THỐNG KÊ THỨ TỰ (order statistics)

hàm số với đối số là mẫu dữ liệu và có giá trị bằng số hạng ở vị trí thứ bậc đã xác định trong dãy số được sắp thứ tự từ nhỏ đến lớn của dữ liệu. Ví dụ, hàm $\mathbf{min}$ ứng với thống kê thứ tự bậc thứ nhất, hàm $\mathbf{max}$ ứng với thống kê thứ tự bậc cao nhất, hàm $\mathbf{med}$ (trung vị) ứng với thống kê thứ tự bậc "chính giữa" trong dãy dữ liệu đã sắp thứ tự. Thuật ngữ **thống kê thứ tự** chứa từ "thống kê'" được hiểu theo nghĩa hẹp, liên quan đến khái niệm "mẫu ngẫu nhiên", là một hàm số với đối số là mẫu dữ liệu và thường nhận giá trị là số thực hoặc véc tơ thực. Điều này khác với nghĩa rộng của từ "thống kê" dùng để chỉ ngành khoa học "sử dụng tư duy hợp lý để nghiên cứu các quy luật trong dữ liệu của đám đông". Lý thuyết thống kê thứ tự được nghiên cứu tốt nhất đối với trường hợp khi các thành phần của véc tơ mẫu ngẫu nhiên là các biến ngẫu nhiên độc lập có cùng phân bố, như giả định từ đây trở đi.

Theo định nghĩa, biến ngẫu nhiên là một hàm $X = X(\omega)$ xác định trên Ω và nhận giá trị số thực, trong đó Ω là tập các biến cố cơ sở. Với một biến ngẫu nhiên X cần quan tâm và một số tự nhiên $n \geq 2$ cố định cho trước, mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ là một bộ n bản sao biến ngẫu nhiên độc lập cùng phân bố với X, khi đó n được gọi là cỡ mẫu. Như vậy, mẫu ngẫu nhiên là một véc tơ ngẫu nhiên với các thành phần độc lập cùng phân bố. Giá trị của véc tơ ngẫu nhiên đó tại một biến cố cơ sở cụ thể ω chính là một mẫu thực nghiệm $(x_1, x_2, \ldots, x_n)$, như một dãy n số thực, là một véc tơ trong không gian $\mathbb{R}^n$. Sắp xếp lại tọa độ của véc tơ này theo thứ tự tăng dần về độ lớn, ta thu được một véc tơ mới $(x_{(n1)}, x_{(n2)}, \ldots, x_{(nn)})$ với các thành phần thỏa mãn quan hệ

$$x_{(n1)} \leq x_{(n2)} \leq \cdots \leq x_{(nn)} .$$

Đây chính là thể hiện $(X_{(n1)}(\omega), X_{(n2)}(\omega), \ldots, X_{(nn)}(\omega))$ tại biến cố cơ sở $\omega \in \Omega$ của véc tơ ngẫu nhiên mới $(X_{(n1)}, X_{(n2)}, \ldots, X_{(nn)})$. Khi đó hàm số

$$(X_1, X_2, \ldots, X_n) \mapsto (X_{(n1)}, X_{(n2)}, \ldots, X_{(nn)})$$

được gọi là *chuỗi* (hoặc *véc tơ*) *thống kê thứ tự* và thành phần thứ k của nó $(X_{(nk)})$ được gọi là *thống kê thứ tự thứ* k.

Trong khi các thành phần trong véc tơ mẫu ban đầu là các biến ngẫu nhiên độc lập có cùng phân bố $F(u)$, thì các thành phần trong véc tơ thống

kê thứ tự mới được thành lập như trên không còn độc lập với nhau nữa và hàm phân bố của chúng cũng thay đổi so với hàm phân bố $F(u)$. Hàm phân bố $F_{nk}(u)$ của thống kê thứ tự thứ k có dạng

$$F_{nk}(u) = \mathbb{P}\{X_{(nk)} \le u\} = I_{F(u)}(k, n-k+1)\,, \tag{1}$$

ở đây

$$I_y(a,b) = \frac{1}{B(a,b)} \int\limits_0^y x^{a-1}(1-x)^{b-1}$$

là hàm beta không đầy đủ, $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$. Từ (1) dẫn đến kết luận nếu hàm phân bố $F(u)$ có hàm mật độ xác suất $f(u)$, thì hàm mật độ xác suất $f_{nk}(u)$ của thống kê thứ tự thứ k, là $X_{(nk)}$, $k = 1, 2, \ldots, n$, cũng tồn tại và có dạng

$$f_{nk}(u) = \frac{n!}{(k-1)!(n-k)!}[F(u)]^{k-1}[1-F(u)]^{n-k} f(u)\,,\ -\infty < u < \infty\,. \tag{2}$$

Nếu hàm mật độ xác suất $f(u)$ tồn tại, thì với mọi $1 \le r_1 < \cdots < r_k \le n$, $k \le n$, hàm mật độ xác suất đồng thời $f_{r_1 \cdots r_k}(u_1, \ldots, u_k)$ của các thống kê thứ tự $X_{(n1)}, \ldots, X_{(nk)}$ sẽ được cho bằng biểu thức

$$f_{r_1 \cdots r_k}(u_1, \ldots, u_k) = \frac{n!}{(r_1-1)!(r_2-r_1-1)! \cdots (n-r_k)!}$$

$$\times F^{r_1-1}(u_1) f(u_1)[F(u_2)-F(u_1)]^{r_2-r_1-1} f(u_2) \cdots [1-F(u_k)]^{n-r_k} f(u_k)\,,$$

$$-\infty < u_1 < \cdots < u_k < \infty\,. \tag{3}$$

Các công thức (1) - (3) cho phép tìm ra phân bố các *thống kê thứ tự cực trị* (hai *thống kê giá trị nhỏ nhất mẫu* và *giá trị lớn nhất mẫu*),

$$X_{(n1)} = \min(X_1, \ldots, X_n) \ \text{ và } \ X_{(nn)} = \max(X_1, \ldots, X_n)\,,$$

và phân bố của $W_n = X_{(nn)} - X_{(n1)}$, được gọi là *thống kê độ trải* (hoặc *độ trải mẫu*). Khi hàm phân bố $F(u)$ liên tục, phân bố của W_n bằng

$$\mathbb{P}\{W_n < w\} = n \int\limits_{-\infty}^{\infty} [F(u+w) - F(u)]^{n-1} dF(u)\,, \ \ w \ge 0\,. \tag{4}$$

Các công thức (1) - (4) cho thấy không thể sử dụng các phân bố chính xác của thống kê thứ tự để thu được các suy luận thống kê nếu không biết hàm phân bố $F(u)$. Chính vì vậy, các phương pháp tiệm cận cho các hàm phân bố của thống kê thứ tự, khi số chiều n của véc tơ quan sát tiến tới vô cùng, đã được phát triển rộng rãi trong lý thuyết thống kê thứ tự. Trong lý thuyết tiệm cận của thống kê thứ tự, người ta nghiên cứu sự phân bố giới hạn khi $n \to \infty$ của các chuỗi thống kê thứ tự $\{X_{(nk)}\}$ được chuẩn hóa một cách thích hợp. Số thứ tự k có thể thay đổi như một hàm của cỡ mẫu n. Nếu số thứ tự k thay đổi khi n tăng đến vô cùng sao cho giới hạn $\lim_{n\to\infty}(k/n)$ tồn tại và không bằng 0 hoặc bằng 1, thì thống kê thứ tự $\{X_{(nk)}\}$ tương ứng của dãy cần quan tâm được gọi là *thống kê thứ tự trung tâm*. Khi $\lim_{n\to\infty}(k/n)$ bằng 0 hoặc bằng 1, chúng được gọi là *thống kê thứ tự cực trị*.

Trong thống kê toán học, thống kê thứ tự trung tâm được sử dụng để xây dựng các dãy nhất quán của các ước lượng vững cho các phân vị của một phân bố $F(u)$ chưa biết, dựa trên véc tơ các thể hiện của một biến ngẫu nhiên X, hay nói cách khác, để ước lượng hàm ngược $F^{-1}(u)$. Chẳng hạn, giả sử x_p là phân vị mức p ($0 < p < 1$) của hàm phân bố $F(u)$, biết rằng mật độ xác suất $f(u)$ của nó liên tục và dương ngặt trong một lân cận nào đó của điểm x_p. Lúc này, chuỗi thống kê thứ tự trung tâm $\{X_{(nk}\}$ với các số thứ tự $k = [(n+1)p + 0.5]$, trong đó $[a]$ là phần nguyên của số thực a, là một chuỗi nhất quán các ước lượng vững cho các phân vị x_p, $n \to \infty$. Hơn nữa, chuỗi thống kê thứ tự $\{X_{(nk}\}$ này có phân bố tiệm cận chuẩn với các tham số kỳ vọng và phương sai

$$x_p \quad \text{và} \quad \frac{p(1-p)}{f^2(x_p)(n+1)}\,,$$

tức là với mọi số thực x bất kỳ,

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \frac{X_{(nk)} - x_p}{\sqrt{p(1-p)/(n+1)}} f(x_p) < x \Big\} = \Phi(x)\,,$$

ở đây $\Phi(x)$ là hàm phân bố chuẩn tắc.

Ví dụ 1. Cho $(X_{(n1)}, \ldots, X_{(nn)})$ là véc tơ thống kê thứ tự dựa trên véc tơ ngẫu nhiên $(X_1, \ldots, X_n)$. Các thành phần của véc tơ này được giả định là

các biến ngẫu nhiên độc lập có cùng phân bố xác suất với mật độ xác suất liên tục và dương trong một lân cận của trung vị $x_{1/2}$. Trong trường hợp này, dãy các trung vị mẫu $\{\mu_n\}$, được xác định cho $n \geq 2$ bất kỳ bởi

$$\mu_n = \begin{cases} X_{(n[m+1])} & \text{nếu } n = 2m+1 \text{ (lẻ)}, \\ \frac{1}{2}(X_{(nm)} + X_{(n[m+1])}) & \text{nếu } n = 2m \text{ (chẵn)}, \end{cases}$$

có phân bố tiệm cận chuẩn, với các tham số kỳ vọng và phương sai

$$x_{1/2} \quad \text{và} \quad \{4(n+1)f^2(x_{1/2})\}^{-1},$$

khi $n \to \infty$. Đặc biệt, nếu

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\Big\{-\frac{(x-\alpha)^2}{2\sigma^2}\Big\}, \;\; |\alpha| < \infty, \;\; \sigma > 0,$$

nghĩa là X_i có phân bố chuẩn $N(\alpha; \sigma^2)$, thì chuỗi $\{\mu_n\}$ có phân bố tiệm cận chuẩn với các tham số $x_{1/2} = \alpha$ và $\pi\sigma^2/(2(n+1))$. Nếu so sánh chuỗi thống kê $\{\mu_n\}$ với chuỗi các ước lượng không chệch tốt nhất

$$\{\bar{X}_n\}, \;\; \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

để ước lượng kỳ vọng α của phân bố chuẩn, thì nên dùng chuỗi $\{\bar{X}_n\}$, vì

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} < \frac{\pi\sigma^2}{2(n+1)} \approx \mathrm{Var}(\mu_n)$$

với mọi $n \geq 2$.

Ví dụ 2. Cho $(X_{(n1)}, \ldots, X_{(nn)})$ là véc tơ thống kê thứ tự dựa trên véc tơ ngẫu nhiên $(X_1, \ldots, X_n)$ có các thành phần độc lập và phân bố đều trên đoạn thẳng $[\alpha - h; \alpha + h]$. Hơn nữa, giả sử các tham số α và h chưa biết. Trong trường hợp này, các dãy thống kê $\{Y_n\}$ và $\{Z_n\}$, trong đó

$$Y_n = \frac{1}{2}(X_{(n1)} + X_{(nn)}) \;\; \text{và} \;\; Z_n = \frac{n+1}{2}(n-1)(X_{(nn)} - X_{(n1)}), \;\; n \geq 2,$$

tương ứng là các chuỗi nhất quán các ước lượng không chệch hiệu quả cho α và h. Hơn thế nữa,

$$\mathrm{Var}(Y_n) = \frac{2h^2}{(n+1)(n+2)} \text{ và } \mathrm{Var}(Z_n) = \frac{2h^2}{(n-1)(n+2)}\,.$$

Có thể chỉ ra rằng các dãy $\{Y_n\}$ và $\{Z_n\}$ xác định các ước lượng tốt nhất cho α và h theo nghĩa có rủi ro bình phương nhỏ nhất trong lớp các ước lượng tuyến tính không chệch được thể hiện dưới dạng thống kê thứ tự.

HỒ ĐĂNG PHÚC

Tài liệu tham khảo

1. H. A. David, *Order Statistics*, Wiley, New York-London-Sydney, 1970.
2. E. J. Gumble, *Statistics of Extremes*, Columbia Univ. Press, New York, 1958.
3. J. Hájek, Z. Sidák, *Theory of Rank Tests*, Acad. Press, New York-London, 1967.
4. S. S. Wilks, *Mathematical Statistics*, Princeton Univ. Press, NJ, 1950.