16
Ước tính cỡ mẫu
(Sample size estimation)

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, số tình nguyện viên, số mẫu ruộng, cây trồng, thiết bị, v.v… 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 ba yếu tố chính:

- Sai sót mà nhà nghiên cứu chấp nhận, cụ thể là sai sót loại I và II;
- Độ dao động (variability) của đo lường, mà cụ thể là độ lệch chuẩn;
- Mức độ khác biệt hay ảnh hưởng mà nhà nghiên cứu muốn phát hiện;

Không có số liệu về ba yếu tố này thì không thể nào ước tính cỡ mẫu. Trong chương này tôi sẽ bàn qua ba yếu tố trên.

16.1 Khái niệm về "power"

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, gen có chức năng làm cho cây lá có màu xanh, sở thích của con người, 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.

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, khách hàng thích hay không thích một

loại nước giải khát. 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 H-) và *giả thiết chính* (hiện tượng tồn tại, kí hiệu H+).

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, như đề cập trong Chương 7, thường dựa vào trị số P: 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 có thể đặt giả thiết H. Do đó, khi xem xét một giả thiết và kết quả kiểm định thống kê, chúng ta có bốn tình huống:

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

(b) Giả thuyết H đúng, nhưng kết quả kiểm định thống kê không có ý nghĩa thống kê (tức $P>0.05$);

(c) Giả thuyết H sai (thuốc risedronate không có hiệu nghiệm) nhưng kết quả kiểm định thống kê có ý nghĩa thống kê ($P<0.05$);

(d) Giả thuyết H sai và kết quả kiểm định thống kê không có ý nghĩa thống kê ($P>0.05$).

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 sót 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 H là thật. Nói cách khác: *power* = 1-β ;

- Xác suất của tình huống (c) được gọi là *sai sót 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 kê cho ra kết quả p<0.05 với điều kiện giả thiết H 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 H**	
	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 | H+)	Sai sót loại I (**Type I error**) α = P(s | H-)
Không có ý nghĩa thống kê (p>0,05)	Sai sót loại II (**Type II error**) β = P(ns | H+)	Âm tính thật (**True negative**) 1-α = P(ns | H-)

Chú thích: *s* trong biểu đồ này có nghĩa là significant; *ns:* non-significant; *H+* là giả thuyết đúng; *H-* là giả thuyết sai. Do đó, có thể mô tả 4 tình huống trên bằng ngôn ngữ xác suất có điều kiện như sau: Power = 1 – β = P(*s* | H+); β = P(*ns* | H+); và α = P(*s* | H-).

16.2 Thử nghiệm giả thiết thống kê và chẩn đoán y khoa

Có lẽ những lí giải trên, 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 thống kê như là một qui trình chẩn đoán bệnh. 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 (quang tuyến X, 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 H-) và bệnh nhân mắc bệnh (H+). 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 sẽ được 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 gián tiếp thử nghiệm 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 (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*);

- 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 (**Sensitivity**),	Dương tính giả (**False positive)**
-ve (Âm tính)	Âm tính giả (**False negative**),	Độ đặc hiệu (**Specificity**),

Đến đây, chúng ta có thể thấy qua mối tương quan song song giữa chẩn đoán y khoa và thử nghiệm thống kê. Trong chẩn đoán y khoa có chỉ số dương tính thật, tương đương với khái niệm "power" trong nghiên cứu. 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	**Thử nghiệm giả thiết khoa học**
Chẩn đoán bệnh	Thử nghiệm một giả thiết khoa học
Bệnh trạng (có hay không)	Giả thiết khoa học (H+ hay H-)
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 | H+)
Dương tính giả (false positive)	Sai sót loại I; trị số p; α; P(s | H-)
Âm tính giả (false negative)	Sai sót loại II; β; β = P(ns | H+)
Âm tính thật (đặc hiệu, hay specificity)	Âm tính thật; $1-\alpha$ = P(ns | H-)

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.

16.3 Số liệu để ước tính cỡ mẫu

Như đã đề cập trong phần đầu của chương này, để ước tính số đối tượng cần thiết cho một công trình nghiên cứu, chúng ta cần phải có 3 số liệu: xác suất sai sót loại I và II, độ dao động của đo lường, và độ ảnh hưởng.

- Về xác suất sai sót, thông thường một nghiên cứu chấp nhận sai sót loại I khoảng 1% hay 5% (tức $\alpha = 0.01$ hay 0.05), và xác suất sai sót loại II khoảng $\beta = 0.1$ đến $\beta = 0.2$ (tức power phải từ 0.8 đến 0.9).

- Độ dao động chính là độ lệch chuẩn (standard deviation) của đo lường mà công trình nghiên cứu dựa vào để phân tích. Chẳng hạn như, nếu nghiên cứu về cao huyết áp, thì nhà nghiên cứu cần phải có độ lệch chuẩn của áp suất máu. Chúng ta tạm gọi độ dao động là σ.

- Độ ảnh hưởng, nếu là công trình nghiên cứu so sánh hai nhóm, là độ khác biệt trung bình giữa hai nhóm mà nhà nghiên cứu muốn phát hiện. Chẳng hạn như nhà nghiên cứu có thể giả thiết rằng bệnh nhân được điều trị bằng thuốc A có áp suất máu giảm 10 mmHg so với nhóm giả dược. Ở đây, 10 mmHg được xem là độ ảnh hưởng. Chúng ta tạm gọi độ ảnh hưởng là Δ.

Một nghiên cứu có thể có một nhóm đối tượng hay nhiều nhóm đối tượng. Và ước tính cỡ mẫu cũng tùy thuộc vào các trường hợp này.

Trong trường hợp một nhóm đối tượng, số lượng đối tượng (n) cần thiết cho nghiên cứu có thể tính toán một cách "thủ công" như sau:

$$n = \frac{C}{\left(\Delta/\sigma\right)^2} \quad [1]$$

Trong trường hợp có hai nhóm đối tượng, số lượng đối tượng (n) cần thiết cho nghiên cứu có thể tính toán như sau:

$$n = 2 \times \frac{C}{\left(\Delta/\sigma\right)^2} \quad [2]$$

Trong đó, hằng số *C* được xác định từ xác suất sai sót loại I và II (hay power) như sau:

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

α	$\beta = 0.20$ (Power = 0.80)	$\beta = 0.10$ (Power = 0.90)	$\beta = 0.05$ (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

16.4 Ước tính cỡ mẫu

16.4.1 Ước tính cỡ mẫu cho một chỉ số trung bình

Ví dụ 1: Chúng ta muốn ước tính chiều cao của 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. Chúng ta có thể áp dụng công thức [1] để ước tính cỡ mẫu cần thiết cho nghiên cứu:

$$n = \frac{C}{\left(\Delta/\sigma\right)^2} = \frac{7.85}{\left(1/4.6\right)^2} = 166$$

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: $n = \frac{7.85}{\left(0.5/4.6\right)^2} = 664$. 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.

Trong R có hàm `power.t.test` có thể áp dụng để ước tính cỡ mẫu cho ví dụ trên như sau. Chú ý chúng ta cho R biết vấn đề là một nhóm tức `type="one.sample"`:

```
# sai số 1 cm, độc lệch chuẩn 4.6, a=0.05, power=0.8
> power.t.test(delta=1, sd=4.6, sig.level=.05, power=.80,
```

```
              type='one.sample')

     One-sample t test power calculation

              n = 168.0131
          delta = 1
             sd = 4.6
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Kết quả tính toán từ R là 168, khác với cách tính thủ công 2 đối tượng, vì cố nhiên R sử dụng nhiều số lẻ hơn và chính xác hơn cách tính thủ công. Với sai số 0.5 cm:

```
# sai số 0.5 cm, độc lệch chuẩn 4.6, a=0.05, power=0.8
> power.t.test(delta=0.5, sd=4.6, sig.level=.05, power=.80,
              type='one.sample')

  One-sample t test power calculation

              n = 666.2525
          delta = 0.5
             sd = 4.6
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Ví dụ 2: 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 I $\alpha = 0.05$ 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 trị số tăng trung bình là 5 U/l và độ lệch chuẩn là 15 U/l, hay nói theo ngôn ngữ R, `delta=5, sd=15, sig.level=.05, power=.80,` và lệnh:

```
> power.t.test(delta=3, sd=15, sig.level=.05, power=.80,
              type='one.sample')

     One-sample t test power calculation
```

```
           n = 198.1513
       delta = 3
          sd = 15
   sig.level = 0.05
       power = 0.8
 alternative = two.sided
```

Như vậy, chúng ta cần phải có 198 bệnh nhân để đạt các mục tiêu trên.

16.4.2 Ước tính cỡ mẫu cho so sánh hai số trung bình

Trong thực tế, rất nhiều nghiên cứu nhằm so sánh hai nhóm với nhau. Cách ước tính cỡ mẫu cho các nghiên cứu này chủ yếu dựa vào công thức [2] như trình bày phần 15.3.1.

Ví dụ 3: 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_1 = 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. Với power = 0.90 và α = 0.05, cỡ mẫu cần thiết là:

$$n = \frac{2C}{\left(\Delta/\sigma\right)^2} = \frac{2\times 10.51}{\left(0.04/0.12\right)^2} = 189$$

Và lời giải từ R qua hàm `power.t.test` như sau:

```
> power.t.test(delta=0.04, sd=0.12, sig.level=0.05,
               power=0.90, type="two.sample")

     Two-sample t test power calculation

              n = 190.0991
```

```
          delta = 0.04
             sd = 0.12
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Chú ý trong hàm `power.t.test`, ngoài các thông số thông thường như `delta` (độ ảnh hưởng hay khác biệt theo giả thiết), `sd` (độ lệch chuẩn), `sig.level` xác suất sai sót loại I, và `power`, chúng ta còn phải cụ thể chỉ ra rằng đây là nghiên cứu gồm có hai nhóm với thông số `type="two.sample"`.

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 $\alpha = 0.05$ 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 $p < 0.05$.

16.4.3 Ước tính cỡ mẫu cho phân tích phương sai

Phương pháp ước tính cỡ mẫu cho so sánh giữa hai nhóm cũng có thể khai triển thêm để ước tính cỡ mẫu cho trường hợp so sánh hơn hai nhóm. Trong trường hợp có nhiều nhóm, như đề cập trong Chương 11, phương pháp so sánh là phân tích phương sai. Theo phương pháp này, số trung bình bình phương phần dư (residual mean square, RMS) chính là ước tính của độ dao động của đo lường trong mỗi nhóm, và chỉ số này rất quan trọng trong việc ước tính cỡ mẫu.

Chi tiết về lí thuyết đằng sau cách ước tính cỡ mẫu cho phân tích phương sai khá phức tạp, và không nằm trong phạm vi của chương này. Nhưng nguyên lí chủ yếu vẫn không khác so với lí thuyết so sánh giữa hai nhóm. Gọi số trung bình của k nhóm là $\mu_1, \mu_2, \mu_3, \ldots, \mu_k$, chúng ta có thể tính tổng bình phương giữa các nhóm bằng SS $SS = \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$, trong đó, $\bar{\mu} = \sum_{i=1}^{k} \mu_i / k$.

Cho $\lambda = \dfrac{SS}{(k-1)\,RMS}$, vấn đề đặt ra là tìm số lượng cỡ mẫu n sao cho z_β đáp ứng yêu cầu power = 0.80 hay 0.9, mà

$$z_\beta = \frac{1}{\sqrt{(k-1)(1+n\lambda)F + k(n-1)(1+2n\lambda)}} \times$$

$$\left(\sqrt{k(n-1)\left[2(k-1)(1+n\lambda)^2-(1|2n\lambda)\right]}-\sqrt{F(k-1)(1+n\lambda)(2k(n-1)-1)}\right)$$

Trong đó F là kiểm định F. (Xem J. Fleiss, "The Design and Analysis of Clinical Experiments", John Wiley & Sons, New York 1986, trang 373).

Ví dụ 4. Để so sánh độ ngọt của một loại nước uống giữa 4 nhóm đối tượng khác nhau về giới tính và độ tuổi (tạm gọi 4 nhóm là A, B, C và D), các nhà nghiên cứu giả thiết rằng độ ngọt trong nhóm A, B. C và D lần lược là 4.5, 3.0, 5.6, và 1.3. Qua xem xét nhiều nghiên cứu trước, các nhà nghiên cứu còn biết rằng RMS về độ ngọt trong mỗi nhóm là khoảng 8.7. Vấn đề đặt ra là bao nhiêu đối tượng cần nghiên cứu để phát hiện sự khác biệt có ý nghĩa thống kê ở mức độ $\alpha = 0.05$ và power = 0.9.

Hàm `power.anova.test` trong R có thể ứng dụng để giải quyết vấn đề. Chúng ta chỉ cần đơn giản cung cấp 4 số trung bình theo giả thiết và số RMS như sau:

```
# trước hết cho 4 số trung bình vào một vector

> groupmeans <- c(4.5, 3.0, 5.6, 1.3)

# sau đó, "gọi" hàm power.anova.test:

> power.anova.test(groups = length(groupmeans),
                   between.var=var(groupmeans),
                                within.var=8.7,  power=0.90,
sig.level=0.05)

          Balanced  one-way  analysis  of  variance  power
calculation

         groups = 4
              n = 12.81152
    between.var = 3.486667
     within.var = 8.7
      sig.level = 0.05
          power = 0.9

 NOTE: n is number in each group
```

Kết quả cho thấy các nhà nghiên cứu cần khoảng 13 đối tượng cho mỗi nhóm (tức 52 đối tượng cho toàn bộ nghiên cứu).

16.4.4 Ước tính cỡ mẫu để ước tính một tỉ lệ

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 một bệnh trong cộng đồng, hay giới thăm dò ý kiến và thị trường thường tìm hiểu tỉ lệ dân số ưa thích một sản phẩm. Trong các 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, thích / không thích, v.v… Và cách ước tính cỡ mẫu cũng khác với ba ví dụ trên.

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. Để trả lời câu hỏi này, chúng ta thử xem qua một vài lí thuyết về ước tính cỡ mẫu cho một tỉ lệ.

Chúng ta biết qua Chương 6 và 9 rằng nếu $\hat{p}$ được ước tính từ n đối tượng, thì khoảng tin cậy 95% của một tỉ lệ p [trong dân số] là: $\hat{p} \pm 1.96 \times SE(\hat{p})$, trong đó $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$.

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

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

Chúng ta muốn tìm số lượng đối tượng n để đạt yêu cầu trên. Qua cách diễn đạt trên, có thể thấy rằng:

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

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.

Ví dụ 5: Chúng ta muốn ước tính tỉ lệ đàn ông hút thuốc 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 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 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.

R không có hàm cho ước tính cỡ mẫu cho một tỉ lệ, nhưng với công thức trên, bạn đọc có thể viết một hàm để tính rất dễ dàng.

16.4.5 Ước tính cỡ mẫu cho so sánh hai tỉ lệ

Nhiều nghiên cứu mang tính suy luận thường có hai hay nhiều hơn hai nhóm để so sánh. Trong phần 15.4.2 chúng ta đã làm quen với phương pháp ước tính cỡ mẫu để so sánh hai số trung bình bằng kiểm định t. Đó là những nghiên cứu mà tiêu chí là những biến số liên tục. Nhưng có nghiên cứu biến số không liên tục mà mang tính nhị phân như vừa bàn trong phần 15.4.3. Để so sánh hai tỉ lệ, phương pháp kiểm định thông dụng nhất là kiểm định nhị phân (binomial test) hay Khi bình phương (χ^2 test). Trong phần này, chúng ta sẽ bàn qua cách tính cỡ mẫu cho hai loại kiểm định thống kê này.

Gọi hai tỉ lệ (Mà chúng ta không biết nhưng muốn tìm hiểu) là p_1 và p_2, gọi $\Delta = p_1 - p_2$. Giả thiết chúng ta muốn kiểm định là Δ = 0. Lí thuyết để ước tính cỡ mẫu cho kiểm định giả thiết này khá phức tạp, nhưng có thể tóm gọn bằng công thức sau đây:

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

Trong đó, $\bar{p}$ = ($p_1 + p_2$)/2, $z_{\alpha/2}$ là trị số z của phân phối chuẩn cho xác suất α/ 2 (Chẳng hạn như khi α = 0.05, thì $z_{\alpha/2}$ = 1.96; Khi α = 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 β = 0.10, thì z_β = 1.28; khi β = 0.20, thì z_β = 0.84).

Ví dụ 6: 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 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 bệnh nhân để kiểm định giả thiết trên.

Hàm `power.prop.test` R có thể ứng dụng để tính cỡ mẫu cho trường hợp trên. Hàm `power.prop.test` cần những thông tin như `power`, `sig.level`, `p1`, và `p2`. Trong ví dụ trên, chúng ta có thể viết:

```
> power.prop.test(p1=0.10, p2=0.06, power=0.90,
                  sig.level=0.01)

Two-sample comparison of proportions power calculation

              n = 1366.430
             p1 = 0.1
             p2 = 0.06
      sig.level = 0.01
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

Chú ý kết quả từ R có phần chính xác hơn (1366 đối tượng cho mỗi nhóm) vì R dùng nhiều số lẽ cho tính toán hơn là tính "thủ công".

Xin nhấn mạnh một lần nữa rằng ước tính cỡ mẫu cho nghiên cứu là một bước cực kì quan trọng trong việc thiết kế một nghiên cứu 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.