**Lâm sàng thống kê**

# Khoảng tin cậy 95% của trung vị

Nguyễn Văn Tuấn

Hỏi*: "Em đo một biến số lâm sàng, nhưng vì biến số này không tuân theo luật phân phối chuẩn, nên em phải dùng số trung vị để mô tả biến số. Em muốn biết cách tính khoảng tin cậy 95% của nó. Tìm trong sách giáo khoa không thấy sách nào chỉ cách tính này. Mong thầy chỉ cách tính khoảng tin cậy 95% của số trung vị."*

Đây là một vấn đề thú vị! Đối với các biến không tuân theo luật phân phối chuẩn, chúng ta không thể sử dụng số trung bình và độ lệch chuẩn để mô tả biến. Thay vào đó, chúng ta phải áp dụng các phương pháp thống kê phi tham số (non-parametric statistics) để tính. Một trong những chỉ số để mô tả trung bình của biến là số trung vị (median).

Đúng như bạn đọc viết, các sách giáo khoa không mô tả cách tính khoảng tin cậy 95% của số trung vị. Đơn giản vì … không có công thức nào để tính. Tuy nhiên, trong ba thập niên trở lại đây, với sự phát triển của máy tính, một cuộc cách mạng thống kê đã xảy ra. Phương pháp cách mạng đó có tên là "bootstrap method" do nhà thống kê học Bradley Efron phát triển vào năm 1979. Phương pháp bootstrap đã được ứng dụng rộng rãi trong nhiều lĩnh vực khoa học, và đến nay có thể xem là một phương pháp chuẩn. Trong bài này, tôi sẽ "lợi dụng" câu hỏi để giới thiệu phương pháp này. Vì phải sử dụng máy tính, cho nên bạn đọc cần phải biết qua một ngôn ngữ thống kê, chẳng hạn như R để tiện việc theo dõi. Chúng ta sẽ bắt đầu bằng một ví dụ cụ thể.

## Phương pháp ước tính số trung vị

**Ví dụ 1**. Số liệu về chỉ số đau (pain index) ở 11 bệnh nhân thấp khớp như sau:

```
0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, và 0.25
```

(Chú ý chỉ số càng cao, độ đau càng nghiêm trọng). Số trung bình của 11 bệnh nhân là 0.163 và độ lệch chuẩn 0.112. Vì số trung bình thấp hơn 2 lần độ lệch chuẩn, chúng ta có thể kết luận rằng biến số này không tuân theo luật phân phối chuẩn. Cách tính median có thể tiến hành qua hai bước đơn giản sau đây:

- Bước 1: Sắp xếp dữ liệu theo thứ tự từ thấp nhất đến cao nhất:

```
0.05, 0.05, 0.05, 0.05,  0.10, 0.15, 0.20, 0.25, 0.25, 0.30, 0.35
(1)   (2)   (3)   (4)    (5)   (6)   (7)   (8)   (9)   (10)  (11)
```

Chú ý: hàng thứ 2 (số trong ngoặc) là số thứ tự từ thấp đến cao.

- Bước 2: Xác định số giữa. Vì có 11 bệnh nhân, số giữa phải là số hàng thứ 6. Số hàng thứ 6 là 0.15 và đây chính là số trung vị:

```
0.05, 0.05, 0.05, 0.05,  0.10, 0.15, 0.20, 0.25, 0.25, 0.30, 0.35
```

## Phương pháp bootstrap

Vấn đề bây giờ là xác định khoảng tin cậy 95% của số trung vị. Nói cách khác, nếu nghiên cứu được lặp lại 1000 lần, và mỗi lần chọn 11 đối tượng, thì khoảng tin cậy của số trung vị ra sao. Phương pháp bootstrap rất có ích để giải quyết vấn đề. Phương pháp này được tiến hành như sau:

- Bước 1: Bắt đầu bằng mẫu gốc $x_1, x_2, x_3, \ldots, x_n$. Trong ví dụ trên:

```
0.05, 0.05, 0.05, 0.05,  0.10, 0.15, 0.20, 0.25, 0.25, 0.30, 0.35
```

- Bước 2: Chọn ngẫu nhiên *n* cá nhân từ mẫu gốc với qui trình *lấy mẫu có hoàn lại* (replacement sample). Mỗi lần chọn mẫu, tính số trung vị và tạm gọi số này là $m_i$.

  Cần giải thích thêm ở đây về phương pháp lấy mẫu có hoàn lại có nghĩa là một cá nhân có thể được hơn một lần trong một lần chọn mẫu. Chẳng hạn như từ quần thể 2, 3, 4, 5, lấy mẫu có hoàn lại có nghĩa là lần chọn mẫu thứ nhất có thể là 2, 4, 5, 2 (tức đối tượng thứ hai được chọn hai lần); lần thứ hai có thể là 4, 4, 2, 2, 5 (tức đối tượng thứ hai và thứ tư được chọn hai lần); lần thứ ba có thể là 2, 5, 2, 3; v.v...

- Bước 3: Lặp lại bước hai *N* lần (*N* thường là 1000 hay 10000 hay thậm chí 1 triệu – tùy theo nhu cầu). Trong trường hợp trên, 10 mẫu đầu tiên có thể là:

```
Mẫu 1: 0.05 0.05 0.10 0.05 0.20 0.20 0.05 0.25 0.10 0.10 0.30  →  0.10
Mẫu 2: 0.05 0.25 0.30 0.05 0.30 0.30 0.05 0.05 0.25 0.05 0.35  →  0.25
Mẫu 3: 0.35 0.10 0.05 0.25 0.05 0.05 0.20 0.25 0.15 0.25 0.10  →  0.15
Mẫu 4: 0.05 0.05 0.10 0.25 0.15 0.05 0.20 0.05 0.10 0.25 0.05  →  0.10
Mẫu 5: 0.30 0.25 0.05 0.25 0.25 0.05 0.20 0.05 0.25 0.05 0.05  →  0.20
Mẫu 6: 0.05 0.25 0.10 0.05 0.05 0.15 0.25 0.05 0.05 0.05 0.05  →  0.05
Mẫu 7: 0.05 0.15 0.25 0.05 0.05 0.30 0.20 0.25 0.30 0.05 0.35  →  0.20
Mẫu 8: 0.05 0.05 0.20 0.05 0.10 0.05 0.05 0.10 0.20 0.10 0.05  →  0.05
Mẫu 9: 0.05 0.05 0.10 0.25 0.20 0.25 0.25 0.20 0.35 0.25 0.35  →  0.25
Mẫu 10: 0.05 0.05 0.05 0.25 0.35 0.25 0.25 0.15 0.20 0.20 0.15 →  0.20
```

v.v...

(Chú ý: số phía sau → có nghĩa là số trung vị cho từng mẫu)

- Bước 4: Đến đây chúng ta có *N* số trung vị. Sắp xếp N số từ thấp đến cao và đánh số: 1, 2, 3, …, *N*. Chọn số trung vị ở hạng 2.5% và 97.5% của *N* số trung vị, và đó chính là khoảng tin cậy 95%. Chẳng hạn như nếu *N* = 1000 lần, thì khoảng tin cậy 95% chính là số trung vị hàng thứ 25 và 975.

Các bước tính toán trên có thể thực hiện bằng ngôn ngữ R (hay một ngôn ngữ hay phần mềm nào mà bạn đọc quen thuộc) rất dễ dàng. Đối với R, các mã sử dụng (và giải thích kèm theo) như sau:

```
# nhập các số liệu gốc vào một vector có tên là x
x <- c(0.05,0.15,0.35,0.25,0.20,0.05,0.10,0.05,0.30,0.05,0.25)

# bước 2 – xác định xem có bao nhiêu số liệu trong vector x
n = length(x)

# muốn lấy 1000 mẫu từ số liệu gốc
B = 1000

# tạo một vector mới để chứa số trung vị
median = numeric(B)

# bắt đầu lấy B mẫu và mỗi mẫu tính toán số trung vị
for (i in 1:B)
  {
    bs.sample <- sample(x, n, replace=T)
    median[i] = median(bs.sample)
 }
# ước tính khoảng tin cậy 95%
quantile(median, probs=c(0.025, 0.975))
```

Chương trình trên sẽ báo cho chúng ta biết khoảng tin cậy 95% của số trung vị là 0.05 đến 0.25.

## Tóm tắt

Phương pháp bootstrap có thể áp dụng để tính toán khoảng tin cậy 95% (hay bất cứ độ tin cậy nào) cho nhiều thông số "bất thường" khác, chứ chẳng riêng gì số trung vị. Đây là một phương pháp rất hữu hiệu và, như đề cập trên, được đánh giá là một cuộc cách mạng quan trọng trong khoa học thống kê.

Vì phương pháp bootstrap đòi hỏi có máy tính, và do đó, người sử dụng phải am hiểu một ngôn ngữ hay phần mềm thống kê. Trong bài này, tôi sử dụng ngôn ngữ R để thực hiện phương pháp bootstrap, vì R là một ngôn ngữ tương đối dễ sử dụng nhưng rất linh hoạt để tính toán các vấn đề khó trong thực tế nghiên cứu lâm sàng. Bạn đọc muốn biết thêm về ngôn ngữ R có thể tìm đọc cuốn sách "Phân tích số liệu và tạo biểu đồ bằng R" của tôi, do Nhà xuất bản Khoa học Kỹ thuật phát hành đầu năm 2007. Trong đó có phần hướng dẫn cách chọn mẫu như sử dụng trong bài viết này.

## Thuật ngữ sử dụng trong bài viết

| **Tiếng Việt** | **Tiếng Anh** |
|---|---|
| Thống kê phi tham số | Non-parametric statistics |
| Trung vị | Median |
| Khoảng tin cậy 95% | 95% confidence interval |