GIỚI THIỆU MỘT SỐ PH□ƠNG PHÁP SẮP XẾP THỐNG KÊ

Phạm Sơn

Đặt vấn đề

Trong thực tế cuộc sống một vấn đề thường gặp đòi hỏi thống kê phải giải quyết là sắp xếp thứ hạng của nhiều tổng thể thống kê khi chúng ta đã quan sát và thu thập được số liệu của một số tin tức có liên quan. Chẳng hạn trong thế vận hội Bắc Kinh năm 2008 có trên 200 đoàn tham gia thi đấu 35 môn để tranh 3 loại huy chương: Vàng, bạc và đồng. Trong 18 ngày thi đấu, chúng ta đều được ban tổ chức công bố thứ hạng của các đoàn. Đó thực chất là một bài toán sắp xếp thống kê.

Trong Thông tin khoa học thống kê số 5/2007 có giới thiệu bài viết của TS. Đặng Quảng *"Phương pháp sắp thứ ý kiến liên kết đồng thời hai đối tượng trở lên"*. Đây là cách sắp xếp đặc thù trong một số nghiên cứu thống kê. Trong bài này sẽ giới thiệu một số phương pháp sắp xếp được sử dụng khá phổ biến trong thực tế.

Một cách khái quát, chúng ta có N tổng thể thống kê: chẳng hạn có N doanh nghiệp, N địa phương hoặc N người. Và ở mỗi tổng thể chúng ta thu thập được thông tin theo n tiêu thức ta ký hiệu là $x_1^k, x_2^k, ..., x_n^k \ (k=1 \div N)$. Bài toán thống kê định ra là sắp xếp thứ tự của N tổng thể, hay cụ thể hơn sắp xếp tập hợp: $(x_1^k, x_2^k, ..., x_n^k)$ với $k=1 \div N$ theo thứ tự tăng dần và giảm dần.

Để giải quyết bài toán trên, các nhà thống kê học đưa ra nhiều phương pháp nhưng phổ biến nhất là 4 phương pháp sau:

1. Phương pháp sắp xếp theo thứ tự ưu tiên của các chỉ tiêu

Thực chất của phương pháp này như sau: Trước hết cơ quan nghiên cứu hoặc nhà nghiên cứu phải sắp xếp thứ tự quan trọng của các chỉ tiêu. Để đơn giản, trong bài viết này chúng tôi lấy thứ tự ưu tiên của các chỉ tiêu là: $x_1, \ldots, x_n$. Sau đó tiến hành sắp xếp thứ tự của các tổng thể theo các bước sau:

Bước 1: Sắp xếp các tổng thể theo chỉ tiêu x_1 theo công thức:

$(x_1^i, x_2^i, ..., x_n^i)$ xếp trước $(x_1^j, x_2^j, ..., x_n^j)$

Nếu $x_1^i > x_2^i$ nếu $x_1^i = x_1^j$ chuyển sang bước 2.

Bước 2: So sánh chỉ tiêu x_2^i với x_2^j, nếu như x_2^i và x_2^j cùng giá trị, chúng ta tiếp tục so sánh x_3^i với x_3^j và quá trình cứ tiếp diễn cho đến x_n^i và x_n^j.

Nếu sau khi so sánh hết tất cả chỉ tiêu mà 2 tổng thể không hơn kém, chúng ta nói đó là 2 tổng thể đồng hạng.

Để minh hoạ, chúng tôi thí dụ như sắp xếp thành tích của cả đoàn tham gia thế vận hội Bắc Kinh năm 2008 theo thứ tự số huy chương vàng, bạc và đồng đạt được.

Ưu điểm của phương pháp này là việc sắp xếp đơn giản và thuận tiện. Nhưng nhược điểm là đánh đồng thành tích các môn. Do vậy có trường hợp như 1 vận động viên bơi lội của Mỹ đạt 8 huy chương vàng, trong khi đó nhiều môn có nhiều vận động viên thi đấu như bóng đá, bóng chuyền, bóng rổ... cũng chỉ có 1 huy chương vàng, 1

bạc và 1 đồng và điều đó ảnh hưởng rất lớn đến thành tích thi đấu của các đoàn.

2. Phương pháp tính khoảng cách (hay trong toán học gọi là độ dài Oclit)

Cơ sở khoa học của phương pháp này là chuyển các đại lượng của không gian n chiều về không gian 1 chiều theo độ đo khoảng cách của không gian Oclit. Về việc so sánh trong không gian nhiều chiều chỉ thực hiện được trong một bộ phận. Ngược lại trong không gian 1 chiều, việc sắp xếp được thực hiện đầy đủ. Tuy nhiên, trong thống kê kinh tế xã hội, các chỉ tiêu thống kê không phải là một đại lượng vô hướng, mà thường gắn với các số đo khác nhau. Do vậy, muốn áp dụng phương pháp này, trước hết phải chuyển các chỉ tiêu theo các số đo khác nhau về cùng một số đo theo phương pháp sau:

$$P_i^k = \frac{x_i^k}{\sum_{i=1}^{n} x_i^k} \times 100 \qquad (1)$$

Lúc đó P_i^k chính là tỷ trọng của x_i^k trong tổng số. Bằng công thức (1), chúng ta đã chuyển đổi các số đo về cùng một số đo. Đó là tỷ lệ phần trăm. Sau đó chúng ta sử dụng công thức:

$$I^k = \sqrt{\sum_{i=1}^{n} \left(P_i^k\right)^2} \qquad (2)$$

Trong đó I^k là khoảng cách Oclit trong không gian n chiều.

Lúc đó chúng ta chỉ tiến hành sắp xếp N tổng thể theo giá trị của $I^k \ (k = 1 \div N)$.

3. Phương pháp quy đổi (trong thống kê thường gọi là phương pháp cho điểm hay là định giá)

Thực chất của phương pháp trên như sau: Đối với từng chi tiết chúng ta gán cho 1 quyền số nhất định (hay cho 1 số điểm nhất định). Như vậy có n chỉ tiêu có n bộ đếm ký hiệu là $\alpha_1, \alpha_2, ..., \alpha_n$: Lúc đó ta có tổng giá trị của từng tổng thể như sau:

$$M^k = \sum_{i=1}^{n} x_i^k \times \alpha_i \qquad (3) \quad (k = 1 \div N)$$

Tính $\overline{M}^k = \frac{M_k}{n}$ (4)

Việc sắp xếp k tổng thể có thể căn cứ vào giá trị của M^k hoặc $\overline{M}^k$.

Đây là phương pháp phổ biến trong thống kê. Chẳng hạn khi xem xét doanh thu của 2 điểm bán hàng hoặc giá trị sản lượng của một tỉnh, huyện hoặc khi chấm điểm thi đua cho các đơn vị…

Trong thống kê kinh tế, phương pháp này thường áp dụng trong thống kê thương mại để tính tổng mức bán/mua các hàng hóa, hoặc trong tính các chỉ tiêu giá trị sản lượng công nghiệp, nông nghiệp… theo công thức:

$$Q = \sum_{i=1}^{N} p_i q_i \qquad (5)$$

Trong đó:

pi là mức giá của sản phẩm thứ i (i=1÷N)

qi là lượng sản phẩm thứ i (i=1÷N)

Nhược điểm lớn nhất của phương pháp này là cách cho điểm cho từng chỉ tiêu hay là định giá cho từng loại sản phẩm.

4. Phương pháp tính số bình quân nhiều chiều

Đây là một phương pháp sắp xếp có nhiều ưu điểm (Phương pháp này sẽ được trình bày cụ thể hơn trong bài báo: HDI, chỉ số bình quân nhiều chiều đăng trên số tiếp theo).

Cần lưu ý rằng, các phương pháp sắp xếp giới thiệu trên đây chỉ là những nội dung cơ bản. Trong việc ứng dụng, đã có nhiều cải tiến. Độc giả có thể tham khảo thêm cách xếp hạng các đội bóng, các kỳ thủ hoặc các vận động viên. Đồng thời cũng cần nhấn mạnh rằng, bất kỳ phương pháp so sánh nào cũng có những ưu nhược điểm của nó. Tuy nhiên, cố nhân thường dạy: "méo mó có hơn không". Và do đó thế giới vẫn chấp nhận các phương pháp sắp xếp thống kê trên đây■