

VỀ VIỆC ÁP DỤNG MỘT SỐ BẤT ĐẲNG THỨC TRONG LÝ THUYẾT XÁC SUẤT VÀO THỐNG KÊ

Thạc sĩ NGUYỄN VĂN SĨ

Vấn đề đánh giá, ước lượng tỷ lệ của những phần tử đặc biệt nào đó trong một tổng thể là một vấn đề quan trọng trong kinh tế và xã hội: dự đoán tỷ lệ sản phẩm hỏng trong sản xuất hay trong kinh doanh, dự đoán tỷ lệ những người mắc một căn bệnh nào đó trong xã hội, dự đoán tỷ lệ những cử tri bỏ phiếu cho một người nào đó trước một cuộc bầu cử...

Thông thường, ta dự đoán tỷ lệ p các phần tử có tính chất A nào đó mà ta quan tâm bằng tỷ lệ mẫu $f_n = \frac{m}{n}$ trong đó n là số phần tử mà ta quan sát từ đám đông (tổng thể) còn m là số phần tử có tính chất A ta phát hiện được trong mẫu điều tra này.

Vấn đề đặt ra là: ta phải tiến hành quan sát ít nhất bao nhiêu phần tử để: $P(|f_n - p| > \varepsilon) \leq \alpha$ (*)

Từ lâu người ta đã tính số n từ các bất đẳng thức sau trong xác suất:

1. Bất đẳng thức Chebyshev

Nếu đại lượng ngẫu nhiên X có kỳ vọng $M(X) = \mu$ và

phương sai $D(X) = \sigma^2$ thì $P(|X - \mu| > \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$

Ta có thể tính được: $M(f_n) = p$ và $D(f_n) = \frac{pq}{n}$ (với $q = 1 - p$)

Do đó:

$$P\left(\left|\frac{m}{n} - p\right| > \varepsilon\right) \leq \frac{pq}{n\varepsilon^2} \leq \alpha$$

Từ đó: $n \geq \frac{pq}{\alpha\varepsilon^2}$ (1)

2. Sử dụng định lý giới hạn trung tâm

Với n "khá lớn" ta có thể xem f_n có phân phối chuẩn $f_n \in N(p, \frac{pq}{n})$. Vì vậy:

$$P(|f_n - p| \leq t_2 \sqrt{\frac{pq}{n}}) = 1 - \alpha \quad (2)$$

Trong đó t_2 xác định từ biểu thức:

$$\frac{1}{\sqrt{2\pi}} \int_{-t_2}^{t_2} e^{-\frac{x^2}{2}} dx = 1 - \alpha$$

Từ bất đẳng thức (2) ở trên ta có: $\varepsilon \geq t_2 \sqrt{\frac{pq}{n}}$ hay

$$n \geq \frac{t_2^2 pq}{\varepsilon^2} \quad (3)$$

3. M. Talagrand đã dựa vào bất đẳng thức tập trung độ đo trong không gian Banach để chứng minh được một bất đẳng thức mà áp dụng vào bất đẳng thức (*) ở trên thì nó có dạng:

$$P(|f_n - p| > \varepsilon) \leq 2 \exp(-2n\varepsilon^2) \leq \alpha$$

từ đó:

$$n \geq \left\lceil \frac{\ln\left(\frac{2}{\alpha}\right)}{2\varepsilon^2} \right\rceil \quad (4)$$

Như vậy là ta có 3 bất đẳng thức là (1) (3) và (4) để tìm số n cần thiết.

Bất đẳng thức (1) và (3) có nhược điểm là trong khi tính toán n , công thức lại phụ thuộc vào tham số p chưa biết. Nhưng vì $pq = p(1 - p)$ đạt cực đại với $p = \frac{1}{2}$ nên để

"chắc chắn" ta sẽ tính với $p = \frac{1}{2}$. Khi đó p bằng bao nhiêu vẫn đảm bảo mức chính xác và độ tin cậy $1 - \varepsilon$:

Ta lấy ví dụ khi cần $\alpha = 0,05$ và $\varepsilon = 0,02$

áp dụng bất đẳng thức (1) ta tính được $n_1 = 12500$ và dùng bất đẳng thức (3) (với $t_2 = 1,96$) ta tính được $n_2 = 2500$.

Nếu áp dụng công thức (4) ta tính được $n_3 = 4650$

Chú ý rằng ở đây n_2 có nhỏ hơn n_1 và n_3 nhưng thực ra, ta đã bỏ qua một "tảng" của sự chính xác, ở trên ta nói: với n "đủ lớn" thì f_n xem như có phân phối chuẩn, lẽ ra ta phải dùng bất đẳng thức Bery Essen:

$$\sup_{-\infty < x < \infty} |F_n(x) - \phi(x)| \leq 0,8 \frac{p^2 + q^2}{\sqrt{pqn}} \quad (5)$$

để tìm n'' sao cho hàm phân phối của f_n xấp xỉ với hàm phân phối chuẩn ở một mức nào đó mà ta không bàn ở đây. Sau đó ta tính $n_2'' = \max(n_2', n_2)$.

Vấn đề đặt ra sao cho các nhà thống kê là ý nghĩa thực tiễn của các con số tính ra ở trên, nói chung n_1, n_2, n_3 đều rất lớn. Ở các nước giàu, kỹ thuật tiên tiến, như nước Mỹ chẳng hạn, trước một cuộc tổng tuyển cử cần thăm dò xem tỷ lệ phần trăm những người bỏ phiếu cho một vị tổng thống nào đó mà cần điều tra thì 4.650 cử tri là việc không khó khăn. Nhưng ở nước ta, đã gọi điều tra mẫu mà tiến hành số quan sát quá lớn như vậy là không thực tiễn đôi khi dẫn tới vô lý. Trong thực tế bất đẳng thức (4) không phụ thuộc vào p . Đối với bất đẳng thức (1) và (3) ta có thể giảm con số mẫu bằng cách tiến hành lấy mẫu hai lần:

Lần 1 điều tra $n^{(1)}$ phần tử và tính được $f_n^{(1)} = \frac{m^{(1)}}{n^{(1)}}$.

Dùng $f_n^{(1)}$ này thay cho p trong (1) và (3) ta sẽ tính gần đúng được n cần thiết và lần 2 sẽ điều tra thêm $n^{(2)} = n - n^{(1)}$ quan sát nữa. Chẳng hạn điều tra sơ bộ $n^{(1)} = 400$ phần tử, ta tính được $f_n^{(1)} = 0,15$ nếu vẫn lấy $\varepsilon = 0,02$ và $\alpha = 0,05$ ta tính được $n_1 = 6375$; $n_2 = 1225$.

Ta thấy số phần tử cần điều tra giảm đi rất nhiều. Tuy nhiên, ở đây ta vẫn còn bỏ qua một "tảng" của sự xấp xỉ và cái chủ yếu là con số vẫn khá lớn. Đó là điều tiếp tục phải đặt ra cho các nhà thống kê đặc biệt là khi tiến hành lấy số liệu khó khăn và tốn kém.