Phần II

THỐNG KÊ

Chương 4.

ƯỚC LƯỢNG THAM SỐ

§1. ĐẠI CƯƠNG VỀ MẪU

1. Phương pháp mẫu

1.1. Đám đông

Trong thống kê người ta đưa ra khái niệm đám đông như sau :

Định nghĩa 4.1 : Đám đông hay tổng thể (thường được kí hiệu là *C*) là tập hợp các phần tử có chung một hay một vài dấu hiệu mà ta quan tâm nghiên cứu. Đám đông có thể hữu hạn hay vô hạn phần tử.

Ví dụ 4.1 :

	Đám đông C	*Dấu hiệu nghiên cứu trên từng phần tử*
1	Các sản phẩm của một nhà máy sau một ca sản xuất (hữu hạn phần tử)	Sản phẩm có đạt tiêu chuẩn không
2	Số ngày bán hàng của một cửa hàng trong 1 năm (hữu hạn phần tử)	Ngày có đạt doanh thu theo chỉ tiêu không
3	Các số đo mực nước sông ở một trạm quan trắc tại một thời điểm vào mỗi ngày (vô hạn phần tử)	Số đo mực nước đó có thuộc khoảng cho trước không

Dấu hiệu nghiên cứu của đám đông *C* thể hiện và thay đổi khác nhau trên mỗi phần tử của đám đông, được mô tả như một biến ngẫu nhiên X có mật độ phân phối xác suất f(x) nào đó. X có thể có phân phối nhị thức, phân phối chuẩn,... Chẳng hạn, với đám đông là tập hợp các sản phẩm của nhà máy có tỉ lệ phế phẩm là p, người ta mô tả bằng biến ngẫu nhiên X như sau : Nếu phần tử của đám đông là phế phẩm thì X nhận giá trị 1 với xác suất p, còn phần tử của đám đông không là phế phẩm thì X nhận giá trị là 0 với xác suất 1 – p. Khi đó X có phân phối nhị thức B(1 ; p).

Như vậy nghiên cứu đám đông là nghiên cứu phân phối xác suất của một biến ngẫu nhiên.

Trong thực tế vì nhiều nguyên nhân, chẳng hạn số phần tử của đám đông rất lớn nhưng khi chi phí và thời gian điều tra các phần tử của đám đông có hạn, không thể khảo sát hết các phần tử của đám đông... Do đó việc điều tra toàn bộ các phần tử của đám đông để tìm luật phân phối xác suất của nó là không thể. Vì vậy người ta cần sử dụng phương pháp mẫu.

1.2. Phương pháp mẫu

Phương pháp mẫu là phương pháp chọn ra n phần tử đại diện cho đám đông (hay còn gọi là thiết lập một mẫu kích thước n). Sử dụng các công cụ của thống kê nghiên cứu mẫu này và dựa vào đó cho kết luận về đám đông. Tùy theo đặc trưng nghiên cứu của đám đông có nhiều phương pháp chọn mẫu.

1.2.1. Chọn mẫu ngẫu nhiên

Với phương pháp này, mỗi phần tử của đám đông được chọn vào mẫu có xác suất được biết và khác 0.

a) Chọn mẫu ngẫu nhiên đơn giản

Là phương pháp chọn mẫu thoả mãn các điều kiện :

– Mỗi lần chỉ được chọn một phần tử vào mẫu. Mỗi phần tử của đám đông đều có thể được chọn vào mẫu với cùng khả năng như nhau ;

– Các mẫu cùng kích thước có cùng xác suất được chọn.

Việc chọn mẫu kiểu này có thể tiến hành theo cách bốc thăm hay dùng bảng số ngẫu nhiên. Có hai phương thức chọn : chọn hoàn lại, chọn không hoàn lại (mỗi phần tử chỉ được chọn một lần). Khi số phần tử N của đám đông rất lớn so với kích thước mẫu n ta có thể coi hai phương thức chọn mẫu này là như nhau.

Mẫu ngẫu nhiên đơn giản được lập có tính chất đại diện cao tuy nhiên nó đòi hỏi phải biết toàn bộ đám đông và chi phí chọn mẫu lớn.

b) Chọn mẫu phân nhóm

Là phương pháp chọn mẫu chia đám đông thành các nhóm tương đối thuần nhất, từ mỗi nhóm lấy ra một mẫu ngẫu nhiên. Tập hợp các mẫu đó lập thành một mẫu ngẫu nhiên phân nhóm. Phương pháp này được dùng khi trong đám đông có những sai khác lớn.

c) Chọn mẫu chùm

Là phương pháp chọn mẫu ngẫu nhiên từ các tập con của đám đông, được gọi là các chùm. Mỗi phần tử của đám đông chỉ được chọn vào một chùm, mỗi chùm cố gắng sao cho có độ phân tán cao như đám đông và đồng đều về quy mô. Phương pháp này có thể tiết kiệm chi phí và thời gian, nhưng sai số chọn mẫu cao hơn các phương pháp trên.

1.2.2. Chọn mẫu có suy luận

Phương pháp chọn mẫu này dựa trên ý kiến chuyên gia về đối tượng nghiên cứu. Nhược điểm của phương pháp này là khó đảm bảo tính khách quan.

1.2.3. Thang đo

Để lượng hoá các đặc trưng cần nghiên cứu người ta đưa vào các loại thang đo. Các thang đo cho đặc trưng định tính :

– Thang đo định danh là thang đo dùng để phân loại, nhằm đánh giá các đặc trưng dùng để đếm tần số cho mỗi loại, không tính toán số học được. Ví dụ : nam [0], nữ : [1].

– Thang đo thứ bậc là thang đo định danh nhưng giữa các đặc trưng đã có quan hệ hơn kém, tuy nhiên khoảng cách giữa các bậc không nhất thiết đều nhau. *Ví dụ :* Đặc trưng trình độ : Mù chữ [0] ; Trung học [1] ; Đại học [2].

– Thang đo khoảng là thang đo thứ bậc có khoảng cách đều nhau, do đó có thể dùng để tính toán cộng trừ được. Với đặc trưng định lượng, người ta dùng thang đo tỉ lệ là thang đo khoảng với điểm gốc được xác định tuyệt đối.

1.3. Mẫu ngẫu nhiên, mẫu cụ thể

1.3.1. Khái niệm

Cho đám đông ***C*** có dấu hiệu được mô tả bằng biến ngẫu nhiên X với phân phối xác suất f(x) (ngắn gọn ta nói : Cho đám đông X). Giả sử một mẫu kích thước n được lập từ ***C*** theo phương pháp chọn mẫu ngẫu nhiên đơn giản. Với cách chọn mẫu này, ta kí hiệu X_i, $i = 1, 2, \ldots, n$ là giá trị quan sát thứ i của X (hay thành phần thứ i) của mẫu. X_i nhận được bằng cách lặp lại phép thử trong cùng điều kiện như nhau, độc lập nên được coi như là các biến ngẫu nhiên độc lập, có cùng luật phân phối của X. Từ đó ta có định nghĩa :

Định nghĩa 4.2 : Mẫu ngẫu nhiên kích thước n của đám đông X có phân phối xác suất f(x) là biến ngẫu nhiên n chiều $(X_1, X_2, \ldots, X_n)$, trong đó các thành phần X_i độc lập có cùng phân phối xác suất của X.

Định nghĩa 4.3 : Mẫu cụ thể là một giá trị của mẫu ngẫu nhiên.

Như vậy khi mẫu ngẫu nhiên kích thước n : $(X_1, X_2, \ldots, X_n)$ nhận một giá trị là : $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ thì $(x_1, x_2, \ldots, x_n)$ là một mẫu cụ thể kích thước n.

Ví dụ 4.2 : Nghiên cứu về tuổi thọ X (tính bằng tháng) của một loại bóng đèn được sản xuất bởi một dây chuyền công nghệ, người ta chọn ngẫu nhiên lần lượt 10 bóng của dây chuyền này khảo sát tuổi thọ của chúng (coi lần chọn thứ i, $i = 1, 2, \ldots, 10$ như là làm phép thử thứ i khảo sát về X, nó được mô tả bằng biến ngẫu nhiên X_i). Khi đó ta được $(X_1, X_2, \ldots, X_{10})$ là một mẫu ngẫu nhiên kích thước n = 10 của X. Với bóng đèn thứ nhất có tuổi thọ $x_1 = 15$ (một giá trị của X_1), bóng đèn thứ hai có tuổi thọ $x_2 = 22$ (một giá trị của X_2),..., bóng đèn thứ 10 có tuổi thọ $x_{10} = 18$ (một giá trị của X_{10}), khi đó $(x_1, x_2, \ldots, x_{10}) = (15, 22, \ldots, 18)$ là một mẫu cụ thể của X. Đương nhiên lấy mẫu 10 bóng đèn khác của dây chuyền công nghệ để khảo sát tuổi thọ của loại bóng đèn này sẽ cho một mẫu cụ thể khác. Như vậy một mẫu số liệu điều tra của đám đông là một mẫu cụ thể.

1.3.2. Cách trình bày một mẫu cụ thể

a) Bảng phân phối thực nghiệm

Cho mẫu cụ thể $(x_1, x_2, \ldots, x_n)$ trong đó :

- x_1 có n_1 giá trị ; x_2 có n_2 giá trị, ..., x_k có n_k giá trị,
- $x_1 < x_2 < \ldots < x_k$ và $n_1 + n_2 + \ldots + n_k = n$.

Khi đó n_i, f_i tương ứng được gọi là tần số, tần suất của x_i. Các bảng mô tả số liệu sau được gọi là bảng phân phối thực nghiệm

Bảng phân phối tần số thực nghiệm

x_i	x_1	x_2 ...	x_k
n_i	n_1	n_2 ...	n_k

với $n_1 + n_2 + \ldots + n_k = n$

Bảng phân phối tần suất thực nghiệm

x_i	x_1	x_2 ...	x_k
f_i	f_1	f_2 ...	f_k

$f_i = \dfrac{n_i}{n}$, $f_1 + f_2 + \ldots + f_k = 1$

Chú ý : Từ mẫu cụ thể $(x_1, \ldots, x_n)$ của mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ lập từ đám đông X, đặt :

$$F_n(x) = \frac{\text{Số phần tử } x_i \le x}{n} = \frac{\sum\limits_{x_i \le x} n_i}{n} = \sum_{x_i \le x} f_i .$$

Ta có $F_n(x) \xrightarrow{P} F(x) = P\{X \le x\}$ (theo luật số lớn Bernoulli).

Như vậy với n đủ lớn : $F_n(x) \approx F(x)$.

$F_n(x)$ được gọi là hàm phân phối mẫu. $F_n(x)$ cho biết tỉ lệ các giá trị của mẫu không vượt quá x. Dáng điệu đồ thị của $F_n(x)$ (có thể thấy qua các phần mềm thống kê) cho biết dáng điệu đồ thị của F(x).

Ví dụ 4.3 : Để điều tra thời gian đợi phục vụ của khách hàng tại một ngân hàng (đơn vị : phút) người ta khảo sát ngẫu nhiên 10 người, kết quả thu được như sau : 9, 8, 10, 10, 12, 6, 11, 10, 12, 8.

Lập các bảng phân phối thực nghiệm thời gian đợi của khách hàng.

GIẢI : Từ mẫu (9, 8, 10, 10, 12, 6, 11, 10, 12, 8) ta có : $x_1 = 6$ có 1 giá trị trong mẫu do đó $n_1 = 1$, $x_2 = 8$ có 2 giá trị trong mẫu do đó $n_2 = 2, \ldots, x_6 = 12$ có 2 giá trị trong mẫu do đó $n_6 = 2$. Từ đó ta thiết lập được các bảng :

Bảng phân phối tần số thực nghiệm

x_i	6	8	9	10	11	12
n_i	1	2	1	3	1	2

Bảng phân phối tần suất

x_i	6	8	9	10	11	12
f_i	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

Khi kích thước mẫu lớn, các giá trị của mẫu khá gần nhau người ta chia các giá trị mẫu thành các lớp và lập bảng phân phối thực nghiệm ghép lớp.

Ví dụ 4.4 : Phân phối thực nghiệm tỉ lệ (%) lợi nhuận của 39 cửa hàng cho trong bảng

Lớp (%)	**Tần số**	**Tần suất**
0 – 10	3	0,0769
10 – 20	16	0,4103
20 – 30	13	0,3333
30 – 40	2	0,0513
40 – 50	3	0,0769
50 – 60	2	0,0513
Tổng	39	1,00

Bảng trên được hiểu như sau : Có bao nhiêu số liệu điều tra về lợi nhuận của các cửa hàng nằm trong lớp, ta ghi số đó làm tần số của lớp. Chẳng hạn, nếu chỉ có 3 cửa hàng trong 39 cửa hàng có số liệu lợi nhuận 5%, 7,5%, 9,5% thuộc lớp 0 – 10, ta ghi 3 là tần số của lớp này.

Bảng trên cho ta biết các tỉ lệ cửa hàng có lợi nhuận theo các mức khác nhau, chẳng hạn : tỉ lệ cửa hàng có lợi nhuận từ 0% đến 10 % là 7,69% ; tỉ lệ cửa hàng có lợi nhuận không quá 20% là 48,72% (0,0769 + 0,4103 = 0,4872). Từ định nghĩa xác suất theo thống kê, khi kích thước mẫu n đủ lớn, các tỉ lệ này được coi là xác suất.

Quy ước : Hai lớp liền nhau $x_{i-1} - x_i$, $x_i - x_{i+1}$ thì x_i thuộc lớp trước $x_{i-1} - x_i$.

Chú ý : Người ta có thể lấy số lớp là k sao cho nó là số nguyên dương nhỏ nhất để $2^k \geq n$ hay $k \geq \log_2(n)$.

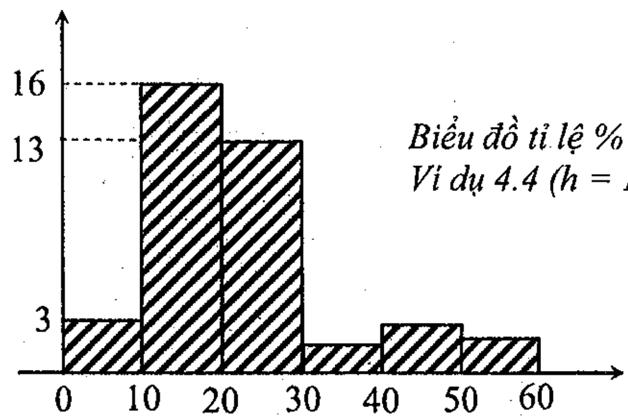
Ở Ví dụ 4.4 : $2^5 = 32 < n = 39 < 2^6$, lấy k = 6 hay $\log_2(39) = 5{,}3$, lấy k = 6.

$$\text{Độ dài mỗi lớp} = \frac{\text{Giá trị mẫu lớn nhất} - \text{giá trị mẫu nhỏ nhất}}{k}.$$

b) Các biểu đồ

Người ta dùng các loại đồ thị để mô tả số liệu mẫu nhằm cho những thông tin sơ bộ ban đầu về đám đông :

- Đa giác tần số (tần suất) là đường gấp khúc gồm các đoạn thẳng nối các điểm (x_i, n_i) (hay các điểm (x_i, f_i)) trên mặt phẳng toạ độ.
- Các biểu đồ tần số (tần suất) là một hình gồm các hình chữ nhật đứng cạnh nhau có đáy là h, chiều cao là $\frac{n_i}{h}$, diện tích của hình chữ nhật là n_i (đáy h, chiều cao $\frac{f_i}{h}$). Ở đây h là chiều dài của một khoảng, nó có được khi ta chia chiều dài khoảng chứa các giá trị quan sát của mẫu thành các đoạn bằng nhau.



Biểu đồ tỉ lệ % lợi nhuận trong Ví dụ 4.4 (h = 1, chia 6 đoạn)

Nhìn vào biểu đồ ta thấy : Tỉ lệ cửa hàng có lợi nhuận từ 10 đến 20 (%) là lớn nhất ; tỉ lệ cửa hàng có lợi nhuận không quá 30% là (3 + 13 + 16)/39 = 0,8205 = 82,05%.

2. Các đặc trưng mẫu

2.1. Thống kê

Định nghĩa 4.4 : Cho mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ của đám đông X. Một hàm của mẫu ngẫu nhiên $T = T(X_1, X_2, ..., X_n)$ được gọi là một thống kê của X.

Các vấn đề của thống kê toán được giải quyết chủ yếu nhờ vào việc xây dựng các thống kê $T(X_1, ..., X_n)$ chỉ phụ thuộc vào mẫu ngẫu nhiên, không phụ thuộc vào tham số.

Ví dụ 4.5 : Xét mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ của đám đông X. Ta có các thống kê :

• Trung bình mẫu : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

• Phương sai mẫu hiệu chỉnh (ta gọi ngắn gọn là phương sai mẫu) :

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

• Nếu biến ngẫu nhiên gốc $X \sim P(\lambda)$, $\lambda > 0$, ta có thống kê :

$$T = P(x_1, x_2,..., x_n) = \prod_{i=1}^{n} P\{X < x_i\} = \frac{e^{-n\lambda}\lambda^{x_1+...+x_n}}{x_1!x_2!...x_n!},\ x_i = 0, 1, 2, ..., n.$$

2.2. Các đặc trưng mẫu

Giả sử đám đông X có $E(X) = \mu$, $V(X) = \sigma^2$. Trường hợp riêng, khi tỉ lệ phần tử đám đông có thuộc tính nghiên cứu là p, người ta mô tả đám đông bằng biến ngẫu nhiên X nhận hai giá trị : X = 1 nếu phần tử có thuộc tính nghiên cứu, X = 0 nếu phần tử không có thuộc tính nghiên cứu và X có bảng phân phối xác suất

X	0	1
P	1 – p	p

Với trường hợp này, ta có kì vọng EX = p (tỉ lệ của đám đông) và mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ (và mẫu cụ thể $(x_1, ..., x_n)$) của X có thành phần X_i (và x_i) chỉ nhận hai giá trị là 0 và 1.

Xét mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ và mẫu cụ thể $(x_1, ..., x_n)$ của đám đông X. Ta có các đặc trưng mẫu của X như sau :

2.2.1. Trung bình mẫu

• Trung bình mẫu của X là $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. (4.1)

• Số đặc trưng của trung bình mẫu :

$$E\left(\overline{X}_n\right) = \mu;\ V\left(\overline{X}_n\right) = \frac{\sigma^2}{n}. \quad (4.2)$$

• Trung bình mẫu cụ thể (mẫu thực nghiệm) :

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (4.3)$$

2.2.2. Tỉ lệ mẫu

• Tỉ lệ mẫu :

$$F_n = \frac{1}{n}\sum_{i=1}^{n} X_i\ ,\ X_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \left(\text{vì } X = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) \tag{4.4}$$

• Tỉ lệ mẫu cụ thể (mẫu thực nghiệm) :

$$f_n = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\ x_i = \begin{bmatrix} 0 \\ 1. \end{bmatrix} \tag{4.5}$$

• Số đặc trưng của tỉ lệ mẫu :

$$E(F_n) = p\ ;\ V(F_n) = \frac{p(1-p)}{n}. \tag{4.6}$$

Ví dụ 4.6 : Mẫu điều tra về sản phẩm đạt tiêu chuẩn của một lô hàng có kích thước $n = 10$ như sau : (0, 1, 1, 1, 0, 1, 1, 0, 1, 0) (sản phẩm đầu tiên không đạt tiêu chuẩn $x_1 = 0$; sản phẩm thứ hai, thứ ba, thứ tư đạt tiêu chuẩn do đó $x_2 = x_3 = x_4 = 1, \ldots, x_{10} = 0$). Tỉ lệ mẫu thực nghiệm : $f_{10} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{6}{10} = 0,6.$

2.2.3. Phương sai mẫu

• Phương sai mẫu :

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]. \tag{4.7}$$

• Kì vọng của phương sai mẫu $E\left(S_n^2\right) = \sigma^2$ (4.8)

• Phương sai mẫu cụ thể : $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2.$ (4.9)

Trong tính toán ta sử dụng công thức :

$$s_n^2 = \frac{n}{n-1}\left(\overline{x}_n^2 - \left(\overline{x}_n\right)^2\right) = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \tag{4.10}$$

với $\overline{x}_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$

• Độ lệch chuẩn của mẫu : $S = \sqrt{S^2}$, $s = \sqrt{s^2}$.

Ngoài ra ta còn có :

• Phương sai mẫu chưa hiệu chỉnh : $\hat{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$, và phương sai mẫu chu

hiệu chỉnh thực nghiệm

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2. \qquad (4.11)$$

• Kì vọng của phương sai mẫu chưa hiệu chỉnh

$$E\left(\hat{S}_n^2\right) = \frac{n-1}{n}\sigma^2; \ (E\left(\hat{S}_n^2\right) \neq \sigma^2). \qquad (4.12)$$

Chú ý : Với một mẫu có kích thước n cho trước, nếu không gây hiểu lầm, ta có th viết các đặc trưng mẫu ngắn gọn như sau : Trung bình mẫu $\overline{X}, \overline{x}$. Tỉ lệ mẫu F, f Phương sai mẫu S^2, s^2.

2.3. Cách tính các đặc trưng mẫu cụ thể

a) Bài toán : Từ một mẫu cụ thể $(x_1, x_2, ..., x_n)$ có bảng phân phối thực nghiệm ha phân phối ghép lớp, tính $\overline{x}$, s^2.

• Cách tính tương tự như các ví dụ sau.

Ví dụ 4.7 : Cho bảng phân phối thực nghiệm

x_i	−2	1	2	3	4	5
n_i	2	1	2	2	2	1

Tính trung bình, phương sai, độ lệch tiêu chuẩn của mẫu.

GIẢI : Ta lập bảng

x_i	n_i	$n_i x_i$	$n_i x_i^2$
−2	2	−4	8
1	1	1	1
2	2	4	8
3	2	6	18
4	2	8	32
5	1	5	25
Tổng	n = 10	$\sum n_i x_i = 20$	$\sum n_i x_i^2 = 92$

• $\overline{x} = \dfrac{\sum n_i x_i}{n} = \dfrac{20}{10} = 2$; $\overline{x^2} = \dfrac{\sum n_i x_i^2}{n} = \dfrac{92}{10} = 9{,}2$;

• $s^2 = \dfrac{n}{n-1}\left(\overline{x^2} - (\overline{x})^2\right) = \dfrac{10}{9}\left(9{,}2 - (2)^2\right) \approx 5{,}7778$;

• $s = \sqrt{s^2} = 2{,}4037$.

Ví dụ 4.8 : Lượng xăng hao phí của một ôtô đi từ A đến B sau 30 lần chạy, kết quả cho trong bảng :

Lượng xăng hao phí (lít)	9,6 – 9,8	9,8 – 10	10 – 10,2	10,2 – 10,4	10,4 – 10,6
Số lần tương ứng	3	5	10	8	4

Tính $\overline{x}, s^2$.

Với bảng phân phối ghép lớp ta thay lớp $x_{i-1} - x_i$ bằng một đại diện $x_i' = \dfrac{x_{i-1} + x_i}{2}$ (trung bình cộng của hai đầu mút), đưa về bảng phân phối thông thường để tính $\overline{x}, s^2$.

GIẢI : Lập bảng

Lớp	x_i	n_i	$n_i x_i$	$n_i x_i^2$
9,6 – 9,8	9,7	3	29,1	282,27
9,8 – 10,0	9,9	5	49,5	490,05
10,0 – 10,2	10,1	10	101	1020,1
10,2 – 10,4	10,3	8	82,4	848,72
10,4 – 10,6	10,5	4	42	441
Tổng		$n = 30$	$\sum n_i x_i = 304$	$\sum n_i x_i^2 = 3082{,}14$

$$\overline{x} = \frac{304}{30} = 10{,}1333 \text{ ; } \overline{x^2} = \frac{3082{,}14}{30} = 102{,}738 \text{ ;}$$

$$s^2 = \frac{30}{29}\left(102{,}738 - (10{,}1333)^2\right) = 0{,}0561 \text{ và } s = 0{,}2369.$$

b) Mẫu cụ thể 2-chiều.

Tương tự như trường hợp một chiều, ta có mẫu ngẫu nhiên 2-chiều kích thước n là n biến ngẫu nhiên độc lập (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) được lập từ biến ngẫu nhiên 2-chiều (X, Y) và cùng phân phối xác suất của (X, Y).

Mẫu cụ thể kích thước n là các cặp

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \ ((X_i = x_i \ ; \ Y_i = y_i) \ ; \ i = \overline{1, n}).$$

Người ta cũng mô tả bằng bảng phân phối tần suất thực nghiệm như sau :

X \ Y	y_1	y_2	...	y_j	...	y_k	n_i
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_1
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_2
$\vdots$	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_i
$\vdots$	...	...	...	...	...	...	...
x_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	n_h
m_j	m_1	m_2	...	m_j	...	m_k	$\sum = n$

Trong đó : n_{ij} là tần số xuất hiện (x_i, y_j) trong mẫu.

n_i, m_j tương ứng là tần số xuất hiện x_i, y_j.

Từ bảng 2-chiều ta có thể lập các bảng tần số thực nghiệm của X, của Y (tương tự như bảng phân phối 2-chiều [2.19, §3, Ch.2]).

Ví dụ 4.9 : Điều tra về hai chỉ tiêu chiều dài X (cm) và hàm lượng chất Y (%) các sản phẩm của một xí nghiệp cho kết quả :

X \ Y	8	10	12	14	16
100	5	5			
110	4	6	7		
120		5	9	8	
130			4	6	9
140				5	7

Từ bảng trên, cộng theo dòng, ta được bảng phân phối tần số thực nghiệm của X :

X	100	110	120	130	140
n_i	10	17	22	19	12

và cộng theo cột, ta được bảng phân phối tần số thực nghiệm của Y :

Y	8	10	12	14	16
m_j	9	16	20	19	16

...c sản phẩm có chỉ tiêu $X \leq 110$ cm và $Y \leq 12\%$ là loại II. Ta có bảng phân phối ...ực nghiệm của chỉ tiêu Y các sản phẩm loại II :

Y	8	10	12
m_j	9	11	7

và bảng phân phối thực nghiệm của chỉ tiêu X các sản phẩm loại II :

X	100	110
n_i	10	17

§2. ƯỚC LƯỢNG ĐIỂM

1. Mô tả bài toán

Giả sử đám đông ***C*** được mô tả bằng biến ngẫu nhiên X có dạng phân phối xác suất $P_X(\theta)$ đã biết, nhưng phụ thuộc vào một hay một vài tham số θ chưa biết, chẳng hạn X có phân phối Poisson $P(\lambda)$, $\theta = \lambda$; hay X có phân phối chuẩn $N(\mu ; \sigma^2)$, $\theta = (\mu, \sigma^2)$. Phân phối xác suất của X được xác định nếu ta tìm được hay ước lượng được giá trị của θ.

Cho mẫu ngẫu nhiên $(X_1, X_2,\ldots,X_n)$ của X. Mỗi giá trị $\hat{\theta} = \hat{\theta}(x_1,\ldots, x_n)$ của thống kê $\Theta = \Theta(X_1, X_2,\ldots,X_n)$ được gọi là một *ước lượng điểm* của tham số θ của đám đông X.

Bài toán đặt ra là tìm một thống kê Θ để mỗi giá trị $\hat{\theta}$ của nó là ước lượng điểm (hay thay thế) tham số chưa biết θ một cách "đủ tốt". Do giá trị đúng của θ chưa biết nên không thể so sánh $\hat{\theta}$ với θ để đánh giá chất lượng (ít sai số) của $\hat{\theta}$, vì vậy người ta đưa ra các tiêu chuẩn ước lượng điểm theo các phương pháp sau.

2. Các phương pháp ước lượng điểm

2.1. Phương pháp sử dụng các đặc trưng mẫu

Để ước lượng tham số chưa biết, người ta tìm một thống kê Θ sao cho mỗi giá trị $\hat{\theta}$ của nó là ước lượng điểm của tham số chưa biết θ một cách "đủ tốt" theo một trong các tiêu chuẩn sau.

2.1.1. Các tiêu chuẩn ước lượng điểm

a) Ước lượng không chệch

Định nghĩa 4.5 : Thống kê Θ được gọi là ước lượng không chệch của θ nếu $E\Theta = \theta$ (4.13)

Ý nghĩa : Từ (4.13) ta có $E(\Theta - \theta) = 0$ (trung bình của độ lệch (sai số) giữa ước lượng với giá trị thật bằng 0). Sai số trung bình bằng 0 được gọi là *sai số ngẫu nhiên*, ngược lại gọi là *sai số hệ thống*. Như vậy Θ là ước lượng không chệch của θ khi sai số ước lượng là sai số ngẫu nhiên.

Cho đám đông X có các tham số $\mu = E(X)$ (hay $p = E(X)$ trong trường hợp tỉ lệ), $\sigma^2 = V(X)$.

Ví dụ 4.10 : Tỉ lệ mẫu F, trung bình mẫu $\overline{X}$, phương sai mẫu S^2 tương ứng là ước lượng không chệch của p, μ, σ^2. Còn phương sai mẫu chưa hiệu chỉnh $\hat{S}^2$ là ước lượng bị chệch của σ^2.

Thật vậy : $E\overline{X} = E\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\sum_{i=1}^{n} EX_i}{n} = \frac{n\mu}{n} = \mu$, và khi $EX = p$ thì $EX = p$.

Ta có : • $\sum_{i=1}^{n}(X_i - \mu)(\overline{X} - E\overline{X}) = \left(n\frac{\sum_{i=1}^{n}(X_i - \mu)}{n} \right)(\overline{X} - E\overline{X})$

$$= n(\overline{X} - \mu)(\overline{X} - \mu) = n(\overline{X} - \mu)^2.$$

• $\sum(X_i - \overline{X})^2 = \sum\left[(X_i - \mu) - (\overline{X} - \mu)\right]^2$

$$= \sum(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum(X_i - \mu) + n(\overline{X} - \mu)^2$$
$$= \sum(X_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

Từ đó $E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i - \overline{X})^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2\right]$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Nhưng $E\left(\widehat{S}^2\right) = \frac{n-1}{n}\left[\frac{\sum_{i=1}^{n} E\left(X_i - \overline{X}\right)^2}{n-1}\right] = \frac{n-1}{n} E\left(S^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$.

b) Ước lượng vững

Định nghĩa 4.6 : Thống kê Θ được gọi là ước lượng vững của θ nếu Θ hội tụ theo xác suất về θ (hay $\Theta(X_1, ..., X_n) \xrightarrow{p} \theta$).

Ý nghĩa thực hành : Trên tập có xác suất gần 1, với n đủ lớn ta có : $\hat{\theta} \approx \theta$.

Từ luật số lớn [Định lí 2.3, §5, Ch.2] cho ta kết quả sau :

Ví dụ 4.11 : F, $\overline{X}$, S^2 tương ứng là ước lượng vững cho p, μ, σ^2 và $\hat{S}^2$ cũng là ước lượng vững cho σ^2.

c) Ước lượng hiệu quả

Định nghĩa 4.7 : Thống kê Θ được gọi là ước lượng hiệu quả của θ nếu nó là ước lượng không chệch và có phương sai bé nhất trong các ước lượng không chệch của θ.

Nếu hàm mật độ xác suất $f(x, \theta)$ của X thoả mãn một số điều kiện nhất định thì ta có bất đẳng thức Cramer-Rao :

$$V(\theta^*) \geq \frac{1}{nE\left(\frac{\partial \ln f(X,\theta)}{\partial\theta}\right)^2} \quad \text{với mọi } \theta^* : E\theta^* = \theta,$$

và $\hat{\theta}$ là ước lượng hiệu quả của θ khi :

$$V\left(\hat{\theta}\right) = \frac{1}{nE\left(\frac{\partial \ln f(X,\theta)}{\partial\theta}\right)^2} \tag{4.14}$$

Ví dụ 4.12 : Nếu $X \sim N(\mu, \sigma^2)$ thì $\overline{X}$ là ước lượng hiệu quả của μ.

GIẢI : Ta có $E\left(\overline{X}\right) = \mu$, $V\left(\overline{X}\right) = \frac{\sigma^2}{n}$. Đặt $\theta = (\mu, \sigma^2)$ và

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

$$\frac{\partial \ln f(x,\theta)}{\partial\mu} = \frac{\partial}{\partial\mu}\left[-\frac{(x-\mu)^2}{2\sigma^2} + \ln\frac{1}{\sigma\sqrt{2\pi}}\right] = \frac{x-\mu}{\sigma^2}.$$

Vậy $nE\left[\dfrac{\partial \ln f(X,\theta)}{\partial \mu}\right]^2 = nE\left(\dfrac{X-\mu}{\sigma^2}\right)^2 = \dfrac{n}{\sigma^4}E(X-\mu)^2 = \dfrac{nVX}{\sigma^4} = \dfrac{n}{\sigma^2}$.

Từ đó : $V(\overline{X}) = \dfrac{\sigma^2}{n} = \dfrac{1}{nE\left(\dfrac{\partial \ln f(X,\theta)}{\partial \mu}\right)}$, theo bất đẳng thức Cramer-Rao, và t

(4.14) suy ra $\overline{X}$ là ước lượng hiệu quả của μ.

Ví dụ 4.13 : Nếu $X \sim B(1, p)$ thì tỉ lệ mẫu F là ước lượng hiệu quả của p.

GIẢI : Ta có $E(F) = p$ và $V(F) = p(1-p)/n$;

$f(x, \theta) = P(X = x) = p^x(1-p)^x$, $\ln f(x, \theta) = x\ln p + (1-x)\ln(1-p)$

$$\Rightarrow \frac{\partial \ln f(x,\theta)}{\partial \theta} = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)}$$

$$\Rightarrow nE\left(\frac{\partial \ln f(x,\theta)}{\partial \theta}\right)^2 = nE\left(\frac{x-p}{p(1-p)}\right)^2 = \frac{n}{p^2(1-p)^2}E(x-p)^2$$

$$= \frac{n}{p^2(1-p)^2}V(X) = \frac{n}{p(1-p)}.$$

Theo đẳng thức (4.14), F là ước lượng hiệu quả của p.

2.1.2. Ước lượng điểm bằng các đặc trưng mẫu.

Cho đám đông X có kì vọng $E(X) = \mu$ hay tỉ lệ $p = E(X)$, phương sai $V(X) = \sigma^2$.

a) Ước lượng điểm của kì vọng

Mệnh đề 4.1 : Trung bình mẫu $\overline{X}$ là ước lượng không chệch và vững của μ. Nế đám đông X có phân phối chuẩn thì $\overline{X}$ là ước lượng hiệu quả của μ.

b) Ước lượng điểm của tỉ lệ

Mệnh đề 4.2 : Tỉ lệ mẫu F là ước lượng không chệch, vững và hiệu quả của tỉ lệ p.

c) Ước lượng điểm của phương sai

Mệnh đề 4.3 : Phương sai mẫu là ước lượng không chệch, vững cho phương sai σ Phương sai mẫu chưa hiệu chỉnh là ước lượng vững, bị chệch cho phương sai σ^2.

2.1.3. Quy tắc thực hành

Khi có mẫu điều tra $(x_1, ..., x_n)$ của đám đông X, ta lấy các ước lượng điểm của l vọng μ, tỉ lệ p, phương sai σ^2 của X như sau :

Trung bình mẫu $\overline{x}$ làm ước lượng điểm cho trung bình μ.

– Tỉ lệ mẫu f làm ước lượng điểm cho tỉ lệ p.

– Phương sai mẫu s^2 làm ước lượng điểm cho σ^2.

Ví dụ 4.14 : Mẫu điều tra về trọng lượng của một loại trái cây (đơn vị : gam) của một vùng cho kết quả trong bảng :

x_i (gam)	0 – 5	5 –10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
Số trái (n_i)	3	7	12	23	30	15	10

a) Những trái cây có trọng lượng nhỏ hơn 10 gam là trái cây loại B. Hãy ước lượng tỉ lệ trái cây loại B của vùng.

b) Hãy ước lượng trọng lượng trung bình của trái cây và phương sai của trọng lượng trái cây của vùng.

GIẢI : Gọi p, μ, σ^2 tương ứng là tỉ lệ trái cây loại B, trung bình trọng lượng trái cây, phương sai trọng lượng trái cây của vùng. Ta có các ước lượng điểm của :

a) Tỉ lệ trái cây loại B là : p = (3 + 7)/100 = 0,1 = 10%.

b) Trọng lượng trung bình của trái cây của vùng là : $\mu = \overline{x} = 19{,}902$ (gam). Phương sai của trọng lượng trái cây của vùng : $\sigma^2 = 59{,}7675$.

2.2. Phương pháp hợp lí tối đa

Cho đám đông X có hàm mật độ xác suất (trường hợp liên tục) hay phân phối xác suất (trường hợp rời rạc) là f(x, θ) ; θ là tham số của phân phối này. Ta lập hàm sau (còn được gọi là hàm hợp lí)

$$L(x_1, x_2,..., x_n, \theta) = f(x_1,\theta).f(x_2,\theta)...f(x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta).$$

Ta coi $L(x_1, x_2,..., x_n, \theta)$ là hàm của θ. Chẳng hạn, trong trường hợp rời rạc, hàm hợp lí

$L(x_1, x_2,..., x_n,\theta) = P\{X_1 = x_1, X_2 = x_2,.., X_n = x_n|\theta\} = \prod_{i=1}^{n} f(x_i,\theta)$ là xác suất của biến cố "Mẫu ngẫu nhiên của X nhận giá trị $(x_1, x_2,..., x_n)$". Ta tìm ước lượng $\hat{\theta}$ của θ sao cho khi $(x_1, x_2,..., x_n)$ cố định, xác suất này là lớn nhất (nghĩa là nhiều khả năng nhận giá trị mẫu này nhất). Ta đưa vào định nghĩa ước lượng hợp lí tối đa như sau :

Định nghĩa 4.8 : Cho mẫu cụ thể $(x_1, x_2,..., x_n)$ của đám đông X. X có hàm mật độ xác suất (trường hợp liên tục) hay phân phối xác suất (trường hợp rời rạc) là f(x, θ). Giá trị $\hat{\theta}$ được gọi là ước lượng hợp lí tối đa (hay hợp lí cực đại) của θ nếu mọi θ (thuộc một tập tham số Θ nào đó) :

$$L(x_1, x_2,..., x_n, \hat{\theta}) = \max_{\theta\in\Theta} L(x_1, x_2,..., x_n,\theta).$$

Đặt $x = (x_1, x_2,..., x_n)$, nếu hàm hợp lí $L(x, \theta)$ khả vi đến cấp 2 và do hàm $L(x, \theta)$, $\ln L(x, \theta)$ cùng đạt cực trị tại một giá trị nên ta có thể giải phương trình sau (gọi là phương trình hợp lí) :

$\dfrac{\partial \ln L(x,\theta)}{\partial\theta} = 0$, tìm nghiệm và thử điều kiện $\dfrac{\partial^2 \ln L(x,\theta)}{\partial\theta^2} < 0$ để có cực đại.

Người ta chứng minh được nếu phương trình hợp lí có nghiệm duy nhất thì hàm hợp lí đạt cực đại tại đó.

Ví dụ 4.15 : Cho X ~ B(1, p), tìm ước lượng hợp lí tối đa cho p.

GIẢI : Ta có X ~ B(1, p) thì $p(x_i, p) = p^{x_i}(1-p)^{1-x_i}$, $x \in \{0 ; 1\}$ và

$$L(x, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i},\ \ln L(x,p) = \sum_{i=1}^{n}\left[x_i \ln p + (1-x_i)\ln(1-p)\right];$$

$$\frac{d\ln L}{dp} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-x_i) = 0 \Leftrightarrow p = \overline{x} = f \text{ (tỉ lệ mẫu).}$$

Với chú ý $\sum_{i=1}^{n} x_i = n\overline{x}$, ta có

$$\left.\frac{d^2 \ln L}{dp^2}\right|_{p=\overline{x}} = \left.\left(-\frac{1}{p^2}\sum_{i=1}^{n} x_i - \frac{1}{(1-p)}\sum_{i=1}^{n}(x_i - 1)\right)\right|_{p=\overline{x}}$$

$$= -n\left[\frac{1}{\overline{x}(1-\overline{x})}\right] < 0.$$

Tại $\overline{x} = f$ (tỉ lệ mẫu) hàm $\ln L(x, p)$ đạt cực đại. Vậy f là ước lượng hợp lí tối đa cho p.

Hoàn toàn tương tự (coi như một bài tập), ta có : Nếu X có phân phối Poisson với tham số λ (X ~ P(λ)) thì $\overline{x}$ là ước lượng hợp lí tối đa cho λ .

Ví dụ 4.16 : Nếu $X \sim N(\mu, \sigma^2)$, tìm ước lượng hợp lí tối đa cho $\theta = (\mu, \sigma^2)$.

GIẢI : Hàm hợp lí $L(x, \theta) = \dfrac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$.

Ta có : $\begin{cases} \dfrac{\partial \ln L(x,\theta)}{\partial \mu} = \dfrac{n\overline{x} - n\mu}{\sigma^2} = 0 \\ \\ \dfrac{\partial \ln L(x,\theta)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum\limits_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4} = 0. \end{cases}$

Giải hệ phương trình trên ta có nghiệm $\hat{\mu} = \overline{x}$ và $\hat{\sigma}^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2 = \hat{s}^2$.

Nghiệm này là điểm cực đại của hàm $\ln L(x, \theta)$ và là điểm cực đại của hàm $L(x, \theta)$.

Vậy ước lượng hợp lí tối đa của μ là trung bình mẫu $\overline{x}$ và ước lượng hợp lí tối đa của σ^2 là phương sai mẫu chưa hiệu chỉnh $\hat{s}^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$.

§3. ƯỚC LƯỢNG KHOẢNG

1. Bài toán ước lượng khoảng

Bài toán : Cho xác suất $1 - \alpha$. Từ mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ của đám đông X có phân phối xác suất $f(x, \theta)$ (ở đây tham số θ chưa biết) tìm các thống kê θ_1, θ_2 sao cho tham số θ nằm trong khoảng (θ_1, θ_2) với xác suất $1 - \alpha$, nghĩa là :

$$P\{\theta_1 < \theta < \theta_2\} = 1 - \alpha. \qquad (4.15)$$

Ta gọi :

- $1 - \alpha$: độ tin cậy của ước lượng,
- (θ_1, θ_2) : khoảng tin cậy của ước lượng,
- $\theta_2 - \theta_1 = 2\varepsilon$ (độ dài khoảng tin cậy) thì ε được gọi là độ chính xác hay sai số của ước lượng.

Bài toán ước lượng khoảng với độ tin cậy $1 - \alpha$ còn được gọi là bài toán tìm khoảng tin cậy $1 - \alpha$.

Các bước giải bài toán tìm khoảng tin cậy :

1) Tìm một thống kê $G = G(X_1, X_2, ..., X_n)$ sao cho phân phối xác suất của nó được xác định, không phụ thuộc vào tham số.

2) Cho độ tin cậy $1 - \alpha$, lấy các cặp số α_1, α_2 đều không âm : $\alpha_1 + \alpha_2 = \alpha$. Tìm các giá trị tới hạn g_{α_2}, $g_{1-\alpha_1}$ của G thoả mãn :

$$P\left\{G > g_{\alpha_2}\right\} = \alpha_2 ; P\left\{G > g_{1-\alpha_1}\right\} = 1 - \alpha_1 \tag{4.16}$$

Khi đó : $P\left\{g_{1-\alpha_1} < G < g_{\alpha_2}\right\} = P\left\{G < g_{\alpha_2}\right\} - P\left\{G < g_{1-\alpha_1}\right\} = 1 - \alpha_2 - \alpha_1 = 1 - \alpha.$

3) Biến đổi tương đương biến cố :

$$\left\{g_{1-\alpha_1} < G < g_{\alpha_2}\right\} = \left\{\theta_1 < \theta < \theta_2\right\}. \tag{4.17}$$

Vì vậy $P\left\{\theta_1 < \theta < \theta_2\right\} = 1 - \alpha.$

Chú ý : 1) Có vô số cách chọn $\alpha_1, \alpha_2 : \alpha_1 + \alpha_2 = \alpha$ nên người ta thường tìm khoảng tin cậy với $\alpha_1 = \alpha_2 = \dfrac{\alpha}{2}$.

2) Vì xác suất $1 - \alpha$ lớn, theo nguyên lí xác suất lớn [3, §2, Ch. 1], suy ra trong một phép thử $(X_1, X_2, ..., X_n)$ thì $\theta_1(X_1, ..., X_n) < \theta < \theta_2(X_1, ..., X_n)$ được coi là xảy ra. Nghĩa là với xác suất $(1 - \alpha)\%$, từ mẫu cụ thể $(x_1, x_2, ..., x_n)$, θ nằm giữa hai số :

$$\theta_1(x_1, x_2, ..., x_n) < \theta < \theta_2(x_1, x_2, ..., x_n).$$

2. Khoảng tin cậy cho các tham số

2.1. Khoảng tin cậy cho tỉ lệ

2.1.1. Bài toán tìm khoảng tin cậy đối xứng

a) Bài toán tìm khoảng tin cậy đối xứng của tỉ lệ : Giả sử ta cần tìm tỉ lệ của các phần tử có tính chất A trong các phần tử của một đám đông. Ta kí hiệu tỉ lệ này là p và mô tả đám đông này bằng biến ngẫu nhiên X nhận 2 giá trị 0 và 1 ; X = 0 khi phần tử không có tính chất A, X = 1 khi phần tử có tính chất A và $P(X = 1) = p$.

Trong mục này ta cần giải quyết bài toán : Cho độ tin cậy $1 - \alpha$, tìm khoảng tin cậy của tỉ lệ p.

Để tìm khoảng tin cậy cho p, ta sử dụng thống kê trong định lí sau (suy ra từ Định lí giới hạn trung tâm [Định lí 3.2, §1, Ch.3]).

Định lí 4.1 : Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên của đám đông X có tỉ lệ p và F là tỉ lệ mẫu của X. Khi đó thống kê

$$G = \frac{F - p}{\sqrt{p(1-p)/n}}$$ hội tụ theo phân phối về $Z \sim N(0, 1)$.

Từ Định lí 4.1, với n đủ lớn $G = \dfrac{F-p}{\sqrt{p(1-p)/n}} \approx N(0, 1)$ và với độ tin cậy $1 - \alpha$, ta tìm các giá trị tới hạn chuẩn $z_{\frac{\alpha}{2}}, -z_{\frac{\alpha}{2}}$:

$$P\left\{\frac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}} > z_{\frac{\alpha}{2}}\right\} = \frac{\alpha}{2}; \ P\left\{\frac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}} > -z_{\frac{\alpha}{2}}\right\} = 1 - \frac{\alpha}{2}$$

$$\Rightarrow \quad P\left\{-z_{\frac{\alpha}{2}} < \frac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right\} = 1 - \alpha \qquad (4.18)$$

Trong (4.18), khi n đủ lớn thay p (ở mẫu số) bằng ước lượng điểm của p là F, ta có :

$$P\left\{F - z_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)}{n}} < p < F + z_{\frac{\alpha}{2}}\sqrt{\frac{F(1-F)}{n}}\right\} = 1 - \alpha. \qquad (4.19)$$

Từ (4.19), trong thực hành, trên mẫu cụ thể $(x_1, x_2, \ldots, x_n)$ khi $nf \geq 10$ và $n(1-f) \geq 10$ ta có khoảng tin cậy đối xứng cho p :

$$f - z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < p < f + z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \qquad (4.20)$$

Quy tắc thực hành. Tìm khoảng tin cậy đối xứng cho p theo các bước :

• Tính $f = \dfrac{m}{n}$ (n là kích thước mẫu, m là số phần tử trong mẫu có tính chất A đang nghiên cứu)

• Tìm $\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}$

• Khoảng tin cậy của tỉ lệ p : $(f - \varepsilon;\ f + \varepsilon)$.

Ở đây $z_{\frac{\alpha}{2}}$ là giá trị tới hạn chuẩn mức $\dfrac{\alpha}{2}$ tra bảng hàm Laplace [Bảng 2, Phụ lục].

Chú ý : 1) Để đơn giản, nếu không gây ra sự hiểu lầm, ta gọi "khoảng tin cậy" thay cho "khoảng tin cậy đối xứng".

2) Hai bài toán liên quan đến bài toán ước lượng tỉ lệ :

Bài toán 1 : Cho độ tin cậy $1-\alpha$ và số phần tử của đám đông là N. Hãy ước lượng số phần tử của đám đông có tính chất A, kí hiệu N_A.

GIẢI : Ta có $p = N_A/N$

+ Bước 1. Tìm khoảng tin cậy của p với độ tin cậy $1-\alpha$:

$$\left(f - z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < p = \frac{N_A}{N} < f + z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right) \Leftrightarrow p \in (f_1, f_2) \quad (4.21)$$

Từ (4.21) nhân 2 vế với N, ta suy ra ước lượng khoảng của N_A.

+ Bước 2. Ước lượng khoảng của N_A : $(N.f_1 , N.f_2)$.

Bài toán 2 : Cho độ tin cậy $1-\alpha$ và số phần tử N_A của đám đông có tính chất A. Hãy ước lượng số phần tử N của đám đông.

GIẢI : Ta có $p = N_A/N$.

+ Bước 1. Tìm khoảng tin cậy của p với độ tin cậy $1-\alpha$:

$$\left(f - z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < p = \frac{N_A}{N} < f + z_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right) \Leftrightarrow p \in (f_1, f_2) \quad (4.22)$$

Từ (4.22) ta suy ra :

+ Bước 2. Ước lượng số phần tử của đám đông N : $\left(\frac{N_A}{f_2}, \frac{N_A}{f_1}\right)$.

3) Để có khoảng tin cậy (4.20), kích thước mẫu n cần đủ lớn ($n \geq 100$) .

Thật vậy, từ (4.18), trên mẫu cụ thể $(x_1, x_2, ..., x_n)$, ta có $F = f$:

$$\left\{-z_{\frac{\alpha}{2}} < \frac{f-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right\} = \left\{(f-p)^2 < \frac{z^2_{\frac{\alpha}{2}}p(1-p)}{n}\right\} \quad (4.23)$$

$$= \left\{\left(n + z^2_{\frac{\alpha}{2}}\right)p^2 - (2nf + z^2_{\frac{\alpha}{2}})p + nf^2 < 0\right\}.$$

Giải bất phương trình bậc 2 trong (4.23) ta tìm được khoảng tin cậy của p là (f_1, f_2) với

$$f_1, f_2 = \frac{f + \dfrac{z_{\frac{\alpha}{2}}^2}{2n} \pm z_{\frac{\alpha}{2}} \sqrt{\dfrac{f(1-f)}{n} + \dfrac{z_{\frac{\alpha}{2}}^2}{4n^2}}}{1 + \left(z_{\frac{\alpha}{2}}^2 \right) / n}. \tag{4.24}$$

Khi n đủ lớn từ (4.24) ta lấy xấp xỉ sẽ được (4.20).

Ví dụ 4.17 : Để đánh giá chất lượng sản phẩm của một nhà máy, người ta kiểm tra ngẫu nhiên 600 sản phẩm của nhà máy này sau ca sản xuất, thấy có 50 phế phẩm. Với độ tin cậy 95%.

1) Hãy ước lượng tỉ lệ phế phẩm của nhà máy sau ca sản xuất.

2) Hãy ước lượng số phế phẩm của nhà máy sau ca sản xuất ; biết rằng mỗi ca nhà máy sản xuất được 5500 sản phẩm.

GIẢI : 1) Gọi p là tỉ lệ phế phẩm mỗi ca sản xuất của nhà máy

Ta có : $f = \dfrac{50}{600} = 0{,}0833$.

- $1 - \alpha = 0{,}95 \Rightarrow 0{,}5 - \dfrac{\alpha}{2} = 0{,}475 = \varphi(1{,}96) \Rightarrow z_{\alpha/2} = 1{,}96$ [Bảng 2]).
- $\varepsilon = 1{,}96\sqrt{\dfrac{0{,}0833(1-0{,}0833)}{600}} = 0{,}0221$.
- Khoảng tin cậy của tỉ lệ phế phẩm p :

$$(f - \varepsilon\ ;\ f + \varepsilon) = (0{,}0612\ ;\ 0{,}1054).$$

Kết luận : Với độ tin cậy 95%, tỉ lệ phế phẩm của mỗi ca sản xuất của nhà máy từ 6,12% đến 10,54%.

2) Gọi M là số phế phẩm của một ca sản xuất, ta có $p = M/5500$. Vì $0{,}0612 < p < 0{,}1054$ suy ra : $0{,}0612 < M/5500 < 0{,}1054$ hay $5500.0{,}0612 < M < 5500.0{,}1054$. Vậy $336{,}6 < M < 579{,}7$.

Kết luận : Với độ tin cậy 95%, số phế phẩm của mỗi ca sản xuất của nhà máy từ 337 đến 579 sản phẩm.

b) Bài toán về chỉ tiêu của ước lượng khoảng đối xứng.

Có ba chỉ tiêu của bài toán khoảng tin cậy đối xứng là : độ chính xác (hay còn gọi là sai số) của ước lượng, độ tin cậy của ước lượng, kích thước mẫu điều tra. Biết 2 trong 3 chỉ tiêu sẽ tìm được chỉ tiêu còn lại. Ta có 3 bài toán suy từ công thức sau

$$\varepsilon = z_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \tag{4.25}$$

Bài toán 1. Cho độ tin cậy $1-\alpha$ và mẫu điều tra (cho n) tìm độ chính xác ε của ước lượng :

Rõ ràng $\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}$. (4.26)

Bài toán 2. Cho độ chính xác ε của ước lượng và mẫu điều tra (cho n) tìm độ tin cậy $1-\alpha$.

Từ (4.25) : $z_{\frac{\alpha}{2}} = \varepsilon\sqrt{\dfrac{n}{f(1-f)}} \xrightarrow{\text{(Bảng 2)}} \varphi(z_{\frac{\alpha}{2}})$.

Độ tin cậy cần tìm $1-\alpha = 2\varphi(z_{\frac{\alpha}{2}})$. (4.27)

Bài toán 3. Cho độ tin cậy $1-\alpha$ và độ chính xác ε tìm :

– Kích thước mẫu tối thiểu cần điều tra m :

$$m \geq \left(f(1-f)\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2\right). \quad (4.28)$$

Nếu $\left(f(1-f)\left(\dfrac{z_{\alpha/2}}{\varepsilon}\right)^2\right)$ là số thập phân ta lấy

$$m = \left[\left(f(1-f)\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2\right)\right] + 1. \quad (4.29)$$

Ở đây kí hiệu [ ] là lấy phần nguyên, ví dụ [152,789] = 152.

– *Kích thước mẫu cần điều tra thêm* : $K = m - n$ (trong trường hợp $m > n$ với n là kích thước mẫu hiện có).

Ví dụ 4.18 : Sử dụng Ví dụ 4.17.

a) Nếu sử dụng mẫu điều tra để ước lượng tỉ lệ phế phẩm trong một ca sản xuất của nhà máy đạt độ chính xác là 2% thì đảm bảo độ tin cậy là bao nhiêu ?

b) Nếu muốn ước lượng tỉ lệ phế phẩm của nhà máy trong một ca sản xuất đạt độ tin cậy là 96% và độ chính xác là 1,8% thì cần điều tra thêm ít nhất bao nhiêu sản phẩm nữa.

GIẢI : a) Ta có $f = \dfrac{50}{600} = 0{,}0833$; $\varepsilon = 0{,}02 = z_{\frac{\alpha}{2}}\sqrt{\dfrac{0{,}0833(1-0{,}0833)}{600}} \Rightarrow z_{\frac{\alpha}{2}} = 1{,}77$.

Từ đó $1-\alpha = 2.\varphi(1{,}77) = 2.0{,}46164 = 0{,}92328$.

b) Ta có $\varepsilon = 0{,}018$ và $1-\alpha = 0{,}96 \Rightarrow z_{\frac{\alpha}{2}} = 2{,}06$.

Kích thước mẫu tối thiểu cần điều tra :

$$m \geq \left(0,0833(1-0,0833)\left(\frac{2,06}{0,018}\right)^2\right).$$

Vậy m = 1001. Kích thước mẫu đã có n = 600, vậy cần điều tra thêm ít nhất K = 1001 – 600 = 401 sản phẩm nữa.

2.1.2. Khoảng tin cậy một phía

Trong thực tế có yêu cầu ước lượng tỉ lệ p một phía với độ tin cậy $1-\alpha$ như sau :

• *Ước lượng tối đa cho* p (phía trái) : $p < f + z_\alpha\sqrt{\frac{f(1-f)}{n}}$ (4.30)

• *Ước lượng tối thiểu cho* p (phía phải) : $f - z_\alpha\sqrt{\frac{f(1-f)}{n}} < p$ (4.31)

Ở đây z_α là giá trị tới hạn chuẩn mức α.

Với chú ý khi n đủ lớn p = F thì (4.30), (4.31) được suy ra như sau :

$$1-\alpha = P\left\{\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} > -z_\alpha\right\} \approx P\left\{p < F + z_\alpha\sqrt{\frac{F(1-F)}{n}}\right\}. \quad (4.32)$$

Từ (4.31) suy ra (4.29).

$$P\left\{\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} > z_\alpha\right\} = \alpha \Leftrightarrow$$

$$1-\alpha = P\left\{\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} < z_\alpha\right\} \approx P\left\{p > F - z_\alpha\sqrt{\frac{F(1-F)}{n}}\right\} \quad (4.33)$$

Từ (4.33) ta suy ra (4.31).

Ví dụ 4.19 : Sử dụng Ví dụ 4.17. Với độ tin cậy 95% hãy ước lượng tỉ lệ phế phẩm tối đa của một ca sản xuất của nhà máy.

GIẢI : Gọi p là tỉ lệ phế phẩm của một ca sản xuất của nhà máy.

$$f = \frac{50}{600} = 0,0833 \ ; \ 1-\alpha = 0,95 : \ z_\alpha = 1,65.$$

Khoảng tin cậy phía trái : $p < 0{,}0833 + 1{,}65\sqrt{\dfrac{0{,}0833\cdot(1-0{,}0833)}{600}} = 0{,}1019.$

Với độ tin cậy 95% tỉ lệ phế phẩm tối đa một ca sản xuất của nhà máy là 10,19%.

2.2 Khoảng tin cậy cho trung bình

2.2.1. Bài toán tìm khoảng tin cậy đối xứng

Bài toán : Giả sử đám đông X có trung bình $E(X) = \mu$ chưa biết. Với độ tin cậy $1 - \alpha$, tìm khoảng tin cậy cho μ.

Theo (4.15), nếu M_1, M_2 thoả mãn $P(M_1 < \mu < M_2) = 1 - \alpha$ thì (M_1, M_2) là khoảng tin cậy của μ.

a) Cơ sở lí thuyết :

1) Trường hợp : Khi đám đông X có phân phối chuẩn, phương sai σ^2 đã biết.

Từ Hệ quả 2, Định lí 3.1, [Ch.3] ta có định lí sau :

Định lí 4.2 : Cho đám đông X có phân phối chuẩn : $X \sim N(\mu, \sigma^2)$. Giả sử $(X_1, X_2,\ldots, X_n)$ là mẫu ngẫu nhiên của đám đông X. Khi đó thống kê $Z = \dfrac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$.

Sử dụng thống kê $Z = \dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}$ của Định lí 4.2, tương tự như (4.18), (4.19) trong phần ước lượng tỉ lệ, ta có

$$P\left\{\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha .$$

Như vậy cho độ tin cậy $1 - \alpha$, ta có :

- Khoảng tin cậy đối xứng của trung bình μ :

$$\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{4.34}$$

Tương tự, ta có $P\left\{\mu < \overline{X} + z_{\alpha}\dfrac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$, $P\left\{\mu > \overline{X} - z_{\alpha}\dfrac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$.

Ta có khoảng tin cậy một phía :

- Ước lượng tối đa cho μ : $\left(\mu < \overline{X} + z_{\alpha}\dfrac{\sigma}{\sqrt{n}}\right)$ (4.35)

• Ước lượng tối thiểu cho μ : $\left(\mu > \bar{X} - z_\alpha \dfrac{\sigma}{\sqrt{n}}\right)$ (4.36)

2) Trường hợp : Khi đám đông X có phân phối chuẩn, phương sai σ^2 chưa biết.

Định lí 4.3 : Giả sử đám đông X có phân phối chuẩn : $X \sim N(\mu, \sigma^2)$ và $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên của đám đông X. Khi đó :

i) Nếu trung bình EX = μ thì thống kê $\chi^2 = \sum_{i=1}^{n} \dfrac{(X_i - \mu)^2}{\sigma^2}$ có phân phối *Khi bình phương với n bậc tự do.*

ii) Nếu trung bình EX chưa biết thì thống kê

$\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{\sigma^2}$ có phân phối *Khi bình phương với n – 1 bậc tự do.*

Ở đây $S = \sqrt{S^2}$; $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$.

Chứng minh i) Do $\dfrac{X_i - \mu}{\sigma} \sim N(0,1)$ [Mệnh đề 3.2, Ch.3] và cấu trúc của phân phối Khi bình phương [Mệnh đề 3.3, Ch.3] suy ra phần i) của Định lí 4.3. Từ các hệ thức :

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n} \qquad (4.37)$$

và vế trái của (4.37) có phân phối Khi bình phương n bậc tự do, thành phần thứ hai vế phải có phân phối Khi bình phương 1 bậc tự do, ta suy ra khẳng định ii) của Định lí 4.3.

Định lí 4.4 : Giả sử đám đông X có phân phối chuẩn : $X \sim N(\mu, \sigma^2)$, σ^2 phương sai chưa biết và $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên của đám đông X. Khi đó thống kê $T = \dfrac{\bar{X} - \mu}{S}\sqrt{n} \sim t(n-1)$ (phân phối Student với n-1 bậc tự do). Ở đây $S = \sqrt{S^2}$; $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$.

Chứng minh :

Do $\dfrac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$ [Mệnh đề 3.2, Ch.3] và $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ [Định lí 4.3], từ cấu trúc của phân phối Student [Mệnh đề 3.4, Ch.3] suy ra Định lí 4.4.

Tương tự như (4.34), (4.35), (4.36), từ Định lí 4.4, để tìm khoảng tin cậy $1-\alpha$ cho trung bình μ, trước hết tìm các giá trị tới hạn phân phối Student $T = \frac{\overline{X}-\mu}{S}\sqrt{n} \sim t(n-1)$ là :

$t\left(n-1, \frac{\alpha}{2}\right)$ và $t\left(n-1, 1-\frac{\alpha}{2}\right) = -t\left(n-1, \frac{\alpha}{2}\right)$ ta nhận được :

$$P\left\{\overline{X} - t\left(n-1, 1-\frac{\alpha}{2}\right)\frac{S}{\sqrt{n}} < \mu < \overline{X} - t\left(n-1, 1-\frac{\alpha}{2}\right)\frac{S}{\sqrt{n}}\right\} \tag{4.38}$$

Hệ thức (4.38) cho khoảng tin cậy đối xứng của μ. Tương tự :

- $P\left\{\mu < \overline{X} + t(n-1,\alpha)\frac{S}{\sqrt{n}}\right\} = 1-\alpha$ (4.39)

- $P\left\{\mu > \overline{X} - t(n-1,\alpha)\frac{S}{\sqrt{n}}\right\} = 1-\alpha$ (4.40)

Các hệ thức (4.39), (4.40) cho khoảng tin cậy một phía của μ.

3) Trường hợp : Khi kích thước mẫu lớn (đám đông X có thể không có phân phối chuẩn)

• *Nếu biết phương sai σ^2 :* Từ Định lí 3.2, [Ch.3] ta có :

$$G = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \xrightarrow{F} Z \sim N(0,1).$$

Do đó khi n khá lớn, ta coi :

$$\frac{\overline{X}-\mu}{\sigma}\sqrt{n} \approx N(0,\ 1) \text{ (xấp xỉ phân phối chuẩn chuẩn tắc)} \tag{4.41}$$

Trường hợp này chuyển về trường hợp 1).

• *Nếu chưa biết phương sai σ^2. Ta có :*

Định lí 4.5 : Giả sử đám đông X có trung bình μ và phương sai σ^2 chưa biết, $(X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên của đám đông X. Khi đó thống kê $G = \frac{\overline{X}-\mu}{S}\sqrt{n} \xrightarrow{F} Z \sim N(0,1)$.

Ở đây $S = \sqrt{S^2}$; $S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$.

Từ Định lí 4.5, khi n khá lớn, ta coi :

$$\frac{\overline{X}-\mu}{S}\sqrt{n} \approx N(0,\ 1) \tag{4.42}$$

Trường hợp này chuyển về trường hợp 1) ta thay σ bằng S.

Trong thực hành, nhiều trường hợp khi $n \geq 30$ các xấp xỉ (4.41), (4.42) được áp dụng tốt.

Chú ý : Trong sách này để tiện sử dụng và nếu không gây hiểu nhầm, ta quy ước khi kích thước mẫu tối thiểu $n \geq 30$ thì các xấp xỉ chuẩn chuẩn tắc (4.41), (4.42) được thỏa mãn. Khi đó, ta coi trường hợp 3) có quy tắc thực hành như trường hợp 1).

b) Quy tắc thực hành :

Cho mẫu cụ thể $(x_1, ..., x_n)$ của đám đông X và độ tin cậy $1 - \alpha$.

Trường hợp 1	**Trường hợp 2**
– *biết phương sai ; hoặc* – *kích thước mẫu lớn* $(n \geq 30)$. *(Bảng giá trị tới hạn chuẩn)* [Bảng 2]	– đám đông *có phân phối chuẩn, phương sai chưa biết.* *(Bảng giá trị tới hạn Student)* [Bảng 3]
Ước lượng trung bình (4.43) $\overline{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \overline{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$	Ước lượng trung bình (4.44) $\overline{x} - t(n-1,\frac{\alpha}{2})\frac{s}{\sqrt{n}} < \mu < \overline{x} + t(n-1,\frac{\alpha}{2})\frac{s}{\sqrt{n}}$
Ước lượng tối đa $\mu < \overline{x} + z_\alpha\frac{s}{\sqrt{n}}$ (4.45)	Ước lượng tối đa $\mu < \overline{x} + t(n-1, \alpha)\frac{s}{\sqrt{n}}$ (4.46)
Ước lượng tối thiểu $\overline{x} - z_\alpha\frac{s}{\sqrt{n}} < \mu$ (4.47)	Ước lượng tối thiểu $\overline{x} - t(n-1, \alpha)\frac{s}{\sqrt{n}} < \mu$ (4.48)

Ở đây $\overline{x}$ là trung bình mẫu, $s = \sqrt{s^2}$ là độ lệch tiêu chuẩn mẫu khi chưa biết phương sai và $s = \sqrt{\sigma^2}$ khi biết phương sai ; $z_{\frac{\alpha}{2}}, z_\alpha$ giá trị tới hạn tương ứng chuẩn mức $\frac{\alpha}{2}$, α [Bảng 2]. $t(n-1, \alpha/2)$, $t(n-1, \alpha)$ là các giá trị tới hạn Student, bậc tự do $n - 1$, tương ứng mức $\frac{\alpha}{2}$, α [Bảng 3].

Chú ý : Trong Trường hợp 2 khi kích thước mẫu lớn $(n \geq 30)$, có thể thực hiện như Trường hợp 1.

Các bước tìm (4.43), (4.44) như sau :

(4.43) [Bảng 2]	(4.44) [Bảng 3]
$+\overline{x}, s$ $+1-\alpha \rightarrow \frac{1-\alpha}{2} = \varphi\left(z_{\frac{\alpha}{2}}\right) \rightarrow z_{\frac{\alpha}{2}}$ $+\varepsilon = z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ $+(\overline{x}-\varepsilon < \mu < \overline{x}+\varepsilon)$	$+\overline{x}, s$ $+1-\alpha \rightarrow \frac{\alpha}{2} \rightarrow t(n-1\,;\frac{\alpha}{2})$ $+\varepsilon = t(n-1\,;\frac{\alpha}{2}).\frac{s}{\sqrt{n}}$ $+(\overline{x}-\varepsilon < \mu < \overline{x}+\varepsilon)$

Ví dụ 4.20 : Chủ một kho cung cấp sơn muốn ước lượng lượng sơn chứa trong một thùng được sản xuất từ một dây chuyền công nghệ quốc gia. Biết rằng theo tiêu chuẩn của dây chuyền công nghệ đó, độ lệch tiêu chuẩn của lượng sơn là 0,08 thùng. Điều tra một mẫu 50 thùng tính được lượng sơn trung bình chứa trong một thùng là 0,97 (thùng). Với độ tin cậy 99%, hãy ước lượng :

a) Lượng sơn trung bình chứa trong một thùng.

b) Lượng sơn trung bình tối thiểu chứa trong một thùng.

GIẢI : a) Gọi μ là lượng sơn trung bình có trong một thùng.

Ta có $\overline{x} = 0{,}97$; $\sigma = 0{,}08$ (trường hợp 1))

$+\ 1-\alpha = 0{,}99 \rightarrow z_{\frac{\alpha}{2}} = 2{,}58$;

$+\ \varepsilon = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2{,}58 \cdot \frac{0{,}08}{\sqrt{50}} \approx 0{,}0292$;

+ Khoảng tin cậy của μ : $(\overline{x}-\varepsilon; \overline{x}+\varepsilon) = (0{,}9408\ ;\ 0{,}9992)$ (thùng).

Với độ tin cậy 99% lượng sơn trung bình có trong một thùng từ 0,9408 (thùng) đến 0,9992 (thùng).

b) $1-\alpha = 0{,}99 :\ z_{\alpha} = 2{,}33$.

+ Với $\varepsilon = z_{\alpha} \frac{\sigma}{\sqrt{n}} = 2{,}33 \cdot \frac{0{,}08}{\sqrt{50}} = 0{,}0264$;

+ Khoảng tin cậy bên phải (ước lượng giá trị tối thiểu của μ) :

$$\mu > \overline{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}} = 0{,}97 - 0{,}0264 = 0{,}9436.$$

Với độ tin cậy 99% lượng sơn trung bình tối thiểu có trong một thùng là 0,9436 (thùng).

Ví dụ 4.21 : Trở lại Ví dụ 4.8. Với độ tin cậy 95%, hãy ước lượng :

a) Lượng xăng hao phí trung bình của ôtô đi từ A đến B.

b) Lượng xăng hao phí trung bình tối đa của ôtô đi từ A đến B.

GIẢI : Gọi μ là lượng xăng hao phí trung bình của ô tô đi từ A tới B. Ta đã tính được $\overline{x} = 10{,}1333$; $s^2 = 0{,}0561$ (phương sai σ^2 của lượng xăng hao phí chưa biết).

a) Với $1 - \alpha = 0{,}95$: $z_{\frac{\alpha}{2}} = 1{,}96$.

$$+ \ \varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}} = 1{,}96\sqrt{\frac{0{,}0561}{30}} \approx 0{,}0848.$$

+ Khoảng tin cậy của μ : $(\overline{x} - \varepsilon\, ;\, \overline{x} + \varepsilon) = (10{,}0485\, ;\, 10{,}2181)$ (lít).

b) Với $1 - \alpha = 0{,}95$: $z_\alpha = 1{,}65$.

+ Khoảng tin cậy phía trái (ước lượng giá trị tối đa của μ)

$$\mu < \overline{x} + z_\alpha\sqrt{\frac{s^2}{n}} = 10{,}1333 + 1{,}65\sqrt{\frac{0{,}0561}{30}} = 10{,}2047.$$

Với độ tin cậy 95% lượng xăng hao phí trung bình tối đa của ôtô đi từ A đến B là 10,2047 lít.

Ví dụ 4.22 : Giá bán của một loại thiết bị (đơn vị USD) trên thị trường là biến ngẫu nhiên có phân phối chuẩn. Một người định mua một thiết bị loại này, khảo sát ngẫu nhiên tại 18 cửa hàng tính được bán giá trung bình của thiết bị là 137,75 USD và độ lệch tiêu chuẩn 7,98 USD. Với độ tin cậy 90%, hãy ước lượng giá bán trung bình của thiết bị.

GIẢI : Gọi μ là giá bán trung bình của thiết bị. Ta có

+ $\overline{x} = 137{,}75$; $s = 7{,}98$.

$$+ \ 1 - \alpha = 0{,}9 \rightarrow \frac{\alpha}{2} = 0{,}05 \xrightarrow{\text{(Bảng 3)}} t(17\, ;\, 0{,}05) = 1{,}74.$$

$$+ \ \varepsilon = t(17\, ;\, 0{,}05) \cdot \frac{7{,}98}{\sqrt{18}} = 3{,}2728.$$

Khoảng tin cậy 90% cho giá bán trung bình của thiết bị là

$$(\overline{x} - \varepsilon\, ;\, \overline{x} + \varepsilon) = (134{,}4772\, ;\, 141{,}0228) \text{ (USD)}.$$

Vậy với độ tin cậy 90%, giá bán trung bình của một thiết bị từ 134,4772 USD đến 141,0228 USD.

2.2.2. Các chỉ tiêu của bài toán khoảng tin cậy đối xứng

Với khoảng tin cậy đối xứng $(\overline{x}-\varepsilon\,;\overline{x}+\varepsilon)$, $\varepsilon = z_{\frac{\alpha}{2}}\dfrac{s}{\sqrt{n}}$ và độ tin cậy $1-\alpha$, kích thước mẫu n, độ chính xác ε. Ta có ba bài toán.

Bài toán 1 : Cho độ tin cậy, mẫu điều tra, tìm độ chính xác, nghĩa là biết $1-\alpha$ và n tìm ε.

Rõ ràng $\varepsilon = z_{\frac{\alpha}{2}}\dfrac{s}{\sqrt{n}}$. (4.49)

Bài toán 2 : Cho độ chính xác, mẫu điều tra, tìm độ tin cậy, nghĩa là biết ε, n tìm $1-\alpha$.

Từ $\varepsilon = z_{\frac{\alpha}{2}}\dfrac{s}{\sqrt{n}}$ ta nhận được $z_{\frac{\alpha}{2}} = \dfrac{\varepsilon\sqrt{n}}{s}$, (4.50)

$$z_{\frac{\alpha}{2}} \xrightarrow{\text{(Bảng 2)}} \varphi(z_{\frac{\alpha}{2}}) \Rightarrow 1-\alpha = 2\varphi(z_{\frac{\alpha}{2}}). \qquad (4.51)$$

Bài toán 3 : Cho độ tin cậy $1-\alpha$ và độ chính xác ε tìm

– Kích thước mẫu tối thiểu cần điều tra m : $m \geq \left(\dfrac{z_{\frac{\alpha}{2}}\cdot s}{\varepsilon}\right)^2$.

Khi $\left(\dfrac{z_{\frac{\alpha}{2}}\cdot s}{\varepsilon}\right)^2$ là số thập phân, ta lấy $m = \left[\left(\dfrac{z_{\frac{\alpha}{2}}\cdot s}{\varepsilon}\right)^2\right]+1$,

ở đây kí hiệu [ ] là lấy phần nguyên, ví dụ [178,589] = 178.

– Kích thước mẫu cần điều tra thêm : $K = m - n$ (trong trường hợp $m > n$ với n là kích thước mẫu hiện có) .

Chú ý : Các bài toán chỉ tiêu của các trường hợp khác có độ chính xác $\varepsilon = z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}$; $\varepsilon = t(n-1, \dfrac{\alpha}{2})\dfrac{s}{\sqrt{n}}$ cũng được thiết lập tương tự.

Ví dụ 4.23 : Trở lại Ví dụ 4.20. Nếu chủ kho muốn ước lượng lượng sơn trung bình trong một thùng với độ tin cậy 99% và độ chính xác 2%, thì cần điều tra thêm bao nhiêu thùng nữa ?

GIẢI : $1-\alpha = 0,99 \to z_{\frac{\alpha}{2}} = 2,58$; $\varepsilon = 0,02$. Ta có kích thước mẫu cần điều tra :

$$m = \left[\left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{\varepsilon}\right)^2\right] + 1 = \left[\left(\frac{2,58 \cdot 0,08}{0,02}\right)^2\right] + 1 = [106,5024] + 1 = 106 + 1 = 107.$$

Kích thước mẫu đã có n = 50, kích thước mẫu cần điều tra thêm : 107 – 50 = 57.

Ví dụ 4.24 : Trở lại Ví dụ 4.21. Nếu sử dụng mẫu này để ước lượng lượng xăng hao phí trung bình của ô tô đi từ A đến B đạt mức chính xác 0,1 lít thì đảm bảo độ tin cậy là bao nhiêu ?

GIẢI : Từ (4.49), $z_{\frac{\alpha}{2}} = 0,1\sqrt{\dfrac{30}{0,0561}} \approx 2,31$

$\Rightarrow \varphi(2,31) = 0,48956$ từ đó $1-\alpha = 2\varphi(2,31) \approx 0,9791$.

2.3. Khoảng tin cậy cho phương sai

Bài toán. Giả sử đám đông X có phân phối chuẩn : $X \sim N(\mu, \sigma^2)$, σ^2 chưa biết. Với độ tin cậy $1-\alpha$, tìm khoảng tin cậy cho σ^2.

a) Trường hợp trung bình μ chưa biết.

Sử dụng thống kê $\chi^2 = \dfrac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$ [xem ii) Định lí 4.3], cho độ tin cậy $1-\alpha$ ta tìm

• Khoảng tin cậy của phương sai : $\dfrac{(n-1)S^2}{\chi^2(n-1, \frac{\alpha}{2})} < \sigma^2 < \dfrac{(n-1)S^2}{\chi^2(n-1, 1-\frac{\alpha}{2})}$ (4.25)

(4.46) được tìm như sau : Tìm các giá trị tới hạn $\chi_1^2 = \chi^2(n-1, \frac{\alpha}{2})$, $\chi_2^2 = \chi^2\left(n-1, 1-\frac{\alpha}{2}\right)$ [Bảng 4]. Ta có :

$$P\left\{\frac{n-1}{\sigma^2}S^2 > \chi^2(n-1, 1-\frac{\alpha}{2})\right\} = 1-\frac{\alpha}{2} \; ; P\left\{\frac{n-1}{\sigma^2}S^2 > \chi^2(n-1, \frac{\alpha}{2})\right\} = \frac{\alpha}{2}.$$

$$\Rightarrow P\left\{\chi^2(n-1, 1-\frac{\alpha}{2}) < \frac{n-1}{\sigma^2}S^2 < \chi^2(n-1, \frac{\alpha}{2})\right\}$$

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

• Tương tự : Ước lượng tối đa của phương sai có từ đẳng thức (4.52)

$$1-\alpha = P\left(\frac{n-1}{\sigma^2}S^2 > \chi^2(n-1,\, 1-\alpha)\right) = P\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2(n-1\,;\, 1-\alpha)}\right). \qquad (4.52)$$

• Ước lượng tối thiểu của phương sai có từ (4.48)

$$P\left(\frac{n-1}{\sigma^2}S^2 > \chi^2(n-1,\, \alpha)\right) = \alpha \Rightarrow P\left(\frac{n-1}{\sigma^2}S^2 < \chi^2(n-1,\, \alpha)\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{n-1}{\chi^2(n-1,\, \alpha)}S^2 < \sigma^2\right) = 1-\alpha \qquad (4.53)$$

Quy tắc thực hành : Từ mẫu cụ thể $(x_1, x_2, ..., x_n)$.

• Tính $(n-1)s^2 = \sum n_i\left(x_i - \overline{x}\right)^2$

• Cho $1-\alpha$: (Bảng 4, bậc tự do $n-1$)

Với $\frac{\alpha}{2}$: $\chi_1^2 = \chi^2_{\left(n-1,\frac{\alpha}{2}\right)}$ và $1-\frac{\alpha}{2}$: $\chi_2^2 = \chi^2\left(n-1, 1-\frac{\alpha}{2}\right)$

• Khoảng tin cậy của phương sai : $\left(\frac{(n-1)s^2}{\chi_1^2}; \frac{(n-1)s^2}{\chi_2^2}\right)$ (4.54)

Tương tự ta có các khoảng tin cậy một phía :

• Ước lượng tối đa của phương sai :

Với $\chi^2 = \chi^2(n-1,\, 1-\alpha)$: $\left(0\,;\, \frac{(n-1)s^2}{\chi^2}\right)$ (4.55)

• Ước lượng tối thiểu của phương sai :

Với $\chi^2 = \chi^2(n-1,\alpha)$: $\left(\frac{(n-1)s^2}{\chi^2}\,;\, +\infty\right)$ (4.56)

Trường hợp trung bình μ biết

Tương tự trường hợp a), sử dụng thống kê $\chi^2 = \sum_{i=1}^{n}\frac{\left(X_i-\mu\right)^2}{\sigma^2} \sim \chi^2(n)$ từ i) Định lí 4.3, cho độ tin cậy $1-\alpha$, tìm các giá trị tới hạn Khi bình phương với n bậc tự do $\chi_1^2 = \chi^2\left(n, \frac{\alpha}{2}\right)$; $\chi_2^2 = \chi^2\left(n, 1-\frac{\alpha}{2}\right)$, ta có :

Quy tắc thực hành : Từ mẫu cụ thể $(x_1, x_2,..., x_n)$.

• Tính $\sum_{i=1}^{n}(x_i - \mu)^2$.

• Cho $1 - \alpha$: (Bảng 4, bậc tự do n)

Với $\frac{\alpha}{2}$: $\chi_1^2 = \chi^2_{\left(n, \frac{\alpha}{2}\right)}$; và $1 - \frac{\alpha}{2}$: $\chi_2^2 = \chi^2\left(n, 1 - \frac{\alpha}{2}\right)$.

• Khoảng tin cậy của phương sai : $$\left(\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\chi_1^2} ; \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\chi_2^2}\right) \qquad (4.57)$$

Tương tự, ta có các khoảng tin cậy một phía :

• Ước lượng tối đa của phương sai :

Với $\chi^2 = \chi^2(n, 1 - \alpha)$: $$\left(0; \frac{\sum(x_i - \mu)^2}{\chi^2}\right) \qquad (4.58)$$

• Ước lượng tối thiểu của phương sai :

Với $\chi^2 = \chi^2(n, \alpha)$: $$\left(\frac{\sum(x_i - \mu)^2}{\chi^2}; +\infty\right) \qquad (4.59)$$

Ví dụ 4.25 : Mức hao phí nguyên liệu X để sản xuất một sản phẩm của nhà máy (tính bằng kilôgam) tuân theo quy luật chuẩn. Mẫu điều tra về mức hao phí nguyên liệu để sản xuất 25 sản phẩm loại này cho kết quả trong bảng :

Mức hao phí nguyên liệu hao phí (kg)	19,5	20,0	20,5
Số sản phẩm	5	18	2

Với độ tin cậy 95%

1) Khi không biết kì vọng EX, hãy :

a) Ước lượng phương sai mức hao phí nguyên liệu để sản xuất một sản phẩm của nhà máy.

b) Ước lượng tối đa cho phương sai của mức hao phí nguyên liệu để sản xuất một sản phẩm của nhà máy.

2) Khi biết kì vọng EX = 20, hãy ước lượng phương sai mức hao phí nguyên liệu để sản xuất một sản phẩm của nhà máy.

GIẢI : 1) a) Ta có : n = 25 ; $\overline{x} = 19,94$; $(n-1)s^2 = 1,66$.

Với $1-\alpha = 0,95$ ta có $\frac{\alpha}{2} = 0,025$, $\chi_1^2 = \chi^2(24\,;0,025) = 39,364$; $1-\frac{\alpha}{2} = 0,975$, $\chi_2^2 = \chi^2(24\,;0,975) = 12,401$.

Khoảng tin cậy 95% của σ^2 : $\left(\frac{1,66}{39,364}\,;\frac{1,66}{12,401}\right) = (0,0422\,;0,1339)\,(kg^2)$.

b) Với $1-\alpha = 0,95$, ta có : $\chi^2 = \chi^2(24\,;0,95) = 13,848$.

Khoảng tin cậy 95% tối đa của σ^2 : $\left(0\,;\frac{1,66}{13,848}\right) = (0\,;0,1199)\,(kg^2)$.

2) Khi biết EX = 20 :

$$\sum n_i(x_i-\mu)^2 = 5.(19,5-20)^2 + 18.(20-20)^2 + 2.(20,5-20)^2 = 1,75\,;$$

$1-\alpha = 0,90$ ta có $\frac{\alpha}{2} = 0,05$, $\chi^2{}_1 = \chi^2(25\,;0,05) = 37,652$;

$1-\frac{\alpha}{2} = 0,95$, $\chi^2{}_2 = \chi^2(25\,;0,95) = 14,611$.

Khoảng tin cậy 95% của σ^2 : $\left(\frac{1,75}{37,652}\,;\frac{1,75}{14,611}\right) = (0,0465\,;0,1198)(kg^2)$.

BÀI TẬP CHƯƠNG 4

4.1. Mẫu số liệu điều tra về lãi suất tiền gửi tiết kiệm X (đơn vị %) của 50 ngân hàng một kì trong một vùng cho như sau :

7	9	8	6	12	6	9	15	9	16
8	5	14	8	7	6	10	8	11	4
10	6	16	5	10	12	7	10	15	7
10	8	8	10	18	8	10	11	7	10
7	8	8	23	13	9	8	9	9	13

a) Lập bảng phân phối tần số, tần suất thực nghiệm của X.

b) Vẽ các đa giác tần số, tần suất của X.

c) Lãi suất tiền gửi tiết kiệm trên 10% một kì là lãi suất cao. Tính tỉ lệ ngân hàng có lãi suất tiền gửi tiết kiệm cao của vùng.

d) Tính trung bình, phương sai, độ lệch tiêu chuẩn của mẫu.

4.2. Điều tra trọng lượng của một loại sản phẩm (đơn vị : gam) kết quả cho trong bảng :

x_i	2 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
n_i	5	25	48	35	17	10

1) Những sản phẩm có trọng lượng lớn hơn 25 gam là loại 1.

Hãy ước lượng tỉ lệ các sản phẩm loại 1.

b) Hãy ước lượng tỉ lệ các sản phẩm loại 1 với độ tin cậy 95%.

c) Nếu sử dụng mẫu này để ước lượng tỉ lệ sản phẩm loại 1 chính xác tới 5% thì độ tin cậy đạt được là bao nhiêu ?

2) Những sản phẩm có trọng lượng nhỏ hơn 15 gam là phế phẩm.

a) Hãy ước lượng tỉ lệ phế phẩm với độ tin cậy 96%.

b) Nếu muốn ước lượng tỉ lệ phế phẩm với sai số không quá 5% và độ tin cậy 99% thì cần điều tra thêm tối thiểu bao nhiêu sản phẩm nữa ?

4.3. Một khách hàng nhận được lô hàng từ một nhà máy sản xuất bút bi rẻ tiền. Để ước lượng tỉ lệ bút bi bị hỏng, khách hàng lấy ngẫu nhiên 300 bút kiểm tra và nhận được 20 chiếc hỏng. Với độ tin cậy 95%, hãy :

a) Ước lượng tỉ lệ bút bi hỏng của lô hàng.

b) Ước lượng tỉ lệ bút bi bị hỏng tối thiểu của lô hàng.

c) Ước lượng số bút bi hỏng, số bút bi hỏng tối đa của lô hàng, biết lô hàng có 5000 bút bi.

d) Nếu muốn ước lượng tỉ lệ bút bi hỏng đạt độ tin cậy 96% và độ chính xác 2% thì cần kiểm tra thêm bao nhiêu bút bi nữa ?

4.4. Một nông dân muốn ước lượng tỉ lệ nảy mầm cho một giống lúa mới, anh ta lấy 1000 hạt lúa này đem gieo và thấy có 640 hạt nảy mầm.

a) Với độ tin cậy 95% hãy ước lượng tỉ lệ hạt nảy mầm của giống lúa này.

b) Nếu muốn sai số của ước lượng tỉ lệ không vượt quá 0,02 và với độ tin cậy 95% thì cần gieo tối thiểu bao nhiêu hạt ?

4.5. Để đánh giá trữ lượng cá trong hồ người ta đánh bắt 2000 con cá đánh dấu rồi thả xuống hồ. Sau đó bắt lại 400 con thì thấy có 80 con có dấu. Với độ tin cậy 95% :

a) Hãy ước lượng trữ lượng cá có trong hồ.

b) Nếu muốn sai số của ước lượng giảm đi một nửa thì lần sau phải đánh bắt bao nhiêu con cá ?

4.6. Để điều tra thị trường về một loại sản phẩm mới, người ta phỏng vấn ngẫu nhiên 300 khách hàng trong đó có 90 người thích sản phẩm này.

a) Với độ tin cậy 95%, hãy ước lượng tỉ lệ khách hàng thích sản phẩm này.

b) Với độ tin cậy 95%, hãy ước lượng tỉ lệ tối đa khách hàng thích sản phẩm này.

c) Nếu muốn ước lượng tỉ lệ khách hàng thích sản phẩm này đạt độ tin cậy 95% và độ chính xác là 0,03 thì cần phỏng vấn thêm bao nhiêu người nữa.

d) Với mẫu điều tra trên, nếu muốn ước lượng tỉ lệ khách hàng thích sản phẩm này chính xác 4,36% thì đảm bảo độ tin cậy là bao nhiêu ?

4.7. Tỉ lệ phế phẩm cho phép của một lô hàng là 10%. Với xác suất 95% nếu điều tra ngẫu nhiên một mẫu 100 sản phẩm của lô hàng thì tỉ lệ phế phẩm tối đa của mẫu sản phẩm đó là bao nhiêu để có thể chấp nhận lô hàng ?

4.8. Độ dài của một chi tiết máy được sản xuất trên dây chuyền tự động là biến ngẫu nhiên có phân phối chuẩn trung bình 200mm và độ lệch tiêu chuẩn 20mm. Một mẫu gồm 25 chi tiết được kiểm tra.

a) Tìm giá trị tới hạn mức 5% của độ dài trung bình các chi tiết máy được kiểm tra.

b) Tính xác suất để độ dài trung bình các chi tiết máy được kiểm tra không vượt quá 212mm.

c) Tìm giá trị tới hạn mức 5% của phương sai mẫu.

4.9. Mẫu số liệu điều tra về thời gian chờ phục vụ của 30 khách hàng tại một ngân hàng (tính bằng phút) cho như sau :

4,6	9,8	5,6	7,7	4,0	6,5	2,9	4,4	4,0	5,7
10,9	4,5	1,4	2,4	8,6	4,7	6,7	7,8	9,2	4,2
5,2	5,0	5,8	3,2	4,3	8,4	7,2	3,4	6,5	2,2

a) Hãy ước lượng thời gian chờ phục vụ trung bình của một khách hàng tại ngân hàng với độ tin cậy 97%.

b) Nếu muốn ước lượng thời gian chờ phục vụ trung bình của khách hàng với sai số không quá 0,8 phút với độ tin cậy 98% thì cần phải điều tra thêm ít nhất bao nhiêu khách hàng nữa ?

c) Với độ tin cậy 96%, hãy ước lượng thời gian chờ phục vụ trung bình tối đa của một khách hàng.

4.10. Tỉ lệ nợ xấu tại một ngân hàng là tỉ số của tổng số nợ quá hạn và tổng số nợ cho vay đang được thực hiện. Điều tra ngẫu nhiên 17 ngân hàng ở vùng A có tỉ lệ nợ xấu (tính bằng %) là : 7, 4, 6, 7, 5, 4, 9, 3, 4, 5, 6, 8, 5, 2, 4, 5, 6. Giả sử tỉ lệ nợ xấu có phân phối xấp xỉ chuẩn. Với độ tin cậy 95%

a) Ước lượng tỉ lệ nợ xấu trung bình của các ngân hàng vùng A.

b) Ước lượng tỉ lệ nợ xấu trung bình tối đa của các ngân hàng vùng A.

4.11. Giám đốc chi nhánh của một ngân hàng muốn ước lượng lượng tiền gửi trung bình của mỗi khách hàng tại ngân hàng. Chọn ngẫu nhiên 50 khách hàng gửi tiền tại ngân hàng, tính được lượng tiền gửi trung bình của một người là 475 triệu và độ lệch tiêu chuẩn là 12 triệu.

a) Với độ tin cậy 95% hãy ước lượng lượng tiền gửi trung bình của mỗi khách hàng tại ngân hàng.

b) Với độ tin cậy 95% hãy ước lượng lượng tiền gửi trung bình của khách hàng tại ngân hàng. Biết ngân hàng có 1200 khách hàng.

c) Nếu sử dụng mẫu điều tra để ước lượng lượng tiền gửi trung bình của một khách hàng tại ngân hàng chính xác tới 2,5 triệu thì đạt độ tin cậy là bao nhiêu.

d) Nếu muốn ước lượng lượng tiền gửi trung bình của một khách hàng tại ngân hàng với độ tin cậy là 99% và độ chính xác là 3 triệu thì cần điều tra thêm bao nhiêu khách hàng nữa.

4.12. Để ước lượng lợi nhuận của một tổng công ty gồm 380 cửa hàng trên toàn quốc trong một tháng, người ta chọn ngẫu nhiên 10% số cửa hàng của tổng công ty điều tra doanh thu được kết quả cho trong bảng :

Doanh thu (triệu đồng/tháng)	20	40	60	80
Số cửa hàng	8	16	12	2

a) Với độ tin cậy 95% hãy ước lượng doanh thu trung bình của một cửa hàng của tổng công ty.

b) Với độ tin cậy 95% hãy ước lượng lợi nhuận trung bình của công ty một năm. Biết lợi nhuận của tổng công ty khoảng 25% tổng doanh thu.

4.13. Điều tra chỉ tiêu X (tính bằng %) của một số sản phẩm cùng loại được kết quả cho trong bảng :

x_i	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40
n_i	7	12	20	25	18	12	5	1

a) Những sản phẩm có chỉ tiêu X không quá 10% là loại 2. Hãy ước lượng tỉ lệ sản phẩm loại 2 với độ tin cậy 96%.

b) Nếu muốn ước lượng lệ sản phẩm loại 2 đạt độ tin cậy 96% và độ chính xác 6,5% thì cần điều tra thêm bao nhiêu sản phẩm nữa.

c) Hãy ước lượng trung bình chỉ tiêu X với độ tin cậy 95%.

d) Nếu muốn ước lượng trung bình chỉ tiêu X với độ tin cậy 95% và độ chính xác 0,95% thì cần điều tra thêm bao nhiêu sản phẩm nữa.

e) Nếu muốn ước lượng lệ sản phẩm loại 2 đạt độ tin cậy 96%, độ chính xác 6,5% và ước lượng trung bình chỉ tiêu X với độ tin cậy 95%, độ chính xác 0,95% thì cần điều tra thêm bao nhiêu sản phẩm nữa.

4.14. Một nhân viên nghiên cứu thị trường của một công ty buôn bán điện tử điều tra về sở thích xem TV của dân cư một thành phố. Điều tra ngẫu nhiên 140 người của thành phố cho thấy số giờ xem TV trung bình của mỗi người trong một tuần lễ là : $\overline{x} = 15,3$; độ lệch tiêu chuẩn $s = 3,8$ và có 30 người xem tin đêm ít nhất 3 lần trong một tuần. Với độ tin cậy 95%, hãy ước lượng :

a) Thời gian xem TV trung bình của mỗi người dân thành phố đó trong một tuần.

b) Tỉ lệ người dân xem tin đêm ít nhất 3 buổi một tuần.

c) Kích thước mẫu điều tra là bao nhiêu nếu với độ tin cậy 95% công ty muốn ước lượng thời gian xem TV trung bình của mỗi người dân ở đó chính xác đến 0,3 giờ và tỉ lệ người xem tin đêm ít nhất 3 buổi một tuần chính xác tới 3,5%.

4.15. Thời gian của một cuộc điện thoại đường dài có phân phối chuẩn trung bình 8 phút, độ lệch tiêu chuẩn 2 phút. Chọn ngẫu nhiên một mẫu 25 cuộc điện thoại đường dài ở một tổng đài.

a) Tìm độ lệch tiêu chuẩn của trung bình mẫu.

b) Trung bình 25 cuộc điện thoại này nằm trong khoảng từ 7,8 đến 8,2 (phút) chiếm tỉ lệ là bao nhiêu.

4.16. Để nghiên cứu tuổi thọ của một thiết bị (tính bằng tháng) người ta điều tra ngẫu nhiên 15 thiết bị loại này kết quả như sau : 114, 78, 96, 137, 78, 103, 126, 86, 99, 114, 72, 104, 73, 86, 117. Giả sử tuổi thọ của thiết bị có phân phối chuẩn.

a) Tìm ước lượng điểm cho tuổi thọ trung bình của thiết bị. Cho biết tính chất của ước lượng này.

b) Tìm khoảng tin cậy 95% cho tuổi thọ trung bình của thiết bị.

c) Nếu muốn ước lượng tuổi thọ trung bình của thiết bị đạt độ tin cậy 95% và độ chính xác là 5 tháng thì cần điều tra thêm bao nhiêu thiết bị nữa ?

d) Hãy ước lượng phương sai của tuổi thọ thiết bị với độ tin cậy 95%.

e) Hãy ước lượng mức phân tán tối đa của tuổi thọ thiết bị với độ tin cậy 95%.

4.17. Lãi suất cổ phiếu của một công ty trong 15 năm qua (đơn vị %) là 5, 6, 7, 4, 8, 10, 8, 7, 6, 10, 15, 10, 20, 7, 14. Với độ tin cậy 95%, hãy ước lượng : độ phân tán ; độ phân tán tối đa ; độ phân tán tối thiểu của lãi suất cổ phiếu của công ty đó, biết lãi suất cổ phiếu là biến ngẫu nhiên có phân phối chuẩn.

4.18. Để nghiên cứu độ ổn định của một loại máy tiện người ta lấy ngẫu nhiên 24 trục máy do máy tiện loại này sản xuất ra và đo đường kính (đơn vị mm) của chúng cho kết quả :

24,1	27,2	26,7	23,6	24,6	24,5	26,4	26,1
25,8	27,3	23,2	26,9	27,1	25,4	23,3	25,9
22,7	26,9	24,8	24,0	23,4	23,0	24,3	25,4

Với độ tin cậy 95% hãy ước lượng : độ phân tán ; độ phân tán tối đa của đường kính trục máy.

Rủi ro đầu tư thường được đo bằng phương sai của tỉ lệ thu hồi vốn của dự án. Từ số liệu mẫu về tỉ lệ thu hồi vốn của hai dự án trong 10 năm tính được kết quả cho trong bảng :

	Dự án 1	***Dự án 2***
Kích thước mẫu	10	10
Tỉ lệ thu hồi vốn trung bình (%)	13,2	14,6
Phương sai mẫu $(\%)^2$	10,9	25,6

Với độ tin cậy 95% hãy ước lượng :

a) Tỉ lệ thu hồi vốn trung bình của hai dự án,

b) Phương sai của tỉ lệ thu hồi vốn của hai dự án.

Biết rằng tỉ lệ thu hồi vốn của các dự án là biến ngẫu nhiên có phân phối chuẩn.

4.20. Để đánh giá mức tiêu thụ điện của mỗi hộ gia đình ở vùng A trong mùa khô, công ty điện lực vùng này tiến hành điều tra 400 hộ kết quả cho trong bảng :

Mức tiêu thụ (tính bằng 100kw/tháng)	0 – 1	1 – 2	2 – 3	3 – 4	4 – 5	5 – 6
Số hộ gia đình	20	110	150	64	46	10

a) Hãy ước lượng mức tiêu thụ điện trung bình của mỗi hộ gia đình ở vùng A trong 6 tháng mùa khô với độ tin cậy 95%.

b) Những hộ có mức tiêu thụ điện trên 400 kwh/tháng là những hộ có mức tiêu thụ điện cao. Hãy ước lượng hộ có mức tiêu thụ điện cao với độ tin cậy 95%, biết vùng này có 10000 hộ.

c) Với độ tin cậy 95%, hãy ước lượng mức tiêu thụ điện trung bình tối đa một tháng của vùng. Biết vùng này có 10000 hộ.

4.21. Điều tra hai chỉ tiêu X (tính bằng %), Y (tính bằng gam) của một loại sản phẩm được kết quả cho trong bảng :

X \ Y	0,75	2	3,2	5
0 – 5	17	5		
5 – 10		20	15	
10 – 15	4	11	4	6
15 – 20			12	8

a) Những sản phẩm có chỉ tiêu X ≥ 10%, Y ≥ 2 gam là loại 1. Với độ tin cậy 95%, hãy ước lượng tỉ lệ sản phẩm loại 1.

b) Với độ tin cậy 95%, hãy ước lượng trung bình chỉ tiêu X.

4.22. Chất lượng của một loại sản phẩm được đánh giá qua hai chỉ tiêu : X (đơn vị %), Y (đơn vị : kg/cm^2). Mẫu điều tra các sản phẩm loại này cho kết quả trong bảng :

Y \ X	30 – 35	35 – 40	40 – 45	45 – 50
130 – 135	3			
135 – 140	3	14	18	
140 – 145		11	20	
145 – 150		3	17	5
150 – 160			2	4

a) Những sản phẩm có chỉ tiêu $X \leq 40\%$, $Y \leq 140$ kg/cm^2 là loại B. Hãy ước lượng tỉ lệ sản phẩm loại B với độ tin cậy 95%.

b) Với độ tin cậy 95%, hãy ước lượng trung bình chỉ tiêu X, trung bình chỉ tiêu Y của các sản phẩm.

c) Nếu muốn ước lượng tỉ lệ sản phẩm loại B đạt độ chính xác là 5% và ước lượng trung bình chỉ tiêu Y đạt độ chính xác là 0,8 kg/cm^2 và cả hai ước lượng cùng có độ tin cậy 90% thì phải điều tra thêm bao nhiêu sản phẩm nữa ?

Chương 5

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

§1. GIẢ THUYẾT THỐNG KÊ VÀ KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

1. Giả thuyết thống kê

Giả thuyết thống kê là một khẳng định nói về các vấn đề của đám đông. Ví dụ giả thuyết thống kê nói về :

• Các tham số đặc trưng của một đám đông như trung bình μ, tỉ lệ p, phương sai σ^2 ; Các số đặc trưng tương ứng về trung bình, tỉ lệ, phương sai của hai đám đông có bằng nhau hay không...

• Dạng quy luật phân phối xác suất của đám đông ;

• Tính độc lập của các đám đông,...

Giả thuyết thống kê được phát biểu trong một mệnh đề, kí hiệu là H_0. Mệnh đề đối lập với nó (được gọi là đối thuyết) kí hiệu là H_1. H_0, H_1 tạo thành một cặp giả thuyết thống kê, được nghiên cứu đồng thời để cho kết luận : hoặc chấp chận H_0, hoặc bác bỏ H_0 chấp nhận H_1.

Ví dụ 5.1 :

Giả thuyết gốc : H_0	Đối thuyết : H_1
• Nhu cầu về một loại hàng hoá của thị trường, kí hiệu là X, có phân phối chuẩn.	• X không có phân phối chuẩn.
• X là sản lượng của một giống lúa, có trung bình 50 (tạ/ha), kí hiệu EX = μ, ta có $H_0 : \mu = 50$.	• $\mu \neq 50$ (hoặc $\mu > 50$, hoặc $\mu < 50$) (tạ/ha).
• Lương trung bình của nhân viên nam và nữ cùng trình độ, cùng thâm niên của tổng công ty là như nhau	• Lương trung bình của nhân viên nam và nữ cùng trình độ, cùng thâm niên của tổng công ty là khác nhau
• Nhu cầu hàng hoá X của thị trường và thu nhập Y của người tiêu dùng là độc lập.	• X và Y phụ thuộc.

2. Thủ tục kiểm định giả thuyết thống kê

1) Kiểm định giả thuyết thống kê là phương pháp của Thống kê, sử dụng số liệu mẫu điều tra cho kết luận về việc chấp nhận hay bác bỏ một giả thuyết thống kê. Chẳng hạn, để chấp nhận hay bác bỏ một báo cáo cho rằng : "Năng suất trung bình μ của giống lúa là 50 tạ/ha", người ta thiết lập cặp giả thuyết $H_0 : \mu = 50$, đối thuyết $H_1 : \mu \neq 50$ và sử dụng phương pháp kiểm định giả thuyết thống kê cho kết luận chấp nhận hay bác bỏ báo cáo từ mẫu số liệu về năng suất của giống lúa này.

2) Tiêu chuẩn kiểm định giả thuyết thống kê.

Để kiểm định giả thuyết thống kê, người ta xây dựng một tiêu chuẩn kiểm định giả thuyết thống kê từ mẫu $(X_1, X_2, ..., X_n)$ của đám đông X. Tiêu chuẩn này là một thống kê $G = G(X_1, X_2, ..., X_n)$, lập từ mẫu ngẫu nhiên kích thước n của đám đông X (có thể phụ thuộc vào tham số đã biết trong H_0), thoả mãn điều kiện : *Khi giả thiết H_0 đúng* thì *luật phân phối xác suất của G hoàn toàn được xác định.* Ví dụ : Gọi p là tỉ lệ phế phẩm của một dây chuyền công nghệ. Để kiểm định $H_0 : p = 0,1$; $H_1 : p < 0,1$ người ta sử dụng tiêu chuẩn kiểm định $G = \dfrac{F - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$ có phân phối chuẩn khi H_0 đúng (nghĩa là khi $p_0 = 0,1$) ; ở đây F là tỉ lệ mẫu và n là kích thước mẫu.

3) Các loại sai lầm. Trừ trường hợp kiểm tra được toàn bộ các phần tử của đám đông (việc kiểm tra này trong nhiều tình huống thực tế là không cần thiết hoặc không thể, chẳng hạn khi đám đông có vô hạn phần tử), còn lại nếu từ số liệu mẫu sử dụng tiêu chuẩn kiểm định giả thuyết thống kê quyết định chấp nhận hay bác bỏ giả thiết H_0, chúng ta mắc phải các sai lầm sau :

– *Sai lầm loại I* là bác bỏ giả thiết H_0 khi nó đúng.

– *Sai lầm loại II* là chấp nhận giả thiết H_0 khi nó sai.

4) Mức ý nghĩa của kiểm định giả thuyết.

Chúng ta mong muốn tìm một tiêu chuẩn kiểm định giả thuyết sao cho với tiêu chuẩn này các xác suất sai lầm loại I và xác suất sai lầm loại II đồng thời nhỏ nhất (ít sai lầm nhất). Tuy nhiên trong thực tế điều đó khó thực hiện đồng thời, do xác suất sai lầm này giảm thì sẽ tăng xác suất sai lầm kia. Vì vậy người ta đưa ra khái niệm :

– *Mức ý nghĩa* α là mức ấn định xác suất sai lầm loại I :

P{sai lầm loại I} $\leq \alpha$ (không được sai quá mức α).

5) Miền bác bỏ giả thuyết thống kê là tập hợp tất cả các giá trị của tiêu chuẩn kiểm định giả thiết G bác bỏ giả thiết H_0, với mức ý nghĩa đã cho α. Ta kí hiệu miền này là W_α.

W_α được tìm như sau : Với mức ý nghĩa α đã cho, từ phân phối xác suất của tiêu chuẩn kiểm định G, W_α là miền thoả điều kiện :

$$P\{G \in W_\alpha \mid H_0 \text{ đúng}\} = \alpha. \qquad (5.1)$$

Từ (5.1) nếu α nhỏ (α có thể : 0,1 ; 0,05 ; 0,01 ; ...) thì theo nguyên lí xác suất nhỏ [xem 3, §2, Ch.1], biến cố $\{G \in W_\alpha\}$ không xảy ra trong một phép thử của $(X_1, X_2, ..., X_n)$ với $X_1 = x_1, X_2 = x_2,..., X_n = x_n$. Vì vậy với một phép thử $(x_1, x_2, ..., x_n)$ nếu biến cố $\{G(x_1, x_2,..., x_n) \in W_\alpha\}$ xảy ra, nghĩa là H_0 sai. Khi đó ta bác bỏ H_0, chấp nhận H_1. Còn nếu $\{G(x_1, x_2,..., x_n) \in W_\alpha\}$ không xảy ra, khi đó chưa có cơ sở bác bỏ H_0 vì vậy ta chấp nhận H_0.

W_α được gọi là miền bác bỏ giả thuyết H_0 còn $\overline{W_\alpha} = \{G(x_1, x_2,..., x_n) \notin W_\alpha\}$ là miền chấp nhận H_0. Với mức ý nghĩa α, ta có :

– Xác suất sai lầm loại I là : $P\{G \in W_\alpha \mid H_0 \text{ đúng}\} = \alpha$

– Xác suất sai lầm loại II là : $P\{G \in \overline{W_\alpha} \mid H_0 \text{ sai}\} = \beta$.

$$P\left\{G > g_{\frac{\alpha}{2}} \middle| H_0 \text{ đúng}\right\} = \frac{\alpha}{2} \; ; \; P\left\{G > g_{1-\frac{\alpha}{2}} \middle| H_0 \text{ đúng}\right\} = 1 - \frac{\alpha}{2}.$$

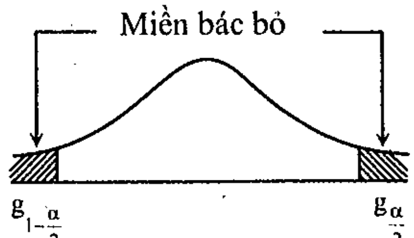
Xác suất bác bỏ giả thuyết H_0 khi nó sai là $1 - \beta$. Xác suất $1 - \beta$ được gọi là *lực lượng* của kiểm định giả thuyết.

Khi có mức ý nghĩa α, người ta tìm tiêu chuẩn kiểm định sao cho xác suất sai lầm loại II nhỏ nhất (hay lực lượng của kiểm định giả thuyết là lớn nhất). Khi này tiêu chuẩn kiểm định được gọi là *mạnh nhất.*

Với mức ý nghĩa và kích thước mẫu xác định, những miền bác bỏ được thiết lập trong các phần sau đây đều làm cho tiêu chuẩn kiểm định là mạnh nhất.

Tùy theo đối thuyết H_1, cho mức ý nghĩa α, với giả thuyết $H_0 : \theta = \theta_0$ ta chọn một trong ba dạng miền bác bỏ W_α của tiêu chuẩn kiểm định G như sau :

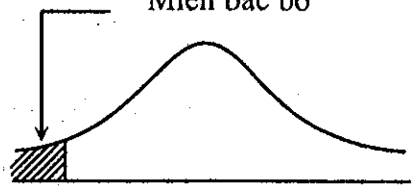
• Trường hợp đối thuyết hai phía : $H_1 : \theta \neq \theta_0$. Khi H_0 đúng, từ phân phối xác suất của G ta tìm được các giá trị tới hạn $g_{\frac{\alpha}{2}}$, $g_{1-\frac{\alpha}{2}}$ của G thoả mãn :



Hình 5.1

Miền bác bỏ $W_\alpha = (G < g_{1-\frac{\alpha}{2}}) \cup (g_{\frac{\alpha}{2}} < G)$ (H 5.1) (5.2)

Trường hợp đối thuyết một phía :



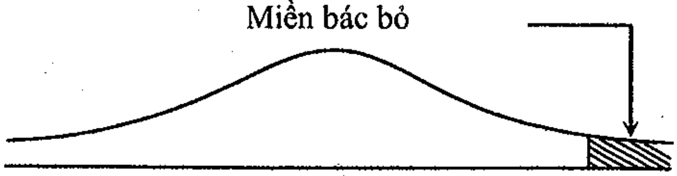
Hình 5.2

• $H_1 : \theta < \theta_0$ (lệch bên trái)

Ta lấy $g_{1-\alpha} : P\left\{G > g_{1-\alpha} \middle| H_0\right\} = 1 - \alpha$

từ đó suy ra $P\{G < g_{1-\alpha} \mid H_0\} = \alpha$.

$W_\alpha = (G < g_{1-\alpha})$. (H 5.2) (5.3)



Hình 5.3

• $H_1 : \theta > \theta_0$ (lệch bên phải)

Lấy $g_\alpha : P\{G > g_\alpha \mid H_0\} = \alpha$.

Ta có $W_\alpha = (g_\alpha < G)$. (H 5.3) (5.4)

Chú ý : Nếu **G** có phân phối chuẩn chuẩn tắc hay phân phối Student thì $g_{1-\alpha} = -g_\alpha$. Khi đó miền bác bỏ (5.2) là $W_\alpha = (G < -g_{\frac{\alpha}{2}}) \cup (g_{\frac{\alpha}{2}} < G) = \left\{|G| > g_{\frac{\alpha}{2}}\right\}$; Miền bác bỏ (5.3) là $W_\alpha = (G < -g_\alpha)$; Miền bác bỏ (5.4) là $W_\alpha = (G > g_\alpha)$.

Ví dụ 5.2 : Khi nghiên cứu tỉ lệ p các phần tử của đám đông X bằng hay không số p_0 (đã cho). Người ta lập giả thuyết $H_0 : p = p_0$ và dùng tiêu chuẩn $Z = (F - p_0)\Big/\sqrt{\dfrac{p_0(1-p_0)}{n}} \approx N(0, 1)$ (khi kích thước mẫu n đủ lớn) để kiểm định giả thuyết này. Với mức ý nghĩa α, miền bác bỏ được tìm theo các đối thuyết như sau :

+ Với đối thuyết hai phía $H_1 : p \neq p_0$, tìm giá trị tới hạn chuẩn mức $\alpha/2$ của Z là $z_{\alpha/2}$: $P\left\{Z = (F - p_0)\Big/\sqrt{\dfrac{p_0(1-p_0)}{n}} > z_{\alpha/2}\right\} = \dfrac{\alpha}{2}$ và giá trị tới hạn chuẩn mức $1 - \alpha/2$ là $z_{1-\alpha/2} = -z_{\alpha/2}$.

$$P\left\{Z = (F - p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} > z_{1-\alpha/2}\right\} = 1 - \frac{\alpha}{2}$$

$$= P\left\{Z = (F - p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} > -z_{\alpha/2}\right\}.$$

Ta có :

$$P\left\{(F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} < -z_{\alpha/2}\right\} + P\left\{(F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} > z_{\alpha/2}\right\} = \alpha .$$

Vậy miền bác bỏ H_0 là :

$$W_\alpha = \left\{|Z| > z_{\alpha/2} : Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}}\right\} \qquad (5.5)$$

+ Với đối thuyết một phía lệch bên trái $H_1 : p < p_0$, tìm giá trị tới hạn chuẩn mức $1-\alpha$ là $z_{1-\alpha} = -z_\alpha$ như sau

$$P\left\{Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} > z_{1-\alpha}\right\} = 1-\alpha$$

$$= P\left\{Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} > -z_\alpha\right\} \Rightarrow P\left\{Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} < -z_\alpha\right\} = \alpha.$$

Vậy miền bác bỏ H_0 là :

$$W_\alpha = \left\{Z < -z_\alpha : Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}}\right\} \qquad (5.6)$$

+ Với đối thuyết một phía lệch bên phải : $H_1 : p > p_0$, tìm giá trị tới hạn chuẩn mức α là z_α như sau

$$P\left\{Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}} > z_\alpha\right\} = \alpha .$$

Vậy miền bác bỏ H_0 là : $W_\alpha = \left\{Z > z_\alpha : Z = (F-p_0)\Big/\sqrt{\frac{p_0(1-p_0)}{n}}\right\}$ (5.7)

3. Các bước tiến hành kiểm định giả thuyết thống kê

1) Phát biểu giả thuyết H_0 và đối thuyết H_1 ;

2) Định mức ý nghĩa α ;

3) Chọn tiêu chuẩn kiểm định G ;

4) Thiết lập miền bác bỏ H_0 : W_α ;

5) Kết luận : Từ mẫu cụ thể $(x_1, ..., x_n)$ tính $g = G(x_1, ..., x_n)$

+ $g = G(x_1, ..., x_n) \in W_\alpha$: Bác bỏ H_0 chấp nhận H_1.

+ $g = G(x_1, ..., x_n) \notin W_\alpha$: Chấp nhận H_0.

4. Quan hệ giữa bài toán kiểm định giả thuyết và bài toán ước lượng khoảng

Bài toán kiểm định giả thuyết $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$ với mức ý nghĩa α tương đương với bài toán tìm khoảng tin cậy $100(1 - \alpha)\%$ cho θ, nếu tiêu chuẩn kiểm định giả thuyết G là biến ngẫu nhiên liên tục. Bác bỏ giả thuyết H_0 khi θ_0 nằm ngoài khoảng tin cậy này, chấp nhận H_0 khi θ_0 nằm trong khoảng tin cậy. Chẳng hạn, với cặp giả thuyết $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ và tiêu chuẩn kiểm định là $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$, khi H_0 đúng với mức ý nghĩa α, theo (5.2) nếu chấp nhận H_0, nghĩa là : $-z_{\frac{\alpha}{2}} \leq Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\frac{\alpha}{2}}$ hay

$$\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

suy ra μ_0 thuộc khoảng tin cậy của trung bình μ.

Tương tự, kiểm định giả thuyết với đối thuyết một phía $H_1 : \theta > \theta_0$ (hay $H_1 : \theta < \theta_0$) và mức ý nghĩa α, tương đương với bài toán tìm khoảng tin cậy một phía với độ tin cậy $100(1 - \alpha)\%$ cho θ, bác bỏ H_0 khi θ_0 không thuộc khoảng tin cậy tối thiểu cho θ (hay θ_0 không thuộc khoảng tin cậy tối đa cho θ).

5. Phương pháp P – giá trị kiểm định giả thuyết

Khi kiểm định giả thuyết thống kê nếu tăng (giảm) mức ý nghĩa thì xác suất mắc sai lầm loại II giảm (tăng). Vì vậy với cùng một cặp giả thuyết, cùng một tiêu chuẩn kiểm định và số liệu mẫu nhưng các nhà nghiên cứu chấp nhận mức xác suất sai lầm loại II khác nhau, nên họ sử dụng mức ý nghĩa khác nhau tùy theo mục đích nghiên cứu của mình. Ngoài phương pháp truyền thống để kiểm định giả thuyết như trình bày trong mục 3, người ta còn sử dụng phương pháp khác gọi là P – giá trị như sau :

Khi H_0 đúng, tiêu chuẩn G có phân phối xác suất xác định, từ số liệu của mẫu người ta xác định mức xác suất nhỏ nhất để bác bỏ giả thuyết H_0. Mức xác suất này được gọi là P – giá trị.

Định nghĩa 5.1 : P – giá trị (hay mức ý nghĩa của quan sát) của thống kê G là mức xác suất nhỏ nhất bác bỏ giả thuyết H_0 khi sử dụng tiêu chuẩn kiểm định G, dựa trên số liệu mẫu đã cho.

Khi P – giá trị của tiêu chuẩn G được xác định, với mức ý nghĩa α bất kì, ta có kết luận :

1) Nếu P – giá trị $\leq \alpha$ thì bác bỏ H_0 với mức ý nghĩa α.

2) Nếu P – giá trị $> \alpha$ thì chấp nhận H_0 với mức ý nghĩa α.

Ngày nay, hầu hết các phần mềm thống kê đều tính được P – giá trị.

Ví dụ 5.3 : Cho thống kê G có phân phối xác suất xác định. Giá trị của G được tính từ mẫu cụ thể $(x_1, ..., x_n)$ của đám đông X là $g = G(x_1, ..., x_n)$. P – giá trị của thống kê G được xác định như sau :

Với đối thuyết $H_1 : \theta > \theta_0$. P – giá trị = $P(G \geq g)$ (5.8)

Với đối thuyết $H_1 : \theta < \theta_0$. P – giá trị = $P(G \leq g)$ (5.9)

Với đối thuyết hai phía $H_1 : \theta \neq \theta_0$,

$$P - \text{giá trị} = \begin{cases} 2P(G \geq g), \text{ khi } P(G \geq g) < 0{,}5 \\ 2P(G \leq g), \text{ khi } P(G \geq g) > 0{,}5 \end{cases} \quad (5.10)$$

Chú ý : Với đối thuyết hai phía $H_1 : \theta \neq \theta_0$, nếu G có phân phối chuẩn chuẩn tắc hay phân phối Student thì P – giá trị = $2P(G \geq |g|)$ (5.11)

GIẢI : Từ số liệu mẫu $(x_1, ..., x_n)$, ta tính được giá trị của G là g. Ta xét các trường hợp :

– Với đối thuyết $H_1 : \theta > \theta_0$.

Cho mức ý nghĩa α ta tìm giá trị tới hạn mức α của G là g_α.

Nếu $P(G \geq g) \leq \alpha = P(G > g_\alpha)$ thì $(G > g) \subset (G > g_\alpha) \Leftrightarrow g > g_\alpha$

$\Leftrightarrow g \in W_\alpha = (g_\alpha < G)$: Bác bỏ H_0. Theo định nghĩa 5.1 P – giá trị của G là $P(G > g)$.

– Với đối thuyết : $H_1 : \theta < \theta_0$.

Nếu $P(G \leq g) \leq \alpha = P(G > g_\alpha)$

$\Leftrightarrow P(G > g) \geq 1 - \alpha = P(G > g_{1-\alpha}) \Leftrightarrow g < g_{1-\alpha}$

$\Leftrightarrow g \in W_\alpha = (G < g_{1-\alpha})$: Bác bỏ H_0.

Theo định nghĩa 5.1, P – giá trị của G là $P(G \leq g)$.

– Với đối thuyết hai phía : $H_1 : \theta \neq \theta_0$.

Ta có : $g \in W_\alpha = (G < g_{1-\frac{\alpha}{2}}) \cup (G > g_{\frac{\alpha}{2}}) \Leftrightarrow g \in (G < g_{1-\frac{\alpha}{2}})$ hay $g \in (G > g_{\frac{\alpha}{2}})$

Khi $P(G \geq g) < 0{,}5$ nếu : $2P(G \geq g) \leq \alpha$

$$\Leftrightarrow P(G \ge g) \le \frac{\alpha}{2} = P(G > g_{\frac{\alpha}{2}}) \Leftrightarrow g > g_{\frac{\alpha}{2}} \Leftrightarrow g \in \left\{G > g_{\frac{\alpha}{2}}\right\}.$$

Khi $P(G \ge g) > 0{,}5$ nếu : $2P(G \le g) \le \alpha$

$$\Leftrightarrow P(G \le g) \le \frac{\alpha}{2} = P(G > g_{\frac{\alpha}{2}}) \Leftrightarrow P(G \ge g) \ge 1 - \frac{\alpha}{2} = P(G > g_{1-\frac{\alpha}{2}})$$

$$\Leftrightarrow g < g_{1-\frac{\alpha}{2}} \Leftrightarrow g \in \left\{G < g_{1-\frac{\alpha}{2}}\right\}.$$

Từ các trường hợp trên suy ra (5.10).

Khi G có phân phối chuẩn chuẩn tắc hay phân phối Student, từ (5.10) ta suy ra (5.11) như sau :

Nếu $P(G \ge g) < 0{,}5$ thì $g > 0$ suy ra $2P(G \ge g) = 2P(G \ge |g|)$;

Nếu $P(G \ge g) > 0{,}5$ thì $g < 0$ vì vậy $2P(G \le g) = 2P(G \ge -g) = 2P(G \ge |g|)$.

§2. KIỂM ĐỊNH GIẢ THUYẾT VỀ THAM SỐ

1. Kiểm định giả thuyết về tỉ lệ đám đông

a) Bài toán : Giả sử p là tỉ lệ các phần tử có tính chất nào đó của đám đông X chưa biết. Với mức ý nghĩa α, kiểm định giả thuyết $H_0 : p = p_0$ (p_0 đã biết).

Trong điều kiện giả thuyết H_0 đúng, khi n đủ lớn (lấy n sao cho $np_0 \ge 10$ và $n(1 - p_0) \ge 10$) tiêu chuẩn kiểm định sau có phân phối xấp xỉ chuẩn chuẩn tắc

$$Z = (F - p_0) \Big/ \sqrt{\frac{p_0(1-p_0)}{n}} \approx N(0 ; 1).$$

Theo Mục §1, Mục §3 và Ví dụ 2, ta suy ra quy tắc thực hành :

b) Quy tắc thực hành : Trên mẫu cụ thể $(x_1, x_2, ..., x_n)$.

• Giả thuyết $H_0 : p = p_0$ và đối thuyết H_1.

• Tính giá trị của tiêu chuẩn kiểm định Z từ số liệu mẫu :

+ $f = \dfrac{m}{n}$, (m là số phần tử của mẫu có tính chất cần nghiên cứu, n là kích thước mẫu).

$+ z = (f - p_0) \Big/ \sqrt{\dfrac{p_0(1-p_0)}{n}}$.

• Tìm giá trị tới hạn chuẩn : $z_{\frac{\alpha}{2}}$ (với đối thuyết 2 phía) hay z_α (với đối thuyết 1 phía).

• Kết luận : (So sánh giá trị z và các giá trị tới hạn chuẩn)

Đối thuyết	**Bác bỏ H_0, chấp nhận H_1**	**Chấp nhận H_0**	
• $H_1 : p \neq p_0$	$W_\alpha = \{\lvert z\rvert > z_{\alpha/2}\}$	$\overline{W}_\alpha = \{\lvert z\rvert \le z_{\alpha/2}\}$	(5.12)
• $H_1 : p < p_0$	$W_\alpha = \{z < -z_\alpha\}$	$\overline{W}_\alpha = \{z \ge -z_\alpha\}$	(5.13)
• $H_1 : p > p_0$	$W_\alpha = \{z > z_\alpha\}$	$\overline{W}_\alpha = \{z \le z_\alpha\}$	(5.14)

Ở đây $z_{\alpha/2}$, z_α là giá giá trị tới hạn chuẩn mức $\frac{\alpha}{2}$, α [Bảng 2].

• *Chú ý :* 1) Với đối thuyết hai phía (5.12), khi bác bỏ H_0 chấp nhận H_1, ta so sánh f và p_0 để kết luận : Nếu $f < p_0$ (hay $f > p_0$) thì $p < p_0$ ($p > p_0$).

2) Quy tắc thực hành (5.13) được áp dụng cho cả cặp giả thuyết $H_0 : p \ge p_0$ và đối thuyết $H_1 : p < p_0$; Quy tắc thực hành (5.14) được áp dụng cho cả cặp giả thuyết H_0 : $p \le p_0$ và đối thuyết $H_1 : p > p_0$.

Ví dụ 5.4 : Ở một nước, một đảng chính trị tuyên bố rằng 45% cử tri sẽ bỏ phiếu bầu cho ông A là ứng cử viên của họ. Chọn ngẫu nhiên 200 người hỏi ý kiến có 80 người sẽ bầu cho ông A. Với mức ý nghĩa 5% hãy cho nhận xét về tuyên bố trên.

GIẢI : Cách 1. Dùng phương pháp miền bác bỏ.

Gọi p là tỉ lệ cử tri sẽ bỏ phiếu bầu cho ông A. Ta có :

$$H_0 : p = 0{,}45 \ ; \ \ H_1 : p \neq 0{,}45$$

• $f = \dfrac{80}{200} = 0{,}4$; $z = \dfrac{(0{,}4 - 0{,}45)}{\sqrt{0{,}45.0{,}55/200}} = -1{,}43$.

• $\alpha = 0{,}05 \rightarrow \dfrac{1-\alpha}{2} = 0{,}5 - \dfrac{\alpha}{2} = 0{,}475 = \varphi(1{,}96)$: Giá trị tới hạn chuẩn $z_{\alpha/2} = 1{,}96$

• $\lvert z\rvert < z_{\alpha/2}$ chấp nhận H_0.

Với mức ý nghĩa 5% chưa có cơ sở bác bỏ tuyên bố trên.

Cách 2. Dùng phương pháp P – giá trị.

Sử dụng thống kê $Z = (F-0,4)\Big/\sqrt{\dfrac{0,4(1-0,4)}{200}} \approx N(0\ ;\ 1)$, từ mẫu ta tính được

$$z = \frac{(0,4-0,45)}{\sqrt{0,45.0,55/200}} = -1,43.$$

Từ Ví dụ 5.3, với đối thuyết hai phía, P – giá trị của Z bằng :

$$2P\{Z \geq 1,43\} = 1 - 2\varphi(1,43) = 0,15272.$$

Vì P – giá trị của Z bằng 0,15272 > $\alpha = 0,05$ nên ta chấp nhận H_0.

Ví dụ 5.5 : Một công ty tuyên bố rằng 40% người tiêu dùng ưa thích sản phẩm của công ty. Để xem xét tuyên bố đó, người ta điều tra 400 người tiêu dùng thấy có 120 người ưa thích sản phẩm của công ty. Với mức ý nghĩa 5%, hãy xem tỉ lệ trong tuyên bố của công ty có cao hơn thực tế không?

GIẢI : Cách 1. Gọi p là tỉ lệ người tiêu dùng ưa thích sản phẩm của công ty (p là tỉ lệ thực tế). Ta có :

$$H_0 : p = 0,4\ ;\ H_1 : p < 0,4$$

- $f = \dfrac{120}{400} = 0,3\ ;\quad z = \dfrac{0,3-0,4}{\sqrt{\dfrac{0,4.0,6}{400}}} \approx -4,082$

- $\alpha = 0,05 \rightarrow 0,5 - \alpha = 0,45 = \varphi(1,65) \rightarrow z_\alpha = 1,65$

- $z < -z_\alpha$: Bác bỏ H_0 chấp nhận H_1.

Với mức ý nghĩa 5%, tỉ lệ trong tuyên bố của công ty là cao hơn thực tế.

Cách 2. Từ Ví dụ 5.3, P – giá trị của

$$Z = (F-0,4)\Big/\sqrt{\frac{0,4(1-0,4)}{200}} \approx N(0\ ;\ 1) \text{ là}$$

$P\{Z \leq -4,082\} = \varphi(-4,082) + 0,5 = 0,5 - \varphi(4,082) \approx 0,000032.$

Vì P – giá trị của $Z = 0,000032 < \alpha = 0,05$ ta bác bỏ H_0, chấp nhận H_1.

2. Kiểm định giả thuyết về trung bình của đám đông

a) Bài toán : Giả sử đám đông X có $EX = \mu$ chưa biết. Với mức ý nghĩa α, kiểm định giả thuyết $H_0 : \mu = \mu_0$ (μ_0 đã biết).

Khi giả thuyết H_0 đúng, tiêu chuẩn kiểm định G được xác định theo các trường hợp sau [xem Định lí 4.2, Định lí 4.4, Ch.4].

• **Trường hợp 1 :** Đám đông X có phân phối chuẩn, biết phương sai $VX = \sigma^2$ khi đó tiêu chuẩn : $Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0\,;1)$.

• **Trường hợp 2 :** Đám đông X có phân phối chuẩn chưa biết, phương sai σ^2 khi đó tiêu chuẩn : $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n} \sim t(n-1)$ (thống kê T có phân phối Student n – 1 bậc tự do), ở đây $S = \sqrt{S^2}$ là độ lệch chuẩn của mẫu.

Chú ý. Khi kích thước mẫu n đủ lớn, các thống kê $\dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$ hay $\dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$ đều có phân phối xấp xỉ chuẩn chuẩn tắc (dù đám đông không có phân phối chuẩn) [xem Định lí 4.5, Ch.4] và thống kê $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n} \sim t(n-1)$ thì $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$ có phân phối xấp xỉ chuẩn chuẩn tắc [xem 3.2, 3, §1, Ch.3]. Trong thực hành khi $n \geq 30$, với nhiều trường hợp người ta thấy các xấp xỉ chuẩn này là đủ tốt.

b) Quy tắc thực hành : Trên một mẫu cụ thể $(x_1, ..., x_n)$.

• Lập giả thuyết $H_0 : \mu = \mu_0$ (μ_0 đã biết), đối thuyết H_1.

• Tính giá trị của tiêu chuẩn kiểm định G từ số liệu mẫu : $g = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$ (trường hợp chưa biết phương sai) hoặc $g = \dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$ (khi biết phương sai).

• Kết luận :

Trường hợp 1 : Tìm giá trị tới hạn chuẩn

– Đám đông X có phân phối chuẩn hoặc

– Đám đông X không có phân phối chuẩn nhưng kích thước mẫu lớn ($n \geq 30$).

Đối thuyết	**Bác bỏ H_0, chấp nhận H_1**	**Chấp nhận H_0**	
• $H_1 : \mu \neq \mu_0$	$W_\alpha = \{\lvert g\rvert > z_{\alpha/2}\}$	$\overline{W}_\alpha = \{\lvert g\rvert \leq z_{\alpha/2}\}$	(5.15)
• $H_1 : \mu < \mu_0$	$W_\alpha = \{g < -z_\alpha\}$	$\overline{W}_\alpha = \{g \geq -z_\alpha\}$	(5.16)
• $H_1 : \mu > \mu_0$	$W_\alpha = \{g > z_\alpha\}$	$\overline{W}_\alpha = \{g \leq z_\alpha\}$	(5.17)

Ở đây $z_{\alpha/2}$, z_α là giá trị tới hạn chuẩn mức $\dfrac{\alpha}{2}$, α [Bảng 2].

Trường hợp 2 : Tìm giá trị tới hạn Student.

Đám đông có phân phối chuẩn, phương sai chưa biết ($n < 30$).

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0	
• $H_1 : \mu \neq \mu_0$	$W_\alpha = \left\{ \lvert g \rvert > t\left(n-1, \frac{\alpha}{2}\right) \right\}$	$\overline{W}_\alpha = \left\{ \lvert g \rvert \le t\left(n-1, \frac{\alpha}{2}\right) \right\}$	(5.18)
• $H_1 : \mu < \mu_0$	$W_\alpha = \{g < -t(n-1;\alpha)\}$	$\overline{W}_\alpha = \{g \ge -t(n-1;\alpha)\}$	(5.19)
• $H_1 : \mu > \mu_0$	$W_\alpha = \{g > t(n-1\,;\,\alpha)\}$	$\overline{W}_\alpha = \{g \le t(n-1\,;\,\alpha)\}$	(5.20)

Ở đây $t\left(n-1, \frac{\alpha}{2}\right)$; $t(n-1,\alpha)$ tương ứng là các giá trị tới hạn Student mức $\frac{\alpha}{2}$, α [Bảng 3].

• *Chú ý :* 1) Trường hợp 2 có thể áp dụng cho kích thước mẫu bất kì. Tuy nhiên theo Định lí 4.5 [Ch.4] về xấp xỉ chuẩn, khi mẫu có kích thước lớn $n \ge 30$ người ta có thể thực hành theo Trường hợp 1 với quy ước khi đó các xấp xỉ chuẩn này là đủ tốt.

2) Với đối thuyết hai phía (5.15), (5.18), khi bác bỏ H_0 chấp nhận H_1, ta so sánh $\overline{x}$ và μ_0 để kết luận : Nếu $\overline{x} < \mu_0$ (hay $\overline{x} > \mu_0$) thì $\mu < \mu_0$ ($\mu > \mu_0$).

3) Quy tắc thực hành cho đối thuyết một phía (5.16), (5.19) được áp dụng cho cả cặp giả thuyết $H_0 : \mu \ge \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$; tương tự (5.17), (5.20) áp dụng cho cặp giả thuyết $H_0 : \mu \le \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$.

Ví dụ 5.6 : Trọng lượng của một hộp sản phẩm do một máy tự động đóng theo quy định là 6kg. Sau một thời gian máy sản xuất, người ta tiến hành kiểm tra 121 hộp sản phẩm tính được trọng lượng trung bình $\overline{x} = 5{,}975$kg và phương sai mẫu là 5,7596. Với mức ý nghĩa 5% hãy cho kết luận về sự hoạt động của máy.

GIẢI : Cách 1. Gọi μ (kg) là trọng lượng trung bình của một hộp sản phẩm do máy đóng. Ta kiểm định cặp giả thuyết và đối thuyết :

H_0 : Máy hoạt động bình thường ;

H_1 : Máy hoạt động không bình thường

Các điều trên tương đương với $\Leftrightarrow \begin{cases} H_0 : \mu = 6 \\ H_1 : \mu \neq 6 \end{cases}$

Kích thước mẫu $n = 121$ ($n \ge 30$), $\overline{x} = 5{,}975$; $s^2 = 5{,}7596$.

$$g = \frac{5{,}975 - 6}{\sqrt{5{,}7596}}\sqrt{121} = -0{,}1146$$

• $\alpha = 0{,}05$; $z_{\frac{\alpha}{2}} = 1{,}96$

• $|g| < z_{\frac{\alpha}{2}}$ chấp nhận H_0, nghĩa là : Với mức ý nghĩa 5% chưa có cơ sở để cho rằng máy hoạt động không bình thường.

Cách 2. Ta tìm khoảng tin cậy (1 – 0,05).100% = 95% của μ là :

$$\left(\overline{x} - 1{,}96\sqrt{\frac{5{,}7596}{121}}\ ;\ \overline{x} + 1{,}96\sqrt{\frac{5{,}7596}{121}}\right) = (5{,}5474\ ;\ 6{,}4026).$$

Vì $\mu_0 = 6 \in (5{,}5474\ ;\ 6{,}4026)$, ta chấp nhận H_0.

Chú ý. Bài toán này có thể giải bằng phương pháp P – giá trị.

Ví dụ 5.7 : Trọng lượng của một bao gạo là biến ngẫu nhiên có phân phối chuẩn trung bình 50kg. Nghi ngờ máy đóng bao gạo làm việc không bình thường làm cho trọng lượng của bao gạo có xu hướng giảm sút. Người ta cân thử 25 bao tính được $\overline{x}$ = 49,27kg và độ lệch tiêu chuẩn mẫu s = 0,49kg. Với mức ý nghĩa 1% hãy cho kết luận về nghi ngờ trên.

GIẢI : Gọi μ (kg) là trọng lượng trung bình của bao gạo do máy đóng gói.

Ta kiểm định cặp giả thuyết và đối thuyết :

H_0 : Máy hoạt động bình thường ;

H_1 : Máy hoạt động không bình thường làm cho trọng lượng bao gạo có xu hướng giảm sút.

Các điều trên tương đương với $H_0 : \mu = 50$; $H_1 : \mu < 50$.

Kích thước mẫu n = 25 (n < 30), phương sai của đám đông chưa biết.

Cách 1. Phương pháp miền bác bỏ

$$g = \frac{(49{,}27 - 50)}{0{,}49}\sqrt{25} = -7{,}45\ ;$$

$\alpha = 0{,}01$; $t(24\ ;\ \alpha) = 2{,}492$ (n = 25) ;

$g < -t(24\ ;\ \alpha)$ bác bỏ H_0 chấp nhận H_1.

Với mức ý nghĩa 1%, nghi ngờ trên là có cơ sở : Trọng lượng của bao gạo có xu hướng giảm sút.

Cách 2. Phương pháp P – giá trị giá trị của $G = \frac{\overline{X} - 50}{S}\sqrt{25} \sim t(24)$. Từ Ví dụ 5.3, sử dụng Bảng 3, ta có $t(24\ ;\ 0{,}005) = 2{,}797$ và

$$P\{G \le -7{,}45\} = P\{G \ge 7{,}45\} < P\{G > 2{,}797\} = 0{,}005.$$

Suy ra $\alpha = 0{,}01 >$ P – giá trị của G : Bác bỏ H_0 chấp nhận H_1.

Cách 3. Khoảng tin cậy $(1 - 0{,}05).100\% = 0{,}95\%$ bên trái (ước lượng tối đa) của μ là : $\mu < 49{,}27 + 2{,}492.\frac{0{,}49}{\sqrt{25}} = 49{,}5142$.

Do $\mu_0 = 50 \notin (-\infty\ ;\ 49{,}5142)$: Bác bỏ H_0.

c) Các bài toán liên quan tới bài toán kiểm định giả thuyết

** Tìm xác suất sai lầm loại 2*

Với giả thuyết $H_0 : \mu = \mu_0$; $H_1 : \mu < \mu_0$ (hoặc $H_1 : \mu > \mu_0$).

Nếu μ_1 là giá trị đúng của μ thì từ tiêu chuẩn kiểm định G ta tìm được xác suất sai lầm loại hai β.

Chẳng hạn : $G = \frac{\overline{X} - \mu_1}{\sigma}\sqrt{n} \approx N(0\ ;\ 1)$ thì

+ Với đối thuyết $H_1 : \mu < \mu_0$. Xác suất chấp chận H_0 khi nó sai :

$$\beta = P\left\{\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \ge -z_\alpha\right\} = P\left\{\frac{\overline{X} - \mu_1}{\sigma}\sqrt{n} \ge -z_\alpha - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right\}$$

$$= P\left\{\frac{\overline{X} - \mu_1}{\sigma}\sqrt{n} \le z_\alpha + \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right\} = 0{,}5 + \varphi\left(z_\alpha + \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right) \tag{5.21}$$

+ Tương tự với $H_1 : \mu > \mu_0$. Xác suất chấp chận H_0 khi nó sai :

$$\beta = P\left\{\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \le z_\alpha\right\} = P\left\{\frac{\overline{X} - \mu_1}{\sigma}\sqrt{n} \le z_\alpha - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right\}$$

$$= 0{,}5 + \varphi\left(z_\alpha - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right) \tag{5.22}$$

z_α là giá trị tới hạn chuẩn mức α của G.

+ Với giả thuyết : $H_0 : \mu = \mu_0$; $H_1 : \mu \ne \mu_0$ ta có khi

$$G = \frac{\overline{X} - \mu_1}{\sigma}\sqrt{n} \approx N(0\,;\,1) \text{ thì } \beta = P\left\{-z_{\frac{\alpha}{2}} \le \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \le z_{\frac{\alpha}{2}}\right\} \quad (5.23)$$

$$= P\left\{-z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n} \le G \le z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right\}$$

$$= \varphi\left\{z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right\} - \varphi\left\{-z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right\}. \quad (5.24)$$

$z_{\frac{\alpha}{2}}$ là giá trị tới hạn chuẩn mức $\frac{\alpha}{2}$ của G.

Với các tiêu chuẩn kiểm định G khác, ta làm tương tự.

Ví dụ 5.8. (Trở lại Ví dụ 5.6) : Giả sử giá trị đúng của μ là 6,2. Tìm xác suất sai lầm loại II và lực lượng của kiểm định.

GIẢI : Với $H_0 : \mu = 6$; $H_1 : \mu \neq 6$ và $\alpha = 0{,}05$.

Tiêu chuẩn $G = \frac{\overline{X} - 6{,}2}{s}\sqrt{n} \approx N(0\,;\,1)$ với $\mu_0 = 6$; $s^2 = 5{,}7596$.

Ta có $z_{\frac{\alpha}{2}} = z_{0,025} = 1{,}96$.

$$\beta = \varphi(z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}) - \varphi(-z_{\frac{\alpha}{2}} - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n})$$

$$= \varphi(1{,}04) - \varphi(-2{,}88) = \varphi(1{,}04) + \varphi(2{,}88) = 0{,}84884.$$

Lực lượng của kiểm định : $1 - \beta = 0{,}15116$.

* *Tìm kích thước mẫu khi cho mức ý nghĩa α, xác suất sai lầm loại hai β.*

Bài toán : Tìm kích thước mẫu điều tra n tối thiểu sao cho khi kiểm định cặp $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (hay $H_1 : \mu < \mu_0$), xác suất mắc sai lầm loại I là α và xác suất mắc sai lầm loại II không vượt quá β nhưng giá trị thực của μ sai lệch so với μ_0 không vượt quá Δ (trong đó α, β, Δ là các số cho trước).

Giả sử ta có tiêu chuẩn $G = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \approx N(0\,;\,1)$. Khi đó gọi z_α, z_β là các giá trị tới hạn chuẩn mức α, β ta có :

$$z_\alpha + z_\beta = z_\alpha - z_{1-\beta} = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} - \frac{\overline{X}-\mu_1}{\sigma}\sqrt{n} = \frac{\mu_1-\mu_0}{\sigma}\sqrt{n} \qquad (5.25)$$

(Ở đây ta giả sử $\overline{X}$ là trung bình mẫu của đám đông X có kích thước n thoả mãn cá yêu cầu bài toán và μ_0, μ_1 là giá trị thực của μ tương ứng với xác suất α, β).

Từ (5.25) suy ra :

$$n = \frac{\sigma^2\left(z_\alpha + z_\beta\right)^2}{\left(\mu_1-\mu_0\right)^2} \text{ và } n \geq \frac{\sigma^2\left(z_\alpha + z_\beta\right)^2}{\Delta^2}, \text{ nếu } \left|\mu_1 - \mu_0\right| \leq \Delta. \qquad (5.26)$$

+ Tương tự với đối thuyết hai phía $H_0 : \mu = \mu_0$ và $H_1 : \mu \neq \mu_0$

ta có : $$n \geq \frac{\sigma^2\left(z_{\alpha/2} + z_\beta\right)^2}{\Delta^2} \qquad (5.27)$$

z_α, $z_{\alpha/2}$, z_β các giá trị tới hạn chuẩn của G.

Ví dụ 5.9 : Người ta muốn kiểm định giả thuyết $H_0 : \mu = 370$ và $H_1 : \mu \neq 370$ đí kiểm tra một kết luận trong một tài liệu cho rằng : "tuổi thọ trung bình của loại bóng đè A là 370 giờ với độ lệch tiêu chuẩn 20 giờ", biết rằng tuổi thọ trung bình của bóng đè không vượt quá 380 giờ. Nếu muốn mức ý nghĩa của kiểm định $\alpha = 0{,}01$ và xác suất sa lầm loại II là β không vượt quá 0,05, thì cần điều tra tối thiểu bao nhiêu bóng đèn A ?

GIẢI : $\alpha = 0{,}01 \Rightarrow z_{\alpha/2} = 2{,}58$; $\beta = 0{,}05 \Rightarrow z_\beta = 1{,}65$;

$$\left|\mu_1 - \mu_0\right| \leq 10 \text{ từ đó : } n \geq \frac{(20)^2(2{,}58+1{,}65)^2}{10^2} = 71{,}5716.$$

Vì n là số nguyên nên kích thước mẫu n tối thiểu bằng 72.

Vậy cần điều tra tối thiểu 72 bóng đèn.

3. Kiểm định giả thuyết về phương sai của đám đông

a) Bài toán : Giả sử đám đông X có phân phối chuẩn $N(\mu ; \sigma^2)$ phương sai $V(X) = \sigma^2$ chưa biết. Với mức ý nghĩa α, kiểm định giả thuyết $H_0 : \sigma^2 = \sigma_0^2$ (σ_0^2 đã biết) và mộ trong các đối thuyết $H_1 : \sigma^2 \neq \sigma_0^2$; $H_1 : \sigma^2 < \sigma_0^2$; $H_1 : \sigma^2 > \sigma_0^2$.

• *Trường hợp 1 :* Trung bình của đám đông EX = μ *chưa biết.*

Khi giả thuyết H_0 đúng, theo Định lí 4.3[Ch.4], tiêu chuẩn kiểm định

$$\chi^2 = \frac{\sum_{i=1}^{k} n_i\left(X_i - \overline{X}\right)^2}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

• *Trường hợp 2 :* Trung bình của đám đông EX = μ *đã biết.*

Khi giả thuyết H_0 đúng, theo Định lí 4.3 [Ch.4], tiêu chuẩn kiểm định

$$\chi^2 = \frac{\sum (X_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n).$$

Như vậy trường hợp 2, ta làm tương tự như trường hợp 1 với chú ý bậc tự do của phân phối là n.

b) Quy tắc thực hành : Trên mẫu cụ thể $(x_1, x_2, ..., x_n)$.

+ Giả thuyết $H_0 : \sigma^2 = \sigma_0^2$ (σ_0^2 đã biết) ; đối thuyết H_1

Trường hợp 1 : Trung bình của đám đông EX = μ *chưa biết*

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0
$H_1 : \sigma^2 \neq \sigma_0^2$	$W_\alpha = \left\{\chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n-1); \chi^2 > \chi^2_{\frac{\alpha}{2}}(n-1)\right\}$	$\overline{W}_\alpha = \left\{\chi^2_{1-\frac{\alpha}{2}}(n-1) \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2}}(n-1)\right\}$ (5.28)
$H_1 : \sigma^2 < \sigma_0^2$	$W_\alpha = \left\{\chi^2 < \chi^2_{1-\alpha}(n-1)\right\}$	$\overline{W}_\alpha = \left\{\chi^2 \geq \chi^2_{1-\alpha}(n-1)\right\}$ (5.29)
$H_1 : \sigma^2 > \sigma_0^2$	$W_\alpha = \left\{\chi^2 > \chi^2_\alpha(n-1)\right\}$	$\overline{W}_\alpha = \left\{\chi^2 \leq \chi^2_\alpha(n-1)\right\}$ (5.30)

Tìm $\chi^2_\alpha(n-1) : \alpha \xrightarrow[\text{bậc tự do } n-1]{\text{Bảng 5 (Khi bình phương)}} \chi^2_\alpha(n-1)$.

Trường hợp 2 : Biết trung bình của đám đông EX = μ.

+ $\chi^2 = \dfrac{\sum n_i (x_i - \mu)^2}{\sigma_0^2}$,

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0
$H_1 : \sigma^2 \neq \sigma_0^2$	$W_\alpha = \left\{\chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n)\ ;\ \chi^2 > \chi^2_{\frac{\alpha}{2}}(n)\right\}$	$\overline{W}_\alpha = \left\{\chi^2_{1-\frac{\alpha}{2}}(n) \le \chi^2 \le \chi^2_{\frac{\alpha}{2}}(n)\right\}$ (5.31)
$H_1 : \sigma^2 < \sigma_0^2$	$W_\alpha = \left\{\chi^2 < \chi^2_{1-\alpha}(n)\right\}$	$\overline{W}_\alpha = \left\{\chi^2 \ge \chi^2_{1-\alpha}(n)\right\}$ (5.32)
$H_1 : \sigma^2 > \sigma_0^2$	$W_\alpha = \left\{\chi^2 > \chi^2_{\alpha}(n)\right\}$	$\overline{W}_\alpha = \left\{\chi^2 \le \chi^2_{\alpha}(n)\right\}$ (5.33)

$$\chi^2_\alpha(n) : \alpha \xrightarrow[\text{bậc tự do } n-1]{\text{Bảng 5 (Khi bình phương)}} \chi^2_\alpha(n).$$

• *Chú ý :* 1) Với đối thuyết hai phía (5.28), (5.31), khi bác bỏ H_0 chấp nhận H_1, ta so sánh s^2 và σ^2 để kết luận : Nếu $s^2 < \sigma_0{}^2$ (hay $s^2 > \sigma_0{}^2$) thì $\sigma^2 < \sigma_0{}^2$ ($\sigma^2 > \sigma_0{}^2$).

2) Quy tắc (5.29), (5.32) vẫn áp dụng được cho các cặp giả thuyết, đối thuyết $H_0 : \sigma^2 \ge \sigma_0{}^2$, $H_1 : \sigma^2 < \sigma_0{}^2$. Quy tắc (5.30), (5.33) vẫn áp dụng được cho cặp $H_0 : \sigma^2 \le \sigma_0{}^2$, $H_1 : \sigma^2 > \sigma_0{}^2$.

Ví dụ 5.10 : Chủ hãng sản xuất một loại thiết bị đo cho biết sai số đo của thiết bị loại này có độ lệch tiêu chuẩn bằng 5mm. Kiểm tra một mẫu 19 thiết bị loại này tính được phương sai mẫu $s^2 = 33$. Với mức ý nghĩa 5%, hãy cho nhận xét về ý kiến trên của chủ hãng. Biết sai số đo của thiết bị có phân phối chuẩn.

GIẢI : Gọi độ lệch tiêu chuẩn của sai số đo của thiết bị là σ

+ Ta có $H_0 : \sigma = 5$, $H_1 : \sigma \neq 5 \Leftrightarrow H_0 : \sigma^2 = 5^2$, $H_1 : \sigma^2 \neq 5^2$;

+ $\chi^2 = \dfrac{(19-1).33}{25} \approx 23,76$;

Cách 1 : Dùng phương pháp miền bác bỏ $\alpha = 0,05$:

$$1 - \frac{\alpha}{2} = 0,975 \to \chi_1 = \chi^2_{(18\,;\,0,975)} = 8,231\ ;$$

$$\frac{\alpha}{2} = 0,025 \to \chi_2 = \chi^2_{(18\,;\,0,025)} = 31,526.$$

+ Do $\chi_1 < \chi^2 = 23,76 < \chi_2$: Chấp nhận H_0.

Với mức ý nghĩa 5%, chưa có cơ sở bác bỏ ý kiến của chủ hãng..

Cách 2 : Dùng phương pháp P – giá trị của thống kê χ^2 với bậc tự do 18. Giá trị của thống kê này trên mẫu $\chi^2 = 23,76$. Từ Ví dụ 5.3 và Bảng 4, ta suy ra :

$$P - \text{giá trị} = 2P\left(\chi^2 > 23,76\right) > 2P\left(\chi^2 > 31,526\right) = 0,05 = \alpha.$$

Do đó, ta chấp nhận H_0.

Ví dụ 5.11 : Nếu độ biến động về đường kính của các sản phẩm được sản xuất bởi một dây chuyền tự động vượt quá 0,2 (mm) thì dây chuyền phải dừng lại để điều chỉnh. Kiểm tra ngẫu nhiên 12 sản phẩm của dây chuyền đo được độ lệch tiêu chuẩn của đường kính s = 0,3. Với mức ý nghĩa 5%, hãy xem dây chuyền có phải dừng lại để điều chỉnh không, biết đường kính của các sản phẩm có phân phối chuẩn.

GIẢI : Độ biến động được đặc trưng bằng độ lệch chuẩn hay phương sai. Gọi độ lệch chuẩn và phương sai của đường kính sản phẩm tương ứng là σ, σ^2.

Ta xét $H_0 : \sigma = 0,2$; $H_1 : \sigma > 0,2 \Leftrightarrow H_0 : \sigma^2 = (0,2)^2$, $H_1 : \sigma^2 > (0,2)^2$

- $\chi^2 = \dfrac{11.(0,3)^2}{(0,2)^2} = 24,75$;
- $\alpha = 0,05 \rightarrow \chi = \chi^2_{(11\,;\,0,05)} = 19,675$;
- $\chi^2 > \chi$: Bác bỏ H_0, chấp nhận H_1.

Với mức ý nghĩa 5%, dây chuyền cần điều chỉnh vì độ biến động của đường kính sản phẩm lớn hơn mức cho phép.

Cách 2. Dùng phương pháp P – giá trị của thống kê χ^2 với bậc tự do 11. Giá trị của thống kê này trên mẫu $\chi^2 = 24,75$. Từ Ví dụ 5.3 và Bảng 5, ta suy ra :

$$\text{P – giá trị} = P\left(\chi^2 > 24,75\right) < P\left(\chi^2 > 19,675\right) = 0,05 = \alpha.$$

Kết luận : Bác bỏ H_0, chấp nhận H_1.

Ví dụ 5.12 : Mức hao phí nguyên liệu X để sản xuất một sản phẩm của nhà máy tuân theo quy luật chuẩn có trung bình 20 kg. Một mẫu điều tra mức hao phí nguyên liệu để sản xuất 25 sản phẩm loại này được kết quả cho trong bảng :

Mức hao phí nguyên liệu hao phí (kg)	19,5	20,0	20,5
Số sản phẩm	5	18	2

Với mức ý nghĩa 5% có thể cho rằng phương sai của mức hao phí nguyên liệu là 0,06 kg^2 hay không.

GIẢI : Gọi σ^2 là phương sai của mức hao phí nguyên liệu để sản xuất một sản phẩm. Ta xét $H_0 : \sigma^2 = 0,06$, $H_1 : \sigma^2 \neq 0,06$

+ $\sum n_i(x_i - \mu)^2 = 5.(19,5-20)^2 + 18.(20-20)^2 + 2.(20,5-20)^2 = 1,75$;

+ $\chi^2 = \dfrac{\sum n_i(x_i - \mu)^2}{\sigma_0^2} = \dfrac{1,75}{0,06} = 29,1667$;

$+\ \alpha = 0{,}05$ ta có $1 - \dfrac{\alpha}{2} = 0{,}975$, $\chi_1 = \chi^2(25,\ 0{,}975) = 13{,}12$;

$\dfrac{\alpha}{2} = 0{,}025$, $\chi_2 = \chi^2(25,\ 0{,}025) = 40{,}646$;

$+\ \chi_1 < \chi^2 < \chi_2$: Chấp nhận H_0.

Với mức ý nghĩa 5% có thể cho rằng phương sai của mức hao phí nguyên liệu là 0,06 kg^2.

§3. SO SÁNH CÁC THAM SỐ

Trong mục này ta xét các bài toán so sánh tham số của các đám đông. Thủ tục kiểm định giả thuyết tương tự như trường hợp một tham số.

1. So sánh hai tỉ lệ

Xét hai đám đông X, Y có cùng đặc trưng định tính A. Giả sử p_1, p_2 tương ứng là tỉ lệ phần tử của X, của Y có tính chất A. Ta muốn so sánh tỉ lệ p_1 và p_2.

a) Bài toán : Từ hai mẫu độc lập $(X_1, ..., X_m)$ của X và $(Y_1, .., Y_n)$ của Y, với mức ý nghĩa α, kiểm định giả thuyết $H_0 : p_1 = p_2$ với một trong các đối thuyết $H_1 : p_1 \neq p_2$; $H_1 : p_1 < p_2$; $H_1 : p_1 > p_2$.

Ta sử dụng kết quả sau : Với m, n đủ lớn, khi H_0 đúng, tiêu chuẩn kiểm định $Z = (F_1 - F_2)\Big/\sqrt{F(1-F)\left(\dfrac{1}{m} + \dfrac{1}{n}\right)} \approx N(0,\ 1)$ với $F = \dfrac{mF_1 + nF_2}{m+n}$, còn F_1, F_2 tương ứng là tỉ lệ mẫu của X, của Y có tính chất A.

b) Quy tắc thực hành : Trên mẫu cụ thể $(x_1, ..., x_m)$ của X, $(y_1, ..., y_n)$ của Y với $Kf \geq 10$ và $K(1 - f) \geq 10$, $K = m + n$.

Giả thuyết $H_0 : p_1 = p_2$, đối thuyết H_1 ;

$$f_1 = \frac{K_1}{m}\ ;\ f_2 = \frac{K_2}{n}\ ;\ f = \frac{mf_1 + nf_2}{m+n} = \frac{K_1 + K_2}{m+n}\ ;$$

$$z = (f_1 - f_2)\Big/\sqrt{f(1-f)\left(\frac{1}{m} + \frac{1}{n}\right)}.$$

t luận :

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0	
$H_1 : p_1 \neq p_2$	$W_\alpha = \{\lvert z\rvert > z_{\alpha/2}\}$	$\overline{W}_\alpha = \{\lvert z\rvert \le z_{\alpha/2}\}$	(5.34)
$H_1 : p_1 < p_2$	$W_\alpha = \{z < -z_\alpha\}$	$\overline{W}_\alpha = \{z \ge -z_\alpha\}$	(5.35)
$H_1 : p_1 > p_2$	$W_\alpha = \{z > z_\alpha\}$	$\overline{W}_\alpha = \{z \le z_\alpha\}$	(5.36)

Ở đây $z_{\alpha/2}$; z_α là các giá trị tới hạn chuẩn.

Chú ý. Trong (5.34) khi bác bỏ H_0 chấp nhận $H_1 : p_1 \neq p_2$ nếu $f_1 < f_2$ ($f_1 > f_2$) ta kết luận $p_1 < p_2$ ($p_1 > p_2$).

Ví dụ 5.13 : Kiểm tra chất lượng sản phẩm về một loại hàng do hai nhà máy A và B sản xuất cho kết quả : trong 500 sản phẩm của A có 50 phế phẩm và trong 400 sản phẩm của B có 60 phế phẩm. Với mức ý nghĩa 5%, hãy xem

a) Chất lượng sản phẩm của A và B có khác nhau không ?

b) Chất lượng sản phẩm của A có tốt hơn B không?

GIẢI : Gọi p_1, p_2 tương ứng là tỉ lệ phế phẩm của A, B

a) Đặt $H_0 : p_1 = p_2$; $H_1 : p_1 \neq p_2$

$$f_1 = \frac{50}{500} = 0,1 \; ; \; f_2 = \frac{60}{400} = 0,15 \; ; \; f = \frac{50+60}{500+400} = 0,1222 \; ;$$

$$z = (0,1-0,15)\Big/\sqrt{0,1222.0,8778.\left(\frac{1}{500}+\frac{1}{400}\right)} = -2,2758 \; ;$$

$\alpha = 0,05 : z_{\alpha/2} = 1,96$;

$|z| > z_{\alpha/2}$: Bác bỏ H_0, chấp nhận H_1.

Với mức ý nghĩa 5% chất lượng sản phẩm của A và B khác nhau.

b) Xét $H_0 : p_1 = p_2$, $H_1 : p_1 < p_2$

Vì $\alpha = 0,05 : z_\alpha = 1,65$;

$Z < -z_\alpha$: Bác bỏ H_0, chấp nhận H_1.

Vậy với mức ý nghĩa 5% chất lượng của A tốt hơn của B.

2. So sánh hai trị trung bình

a) Bài toán : Giả sử có hai đám đông X và Y có kì vọng EX = μ_1, EY = μ_2. Từ hai mẫu độc lập : $(X_1, ..., X_m)$ của X và $(Y_1, Y_2, ..., Y_n)$ của Y, với mức ý nghĩa α, kiểm định giả thuyết $H_0 : \mu_1 = \mu_2$ với một trong các đối thuyết $H_1 : \mu_1 \neq \mu_2$; $H_1 : \mu_1 < \mu_2$; $H_1 : \mu_1 > \mu_2$.

Giả sử phương sai của các đám đông là $VX = \sigma_1^2$; $VY = \sigma_2^2$. Cho hai mẫu độc lập : $(X_1, ..., X_m)$ của X và $(Y_1, Y_2, ..., Y_n)$ của Y, tương tự như các trường hợp kiểm định một tham số, ta sử dụng các kết quả sau :

• **Trường hợp 1 :** Các tiêu chuẩn kiểm định có phân phối chuẩn hay xấp xỉ chuẩn.

1) Nếu các đám đông X và Y có phân phối chuẩn, biết σ_1^2, σ_2^2, thì thống kê :

$$Z = \left(\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)\right) \Big/ \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \approx N(0 ; 1) \qquad (5.37)$$

2) Nếu các đám đông X và Y chưa biết phân phối xác suất nhưng kích thước mẫu của các đám đông đủ lớn :

a) Chưa biết σ_1^2, σ_2^2 : thì thống kê

$$Z = \left(\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)\right) \Big/ \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}} \approx N(0 ; 1) \qquad (5.38)$$

b) Biết σ_1^2, σ_2^2 : thì thống kê

$$Z = \left(\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)\right) \Big/ \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \approx N(0 ; 1) \qquad (5.39)$$

• **Trường hợp 2 :** Các tiêu chuẩn kiểm định có phân phối Student

1) X và Y có phân phối chuẩn, chưa biết σ_1^2, σ_2^2 nhưng $\sigma_1^2 = \sigma_2^2$ thì thống kê

$$T = \left(\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)\right) \Big/ \sqrt{S^2\left(\frac{1}{m} + \frac{1}{n}\right)} \sim t(m + n - 2) \qquad (5.40)$$

ở đây S_1^2, S_2^2 là phương sai mẫu tương ứng của X và Y

$$S^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m + n - 2}.$$

2) X và Y có phân phối chuẩn, chưa biết σ_1^2, σ_2^2 và $\sigma_1^2 \neq \sigma_2^2$ thì thống kê

$$T = \left(\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)\right) \Big/ \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}} \sim t(k) \tag{5.41}$$

với
$$k = \frac{\left(S_1^2/m + S_2^2/n\right)^2}{\left(S_1^2/m\right)^2/(m-1) + \left(S_2^2/n\right)^2/(n-1)}.$$

Chú ý : Khi kích thước mẫu của các đám đông đủ lớn, thống kê T có phân phối xấp xỉ chuẩn chuẩn tắc, ta có thể chuyển trường hợp 2 này về trường hợp 1. Để tiện sử dụng và nếu không gây hiểu nhầm, trong sách này ta quy ước xấp xỉ này thỏa mãn khi kích thước mẫu $m \geq 30$, $n \geq 30$. Vì vậy trong thực hành ta chia các trường hợp sau :

b) Quy tắc thực hành : Cho hai mẫu $(x_1, ..., x_m)$, $(y_1, ..., y_n)$.

- Lập giả thuyết $H_0 : \mu_1 = \mu_2$, đối thuyết H_1.
- Tính giá trị của tiêu chuẩn kiểm định G từ số liệu mẫu.

Trường hợp 1 : (tiêu chuẩn kiểm định $G = Z \sim N(0 ; 1)$).

Từ (5.37), (5.38), (5.39) ta có : Khi các kích thước mẫu $m \geq 30$, $n \geq 30$:

– Các phương sai σ_1^2, σ_2^2 chưa biết $z = \left(\overline{x} - \overline{y}\right) \Big/ \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$, ở đây $s_1^2 ; s_2^2$ là phương sai mẫu tương ứng của X, của Y.

– Các phương sai σ_1^2, σ_2^2 đã biết $z = \left(\overline{x} - \overline{y}\right) \Big/ \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$.

- Tìm các giá trị tới hạn chuẩn và kết luận dựa vào miền bác bỏ

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0	
• $H_1 : \mu_1 \neq \mu_2$	$W_\alpha = \{\lvert z\rvert > z_{\alpha/2}\}$	$\overline{W}_\alpha = \{\lvert z\rvert \leq z_{\alpha/2}\}$	(5.42)
• $H_1 : \mu_1 < \mu_2$	$W_\alpha = \{z < -z_\alpha\}$	$\overline{W}_\alpha = \{z \geq -z_\alpha\}$	(5.43)
• $H_1 : \mu_1 > \mu_2$	$W_\alpha = \{z > z_\alpha\}$	$\overline{W}_\alpha = \{z \leq z_\alpha\}$	(5.44)

Ở đây $z_{\alpha/2}; z_\alpha$ là các giá trị tới hạn chuẩn.

Trường hợp 2 : (Tiêu chuẩn kiểm định có phân phối Student)

Khi m < 30, n < 30 , X và Y có phân phối chuẩn :

1) Phương sai σ_1^2, σ_2^2 chưa biết nhưng $\sigma_1^2 = \sigma_2^2$:

+ $T = \left(\overline{x} - \overline{y}\right) \Big/ \sqrt{s^2\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}$; ở đây $s^2 = \dfrac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$.

+ Tìm các giá trị tới hạn Student và kết luận :

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0	
• $H_1 : \mu_1 \neq \mu_2$	$W_\alpha = \left\{\lvert T\rvert > t\left(k; \dfrac{\alpha}{2}\right)\right\}$	$\overline{W}_\alpha = \left\{\lvert T\rvert \le t\left(k ; \dfrac{\alpha}{2}\right)\right\}$	(5.45)
• $H_1 : \mu_1 < \mu_2$	$W_\alpha = \{T < -t(k;\alpha)\}$	$\overline{W}_\alpha = \{T \ge -t(k ; \alpha)\}$	(5.46)
• $H_1 : \mu_1 > \mu_2$	$W_\alpha = \{T > t(k ; \alpha)\}$	$\overline{W}_\alpha = \{T \le t(k ; \alpha)\}$	(5.47)

Ở đây k = m + n – 2 và $t(k,\alpha)$ là giá trị tới hạn Student mức α.

2) Phương sai chưa biết và có thể khác nhau :

+ Tiêu chuẩn kiểm định : $T = \left(\overline{x} - \overline{y}\right) \Big/ \sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}$

+ Quy tắc thực hành như (5.45), (5.46), (5.47)

+ Giá trị tới hạn Student với bậc tự do

$$k = \left[\frac{\left(s_1^2/m + s_2^2/n\right)^2}{\left(s_1^2/m\right)^2/(m-1) + \left(s_2^2/n\right)^2/(n-1)}\right] \qquad (5.48)$$

Trong (5.48), k là phần nguyên của số tính được trong ngoặc vuông, chẳng hạn [23,75] = 23 và s_1^2 ; s_2^2 tương ứng là phương sai mẫu của X và Y.

Chú ý. Trong (5.42) hay (5.45) khi bác bỏ H_0 chấp nhận $H_1 : \mu_1 \neq \mu_2$ nếu $\overline{x} < \overline{y}$ $\left(\overline{x} > \overline{y}\right)$ ta kết luận $\mu_1 < \mu_2$ ($\mu_1 > \mu_2$)..

Ví dụ 5.14 : Giám đốc một hãng sản xuất thép muốn xác định xem có sự khác nhau về năng suất giữa ca ngày và ca tối hay không. Điều tra một mẫu 100 công nhân của ca ngày sản xuất được trung bình 74,3 (đơn vị : phần/1 giờ/người) với độ lệch tiêu chuẩn 16 ;

...u khác 100 công nhân của ca tối sản xuất được trung bình 69,7 (đơn vị : phần/1 giờ/người) với độ lệch tiêu chuẩn 18. Hãy xem có sự khác nhau về năng suất giữa hai ca không, với mức ý nghĩa 5%.

GIẢI : Gọi μ_1, μ_2 tương ứng là năng suất trung bình của mỗi công nhân làm việc thuộc ca sáng, thuộc ca tối.

Ta xét $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$.

$$z = (74,3 - 69,7) \Big/ \sqrt{\frac{16^2}{100} + \frac{18^2}{100}} = 1,91 \text{ (ở đây } m, n > 30).$$

Với $\alpha = 0,05$: $z_{\alpha/2} = 1,96$;

$|z| < z_{\alpha/2}$: Chấp nhận H_0.

Với mức ý nghĩa 5% chưa có bằng chứng về sự khác nhau về năng suất trung bình của hai ca.

Ví dụ 5.15 : Giám đốc của công ty thực phẩm muốn xác định liệu một kiểu đóng gói mới có làm tăng lượng hàng bán được hay không. Chọn ngẫu nhiên 30 quầy tương đương nhau gồm 15 quầy bán hàng được đóng theo gói mới và 15 quầy khác bán hàng được đóng theo gói cũ, tính được trong thời gian nghiên cứu lượng hàng bán được tại mỗi quầy hàng đóng gói mới có trung bình $\overline{x} = 130$ hộp với $s_1 = 10$; hàng đóng gói cũ có trung bình : $\overline{y} = 117$ hộp với $s_2 = 12$.

Với mức ý nghĩa 5% hãy xem kiểu đóng gói mới có làm tăng lượng hàng bán được không. Giả sử lượng hàng bán được theo hai kiểu đóng gói có phân phối chuẩn và cùng phương sai.

GIẢI : Gọi lượng hàng trung bình bán được theo kiểu đóng gói mới và kiểu đóng gói cũ tại mỗi quầy tương ứng là μ_1, μ_2.

Xét $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 > \mu_2$.

$$s^2 = \left(14.10^2 + 14.12^2\right) \big/ (15 + 15 - 2) = 122 ;$$

$$T = (130 - 117) \Big/ \sqrt{122.\left(\frac{1}{15} + \frac{1}{15}\right)} \approx 3,2233 ;$$

$\alpha = 0,05 \rightarrow t = t_{(28\,;\,0,05)} = 1,701$;

$T > t$: Bác bỏ H_0, chấp nhận H_1.

Với mức ý nghĩa 5% kiểu đóng gói mới làm tăng lượng hàng trung bình bán được.

Ví dụ 5.16 : Điều tra doanh số bán hàng tại các cửa hàng ở hai vùng A và B của một công ty trong thời gian T, kết quả cho như sau : (đơn vị : triệu đồng)

Vùng A	Vùng B
$\overline{x} = 121$	$\overline{y} = 89{,}17$
$s_1^2 = 76{,}57$	$s_2^2 = 17{,}37$
$m = 16$	$n = 10$

Với mức ý nghĩa 10%, hãy xem có sự khác nhau về doanh số bán hàng trung bình của mỗi cửa hàng ở hai vùng không. Giả sử doanh số bán hàng của mỗi cửa hàng có phân phối chuẩn.

GIẢI : Giả sử μ_1, μ_2 tương ứng là doanh số bán hàng trung bình của mỗi cửa hàng ở hai vùng A, B ; σ_A^2, σ_B^2 tương ứng là phương sai của doanh số bán hàng của mỗi cửa hàng ở hai vùng A, B.

Với $\alpha = 0{,}1$ ta có $\sigma_A^2 \neq \sigma_B^2$ (xem Ví dụ 5.20).

+ Đặt $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$

$$+ T = (121 - 89{,}17) \Big/ \sqrt{\frac{76{,}57}{16} + \frac{17{,}37}{10}} = 12{,}4631 \ ;$$

$$\alpha = 0{,}1 \ ; \ k = \left[\frac{\left(\frac{76{,}57}{16} + \frac{17{,}37}{10} \right)^2}{\left(\frac{76{,}57}{16} \right)^2 / 15 + \left(\frac{17{,}37}{10} \right)^2 / 9} \right] = 22 \ ;$$

$t(22 \ ; 0{,}05) = 1{,}717$.

+ $|T| > t(22 \ ; 0{,}05)$: Bác bỏ H_0, chấp nhận H_1. Khi này $\overline{x} > \overline{y}$.

Kết luận : Với mức ý nghĩa 10% doanh số bán hàng trung bình của mỗi cửa hàng tại hai vùng khác nhau và vùng A cao hơn vùng B.

Chú ý : 1) Các quy tắc kiểm định trình bày trong mục trên *không* sử dụng được khi các đám đông X, Y *không độc lập*. Trường hợp này ta chuyển về bài toán so sánh từng cặp như sau :

Để kiểm định $H_0 : \mu_1 = \mu_2 \Leftrightarrow H_0 : \mu_1 - \mu_2 = 0$, ta xét đám đông : $D = X - Y$ có $ED = \mu_1 - \mu_2 = \mu_d$ và mẫu các quan sát của D : $(d_1, d_2, ..., d_n)$ với $d_i = x_i - y_i$, ở đây $(x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n)$ mẫu của (X, Y). Bài toán trở thành : Với mức ý nghĩa α, kiểm định $H_0 : \mu_d = 0$ (bài toán này được xét trong §2).

Ví dụ 5.17 : Để khảo sát giá bán của hai cửa hiệu thực phẩm lớn A và B, người ta điều tra ngẫu nhiên giá bán của 12 mặt hàng thông dụng nhất như sau :

Mặt hàng	Cửa hiệu A	Cửa hiệu B	Mặt hàng	Cửa hiệu A	Cửa hiệu B
1	0,89	0,95	7	0,99	0,99
2	0,59	0,55	8	1,99	1,79
3	1,29	1,49	9	2,25	2,39
4	1,50	1,69	10	0,50	0,59
5	2,49	2,39	11	1,99	2,19
6	0,65	0,79	12	1,79	1,99

Với mức ý nghĩa 2% hãy xem có sự khác nhau về giá bán ở hai cửa hiệu hay không.

GIẢI :

Mặt hàng	1	2	3	4	5	6	7
d_i	–0,06	0,04	–0,2	–0,19	0,1	–0,14	0

Mặt hàng	8	9	10	11	12
d_i	0,2	–0,14	–0,09	–0,2	–0,2

Từ bảng tìm được $\overline{d} = -0,0733$, $s^2 = 0,0176$.

Đặt $H_0 : \mu_d = 0$; $H_1 : \mu_d \neq 0$.

Do $n = 12$; $T = (-0,0733 - 0)\Big/\sqrt{\dfrac{0,0176}{12}} = -1,91398$;

$\alpha = 0,02 : t(11, 0,01) = 2,718$;

$|T| < t(11 ; 0,01)$, chấp nhận H_0.

Với mức ý nghĩa 2%, chưa có cơ sở để cho rằng giá bán trung bình của hai cửa hiệu là khác nhau.

2) Tương tự sử dụng các tiêu chuẩn thống kê (5.37), (5.38), (5.39), (5.40), (5.41), ta kiểm định giả thuyết

$H_0 : \mu_1 - \mu_2 = m$ với một trong các đối thuyết ;

$H_1 : \mu_1 - \mu_2 \neq m$; $H_1 : \mu_1 - \mu_2 < m$; $H_1 : \mu_1 - \mu_2 > m$, ở đây m là một số đã cho.

Ví dụ 5.18 : Trở lại Ví dụ 5.15. Với mức ý nghĩa 5%, hãy kiểm định $H_0 : \mu_1 - \mu_2 = 2$, $H_1 : \mu_1 - \mu_2 > 2$.

GIẢI : $T = \left((130-117)-2\right) \Big/ \sqrt{122.\left(\frac{1}{15}+\frac{1}{15}\right)} \approx 2,7272$;

$\alpha = 0,05 : t = t(28, 0,05) = 1,701$;

$T > t$: Bác bỏ H_0, chấp nhận H_1.

3) Khoảng tin cậy cho hiệu số $\mu_1 - \mu_2$.

Khi bác bỏ $H_0 : \mu_1 = \mu_2$, chấp nhận $H_1 : \mu_1 \neq \mu_2$, ta xét bài toán :

Bài toán : Với độ tin cậy $1 - \alpha$, tìm khoảng tin cậy cho $\mu_1 - \mu_2$.

Áp dụng các thống kê (5.37), (5.38), (5.39), (5.40), (5.41), cho các trường hợp ta tìm được khoảng tin cậy cho $\mu_1 - \mu_2$.

Chẳng hạn :

• Khi $m, n \geq 30$, σ_1^2 và σ_2^2 đã biết (hay σ_1^2 và σ_2^2 chưa biết), sử dụng thống kê

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \approx N(0 ; 1)$$

(hay $Z = \left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2) \Big/ \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}} \approx N(0 ; 1)$), ta có

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \quad \left(\text{hay } \varepsilon = z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}\right).$$

Khoảng tin cậy cho $\mu_1 - \mu_2$: $\left(\overline{x} - \overline{y} - \varepsilon ; \overline{x} - \overline{y} + \varepsilon\right)$.

• Khi X và Y có phân phối chuẩn, $\sigma_1^2 = \sigma_2^2$ chưa biết.

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t(m+n-2) ; \quad S^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}.$$

Với $\varepsilon = t(m+n-2 ; \alpha/2).\sqrt{s^2\left(\frac{1}{m} + \frac{1}{n}\right)}$; $s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$.

Khoảng tin cậy : $\left(\overline{x} - \overline{y} - \varepsilon ; \overline{x} - \overline{y} + \varepsilon\right)$.

Tương tự, cho các trường hợp khác, ta tìm được khoảng tin cậy đối xứng và tìm được khoảng tin cậy một bên cho $\mu_1 - \mu_2$.

Ví dụ 5.19 : Trở lại Ví dụ 5.16. Với độ tin cậy 95% hãy ước lượng mức chênh lệch doanh thu trung bình mỗi cửa hàng ở hai vùng trong khoảng thời gian T.

GIẢI : $\overline{x} = 121$; $s_1^2 = 76{,}57$; $m = 16$; $\overline{y} = 89{,}17$; $s_2^2 = 17{,}37$; $n = 10$;

$\overline{x} - \overline{y} = 121 - 89{,}17 = 31{,}83$;

$$k = \left[\frac{\left(\frac{76{,}57}{16} + \frac{17{,}37}{10}\right)^2}{\left(\frac{76{,}57}{16}\right)^2/15 + \left(\frac{17{,}37}{10}\right)^2/9}\right] = 22\ ;$$

$1 - \alpha = 0{,}95$, $\alpha/2 = 0{,}025$: $t(22;\ 0{,}025) = 2{,}08$;

$$\varepsilon = 2{,}08\sqrt{\frac{76{,}57}{16} + \frac{17{,}37}{10}} = 5{,}3122.$$

Khoảng tin cậy 95% của $\mu_1 - \mu_2$:

$$\left(\overline{x} - \overline{y} - \varepsilon;\ \overline{x} - \overline{y} + \varepsilon\right) = (26{,}5178\ ;\ 37{,}1422)\ \text{(triệu đồng)}.$$

3. So sánh hai phương sai

Cho hai đám đông $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$. Từ hai mẫu độc lập $(X_1, ..., X_m)$ của X, $(Y_1, ..., Y_n)$ của Y, ta xét bài toán sau :

a) Bài toán : Với mức ý nghĩa α, kiểm định giả thuyết

$H_0 : \sigma_1^2 = \sigma_2^2$ với một trong các đối thuyết

$H_1 : \sigma_1^2 \neq \sigma_2^2$; $H_1 : \sigma_1^2 < \sigma_2^2$; $H_1 : \sigma_1^2 > \sigma_2^2$

Khi giả thuyết H_0 đúng, ta sử dụng tiêu chuẩn :

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F(m-1, n-1).$$

Giá trị tới hạn của $F(m - 1, n - 1)$ cho trong Bảng 5.

b) Quy tắc thực hành : Từ các mẫu $(x_1, ..., x_m)$ của X ; $(y_1, ..., y_n)$ của Y ta tính :

- $F = \dfrac{s_1^2}{s_2^2}$ (giả sử $s_1^2 > s_2^2$), ở đây s_1^2, s_2^2 là phương sai mẫu của X, Y.

• Tìm giá trị tới hạn F (Bảng 5, bậc tự do m – 1, n – 1) : Với mức ý nghĩa α

$$\frac{\alpha}{2} \Rightarrow f_{\alpha/2}(m-1, n-1) = f_{\alpha/2} \ ;$$

$$1-\frac{\alpha}{2} \Rightarrow f_{1-\alpha/2}(m-1, n-1) = \frac{1}{f_{\alpha/2}(n-1, m-1)} = f_{1-\alpha/2} \ ;$$

(Nếu đối thuyết một phía $\alpha \Rightarrow f_{\alpha}(m-1, n-1) = f_{\alpha}$).

• Kết luận

Đối thuyết	Bác bỏ H_0, chấp nhận H_1	Chấp nhận H_0	
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$F \notin W_{\alpha} = [f_{1-\alpha/2}, f_{\alpha/2}]$	$F \in \overline{W}_{\alpha} = [f_{1-\alpha/2}, f_{\alpha/2}]$	(5.49)
$H_1 : \sigma_1^2 > \sigma_2^2$	$W_{\alpha} = \{F > f_{\alpha}\}$	$\overline{W}_{\alpha} = \{F \leq f_{\alpha}\}$	(5.50)

Chú ý : 1) Nếu $s_2^2 > s_1^2$ ta lấy $F = \frac{s_2^2}{s_1^2}$ lúc này bậc tự do là (n – 1, m – 1).

2) Cũng như so sánh hai tỉ lệ hay so sánh hai trung bình, với đối thuyết hai phía (5.49) khi bác bỏ H_0, chấp nhận H_1 ta so sánh ${s_1}^2$ và ${s_2}^2$ để kết luận.

Ví dụ 5.20 : Trở lại Ví dụ 5.16. Với mức ý nghĩa 10%, hãy xem phương sai của doanh số bán hàng của mỗi cửa hàng ở hai vùng có khác nhau không.

GIẢI : Gọi σ_A^2, σ_B^2 tương ứng là phương sai của doanh số bán hàng của mỗi cửa hàng ở hai vùng A, B. Xét

$$H_0 : \sigma_A^2 = \sigma_B^2 \ ; H_1 : \sigma_A^2 \neq \sigma_B^2$$

$$F = \frac{s_1^2}{s_2^2} = \frac{945}{17,37} = 54,4041 \ ; \ m = 7 \text{ và } n = 6.$$

$$\alpha = 0,1 : f^* = f_{0,05}(6, 5) = 4,95 \ ;$$

$$f_* = f_{0,95}(6,5) = \frac{1}{f_{0,05}(5,6)} = \frac{1}{4,39} = 0,2278 \ ;$$

$$F \notin [f_*, f^*] : \text{Bác bỏ } H_0, \text{ chấp nhận } H_1.$$

Với mức ý nghĩa 10%, phương sai của doanh số bán hàng của mỗi cửa hàng ở hai vùng A, B là khác nhau.

Cách 2. Dùng phương pháp P – giá trị của tiêu chuẩn F bậc tự do (6, 5).

Giá trị F tính từ mẫu : F = 54,404. Sử dụng Bảng 5, với chú ý đối thuyết 2 – phía, từ Ví dụ 5.3 ta có :

$$P - \text{giá trị} = 2P(F > 54{,}404) < 2P(F > 4{,}95) = 2.0{,}05 = 0{,}1.$$

Kết luận : Bác bỏ H_0, chấp nhận H_1.

Ví dụ 5.21 : Người ta dùng phương sai hay độ lệch tiêu chuẩn làm độ đo đánh giá sự rủi ro của cổ phiếu. Điều tra ngẫu nhiên giá cổ phiếu của công ty A trong 25 ngày tính được $s_1^2 = 6{,}52$; của công ty B trong 22 ngày tính được $s_2^2 = 3{,}47$. Với mức ý nghĩa 5% có thể cho rằng độ rủi ro cổ phiếu của công ty A cao hơn của công ty B không ?

GIẢI : Cách 1. Gọi σ_1^2, σ_2^2 tương ứng là phương sai giá cổ phiếu của các công ty A, B. Đặt $H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 > \sigma_2^2$

$$F = \frac{6{,}52}{3{,}47} = 1{,}879 \,;$$

$\alpha = 0{,}05 \Rightarrow f = f_{0,05}(24, 21) = 2{,}05$;

$F < f$: Chấp nhận H_0.

Với mức ý nghĩa 5%, chưa có cơ sở để cho rằng cổ phiếu của A nhiều rủi ro hơn cổ phiếu của B.

Cách 2. Dùng phương pháp P – giá trị của tiêu chuẩn F, bậc tự do (24, 21). Giá trị F tính từ mẫu : F = 1,879. Sử dụng Bảng 6, với chú ý đối thuyết 1 – phía, từ Ví dụ 5.3 ta có :

$$P - \text{giá trị} = P(F > 1{,}879) > P(F > 2{,}05) = 0{,}05.$$

Kết luận : Chấp nhận H_0.

§4. KIỂM ĐỊNH PHI THAM SỐ

Các kiểm định giả thuyết thống kê liên quan tới tham số của đám đông đã được xét trong các mục trước (của chương này). Ở mục này ta xét các thủ tục kiểm định giả thuyết thống kê không liên quan tới tham số của đám đông, chúng được gọi là kiểm định phi tham số.

1. Kiểm định về tính độc lập

Giả sử cần nghiên cứu đồng thời hai dấu hiệu định tính A và B trên cùng một đám đông. Dấu hiệu A có k dấu hiệu thành phần $A_1, \ldots, A_k$; B có m dấu hiệu thành phần $B_1, \ldots, B_m$.

Bài toán : Với mức ý nghĩa α, kiểm định giả thuyết H_0 : A và B độc lập ; đối thuyết H_1 : A và B phụ thuộc.

Từ mẫu kích thước n quan sát đồng thời hai dấu hiệu A, B ta có :

A \ B	B_1	B_2	...	B_j	...	B_m	$\sum$
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	n_1
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	n_2
$\vdots$							$\vdots$
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	n_i
$\vdots$							$\vdots$
A_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	n_k
$\sum$	r_1	r_2	...	r_j	...	r_m	n

$n_i = \sum_{j=1}^{m} n_{ij}$: tổng các tần số ứng với dấu hiệu thành phần A_i ;

$r_j = \sum_{i=1}^{k} n_{ij}$: tổng các tần số ứng với dấu hiệu thành phần B_j ;

n_{ij} : tần số ứng với các phần tử đồng thời mang dấu hiệu A_i và B_j.

Với n đủ lớn : Xác suất xuất hiện đồng thời $(A_i.B_j)$

$$P(A_i.B_j) \approx \frac{n_{ij}}{n},$$

và $P(A_i)$, $P(B_j)$ tương ứng là xác suất xuất hiện A_i, B_j :

$$P(A_i) \approx \frac{n_i}{n} \;;\; P(B_j) \approx \frac{n_j}{n}.$$

Khi H_0 đúng : A, B độc lập thì $P(A_i.B_j) = P(A_i)P(B_j)$ do đó :

$$\frac{n_{ij}}{n} = \frac{n_i r_j}{n^2} \;;\; i = \overline{1, k};\; j = \overline{1, m}.$$

Ta sử dụng tiêu chuẩn :

$$\chi^2 = n\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\left(\frac{n_{ij}}{n} - \frac{n_i.r_j}{n^2}\right)^2}{\frac{n_i}{n}.\frac{r_j}{n}} \sim \chi^2((k-1)(m-1)) \qquad (5.51)$$

$\chi^2((k-1)(m-1))$: Phân phối Khi bình phương với (k – 1)(m – 1) bậc tự do.

Biến đổi (5.51) ta có công thức tiện sử dụng trong tính toán :

$$\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_i.r_j} - 1\right] \qquad (5.52)$$

Miền bác bỏ : $W_\alpha = \left\{\chi^2 > \chi^2_{((k-1)(m-1),\alpha)} : \chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_i r_j} - 1\right]\right\}$.

Quy tắc thực hành :

- Lập giả thuyết H_0 : A và B độc lập ; đối thuyết H_1 : A và B phụ thuộc.
- Tính $\chi^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_i r_j} - 1\right]$.
- Cho α :

Cách 1. Dùng phương pháp miền bác bỏ

+ Tìm $\chi^2_\alpha = \chi^2((k-1)(m-1),\alpha)$

(Bảng 5 : Khi bình phương bậc tự do (k – 1)(m – 1)).

+ $\chi^2 > \chi^2_\alpha$: Bác bỏ H_0, chấp nhận H_1 ;

$\chi^2 \le \chi^2_\alpha$: Chấp nhận H_0.

Cách 2. Dùng phương pháp P – giá trị của $\chi^2((k-1)(m-1))$

P – giá trị = $P(\chi^2((k-1)(m-1)) > \chi^2) \le \alpha$: Bác bỏ H_0, chấp nhận H_1.

Ví dụ 5.22 : Ở một nước, để xác định xem sự ủng hộ của những công dân nước đó về một sắc thuế mới có thụ thuộc vào mức thu nhập của họ hay không, người ta tiến hành điều tra ngẫu nhiên 1000 người và phân loại theo các mức thu nhập của họ cho trong bảng sau :

Sắc thuế mới	Mức thu nhập			
	Thấp	Trung bình	Cao	Tổng
Ủng hộ	182	213	203	598
Phản đối	154	138	110	402
Tổng	336	351	313	1000

Với mức ý nghĩa 5%, hãy cho kết luận về việc xác định trên.

GIẢI : Ta có

H_0 : Sự ủng hộ không phụ thuộc vào mức thu nhập

H_1 : Sự ủng hộ phụ thuộc vào mức thu nhập

$$\chi^2 = 1000\left[\frac{182^2}{336.598} + \frac{213^2}{351.598} + \frac{203^2}{313.598} + \frac{154^2}{336.402} + \frac{138^2}{351.402} + \frac{110^2}{313.402} - 1\right]$$

$$= 7{,}8782\ ;$$

$\alpha = 0{,}05 : \chi^2 = \chi^2_\alpha = \chi^2_{((2-1)(3-1),\alpha)} = 5{,}991$. Vậy $\chi^2 > \chi^2_\alpha$: Bác bỏ H_0, chấp nhận H_1.

Với mức ý nghĩa 0,05, có sự phụ thuộc giữa sự ủng hộ sắc thuế mới và thu nhập của công dân.

Cách 2. Dùng phương pháp P – giá trị.

P – giá trị $= P(\chi^2_{((2-1)(3-1))} \geq 7{,}8782) < P(\chi^2_{((2-1)(3-1))} > 5{,}991) = 0{,}05$.

Bác bỏ H_0 , chấp nhận H_1.

2. Kiểm định giả thuyết về luật phân phối

Cho biến ngẫu nhiên X nhận giá trị trên một tập số S nào đó. Giả sử S được chia thành k phần rời nhau $S_1, S_2, \ldots, S_k$ và $S = S_1 \cup S_2 \cup \ldots \cup S_k$; $p_i = P\{X \in S_i\}$; mẫu $(x_1, x_2, \ldots, x_n)$ của X và n_i là số giá trị của mẫu này thuộc vào S_i.

Bài toán : Với mức ý nghĩa α, cho kết luận về các xác suất p_i có bằng các số p^*_i đã cho hay không.

Ta tiến hành kiểm định :

Giả thuyết $H_0 : p_1 = p^*_1, p_2 = p^*_2, \ldots, p_k = p^*_k$ (5.53)

Đối thuyết H_1 : tồn tại $p_i \neq p^*_i$, $i = 1, \ldots, k$.

Ở đây $p^*_i : 0 < p^*_i < 1, \sum_{i=1}^{k} p^*_i = 1$.

Giả sử, từ mẫu cụ thể $(x_1, x_2, \dots, x_n)$ của X ta có bảng phân phối tần số thực nghiệm :

– Trường hợp biến ngẫu nhiên X rời rạc.

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

– Trường hợp biến ngẫu nhiên X liên tục

Lớp	$x_0 - x_1$	$x_1 - x_2$	...	$x_{k-1} - x_k$
n_i	n_1	n_2	...	n_k

Với $n_1 + n_2 + \dots + n_k = n$, n đủ lớn, $n_i \geq 5$ (Nếu $n_i < 5$ ta ghép các thành phần của mẫu lại để có $n_i \geq 5$).

Có hai trường hợp :

a) Trường hợp các số p^*_i được tính từ một phân phối xác suất $F^*(x)$ đã cho có tham số chưa biết

Bài toán kiểm định cặp giả thuyết đối thuyết (5.53) trở thành :

Bài toán : Cho mức ý nghĩa α, kiểm định.

Giả thuyết H_0 : X có phân phối xác suất $F^*(x)$ (5.54)

Đối thuyết H_1 : X không có phân phối xác suất $F^*(x)$.

Quy tắc thực hành giải bài toán :

• Từ mẫu ta ước lượng tham số của phân phối $F^*(x)$ đã cho bằng phương pháp hợp lí tối đa [xem 1.2.2, §2, Ch.4].

• Tính theo phân phối $F^*(x)$ các xác suất

$$p^*_i = \begin{cases} P\{X = x_i\} & \text{nếu X là rời rạc} \\ P\{x_{i-1} < X < x_i\} & \text{nếu X là liên tục.} \end{cases}$$

(Trường hợp rời rạc ta tính xác suất p^*_i cho tất cả giá trị của X ; Trường hợp liên tục chọn $x_0 = -\infty$, $x_k = +\infty$).

+ Sau đó tính các số $n^*_i = np^*_i$ (được gọi là các tần số lí thuyết).

• Tính giá trị của tiêu chuẩn kiểm định từ mẫu

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np^*_i)^2}{np^*_i}.$$

(Với n đủ lớn và $n_i \geq 5$, thống kê $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np^*_i)^2}{np^*_i} \sim \chi^2(k - r - 1)$) (5.55)

• Cho mức ý nghĩa α, tìm $\chi^2_\alpha = \chi^2(k - r - 1\,;\, \alpha)$.

(Bảng 4 : Khi bình phương bậc tự do (k – r – 1)). Ở đây r là số tham số của phân phối $F^*(x)$. Chẳng hạn : Phân phối chuẩn thì r = 2 ; phân phối nhị thức và Poisson thì r = 1.

• Kết luận :

+ $\chi^2 > \chi^2_\alpha$: Bác bỏ H_0, chấp nhận H_1 ;

+ $\chi^2 \leq \chi^2_\alpha$: Chấp nhận H_0.

b) Trường hợp các số p^*_i đã biết

Ta trở về bài toán kiểm định (5.53) :

Bài toán : Cho mức ý nghĩa α, kiểm định các xác suất p_i có bằng các số p^*_i đã cho hay không. Xét :

Giả thuyết $H_0 : p_1 = p^*_1, p_2 = p^*_2, \ldots, p_k = p^*_k$ (5.56)

Đối thuyết H_1 : tồn tại $p_i \neq p^*_i$, i = 1,…, k.

Quy tắc thực hành giải bài toán :

• Tính các số $n^*_i = np^*_i$.

Tính giá trị của tiêu chuẩn kiểm định từ mẫu :

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np^*_i)^2}{np^*_i}.$$

(Với n đủ lớn và $n_i \geq 5$, thống kê $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np^*_i)^2}{np^*_i} \sim \chi^2(k - 1)$) (5.57)

• Cho mức ý nghĩa α, tìm giá trị tới hạn của phân phối Khi bình phương [Bảng 4] : $\chi^2(k - 1, \alpha)$.

• Kết luận :

Nếu $\chi^2 > \chi^2(k-1, \alpha)$: Bác bỏ H_0 chấp nhận H_1 ;

Nếu $\chi^2 \le \chi^2(k-1, \alpha)$: Chấp nhận H_0.

Chú ý : Ta cũng có thể dùng phương pháp P – giá trị của thống kê (5.55), (5.57) để kiểm định.

Ví dụ 5.23 : Sản phẩm sản xuất ra trên một dây chuyền tự động được đóng gói một cách ngẫu nhiên cứ 3 sản phẩm trong một hộp. Mẫu kiểm tra ngẫu nhiên 100 hộp sản phẩm loại này cho kết quả :

Số sản phẩm loại I	0	1	2	3
Số hộp	0	10	40	50

Với mức ý nghĩa 5%, có thể xem số sản phẩm loại I có trong một hộp là biến ngẫu nhiên có phân phối nhị thức không?

GIẢI : Gọi X là số sản phẩm loại I có trong một hộp. Nếu X có phân phối nhị thức thì X~ B(3 ; p) và p được ước lượng bằng phương pháp hợp lí tối đa [xem Ví dụ 4.13, §2,Ch. 4] :

$$p = (1.10 + 2.40 + 3.50)/(3.100) = 0{,}8.$$

Ta xét H_0 : X có phân phối xác suất nhị thức B(3 ; 0,8)

H_1 : X có không có phân phối xác suất nhị thức B(3 ; 0,8).

Với $P_i = P\{X = x_i\} = P\{X_i = x_i\} = C_3^i 0{,}8^i 0{,}2^{3-i}$, $i = 0,..,3$,

Vì $n_1 = 0 < 5$, ghép hai giá trị $(X = 0)$ và $(X = 1)$ thành $(X \le 1)$ ta lập bảng

x_i	n_i	p^*_i	np^*_i	$(n_i - np^*_i)^2/np^*_i$
≤ 1	10	0,104	10,4	0,0154
2	40	0,384	38,4	0,0667
3	50	0,512	51,2	0,0281
				$\chi^2 = 0{,}1102$

Với mức ý nghĩa 5%, áp dụng (5.54) k = 3, r = 1, ta tìm

$$\chi^2_{0,05}(3-1-1) = \chi^2_{0,05}(1) = 3{,}841;$$

$\chi^2 = 0{,}1102 < 3{,}841$: Chấp nhận H_0, nghĩa là X ~ B(3 ; 0,8).

Ví dụ 5.24 : Để tìm hiểu số thiết bị bị hỏng trong một tháng của một hệ thống máy, người ta theo dõi 50 tháng liền và được số thiết bị bị hỏng cho trong bảng :

x_i	0	1	2	3	4	6	8
n_i	10	4	12	8	7	6	3

Với mức ý nghĩa 5%, có thể cho rằng số thiết bị bị hỏng X tuân theo quy luật Poisson P(λ) hay không ?

GIẢI : Do λ chưa biết ta ước lượng λ theo phương pháp hợp lí tối đa $\lambda = \overline{x} = 2,8$ [xem 1.2.2, §2, Ch. 4]. Ta kiểm định :

H_0 : Phân phối xác suất của X là P(2,8) ;

H_1 : Phân phối xác suất của X không là Poisson P(2,8).

Do có $n_i < 5$ ta sắp xếp lại mẫu :

x_i	≤ 1	2	3	4 – 5	≥ 6
n_i	14	12	8	7	9

Ta có bảng sau

	n_i	p^*_i	np^*_i	$(n_i - np^*_i)^2/np^*_i$
≤ 1	14	0,2311	11,555	0,5174
2	12	0,2384	11,92	0,0005
3	8	0,2225	11,125	0,8778
4 – 5	7	0,2429	12,145	2,1796
≥ 6	9	0,0651	3,255	10,1398
				$\chi^2 = 13,7691$

Áp dụng (5.54) : k = 5, r = 1, với $\alpha = 0,05$:

$$\chi^2(5 - 1 - 1,\ 0,05) = 7,815 < \chi^2 = 13,7691 : \text{Bác bỏ } H_0, \text{ chấp nhận } H_1.$$

Với mức ý nghĩa 5%, chưa có cơ sở cho rằng số thiết bị bị hỏng X tuân theo quy luật Poisson.

Ví dụ 5.25 : Đo chỉ tiêu X (đơn vị là gam) của một loại sản phẩm kết quả cho trong bảng

Lớp	18–20	20–22	22–24	24–26	26–28	28–30
n_i	3	4	14	33	27	19

Với mức ý nghĩa 5% có thể cho rằng X có phân phối chuẩn hay không ?

GIẢI : Giả sử phân phối chuẩn đã cho có trung bình μ và phương sai σ^2. Theo Ví dụ 4.14, §2, Ch.4, các tham số này được ước lượng theo phương pháp hợp lí tối đa là :

$$\mu = \overline{x} = 25,68 \; ; \sigma^2 = \widehat{s}^2 = \frac{1}{100}\sum_{i=1}^{6}(x_i - \overline{x})^2 n_i = 5,9376 \; ; \sigma = 2,4367.$$

Với mức ý nghĩa 5%, ta kiểm định :

$H_0 : X \sim N(25,68,\ 5,9376)$; H_1 : X không có phân phối chuẩn.

Vì bảng phân phối thực nghiệm trên có $n_i < 5$, ta sắp xếp lại mẫu theo bảng sau :

x_i	18–22	22–24	24–26	26–28	28–30
n_i	7	14	33	27	19

Ta tính : $p_i = P\{x_{i-1} < X < x_i\} = \varphi\left(\dfrac{x_{i-1} - 25,68}{2,4367}\right) - \varphi\left(\dfrac{x_i - 25,68}{2,4367}\right)$.

Các lớp	n_i	p_i	np_i	$(n_i - np_i)^2/np_i$
≤ 22	7	0,0655	6,55	0,0309
22–24	14	0,1796	17,96	0,8731
24–26	33	0,3066	30,66	0,1786
26–28	27	0,2772	27,72	0,0187
≥ 28	19	0,171	17,1	0,2111
				$\chi^2 = 1,3124$

$\alpha = 0,05$, áp dụng (5.52) r = 2, ta có : $\chi^2(5 - 2 - 1, \alpha) = 5,991$.

Do $\chi^2 < 5,991$: Chấp nhận H_0.

Vậy với mức ý nghĩa 5%, ta coi X có phân phối chuẩn.

Chú ý : Để kiểm định H_0 : X có phân phối chuẩn ; H_1 : X không có phân phối chuẩn, người ta còn sử dụng tiêu chuẩn Jarque – Bera (JB) như sau :

Từ mẫu từ mẫu cụ thể $(x_1, x_2, \ldots, x_n)$ của X, tính tiêu chuẩn

$$JB = n[S^2/6 + (K - 3)^2/24]$$

với $S = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^3/n}{S_X^3}$; $K = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^4/n}{S_X^4}$ và $S_X = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$.

Với n đủ lớn JB có phân phối xấp xỉ χ^2 (2). Do đó P – giá trị của JB xấp xỉ P – giá trị của χ^2 (2) và miền bác bỏ của thống kê này là $W_\alpha = \left\{JB > \chi^2(2;\alpha)\right\}$.

Ví dụ 5.26 : Một báo cáo cho rằng "Sự phân chia thị trường năm ngoái về tỉ lệ khách hàng sử dụng sản phẩm của ba công ty A, B, C tương ứng là $p_A = 30\%$; $p_B = 50\%$; $p_C = 20\%$ đã bị thay đổi trong năm nay khi công ty C cho ra một loại sản phẩm mới". Hãy cho nhận xét về báo cáo đó với mức ý nghĩa 5%. Biết rằng mẫu điều tra thị trường năm nay của các công ty cho kết quả : Trong 200 khách hàng, số người dùng sản phẩm của các công ty A, B, C tương ứng là 48 ; 98 ; 54 (người).

GIẢI : Gọi H_0 : Thị trường sản phẩm của các công ty A, B, C tương ứng là $p_A = 0{,}3$; $p_B = 0{,}5$; $p_C = 0{,}2$.

H_1 : Thị trường sản phẩm của các công ty A, B, C thay đổi.

Số người dùng sản phẩm	n_i	p_i	$n\,p^*_i$	$(n_i - np^*_i)^2/np^*_i$
A	48	0,3	60	2,4
B	98	0,5	100	0,04
C	54	0,2	40	4,9
Tổng	200		200	$\chi^2 = 7{,}34$

Ta có $\chi^2_\alpha(3-1) = \chi^2_\alpha(2) = 5{,}99$ và $\chi^2 > 5{,}99$: Bác bỏ H_0, chấp nhận H_1. Với mức ý nghĩa 5%, chấp nhận báo cáo.

BÀI TẬP CHƯƠNG 5

5.1. Điều tra trọng lượng của một loại sản phẩm (đơn vị : gam) kết quả cho trong bảng :

x_i	2 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
n_i	5	25	48	35	17	10

Những sản phẩm có trọng lượng không nhỏ hơn 25 gam là loại 1.

a) Có báo cáo cho rằng tỉ lệ sản phẩm loại 1 là 21%. Cho nhận xét về báo cáo đó với mức ý nghĩa 5%.

b) Nếu bác bỏ báo cáo câu a) thì mức ý nghĩa có thể tối thiểu là bao nhiêu.

c) Trước đây tỉ lệ sản phẩm loại 1 là 20%. Có ý kiến cho rằng nay do thiếu đầu tư sản xuất nên tỉ lệ sản phẩm loại 1 giảm. Cho nhận xét về ý kiến đó với mức ý nghĩa 3%.

5.2. Để nhận một lô hàng gồm toàn bút bi rẻ tiền từ một nhà máy sản xuất, một khách hàng kiểm tra ngẫu nhiên 300 bút bi loại này từ lô hàng thấy có 30 bút phế phẩm. Lô hàng sẽ bị từ chối nếu có trên 5% bút phế phẩm. Với mức ý nghĩa 5%, lô hàng có bị từ chối hay không ?

5.3. Một công ty tuyên bố rằng 70% khách hàng ưa thích sản phẩm của công ty. Điều tra ngẫu nhiên 400 khách hàng có 260 người ưa thích sản phẩm của công ty. Với mức ý nghĩa 5%

a) Hãy xem tỉ lệ trong tuyên bố trên của công ty có chấp nhận được không.

b) Hãy xem tỉ lệ trong tuyên bố trên của công ty có cao hơn thực tế không.

5.4. Lô hàng là đủ tiêu chuẩn xuất khẩu nếu tỉ lệ phế phẩm của nó không vượt quá 3%. Kiểm tra ngẫu nhiên 400 sản phẩm của lô hàng thấy có 14 phế phẩm.

a) Với mức ý nghĩa 5% lô hàng có được phép xuất khẩu không ?

b) Nếu lô hàng không được xuất khẩu thì mức ý nghĩa có thể thấp nhất là bao nhiêu ?

5.5. Một nhà máy chế tạo TV tuyên bố rằng do áp dụng công nghệ sản xuất mới nên tỉ lệ TV của nhà máy cần có sự sửa chữa trong thời gian 2 năm đầu hoạt động giảm còn 9%. Để kiểm tra tuyên bố trên người ta điều tra ngẫu nhiên 100 TV của nhà máy sản xuất thấy có 14 TV bị sửa chữa trong thời gian 2 năm đầu hoạt động. Với mức ý nghĩa 2% cho kết luận về tuyên bố trên.

5.6. Nếu áp dụng phương pháp công nghệ thứ nhất thì tỉ lệ phế phẩm là 6%, còn áp dụng phương pháp công nghệ thứ hai thì trong 100 sản phẩm có 5 phế phẩm. Với mức ý nghĩa 5% có thể cho rằng áp dụng phương pháp công nghệ thứ hai cho tỉ lệ phế phẩm thấp hơn không ?

5.7. Một công ty tuyên bố rằng : Nhu cầu đặt hàng của công ty được công ty giải quyết trong vòng 2 tháng, tăng lên là 50%. Điều tra ngẫu nhiên 200 khách hàng có nhu cầu đặt hàng của công ty thấy có 80 khách có nhu cầu đặt hàng được giải quyết trong thời gian 2 tháng. Với mức ý nghĩa 5%, hãy cho kết luận về tuyên bố trên.

5.8. Một dây chuyền sản xuất bóng đèn được gọi là hoạt động bình thường nếu tuổi thọ trung bình của bóng đèn sản xuất ra từ dây chuyền không dưới 375 giờ. Mẫu điều tra 50 bóng đèn do dây chuyền này sản xuất ra tính được tuổi thọ trung bình của một bóng đèn là 360 giờ và độ lệch chuẩn 100 giờ.

a) Với mức ý nghĩa 5% hãy xem dây chuyền có hoạt động bình thường không.

b) Có báo cho rằng tuổi thọ trung bình của một bóng đèn do dây chuyền sản ruất ra là 390 giờ. Cho nhận xét về báo cáo đó với mức ý nghĩa 5%.

c) Hãy giải câu a), b) bằng phương pháp P – giá trị.

d) Giả sử tuổi thọ trung bình của một bóng đèn do dây chuyền sản xuất ra không dưới 385 giờ. Nếu muốn kiểm định trong câu a) (hay trong câu b)) có mức ý nghĩa $\alpha = 5\%$ và mức sai lầm loại hai β không vượt quá 3% thì cần phải điều tra bao nhiêu bóng đèn ?

Giả sử tuổi thọ của một bóng đèn do dây chuyền sản xuất ra có phân phối chuẩn.

5.9. Để kiểm tra lượng nước ngọt được đóng vào chai 2 lít được sản xuất tại một nhà máy người ta kiểm tra ngẫu nhiên 100 chai nước ngọt loại này của nhà máy và tính được lượng nước ngọt trung bình 1,99 lít ; độ lệch tiêu chuẩn 0,05 lít. Với mức ý nghĩa 2% hãy xem lượng nước ngọt đóng trong chai loại này có bị thiếu không.

5.10. Một tổ kiểm tra muốn xác định thời gian trung bình từ lúc công ty A nhận được đơn khiếu kiện của khách hàng đến lúc giải quyết là bao nhiêu ngày. Chọn ngẫu nhiên 15 trường hợp khiếu kiện trong năm qua, kết quả cho như sau :

114	78	96	137	78	103	117	126	86	99
114	72	104	73	96.					

a) Với mức ý nghĩa 5%, có thể cho rằng thời gian trung bình để một khiếu kiện được giải quyết bởi công ty A vượt quá 90 ngày không.

b) Hãy kết luận vấn đề trên bằng phương pháp P – giá trị.

Giả sử thời gian giải quyết một khiếu kiện có phân phối chuẩn.

5.11. Một máy đóng hộp đựng một loại sản phẩm được gọi là làm việc bình thường nếu trọng lượng trung bình mỗi hộp được đóng là 368 gam. Giả sử trọng lượng của các hộp được sản xuất có phân phối chuẩn với độ lệch tiêu chuẩn 30 gam. Người ta sẽ dừng máy điều chỉnh khi trọng lượng trung bình của hộp nhỏ hơn 368 gam. Điều tra ngẫu nhiên 25 hộp do máy đóng hộp tính được trọng lượng trung bình là 360 gam.

a) Với mức ý nghĩa $\alpha = 5\%$, hãy cho biết máy có phải dừng để điều chỉnh hay không.

b) Giải câu a) bằng phương pháp P – giá trị.

c) Nếu người phụ trách máy muốn lực lượng của kiểm định $(1 - \beta)$ là 90% và mức ý nghĩa $\alpha = 5\%$, biết trọng lượng trung bình thực của hộp giảm nằm trong khoảng (360, 368) (gam) thì kích thước mẫu điều tra cần là bao nhiêu.

5.12. Thời gian trước số tiền gửi tiết kiệm bằng ngoại tệ trung bình của mỗi khách hàng là 1000 USD. Để đánh giá xem xu hướng này còn giữ nguyên hay không người ta kiểm tra ngẫu nhiên 64 sổ tiết kiệm và tính được số tiền gửi trung bình là 990 USD, độ lệch tiêu chuẩn 100 USD.

a) Với mức ý nghĩa 5%, hãy xem số tiền gửi tiết kiệm có thay đổi không.

b) Với mức ý nghĩa 5%, hãy xem số tiền gửi tiết kiệm có giảm so với trước không.

c) Tìm xác suất sai lầm loại II nếu số tiền tiết kiệm trung bình của mỗi khách hàng thực sự bằng 1050 USD.

d) Tìm lực lượng của kiểm định.

e) Nếu đòi hỏi mức ý nghĩa α và xác suất sai lầm loại II là β đều bằng 0,05, sai lệch giữa trung bình thực sự và giá trị giả thuyết không quá 30 USD thì cần mẫu điều tra có kích thước tối thiểu là bao nhiêu.

5.13. Điều tra trọng lượng của một loại trái cây (đơn vị : gam) kết quả cho trong bảng :

x_i	2 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
n_i	5	25	48	35	17	10

a) Có báo cáo cho rằng trọng lượng trung bình của một trái cây là 20 gam. Cho nhận xét về báo cáo đó với mức ý nghĩa 5%.

b) Có ý kiến cho rằng trước đây trọng lượng trung bình của một trái cây là 15 gam, nay do áp dụng phương pháp sản xuất mới làm cho trọng lượng trung bình của các trái cây tăng lên. Cho nhận xét về ý kiến đó với mức ý nghĩa 5%.

c) Trái cây có trọng lượng lớn hơn 20 gam là cây loại A. Có ý kiến cho rằng tỉ lệ trái cây loại A thấp hơn tỉ lệ trái cây còn lại. Nếu chấp nhận ý kiến đó thì mức ý nghĩa có thể tối thiểu là bao nhiêu.

5.14. Tỉ lệ nợ xấu tại một ngân hàng vùng A là biến ngẫu nhiên có phân phối xấp xỉ chuẩn. Điều tra ngẫu nhiên 17 ngân hàng ở vùng A có tỉ lệ nợ xấu (tính bằng %) là : 7 ; 4 ; 6 ; 7 ; 5 ; 4 ; 9 ; 6 ; 7 ; 3 ; 8 ; 7 ; 6 ; 8 ; 9 ; 4 ; 8.

a) Có báo cáo cho rằng tỉ lệ nợ xấu trung bình của ngân hàng vùng A là 6,2%. Hãy cho nhận xét về báo cáo đó với mức ý nghĩa 5%.

b) Nhân viên thanh tra nhận xét rằng tỉ lệ nợ xấu của các ngân hàng vùng A cao hơn tỉ lệ nợ xấu của các ngân hàng vùng B vì ở vùng B trung bình tỉ lệ này chỉ khoảng 5%. Với mức ý nghĩa 5%, hãy cho ý kiến về nhận xét này.

c) Có báo cáo cho rằng độ phân tán tỉ lệ nợ xấu của một ngân hàng vùng A là 5. Với mức ý nghĩa 5%, hãy cho nhận xét về báo cáo đó.

d) Hãy giải các câu a), b) bằng cách sử dụng khoảng tin cậy.

e) Hãy giải các câu a), b) c) bằng phương pháp P – giá trị.

5.15. Một máy đóng chai nước uống phải dừng để điều chỉnh nếu sai lệch của lượng nước uống được máy đóng vào chai so với quy định vượt quá 1,15 cm^3. Mẫu kiểm tra 25 chai nước uống do máy đóng thu được sai lệch trung bình là 2,03 cm^3 và độ lệch chuẩn là 0,95 cm^3. Với mức ý nghĩa 5%, hãy xem máy có phải dừng để điều chỉnh hay không. Giả sử lượng nước uống do máy đóng vào chai có phân phối chuẩn.

5.16. Mẫu điều tra về lợi nhuận năm nay của 36 công ty chứng khoán lớn trong vùng A cho độ lệch tiêu chuẩn là 15,2%.

a) Với mức ý nghĩa 5%, có thể cho rằng lợi nhuận của các công ty chứng khoán trong vùng A năm nay là ổn định hơn không, biết rằng năm trước độ lệch chuẩn lợi nhuận của các công ty chứng khoán trong vùng A là 18,2% ?

b) Giải câu a) bằng phương pháp P – giá trị.

Giả sử lợi nhuận của công ty chứng khoán trong vùng A có phân phối chuẩn.

5.17. Trọng lượng của một sản phẩm do máy sản xuất ra là biến ngẫu nhiên có phân phối chuẩn. Nghi ngờ độ đồng đều về trọng lượng sản phẩm có xu hướng giảm sút người ta cân thử 12 sản phẩm được $s^2 = 11,41\ (\text{gam})^2$. Với mức ý nghĩa 5% hãy cho kết luận về nghi ngờ trên, biết rằng bình thường độ phân tán của trọng lượng sản phẩm là $10\ (\text{gam})^2$.

5.18. Một cuộc thăm dò trước bầu cử, 40 trong số 100 cử tri được hỏi cho biết họ sẽ bỏ phiếu cho ông A. Một tuần sau một cuộc thăm dò khác được tiến hành và có 68 trong số 150 người được hỏi cho biết sẽ bỏ phiếu cho ông A. Với mức ý nghĩa 5% tỉ lệ cử tri bỏ phiếu cho ông A có : a) thay đổi không ; b) tăng lên không.

5.19. Điều tra ngẫu nhiên 300 sinh viên tốt nghiệp của trường đại học A sau một năm ra trường, thu được kết quả : 180 sinh viên đã đi làm, 36 sinh viên học tiếp sau đại học và 84 sinh viên chưa có việc làm. Điều tra ngẫu nhiên 200 sinh viên tốt nghiệp của trường đại học B sau một năm ra trường, thu được kết quả : 110 sinh viên đã đi làm, 20 sinh viên học tiếp sau đại học và 70 sinh viên chưa có việc làm. Nếu kết luận tỉ lệ sinh viên đã đi làm sau khi tốt nghiệp một năm của hai trường đại học A và B bằng nhau thì mức ý nghĩa tối đa là bao nhiêu ?

5.20. Giám đốc chất lượng của một nhà máy sản xuất bóng đèn muốn xác định có sự khác nhau hay không về tuổi thọ của bóng đèn được sản xuất ra bởi hai kiểu máy

Biết rằng độ lệch tiêu chuẩn tuổi thọ của bóng đèn do kiểu máy thứ nhất, thứ hai sản xuất ra tương ứng là 110 giờ, 125 giờ. Điều tra 25 bóng đèn do từng kiểu máy sản xuất ra tính được tuổi thọ trung bình của chúng tương ứng là 375 giờ, 362 giờ. Với mức ý nghĩa 1% :

a) Hãy cho kết luận về sự khác nhau của tuổi thọ bóng đèn trung bình do hai kiểu máy sản xuất.

b) Giải bằng phương pháp P – giá trị.

5.21. Giám đốc tài chính của một doanh nghiệp muốn so sánh chi phí tiếp khách giữa hai bộ phận bán hàng và sản xuất. Một mẫu điều tra 15 biên nhận chi phí mỗi nơi, kết quả cho như sau :

Bộ phận bán hàng	*Bộ phận sản xuất*
$\overline{x}_1 = 42{,}5$ USD	$\overline{x}_2 = 31{,}75$ USD
$s_1 = 9{,}50$ USD	$s_2 = 8{,}20$ USD

a) Với mức ý nghĩa 5%, có sự khác nhau hay không về chi phí trung bình giữa hai bộ phận trên. Xét hai trường hợp khi các mức chi phí có cùng, có khác độ lệch tiêu chuẩn.

b) Với mức ý nghĩa 10%, với số liệu trên hãy cho kết luận về sự khác nhau về phương sai của chi phí của hai bộ phận.

Giả sử chi phí của hai bộ phận có phân phối chuẩn.

5.22. Người ta muốn so sánh chất lượng đào tạo tại hai cơ sở đào tạo A và B bằng cách căn cứ trên điểm trung bình (thang điểm 10) ở kì thi quốc gia. Một mẫu 100 thí sinh được đào tạo tại cơ sở A có điểm trung bình 9,4, độ lệch tiêu chuẩn 0,8 và một mẫu 80 thí sinh được đào tạo tại cơ sở B có điểm trung bình 9 với độ lệch tiêu chuẩn 1. Với mức ý nghĩa 1% hãy phát biểu và kiểm định bài toán so sánh điểm thi trung bình của hai cơ sở A, B biết rằng A có cơ sở và đội ngũ giáo viên tốt hơn B.

5.23. Để so sánh thời gian sản xuất ra 1 sản phẩm cùng loại của hai máy (đơn vị là giây) người ta điều tra và ghi lại kết quả như sau :

Máy 1	58	58	56	38	70	38	42	75	68	67
Máy 2	57	55	63	24	67	43	33	68	56	54

Với mức ý nghĩa 5% có thể cho rằng thời gian trung bình của hai máy để sản xuất một sản phẩm là như nhau không.

Giải bài toán khi thời gian sản xuất mỗi sản phẩm của hai máy có phân phối chuẩn, có phương sai như nhau và có phương sai khác nhau.

5.24. Để so sánh năng suất của hai loại máy H1, H2, người ta điều tra thời gian (tính bằng phút) cần để sản xuất ra một sản phẩm cùng loại của hai loại máy này kết quả cho trong bảng sau

Loại máy	Số sản phẩm	Trung bình mẫu	Độ lệch chuẩn mẫu
H1	100	57	13,59
H2	100	52	14,46

a) Dựa vào kết quả điều tra nói trên, với mức ý nghĩa 5% hãy cho biết loại máy H2 có cho năng suất cao hơn loại máy H1 không.

b) Nếu bác bỏ báo cáo : Năng suất của hai loại máy là như nhau, thì mức ý nghĩa có thể tối thiểu là bao nhiêu.

5.25. Kiểm tra chất lượng sản phẩm của cùng một loại do hai nhà máy A và B sản xuất ra, kết quả cho trong bảng :

Nhà máy	*Số sản phẩm được kiểm tra*	*Số phế phẩm*
A	1800	54
B	1200	30

Với mức ý nghĩa 5% có thể cho rằng chất lượng sản phẩm của hai nhà máy là như nhau không ?

5.26. Hai máy cùng gia công một loại chi tiết. Để kiểm tra độ chính xác của hai máy người ta lấy ngẫu nhiên từ mỗi máy 7 chi tiết đem đo và được kết quả :

Máy 1	135	138	136	140	138	135	139
Máy 2	140	135	140	138	135	138	140

Với mức ý nghĩa 5% có thể cho rằng hai máy có độ chính xác như nhau không. Biết kích thước chi tiết có phân phối chuẩn.

5.27. Một nhà kinh tế cho rằng độ phân tán của thị phần trong các công ty hoạt động có cạnh tranh về giá cả cao hơn trong các công ty độc quyền. Để kết luận về điều đó người ta đã điều tra thị phần của một công ty cạnh tranh về giá cả trong 4 năm và tìm thấy phương sai mẫu $s^2 = 85,576$. Đồng thời điều tra thị phần của một công ty độc quyền trong 7 năm tìm được phương sai mẫu 13,78. Với mức ý nghĩa 5% hãy cho kết luận về ý kiến trên. Giả sử thị phần của các công ty là các biến ngẫu nhiên có phân phối chuẩn.

5.28. Nhà phân tích tài chính thường dùng độ lệch chuẩn hay phương sai làm thước đo về rủi ro đầu tư. Mẫu 26 số liệu điều tra về tỉ lệ tăng lợi nhuận hàng năm của hai công ty chứng khoán A và B tính được độ lệch chuẩn tương ứng là 8,89% ; 13,03%. Với mức ý nghĩa 5%, có thể cho rằng rủi ro của A thấp hơn B hay không ? Giả sử tỉ lệ tăng lợi nhuận của các công ty có phân phối chuẩn.

..ẫu điều tra ngẫu nhiên 90 người trưởng thành phân loại theo giới tính và số giờ họ xem TV trong 1 tuần như sau

	Nam	Nữ
Không dưới 25 giờ	15	29
Dưới 25	27	19

Với mức ý nghĩa 1%, hãy xem thời gian xem TV của một người có phụ thuộc vào giới tính của họ không.

5.30. Điều tra ngẫu nhiên thu nhập của 400 công nhân ở hai thành phố A và B kết quả như sau (đơn vị triệu đồng/1 năm)

Thu nhập	0–5	5–10	10–15	Trên 15
Thành phố A	28	42	30	24
Thành phố B	44	78	78	76

Với mức ý nghĩa 5%, hãy xem thu nhập của công nhân có phụ thuộc vào thành phố mà họ làm việc hay không.

5.31. Quan sát 100 sản phẩm được sản xuất ở ba ca : sáng, chiều, tối với chất lượng được chia làm hai loại kết quả cho trong bảng :

	Sáng	Chiều	Tối
Tốt	25	32	28
Xấu	7	6	2

Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận chất lượng sản phẩm phụ thuộc vào ca sản xuất không.

5.32. Kết quả điều tra số lần khách hàng gọi phục vụ mỗi ngày tại một cơ sở dịch vụ máy tính trong 500 ngày cho trong bảng :

Số lần gọi phục vụ	0	1	2	3	4	5	6
Số ngày	160	175	86	41	18	12	8

Với mức ý nghĩa 1% có thể cho rằng số lần gọi phục vụ của khách hàng trong một ngày có phân phối Poisson không ?

5.33. Số liệu X (tính bằng phút) của 500 cuộc điện thoại đường dài được cho như sau :

X (tính bằng phút)	0–5	5–10	10–15	15–20	20–25	25–30
Tần số	48	84	164	126	50	28

Với mức ý nghĩa 1% có thể coi X có phân phối chuẩn được không ?

5.34. Tung một đồng tiền 100 lần thấy có 63 lần xuất hiện mặt sấp, 37 lần xuất hiện mặt ngửa. Với mức ý nghĩa 5%, đồng tiền có cân xứng không ?

Chương 6

HỒI QUY VÀ TƯƠNG QUAN

§1. PHÂN TÍCH HỒI QUY

1. Mở đầu

Trong nhiều bài toán thực tế người ta quan tâm tới mối quan hệ của hai hay nhiều biến ngẫu nhiên. Giả sử có hai biến ngẫu nhiên X và Y. Vấn đề đặt ra là có hay không mối quan hệ phụ thuộc giữa X và Y. Nếu X và Y độc lập ta có thể xét riêng từng biến ; còn nếu X và Y phụ thuộc thì sự phụ thuộc và mức độ phụ thuộc là như thế nào. Ta coi giữa X và Y có các loại phụ thuộc sau :

– Sự phụ thuộc hàm số : tồn tại một hàm f(x) sao cho Y = f(X).

– Sự phụ thuộc thống kê : X thay đổi thì phân phối xác suất của Y cũng thay đổi.

– Sự phụ thuộc tương quan : X thay đổi thì trung bình điều kiện $E(Y|X)$ cũng thay đổi, nghĩa là $E(Y|X) = \varphi(X) \neq$ hằng số. Sự phụ thuộc tương quan là sự phụ thuộc thống kê. Trong chương này ta xét sự phụ thuộc tương quan. Người ta chứng minh được rằng trong tất cả các hàm h(X) được dùng để ước lượng Y thì $\varphi(X) = E(Y|X)$ sẽ làm sai số bình phương trung bình $E(Y - h(X))^2$ đạt cực tiểu nghĩa là :

$$E(Y - \varphi(X))^2 = \min E(Y - h(X))^2.$$

Phương trình $E(Y|X) = \varphi(X)$ được gọi là phương trình hồi quy của Y theo X.

Nếu $\varphi(X) = A + BX$ ta nói giữa Y và X có sự phụ thuộc tuyến tính hay có tương quan tuyến tính.

Nếu $\varphi(X)$ là hàm không tuyến tính ta nói giữa Y và X có tương quan phi tuyến.

Có hai loại bài toán được nghiên cứu ở đây :

– Tìm dạng hàm $\varphi(X)$ và sau đó từ mẫu quan sát tìm $\varphi(X)$ để dự báo giá trị trung bình của Y. Đây là nội dung của phân tích hồi quy.

– Xét xem có hay không sự phụ thuộc tương quan giữa Y và X và đánh giá mức độ chặt chẽ của sự phụ thuộc tương quan (nếu có). Đây là nội dung của phân tích tương quan.

2. Hồi quy tuyến tính

2.1. Mô hình hồi quy và phương tình hồi quy tuyến tính

Trong phần này ta xét mô hình đơn giản với X là biến độc lập không ngẫu nhiên. Y là biến phụ thuộc và giữa chúng có quan hệ cho bởi :

Mô hình hồi quy tuyến tính : $Y = A + BX + \varepsilon$, (6.1)

với A, B là các hằng số ; ε là biến ngẫu nhiên được gọi là nhiễu hay sai số ngẫu nhiên.

Mô hình (6.1) còn được cho trong dạng phương trình hồi quy tuyến tính :

$$E(Y \mid X) = A + BX \qquad (6.2)$$

A, B được gọi là các hệ số hồi quy lí thuyết ; A được gọi là hệ số chặn, B được gọi là hệ số góc.

Từ mẫu quan sát $(X_1, Y_1), \ldots, (X_n, Y_n)$ của (X, Y) cho bởi mô hình (6.1), nghĩa là (X_i, Y_i) $(i = \overline{1, n})$ thoả mãn phương trình :

$$Y_i = A + BX_i + \varepsilon_i \qquad (6.3)$$

ta ước lượng các hệ số A, B khi nhiễu ε_i thỏa mãn các điều kiện sau :

i) $E(\varepsilon_i) = 0$.

ii) $Var(\varepsilon_i) = \sigma^2$ (phương sai không đổi).

iii) ε_i độc lập và có phân phối chuẩn $\varepsilon_i \sim N(0, \sigma^2)$.

2.2. Phương pháp bình phương bé nhất

Vấn đề đặt ra là từ mẫu quan sát $(x_1, y_1), \ldots, (x_n, y_n)$ của (X, Y) hãy ước lượng các hệ số A, B sao cho tổng bình phương sai lệch giữa các giá trị quan sát y_i và giá trị cho bởi phương trình hồi quy $A + Bx_i$ là nhỏ nhất.

Người ta dùng phương pháp bình phương bé nhất để để ước lượng A, B như sau :

Xét hàm hai biến $F(K, L) = \sum_{i=1}^{n}\left(y_i - (K + Lx_i)\right)^2$, ước lượng của các tham số A, B của (6.1) hay (6.2) là (a, b) được tìm từ bài toán cực tiểu của F(K, L) :

$$F(a, b) = \min_{(K, L)} F(K, L).$$

Để tìm cực tiểu của F(K, L) ta giải hệ :

$$\begin{cases} \dfrac{\partial F(K,L)}{\partial K} = -2\sum_{i=1}^{n}(y_i - K - Lx_i) = 0 \\ \dfrac{\partial F(K,L)}{\partial L} = -2\sum_{i=1}^{n}(y_i - K - Lx_i)x_i = 0 \end{cases} \Leftrightarrow \begin{cases} nK + \left(\sum_{i=1}^{n} x_i\right)L = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)K + \left(\sum_{i=1}^{n} x_i^2\right)L = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Đây là hệ phương trình tuyến tính bậc nhất đối với K, L. Nghiệm a, b của hệ cho như sau :

$$\begin{cases} b = \dfrac{n\sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \dfrac{S_{yx}}{S_{xx}}, & (6.4) \\ a = \dfrac{\sum y_i - b\sum x_i}{n} & (6.5) \end{cases}$$

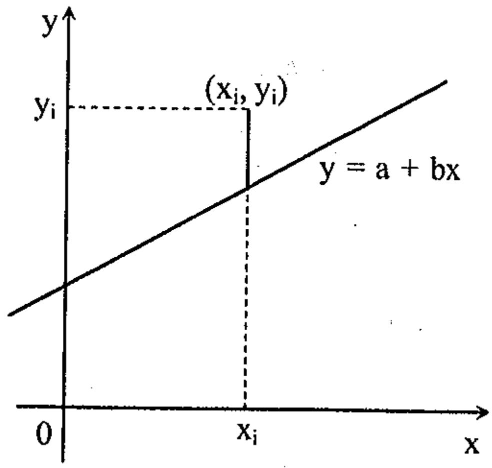
Trong (6.4) : $$S_{yx} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} \qquad (6.6)$$

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \qquad (6.7)$$

Các số a, b cho bởi các biểu thức (6.4), (6, 5) được gọi là các hệ số hồi quy tuyến tính mẫu và là ước lượng điểm của A, B bằng phương pháp bình phương bé nhất.

Phương trình $\hat{y} = a + bx$ là ước lượng của phương trình (6.1) được gọi là *phương trình hồi quy tuyến tính mẫu.* Có phương trình hồi quy tuyến tính mẫu ta sẽ dự báo được trung bình của Y khi biết $X = x_0$ là $\hat{y}_0 = a + bx_0$.

Đường thẳng $\hat{y} = a + bx$ được gọi là đường thẳng hồi quy (H.6.1).



Hình 6.1

Chú ý :

1) $\left|y_i - \widehat{y_i}\right|^2 = \left|y_i - a - bx_i\right|^2$ là bình phương sai lệch giữa giá trị quan sát y_i và giá trị của phương trình hồi quy $\widehat{y_i} = a + bx_i$. Như vậy tổng các bình phương sai lệch của các giá trị quan sát từ mẫu (x_i, y_i) tới đường thẳng hồi quy là nhỏ nhất.

2) a, b là ước lượng không chệch và hiệu quả của A, B. Hệ số a là giá trị trung bình của Y khi X = 0 ; hệ số b cho biết khi X tăng lên 1 đơn vị thì Y tăng (hay giảm) b đơn vị nếu b dương (nếu b âm).

Ví dụ 6.1 : Một công ty tiến hành phân tích hiệu quả quảng cáo của mình và thu thập số liệu trong thời gian 5 tháng được kết quả :

X	5	8	10	15	22
Y	6	15	20	30	39

Trong đó X là số tiền chi cho quảng cáo (đơn vị : triệu đồng) còn Y là tổng doanh thu (đơn vị : chục triệu đồng).

a) Viết phương trình hồi quy tuyến tính mẫu của tổng doanh thu theo chi phí quảng cáo.

b) Dự báo trung bình tổng doanh thu khi chi phí quảng cáo là 13 triệu đồng.

GIẢI : Ta đã tính được $\sum x_i = 60$; $\sum x_i^2 = 898$;

$\sum y_i = 110$; $\sum x_i y_i = 1658$; $n = 5$.

Từ (6.5) ta có :

$$b = \frac{5 \times 1658 - 60 \times 110}{5 \times 898 - (60)^2} = 1{,}8989 ;$$

$$a = \frac{110 - 1{,}8989 \times 60}{5} = -0{,}7868.$$

Phương trình hồi quy tuyến tính mẫu của tổng doanh thu theo chi phí quảng cáo $\hat{y} = -0{,}7868 + 1{,}8989x$.

Khi chi phí quảng cáo 13 triệu đồng thì trung bình tổng doanh thu là

$$\hat{y} = -0{,}7868 + 1{,}8989.13 = 23{,}8989 \text{ (chục triệu).}$$

2.3. Một số bài toán liên quan

Ngoài việc ước lượng các hệ số hồi quy ta quan tâm tới các bài toán sau :

1. Tìm hệ số xác định.

a) Tổng bình phương phần dư - RSS (Residual Sum of Squares)

$$RSS = \sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - a\sum y_i - b\sum x_i y_i \qquad (6.8)$$

RSS đo độ sai lệch khi dùng phương trình hồi quy để ước lượng giá trị của biến phụ thuộc Y từ mẫu.

Ước lượng điểm phương sai nhiễu $\text{Var}\varepsilon = \sigma^2$ là $\hat{\sigma}^2 = \dfrac{RSS}{n-2}$ (6.9)

Sai số chuẩn của đường hồi quy, kí hiệu $S_{Y,X} = \sqrt{\hat{\sigma}^2} = \sqrt{\dfrac{RSS}{n-2}}$ (6.10)

b) Tổng bình phương sai lệch được giải thích-ESS (Explained Sum of Squares)

$$ESS = \sum(\hat{y}_i - \bar{y})^2 = b^2\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right) \quad (6.11)$$

ESS đo độ sai lệch của biến phụ thuộc y được tính bởi phương trình hồi quy với giá trị trung bình mẫu quan sát.

c) Tổng bình phương tất cả sai lệch-TSS (Total Sum of Squares)

$$TSS = \sum(y_i - \bar{y})^2 = RSS + ESS \quad (6.12)$$

TSS đo độ sai lệch giữa giá trị quan sát của biến phụ thuộc Y với giá trị trung bình mẫu quan sát của Y.

d) Hệ số xác định (Coefficient of determination)

$$r^2 = \frac{ESS}{TSS} = \frac{\text{tổng các bình phương sai lệch do hồi quy}}{\text{tổng các bình phương sai lệch}} \quad (6.13)$$

Ý nghĩa : Hệ số r^2 cho biết : Mô hình hồi quy đã giải thích được (r^2.100)% sự thay đổi của Y quanh giá trị trung bình của nó là do biến giải thích X gây nên, số phần trăm còn lại do sai số ngẫu nhiên và do các yếu tố khác (nếu có) chưa đưa vào mô hình.

Hệ số xác định là đại lượng đo sự phù hợp của mô hình hồi quy với số liệu mẫu quan sát.

2. Xét ảnh hưởng của X đối với Y.

Kiểm định giả thuyết $H_0 : B = 0$; $H_1 : B \neq 0$ với mức ý nghĩa α.

$$\text{Đặt } S_b = \frac{S_{Y,X}}{S_{XX}} = \frac{S_{Y,X}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}} \quad (6.14)$$

$S_{Y,X}$ cho bởi (6.10).

Tiêu chuẩn kiểm định : $T = \dfrac{b}{S_b} \sim t(n-2)$ và miền bác bỏ

$$W_\alpha = \left\{|T| > t_{(n-2,\alpha/2)}\right\} \quad (6.15)$$

3. Tìm khoảng tin cậy của hệ số góc B.

Với thống kê $T = \dfrac{b}{S_b}$ ta có khoảng tin cậy $1-\alpha$ cho hệ số B :

$$b \pm t(n-2,\alpha/2).S_b, \quad (6.16)$$

ở đây b cho bởi (6.4).

4. Dự báo

a) Dự báo trung bình của Y khi $X = x_0$ với độ tin cậy $1 - \alpha$

Mệnh đề 6.1 : Đặt $\hat{y}_0 = a + bx_0$, $E(Y|X = x_0) = \mu_{x_0}$, $\bar{x} = \dfrac{\sum x_i}{n}$;

$$h = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}} \tag{6.17}$$

thì thống kê
$$T = \frac{\hat{y}_0 - \mu_{x_0}}{S_{Y,X}\sqrt{h}} \sim t(n-2). \tag{6.18}$$

Để dự báo μ_{x_0} ta sử dụng (6.18) tìm ước lượng khoảng của nó :

• Tìm $\hat{y}_0 = a + bx_0$.

• Tìm khoảng tin cậy $1 - \alpha$ của $E(Y|X = x_0) = \mu_{x_0}$

$$\hat{y}_{x_0} \pm t(n-2, \alpha/2)S_{Y,X}\sqrt{h}. \tag{6.19}$$

Kết luận : Với độ tin cậy $1 - \alpha$ trung bình của Y khi $X = x_0$ từ

$$\hat{y}_{x_0} - t(n-2, \alpha/2)S_{Y,X}\sqrt{h} \text{ đến } \hat{y}_{x_0} + t(n-2, \alpha/2)S_{Y,X}\sqrt{h}.$$

b) Dự báo giá trị của Y khi $X = x_0$ với độ tin cậy $1 - \alpha$.

Mệnh đề 6.2 : Gọi giá trị của Y khi $X = x_0$ là Y_0. Thống kê

$$T = \frac{\hat{y}_{x_0} - Y_0}{S_{Y,X}\sqrt{1+h}} \sim t(n-2). \tag{6.20}$$

Dự báo giá trị của Y khi $X = x_0$ là tìm

– Khoảng tin cậy của Y khi $X = x_0$:

$$\hat{y}_{x_0} \pm t(n-2, \alpha/2)S_{Y,X}\sqrt{1+h} \tag{6.21}$$

$t(n-2, \alpha/2)$ là giá trị tới hạn mức $\dfrac{\alpha}{2}$ của phân phối Student với $n - 2$ bậc tự do.

Ví dụ 6.2 : (Trở lại Ví dụ 6.1)

a) Tìm hệ số xác định. Cho biết ý nghĩa của nó.

b) Tính sai số tiêu chuẩn của đường hồi quy : $S_{Y,X}$.

c) Với mức ý nghĩa 5%, chi phí quảng cáo có ảnh hưởng tới tổng doanh thu không ?

d) Dự báo trung bình tổng doanh thu khi tiền chi cho quảng cáo là 14 triệu đồng với độ tin cậy 95%.

GIẢI : a) Từ Ví dụ 6.1 ta có :

$$\sum x_i = 60 \; ; \; \sum x_i^2 = 898 \; ; \; \sum y_i = 110 \; ; \; \sum x_i y_i = 1658, \; \sum y_i^2 = 3082, \; n = 5 \; ;$$

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 178.$$

Từ (6.8) : $RSS = 3082 + 0,7868 \times 110 - 1,8989 \times 1658 = 20,1718$;

Từ (6.11) : $ESS = (-1,8989)^2\, 178 = 641,8362$; Từ (6.12) $TSS = 662,008$;

Từ (6.13) : $r^2 = 641,8362/662,008 = 0,9695$.

Kết luận : 96,95% sự thay đổi của tổng doanh thu là do quảng cáo gây nên, số phần trăm còn lại là do yếu tố khác.

b) Tính sai số chuẩn của đường hồi quy

$$S_{Y,X}^2 = \frac{\sum y_i^2 - b\sum x_i y_i - a\sum y_i}{n-2}$$

$$= \frac{3082 - 1,8989 \times 1658 + 0,7868 \times 110}{3} = 6,7239 \Rightarrow S_{Y,X} = 2,593.$$

c) Với mức ý nghĩa 5%, ta kiểm định $H_0 : B = 0$; $H_1 : B \neq 0$.

Theo (6.14) $$S_b = \frac{S_{Y,X}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}} = \frac{2,593}{\sqrt{898 - \frac{60^2}{5}}} = 0,1944 .$$

Theo (6.15) : $$T = \frac{1,8989}{0,1944} = 9,768 > t(3 \; ; 0,025) = 3,182.$$

Bác bỏ H_0, chấp nhận H_1, nghĩa là chi phí quảng cáo ảnh hưởng tới tổng doanh thu với mức ý nghĩa 5%.

Cách 2. Từ Ví dụ 5.3 (§1, Ch.5), P-giá trị của thống kê $T = t(3)$ là :

$$2P(T > 9,768) < 2P(T > 5,841) = 2.0,005 = 0,01 < \alpha = 0,05.$$

Do đó ta bác bỏ H_0, chấp nhận H_1.

d) Từ Ví dụ 6.1, khi X = 14 ;

$$\hat{y}_{14} = 1,8989 \times 14 - 0,7868 = 25,7978 ;$$

$$\sum x_i = 60 \Rightarrow \overline{x} = \frac{60}{5} = 12 ; t(3 ; 0,025) = 3,182.$$

Từ (6.17) : $h = \frac{1}{5} + \frac{(14-12)^2}{898 - \frac{60^2}{5}} = 0,2225$;

$$\hat{y}_{14} \pm t(5-2, \alpha/2) S_{Y,X}\sqrt{h} = 25,7978 \pm 3,182 \times 2,593\sqrt{0,2225}.$$

Khoảng tin cậy 95% : (21,9058 ; 29,6898) (chục triệu đồng).

Kết luận : Với độ tin cậy 95%, khi tiền chi cho quảng cáo là 14 triệu đồng thì trung bình tổng doanh thu từ 219,058 triệu đến 296,898 triệu đồng.

§2. PHÂN TÍCH TƯƠNG QUAN

1. Hệ số tương quan tuyến tính mẫu

Như ta đã biết trong [2, §4, Ch.2], hệ số tương quan $\rho(X, Y)$ dùng để đo sự phụ thuộc tương quan tuyến tính giữa hai biến ngẫu nhiên X và Y. Tuy nhiên để tính được $\rho(X, Y)$ ta cần biết phân phối xác suất của $\rho(X, Y)$. Điều này trong thực tế là khó. Vì vậy chúng ta sẽ căn cứ từ mẫu quan sát $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ của (X, Y) để ước lượng $\rho(X, Y)$ bằng hệ số sau :

Định nghĩa 6.1 : *Hệ số tương quan mẫu,* kí hiệu là r, được xác định như sau :

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}. \qquad (6.22)$$

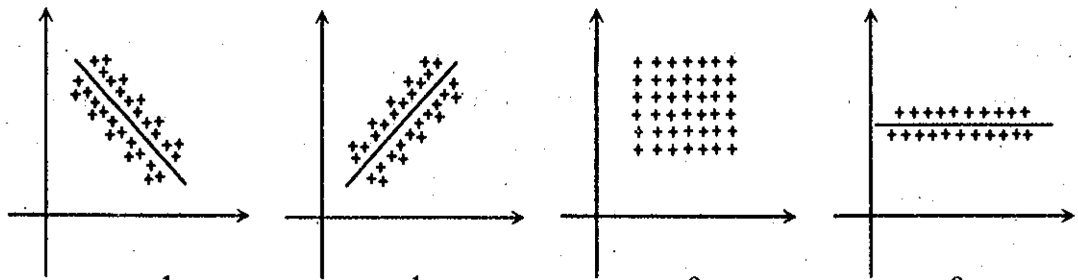
Biến đổi (6.22) nhận được :

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} . \qquad (6.23)$$

Chú ý : • $|r| \le 1$.

• $r = (\text{dấu của b}).\sqrt{r^2}$, r^2 là hệ số xác định. (6.24)

Hình biểu diễn các số liệu mẫu $(x_1, y_1), (x_2, y_2), \ldots ,(x_n, y_n)$ trên mặt phẳng toạ độ được gọi là đám mây điểm (H.6.2). Hình ảnh của chúng cũng có thể cho nhận xét sơ bộ về mối quan hệ giữa X và Y. Nếu đám mây điểm có xu hướng tập trung quanh một đường thẳng nào đó thì $|r|$ khá gần 1 ; còn nếu nó phân tán thành hình tròn hay hình vuông thì $|r|$ rất gần 0.



$r \approx -1$ $r \approx 1$ $r \approx 0$ $r \approx 0$

Tương quan âm Tương quan dương

Hình 6.2

Ví dụ 6.3 : Với Ví dụ 6.1. Hãy xác định hệ số tương quan mẫu.

GIẢI : Tính $\begin{cases} kx(1-x) & \text{khi } x \in [0\,;1] \\ 0 & \text{khi } x \notin [0\,;1] \end{cases}$; $\sum y_i = 110$; $\sum x_i^2 = 898$;

$\sum y_i^2 = 3082$; $\sum x_i y_i = 1658$;

$n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right) = 5.(1658) - 60.110 = 1690$;

$n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = 5.898 - (60)^2 = 890$;

$n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2 = 5.3083 - (110)^2 = 3310$;

Theo công thức (1.2) : $r = \dfrac{1690}{\sqrt{890}\sqrt{3310}} = 0{,}9846.$

2. Sự tương quan giữa các biến

a) Kiểm định giả thuyết về sự tương quan giữa các biến

Mệnh đề 6.3 : Nếu (X, Y) có phân phối chuẩn 2 chiều và $\rho = \rho(X, Y)$ là hệ số tương quan và r là hệ số tương quan mẫu của (X, Y) thì khi giả thuyết

$H_0 : \rho = 0$ đúng, thống kê

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2). \tag{6.25}$$

Bài toán 1 : Với mức ý nghĩa α, kiểm định về sự tương quan giữa X và Y.

Giải bài toán này bằng cách xét giả thuyết $H_0 : \rho = 0$; đối thuyết $H_1 : \rho \neq 0$.

Từ (6.25) ta có miền bác bỏ giả thuyết H_0

$$W_\alpha = \left\{ |T| > t\left(n-2, \frac{\alpha}{2}\right), T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right\}.$$

Ví dụ 6.4 : Một nhà xã hội học tuyên bố rằng kết quả học tập X của sinh viên (đo bằng điểm thi tốt nghiệp) không có liên quan tới thu nhập Y của gia đình họ. Lấy một mẫu ngẫu nhiên 20 sinh viên người ta tính được hệ số tương quan mẫu r = 0,4. Với mức ý nghĩa 1%, hãy cho kết luận về tuyên bố trên.

GIẢI : Gọi ρ là hệ số tương quan giữa X, Y. Xét

$H_0 : \rho = 0 ; H_1 : \rho \neq 0$

$$T = \frac{0,4\sqrt{20-2}}{\sqrt{1-(0,4)^2}} = 1,852 ;$$

$\alpha = 0,01 : t(18 ; 0,005) = 2,861.$

$|T| < t(18 ; 0,005)$ chấp nhận H_0, nghĩa là X và Y độc lập với nhau.

b) Kiểm định giả thuyết về mức độ tương quan giữa hai biến

Mệnh đề 6.5 : Với giả thuyết $H_0 : \rho = \rho_0 ; H_1 : \rho \neq \rho_0$

khi H_0 đúng tiêu chuẩn $Z = \dfrac{u-m}{\sigma} \approx N(0 ; 1)$ (6.26)

Với $u = \dfrac{1}{2}\ln\dfrac{1+r}{1-r}$; $m = \dfrac{1}{2}\ln\dfrac{1+\rho_0}{1-\rho_0}$; $\sigma = \dfrac{1}{\sqrt{n-3}}$.

b) Với mức ý nghĩa 5%, hãy cho biết lượng cung tiền của các nước này có ảnh hưởn tới tốc độ lạm phát không.

c) Tìm hệ số xác định, cho biết ý nghĩa của nó.

d) Với mức ý nghĩa 5%, lượng cung tiền và tốc độ lạm phát có tương quan vớ nhau không ?

6.4. Một nhà thống kê của một hãng ôtô lớn xây dựng mô hình dự báo về thời gian mu được hàng (tính bằng ngày từ lúc đặt hàng đến khi chính thức nhận được một loại ôt mới) của khách hàng qua số đơn đặt hàng. Kết quả điều tra ngẫu nhiên từ 16 loại ôt cho trong bảng :

Loại ôtô	Số đơn đặt hàng X	Thời gian mua được hàng Y
1	3	25
2	4	32
3	4	26
4	7	38
5	7	34
6	8	41
7	9	39
8	11	46
9	12	44
10	12	51
11	14	53
12	16	58
13	17	61
14	20	64
15	23	66
16	25	70

a) Ước lượng các hệ số hồi quy và lập phương trình hồi quy tuyến tính mẫu của theo X.

b) Nếu một loại ôtô có 19 đơn đặt hàng thì sau bao nhiêu ngày khách hàng sẽ có hàng.

c) Tính sai số tiêu chuẩn của đường hồi quy.

d) Tính hệ số xác định r^2 và nêu ý nghĩa trong bài toán này.

e) Tính hệ số tương quan mẫu r.

f) Với độ tin cậy 95%, hãy dự báo thời gian nhận được hàng cho những ôtô có 1 đơn đặt hàng.

i) Với độ tin cậy 95%, hãy dự báo thời gian nhận được hàng trung bình cho loại ôt có 19 đơn đặt hàng.

j) Với mức ý nghĩa 5% hãy xem có sự phụ thuộc tuyến tính giữa số đơn đặt hàng v thời gian nhận được hàng không.