Lâm sàng thống kê

Làm cách nào để chọn ngẫu nhiên

Hỏi*: "Em muốn làm một nghiên cứu trong những bệnh nhân em khám hàng tuần, Thầy nói phải chọn ngẫu nhiên thì kết quả mới có ý nghĩa khoa học. Vậy xin Thầy chỉ cách chọn ngẫu nhiên. Nếu em chọn mỗi bệnh nhân thứ 3 hay thứ năm có thể xem là ngẫu nhiên không?"*

Đây là một câu hỏi liên quan đến vấn đề thiết kế nghiên cứu. Liên quan đến phần hai của câu hỏi, trả lời ngắn gọn là: "không". Cách chọn theo thứ tự bệnh nhân thứ 3, 6, 9, … (hay 5, 10, 15, 20, …) thì không thể xem là ngẫu nhiên được, bởi vì cách chọn đã nói lên rằng đây là cách chọn có hệ thống!

Thế thì định nghĩa "chọn ngẫu nhiên" là gì? Chọn ngẫu nhiên có nghĩa là chọn đối tượng sao cho tất cả các đối tượng trong một quần thể có xác suất được chọn như nhau. Nếu chúng ta có 10 đối tượng, thì mỗi đối tượng có xác suất được chọn là 1/10. Nếu chúng ta có hai nhóm A và B, và chọn ngẫu nhiên có nghĩa là đối tượng được chọn vào nhóm A có xác suất bằng với đối tượng được chọn vào nhóm B (tức là 50%).

Ý nghĩa của việc chọn ngẫu nhiên rất quan trọng trong nghiên cứu y học và triết lí của nghiên cứu khoa học. Tất cả các mô hình phân tích thống kê đều giả định rằng mẫu được chọn phải là mẫu ngẫu nhiên. Chỉ khi nào mẫu ngẫu nhiên thì kết quả phân tích mới có giá trị khoa học cao. Ngoài ra, trong các nghiên cứu bệnh chứng (case-control study) khi so sánh hai nhóm, chúng ta cần phải đảm bảo hai nhóm tương đương nhau về các yếu tố lâm sàng có thể có ảnh hưởng đến kết quả nghiên cứu. Chẳng hạn như nếu chúng ta muốn tìm hiểu ảnh hưởng của thói quen hút thuốc lá đến nguy cơ ung thư phổi, chúng ta có thể so sánh tỉ lệ ung thư giữa nhóm hút thuốc lá và nhóm không hút thuốc lá. Nhưng như thế vẫn chưa đủ, vì các yếu tố khác như độ tuổi, hormone, môi trường sống, v.v… (gọi chung bằng thuật ngữ "covariates") cũng có thể gây ung thư. Do đó, vấn đề là phải chọn hai nhóm tương đương nhau về những covariates này. Chỉ khi nào hai nhóm có cùng (hay tương đương) về các yếu tố covariates thì kết luận về mối liên hệ giữa hút thuốc lá và ung thư mới đáng tin cậy.

Nhưng cách phân chia đối tượng sao cho hai nhóm tương đương nhau rất khó làm bằng phương pháp thủ công, bởi vì chúng ta hoàn toàn có thể chọn hai nhóm tương đương nhau về độ tuổi, nhưng có thể lại khác nhau về hormone. Hay chúng ta có thể phân chia đối tượng sao cho hai nhóm tương đương nhau về độ tuổi và hormone, nhưng có thể hai nhóm không tương đương về môi trường sống. Số lượng covariates càng nhiều, cách phân chia càng phức tạp. Chỉ có cách duy nhất là ngẫu nhiên hóa (randomization) thì mới đảm bảo tương đương giữa hai nhóm.

Mỗi chúng ta (trong thế giới 4 tỉ người) đều là những cá thể duy nhất, hiểu theo nghĩa không có ai giống ai, và sự "độc nhất vô nhị" đó được định nghĩa bằng những những đặc điểm và những đặc tính liên quan đến mỗi cá nhân. Có thể hai người có cùng chiều cao, cùng cân nặng, cùng độ tuổi, nhưng hai người đó có thể khác nhau về các đặc điểm lâm sàng khác, và nhất là khác nhau về môi trường sống. Vì thế, nếu chúng ta chọn đối tượng dựa vào một hay hai đặc tính thì vẫn chưa đủ, mà phải chọn sao cho hoàn toàn ngẫu nhiên. Đây là triết lí đằng sau của các nghiên cứu lâm sàng đối chứng ngẫu nhiên (randomized clinical trial). Qua nhiều năm kinh nghiệm, y học đã hoàn thiện và chứng minh rằng cách ngẫu nhiên hóa thực sự tương đồng hóa các nhóm.

Máy tính có thể giúp chúng ta chọn hay phân chia ngẫu nhiên. Điều cần thiết là chúng ta phải có một phần mềm thống kê. Ở đây, tôi sẽ sử dụng phần mềm R để ngẫu nhiên hóa. Bạn đọc muốn biết thêm về R có thể tham khảo cuốn sách "Phân tích số liệu và tạo biểu đồ bằng R" của tôi do Nhà xuất bản Khoa học Kỹ thuật vừa mới phát hành năm 2007.

Phương pháp chọn ngẫu nhiên

Quay lại câu hỏi trên, giả sử bạn đọc biết rằng mỗi tháng số bệnh nhân đến khám là 500 người, và công trình nghiên cứu cần 100 người. Cách chọn ngẫu nhiên 100 bệnh nhân có thể tiến hành từng bước như sau:

- Bước 1: lên danh sách từ 1 đến 500 (tức quần thể nghiên cứu). Đối với R việc này cực kì đơn giản với lệnh:

  ```
  population <- 1:500
  ```

  Lệnh này có nghĩa rằng chúng ta làm một danh sách từ 1, 2, 3, … đến 500, và chứa danh sách này trong biến có tên là `population`.

- Bước 2: sử dụng hàm `sample` để lấy mẫu ngẫu nhiên. Nên nhớ, chúng ta muốn chọn 100 bệnh nhân từ `population`, và hàm `sample` đơn giản như sau:

  ```
  selected <- sample(population, 100)
  ```

  Lệnh này có nghĩa rằng chúng ta muốn chọn ngẫu nhiên 100 người từ quần thể có tên là `population` và lưu trữ danh sách này trong biến có tên là `selected`.

- Bước 3: In ra 100 đối tượng vừa mới chọn đó:

```
selected
```

và R sẽ cho chúng ta biết:

```
  [1]  42 172  31  22 234 432  75 190 386 183  64 291 139 323 356  68 462 485
 [19]  61 253 456 484 337 363 488 136 498 113 117 197 378 406 256 476 466 351
 [37]  95   1 218 300 219  69  28  43 250 239 326 303  84 210   3 162 493  36
 [55] 425 368 182 233  57 311  51 282  93 100 130  70  18  74 446 376 321 103
 [73] 125 344 500 391  34 161  78 349 252 265 147 289   9 342 231 395  73  13
 [91] 180 400   6 414 367 137  81 155 360 187
```

(Bạn đọc có thể không cần lưu ý đến những số như [1], [19], [37], v.v… vì đây là những số cho chúng ta biết vị trí khởi đầu của từng dòng số liệu).

Theo kết quả trên, chúng ta nên chọn các bệnh nhân số 42, 172, 31, v.v… Nhưng danh sách này khó sử dụng, vì chúng ta biết rằng bệnh nhân đến khám theo thứ tự với mã 1, 2, 3, …, 500. Vì thế, cần phải sắp xếp biến `selected` theo thứ tự, và hàm `sort` giúp chúng ta làm việc này rất hữu hiệu:

```
sort(selected)
```

và R sẽ cho chúng ta biết:

```
  [1]   1   3   6   9  13  18  22  28  31  34  36  42  43  51  57  61  64  68
 [19]  69  70  73  74  75  78  81  84  93  95 100 103 113 117 125 130 136 137
 [37] 139 147 155 161 162 172 180 182 183 187 190 197 210 218 219 231 233 234
 [55] 239 250 252 253 256 265 282 289 291 300 303 311 321 323 326 337 342 344
 [73] 349 351 356 360 363 367 368 376 378 386 391 395 400 406 414 425 432 446
 [91] 456 462 466 476 484 485 488 493 498 500
```

Bây giờ thì chúng ta đã có một danh sách ngẫu nhiên. Theo danh sách này, bệnh nhân đầu tiên (sô 1), tiếp theo là bệnh nhân số 3, 6, … và 500 nên được chọn.

Cần chú ý rằng vì đây là cách chọn hoàn toàn ngẫu nhiên, cho nên cứ mỗi lần chúng ta ra 3 lệnh trên thì R cung cấp một danh sách hoàn toàn mới. Bạn đọc có thể kiểm tra phát biểu này bằng cách ra 3 lệnh trên như sau:

```
population <- 1:500
selected <- sample(population, 100)
sort(selected)
```

Vì lí do này, chúng ta chỉ cần chọn một lần, không cần phải chọn nhiều lần. Sau khi chọn, chúng ta lên danh sách và lưu trữ trong một phần mềm khác như Excel chẳng hạn để sử dụng và theo dõi sau này.

Phương pháp phân nhóm ngẫu nhiên

Trong các nghiên cứu lâm sàng đối chứng ngẫu nhiên, chúng ta thường có hai nhóm đối tượng. Với số lượng cỡ mẫu định trước là *n,* mục tiêu là chia *n*/2 đối tượng vào nhóm 1 và *n*/2 vào nhóm 2. Có vài phương pháp để chia ngẫu nhiên. Cách đơn giản nhất là lấy số chẵn hay lẻ để quyết định phân nhóm. Chẳng hạn nếu đối tượng được chọn [ngẫu nhiên] là số chẵn thì sẽ cho vào nhóm 1 và số lẻ vào nhóm 2 (hay ngược lại). Với R chúng ta có thể tiến hành phân nhóm cực kì đơn giản.

Ví dụ 1: **Phân nhóm tổng thể**. Giả sử chúng ta có 100 bệnh nhân và muốn phân 50 vào nhóm can thiệp (A) và 50 vào nhóm đối chứng (P). Chúng ta tiến hành theo trình tự sau đây:

- Bước 1: Cho biết chúng ta có 100 đối tượng và tạo 100 mã số và cho vào biến `id`.

```
n <- 100
id <- 1:n
```

- Bước 2: Dùng hàm `runif` để tạo một biến ngẫu nhiên mới với 100 đối tượng. Hàm `runif` cho ra những số từ 0 đến 1 (với nhiều số thập phân), cho nên chúng ta cần phải hoán chuyển thành số nguyên (integer) bằng cách nhân cho 100 và sư dụng hàm `as.integer`:

```
random <- runif(n)
int <- as.integer(random*100)
```

Có thể in `random` và `int` ra để hiểu lệnh trên:

`random`

```
 [1] 0.0165335056 0.5482203555 0.7691326942 0.9717108703 0.7892011970
 [6] 0.3479388587 0.2547544581 0.2909628002 0.8007796723 0.9694102113
 ...
[96] 0.6618360400 0.4355043718 0.2979350316 0.9742071696 0.3063064239
```

`int`

```
 [1]  1 54 76 97 78 34 25 29 80 96 17  3 22 31 68 46 64 50 92 60 53 61 92 70
...
[97] 43 29 97 30
```

- Bước 3: Xác định `int` là số chẵn hay lẻ bằng hàm `%%` và cho vào biến `odd`. Dùng hàm `replace` để chia nhóm: nếu `odd` là số lẻ, cho vào nhóm A; nếu odd là số chẵn, cho vào nhóm P, và gọi nhóm bằng tên mới là `group`:

```
odd <- int%%2
group <- odd
group <- replace(group, odd == 1, "A")
group <- replace(group, odd == 0, "P")
```

(Trong lệnh số 1, `int%%2` chúng ta muốn biết là `int` là số chẵn hay lẻ. Nếu `int` là số chẵn thì có thể chia cho 2 và phần dư là 0; nếu `int` là số lẻ thì phần dư không phải là 0. Giá trị của `odd` sẽ là 0 (chẵn) hay 1 (lẻ).

Lệnh thứ hai, tạm gọi cho `group` lấy giá trị của `odd`.

Lệnh thứ ba: nếu `odd` bằng 1, thay thế `group` bằng giá trị "A"; nếu `odd` bằng 0, thay thế `group` bằng giá trị "P").

- Bước 4: Dùng hàm `data.frame` để chứa tất cả các số liệu liên quan như `id` và `group` vào một dữ liệu có tên là `grouping` và in ra:

```
grouping <- data.frame(id, group)
grouping
```

```
     id group
1     1     A
2     2     A
3     3     A
4     4     P
5     5     A
6     6     P
7     7     P
8     8     P
9     9     P
10   10     A
...
98   98     A
99   99     P
100 100     P
```

Theo kết quả trên, chúng ta sẽ xếp bệnh nhân số 1, 2, 3 vào nhóm can thiệp (A), bệnh nhân 4 vào nhóm đối chứng (P), v.v… Để kiểm tra xem chúng ta có bao nhiêu bệnh nhân trong nhóm A và B, chúng ta sử dụng hàm `table` như sau:

```
table(group)
```

Và kết quả là có 45 bệnh nhân được phân chia vào nhóm A, 55 vào nhóm P.

```
group
 A  P
45 55
```

Vì phân chia hoàn toàn ngẫu nhiên, cho nên số lượng bệnh nhân không hoàn toàn cân đối, nhất là những nghiên cứu có số lượng đối tượng không nhiều. Nhưng đối với những nghiên cứu với hàng ngàn đối tượng thì phân chia theo ngẫu nhiên hóa có thể cân đối rất hữu hiệu.

Tất nhiên, chúng ta có thể "chạy" (lặp lại các lệnh trên) qui trình trên cho đến khi nào số lượng đối tượng của hai nhóm cân bằng thì ngưng (bạn đọc chỉ đơn giản cắt (cut) toàn bộ lệnh và dán (paste) vào R các lệnh dưới đây):

```
n <- 100
id <- 1:n
random <- runif(n)
int <- as.integer(random*100)
odd <- int%%2
group <- odd
group <- replace(group, odd == 1, "A")
group <- replace(group, odd == 0, "P")
grouping <- data.frame(id, group)
table(group)
```

Tôi đã thử chạy các lệnh trên, và chỉ 3 lần là đáp ứng yêu cầu 50:50! Khi đã đạt yêu cầu, chúng ta chỉ đơn giản in ra (với lệnh `grouping`) và lưu trữ kết quả trong một hồ sơ Excel để tham chiếu sau này.

Ví dụ 2: **Phân nhóm theo từng cụm (block)**. Vì khả năng thiếu cân đối trong cách phân nhóm tổng thể như trên, chúng ta cần một phương pháp khác để đảm bảo cân đối ngay cả với những nghiên cứu ít đối tượng (như 40 đối tượng chẳng hạn). Một phương pháp có thể đáp ứng yêu cầu này là phân chia theo từng cụm. Chẳng hạn như với 40 bệnh nhân, chúng ta chia thành 4 cụm, mỗi cụm gồm 10 bệnh nhân, và tiến hành phân chia ngẫu nhiên từng cụm. Trong trường hợp này, lệnh `n <- 100` trên sẽ được thay thế bằng `n <- 10` và các lệnh khác không thay đổi.

Tóm lại, ngẫu nhiên hóa đóng vai trò quan trọng trong suy luận khoa học và kiểm định giả thuyết khoa học. Trong nghiên cứu y học, ý tưởng ngẫu nhiên hóa đã làm

nên một cuộc cách mạng trong việc thẩm định các thuật điều trị trong vòng nửa thế kỉ qua. Đứng trên phương diện "kĩ thuật", việc chọn ngẫu nhiên và phân nhóm ngẫu nhiên rất đơn giản nếu bạn đọc có sẵn phần mềm R (có thể tải về máy tính cá nhân hoàn toàn miễn phí). Bạn đọc nên tự mình kiểm tra các lệnh trên đây bằng cách thay đổi các thông số để hiểu thêm về cơ chế của chọn và phân nhóm ngẫu nhiên.

Ghi chú kĩ thuật:

Các lệnh R trong ví dụ 1 có thể đơn giản hóa thành một hàm (function). Gọi hàm bằng tên `grp`, chúng ta có thể viết như sau:

```
grp <- function(k)
  {
      n <- k
      id <- 1:n
      random <- runif(n)
      int <- as.integer(random*100)
      odd <- int%%2
      group <- odd
      group <- replace(group, odd == 1, "A")
      group <- replace(group, odd == 0, "P")
      grouping <- data.frame(id, group)
      grouping[,1:2]
      table(group)
  }
```

Để chạy hàm `grp` này, chúng ta trước hết copy toàn bộ hàm vào R, và sau đó chỉ đơn giản ra lệnh `grp(k)`, trong đó, k là số lượng đối tượng chúng ta muốn phân chia:

```
grp(1000)
```

Lệnh này có thể lặp lại nhiều lần, cho đến khi nào hai nhóm cân đối nhau.