Phương pháp ước tính cỡ mẫu cho một nghiên cứu y học

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

Một công trình nghiên cứu thường dựa vào một mẫu (sample). Một trong những câu hỏi quan trọng nhất trước khi tiến hành nghiên cứu là cần bao nhiêu mẫu hay bao nhiêu đối tượng cho nghiên cứu. "Đối tượng" ở đây là đơn vị căn bản của một nghiên cứu, là số bệnh nhân hay số tình nguyện viên. Ước tính số lượng đối tượng cần thiết cho một công trình nghiên cứu đóng vai trò cực kì quan trọng, vì nó có thể là yếu tố quyết định sự thành công hay thất bại của nghiên cứu. Nếu số lượng đối tượng không đủ thì kết luận rút ra từ công trình nghiên cứu không có độ chính xác cao, thậm chí không thể kết luận gì được. Ngược lại, nếu số lượng đối tượng quá nhiều hơn số cần thiết thì tài nguyên, tiền bạc và thời gian sẽ bị hao phí. Do đó, vấn đề then chốt trước khi nghiên cứu là phải ước tính cho được một số đối tượng vừa đủ cho mục tiêu của nghiên cứu. Số lượng đối tượng "vừa đủ" tùy thuộc vào loại hình nghiên cứu và hai thông số chính:

- Phương pháp thiết kế nghiên cứu và tiêu chí lâm sàng (outcome measure).
- Hệ số ảnh hưởng (effect size);
- Sai lầm mà nhà nghiên cứu chấp nhận, cụ thể là sai lầm loại I và II (power);

Không biết [hay chưa quyết định] được thiết kế nghiên cứu và không có số liệu về hai thông số trên thì không thể nào ước tính cỡ mẫu. Kinh nghiệm của người viết cho thấy rất nhiều người khi tiến hành nghiên cứu thường không có ý niệm gì về các số liệu này, cho nên khi đến tham vấn các chuyên gia về thống kê học, họ chỉ nhận câu trả lời: "không thể tính được"! Trong bài này tôi sẽ bàn qua hai thông số trên và trình bày một số ví dụ nghiên cứu lâm sàng cụ thể về ước tính cỡ mẫu.

1. Thiết kế nghiên cứu và tiêu chí lâm sàng

1.1 Thiết kế nghiên cứu

Thông tin thứ nhất trong qui trình ước tính cỡ mẫu là thể loại nghiên cứu, bởi vì yếu tố này có ảnh hưởng đến phương pháp phân tích thống kê và vì thế phương pháp ước tính cỡ mẫu. Có thể phân biệt các thể loại nghiên cứu này dựa vào hai tiêu chí: thời gian và đặc tính. Về thời gian, các nghiên cứu thu thập dữ liệu tại một thời điểm *hiện tại* (present) được gọi là *cross-sectional study* (nghiên cứu tiêu biểu một thời điểm); các

nghiên cứu có định hướng theo dõi tình trạng sức khỏe của đối tượng trong một thời gian, tức thu thập dữ liệu từng đối tượng nhiều lần (hiện tại và tương lai) được gọi là *prospective (*hay *longitudinal) study* (nghiên cứu theo thời gian); và các nghiên cứu được tiến hành hiện tại nhưng có định hướng tìm hiểu quá khứ (past) được gọi là *retrospective study*.

Nghiên cứu tại một thời điểm hay *cross-sectional study* (được dịch theo nghĩa đen là "nghiên cứu cắt ngang"). Đây là một thiết kế mà các nhà nghiên cứu chọn một quần thể một cách ngẫu nhiên nhưng tiêu biểu cho một cộng đồng, tại một thời điểm nào đó. Nói cách khác, nhà nghiên cứu thu thập dữ liệu chỉ một lần duy nhất của các đối tượng ngay tại thời điểm đó (hiện tại). Mục đích chính của các nghiên cứu này là tìm hiểu tỉ lệ hiện hành (prevalence) của một bệnh nào đó, hay tìm hiểu mối tương quan giữa một yếu tố nguy cơ và một bệnh.

Nghiên cứu đối chứng hay *case-control study*. Trong các nghiên cứu này, mục đích chính là tìm hiểu mối liên hệ giữa một (hay nhiều) yếu tố nguy cơ (risk factors) và một bệnh rất cụ thể. Để tiến hành nghiên cứu này, nhà nghiên cúu bắt đầu bằng một nhóm bệnh nhân và một nhóm đối tượng không bệnh (đối chứng), và "đi ngược thời gian" tìm hiểu những yếu tố nguy cơ mà cả hai nhóm phơi nhiễm trong quá khứ.

Nghiên cứu xuôi thời gian (longitudinal studies hay prospective study). Ngược lại với nghiên cứu đối chứng (trường hợp nhà nghiên cứu biết ai mắc bệnh và ai không mắc bệnh), với các nghiên cứu theo thời gian nhà nghiên cứu bắt đầu bằng một nhóm không mắc bệnh, và theo dõi một thời gian sau để quan sát ai mắc bệnh hay không mắc bệnh trong thời gian đó. Ngược lại với nghiên cứu đối chứng (trường hợp nhà nghiên cứu đi ngược về quá khứ để tìm hiểu ai bị phơi nhiễm yếu tố nguy cơ), với các nghiên cứu theo thời gian, nhà nghiên cứu biết ngay từ lúc ban đầu ai bị phơi nhiễm hay không phơi nhiễm yếu tố nguy cơ. Mục đích của các nghiên cứu xuôi thời gian thường là ước tính tỉ lệ phát sinh (incidence) bệnh trong một thời gian (điều này khác với mục đích của nghiên cứu tại một thời điểm là ước tính tỉ lệ hiện hành – tức prevalence – của bệnh). Ngoài ra, các nghiên cứu theo thời gian còn cho phép nhà nghiên cứu tìm hiểu mối liên hệ giữa một hay nhiều yếu tố nguy cơ và nguy cơ phát sinh bệnh tật. Khác với nghiên cứu cross-section chỉ ghi nhận sự kiện tại một thời điểm, các nghiên cứu longitudinal phải theo dõi đối tượng trong một thời gian có thể là nhiều năm tháng.

1.2 Tiêu chí lâm sàng

Sau khi đã xác định thể loại nghiên cứu, nhà nghiên cứu cần phải quyết định chọn một tiêu chí lâm sàng chính (primary outcome measure) để căn cứ vào đó mà ước tính cỡ mẫu. Quyết định chọn tiêu chí lâm sàng là một quyết định vừa mang tính lâm sàng, vừa

mang tính khoa học. Bởi vì mục tiêu tối hậu của nghiên cứu y khoa là đem lại lợi ích cho bệnh nhân hay cộng đồng, cho nên tiêu chí được chọn phải có ý nghĩa thực tế đối với bệnh nhân. Chẳng hạn như trong việc thẩm định hiệu quả của các phương pháp truy tìm ung thư, thì tỉ lệ phát hiện ung thư và điều trị không phải là tiêu chí có ý nghĩa thực tế, nhưng tỉ lệ tử vong và thời gian sống sót sau khi truy tìm ung thư mới là tiêu chí có ý nghĩa lâm sàng và thực tế. Mặc khác, tiêu chí phải đáp ứng các tiêu chuẩn khoa học về độ tin cậy và độ chính xác. Nếu một nghiên cứu có mục tiêu tìm hiểu hiệu quả của một loại thuốc phòng chống bệnh xơ vữa động mạch, thì độ cholesterol trong máu không thể được xem là tiêu chí có ý nghĩa lâm sàng, dù nó đáp ứng yêu cầu khoa học tính. Do đó, việc chọn một tiêu chí lâm sàng cho nghiên cứu cần phải cân nhắc rất cẩn thận.

Quyết định chọn tiêu chí lâm sàng là một quyết định quan trọng, bởi vì nó có ảnh hưởng đến cỡ mẫu rất lớn. Chẳng hạn như trong các nghiên cứu loãng xương, các nhà nghiên cứu có thể so sánh mật độ xương hay tỉ lệ gãy xương giữa hai nhóm can thiệp để biết hiệu quả của thuốc. Nếu chọn mật độ xương làm tiêu chí lâm sàng thì số lượng cỡ mẫu có thể sẽ là con số vài trăm bệnh nhân, nhưng nếu chọn tỉ lệ gãy xương con số cỡ mẫu có thể lên đến vài chục ngàn đối tượng.

2. Khái niệm về "hệ số ảnh hưởng" (effect size)

Hệ số ảnh hưởng, nói một cách đơn giản, là một chỉ số về độ ảnh hưởng của một thuật can thiệp. Vì phản ảnh mức độ khác biệt, hệ số ảnh hưởng cho phép chúng ta tránh khỏi cách diễn dịch giới hạn bởi ngôn ngữ nhị phân (như "có hay không có ảnh hưởng?"), và tập trung vào một cách diễn dịch mang tính khoa học hơn (như "mức độ ảnh hưởng cao hay thấp cỡ nào?") Ba trường hợp đơn giản sau đây sẽ minh họa cho khái niệm về hệ số ảnh hưởng:

Trường hợp 1: Trong một nghiên cứu gồm 50 bệnh nhân cao huyết áp được điều trị bằng một thuốc trong nhóm beta-blocker. Trước khi điều trị, huyết áp tâm thu (SBP) trung bình cho cả nhóm là 140 mmHg và độ lệch chuẩn là 22 mmHg. Sau khi điều trị, huyết áp tâm thu giảm xuống còn 125 mmHg.

Trường hợp 2: Một nghiên cứu khác thẩm định hiệu quả của một thuốc chống loãng xương trong nhóm bisphosphonate. Nghiên cứu được tiến hành trên 50 bệnh nhân. Trước khi điều trị, mật độ xương ở cổ xương đùi (femoral neck bone mineral density, viết tắt là BMD) trung bình là 0.68 g/cm^2 với độ lệch chuẩn 0.12 g/cm^2. Sau 6 tháng điều trị, BMD trung bình cho cả nhóm tăng lên 0.72 g/cm^2 với độ lệch chuẩn 0.13 g/cm^2.

Trường hợp 3: Một nghiên cứu bệnh – chứng (case-control study) nhằm thẩm định ảnh hưởng của thói quen hút thuốc lá đến độ glucose trong máu. Nhóm hút thuốc lá gồm 30 người có độ glucose trung bình là 130 mg/dL với độ lệch chuẩn 35 mg/dL. Nhóm không hút thuốc lá gồm 70 người có độ glucose trung bình là 110 mg/dL với độ lệch chuẩn 50 mg/dL.

Trong trường hợp 1, chúng ta có thể ước tính mức độ ảnh hưởng bằng cách lấy huyết áp sau khi điều trị trừ cho huyết áp trước khi điều trị: $d_1 = 125 – 140 = -15$ mmHg. Tương tự, ảnh hưởng của thuốc bisphosphonate cho trường hợp 2 là $d_2 = 0.72 – 0.68 = 0.04$ g/cm^2. Và trường hợp 3, độ ảnh hưởng của hút thuốc lá có thể ước tính bằng d3 = 130 – 110 = 20 mg/dL.

Khó khăn trong cách ước tính độ ảnh hưởng trên đây là không thể so sánh trực tiếp được độ ảnh hưởng, bởi vì đơn vị đo lường khác nhau. Và, quan trọng hơn nữa, độ dao động (phản ảnh bằng độ lệch chuẩn) giữa 3 trường hợp cũng rất khác nhau. Phương pháp so sánh trực tiếp ảnh hưởng lí tưởng là hoán chuyển sao cho cả ba trường hợp có cùng một đơn vị đo lường. Để đạt được mục đích này, cách đơn giản nhất là lấy độ ảnh hưởng chia cho độ lệch chuẩn. Tỉ số này có tên tiếng Anh là *effect size* (có khi còn gọi là *standardized difference*) mà tôi tạm dịch là *hệ số ảnh hưởng*. Công thức chung cho ước tính hệ số ảnh hưởng (sẽ viết tắt bằng ES) là:

$$ES = \frac{\overline{x}_1 - \overline{x}_0}{s_0} \qquad [1]$$

Trong đó:

- $\overline{x}_1$ là số trung bình của nhóm can thiệp;
- $\overline{x}_0$ là số trung bình của nhóm đối chứng; và
- s_0 là độ lệch chuẩn của nhóm đối chứng.

Hệ số ảnh hưởng của 3 trường hợp trên là:

- Trường hợp 1: $ES_1 = -15 / 22 = 0.68$
- Trường hợp 2: $ES_2 = 0.04 / 0.12 = 0.33$
- Trường hợp 3: $ES_3 = 20 / 50 = 0.40$

Nên nhớ rằng độ lệch chuẩn có cùng đơn vị đo lường với độ ảnh hưởng trung bình, cho nên hệ số ảnh hưởng không có đơn vị. Nói cách khác, đơn vị đo lường độ ảnh hưởng bây giờ là độ lệch chuẩn. Chẳng hạn như trong trường hợp 1, thuốc beta-blocker có tác dụng giảm huyết áp tâm thu khoảng 0.68 độ lệch chuẩn, còn trong trường hợp 2, thuốc bisphosphonate tăng mật độ xương chỉ 0.33 độ lệch chuẩn. Vì có cùng đơn vị so sánh, có

thể nói [đơn giản] rằng hệ số ảnh hưởng của thuốc beta-blocker cao hơn thuốc bisphosphonate.

Theo một qui ước [không có cơ sở khoa học mấy], một hệ số ảnh hưởng bằng 0.2 được xem là "thấp", 0.5 là "trung bình", và >0.8 là "cao" [1]. Một hệ số ảnh hưởng 0.2 tương đương với độ khác biệt về chiều cao của một em bé 15 tuổi và một em bé 16 tuổi. Một hệ số ảnh hưởng 0.5 tương đương với độ khác biệt về chiều cao của một em bé 14 tuổi và một em bé 18 tuổi. Một hệ số ảnh hưởng 0.8 tương đương với độ khác biệt về chỉ số thông minh (IQ) của một sinh viên năm thứ nhất và một tiến sĩ.

3. Sai lầm loại I, II và khái niệm về "power"

3.1 Sai lầm loại I và II

Thống kê học là một phương pháp khoa học có mục đích phát hiện, hay đi tìm những cái có thể gộp chung lại bằng cụm từ "chưa được biết" (unknown). Cái chưa được biết ở đây là những hiện tượng chúng ta không quan sát được, hay quan sát được nhưng không đầy đủ. Cái chưa biết có thể là một ẩn số (như chiều cao trung bình ở người Việt Nam, hay trọng lượng một phần tử), hiệu quả của một thuật điều trị, tỉ lệ lưu hành (prevalence), tỉ lệ phát sinh (incidence) của bệnh, v.v… Chúng ta có thể đo chiều cao, hay tiến hành xét nghiệm để biết hiệu quả của thuốc, nhưng các nghiên cứu như thế chỉ được tiến hành trên một nhóm đối tượng, chứ không phải toàn bộ quần thể của dân số. Vấn đề là sử dụng kết quả của một nhóm đối tượng để suy luận cho một quần thể lớn hơn. Mục đích của ước tính cỡ mẫu là tìm số lượng đối tượng sao cho suy luận đó đạt độ chính xác cao nhất và đầy đủ nhất.

Ở mức độ đơn giản nhất, những cái chưa biết này có thể xuất hiện dưới hai hình thức: hoặc là có, hoặc là không. Chẳng hạn như một thuật điều trị có hay không có hiệu quả chống gãy xương. Bởi vì không ai biết hiện tượng một cách đầy đủ, chúng ta phải đặt ra giả thiết. Giả thiết đơn giản nhất là *giả thiết đảo* (hiện tượng không tồn tại, kí hiệu Ho) và *giả thiết chính* (hiện tượng tồn tại, kí hiệu Ha).

Chúng ta sử dụng các phương pháp kiểm định thống kê (statistical test) như kiểm định t, F, z, χ^2, v.v… để đánh giá khả năng của giả thiết. Kết quả của một kiểm định thống kê có thể đơn giản chia thành hai giá trị: hoặc là *có ý nghĩa thống kê* (statistical significance), hoặc là *không có ý nghĩa thống kê* (non-significance). Có ý nghĩa thống kê ở đây thường dựa vào trị số P: thông thường, nếu $P < 0.05$, chúng ta phát biểu kết quả có ý nghĩa thống kê; nếu $P > 0.05$ chúng ta nói kết quả không có ý nghĩa thống kê. Cũng có thể xem có ý nghĩa thống kê hay không có ý nghĩa thống kê như là có tín hiệu hay không

có tín hiệu. Hãy tạm đặt kí hiệu T+ là kết quả có ý nghĩa thống kê, và T- là kết quả kiểm định không có ý nghĩa thống kê.

Hãy xem xét một ví dụ cụ thể: để biết thuốc risedronate có hiệu quả hay không trong việc điều trị loãng xương, chúng ta tiến hành một nghiên cứu gồm 2 nhóm bệnh nhân (một nhóm được điều trị bằng risedronate và một nhóm chỉ sử dụng giả dược placebo). Chúng ta theo dõi và thu thập số liệu gãy xương, ước tính tỉ lệ gãy xương cho từng nhóm, và so sánh hai tỉ lệ bằng một kiểm định thống kê. Kết quả kiểm định thống kê hoặc là *có ý nghĩa thống kê* ($P<0.05$) hay không có ý nghĩa thống kê ($P>0.05$). Xin nhắc lại rằng chúng ta không biết risedronate thật sự có hiệu nghiệm chống gãy xương hay không; chúng ta chỉ có thể đặt giả thiết. Do đó, khi xem xét một giả thiết và kết quả kiểm định thống kê, chúng ta có bốn tình huống:

(a) Giả thuyết Ha đúng (thuốc risedronate có hiệu nghiệm) và kết quả kiểm định thống kê $P<0.05$.

(b) Giả thuyết Ha đúng, nhưng kết quả kiểm định thống kê không có ý nghĩa thống kê;

(c) Giả thuyết Ha sai (thuốc risedronate không có hiệu nghiệm) nhưng kết quả kiểm định thống kê có ý nghĩa thống kê;

(d) Giả thuyết Ha sai và kết quả kiểm định thống kê không có ý nghĩa thống kê.

Ở đây, trường hợp (a) và (d) không có vấn đề, vì kết quả kiểm định thống kê nhất quán với thực tế của hiện tượng. Nhưng trong trường hợp (b) và (c), chúng ta phạm sai lầm, vì kết quả kiểm định thống kê không phù hợp với giả thiết. Trong ngôn ngữ thống kê học, chúng ta có vài thuật ngữ:

- xác suất của tình huống (b) xảy ra được gọi là *sai lầm loại II* (type II error), và thường kí hiệu bằng β.

- xác suất của tình huống (a) được gọi là *Power*. Nói cách khác, *power* chính là xác suất mà kết quả kiểm định thống cho ra kết quả $p<0.05$ với điều kiện giả thiết Ha là thật. Nói cách khác: $power = 1-\beta$;

- xác suất của tình huống (c) được gọi là *sai lầm loại I* (type I error, hay significance level), và thường kí hiệu bằng α. Nói cách khác, α chính là xác suất mà kết quả kiểm định thống cho ra kết quả $p<0.05$ với điều kiện giả thiết Ha sai;

- xác suất tình huống (d) không phải là vấn đề cần quan tâm, nên không có thuật ngữ, dù có thể gọi đó là kết quả *âm tính thật* (hay true negative).

Có thể tóm lược 4 tình huống đó trong một Bảng 1 sau đây:

Bảng 1. Các tình huống trong việc thử nghiệm một giả thiết khoa học

Kết quả kiểm định thống kê	**Giả thuyết Ha**	
	Đúng (thuốc có hiệu nghiệm)	**Sai** (thuốc không có hiệu nghiệm)
Có ý nghĩa thống kê (p<0,05)	Dương tính thật (**power**), 1-β= P(S | Ha)	Sai lầm loại I (**type I error**) α = P(S | Ho)
Không có ý nghĩa thống kê (p>0,05)	Sai lầm loại II (**type II error**) β = P(NS | Ha)	Âm tính thật (**true negative**) 1-α = P(NS | Ho)

Chú thích: kí hiệu *S* trong bảng này có nghĩa là "significant" (tức p<0.05); *NS* = "non-significant" (tức p>0.05). Do đó, có thể mô tả 4 tình huống trên bằng ngôn ngữ xác suất có điều kiện như sau: Power $= 1 - \beta =$ P(*S* | Ha); β = P(*NS* | Ha); và α = P(*S* | Ho). Xin nhắc lại rằng kí hiệu toán học "P(A | B)" có nghĩa là một xác suất có điều kiện, cụ thể hơn kí hiệu P(*S* | Ha) có nghĩa là "xác suất *S* xảy ra nếu (hay với điều kiện) Ha là đúng."

3.2 Kiểm định giả thiết thống kê và chẩn đoán y khoa

Có lẽ những lí giải trên đây, đối với một số bạn đọc, vẫn còn khá trừu tượng. Một cách để minh họa các khái niệm *power* và trị số P là qua chẩn đoán y khoa. Thật vậy, có thể ví nghiên cứu khoa học và suy luận khoa học như là một qui trình chẩn đoán bệnh. Trong chẩn đoán, thoạt đầu chúng ta không biết bệnh nhân mắc bệnh hay không, và phải thu thập thông tin (như tìm hiểu tiền sử bệnh, cách sống, thói quen, v.v…) và làm xét nghiệm (như quang tuyến X, như siêu âm, phân tích máu, nước tiểu, v.v…) để đi đến kết luận.

Có hai giả thiết: bệnh nhân không có bệnh (kí hiệu Ho) và bệnh nhân mắc bệnh (Ha). Ở mức độ đơn giản nhất, kết quả xét nghiệm có thể là *dương tính* (+ve) hay *âm tính* (-ve). Trong chẩn đoán cũng có 4 tình huống và tôi sẽ bàn trong phần dưới đây, nhưng để vấn đề rõ ràng hơn, chúng ta hãy xem qua một ví dụ cụ thể như sau:

Trong chẩn đoán ung thư, để biết chắc chắn có ung thư hay không, phương pháp chuẩn là dùng sinh thiết (tức giải phẫu để xem xét mô dưới ống kính hiển vi để xác định xem *có ung thư* hay *không có ung thư*. Nhưng sinh thiết là một phẫu thuật có tính cách xâm phạm vào cơ thể bệnh nhân, nên không thể áp dụng phẫu thuật này một cách đại trà cho mọi người. Thay vào đó, y khoa phát triển những phương pháp xét nghiệm không mang tính xâm phạm để thử nghiệm ung thư. Các phương pháp này bao gồm quang tuyến X hay thử máu. Kết quả của một xét nghiệm bằng quang tuyến X hay thử máu có thể tóm tắt bằng hai giá trị: hoặc là *dương tính* (+ve), hoặc là *âm tính* (-ve).

Nhưng không có một phương pháp thử nghiệm gián tiếp nào, dù tinh vi đến đâu đi nữa, là hoàn hảo và chính xác tuyệt đối. Một số người có kết quả dương tính, nhưng thực sự không có ung thư. Và một số người có kết quả âm tính, nhưng trong thực tế lại có ung thư. Đến đây thì chúng ta có bốn khả năng:

- Bệnh nhân có ung thư, và kết quả thử nghiệm là dương tính. Đây là trường hợp *dương tính thật* (danh từ chuyên môn là *độ nhạy,* tiếng Anh gọi là *sensitivity*);

- bệnh nhân không có ung thư, nhưng kết quả thử nghiệm là dương tính. Đây là trường hợp *dương tính giả* (*false positive*);

- bệnh nhân không có ung thư, nhưng kết quả thử nghiệm là âm tính. Đây là trường hợp của *âm tính thật* (*specificity*); và,

- bệnh nhân có ung thư, và kết quả thử nghiệm là âm tính. Đây là trường hợp *âm tính giả* hay *độ đặc hiệu* (*false negative*).

Có thể tóm lược 4 tình huống đó trong Bảng 2 sau đây:

Bảng 2. Các tình huống trong việc chẩn đoán y khoa: kết quả xét nghiệm và bệnh trạng

Kết quả xét nghiệm	**Bệnh trạng**	
	Có bệnh	**Không có bệnh**
+ve (dương tính)	Độ nhạy hay dương tính thật (**sensitivity**),	Dương tính giả (**false positive)**
-ve (âm tính)	Âm tính giả (**false negative**),	Độ đặc hiệu hay âm tính thật (**Specificity**),

Đến đây, chúng ta có thể nhận ra mối tương quan song song giữa chẩn đoán y khoa và kiểm định một giả thiết khoa học. Trong chẩn đoán y khoa có chỉ số dương tính thật, tương đương với khái niệm "power" trong nghiên cứu khoa học. Trong chẩn đoán y khoa có xác suất dương tính giả, và xác suất này chính là trị số p trong suy luận khoa học. Bảng sau đây sẽ cho thấy mối tương quan đó:

Bảng 3. Tương quan giữa chẩn đoán y khoa và suy luận trong khoa học

Chẩn đoán y khoa	**Kiểm định giả thiết khoa học**
Chẩn đoán bệnh	Thử nghiệm một giả thiết khoa học
Bệnh trạng (có hay không)	Giả thiết khoa học (Ha hay Ho)
Phương pháp xét nghiệm	Kiểm định thống kê
Kết quả xét nghiệm +ve	Trị số $p < 0.05$ hay "có ý nghĩa thống kê"
Kết quả xét nghiệm –ve	Trị số $p > 0.05$ hay "không có ý nghĩa thống kê"
Dương tính thật (sensitivity)	Power; $1-\beta$; P(s | Ha)
Dương tính giả (false positive)	Sai lầm loại I; trị số p; α; P(S | Ho)
Âm tính giả (false negative)	Sai lầm loại II; β; β = P(NS | Ha)
Âm tính thật (đặc hiệu, hay specificity)	Âm tính thật; $1-\alpha$ = P(NS | Ho)

Cũng như các phương pháp xét nghiệm y khoa không bao giờ hoàn hảo, các phương pháp kiểm định thống kê cũng có sai sót. Và do đó, kết quả nghiên cứu lúc nào cũng có độ bất định (như sự bất định trong một chẩn đoán y khoa vậy). Vấn đề là chúng ta phải thiết kế nghiên cứu sao cho *sai sót* loại I và II thấp nhất.

4. Phương pháp ước tính cỡ mẫu

Như đã đề cập trong phần đầu của bài viết, để ước tính số đối tượng cần thiết cho một công trình nghiên cứu, ngoài thể loại nghiên cứu, chúng ta cần phải có 3 số liệu: xác suất sai sót loại I và power, và hệ số ảnh hưởng. Số lượng cỡ mẫu là hàm số của ba thông số này. Gọi n là số lượng cỡ mẫu cần thiết, α là sai sót loại I, β là sai sót loại II (tức $1-\beta$ là power), hệ số ảnh hưởng là *ES*, thì công thức chung để ước tính cỡ mẫu là:

$$n = \frac{\left(z_{\alpha/2} + z_b\right)^2}{\left(ES\right)^2}$$

Trong đó, $z_{\alpha/2}$ và z_{β} là những hằng số (thật ra là số độ lệch chuẩn) từ phân phối chuẩn (standardized normal distribution) cho xác suất sai sót α và β. Bởi vì, trong công thức trên *ES* là mẫu số, cho nên nếu *ES* thấp thì số lượng cỡ mẫu sẽ tăng; ngược lại, nếu *ES* cao thì số lượng cỡ mẫu sẽ giảm.

Vì ảnh hưởng như thế, hệ số ảnh hưởng phải được giả định trước khi tính toán. Đây là thông số không phải lúc nào cũng có sẵn, cho nên nhà nghiên cứu cần phải xem xét các nghiên cứu trước hay độ ảnh hưởng có ý nghĩa lâm sàng để tính toán cỡ mẫu.

Về xác suất sai sót, thông thường một nghiên cứu chấp nhận sai sót loại I khoảng 1% hay 5% (tứcα = 0.01 hay 0.05), và xác suất sai sót loại II khoảng β = 0.1 đến β = 0.2 (tức power phải từ 0.8 đến 0.9). Mỗi trường hợp gắn liền với một hằng số $z_{\alpha/2}$ và z_{β} như vừa đề cập. Hai hằng số này có thể tóm gọn bằng công thức $C = \left(z_{\alpha/2} + z_b\right)^2$. *C* được xác định bởi luật phân phối chuẩn như trình bày trong **Bảng 3** dưới đây. Chẳng hạn như nếu muốn α = 0.05 và power = 0.80, thì hằng số *C* là 7.85.

Bảng 3: Hằng số C liên quan đến sai sót loại I và II

α =	β = 0.20 (Power = 0.80)	β = 0.10 (Power = 0.90)	β = 0.05 (Power = 0.95)
0.10	6.15	8.53	10.79
0.05	7.85	10.51	13.00
0.01	13.33	16.74	19.84

4.1 Các nghiên cứu với tiêu chí là biến liên tục (continuous variable)

4.1.1 Trong trường hợp nghiên cứu chỉ có một nhóm đối tượng, và mục tiêu là ước tính một chỉ số trung bình (kí hiệu μ) với một sai số định trước là σ. Với nghiên cứu như thế, hệ số ảnh hưởng có thể ước tính bằng ES = μ /σ. Và số đối tượng (*n*) cần thiết cho nghiên cứu có thể tính toán theo công thức sau đây:

$$n = \frac{C}{\left(ES\right)^2} \qquad [2]$$

Trong đó, C là hằng số từ Bảng 3.

4.1.2 Trong trường hợp nghiên cứu "trước-sau" (before-after studies). Nhiều nghiên cứu can thiệp trên một nhóm bệnh nhân, mà theo đó tiêu chí lâm sàng ở mỗi bệnh nhân được đo lường hai lần: trước khi can thiệp và sau khi can thiệp. Trong thuật ngữ dịch tễ học, người ta gọi là nghiên cứu trước-sau (before-after study). Chẳng hạn như để đánh giá hiệu quả của một loại thuốc điều trị cao huyết áp, các nhà nghiên cứu có thể chọn một nhóm bệnh nhân thích hợp, sau đó đo lường huyết áp trước khi điều trị và sau khi điều trị. Hệ số ảnh hưởng có thể tính từ khác biệt giữa hai thời điểm, nhưng ở đây còn một thông số liên quan khác: đó là hệ số tương quan giữa hai lần đo lường.

Gọi đo lường trước khi điều trị của bệnh nhân *i* là X_i và sau khi điều trị là Y_i. Ảnh hưởng của thuật điều trị có thể ước tính cho mỗi bệnh nhân *i* bằng $\Delta_i = Y_i - X_i$. Từ đó, chúng ta có thể tính độ ảnh hưởng trung bình và độ lệch chuẩn của Δ_i. Trong thực tế, chúng ta không biết Δ_i, cho nên phải dựa vào một mẫu. Nếu gọi ước số mẫu của Δ_i là d_i, chúng ta có thể ước tính độ ảnh hưởng trung bình và độ lệch chuẩn của d_i. Gọi chỉ số trung bình đó là $\overline{d}$ và độ lệch chuẩn là *s*. Hệ số ảnh hưởng có thể ước tính bằng công thức:

$$ES = \frac{\overline{d}}{s}$$

Ngoài ra, gọi *r* là hệ số tương quan giữa hai đo lường. Với các thông số này, số lượng cỡ mẫu cần thiết cho nghiên cứu là:

$$n = \frac{2C(1-r)}{(ES)^2} \qquad [3]$$

4.1.3 Trong trường hợp nghiên cứu với hai nhóm đối tượng, mục tiêu thường là so sánh hai chỉ số trung bình. Gọi chỉ số trung bình của nhóm 1 và 2 là μ_1 và μ_2. Gọi độ lệch chuẩn của hai nhóm là σ_1 và σ_2. Nếu hai độ lệch chuẩn không khác nhau, hệ số ảnh hưởng có thể ước tính từ công thức [1] như sau:

$$ES = \frac{\mu_1 - \mu_2}{\sigma_1}$$

Số lượng đối tượng cho **mỗi nhóm** (*n*) cần thiết cho nghiên cứu có thể tính toán như sau (giá trị của hằng số *C* được xác định từ xác suất sai sót loại I và II (hay power) trong Bảng 3):

$$n = \frac{2C}{(ES)^2} \qquad [4]$$

4.1.4 Trong trường hợp nghiên cứu với hai nhóm đối tượng nhưng mục tiêu kiểm định độ ảnh hưởng tương đương (equivalence studies). Trong nhiều nghiên cứu, chúng ta muốn đánh giá xem hai thuật can thiệp hay điều trị có hiệu quả như nhau. Gọi chỉ số trung bình của nhóm 1 và 2 là μ_1 và μ_2. Nếu $|\mu_1 - \mu_2| < d$ (trong đó d là độ khác biệt không có ý nghĩa lâm sàng), thì chúng ta tuyên bố rằng hai thuật điều trị có ảnh hưởng tương đương. Trong trường này, hệ số ảnh hưởng sẽ là (tôi sẽ dùng kí hiệu *H* thay vì ES để không nhầm lẫn với công thức [1]):

$$H = \frac{|\mu_1 - \mu_2| - d}{\sigma}$$

Và số lượng cỡ mẫu cần thiết cho **mỗi nhóm** là:

$$n = \frac{2C}{H^2} \qquad [5]$$

4.2 Các nghiên cứu với tiêu chí là biến nhị phân (binomial variable)

Trong phần trước chúng ta đã làm quen với phương pháp ước tính cỡ mẫu để so sánh hai số trung bình bằng kiểm định t. Nhưng có nghiên cứu biến số không liên tục mà mang tính nhị phân (như có / không, sống / chết, dứt bệnh / không dứt bệnh, v.v…), chỉ số tóm lược [dĩ nhiên] không thể là số trung bình, mà là tỉ lệ (proportion). Nhiều nghiên cứu mô tả có mục đích khá đơn giản là ước tính một tỉ lệ. Chẳng hạn như giới y tế thường hay tìm hiểu tỉ lệ lưu hành bệnh trong một cộng đồng. Trong trường hợp này, chúng ta không có những đo lường mang tính liên tục, nhưng kết quả chỉ là những giá trị nhị phân như có / không. Phương pháp ước tính cỡ mẫu cũng khác với các phương pháp cho các nghiên cứu với biến số liên tục.

Năm 1991, một cuộc thăm dò ý kiến ở Mĩ cho thấy 45% người được hỏi sẵn sàng khuyến khích con họ nên hiến một quả thận cho những bệnh nhân cần thiết. Khoảng tin cậy 95% của tỉ lệ này là 42% đến 48%, tức một khoảng cách đến 6%! Kết quả này [tương đối] thiếu chính xác, dù số lượng đối tượng tham gia lên đến 1000 người. Tại sao? Để trả lời câu hỏi này, chúng ta thử xem qua một vài lí thuyết về ước tính cỡ mẫu cho một tỉ lệ.

4.2.1 Trong trường hợp nghiên cứu chỉ có một nhóm đối tượng, và mục tiêu là ước tính một tỉ lệ (kí hiệu π) về một biến cố lâm sàng. Qua lí thuyết xác suất, chúng ta biết rằng nếu trong *n* đối tượng, có *k* biến cố thì ước số của π là $\hat{p} = x / n$, với sai số

chuẩn $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$. Khoảng tin cậy 95% của một tỉ lệ π [trong quần thể] là: $\hat{p} \pm 1.96 \times SE(\hat{p})$.

Bây giờ, thử lật ngược vấn đề: chúng ta muốn ước tính π sao khoảng tin cậy $2 \times 1.96 \times SE(\hat{p})$ không quá một hằng số *m*. Nói cách khác, chúng ta muốn:

$$1.96 \times \sqrt{\hat{p}(1-\hat{p})/n} \leq m$$

Chúng ta muốn tìm số lượng đối tượng *n* để đạt yêu câu trên. Qua cách diễn đạt trên, dễ dàng thấy rằng:

$$n \geq \left(\frac{1.96}{m}\right)^2 \hat{p}(1-\hat{p}) \qquad [6]$$

Do đó, số lượng cỡ mẫu tùy thuộc vào độ sai số *m* và tỉ lệ *p* mà chúng ta muốn ước tính. Độ sai số càng thấp, số lượng cỡ mẫu càng cao.

4.2.2 Trong trường hợp nghiên cứu có hai nhóm đối tượng, và mục tiêu nghiên cứu là so sánh hai tỉ lệ. Để so sánh hai tỉ lệ, phương pháp kiểm định thông dụng nhất là kiểm định nhị phân (binomial test) hay Chi bình phương (χ^2 test). Gọi hai tỉ lệ [mà chúng ta không biết nhưng muốn tìm hiểu] là π_1 và π_2, và gọi $\Delta = \pi_1 - \pi_2$. Giả thiết mà chúng ta muốn kiểm định là $\Delta = 0$.

Nhưng trong thực tế, chúng ta không biết π_1 và π_2, mà chỉ ước tính qua hai tỉ lệ p_1 và p_2. Lí thuyết đằng sau để ước tính cỡ mẫu cho kiểm định giả thiết này khá rườm rà, nhưng có thể tóm gọn bằng công thức sau đây:

$$n = \frac{\left(z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right)^2}{\Delta^2} \qquad [7]$$

Trong đó, $\bar{p} = (p_1 + p_2)/2$, $z_{\alpha/2}$ là trị số *z* của phân phối chuẩn cho xác suất $\alpha/2$ (chẳng hạn như khi $\alpha = 0.05$, thì $z_{\alpha/2} = 1.96$; khi $\alpha = 0.01$, thì $z_{\alpha/2} = 2.57$), và z_{β} là trị số *z* của phân phối chuẩn cho xác suất β (chẳng hạn như khi $\beta = 0.10$, thì $z_{\beta} = 1.282$; khi $\beta = 0.20$ hay power = 0.80, thì $z_{\beta} = 0.842$).

4.2.3 Trong trường hợp nghiên cứu có hai nhóm đối tượng với mục tiêu nghiên cứu là nhằm "chứng minh" hai tỉ lệ tương đương nhau. Với các nghiên cứu thể loại này, giả thiết đặt ra là nếu độ khác biệt giữa p_1 và p_2 thấp hơn d thì có thể chấp nhận rằng π_1 và π_2 tương đương nhau; nếu $|p_1 - p_2| > d$, thì hai tỉ lệ không tương đương nhau. Để kiểm định giả thiết tương đương này, số lượng cỡ mẫu cần thiết cho **mỗi nhóm** là:

$$n = \frac{2C\left[p_1(1-p_1) + p_2(1-p_2)\right]}{\left(|p_1 - p_2| - d\right)^2} \qquad [8]$$

4.3 Các nghiên cứu ước tính hệ số tương quan (coefficient of correlation)

4.3.1 Trường hợp chỉ có hai biến liên tục. Trong các nghiên cứu quan sát (observational studies), kể cả các nghiên cứu một thời điểm (cross-sectional studies), đôi khi mục tiêu chính là ước tính một hệ số tương quan giữa hai biến liên tục (chẳng hạn như hệ số tương quan giữa độ tuổi và nồng độ cholesterol). Gọi hệ số tương quan giữa hai biến là ρ, giả thiết đặt ra là: $H_o: \rho = 0$ hoặc $H_1: \rho \neq 0$. (Nếu $\rho = 0$, hai biến hoàn toàn độc lập với nhau, tức không có mối liên hệ).

Trong thực tế, chúng ta không biết ρ, nhưng có thể ước tính qua hệ số tương quan quan sát được là r, có khi còn gọi là hệ số Pearson. Giả thiết có thể kiểm định bằng chỉ số thống kê t như sau:

$$t = \frac{1}{2}\log_e\left[\frac{1+r}{1-r}\right]\sqrt{n-3}$$

Trong đó n là số cỡ mẫu. Chỉ số t phân phối theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Do đó, vấn đề là tìm n sao cho t có ý nghĩa thống kê, và đáp số của n là:

$$n = \frac{C}{\frac{1}{4}\left[\log_e\left(\frac{1+\rho}{1-\rho}\right)\right]^2} + 3 \qquad [9]$$

4.3.2 Trường hợp nghiên cứu có nhiều biến liên tục. Với những nghiên cứu có một biến phụ thuộc (dependent variable) và nhiều biến độc lập (independent

variables), mục tiêu thường là xác định các biến độc lập có thể "giải thích" bao nhiêu phần trăm phương sai của biến phụ thuộc. Phương pháp phân tích chính là mô hình hồi qui tuyến tính đa biến (multiple linear regression). Trong mô hình này, chỉ số phản ảnh mối liên hệ đa chiều này là *hệ số xác định bội* (coefficient of determination), kí hiệu R^2.

Phương pháp ước tính cỡ mẫu cho hệ số R^2 tương đối phức tạp, và thường phải sử dụng đến thuật mô phỏng (simulation). Tuy nhiên, một số qui ước khá tin cậy có thể áp dụng như sau:

- Với power = 0.80 và $\alpha = 0.05$, nghiên cứu cần tối thiểu 50 đối tượng để ước tính $R^2 \geq 0.23$; hay tối thiểu 100 để ước tính $R^2 \geq 0.12$ [2].

- Với *m* biến độc lập và 1 biến phụ thuộc, số lượng cỡ mẫu cần thiết tối thiểu là $n > 104 + m$ [3].

- Với $m \leq 5$, số lượng cỡ mẫu cần thiết tối thiểu là $n > 50 + m$ [4].

4.4 Các nghiên cứu ước tính tỉ số nguy cơ (odds ratio)

Trong các nghiên cứu đối chứng (case-control study), nhà nghiên cứu thường muốn tìm hiểu mối liên hệ giữa một yếu tố nguy cơ (risk factor) và một bệnh cụ thể. Mối liên hệ này thường được "đo lường" bằng odds ratio (OR) mà tôi tạm dịch là "tỉ số nguy cơ" (chứ không phải "tỉ số chênh" mà đồng nghiệp trong nước hay sử dụng). Chẳng hạn như nếu tỉ số nguy cơ giữa hút thuốc lá và gãy xương là 2, thì điều này có nghĩa là những người hút thuốc lá có nguy cơ bị gãy xương tăng khoảng 2 lần so với những người không hút thuốc lá.

Do đó, ước tính cỡ mẫu để thử nghiệm một giả thiết về mối liên hệ giữa một yếu tố nguy cơ và bệnh thường dựa vào tỉ số nguy cơ. Để ước tính cở mẫu cho các nghiên cứu như thế, nhà nghiên cứu cần phải có trong tay 3 số liệu:

- Tỉ lệ lưu hành (prevalence) của yếu tố nguy cơ trong một quần thể (gọi tắt là *p*);
- Tỉ số nguy cơ mà nhà nghiên cứu muốn biết; và
- Các sai số thống kê thể hiện qua xác suất $\boldsymbol{\alpha}$ và power.

Với các số liệu trên, công thức sau đây sẽ cung cấp cho nhà nghiên cứu một ước tính số lượng đối tượng cần thiết cho nghiên cứu (*N*):

$$N = \frac{(1+r)^2 C}{r(\ln OR)^2 p(1-p)} \qquad [10]$$

Trong đó, *r* là tỉ số cỡ mẫu giữa hai nhóm (vì trong các nghiên cứu đối chứng, không nhất thiết hai nhóm phải có cùng cỡ mẫu). Nếu *r* = 1 (tức hai nhóm có cùng số lượng cỡ mẫu), thì công thức trên sẽ đơn giản thành:

$$N = \frac{4C}{(\ln OR)^2 p(1-p)} \qquad [11]$$

4.5 Các nghiên cứu với biến phụ thuộc là thời gian dẫn đến sự kiện (survival studies)

Trong nhiều nghiên cứu khoa học, kể cả nghiên cứu lâm sàng, các nhà nghiên cứu thường theo dõi đối tượng trong một thời gian, có khi lên đến vài mươi năm. Biến cố xảy ra trong thời gian đó như có bệnh hay không có bệnh, sống hay chết, v.v… là những biến cố có ý nghĩa lâm sàng nhất định, nhưng thời gian dẫn đến bệnh nhân mắc bệnh hay chết còn quan trọng hơn cho việc đánh giá ảnh hưởng của một thuật điều trị hay một yếu tố nguy cơ. Nhưng thời gian này khác nhau giữa các bệnh nhân. Chẳng hạn như thời điểm từ lúc điều trị ung thư đến thời điểm bệnh nhân chết rất khác nhau giữa các bệnh nhân, và do đó tiêu chí lâm sàng thường là thời gian sống sót của bệnh nhân tính từ khi được điều trị (hay từ khi được chẩn đoán bệnh).

Nghiên cứu tiêu biểu thường có 2 nhóm bệnh nhân: một nhóm đối chứng và một nhóm can thiệp. Phương pháp tính cỡ mẫu cho các nghiên cứu thể loại này khá phức tạp, nhưng một cách tính đơn giản cũng có thể ứng dụng. Nếu thời gian theo dõi đối tượng được định trước, và tỉ lệ phát sinh của hai nhóm trong thời gian đó là p_1 và p_2, thì tỉ số nguy cơ (hazards ratio) có thể ước tính như sau [5,6]:

$$h = \frac{\log_e(p_1)}{\log_e(p_2)}$$

Và số cỡ mẫu cần thiết cho từng nhóm là:

$$n = \frac{C(h+1)^2}{(2-p_1-p_2)(h-1)^2} \qquad [12]$$

4.5 Các nghiên cứu về chẩn đoán (diagnostic studies)

Nghiên cứu về chẩn đoán thường xoay quanh hai chỉ số: độ nhạy (sensivity) và đặc hiệu (specificity) như trình bày trong Bảng 2. Một phương pháp chẩn đoán được xem là đáng tin cậy và có thể sử dụng trong thực hành lâm sàng cần phải đạt độ nhạy và đặc hiệu tối thiểu 0.75 (hay tốt hơn nữa là 0.80). Việc phát hiện bệnh qua chẩn đoán còn tùy thuộc vào tỉ lệ lưu hành (prevalence) của bệnh trong một quần thể. Do đó, phương pháp ước tính cỡ mẫu phải dựa vào các chỉ số này. Một cách cụ thể, nhà nghiên cứu cần phải xác định các số liệu sau đây:

- Xác suất dương tính thật (hay độ nhạy – kí hiệu p_{se}) tối thiểu là bao nhiêu?
- Xác suất âm tính thật (hay độ đặc hiệu – kí hiệu p_{sp}) tối thiểu là bao nhiêu?
- Sai số của hai xác suất dương tính thật và âm tính thật là bao nhiêu (kí hiệu *w*)?
- Tỉ lệ lưu hành của bệnh trong quần thể là bao nhiêu (kí hiệu p_{dis})

Với các thông số này, số lượng cỡ mẫu để ước tính độ nhạy có thể ước tính bằng công thức sau đây [7]:

- Trước hết, ước tính TP+FN (tức là số dương tính thật – true positive và âm tính giả - false negative)

$$TP + FN = \frac{Z_\alpha^2 \times p_{se} \times (1 - p_{se})}{w^2}$$

- Trong đó, Z_α^2 là hằng số của phân phối chuẩn. Nếu $\alpha = 0.05$, hằng số Z_α^2 bằng 1.96. Sau đó, ước tính số lượng cỡ mẫu (tôi sẽ dùng kí hiệu n_{se} để chỉ rõ đây là số cỡ mẫu cho độ nhạy):

$$n_{se} = \frac{TP + FN}{p_{dis}} \qquad [13]$$

Tương tự, số lượng cỡ mẫu để ước tính độ đặc hiệu có thể ước tính qua hai bước như sau:

- Trước hết, ước tính FP+TN (tức là số dương tính giả - false positive và âm tính thật – true negative)

$$FP + TN = \frac{Z_\alpha^2 \times p_{sp} \times (1 - p_{sp})}{w^2}$$

- Sau đó, ước tính số lượng cỡ mẫu (tôi sẽ dùng kí hiệu n_{sp} để chỉ rõ đây là số cỡ mẫu cho độ nhạy):

$$n_{sp} = \frac{FP + TN}{1 - p_{dis}} \qquad [14]$$

5. Ví dụ

Trong phần này, tôi sẽ nêu nhiều ví dụ về ước tính cỡ mẫu để minh họa cho phần "lí thuyết" vừa trình bày trong phần trên. Tôi sẽ tập trung các ví dụ liên quan đến nghiên cứu lâm sàng để bạn đọc tạp chí dễ theo dõi.

5.1 Ước tính cỡ mẫu cho một chỉ số trung bình

Ví dụ 1 – Ước tính một chỉ số trung bình: Chúng ta muốn ước tính chiều cao ở đàn ông người Việt, và chấp nhận sai số trong vòng 1 cm (d = 1) với khoảng tin cậy 0.95 (tức α=0.05) và power = 0.8 (hay β = 0.2). Các nghiên cứu trước cho biết độ lệch chuẩn chiều cao ở người Việt khoảng 4.6 cm. Như vậy, hệ số ảnh hưởng là: ES = 1/4.6 = 0.217, và hằng số C = 7.85. Chúng ta có thể áp dụng công thức [2] để ước tính cỡ mẫu cần thiết cho nghiên cứu:

$$n = \frac{C}{(ES)^2} = \frac{7.85}{(0.217)^2} = 166$$

Nói cách khác, chúng ta cần phải đo chiều cao ở 166 đối tượng để ước tính chiều cao đàn ông Việt với sai số trong vòng 1 cm.

Nếu sai số chấp nhận là 0.5 cm (thay vì 1 cm), số lượng đối tượng cần thiết là: $n = \frac{7.85}{(0.5/4.6)^2} = 664$. Nếu độ sai số mà chúng ta chấp nhận là 0.1 cm thì số lượng đối tượng nghiên cứu lên đến 16610 người! Qua các ước tính này, chúng ta dễ dàng thấy cỡ mẫu tùy thuộc rất lớn vào độ sai số mà chúng ta chấp nhận. Muốn có ước tính càng chính xác, chúng ta cần càng nhiều đối tượng nghiên cứu.

Ví dụ 2 – Ước tính cỡ mẫu cho nghiên cứu "trước – sau": Một loại thuốc điều trị có khả năng tăng độ alkaline phosphatase ở bệnh nhân loãng xương. Độ lệch chuẩn của alkaline phosphatase là 15 U/l. Một nghiên cứu mới sẽ tiến hành trong một quần thể bệnh nhân ở Việt Nam, và các nhà nghiên cứu muốn biết bao nhiêu bệnh nhân cần tuyển để chứng minh rằng thuốc có thể alkaline phosphatase từ 60 đến 65 U/l sau 3 tháng điều trị, với sai số α = 0.05 và power = 0.8.

Đây là một loại nghiên cứu "trước – sau" (before-after study); có nghĩa là trước và sau khi điều trị. Ở đây, chúng ta chỉ có một nhóm bệnh nhân, nhưng được đo hai lần (trước khi dùng thuốc và sau khi dùng thuốc). Chỉ tiêu lâm sàng để đánh giá hiệu nghiệm của thuốc là độ thay đổi về alkaline phosphatase. Trong trường hợp này, chúng ta có thể ước tính hệ số ảnh hưởng như sau:

$$ES = \frac{5}{15} = 0.3333$$

Vì là nghiên cứu trước – sau, chúng ta cần một thông tin khác nữa: đó là hệ số tương quan giữa hai lần đo lường alkaline phosphatase. Chúng ta không biết hệ số này, nhưng có thể giả định nó dao động khoảng 0.6 đến 0.8. Với hệ số tương quan 0.6, và sử dụng công thức [3], chúng ta có thể ước tính số cỡ mẫu như sau:

$$n = \frac{2 \times C \times (1-r)}{(ES)^2} = \frac{2 \times 7.85 \times (1-0.6)}{(0.3333)^2} = 56$$

Nhưng nếu hệ số tương quan là 0.8, thì số cỡ mẫu trở thành:

$$n = \frac{2 \times 7.85 \times (1-0.8)}{(0.3333)^2} = 28$$

Nói cách khác, khi hệ số tương quan càng cao (tức độ tin cậy của đo lường cao), số lượng cỡ mẫu càng thấp.

5.2 Ước tính cỡ mẫu cho so sánh hai số trung bình (hai nhóm)

Ví dụ 3 – Nghiên cứu so sánh hai chỉ số trung bình: Một nghiên cứu được thiết kế để thử nghiệm thuốc alendronate trong việc điều trị loãng xương ở phụ nữ sau thời kì mãn kinh. Có hai nhóm bệnh nhân được tuyển: nhóm 1 là nhóm can thiệp (được điều trị bằng alendronate), và nhóm 2 là nhóm đối chứng (tức không được điều trị). Tiêu chí để đánh giá hiệu quả của thuốc là mật độ xương (bone mineral density – BMD). Số liệu từ nghiên cứu dịch tễ học cho thấy giá trị trung bình của BMD trong phụ nữ sau thời kì mãn kinh là 0.80 g/cm^2, với độ lệch chuẩn là 0.12 g/cm^2. Vấn đề đặt ra là chúng ta cần phải nghiên cứu ở bao nhiêu đối tượng để "chứng minh" rằng sau 12 tháng điều trị BMD của nhóm 1 tăng khoảng 5% so với nhóm 2?

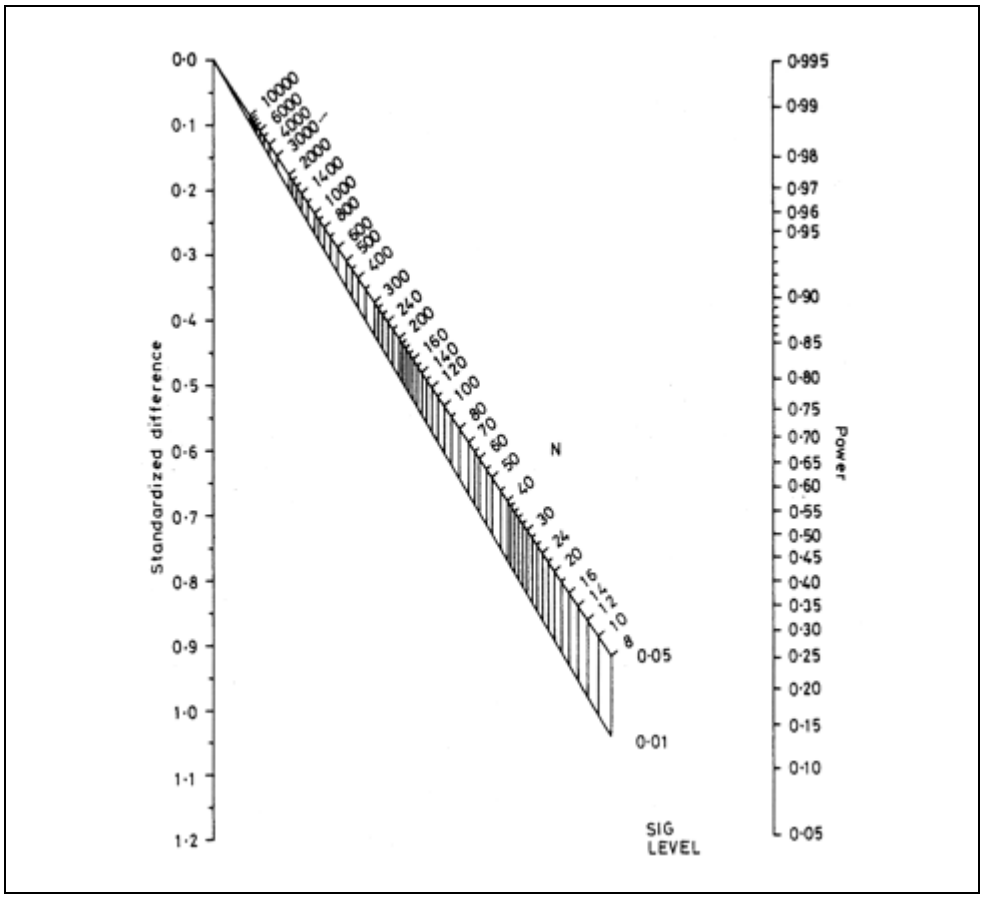
Trong ví dụ trên, tạm gọi trị số trung bình của nhóm 2 là μ_2 và nhóm 1 là μ_1, chúng ta có: μ_2 = 0.8*1.05 = 0.84 g/cm^2 (tức tăng 5% so với nhóm 1), và do đó, Δ = 0.84

– 0.80 = 0.04 g/cm^2. Độ lệch chuẩn là σ = 0.12 g/cm^2. Như vậy, hệ số ảnh hưởng là: $ES = 0.04 / 0.12 = 0.3333$. Với power = 0.90 và α = 0.05, hằng số C = 10.51, và số cỡ mẫu cần thiết là:

$$n = \frac{2C}{(ES)^2} = \frac{2 \times 10.51}{(0.333)^2} = 189$$

Kết quả trên cho biết chúng ta cần 190 bệnh nhân **cho mỗi nhóm** (hay 380 bệnh nhân cho công trình nghiên cứu). Trong trường hợp này, power = 0.90 và α = 0.05 có nghĩa là gì ? Trả lời: hai thông số đó có nghĩa là nếu chúng ta tiến hành thật nhiều nghiên cứu (ví dụ 1000) và mỗi nghiên cứu với 380 bệnh nhân, sẽ có 90% (hay 900) nghiên cứu sẽ cho ra kết quả trên với trị số p < 0.05.

Vì đây là thể loại nghiên cứu thông dụng, cho nên có người vẽ một biểu đồ (xem biểu đồ 1 dưới đây) để ước tính cỡ mẫu cho những ai không thích tính toán. Biểu đồ này đòi hỏi người sử dụng phải biết được hệ số ảnh hưởng (mà biểu đồ viết là "standardised difference") và power.



Biểu đồ 1. Biểu đồ (nomogram) cho ước tính cỡ mẫu và power cho các nghiên cứu hai nhóm. (Nguồn: *British Medical Journal*, 1980, **281,** 1336–1338).
Cách sử dụng: Lấy ví dụ 3, chúng ta có standardised difference là 0.33 (tức ES), power = 0.9. Đánh dấu 0.33 ở cột trái, 0.9 cột phải; kẻ nối hai điểm đã đánh dấu bằng một thước thẳng. Điểm giao chéo giữa đường kẻ thẳng và cột giữa chính là số cỡ mẫu cần thiết cho $\alpha = 0.05$ hay 0.01.

Ví dụ 4 – Ước tính cỡ mẫu để "chứng minh" hai thuật điều trị có hiệu quả tương đương nhau: Một nghiên cứu được thiết kế để "chứng minh" rằng hiệu quả của thuốc A và B tương đương nhau. Các nhà nghiên cứu chọn mật độ xương (BMD) làm tiêu chí lâm sàng. Nghiên cứu gồm 2 nhóm bệnh nhân loãng xương được phân chia ngẫu nhiên: nhóm 1 được điều trị bằng A, và nhóm 2 được điều trị bằng B. Các nghiên cứu trước cho thấy sau 6 tháng điều trị, A có thể tăng BMD khoảng 7%, và B có tác dụng tăng BMD khoảng 4%. Theo các nghiên cứu đó, độ lệch chuẩn của tăng BMD là 10%. Các nhà nghiên cứu quyết định rằng nếu độ khác biệt về BMD giữa hai nhóm trong vòng 2 g/cm^2 thì sẽ xem là hai loại thuốc có hiệu quả như nhau. Vấn đề đặt ra là cần bao nhiêu đối tượng cho nghiên cứu với $\alpha = 0.05$ và power = 0.8?

Với các số liệu trên, chúng ta có thể ước tính hệ số ảnh hưởng (xem phần 4.1.4) như sau: $H = \dfrac{|\mu_1 - \mu_2| - d}{\sigma} = \dfrac{|7-4|-2}{10} = 0.1$. Với $\alpha = 0.05$ và power = 0.8, hằng số C = 7.85 (Bảng 3). Số lượng cỡ mẫu cần thiết cho mỗi nhóm (theo công thức [5]) là:

$$n = \frac{2C}{H^2} = \frac{2 \times 7.85}{(0.1)^2} = 1570$$

Nói cách khác, công trình cần tuyển chọn 3140 đối tượng để đạt được yêu cầu và mục tiêu của nghiên cứu.

5.3 Ước tính cỡ mẫu cho các nghiên cứu ước tính hệ số tương quan

Ví dụ 5 – Nghiên cứu tìm hiểu độ tương quan. Mối tương quan giữa lượng đường trong máu (fasting plasma glucose - FPG) và tỉ trọng cơ thể (body mass index – BMI) thường không nhất quán giữa các sắc dân, và ngay cả ở một sắc, cũng không nhất quán giữa các quần thể. Một nghiên cứu được thiết kế để ước tính hệ số tương quan giữa BMI và FPG. Số liệu từ các nghiên cứu trước cho thấy hệ số tương quan thường dao động từ 0.08 đến 0.30. Vấn đề đặt ra là trong độ dao động đó, nghiên cứu cần đo lường bao nhiêu đối tượng để có kết quả với độ tin cậy 99% (tức $\alpha = 0.01$) và power = 0.80?

Công thức [9] có thể ứng dụng để ước tính cỡ mẫu cho nghiên cứu. Giả dụ như hệ số tương quan thật là 0.15, và với $\alpha = 0.01$ và power = 0.80, hằng số C = 13.33 (Bảng 3). Số cỡ mẫu do đó là:

$$n = 3 + \frac{C}{0.25\left[\log\left(\frac{1+r}{1-r}\right)\right]^2} = 3 + \frac{11.33}{0.25 \times \left[\log\left(\frac{1+0.15}{1-0.15}\right)\right]^2} = 499$$

Nói cách khác, công trình này cần phải tuyển khoảng 500 đối tượng để đạt được yêu cầu và mục tiêu của nghiên cứu. Bảng sau đây cho thấy số lượng cỡ mẫu có thể dao động khá cao tùy thuộc vào hệ số tương quan.

Bảng 4. Số cỡ mẫu cần thiết để ước tính hệ số tương quan với $\alpha = 0.01$ hay 0.05 và power = 0.80

Hệ số tương quan	**Số cỡ mẫu cần thiết cho power = 0.80 và**	
	$\alpha = 0.01$	**$\alpha = 0.05$**
0.05	4527	3138
0.10	1128	783
0.15	499	347
0.20	279	194
0.25	177	123
0.30	121	85
0.35	88	62
0.40	66	47
0.45	51	36
0.50	41	29

5.3 Ước tính cỡ mẫu để ước tính một tỉ lệ

Ví dụ 6 – Nghiên cứu ước tính tỉ lệ lưu hành: Chúng ta muốn ước tính tỉ lệ đàn ông hút thuốc lá ở Việt Nam sao cho ước số không cao hơn hay thấp hơn 2% so với tỉ lệ thật trong toàn dân số. Một nghiên cứu trước đây cho thấy tỉ lệ hút thuốc trong đàn ông người Việt có thể lên đến 70%. Câu hỏi đặt ra là chúng ta cần nghiên cứu trên bao nhiêu đàn ông để đạt yêu cầu trên.

Trong ví dụ này, chúng ta có sai số $m = 0.02$, $\hat{p} = 0.70$, và số lượng cỡ mẫu cần thiết cho nghiên cứu (theo công thức [6]) là:

$$n \geq \left(\frac{1.96}{0.02}\right)^2 0.7 \times 0.3$$

Nói cách khác, chúng ta cần nghiên cứu ít nhất là 2017. Nếu chúng ta muốn giảm sai số từ 2% xuống 1% (tức m = 0.01) thì số lượng đối tượng sẽ là 8067! Chỉ cần thêm độ chính xác 1%, số lượng mẫu có thể thêm hơn 6000 người. Do đó, vấn đề ước tính cỡ mẫu phải rất thận trọng, xem xét cân bằng giữa độ chính xác thông tin cần thu thập và chi phí.

5.4 Ước tính cỡ mẫu cho so sánh hai tỉ lệ

Ví dụ 7 – Nghiên cứu so sánh hai tỉ lệ phát sinh (incidence rate): Một thử nghiệm lâm sàng đối chứng ngẫu nhiên được thiết kế để đánh giá hiệu quả của một loại thuốc chống gãy xương sống. Hai nhóm bệnh nhân sẽ được tuyển. Nhóm 1 được điều trị bằng thuốc, và nhóm 2 là nhóm đối chứng (không được điều trị). Các nhà nghiên cứu giả thiết rằng tỉ lệ gãy xương trong nhóm 2 là khoảng 10%, và thuốc có thể làm giảm tỉ lệ này xuống khoảng 6%. Nếu các nhà nghiên cứu muốn thử nghiệm giả thiết này với sai sót I là α = 0.01 và power = 0.90, bao nhiêu bệnh nhân cần phải được tuyển mộ cho nghiên cứu?

Ở đây, chúng ta có Δ = 0.10 – 0.06 = 0.04, và $\bar{p}$ = (0.10 + 0.06)/2 = 0.08. Với α = 0.01, $z_{\alpha/2}$ = 2.57 và với power = 0.90, z_{β} = 1.28. Do đó, số lượng bệnh nhân cần thiết cho mỗi nhóm (theo công thức [7]) là:

$$n = \frac{\left(2.57\sqrt{2 \times 0.08 \times 0.92} + 1.28\sqrt{0.1 \times 0.90 + 0.06 \times 0.94}\right)^2}{\left(0.04\right)^2} = 1361$$

Như vậy, công trình nghiên cứu này cần phải tuyển ít nhất là 2722 (1361 x 2) bệnh nhân để kiểm định giả thiết trên.

Ví dụ 8 – Nghiên cứu "chứng minh" hai tỉ lệ tương đương: Quay lại với **ví dụ 4** về nghiên cứu nhằm "chứng minh" hai loại thuốc A và B có hiệu quả như nhau. Nhưng lần này, tiêu chí lâm sàng để đánh giá hiệu quả của thuốc là tỉ lệ phát sinh gãy xương cột sống (incidence of vertebral fracture), chứ không phải là sự thay đổi mật độ xương. Số liệu từ các nghiên cứu trước đây cho thấy tỉ lệ gãy xương mới ở các bệnh nhân được điều trị bằng A là khoảng 2% và B là 3%. Đứng trên quan điểm lâm sàng, các nhà nghiên cứu

cho rằng nếu hai tỉ lệ khác nhau trong vòng 0.5% thì có thể xem là tương đương. Vấn đề đặt ra là cần bao nhiêu đối tượng cho nghiên cứu để đạt được $\alpha = 0.05$ (tức độ tin cậy 0.95) và power = 0.80?

Với các số liệu trên ($p_1 = 0.02$, $p_2 = 0.03$, và $d = 0.005$ tức 0.5%) và áp dụng công thức [8], chúng ta có thể ước tính số cỡ mẫu cần thiết cho mỗi nhóm như sau:

$$n = \frac{2C\left[p_1(1-p_1) + p_2(1-p_2)\right]}{\left(|p_1 - p_2| - d\right)^2} = \frac{2 \times 7.85 \times \left[(0.02 \times 0.98) + (0.03 \times 0.97)\right]}{\left(|0.02 - 0.03| - 0.005\right)^2} = 15291$$

Do đó, công trình này cần tuyển 30582 đối tượng để đạt yêu cầu của nghiên cứu. Như có thể thấy được trong công thức trên, yếu tố quan trọng trong cách tính cỡ mẫu cho nghiên cứu loại này là độ khác biệt *d* để xem là hiệu quả hai loại thuốc tương đương. Số cỡ mẫu trên dựa vào tiêu chuẩn khác biệt 0.5% để kết luận "tương đương". Nhưng nếu tiêu chuẩn "dễ dãi" hơn một chút (như 1%) thì số lượng cỡ mẫu giảm xuống cho mỗi nhóm giảm xuống còn 7646 đối tượng – vẫn là một con số lớn. So sánh với kết quả ước tính từ **ví dụ 4**, chúng ta thấy cùng một mục tiêu nghiên cứu, nhưng vấn đề chọn tiêu chí lâm sàng ở đây rất quan trọng và có ảnh hưởng lớn đến số cỡ mẫu.

5.5 Ước tính cỡ mẫu cho một tỉ số nguy cơ (odds ratio)

Ví dụ 9 – Nghiên cứu bệnh-chứng (case-control study): Nhà nghiên cứu muốn tìm hiểu mối liên hệ giữa hút thuốc lá và nguy cơ gãy xương cột sống (vertebral fracture). Hai nhóm đối tượng được chọn: Nhóm 1 là những bệnh nhân mới gãy xương cột sống, và nhóm 2 là những đối tượng không gãy xương, nhưng có cùng độ tuổi, giới với nhóm bệnh nhân. Sau khi có hai nhóm đối tượng, các nhà nghiên cứu sẽ phỏng vấn xem trong hai nhóm, có bao nhiêu người hút thuốc lá. Các nhà nghiên cứu giả thiết rằng tỉ số nguy cơ gãy xương ở những người hút thuốc lá là 2. Nếu các nhà nghiên cứu muốn thử nghiệm giả thiết này với sai sót I là $\alpha = 0.05$ và power = 0.80, bao nhiêu đối tượng cần phải được tuyển mộ cho nghiên cứu? Áp dụng công thức [11], chúng ta có:

$$n = \frac{4C}{(\ln OR)^2\, p(1-p)} = \frac{4 \times 7.85}{(\ln 2)^2 \times 0.25 \times 0.75} = 349$$

Công trình nghiên cứu cần phải tuyển ít nhất là 350 đối tượng (175 bệnh nhân và 175 đối chứng) để kiểm định giả thiết trên.

5.6 Ước tính cỡ mẫu cho nghiên cứu về sống sót

Ví dụ 10 – Nghiên cứu so sánh thời gian sống sót: Như đề cập trong mục 4.5, nhiều nghiên cứu y khoa có mục đích so sánh thời gian sống sót (survival time) giữa hai nhóm. Cụm từ "sống sót" ở đây phải được hiểu rộng hơn, không chỉ phản ảnh tử vong và còn sống, mà bao gồm thời gian dẫn đến một sự kiện lâm sàng (có thể là ung thư, đột quị, gãy xương, đái tháo đường, v.v…) Chẳng hạn như nghiên cứu tìm hiểu xem một thuốc mới có thể kéo dài thời gian sống của bệnh nhân hay không, các nhà nghiên cứu theo dõi 2 nhóm bệnh nhân (nhóm 1 được điều trị và nhóm 2 là nhóm đối chứng) trong vòng 2 năm. Theo y văn, tỉ lệ sống sót trong vòng 2 năm của nhóm đối chứng là 25%, các nhà nghiên cứu hi vọng thuốc mới có thể kéo dài thời gian sống cao hơn nhóm đối chứng khoảng 1.5 lần. Để đạt được ý nghĩa thống kê $\alpha = 0.05$ và power = 0.80, nghiên cứu cần bao nhiêu bệnh nhân?

Trong trường hợp này, chúng ta biết được $p_1 = 0.25$, và $h = 1.5$, do đó, có thể suy ra từ công thức $h = \frac{\ln(0.25)}{\ln(p_2)} = 1.5$, và theo đó: $p_2 = 0.397$. Với $\alpha = 0.05$ và power = 0.80, hằng số C = 7.85. Thay thế các số này vào công thức [12], chúng ta có số cỡ mẫu cho mỗi nhóm là:

$$n = \frac{C(h+1)^2}{(2-p_1-p_2)(h-1)^2} = \frac{7.85(1.5+1)^2}{(2-0.25-0.397)(1.5-1)^2} = 195.$$

Như vậy, công trình cần phải tuyển chọn 390 bệnh nhân để đạt các yêu cầu và mục đề ra.

5.7 Ước tính cỡ mẫu cho nghiên cứu chẩn đoán

Ví dụ 11 – Nghiên cứu chẩn đoán bệnh lao phổi: Hiện nay chẩn đoán bệnh lao phổi AFB âm tính (soi đờm trực tiếp bằng phương pháp Ziehl-Neelsen âm tính) chưa có tiêu chuẩn thống nhất, chủ yếu dựa vào kết quả X quang phổi và việc đáp ứng điều trị thử bằng thuốc chống lao. Ở Việt Nam các kĩ thuật hiện đại phát hiện nhanh vi khuẩn lao chỉ tiến hành ở các bệnh viện lớn, các trung tâm ở thành phố, còn các bệnh viện cơ sở hay cấp tỉnh chưa có điều kiện áp dụng được. Do đó, các nhà nghiên cứu phát triển một phương pháp chẩn đoán dựa vào cận lâm sàng. Các nhà nghiên cứu hi vọng rằng phương pháp cận lâm sàng sẽ có độ nhậy khoảng 0.80 và độ đặc hiệu khoảng 0.95, và muốn thiết kế nghiên cứu sao cho hai chỉ số này dao động trên dưới 5%. Biết rằng tỉ lệ hiện hành của bệnh lao phổi trong cộng đồng khoảng 20%. Câu hỏi đặt ra là nghiên cứu cần bao nhiêu đối tượng để đạt được độ tin cậy thống kê 95% (tức $\alpha = 0.05$).

Với các số liệu trên, chúng ta có thể ước tính TP+FN:

$$TP + FN = \frac{Z_\alpha^2 \times p_{se} \times (1 - p_{se})}{w^2} = \frac{(1.96)^2 \times 0.80 \times 0.20}{(0.05)^2} = 246$$

Với tỉ lệ hiện hành của bệnh là 20%, số lượng cỡ mẫu cần thiết để ước tính độ nhậy (theo công thức [13]) là: 246 / 0.20 = 1229.

Mặt khác, các số liệu trên, chúng ta có thể ước tính FP+TN:

$$FP + TN = \frac{Z_\alpha^2 \times p_{sp} \times (1 - p_{sp})}{w^2} = \frac{(1.96)^2 \times 0.95 \times 0.05}{(0.05)^2} = 73$$

Với tỉ lệ hiện hành của bệnh là 20%, số lượng cỡ mẫu cần thiết để ước tính độ đặc hiệu (theo công thức [14]) là: 73 / (1 – 0.20) = 91.

Trong trường hợp này, chúng ta có hai ước tính khá khác nhau. Nếu dựa vào tiêu chí độ nhạy, nghiên cứu cần đến 1229 đối tượng, nhưng nếu lấy tiêu chí đặc hiệu nghiên cứu chỉ cần 91 đối tượng. Vấn đề còn lại là phải xác định giữa hai chỉ số nhạy và đặc hiệu, chỉ số nào quan trọng hơn để lấy làm tiêu chí cho nghiên cứu. Đây là một quyết định lâm sàng cần phải cân nhắc trước khi ước tính cỡ mẫu!

6. Một số vấn đề về ước tính cỡ mẫu

Điều chỉnh cho hiện tượng "bỏ cuộc". Các phương pháp ước tính cỡ mẫu trình bày trên đây dựa vào một giả định quan trọng là nghiên cứu sẽ tiến hành suông sẻ, tức là không có đối tượng bỏ cuộc và số liệu thu thập đầy đủ cho tất cả mọi đối tượng. Nhưng trong thực tế, chúng ta biết rằng không có một nghiên cứu nào hoàn hảo như thế cả. Hiện tượng bỏ cuộc (withdrawal) rất phổ biến trong nghiên cứu lâm sàng. Vì nhiều lí do, đối tượng không thể tham gia từ đầu đến cuối công trình nghiên cứu, và phải "bỏ cuộc". Vì bỏ cuộc, cho nên số liệu của đối tượng không đầy đủ. Tùy theo thể loại và tính can thiệp hay xâm phạm của công trình nghiên cứu, tỉ lệ bỏ cuộc và không đầy đủ số liệu có thể dao động từ 10% đến 30%.

Vì thế, ước tính cỡ mẫu cũng phải xem xét đến khả năng trên bằng cách điều chỉnh cho tỉ lệ bỏ cuộc. Nếu theo lí thuyết ước tính, nghiên cứu cần *n* đối tượng, và nếu tỉ lệ bỏ cuộc là *q* thì số lượng đối tượng thực tế cần phải là *n/(*1-*q)*. Chẳng hạn như trong ví dụ 8, số lượng bệnh nhân cần thiết cho nghiên cứu theo lí thuyết là 30582 người, nếu tỉ

lệ bỏ cuộc là 20%, thì trong thực tế nhà nghiên cứu phải cần tuyển đến 30582/(1-0.20) = 34280 bệnh nhân.

Điều chỉnh cho trường hợp không cân đối giữa hai nhóm. Các công thức tính toán cỡ mẫu tôi trình bày trong phần trước còn giả định rằng trong các nghiên cứu gồm hai nhóm thì hai nhóm phải có cùng số lượng đối tượng. Nhưng trong thực tế, nhiều nghiên cứu (nhất là nghiên cứu bệnh chứng – case-control study) số lượng bệnh nhân khó mà bằng số lượng đối chứng, nếu bệnh thuộc vào dạng hiếm trong cộng đồng. Vì thế, nhóm đối chứng cần phải có nhiều đối tượng hơn nhóm bệnh.

Trong trường hợp mất cân đối giữa hai nhóm, số lượng cỡ mẫu cũng cần phải được điều chỉnh. Gọi tổng số cỡ mẫu lí thuyết (của hai nhóm cộng lại) là N; gọi tổng số cỡ mẫu điều chỉnh là N^*, nếu chúng ta kì vọng rằng tỉ số cỡ mẫu của nhóm 1 và nhóm 2 là k, thì N^* có thể xác định bằng công thức sau đây:

$$N^* = \frac{N(1+k)^2}{4k}$$

Dễ thấy rằng nếu hai nhóm cân đối (có cùng cỡ mẫu, hay $k = 1$) thì $N = N^*$. Quay lại ví dụ 10, số lượng cỡ mẫu cần thiết cho mỗi nhóm là 195 (do đó $N = 390$). Nhưng nếu chúng ta kì vọng rằng số cỡ mẫu của nhóm 1 phải hơn nhóm 2 khoảng 1.5 lần, thì tổng số lượng đối tượng cần phải tuyển chọn là: $N = [390\times (1+1.5)^2] / (4\times 1.5) = 406$, tức tăng khoảng 16 đối tượng so với ước tính lí thuyết.

Ước tính cỡ mẫu là bước đầu trong nghiên cứu. Trước khi kết thúc bài viết này, tôi muốn nhân cơ hội để nhấn mạnh một lần nữa, ước tính cỡ mẫu cho nghiên cứu là một bước cực kì quan trọng trong việc thiết kế một nghiên cứu cho có ý nghĩa khoa học, vì nó có thể quyết định thành bại của nghiên cứu. Trước khi ước tính cỡ mẫu nhà nghiên cứu cần phải biết trước (hay ít ra là có vài giả thiết *cụ thể*) về vấn đề mình quan tâm.

Ước tính cỡ mẫu cần một số thông số như đề cập đến trong phần đầu của chương, và nếu các thông số này không có thì không thể ước tính được. Trong trường hợp một nghiên cứu hoàn toàn mới, tức chưa ai từng làm trước đó, có thể các thông số về độ ảnh hưởng và độ dao động đo lường sẽ không có, và nhà nghiên cứu cần phải tiến hành một số mô phỏng (simulation) hay một nghiên cứu sơ khởi để có những thông số cần thiết. Cách ước tính cỡ mẫu bằng mô phỏng là một lĩnh vực nghiên cứu khá chuyên sâu, không nằm trong đề tài của sách này, nhưng bạn đọc có thể tìm hiểu thêm phương pháp này trong các sách giáo khoa về thống kê học cấp cao hơn.

Tài liệu tham khảo:

1. Cohen J. *Statistical power analysis for the behavioral science*. NY: Academic Press, 1969.

2. Hair JF, Anderson RE, et al. *Multivariate data analysis*, 5th Ed. New Jersey:Prentice-Hall, 1998.

3. Green SB. How many subjects does it take to do a regression analysis. *Multivariat Behav Res* 1991; 26:499-510.

4. Harris RJ. *A primer of multivariate analysis*, 2nd Ed. New York: Academic Press, 1985.

5. Freeman LS. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982; 1:121-129.

6. Lee ET. *Statistical methods for survival analysis*. Page 320. New York: Wiley, 1992.

7. Jones SJ, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J* 2003; 20:453-458.

Tài liệu đọc thêm: Các công thức trình bày trong bài viết này có thể tìm thấy trong các sách giáo khoa về dịch tễ học và thống kê học. Ba cuốn sách sau đây có thể xem như loại sách dẫn nhập:

1. Machin JM. *Biostatistical Methods – The Assessment of Relative Risks*. New York: John Wiley & Sons, 2000.

2. Kahn HA, Sempos CT. *Statistical Methods in Epidemiology*. New York: Oxford University Press, 1989.

3. *Phân tích số liệu và tạo biểu đồ bằng R - hướng dẫn thực hành* của tôi (tác giả bài này) do Nhà xuất bản Khoa học và Kỹ thuật phát hành, Thành phố Hồ Chí Minh, 2006. Trong sách có hướng dẫn cách tính cỡ mẫu (và phân tích số liệu) bằng máy tính qua ngôn ngữ thống kê R.

Ngoài ra, bạn đọc muốn tìm hiểu thêm về các phương pháp tính cỡ mẫu có thể tìm đọc các bài báo quan trọng sau đây:

1. Florey CD. Sample size for beginners. *BMJ* 1993;306(6886):1181-4.

2. Day SJ, Graham DF. Sample size and power for comparing two or more treatment groups in clinical trials. *BMJ* 1989;299(6700):663-5.

3. Kieser M, Hauschke D. Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *J Biopharm Stat* 1999;9(4):641-50.

4. Miller DK, Homan SM. Graphical aid for determining power of clinical trials involving two groups. *BMJ* 1988;297(6649):672-6.

5. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995;311(7013):1145-8.

6. Sahai H, Khurshid A. Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review. *Stat Med* 1996;15(1):1-21.