THỐNG KÊ ỨNG DỤNG

Giảng viên: TS. Bùi Thanh Hùng
Bộ môn Khoa học dữ liệu, Khoa Công nghệ thông tin
Đại học Công nghiệp thành phố Hồ Chí Minh
Email: buithanhhung@iuh.edu.vn
Website: https://sites.google.com/site/hungthanhbui1980/

Mỗi bài tập đều làm theo 2 cách: Tự tính bằng tay và viết code cho máy tính tính và trực quan hóa kết quả đó.

Bài 1:

Nhận thức được rằng phần thưởng cho việc thi hành nghĩa vụ pháp lý của tòa án thay đổi theo thời gian, một công ty bảo hiểm muốn so sánh mức trung bình của phần thưởng cho việc thi hành nghĩa vụ pháp lý cá nhân hiện hành với mức của một năm trước đó. Một mẫu ngẫu nhiên gồm $n = 30$ vụ kiện được chọn lựa trong số các vụ kiện được phân xử trong từng năm trong số hai thời kỳ hàng năm này. Các số trung bình và phương sai mẫu của các phần thưởng cho việc thi hành nghĩa vụ pháp lý (tính bằng triệu đôla Mỹ) cho mỗi trong số hai năm này được cho trong Bảng 7.4.

a. Hãy tìm một ước lượng điểm cho chênh lệch trong mức trung bình về phần thưởng cho việc thi hành nghĩa vụ pháp lý giữa năm hiện tại và năm trước đó. Cho biết biên sai số.

b. Tìm một khoảng tin cậy 90% cho chênh lệch trong mức trung bình về phần thưởng cho việc thi hành nghĩa vụ pháp lý giữa năm hiện tại và năm trước đó.

Các số trung bình và phương sai mẫu

Năm	Cỡ Mẫu	Trung bình Mẫu (triệu $)	Phương sai Mẫu (triệu $)^2
Hiện tại	$n_1 = 30$	$\bar{x}_1 = 1.32$	$s_1^2 = 0.9734$
Trước đó	$n_2 = 30$	$\bar{x}_2 = 1.04$	$s_2^2 = 0.7291$

Bài 2:

Một cửa hàng thực phẩm nhận thấy thời gian vừa qua trung bình một khách hàng mua 25 nghìn đồng thực phẩm trong ngày. Nay cửa hàng chọn ngẫu nhiên 15 khách hàng thấy trung bình một khách hàng mua 24 nghìn đồng trong ngày và phương sai mẫu điều chỉnh là $s^2 = (2 \text{ nghìn đồng})^2$. Với mức ý nghĩa là 5% , thử xem có phải sức mua của khách hàng hiện nay thực sự giảm sút.

Bài 3:

Giám đốc một xí nghiệp cho biết lương trung bình của một công nhân thuộc xí nghiệp là 380 nghìn đồng/ tháng. Chọn ngẫu nhiên 36 công nhân thấy lương trung bình là 350 nghìn đồng/ tháng, với độ lệch chuẩn $\sigma = 40$ nghìn. Lời báo cáo của giám đốc có tin cậy được không, với mức ý nghĩa là 5%.

Bài 4:

Một công ty muốn dự đoán số phút gọi vào tổng đài dịch vụ của khách dựa vào số linh kiện cần sửa chữa, hãy xây dựng mô hình hồi quy theo mô hình sau:

$$\text{Minutes} = \beta_0 + \beta_1 \text{Units} + \varepsilon,$$

Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

GIẢ THUYẾT

Giả thuyết là một phát biểu về một thuộc tính của một quần thể.
Một kiểm định giả thuyết thống kê là một thủ tục để kiểm định một phát biểu về một thuộc tính của một quần thể.

Giả thuyết null

- Giả thuyết null (ký hiệu bởi H_0) là một phát biểu rằng giá trị của một tham số quần thể (chẳng hạn như tỷ lệ, trung bình,…) bằng với một số giá trị được phát biểu.
- Chúng ta kiểm định giả thuyết null trực tiếp theo nghĩa là chúng ta giả định nó là đúng và đi đến một kết luận hoặc bác bỏ H_0 hoặc không bác bỏ H_0.

Giả thuyết thay thế

- Giả thuyết thay thế (được ký hiệu bởi H_1 hay H_A) là phát biểu rằng tham số có giá trị nào đó khác với giả thuyết null.
- Ký hiệu của giả thuyết thay thế phải sử dụng một trong các ký hiệu này: <, >, ≠.

Giả thuyết đuôi trái, phải, hai đuôi

- Giả thuyết đuôi trái: liên quan đến giả thuyết tham số quần thể nhỏ hơn một giá trị.
  H_0: m=10, H_1: m < 10
- Giả thuyết đuôi phải: liên quan đến giả thuyết tham số quần thể lớn hơn một giá trị.
  H_0: m=10, H_1: m > 10
- Giả thuyết hai đuôi: liên quan đến giả thuyết tham số quần thể không bằng một giá trị.
  H_0: m=10, H_1: m ≠ 10

Sai lầm loại I là bác bỏ giả thuyết null khi nó thực sự đúng

- Ký hiệu *α (alpha)* được dùng để biểu diễn cho xác suất mắc sai lầm loại I.

Sai lầm loại II là từ chối bác bỏ giả thuyết null khi nó thực sự sai

- Ký hiệu *β (beta)* được dùng để biểu diễn cho xác suất mắc sai lầm loại II.

		True State of Nature	
		The null hypothesis is true	**The null hypothesis is false**
Decision	We decide to reject the null hypothesis	**Type I error** (rejecting a true null hypothesis) $P(\text{type I error}) = \alpha$	Correct decision
	We fail to reject the null hypothesis	Correct decision	**Type II error** (failing to reject a false null hypothesis) $P(\text{type II error}) = \beta$

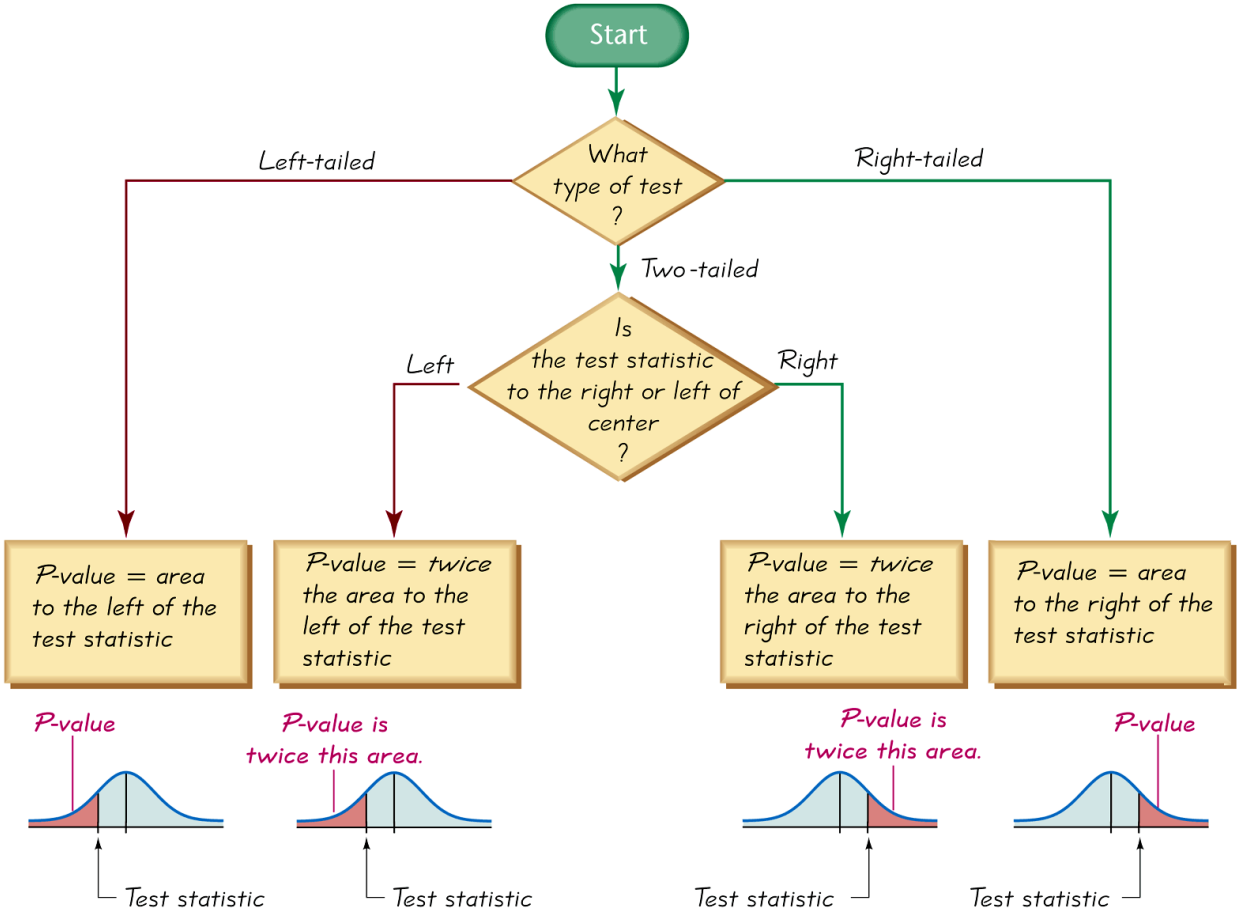
Thủ tục kiểm định

1. Tính giá trị thống kê (test statistic) dùng để kiểm định
2. Dựa vào loại phân phối của biến ngẫn nhiên là giá trị thống kê dùng để kiểm định để tính trị số p (p-value)
3. So sánh trị số p với mức có ý nghĩa α
    - Nếu p-value< α: bác bỏ H_0
    - Nếu p-value> α: chấp nhận H_0
    - Nếu p-value= α: tùy người thực hiện kiểm định

Mức có ý nghĩa

Mức có ý nghĩa (được biểu thị bởi α) là xác suất của số liệu thống kê (kiểm định giả thuyết) sẽ rơi vào vùng critical khi giả thuyết null thực sự là đúng (làm cho sai lầm của việc bác bỏ giả thuyết null khi nó là đúng).
Nó cũng giống như α đã giới thiệu trong chương 7.
Sự lựa chọn phổ biến cho α là 0,05, 0,01, và 0,10.

Thủ tục tìm giá trị P

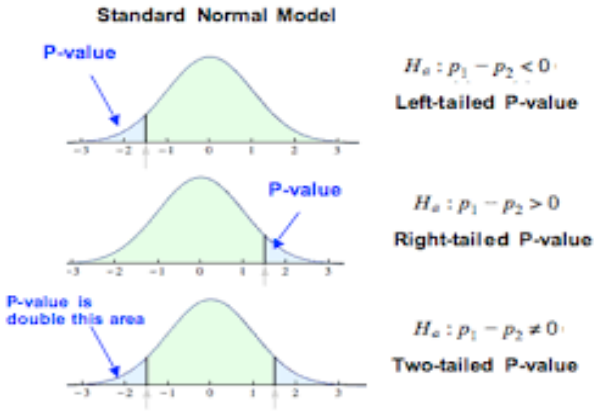


Giá trị P (hoặc giá trị xác suất) là xác suất nhận được một giá trị thống kê để kiểm định ít nhất là cực đại với giá trị đại diện cho dữ liệu mẫu, giả thiết rằng giả thiết null là đúng.

Vùng critical ở đuôi trái

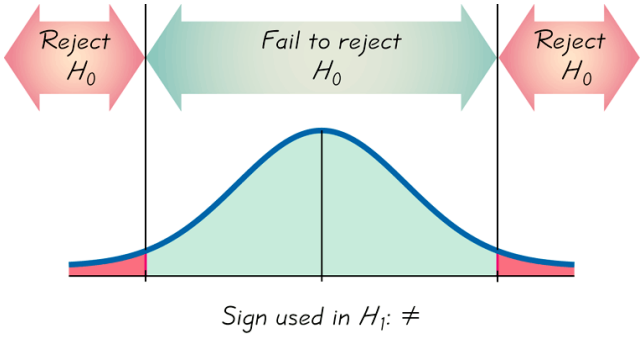
Vùng critical ở đuôi phải

Vùng critical ở hai đuôi



Các loại kiểm định giả thuyết: Hai đuôi, đuôi bên trái, đuôi phải
Các đuôi trong phân phối là các vùng cực đại bị giới hạn bởi các critical values.
Việc xác định giá trị P và các critical value bị ảnh hưởng bởi vùng critical ở hai đuôi, đuôi trái hay đuôi phải.

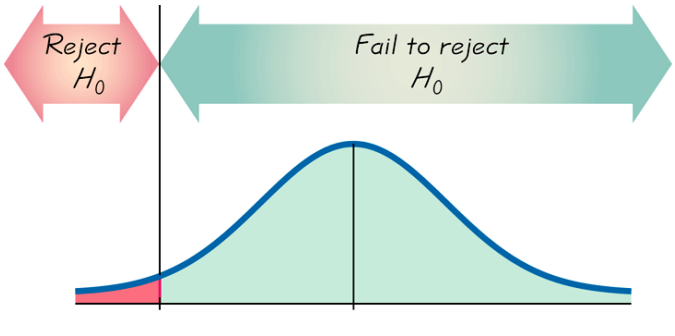
Kiểm định 2 đuôi



$H_0 :=$

$H_1 :\neq$

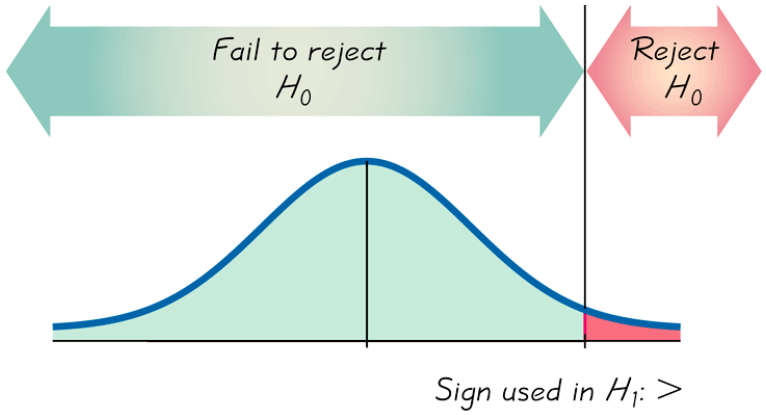
α được chia đều giữa hai đuôi của vùng critical



$H_0 :=$

$H_1 :<$

Kiểm định đuôi trái: α nằm ở đuôi trái



H_0 :=

H_1 :> ***Kiểm định đuôi phải: α nằm ở đuôi phải***

Trường hợp phương sai biết trước hay mẫu lớn (n>30)

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

Giá trị thống kê:

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\frac{\sigma}{\sqrt{n}}}$$

Phân phối chuẩn chính tắc

Trường hợp mẫu lớn (n>30) thì phương sai của mẫu s^2 có thể được xem như là phương sai của quần thể σ^2

Phương pháp phổ biến để kiểm định một phát biểu về tỷ lệ quần thể là sử dụng phân phối chuẩn xấp xỉ với phân phối nhị thức

n = cỡ mẫu hoặc số thử nghiệm

$\hat{p} = \frac{x}{n}$

p = tỉ lệ quần thể

q = $1 - p$

Yêu cầu đối với kiểm định phát biểu về tỷ lệ quần thể p

1. Các quan sát mẫu là một mẫu ngẫu nhiên đơn giản.
2. Các điều kiện cho phân phối nhị thức được thỏa mãn.
3. Các điều kiện $np \geq 5$ và $nq \geq 5$ đều thỏa mãn, vì vậy sự phân phối nhị thức của tỷ lệ mẫu có thể xấp xỉ bằng phân phối chuẩn.
4. Lưu ý: p là tỷ lệ giả định chứ không phải tỷ lệ mẫu.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Đừng nhầm lẫn giữa P-value với tỷ lệ p

- *P*-value = là xác suất nhận được một giá trị thống kê để kiểm định ít nhất là cực đại với giá trị đại diện cho dữ liệu mẫu.
- *p* = tỉ lệ quần thể

COVARIANCE VÀ CORRELATION COEFFICIENT

$$\mathrm{Cov}(Y,X) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

$$\begin{aligned}\mathrm{Cor}(Y,X) &= \frac{\mathrm{Cov}(Y,X)}{s_y s_x} \\ &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(x_i - \bar{x})^2}}\end{aligned}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$-1 \leq \mathrm{Cor}(Y,X) \leq 1$$

***Covariance** giữa Y* và X thể hiện hướng **của mối quan hệ tuyến tính giữa Y và X**.
Cov(Y, X) **không cho ta biết độ mạnh** của mối quan hệ giữa Y và X
Cor(Y,X): cho ta biết hướng và độ mạnh mối quan hệ giữa X và Y

Hồi quy tuyến tính đơn biến

- Mối quan hệ giữa biến phản hồi Y và biến dự đoán X được quy định bởi mô hình tuyến tính:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Trong đó β_0 và β_1*: hệ số tương quan hồi quy của mô hình hay còn gọi là các tham số.*

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, 2, \cdots, n$$

The sum of squares of these distances can then be written as

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $S(\beta_0, \beta_1)$ are given by

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- H_0: $\beta_1 = \beta_1^0$ (β_1^0 : constant chosen by investigator)
- H_1: $\beta_1 \# \beta_1^0$
- t-Test

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)} \qquad \text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}},$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

- H_0 bị bác bỏ tại mức có ý nghĩa α nếu:

$$|t_1| \geq t_{(n-2,\alpha/2)}$$

hoặc:

$$p(|t_1|) \leq \alpha.$$

Phương trình hồi quy đơn giản có thể được sử dụng để dự đoán giá trị của biến phản hồi (response) bằng các giá trị cụ thể của biến dự báo (predictor)

Giá trị dự báo là $\hat{y}_0$ tương ứng với x_0 theo công thức sau:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Khoảng tin cậy cho $\hat{y}_0$ với hệ số tin cậy 1-α là:

$$\hat{y}_0 \pm t_{(n-2,\alpha/2)}\, \text{s.e.}(\hat{y}_0)$$