7
Kiểm định giả thiết thống kê và ý nghĩa của trị số P (P-value)

7.1 Trị số P

Trong nghiên cứu khoa học, ngoài những dữ kiện bằng số, biểu đồ và hình ảnh, con số mà chúng ta thường hay gặp nhất là trị số P (mà tiếng Anh gọi là P-value). Trong các chương sau đây, bạn đọc sẽ gặp trị số P rất nhiều lần, và đại đa số các suy luận phân tích thống kê, suy luận khoa học đều dựa vào trị số P. Do đó, trước khi bàn đến các phương pháp phân tích thống kê bằng R, tôi thấy cần phải có đôi lời về ý nghĩa của trị số này.

Trị số P là một con số xác suất, tức là viết tắt chữ "probability value". Chúng ta thường gặp những phát biểu được kèm theo con số, chẳng hạn như "Kết quả phân tích cho thấy tỉ lệ gãy xương trong nhóm bệnh nhân được điều trị bằng thuốc Alendronate là 2%, thấp hơn tỉ lệ trong nhóm bệnh nhân không được chữa trị (5%), và mức độ khác biệt này có ý nghĩa thống kê ($p = 0.01$)", hay một phát biểu như "Sau 3 tháng điều trị, mức độ giảm áp suất máu trong nhóm bệnh nhân là 10% ($p < 0.05$)". Trong văn cảnh trên đây, đại đa số nhà khoa học hiểu rằng trị số P phản ánh xác suất sự hiệu nghiệm của thuốc Alendronate hay một thuật điều trị, họ hiểu rằng câu văn trên có nghĩa là "xác suất mà thuốc Alendronate tốt hơn giả dược là 0.99" (lấy 1 trừ cho 0.01). Nhưng cách hiểu đó hoàn toàn sai!

Trong "Từ điển toán kinh tế thống kê, kinh tế lượng Anh – Việt" (Nhà xuất bản Khoa học và Kĩ thuật, 2004), tác giả định nghĩa trị số P như sau: "*P – giá trị (hoặc giá trị xác suất). P giá trị là mức ý nghĩa thống kê thấp nhất mà ở đó giá trị quan sát được của thống kê kiểm định có ý nghĩa*" (trang 690). Định nghĩa này thật là khó hiểu! Thật ra đó cũng là định nghĩa chung mà các sách khoa Tây phương thường hay viết. Lật bất cứ sách giáo khoa nào bằng tiếng Anh, chúng ta sẽ thấy một định nghĩa về trị số P na ná giống nhau như "Trị số P là xác suất mà mức độ khác biệt quan sát do các yếu tố ngẫu nhiên gây ra (P value is the probability that the observed difference arose by chance)". Thật ra định nghĩa này chưa đầy đủ, nếu không muốn nói là … sai. Chính vì sự mù mờ của định nghĩa cho nên rất nhiều nhà khoa học hiểu sai ý nghĩa của trị số P.

Thật vậy, rất nhiều người, không chỉ người đọc mà ngay cả chính các tác giả của những bài báo khoa học, không hiểu ý nghĩa của trị số P. Theo một nghiên cứu được công bố trên tập san danh tiếng *Statistics in Medicine* [1], tác giả cho biết 85% các tác giả khoa học và bác sĩ nghiên cứu không hiểu hay hiểu sai ý nghĩa của trị số P. Đọc đến đây có lẽ bạn đọc rất ngạc nhiên, bởi vì điều này có nghĩa là nhiều nhà nghiên cứu khoa học có khi không hiểu hay hiểu sai những gì chính họ viết ra có nghĩa gì! Thế thì, câu hỏi cần đặt ra một cách nghiêm chỉnh: *Ý nghĩa của trị số P là gì*? Để trả lời cho câu hỏi này,

chúng ta cần phải xem xét qua khái niệm phản nghiệm và tiến trình của một nghiên cứu khoa học.

7.2 Giả thiết khoa học và phản nghiệm

Một giả thiết được xem là mang tính "khoa học" nếu giả thiết đó có khả năng "phản nghiệm". TheoKarl Popper, nhà triết học khoa học, đặc điểm duy nhất để có thể phân biệt giữa một lí thuyết khoa học thực thụ với ngụy khoa học (pseudoscience) là thuyết khoa học luôn có đặc tính có thể " bị bác bỏ" (hay bị phản bác – falsified) bằng những thực nghiệm đơn giản. Ông gọi đó là "khả năng phản nghiệm" (falsifiability, có tài liệu ghi là falsibility). Phép phản nghiệm là phương cách tiến hành những thực nghiệm không phải để xác minh mà để phê phán các lí thuyết khoa học, và có thể coi đây như là một nền tảng cho khoa học thực thụ. Chẳng hạn như giả thiết "Tất cả các quạ đều màu đen" có thể bị bác bỏ nếu ta tìm ra có một con quạ màu đỏ.

Có thể xem qui trình phản nghiệm là một cách học hỏi từ sai lầm! Thật vậy, trong khoa học chúng ta học hỏi từ sai lầm. Khoa học phát triển cũng một phần lớn là do học hỏi từ sai lầm mà giới khoa học không ai chối cãi. Sai lầm là điểm mạnh của khoa học. Có thể xác định nghiên cứu khoa học như là một qui trình thử nghiệm giả thuyết, theo các bước sau đây:

Bước 1, nhà nghiên cứu cần phải định nghĩa một giả thuyết đảo (null hypothesis), tức là một giả thuyết ngược lại với những gì mà nhà nghiên cứu tin là sự thật. Thí dụ trong một nghiên cứu lâm sàng, gồm hai nhóm bệnh nhân: một nhóm được điều trị bằng thuốc A, và một nhóm được điều trị bằng placebo, nhà nghiên cứu có thể phát biểu một giả thuyết đảo rằng sự hiệu nghiệm thuốc A tương đương với sự hiệu nghiệm của placebo (có nghĩa là thuốc A không có tác dụng như mong muốn).

Bước 2, nhà nghiên cứu cần phải định nghĩa một giả thuyết phụ (alternative hypothesis), tức là một giả thuyết mà nhà nghiên cứu nghĩ là sự thật, và điều cần được "chứng minh" bằng dữ kiện. Chẳng hạn như trong ví dụ trên đây, nhà nghiên cứu có thể phát biểu giả thuyết phụ rằng thuốc A có hiệu nghiệm cao hơn placebo.

Bước 3, sau khi đã thu thập đầy đủ những dữ kiện liên quan, nhà nghiên cứu dùng một hay nhiều phương pháp thống kê để kiểm tra xem trong hai giả thuyết trên, giả thuyết nào được xem là khả dĩ. Cách kiểm tra này được tiến hành để trả lời câu hỏi: nếu giả thuyết đảo đúng, thì xác suất mà những dữ kiện thu thập được phù hợp với giả thuyết đảo là bao nhiêu. Giá trị của xác suất này thường được đề cập đến trong các báo cáo khoa học bằng kí hiệu "P value". Điều cần chú ý ở đây là nhà nghiên cứu không thử nghiệm giả thuyết khác, mà chỉ thử nghiệm giả thuyết đảo mà thôi.

Bước 4, quyết định chấp nhận hay loại bỏ giả thuyết đảo, bằng cách dựa vào giá trị xác suất trong bước thứ ba. Chẳng hạn như theo truyền thống lựa chọn trong một nghiên cứu y học, nếu giá trị xác suất nhỏ hơn 5% thì nhà nghiên cứu sẵn sàng bác bỏ giả thuyết đảo: sự hiệu nghiệm của thuốc A khác với sự hiệu nghiệm của placebo. Tuy nhiên, nếu giá trị xác suất cao hơn 5%, thì nhà nghiên cứu chỉ có thể phát biểu rằng chưa

có bằng chứng đầy đủ để bác bỏ giả thuyết đảo, và điều này không có nghĩa rằng giả thuyết đảo là đúng, là sự thật. Nói một cách khác, thiếu bằng chứng không có nghĩa là không có bằng chứng.

Bước 5, nếu giả thuyết đảo bị bác bỏ, thì nhà nghiên cứu mặc nhiên thừa nhận giả thuyết phụ. Nhưng vấn đề khởi đi từ đây, bởi vì có nhiều giả thuyết phụ khác nhau. Chẳng hạn như so sánh với giả thuyết phụ ban đầu (A khác với Placebo), nhà nghiên cứu có thể đặt ra nhiều giả thuyết phụ khác nhau như thuốc sự hiệu nghiệm của thuốc A cao hơn Placebo 5%, 10% hay nói chung X%. Nói tóm lại, một khi nhà nghiên cứu bác bỏ giả thuyết đảo, thì giả thuyết phụ được mặc nhiên công nhận, nhưng nhà nghiên cứu không thể xác định giả thuyết phụ nào là đúng với sự thật.

7.3 Ý nghĩa của trị số P qua mô phỏng

Để hiểu ý nghĩa thực tế của trị số P, tôi sẽ nêu một ví dụ đơn giản như sau:

Ví dụ 1. Một thí nghiệm được tiến hành để tìm hiểu sở thích của người tiêu thụ đối với hai loại cà phê (hãy tạm gọi là cà phê A và B). Các nhà nghiên cứu cho 50 khách hàng uống thử hai loại cà phê trong cùng một điều kiện, và hỏi họ thích loại cà phê nào. Kết quả cho thấy 35 người thích cà phê A, và 15 người thích cà phê B. Vấn để đặt ra là qua kết quả này, các nhà nghiên cứu có thể kết luận rằng cà phê loại A được ưa chuộng hơn cà phê B, hay kết quả trên chỉ là do ngẫu nhiên mà ra?

"Do ngẫu nhiên mà ra" có nghĩa là theo luật nhị phân, khả năng mà kết quả trên xảy ra là bao nhiêu? Do đó, lí thuyết xác suất nhị phân có phần ứng dụng trong trường hợp này, bởi vì kết quả của nghiên cứu chỉ có hai "giá trị" (hoặc là thích A, hoặc thích B).

Nói theo ngôn ngữ của phản nghiệm, giả thiết đảo là nếu không có sự khác biệt về sở thích, xác suất mà một khách hàng ưa chuộng một loại cà phê là 0.5. Nếu giả thiết này là đúng (tức p = 0.5, p ở đây là xác suất thích cà phê A), và nếu nghiên cứu trên được lặp đi lặp lại (chẳng hạn như) 1000 lần, và mỗi lần vẫn 50 khách hàng, thì có bao nhiêu lần với 35 khách hàng ưa chuộng cà phê A? Gọi số lần nghiên cứu mà 35 (hay nhiều hơn) trong số 50 thích cà phê A là "biến cố" X, nói theo ngôn ngữ xác suất, chúng ta muốn tìm P(X | p=0.50) =?

Để trả lời câu hỏi này, chúng ta có thể ứng dụng hàm `rbinom` để mô phỏng vì như nói trên thực chất của vấn đề là một phân phối nhị phân:

```
> bin <- rbinom(1000, 50, 0.5)
```

Trong lệnh trên, chúng ta yêu cầu R mô phỏng 1000 lần nghiên cứu, mỗi lần có 50 khách hàng, và theo giả thiết đảo, xác suất thích A là 0.50. Để biết kết quả của mô phỏng đó, chúng ta sử dụng hàm `table` như sau:

```
> table(bin)
```

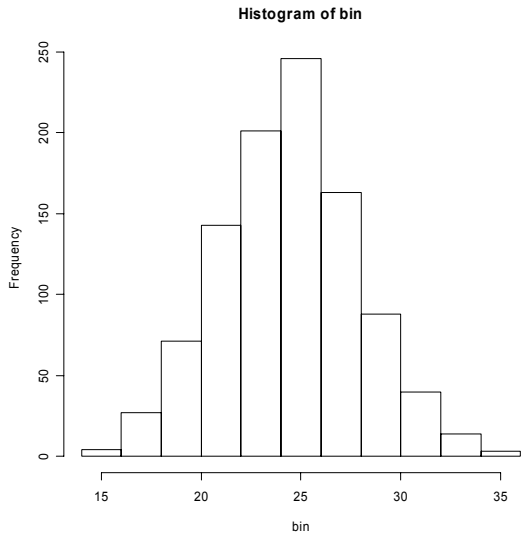
```
bin
 14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33
  1   1   2  11  16  24  47  60  83  94 107 132 114  98  65  44  44  26  14  12

 34  35
  2   3
```

Qua kết quả trên, chúng ta thấy trong số 1000 "nghiên cứu" đó, chỉ có 3 nghiên cứu mà số khách hàng thích cà phê A là 35 người (với điều kiện không có khác biệt giữa hai loại cà phê, hay nói đúng hơn là nếu p =0.5). Nói cách khác:

$$P(X \geq 35 \mid p=0.50) = 3/1000 = 0.003$$

Chúng ta cũng có thể thể hiện tần số trên bằng một biểu đồ tần số như sau:



Tất nhiên chúng ta có thể làm một mô phỏng khác với số lần tái thí nghiệm là 100.000 lần (thay vì 1000 lần) và tính xác suất $P(X \geq 35 \mid p=0.50)$.

```
bin <- rbinom(100000, 50, 0.5)
> bin <- rbinom(100000, 50, 0.5)
> table(bin)
bin
   11    12    13    14    15    16    17    18    19    20    21    22    23
    4    17    40    83   197   462   946  1592  2719  4098  5892  7937  9733

   24    25    26    27    28    29    30    31    32    33    34    35    36
10822 11191 10799  9497  7925  5904  4185  2682  1562   893   455   223    98

   37    38    39    40
   31     5     7     1
```

Lần này, chúng ta có nhiều khả năng hơn (vì số lần mô phỏng tăng lên). Chẳng hạn như có thể có nghiên cứu cho ra 11 khách hàng (tối thiểu) hay 40 khách hàng (tối đa) thích cà

phê A. Nhưng chúng ta muốn biết số lần nghiên cứu mà 35 khách hàng trở lên thích cà phê A, và kết quả trên cho chúng ta biết, xác suất đó là:

```
> (223+98+21+5+7+1)/100000
[1] 0.00355
```

Nói cách khác, xác suất $P(X \geq 35 \mid p=0.50)$ quá thấp (chỉ 0.3%), chúng ta có bằng chứng để cho rằng kết quả trên có thể không do các yếu tố ngẫu nhiên gây nên; tức có một sự khác biệt về sở thích của khách hàng đối với hai loại cà phê.

Con số P = 0.0035 chính là trị số P. Theo một qui ước khoa học, tất cả các trị số P thấp hơn 0.05 (tức thấp hơn 5%) được xem là "significant", tức là "có ý nghĩa thống kê".

Cần phải nhấn mạnh một lần nữa để hiểu ý nghĩa của trị số P như sau: Mục đích của phân tích trên là nhằm trả lời câu hỏi: *nếu hai loại cà phê có xác suất ưa chuộng bằng nhau (p = 0.5, giả thuyết đảo), thì xác suất mà kết quả trên (35 trong số 50 khách hàng thích A) xảy ra là bao nhiêu?* Nói cách khác, đó chính là phương pháp đi tìm trị số P. Do đó, diễn dịch trị số P phải có điều kiện, và điều kiện ở đây là p = 0.50. bạn đọc có thể làm thí nghiệm thêm với p = 0.6 hay p = 0.7 để thấy kết quả khác nhau ra sao.

Trong thực tế, trị số P có một ảnh hưởng rất lớn đến số phận của một bài báo khoa học. Nhiều tập san và nhà khoa học xem một nghiên cứu khoa học với trị số P cao hơn 0.05 là một "kết quả tiêu cực" ("negative result") và bài báo có thể bị từ chối cho công bố. Chính vì thế mà đối với đại đa số nhà khoa học, con số "P < 0.05" đã trở thành một cái "giấy thông hành" để công bố kết quả nghiên cứu. Nếu kết quả với P < 0.05, bài báo có cơ may xuất hiện trên một tập san nào đó và tác giả có thể sẽ nổi tiếng; nếu kết quả P > 0.05, số phận bài báo và công trình nghiên cứu có cơ may đi vào lãng quên!

7.4 Vấn đề logic của trị số P

Nhưng đứng trên phương diện lí trí và khoa học nghiêm chỉnh, chúng ta có nên đặt tầm quan trọng vào trị số P như thế hay không? Theo tôi, câu trả lời là không. Trị số P có nhiều vấn đề, và việc phụ thuộc vào nó trong quá khứ (cũng như hiện nay) đã bị rất nhiều người phê phán gay gắt. Cái khiếm khuyết số 1 của trị số P là nó thiếu tính logic. Thật vậy, nếu chúng ta chịu khó xem xét lại ví dụ trên, chúng ta có thể khái quát tiến trình của một nghiên cứu y học (dựa vào trị số P) như sau:

- Đề ra một giả thuyết chính (H+)
- Từ giả thuyết chính, đề ra một giả thuyết đảo (H-)
- Tiến hành thu thập dữ kiện (D)
- Phân tích dữ kiện: tính toán xác suất D xảy ra nếu H- là sự thật. Nói theo ngôn ngữ toán xác suất, bước này xác định $P(D \mid H-)$.

Vì thế, con số P có nghĩa là xác suất của dữ kiện D xảy ra *nếu* (nhấn mạnh: "nếu") giả thuyết đảo H- là sự thật. Như vậy, con số P không trực tiếp cho chúng ta một ý niệm gì về sự thật của giả thuyết chính H; nó chỉ gián tiếp cung cấp bằng chứng để chúng ta chấp nhận giả thuyết chính và bác bỏ giả thuyết đảo.

Cái logic đằng sau của trị số P có thể được hiểu như là một tiến trình *chứng minh đảo ngược* (proof by contradiction):

- Mệnh đề 1: Nếu giả thuyết đảo là sự thật, thì dữ kiện này không thể xảy ra;
- Mệnh đề 2: Dữ kiện xảy ra;
- Mệnh đề 3 (kết luận): Giả thuyết đảo không thể là sự thật.

Nếu bạn đọc cảm thấy khó hiểu cách lập luận trên, tôi xin lấy thêm một ví dụ trong y khoa để minh họa cho tiến trình này:

- Nếu ông Tuấn bị cao huyết áp, thì ông không thể có triệu chứng rụng tóc (hai hiện tượng sinh học này không liên quan với nhau, ít ra là theo kiến thức y khoa hiện nay);
- Ông Tuấn bị rụng tóc;
- Do đó, ông Tuấn không thể bị cao huyết áp.

Trị số P, do đó, gián tiếp phản ánh xác suất của mệnh đề 3. Và đó cũng chính là một khiếm khuyết quan trọng của trị số P, bởi vì con số P nó ước tính mức độ khả dĩ của dữ kiện, chứ không nói cho chúng ta biết mức độ khả dĩ của một giả thuyết. Điều này làm cho việc suy luận dựa vào trị số P rất xa rời với thực tế, xa rời với khoa học thực nghiệm. Trong khoa học thực nghiệm, điều mà nhà nghiên cứu muốn biết là với dữ kiện mà họ có được, xác suất của giả thuyết chính là bao nhiêu, chứ họ không muốn biết nếu giả thuyết đảo là sự thật thì xác suất của dữ kiện là bao nhiêu. Nói cách khác và dùng kí hiệu mô tả trên, nhà nghiên cứu muốn biết P(H+ | D), chứ không muốn biết P(D | H+) hay P(D | H-).

7.5. Vấn đề kiểm định nhiều giả thuyết (multiple tests of hypothesis)

Như đã nói trên, nghiên cứu y học là một qui trình thử nghiệm giả thuyết. Trong một nghiên cứu, ít khi nào chúng ta thử nghiệm chỉ một giả thuyết duy nhất, mà rất nhiều giả thuyết một lược. Chẳng hạn như trong một nghiên cứu về mối liên hệ giữa vitamin D và nguy cơ gãy xương đùi, các nhà nghiên cứu có thể phân tích mối liên hệ tương quan giữa vitamin D và mật độ xương (bone mineral density), giữa vitamin D và nguy cơ gãy xương theo từng giới tính, từng nhóm tuổi, hay phân tích theo các đặc tính lâm sàng của bệnh nhân, v.v… (Xem ví dụ dưới đây). Mỗi một phân tích như thế có thể xem là một thử nghiệm giả thuyết. Ở đây, chúng ta phải đối diện với vấn đề nhiều giả thuyết (multiple tests of hypothesis hay còn gọi là **multiple comparisons**).

Bảng 2. Phân tích hiệu quả của vitamin D và calcium theo đặc tính của bệnh nhân

Đặc tính bệnh nhân	Nhóm được điều trị bằng calcium và vitamin D [1]	Nhóm giả dược (placebo) [1]	Tỉ số nguy cơ (relative risk) và khoảng tin cậy 95% [2]
Độ tuổi			
50-59	29 (0.06)	13 (0.03)	2,17 (1.13-4.18)
60-69	53 (0.09)	71 (0.13)	0.74 (0.52-1.06)
70-79	93 (0.44)	115 (0.54)	0.82 (0.62-1.08)
Body mass index			
<25	69 (0.20)	66 (0.19)	1.05 (0.75-1.47)
25-30	63 (0.14)	74 (0.16)	0.87 (0.62-1.22)
≥30	43 (0.09)	59 (0.13)	0.73 (0.49-1.09)
Hút thuốc lá			
Không hút thuốc	159 (0.14)	178 (0.15)	0.90 (0.71-1.11)
Hiện hút thuốc	14 (0.14)	16 (0.17)	0.85 (0.41-1.74)

Chú thích: [1] số ngoài ngoặc là số bệnh nhân bị gãy xương đùi trong thời gian theo dõi (7 năm) và số trong ngoặc là tỉ lệ gãy xương tính bằng phần trăm mỗi năm. [2] Tỉ số nguy cơ tương đối (hay relative risk – RR – sẽ giải thích trong một chương sau) được ước tính bằng cách lấy tỉ lệ gãy xương trong nhóm can thiệp chia cho tỉ lệ trong nhóm giả dược; nếu khoảng tin cậy 95% bao gồm 1 thì mức độ khác biệt giữa 2 nhóm không có ý nghĩa thống kê; nếu khoảng tin cậy 95% *không* bao gồm 1 thì mức độ khác biệt giữa 2 nhóm được xem là có ý nghĩa thống kê (hay $p<0.05$).

Xin nhắc lại rằng trong mỗi lần thử nghiệm một giả thuyết, chúng ta chấp nhận một sai sót 5% (giả dụ chúng ta chấp nhận tiêu chuẩn $p = 0.05$ để tuyên bố có ý nghĩa hay không có ý nghĩa thống kê). Vấn đề đặt ra là trong bối cảnh thử nghiệm nhiều giả thuyết là như sau: **nếu trong số *n* thử nghiệm, chúng ta tuyên bố *k* thử nghiệm "có ý nghĩa thống kê" (tức là $p<0.05$), thì xác suất có ít nhất một giả thuyết sai là bao nhiêu?**

Để trả lời câu hỏi này tôi sẽ bắt đầu bằng một ví dụ đơn giản. Mỗi thử nghiệm chúng ta chấp nhận một xác suất sai lầm là 0.05. Nói cách khác, chúng ta có xác suất đúng là 0.95. Nếu chúng ta thử nghiệm 3 giả thuyết, xác suất mà chúng ta đúng cả ba là [dĩ nhiên]: $0.95 \times 0.95 \times 0.95 = 0.8574$. Như vậy, xác xuất có ít nhất một sai lầm trong ba tuyên bố "có ý nghĩa thống kê" là: $1 - 0.8574 = 0.1426$ (tức khoảng 14%).

Nói chung, nếu chúng ta thử nghiệm *n* giả thuyết, và mỗi lần thử nghiệm chúng ta chấp nhận một xác suất sai lầm là *p,* thì xác suất có ít nhất 1 sai lầm trong *n* lần thử nghiệm đó là $1-(1-p)^n$. Khi $n = 10$ và $p = 0.05$ thì xác suất có ít nhất một sai lầm lên đến: 40%.

"Bài học" rút ra từ cách lí giải trên là như sau: nếu chúng ta đọc một bài báo khoa học mà trong đó nhà nghiên cứu tiến hành nhiều thử nghiệm khác nhau với các kết quả trị số $p < 0.05$, chúng ta có lí do để cho rằng xác suất mà một trong những cái-gọi-là

"significant" (hay "có ý nghĩa thống kê") đó rất cao. Chúng ta cần phải dè dặt với những kết quả phân tích như thế.

Đối với một người làm nghiên cứu, ý nghĩa của vấn đề thử nghiệm nhiều giả thuyết là: không nên "câu cá". Tôi xin nói thêm về khái niệm "câu cá" trong khoa học. Hãy tưởng tượng, một nhà nghiên cứu muốn tìm hiểu hiệu quả của một thuật điều trị mới cho các bệnh nhân đau khớp. Sau khi xem xét các nghiên cứu đã công bố trong y văn, nhà nghiên cứu quyết định tiến hành một nghiên cứu trên 300 bệnh nhân: phân nửa được điều trị bằng thuật mới, phân nửa chỉ sử dụng giả dược. Sau thời gian theo dõi, thu thập dữ liệu, nhà nghiên cứu phân tích và phát hiện sự khác biệt giữa hai nhóm không có ý nghĩa thống kê. Nói cách khác, thuật điều trị không có hiệu quả. Nhà nghiên cứu không chịu "đầu hàng", nên tìm cách tìm cho được một kết quả có ý nghĩa thống kê. Ông chia bệnh nhân thành nhiều nhóm theo độ tuổi (trên 50 hay dưới 50), theo giới tính (nam hay nữa), thành phần kinh tế (có thu nhập cao hay thấp), và thói quen (chơi thể thao hay không). Tính chung, ông có 16 nhóm khác nhau, và có thể thử nghiệm 16 lần. Ông "khám phá" thuật điều trị có ý nghĩa thống kê trong nhóm phụ nữ tuổi trên 50 và có thu nhập cao. Và, ông công bố kết quả. Đó là một qui trình làm việc mà giới nghiên cứu khoa học gọi là "fishing expedition" (một chuyến đi câu cá). Tất nhiên, một kết quả như thế không có giá trị khoa học và không thể tin được. (Với 16 thử nghiệm khác nhau và với p = 0.05, xác suất mà một thử nghiệm có kết quả "significant" lên đến 55%, do đó chúng ta chẳng ngạc nhiên khi thấy có một "con cá" được bắt!)

Để cho kết quả trị số P có ý nghĩa nguyên thủy của nó trong bối cảnh thử nghiệm nhiều giả thuyết, các nhà nghiên cứu đề nghị sử dụng thuật điều chỉnh Bonferroni (tên của một nhà thống kê học người Ý từng đề nghị cách làm này). Theo đề nghị này, **trước khi** tiến hành nghiên cứu, nhà nghiên cứu phải xác định rõ giả thuyết nào là chính, và giả thuyết nào là phụ. Ngoài ra, nhà nghiên cứu còn phải đề ra kế hoạch sẽ thử nghiệm bao nhiêu giả thuyết **trước khi bắt tay vào phân tích dữ liệu**. Chẳng hạn như nếu nhà nghiên cứu có kế hoạch thử nghiệm 20 so sánh và muốn giữ cho trị số p ở 0.05, thì thay vì dựa vào 0.05 là tiêu chuẩn để tuyên bố"significant", nhà nghiên cứu phải dựa vào tiêu chuẩn 0.0025 (tức lấy 0.05 chia cho 20) để tuyên bố "significant". Nói cách khác, chỉ khi nào một kết quả có trị số p thấp hơn 0.0025 (hay nói chung là *p/n*) thì nhà nghiên cứu mới có "quyền" tuyên bố kết quả đó có ý nghĩa thống kê.

Trị số P, dù cực kì thông dụng trong nghiên cứu khoa học, không phải là một phán xét cuối cùng của một công trình nghiên cứu hay một giả thuyết. Thế nhưng trong thực tế, các nhà khoa học đã quá lệ thuộc vào trị số P để suy luận trong nghiên cứu và tuyên bố những khám phá mà sau này được chứng minh là sai lầm. Có thể nói không ngoa rằng chính vì sự lạm dụng và phụ thuộc một cách mù quáng vào trị số P mà khoa học, nhất là y sinh học, đã trở nên nghèo nàn. Hàng ngày chúng ta đọc hay nghe những phát hiện khoa học trái ngược nhau (như lúc thì có nghiên cứu cho thấy cà phê có tác dụng tốt cho sức khỏe, lúc khác có nghiên cứu cho biết cà phê có hại cho sức khỏe; hay lúc thì thuốc giảm đau aspirin có hiệu năng làm giảm nguy cơ ung thư, nhưng mới đây có nghiên cứu cho thấy aspirin có thể làm tăng nguy cơ bị ung thư vú, v.v…). Có khi công chúng không biết phát hiện nào là thực và phát hiện nào là "dương tính giả". Theo phân

tích của Berger và Sellke, khoảng 25% các phát hiện với "$p < 0.05$" là các phát hiện dương tính giả [2].

Do đó, chúng ta không nên quá phụ thuộc vào trị số P. Không phải cứ nghiên cứu nào với $p<0.05$ là thành công và $p>0.05$ là thất bại. Có khi một phát hiện với $p>0.05$ nhưng lại là một phát hiện có ý nghĩa. Vấn đề quan trọng là làm sao để ước tính mức độ khả dĩ của một giả thuyết một khi có dữ kiện thật trong tay, tức là ước tính $P(H+ \mid D)$. Để ước tính $P(H+ \mid D)$, chúng ta phải áp dụng Định lí Bayes, và cách tiếp cận định lí này không nằm trong phạm trù của cuốn sách này. Bạn đọc muốn tham khảo thêm có thể đọc một vài bài báo của tôi hay các các bài báo của James Berger mà tài liệu tham khảo dưới đây có thể cung cấp thêm.

Tài liệu tham khảo:

[1] Wulff et al., *Statistics in Medicine* 1987; 6:3-10.

[2] Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P-values and evidence. *Journal of the American Statistical Association* 1987; 82:112-20.