Lâm sàng thống kê

Phân tích các biến không thể hoán chuyển

Nguyễn Văn Tuấn

Trong hai bài trước, tôi có mô tả cách hoán chuyển số liệu sao cho tuân theo luật phân phối chuẩn (Normal distribution) để tiện cho việc ứng dụng các phương pháp phân tích như kiểm định t, phân tích phương sai (analysis of variance). Tuy nhiên cũng có trường hợp chúng ta không thể hoán chuyển số liệu bằng các hàm thông dụng như logarít hay hàm mũ. Trong trường hợp này, chúng ta có hai phương án để phân tích:

Phương án thứ nhất là sử dụng các phương pháp phân tích phi thông số (non-parametric methods). Như tên gọi, các phương pháp phi tham số không đòi hỏi các biến số phải tuân theo luật phân phối chuẩn, và cách tính cũng tương đối đơn giản hơn các phương pháp có tham số. Phần lớn các phương pháp này hoán chuyển các biến liên tục (contonuous measurement) thành các biến thứ hạng (rank), và phân tích trên các biến thứ hạng này. Chẳng hạn như biến {79, 23, 5, 7, 56, } trước khi phân tích sẽ được hoán chuyển thành số thứ hạng như {5, 3, 1, 2, 4}. Như thấy qua ví dụ đơn giản trên, phương cách hoán chuyển từ số liên tục sang số thứ hạng trên có thể gây nên tình trạng mất thông tin (loss of information). Nhưng may mắn thay, trong nhiều trường hợp, vấn đề mất thông tin không gây ảnh hưởng lớn đến việc kiểm định các giả thiết khoa học.

Nếu chúng ta muốn kiểm định số liệu từ hai nhóm độc lập, thay vì sử dụng kiểm định t, phương pháp phi tham số tương đương là phương pháp Wilcoxon (còn gọi là **Wilcoxon's rank sum test,** hay có khi còn gọi là **Wilcoxon-Mann-Whitney test**). Nếu có hơn hai nhóm, thay vì sử dụng phân tích phương sai, phương pháp phi tham số tương đương là kiểm định Kruskal-Wallis (còn gọi là **Kruskal-Wallis test**).

Phương án thứ hai là ứng dụng phương pháp bootstrap (mà tôi đã giải thích trong bài trả lời về cách ước tính khoảng tin cậy 95% cho số trung vị trước đây).

1. Kiểm định Wilcoxon

Phương pháp kiểm định Wilcoxon có thể minh hoạ bằng một ví dụ đơn giản như sau: giả dụ chúng ta có số liệu từ 2 nhóm (A và B) như sau:

```
Nhóm A (2 đối tượng):  4, 12
Nhóm B (3 đối tượng): 14, 10, 17
```

và chúng ta muốn biết xem hai nhóm này có khác biệt hay không. Với phương pháp kiểm định Wilcoxon, chúng ta sẽ hoán chuyển số liệu thành hạng (rank). Trước kết, chúng ta tập hợp số liệu hai nhóm thành một vector, và sắp xếp theo tứ tự từ số thấp nhất đến cao nhất như sau:

```
Nhóm A + B:  4, 10, 12, 14, 17
Hạng (rank): 1, 2,  3,  4,  5
```

Chú ý rằng những chữ số được tô đậm thuộc nhóm B. chúng ta thấy tổng số hạng của nhóm B là:

$$S = 2 + 4 + 5 = 11 \qquad [1]$$

Tổng số hạng đó có ý nghĩa gì? Trước khi trả lời câu hỏi đó, chúng ta dành vài phút suy nghĩ. Có tất cả 5 đo lường; trong đó, nhóm B có 3 đo lường. Do đó, nếu3 đo lường của nhóm B hoàn toàn thấp hơn 2 đo lường nhóm A, thì **tổng số hạng** (sum of ranks) của nhóm B phải có giá trị tối thiểu là: 1 + 2 + 3 = 6. Ngược lại, nếu 3 đo lường của nhóm B hoàn toàn cao hơn 2 đo lường nhóm A, thì tổng số hạng của nhóm B phải có giá trị tối đa là 3 + 4 + 5 = 12.

Nói chung, nếu chúng ta có n_1 đối tượng trong nhóm A và n_2 đối tượng trong nhóm B, **tối thiểu tổng số hạng** của nhóm B là: $\frac{n_1(n_2+1)}{2}$, và **tối đa tổng số hạng** của nhóm B là: $n_1 n_2\left[\frac{n_2(n_2+1)}{2}\right]$. Bạn đọc có thể thay thế $n_1 = 2$ và $n_2 = 3$ để kiểm tra kết quả trên.

Nếu hai nhóm không khác nhau, Wilcoxon (từng là chủ tịch Hiệp hội khoa học thống kê Mĩ trong thập niên 1950s) chỉ ra rằng số hạng trung bình của nhóm B là:

$$\mu_T = \frac{n_2(n_1+n_2+1)}{2} \qquad [2]$$

Chú ý rằng trong công thức trên phải tuân theo thứ tự $n_2 > n_1$. Và phương sai là:

$$\sigma_T^2 = \frac{n_1 n_2(n_1+n_2+1)}{12} \qquad [3]$$

(Nói cách khác, độ lệch chuẩn là: $\sigma_T = \sqrt{\sigma_T^2}$).

Và nếu số đối tượng của hai nhóm tương đối đầy đủ (như trên 10 chẳng hạn), thì chỉ số thống kê $T = \frac{S - \mu_T}{\sigma_T}$ tuân theo luật phân phối chuẩn. Nói cách khác, nếu hai nhóm không khác nhau, thì 95% trị số của *T* sẽ dao động trong khoảng -2 đến 2. Tức là, nếu *T* thấp hơn -2 hay cao hơn 2, chúng ta có bằng chứng để phát biểu rằng độ khác biệt giữa hai nhóm có ý nghĩa thống kê.

Một cách khác là ước tính khoảng tin cậy 95% của μ_T như sau: $\mu_T \pm 1.96 \times \sigma_T$. Nếu tổng số hạng *S* nằm trong khoảng tin cậy 95%, chúng ta có lí do để phát biểu rằng hai nhóm không khác nhau; nếu *S* nằm ngoài khoảng tin cậy 95%, đó là tín hiệu cho thấy hai nhóm khác nhau có ý nghĩa thống kê.

2. Kiểm định Wilcoxon: một ví dụ

Ví dụ 1: Số liệu sau đây (do một bạn đọc cung cấp) về một nghiên cứu so sánh tỉ lệ thải sodium qua đường nước tiểu (fractional excretion – FE) giữa hai nhóm bệnh nhân, tạm gọi là nhóm 1 và nhóm 2. Số liệu FE (trích dẫn để minh họa cho ví dụ) của hai nhóm như sau:

Bảng 1. Tỉ lệ thải sodium (tính bằng phần trăm) qua đường nước tiểu của 25 đối tượng

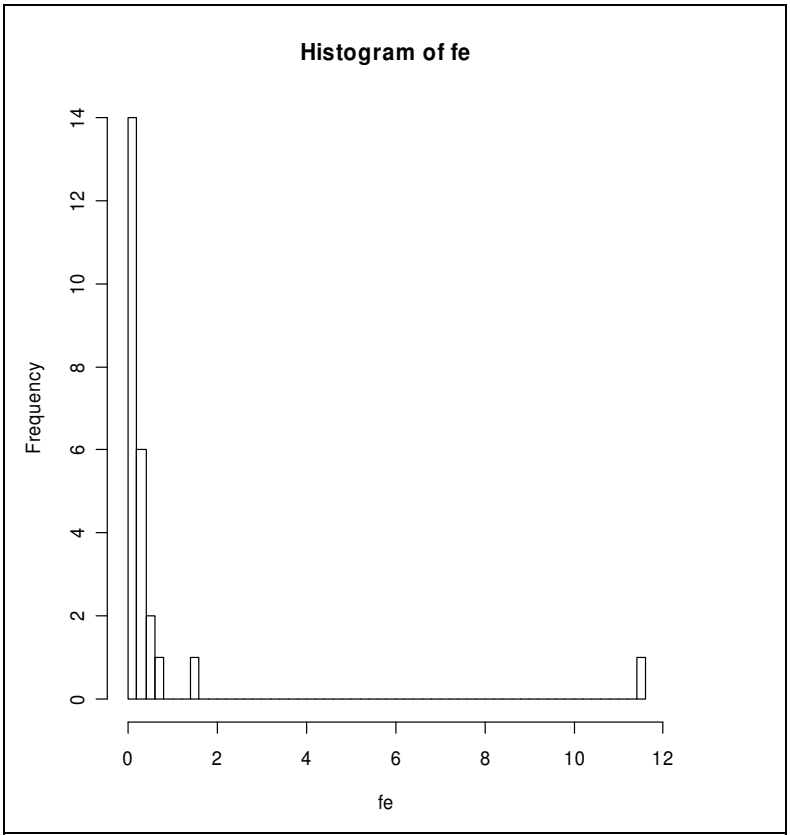
Nhóm 1 (gồm 10 bệnh nhân):

1.51, 0.07, 0.21, 0.29, 0.04, 0.03, 0.11, 0.00, 0.05, 0.00

Nhóm 2: (gồm 15 bệnh nhân):

0.08, 0.41, 11.60, 0.09, 0.00, 0.44, 0.03, 0.21, 0.28,0.36,
0.73, 0.05, 0.23, 0.06, 0.14

Chúng ta muốn kiểm định giả thiết FE của hai nhóm tương đương nhau. Một phương pháp "chuẩn" để xét nghiệm giả định này là kiểm định t (t-test). Nhưng biểu đồ sau đây cho thấy số liệu rất rời rạc. Phần lớn (75%) bệnh nhân có phần trăm FE thấp hơn 0.30%, và có một bệnh nhân có giá trị FE cao nhất (11.6%). Trong tất cả các hàm hoán chuyển như arsine, logarít, hàm mũ, v.v… đều không thể chuẩn hóa (normalize) số liệu. Do đó, phương pháp kiểm định t không thể ứng dụng trong trường hợp này.



Biểu đồ 1. Phân phối của FE trong hai nhóm bệnh nhân

Như đề cập trong phần đầu, một phương pháp thay thể cho kiểm định t là kiểm định Wilcoxon trong nhóm các phương pháp phi tham số. Giả định chính của phương pháp Wilcoxon là đối tượng của hai nhóm được chọn một cách ngẫu nhiên và độc lập nhau. Phương pháp Wilcoxon không đòi hỏi số liệu FE phải tuân theo luật phân phối chuẩn.

Phương pháp tính toán của kiểm định Wilcoxon có thể mô tả bằng các bước cụ thể như sau:

- **Bước 1**: tổng hợp số liệu của hai nhóm trong **Bảng 1** với nhau thành một nhóm chung:

```
1.51, 0.07, 0.21, 0.29, 0.04, 0.03, 0.11, 0.00, 0.05, 0.00
0.08, 0.41, 11.60, 0.09, 0.00, 0.44, 0.03, 0.21, 0.28,0.36,
0.73, 0.05, 0.23, 0.06, 0.14
```

- **Bước 2**: sắp xếp các giá trị từ thấp nhất đến cao nhất, cho hạng (rank), và tính tổng số của hạng:

0.00 0.00 0.00 0.03 0.03 0.04 0.05 0.05 0.06 0.07 0.08 0.09
0.11 0.14 0.21 0.21 0.23 0.28 0.29 0.36 0.41 0.44 0.73 1.51
11.60

cho hạng như sau:

FE	Hạng	Hạng liên kết
0.00	**1**	**2**
0.00	1	2
0.00	1	2
0.03	**2**	**4.5**
0.03	2	4.5
0.04	3	6
0.05	**4**	**7.5**
0.05	4	7.5
0.06	**5**	**9**
0.07	6	10
0.08	**7**	**11**
0.09	**8**	**12**
0.11	9	13
0.14	**10**	**14**
0.21	**11**	**15.5**
0.21	11	15.5
0.23	**12**	**17**
0.28	**13**	**18**
0.29	14	19
0.36	**15**	**20**
0.41	**16**	**21**
0.44	**17**	**22**
0.73	**18**	**23**
1.51	19	24
11.60	**20**	**25**
Tổng số nhóm 2		**221.5**

Chú ý: số liệu tô đậm là thuộc nhóm 2.

Chúng ta chú ý trong bảng trên, có 2 loại hạng: hạng đơn giản và **hạng liên kết** (**tied rank**). Hạng đơn giản là số hạng từ thấp nhất đến cao nhất, theo đó các đối tượng với FE 0 chúng ta cho hạng 1, và đối tượng với FE bằng 11.6 có hạng 20.

Nhưng ở đây có 3 đối tượng với FE bằng 0, 2 đối tượng với 0.3, 0.5 và 0.21; do đó, chúng ta cần phải "điều chỉnh" hạng cho các đối tượng này. Có nhiều phương pháp điều chỉnh, nhưng phương pháp phổ biến nhất là phương pháp bình quân. Theo phương pháp bình quân, khi *n* đối tượng có cùng giá trị chúng ta lấy tổng số vị trí chia cho *n*. Cụ thể trong trường hợp trên, có 3 đối tượng với FE 0, và vị trí của họ là 1, 2 và 3, cho nên tổng số vị trí là 1+2+3 = 6, và "hạng liên kết" do đó bằng 6/3 = 2. Tương tự, 2 đối tượng với FE 0.03, và tổng vị trí là 4+5 = 9, cho nên số hạng là 9/2 = 4.5. Vân vân ...

và tính tổng số hạng (gọi là S) cho nhóm 2 theo công thức [1] như sau:

```
S = 2 + 4.5 + 7.5 + 9 + ... + 22 + 23 + 25 = 221.5
```

- **Bước 3**: Ước tính chỉ số trung bình và phương sai của hạng theo công thức sau đây (xin nhắc lại một lần nữa – **<u>rất quan trọng</u>** – rằng, n_1 = 10 và n_2 = 15, cho nên chúng ta phải sắp xếp n_2 nằm ngoài ngoặc kép):

$$\mu_T = \frac{n_2\left(n_1+n_2+1\right)}{2} = \frac{15\left(10+15+1\right)}{2} = 195$$

và $$\sigma_T^2 = \frac{n_1 n_2\left(n_1+n_2+1\right)}{12} = \frac{10\times 15\left(10+15+1\right)}{12} = 325$$

Nói cách khác, độ lệch chuẩn là: $\sigma_T = \sqrt{325} = 18.0$.

- **Bước 4**: Ước tính chỉ số thống kê $T = \frac{S-\mu_T}{\sigma_T}$. Trong trường hợp trên, chúng ta có T = (221.5 – 130) / 18 = 1.47

  Theo lí thuyết phân phối chuẩn, nếu số đối tượng của hai nhóm trên 20, T có số trung bình là 0 và độ lệch chuẩn là 1. Nói cách khác, nếu hai nhóm hoàn toàn không khác nhau, thì chúng ta kì vọng rằng 95% trị số của T sẽ dao động trong khoảng -2 đến 2.

Nhưng trong thực tế, nghiên cứu này cho thấy trị số *T* là 1.47, tức cao hơn trị số kì vọng 2 gấp năm lần. Do đó, chúng ta kết luận rằng độ thải sodium giữa hai nhóm đối tượng không khác nhau.

Cũng có thể tính toán khoảng tin cậy 95% của thứ hạng như sau:

$$\mu_T \pm 1.96 \times \sigma_T = 195 \pm 1.96 \times 18$$
$$= 160 \text{ to } 230$$

Vì S = 221.5 nằm trong khoảng tin cậy này, chúng ta kết luận rằng hai nhóm tương đương nhau.

Các bước tính toán trên đây có vẻ rắc rối, nhưng trong thực tế, với phần mềm R, chỉ cần một lệnh duy nhất là chúng ta có kết quả và trị số p. (Xem **Chú thích 1** dưới đây)

3. Phương pháp bootstrap

Phương pháp bootstrap cần đến máy tính và phần mềm R, vì nó dựa vào lí thuyết **chọn mẫu ngẫu nhiên có hoàn lại** (sampling with replacement). Để giải thích khái niệm chọn mẫu ngẫu nhiên có hoàn lại, tôi sẽ lấy một ví dụ đơn giản như sau: Giả dụ chúng ta có số đo lường huyết áp từ 5 bệnh nhân. Hãy tạm xem đây là một quần thể. Chúng ta muốn tiến hành chọn ngẫu nhiên từ "quần thể" này 1000 lần, mỗi lần chọn 3 bệnh nhân. Chúng ta có thể làm thủ công như sau:

- Bước 1: đánh dấu số hiệu bệnh nhân: 1, 2, 3, 4, và 5. Bỏ các số hiệu này vào một cái rổ (xin lỗi bạn đọc nếu ngôn từ của tôi nghe hơi … thiếu khoa học, nhưng sự thật thì cái rổ là … cái rổ!);

- Bước 2: đưa tay vào rổ, chọn một số hiệu, ghi nhận số hiệu đó trong một tờ giấy, và bỏ lại số hiệu đó vào cái rổ;

- Bước 3: chọn số hiệu lần thứ 2, ghi nhận số hiệu vào tờ giấy, bỏ lại số hiệu vào rổ; tiếp tục chọn số hiệu lần thứ 3, ghi nhận số hiệu vào tờ giấy, bỏ lại số hiệu vào rổ;

- Bước 4: lặp lại bước 2 và 3 1000 lần.

Kết quả của việc chọn mẫu ngẫu nhiên có hoàn lại có thể như sau:

Lần đầu tiên, chọn: 3, 2, 1
Lần thứ hai, chọn: 1, 4, 3
Lần thứ 3: 2, 3, 4
Lần thứ 4: 2, 1, 1
Lần thứ 5: 2, 1, 2
…
Lần thứ 1000: 5, 4, 1

Chú ý rằng trong cách chọn ngẫu nhiên có hoàn lại như trên, có thể có một số mẫu bệnh nhân được chọn hơn 1 lần. Chẳng hạn như trong ví dụ trên, bệnh nhân 1 được chọn 2 lần trong lần chọn mẫu thứ 4, và bệnh nhân 2 được chọn 2 lần trong lần chọn mẫu thứ 5.

Và cứ mỗi lần chọn mẫu 3 bệnh nhân, chúng ta tính một chỉ số thống kê (như số trung bình, trung vị, phương sai, v.v…). Sau khi có 1000 mẫu, chúng ta có 1000 chỉ số thống kê, và qua đó, có thể ước tính khoảng tin cậy 95% của chỉ số này.

Phương pháp chọn mẫu như thế nhằm mục đích tạo ra nhiều mẫu ngẫu nhiên từ một mẫu, và qua cách chọn này, tập hợp những mẫu có thể đại diện cho một quần thể. Vì tính đại diện đó, việc suy luận thống kê cũng mang tính hợp lí của nó. Chính vì thế mà việc phát triển của phương pháp bootstrap được xem là một cuộc cách mạng quan trọng nhất trong khoa học thống kê ở thế kỉ 20 và 21.

4. Một ví dụ ứng dụng phương pháp bootstrap để kiểm định khác biệt giữa hai nhóm

Ví dụ 1 (tiếp tục): Chúng ta có số liệu FE từ hai nhóm đối tượng (nhóm 1 gồm 10 người và nhóm 2 gồm 15 người). Vì số liệu không tuân theo luật phân phối chuẩn, nên chúng ta sẽ tiến hành phân tích bằng phương pháp bootstrap. Để tiện cho việc theo dõi, tôi trình bày số liệu của 25 bệnh nhân dưới đây, và mô tả các bước phân tích cụ thể:

Nhóm 1: `1.51, 0.07, 0.21, 0.29, 0.04, 0.03, 0.11, 0.00, 0.05, 0.00`
Nhóm 2: `0.08, 0.41, 11.60, 0.09, 0.00, 0.44, 0.03, 0.21, 0.28,0.36, 0.73, 0.05, 0.23, 0.06, 0.14`

- **Bước 1**: chọn 10 đối tượng một cách ngẫu nhiên từ nhóm 1, và ước tính số trung vị cho nhóm 1 (tạm gọi là m_1);

- **Bước 2**: chọn 15 đối tượng một cách ngẫu nhiên từ nhóm 2, và ước tính số trung vị cho nhóm 2 (tạm gọi là m_2);

- **Bước 3**: tính độ khác biệt giữa hai nhóm (tạm gọi là *d*): $d = m_1 - m_2$

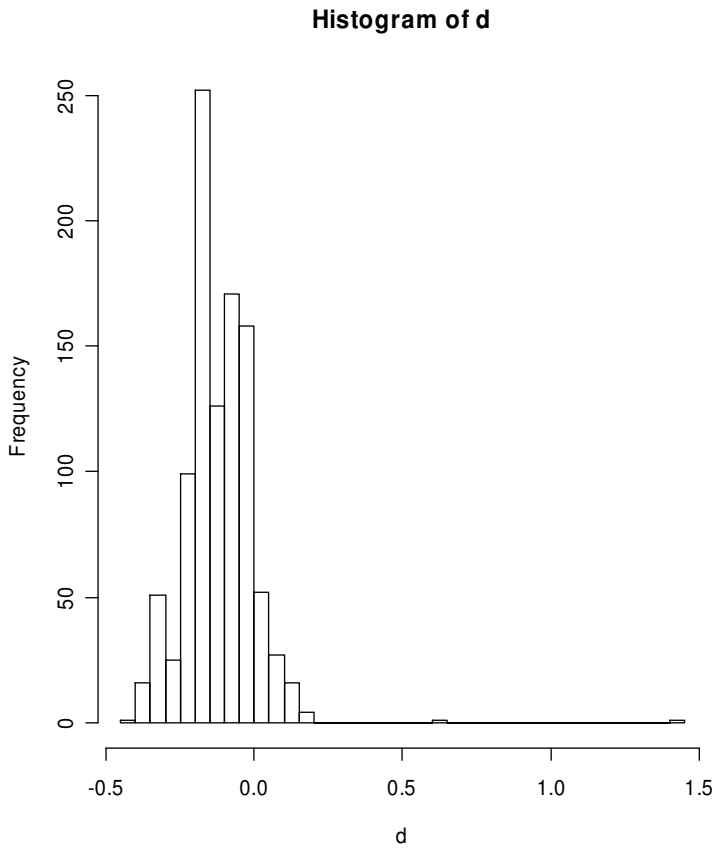
- **Bước 4**: lặp lại bước 1, 2 và 3 đến *n* lần (*n* có thể là 1000 lần hay 1 triệu lần – cẩn thận coi chứng máy tính bị "ngất" nếu cho *n* lớn quá!)

- **Bước 5**: sau khi xong bước 4, chúng ta đã có *n* số trung vị, do đó, bước sau cùng là ước tính số trung vị của *d* và khoảng dao động 95% của *d.*

Trong thực tế, các bước trên có thể cho ra kết quả như sau:

Lần chọn mẫu:	m1	m2	d
1	0.06	0.09	-0.03
2	0.04	0.21	-0.17
3	0.07	0.08	-0.01
4	0.035	0.21	-0.175
5	0.06	0.14	-0.08
6	0.29	0.21	0.08
7	0.09	0.36	-0.27
8	0.04	0.21	-0.17
9	0.06	0.14	-0.08
10	0.07	0.06	0.01
...	...	...	...

Sau 1000 lần chọn mẫu, chúng ta có số trung vị của *d* là -0.12 và khoảng tin cậy 95% từ -0.32 đến 0.08. Nói cách khác, xác suất 95% là FE ở bệnh nhân nhóm 1 có thể thấp hơn FE ở bệnh nhân nhóm 2 khoảng 0.32%, hay cao hơn nhóm 2 khoảng 0.08%. Biểu đồ phân phối của *d* như sau:



Biểu đồ 2. Phân phối của độ khác biệt giữa hai trung vị của nhóm 1 và nhóm 2 sau 1000 lần tái chọn mẫu qua phương pháp bootstrap

Vì khoảng tin cậy 95% bao gồm cả thấp hơn (số âm) và cao hơn (số dương), chúng ta phải kết luận rằng độ khác biệt về FE giữa hai nhóm không khác nhau. Kết luận này cũng nhất quán với kết quả phân tích phi tham số như trình bày trên. (Xem **Chú thích 2** về các bước trên bằng R).

5. Tóm lược

Trong tình huống số liệu không thể hoán chuyển để tuân theo luật phân phối chuẩn, có hai phương pháp phân tích để kiểm định độ khác biệt giữa hai mẫu: đó là phương pháp kiểm định Wilcoxon, và phương pháp bootstrap. Ngày nay, phương pháp bootstrap được ưa chuộng cho các phân tích mà số liệu không tuân theo luật phân phối chuẩn. Tuy "kĩ thuật" tính toán của phương pháp bootstrap có vẻ phức tạp, nhưng với phần mềm như R thì rất đơn giản. Bạn đọc có thể cắt và dán các lệnh dưới đây vào R và sẽ có kết quả trong vòng 1 phút!

Chú thích kĩ thuật:

Chú thích 1: Các mã R sau đây đã được sử dụng cho phân tích vừa trình bày trong bài viết.

```
# nhập số liệu của từng nhóm
fe1 <- c(1.51, 0.07, 0.21, 0.29, 0.04, 0.03, 0.11, 0.00, 0.05, 0.00)
fe2 <- c(0.08, 0.41, 11.60, 0.09, 0.00, 0.44, 0.03, 0.21, 0.28,0.36,
         0.73, 0.05, 0.23, 0.06, 0.14)

# tổng hợp số liệu 2 nhóm thành một nhóm chung, gọi biến đó là fe.
# ngoài ra, tạo thêm một biến group để phân biệt nhóm 1 có 10 bệnh nhân
# và nhóm 2 có 15 bệnh nhân
fe <- c(fe1, fe2)
group <- c(rep(1,10), rep(2,15))

# thử vẻ biểu đồ 1
hist(fe)

# sau đó đơn giản lệnh wilcox.test như sau:
wilcox.test(fe ~ group)

# Ghi chú thêm: Nếu bạn đọc muốn tính số hạng (rank) thì có thể
# sử dụng lệnh sort và ranks như sau:
# trước hết, sắp xếp fe theo thứ tự thấp đến cao:
sort(fe)
# sau đó lệnh rank
rank(sort(fe), ties.method="average")
```

Chú thích 2: Các mã R sau đây đã được sử dụng cho phân tích bootstrap trình bày trong phần 4 của bài viết:

```
# nhập số liệu vào hai biến gọi là fe1 và fe2 (cho hai nhóm)
fe1 <- c(1.51, 0.07, 0.21, 0.29, 0.04, 0.03, 0.11, 0.00, 0.05, 0.00)
fe2 <- c(0.08, 0.41, 11.60, 0.09, 0.00, 0.44, 0.03, 0.21, 0.28,0.36,
         0.73, 0.05, 0.23, 0.06, 0.14)
```

```
# xác định số đối tượng trong nhóm 1 và nhóm 2
n1 <- length(fe1)
n2 <- length(fe2)

# chúng ta sẽ chọn mẫu 1000 lần
N <- 1000

# tạo 3 vectors trống, gọi là m1, m2 và d, để chứa số trung vị
# của nhóm 1, nhóm 2, và khác biệt giữa 2 số trung vị
m1 <- numeric(N)
m2 <- numeric(N)
d <- numeric(N)

# Bắt đầu chọn mẫu, cứ mỗi mẫu, chúng ta tính độ khác biệt giữa hai
# nhóm và cho vào m1, m2 và d vừa tạo trên
for (i in 1:N)
{
  bs.fe1 <- sample(fe1, n1, replace=T)
  bs.fe2 <- sample(fe2, n2, replace=T)
  m1[i] <- median(bs.fe1)
  m2[i] <- median(bs.fe2)
  d[i] <- median(bs.fe1) - median(bs.fe2)
}

# ước tính khoảng tin cậy 95% và trung vị của d
quantile(d, probs=c(0.025, 0.50, 0.975))

# Vẽ biểu đồ 2 - phân phối của 1000 số d:
hist(d, breaks=50)
```