Y học thực chứng

Diễn giải trị số P và khoảng tin cậy 95%

Nguyễn Văn Tuấn
Viện nghiên cứu Y khoa Garvan, Sydney, Australia

Trị số P có lẽ là một chỉ số phổ biến nhất trong các công trình nghiên cứu lâm sàng, và cũng chính vì sự phổ biến mà nó cũng dễ bị hiểu lầm và lạm dụng. Một nghiên cứu ở một nhóm bác sĩ chuyên khoa và có kinh nghiệm trong nghiên cứu y học cho thấy có đến 85% không hiểu ý nghĩa của trị số P [1]. 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!

Vậy thì ý nghĩa thật của trị số *P* là gì? Để trả lời câu hỏi này, chúng ta phải điểm qua triết lí khoa học, bởi vì mô hình nghiên cứu lâm sàng đối chứng ngẫu nhiên (randomized controlled trial – RCT) dựa vào triết lí phản nghiệm (falsificationism).

Theo Theo Karl Popper [3], cha đẻ của triết lí phản nghiệm, một giả thuyết được xem là mang tính "khoa học" nếu giả thuyết đó có khả năng "phản nghiệm". 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] "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.

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 Ramipril làm giảm nguy cơ tử vong, thì giả thuyết vô hiệu sẽ phát biểu là "tỉ lệ tử vong ở bệnh nhân được điều trị bằng Ramipril **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.

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 α = 5% và β = 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 nổi 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!

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 Ramipril trong việc phòng chống tử vong và đột quị [1]. Với giả thuyết rằng thuốc có hiệu nghiệm giảm nguy cơ tử vong và đột quị, các nhà nghiên cứu so sánh tỉ lệ tử vong và độ quị giữa hai nhóm bệnh nhân: nhóm 1 được điều trị bằng Ramipril và nhóm 2 là nhóm giả được (placebo). 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 1. Hiệu quả của Ramipril giảm nguy cơ nhồi máu cơ tim, đột quị, tử vong và tiểu đường thể hiện qua tỉ số nguy cơ và trị số P

	Tỉ số nguy cơ (relative risk) và khoảng tin cậy 95%	Trị số P
Nhồi máu cơ tim, đột quị, tử vong	0.83 (0.75 – 0.91)	0.0002
Tỉ lệ mới mắc bệnh tiểu đường	0.69 (0.56 – 0.85)	0.0006

HOPE/HOPE-TOO Study Investigators. Long-term effects of Ramipril on cardiovascular events and diabetes. Results of the HOPE Study Extension. Circulation 2005; 112:1339-1346.

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 tử vong và đột quị giữa hai nhóm có ý nghĩa thống kê. Tất nhiên, *trị số P trên không có nghĩa là nghiên cứu đã chứng minh rằng thuốc Ramipril có hiệu quả giảm nguy cơ tử vong và đột quị*. Nó có nghĩa là nếu thật sự thuốc Ramipril không có hiệu quả giảm nguy cơ tử vong và đột quị thì xác suất mà các nhà nghiên cứu quan sát các số liệu trên là 0.0002.

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*.

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 | 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 tăng 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 tăng 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).

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 hiệu quả của Ramipril, các nhà nghiên cứu có thể phân tích hiệu quả của Ramipril qua nhiều tiêu chí lâm sàng như nguy cơ tử vong, nhồi máu cơ tim, đột quị, tỉ lệ mắc bệnh tiểu đường, v.v… từng giới tính, từng nhóm tuổi, hay phân tích theo các đặc tính lâm sàng của bệnh nhân. 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 kiểm định giả thuyết và với $P = 0.05$, xác suất mà một kiểm định 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.

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 ngưỡng chuẩn "có ý nghĩa thống kê"* (tức $P > 0.05$).

Bảng 2 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 2. 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 ngưỡng 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!

Khoảng tin cậy 95%

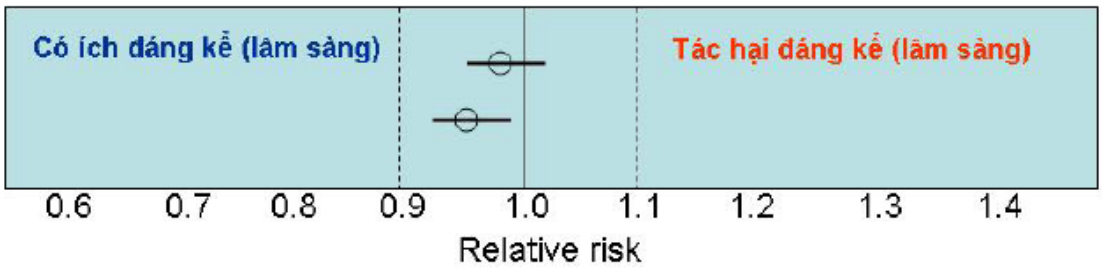
Trong vài năm gần đây, càng ngày giới nghiên cứu y khoa càng nhận thức được những khiếm khuyết quan trọng của trị số *P*, nên có một xu hướng mới không đặt nặng diễn giải kết quả nghiên cứu qua trị số *P* và tập trung vào ước tính hiệu quả. Hiệu quả của một thuốc thường được mô tả qua chỉ số tỉ số nguy cơ – *relative risk* (hay viết tắt là *RR*) và khoảng tin cậy 95% (95% confidence interval, CI). Chẳng hạn như trong ví dụ (bảng 1) về hiệu quả của Ramipril giảm nguy cơ tử vong, đột quị và nhồi máu cơ tim, các nhà nghiên cứu báo cáo rằng tỉ số nguy cơ là 0.83 (với khoảng tin cậy 95% dao động trong khoảng 0.75 đến 0.91). Kết quả này nên được hiểu như thế nào?

Trước hết, bởi vì *RR* lấy nguy cơ của nhóm điều trị làm tử số và nguy cơ của nhóm chứng làm mẫu số, cho nên khi *RR* < 1 có nghĩa là thuốc giảm nguy cơ bệnh / biến cố lâm sàng. Chẳng hạn như trong trưởng hợp trên, nguy cơ tử vong, đột quị và nhồi máu cơ tim trong nhóm được điều trị bằng Ramipril bằng 0.83 nguy cơ nhóm chứng. Nói cách khác, Ramipril giảm nguy cơ tử vong, đột quị và nhồi máu cơ tim 17% (lấy 1 trừ cho 0.83).

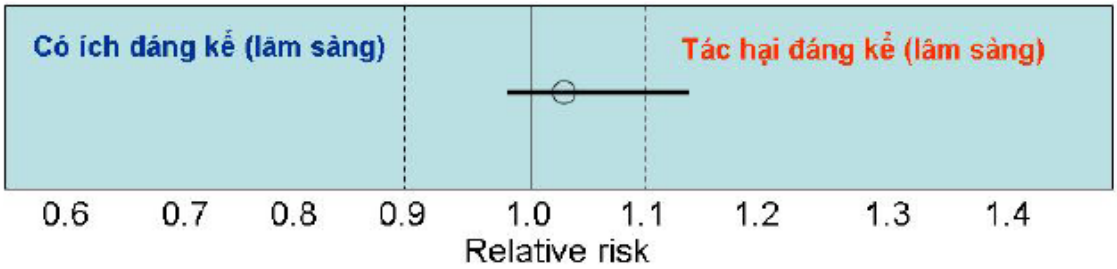
Thứ hai, dưới vài giả định cơ bản, có thể phát biểu rằng khoảng tin cậy 95% dao động trong khoảng 0.75 đến 0.91 có nghĩa là xác suất mà *RR* dao động trong khoảng khoảng 0.75 đến 0.91 bằng 95%. Cũng có thể nói cách khác: nếu nghiên cứu được lặp lại 100 lần trong điều kiện giống như nghiên cứu Ramipril thì hiệu quả trung bình của 100 nghiên cứu sẽ là 0.83, và có 95 nghiên cứu sẽ có *RR* thấp 0.91 hay cao đến 0.75.

Thứ ba, một khoảng tin cậy 95% [KTC95%] của *RR* không bao gồm 1 cũng có nghĩa là trị số *P* phải thấp hơn 0.05. Nói cách khác, một *RR* với khoảng tin cậy 95% không bao gồm 1 cũng có nghĩa là kết quả đó có ý nghĩa thống kê, hiểu theo nghĩa *P* < 0.05.

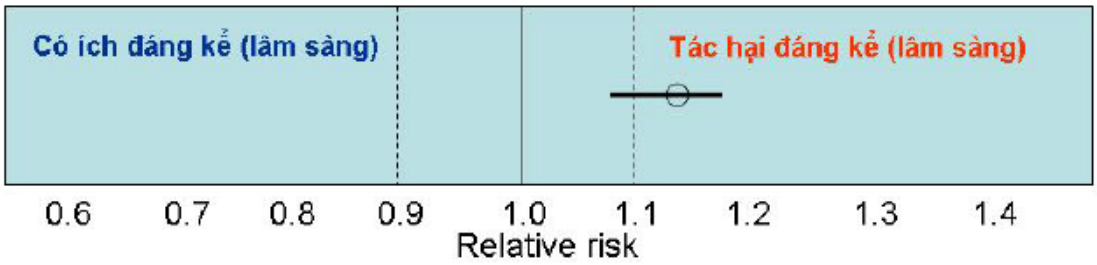
Tuy nhiên, khoảng tin cậy 95% cũng có thể diễn giải với ý nghĩa lâm sàng. Như đề cập trên, một kết quả có ý nghĩa thống kê (*P*<0.05 hay KTC95% không hàm chứa 1) nhưng không có ý nghĩa lâm sàng; ngược lại, một kết quả không có ý nghĩa thống kê (*P* > 0.05 hay KTC95% hàm chứa 1) nhưng có thể có ý nghĩa lâm sàng. Để diễn giải KTC95%, chúng ta cần xác định độ ảnh hưởng *K* mà ảnh hưởng trên hoặc dưới *K* được xem là có ý nghĩa lâm sàng. Chẳng hạn như *K* = 0.1 thì bất cứ *RR* thấp hơn 0.9 xem là tác dụng có ý nghĩa lâm sàng, và *RR* cao hơn 1.1 có thể xem là tác hại có ý nghĩa lâm sàng. Ý nghĩa của khoảng tin cậy có thể diễn giải theo tiêu chí này như sau:



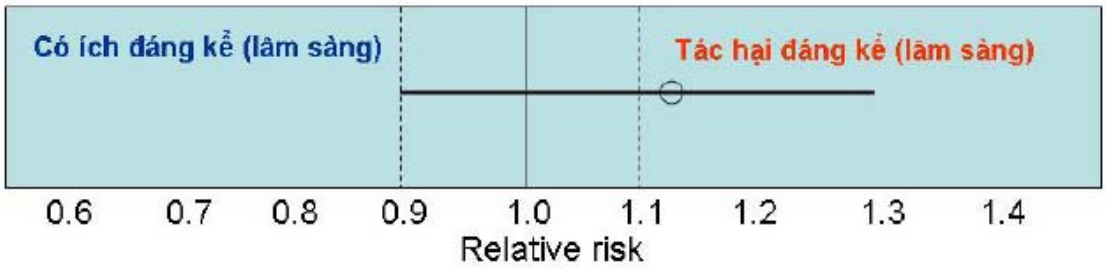
Ý nghĩa: có hiệu quả (*RR* thấp hơn 1), nhưng không có tác hại lâm sàng, bởi vì tất cả khoảng tin cậy 95% dao động trong khoảng 0.9 đến 1.1.



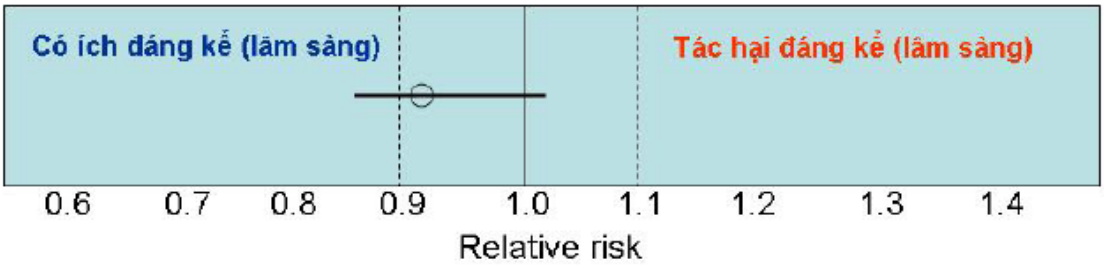
Ý nghĩa: Có thể có tác hại (vì KTC95% hàm chứa một xác suất nhỏ thuốc có thể có hại. Nhưng rõ ràng là thuốc không có lợi ích lâm sàng.



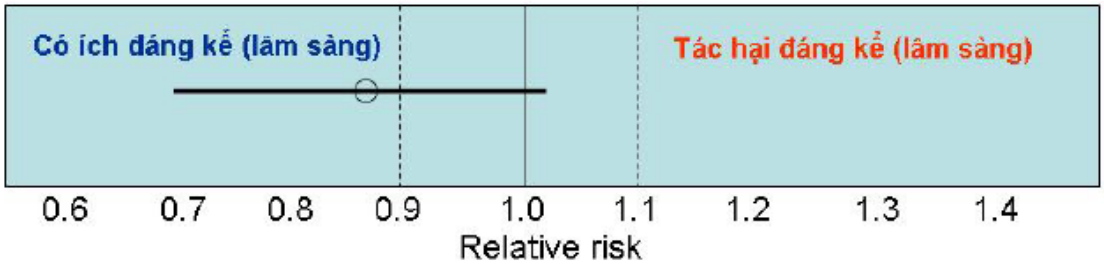
Ý nghĩa: Rất có khả năng có hại. KTC95% nghiêng hẳn về phía *RR* > 1 và xác suất có hại khá cao.



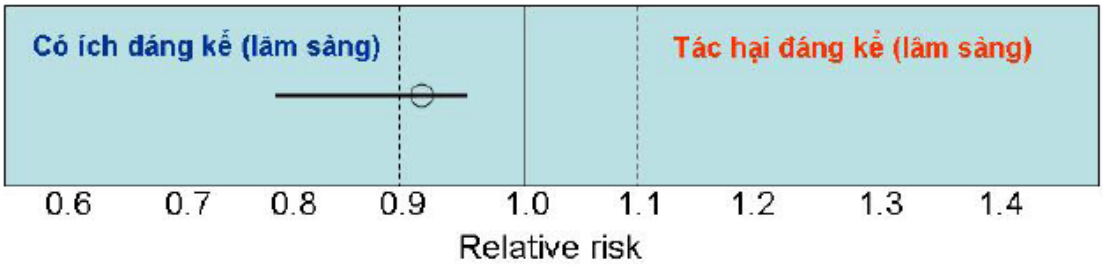
Ý nghĩa: Kết quả này cho thấy thuốc có thể có tác hại nghiêm trọng, nhưng vì KTC95% quá rộng, chúng ta không thể kết luận dứt khoát. Tuy nhiên kết quả cho thấy thuốc chẳng có ích lợi lâm sàng vì *RR* thấp nhất chỉ 0.9.



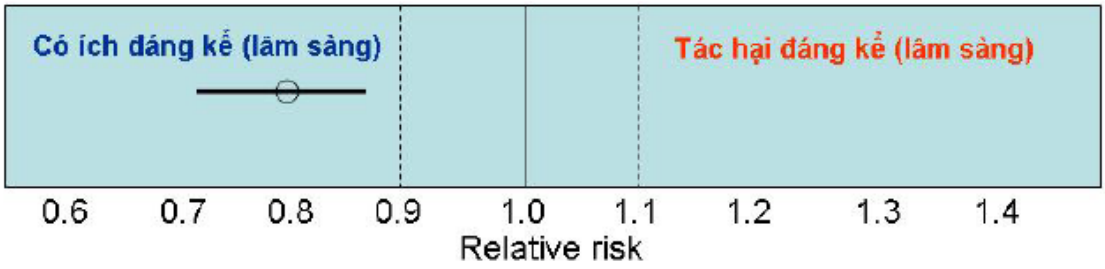
Ý nghĩa: Có thể thuốc có lợi lâm sàng, nhưng chưa chắc chắn, bởi vì KTC95% bao gồm 1. Có bằng chứng cho thấy thuốc không gây tác hại lâm sàng.



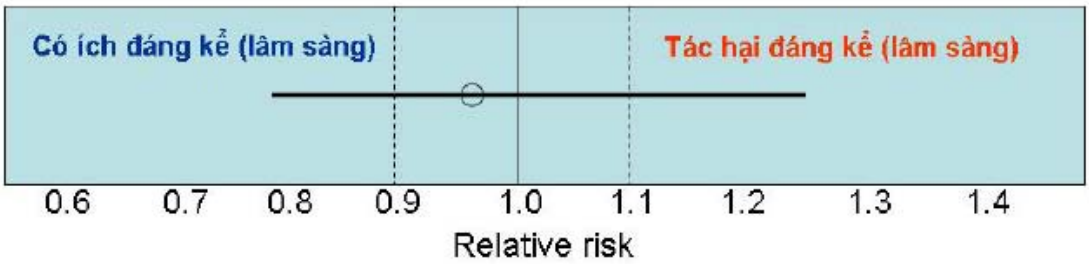
Ý nghĩa: kết quả này cho thấy thuốc có thể đem lại lợi ích đáng kể, nhưng KTC95% quá rộng nên tình trạng bất định còn quá cao.



Ý nghĩa: kết quả này cho thấy thuốc có lợi ích (và có ý nghĩa thống kê), nhưng tác dụng lâm sàng chưa rõ ràng



Ý nghĩa: kết quả này cho thấy thuốc rõ ràng đem lại lợi ích cho bệnh nhân.



Ý nghĩa: chúng ta không thể rút được gì từ kết quả trên!

Vấn đề then chốt trong các diễn giải kết quả với khoảng tin cậy 95% là xác định *K,* mức độ ảnh hưởng có ý nghĩa lâm sàng. Trong các kết quả vừa trình bày trên, nếu thuốc tăng nguy cơ biến cố 10% hay giảm 10% đều được xem là không có ý nghĩa lâm sàng. Tuy nhiên, ngưỡng *K* còn tùy thuộc vào biến cố lâm sàng mà chúng ta nghiên cứu và các chỉ số kinh tế. Chẳng hạn như nếu tiêu chí lâm sàng là tử vong thì dù tăng 5% cũng có thể xem là có ý nghĩa lâm sàng, hay giảm 5% cũng có thể xem là có hiệu quả lâm sàng, nhưng đối với các biến như đau nhức thì tăng hay giảm 20% có thể không có ý nghĩa lâm sàng.

Kết luận

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 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". Khoảng 25% các phát hiện với "$P < 0.05$" là các phát hiện dương tính giả.

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. Xu hướng chung hiện nay là xem xét kĩ đến các giá trị của khoảng tin cậy 95% (thay vì tùy thuộc vào trị số *P*), và những chỉ dẫn về cách diễn giải khoảng tin cậy 95% vừa trình bày trên hi vọng sẽ giúp kết quả nghiên cứu lâm sàng có ý nghĩa hơn là những con số thống kê.

Tài liệu tham khảo

[1] HOPE/HOPE-TOO Study Investigators. Long-term effects of Ramipril on cardiovascular events and diabetes. Results of the HOPE Study Extension. *Circulation* 2005; 112:1339-1346.

[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.

Những ngộ nhận thông thường về trị số P

1. Nếu $P = 0.05$, xác suất mà giả thuyết vô hiệu đúng là 5%.
2. Một kết quả với $P > 0.05$ (không có ý nghĩa thống kê) có nghĩa là không có khác biệt giữa hai nhóm.
3. Một kết quả có ý nghĩa thống kê cũng có ý nghĩa lâm sàng.
4. Những kết quả có cùng trị số P là bằng chứng phản biện lại giả thuyết vô hiệu.
5. $P = 0.05$ có nghĩa là dữ liệu quan sát được chỉ xảy ra 5% khi giả thuyết vô hiệu đúng.
6. $P = 0.05$ và $P < 0.05$ có ý nghĩa như nhau.
7. Trị số P phải viết một cách nghiêm chỉnh là $P < 0.02$ khi $P = 0.015$.
8. $P = 0.05$ có nghĩa là nếu chúng ta bác bỏ giả thuyết vô hiệu thì xác suất sai sót loại I là 5%.
9. Với ngưỡng $P = 0.05$, xác suất sai sót loại I là 5%.
10. Một kết luận khoa học nên dựa vào trị số P có ý nghĩa thống kê hay là không.