Trường Đại Học Khoa Học Tự Nhiên TP.HCM

XÁC SUẤT THỐNG KÊ

Lưu hành nội bộ
2020

Giả sử biến ngẫu nhiên X có tham số θ chưa biết. Ước lượng tham số θ là dựa vào mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$ ta đưa ra thống kê $\hat{\theta}$ để ước lượng (dự đoán) θ.

Ví dụ:
∘ **Ước lượng điểm**: chỉ ra $\hat{\theta} = \theta_0$ nào đó để ước lượng θ.
∘ **Ước lượng khoảng**: chỉ ra một khoảng $(\hat{\theta}_L, \hat{\theta}_U)$ chứa θ sao cho $\mathbb{P}(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$ cho trước, ($1 - \alpha$ đgl độ tin cậy của ước lượng).

Định nghĩa

Một ước lượng điểm cho tham số tổng thể θ là một giá trị đơn $\hat{\theta}$ của một thống kê $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$.

Nhận xét

Thông thường giá trị được chọn này là giá trị cụ thể của một thống kê $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$ nào đó của mẫu ngẫu nhiên.

Ví dụ: Giá trị $\bar{x}$ của thống kê $\bar{X}$ được tính toán từ một mẫu ngẫu nhiên kích thước n là một ước lượng điểm của tham số trung bình tổng thể μ .

Cùng một mẫu ngẫu nhiên ta có thể xây dựng được nhiều thống kê $\hat{\theta}$ khác nhau để ước lượng cho tham số tổng thể θ. Vì vậy ta cần lựa chọn thống kê tốt nhất để ước lượng cho tham số θ dựa vào các tiêu chuẩn sau:

a. Ước lượng không chệch

Thống kê $\hat{\theta}$ của mẫu đgl ước lượng không chệch của tham số θ của biến ngẫu nhiên gốc X nếu:

$$\mathbb{E}(\hat{\theta}) = \theta$$

b. Ước lượng hiệu quả

Thống kê $\hat{\theta}$ của mẫu đgl ước lượng hiệu quả nhất của tham số θ của biến ngẫu nhiên gốc X nếu nó là ước lượng không chệch và có phương sai nhỏ nhất so với mọi ước lượng không chệch khác được xây dựng trên cùng mẫu đó.

c. Ước lượng vững

Thống kê $\hat{\theta}$ của mẫu đgl ước lượng vững của tham số θ của biến ngẫu nhiên gốc X nếu: $\hat{\theta}$ hội tụ theo xác suất đến θ khi $n \to \infty$

d. Ước lượng đủ

Một ước lượng $\hat{\theta}$ đgl ước lượng đủ nếu nó chứa đựng toàn bộ các thông tin trong mẫu về tham số θ của ước lượng.

- Giả sử cần khảo sát một đặc tính X trên một tổng thể xác định.
- BNN X có phân phối $F(x;\theta)$, tham số θ chưa biết.
- Chọn một mẫu ngẫu nhiên cỡ n: $X = (X_1, \dots, X_n)$.

Định nghĩa 1

Một ước lượng khoảng (interval estimator) của một tham số θ là một cặp các thống kê $L(X_1, \dots, X_n)$ và $U(X_1, \dots, X_n)$ của một mẫu ngẫu nhiên thỏa $L(\mathbf{X}) \le U(\mathbf{X})$, và $L(\mathbf{X}) \le \theta \le U(\mathbf{X})$. Nếu một mẫu thực nghiệm $\mathbf{x} = (x_1, \dots, x_n)$ được quan trắc, $[l(\mathbf{x}), u(\mathbf{x})]$ gọi là một khoảng ước lượng (interval estimate) cho θ.

Khoảng tin cậy

Định nghĩa 2

Xét vector ngẫu nhiên $\mathbf{X} = (X_1, \dots, X_n)$ có hàm mật độ đồng thời phụ thuộc vào tham số $\theta \in \Theta$ và $L(\mathbf{X})$ và $U(\mathbf{X})$ là hai thống kê sao cho $L(\mathbf{X}) \le U(\mathbf{X})$. Khi đó, khoảng ngẫu nhiên $[L(\mathbf{X}), U(\mathbf{X})]$ gọi là *khoảng tin cậy cho tham số θ với độ tin cậy* $100(1-\alpha)\%$ nếu

$$\mathbb{P}\{L(\mathbf{X}) \le \theta \le U(\mathbf{X})\} = 1 - \alpha \quad (1)$$

Chú ý

- Đôi khi, "khoảng tin cậy cho tham số θ với độ tin cậy $100(1-\alpha)\%$" thường được viết gọn là "khoảng tin cậy $100(1-\alpha)\%$ cho tham số θ".
- Với mẫu thực nghiệm $\mathbf{x} = (x_1, \dots, x_n)$, ta có khoảng tin cậy cụ thể cho tham số θ là
$$l(\mathbf{x}) \le \theta \le u(\mathbf{x})$$

- **Ý nghĩa**: Với 100% lần lấy mẫu cỡ n thì
  - Có $100(1-\alpha)\%$ lần lấy giá trị tham số $\theta \in [l, u]$
  - Có $100\alpha\%$ lần lấy giá trị tham số $\theta \notin [l, u]$
  - $1 - \alpha$ đgl hệ số tin cậy hay độ tin cậy.
  - Khi $\alpha = 0.05$ ta có khoảng tin cậy 95%
    Khi $\alpha = 0.01$ ta có khoảng tin cậy 99%

Ước lượng trung bình của tổng thể.

Bài toán

Cho tổng thể với trung bình μ với phương sai có thể đã biết hoặc chưa biết. Từ mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ hãy ước lượng μ với độ tin cậy $1 - \alpha$.

Cách giải quyết

Ta chia bài toán thành 3 trường hợp (TH) sau:

- TH1 Kích thước mẫu $n \geq 30$ (hoặc $n < 30$ nhưng X có phân phối chuẩn), σ^2 đã biết
- TH2 Kích thước mẫu $n \geq 30$, σ^2 chưa biết
- TH3 Kích thước mẫu $n < 30$, σ^2 chưa biết, X tuân theo quy luật phân phối chuẩn.

a. TH 1 : $n \geq 30$ (hoặc $n < 30$ nhưng X có phân phối chuẩn); σ^2 đã biết.

Mệnh đề 1

Trong trường hợp này, thống kê

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

có phân phối chuẩn $N(0,1)$.

a. TH 1 :$n \geq 30$ (hoặc $n < 30$ nhưng X có phân phối chuẩn); σ^2 đã biết.

Với độ tin cậy $1 - \alpha$, ta có

$$\mathbb{P}\left\{-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right\} = 1 - \alpha$$

hay

$$\mathbb{P}\left\{\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

với $z_{1-\frac{\alpha}{2}}$ là phân vị mức $1 - \alpha/2$ của phân phối chuẩn hóa $N(0,1)$.

a. TH 1 :$n \geq 30$ (hoặc $n < 30$ nhưng X có phân phối chuẩn); σ^2 đã biết.

- Với mẫu ngẫu nhiên, khoảng tin cậy cho tham số μ với độ tin cậy $1 - \alpha$ là
$$\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$
- Với mẫu thực nghiệm, khoảng tin cậy cho tham số μ với độ tin cậy $1 - \alpha$ là
$$\left[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$
- Đại lượng $\epsilon = z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ được gọi là dung sai (sai số giới hạn) của khoảng tin cậy.

a. TH 1 : $n \geq 30$ (hoặc $n < 30$ nhưng X có phân phối chuẩn); σ^2 đã biết.

Ví dụ: Hàm lượng kẽm trung bình thu hồi được từ một mẫu các giá trị đo kẽm tại 36 điểm đo khác nhau được xác định là 2.6g/ml. Xác định các khoảng tin cậy 95% và 99% cho mật độ kẽm trung bình ở sông. Giả thiết độ lệch tiêu chuẩn tổng thể là 0.3.

b. TH 2 : $n \geq 30$, σ^2 chưa biết.

Ta có thể dùng ước lượng của $Var(X)$ là S^2 để thay thế cho σ^2. Định lí giới hạn trung tâm nói rằng

Mệnh đề 2

Trong trường hợp này, thống kê

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

có phân phối chuẩn $N(0,1)$.

b. TH 2 : $n \geq 30$, σ^2 chưa biết.

Với độ tin cậy $1-\alpha$, ta có

$$\mathbb{P}\left\{-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right\} = 1-\alpha$$

hay

$$\mathbb{P}\left\{\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$

với $z_{1-\frac{\alpha}{2}}$ là phân vị mức $1-\alpha/2$ của phân phối chuẩn hóa $N(0,1)$.

b. TH 2 : $n \geq 30$, σ^2 chưa biết.

- Với mẫu ngẫu nhiên, khoảng tin cậy cho tham số μ với độ tin cậy $1-\alpha$ là
$$\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$
- Với mẫu thực nghiệm, khoảng tin cậy cho tham số μ với độ tin cậy $1-\alpha$ là
$$\left[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$$
- Đại lượng $\epsilon = z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$ được gọi là dung sai (sai số giới hạn) của khoảng tin cậy.

c. TH 3 : $n < 30$; X có phân phối chuẩn; σ^2 chưa biết.

Mệnh đề 3

Trong trường hợp này, thống kê

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

có phân phối Student với $n-1$ bậc tự do.

XÁC SUẤT THỐNG KÊ
N.T.M.Ngọc
4. Ước lượng tham số thống kê
4.1 Ước lượng điểm.
4.2 Ước lượng khoảng
4.3 Ước lượng trung bình
4.4 Ước lượng tỷ lệ

c. TH 3 : $n < 30$; X có phân phối chuẩn; σ^2 chưa biết.

Với độ tin cậy $1-\alpha$, ta có

$$\mathbb{P}\left\{-t_{1-\frac{\alpha}{2}}^{n-1} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le t_{1-\frac{\alpha}{2}}^{n-1}\right\} = 1 - \alpha$$

hay

$$\mathbb{P}\left\{\bar{X} - t_{1-\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{1-\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}}\right\} = 1 - \alpha$$

với $t_{1-\frac{\alpha}{2}}^{n-1}$ là phân vị mức $1-\frac{\alpha}{2}$ của luật phân phối Student với $n-1$ bậc tự do.

XÁC SUẤT THỐNG KÊ
N.T.M.Ngọc
4. Ước lượng tham số thống kê
4.1 Ước lượng điểm.
4.2 Ước lượng khoảng
4.3 Ước lượng trung bình
4.4 Ước lượng tỷ lệ

c. TH 3 : $n < 30$; X có phân phối chuẩn; σ^2 chưa biết.

- Với mẫu ngẫu nhiên, khoảng tin cậy cho tham số μ với độ tin cậy $1-\alpha$ là
$$\left[\bar{X} - t_{1-\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}}^{n-1}\frac{S}{\sqrt{n}}\right]$$
- Với mẫu thực nghiệm, khoảng tin cậy cho tham số μ với độ tin cậy $1-\alpha$ là
$$\left[\bar{x} - t_{1-\frac{\alpha}{2}}^{n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\frac{\alpha}{2}}^{n-1}\frac{s}{\sqrt{n}}\right]$$
- Đại lượng $\epsilon = t_{1-\frac{\alpha}{2}}^{n-1}\frac{s}{\sqrt{n}}$ được gọi là dung sai (sai số giới hạn) của khoảng tin cậy.

XÁC SUẤT THỐNG KÊ
N.T.M.Ngọc
4. Ước lượng tham số thống kê
4.1 Ước lượng điểm.
4.2 Ước lượng khoảng
4.3 Ước lượng trung bình
4.4 Ước lượng tỷ lệ

c. TH 3 : $n < 30$; X có phân phối chuẩn; σ^2 chưa biết.

Ví dụ: Các hàm lượng của 7 container axit sulfuric là $9.8, 10.2, 10.4, 9.8, 10.0, 10.2, 9.6$ lít. Tìm khoảng tin cậy 95% cho giá trị trung bình của tất cả các container đó, giả sử có phân phối chuẩn ước lượng.

Ước lượng trung bình của tổng thể

Tóm lại,

Các bước thực hiện

B1 Tìm trung bình mẫu $\bar{x}$ và phương sai mẫu s^2.

B2 Xác định trường hợp áp dụng:

- TH1 $n \geq 30$ (hoặc $n < 30$, X có phân phối chuẩn) và σ^2 đã biết.
- TH2 $n \geq 30$, σ^2 chưa biết.
- TH3 $n < 30$, X có phân phối chuẩn, và σ^2 chưa biết.

B3 Tìm phân vị: $z_{1-\alpha/2}$ nếu là TH1 và TH2; hoặc $t^{n-1}_{1-\alpha/2}$ nếu là TH3.

B4 Tìm dung sai:

$$\epsilon = \begin{cases} z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} & \text{nếu TH1} \\ z_{1-\alpha/2}\frac{s}{\sqrt{n}} & \text{nếu TH2} \\ t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}} & \text{nếu TH3} \end{cases}$$

KL Khoảng tin cậy $100(1-\alpha)\%$ cho trung bình của tổng thể là $[\bar{x}-\epsilon, \bar{x}+\epsilon]$.

Ví dụ:

Biết lương tháng của công nhân (Đv: triệu đồng) trong một nhà máy có phân phối chuẩn. Chọn ngẫu nhiên 16 công nhân khảo sát.

Lương tháng	0.8	1.0	1.2	1.3	1.5	1.7	2	2.3	2.5
Số công nhân	1	1	2	2	2	3	2	2	1

a. Giả sử $\sigma = 0.63$, tìm KTC 96% cho mức lương trung bình hàng tháng của một công nhân.

b. Lập KTC 99% cho mức lương trung bình.

Ước lượng tỷ lệ của tổng thể

Bài toán

Cho tổng thể X, trong đó tỷ lệ cá thể mang đặc tính $\mathcal{A}$ nào đó trong tổng thể là p. Từ mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ hãy tìm khoảng tin cậy cho p với độ tin cậy $1-\alpha$.

- Gọi Y là số phần tử thỏa tính chất $\mathcal{A}$ trong n phần tử khảo sát, thì $Y \sim B(n,p)$.
- Đặt
$$\hat{P} = \frac{Y}{n} \qquad (2)$$
- Thống kê $\hat{P}$ có kỳ vọng và phương sai lần lượt là
$$\mathbb{E}(\hat{P}) = \mu_{\hat{P}} = p, \quad \mathbb{V}\mathrm{ar}(\hat{P}) = \sigma^2_{\hat{P}} = \frac{p(1-p)}{n}$$

Ước lượng tỷ lệ của tổng thể

Mệnh đề 4

Thống kê

$$Z = \frac{\hat{P} - \mu_{\hat{P}}}{\sigma_{\hat{P}}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0,1) \quad (3)$$

và

$$W = \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \rightsquigarrow \mathcal{N}(0,1) \quad (4)$$

XÁC SUẤT THỐNG KÊ
N.T.M.Ngọc

Do đó, khi kích thước mẫu đủ lớn,

$$\mathbb{P}\left\{-z_{1-\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \leq z_{1-\alpha/2}\right\} = 1 - \alpha \quad (5)$$

hay

$$\mathbb{P}\left\{\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right\} = 1 - \alpha \quad (6)$$

XÁC SUẤT THỐNG KÊ
N.T.M.Ngọc

Vậy

- Với mẫu ngẫu nhiên, khoảng tin cậy $100(1-\alpha)\%$ cho p là

$$\left[\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right]$$

- Với mẫu thực nghiệm, khoảng tin cậy $100(1-\alpha)\%$ cho p là

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

XÁC SUẤT THỐNG KÊ
N.T.M.Ngọc

Ước lượng tỷ lệ của tổng thể

Trong thực hành, ta thực hiện theo các bước sau đây:

Các bước thực hiện

B1 Tìm tỉ lệ mẫu: $\hat{p}$.

B2 Kiểm tra điều kiện: $n\hat{p} \geq 5$ và $n(1-\hat{p}) \geq 5$.

B3 Tìm phân vị: $z_{1-\alpha/2}$ bằng cách tra bảng.

B4 Tìm dung sai: $\epsilon = z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

KL Khoảng tin cậy $100(1-\alpha)\%$ cho tỷ lệ của tổng thể là $[\hat{p} - \epsilon, \hat{p} + \epsilon]$.

Biết lương tháng của công nhân (Đv: triệu đồng) trong một nhà máy có phân phối chuẩn. Chọn ngẫu nhiên 16 công nhân khảo sát.

Lương tháng	0.8	1.0	1.2	1.3	1.5	1.7	2	2.3	2.5
Số công nhân	1	1	2	2	2	3	2	2	1

Công nhân gọi là có thu nhập cao nếu lương tháng từ 2 triệu đồng trở lên. Hãy lập khoảng tin cậy 95% cho tỉ lệ công nhân có thu nhập cao.

Xác định kích thước mẫu.

Nhận xét

- Chất lượng của ước lượng phản ánh qua độ tin cậy $1-\alpha$ và dung sai ϵ.
- Độ tin cậy càng cao và dung sai càng nhỏ thì ước lượng đó càng tốt.
- Tuy nhiên, dung sai ϵ lại phụ thuộc vào kích thước mẫu n và độ tin cậy $1-\alpha$.

Câu hỏi

Với độ tin cậy $1-\alpha$, nếu ta muốn dung sai ϵ đạt được ở một mức nào đó cho trước thì kích thước mẫu n tối thiểu là bao nhiêu?

Bài toán

Tìm giá trị nhỏ nhất của n sao cho $\epsilon \le \epsilon_0$, với ϵ_0 và α cho trước.

Xác định kích thước mẫu

Khi ước lượng trung bình tổng thể

a. Nếu biết $\mathbb{V}ar(X) = \sigma^2$, từ công thức

$$\epsilon = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Để $\epsilon \le \epsilon_0$ ta cần chọn

$$n \ge (z_{1-\alpha/2})^2\frac{\sigma^2}{\epsilon_0^2}$$

b. Nếu chưa biết σ^2, ta căn cứ vào mẫu đã cho để tính s^2. Từ đó ta xác định được kích thước mẫu tối thiểu:

$$n \ge (z_{1-\alpha/2})^2\frac{s^2}{\epsilon_0^2}$$

Xác định kích thước mẫu

Khi ước lượng tỷ lệ tổng thể

a. Khi đã biết $\hat{p}$, để $\epsilon \le \epsilon_0$ thì từ công thức

$$\epsilon = z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n \ge (z_{1-\alpha/2})^2\frac{\hat{p}(1-\hat{p})}{\epsilon_0^2}$$

b. Khi chưa biết $\hat{p}$, ta có $\epsilon = z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Do $\hat{p}(1-\hat{p})$ đạt giá trị cực đại 0.25 khi $\hat{p} = 0.5$ nên

$$\epsilon \le z_{1-\alpha/2}\sqrt{\frac{0.25}{n}}$$

Do đó, để $\epsilon \le \epsilon_0$ ta chọn n sao cho $z_{1-\alpha/2}\sqrt{\frac{0.25}{n}} \le \epsilon_0$ tức là

$$n \ge \frac{0.25(z_{1-\alpha/2})^2}{\epsilon_0^2}$$