Chương 5:

ƯỚC LƯỢNG THAM SỐ THỐNG KÊ

I. Mẫu thống kê

Ký hiệu X là đặc tính cần nghiên cứu trên các phần tử của tập hợp M.

M gọi là *tổng thể* (population), số phần tử của M ký hiệu là N.

Thông thường không thể lấy hết các phần tử của M để quan sát X vì những lý do sau

- Số N quá lớn.
- Thời gian và kinh phí không cho phép.
- Có thể làm hư hại hết các phần tử của M.

Vì vậy người ta thường lấy một số phần tử của M để quan sát X, các phần tử này gọi là *mẫu* (sample) lấy từ M.

Số phần tử của mẫu gọi là *cỡ mẫu* (sample size), ký hiệu là n.

Điều kiện để chọn mẫu

- Các phần tử của mẫu lấy ngẫu nhiên từ M.
- Các phần tử của M có đồng khả năng được chọn làm mẫu.
- Các phần tử của mẫu được lấy một cách độc lập với nhau.

Phân phối lý thuyết của X

X	x_1	x_2	...	x_N
P	$\frac{1}{N}$	$\frac{1}{N}$	...	$\frac{1}{N}$

Trong đó x_i là giá trị quan sát X trên phần tử thứ i của $\boldsymbol{M}$.

- <u>Giai đoạn lý thuyết</u> (Chưa lấy mẫu cụ thể)

Ký hiệu Xj là giá trị quan sát X trên phần tử thứ j của mẫu. Khi đó ta có một bộ n biến ngẫu nhiên $(X_1, ..., X_n)$ gọi là <u>mẫu lý thuyết</u> lấy từ $\boldsymbol{M}$.

<u>Tính chất mẫu lý thuyết</u>

1) Các Xj có cùng phân phối như X.
2) Các Xj độc lập với nhau.

Giai đoạn thực nghiệm

(Đã lấy mẫu cụ thể)

Khi đã lấy mẫu cụ thể xong ta có các số liệu $(x_1, ..., x_n)$ và gọi là mẫu thực nghiệm lấy từ X.

Phương pháp lấy mẫu ngẫu nhiên đơn giản

- Đánh số các phần tử của $\boldsymbol{M}$ từ 1 đến N.
- Lập các phiếu cũng đánh số từ 1 đến N.
- Trộn đều các phiếu, sau đó lấy lần lượt có hoàn lại n phiếu.

Các phần tử của $\boldsymbol{M}$ có số thứ tự trong các phiếu lấy ra sẽ được chọn làm mẫu.

II. Các đặc trưng mẫu

X là biến ngẫu nhiên cần nghiên cứu.

Ký hiệu $EX = \mu$ (trung bình tổng thể)

$DX = \sigma^2$ (phương sai tổng thể)

1. Trung bình mẫu

Giai đoạn lý thuyết

Xét mẫu lý thuyết $(X_1, ..., X_n)$. Trung bình mẫu được xác định như sau.

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

và $\overline{X}$ là một biến ngẫu nhiên.

Ta tính kỳ vọng, phương sai và độ lệch chuẩn của $\overline{X}$:

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{n\mu}{n} = \mu$$

$$D\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\sqrt{D\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Giai đoạn thực nghiệm

Xét mẫu thực nghiệm : $(x_1 , ... , x_n)$

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

a) Mẫu có lặp

X	$x_1 \quad . . . \quad x_k$	Tổng
n_i	$n_1 \quad . . . \quad n_k$	n

Trong đó, n_i là tần số giá trị x_i trong mẫu và $n_1 + ... + n_k = n$.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$

b) Mẫu chia khoảng

X	$(a_1, a_2]$	. . .	$(a_k, a_{k+1}]$	Tổng
n_i	n_1	. . .	n_k	n

Trong đó n_i là tần số giá trị trong mẫu rơi vào $(a_i, a_{i+1}]$ và $n_1 + ... + n_k = n$.

$$\theta_i = \frac{a_i + a_{i+1}}{2}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i \theta_i$$

2. Phương sai mẫu

Giai đoạn lý thuyết

Xét mẫu lý thuyết $(X_1,..., X_n)$.

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

$$Es^2 = \frac{n-1}{n}\sigma^2$$

$$s^2 = \overline{X^2} - (\overline{X})^2$$

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

$s = \sqrt{s^2}$ - độ lệch chuẩn mẫu

Phương sai mẫu điều chỉnh

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

$$S^2 = \frac{n}{n-1}s^2$$

$$ES^2 = \sigma^2$$

$S = \sqrt{S^2}$ - độ lệch chuẩn mẫu điều chỉnh

Giai đoạn thực nghiệm

Xét mẫu thực nghiệm $(x_1, \ldots, x_n)$

a) Mẫu có lặp

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{X})^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \overline{X})^2$$

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2$$

b) Mẫu chia khoảng

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\theta_i - \overline{X})^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(\theta_i - \overline{X})^2$$

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{k} n_i\theta_i^2$$

3. Tỷ lệ mẫu

Giả sử tham số *p* là tỷ lệ các phần tử loại L trên tổng thể M. Xét mẫu (X_1 , ..., X_n), với

X_i = 1 nếu phần tử thứ i của mẫu thuộc loại L,

X_i = 0 nếu ngược lại.

Gọi m số phần tử loại L trên mẫu, khi đó

$$m = X_1 + \ldots + X_n$$

$$f = \frac{m}{n}$$

và được gọi là tỷ lệ mẫu (tần suất) của các phần tử loại L trên mẫu. Và f cũng là biến ngẫu nhiên.

$$Ef = E\frac{m}{n} = \frac{1}{n}(EX_1 + \ldots + EX_n) = \frac{np}{n} = p$$

$$Df = D\frac{m}{n} = \frac{1}{n^2}(DX_1 + \ldots + DX_n) = \frac{npq}{n^2} = \frac{pq}{n}$$

Độ lệch chuẩn của tỷ lệ mẫu f :

$$\sqrt{Df} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

III. Ước lượng điểm

Giả sử θ là tham số chưa biết của biến ngẫu nhiên X. Dựa vào mẫu $(X_1, ..., X_n)$ cần tìm đại lượng $\hat{\theta}(X_1, \ldots, X_n)$ làm xấp xỉ cho θ, gọi là ước lượng điểm của θ.

1. Ước lượng không chệch

$\hat{\theta}(X_1, \ldots, X_n)$ được gọi là ước lượng không chệch của θ nếu $E\hat{\theta}(X_1, \ldots, X_n) = \theta$.

Khi đó trung bình sai số của ước lượng bằng

$$E[\hat{\theta}(X_1, \ldots, X_n) - \theta] = 0$$

2. Các phương pháp tìm ước lượng điểm

Hợp lý cực đại, Bình phương nhỏ nhất …

Thí dụ : Các tham số của biến X là

$$\mu = EX \text{ và } \sigma^2 = DX$$

- $\overline{X}$ là ước lượng không chệch của μ .
- s^2 là ước lượng chệch của σ^2.
- S^2 là ước lượng không chệch của σ^2.
- f - ước lượng không chệch của tỷ lệ p.

IV. Ước lượng tham số bằng khoảng tin cậy

1. Khái niệm chung

Giả sử θ là tham số chưa biết của biến ngẫu nhiên X. Dựa vào mẫu $(X_1, ..., X_n)$ cần tìm hai đại lượng $\theta_1(X_1, ..., X_n)$, $\theta_2(X_1, ..., X_n)$ sao cho

$$P(\theta_1 \le \theta \le \theta_2) = \gamma \qquad (*)$$

với γ đủ lớn cho trước, thường $\gamma = 95\%$ hay 99%.

Xác suất γ gọi là ***độ tin cậy*** của ước lượng khoảng.

Khoảng $[\theta_1 , \theta_2]$ gọi là ***khoảng tin cậy*** cho θ.

Ý nghĩa của (*) :

Có γ 100% số lần lấy mẫu cỡ n thì

$$\theta \in [\theta_1 , \theta_2]$$

Có $(1-\gamma)$100% số lần lấy mẫu cỡ n thì

$$\theta \notin [\theta_1 , \theta_2]$$

2. Khoảng tin cậy cho trung bình

Giả sử tham số là $\mu = EX$ chưa biết của biến ngẫu nhiên X và $\sigma^2 = DX$. Dựa vào mẫu $(X_1, ..., X_n)$ cần tìm hai đại lượng

$$\mu_1(X_1, ..., X_n), \mu_2(X_1, ..., X_n)$$

sao cho

$$P(\mu_1 \le \mu \le \mu_2) = \gamma$$

1) Khi $n \geq 30,\ \sigma^2$ đã biết.

Xét thống kê

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

theo định lý giới hạn trung tâm.

Dựa vào luật phân phối đã biết của Z ta tìm z sao cho

$$P(|Z| \leq z) = \gamma$$

Từ đó ta có

$$\mu_{1,2} = \overline{X} \pm z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}$$

Trong đó là $z_{\frac{1+\gamma}{2}}$ phân vị mức $\frac{1+\gamma}{2}$ của

$\Phi(x)$, tức là

$$\Phi\left(z_{\frac{1+\gamma}{2}}\right) = \frac{1+\gamma}{2}$$

2) Khi $n \geq 30$, σ^2 không biết

$$\mu_{1,2} = \overline{X} \pm z_{\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}$$

3) Khi $n < 30$, σ^2 đã biết và $X \sim N(\mu, \sigma^2)$

$$\mu_{1,2} = \overline{X} \pm z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}$$

4) Khi $n < 30, \sigma^2$ không biết và $X \sim N(\mu, \sigma^2)$

Xét thống kê

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

Dựa vào luật phân phối Student với

$(n\text{ -}1)$ bậc tự do của T ta có

$$\mu_{1,2} = \overline{X} \pm t^{n-1}_{\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}$$

Trong đó $t^{n-1}_{\frac{1+\gamma}{2}}$ là phân vị mức $\frac{1+\gamma}{2}$ của luật phân phối Student với $(n-1)$ bậc tự do. (Bảng ở trang 229)

$$P\left(T \le t_q^v\right) = q$$

$$P\left(T \le t^{n-1}_{\frac{1+\gamma}{2}}\right) = \frac{1+\gamma}{2}$$

Thí dụ 1: Điều tra đường huyết **100** người, có kết quả sau:

Trung bình mẫu: **100mg%**

Độ lệch chuẩn mẫu điều chỉnh: **10,08mg%**

Tìm KTC **95%** cho đường huyết trung bình của dân số.

Giải: Gọi *X* - đường huyết của dân số.

$EX=\mu$ - đường huyết trung bình của dân số

$DX=\sigma^2$ - phương sai đường huyết của dân số

Cần tìm KTC cho μ..

- Cỡ mẫu n=100 >30, σ^2 không biết: TH2

γ=0,95 ; $z_{\frac{1+\gamma}{2}} = z_{0,975} = 1,96$

$$\mu_{1,2} = \overline{X} \pm z_{\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}$$

$$= 100 \pm 1,96 \frac{10,08}{\sqrt{100}} = 100 \pm 1,98$$

KTC: [98,02 ; 101,98] mg%

Thí dụ 2: Đo urê máu của **10** người, tính được:
Trung bình mẫu: **30,7cg/l**
Phương sai mẫu điều chỉnh: **103,3**
Tìm KTC **95%** cho urê máu trung bình của dân số. Biết rằng urê máu của dân số có phân phối chuẩn.

Giải: Gọi *X* - urê máu của dân số.

$EX=\mu$ - urê máu trung bình của dân số

$DX=\sigma^2$ - phương sai urê máu của dân số

Cần tìm KTC cho μ..

- Cỡ mẫu n=10 < 30, σ^2 không biết,

$X \sim N(\mu, \sigma^2)$: TH4

γ=0,95 ; $t^{n-1}_{\frac{1+\gamma}{2}} = t^{9}_{0,975} = 2,262$

$$\mu_{1,2} = \overline{X} \pm t^{n-1}_{\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}$$

$$= 30,7 \pm 2,262 \frac{\sqrt{103,3}}{\sqrt{10}} = 30,7 \pm 7,2$$

KTC: [23,5 ; 37,9] cg/l

3. Khoảng tin cậy cho tỷ lệ

Giả sử tham số p là tỷ lệ các phần tử loại L trên tổng thể M.

Xét mẫu $(X_1, ..., X_n)$ với

$X_i = 1$ nếu phần tử thứ i của mẫu thuộc loại L

$X_i = 0$ nếu ngược lại.

Cần tìm hai đại lượng

$$p_1(X_1, ..., X_n),\ p_2(X_1, ..., X_n)$$

sao cho

$$P(p_1 \le p \le p_2) = \gamma$$

- Xét mẫu cỡ lớn : $nf \geq 10$, $n(1-f) \geq 10$ và thống kê

$$Z = \frac{f - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Trong đó f là tỷ lệ mẫu.

Từ đó

$$p_{1,2} = f \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{f(1-f)}{n}}$$

Thí dụ 3: Kiểm tra **100** sản phẩm ở một kho hàng thấy có **10** phế phẩm.

Ước lượng tỷ lệ phế phẩm của kho hàng với ĐTC **95%, 99%.**

Giải: Gọi p - tỷ lệ phế phẩm của kho hàng.

Cần tìm KTC cho p .

Tỷ lệ mẫu:

$$f = 10/100=0,1$$

$$nf \geq 10,\ n(1\text{-}f\,) \geq 10$$

- Xét γ =0,95 ; $z_{\frac{1+\gamma}{2}} = z_{0,975} = 1,96$

$$p_{1,2} = f \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{f(1-f)}{n}}$$

$$= 0,1 \pm 1,96 \sqrt{\frac{0,1(1-0,1)}{100}}$$

$$= 0,1 \pm 0,059$$

KTC: [0,04 ; 0,16]

- Xét γ =0,99 ; $z_{\frac{1+\gamma}{2}} = z_{0,995} = 2,58$

$$p_{1,2} = f \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{f(1-f)}{n}}$$

$$= 0,1 \pm 2,58 \sqrt{\frac{0,1(1-0,1)}{100}}$$

$$= 0,1 \pm 0,077$$

KTC: [0,02 ; 0,18]

Nhận xét:

- ĐTC càng tăng thì KTC càng rộng
- Cỡ mẫu càng tăng thì KTC càng hẹp

4. Độ chính xác của ước lượng và xác định cỡ mẫu

1) Trường hợp kỳ vọng

Độ chính xác (sai số) của ước lượng $\overline{X}$ cho tham số $\mu = EX$ với độ tin cậy γ là số $\varepsilon > 0$ sao cho

$$P(|\overline{X} - \mu| \le \varepsilon) = \gamma$$

Từ đó

$$P(-\varepsilon \le \overline{X} - \mu \le \varepsilon) = \gamma$$

$$P(\overline{X} - \varepsilon \le \mu \le \overline{X} + \varepsilon) = \gamma$$

So sánh với KTC cho kỳ vọng ta có

$$\varepsilon = z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}$$, nếu σ^2 đã biết

$$\varepsilon = z_{\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}$$, nếu σ^2 không biết

- **Cho ε và γ tìm cỡ mẫu n**

$$n \ge \left[z_{\frac{1+\gamma}{2}} \frac{\sigma}{\varepsilon} \right]^2$$, nếu σ^2 đã biết.

$$n \ge \left[z_{\frac{1+\gamma}{2}} \frac{S}{\varepsilon} \right]^2$$, nếu σ^2 không biết.

2) Trường hợp tỷ lệ

Độ chính xác (sai số) của ước lượng f cho tham số p với ĐTC γ là số $\varepsilon > 0$ sao cho

$$P(|f - p| \le \varepsilon) = \gamma$$

Tương tự trên, so sánh với KTC cho tỷ lệ ta có

$$\varepsilon = z_{\frac{1+\gamma}{2}} \sqrt{\frac{f(1-f)}{n}}$$

- **Cho ε và γ tìm cỡ mẫu n**

$$n \geq \left[z_{\frac{1+\gamma}{2}} \frac{\sqrt{f(1-f)}}{\varepsilon} \right]^2$$

Thí dụ 4: Điều tra ban đầu trên **30** người thấy có **5** người bị bệnh B. Muốn tìm KTC 95% của tỷ lệ bệnh B trong dân số thì phải lấy mẫu cỡ bao nhiêu, nếu chấp nhận sai số là **0,05**.

Giải: Tỷ lệ mẫu $f = 5/30$.

γ =0,95 ; $z_{\frac{1+\gamma}{2}} = z_{0,975} = 1,96$

$$n \ge \left[z_{\frac{1+\gamma}{2}} \frac{\sqrt{f(1-f)}}{\varepsilon} \right]^2 =$$

$$= 1,96^2 \frac{\frac{5}{30}\left[1-\frac{5}{30}\right]}{0,05^2}$$

$$= 213,42$$

n=214

Thí dụ 5: Đo đường huyết trên **10** người tính được phương sai mẫu điều chỉnh là **0,04**. Muốn cho chênh lệch tuyệt đối giữa TB mẫu và trung bình dân số là **0,02g/l** với ĐTC **95%** thì phải lấy mẫu cỡ bao nhiêu ?

Giải: Ta có chênh lệch

$$\left|\overline{X} - \mu\right| = 0{,}02 = \varepsilon$$

γ =0,95 ; $z_{\frac{1+\gamma}{2}} = z_{0,975} = 1{,}96$

$$n \geq \left[z_{\frac{1+\gamma}{2}} \frac{S}{\varepsilon} \right]^2$$

$$= 1.96^2 \frac{0,04}{0,02^2} = 384,16$$

$n = 385$