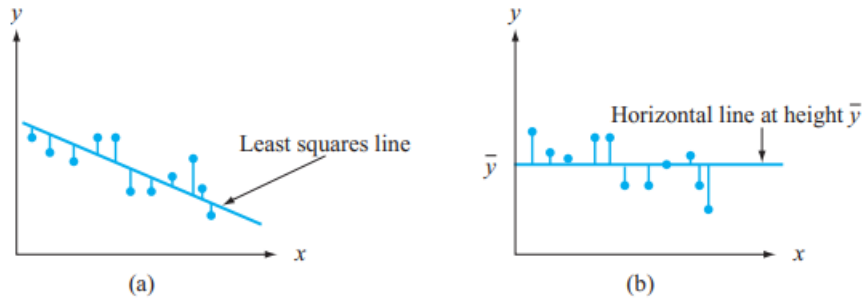
Tổng bình phương các sai số có thể được hiểu như một độ đo. Đo xem có bao nhiêu sự biến thiên của y không giải thích được bởi mô hình. Trong hình 10.7(a) $SSE = 0$ tức là không có sự biến thiên nào không được giải thích, SSE là nhỏ trong hình 10.7(b) khi có số ít sự biến thiên của y không giải thích được bởi mô hình và SSE lớn trong hình 10.7(c) khi hầu hết sự biến thiên của y không giải thích được bởi mô hình.

Một giá trị đo tổng lượng thay đổi các giá trị quan sát y cho bởi **tổng các bình phương**

$$SST = S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2.$$

Tổng các bình phương là tổng của các bình phương độ lệch với trung bình mẫu và các giá trị quan sát y. Giống như SSE là tổng của bình phương các độ lệch với đường bình phương tối thiểu $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$, SST là tổng các bình phương tối thiểu so với đường nằm ngang tại $\overline{y}$ (do đó các độ lệch thẳng đứng là $y_i - \overline{y}$) như hình 10.8. Hơn nữa vì tổng các bình phương độ lệch so với đường bình phương tối thiểu là nhỏ hơn tổng các bình phương tối thiểu so với các đường khác, SSE $<$ SST trừ khi đường nằm ngang là đường bình phương tối thiểu.

Tỷ số $\dfrac{SSE}{SST}$ là tỷ lệ của tổng các biến thiên mà không thể giải thích bởi mô hình hồi quy tuyến tính đơn giản và $1 - SSE/SST$ (là một số nằm giữa 0 và 1) là tỷ lệ của các biến thiên của y được giải thích bởi mô hình.



Hình 10.8: Minh họa tổng các bình phương: a) SSE: tổng các bình phương độ lệch so với đường bình phương tối thiểu b) SST: tổng các bình phương độ lệch so với đường nằm ngang.

Định nghĩa 10.4. **Hệ số xác định** ký hiệu là r^2, cho bởi công thức

$$r^2 = 1 - \frac{SSE}{SST}$$

Hệ số xác định như tỷ lệ sự thay đổi (biến thiên) của quan sát y có thể giải thích bởi mô hình hồi quy tuyến tính đơn giản.

Hệ số r^2 càng cao mô hình hồi quy đơn giản giải thích cho sự thay đổi của y càng tốt (thích hợp). Khi sử dụng phân tích hồi quy trong các phần mềm thống kê, r^2 hay $100r^2$ (số phần trăm sự biến thiên có thể giải thích bởi mô hình) là một kết quả sẽ được đưa ra. Nếu r^2 nhỏ nhà nghiên cứu sẽ mong muốn sử dụng mô hình khác (như hồi quy phi tuyến hay mô hình hồi bội có nhiều hơn một biến độc lập) để giải thích cho sự thay đổi của y.

***Ví dụ* 10.4.** Tiếp tục ví dụ trên ta có:

$$\begin{aligned} SST &= \sum y_i^2 - \frac{1}{n}(\sum y_i)^2 \\ &= 266 - \frac{1}{19}69,4^2 = \frac{5941}{475} \\ &= 12,507368421052631578 44 \end{aligned}$$

$$\begin{aligned} SSE &= \sum y_i^2 - \widehat{\beta}_0 \sum y_i - \widehat{\beta}_1 \sum x_i y_i \\ &= 266 - (5,386424685).(69,4) + 0,0215307640650345 \\ &= 0,535696993 \end{aligned}$$

$$r^2 = 1 - \frac{SSE}{SST} = 0,957169488$$

Bài tập 12.2

7. (a) Xác định phương trình đường bình phương tối thiểu cho số liệu ghép cặp trong bài 2.

   (b) Dự đoán giá trị của y khi biết $x = 35$ và tính giá trị phần dư tương ứng.

   (c) Tính SSE và một giá trị ước lượng điểm của σ.

   (d) Tỷ lệ sự biến thiên thay đổi của y có thể giải thích được quan hệ xấp xỉ tuyến tính giữa hai biến là bao nhiêu?

8. Bài báo "Characterization of Highway Runoff in Austin, Texas, Area" (J. of Envir. Engr., 1998: 131–137). Khảo sát về lượng mưa $x(m^3)$ và lượng nước thoát $y(m^3)$ tại một điểm cụ thể được bộ số liệu tương ứng

x	5	12	14	17	23	30	40	47
y	4	10	13	15	15	25	27	46
x	55	67	72	81	96	112	127	
y	38	46	53	70	82	99	100	

   (a) Vẽ biểu đồ chấm cho dữ liệu này. Biểu đồ chấm có đưa tới việc dùng mô hình hồi quy tuyến tính đơn giản?

   (b) Hãy tính một giá trị ước lượng điểm cho hệ số dốc và hệ số tự do của đường hồi quy tổng thể.

   (c) Tính một giá trị ước lượng điểm cho lượng nước thoát trung bình khi lượng nước mưa là $50m^3$.

   (d) Tính một giá trị ước lượng điểm cho độ lệch chuẩn.

   (e) Tỷ lệ sự thay đổi của lượng nước thoát có thể giải thích bởi mối quan hệ hồi quy tuyến tính giữa lượng nước mưa x và lượng nước thoát y?

9. Cho bộ số liệu ghép cặp

x	102,3	87	82	76	92	89,2	85,5	93,5	79	76,7
y	85	81	67,7	58,7	84,3	83,3	78	69	67,5	58,3

   (a) Xác định đường bình phương tối thiểu của dữ liệu và giải thích các hệ số.

(b) Tính hệ số xác định r^2 và giải thích.

(c) Tính và giải thích một giá trị ước lượng điểm cho độ lệch chuẩn σ trong mô hình hồi quy tuyến tính đơn giản.

10. Theo dữ liệu công bố trong bài báo "An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data" (J. of Engr. for Power, July 1973: 165–170) với x là tốc độ lan của đám cháy ($MBtu/hr - ft^2$) và y là tốc độ giải phóng ra khí $NO_x(ppm)$

x	100	125	125	150	150	200	200
y	150	140	180	210	190	320	280
x	250	250	300	300	350	400	400
y	400	430	440	390	600	610	670

(a) Giả sử mô hình hồi quy tuyến tính đơn giản có hiệu lực hãy xác định một ước lượng của đường hồi quy đúng.

(b) Ước lượng tốc độ giải phóng khí NO_x trung bình khi tốc độ lan tỏa của đám cháy là 225.

(c) Khi tốc độ lan tỏa của đám cháy giảm 50 thì tốc độ giải phóng khí NO_x thay đổi trung bình là bao nhiêu?

(d) Có thể dùng đường hồi quy ước lượng để dự đoán tốc độ giải phóng khí NO_x. Khi tốc độ lan của đám cháy là 500? Tại sao?

11. Cho bộ số liệu ghép cặp

x	0,05	0,1	0,15	0,21	0,28	0,32
y	0,45	0,53	0,52	0,59	0,67	0,89
x	0,37	0,41	0,46	0,51	0,57	0,63
y	0,88	0,93	1,05	1,36	1,48	1,72
x	0,7	0,76	0,82	0,89	0,92	0,98
y	1,69	1,85	1,81	2,01	2,38	2,39

(a) Tính giá trị ước lượng bình phương tối thiểu cho β_0, β_1 trong mô hình hồi quy tuyến tính đơn giản cho bộ số liệu ghép cặp này.

(b) Dự đoán giá trị của y khi $x = 0,5$.

(c) Tính một giá trị ước lượng cho σ.

(d) Tính giá trị của tổng các bình phương thay đổi SST và giá trị hệ số xác định r^2 và đưa ra bình luận về các giá trị này.

(e) Giả sử rằng thay vì tìm đường bình phương tối thiểu qua các điểm $(x_1, y_1), \ldots, (x_n, y_n)$ ta tìm đường bình phương tối thiểu qua các điểm $(x_1 - \overline{x}, y_1), \ldots, (x_n - \overline{x}, y_n)$. Vẽ biểu đồ chấm cho các điểm $(x_1, y_1), \ldots,$ (x_n, y_n) rồi vẽ biểu đồ chấm cho các điểm $(x_1 - \overline{x}, y_1), \ldots, (x_n - \overline{x}, y_n)$. Dùng các biểu đồ này để giải thích bằng trực quan mối quan hệ giữa các biểu đồ chấm và đường phương tối thiểu tương ứng.

(f) Giả sử thay vì tìm mô hình $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i (i = 1, 2, \ldots, n)$ ta tìm mô hình $y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \varepsilon_i \ (i = 1, 2, \ldots, n)$. Tìm ước lượng bình phương tối thiểu cho β_0^* và mối quan hệ với $\widehat{\beta}_0$ và $\widehat{\beta}_1$.

10.3 Hệ số tương quan

Có nhiều nghiên cứu cần chỉ ra có hay không mối quan hệ giữa hai biến hay đúng hơn là có thể sử dụng biến này để dự đoán biến kia hay không. Trong mục này trước tiên ta sẽ xét hệ số tương quan mẫu r như là một độ đo về mối quan hệ giữa hai biến x và y trong mẫu, sau đó xét về mối quan hệ giữa r và hệ số tương quan ρ đã định nghĩa trong chương 5.

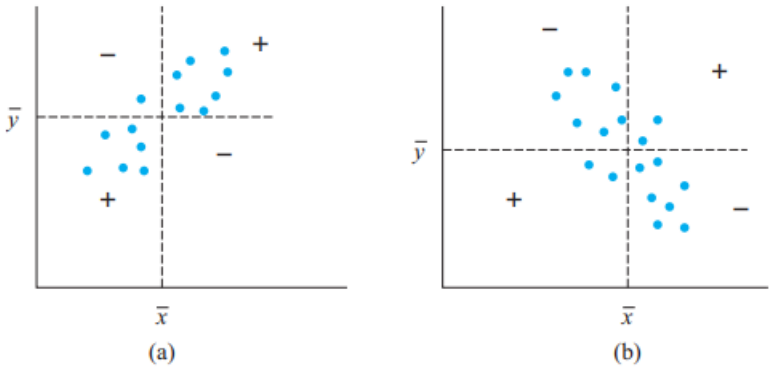
Hệ số tương quan mẫu

Cho n cặp số $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Rất tự nhiên khi ta nói x và y có mối quan hệ dương (đồng biến) nếu giá trị lớn của x ghép cặp với giá trị lớn của y và giá trị nhỏ của x ghép cặp với giá trị nhỏ của y. Còn x và y có mối quan hệ âm (nghịch biến) nếu giá trị lớn của x ghép cặp với giá trị nhỏ của y và giá trị nhỏ của x ghép cặp với giá trị lớn của y. Xét đại lượng:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Nếu mối quan hệ dương là mạnh thì x_i lớn hơn $\overline{x}$ giá trị ghép cặp y_i tương ứng lớn hơn $\overline{y}$, do đó $(x_i - \overline{x})(y_i - \overline{y}) > 0$ là tích này cũng sẽ âm khi cả hai giá trị x_i, y_i nhỏ

hơn các trung bình $\overline{x}, \overline{y}$ tương ứng. Suy ra x, y có mối quan hệ dương (đồng biến) thì S_{xy} sẽ dương. Lập luận tương tự chỉ ra rằng nếu x, y có mối quan hệ âm thì tích $(x_i - \overline{x})(y_i - \overline{y})$ sẽ âm. Minh họa trong hình 10.9.



Hình 10.9: (a) biểu đồ chấm với S_{xy} dương, (b) biểu đồ chấm với S_{xy} âm. [+ nghĩa là trung bình $(x_i - \overline{x})(y_i - \overline{y}) > 0$ và - nghĩa là trung bình $(x_i - \overline{x})(y_i - \overline{y}) < 0$]

Mặc dù S_{xy} có vẻ như là một độ đo tin cậy đo mối quan hệ giữa x, y tuy nhiên khi thay đổi đơn vị của x hay y thì độ lớn sẽ thay đổi hoặc rất lớn hoặc rất gần 0. Ví dụ như bằng $S_{xy} = 25$ khi x có độ đo là m và $S_{xy} = 25000$ khi x có đơn vị là mm và $S_{xy} = 0,205$ khi x có đơn vị là km. Vì vậy ta cần một độ đo mối quan hệ giữa x và y mà không phụ thuộc vào đơn vị độ đo được sử dụng để đo các biến này. Từ đó hệ số tương quan mẫu được đưa ra từ việc điều chỉnh S_{xy}.

Định nghĩa 10.5. **Hệ số tương quan mẫu** cho n cặp giá trị $(x_1, y_1), \ldots, (x_n, y_n)$ là

$$r = \frac{S_{xy}}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Tính chất của r

Ta có một số tính chất quan trọng của r như sau:

1. Giá trị của r không phụ thuộc vào biến nào đánh nhãn là x hay y.
2. Giá trị của r độc lập với các đơn vị dùng để đo biến x, y.
3. $-1 \leq r \leq 1$
4. $r = 1$ nếu và chỉ nếu mọi cặp (x_i, y_i) nằm trên một đường thẳng với hệ số góc dương và $r = -1$ nếu mọi cặp (x_i, y_i) cùng nằm trên một đường thẳng với hệ số góc âm.

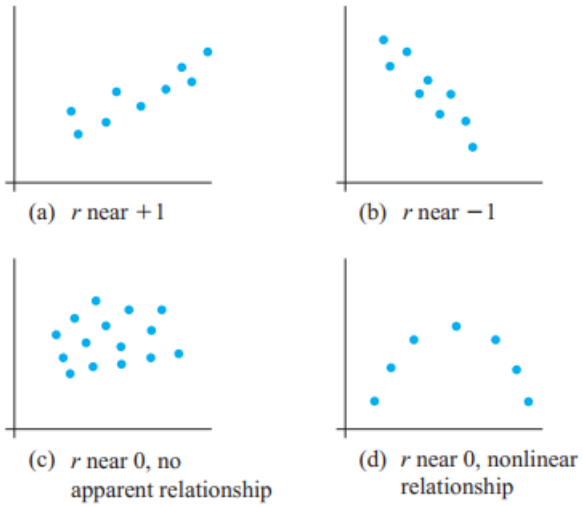
5. Bình phương của hệ số tương quan mẫu là giá trị của hệ số xác định sự thích hợp của mô hình hồi quy tuyến tính đơn giản, theo ký hiệu $(r)^2 = r^2$.

Tính chất 1 chỉ ra có sự tương phản trong phân tích hồi quy khi mọi đại lượng ta quan tâm (hệ số dốc, hệ số tự do, s^2, ...) phụ thuộc vào việc trong hai biến ta xem biến nào là biến phụ thuộc. Tuy nhiên tính chất 5 chỉ ra rằng tỷ lệ sự thay đổi (biến thiên) của biến phụ thuộc có thể giải thích được bởi mô hình hồi quy tuyến tính đơn giản không phụ thuộc vào việc biến nào có vai trò gì.

Cách phát biểu khác của tính chất 2 là giá trị của r không thay đổi nếu mỗi x_i được thay bởi cx_i và nếu mỗi y_i được thay bởi dy_i cũng như nếu mỗi x_i được thay bởi $x_i - a$ và mỗi y_i được thay bởi $y_i - b$.

Tính chất 3 cho rằng giá trị lớn nhất của r, tương ứng với quan hệ cực dương là 1 còn tương ứng với quan hệ âm nhất là $r = -1$. Theo tính chất 4, tương quan dương và âm nhất chỉ khi mọi điểm (x_i, y_i) đều nằm trên một đường thẳng. Các hình dạng khác của đám mây điểm sẽ gợi ý về mối quan hệ số giữa các biến nếu trị tuyệt đối của r gần 1, suy ra đo mức độ phụ thuộc tuyến tính giữa các biến.

Một giá trị r gần bằng 0 ta có thể nói giữa các biến không có quan hệ tuyến tính nhưng không thể khẳng định không có mối quan hệ nào giữa các biến này. Hình 10.10 minh họa các hình dạng khác nhau của đám mây điểm tương ứng với giá trị khác nhau của r.



Hình 10.10: Biểu đồ chấm của các dữ liệu tương ứng với các giá trị r khác nhau.

Quy tắc kết luận về mức độ tương quan giữa hai biến mạnh hay yếu dựa trên giá trị của hệ số tương quan mẫu r.

Yếu	**Trung bình**	**Mạnh**
$-0,5 \leq r \leq 0,5$	$-0,8 < r < -0,5$ hoặc $0,5 < r < 0,8$	$r \geq 0,8$ hay $r \leq -0,8$

Kết luận về hệ số tương quan của tổng thể

Hệ số tương quan mẫu r là một độ đo mức độ của quan hệ giữa x và y trong mẫu quan sát. Ta có thể xem các cặp giá trị (x_i, y_i) được rút ra từ tổng thể ghép cặp với (x_i, y_i) có hàm sát xuất hay mật độ đồng thời. Trong chương 5 ta định nghĩa hệ số tương quan

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

Trong đó

$$Cov(X,Y) = \begin{cases} \sum\limits_x \sum\limits_y (x-\mu_x)(y-\mu_y)p(x,y) & \text{với } (X,Y) \text{ rời rạc} \\ \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} (x-\mu_x)(y-\mu_y)f(x,y) & \text{với } (X,Y) \text{ liên tục} \end{cases}$$

Nếu $p(x,y)$ hay $f(x,y)$ mô tả phân phối của cả tổng thể ghép cặp thì ρ là độ do mức độ mạnh của quan hệ giữa x và y trong tổng thể.

Hệ số tương quan tổng thể ρ là một tham số hay một đặc tính của tổng thể như μ_x, μ_y, σ_x và σ_y, do đó ta có thể sử dụng hệ số tương quan mẫu để đưa ra kết luận về ρ. Cụ thể r là 1 giá trị ước lượng điểm cho ρ và giá trị ước lượng tương ứng là

$$\widehat{\rho} = r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

***Ví dụ* 10.5.** Cho mẫu ghép cặp

x	562	552	562	537	526	515	505	480	460	450	445	435	421	418	400
y	2,1	2,3	2,2	2,5	2,5	2,7	2,8	2,8	3,2	3,3	3,3	3,5	3,6	3,6	3,7

$$n = 15$$
$$\sum x_i = 7268$$
$$\sum x_i^2 = 3565402$$
$$\sum y_i = 44,1$$
$$\sum y_i^2 = 133,89$$
$$\sum x_i y_i = 20940,6$$

Hệ số tương qua mẫu

$$r = \frac{20940,6 - (7268)(44,1)/15}{\sqrt{3565402 - (7268)^2/15}\sqrt{133,89 - (44,1)^2/15}}$$
$$= -0,9919050782$$

Một giá trị ước lượng điểm cho hệ số tương quan tổng thể ρ là $\hat{\rho} = r = -0,9919050782$.

Bài tập

Bài 1 Quan sát việc tổng hợp sinh khối ở một nhà máy từ năng lượng bức xạ mặt trời sau 8 tuần người ta thu được bảng số liệu sau:

Bức xạ mặt trời	30	68	121	217	314	419	536	642
Trọng lượng sinh khối (gram)	17	49	122	220	376	571	648	756

Dựa vào số liệu này có thể dự đoán được trọng lượng sinh khối qua bức xạ mặt trời bằng hàm hồi quy tuyến tính thực nghiệm hay không? Nếu được hãy dự báo xem khi bức xạ mặt trời ở mức 600 thì trung bình sinh khối được sản xuất là bao nhiêu?
Bài 2 (Hồi quy) Để nghiên cứu sự phát triển của một loại cây trồng, người ta tiến hành đo chiều cao Y (m) và đường kính X (cm) của một số cây. Kết quả được ghi ở bảng sau đây:

X Y	3	4	5	6	7	8
21	2	5				
23		3	11			
25			8	15	10	
27			4	17	3	
29					7	12

Tìm hệ số tương quan mẫu và phương trình hồi quy tuyến tính mẫu Y theo X.
Bài 3 Một công ty ấn định giá bán X của một loại sản phẩm tại 10 miền khác nhau. Bảng sau đây cho biết số lượng Y bán được trong một tháng ứng với giá bán:

X	34	35	36	36	35	37	38	40
Y	6	5,9	5,7	5,7	6,2	6,7	5,6	5,5

1. Có thể biểu diễn só lượng theo giá bán bằng phương trình hồi quy tuyến tính không? Vì sao?

2. Viết phương trình đường thẳng hồi quy mẫu của Y theo X.

Bài 4 Giả sử giá trị quan sát trên một mẫu của (X,Y) tuân theo quy luật phân phối chuẩn hai chiều được cho trong bảng sau:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Viết phương trình đường thẳng hồi quy mẫu của Y theo X. Dự đoán giá trị trung bình của Y khi $X = 12$. Khi X tăng 2 đơn vị thì Y tăng trung bình bao nhiêu?

Bài 5 Do chiều cao X (đơn vị: cm) và trọng lượng Y (kg) của một số học sinh chọn ngẫu nhiên được:

X	155	156	158	159	159	160	160	162
Y	48	47	48	49	48	50	51	51

Có thể dự đoán cân nặng trung bình khi biết chiều cao X bằng hàm hồi quy tuyến tính được không? Nếu có hãy dự báo cân nặng trung bình của các học sinh có chiều cao 165 cm. Khi chiều cao học sinh tăng 5 cm thì cân nặng trung bình tăng bao nhiêu?