Chương 6

Lý thuyết mẫu

6.1 Tổng thể, mẫu

Ta cần nghiên cứu đặc tính X (cân nặng, chiều cao ...) của tập lớn gồm N phần tử (N phần tử này được gọi là tổng thể). Thông thường ta không quan sát hết tất cả các phần tử của tập hợp này bởi vì các lý do:

- Làm hư hại tất cả các phần tử (kiểm tra đồ hộp, bắn thử đạn)
- Thời gian và kinh phí không cho phép – Số phần tử quá lớn (Nghiên cứu một đặc điểm nào của trẻ ta không thể đợi nghiên cứu toàn bộ trẻ em trên thế giới rồi mới đưa ra kết luận).

Do đó người ta lấy từ tổng thể này ra n phần tử (n phần tử này được gọi là mẫu) và quan sát đặc tính X để tính các đặc trưng trên mẫu sau đó sử dụng công cụ toán học để đưa ra kết luận cho tổng thể mà ta không có điều kiện khảo sát tất cả các phần tử.

Muốn mẫu lấy ra đại diện tốt cho tổng thể thì mẫu phải thỏa mãn hai điều kiện chính:

- Mẫu phải chọn ngẫu nhiên từ tổng thể.
- Các phân phối của mẫu phải được chọn độc lập nhau.

Khi quan sát phần tử thứ i, ta gọi X_i là biến ngẫu nhiên giá trị quan sát đặc tính X trên phần tử thứ i. Trong trường hợp cụ thể, giả sử X_i có giá trị x_n thì bộ n giá trị cụ thể $(x_1, \ldots, x_n)$ được gọi là mẫu cụ thể, cỡ mẫu cụ thể là n. Bộ n biến ngẫu nhiên độc lập $(X_1, \ldots, X_n)$ gọi là mẫu ngẫu nhiên.

Ví dụ 6.1. Khảo sát điểm môn xác suất thống kê của sinh viên lớp A có 100 sinh viên, tiến hành lấy mẫu có cỡ mẫu là 5. Gọi $X_i, i = 1, \ldots, 5$ là điểm của sinh viên thứ i trong 5 sinh viên được khảo sát. Nếu $X_1 = 3, X_2 = 7, X_3 = 8, X_4 = 5, X_5 = 7$ thì ta có mẫu cụ thể $(3, 7, 8, 5, 7)$. □

Tính chất 6.1 (Mẫu ngẫu nhiên). *Cho ngẫu nhiên* $(X_1, \ldots, X_n)$, *trong đó* X_i *giá trị quan sát đặc tính* X *trên phần tử thứ i. Khi đó:*

i. Các X_i *có cùng phân phối như* X.

ii. Các X_i *độc lập nhau.*

6.2 Mô tả dữ liệu

6.2.1 Phân loại mẫu ngẫu nhiên

Mẫu ngẫu nhiên còn được phân làm 2 loại:

- Mẫu chỉ quan tâm các phần tử của nó có tính chất A hay không gọi là *mẫu định tính*. Giả sử tỷ lệ phần tử A trên tổng thể là p, ta đặt

$$X_i = \begin{cases} 1 & \text{Nếu phần tử thứ } i \text{ loại } A \\ 0 & \text{Nếu phần tử thứ } i \text{ khác loại } A \end{cases}, \quad i = 1, \ldots, n$$

  Khi đó các X_i độc lập và cùng phân phối xác suất với X, $X_i \sim B(p)$.

- Mẫu mà ta quan tâm đến các yếu tố về lượng như là chiều cao, cân nặng, mức hao phí nhiên liệu của một loại động cơ,... gọi là *mẫu định lượng*.

6.2.2 Sắp xếp số liệu

Giả sử mẫu cụ thể $(x_1, \ldots, x_n)$ có k giá trị khác nhau $x_1, \ldots, x_k, (k \leq n)$ và x_i có tần số n_i (với $n_1 + \cdots + n_k = n$). khi đó, số liệu được sắp xếp theo thứ tự tăng dần của x_i như sau:

X	x_1	x_2	$\cdots$	x_k
n_i	n_1	n_2	$\cdots$	n_k

Bảng này gọi là *bảng tần số dạng điểm*.

Ví dụ 6.2. Khảo sát tuổi (X) trẻ bắt đầu đến trường ở một địa phương, lấy mẫu cỡ 10 ta có mẫu cụ thể như sau:

$$4, 5, 6, 7, 6, 6, 5, 5, 6, 6$$

Có bảng tần số dạng điểm:

X	4	5	6	7
n_i	1	3	5	1

□

Giả sử mẫu cụ thể $(x_1, \ldots, x_n)$ có nhiều giá trị khác nhau (quan sát từ biến ngẫu nhiên liên tục) thường người ta phân dữ liệu theo khoảng:

X	$a_0 - a_1$	$a_1 - a_2$	$\cdots$	$a_{k-1} - a_k$
n_i	n_1	n_2	$\cdots$	n_k

Bảng này gọi là *bảng tần số dạng khoảng.* Trong đó n_k là số quan sát có giá trị thuộc khoảng $(a_{k-1}; a_k]$. Khi tính toán ta đưa về *bảng tần số dạng điểm* bằng cách lấy giá trị chính giữa của mỗi khoảng $x_k = \dfrac{x_{k-1} + x_k}{x}$.

Ví dụ 6.3. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

□

Bảng tần số dạng điểm có dạng:

Thời gian	35	37	39	41	43
Số thai phụ	7	10	59	41	4

6.3 Các đặc trưng của mẫu

Giả sử ta cần nghiên cứu đặc tính X. Ký hiệu các tham số $\mu = \mathbb{E}X$ và $\sigma^2 = \mathbb{V}arX$. Trong thống kê các tham số này là các *tham số lý thuyết.*

6.3.1 Trung bình mẫu

Xét mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ lấy từ X.

Định nghĩa 6.2 (Trung bình mẫu)**.** *Biến ngẫu nhiên*

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

được gọi là trung bình mẫu.

Từ các tính chất của mẫu ngẫu nhiên, ta có:

Tính chất 6.3. *Trung bình mẫu có tính chất:*

i. $\mathbb{E}\bar{X} = \frac{1}{n}(\mathbb{E}X_1 + \cdots + \mathbb{E}X_n) = \frac{n\mu}{n} = \mu.$

ii. $\mathbb{V}ar\bar{X} = \frac{1}{n^2}(\mathbb{V}arX_1 + \cdots + \mathbb{V}arX_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

Cho mẫu cụ thể $(x_1, \ldots, x_n)$, trung bình mẫu $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ và trung bình của bình phương $\overline{x^2} = \frac{1}{n}(x_1^2 + \cdots + x_n^2)$

Chú ý. Khi số liệu cho dưới dạng bảng tần số thì $\bar{x} = \frac{1}{n}(x_1n_1 + \cdots x_kn_k)$ và trung bình của bình phương là $\overline{x^2} = \frac{1}{n}(x_1^2n_1 + \cdots x_k^2n_k)$

6.3.2 Phương sai mẫu

Xét mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ lấy từ X.

Định nghĩa 6.4 (Phương sai mẫu)**.** *Biến ngẫu nhiên*

$$\hat{S}^2 = \frac{1}{n}\left((X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right)$$

được gọi là phương sai mẫu.

Tính chất 6.5. *Phương sai mẫu có các tính chất*

i. $\hat{S}^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$

ii. $\mathbb{E}\hat{S}^2 = \frac{n-1}{n}\sigma^2.$

Cho mẫu cụ thể $(x_1, \ldots, x_n)$, phương sai mẫu $\hat{s}^2 = \overline{x^2} - \bar{x}^2$.

6.3.3 Phương sai mẫu có hiệu chỉnh

Xét mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ lấy từ X.

Định nghĩa 6.6 (Phương sai mẫu có hiệu chỉnh)**.** *Biến ngẫu nhiên*

$$S^2 = \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right)$$

được gọi là phương sai mẫu có hiệu chỉnh.

Tính chất 6.7. *Phương sai mẫu có các tính chất*

i. $S^2 = \dfrac{n}{n-1}\hat{S}^2$
ii. $\mathbb{E}S^2 = \sigma^2$.

Cho mẫu cụ thể $(x_1, \ldots, x_n)$, phương sai mẫu có hiệu chỉnh $s^2 = \dfrac{n}{n-1}\hat{s}^2$.

Ta thấy phương sai mẫu và phương sai mẫu có đơn vị đo bằng bình phương đơn vị đo của đặc tính X. Để chuyển về cùng đơn vị ta có khái niệm:

- Độ lệch chuẩn của mẫu, $\hat{s} = \sqrt{\hat{s}^2}$
- Độ lệch chuẩn của mẫu có hiệu chỉnh, $s = \sqrt{s^2}$

Ví dụ 6.4. Khảo sát chiều cao (cm) của nữ sinh trong một trường đại học ta có số liệu như sau

$$153; \quad 160; \quad 145; \quad 162; \quad 165; \quad 158$$

Tính: $\bar{x}, \hat{s}^2, s^2, \hat{s}, s$

Giải. Trung bình mẫu

$$\bar{x} = \frac{1}{6}(153 + 160 + 145 + 162 + 165 + 158) = 157,1666$$

Trung bình của bình phương

$$\overline{x^2} = \frac{1}{6}(153^2 + 160^2 + 145^2 + 162^2 + 165^2 + 158^2) = 24744,5$$

Phương sai mẫu

$$\hat{s}^2 = \overline{x^2} - \bar{x}^2 = 24744,5 - 157,1666^2 = 43,1598$$

Phương sai mẫu có hiệu chỉnh $s^2 = \dfrac{n}{n-1}\hat{s}^2 = \dfrac{6}{5}43,1598 = 51,7907$

Độ lệch chuẩn của mẫu $\hat{s} = \sqrt{\hat{s}^2} = \sqrt{43,1598}$

Độ lệch chuẩn của mẫu có hiệu chỉnh $s = \sqrt{s^2} = \sqrt{51,7907}$

Chú ý. Hướng dẫn sử dụng máy tính cầm tay tính các đặc trưng mẫu

a. Máy FX500MS (tương tự cho máy FX570MS)

- Bước 1: Ấn phím **Mod** đến khi màn hình xuất hiện chữ **SD** và chọn số tương ứng với mục **SD**
- Bước 2: Nhập số liệu
  153; M+; 160; M+; 145; M+; 162; M+; 165; M+; 158; M+
- Bước 3: Sau khi đã nhập hết các số liệu tiếp theo bạn nhấn phím **on**
- Bước 4: Xuất kết quả nhấn **Shift -> 2**
  - Tính $\bar{x}(\bar{x})$: **1; =**
  - Tính $\hat{s}(x\sigma n)$: **2; =**
  - Tính $s(x\sigma n-1)$: **3; =**

b. Máy FX500ES (tương tự cho FX570ES)

- Bước 1: Shift; Mode; ↓; chọn (Stat); chọn (Off) (Số liệu nhập vào không có tần số)
- Bước 2: Mod; chọn (Stat); chọn (1-Var)
- Bước 3: Nhập số liệu
  153; =; 160; =; 145; =; 162; =; 165; =; 158; =
- Sau khi đã nhập hết các số liệu tiếp theo bạn nhấn phím **on**
- Xuất kết quả **Shift; 1; chọn (Var)**
  - Tính $n(n)$: **1; =**
  - Tính $\bar{x}(\bar{x})$: **2; =**
  - Tính $\hat{s}(x\sigma n)$: **3; =**
  - Tính $s(x\sigma n-1)$: **4; =**

Ví dụ 6.5. Điểm môn xác suất thống kê của một số sinh viên khoa A cho như sau

Điểm	5	6	7	8	9	10
Số SV	2	4	12	15	6	2

a. Tính $\bar{x}$.

$$\bar{x} = \frac{1}{41}(5 \cdot 2 + 6 \cdot 4 + 7 \cdot 12 + 8 \cdot 15 + 9 \cdot 6 + 10 \cdot 2) = 7,6097$$

b. Tính $\hat{s}^2$.

$$\overline{x^2} = \frac{1}{41}(5^2 \cdot 2 + 6^2 \cdot 4 + 7^2 \cdot 12 + 8^2 \cdot 15 + 9^2 \cdot 6 + 10^2 \cdot 2) = 59,2195$$

suy ra $\hat{s}^2 = \overline{x^2} - \bar{x}^2 = 59,2195$ -$7,6097^2 = 1,3119$.

□

Chú ý. Hướng dẫn sử dụng máy tính cầm ta tính các đặc trưng mẫu (mẫu có tần số)

a. Máy FX500MS (tương tự cho máy FX570MS)

- Bước 1: Ấn phím **Mod** đến khi màn hình xuất hiện chữ **SD** và chọn số tương ứng với mục **SD**
- Bước 2: Nhập số liệu
  5; Shift;, ; 2; M+;
  6; Shift;, ; 4; M+;
  7; Shift;, ; 12; M+;
  8; Shift;, ; 15; M+;
  9; Shift;, ; 6; M+;
  10; Shift;, ; 2; M+
- Bước 4: Sau khi đã nhập hết các số liệu tiếp theo bạn nhấn phím **on**
- Bước 3: Xuất kết quả nhấn **Shift; 2**
  * Tính $\bar{x}(\bar{x})$: **1; =**
  * Tính $\hat{s}(x\sigma n)$: **2; =**
  * Tính $s(x\sigma n-1)$: **3; =**

b. Máy FX500ES (tương tự cho FX570ES)

– Bước 1: Shift; Mode; $\downarrow$; chọn (Stat); chọn (On) (Số liệu nhập vào có tần số)
– Bước 2: Mod; chọn (Stat); chọn (1-Var)
– Bước 3: Nhập số liệu

  Cột x: 5 ; =; 6; =; 7; =; 8; =; 9; =; 10; =

  Cột Freq: 2; =; 4; =; 12; =; 15; =; 6; =; 2; =
– Sau khi đã nhập hết các số liệu tiếp theo bạn nhấn phím **on**
– Xuất kết quả **Shift; 1; chọn (Var)**
  * Tính $n(n)$: **1; =**
  * Tính $\bar{x}(\bar{x})$: **2; =**
  * Tính $\hat{s}(x\sigma n)$: **3; =**
  * Tính $s(x\sigma n-1)$: **4; =**

Ví dụ 6.6. Năng suất lúa trong 1 vùng là đại lượng ngẫu nhiên có phân phối chuẩn. Gặt ngẫu nhiên 115 ha của vùng này, người ta thu được bảng số liệu:

Năng suất (tạ / ha)	40-42	42 – 44	44 – 46	46 – 48	48 – 50	50 – 52
Diện tích (ha)	7	13	25	35	30	5

Tính $\bar{x}$; $\hat{s}^2$. □

6.4 Phân phối xác suất của trung bình mẫu

a. Trường hợp $X \sim N(\mu;\sigma^3)$

Gọi $(X_1,\ldots,X_n)$ là mẫu ngẫu nhiên lấy từ X, khi đó $X_i \sim N(\mu;\sigma^2)$ và

$$\bar{X} \sim N\left(\mu;\frac{\sigma^2}{n}\right) \tag{6.1}$$

Trong trường hợp chưa biết σ^2 ta có

$$\frac{\bar{X}-\mu}{\dfrac{S}{\sqrt{n}}} \sim T^{n-1} \tag{6.2}$$

b. Trường hợp cỡ mẫu lớn*

$$\bar{X} \dot{\sim} N\left(\mu; \frac{\sigma^2}{n}\right) \tag{6.3}$$

Trong trường hợp chưa biết σ^2 ta có

$$\bar{X} \dot{\sim} N\left(\mu; \frac{S^2}{n}\right) \tag{6.4}$$

Chú ý. Khi mẫu $(X_1, \ldots, X_n)$ là mẫu định tính, tỷ lệ phần tử A trên tổng thể là p.

$$X_i = \begin{cases} 1 & \text{Nếu phần tử thứ } i \text{ loại } A, \mathbb{P}(X_i = 1) = p \\ 0 & \text{Nếu phần tử thứ } i \text{ khác loại } A, \mathbb{P}(X_i = 0) = q \end{cases}, \quad i = 1, \ldots, n$$

Các biến ngẫu nhiên X_i độc lập và $X_i \sim B(p)$, theo 4.5.2 ta có

$$X = X_1 + \ldots + X_n \dot{\sim} N\left(np; \sqrt{npq}^2\right) \text{ hay } \frac{X/n - p}{\sqrt{\frac{npq}{n}}} \dot{\sim} N(0; 1) \tag{6.5}$$

Trong đó X/n gọi là tỷ lệ phần tử A của mẫu, thường được ký hiệu F.

6.5 Đại lượng thống kê

Giả sử có mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ từ biến ngẫu nhiên X.

Định nghĩa 6.8. *Hàm số $\theta(X_1, \ldots, X_n)$ phụ thuộc vào mẫu được gọi là đại lượng thống kê. (Người ta còn gọi ngắn gọn là thống kê).*

Ví dụ 6.7. Trung bình mẫu, phương sai mẫu, tỷ lệ mẫu là các thống kê. □

*Trong thống kê, cỡ mẫu gọi là lớn khi $n \geq 30$.

Chương 7

Ước lượng tham số

7.1 Khái niệm chung

Giả sử biến ngẫu nhiên X có tham số θ chưa biết, dựa vào mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ ta đưa ra thống kê $\hat{\theta} = \theta(X_1, \ldots, X_n)$ để ước lượng giá trị của θ. Có hai phương pháp:

- Ước lượng điểm: Chỉ ra giá trị θ_0 để ước lượng cho θ.
- Ước lượng khoảng: Chỉ ra một khoảng $(\theta_1; \theta_2)$ chứa θ sao cho $\mathbb{P}\left(\theta_1 < \theta < \theta_2\right) = 1 - \alpha$ cho trước, trong đó $1 - \alpha$ gọi là độ tin cậy của ước lượng.

7.2 Ước lượng điểm

Định nghĩa 7.1 (Ước lượng không chệch)**.** *Thống kê $\hat{\theta}$ được gọi là ước lượng không chệch cho tham số θ nếu $\mathbb{E}(\hat{\theta}) = \theta$.*

Ví dụ 7.1. Giả sử biến ngẫu nhiên X có giá trị trung bình là μ. Từ X ta lập mẫu ngẫu nhiên $(X_1, \ldots, X_n)$. Khi đó $\bar{X}$ là ước lượng không chệch* cho μ □

Ta nhận thấy thống kê $\hat{\theta} = \dfrac{1}{2}(X_1 + X_n)$ cũng là một ước lượng không chệch cho θ. Vì vậy có thể nói có nhiều ước lượng không chệch cho θ. Vấn đề cần một tiêu chuẩn để chọn một thống kê $\hat{\theta}$ trong lớp các ước lượng không chệch cho θ.

*Theo tính chất 6.3

Định nghĩa 7.2 (Ước lượng hiệu quả). *Ước lượng không chệch $\hat{\theta}$ được gọi là ước lượng có hiệu quả của tham số θ nếu $\mathbb{V}ar\left(\hat{\theta}\right)$ nhỏ nhất trong các ước lượng không chệch của θ.*

Chú ý. Người ta chứng minh được rằng nếu $\hat{\theta}$ là ước lượng hiệu quả của θ thì phương sai của nó là

$$\mathbb{V}ar\left(\hat{\theta}\right) = \frac{1}{n\mathbb{E}\left(\frac{\partial \ln f(x,0)}{\partial \theta}\right)}$$

Trong đó $f(x,\theta)$ là hàm mật độ xác suất của biến ngẫu nhiên gốc.

Các thống kê $\bar{X}, S^2, F$ là ước lượng hiệu quả cho tham số μ, σ^2, p. Ta có quy tắc thực hành ước lượng điểm như sau:

Tham số lý thuyết	Đặc trưng mẫu	Ước lượng
$\mathbb{E}X = \mu$	$\bar{x}$	$\mu \approx \bar{x}$
$\mathbb{V}arX = \sigma^2$	s^2	$\sigma^2 \approx s^2$
p (tỷ lệ phần tử A)	f=tỷ lệ phần tử A trên mẫu	$p \approx f$

7.3 Ước lượng khoảng

7.3.1 Mô tả phương pháp.

Gọi θ là tham số của X chưa biết. Với mẫu cụ thể $(x_1, \ldots, x_n)$ ta tìm khoảng $(\theta_1; \theta_2)$ chứa θ sao cho $\mathbb{P}(\theta_1 < \theta < \theta_2) = 1 - \alpha$ cho trước.

- Khoảng $(\theta_1; \theta_2)$ gọi là khoảng tin cậy.
- $|\theta_1 - \theta_2|$ gọi là độ dài khoảng tin cậy.
- $1 - \alpha$ gọi là độ tin cậy.

7.3.2 Ước lượng khoảng cho trung bình

Gọi μ là trung bình của X chưa biết ta tìm khoảng $(\mu_1; \mu_2)$ chứa μ sao cho $\mathbb{P}(\mu_1 < \mu < \mu_2) = 1 - \alpha$. Khoảng tin cậy $(\mu_1; \mu_2) = (\bar{x} - \varepsilon; \bar{x} + \varepsilon)$, với ε gọi là độ chính xác của ước lượng. Trong đó ε tính như sau[†]

[†]Công thức tính độ chính xác được giải thích ở phụ lục B.1.1

Cỡ mẫu / $\mathbb{V}arX$	$n \geq 30$	$n < 30, X \sim N(\mu; \sigma^2)$
Biết σ^2	$\varepsilon = \frac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}}$ ($t_{\frac{1-\alpha}{2}}$ tra bảng A.2)	$\varepsilon = \frac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}}$ ($t_{\frac{1-\alpha}{2}}$ tra bảng A.2)
Không biết σ^2	$\varepsilon = \frac{s}{\sqrt{n}} t_{\frac{1-\alpha}{2}}$ ($t_{\frac{1-\alpha}{2}}$ tra bảng A.2)	$\varepsilon = \frac{s}{\sqrt{n}} t_{\alpha}^{n-1}$ (t_{α}^{n-1} tra bảng A.3).

Ví dụ 7.2. Khảo sát về thời gian tự học X (giờ/tuần) trong tuần của một số sinh viên hệ chính quy ở trường đại học A trong thời gian gần đây, người ta thu được bảng số liệu

X	5	6	7	8	9	10
Số SV	10	35	45	36	10	8

Ước lượng thời gian tự học trung bình của một sinh viên với độ tin cậy 95% cho hai trường hợp:

- **a.** Biết $\sigma = 2$
- **b.** Chưa biết σ

Giải. Từ mẫu ta tính được $n = 144; \bar{x} = 7,1736; s = 1,2366$.

Gọi μ là thời gian tự học trung bình của sinh viên. Khoảng ước lượng cho μ với độ tin cậy 95% có dạng

$$(\mu_1; \mu_2) = (\bar{x} - \varepsilon; \bar{x} + \varepsilon)$$

Tiếp theo ta tính ε cho từng trường hợp:

a. Biết $\sigma = 2$

$$\varepsilon = \frac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}} = \frac{2}{\sqrt{144}} 1,96 = 0,3267$$

Vậy khoảng ước lượng

$$(\mu_1; \mu_2) = (7,1736 - 0,3267; 7,1736 + 0,3267) = (6,8469; 7,5003)$$

Chú ý. Cho trước độ tin cậy là $1-\alpha = 0,95$ cho nên ta có $\frac{1-\alpha}{2} = 0,475$. Tra bảng A.2 ta có $t_{0,475} = 1,96$.

b. Không biết σ

$$\varepsilon = \frac{s}{\sqrt{n}} t_{\frac{1-\alpha}{2}} = \frac{1,2366}{\sqrt{144}} 1,96 = 0,202$$

Vậy khoảng ước lượng $(\mu_1; \mu_2) = (7,1736 - 0,202; 7,1736 + 0,202) = (6,9716; 7,3756)$

Chú ý. Với $t_{0,475} = 1,96$ được tính như câu a. $\square$

Ví dụ 7.3. Khảo sát cân nặng (kg) của gà khi xuất chuồng, người ta cân một số con và kết quả cho như sau:

2,1; 1,8; 2,0; 2,3; 1,7; 1,5; 2,0; 2,2; 1,8

Giả sử cân nặng của gà là biến ngẫu nhiên có phân phối chuẩn. Với độ tin cậy 95% ước lượng cân nặng trung bình của gà khi xuất chuồng:

a. Biết $\sigma = 0,3$.

b. Không biết σ.

Giải. Từ mẫu ta tính được $n = 9; \bar{x} = 1,9333; s = 0,2549$.

Gọi μ là cân nặng trung bình của gà khi xuất chuồng.

a. Cho biết $\sigma = 0,3$

$$\varepsilon = \frac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}} = \frac{0,3}{\sqrt{9}} 1,96 = 0,196$$

Vậy khoảng ước lượng

$$(\mu_1; \mu_2) = (1,9333 - 0,196; 1,9333 + 0,196) = (1,7373; 2,1293)$$

b. Không biết σ

$$\varepsilon = \frac{s}{\sqrt{n}} t_{\alpha}^{n-1} = \frac{0,2549}{\sqrt{9}} 2,306 = 0,1959$$

Vậy khoảng ước lượng $(\mu_1; \mu_2) = (1,9333 - 0,1959; 1,9333 + 0,1959) = (1,7374; 2,1292)$

Chú ý. Cho trước độ tin cậy là $1-\alpha = 0,95$ cho nên ta có $\alpha = 0,05$. Tra bảng A.3 ta có $t^8_{0,05} = 2,306$. □

Chú ý. Các chỉ tiêu ước lượng trung bình. Ta nhận thấy trong ước lượng trung bình có 3 chỉ tiêu chính $\varepsilon, 1-\alpha, n$. Nếu biết hai chỉ tiêu thì sẽ xác định được chỉ tiêu thứ 3.

a. Xác định cỡ mẫu n nhỏ nhất sao cho độ chính xác không lớn hơn ε và độ tin cậy là $1-\alpha$ (ở đây ta luôn giả sử cỡ mẫu lớn). Ta có

$$n \geq \left(\frac{\sigma}{\varepsilon} t_{\frac{1-\alpha}{2}}\right)^2 \left(\text{ hoặc } n \geq \left(\frac{s}{\varepsilon} t_{\frac{1-\alpha}{2}}\right)^2 \right)$$

n nhỏ nhất thỏa điều kiện trên là

$$n = \left| \left(\frac{\sigma}{\varepsilon} t_{\frac{1-\alpha}{2}}\right)^2 \right| + 1 \left(\text{ hoặc } n = \left| \left(\frac{s}{\varepsilon} t_{\frac{1-\alpha}{2}}\right)^2 \right| + 1 \right)$$

b. Xác định độ tin cậy của ước lượng khi biết độ chính xác của ước lượng. Trước hết xác định giá trị $t_{\frac{1-\alpha}{2}} = \dfrac{\varepsilon\sqrt{n}}{s}$. Và từ đây dễ dàng tính được $1-\alpha$.

Ví dụ 7.4. Cân thử 121 sản phẩm (đơn vị tính bằng kg) ta tính được $s^2 = 5,76$.

a. Xác định độ chính xác nếu muốn ước lượng trọng lượng trung bình với độ tin cậy 95%.

b. Xác định cỡ mẫu nhỏ nhất để lượng trọng lượng trung bình với độ tin cậy 95% và độ chính xác nhỏ hơn 0,4.

c. Xác định độ tin cậy nếu muốn ước lượng trung bình với độ chính xác là $\varepsilon = 0,5$.

Giải.

a. Xác định độ chính xác:

$$\varepsilon = \frac{s}{\sqrt{n}} t_{\frac{1-\alpha}{2}} = \frac{2,4}{\sqrt{121}} 1,96 = 0,4276$$

b. Xác định cỡ mẫu n.

$$n = \left| \left(\frac{s}{\varepsilon} t_{\frac{1-\alpha}{2}} \right)^2 \right| + 1 = \left| \left(\frac{2,4}{0,4} 1.96 \right)^2 \right| + 1 = 139$$

c. Xác định độ tin cậy, trước hết ta tính

$$t_{\frac{1-\alpha}{2}} = \frac{\varepsilon\sqrt{n}}{s} = \frac{0,5\sqrt{121}}{2,4} = 2,29$$

Tra bảng A.2 ta tính được $\frac{1-\alpha}{2} = 0,489$. Từ đó suy ra $1 - \alpha = 0,978$ □

7.3.3 Ước lượng khoảng cho tỷ lệ

Gọi p là tỷ lệ phần tử A chưa biết ta tìm khoảng $(p_1; p_2)$ chứa p sao cho $\mathbb{P}(p_1 < p < p_2) = 1 - \alpha$. Khoảng tin cậy

$$(p_1; p_2) = (f - \varepsilon; f + \varepsilon)$$

trong đó

- f là tỷ lệ phần tử A tính trên mẫu.
- ε gọi là độ chính xác của ước lượng được tính như sau[‡]

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

Ví dụ 7.5. Khảo sát tỷ lệ phế phẩm do một nhà máy sản xuất ra, người ta quan sát 800 sản phẩm thấy có 8 phế phẩm. Với độ tin cậy 95% hãy ước lượng tỷ lệ phế phẩn của nhà máy.

Giải. Gọi

f là tỷ lệ phế phẩm trên mẫu. $\left(f = \frac{8}{800} = 0,01 \right)$.

p là tỷ lệ phế phẩm của nhà máy.

[‡]Công thức tính độ chính xác được giải thích ở phụ lục B.1.2

Độ chính xác của ước lượng tỷ lệ

$$\varepsilon = \sqrt{\frac{f(1-f)}{n}} t_{\frac{1-\alpha}{2}} = \sqrt{\frac{0,01(1-0,01)}{800}} 1,96 = 0,0069$$

Vậy khoảng ước lượng cho p với độ tin cậy 95% là

$$(p_1; p_2) = (0,01 - 0,0069; 0,01 + 0,0069) = (0,0031; 0,0169)$$

□

Chú ý. Xác định các chỉ tiêu ước lượng

a Xác định cỡ mẫu n nhỏ nhất sao cho độ chính xác không lớn hơn ε và độ tin cậy là $1-\alpha$ Ta có $n \geq \frac{f(1-f)}{\varepsilon^2}\left(t_{\frac{1-\alpha}{2}}\right)^2$. n nhỏ nhất thỏa điều kiện trên là

$$n = \left|\frac{f(1-f)}{\varepsilon^2}\left(t_{\frac{1-\alpha}{2}}\right)^2\right| + 1$$

b Xác định độ tin cậy của ước lượng khi biết độ chính xác của ước lượng. Trước hết xác định giá trị

$$t_{\frac{1-\alpha}{2}} = \varepsilon\sqrt{\frac{n}{f(1-f)}}.$$

Và từ đây dễ dàng tính được $1-\alpha$ bằng bảng A.2.

Ví dụ 7.6. Quan sát 800 sản phẩm do một xí nghiệp sản xuất ra thấy có 128 mẫu loại A.

a. Xác định độ chính xác nếu muốn ước lượng tỷ lệ sản phẩm loại A với độ tin cậy 95%.

b. Xác định cỡ mẫu nhỏ nhất để ước lượng tỷ lệ sản phẩm loại A với độ chính xác nhỏ hơn 0,023 và độ tin cậy 95%.

c. Xác định độ tin cậy nếu muốn ước lượng tỷ lệ sản phẩm A với độ chính xác là 0,022.

Giải. Gọi:

f là tỷ lệ sản phẩm loại A tính trên mẫu $\left(f = \frac{128}{800} = 0,16\right)$.

p là tỷ lệ sản phẩm loại A do xí nghiệp sản xuất ra.

a. Độ chính xác của ước lượng

$$\varepsilon = \sqrt{\frac{f(1-f)}{n}} t_{\frac{1-\alpha}{2}} = \sqrt{\frac{0,16(1-0,16)}{800}} 1,96 = 0,0254$$

b. Xác định n

$$n = \left| \frac{f(1-f)}{\varepsilon^2} \left(t_{\frac{1-\alpha}{2}} \right)^2 \right| + 1 = \left| \frac{0,16(1-0,16)}{0,023^2} 1,96^2 \right| + 1 = 977$$

c. Xác định độ tin cậy $1-\alpha$

$$t_{\frac{1-\alpha}{2}} = \varepsilon \sqrt{\frac{n}{f(1-f)}} = 0,022 \sqrt{\frac{800}{0,016(1-0,016)}} = 1,69$$

Tra bảng A.2 ta tính được $\frac{1-\alpha}{2} = 0,4545$. Từ đó suy ra $1-\alpha = 0,909$ $\square$

7.4 Bài tập chương 7

Bài tập 7.1. Kiểm tra ngẫu nhiên 25 bóng đèn của một hãng điện tử, thấy tuổi thọ trung bình là 5000 giờ, độ lệch chuẩn của mẫu có hiệu chỉnh là 200 giờ. Giả sử tuổi thọ của bóng đèn có phân phối chuẩn. Tính khoảng ước lượng tuổi thọ trung bình của loại bóng đèn trên với độ tin cậy 95%. **(4917,44 giờ; 5082,56 giờ)**

Bài tập 7.2. Kiểm tra ngẫu nhiên 25 bóng đèn của một hãng điện tử, thấy độ lệch chuẩn của mẫu có hiệu chỉnh là 200 giờ. Giả sử tuổi thọ của bóng đèn có phân phối chuẩn. Sử dụng mẫu trên để ước lượng tuổi thọ trung bình của loại bóng đèn trên với độ chính xác là 73,12 giờ thì đảm bảo độ tin cậy bao nhiêu? **92%**

Bài tập 7.3. Thăm dò 25 người đang sử dụng điện thoại di động về số tiền phải trả trong 1 tháng, thấy số tiền trung bình một người phải trả là 200 ngàn đồng, độ lệch chuẩn của mẫu có hiệu chỉnh là 50 ngàn đồng. Giả sử số tiền phải trả trong một tháng có phân phối chuẩn. Với độ tin cậy là 95% tính khoảng ước lượng số tiền trung bình một người sử dụng điện thoại di động phải trả. **(179,36 ngàn đồng; 220,64 ngàn đồng)**

Bài tập 7.4. Thăm dò 25 người đang sử dụng điện thoại di động về số tiền phải trả trong 1 tháng, thấy độ lệch chuẩn của mẫu có hiệu chỉnh là 50 ngàn đồng. Giả sử số tiền phải trả trong một tháng có phân phối chuẩn. Với độ chính xác là 19,74 ngàn đồng thì độ tin cậy bao nhiêu? **94%**

Bài tập 7.5. Biết chiều dài của một loại sản phẩm là biến ngẫu nhiên có phân phối chuẩn. Đo ngẫu nhiên 10 sản phẩm loại này thì được chiều dài trung bình là 10,02m và độ lệch chuẩn của mẫu chưa hiệu chỉnh là 0,04m. Tính khoảng ước lượng chiều dài trung bình của loại sản phẩm này với độ tin cậy 95%. **(9,9898m; 10,0502m)**

Chương 8

Kiểm định giả thiết

8.1 Bài toán kiểm định giả thiết

8.1.1 Giả thiết không, đối thiết

Trong chương này chúng ta sẽ đề cập đến bài toán thống kê liên quan đến tham số θ, với giá trị của nó không biết thuộc không gian tham Θ. Tuy nhiên chúng ta sẽ giả sử Θ có thể được phân chia thành hai tập tách biệt Θ_0 và Θ_1 và nhiệm vụ của người làm thống kê phải quyết định xem θ thuộc Θ_0 hay Θ_1.

Chúng ta đặt H_0 để ký hiệu giả thiết $\theta \in \Theta_0$, và H_1 ký hiệu giả thiết $\theta \in \Theta_1$. Bởi vì Θ_0 và Θ_1 tách biệt và $\Theta_0 \cap \Theta_1 = \Theta$, chính xác chỉ có giả thiết H_0 hoặc H_1 là đúng. Chúng ta phải quyết định chấp nhận H_0 để bác bỏ H_1 hoặc ngược lại. Bài toán thuộc dạng này được gọi là kiểm định giả thiết.

Đến đây, chúng ta thấy vai trò của giả thiết H_0 và H_1 cơ bản giống nhau. Trong hầu hết các bài toán kiểm định, hai giả thiết này hơi khác. Để phân biệt giữa hai giả thiết này ta gọi H_0 gọi là *giả thiết không* và H_1 gọi là *đối thiết*. Chúng ta sẽ dùng các thuật ngữ này trong phần còn lại của chương.

8.1.2 Miền tới hạn

Ta xét bài toán với giả thiết có dạng như sau:

$$\begin{cases} \text{Giả thiết không } H_0 : & \theta \in \Theta_0 \\ \text{Đối thiết } H_1 : & \theta \in \Theta_1 \end{cases}$$

Giả sử trước khi chúng ta quyết định giả thiết nào sẽ được chấp nhận, chúng ta có mẫu ngẫu nhiên $X_1, \ldots, X_n$ được trích từ phân phối của đặc tính X với tham số θ chưa biết. Chúng ta ký hiệu Ω là không gian mẫu, Ω chứa tất cả các kết quả có thể xảy ra khi lấy mẫu ngẫu nhiên.

Trong quá trình kiểm định, chúng ta sẽ chia Ω thành hai tập con. Một tập chứa tất cả các giá trị của X sao cho ta chấp nhận H_0, và tập còn lại chứa tất cả các giá trị của X sao cho ta bác bỏ H_0 và chấp nhận H_1. Tập các giá trị của X để H_0 bị bác bỏ gọi là *miền tới hạn*, ký hiệu C .

Với mỗi giá trị $\theta \in \Theta$ ta đặt hàm *lực lượng* $\pi(\theta)$ là xác suất dẫn đến bác bỏ H_0, ngược lại $1 - \pi(\theta)$ là xác suất dẫn đến chấp nhận H_0. Nếu ký hiệu C là miền tới hạn của kiểm định, hàm $\pi(\theta)$ được xác định bởi quan hệ

$$\pi(\theta) = \mathbb{P}\left(X \in C|\theta\right), \quad \forall\theta \in \Theta$$

Bởi vì $\pi(\theta)$ là xác suất ứng với mỗi θ thì H_0 bị bác bỏ, trong trường hợp lý tưởng hàm $\pi(\theta) = 0$ với mọi $\theta \in \Theta_0$ và $\pi(\theta) = 1$ với mọi $\theta \in \Theta_1$. Nếu hàm $\pi(\theta)$ có các giá trị này thì bất chấp giá trị thực tế θ nào ta luôn có kết luận đúng với xác suất 1.

8.1.3 Hai loại sai lầm

Khi chọn một trong hai quyết định trên sẽ nẩy sinh ra hai sai lầm:

- Sai lầm loại I: Bác bỏ H_0 khi H_0 đúng, xác suất sai lầm loại I là
$$\mathbb{P}\left(C|H_0\right) = \mathbb{P}\left((X_1, \ldots, X_n) \in C|H_0\right)$$
- Sai lầm loại II: Chấp nhận H_0 khi H_0 sai, xác suất sai lầm loại II là
$$\mathbb{P}\left(\bar{C}|H_1\right) = \mathbb{P}\left((X_1, \ldots, X_n) \notin C|H_1\right)$$

Ví dụ 8.1. Cần nghiên cứu tác dụng phụ của một loại thuốc mới vừa được nghiên cứu ta đặt giả thiết và đối thiết như sau

$$\begin{cases} \text{Giả thiết } H_0: & \text{Thuốc có tác dụng phụ} \\ \text{Đối thiết } H_1: & \text{Thuốc không có tác dụng phụ} \end{cases}$$

Kết luận \ Thực tế	Thuốc có tác dụng phụ	Thuốc không có tác dụng phụ
Chấp nhận H_0	Kết luận đúng	Sai lầm loại II
Bác bỏ H_0	Sai lầm loại I	Kết luận đúng

Việc đặt giả thiết như trên khi sai lầm loại I xảy ra là tai hại hơn sai lầm loại II (thuốc có tác dụng phụ mà kết luận thuốc không có tác dụng phụ). □

Lẽ tự nhiên là ta chọn miền C sao cho cực tiểu cả hai xác suất phạm sai lầm. Song không thể cực tiểu đồng thời cả hai sai lầm khi cỡ mẫu cố định, bởi vì hai xác suất trên hiên hệ nhau bởi:

$$\mathbb{P}\left(C|H_0\right) + \mathbb{P}\left(\bar{C}|H_0\right) = 1; \mathbb{P}\left(C|H_1\right) + \mathbb{P}\left(\bar{C}|H_1\right) = 1.$$

Do đó C cực tiểu $\mathbb{P}\left(C|H_0\right)$ chưa chắc đã cực tiểu $\mathbb{P}\left(\bar{C}|H_1\right)$

8.1.4 Phương pháp chọn miền tới hạn

Ta cố định một loại xác suất sai lầm và tìm miền C sao cho xác suất phạm sai lầm kia đạt giá trị nhỏ nhất. Thông thường ta cố định xác suất sai lầm loại I: $\mathbb{P}\left(C|H_0\right) \leq \alpha$, ta sẽ chọn miền C sao cho $\mathbb{P}\left(\bar{C}|H_1\right)$ đạt cực tiểu hay $\mathbb{P}\left(C|H_1\right)$ cực đại, nghĩa là tim C sao cho:

$$\begin{cases} \mathbb{P}\left(C|H_0\right) \leq \alpha \\ \mathbb{P}\left(C|H_1\right) \text{ đạt cực đại} \end{cases} \text{hay} \begin{cases} \pi(\theta) \leq \alpha \text{ với } \theta \in \Theta_0 \\ \pi(\theta) \text{ đạt cực đại với } \theta \in \Theta_1 \end{cases} \tag{8.1}$$

Ta gọi α là *mức ý nghĩa* của kiểm định, khi cố định α và có hàm lực lượng $\pi(\theta), \forall\theta \in \Theta_1$ lớn nhất thì qui tắc này gọi là qui tắc mạnh nhất.

8.2 Kiểm định giả thiết về trung bình

Giả sử μ (chưa biết) là trung bình của biến ngẫu nhiên X, cần kiểm định*

$$\begin{cases} \text{Giả thiết } H_0: & \mu = \mu_0 \\ \text{Đối thiết } H_1: & \mu \neq \mu_0 \end{cases}$$

*Xem giải thích phụ lục B.2.1

Cỡ mẫu / $\mathbb{V}arX$	$n \geq 30$	$n < 30, X \sim N(\mu;\sigma^2)$
Biết σ^2	$t = \dfrac{\lvert\bar{x}-\mu_0\rvert}{\sigma}\sqrt{n}$ $t_{\frac{1-\alpha}{2}}$ (Bảng A.2)	$t = \dfrac{\lvert\bar{x}-\mu_0\rvert}{\sigma}\sqrt{n}$ $t_{\frac{1-\alpha}{2}}$ (Bảng A.2)
Không biết σ^2	$t = \dfrac{\lvert\bar{x}-\mu_0\rvert}{s}\sqrt{n}$ $(t_{\frac{1-\alpha}{2}}$ (Bảng A.2)	$t = \dfrac{\lvert\bar{x}-\mu_0\rvert}{s}\sqrt{n}$ t_{α}^{n-1} (Bảng A.3)

Kết luận

- Chấp nhận giả thiết H_0 khi $t \leq t_{\frac{1-\alpha}{2}}$ hoặc $\left(t \leq t_{\alpha}^{n-1}\right)$
- Bác bỏ giả thiết H_0 khi $t > t_{\frac{1-\alpha}{2}}$ hoặc $\left(t > t_{\alpha}^{n-1}\right)$

Ví dụ 8.2. Cân thử 15 con gà tây ở 1 trại chăn nuôi khi xuất chuồng ta tính được $\bar{x} = 3,62kg$. Cho biết $\sigma^2 = 0,01$.

a. Giám đốc trại tuyên bố trọng lương trung bình của gà tây là $3,5kg$ thì có tin được không với mức ý nghĩa α =1%.

b. Giả sử người ta dùng thức ăn mới và khi xuất chuồng trọng lượng trung bình của gà tây là 3,9 kg. Cho kết luận về loại thức ăn này với mức ý nghĩa $\alpha = 1\%$.

Giải.

a. Gọi μ cân nặng trung bình của gà khi xuất chuồng. Cần kiểm định:

$$\begin{cases} \text{Giả thiết } H_0: & \mu = 3,5kg \\ \text{Đối thiết } H_1: & \mu \neq 3,5kg \end{cases}$$

$$t = \frac{|\bar{x}-\mu_0|}{\sigma}\sqrt{n} = \frac{|3,62-3,5|}{0,1}\sqrt{15} = 4,6 \text{ và } t_{\frac{1-\alpha}{2}} = 2,58$$

$\left(t > t_{\frac{1-\alpha}{2}}\right)$ nên bác bỏ giả thiết. Vậy giám đốc báo cáo sai.

b. Gọi μ cân nặng trung bình của gà tây khi xuất chuồng (trước khi sử dụng thức ăn mới)

$$\begin{cases} \text{Giả thiết } H_0: & \mu = 3,9kg \\ \text{Đối thiết } H_1: & \mu \neq 3,9kg \end{cases}$$

$$t = \frac{|\overline{x}_n - \mu_0|}{\sigma}\sqrt{n} = \frac{|3,62 - 3,9|}{0,1}\sqrt{15} = 10,84$$

$\left(t > t_{\frac{1-\alpha}{2}}\right)$ nên bác bỏ giả thiết. Vậy thức ăn mới có tác dụng tốt. □

8.3 Kiểm định giả thiết về tỷ lệ

Giả sử p(chưa biết) là tỷ lệ phần tử loại A, cần kiểm định†

$$\begin{cases} \text{Giả thiết } H_0: & p = p_0 \\ \text{Đối thiết } H_1: & p \neq p_0 \end{cases}$$

Qui tắc thực hành như sau: Tính giá trị

$$t = \frac{|f - p_0|}{\sqrt{p_0(1-p_0)}}\sqrt{n} \text{ và } t_{\frac{1-\alpha}{2}} \text{(Bảng A.2)}$$

Trong đó f là tỷ lệ phần tử A trên mẫu

Kết luận:

- Chấp nhận giả thiết H_0 khi $t \leq t_{\frac{1-\alpha}{2}}$.
- Bác bỏ giả thiết H_0 khi $t > t_{\frac{1-\alpha}{2}}$.

Ví dụ 8.3. Để kiểm tra một loại súng thể thao, người ta cho bắn 1000 viên đạn vào bia thấy có 540 viên trúng mục tiêu. Sau đó, bằng cải tiến kỹ thuật người ta tính được tỷ lệ trúng mục tiêu là 70%. Hãy cho kết luận về cải tiến với mức ý nghĩa 1%.

Giải. Gọi

- p là tỷ lệ bắn trúng trước cải tiến.
- f là tỷ lệ bắn trúng trên mẫu (trước cải tiến).

†Xem giải thích ở phụ lục B.2.2

Cần kiểm định giả thiết

$$\begin{cases} \text{Giả thiết } H_0: & p = 0,7 \\ \text{Đối thiết } H_1: & p \neq 0,7 \end{cases}$$

Tiến hành kiểm tra giả thiết

$$t = \frac{|f - p_0|}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{|0,54 - 0,7|}{\sqrt{0,7.0,3}}\sqrt{1000} = 11,04$$

$1-\alpha = 0,99$ tra bảng A.2 ta được $t_{\frac{1-\alpha}{2}} = 2,58$. Kết luận cải tiến có tác dụng tốt. □

Ví dụ 8.4. Kiểm tra 800 sinh viên thấy có 128 sinh viên giỏi. Trường báo cáo tổng kết là có 40% sinh viên giỏi thì có thể chấp nhận được không với mức ý nghĩa 5%.

Giải. Gọi

- p tỷ lệ sinh viên giỏi thực tế (chưa biết)
- f tỷ lệ sinh viên giỏi tính trên mẫu $f = \dfrac{128}{800} = 0,16$

$$\begin{cases} \text{Giả thiết } H_0: & p = 40\% \\ \text{Đối thiết } H_1: & p \neq 40\% \end{cases}$$

Tiến hành kiểm tra giả thiết

$$t = \frac{|f - p_0|}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{|0,16 - 0,4|}{\sqrt{0,4.0,6}}\sqrt{800} = 13,871$$

$1-\alpha = 0,95$ tra bảng A.2 ta được $t_{\frac{1-\alpha}{2}} = 1,96$. Kết luận báo cáo là sai sự thật, tỷ lệ sinh viên giỏi trong thực tế thấp hơn nhiều. □

8.4 So sánh hai giá trị trung bình

Giả sử X_1 và X_2 là hai biến ngẫu nhiên độc lập có giá trị trung bình là μ_1 và μ_2 . Cần kiểm định

$$\begin{cases} \text{Giả thiết } H_0: & \mu_1 = \mu_2 \\ \text{Đối thiết } H_1: & \mu_1 \neq \mu_2 \end{cases}$$

Ký hiệu các đặc trưng của mẫu 1, 2 lấy từ tổng thể 1, tổng thể 2.

Mẫu	Cỡ mẫu	Trung bình mẫu	Độ lệch chuẩn có hiệu chỉnh
I	n_1	$\bar{x}_1$	s_1
II	n_2	$\bar{x}_2$	s_2

Cỡ mẫu / $\mathbb{V}arX_1; \mathbb{V}arX_2$	$n_1; n_2 \geq 30$	$n_1 < 30; X_1 \sim N(\mu_1; \sigma_1^2)$ $n_2 < 30; X_2 \sim N(\mu_2; \sigma_2^2)$
Biết $\sigma_1^2; \sigma_2^2$	$t = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ $t_{\frac{1-\alpha}{2}}$ (Bảng A.2)	$t = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ $t_{\frac{1-\alpha}{2}}$ (Bảng A.2)
Không biết $\sigma_1^2; \sigma_2^2$	$t = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ $t_{\frac{1-\alpha}{2}}$ (Bảng A.2)	$t = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}}$ $t_{\alpha}^{n_1+n_2-2}$ (Bảng A.3)

Trong đó $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ gọi là phương sai gộp.

Kết luận:

- Chấp nhận giả thiết H_0 khi $t \leq t_{\frac{1-\alpha}{2}}$ hoặc $\left(t \leq t_{\alpha}^{n_1+n_2-2}\right)$
- Bác bỏ giả thiết H_0 khi $t > t_{\frac{1-\alpha}{2}}$ hoặc $\left(t > t_{\alpha}^{n_1+n_2-2}\right)$

Ví dụ 8.5. Cân thử 100 trái cây ở nông trường I ta tính được $\bar{x}_1 = 101,2$; $s_1^2 = 571,7$ và 361 trái cây ở nông trường II tính được $\bar{x}_2 = 66,39$; $s_2^2 = 29,72$. So sánh trọng lượng trung bình của trái cây ở hai nông trường với mức ý nghĩa 1%.

Giải. Gọi μ_1, μ_2 cân nặng trung bình của trái cây ở nông trường I và II.

Cần kiểm định

$$\begin{cases} \text{Giả thiết } H_0: & \mu_1 = \mu_2 \\ \text{Đối thiết } H_1: & \mu_1 \neq \mu_2 \end{cases}$$

Mẫu	Cỡ mẫu	Trung bình mẫu	Độ lệch chuẩn có hiệu chỉnh
I	$n_1 = 100$	$\bar{x}_1 = 101,2$	$s_1^2 = 571,7$
II	$n_2 = 361$	$\bar{x}_2 = 66,39$	$s_2^2 = 29,72$

Tính giá trị

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{|101,2 - 66,39|}{\sqrt{\dfrac{571,7}{100} + \dfrac{29,72}{361}}} = 14,4549$$

Tra bảng A.2 ta được $t_{\frac{1-\alpha}{2}} = t_0,495 = 2,58$. Vậy $t > t_{\frac{1-\alpha}{2}}$ cho nên bác bỏ giả thiết H_0 hay cân nặng trung bình của trái cây ở hai địa phương không bằng nhau. □

Ví dụ 8.6. Đo đường kính 20 trục máy do máy I sản xuất và 22 trục máy do máy II sản xuất ta tính được $\bar{x}_1 = 251,7$; $s_1^2 = 52,853$ và $\bar{x}_2 = 249,8$; $s_2^2 = 56,2$. Có thể xem đường kính trung bình của các trục máy ở 2 máy như nhau với mức ý nghĩa 1% không?

Giải. ______________________________

8.5 So sánh hai tỷ lệ

Gọi $p_1; p_2$ tỷ lệ phần tử A trên tổng thể 1 và 2 chưa biết. Ta cần kiểm định

$$\begin{cases} \text{Giả thiết } H_0: & p_1 = p_2 \\ \text{Đối thiết } H_1: & p_1 \neq p_2 \end{cases}$$

Tính: $f = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ (Tỷ lệ phần tử A chung của 2 mẫu), trong đó $f_1; f_2$ tỷ lệ phần tử A trên mẫu 1, 2.

$$t = \frac{|f_1 - f_2|}{\sqrt{f(1-f)\left(\dfrac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Kết luận:

- Chấp nhận giả thiết H_0 khi $t \leq t_{\frac{1-\alpha}{2}}$.
- Bác bỏ giả thiết H_0 khi $t > t_{\frac{1-\alpha}{2}}$.

Ví dụ 8.7. Từ hai đám đông tiến hành 2 mẫu với $n_1 = 100, n_2 = 120$ tính được tỷ lệ phần tử loại A trên mẫu 1, 2 lần lượt $f_1 = 0,2$ và $f_2 = 0,3$. Với mức ý nghĩa $\alpha = 1\%$ cho kết luận tỷ lệ phần tử A của 2 đám đông có như nhau không.

Giải. Tính $f = \dfrac{20 + 36}{100 + 120} = 0,255$.

Gọi p_1, p_2 (chưa biết) tỷ lệ phần tử A trên tổng thể 1, 2. Cần kiểm định giả thiết

$$\begin{cases} \text{Giả thiết } H_0: & p_1 = p_2 \\ \text{Đối thiết } H_1: & p_1 \neq p_2 \end{cases}$$

$$t = \frac{|0,2 - 0,3|}{\sqrt{0,255.0,745\left(\dfrac{1}{100} + \dfrac{1}{120}\right)}} = 1,695$$

Với $\alpha = 1\%$ tra bảng A.2 tính được $t_{\frac{1-\alpha}{2}} = 2,58$. Kết luận chấp nhận giả thiết H_0 hay tỷ lệ phần tử A trên 2 mẫu như nhau. □

Ví dụ 8.8. Kiểm tra 120 sinh viên trường A thấy có 80 sinh viên giỏi, 150 sinh viên trường B có 90 sinh viên giỏi. Hỏi tỷ lệ sinh viên giỏi của 2 trường như nhau không? Biết mức ý nghĩa là 5%.

Giải.

Ví dụ 8.9. Kiểm tra 230 sản phẩm của ca ngày thấy có 4 sản phẩm hỏng. Còn kiểm tra 160 sản phẩm của ca đêm thấy có 3 sản phẩm hỏng. Kết luận tỷ lệ sản phẩm hỏng phụ thuộc vào ca có đúng không với mức ý nghĩa 1%.

Giải.

8.6 Bài tập chương 8

Bài tập 8.1. Biết chiều dài của một loại sản phẩm là biến ngẫu nhiên có phân phối chuẩn. Đo ngẫu nhiên 10 sản phẩm loại này thì được chiều dài trung bình là 10,02m và độ lệch chuẩn của mẫu chưa hiệu chỉnh là 0,04m. Kiểm định giả thuyết H: "chiều dài trung bình của loại sản phẩm này là 10,0543m" có giá trị kiểm định t là bao nhiêu và cho kết luận với mức ý nghĩa 3%. **t = 2,5703; chiều dài trung bình của loại sản phẩm này là 10,0543m với mức ý nghĩa 3%**

Bài tập 8.2. Khảo sát về thời gian tự học (giờ/tuần) của sinh viên hệ chính quy ở trường đại học A trong học kỳ này. Tiến hành lấy mẫu, người ta thu được bảng số liệu:

Thời gian	$3-5$	$5-7$	$7-9$	$9-11$	$11-13$
Số sinh viên	5	14	16	8	6

a. Tìm khoảng ước lượng thời gian tự học trung bình trong tuần của sinh viên trường A với độ tin cậy 95%. **(7,1817giờ/tuần; 8,4917giờ/tuần)**

b. Để ước lượng thời gian tự học trung bình trong tuần với độ tin cậy 95% và độ chính xác nhỏ hơn $\varepsilon = 0,6$(giờ/tuần) thì cỡ mẫu nhỏ nhất là bao nhiêu? **59**

c. Sử dụng mẫu ban đầu để ước lượng thời gian tự học trung bình trong tuần với độ chính xác $\varepsilon = 0,6$(giờ/tuần) thì đảm bảo độ tin cậy là bao nhiêu? **92,82%**

d. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Với độ tin cậy 95% khoảng ước lượng tỷ lệ sinh viên chăm học là bao nhiêu? **(15,92%; 41,22%)**

e. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Để ước lượng tỷ lệ sinh viên "chăm học" với độ tin cậy

95% và độ chính xác nhỏ hơn $\varepsilon = 0,12$ thì cỡ mẫu nhỏ nhất là bao nhiêu? **55**

f. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Sử dụng mẫu trên để ước lượng tỷ lệ sinh viên "chăm học" với độ chính xác $\varepsilon = 0,12$ thì đảm bảo độ tin cậy là bao nhiêu? **93,71%**

g. Tính giá trị thống kê t để kiểm định giả thuyết H: "thời gian tự học trung bình của sinh viên trường A là 8,4(giờ/tuần)" và cho kết luận với mức ý nghĩa 5%. **t = 1,6855; thời gian tự học trung bình của sinh viên trường A là 8,4(giờ/tuần) với mức ý nghĩa 5%**

h. Trong kiểm định giả thuyết H: "thời gian tự học trung bình của sinh viên trường A là 8,4(giờ/tuần)", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là bao nhiêu? **9,1%**

i. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Tính giá trị thống kê t để kiểm định giả thuyết H: "tỷ lệ sinh viên chăm học ở trường A là 18%" và cho kết luận với mức ý nghĩa 5%. **t = 1,9261; tỷ lệ sinh viên chăm học ở trường A là 18% với mức ý nghĩa 5%**

j. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Trong kiểm định giả thuyết H: "tỷ lệ sinh viên chăm học ở trường A là 18%", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là bao nhiêu? **5,36%**

k. Trường B khảo sát 64 sinh viên về thời gian tự học. Người ta tính được độ lệch chuẩn của mẫu có hiệu chỉnh là 2(giờ/tuần) và trung bình mẫu là 8,5(giờ/tuần). Tính giá trị thống kê t để kiểm định giả thuyết H: "thời gian tự học trung bình trong tuần của sinh viên hai trường là như nhau" và cho kết luận với mức ý nghĩa 5%. **t = 1,5893; thời gian tự học trung bình trong tuần của sinh viên hai trường là như nhau mức ý nghĩa 5%** ____________

l. Trường B khảo sát 64 sinh viên về thời gian tự học. Người ta tính được độ lệch chuẩn của mẫu có hiệu chỉnh là 2(giờ/tuần) và trung bình mẫu là 8,5(giờ/tuần). Trong kiểm định giả thuyết H: "thời gian tự học trung bình trong tuần của sinh viên hai trường là như nhau", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là bao nhiêu? **11,18%** ____________

m. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Trường B khảo sát 64 sinh viên về thời gian tự học thấy có 28 sinh viên "chăm học". Tính giá trị thống kê t để kiểm định giả thuyết H: "tỷ lệ sinh viên "chăm học" của hai trường là như nhau" và cho kết luận với mức ý nghĩa 5%. **t = 1,6546; tỷ lệ sinh viên chăm học của hai trường là như nhau với mức ý nghĩa 5%** ____________

n. Những sinh viên có thời gian tự học từ 9(giờ/tuần) trở lên gọi là sinh viên "chăm học". Trường B khảo sát 64 sinh viên về thời gian tự học thấy có 28 sinh viên "chăm học". Trong kiểm định giả thuyết H: "tỷ lệ sinh viên "chăm học" của hai trường là như nhau", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là bao nhiêu? **9,7%**

Bài tập 8.3. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Khoảng ước lượng thời gian mang thai trung bình của thai phụ với độ tin cậy 95% là:

A. (39,1049 tuần; 39,7215 tuần).
B. (38,1049 tuần; 38,7215 tuần).
C. (37,1049 tuần; 37,7215 tuần).
D. (40,1049 tuần; 40,7215 tuần).

Bài tập 8.4. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Để ước lượng thời gian mang thai trung bình của thai phụ với độ tin cậy 95% và độ chính xác nhỏ hơn $\varepsilon = 0,25$(tuần) thì cỡ mẫu nhỏ nhất là:

A. 175. B. 185. C. 195. D. 165.

Bài tập 8.5. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Sử dụng mẫu trên để ước lượng thời gian mang thai trung bình của thai phụ với độ chính xác $\varepsilon = 0,25$(tuần) thì đảm bảo độ tin cậy:

A. 86,82%. B. 87,82%. C. 88,82%. D. 89,82%.

Bài tập 8.6. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Với độ tin cậy 95% khoảng ước lượng tỷ lệ thai phụ sinh non:

A. (2,63%; 10,95%). B. (3,63%; 11,95%).
C. (4,63%; 12,95%). D. (1,63%; 9,95%).

Bài tập 8.7. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Để ước lượng tỷ lệ thai phụ sinh non với độ tin cậy 95% và độ chính xác nhỏ hơn $\varepsilon = 0,04$ thì cỡ mẫu nhỏ nhất là:

A. 121. B. 141. C. 151. D. 131.

Bài tập 8.8. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	$34-36$	$36-38$	$38-40$	$40-42$	$42-44$
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Sử dụng mẫu trên để ước lượng tỷ lệ thai phụ sinh non với độ chính xác $\varepsilon = 0,04$ thì đảm bảo độ tin cậy:

A. 91,99%. B. 95,99%. C. 93,99%. D. 97,99%.

Bài tập 8.9. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Giá trị thống kê t để kiểm định giả thuyết H: "thời gian mang thai trung bình của thai phụ là 39,7 tuần" là:

A. t = 1,8231; thời gian mang thai trung bình của thai phụ là 39,7 tuần với mức ý nghĩa 7%.

B. t = 1,8231; thời gian mang thai trung bình của thai phụ là 39,7 tuần với mức ý nghĩa 5%.

C. t = 2,8231; thời gian mang thai trung bình của thai phụ lớn hơn 39,7 tuần với mức ý nghĩa 5%.

D. t = 2,8231; thời gian mang thai trung bình của thai phụ nhỏ hơn 39,7 tuần với mức ý nghĩa 3%.

Bài tập 8.10. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Trong kiểm định giả thuyết H: "thời gian mang thai trung bình của thai phụ là 39,7 tuần", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là:

A. 6,72%. B. 7,72%. C. 8,72%. D. 9,72%.

Bài tập 8.11. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 − 36	36 − 38	38 − 40	40 − 42	42 − 44
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Giá trị thống kê t để kiểm định giả thuyết H: "tỷ lệ thai phụ sinh non là 12%" là:

A. t = 2,1037; tỷ lệ thai phụ sinh non thấp hơn 12% với mức ý nghĩa 5%.

B. t = 2,1037; tỷ lệ thai phụ sinh non lớn hơn 12% với mức ý nghĩa 5%.

C. t = 1,1037; tỷ lệ thai phụ sinh non cao hơn 12% với mức ý nghĩa 5%.

D. t = 1,1037; tỷ lệ thai phụ sinh non là 12% với mức ý nghĩa 5%.

Bài tập 8.12. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Trong kiểm định giả thuyết H: "tỷ lệ thai phụ sinh non là 12%", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là:

A. 3,48%. B. 4,48%. C. 5,48%. D. 6,48%.

Bài tập 8.13. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Khảo sát thời gian mang thai của 100 thai phụ có hút thuốc và tính được thời gian mang thai trung bình là 38,5 tuần và độ lệch chuẩn của mẫu có hiệu chỉnh 3,5 tuần. Giá trị thống kê t để kiểm định giả thuyết H: "Thời gian mang thai của thai phụ hút thuốc và không hút thuốc là như nhau" là:

A. t = 1,3798; Thời gian mang thai của thai phụ hút thuốc và không hút thuốc là như nhau với mức ý nghĩa 5%.

B. t = 1,3798; Thời gian mang thai của thai phụ hút thuốc nhỏ hơn với mức ý nghĩa 5%.

C. t = 2,3798; Thời gian mang thai của thai phụ hút thuốc lớn hơn với mức ý nghĩa 5%.

D. t = 2,3798; Thời gian mang thai của thai phụ hút thuốc nhỏ hơn với mức ý nghĩa 5%.

Bài tập 8.14. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Khảo sát thời gian mang thai của 100 thai phụ có hút thuốc và tính được thời gian mang thai trung bình là 38,5 tuần và độ lệch chuẩn của mẫu có hiệu chỉnh 3,5 tuần. Trong kiểm định giả thuyết H: "Thời gian mang thai của thai phụ hút thuốc và không hút thuốc là như nhau", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là

A. 2,74%. B. 3,74%. C. 1,74%. D. 4,74%.

Bài tập 8.15. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Khảo sát thời gian mang thai của 100 thai phụ có hút thuốc và tính được thời gian mang thai thấy có 16 thai phụ sinh non. Giá trị thống kê t để kiểm định giả thuyết H: "tỷ lệ sinh non của thai phụ có hút thuốc và không hút thuốc là như nhau" là:

A. t = 2,4753; tỷ lệ sinh non của thai phụ không hút thuốc lớn hơn với mức ý nghĩa 5%.

B. t = 2,4753; tỷ lệ sinh non của thai phụ có hút thuốc lớn hơn với mức ý nghĩa 5%.

C. t = 1,4753; tỷ lệ sinh non của thai phụ không hút thuốc lớn hơn với mức ý nghĩa 5%.

D. t = 1,4753; tỷ lệ sinh non của thai phụ có hút thuốc lớn hơn với mức ý nghĩa 5%.

Bài tập 8.16. Khảo sát thời gian (tuần) mang thai của thai phụ không hút thuốc. Tiến hành lấy mẫu, người ta có số liệu cho như bảng sau:

Thời gian	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44
Số thai phụ	7	10	59	41	4

Những thai phụ có thời gian mang thai dưới 36 tuần là thai phụ sinh non. Khảo sát thời gian mang thai của 100 thai phụ có hút thuốc và tính được thời gian mang thai thấy có 16 thai phụ sinh non. Trong kiểm định giả thuyết H: "tỷ lệ sinh non của thai phụ có hút thuốc và không hút thuốc là như nhau", mức ý nghĩa tối đa để giả thuyết H được chấp nhận là:

A. 1,32%. B. 2,32%. C. 3,32%. D. 4,32%.

Đáp án câu hỏi trắc nghiệm

8.3 A	**8.5** C	**8.7** D	**8.9** B	**8.11** A	**8.13** D	**8.15** B
8.4 B	**8.6** D	**8.8** C	**8.10** A	**8.12** A	**8.14** C	**8.16** A

Chương 9

Tương quan, hồi qui

9.1 Mở đầu

9.1.1 Số liệu trong phân tích tương quan, hồi qui

Quan trắc n đối tượng và ở mỗi đối tượng chúng ta "đo" 2 đại lượng X, Y. Số liệu cụ thể của n đối tượng cụ thể như sau:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Ví dụ 9.1. Khảo sát chiều cao $Y(cm)$ của 10 đứa trẻ tuổi X(tháng tuổi). Mỗi đứa trẻ ta ghi nhận một cặp $(X; Y)$ và các giá trị như sau:

$$\begin{array}{ccccc} (18; 76,0) & (19; 77,0) & (19; 76,3) & (20; 77,3) & (21; 77,7) \\ (22; 78,8) & (22; 78,2) & (23; 79,0) & (24; 80,2) & (25; 80,6) \end{array}$$

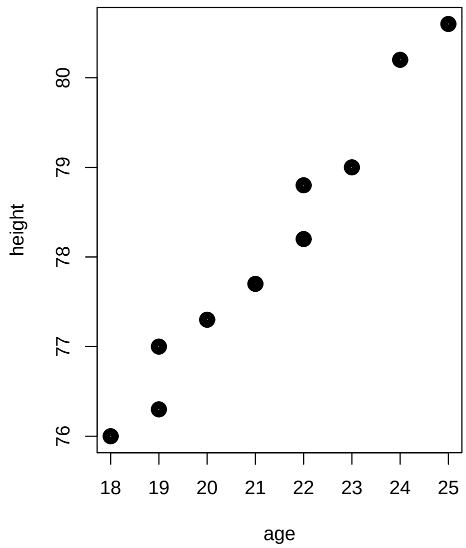
Thông thường các giá trị trên còn được xếp thành bảng như sau

X	18	19	19	20	21	22	22	23	24	25
Y	76,0	77,0	76,3	77,3	77,7	78,8	78,2	79,0	80,2	80,6

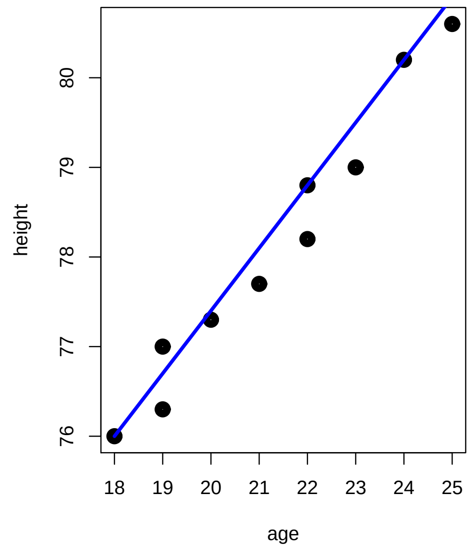
□

9.1.2 Biểu đồ tán xạ

Khi quan sát một đối tượng ta có cặp giá trị $(x_i; y_i)$. Để có được hình ảnh về sự phân tán của các cặp giá trị $(x_i; y_i)$ ta có thể biểu diễn các cặp giá trị này trên hệ trục $0xy$. Để minh họa, với số liệu ..... ta có biểu đồ tán xạ như sau



Hình a



Hình b

Ta nhận thấy hai đứa trẻ bất kỳ mặc dù cùng tuổi nhưng có chiều cao khác nhau (ngẫu nhiên) tuy nhiên xu hướng ở đây là chiều cao tăng theo độ tuổi (tất nhiên) hay chiều cao Y thay đổi một cách có hệ thống theo độ tuổi X.

Biểu đồ trên đây gợi ý cho thấy mối liên hệ giữa độ tuổi (X) và chiều cao (Y) là một đường thẳng (tuyến tính - như hình b). Để "đo lường" mối liên hệ này, chúng ta có thể sử dụng hệ số tương quan

9.2 Hệ số tương quan

Định nghĩa 9.1. *Giả sử ta có mẫu n quan trắc $(x_1, y_1), \ldots, (x_n, y_n)$. Hệ số tương quan* Pearson *được ước tính bằng công thức như sau*

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\hat{s}_x \hat{s}_y}$$

Trong đó $\overline{xy} = \frac{1}{n} \sum\limits_{i=1}^{n} x_i y_i$

Ý nghĩa hệ số tương quan

- r_{xy} đo mức độ quan hệ tuyến tính giữa x; y và $-1 \leq r_{xy} \leq 1$.

- $r_{xy} = 0$ hai biến số không có quan hệ tuyến tính, $r_{xy} = \pm 1$ thì hai biến số có quan hệ tuyến tính tuyệt đối (các cặp $(x_i; y_i)$ thuộc một đường thẳng).
- $r_{xy} < 0$ quan hệ giữa x, y là nghịch biến (có nghĩa là khi x tăng thì y giảm)
- $r_{xy} > 0$ quan hệ giữa x, y là đồng biến (có nghĩa là khi x tăng cao thì y tăng)

Ví dụ 9.2. Nghiên cứu đo lường độ cholesterol (Y) trong máu của 10 đối tượng nam của người độ tuổi (X). Kết quả đo lường như sau:

X	20	52	30	57	28	43	57	63	40	49
Y	1,9	4	2,6	4,5	2,9	3,8	4,1	4,6	3,2	4

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{451}{10} = 45,1; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{35,6}{10} = 3,56$$

$$\widehat{s}_x = 11,785; \quad \widehat{s}_y = 0,8333$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1695,4}{10} = 169,54$$

$$r_{xy} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\widehat{s}_x.\widehat{s}_y} = \frac{169,54 - 33,9 \cdot 3,56}{11,785 \cdot 0.8333} = 0,914$$

□

9.3 Tìm đường thẳng hồi qui

Để tiện việc theo dõi và mô tả mô hình, gọi độ tuổi cho cá nhân $i$là $x_i$và cholesterol là y_i ở đây $i = 1, 2 \ldots 10$. Mô hình hồi tuyến tính phát biểu rằng:

$$y_i = a + bx_i + \varepsilon_i$$

Nói cách khác, phương trình trên giả định rằng độ cholesterol của một cá nhân bằng một hằng số a cộng với một hệ số b liên quan đến độ tuổi, và một sai số ε_i. Trong phương trình trên, $a$là chặn (intercept, tức giá trị lúc x_i=0), và b là độ dốc (slope hay gradient).

Các thông số a, b phải được ước tính từ dữ liệu. Phương pháp để ước tính các thông số này là phương pháp bình phương nhỏ nhất (least squares method). Như tên gọi, phương pháp bình phương nhỏ nhất tìm giá trị a, b sao cho tổng bình phương sai số

$$\sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

là nhỏ nhất. Sau vài thao tác toán, có thể chứng minh dễ dàng rằng, ước lượng cho a, bđáp ứng điều kiện đó là

$$b = \frac{\overline{xy} - \bar{x}.\bar{y}}{\widehat{s}_x^{\,2}}; \quad a = \bar{y} - b\bar{x}$$

Cuối cùng ta được đường hồi qui $y = a + bx$

Chú ý: $\dfrac{y - \overline{y}}{\widehat{s}_y} = r_{xy}\dfrac{x - \overline{x}}{\widehat{s}_x}$

Ví dụ 9.3. xác định phương trình hồi qui mẫu giữa tuổi và cholesterol. Từ

$$\frac{y - \overline{y}}{\widehat{s}_y} = r_{xy}\frac{x - \overline{x}}{\widehat{s}_x}$$

thay các giá trị $\bar{y}, \bar{x}, \widehat{s}_x, \widehat{s}_y, r_{xy}$ được tính ở ví dụ trên vào ta có kết quả

$$y = 0,9311 + 0,05988x$$

□

9.4 Sử dụng máy tính cầm tay

Ví dụ 9.4. Bài toán cho dạng cặp (x_i, y_i) như sau:

X	20	52	30	57	28	43	57	63	40	49
Y	1,9	4	2,6	4,5	2,9	3,8	4,1	4,6	3,2	4

Tìm hệ số tương quan r_{xy}, đường hồi qui mẫu $y = a + bx$.

a. Máy FX500MS (máy FX570MS tương tự)

- Bước 1: Nhấn phím Mod đến lúc màn hình xuất hiện **REG**; chọn (REG); Chọn (Lin)
- Bước 2: Nhập liệu 20; ,; 1.9; M+ $\cdots$
- Bước 3: Xuất kết quả Shift; chọn (S-Var); chọn (mũi tên phải 2 lần); 1(A =a); 2(B=b); 3(r=r_{xy})

b. Máy FX500ES(tương tự FX570ES)

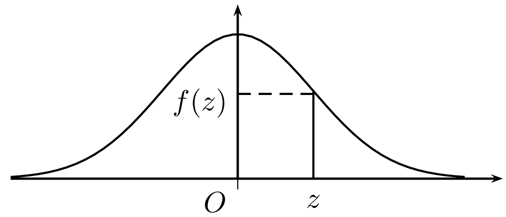
- Bước 1: SHIFT; MODE; $\downarrow$; chọn (Stat); chọn (Off)
- Bước 2: MODE; chọn (stat); chọn (A+Bx); (nhập các giá trị của X, Y vào 2 cột)
  - $*$ Nhập giá trị của X 20= 52= $\cdots$
  - $*$ Nhập giá trị của Y 1.9= 4= $\cdots$
- Bước 3: Xuất kết quả SHIFT; chọn phím (Stat); chọn (Reg); 1(A =a); 2(B=b); 3(r=r_{xy}).

Kết quả $r_{xy} = 0,9729; y = 0,9311 + 0,0599x$.

Phụ lục A

Các bảng giá trị xác suất

A.1 Giá trị hàm mật độ chuẩn đơn giản $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

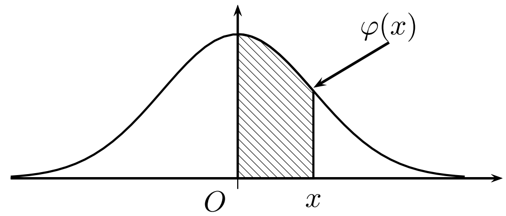


z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3970
0,1	0,3970	0,3965	0,3961	0,3956	0,3951	0,3945	0,3939	0,3932	0,3925	0,3911
0,2	0,3910	0,3902	0,3894	0,3885	0,3876	0,3867	0,3857	0,3847	0,3836	0,3815
0,3	0,3814	0,3802	0,3790	0,3778	0,3765	0,3752	0,3739	0,3725	0,3712	0,3684
0,4	0,3683	0,3668	0,3653	0,3637	0,3621	0,3605	0,3589	0,3572	0,3555	0,3522
0,5	0,3521	0,3503	0,3485	0,3467	0,3448	0,3429	0,3410	0,3391	0,3372	0,3334
0,6	0,3332	0,3312	0,3292	0,3271	0,3251	0,3230	0,3209	0,3187	0,3166	0,3125
0,7	0,3123	0,3101	0,3079	0,3056	0,3034	0,3011	0,2989	0,2966	0,2943	0,2899
0,8	0,2897	0,2874	0,2850	0,2827	0,2803	0,2780	0,2756	0,2732	0,2709	0,2663
0,9	0,2661	0,2637	0,2613	0,2589	0,2565	0,2541	0,2516	0,2492	0,2468	0,2422
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2181
1,1	0,2179	0,2155	0,2131	0,2107	0,2083	0,2059	0,2036	0,2012	0,1989	0,1944
1,2	0,1942	0,1919	0,1895	0,1872	0,1849	0,1826	0,1804	0,1781	0,1758	0,1716
1,3	0,1714	0,1691	0,1669	0,1647	0,1626	0,1604	0,1582	0,1561	0,1539	0,1499
1,4	0,1497	0,1476	0,1456	0,1435	0,1415	0,1394	0,1374	0,1354	0,1334	0,1297
1,5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1111
1,6	0,1109	0,1092	0,1074	0,1057	0,1040	0,1023	0,1006	0,0989	0,0973	0,0942
1,7	0,0940	0,0925	0,0909	0,0893	0,0878	0,0863	0,0848	0,0833	0,0818	0,0791
1,8	0,0790	0,0775	0,0761	0,0748	0,0734	0,0721	0,0707	0,0694	0,0681	0,0657
1,9	0,0656	0,0644	0,0632	0,0620	0,0608	0,0596	0,0584	0,0573	0,0562	0,0541
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0441
2,1	0,0440	0,0431	0,0422	0,0413	0,0404	0,0396	0,0387	0,0379	0,0371	0,0356
2,2	0,0355	0,0347	0,0339	0,0332	0,0325	0,0317	0,0310	0,0303	0,0297	0,0284
2,3	0,0283	0,0277	0,0270	0,0264	0,0258	0,0252	0,0246	0,0241	0,0235	0,0224
2,4	0,0224	0,0219	0,0213	0,0208	0,0203	0,0198	0,0194	0,0189	0,0184	0,0176
2,5	0,0175	0,0171	0,0167	0,0163	0,0158	0,0154	0,0151	0,0147	0,0143	0,0136
2,6	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110	0,0104
2,7	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084	0,0079
2,8	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063	0,0060
2,9	0,0060	0,0058	0,0056	0,0055	0,0053	0,0051	0,0050	0,0048	0,0047	0,0044
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0033

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
3,1	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	0,0025	0,0024
3,2	0,0024	0,0023	0,0022	0,0022	0,0021	0,0020	0,0020	0,0019	0,0018	0,0017
3,3	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013	0,0012
3,4	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010	0,0010	0,0009	0,0009
3,5	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007	0,0006
3,6	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0004
3,7	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003
3,8	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002
3,9	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001

Bảng A.1: Giá trị hàm mật độ chuẩn hóa

A.2 Giá trị hàm $\varphi(x) = \int_0^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)dz$

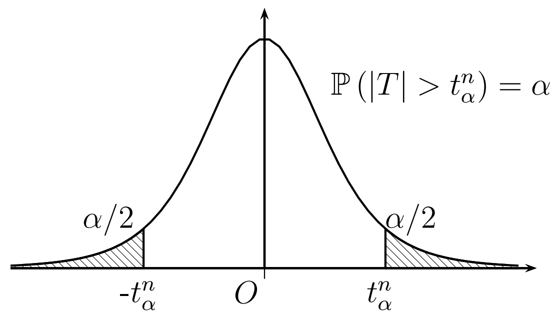


x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	**0,475**	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Bảng A.2: Giá trị hàm φ của phân phối chuẩn đơn giản

A.3 Giá trị phân vị của luật Student ($T \sim T_n$)



n \ α	0,14	0,13	0,12	0,11	0,10	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01
1	4,474	4,829	5,242	5,730	6,314	7,026	7,916	9,058	10,579	12,706	15,895	21,205	31,821	63,657
2	2,383	2,495	2,620	2,760	2,920	3,104	3,320	3,578	3,896	4,303	4,849	5,643	6,965	9,925
3	1,995	2,072	2,156	2,249	2,353	2,471	2,605	2,763	2,951	3,182	3,482	3,896	4,541	5,841
4	1,838	1,902	1,971	2,048	2,132	2,226	2,333	2,456	2,601	2,776	2,999	3,298	3,747	4,604
5	1,753	1,810	1,873	1,941	2,015	2,098	2,191	2,297	2,422	2,571	2,757	3,003	3,365	4,032
6	1,700	1,754	1,812	1,874	1,943	2,019	2,104	2,201	2,313	2,447	2,612	2,829	3,143	3,707
7	1,664	1,715	1,770	1,830	1,895	1,966	2,046	2,136	2,241	2,365	2,517	2,715	2,998	3,499
8	1,638	1,687	1,740	1,797	1,860	1,928	2,004	2,090	2,189	2,306	2,449	2,634	2,896	3,355
9	1,619	1,666	1,718	1,773	1,833	1,899	1,973	2,055	2,150	2,262	2,398	2,574	2,821	3,250
10	1,603	1,650	1,700	1,754	1,812	1,877	1,948	2,028	2,120	2,228	2,359	2,527	2,764	3,169
11	1,591	1,636	1,686	1,738	1,796	1,859	1,928	2,007	2,096	2,201	2,328	2,491	2,718	3,106
12	1,580	1,626	1,674	1,726	1,782	1,844	1,912	1,989	2,076	2,179	2,303	2,461	2,681	3,055
13	1,572	1,616	1,664	1,715	1,771	1,832	1,899	1,974	2,060	2,160	2,282	2,436	2,650	3,012
14	1,565	1,609	1,656	1,706	1,761	1,821	1,887	1,962	2,046	2,145	2,264	2,415	2,624	2,977
15	1,558	1,602	1,649	1,699	1,753	1,812	1,878	1,951	2,034	2,131	2,249	2,397	2,602	2,947

Bảng A.3: Giá trị phân vị của luật Student (tiếp theo)

n \ α	0,14	0,13	0,12	0,11	0,10	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01
16	1,553	1,596	1,642	1,692	1,746	1,805	1,869	1,942	2,024	2,120	2,235	2,382	2,583	2,921
17	1,548	1,591	1,637	1,686	1,740	1,798	1,862	1,934	2,015	2,110	2,224	2,368	2,567	2,898
18	1,544	1,587	1,632	1,681	1,734	1,792	1,855	1,926	2,007	2,101	2,214	2,356	2,552	2,878
19	1,540	1,583	1,628	1,677	1,729	1,786	1,850	1,920	2,000	2,093	2,205	2,346	2,539	2,861
20	1,537	1,579	1,624	1,672	1,725	1,782	1,844	1,914	1,994	2,086	2,197	2,336	2,528	2,845
21	1,534	1,576	1,621	1,669	1,721	1,777	1,840	1,909	1,988	2,080	2,189	2,328	2,518	2,831
22	1,531	1,573	1,618	1,665	1,717	1,773	1,835	1,905	1,983	2,074	2,183	2,320	2,508	2,819
23	1,529	1,570	1,615	1,662	1,714	1,770	1,832	1,900	1,978	2,069	2,177	2,313	2,500	2,807
24	1,526	1,568	1,612	1,660	1,711	1,767	1,828	1,896	1,974	2,064	2,172	2,307	2,492	2,797
25	1,524	1,566	1,610	1,657	1,708	1,764	1,825	1,893	1,970	2,060	2,167	2,301	2,485	2,787
26	1,522	1,564	1,608	1,655	1,706	1,761	1,822	1,890	1,967	2,056	2,162	2,296	2,479	2,779
27	1,521	1,562	1,606	1,653	1,703	1,758	1,819	1,887	1,963	2,052	2,158	2,291	2,473	2,771
28	1,519	1,560	1,604	1,651	1,701	1,756	1,817	1,884	1,960	2,048	2,154	2,286	2,467	2,763
29	1,517	1,558	1,602	1,649	1,699	1,754	1,814	1,881	1,957	2,045	2,150	2,282	2,462	2,756
30	1,516	1,557	1,600	1,647	1,697	1,752	1,812	1,879	1,955	2,042	2,147	2,278	2,457	2,750
40	1,506	1,546	1,589	1,635	1,684	1,737	1,796	1,862	1,936	2,021	2,123	2,250	2,423	2,704
60	1,496	1,535	1,577	1,622	1,671	1,723	1,781	1,845	1,917	2,000	2,099	2,223	2,390	2,660
80	1,491	1,530	1,572	1,616	1,664	1,716	1,773	1,836	1,908	1,990	2,088	2,209	2,374	2,639
100	1,488	1,527	1,568	1,613	1,660	1,712	1,769	1,832	1,902	1,984	2,081	2,201	2,364	2,626
1000	1,477	1,515	1,556	1,600	1,646	1,697	1,752	1,814	1,883	1,962	2,056	2,173	2,330	2,581

Bảng A.3: Giá trị phân vị của luật Student

Phụ lục B

Giải thích lý thuyết

B.1 Ước lượng khoảng

B.1.1 Ước lượng khoảng cho trung bình

Trường hợp $X \sim X(\mu; \sigma^2)$, biết σ

Từ 6.1 trang 99 ta có

$$\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right) \text{ suy ra } T = \frac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0; 1)$$

Gọi $t_{\frac{1-\alpha}{2}}$ là giá trị của T sao cho

$$\mathbb{P}\left(t_{\frac{1-\alpha}{2}} < T < t_{\frac{1-\alpha}{2}}\right) = 1 - \alpha$$

Thay T vào ta được

$$\mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}}\right) = 1 - \alpha$$

Vậy ta có $\mu_1 = \bar{X} - \dfrac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}}$ và $\mu_2 = \bar{X} + \dfrac{\sigma}{\sqrt{n}} t_{\frac{1-\alpha}{2}}$

Các trường hợp còn lại giải tương tự.

B.1.2 Ước lượng khoảng cho tỷ lệ

Từ 6.5 trang 100 ta có

$$X = X_1 + \ldots + X_n \dot{\sim} N\left(np; \sqrt{np(1-p)}^2\right) \text{ hay } \frac{X - np}{\sqrt{np(1-p)}} \dot{\sim} N(0;1) \tag{B.1}$$

Bởi vì $F = X/n$ là ước lượng điểm cho p cho nên $\sqrt{n(X/n)(1-X/n)}$ sẽ xấp xỉ cho $\sqrt{np(1-p)}$, cho nên B.1 trở thành

$$T = \frac{X - np}{\sqrt{n(X/n)(1-X/n)}} \dot{\sim} N(0;1)$$

Gọi $t_{\frac{1-\alpha}{2}}$ là giá trị của T sao cho

$$\mathbb{P}\left(t_{\frac{1-\alpha}{2}} < T < t_{\frac{1-\alpha}{2}}\right) = 1 - \alpha$$

Thay T vào ta được

$$\mathbb{P}\left(X/n - \sqrt{\frac{X/n(1-X/n)}{n}} < t_{\frac{1-\alpha}{2}} < p < X/n + \sqrt{\frac{X/n(1-X/n)}{n}} < t_{\frac{1-\alpha}{2}}\right) = 1$$

Chú ý. Khi có mẫu cụ thể ta thay $F = X/n$ bằng giá trị f, là tỷ lệ phần tử A trên mẫu.

B.2 Kiểm định giả thiết

B.2.1 So sánh trung bình với một số

Gọi μ là trung bình của X, cần kiểm định giả thiết:

$$\begin{cases} \text{Giả thiết không } H_0: & \mu = \mu_0 \\ \text{Đối thiết } H_1: & \mu = \mu_1 \end{cases}$$

Bởi vì $\bar{X}$ là ước lượng điểm cho μ, do đó ta sẽ chấp nhận giả thiết nếu $\bar{X}$ và μ_0 không quá khác nhau. Do đó miền bác bỏ sẽ có dạng

$$C = \left\{(X_1, \ldots, X_n) : |\bar{X} - \mu_0| > c\right\} \tag{B.2}$$

với c là một giá trị nào đó.

Nếu cho trước mức ý nghĩa α, chúng ta sẽ xác định giá trị tới hạn c trong (B.2) sao cho sai lầm loại I bằng với α. Do đó, c phải thoải

$$\mathbb{P}\left(|\bar{X}-\mu_0|>c|H_0\right)=\alpha \text{ hay } \mathbb{P}\left(|\bar{X}-\mu_0|>c|\mu=\mu_0\right)=\alpha \tag{B.3}$$

Ở đây chỉ xét trường hợp à $X\sim N(\mu;\sigma^2)$ và đã biết σ. Khi $\mu=\mu_0$ thì theo (6.1) trang 99 ta có

$$T=\frac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}=\frac{\bar{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}\sim N(0;1)$$

Bây giờ (B.3) trở thành

$$\mathbb{P}\left(|T|>\frac{c\sqrt{n}}{\sigma}\right)=\alpha$$

Ta biết rằng $T\sim N(0;1)$ thì $\mathbb{P}\left(|T|>t_{\frac{1-\alpha}{2}}\right)=\alpha$. Cho nên ta chọn $\dfrac{c\sqrt{n}}{\sigma}=t_{\frac{1-\alpha}{2}}$. Vậy ta bác bỏ H_0 khi

$$T=\frac{|\bar{X}-\mu_0|}{\dfrac{\sigma}{\sqrt{n}}}>t_{\frac{1-\alpha}{2}}$$

B.2.2 So sánh tỷ lệ với một số

Giống như B.2.1, ở đây ta xem thống kê

$$X=X_1+\ldots+X_n \mathrel{\dot\sim} N\left(np;\sqrt{np(1-p)}^2\right) \text{ hay } T=\frac{X-np}{\sqrt{np(1-p)}}\mathrel{\dot\sim} N\left(0;1\right)$$

Tài liệu tham khảo

[1] Nguyễn Phú Vinh. *Xác Suất - Thống Kê Và Ứng Dụng*

[2] Đinh Văn Gắng. (1999). *Lý thuyết xác suất và thống kê toán*. NXB Giáo dục.

[3] Tô Anh Dũng. (2007). *Lý thuyết xác suất và thống kê toán*. NXB ĐHQG TP.HCM.

[4] Nguyễn Bác Văn. (1999). *Xác suất và xử lý số liệu thống kê*. NXB Giáo dục.

[5] Đặng Hấn. (1986). *Xác suất thống kê*. NXB Thống kê.

[6] Sheldon M. Ross. (1987). *Introduction to probability and statistics for engineers and scientists*. A John Wiley & Sons Publication.

[7] F.M. Dekking. (2005). *A modern introduction to Probability and Statistics*. Springer Publication.

[8] T.T. Song. (2004). *Fundamentals of probability and statistics for engineers*. A John Wiley & Sons Publication.

[9] Ronald N. Forthofer. (2007). *Biostatistics: Aguide to design, analysis, and discovery*. Academic Press.

[10] Y. Suhov. (2005). *Volume I: Basic probability and statistics*. Cambridge University Press.

[11] Michaelr. Chernick. (2003). *Introductory biostatistics for the health sciences*. A John Wiley & Sons Publication.

[12] E.L. Lehmann. (2005). *Testing statistical hypotheses: Third Edition*. Springer Publication.