Lâm sàng thống kê 19

Phương pháp kiểm định outlier

Nguyễn Văn Tuấn

Một bạn đọc từ Hà Nội làm thí nghiệm liên quan đến bệnh đái tháo đường (trên chuột). Trong thí nghiệm này, anh có hai nhóm chuột: nhóm thứ nhất được cho uống thuốc và nhóm thứ hai là nhóm chứng (không uống thuốc). Ở mỗi chuột, một chỉ số lâm sàng được đo 4 lần: lúc ban đầu (chưa uống thuốc, tạm gọi là T0), 2 giờ, 3 giờ, và 4 giờ sau khi uống thuốc (tạm kí hiệu T2, T3 và T4). Kết quả của thí nghiệm như sau:

Bảng 1. Nồng độ glucose của nhóm chuột được điều trị và nhóm chứng

Treatment	id	T0	T2	T3	T4
Test	1	5.9	3.9	3.9	3.6
Test	2	5.3	4.7	3.5	3.2
Test	3	4.6	3.7	3.3	3.2
Test	4	6.2	4.6	4.3	3.9
Test	5	6.0	5.4	5.2	4.8
Test	6	6.4	4.7	4.8	4.3
Test	7	7.6	4.1	3.8	4.1
Test	8	5.9	3.1	3.6	3.3
Test	9	7.5	6.1	5.4	4.6
Control	10	6.2	5.3	4.9	4.5
Control	11	6.9	5.6	5.9	5.9
Control	12	5.6	4.7	4.6	4.0
Control	13	5.1	3.9	2.9	2.9
Control	14	5.7	4.7	4.3	4.6
Control	15	5.0	4.0	3.5	3.3
Control	16	5.2	4.2	4.0	3.8
Control	17	7.7	6.2	6.1	5.7
Control	18	8.0	5.8	6.5	6.0
Control	19	7.7	5.0	6.3	6.2

Chú thích: *id* là cột chỉ mã số của chuột.

Bạn đọc tính phần trăm thay đổi từ T2, T3 và T4 so với T0. Kết quả như sau:

Treatment	id	PT0	PT2	PT3	PT4
Test	1	1	0.339	0.339	0.390
Test	2	1	0.113	0.340	0.396

Test	3	1	0.196	0.283	0.304
Test	4	1	0.258	0.306	0.371
Test	5	1	0.100	0.133	0.200
Test	6	1	0.266	0.250	0.328
Test	7	1	0.461	0.500	0.461
Test	8	1	0.475	0.390	0.441
Test	9	1	0.187	0.280	0.387
Control	10	1	0.145	0.210	0.274
Control	11	1	0.188	0.145	0.145
Control	12	1	0.161	0.179	0.286
Control	13	1	0.235	**0.431**	**0.431**
Control	14	1	0.175	0.246	0.193
Control	15	1	0.200	0.300	0.340
Control	16	1	0.192	0.231	0.269
Control	17	1	0.195	0.208	0.260
Control	18	1	0.275	0.188	0.250
Control	19	1	0.351	0.182	0.195

Trong bảng trên, bạn đọc chú ý có hai số liệu đáng ngờ (màu đỏ): Đó là số liệu cho chuột mang mã số 13 (nhóm chứng), mà tỉ lệ giảm đến 43% (từ 5.1 trước khi uống thuốc – thật ra thì không uống thuốc vì đây là nhóm chứng – xuống còn 2.9, trong khi mức thay đổi sau 2 giờ uống thuốc chỉ 23.5%. Bạn đọc viết thư hỏi tôi đó có phải là outlier hay không và cách phân tích như thế nào.

Đây là một vấn đề khá thú vị, và cũng khá phổ biến. Bất cứ ai làm khoa học thực nghiệm cũng từng trải qua các dữ liệu mà thoạt đầu mới nhìn qua thì có vẻ .. lạ lùng. Có người "ăn gian" thì bỏ nó đi; có người gian lận thì sửa nó (đây là hành vi vi phạm đạo đức khoa học nguy hiểm vì nếu bị phát hiện thì sẽ bị đuổi việc hay kỉ luật). Nhưng bạn đọc của chúng ta là một nghiên cứu sinh nghiêm chỉnh, nên anh viết thư hỏi cách xử lí.

Outlier là gì? Trong cuốn sách "Statistical Design and Analysis of Experiments", các tác giả Mason, Gunst, và Hess định nghĩa outlier như sau: "*Outliers are observations that have extreme values relative to other observations observed under the same conditions. Observations may be outliers because of a single large or small value of one variable or because of an unusual combination of values of two or more variables*." Tạm dịch: outlier là các giá trị cực so với các giá trị khác được quan sát trong cùng một điều kiện. Outlier có thể là một giá trị đơn lẻ, nhưng cũng có thể là giá trị từ hai hay nhiều biến số.

Vấn đề ở đây là "giá trị cực" hay extreme values? Thế nào là giá trị cực? Thật là khó trả lời. Không có câu trả lời định tính, nhưng có thể có câu trả lời định lượng. Có nhiều cách để đánh giá xem một số liệu có phải là outlier hay không.

Phương pháp kiểm định outlier

Trước khi bàn về các phương pháp này, chúng ta cần vài kí hiệu để dễ hiểu hơn. Giả dụ chúng ta có một biến số *X* gồm 100 giá trị được sắp xếp từ thấp nhất đến cao nhất như sau:

$$x_1,\ x_2,\ x_3,\ldots,\ x_{100}$$

tức là x_1 là số thấp nhất và x_{100} là số cao nhất.

Phương pháp đơn giản nhất là dựa vào giả định phân phối chuẩn (normal distribution). Chúng ta biết rằng nếu biến số *X* tuân theo luật phân phối chuẩn với trung bình *m* và độ lệch chuẩn *s* thì 99% các giá trị của *X* phải nằm trong khoảng $m-3\times s$ đến $m+3\times s$. Do đó, bất cứ số x_i nào có giá trị thấp hơn $m-3\times s$ hay cao hơn $m+3\times s$ thì có thể nghi ngờ là outlier.

Phương pháp dựa vào số trung vị. Một phương pháp đơn giản khác là dựa vào số trung vị. Phương pháp này có thể tiến hành qua các bước như sau:

- Tính trung vị của biến số; gọi số này là *M* ;
- Tính độ khác biệt tuyệt đối giữa từng số trong biến *X* và *M,* và gọi kết quả là d_i :

$$d_i = x_i - M$$

- Tính trung vị của d_i và gọi là *Md*;
- Lấy d_i chia cho *Md*, và gọi chỉ số này là t_i :

$$t_i = \frac{d_i}{Md}$$

- Nếu t_i cao hơn 4.5, chúng ta có thể gọi đó là outlier.

Có thể sử dụng Excel để thực hiện các bước vừa mô tả. Nhưng các lệnh R dưới đây cũng có thể giải quyết phương pháp này rất đơn giản.

Phương pháp phi tham số. Nhưng nếu *X* không tuân theo luật phân phối (non-normal distribution) thì chúng ta có một **phương pháp phi tham số (non-parametric method)** để kiểm định xem có outlier trong biến số hay không. Phương pháp này khá đơn giản và có thể tóm lược như sau:

- Tìm giá trị bách phân 25 (25th percentile) của biến *X*. Nói cách khác, chúng ta tìm x_{25} trong biến số trên. Gọi trị số này là Q1.

- Tìm giá trị bách phân 75 (75th percentile) của biến *X*. Nói cách khác, chúng ta tìm x_{75}. Gọi trị số này là Q3 (đây là cách gọi thông thường trong các phần mềm thống kê).

- Tính độ khác biệt giữa Q1 và Q3 bằng công thức: IQR = Q3 – Q1.

- Tính giá trị thấp của biến và gọi đó là L (tức lower): $L = Q1 - 1.5 \times IQR$.

- Tính giá trị cao của biến và gọi đó là U (upper): $U = Q3 + 1.5 \times IQR$.

- Nếu trong dãy số x_1, x_2, x_3,...., x_{100} có số nào thấp hơn *L* hay cao hơn *U,* thì có thể xem đó là outlier.

Một phương pháp khác nữa có tên là phương pháp Dixon hay kiểm định Dixon. Phương pháp này có thể thực hiện bằng phần mềm R, nhưng tôi thấy khá phức tạp và cũng không cần thiết cho các vấn đề đơn giản nên không giải thích ở đây. Nếu cần bạn đọc có thể tìm hiểu phương pháp này trong các sách giáo khoa hay phần mềm R.

Trong các phương pháp vừa mô tả, tôi thấy phương pháp phi tham số là phổ biến nhất và cũng dễ ứng dụng nhất. Trong thực tế, các phương pháp này có kết quả rất giống nhau. Do đó, trong bài này tôi sẽ sử dụng phương pháp phi tham số để xét xem số liệu của bạn đọc có outlier hay không.

Kiểm định outlier

Có một vấn đề cần phải thảo luận ở đây: chúng ta nên xem xét outlier ở mỗi chuột hay giữa các chuột với nhau? Trong trường hợp này, mỗi chuột chỉ được đo 4 lần, và số liệu quá ít để có thể tính giá trị bách phân 25 hay 75. Vì thế, chúng ta không thể (và không nên) xác định outlier cho từng chuột.

Có thể xem các giá trị đo lường trong thí nghiệm này là một cách lấy mẫu ngẫu nhiên từ một quần thể với nhiều nhóm và nhiều đặc tính khác nhau. Chúng ta sẽ không quan tâm đến phẩm trăm thay đổi giữa các thời điểm, nhưng chỉ quan tâm đến số liệu gốc. Do đó, cách kiểm định outlier đơn giản nhất là sắp xếp dữ liệu thành một vector hay một biến số. Cách sắp xếp số liệu sao cho mỗi chuột có một dòng, và dòng này phải mã số được chuột thuộc vào nhóm nào và thời điểm nào. Nói cách khác, chúng ta biến số liệu từ Bảng 1 thành Bảng 2 dưới đây:

Bảng 2. Kết quả thí nghiệm được sắp xếp theo dòng

Treatment	id	Time	Y
Test	1	0	3.9
Test	1	2	4.7
Test	1	3	3.7

Test	1	4	4.6
Test	2	0	5.3
Test	2	2	4.7
Test	2	3	3.5
Test	2	4	3.2
vân vân …	…	…	…
Control	10	0	6.2
Control	10	2	5.3
Control	10	3	4.9
Control	10	4	4.5
…	…	…	…
Control	19	0	7.7
Control	19	2	5.0
Control	19	3	6.3
Control	19	4	6.2

Nói cách khác, thay vì mỗi chuột có 4 cột, chúng ta sắp xếp lại mỗi chuột có 4 dòng, nhưng **mỗi** dòng phải được nhận dạng rõ ràng thuộc nhóm nào và thời điểm nào. Chúng ta có 19 x 4 = 76 dòng như trên. Điều này rất cần thiết cho phân tích sau này. (Xin nói thêm rằng phần lớn phân tích sai lầm cũng vì cách tổ chức số liệu không đúng; do đó cách sắp xếp số liệu là một bước cực kì quan trọng trong phân tích thống kê).

Bây giờ chúng ta chỉ có một cột để phân tích (cột đó được kí hiệu là Y như bảng số liệu trên). Chúng ta thử nhập số liệu vào R và sử dụng package `Design` để phân tích:

```
library(Design)

# nhập số liệu và gọi biến y

y = c(5.9, 3.9, 3.9, 3.6, 5.3, 4.7, 3.5, 3.2, 4.6, 3.7,
      3.3, 3.2, 6.2, 4.6, 4.3, 3.9, 6.0, 5.4, 5.2, 4.8,
      6.4, 4.7, 4.8, 4.3, 7.6, 4.1, 3.8, 4.1, 5.9, 3.1,
      3.6, 3.3, 7.5, 6.1, 5.4, 4.6, 6.2, 5.3, 4.9, 4.5,
      6.9, 5.6, 5.9, 5.9, 5.6, 4.7, 4.6, 4.0, 5.1, 3.9,
      2.9, 2.9, 5.7, 4.7, 4.3, 4.6, 5.0, 4.0, 3.5, 3.3,
      5.2, 4.2, 4.0, 3.8, 7.7, 6.2, 6.1, 5.7, 8.0, 5.8,
      6.5, 6.0, 7.7, 5.0, 6.3, 6.2)

# tạo biến số id
# tạo biến số time có giá trị 0,2,3,4 lặp lại 19 lần
# tạo biến số treatment
# mô tả biến y

id = rep(1:19, each=4)
time = rep(c(0,2,3,4), 19)
treatment = rep(1:2, c(9*4, 10*4))
```

```
describe(y)
```

Kết quả như sau:

```
n  missing  unique  Mean  .05     .10     .25     .50     .75     .90     .95
76   0      38    4.959   3.200   3.400   3.975   4.750   5.900   6.350   7.525

lowest : 2.9 3.1 3.2 3.3 3.5, highest: 6.9 7.5 7.6 7.7 8.0
```

Nhìn qua kết quả trên, chúng ta thấy Q1 = 3.975 và Q3 = 5.90, và IQR = 1.925. Do đó, giá trị thấp và cao của biến là:

$$L = 3.975 – IQR*1.5 = 1.09$$

$$U = 5.90 + IQR*1.5 = 8.79$$

Nhưng kết quả trên cho chúng ta thấy số liệu thấp nhất của biến Y là 2.9 và cao nhất là 8.0; tất cả đều nằm trong vòng L và U. Dựa vào phân tích này chúng ta có thể nói không có số liệu nào trong nghiên cứu này được xem là outlier.

Các bạn có thể kiểm định outlier bằng phương pháp dựa vào phân phối chuẩn như tôi vừa mô tả phần trên. Phương pháp này cũng cho kết quả không có outlier. Một vấn đề khác, số liệu này không hẳn tuân theo luật phân phối chuẩn, nhưng cũng không quá lệch. Do đó, các bạn có thể hoán chuyển số liệu sang đơn vị logarithm và kiểm định outlier dựa trên biến số mới này. Nói theo "ngôn ngữ" R:

```
log.y = log(y)
describe(log.y)
```

Nhưng tôi để cho các bạn thực hành. Trong bài sau, tôi sẽ bàn về cách phân tích số liệu của thí nghiệm này. Đây là một trường hợp lí tưởng để các bạn có thể thực hành phân tích theo mô hình mixed-effects, một mô hình được xem là state-of-the-art hiện nay.

Chú thích:

1. Phương pháp phi tham số

Bạn đọc quen với ngôn ngữ R có thể tiến hành các phân tích trên như sau:

```
y = c(5.9, 3.9, 3.9, 3.6, 5.3, 4.7, 3.5, 3.2, 4.6, 3.7,
      3.3, 3.2, 6.2, 4.6, 4.3, 3.9, 6.0, 5.4, 5.2, 4.8,
      6.4, 4.7, 4.8, 4.3, 7.6, 4.1, 3.8, 4.1, 5.9, 3.1,
      3.6, 3.3, 7.5, 6.1, 5.4, 4.6, 6.2, 5.3, 4.9, 4.5,
      6.9, 5.6, 5.9, 5.9, 5.6, 4.7, 4.6, 4.0, 5.1, 3.9,
      2.9, 2.9, 5.7, 4.7, 4.3, 4.6, 5.0, 4.0, 3.5, 3.3,
      5.2, 4.2, 4.0, 3.8, 7.7, 6.2, 6.1, 5.7, 8.0, 5.8,
      6.5, 6.0, 7.7, 5.0, 6.3, 6.2)
```

```
q1 = quantile(y, prob=0.25)
q3 = quantile(y, prob=0.75)
iqr = q3-q1
L = q1-1.5*iqr
U = q3+1.5*iqr
outlier <- ifelse(y < L | y > U, "*", "OK")
cbind(y, outlier)
```

Bạn chỉ cần cắt và dán các lệnh trên vào R là sẽ thấy kết quả.

2. Phương pháp dựa vào trung vị

Các mã sau đây sẽ kiểm định outlier theo phương pháp dựa vào số trung vị:

```
y = c(5.9, 3.9, 3.9, 3.6, 5.3, 4.7, 3.5, 3.2, 4.6, 3.7,
      3.3, 3.2, 6.2, 4.6, 4.3, 3.9, 6.0, 5.4, 5.2, 4.8,
      6.4, 4.7, 4.8, 4.3, 7.6, 4.1, 3.8, 4.1, 5.9, 3.1,
      3.6, 3.3, 7.5, 6.1, 5.4, 4.6, 6.2, 5.3, 4.9, 4.5,
      6.9, 5.6, 5.9, 5.9, 5.6, 4.7, 4.6, 4.0, 5.1, 3.9,
      2.9, 2.9, 5.7, 4.7, 4.3, 4.6, 5.0, 4.0, 3.5, 3.3,
      5.2, 4.2, 4.0, 3.8, 7.7, 6.2, 6.1, 5.7, 8.0, 5.8,
      6.5, 6.0, 7.7, 5.0, 6.3, 6.2)
median = median(y)
abs.dev = abs(y - median)
median.abs.dev = median(abs.dev)
stat <- abs.dev/median.abs.dev
outlier2.5 <- ifelse(stat>=2.5, 'outlier', 'ok')
outlier4.5 <- ifelse(stat>=4.5, 'outlier', 'ok')
cbind(y, outlier2.5, outlier4.5)
```