LTS. Trị số P trong nghiên cứu y khoa vẫn thỉnh thoảng được đem ra thảo luận trên các tập san y học quốc tế, và ý nghĩa của nó vẫn là một đề tài cho chúng ta khai thác để hiểu rõ hơn. Bài viết sau đây một lần nữa bàn về ý nghĩa của trị số P nhưng không phải đứng trên quan điểm thống kê, mà qua cái nhìn chẩn đoán lâm sàng. Có thể các bạn sẽ thấy thú vị về sự tương đương giữa nghiên cứu y khoa và chẩn đoán lâm sàng trong bài viết này. Bài viết đã được đăng trên tập san Thông tin Y học; do đó, bạn đọc có thể tham khảo tập san để biết thêm chi tiết.

Ý nghĩa của trị số P trong nghiên cứu y học

Nguyễn Văn Tuấn

Trong một công trình nghiên cứu đánh giá hiệu quả chống gãy xương của thuốc zoledronate, các nhà nghiên cứu điều trị 1065 bệnh nhân bằng zoledronate và 1062 bệnh nhân không được điều trị bằng zoledronate (placebo), và kết quả được trình bày qua một đoạn văn quan trọng sau đây: "*The rates of any new fracture were 8,6% in the zoledronic acid group and 13,9% in the placebo group, a 35% risk reduction with zoledronic acid (p = 0,001); the respective rates of new vertebral fracture were 1,7% and 3,8% (p = 0,02)*" [1]. Câu văn trên đây gắn liền với trị số p có nghĩa gì?

Khi một câu hỏi tương tự được đem đi hỏi một nhóm bác sĩ chuyên khoa và có kinh nghiệm trong nghiên cứu y học, có đến 85% trả lời sai [2]. Đại đa số những người được hỏi hiểu rằng một kết luận (về sự khác biệt) với trị số p = 0,05 có nghĩa là khả năng mà kết luận đó sai là 5%, hay khả năng mà kết luận đó đúng là 95% (lấy 1 trừ cho 0,05). Nhiều người khác thì hiểu rằng một sự khác biệt với trị số P càng nhỏ thì mức độ ảnh hưởng càng có ý nghĩa và độ tin cậy của kết luận càng cao. Nhưng rất tiếc rằng cả hai cách hiểu này đều sai. Điều đáng ngạc nhiên là không những giới làm nghiên cứu khoa học hiểu sai, mà ngay cả các nhà nghiên cứu có kiến thức thống kê khá như dịch tễ học cũng hiểu sai. Thật ra, một số nhà thống kê chuyên nghiệp cũng hiểu sai ý nghĩa của trị số P bởi vì một số sách giáo khoa giải thích hoặc là sai, hoặc không rõ ràng!

Trong bài viết ngắn này, tôi sẽ giải thích ý nghĩa thật của trị số P, bàn qua những khiếm khuyết của nó, và giới thiệu một trường phái suy luận khoa học có ích cho nghiên cứu lâm sàng.

1. Trị số P và triết lí phản nghiệm (falsificationism)

Khi đọc các bài báo khoa học trên các tập san y học, chúng ta thường hay gặp những trị số P. Một sự khác biệt với trị số $p < 0,05$ thường được hiểu là sự khác biệt đó có ý nghĩa thống kê (statistically significant); ngược lại, khi $p > 0,05$ chúng ta thường hiểu rằng sự khác biệt không có ý nghĩa thống kê, không đáng kể, hay do ngẫu nhiên.

Tuy nhiên, cách hiểu P [là một xác suất phi điều kiện] như thế rất sai lầm. Trị số P là một xác suất có điều kiện. Ý nghĩa của trị số P gắn liền với triết lí phản nghiệm (falsificationism) trong khoa học. Do đó, trước khi bàn về ý nghĩa của trị số P, thiết tưởng chúng ta cần phải hiểu qua về triết lí phản nghiệm.

Một giả thuyết được xem là mang tính "khoa học" nếu giả thuyết đó có khả năng "phản nghiệm". Theo Karl Popper [3], 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 "khả năng phản nghiệm" (falsified) bằng những thực nghiệm đơn giản. Ông gọi đó là "khả năng phản nghiệm" (falsifiability) [4]. 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 thuyết [đơn giản] "Tất cả các quạ đều màu đen" có thể bị bác bỏ nếu chúng ta quan sát được một con quạ màu đỏ. Hay, giả thuyết "vi khuẩn *V. cholerae* gây bệnh dịch tả" có thể bác bỏ nếu có một bệnh nhân dịch tả không nhiễm vi khuẩn *V. cholerae.*

Đứng trên phương diện khoa học, có hai mô hình thực tế để tiếp cận lí thuyết phản nghiệm: đó là mô hình kiểm định thống kê và mô hình kiểm định giả thuyết. Rất nhiều sách giáo khoa thống kê và khoa học đã được viết ra, nhưng rất tiếc, nhiều tác giả không giải thích hay không phân biệt được hai mô hình này. Có tác giả thậm chí còn nhầm lẫn khi diễn dịch, và đó cũng chính là một trong những nguyên nhân dẫn đến tình trạng hiểu lầm ý nghĩa của trị số P. Trong phần này, tôi sẽ giải thích ngắn gọn và cung cấp tài liệu tham khảo của hai mô hình để bạn đọc có thể hiểu qua và nghiên cứu thêm.

1.1 Fisher và mô hình kiểm định ý nghĩa thống kê

Triết lí phản nghiệm rất phổ biến và trở thành một mô hình để giải thích sự tiến bộ của khoa học. Chịu ảnh hưởng bởi triết lí này, Ronald A. Fisher (1890 – 1962), một nhà di truyền học người Anh và cũng là "cha đẻ" của nền thống kê học hiện đại, đề xuất một phương pháp định lượng để phản nghiệm một giả thuyết khoa học. Ông gọi phương pháp này là "**Test of Significance**" [5-6] (tôi tạm dịch là: **phương pháp kiểm định ý nghĩa thống kê**). Fisher quan niệm rằng thống kê là một bộ phận quan trọng của phương pháp suy luận theo phép qui nạp (inductive inference), tức là phương pháp suy luận dựa vào quan sát từ các mẫu (sample) và khái quát cho một quần thể (population). Phương pháp kiểm định ý nghĩa thống kê được tiến hành theo 3 bước như sau:

- *Bước 1, phát biểu một giả thuyết vô hiệu (null hypothesis).* Giả thuyết vô hiệu là giả thuyết ngược lại với giả thuyết mà nhà nghiên cứu muốn kiểm định. Chẳng hạn như nếu giả thuyết điều trị bằng thuốc zoledronate làm giảm nguy cơ tử vong (nhóm được điều trị bằng zoledronate có tỉ lệ tử vong thấp hơn nhóm giả dược),

thì giả thuyết vô hiệu sẽ phát biểu là "tỉ lệ tử vong ở bệnh nhân được điều trị bằng zoledronate **bằng** với nhóm giả dược. Gọi giả thuyết vô hiệu là *H0.*

- *Bước 2, thu thập dữ liệu* liên quan đến giả thuyết. Trong ví dụ trên, số liệu sẽ là số trường hợp tử vong. Gọi dữ liệu là *D*.

- *Bước 3, ước tính xác suất* quan sát dữ liệu *D* nếu giả thuyết *H0* đúng. Nói cách khác và viết theo ngôn ngữ toán, bước này ước tính P(D | H0). Đây chính là **trị số P** (**p-value**).

Fisher đề nghị báo cáo trị số P một cách chính xác. Tức là không có những cách viết như p < 0,05 hay p > 0,01 mà phải là p = 0,043 hay p = 0,002. Fisher còn đề nghị rằng nếu trị số p thấp hơn 0,05 thì giả thuyết H0 (vô hiệu) không phù hợp với số liệu quan sát được. Đối với Fisher, không có chuyện "bác bỏ giả thuyết" hay "chứng minh giả thuyết" mà chỉ có số liệu có phù hợp, có nhất quán với giả thuyết hay không mà thôi. Quan điểm này chịu ảnh hưởng "đậm" của triết lí phản nghiệm của Popper, vì theo triết lí này, chúng ta không thể chứng minh bất cứ một giả thuyết nào, mà chỉ có thể bác bỏ (disprove) một giả thuyết bằng dữ liệu quan sát được.

Ví dụ 1. Có thể minh họa cho các bước trên bằng một ví dụ như sau: chúng ta có 10 bệnh nhân, mỗi bệnh nhân được điều trị bằng 2 loại thuốc (A và B). Sau khi theo dõi một thời gian, có 8 bệnh nhân mà hiệu quả của thuốc A tốt hơn thuốc B. Kết quả này có phù hợp với giả thuyết thuốc A tốt hơn thuốc B?

Để trả lời câu hỏi và cũng là kiểm định giả thuyết trên, chúng ta phát biểu một giả thuyết vô hiệu: nếu hai loại thuốc này có hiệu quả như nhau, thì sẽ có 5 bệnh nhân với kết quả A tốt hơn B, và 5 bệnh nhân với kết quả B tốt hơn A. Gọi π là xác suất mà kết quả thuốc A tốt hơn thuốc B. Giả thuyết vô hiệu này cũng có nghĩa là $\pi = 0,5$. Nếu giả thuyết vô hiệu này đúng (tức $\pi = 0,5$), chúng ta có thể tính toán xác suất quan sát *k* bệnh nhân ($k = 0, 1, 2, 3, \ldots, 10$) với kết quả A tốt hơn B theo luật phân phối nhị phân như sau:

$$P(k \mid \pi = 0,5) = \binom{10}{k}(0,5)^k (1-0,5)^{10-k}$$

Và kết quả có thể trình bày trong bảng sau đây:

Bảng 1. Xác suất quan sát k bệnh nhân (trong số 10 bệnh nhân) với kết quả A>B nếu giả thuyết vô hiệu ($\pi = 0,5$) đúng

k =	Pr(k | π=0,5)

0	0,0009765625
1	0,009765625
2	0,04394531
3	0,1171875
4	0,2050781
5	0,2460938
6	0,2050781
7	0,1171875
8	**0,04394531**
9	**0,009765625**
10	**0,0009765625**
P(k ≥ 8)	**0,054687**

Cố nhiên, tổng số xác suất k = 0, 1, 2, …, 10 phải bằng 1. Theo kết quả trên, *nếu không có sự khác biệt về hiệu quả của hai thuốc, xác suất mà chúng ta quan sát 8 bệnh nhân với kết quả A>B là khoảng 4,39%*. Diễn dịch tương tự, chúng ta ước tính rằng xác suất với 9 bệnh nhân kết quả A>B là 0,97%, và xác suất tất cả 10 bệnh nhân với kết quả A>B là 0,097%. Xác suất mà tối thiểu 8 bệnh nhân với kết quả A>B là 0,055 hay 5,5%. Viết theo kí hiệu toán: P(k ≥ 8) = 0,0547. Đây chính là trị số P.

Sử dụng tiêu chí 0,05, chúng ta có thể nói rằng dù 80% (8 trên 10) bệnh nhân với kết quả A>B, chúng ta vẫn chưa có đầy đủ bằng chứng để khẳng định rằng kết quả này nhất quán với giả thuyết thuốc A tốt hơn B.

1.2 Neyman và Pearson và mô hình Kiểm định giả thuyết

Jerzy Neyman (1894 – 1981) là một nhà toán học xuất sắc gốc Ba Lan và Egon Pearson (1895 – 1980) là một nhà thống kê học (con của giáo sư Karl Pearson, cha đẻ của lí thuyết Chi-square và hệ số tương quan) cùng lúc với Fisher, phát triển một phương pháp rất khác với Fisher, mà hai ông gọi là **Test of Hypothesis** (**Kiểm định giả thuyết**) [7]. Neyman và Pearson bác bỏ khái niệm suy luận theo qui nạp; hai ông nghĩ rằng thống kê học là một phương pháp hay cơ chế để hướng dẫn chúng ta đi đến một quyết định đúng về lâu về dài. Nói cách khác, Neyman và Pearson cho rằng phương pháp của Fisher vô nghĩa!

Một cách đơn giản, mô hình kiểm định giả thuyết của Neyman và Pearson có thể thực hiện qua các bước như sau:

- Bước 1, phát biểu giả thuyết chính (H1) và giả thuyết vô hiệu (H0).

- Bước 2, quyết định mức độ α và β có thể chấp nhận được và ước tính cỡ mẫu cần thuyết. α là xác suất bác bỏ giả thuyết H1 nhưng đó là giả thuyết đúng. β là xác suất bác bỏ H0 trong khi H0 đúng.

- Bước 3, thu thập dữ liệu liên quan đến giả thuyết.

- Bước 4, nếu dữ liệu nằm trong khoảng bác bỏ giả thuyết H0, thì chấp nhận giả thuyết H1; nếu không thì chấp nhận giả thuyết H0. Chú ý rằng "chấp nhận" một giả thuyết không có nghĩa là chúng ta tin vào giả thuyết đó, mà chỉ có nghĩa là chúng ta hành động với điều kiện đó là giả thuyết đúng.

Nguyên lí của mô hình Neyman và Pearson là chúng ta dựa vào dữ liệu để chọn một giả thuyết sao cho về lâu về dài chúng ta không quá sai. Chính vì thế mà ngày nay chúng ta thường chọn $\alpha = 5\%$ và $\beta = 10\%$ đến 20%.

Fisher bác bỏ hoàn toàn mô hình của Neyman và Pearson [8]. Ông cho rằng đó là một mô hình … vô duyên. Fisher nhạo báng rằng các nhà toán học (ám chỉ Neyman và Pearson) "chẳng hiểu gì về thực nghiệm và đề ra một mô hình quá phi thực tế". Trong những năm sau đó (thập niên 1930s) cộng đồng thống kê học chứng kiến một cuộc tranh luận dai dẳng và đôi khi nóng bỏng giữa Fisher và Neyman-Pearson trên các tập san thống kê học ở Anh. Fisher tuy là một người thông minh tuyệt vời, một nhà tư tưởng với những suy nghĩ trừu tượng, nhưng lại là một người rất khó tính và có khi hẹp hòi. Sự hẹp hòi của Fisher thể hiện ở chỗ ông sử dụng chức quyền khoa bảng của mình để gây khó khăn cho Neyman đến nỗi ông này chịu không nỗi và phải di cư sang Mĩ và sau này trở thành giáo sư tại trường Đại học Berkeley. Sau này, Neyman được lịch sử ghi nhận là một nhà thống kê học xuất sắc có công cực kì to lớn cho khoa học hiện đại, sánh vai cùng các "đại thụ" trong khoa học hiện đại. Nước Mĩ quả thật là môi trường cho ông thi thố tài năng!

1.2 Một mô hình hỗn hợp

Trớ trêu thay, mấy mươi năm sau, hai mô hình của Fisher và Neyman-Pearson được "hun đúc" thành một mô hình tổng hợp mà chúng ta ứng dụng ngày nay trong nghiên cứu y học. Mô hình này sử dụng kết quả kiểm định thống kê của Fisher để đi đến quyết định chấp nhận hay bác bỏ giả thuyết vô hiệu H0 hay giả thuyết chính H1 theo mô hình của Neyman và Pearson. Tiêu biểu cho mô hình này là nghiên cứu lâm sàng đối chứng ngẫu nhiên (randomized controlled clinical trial hay RCT). Theo đó, một nghiên cứu lâm sàng được tiến hành theo các bước như sau:

- Bước 1, định nghĩa một giả thuyết vô hiệu và một giả thuyết chính. 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 giả thuyết vô hiệu rằng độ hiệu nghiệm thuốc A tương đương với placebo.

- Bước 2, xác định xác suất α (còn gọi là sai số loại I) và β (còn gọi là sai số loại II), và ước tính cỡ mẫu dựa vào hai xác suất này.

- Bước 3, thu thập dữ liệu liên quan đến giả thuyết. Gọi dữ liệu là D.

- Bước 4, sử dụng phương pháp kiểm định ý nghĩa thống kê của Fisher ước tính xác suất P(D | H0). Gọi trị số này là P.

- Bước 5, nếu P < 0,05, bác bỏ giả thuyết H0. Chú ý, bác bỏ H0 không có nghĩa là chúng ta chấp nhận giả thuyết H1.

Ví dụ 2. Có thể minh họa cho các bước trên bằng một ví dụ về nghiên cứu hiệu quả của thuốc zoledronate trong việc phòng chống loãng xương [1]. Với giả thuyết rằng thuốc có hiệu nghiệm giảm nguy cơ gãy xương, các nhà nghiên cứu so sánh tỉ lệ gãy xương giữa hai nhóm bệnh nhân: nhóm 1 được điều trị bằng zoledronate và nhóm 2 là nhóm giả được (nhận calcium và vitamin D). Bắt đầu bằng cách xác định $\alpha = 0,05$ và $\beta = 0,80$, các nhà nghiên cứu ước tính số lượng bệnh nhân cần thiết. Sau ba năm thu thập số liệu, kết quả có thể tóm lược trong bảng số liệu sau đây:

Bảng 2. Nguy cơ gãy xương ở bệnh nhân được điều trị bằng zoledronate và placebo

Chỉ số	Zoledronate	Placebo	Trị số P
Số bệnh nhân	1065	1062	
Số gãy xương	92	139	
Tỉ lệ gãy xương	8,6%	13,9	0,001

Bởi vì trị số P thấp hơn mức α (0,05) mà các nhà nghiên cứu đề ra từ lúc đầu (trước khi thu thập số liệu); cho nên, các nhà nghiên cứu kết luận rằng sự khác biệt về tỉ lệ gãy xương giữa hai nhóm (8,6% vs 13,9%) có ý nghĩa thống kê. Tất nhiên, *trị số P trên không có nghĩa là nghiên cứu đã chứng minh rằng thuốc zoledronate có hiệu quả giảm nguy cơ gãy xương*. Nó có nghĩa là nếu thật sự thuốc zoledronate không có hiệu quả giảm nguy cơ gãy xương thì xác suất mà các nhà nghiên cứu quan sát các số liệu trên (13,9% so với 8,6%) là 0,001.

2. Vấn đề của trị số P

Có lẽ nói không ngoa rằng trị số P là một con số phổ biến nhất trong khoa học từ khoảng 100 năm qua [9]. Hầu hết các bài báo khoa học đều trình bày trị số P như hàm ý nâng cao tính khoa học và độ tin cậy của bài báo. Tuy nhiên, ngay từ lúc mới "ra đời", trị số P đã bị phê bình dữ dội. Có người cho rằng việc ứng dụng trị số P trong suy luận khoa học là một bước lùi, là một sự thoái hóa của khoa học, nên đề nghị không sử dụng trị số này trong nghiên cứu khoa học. Nhưng dù chịu nhiều chỉ trích và phê bình, ứng dụng phương pháp kiểm định giả thuyết và trị số P vẫn càng ngày càng phổ biến, đơn giản vì chúng ta chưa có một phương pháp khác tốt hơn, hay hợp lí hơn, hay đơn giản hơn. Trong phần này, tôi sẽ không điểm qua tất cả các phê bình trị số P (vì làm như thế cần một cuốn sách), mà chỉ nêu một số vấn đề chúng ta cần lưu ý khi diễn dịch trị số P.

2.1 Vấn đề logic

Như qua minh họa trên, trị số P không cho chúng ta biết gì về sự khả dĩ của một giả thuyết, bởi vì nó là một xác suất có điều kiện. Trị số P cho chúng ta biết xác suất của dữ liệu (data) nếu một giả thuyết là đúng. Cái khiếm khuyết lớn nhất của trị số P là nó thiếu tính logic. Thật vậy, nếu chúng ta chịu khó xem xét lại ví dụ trên, 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 vô hiệu (H0)
- Từ giả thuyết vô hiệu, đề ra một giả thuyết chính (H1)
- Tiến hành thu thập dữ liệu (D)
- Phân tích dữ kiện: tính toán xác suất D xảy ra nếu H0 là thật. Nói theo ngôn ngữ toán xác suất, bước này chính là bước tính toán trị số P hay $P(D \mid H0)$.

Vì thế, con số P có nghĩa là xác suất của dữ liệu D xảy ra *nếu* (nhấn mạnh: "nếu") giả thuyết vô hiệu H0 là đúng. Như vậy, con số P không trực tiếp cho chúng ta một ý niệm gì về sự thật của giả thuyết chính H1; nó chỉ gián tiếp cung cấp bằng chứng để chúng ta chấp nhận giả thuyết chính và bác bỏ giả thuyết vô hiệu.

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

- Mệnh đề 1: Nếu giả thuyết vô hiệu đúng, thì sự kiện này không thể xảy ra;
- Mệnh đề 2: Sự kiện xảy ra;
- Mệnh đề 3 (kết luận): Giả thuyết vô hiệu không thể đúng.

Nếu cách lập luận trên khó hiểu, chúng ta thử xem một ví dụ cụ thể như sau:

- 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 nó *ước tính mức độ khả dĩ của dữ liệu*, 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 liệu 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 liệu là bao nhiêu. Nói cách khác và dùng kí hiệu mô tả trên, nhà nghiên cứu muốn biết P(H1 | D), chứ không muốn biết P(D | H0) hay P(D | H1).

2.2 Ý nghĩa thống kê không tương đương với ý nghĩa lâm sàng

Một sai lầm rất phổ biến trong giới y khoa là xem một khác biệt có "ý nghĩa thống kê" (statistical significance) tương đương với "ý nghĩa lâm sàng" (clinical significance). Có thể xem trị số P được tính toán từ tỉ số của tín hiệu (signal, mức độ khác biệt giữa hai nhóm) và nhiễu (noise hay độ dao động của mẫu). Gọi *T* là kiểm định thống kê, *S* là tín hiệu, và *E* là nhiễu, ý tưởng trên có thể mô tả như sau:

$$T = \frac{S}{E}$$

Khi số lượng cỡ mẫu tăng và nếu *S* bất biến thì *T* sẽ tăng, tức có cơ hội đạt ý nghĩa thống kê. Điều này có nghĩa là chúng ta có thể giảm *E* tối đa bằng cách tăng số lượng cỡ mẫu, và nó cũng có nghĩa là *một khác biệt rất nhỏ chẳng có ý nghĩa gì trong thực tế nhưng vẫn có thể có ý nghĩa thống kê. Ngược lại, một khác biệt hay ảnh hưởng (effect) lớn, nhưng nếu số lượng cỡ mẫu không đầy đủ không thể đạt được cái chuẩn "có ý nghĩa thống kê"* (tức $p > 0,05$).

Bảng 3 sau đây trình bày 4 nghiên cứu (tưởng tượng) với số cỡ mẫu khác nhau, từ 20 đến 2.000.000 bệnh nhân. Cột "Kết quả" trình bày số bệnh nhân được điều trị dứt bệnh và số trong ngoặc là phần trăm. Giả thuyết vô hiệu là xác suất kết quả 0,5 (tức 50%). Tất cả 4 nghiên cứu đều có trị số P = 0,041. Như có thể thấy qua bảng này, nghiên cứu 1 có tỉ lệ ảnh hưởng cao và có ý nghĩa lâm sàng (75%), và chỉ với 20 bệnh nhân, các nhà nghiên cứu có thể bác bỏ giả thuyết H0. Nhưng nghiên cứu 4, mức độ ảnh hưởng rất thấp (chỉ 50,07%, tức chỉ cao hơn giả thuyết vô hiệu 0,07%) nhưng vẫn có ý nghĩa thống kê vì số cỡ mẫu quá lớn !

Bảng 3. Ảnh hưởng của cỡ mẫu đến trị số P

Nghiên cứu	Số lượng đối tượng	Kết quả điều trị thành công (%)	Trị số P
1	20	15 (75%)	0,041
2	200	114 (57%)	0,041
3	2000	1.046 (52,5%)	0,041
4	2.000.000	1.001.445 (50,07%)	0,041

Trong thực tế, có rất nhiều nghiên cứu mà độ khác biệt giữa hai nhóm rất nhỏ, nhưng vẫn có ý nghĩa thống kê [10-11]. Điều đáng quan tâm là kết quả có ý nghĩa thống kê như thế được các nhà nghiên cứu diễn dịch với hàm ý có ý nghĩa lâm sàng.

Ngược lại, có những nghiên cứu mà kết quả có ý nghĩa lâm sàng nhưng vì không đạt cái chuẩn $p < 0,05$, nên các nhà nghiên cứu lại diễn dịch rằng không có ý nghĩa lâm sàng! Chẳng hạn như một nghiên cứu về hiệu quả của bổ sung vitamin C và E ở phụ nữ mang thai [12], các nhà nghiên cứu kết luận rằng "Supplementation with vitamin C and E during pregnancy does not reduce the risk of serious outcomes in their infants" (Bổ sung vitamin E và E không làm giảm các triệu chứng lâm sàng nghiêm trọng). Nhưng khi xét qua số liệu thực tế thì thấy ở trẻ em mà mẹ có bổ sung vitamin C và E, tỉ lệ với triệu chứng lâm sàng giảm đến 21% ($p = 0,06$). Chỉ vì $p = 0,06$ mà các nhà nghiên cứu có xu hướng diễn dịch sai kết quả, và sai lầm này rất nghiêm trọng!

2.2 Vấn đề kiểm định nhiều giả thuyết

Như đã nói trên, nghiên cứu y học là một qui trình kiểm định giả thuyết. Trong một nghiên cứu, ít khi nào chúng ta kiểm định chỉ một giả thuyết duy nhất, mà rất nhiều giả thuyết cùng một lúc. Chẳng hạn như trong một nghiên cứu về mối liên hệ giữa vitamin D và nguy cơ gãy cổ xương đùi, các nhà nghiên cứu có thể phân tích mối liên hệ 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… Mỗi một phân tích như thế có thể xem là một kiểm định 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**).

Vấn đề là như sau: nếu chúng ta kiểm định một giả 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). Nói cách khác, sự thật là *không* thuốc có hiệu quả sai, nhưng kết quả kiểm định thống kê cho ra kết quả có ý nghĩa thống kê, và chúng ta chấp nhận rằng sự

kiện này có thể xảy ra với tần số 5%. Vấn đề đặt ra là trong bối cảnh kiểm định 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 chúng ta sẽ bắt đầu bằng một ví dụ đơn giản. Mỗi kiểm định 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 [dĩ nhiên] là: 0,95 x 0,95 x 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 số lần kiểm định là $n = 10$ và p=0,05 thì xác suất có ít nhất một kết luận 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". 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 cho được một kết quả có ý nghĩa thống kê: chia bệnh nhân thành nhiều nhóm theo độ tuổi (trên 50 hay dưới 50), theo giới tính (nam hay nữ), thành phần kinh tế (có thu nhập cao hay thấp), và thói quen (chơi thể thao hay không). Tính chung, nhà nghiên cứu có 16 nhóm khác nhau, và có thể kiểm định 16 giả thuyết. Nhà nghiên cứu "khám phá" thuật điều trị có ý nghĩa thống kê trong nhóm phụ nữ tuổi trên 50 và có thu nhập cao. Và, kết quả trên được công bố. 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 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.

3. Trị số P và chẩn đoán y khoa

Có một mối tương quan giữa nghiên cứu khoa học và chẩn đoán y khoa, mà tôi thấy giới y học ít khi nào để ý đến để giải thích về ý nghĩa của trị số P:

- Hai lĩnh vực đều có cùng mục đích: đi tìm cái chưa được biết. Trong nghiên cứu y học chúng ta tìm một mối liên hệ (hay ước tính / đánh giá hiệu quả của một thuật can thiệp), còn trong chẩn đoán chúng ta muốn biết bệnh nhân có bệnh hay không có bệnh.

- Nghiên cứu y học sử dụng thống kê học làm phương pháp kiểm định, còn chẩn đoán y khoa sử dụng xét nghiệm lâm sàng hay sinh hóa để định bệnh. Do đó, phương pháp kiểm định thống kê tương đương với phương pháp xét nghiệm sinh hóa / lâm sàng.

- Trong nghiên cứu y học, thuốc thực sự không hiệu quả, nhưng kết quả phân tích thống kê cho rằng có ý nghĩa thống kê. Trong chẩn đoán y khoa, bệnh nhân không có bệnh, nhưng kết quả xét nghiệm là dương tính.

- Tương tự, trong nghiên cứu y học, thuốc thực sự có hiệu quả, nhưng kết quả phân tích thống kê cho rằng không có ý nghĩa thống kê. Trong chẩn đoán y khoa, bệnh nhân có bệnh, nhưng kết quả xét nghiệm là âm tính.

Do đó, để hiểu ý nghĩa và cách diễn dịch trị số P, chúng ta cần bàn qua và quán triệt ý nghĩa của một kết quả chẩn đoán y khoa. Tôi sẽ lấy ví dụ chẩn đoán ung thư làm ví dụ. Để biết một phụ nữ bị ung thư vú hay không, cách chính xác nhất là qua giải phẫu, hay trong trường hợp những người đã chết, là qua giảo nghiệm tử thi. Nhưng giải phẫu là một thuật mang tính xâm phạm cao, và tốn kém. Do đó, các nhà khoa học phát triển nhiều phương pháp để có thể chẩn đoán ung thư mà không cần đến giải phẫu để biết bệnh

trạng của của bệnh nhân. Trong trường hợp ung thư vú, một phương pháp công nghệ cao là chụp quang tuyến X, hay còn gọi là *mammography*.

Kết quả của việc xét nghiệm bằng quang tuyến X có thể là *dương tính* (positive, sẽ viết tắt là +ve), hay *âm tính* (negative, -ve). Một kết quả dương tính có nghĩa rằng bệnh nhân có thể bị ung thư vú, và một kết quả âm tính cho biết bệnh nhân có thể không bị ung thư vú. (Hai chữ "có thể" ở đây rất quan trọng, vì nó nói lên một sự bất định trong việc chẩn đoán ung thư vú bằng quang tuyến X). Do đó, đối chiếu kết quả thử nghiệm của X-quang tuyến với thực trạng của bệnh nhân, chúng ta có 4 khả năng:

Chẩn đoán ung thư vú	**Nghiên cứu y học**
K : bệnh nhân thật sự ung thư	H1 : giả thuyết chính là đúng
N : bệnh nhân không bị ung thư	H0 : giả thuyết vô hiệu đúng
+ve : kết quả xét nghiệm dương tính	S (P<0,05) : có ý nghĩa thống kê
-ve : kết quả xét nghiệm âm tính	NS (P>0,05) : không có ý nghĩa thống kê
Khả năng	**Khả năng**
Bệnh nhân quả thật bị ung thư vú, và kết quả xét nghiệm là dương tính; trong chẩn đoán y khoa, trường hợp này được gọi là ***dương tính thật*** hay ***độ nhạy*** (danh từ chuyên môn tiếng Anh gọi là ***sensitivity***). Phát biểu theo ngôn ngữ xác suất, đây chính P(+ve ǀ K).	Giả thuyết H1 đúng (chẳng hạn như thuốc có hiệu nghiệm), và kết quả phân tích có ý nghĩa thống kê. Đây là trường hợp mà các nhà nghiên cứu đề cập đến là ***power***. Nói theo xác suất: P(H1 ǀ S) = power, tương đương với dương tính thật.
Bệnh nhân quả thật bị ung thư, nhưng kết quả thử nghiệm lại âm tính; đây là trường hợp còn được gọi ngắn gọn là ***âm tính giả*** (***false negative***) hay P(-ve ǀ K).	Giả thuyết H1 đúng, nhưng kết quả phân tích không có ý nghĩa thống kê. Đây là trường hợp mà các nhà nghiên cứu đề cập đến là ***type II error*** (sai sót loại II). Nói theo xác suất: P(NS ǀ H1), tương đương với âm tính giả.
Bệnh nhân không bị ung thư, và kết quả thử nghiệm là âm tính; đây là trường hợp của ***âm tính thật*** hay ***độ đặc hiệu*** (***specificity***) hay P(-ve ǀ N)	Giả thuyết H0 đúng (tức thuốc không có hiệu quả), và kết quả phân tích cũng không có ý nghĩa thống kê. Đây là trường hợp mà các nhà nghiên cứu đề cập đến là ***confidence level***. Nói theo ngôn ngữ xác suất: P(NS ǀ H0), tương đương với âm tính thật.

Bệnh nhân quả thật không có ung thư, nhưng kết quả thử nghiệm là dương tính; đây là trường hợp của ***dương tính giả*** (***false positive***) hay P(+ve | K).	Giả thuyết H0 đúng, nhưng kết quả phân tích có ý nghĩa thống kê. Đây là trường hợp mà các nhà nghiên cứu đề cập đến là ***type I error*** (sai sót loại I). Nói theo xác suất: P(S | H0), tương đương với dương tính giả.

Ý nghĩa của độ nhạy, đặc hiệu, dương tính giả, âm tính giả có thể hiểu qua các giải thích sau đây:

- Độ nhạy (hay sensitivity, dương tính thật) có thể diễn giải như sau: nếu 100 bệnh nhân mắc bệnh đều đi xét nghiệm, có bao nhiêu người có kết quả dương tính.

- Độ đặc hiệu (specificity, âm tính thật) trả lời câu hỏi sau đây: nếu 100 người không mắc bệnh đều đi xét nghiệm, có bao nhiêu người có kết quả âm tính.

- Do đó, dương tính giả (false positive) là số người không mắc bệnh nhưng có kết quả xét nghiệm dương tính.

- Tương tự, âm tính giả (false negative) là số người mắc bệnh nhưng có kết quả xét nghiệm âm tính.

Một phương pháp chẩn đoán hoàn hảo là phương pháp có tỉ lệ dương tính thật và âm tính thật 100% (tức tỉ lệ dương tính giả và âm tính giả là 0%). Nhưng trong thực tế, không có phương pháp thử nghiệm nào là hoàn hảo cả. Thực vậy, bất cứ một phương pháp thử nghiệm y khoa nào, kể cả quang tuyến X, cũng đều có, không ít thì nhiều, tỉ lệ dương tính giả và âm tính giả. Hai sai sót này là đầu mối của nhiều vấn đề trong việc khám nghiệm ung thư vú.

Do đó, một kết quả xét nghiệm dương tính ***không*** có nghĩa là bệnh nhân mắc bệnh ung thư vú. Điều này đúng, bởi vì kết quả xét nghiệm có phản ảnh sai thực trạng của bệnh. Nên nhớ rằng các chỉ số như độ nhạy, độ đặc hiệu chỉ cho chúng ta biết độ chính xác của phương pháp xét nghiệm, chứ không cho biết khả năng mắc bệnh. Đây là một điều rất quan trọng mà rất tiếc rất nhiều bác sĩ không hay chưa nhận thức được.

Tương tự, trong nghiên cứu y học, một kết quả có ý nghĩa thống kê ($p<0,05$) không có nghĩa là giả thuyết đúng, bởi vì trị số P chỉ nói lên độ tin cậy của phương pháp kiểm định thống kê, chứ không phản ảnh độ khả dĩ của một giả thuyết khoa học. Vì không phân biệt được hai khái niệm này, nên rất nhiều nhà nghiên cứu diễn dịch sai ý nghĩa của trị số P và kết quả nghiên cứu.

3.1 Cần phân biệt P(+ve | K) và P(K | +ve)

Xin nhắc lại: P(+ve | K) là xác suất có kết quả xét nghiệm dương tính nếu cá nhân thật sự mắc bệnh (hay tỉ lệ những bệnh nhân mắc bệnh ung thư có kết quả dương tính), còn và P(K | +ve) là xác suất một cá nhân mắc bệnh nếu kết quả xét nghiệm dương tính (tức là trong số những người có kết quả dương tính, bao nhiêu người thật sự mắc bệnh).

Cần phải phân biệt hai chỉ số trên!

Vấn đề đặt ra là chúng ta cần biết chỉ số nào? Chúng ta không muốn biết nếu bệnh nhân mắc bệnh, xác suất mà bệnh nhân có kết quả dương tính là bao nhiêu, tức P(+ve | K), tức là độ nhạy. (Nếu bệnh nhân mắc bệnh thì chúng ta điều trị, chứ không cần hỏi câu hỏi ngược về quá khứ như thế!)

Đối với bác sĩ và bệnh nhân, khi nhận được kết quả xét nghiệm [hãy cho là] dương tính, người ta muốn biết xác suất mà cá nhân mắc bệnh là bao nhiêu. Tức là chúng ta muốn biết P(K | +ve). Trong chẩn đoán y khoa, thuật ngữ cho chỉ số này là *positive predictive value* (PPV), hay *giá trị tiên lượng dương tính*.

3.2 Ước tính P(K | +ve)

Giá trị tiên lượng dương tính tùy thuộc vào ba thông số: độ nhạy, độ đặc hiệu của phương pháp xét nghiệm, và tần số mắc bệnh trong cộng đồng (còn gọi là tỉ lệ lưu hành – prevalence). Theo thông lệ khoa học quốc tế, gọi độ nhạy là *Se*, độ đặc hiệu là *Sp,* và tỉ lệ lưu hành là *P*. Với ba thông số này, chúng ta có thể ước tính giá trị tiên lượng dương tính:

$$P\left(K \mid +ve\right) = \frac{P \times Se}{P \times Se + \left(1-P\right) \times \left(1-Sp\right)} \qquad [1]$$

Ví dụ: Nữ bệnh nhân người Mĩ, 50 tuổi, đi xét nghiệm ung thư vú và kết quả dương tính. Bệnh nhân muốn biết xác suất mà bà thật sự mắc bệnh là bao nhiêu? Y văn cho biết độ nhạy của phương pháp X quang (mammography) là 90% (tức Se = 0,90), và độ đặc hiệu là 95% (hay Sp = 0,95). Y văn cũng cho biết trong những người ở độu tuổi bệnh nhân, có khoảng 1% (hay P = 0,01). Dựa vào công thức trên, chúng ta có thể ước trả lời câu hỏi của bệnh nhân:

$$P\left(K \mid +ve\right) = \frac{0,01 \times 0,90}{0,01 \times 0,9 + \left(1-0,01\right) \times \left(1-0,95\right)} == 0,15$$

Nói cách khác, xác suất mà bệnh nhân thật sự mắc bệnh nếu kết quả xét nghiệm dương tính là 15%. Nói cụ thể hơn, cứ 100 phụ nữ như bệnh nhân có kết quả xét nghiệm dương tính, khoảng 15 người thật sự mắc bệnh ung thư vú. Tuy nhiên, chúng ta vẫn không biết vị phụ nữ đó nằm trong số 15 bệnh nhân hay không!

3.3 Ước tính P(H1 | S)

Tương tự, trong nghiên cứu y học, chúng ta cũng muốn biết nếu kết quả kiểm định có ý nghĩa thống kê (S) thì xác suất mà giả thuyết chính đúng là bao nhiêu. Nói cách khác, chúng ta muốn biết P(H1 | S).

Cũng như trong chẩn đoán y khoa, P(H1 | S) tùy thuộc vào ba thông số: power hay P(S | H1), sai sót loại I, và xác suất mà giả thuyết H1 đúng là bao nhiêu hay P(H1). Gọi sai sót loại I là α, chúng ta có thể ước tính P(H1 | S) như sau:

$$P(H1 \mid S) = \frac{P(H1) \times power}{P(H1) \times power + \left[1 - P(H1)\right] \times \alpha} \qquad [2]$$

Trong công thức trên, hai thông số đầu (power và sai sót loại I) thường được hoạch định trước khi nghiên cứu được thực hiện. Thông thường, power dao động trong khoảng 0,80 đến 0,90, và sai sót loại I thường α = 0,01 đến 0,05. Nhưng P(H1) có lẽ là thông số khó nhất trong nghiên cứu, vì trong nhiều trường hợp chúng ta không biết xác suất H1 là bao nhiêu. Tuy nhiên, tùy trường hợp cụ thể, chúng ta có thể tiếp cận P(H1) qua tần số của một sự kiện. Chẳng hạn như trong nghiên cứu về mối liên hệ giữa một gien và bệnh, trong số 30.000 gien, xác suất mà một gien có liên hệ đến bệnh có thể là 1/30.000, hoặc cao hơn chút ít nếu có bằng chứng khoa học làm cơ sở.

Ví dụ: Một nghiên cứu về mối liên hệ giữa gien VDR và loãng xương, các nhà nghiên cứu ước tính rằng họ cần 1000 đối tượng để có power 90% và sai sót loại I là 1%. Kết quả phân tích thống kê cho thấy mối liên hệ có ý nghĩa thống kê với trị số P = 0,015. Câu hỏi đặt ra là xác suất mà giả thuyết về mối liên hệ giữa VDR và loãng xương là bao nhiêu? Chúng ta tạm thời cho xác suất P(H1) = 1/30000 = 0,0000333. Áp dụng công thức trên, chúng ta có:

$$P(H1 \mid S) = \frac{0,0000333 \times 0,9}{0,0000333 \times 0,9 + \left[1 - 0,0000333\right] \times 0,05} = 0,0006$$

Nói cách khác, cho dù kết quả có ý nghĩa thống kê, nhưng xác suất mà VDR thật sự có liên quan đến loãng xương chỉ 0,06% -- một mối liên hệ còn quá nhiều bất định.

Công thức (1) và (2) vừa trình bày trên chính là Định lí Bayes (Bayesian theorem) rất nổi tiếng trong xác suất học [13]. Định lí Bayes phát biểu rằng có thể ước tính xác suất một sự kiện sau khi đã có dữ liệu quan sát sát hay đo lường được. Nói một cách thực tế hơn, có thể xem Định lí Bayes là qui trình cập nhật hóa kiến thức. Trong ví dụ về chẩn đoán trên, trước khi xét nghiệm, chúng ta biết rằng xác suất mà người phụ nữ đó mắc bệnh là 1% (tỉ lệ lưu hành). Sau khi kết quả xét nghiệm dương tính, xác suất này tăng lên 15% -- hay 15 lần. Tương tự, trước khi làm nghiên cứu, chúng ta có thể nói rằng xác suất gien VDR liên hệ đến loãng xương là 0,0000333, nhưng sau khi có số liệu "dương tính", chúng ta có thể nói xác suất của mối liên hệ này lên 0,0006, tức tăng gần 1800 lần, nhưng vẫn còn nhiều bất định.

4. Yếu tố Bayes

Một trong những khó khăn trong việc ước tính P(H1 | S) theo Định lí Bayes như vừa trình bày vẫn là xác định thông số P(H1), hay còn gọi là xác suất tiền định của một giả thuyết (*prior probability of a hypthesis*). Đây cũng chính là điểm gây ra nhiều tranh luận đậm màu sắc triết học trong suốt 100 năm qua.

Một cách khách quan hơn để đánh giá hai giả thuyết là so sánh trực tiếp khả năng của hai giả thuyết đó. Thay vì ước tính trực tiếp xác suất một giả thuyết, chúng ta có thể ước tính xác suất dữ liệu cho một giả thuyết. Gọi *D* (viết tắt từ data) là dữ liệu, H0 là giả thuyết vô hiệu, và H1 là giả thuyết chính, chúng ta định nghĩa:

- P(*D* | H0) là xác suất dữ liệu quan sát được nếu giả thuyết H0 đúng; và
- P(*D* | H1) là xác suất dữ liệu quan sát được nếu giả thuyết H0 đúng.

Yếu tố Bayes (Bayes Factor – BF) [14-15] được định nghĩa như là tỉ số của hai xác suất trên:

$$BF = \frac{P(D \mid H1)}{P(D \mid H0)} \qquad [3]$$

Nếu chúng ta xem dữ liệu *D* là bằng chứng, thì Yếu tố Bayes chính là một đo lường bằng chứng nghiêng về giả thuyết nào. Nhìn qua công thức trên chúng ta có thể thấy:

- Nếu BF = 1, bằng chứng không nghiêng về một giả thuyết nào cả (hai giả thuyết có xác suất như nhau);
- Nếu BF > 1, bằng chứng nghiêng về (yểm trợ) giả thuyết H1 hơn là H0;

- Ngược lại, nếu BF < 1, bằng chứng nghiêng về (yểm trợ) giả thuyết H0 hơn là H1.

Theo một qui ước chung, cách diễn dịch Yếu tố Bayes như sau:

Yếu tố Bayes (BF)	**Bằng chứng nghiêng về H1 ở mức độ**
BF = 3 đến BF = 10	đáng kể (substantial evidence)
BF = 10 đến BF = 30	thuyết phục (strong evidence)
BF = 30 đến BF = 100	rất thuyết phục (very strong evidence)
BF > 100	gần như xác định

Ví dụ: Trong nạn dịch tiêu chảy vào cuối năm 2007 ở một số tỉnh phía Bắc, một số quan chức y tế cho rằng mắm tôm là nguyên nhân, là nguồn gốc của nạn dịch, vì họ nghi rằng mắm tôm hàm chứa vi khuẩn gây bệnh tả (*Vibrio cholerae*). Viện vệ sinh dịch tễ trung ương xét nghiệm 75 mẫu mắm tôm được chọn ngẫu nhiên từ Hà Nội, Nghệ An, và Thanh Hóa. Kết quả xét nghiệm tất cả đều âm tính (không có vi khuẩn tả). Chúng ta có thể diễn giải bằng chứng này như thế nào?

Gọi π là xác suất mắm tôm chứa vi khuẩn tả. Chúng ta biết rằng theo luật phân phối nhị phân (binomial distribution), nếu xác suất nhiễm tả là π, và nếu chúng ta xét nghiệm n mẫu, thì xác suất có k mẫu bị nhiễm là:

$$P\left(k \mid \pi, n\right) = \binom{n}{k} \pi^k \left(1-\pi\right)^{n-k}$$

Gọi H0 là giả thuyết mắm tôm không hàm chứa vi khuẩn tả, do đó, $\pi = 0$. Với 75 mẫu mắm tôm được xét nghiệm, chúng ta có $k = 0$ (không có kết quả dương tính). Do đó, xác suất $k = 0$ dưới giả thuyết H0 là:

$$P\left(D \mid H0\right) = P\left(0 \mid 0,75\right) = \binom{75}{0} 0^0 \left(1-0\right)^{75-0} = 1$$

Nếu H1 là giả thuyết mắm tôm có vi khuẩn tả, chúng ta hãy cho rằng 20% mắm tôm nhiễm khuẩn, và do đó: $\pi = 0{,}20$. Xác suất dữ liệu ($k = 0$) dưới giả thuyết này là:

$$P(D \mid H1) = P(0 \mid 0{,}2;75) = \binom{75}{0{,}2} (0{,}2)^0 \left(1-0{,}2\right)^{75-0} = (5{,}39)^{-8}$$

Do đó, Yếu tố Bayes, theo định nghĩa (3) là:

$$BF = \frac{P(D \mid H1)}{P(D \mid H0)} = \frac{1}{(5.39)^{-8}} = 18.546.031$$

Nói cách khác, bằng chứng (dữ liệu từ 75/75 âm tính) nghiêng về giả thuyết mắm tôm không nhiễm vi khuẩn tả đến 18,5 triệu lần!

Trên đây là một cách tính tương rất đơn giản để minh hoạ cho ý nghĩa của Yếu tố Bayes. Trong thực tế, các nghiên cứu với các phân tích phức tạp, cách tính Yếu tố Bayes cũng rất phức tạp. Tuy nhiên, chúng ta có thể ước tính giá trị tối thiểu của Yếu tố Bayes có thể ước tính bằng một công thức rất đơn giản, chỉ là hàm số của trị số *p*, mà Sellke và đồng nghiệp [16-17] phát triển như sau:

$$BF_{min} > 1 / (-e\ p\ ln(p)) \qquad [4]$$

Trong đó e = 2,71828. Chẳng hạn như một nghiên cứu với trịo số p = 0,05, Yếu tố Bayes tối thiểu là: 1 / (-2,71828 x 0,05 x log(0,05)) = 2,45. Theo cách hiểu thông thường, khi p <0,05, các nhà nghiên cứu kết luận rằng kết quả "có ý nghĩa thống kê" (significant), nhưng với cách tính khách quan trên, chúng ta thấy bằng chứng vẫn chưa thuyết phục. Nhưng khi trị số p rất thấp như ví dụ trên với p = 0,0009, thì giá trị tối thiểu của BF là 1/(-2,71828 x 0,0009 x log(0,0009)) = 58,3. Nói cách khác, bằng chứng có vẻ nghiêng về giả thuyết H1 nhiều hơn là giả thuyết H0.

Qua định lí Bayes [xem chú thích 3], chúng ta biết rằng $P(H \mid D) = \frac{P(D \mid H) \times P(H)}{P(D)}$. Dùng định lí này và qua vài thao tác đại số, chúng ta có thể diễn tả xác suất tối đa của P(H+ | s) như là hàm số của BF như sau:

$$P(H1 \mid S) = \left[1 + \frac{\left[1 - P(H1)\right]}{BF \times P(H1)}\right]^{-1}$$

Do đó, theo ví dụ trên, với giá trị tối thiểu BF là 58,3, và P(H1) = 0,5, chúng ta có thể ước tính xác suất tối đa của P(H1 | S) là 0,983. Nếu chúng ta chấp nhận xác suất >0,95 để kết luận, thì qua cách tính này, chúng ta có bằng chứng (p =0,0009) để kết luận rằng giả thuyết H1 có xác suất đúng lên đến 98%.

Hãy lấy một ví dụ khác: mới đây báo chí khá quan tâm về một nghiên cứu mà trong đó các nhà nghiên cứu phát hiện rằng tỉ lệ bị ung thư vú trong các phụ nữ dùng thuốc aspirin (giảm đau) cao hơn các phụ nữ không dùng aspirin khoảng 20% [6]. Kết luận này chỉ đơn thuần dựa vào trị số p = 0,022, tức "có ý nghĩa". Các nhà nghiên cứu không giải thích được hiện tượng này, và phát hiện cũng nằm ngoài dự đoán sinh học của

họ. Nói cách khác, ở đây xác suất giả thuyết H1 rất thấp, có thể chỉ 0,01 (tức 1%). Và nếu P(H1) = 0,01, và giá trị tối thiểu BF là 1/[-2,71828 x 0,022 x log(0.022)] = 4,38, xác suất tối đa của P(H1 | S) chỉ 0,042 hay 4,2%.

Cho dù P(H1) = 0,1 đi nữa, xác suất tối đa của P(H1 | S) cũng chỉ 0,33. Vì xác suất P(H1 | S) thấp hơn 0,95, chúng ta có thể phát biểu rằng giả thuyết về mối liên hệ giữa aspirin và ung thư vú chưa có bằng chứng thuyết phục, hay bằng chứng hiện có không nhất quán với giả thuyết đó. Nói cách khác, các nhà nghiên cứu có thể đã đi đến một kết luận sai và phát hiện của họ có thể là một phát hiện dương tính giả!

5. Vài nhận xét và kết luận

Trị số p là một số có ảnh hưởng cực kì lớn đến hoạt động 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!

Nhưng cần phải nhấn mạnh một lần nữa để hiểu ý nghĩa của trị số P như sau: Mục đích của trị số p là nhằm trả lời câu hỏi: *nếu giả thuyết vô hiệu H0 đúng, thì xác suất mà dữ liệu chúng ta quan sát được là bao nhiêu?* Nói cách khác, đó chính là phương pháp chứng minh đảo ngược. Do đó, diễn dịch trị số P phải có điều kiện. Trị số P không cung cấp cho chúng ta một định lượng gì nói đến một giả thuyết.

Trong suốt một thế kỉ qua khoa học thực nghiệm dựa vào trị số p của trường phái thống kê [có khi] gọi là *frequentist* (**trường phái tần số**) để suy luận và đi đến kết các luận khoa học. Cách suy luận này hiện vẫn là cách làm việc chuẩn trong khoa học. Thế nhưng cái logic đằng sau trị số p có rất nhiều vấn đề, kể cả sự phản trực giác (counter-intuitive) và rất khó hiểu, có khi ... phi logic. Theo trường phái tần số, xác suất được định nghĩa chỉ qua "thử nghiệm" (experiments) mà trên lí thuyết các thử nghiệm có thể lặp đi lặp lại nhiều lần đến vô tận, trong những điều kiện giống nhau nhưng độc lập với nhau. Nói "độc lập" có nghĩa là thử nghiệm thứ hai không có liên quan gì đến thử nghiệm thứ nhất hay bất cứ thử nghiệm nào sau đó. Ví dụ như một đồng xu được quăng 1 lần, thì đó cũng chính là một "thử nghiệm", và nếu đồng xu được quăng liên tục 1 triệu lần cũng có nghĩa là 1 triệu thử nghiệm, và các thử nghiệm này độc lập với nhau. Theo cách hiểu này, xác suất có nghĩa là số lần một sự kiện xảy ra trong vô số thử nghiệm đó, và tần số này được diễn đạt qua con số tỉ lệ hay phần trăm. Nói cách khác, xác suất là một *tần số tương đối* (relative frequency).

Nói cho cùng, xác suất là một cảm nhận cá nhân, là mức độ tin tưởng của một cá nhân về một sự kiện hay hiện tượng nào đó. Nói cách khác, xác suất phản ánh kinh nghiệm cá nhân, hay khả năng của cá nhân đó tích lũy và phân tích thông tin từ các nguồn ngoại tại. Do đó, câu phát biểu "xác suất aspirin gây ra ung thư vú là 0.33" phản ánh mức độ tin tưởng của người phát biểu đối với mối liên hệ giữa aspirin và ung thư vú. Vì là cảm nhận cá nhân, con số đó cũng được cảm nhận khác nhau giữa các cá nhân: đối với ông A, 0,33 có thể là mức độ tin tưởng còn thấp; nhưng đối với chị B, 0.33 có thể là một khả dĩ cao. Vì là cảm nhận cá nhân, con số xác suất không phải là một chỉ số khách quan như cách hiểu của trường phái tần số. Theo trường phái tần số, "xác suất nữ thông minh hơn nam là 0,98" có thể được diễn dịch nhiều cách khác nhau: nó có thể có nghĩa là trong 100 cặp nam nữ được chọn một cách ngẫu nhiên, có 98 cặp mà trong đó chỉ số IQ của nữ cao hơn nam; nó cũng có thể có nghĩa là nếu nghiên cứu được lặp lại 100 lần, mỗi lần với đối tượng khác nhau, có 98 nghiên cứu cho thấy số trung bình IQ của phái nữ cao hơn phái nam. Tất nhiên, trong thực tế ít ai – nếu không muốn nói là chẳng ai – chịu khó lặp lại nghiên cứu 100 lần hay 1000 lần; do đó, cách diễn dịch của trường phái tần số rất ư là phi thực tế.

Trong suy luận khoa học, có thể nói không ngoa rằng chỉ có suy luận dựa vào Định lí Bayes là logic nhất. Tuy trị số p = P(D | H0) và trị số P(H1 | D) hay P(H1 | S)đều là xác suất, nhưng trị số p theo trường phái tần số cho chúng ta biết nhiều về tính chính xác của nghiệm toán thống kê, hơn là về mức độ khả dĩ của một giả thuyết nghiên cứu. Đối với nhà nghiên cứu chỉ có P(H1 | S) là có ý nghĩa, cũng như đối với bệnh nhân chỉ có P(K | +ve) là có ý nghĩa. Muốn ước tính mức độ khả dĩ của một giả thuyết nghiên cứu, chúng ta cần phải ứng dụng Định lí Bayes và các phương pháp liên quan đến Định lí Bayes. Qua bài viết mang tính giới thiệu này, tác giả hi vọng thuyết phục bạn đọc, nhất là các nhà nghiên cứu thực nghiệm, nên tìm hiểu và tiếp cận các phương pháp thống kê thuộc trường phái Bayes, hiện đang rất thịnh hành trong lĩnh vực y sinh học, vật lí học, và ngay cả tin học. Hi vọng bạn đọc sẽ có dịp đóng góp vào sự phát triển của các phương pháp Bayes trong tương lai và làm cho suy luận khoa học hoàn hảo hơn và logic hơn.

Chú thích và tài liệu tham khảo:

[1] Lyles KW, et al. Zoledronic acid and clinical fractures and mortality after hip fracture. *N Engl J Med* 2007 Nov 1;357(18):1799-809.

[2] Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? *Statistics in Medicine* 1987; 6:3-10.

[3] Karl Popper (28/07/1902- 17/09/1994), người Áo, Ông được coi là một triết gia khoa học hàng đầu của thế kỉ XX. Tác phẩm chính đầu tiên, Logik der Forschung (The Logic of Reseach), xuất bản năm 1934, được coi như là một tác phẩm kinh điển của phép phản nghiệm, một trường phái phổ biến của chủ nghĩa thực chứng logic (logical positivism), rồi tiếp cận đến khoa học được gọi là "chủ nghĩa phản nghiệm" (falsificationism), mà cơ sở dựa trên phép phê phán hơn là xác minh. Từ đó mà ông đã được thỉnh giảng ở Anh quốc, mà sau này trở thành quê hương thứ hai

của ông. Từ lí thuyết phản nghiệm của ông mà sau này người ta có thể phân định sự khác biệt giữa khoa học với nguỵ khoa học. Ông nhận được rất nhiều giải thưởng vinh dự của cả Hiệp hội Khoa học Chính trị Mĩ, Viện Hàn lâm Anh v.v.. Ông đã được Nữ hoàng Elisabeth II phong tước hiệp sĩ năm 1965, và Huân chương Danh dự năm 1982. Ngoài tác phẩm nổi tiếng nêu trên ông đã công hiến cho khoa học thế giới nhiều tác phẩm vô giá về triết lí khoa học.

[4] Để biết triết lí phản nghiệm trong nghiên cứu lâm sàng, có thể đọc bài của Senn SJ. Falsificationism and clinical trials. *Stat Med* 1991 Nov;10(11):1679-92.

[6] Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 1922; 85(1):87-94.

[6] Fisher RA. *Statistical Methods for research workers*. Oliver and Boyd, 1954.

[7] Neyman J, Pearson E. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 1933; 231: 289-337.

[8] Xem thêm chi tiết về những tranh luận liên quan đến kiểm định ý nghĩa thống kê và kiểm định giả thuyết trong sách *The Significance Test Controversy*, do DE Morrison và RE Henkel biên tập, Nhà xuất bản Aldine, Chicago: 1970.

[9] Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L. *The Empire of Chace: How Probability Changed Science and Everyday Life*. Cambridge University Press, 1989.

[10] Barnard GA. Must clinical trials be large? The interpretation of P-values and the combination of test results. *Stat Med* 1990;9(6):601-14.

[11] Barnard GA. On alleged gains in power from lower P-values. *Stat Med* 1989;8(12):1469-77.

[12] Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS; ACTS Study Group. Vitamins C and E and the risks of preeclampsia and perinatal complications. *N Engl J Med* 2006;354(17):1796-806.

[13] Thomas Bayes (1702 – 1761) là một linh mục sống ở Anh vào thế kỉ 18. Ngoài công việc giảng giáo lí, ông còn là nhà toán học có hạng. Năm 1763 (tức sau khi ông qua đời), một người đồng nghiệp của ông công bố một công thức xác suất mà ngày nay được biết đến là *Định lí Bayes* (Bayesian theorem) do ông viết lúc còn sống như vì quá cẩn thận nên ông không cho xuất bản. Định lí này có một ảnh hưởng cực kì to lớn trong nghiên cứu khoa học và chẩn đoán y khoa, nhưng cũng là một định lí gây ra nhiều tranh cãi gay gắt trong khoa học suốt 2 thế kỉ qua (mà tôi sẽ đề cập đến trong một dịp khác). Để giải thích định lí này ngắn gọn, có lẽ chúng ta cần phải điểm qua vài sự thật cơ bản về xác suất có điều kiện (conditional probability).

Để tiện theo dõi lí giải, tôi sẽ dùng kí hiệu H là giả thuyết và D là dữ kiện như đế cập trong phần đầu của bài viết. Như chúng ta biết, nếu hai hiện tượng H và D độc lập, thì xác suất có điều kiện phát biểu rằng:

$$P(D \cap H) = P(D|H) \times P(H) \qquad [A1]$$

Nói cách khác, P(D|H)= P(D∩H) / P(H), với điều kiện dĩ nhiên là P(H) không phải 0. Đến đây bạn đọc đã thấy P(D|H) chính là sai sót loại I mà tôi đã đề cập. Hay nói cụ thể hơn P(D|H) chính là P(S|H0) sau khi nghiên cứu dữ kiện đã được phân tích bằng một kiểm định thống kê.

Nhưng vấn đề là chúng ta muốn ước tính P(H|D). Một vài sắp xếp công thức [A1] chúng ta sẽ đi đến định lí Bayes:

$$P(H|D) = P(D|H) \times P(H) / P(D) \qquad [A2]$$

Ý nghĩa của định lí Bayes trên đây là muốn ước tính xác suất một giả thuyết H sau khi đã quan sát dữ kiện D, thì chúng ta phải biết xác suất của dữ kiện hay P(D), và quan trọng hơn hết là xác suất của giả thuyết, tức P(H).

Muốn tìm hiểu thêm về lí thuyết và ứng dụng thống kê theo trường phái Bayes (Bayesian Statistics) có thể tham khảo các sách mang tính nhập môn sau đây: (1) sách về lí thuyết: Peter M. Lee, *Bayesian Statistics,* 2nd Edition, London: Arnold, 1997; (2) sách về ứng dụng: Donald A. Berry, *Statistics: A Bayesian Perspective*, Belmont: Duxbury Press, 1996; (3) hay sách cho các nhà vật lí học: Giulio D'Agistini, *Bayesian Reasoning in Data Analysis,* World Scientific, 2003.

[14] Jeffreys H. *The Theory of Probability* (3e), Oxford (1961); trang 432.

[15] Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130 (12): 1005-13.

[16] Sellke T, Bayarri MJ, Berger JO. Calibration of p-values for testing precise null hypothesis. *The American Statistician* 2001.

[17] 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.