Thống kê suy diễn 2

Bởi:

Phạm Trí Cao

Thống kê suy diễn - Kiểm định giả thiết thống kê

Giả thiết

Giả thiết không là một phát biểu về giá trị của tham số hoặc về giá trị của một tập hợp các tham số. Giả thiết ngược phát biểu về giá trị của tham số hoặc một tập hợp tham số khi giả thiết không sai. Giả thiết không thường được ký hiệu là H_0 và giả thiết ngược thường được ký hiệu là H_1.

Kiểm định hai đuôi

Ví dụ 13. Quay lại ví dụ 11 về biến X là chi phí cho học tập của học sinh tiểu học. Chúng ta biết phương sai của X là σ_x^2=100. Với một mẫu với cỡ mẫu n=100 chúng ta đã tính được $\overline{X}_1$=105 ngàn đồng/học sinh/tháng. Chúng ta xem xét khả năng bác bỏ phát biểu cho rằng chi phí cho học tập trung bình của học sinh tiểu học là 106 ngàn đồng/tháng.

Giả thiết

H_0: $\mu = 106 = \mu_0$

H_1: μ $106 = \mu_0$

Chúng ta đã biết $\overline{X}$~N(μ, σ_x^2/n), với độ tin cậy 95% hay mức ý nghĩa a = 5% chúng ta đã xây dựng được ước lượng khoảng của μ là

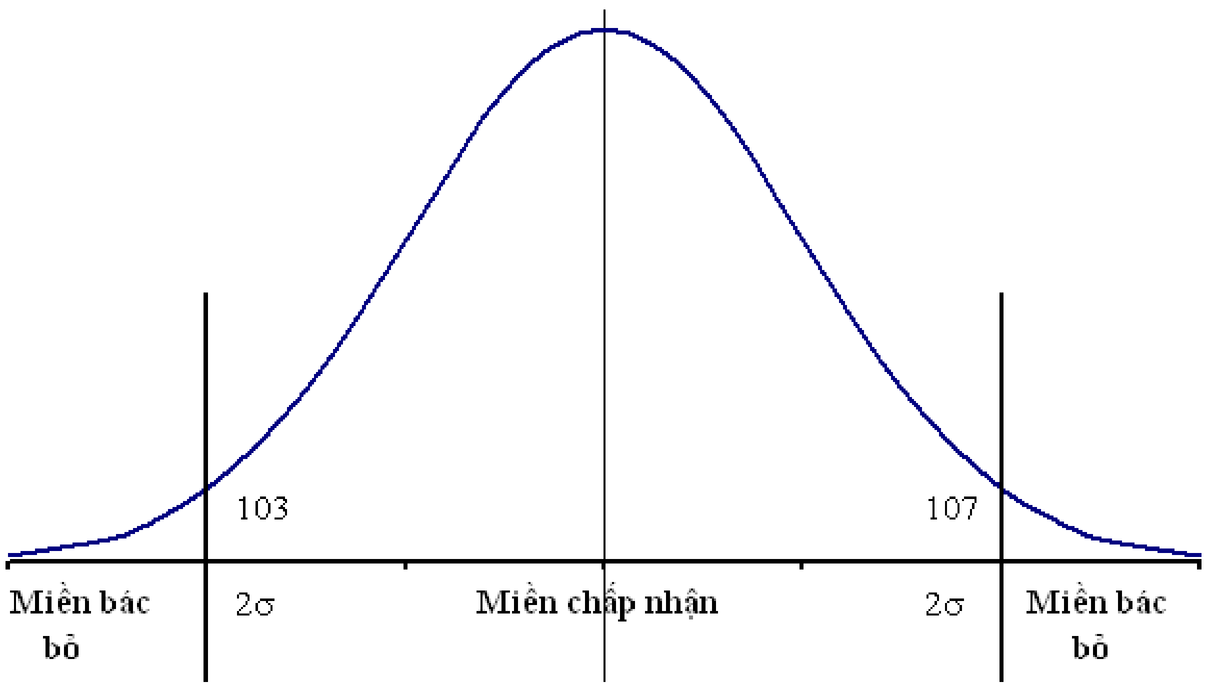
$$\overline{X}_1 \pm 2\frac{\sigma_x}{\sqrt{n}}$$

.

Nếu khoảng này không chứa μ thì ta bác bỏ giả thiết không với độ tin cậy 95%, ngược lại ta không đủ cơ sở để bác bỏ giả thiết H_0.

Ở phần trên chúng ta đã tính được ước lượng khoảng của μ dựa theo $\bar{X}_1$ là (103;107). Khoảng này chứa $\mu_0 = 106$. Vậy ta không thể bác bỏ được giả thiết H_0.

Khoảng tin cậy mà ta thiết lập được được gọi là miền chấp nhận, miền giá trị nằm ngoài miền chấp nhận được gọi là miền bác bỏ.



Hình 2.7. Miền bác bỏ và miền chấp nhận H_0.

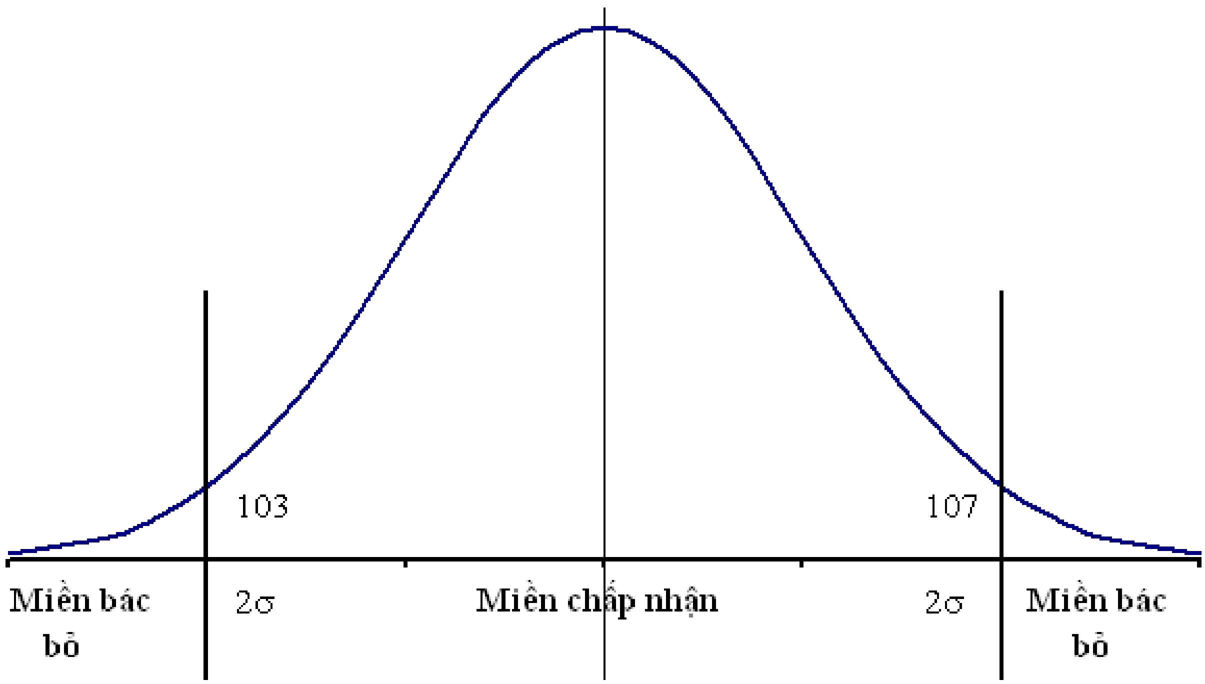
Tổng quát hơn ta có

Z=

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

~N(0,1) hay Z tuân theo phân phối chuẩn hoá.

α/2α/2



Hình 2.8. Miền chấp nhận và miền bác bỏ theo αcủa trị thống kê Z

Ta có tất cả hai miền bác bỏ và do tính chất đối xứng của phân phối chuẩn, nếu mức ý nghĩa là αthì xác suất để Z nằm ở miền bác bỏ bên trái là α/2 và xác suất để Z nằm ở miền bác bỏ bên trái cũng là α/2. Chúng ta đặt giá trị tới hạn bên trái là $Z_{\alpha/2}$ và giá trị tới hạn bên phải là $Z_{1-\alpha/2}$. Do tính đối xứng ta lại có

$$Z_{\alpha/2} = -Z_{1-\alpha/2}.$$

Xác suất để Z nằm trong hai khoảng tới hạn là

$$\mathbf{P(Z_{\alpha/2} \leq Z \leq Z_{1-\alpha/2}) = 1 - \alpha}$$

hay

$$\mathbf{P(-Z_{1-\alpha/2} \leq Z \leq Z_{1-\alpha/2}) = 1 - \alpha}$$

Thay

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

và biến đổi một chút chúng ta nhận được

$$P\left(\overline{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \ (2)$$

Các mệnh đề (2.1) và (2.2) là những mệnh đề xác suất.

Kiểm định giả thiết thống kê theo phương pháp truyền thống

Phát biểu mệnh đề xác suất

$$P\left(\overline{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \middle| \mu = \mu_0\right) = 1-\alpha$$

Nguyên tắc ra quyết định

Nếu

$$\overline{X}_1 - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu_0$$

hoặc

$$\overline{X}_1 + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0$$

thì ta bác bỏ H_0 với độ tin cậy 1- α hay xác suất mắc sai lầm là α.

Nếu

$$\overline{X}_1 - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \overline{X}_1 + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

thì ta không thể bác bỏ H_0.

Với mức ý nghĩa

X=5% thì $Z_{1-1/2} = Z_{97,5\%} = 1,96$ ˜ 2

Ta có

$$\overline{X}_1 - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 105 - 2\frac{10}{10} = 103$$

$$\bar{X}_1 + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 105 + 2\frac{10}{10} = 107$$

Vậy ta không thể bác bỏ giả thiết Ho.

Kiểm định giả thiết thống kê theo trị thống kê Z

Phát biểu mệnh đề xác suất

$$P(Z_{\alpha/2} \le Z \le Z_{1-\alpha/2}) = 1 - \alpha$$

Quy tắc quyết định

Nếu

$$Z_{tt} = \frac{X_1 - \mu_0}{\sigma^2/\sqrt{n}} < Z\alpha_{/2}$$

hoặc

$$Z_{tt} = \frac{X_1 - \mu_0}{\sigma/\sqrt{n}} > Z_{1-}\alpha_{/2}$$

thì ta bác bỏ H_0 với độ tin cậy 1- α hay xác suất mắc sai lầm là α.

Nếu $Z_{\alpha/2} \le Z_{tt} \le Z_{1-\alpha/2}$ thì ta không thể bác bỏ H_0.

Với mức ý nghĩa α =5% ta có

$Z_{1-\alpha/2} = Z_{97,5\%} = 1,96 \approx 2$

và $Z_{\alpha/2} = Z_{2,5\%} = -1,96 \approx -2$

$$Z_{tt} = \frac{X_1 - \mu_0}{\sigma/\sqrt{n}} = \frac{105 - 106}{10/\sqrt{100}} = -1$$

Vậy ta không thể bác bỏ Ho.

Kiểm định giả thiết thống kê theo giá trị p

Đối với kiểm định hai đuôi giá trị p được tính như sau:

$$p = 2P(|Z_{tt}| < Z)$$

Với Ztt = -1 ta có P(1<Z) = 0,16, vậy giá trị p = 0,32.

Quy tắc quyết định

Nếu p α : Bác bỏ Ho.

Nếu p ≥ α : Không thể bác bỏ Ho.

Trong ví dụ trên p = 0,32 > α = 5%. Vậy ta không thể bác bỏ Ho.

Ba cách tiếp cận trên cho cùng một kết quả vì thực ra chỉ từ những biến đổi của cùng một mệnh đề xác suất. Trong kinh tế lượng người ta cũng thường hay sử dụng giá trị p.

Kiểm định một đuôi

Kiểm định đuôi trái

Ví dụ 14. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi cho học tập trung bình của học sinh tiểu học lớn hơn 108 ngàn đồng/học sinh/tháng".

Giả thiết

H_0: $\mu > 108 = \mu_0$

H_1: $\mu \leq 108 = \mu_0$

Phát biểu mệnh đề xác suất

P(Z α<Z) =1- α

Quy tắc quyết định

Nếu $Z_{tt} < Z\ \alpha$: Bác bỏ Ho.

Nếu $Z_{tt} \geq Z\ \alpha$: Không thể bác bỏ Ho.

Với α = 5% ta có $Z_{5\%} = -1{,}644$

Ta có

$$Ztt = \frac{\overline{X}_1 - \mu_0}{\sigma / \sqrt{n}} = \frac{105-108}{10/\sqrt{100}} = -3 < Z_{5\%} = -1{,}644$$

vậy ta bác bỏ Ho.

Kiểm định đuôi phải

Ví dụ 15. Tiếp tục ví dụ 13. Kiểm định phát biểu : “Chi tiêu cho học tập trung bình của học sinh tiểu học nhỏ hơn 108 ngàn đồng/học sinh/tháng”.

Giả thiết

H_0: $\mu < 107 = \mu_0$

H_1: $\mu \geq 107 = \mu_0$

Phát biểu mệnh đề xác suất

$P(Z<Z_{1-\alpha}) = 1-\alpha$

Quy tắc quyết định

Nếu $Z_{tt} > Z\ \alpha$: Bác bỏ Ho.

Nếu $Z_{tt} \leq Z\ \alpha$: Không thể bác bỏ Ho.

Ta có

$$Ztt = \frac{\overline{X}_1 - \mu_0}{\sigma / \sqrt{n}} = \frac{105-108}{10/\sqrt{100}} = -3 < Z_{5\%} = -1{,}644$$

vậy ta không thể bác bỏ Ho.

Một số trường hợp đặc biệt cho ước lượng giá trị trung bình của tổng thể

Tổng thể có phân phối chuẩn, cỡ mẫu lớn, phương sai chưa biết. Chiến lược kiểm định giống như trên nhưng thay phương sai tổng thể bằng phương sai mẫu.

Tổng thể có phân phối chuẩn, phương sai chưa biết, cỡ mẫu nhỏ:

$$\frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim \text{t-stat} \sim t_{(n-1)}$$

Kiểm định trên trị thống kê t cũng tương tự như đối với trị thống kê Z, ta chỉ việc tra t thay cho Z. Khi cỡ mẫu đủ lớn trị thống kê t tương tự trị thống kê Z.

Tổng thể không tuân theo phân phối chuẩn, áp dụng định lý giới hạn trung tâm. Khi cỡ mẫu đủ lớn thì trị thống kê t tính toán như phần trên có phân phối gần với phân phối Z.

Ngoài ra chúng ta còn có thể kiểm định các giả thiết về phương sai, kiểm định sự bằng nhau giữa các phương sai của hai tổng thể và kiểm định sự bằng nhau giữa các trung bình tổng thể. Chúng ta xét kiểm định giả thiết về phương sai vì giả định về phương sai không đổi là một giả định quan trọng trong phân tích hồi quy.

Kiểm định giả thiết về phương sai

Xét giả thiết

Ho : $\sigma^2 = \sigma_0^2$

H1 : $\sigma^2 \neq \sigma_0^2$

Có thể chứng minh được

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Mệnh đề xác suất

$$P\left(\chi^2_{(n-1,\alpha/2)} \leq n-1)\frac{s^2}{\sigma_0^2} \leq \chi^2_{(n-1,1-\alpha/2)}\right) = 1-\alpha$$

Quy tắc quyết định

Nếu

$$(n-1)\frac{s^2}{\sigma_0^2} < \chi^2_{(n-1,\alpha/2)}$$

hoặc

$$(n-1)\frac{s^2}{\sigma_0^2} > \chi^2_{(n-1,\alpha/2)}$$

, thì bác bỏ H_0.

Nếu

$$\chi^2_{(n-1,\alpha/2)} \leq n-1)\frac{s^2}{\sigma_0^2} \leq \chi^2_{(n-1,1-\alpha/2)}$$

, thì không bác bỏ H_0.

Kiểm định sự bằng nhau của phương sai hai tổng thể

Chúng ta có mẫu cỡ n_1 từ tổng thể 1 và mẫu cỡ n_2 từ tổng thể 2.

Xét giả thiết

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$

$H_1 : \sigma_1^2\sigma_2^2$

Chúng ta đã có

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Vậy

$$\frac{(n_1-1)\frac{s_1^2}{\sigma^2} \Big/ (n_1-1)}{(n_2-1)\frac{s_2^2}{\sigma^2} \Big/ (n_2-1)} \sim \frac{\chi^2_{(n_1-1)} \Big/ (n_1-1)}{\chi^2_{(n_2-1)} \Big/ (n_2-1)} \sim F_{(n_1-1,n_2-1)}$$

Hay

$$\frac{s_1^2}{s_2^2} \sim F_{(n_1-1,n_2-1)}$$

Phát biểu mệnh đề xác suất

$$P\left(F_{(n_1-1,n_2-1,\alpha/2)} \le \frac{s_1^2}{s_2^2} \le F_{(n_1-1,n_2-1,1-\alpha/2)}\right) = 1-\alpha$$

Quy tắc quyết định

Nếu

$$\frac{s_1^2}{s_2^2} < F_{(n_1-1,n_2-1,\alpha/2)} \text{ ho?c } \frac{s_1^2}{s_2^2} > F_{(n_1-1,n_2-1,1-\alpha/2)}$$

thì ta bác bỏ H_0.

Nếu

$$F_{(n_1-1,n_2-1,\alpha/2)} \le \frac{s_1^2}{s_2^2} \le F_{(n_1-1,n_2-1,1-\alpha/2)}$$

thì không bác bỏ H_0.

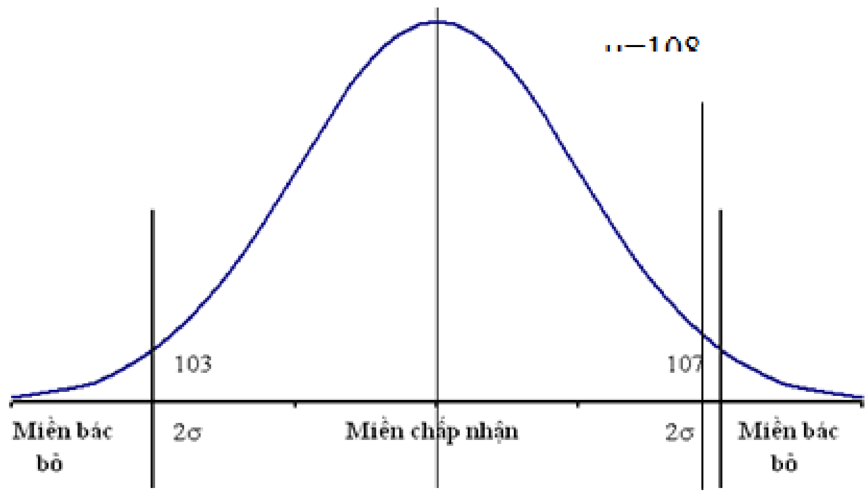
2.4.5. Sai lầm loại I và sai lầm loại II

Khi ta dựa vào một mẫu để bác bỏ một giả thiết, ta có thể mắc phải một trong hai sai lầm như sau:

Sai lầm loại I: Bác bỏ Ho khi thực tế Ho đúng.

Sai lầm loại II : Không bác bỏ Ho khi thực tế nó sai.

	Tính chất	
Quyết định	H_0 đúng	H_0 sai
Bác bỏ	Sai lầm loại I	Không mắc sai lầm
Không bác bỏ	Không mắc sai lầm	Sai lầm loại II



Hình 2.7. Sai lầm loại I-Bác bỏ H_0: $_{H0}$=108 trong khi thực tế H_0 đúng.

Xác suất mắc sai lầm loại I

Ví dụ 16. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi cho học tập trung bình của học sinh tiểu học là 108 ngàn đồng/học sinh/tháng". Trung bình thực $\mu = \mu_0$=108.

Giả thiết

H_0: $\mu = 108 = \mu_0$

H_1: $\mu \neq 108 = \mu_0$

Giả sử giá trị μthực là μ=108. Với ước lượng khoảng cho μ là (103;107) với độ tin cậy 95% chúng ta bác bỏ H_0 trong khi thực sự H_0 là đúng. Xác suất chúng ta mắc sai lầm loại này là μ = 5%.

Xác suất mắc sai lầm loại II

Ví dụ 17. Tiếp tục ví dụ 13. Kiểm định phát biểu : "Chi tiêu cho học tập trung bình của học sinh tiểu học là 108 ngàn đồng/học sinh/tháng". Trung bình thực $\mu = \mu_0$=104.

Giả thiết

H_0: $\mu = 108 = \mu_0$

H_1: $\mu \neq 108 = \mu_0$

Giả sử giá trị μ thực là μ=104. Với ước lượng khoảng cho μ là (103;107) với độ tin cậy 95% chúng ta không bác bỏ H_0 trong khi H_0 sai. Xác suất chúng ta mắc sai lầm loại II này là μ

Lý tưởng nhất là chúng ta tối thiểu hoá cả hai loại sai lầm. Nhưng nếu chúng ta muốn hạn chế sai lầm loại I, tức là chọn mức ý nghĩa μ nhỏ thì khoảng ước lượng càng lớn và xác suất mắc phải sai lầm loại II càng lớn. Nghiên cứu của Newman và Pearson

Damodar N. Gujarati, Basic Econometrics-Third Edition, McGraw-Hill Inc -1995, p 787.

cho rằng sai lầm loại I là nghiêm trọng hơn sai lầm loại II. Do đó, trong thống kê suy diễn cổ điển cũng như trong kinh tế lượng cổ điển, người ta chọn mức ý nghĩa μ hay xác suất mắc sai lầm loại I nhỏ, thông thường nhất là 5% mà không quan tâm nhiều đến μ.

Tóm tắt các bước của kiểm định giả thiết thống kê

Bước 1.Phát biểu giả thiết H_0 và giả thiết ngược H_1.

Bước 2. Lựa chọn trị thống kê kiểm định

Bước 3. Xác định phân phối thống kê của kiểm định

Bước 4. Lựa chọn mức ý nghĩa μhay xác suất mắc sai lầm loại I.

Bước 5. Sử dụng phân phối xác suất của thống kê kiểm định, thiết lập một khoảng tin cậy 1- μ, khoảng này còn được gọi là miền chấp nhận. Nếu trị thống kê ứng với H_0 nằm trong miền chấp nhận thì ta không bác bỏ H_0, nếu trị thông kê ứng với H_0 nằm ngoài miền chấp nhận thì ta bác bỏ H_0. Lưu ý là khi bác bỏ H_0 chúng ta chấp nhận mức độ sai lầm là μ.