Thống kê suy diễn

Bởi:

Phạm Trí Cao

Thống kê suy diễn - vấn đề ước lượng

Ước lượng

Chúng ta tìm hiểu bản chất, đặc trưng và yêu cầu của ước lượng thống kê thông qua một ví dụ đơn giản là ước lượng giá trị trung bình của tổng thể.

Ví dụ 11. Giả sử chúng ta muốn khảo sát chi phí cho học tập của học sinh tiểu học tại trường tiểu học Y. Chúng ta muốn biết trung bình chi phí cho học tập của một học sinh tiểu học là bao nhiêu. Gọi X là biến ngẫu nhiên ứng với chi phí cho học tập của một học sinh tiểu học (X tính bằng ngàn đồng/học sinh/tháng). Giả sử chúng ta biết phương sai của X là $\sigma_x^2=100$. Trung bình thực của X là μ là một số chưa biết. Chúng ta tìm cách ước lượng μ dựa trên một mẫu gồm n=100 học sinh được lựa chọn một cách ngẫu nhiên.

Hàm ước lượng cho μ

Chúng ta dùng giá trị trung bình mẫu $\overline{X}$ để ước lượng cho giá trị trung bình của tổng thể μ. Hàm ước lượng như sau

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$\bar{X}$ là một biến ngẫu nhiên. Ứng với một mẫu cụ thể thì $\bar{X}$ nhận một giá trị xác định.

Ước lượng điểm

Ứng với một mẫu cụ thể, giả sử chúng ta tính được $\bar{X}$ = 105 (ngàn đồng/học sinh). Đây là một ước lượng điểm.

Xác suất để một ước lượng điểm như trên đúng bằng trung bình thực là bao nhiêu? Rất thấp hay có thể nói hầu như bằng 0.

Ước lượng khoảng

Ước lượng khoảng cung cấp một khoảng giá trị có thể chứa giá trị chi phí trung bình cho học tập của một học sinh tiểu học. Ví dụ chúng ta tìm được $\bar{X} = 105$. Chúng ta có thể nói μcó thể nằm trong khoảng $\bar{X} \pm 10$ hay $95 \leq \mu \leq 115$.

Khoảng ước lượng càng rộng thì càng có khả năng chứa giá trị trung bình thực nhưng một khoảng ước lượng quá rộng như khoảng $\bar{X} \pm 100$ hay $5 \leq \mu \leq 205$ thì hầu như không giúp ích được gì cho chúng ta trong việc xác định μ. Như vậy có một sự đánh đổi trong ước lượng khoảng với cùng một phương pháp ước lượng nhất định: khoảng càng hẹp thì mức độ tin cậy càng nhỏ.

Phân phối của $\bar{X}$

Theo định lý giới hạn trung tâm 1 thì $\bar{X}$ là một biến ngẫu nhiên có phân phối chuẩn. Vì $\bar{X}$ có phân phối chuẩn nên chúng ta chỉ cần tìm hai đặc trưng của nó là kỳ vọng và phương sai.

Kỳ vọng của $\bar{X}$

$E\left(\bar{X}\right)$

$$= E\left(\frac{1}{n}(X_1 + X_2 + \_ + X_n)\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} * n\mu = \mu$$

Phương sai của $\bar{X}$

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} n\sigma_x^2 = \frac{\sigma_x^2}{n}$$

Vậy độ lệch chuẩn của $\bar{X}$ là

$$\frac{\sigma_x}{\sqrt{n}}$$

.

Từ thông tin này, áp dụng quy tắc 2 σ thì xác suất khoảng $\bar{X} \pm 2\frac{\sigma_x}{\sqrt{n}}$ chứa μ sẽ xấp xỉ 95%. Ước lượng khoảng với độ tin cậy 95% cho μ là

$$\bar{X} - 2\frac{\sigma_x}{\sqrt{n}} \le \mu \le \bar{X} + 2\frac{\sigma_x}{\sqrt{n}}$$
$$105 - 2\frac{10}{\sqrt{100}} \le \mu \le 105 + 2\frac{10}{\sqrt{100}}$$
$$\hat{\theta}_1 = 103 \le \mu \le 107 = \hat{\theta}_2$$

Lưu ý: Mặc dù về mặt kỹ thuật ta nói khoảng

$$\bar{X} \pm 2\frac{\sigma_x}{\sqrt{n}}$$

chứa μ với xác suất 95% nhưng không thể nói một khoảng cụ thể như (103; 107) có xác suất chứa μ là 95%. Khoảng (103;107) chỉ có thể hoặc chứa μ hoặc không chứa μ.

Ý nghĩa chính xác của độ tin cậy 95% cho ước lượng khoảng cho μ như sau: Với quy tắc xây dựng khoảng là

$$\bar{X} \pm 2\frac{\sigma_x}{\sqrt{n}}$$

và chúng ta tiến hành lấy một mẫu với cỡ mẫu n và tính được một khoảng ước lượng. Chúng ta cứ lặp đi lặp lại quá trình lấy mẫu và ước lượng khoảng như trên thì khoảng 95% khoảng ước lượng chúng ta tìm được sẽ chứa μ.

Tổng quát hơn, nếu trị thống kê cần ước lượng là

và ta tính được hai ước lượng $\hat{\theta}_1$ và $\hat{\theta}_2$ sao cho

$P(\hat{\theta}_1 \le \mu \le \hat{\theta}_1) = 1 - \alpha$ với $0 < \alpha < 1$

hay xác suất khoảng từ $\hat{\theta}_1$ đến $\hat{\theta}_2$ chứa giá trị thật θ là 1- α thì 1- α được gọi là độ tin cậy của ước lượng, α được gọi là mức ý nghĩa của ước lượng và cũng là xác suất mắc sai lầm loại I.

Nếu α= 5% thì 1- α là 95%. Mức ý nghĩa 5% hay độ tin cậy 95% thường được sử dụng trong thống kê và trong kinh tế lượng.

Các tính chất đáng mong đợi của một ước lượng được chia thành hai nhóm, nhóm tính chất của ước lượng trên cỡ mẫu nhỏ và nhóm tính chất ước lượng trên cỡ mẫu lớn.

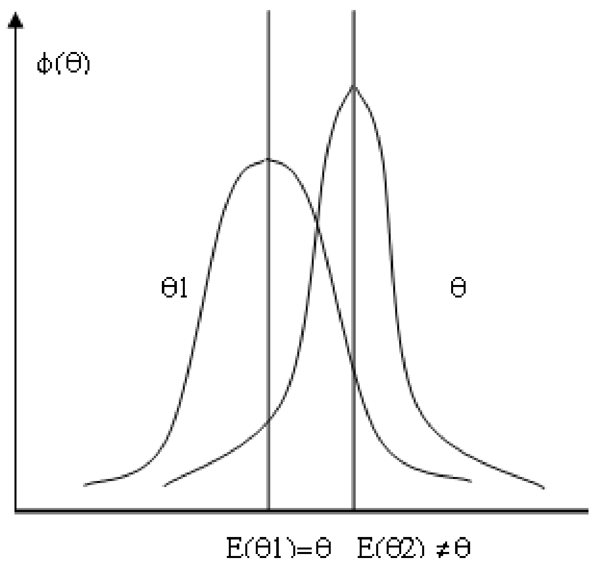
Các tính chất ứng với mẫu nhỏ

Không thiên lệch(không chệch)

Một ước lượng là không thiên lệch nếu kỳ vọng của $\hat{\theta}$ đúng bằng θ.

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

Như đã chứng minh ở phần trên, $\bar{X}$ là ước lượng không thiên lệch của μ.



Hình 2.4. Tính không thiên lệch của ước lượng.

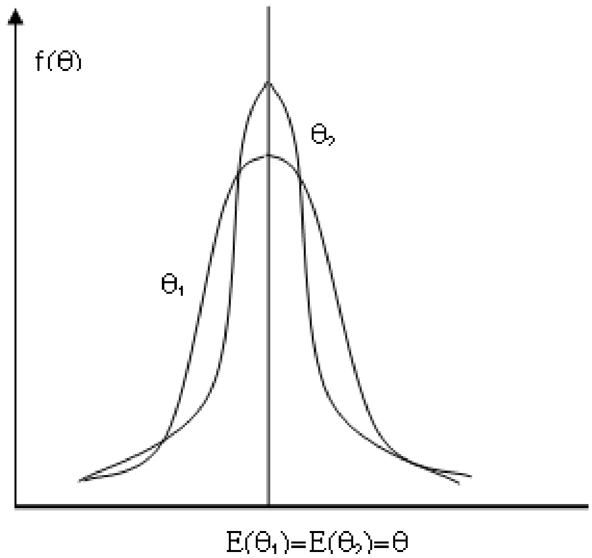
θ_1 là ước lượng không thiên lệch của μ trong khi θ_2 là ước lượng thiên lệch của μ.

Phương sai nhỏ nhất

Hàm ước lượng $\hat{\theta}_1$ có phương sai nhỏ nhất khi với bất cứ hàm ước lượng $\hat{\theta}_2$ nào ta cũng có $var(\hat{\theta}_1) \leq var(\hat{\theta}_2)$.

Không thiên lệch tốt nhất hay hiệu quả

Một ước lượng là hiệu quả nếu nó là ước lượng không thiên lệch và có phương sai nhỏ nhất.



Hình 2.5. Ước lượng hiệu quả. Hàm ước lượng θ_2 hiệu quả hơn θ_1.

Tuyến tính

Một ước lượng $\hat{\theta}$ của θ được gọi là ước lượng tuyến tính nếu nó là một hàm số tuyến tính của các quan sát mẫu.

Ta có

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \_ + X_n)$$

Vậy $\bar{X}$ là ước lượng tuyến tính cho μ.

Ước lượng không thiên lệch tuyến tính tốt nhất (Best Linear Unbiased Estimator-BLUE)

Một ước lượng $\hat{\theta}$ được gọi là BLUE nếu nó là ước lượng tuyến tính, không thiên lệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không thiên lệch của θ. Có thể chứng minh được $\bar{X}$ là BLUE.

Sai số bình phương trung bình nhỏ nhất

Sai số bình phương trung bình: $MSE(\hat{\theta})=E(\hat{\theta} - \theta)^2$

Sau khi biến đổi chúng ta nhận được: $MSE(\hat{\theta})=var(\hat{\theta})+E[E(\hat{\theta})- \theta]^2$

MSE($\hat{\theta}$)=var($\hat{\theta}$)+bias($\hat{\theta}$)

Sai số bình phương trung bình bằng phương sai của ước lượng cộng với thiên lệch của ước lượng. Chúng ta muốn ước lượng ít thiên lệch đồng thời có phương sai nhỏ. Người ta sử dụng tính chất sai số bình phương trung bình nhỏ khi không thể chọn ước lượng không thiên lệch tốt nhất.

Tính chất của mẫu lớn

Một số ước lượng không thoả mãn các tính chất thống kê mong muốn khi cỡ mẫu nhỏ nhưng khi cỡ mẫu lớn đến vô hạn thì lại có một số tính chất thống kê mong muốn. Các tính chất thống kê này được gọi là tính chất của mẫu lớn hay tính tiệm cận.

Tính không thiên lệch tiệm cận

Ước lượng $\hat{\theta}$ được gọi là không thiên lệch tiệm cận của θ nếu $\lim_{n \to \infty} E(\hat{\theta}_n) = \theta$

Ví dụ 2.12. Xét phương sai mẫu của biến ngẫu nhiên X:

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

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}$$

Có thể chứng minh được

$$E[s_x^2] = \sigma_x^2$$

$$E[\hat{\sigma}_x^2] = \sigma_x^2\left(1 - \frac{1}{n}\right)$$

Vậy s_x^2 là ước lượng không thiên lệch của σ_x^2, trong khi $\hat{\sigma}_x^2$ là ước lượng không thiên lệch tiệm cận của σ_x^2.

Nhất quán

Một ước lượng

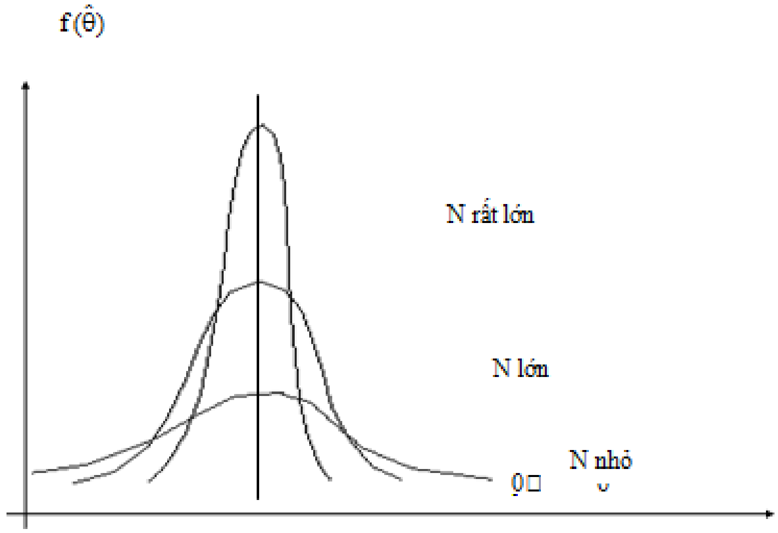
được gọi là nhất quán nếu xác suất nếu nó tiến đến giá trị đúng của

khi cỡ mẫu ngày càng lớn.

là nhất quán thì

$$\lim_{n\to\infty}\left\{\hat{\theta}-\theta\right|<\delta\}=1$$

với 1 là một số dương nhỏ tuỳ ý.



Hình 2.6. Ước lượng nhất quán

Quy luật chuẩn tiệm cận

Một ước lượng

được gọi là phân phối chuẩn tiệm cận khi phân phối mẫu của nó tiến đến phân phối chuẩn khi cỡ mẫu *n* tiến đến vô cùng.

Trong phần trên chúng ta đã thấy biến X có phân phối chuẩn với trung bình μ và phương sai σ^2 thì $\bar{X}$ có phân phối chuẩn với trung bình μ và phương sai σ^2/n với cả cỡ mẫu nhỏ và lớn.

Nếu X là biến ngẫu nhiên có trung bình μ và phương sai σ^2 nhưng không theo phân phân phối chuẩn thì cũng sẽ có phân phối chuẩn với trung bình μ và phương sai σ^2/n khi n tiến đến vô cùng. Đây chính là định lý giới hạn trung tâm 2.