Lâm sàng thống kê

Kiểm định t và hoán chuyển số liệu

Hỏi: "*Tôi nghe nói rằng khi đánh giá sự khác biệt giữa hai nhóm bằng t-test cần phải chuyển đổi số liệu. Tại sao?*"

Để đánh giá độ khác biệt giữa hai nhóm, chúng ta thường sử dụng phương pháp kiểm định t (hay t-test). Kiểm định t có lẽ là một trong những phương pháp đơn giản nhất trong thống kê học, vì có thể tính toán một cách thủ công, mà không cần đến máy tính hay phần mềm phân tích số liệu (nhưng nếu có thì tốt hơn!)

Tuy đơn giản, nhưng phương pháp kiểm định t cũng rất dễ sai lầm. Sai lầm thông thường nhất là không để ý đến những giả định đằng sau phương pháp này. Phương pháp kiểm định t chỉ thích hợp nếu số liệu đáp ứng những điều kiện hay giả định sau đây:

- Hai nhóm so sánh phải hoàn toàn độc lập nhau;
- Biến so sánh phải tuân theo luật phân phối chuẩn (Gaussian distribution);
- Phương sai của hai nhóm bằng nhau, hay gần bằng nhau; và
- Các đối tượng phải được chọn một cách ngẫu nhiên (random sample).

Thế nào là "độc lập"? Khi nói đến độc lập ở đây là nói đến hai nhóm không có tương quan nhau. Chẳng hạn như một nhóm 1 gồm bệnh nhân A, B, C và D; nhóm 2 gồm bệnh nhân E, F, G và H, thì hai nhóm này độc lập nhau. Nhưng nếu có một nhóm bệnh nhân mà đo hai lần, thì hai biến số của hai lần đo đó không độc lập với nhau. Độc lập cũng có nghĩa là không liên hệ nhau. Chẳng hạn như nếu 2 bệnh nhân trong nhóm 1 (A và C) có liên hệ huyết thống, và nếu biến mà chúng ta phân tích có yếu tố di truyền thì đo lường của hai bệnh nhân không được xem là độc lập.

1. Lí thuyết của kiểm định t

Cho hai quần thể độc lập 1 và 2, với chỉ số trung bình μ_1 và μ_2, và phương sai σ^2. Chúng ta muốn đánh giá độ khác biệt giữa hai quần thể. Nhưng chúng ta không biết các giá trị này.

Để tìm hiểu xem μ_1 và μ_2 có khác nhau hay không, chúng ta lấy mẫu từ hai quần thể đó. Giả sử chúng ta lấy ngẫu nhiên n_1 đối tượng từ quần thể 1, và n_2 đối tượng từ quần thể 2. Sau khi đo lường biến số, chúng ta có kết quả như sau:

	Nhóm 1	Nhóm 2
Số đối tượng	n_1	n_2
Trung bình	$\overline{x}_1$	$\overline{x}_2$
Phương sai	s_1^2	s_2^2
Độ lệch chuẩn	s_1	s_2

Xin nhắc lại, chúng ta muốn tìm hiểu độ khác biệt giữa hai quần thể (chứ không phải giữa hai nhóm mẫu). Mục đích này có thể phát biểu bằng hai giả thuyết như sau:

Giả thuyết vô hiệu H_o: $\mu_1 = \mu_2$

Giả thuyết chính H_1: $\mu_1 \neq \mu_2$

Gọi $\Delta = \mu_1 - \mu_2$, hai giả thuyết trên cũng có thể phát biểu như sau:

H_o: $\Delta = 0$

H_1: $\Delta \neq 0$

Trong điều kiện không biết các giá trị của quần thể μ_1 và μ_2, ước số thích hợp nhất quần thể chính là hai số trung bình $\overline{x}_1$ và $\overline{x}_2$ tính từ mẫu 1 và mẫu 2. Và, ước tính độ khác biệt Δ chính là độ khác biệt giữa hai số trung bình:

$$d = \overline{x}_1 - \overline{x}_2 \qquad [1]$$

Nhưng vì lấy mẫu, cho nên d có thể biến thiên từ mẫu này sang mẫu khác, và vấn đề là tìm phương sai của d. Lí thuyết xác suất cho chúng ta biết rằng phương sai của khác biệt giữa hai biến bằng tổng phương sai của hai biến trừ cho 2 lần hiệp biến, tức là:

$$\text{var}(a - b) = \text{var}(a) + \text{var}(b) - 2 \times \text{cov}(a,b)$$

Trong đó, "var" là viết tắt của variance (phương sai), và "covar" là viết tắt của covariance (hiệp biến). Hiệp biến phản ảnh độ tương quan giữa hai biến. Nhưng nếu hai biến hoàn toàn độc lập, thì hiệp biến sẽ là 0, và công thức trên đơn giản thành:

$$\text{var}(a - b) = \text{var}(a) + \text{var}(b)$$

Áp dụng công thức này, chúng ta có thể ước tính phương sai cho *d* trong [1] như sau (Tôi sẽ kí hiệu phương sai bằng *s* bình phương):

$$s_d^2 = s_1^2 + s_2^2 \qquad [2]$$

Từ đó, độ lệch chuẩn của *d* là:

$$s_d = \sqrt{s_1^2 + s_2^2} \qquad [3]$$

Nhưng vì những ước số đều dựa vào số cỡ mẫu, cho nên chúng ta phải "điều chỉnh" bằng cách chia phương sai cho số cỡ mẫu:

$$SE_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad [4]$$

Nếu phương sai của hai nhóm bằng nhau (tức $s_1^2 = s_2^2 = s^2$), phương trình [4] đơn giản thành:

$$SE_d = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad [5]$$

Kiểm định t đơn giản là tỉ số của *d* trên SE_d, hay cụ thể hơn:

$$t = \frac{d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad [6]$$

Có thể xem công thức [5] như là tỉ số của "tín hiệu" (signal) và "nhiễu" (SE_d). Thật vậy, *d* phản ảnh độ khác biệt giữa hai nhóm, và SE_d phản ảnh độ nhiễu của *d*. Thành ra, nếu tỉ số *t* cao, chúng ta có bằng chứng để nói tín hiệu nhiều hơn nhiễu (tức có ý nghĩa thống kê); nếu tỉ số *t* thấp dưới 1 chẳng hạn, chúng ta có bằng chứng để phát biểu tín hiệu thấp hơn nhiễu và do đó độ khác biệt không có ý nghĩa thống kê.

Nhưng "cao" là cao bao nhiêu để có thể nói là có ý nghĩa thống kê? Để trả lời câu hỏi này, chúng ta quay trở về với giả thuyết. Nếu giả thuyết vô hiệu H_o là sự thật (tức không có khác biệt giữa 2 quần thể), thì sự phân phối ngẫu nhiên của t như thế nào. May mắn thay, đã có nhà thống kê học trả lời câu hỏi này: đó là ông William Gossett, người phát kiến kiểm định t. Theo chứng minh của Gossett, nếu hai quần thể không khác nhau, thì giá trị của t tùy thuộc vào số cỡ mẫu (hay nói theo ngôn ngữ thống kê học là *bậc tự do* – degrees of freedom). Số bậc tự do (kí hiệu) được tính bằng công thức sau đây:

$$df = n_1 + n_2 - 2$$

Bảng 1 sau đây trình bày tỉ số *t* cho từng bậc tự do và khoảng xác suất mà tỉ số *t* có thể dao động ngẫu nhiên:

Bảng 1. Tỉ số *t* cho từng bậc tự do nếu giả thuyết vô hiệu Ho đúng

Bậc tự do (df)	Xác suất 95% tỉ số t sẽ dao động trong khoảng	Xác suất 99% tỉ số t sẽ dao động trong khoảng
5	-2.57 đến 2.57	-4.03 đến 4.03
10	-2.23 đến 2.23	-3.17 đến 3.17
14	-2.14 đến 2.14	-2.98 đến 2.98
16	-2.12 đến 2.12	-2.92 đến 2.92
18	-2.10 đến 2.10	-2.88 đến 2.88
20	-2.08 đến 2.08	-2.84 đến 2.84
24	-2.06 đến 2.06	-2.80 đến 2.80
30	-2.04 đến 2.04	-2.75 đến 2.75
34	-2.03 đến 2.03	-2.73 đến 2.73
40	-2.02 đến 2.02	-2.70 đến 2.70
50	-2.01 đến 2.01	-2.68 đến 2.68
60	-2.00 đến 2.00	-2.66 đến 2.66
70	-2.00 đến 2.00	-2.65 đến 2.65
80	-2.00 đến 2.00	-2.64 đến 2.64
90	-1.99 đến 1.99	-2.64 đến 2.64
100	-1.98 đến 1.98	-2.62 đến 2.62
500	-1.96 đến 1.96	-2.58 đến 2.58
1000	-1.96 đến 1.96	-2.58 đến 2.58

Do đó, nếu tỉ số *t* tính toán từ công thức [6] nằm ngoài khoảng tin cậy trên đây, chúng ta có thể nói rằng độ khác biệt giữa hai quần thể có ý nghĩa thống kê (thuật ngữ tiếng Anh là "statistically significant").

2. Kiểm định t với biến được hoán chuyển logarít

Ví dụ 1. Một nghiên cứu nhằm so sánh nồng độ lysozyme giữa hai nhóm bệnh nhân (tạm gọi là nhóm 1 và nhóm 2). Nhóm 1 gồm 29 bệnh nhân, và nhóm 2 gồm 30 bệnh nhân, tuổi từ 20 đến 60. Nồng độ lysozyme (mg/L) như sau và có thể tóm lược trong **Bảng 2**:

Nhóm 1: 0.2, 0.3, 0.4, 1.1, 2.0, 2.1, 3.3, 3.8, 4.5, 4.8, 4.9, 5.0, 5.3, 7.5, 9.8, 10.4, 10.9, 11.3, 12.4, 16.2, 17.6, 18.9, 20.7, 24.0, 25.4, 40.0, 42.2, 50.0, 60.0

Nhóm 2: 0.2, 0.3, 0.4, 0.7, 1.2, 1.5, 1.5, 1.9, 2.0, 2.4, 2.5, 2.8, 3.6, 4.8, 4.8, 5.4, 5.7, 5.8, 7.5, 8.7, 8.8, 9.1, 10.3, 15.6, 16.1, 16.5, 16.7, 20.0, 20.7, 33.0

Bảng 2. Nồng độ lysozyme ở bệnh nhân nhóm 1 và nhóm 2

	Nhóm 1	Nhóm 2
Số đối tượng	$n_1 = 29$	$n_2 = 30$
Trung bình	$\overline{x}_1 = 14.31$	$\overline{x}_2 = 7.68$
Phương sai	$s_1^2 = 247.8$	$s_2^2 = 61.6$
Độ lệch chuẩn	$s_1 = 15.7$	$s_2 = 7.8$

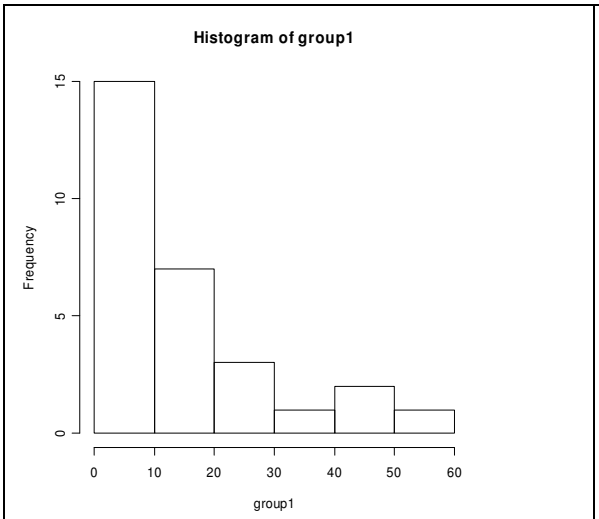
Áp dụng công thức [6], chúng ta có tỉ số t như sau:

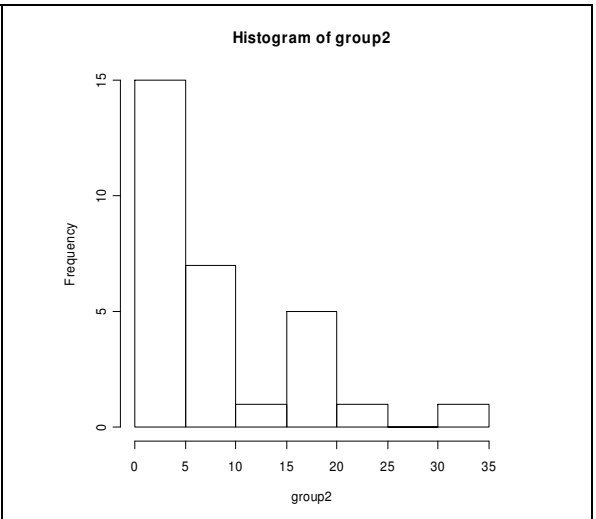
$$t = \frac{d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{14.31 - 7.68}{\sqrt{\frac{14.31}{29} + \frac{7.68}{30}}} = 2.03$$

Với bậc tự do df = 29+30-2 = 57, và nếu hai nhóm không khác nhau, chúng ta kì vọng rằng tỉ số *t* dao động từ -2.00 đến 2.00 (theo **Bảng 1**). Nhưng tỉ số *t* quan sát được nằm ngoài khoảng tin cậy này, nên chúng ta có thể phát biểu rằng độ lysozyme của hai nhóm khác nhau.

Nhưng kết quả và kết luận trên có thể sai! Nhìn qua tóm tắt trình bày trong Bảng 2, chúng ta chú ý phương sai của nhóm 1 cao gấp 4 lần so với nhóm 1. Ngoài ra, phương sai có xu hướng biến thiên theo số trung bình: nhóm có số trung bình cao cũng là nhóm có phương sai cao. Độ lệch chuẩn của nhóm 1 cao hơn nhóm 2 gấp hai lần.

Chúng ta cũng chú ý rằng độ lệch chuẩn của hai nhóm cao hơn số trung bình. Điều này hàm ý cho biết số liệu lysozyme không tuân theo luật phân phối chuẩn, và phân tích trên đã vi phạm giả định thống kê. Chúng ta thử xem qua phân phối của lysozyme trong nhóm 1 và nhóm 2 như sau:





Biểu đồ 1. Phân phối lysozyme của nhóm 1 (biểu đồ bên phải) và nhóm 2 (biểu đồ bên phải)

Rõ ràng độ lysozyme có xu hướng lệch về các giá trị nhỏ. Với xu hướng này, chúng ta có thể sử dụng hàm logarít để hoán chuyển số liệu. Sau khi hoán chuyển bằng logarít, chúng ta có số liệu mới cho nhóm 1 và 2 như sau (và bảng tóm lược 3)

Nhóm 1:

```
-1.60943791 -1.20397280 -0.91629073  0.09531018  0.69314718  0.74193734
 1.19392247  1.33500107  1.50407740  1.56861592  1.58923521  1.60943791
 1.66770682  2.01490302  2.28238239  2.34180581  2.38876279  2.42480273
 2.51769647  2.78501124  2.86789890  2.93916192  3.03013370  3.17805383
 3.23474917  3.68887945  3.74242022  3.91202301  4.09434456
```

Nhóm 2:

```
-1.6094379 -1.2039728 -0.9162907 -0.3566749  0.1823216  0.4054651
 0.4054651  0.6418539  0.6931472  0.8754687  0.9162907  1.0296194
 1.2809338  1.5686159  1.5686159  1.6863990  1.7404662  1.7578579
 2.0149030  2.1633230  2.1747517  2.2082744  2.3321439  2.7472709
 2.7788193  2.8033604  2.8154087  2.9957323  3.0301337  3.4965076
```

Bảng 3. Nồng độ lysozyme ở bệnh nhân nhóm 1 và nhóm 2

	Nhóm 1	Nhóm 2
Số đối tượng	$n_1 = 29$	$n_2 = 30$
Trung bình	$\overline{x}_1 = 1.92$	$\overline{x}_2 = 1.41$
Phương sai	$s_1^2 = 2.19$	$s_2^2 = 1.73$
Độ lệch chuẩn	$s_1 = 1.48$	$s_2 = 1.32$

Bây giờ thì hai phương sai tương đương nhau, và chúng ta có thể áp dụng kiểm định *t* qua công thức [6] như sau:

$$t = \frac{d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1.92 - 1.41}{\sqrt{\frac{2.19}{29} + \frac{1.73}{30}}} = 1.406$$

Như vậy, tỉ số *t* nằm trong khoảng -2.00 đến 2.00, tức là khoảng dao động hoàn toàn do ngẫu nhiên. Do đó, chúng ta kết luận rằng lysozyme của hai nhóm tương đương nhau.

3. Kiểm định t với biến được hoán chuyển căn số bậc 2

Nhiều nghiên cứu lâm sàng, tiêu chí để đánh giá kết quả (outcome measure) chỉ đơn giản là số đếm, và trước khi tiến hành kiểm định t, số liệu cần phải hoán chuyển bằng căn số bậc 2 để làm cho số liệu tuân theo luật phân phối chuẩn.

Ví dụ 2. Trong nghiên cứu trình bày dưới đây, các nhà khoa học đếm số lượng vi khuẩn lactobacilli trong nước bọt của hai nhóm bệnh nhân. Nhóm 1 gồm có 7 bệnh nhân được tiêm vắc-xin, và nhóm 2 gồm 6 đối tượng không được tiêm vắc-xin. Kết quả nghiên cứu như sau:

Nhóm 1		Nhóm 2	
Số vi khuẩn lactobacilli (*k*)	Hoán chuyển $\sqrt{k}$	Số vi khuẩn lactobacilli (*k*)	Hoán chuyển $\sqrt{k}$
7925	89.02	3158	56.20
15643	125.07	3669	60.57
17462	132.14	5930	77.01
10805	103.95	5697	75.48
9300	96.44	8331	91.27
7538	86.82	11822	108.73
6297	79.35		

Số liệu này có thể tóm lược trong **Bảng 4** sau đây:

Bảng 4. Tóm lược số liệu lactobacilli

	Nhóm 1	Nhóm 2
Số đối tượng	$n_1 = 7$	$n_2 = 6$

Trung bình (*x*)	$\overline{x}_1 = 10710$	$\overline{x}_2 = 6434$
Độ lệch chuẩn (*sd*)	$s_1 = 4266$	$s_2 = 3219$
Tỉ số $sd / \sqrt{x}$	41.2	40.1

Chúng ta chú ý rằng tỉ số độ lệch chuẩn trên căn số bậc 2 của số trung bình của hai nhóm là khoảng 40 đến 41 (tức tương đương nhau). Điều này cho thấy, chúng ta cần phải hoán chuyển số liệu bằng hàm căn số bậc 2, và kết quả được trình bày trong cột 2 (màu đỏ) của từng nhóm trong bảng số liệu gốc trên. Sau khi hoán chuyển chúng ta có một bảng tóm lược mới như sau:

Bảng 5. Tóm lược số liệu hoán chuyển lactobacilli bằng căn số bậc 2

	Nhóm 1	Nhóm 2
Số đối tượng	$n_1 = 7$	$n_2 = 6$
Trung bình (*x*)	$\overline{x}_1 = 101.8$	$\overline{x}_2 = 78.2$
Độ lệch chuẩn (*sd*)	$s_1 = 20.0$	$s_2 = 19.5$

Nếu phân tích dựa vào số liệu hoán chuyển, chúng ta có tỉ số *t* như sau:

$$t = \frac{d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{101.8 - 78.2}{\sqrt{\frac{(20)^2}{7} + \frac{(19.5)^2}{6}}} = 2.05$$

Với bậc tự do = 7+6-2 = 11, và nếu hai nhóm không khác nhau, chúng ta kì vọng tỉ số *t* sẽ dao động trong khoảng -2.23 đến 2.23 (Bảng 1) với xác suất 95%. Ở đây, chúng ta có tỉ số t quan sát là 2.05, nằm trong khoảng xác suất ngẫu nhiên này, chúng ta phải kết luận rằng chưa có bằng chứng để kết luận rằng hai nhóm bệnh nhân khác nhau về số lượng vi khuẩn lactobacilli. (Bạn đọc có thể tự làm phân tích trên số liệu chưa được hoán chuyển và sẽ thấy kết quả khác với kết luận vừa trình bày!)

4. Kiểm định t với biến là tỉ lệ

Ví dụ 3. Bảng số liệu sau đây là kết quả của một nghiên cứu lâm sàng đối chứng ngẫu nhiên, với mục tiêu so sánh hai phương pháp tập luyện bệnh nhân với chứng mất trí vì tuổi già. Nhóm một gồm 11 bệnh nhân được tập luyện, và nhóm hai gồm 8 bệnh nhân đối chứng (không tập luyện). Sau hai tuần tập luyện, mỗi bệnh nhân được cho 20 câu hỏi

về những việc trong đời sống hàng ngày (như khóa cửa, buộc giây, quét dọn, mặc quần áo, v.v…). Số câu trả lời đúng được ghi nhận và chia cho 20 (tức tính tỉ lệ trả lời đúng).

Tỉ lệ thành công trong 20 câu hỏi cho 2 nhóm bệnh nhân mất trí

Nhóm 1: `0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, 0.25`

Nhóm 2: `0.0, 0.15, 0.0, 0.05, 0.0, 0.0, 0.05, 0.10`

Bảng 6. Tóm lược số liệu của bệnh nhân mất trí

	Nhóm 1	Nhóm 2
Số đối tượng	11	8
Trung bình (*x*)	0.164	0.044
Độ lệch chuẩn (*sd*)	0.112	0.056

Trong trường hợp này, chúng ta thấy độ lệch chuẩn bằng hay cao hơn số trung bình, và đó là tín hiệu cho thấy biến số không tuân theo luật phân phối chuẩn.

Một trong những hàm hoán chuyển khá hữu hiệu cho các số liệu mang tính tỉ lệ (proportion) là hàm lượng giác arsin của căn số bậc 2 (tức $\arcsin\sqrt{x}$, trong đó x là tỉ lệ). Chẳng hạn như nếu $x = 0.05$, thì $\arcsin\sqrt{x} = \arcsin\sqrt{0.05} = 0.2255$. Sau khi hoán chuyển bằng hàm $\arcsin\sqrt{x}$, chúng ta có số liệu mới như sau.

Số liệu hoán chuyển bằng hàm $\arcsin\sqrt{x}$

Nhóm 1:

```
0.2255134 0.3976994 0.6330518 0.5235988 0.4636476 0.2255134 0.3217506
0.2255134 0.5796397 0.2255134 0.5235988
```

Nhóm 2:

```
0.0000000 0.3976994 0.0000000 0.2255134 0.0000000 0.0000000 0.2255134
0.3217506
```

Bảng 7. Tóm lược số liệu của bệnh nhân mất trí sau khi hoán chuyển

	Nhóm 1	Nhóm 2
Số đối tượng	11	8

Trung bình (*x*)	0.395	0.146
Độ lệch chuẩn (*sd*)	0.158	0.166

Áp dụng công thứ [6] cho số liệu hoán chuyển, chúng ta có:

$$t = \frac{d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0.395 - 0.146}{\sqrt{\frac{(0.158)^2}{11} + \frac{(0.146)^2}{8}}} = 3.30$$

Với bậc tự do 17 (df = 11 + 8 – 2), và nếu không có khác biệt giữa hai nhóm bệnh nhân, chúng ta kì vọng tỉ số *t* dao động trong khoảng -2.10 đến 2.10 với xác suất 95%. Tuy nhiên, ở đây tỉ số t = 3.30, nằm ngoài khoảng dao động ngẫu nhiên trên, chúng ta có bằng chứng để phát biểu rằng độ khác biệt hay ảnh hưởng của tập luyện có ý nghĩa thống kê. Thật ra, trị số P của tỉ số t trên là 0.005.

5. Tóm lược

Như vừa mô tả trong 3 ví dụ trên, chúng ta thấy rằng việc phân tích số liệu bằng phương pháp kiểm định t cực kì đơn giản, không cần đến máy tính. Logic đằng sau của phương pháp kiểm định t (cũng như của nhiều phương pháp khác) là kiểm định một giả thuyết vô hiệu (H_o) như sau:

- Giả thuyết H_o : Không có khác nhau giữa hai nhóm;
- Tính toán tỉ số *t* (độ khác biệt giữa 2 nhóm chia cho độ dao động)
- Nếu H_o đúng, xác định độ biến thiên của t_0 trong vòng 95% hay 99%
- Nếu *t* nằm ngoài khoảng biến thiên của t_0 , chúng ta loại giả thuyết H_o.

Dù phương tính và logic đơn giản như thế, nhưng phương pháp kiểm định t thường bị áp dụng sai, do không chú ý đến các giả định đằng sau của phương pháp. Trong nhiều trường hợp, sai phương pháp dẫn đến kết luận sai. Do đó, ảnh hưởng của việc bất cẩn trong phân tích có khi rất nghiêm trọng. Hi vọng qua các ví dụ này, bạn đọc đã biết qua vài phương pháp hoán chuyển số liệu, và có một cái nhìn mới hơn về phương pháp kiểm định t.

Nguyễn Văn Tuấn

Chú thích:

Tất cả các phân tích trên có thể tiến hành rất đơn giản bằng ngôn ngữ thống kê R. Dưới đây là các mã R mà tôi đã dùng cho các phân tích và biểu đồ trên. Bạn đọc có thể tự

mình kiểm tra bằng cách cắt từng phần và dán vào R để hiểu thêm. (Cách học hay nhất là bắt chước). Nếu muốn tìm hiểu thêm về R, bạn đọc có thể tìm mua quyển sách "**Phân tích số liệu và tạo biểu đồ bằng R**" của tôi do Nhà xuất bản Khoa học Kĩ thuật phát hành năm 2007.

 # Mã R để tìm tỉ số t cho Bảng 1

```
# bậc tự do - degrees of freedom

df <- c(5,10,14,16,20,24,30,34,40,50,60,70,80,90, 100, 500, 1000)

# tính tỉ số t cho xác suất 0.025 đến 0.975 (tức 95%)

t.025 <- qt(0.025, df)
t.975 <- qt(0.975, df)

# tính tỉ số t cho xác suất 0.005 đến 0.995 (tức 99%)

t.005 <- qt(0.005, df)
t.995 <- qt(0.995, df)
```

 # Ví dụ 1

```
# nhập package "epicalc" - chỉ R version 2.4.1

library(epicalc)

# nhập số liệu

group1 <- c(0.2, 0.3, 0.4, 1.1, 2.0, 2.1, 3.3, 3.8, 4.5, 4.8, 4.9, 5.0,
            5.3, 7.5, 9.8, 10.4, 10.9, 11.3, 12.4, 16.2, 17.6, 18.9,
            20.7, 24.0, 25.4, 40.0, 42.2, 50.0, 60.0)

group2 <-  c(0.2, 0.3, 0.4, 0.7, 1.2, 1.5, 1.5, 1.9, 2.0, 2.4, 2.5,
             2.8, 3.6, 4.8, 4.8, 5.4, 5.7, 5.8, 7.5, 8.7, 8.8, 9.1,
             10.3, 15.6, 16.1, 16.5, 16.7, 20.0, 20.7, 33.0)

# Phân tích mô tả (bảng 2)

summ(group1)
summ(group2)

# Kiểm định t - không hoán chuyển

t.test(group1, group2)

# Vẽ biểu đồ 1
```

```
hist(group1)
hist(group2)

# Hoán chuyển số liệu bằng hàm logarít

log.group1 <- log(group1)
log.group2 <- log(group2)

# Kiểm định t – số liệu hoán chuyển

t.test(log.group1, log.group2)

# Ví dụ 2: nhập dữ liệu

g1 <- c(7925, 15643, 17462, 10805, 9300, 7538, 6297)
g2 <- c(3158, 3669, 5930, 5697, 8331, 11822)

# Hoán chuyển bằng căn số bậc 2

t.g1 <- sqrt(g1)
t.g2 <- sqrt(g2)

# Kiểm định t

t.test(t.g1, t.g2)

# Ví dụ 3: nhập dữ liệu

d1 <- c(0.05,0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05,0.25)
d2 <- c(0.0, 0.15, 0.0, 0.05, 0.0, 0.0, 0.05, 0.10)

# Hoán chuyển bằng arcsin(sqrt(x))

t.d1 <- asin(sqrt(d1))
t.d2 <- asin(sqrt(d2))

# Kiểm định t

t.test(t.d1, t.d2)
```