

Xác suất thống kê ứng dụng

Nguyễn Ngọc Tứ

Bài giảng 7

- Kiểm định giả thuyết dựa trên một mẫu đơn
- Kiểm định trung bình
- Kiểm định tỷ lệ
- Kiểm định 2 trung bình
- Kiểm định 2 tỷ lệ
- Kiểm định t cặp
- Hồi quy tuyến tính

Khái niệm về giả thiết thống kê

Một **giả thiết thống kê** là một khẳng định về giá trị của một tham số đơn (đặc trưng cho một phân phối xác suất) hoặc về hình dạng của phân phối xác suất.

Ví dụ:

- Giả thiết $\mu = 6.5$, với μ là giá trị trung bình của điểm XSTK.
- Giả thiết $p < 0.1$, với p là tỉ lệ sinh viên rớt XSTK.

Định nghĩa.

Giả thiết không, kí hiệu H_0 , là phát biểu ban đầu và giả sử giả thiết này đúng.

Giả thiết thay thế hay **giả thiết đối**, kí hiệu H_a , là khẳng định ngược lại với giả thiết H_0 .

Giả thiết không H_0 bị bác bỏ bởi giả thiết đối nếu phân tích mẫu dữ liệu chỉ ra H_0 sai.

Có hai kết luận từ bài toán kiểm định giả thiết là *bác bỏ H_0* hay *chấp nhận H_0* .

Khái niệm về giả thiết thống kê

Kiểm định giả thiết là phương pháp mà sử dụng dữ liệu mẫu để kết luận có từ chối H_0 hay không.

Cách kiểm định bài toán như sau:

1. (Tính) một đại lượng **thống kê kiểm định** là một hàm số của mẫu dữ liệu mà dựa vào đó ta để quyết định bác bỏ H_0 hay chấp nhận H_0 .
2. (Tìm) **Miền bác bỏ** là tập hợp tất cả các giá trị thống kê kiểm định để H_0 bị bác bỏ.

Giả thiết H_0 bị bác bỏ khi các giá trị thống kê kiểm định nằm trong miền bác bỏ. Có hai loại sai lầm thường gặp khi kiểm định giả thiết:

- 1 **Sai lầm loại I:** bác bỏ H_0 khi H_0 đúng.
- 2 **Sai lầm loại II:** không bác bỏ H_0 khi H_0 sai.

Kiểm định trung bình với σ đã biết

Cho $X_1, \dots, X_n$ là những mẫu ngẫu nhiên có phân phối chuẩn với kì vọng μ và phương sai σ^2 . Khi đó,

$$\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \sigma^2/n)$$

Đặt giả thiết $H_0 : \mu = \mu_0$

Tính thống kê kiểm định: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Giả thiết đối	Miền bác bỏ cho kiểm định mức α
$H_a : \mu > \mu_0$	$ z \geq z_\alpha$ (kiểm định bên phải)
$H_a : \mu < \mu_0$	$ z \geq z_\alpha$ (kiểm định bên trái)
$H_a : \mu \neq \mu_0$	$ z \geq z_{\alpha/2}$ (kiểm định hai bên)

Lưu ý: α được gọi là mức ý nghĩa.

Kiểm định trung bình với σ đã biết

Ví dụ. Nhà sản xuất hệ thống vòi phun nước để chống cháy trong chung cư cho rằng nhiệt độ trung bình để hệ thống kích hoạt là $130^{\circ}F$. Một mẫu $n = 9$ hệ thống, khi kiểm tra thấy nhiệt độ trung bình kích hoạt là $131.08^{\circ}F$. Nếu phân phối của thời gian hoạt động là phân phối chuẩn với độ lệch chuẩn $1.5^{\circ}F$. Dữ liệu có mâu thuẫn với phát biểu của nhà sản xuất tại mức ý nghĩa $\alpha = 0.01$ hay không?

Giải

Tham số μ = nhiệt độ trung bình kích hoạt hệ thống.

Đặt giả thiết $H_0 : \mu = 130$ (với $\mu_0 = 130$). Đối thiết $H_a : \mu \neq 130$

Tính giá trị kiểm định thống kê:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}} = \frac{131.08 - 130}{1.5/\sqrt{9}} = 2.16$$

Miền bác bỏ: với $\alpha = 0.01$, ta có: $z_{\alpha/2} = z_{0.005} = 2.58$

So sánh ta có: $|z| < z_{\alpha/2}$. Do đó, ta không thể bác bỏ H_0 (chấp nhận H_0).

Kết luận dữ liệu không làm thay đổi phát biểu của nhà sản xuất.

Kiểm định trung bình với cỡ mẫu lớn ($n > 40$)

Cho $X_1, \dots, X_n$ là những mẫu ngẫu nhiên có phân phối chuẩn với kì vọng μ và phương sai σ^2 . Khi đó,

$$\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \sigma^2/n)$$

Đặt giả thiết $H_0 : \mu = \mu_0$

Tính thống kê kiểm định: $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Giả thiết đối	Miền bác bỏ cho kiểm định mức α
$H_a : \mu > \mu_0$	$ z \geq z_\alpha$ (kiểm định bên phải)
$H_a : \mu < \mu_0$	$ z \geq z_\alpha$ (kiểm định bên trái)
$H_a : \mu \neq \mu_0$	$ z \geq z_{\alpha/2}$ (kiểm định hai bên)

Kiểm định trung bình với cỡ mẫu lớn ($n > 40$)

Ví dụ. Một máy cân đóng gói các sản phẩm có khối lượng 1 kg. Nghi ngờ máy hoạt động không bình thường, người ta chọn ra một mẫu gồm 100 sản phẩm thì thấy như sau:

Khối lượng	0.95	0.97	0.99	1.01	1.03	1.05
Số gói	9	31	40	15	3	2

Với mức ý nghĩa $\alpha = 0.05$, hãy kết luận về nghi ngờ trên?

Giải

Gọi μ là trọng lượng trung bình của một gói sản phẩm.

Đặt giả thiết $H_0 : \mu = 1$. Đối thiết $H_a : \mu \neq 1$

Từ bảng dữ liệu ta có: $n = 100, \bar{x} = 0.9856, s = 0.02$

Tính giá trị kiểm định thống kê: $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.9856 - 1}{0.02/\sqrt{100}} = -7.2$

Với mức ý nghĩa $\alpha = 0.05$, ta có $z_{\alpha/2} = z_{0.025} = 1.96$

So sánh, ta được $|z| > z_{\alpha/2}$. Do đó, bác bỏ giả thiết H_0 tức là nghi ngờ trên là có cơ sở ($\bar{x} = 0.9856 < \mu = 1$. Túi đóng gói nhỏ hơn 1 kg.)

Kiểm định trung bình với cỡ mẫu nhỏ ($n < 40$)

Cho $X_1, \dots, X_n$ là những mẫu ngẫu nhiên có phân phối chuẩn với kì vọng μ và phương sai σ^2 . Khi đó, biến ngẫu nhiên

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

có phân phối t với bậc tự do $n - 1$.

Đặt giả thiết $H_0 : \mu = \mu_0$

Tính thống kê kiểm định: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Giả thiết đối	Miền bác bỏ cho kiểm định mức α
$H_a : \mu > \mu_0$	$ t \geq t_{\alpha, n-1}$ (kiểm định bên phải)
$H_a : \mu < \mu_0$	$ t \geq t_{\alpha, n-1}$ (kiểm định bên trái)
$H_a : \mu \neq \mu_0$	$ t \geq t_{\alpha/2, n-1}$ (kiểm định hai bên)

Kiểm định trung bình với cỡ mẫu nhỏ ($n < 40$)

Ví dụ. Glycerol là sản phẩm chính trong quá trình lên men của rượu vang, góp phần tăng độ ngọt, hương vị, và đậm vị rượu. Người ta lấy 5 mẫu rượu và đo nồng độ (mg/ml) Glycerol trong đó, thu được kết quả: 2.67, 4.62, 4.14, 3.81, 3.83. Giả sử nồng độ Glycerol mong muốn là 4. Dùng dữ liệu thu được kiểm định xem nồng độ Glycerol thu được có đạt như mong muốn không với mức ý nghĩa 0.05?

Giải

Đặt giả thiết $H_0 : \mu = 4$. Đối thiết $H_a : \mu \neq 4$

Từ số liệu cho trước, ta có: $n = 5, \bar{x} = 3.814, \sigma = 0.718$

Tính giá trị kiểm định $t = \frac{\bar{x} - 4}{s/\sqrt{n}} = \frac{3.814 - 4}{0.718/\sqrt{5}} = -0.58$

Với mức ý nghĩa 0.05 ta thu được $t_{\alpha/2, n-1} = t_{0.025, 4} = 2.776$.

So sánh, ta thấy $|t| < t_{\alpha/2, n-1}$ nên ta không bác bỏ H_0 .

Kết luận nồng độ Glycerol thu được đạt như mong muốn.

Kiểm định tỉ lệ

Kiểm định mẫu lớn liên quan đến p là một trường hợp đặc biệt của quá trình mẫu lớn nói chung. Đặt $\hat{p}$ là ước lượng không chệch của p và $\hat{p}$ có xấp xỉ phân phối chuẩn. Khi đó:

Đặt giả thiết $H_0 : p = p_0$.

Tính giá trị kiểm định thống kê:
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Giả thiết đối	Miền bác bỏ
$H_a : p > p_0$	$ z \geq z_\alpha$ (kiểm định bên phải)
$H_a : p < p_0$	$ z \geq z_\alpha$ (kiểm định bên trái)
$H_a : p \neq p_0$	$z \geq z_{\alpha/2}$ (kiểm định hai bên)

Lưu ý: Điều kiện kiểm định: $np_0 \geq 10$ và $n(1 - p_0) \geq 10$.

Kiểm định tỉ lệ

Ví dụ. Nút chai làm từ các thành phần tự nhiên trong những chai rượu vang có thể bị hư hỏng. Vì thế rượu vang trong chai có thể bị nhiễm bẩn. Một báo cáo cho rằng trong quảng cáo về mùi vị của rượu làm từ nho Chardonnay, trong 91 thì có 16 chai bị hư hỏng có liên quan đến nút chai. Với mức ý nghĩa 10%, dữ liệu này có cung cấp bằng chứng mạnh mẽ nào cho kết luận rằng hơn 15% của tất cả các chai bị nhiễm bẩn cũng bằng con đường này không?

Giải

Gọi p = tỉ lệ chai bị hư hỏng do nút chai.

Đặt giả thiết $H_0 : p = 0.15$. Đối thiết $H_a : p > 0.15$

Với $n = 91$, $\hat{p} = \frac{16}{91} = 0.1758$, ta tính giá trị kiểm định thống kê

$$z = \frac{\hat{p} - 0.15}{\sqrt{(0.15)(0.85)/n}} = \frac{0.1758 - 0.15}{\sqrt{(0.15)(0.85)/91}} = 0.69$$

Với mức ý nghĩa $\alpha = 0.1$, ta có: $z_\alpha = 1.28$.

Khi đó, ta có $z \leq z_\alpha$. Do đó không thể bác bỏ H_0 .

Kết luận, chỉ có 15% chai bị nhiễm bẩn do nút chai.

Bài tập

1. Đo chiều dài 1 số sản phẩm do nhà máy A sản xuất, có thống kê sau: (X có phân phối chuẩn).

$X(\text{cm})$	53.80	53.81	53.82	53.83	53.84	53.85	53.86	53.87
n_i	9	14	39	47	40	33	15	12

- a. Tìm khoảng tin cậy 95% cho chiều dài trung bình các sản phẩm do nhà máy A sản xuất.
- b. Sản phẩm đem tiêu thụ nếu có chiều dài trung bình 53,83cm. Với mức ý nghĩa 1% với thống kê trên, sp của nhà máy A đem tiêu thụ được chưa?
- c. Tìm khoảng tin cậy 98% cho tỉ lệ sản phẩm có chiều dài trên 53,84 cm.
- d. Sản phẩm có chiều dài trên 53,86 cm gọi là phế phẩm. Với mức ý nghĩa 5% có thể cho rằng tỉ lệ phế phẩm ở nhà máy A là 5% được không?

Bài tập

2. Biết mức độ rò rỉ khí gas tại van của bình gas GP đã nạp 50 lần là biến ngẫu nhiên X có phân phối chuẩn, khảo sát mức độ rò rỉ khí gas tại van của 200 bình gas GP đã nạp 50 lần thu được bảng số liệu sau:

$X(\text{LEL})$	0.07	0.08	0.11	0.13	0.15	0.17	0.19
Số bình	18	19	38	55	31	23	16

- Tìm khoảng ước lượng với độ tin cậy 95% cho mức độ rò rỉ khí gas trung bình tại van của bình gas GP đã nạp 50 lần
- Nếu mức độ rò rỉ khí gas trung bình tại van của bình gas GP đã nạp 50 lần vượt quá 0,1 LEL thì nhà sản xuất phải tiến hành việc thu hồi những bình gas GP đã nạp 50 lần. Ở mức ý nghĩa 1%, hãy cho biết nhà sản xuất có phải tiến hành việc thu hồi này hay không?
- Nếu tỉ lệ bình gas có mức độ rò rỉ khí gas tại van vượt quá 0,1 LEL trong tổng số bình gas GP đã nạp 50 lần là trên 80% thì nhà sản xuất cần cải tiến quy trình sản xuất. Ở mức ý nghĩa 1%, hãy kiểm định xem nhà sản xuất có cần cải tiến quy trình sản xuất hay không.
- Tìm khoảng ước lượng với độ tin cậy 95% cho tỉ lệ bình gas có mức độ rò rỉ khí gas tại van ít hơn 0,08 LEL trong tổng số bình gas GP đã nạp 50 lần.

Bài tập

3. Công ty M có 3000 đại lý, cho tiến hành điều tra ngẫu nhiên một số đại lý của mình và thu được bảng số liệu sau (X là doanh số, đơn vị: triệu đồng/tháng), biết X có phân phối chuẩn.

X	20 - 25	25 - 30	30 - 35	35 - 40
Số đại lý	7	12	18	27

40 - 45	45 - 50	50 - 55	55 - 60
22	17	13	4

- Những đại lý có $X > 45$ triệu đồng/tháng gọi là đại lý có doanh số cao. Hãy ước lượng số đại lý có doanh số cao với độ tin cậy 95%.
- Có ý kiến cho rằng tỉ lệ đại lý có doanh số cao bằng $1/3$ tỉ lệ đại lý có doanh thu còn lại. Cho nhận xét về ý kiến này với mức ý nghĩa 1%.
- Hãy UL doanh số trung bình/tháng của các đại lý với độ tin cậy 99%.
- Doanh số trung bình/tháng của các đại lý của cty N là 35trđ/tháng. Hãy so sánh doanh số trung bình/tháng của các đại lý của 2 công ty M, N với mức ý nghĩa 5%.

Bài tập

4. Khảo sát mức tiêu thụ điện X của một số hộ gia đình được chọn ngẫu nhiên ở vùng A ta được bảng số liệu sau:

$X(\text{kwh}/\text{tháng})$	50-100	100-150	150-200	200-250
Số hộ	24	36	55	64
	250-300	300-350	350-400	400-450
	50	35	20	15

- a. Ước lượng mức tiêu thụ điện trung bình của các hộ gia đình ở vùng A với độ tin cậy 99%.
- b. Có ý kiến cho rằng tỉ lệ hộ gia đình vùng A có mức tiêu thụ điện dưới 300kwh/tháng gấp 3 lần tỉ lệ hộ gia đình có mức tiêu thụ điện trên 300kwh/tháng. Cho kết luận ý kiến này với mức ý nghĩa 5%.
- c. Hộ có mức tiêu thụ điện dưới 100kwh/tháng gọi là hộ có mức tiêu thụ điện thấp. hãy ước lượng số hộ có mức tiêu thụ điện thấp ở vùng A với độ tin cậy 98% .Biết vùng A có 10.000 hộ dân.

Kiểm định 2 trung bình với σ đã biết

Giả sử, cho $X_1, \dots, X_n$ là những mẫu ngẫu nhiên có phân phối chuẩn với kì vọng μ_1 và phương sai σ_1^2 .

Cho $Y_1, \dots, Y_n$ là những mẫu ngẫu nhiên có phân phối chuẩn với kì vọng μ_2 và phương sai σ_2^2 .

Mẫu X và Y độc lập. Khi đó, giá trị kì vọng của $\bar{X} - \bar{Y}$ là $\mu_1 - \mu_2$ và độ lệch chuẩn của $\bar{X} - \bar{Y}$ là $\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$.

Đặt giả thiết $H_0 : \mu_1 - \mu_2 = \Delta_0$

Tính thống kê kiểm định: $z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$

Giả thiết đối	Miền bác bỏ cho kiểm định mức α
$H_a : \mu_1 - \mu_2 > \Delta_0$	$ z \geq z_\alpha$ (kiểm định bên phải)
$H_a : \mu_1 - \mu_2 < \Delta_0$	$ z \geq z_\alpha$ (kiểm định bên trái)
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$ z \geq z_{\alpha/2}$ (kiểm định hai bên)

Kiểm định 2 trung bình với σ đã biết

Ví dụ. Phân tích 20 mẫu thép cán nguội để xem độ cứng của thép, ta thu được độ cứng trung bình $\bar{x} = 29.8$ ksi và $\sigma_1 = 4.0$. Xét một mẫu khác có cỡ mẫu $n = 25$ và $\bar{y} = 34.7$ ksi, $\sigma_2 = 5.0$. Giả sử 2 mẫu này đều có phân phối chuẩn và người ta cho rằng độ cứng trung bình μ_1 và μ_2 khác nhau. Hãy kiểm định điều đó với mức ý nghĩa 1%.

Giải

Đặt giả thiết $H_0 : \mu_1 - \mu_2 = 0$. Đối thiết $H_a : \mu_1 - \mu_2 \neq 0$

Tìm giá trị kiểm định thống kê

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{29.8 - 34.7}{\sqrt{\frac{16}{20} + \frac{25}{25}}} = -3.66$$

Với mức ý nghĩa $\alpha = 0.01$, ta có $z_{\alpha/2} = z_{0.005} = 2.58$.

So sánh: $|z| \geq z_{\alpha/2}$, ta bác bỏ H_0 .

Kết luận độ cứng trung bình của hai mẫu khác nhau.

Kiểm định 2 trung bình với σ chưa biết

Nếu σ chưa biết, giá trị kiểm định thống kê được tính như sau:

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

Ví dụ. Để nghiên cứu tác dụng việc bón phân đạm theo công thức A đối với sản lượng bắp, người ta làm thí nghiệm trên 5 cặp mảnh đất cạnh nhau, mỗi cặp gồm có 1 mảnh đối chứng (không bón phân đạm) và một mảnh có bón phân đạm theo công thức A. Các sản phẩm thu được như sau (tính theo tạ/ha)

Mảnh đối chứng	55	53	30	37	49
Mảnh bón phân	60	58	28	39	47

Hãy cho kết quả về hiệu quả của việc bón phân đạm theo công thức A, với mức ý nghĩa 5%.

Kiểm định 2 tỷ lệ

Gọi p_1, p_2 là tỉ lệ thành công của đám đông thứ nhất và thứ hai.
Xét $X \sim \text{Bin}(m, p_1)$ và $Y \sim \text{Bin}(n, p_2)$ sao cho X và Y độc lập.
Đặt $\hat{p}_1 = X/m$ và $\hat{p}_2 = Y/n$. Khi đó, ta có

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2,$$

tức là $\hat{p}_1 - \hat{p}_2$ là một ước lượng không chệch của $p_1 - p_2$.
Do đó, ta có bài toán kiểm định 2 tỷ lệ sau:

Đặt giả thiết $H_0 : p_1 - p_2 = 0$

Giá trị kiểm định thống kê: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$ với $\hat{p} = \frac{X+Y}{m+n}$ và $\hat{q} = 1 - \hat{p}$.

Giả thiết đối	Miền bác bỏ cho kiểm định mức α
$H_a : p_1 - p_2 > 0$	$ z \geq z_\alpha$ (kiểm định bên phải)
$H_a : p_1 - p_2 < 0$	$ z \geq z_\alpha$ (kiểm định bên trái)
$H_a : p_1 - p_2 \neq 0$	$ z \geq z_{\alpha/2}$ (kiểm định hai bên)

Kiểm định 2 tỷ lệ

Ví dụ. Nghiên cứu về dự định đăng kí thi đại học về khối ngành kỹ thuật của học sinh phổ thông ở thành thị và nông thôn với câu hỏi: "Em có đăng kí vào khối ngành kỹ thuật hay không?" Ta thu được kết quả sau:

Nơi điều tra	Số học sinh trả lời CÓ	Số học sinh trả lời KHÔNG
Thành thị	240	300
Nông thôn	185	200

- Hãy so sánh tỉ lệ học sinh dự định thi đại học khối ngành kỹ thuật ở thành thị và nông thôn với mức ý nghĩa 5%.
- Có kết luận cho rằng tỉ lệ học sinh phổ thông có dự định thi đại học khối ngành kỹ thuật là 50%. Điều đó có đúng không? (cho kết luận với mức ý nghĩa 5%.)

Kiểm định 2 tỷ lệ - Bài tập

1. Trong 500 sv nam có 45 sv đạt loại giỏi. Trong 400 sv nữ có 50 sv đạt loại giỏi. Với mức ý nghĩa 2%, hãy so sánh tỉ lệ giỏi của nam và nữ.
2. Nghiên cứu khả năng chống cảm cúm của Vitamin C, có kết quả sau:
Trong số 420 người không uống Vitamin C, có 93 người bị cảm cúm.
Trong số 417 người, mỗi ngày uống 1g Vitamin C/ mỗi người, có 51 người bị cảm cúm. Với mức ý nghĩa 1% có thể cho rằng Vitamin C có khả năng chống cảm cúm hay không?
3. Công ty M kiểm tra ngẫu nhiên 1200 sản phẩm do ca sáng sản xuất thấy có 45 sản phẩm không đạt chuẩn, và kiểm tra ngẫu nhiên 1000 sản phẩm do ca chiều sản xuất thấy có 55 sản phẩm không đạt chuẩn.
 - a. Với mức ý nghĩa 5%, hãy so sánh tỷ lệ sản phẩm không đạt chuẩn do ca sáng và ca chiều sản xuất.
 - b. Tính tỷ lệ sản phẩm đạt chuẩn tối đa do ca sáng sản xuất với độ tin cậy 99%.

Kiểm định t cặp

Xét một dữ liệu gồm n cặp được chọn độc lập

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n),$$

trong đó $E(X_i) = \mu_1$ và $E(Y_i) = \mu_2$ với $i = 1, \dots, n$.

Đặt

$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \dots, D_n = X_n - Y_n.$$

Khi đó, D_i là khác nhau từng cặp.

Giả sử D_i có phân phối chuẩn với trung bình μ_D và phương sai σ_D^2 (bởi vì mẫu X_i và mẫu Y_i có phân phối chuẩn).

Kiểm định t cặp

Đặt giả thuyết: $H_0 : \mu_D = \Delta_0$

(trong đó $D = X - Y$ là hiệu giữa quan sát thứ nhất và quan sát thứ hai trong một cặp, và $\mu_D = \mu_X - \mu_Y$).

Tìm giá trị thống kê kiểm định:

$$t = \frac{\bar{d} - \Delta_0}{s_D / \sqrt{n}}$$

(trong đó $\bar{D}$ và s_D là trung bình mẫu và độ lệch chuẩn của mẫu d_i).

Giả thiết đối	Miền bác bỏ cho kiểm định mức α
$H_a : \mu_D > \Delta_0$	$ t \geq t_{\alpha, n-1}$ (kiểm định bên phải)
$H_a : \mu_D < \Delta_0$	$ t \geq t_{\alpha, n-1}$ (kiểm định bên trái)
$H_a : \mu_D \neq \Delta_0$	$ t \geq t_{\alpha/2, n-1}$ (kiểm định hai bên)

Kiểm định t cặp - Ví dụ

1. Đo chiều dài tay trái X và chiều dài tay phải Y của một số người được chọn ngẫu nhiên thu được bảng số liệu:

X (cm)	60	61	61.5	60.8	63.2	62.4	61.5	63.2
Y (cm)	61	63	62.1	61.3	61.9	61.1	61.4	62.8

Với mức ý nghĩa 1%, hãy so sánh chiều dài trung bình của tay trái và tay phải.

2. Khảo sát chiều cao (đơn vị: cm) của một số học sinh tiểu học lớp 4 sau 3 tháng nghỉ hè ở vùng A ta thu được số liệu:

Trước hè (Y)	132.1	135.2	128.5	122.8	128.5	138.1	140.2	136.3
Sau hè (X)	132.8	135.2	128.9	123.5	129.3	138.9	140.5	136.5

Hãy cho kết luận về ý kiến cho rằng sau 3 tháng nghỉ hè, học sinh tiểu học lớp 4 tăng ít nhất 0,3 cm với mức ý nghĩa 5%.

Kiểm định t cặp

1. Gọi μ_X, μ_Y là chiều dài trung bình của tay trái và tay phải.
Đặt $\mu_D = \mu_X - \mu_Y$. Khi đó, ta có:

X (cm)	60	61	61.5	60.8	63.2	62.4	61.5	63.2
Y (cm)	61	63	62.1	61.3	61.9	61.1	61.4	62.8
$d = X - Y$	-1	-2	-0.6	-0.5	1.3	1.3	0.1	0.4

Đặt giả thiết: $H_0 : \mu_D = 0$. Đối thiết: $H_a : \mu_D \neq 0$.

Từ số liệu trên, ta có: $n = 8, \bar{d} = -0.124, s_d = 1.136$.

Giá trị kiểm định:

$$t = \frac{\bar{d} - \Delta_0}{s_D / \sqrt{n}} = \frac{-0.124 - 0}{1.136 / \sqrt{8}} = -0.3112$$

Với $\alpha = 0.01$, ta có $t_{\alpha/2; n-1} = t_{0.005; 7} = 3.499$.

So sánh $|t| < t_{\alpha/2; n-1}$. Do đó, chấp nhận H_0 .

Kết luận: chiều dài 2 tay bằng nhau.

Kiểm định t cặp

2. Gọi μ_X, μ_Y lần lượt là chiều dài trung bình sau hè và trước hè. Đặt $\mu_D = \mu_X - \mu_Y$. Khi đó, ta có

Trước hè (Y)	132.1	135.2	128.5	122.8	128.5	138.1	140.2	136.3
Sau hè (X)	132.8	135.2	128.9	123.5	129.3	138.9	140.5	136.5
$d = X - Y$	0.7	0	0.4	0.7	0.8	0.8	0.3	0.2

Đặt giả thiết: $H_0 : \mu_D = 0.3$. Đối thiết: $H_a : \mu_D > 0.3$.

Từ số liệu trên, ta có: $n = 8, \bar{d} = 0.4875, s_d = 0.3044$.

Giá trị kiểm định:

$$t = \frac{\bar{d} - \Delta_0}{s_d / \sqrt{n}} = \frac{0.4874 - 0.3}{0.3044 / \sqrt{8}} = 1.742$$

Với $\alpha = 0.05$, ta có $t_{\alpha; n-1} = t_{0.05; 7} = 1.895$.

So sánh $t < t_{\alpha; n-1}$. Do đó, chấp nhận H_0 .

Kết luận: ý kiến trên là sai.

Hồi quy tuyến tính

Cho X và Y là hai biến ngẫu nhiên nhận các giá trị tương ứng $x_1, \dots, x_n$ và $y_1, \dots, y_n$

X	x_1	x_2	x_3	$\dots$	x_n
Y	y_1	y_2	y_3	$\dots$	y_n

Ta tìm mối quan hệ của X và Y bằng cách thiết lập mô hình hồi quy tuyến tính.

Hồi quy tuyến tính là phương trình hồi quy: $y = a + bx$, trong đó

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

và

$$a = \frac{\sum y_i - b \sum x_i}{n} = \bar{y} - b\bar{x}$$

Hồi quy tuyến tính

Hệ số tương quan của n cặp $(x_1, y_1), \dots, (x_n, y_n)$ được cho bởi

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

Tính chất.

- i. $r \in [-1; 1]$
- ii. Nếu $r = 0$ thì hai biến ngẫu nhiên không có tương quan.
- iii. Mức phụ thuộc tuyến tính của hai biến cho trong bảng sau

Yếu	Vừa	Mạnh
$ r \leq 0.5$	$-0.8 < r < -0.5, 0.5 < r < 0.8$	$ r \geq 0.8$

Hồi quy tuyến tính

Ví dụ. Giả sử ta có bảng số liệu sau:

X	132.0	129.0	120.0	113.2	105.0	92.0	84.0
Y	46.0	48.0	51.0	52.1	54.0	52.0	59.0
X	83.2	88.4	59.0	80.0	81.5	71.0	69.2
Y	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Khi đó, ta tìm được:

$$b = -0.20938742 \text{ và } a = 75.212432$$

Từ đó, ta có phương trình hồi quy là

$$y = 75.212 - 0.209x$$

Hệ số tương quan là: $r = -0.8892$. Tương quan mạnh.