Lâm sàng thống kê

Độ lệch chuẩn hay sai số chuẩn?

Nguyễn Văn Tuấn

Trong vài năm qua, tôi nhận khá nhiều email hỏi về những vấn đề căn bản trong thống kê sinh học và phương pháp dịch tễ học. Tôi có ý định mở mục ***Lâm sàng thống kê*** *(Statistical Clinic) để trao đổi với bạn đọc về các vấn đề mà tôi thấy quan trọng này. Tôi hân hoan chào đón các câu hỏi của bạn đọc để có cảm hứng trả lời.*

Trong hàng trăm thư hỏi và tham vấn trong thời gian 3 năm qua, tôi đếm có đến 5 thư hỏi về vấn đề mà tôi lấy làm tựa đề cho bài viết này. Chẳng hạn như một bạn đọc ở Hà Nội viết email đến tôi hỏi: "Thưa thầy! Em đọc thấy trong các tập san y học người ta thường hay trình bày số trung bình kèm theo SEM, nhưng cũng có bài báo trình bày số trung bình kèm theo SD. Xin hỏi Thầy cách trình bày nào đúng?"

Đây là một câu hỏi đơn giản nhưng tôi thấy có ý nghĩa ứng dụng khá rộng, nên muốn nhân cột báo ***Lâm sàng thống kê*** *để trả lời bạn đọc*.

***

Trong các tập san y học, chúng ta thường thấy những cột số dưới hình thức $x \pm y$, trong đó x là số trung bình, còn y thì có khi là độ lệch chuẩn (standard deviation – SD) hay sai số chuẩn (standard error – SE). Cũng có tác giả viết SEM (viết tắt từ cụm từ *standard error of the mean*). Cách trình bày như thế thông dụng đến nỗi một số chuyên gia và các ban biên tập tập san y học phải lên tiếng khuyến cáo. Theo khuyến cáo chung và cũng là qui ước nghiên cứu y học: **để mô tả một biến số lâm sàng tuân theo luật phân phối chuẩn, các nhà nghiên cứu nên cách trình bày số trung bình và kèm độ lệch chuẩn (không phải sai số chuẩn; để mô tả một biến số lâm sàng không tuân theo luật phân phối chuẩn, nên trình bày số trung vị và số ở vị trí 25% và 75% (tức là interquartile range)**.

Để hiểu qui ước này, chúng ta cần phải tìm hiểu ý nghĩa của độ lệch chuẩn và sai số chuẩn. Tôi thấy điều này cần thiết, bởi vì hầu hết sách giáo khoa thống kê (ngay cả sách giáo khoa do người Tây phương viết) đều không giải rõ những khác biệt về ý nghĩa của hai chỉ số thống kê này.

Mô tả một biến số theo luật phân phối chuẩn

Xin nhắc lại thuật ngữ: cụm từ "phân phối chuẩn" ở đây chính là "Normal distribution" (hay có sách còn gọi là "Gaussian distribution", lấy từ tên của nhà toán học vĩ đại người Đức Frederick Gauss). Một biến số tuân theo luật phân phối chuẩn, khi vẽ bằng biểu đồ, giống như hình một cái chuông cân đối (**Biểu đồ 1**). Phân phối này được xác định bằng hai thông số: số trung bình và độ lệch chuẩn. Để tiết kiệm chữ nghĩa, tôi sẽ lấy kí hiệu *m* thể hiện số trung bình, và *s* thể hiện độ lệch chuẩn.

Tại sao chúng ta cần độ lệch chuẩn? Để trả lời câu hỏi này, chúng ta thử xem qua ví dụ sau đây:

Ví dụ 1. Một biến số phản ảnh tình trạng của một bệnh trong hai nhóm bệnh nhân (nhóm A gồm 6 bệnh nhân, và nhóm B gồm 4 bệnh nhân) như sau:

Nhóm A: 6, 7, 8, 4, 5, 6
Nhóm B: 10, 2, 3, 9

Có thể dễ dàng thấy rằng số trung bình của nhóm A là 6, bằng với số trung bình của nhóm B. Tuy có cùng số trung bình, chúng ta khó có thể kết luận hai nhóm này tương đương nhau, bởi vì độ khác biệt trong nhóm B cao hơn trong nhóm A. Thật vậy, độ khác biệt giữa số lớn nhất và số nhỏ nhất trong nhóm B là 8 (tức 10 trừ cho 2) gấp hai lần so với nhóm A với độ khác biệt là 4 (lấy 8 trừ cho 4).

Chúng ta cần một chỉ số để phản ảnh sự khác biệt giữa các bệnh nhân (hay nói theo thuật ngữ là *biến thiên*). Cách làm hiển nhiên nhất là lấy kết quả của từng bệnh nhân trừ cho số trung bình và cộng chung lại. Gọi chỉ số này là *D,* và để phân biệt hai nhóm A và B, chúng ta dùng kí hiệu dưới dòng (subscript):

Nhóm A: $D_A = (6-\mathbf{6}) + (7-\mathbf{6}) + (8-\mathbf{6}) + (4-\mathbf{6}) + (5-\mathbf{6}) + (6-\mathbf{6}) = 0$

Nhóm B: $D_B = (10-\mathbf{6}) + (2-\mathbf{6}) + (3-\mathbf{6}) + (9-\mathbf{6}) = 0$

Như thấy trên, vấn đề ở đây là tổng số khác biệt của *D* là 0. Như vậy *D* vẫn chưa phản ảnh được độ biến thiên mà chúng ta muốn. Một cách làm cho *D* có "hồn" hơn là chúng ta lấy bình phương của từng cá nhân và cộng số bình phương lại với nhau. Gọi chỉ số mới này là D^2, chúng ta có:

Nhóm A: $D_A^2 = (6-\mathbf{6})^2 + (7-\mathbf{6})^2 + (8-\mathbf{6})^2 + (4-\mathbf{6})^2 + (5-\mathbf{6})^2 + (6-\mathbf{6})^2 = 10$

Nhóm B: $D_B^2 = (10-\mathbf{6})^2 + (2-\mathbf{6})^2 + (3-\mathbf{6})^2 + (9-\mathbf{6})^2 = 50$

Bây giờ thì D^2 rõ ràng cho thấy nhóm B có độ biến thiên cao hơn nhóm A. Nhưng còn một vấn đề, vì D^2 là tổng số, tức là chịu ảnh hưởng số cỡ mẫu trong từng nhóm. Một cách điều chỉnh hợp lí nhất là chia D^2 cho số cỡ mẫu. Gọi chỉ số mới này là S^2, chúng ta có:

Nhóm A: $S_A^2 = 10 / 6 = 1.67$

Nhóm B: $S_B^2 = 50 / 4 = 12.5$

Nhưng để khách quan hơn nữa, chúng ta còn phải điều chỉnh cho số thông số sử dụng trong tính toán. Chú ý rằng khi tính D hay D^2, chúng ta trừ kết quả mỗi bệnh nhân cho số trung bình (tức là tốn một thông số). Vì thế, thay vì chia D^2 cho số cỡ mẫu, chúng ta phải chia cho số cỡ mẫu trừ 1. Gọi chỉ số mới nhất là s^2, chúng ta có:

Nhóm A: $s_A^2 = \frac{10}{5-1} = 2$

Nhóm B: $s_B^2 = \frac{50}{4-1} = 16.7$

Chỉ số s^2 ở đây chính là *phương sai*.

Nhưng còn một vấn đề nhỏ nữa: bởi vì đơn vị phương sai là bình phương, khác với đơn vị của số trung bình. Vì thế, cách hoán chuyển tốt nhất là chuyển giá trị của phương sai sao cho có cùng đơn vị với số trung bình bằng cách lấy căn số bậc hai, và đây chính là *độ lệch chuẩn* (kí hiệu s).

Nhóm A: $s_A = \sqrt{2} = 1.41$

Nhóm B: $s_B = \sqrt{16.7} = 4.08$

Đến đây, chúng ta có thể thấy nhóm B có độ biến thiên cao hơn nhóm A. Một cách để định lượng hóa độ lệch chuẩn tương quan với số trung bình là lấy độ lệch chuẩn chia cho số trung bình (và nếu cần, nhân cho 100). Kết quả của tính toán này có tên là *hệ số biến thiên* (coefficient of variation – CV):

Nhóm A: $CV_A = 1.41 / 6 \times 100 = 23.5\%$

Nhóm B: $CV_B = 4.08 / 6 \times 100 = 68.3\%$

Lợi thế của hệ số biến thiên là nó cho chúng ta một phép so sánh các biến số không có cùng đơn vị. Chẳng hạn như chúng ta có thể so sánh độ biến thiên của áp suất máu và độ cholesterol trong một quần thể, vì hệ số biến thiên có cùng đơn vị phần trăm.

Đến đây, chúng ta có thể tóm lược sự phân phối của hai nhóm bệnh nhân bằng bảng sau đây:

Nhóm	Số đối tượng (N)	Trung bình	Độ lệch chuẩn	Hệ số biến thiên
A	6	6.0	1.41	23.5%
B	4	6.0	4.08	68.3%

Mô tả sự biến thiên của số trung bình: sai số chuẩn

Các sách giáo khoa thống kê thường mô tả cách tính sai số chuẩn trong phần mở đầu, nhưng không giải thích nó có nghĩa là gì và tại sao phải cần đến chỉ số thống kê này. Công thức tính sai số chuẩn (kí hiệu bằng SE – viết tắt từ *standard error*) rất đơn giản: lấy độ lệch chuẩn chia cho căn số bậc hai của số cỡ mẫu (n):

$$SE = \frac{s}{\sqrt{n}}$$

Áp dụng công thức trên cho ví dụ, *SE* của nhóm A và B lần lược là:

Nhóm A: $SE_A = 1.41/\sqrt{6} = 0.58$

Nhóm B: $SE_A = 4.08/\sqrt{4} = 2.04$

Tại sao chúng ta cần tính *SE* ? Xin nhắc lại nguyên lí và mục đích đằng sau của thống kê học là ước tính những thông số của một quần thể (population). Trong thực tế chúng ta không biết các thông số này, mà chỉ dựa vào những ước tính từ một hay nhiều mẫu để suy luận cho giá trị của quần thể mà các mẫu được chọn. Chẳng hạn như chúng ta không biết chiều cao của người Việt là bao nhiêu (bởi vì đâu có ai đo lường chiều cao của 82 triệu dân); chúng ta phải chọn một mẫu gồm n đối tượng để tính trị số trung bình của mẫu này, và dùng trị số trung bình của mẫu để suy luận cho toàn dân số.

Nhưng chọn mẫu phải ngẫu nhiên thì mới mang tính đại diện cao. Cứ mỗi lần chọn mẫu, chúng ta có một nhóm đối tượng khác. Và, cứ mỗi mẫu, chúng ta có một số trung bình mới. Câu hỏi đặt ra là: nếu chọn mẫu nhiều lần ("nhiều" ở đây có nghĩa là hàng triệu hay tỉ lần) thì các số trung bình này dao động cỡ nào.

Ví dụ 2. Hãy lấy một ví dụ cụ thể (nhưng đơn giản) để minh họa cho ý tưởng vừa trình bày. Giả sử chúng ta có một quần thể chỉ 10 người, và chiều cao tính bằng cm của 10 người này là:

Quần thể: 130, 189, 200, 156, 154, 160, 162, 170, 145, 140

Như vậy chiều cao trung bình của quần thể (chúng ta biết) là 160.6 cm. Gọi chỉ số này là $\mu = 160.6$ cm.

Bây giờ, giả sử chúng ta không có điều kiện và tài lực để đo chiều cao của toàn bộ quần thể, mà chỉ có khả năng lấy mẫu 5 người từ quần thể này để ước tính chiều cao. Chúng ta có thể lấy nhiều mẫu ngẫu nhiên, mỗi lần 5 người:

Lần thứ 1: 140, 160, 200, 140, 145	$x_1 = 157.0$
Lần thứ 2: 154, 170, 162, 160, 162	$x_2 = 161.6$
Lần thứ 3: 145, 140, 156, 140, 156	$x_3 = 147.4$
Lần thứ 4: 140, 170, 162, 170, 145	$x_4 = 157.4$
Lần thứ 5: 156, 156, 170, 189, 170	$x_5 = 168.2$
Lần thứ 6: 130, 170, 170, 170, 170	$x_6 = 162.0$
Lần thứ 7: 156, 154, 145, 154, 189	$x_7 = 159.6$
Lần thứ 8: 200, 154, 140, 170, 170	$x_8 = 166.8$
Lần thứ 9: 140, 170, 145, 162, 160	$x_9 = 155.4$
Lần thứ 10: 200, 200, 162, 170, 162	$x_{10} = 178.8$
….	

Chú ý trong dãy trên, các số x_1, x_2, x_3, … là số trung bình cho mỗi mẫu được chọn. Chúng ta thấy cứ mỗi lần chọn mẫu, số trung bình chiều cao ước tính khác nhau, và biến thiên từ 147.4 cm đến 178.8 cm. Các số trung bình này dao động chung quanh số trung bình của quần thể (tức là 160.6 cm).

Nếu chúng ta chọn mẫu N lần (mỗi lần với n đối tượng), thì chúng ta sẽ có N số trung bình. **Độ lệch chuẩn của *N* số trung bình này chính là sai số chuẩn**. (Nen nhớ N ở đây là hàng triệu hay tỉ lần). Do đó, sai số chuẩn phản ảnh độ dao động hay biến thiên của các số trung bình mẫu (sample averages).

Một số sách giáo khoa thống kê dùng danh từ "Standard error of the mean" (SEM), nhưng đây là một cách dùng từ sai. Như tôi vừa trình bày trên, không có cái gọi là "standard error of the mean", mà chỉ là *standard deviation of the means* (chú ý chữ

"means" số nhiều vì tính từ nhiều số trung bình). Thay vì gọi standard deviation of the means (quá dài dòng), người ta gọi ngắn gọn bằng một thuật ngữ mới: *standard error*.

Ý nghĩa của độ lệch chuẩn và sai số chuẩn

Gọi thông số trung bình của một quần thể là μ (nên nhớ rằng chúng ta không biết giá trị của μ). Gọi ước số trung bình tính từ mẫu là $\bar{x}$ và độ lệch chuẩn là s. Theo lí thuyết xác suất của phân phối chuẩn, chúng ta có thể phát biểu rằng:

- 68% cá nhân trong quần thể đó có giá trị từ $\bar{x} - s$ đến $\bar{x} + s$;
- 95% cá nhân trong quần thể đó có giá trị từ $\bar{x} - 1.96 \times s$ đến $\bar{x} + 1.96 \times s$;
- 99% cá nhân trong quần thể đó có giá trị từ $\bar{x} - 3 \times s$ đến $\bar{x} + 3 \times s$.

Ngoài ra, gọi sai số chuẩn là *SE*, chúng ta còn có thể phát biểu rằng:

- 68% số trung bình tính từ mẫu có giá trị từ $\bar{x} - SE$ đến $\bar{x} + SE$;
- 95% số trung bình tính từ mẫu có giá trị từ $\bar{x} - 1.96 \times SE$ đến $\bar{x} + 1.96 \times SE$;
- 99% số trung bình tính từ mẫu có giá trị từ $\bar{x} - 3 \times SE$ đến $\bar{x} + 3 \times SE$.

Qua trình bày trên, chúng ta thấy rõ ràng độ lệch chuẩn phản ảnh độ biến thiên của một số cá nhân trong một quần thể. Còn sai số chuẩn phản ảnh độ dao động của các số trung bình chọn từ quần thể.

Ví dụ 3. Chẳng hạn như khi nói trọng lượng trung bình của một nhóm bệnh nhân là 55 kg với độ lệch chuẩn 8.2 kg, thì câu nói này có nghĩa rằng nếu ta chọn [một cách ngẫu nhiên] một bệnh nhân từ quần thể, thì xác suất 95% là bệnh nhân này sẽ có trọng lượng từ $55 - 1.96 \times 8.2 =$ **39 kg** đến $55 + 1.96 \times 8.2 =$ **71 kg**. Giá trị 39 kg đến 71 kg được gọi là **khoảng tin cậy 95%** (95% confidence interval).

Trong trường hợp khoảng tin cậy 95% hàm chứa giá trị âm thì sao? Chúng ta biết rằng chiều cao không thể có giá trị âm! Vì thế, nếu khoảng tin cậy 95% hàm chứa giá trị âm thì điều này cho chúng ta biết rằng hoặc là (a) phân phối của biến số không tuân theo luật phân phối chuẩn, và các số trung bình, độ lệch chuẩn, hay phương sai không còn ý nghĩa thực tế nữa, hoặc (b) cách chọn mẫu có vấn đề. Đây là một đề tài thú vị mà tôi sẽ trở lại trong một bài khác.

Về ý nghĩa của sai số chuẩn, chúng ta quay lại với **Ví dụ 2**. Giả sử chúng ta không biết giá trị thật của số trung bình cho toàn quần thể, mà chỉ dựa vào mẫu thứ nhất để ước tính. Lần chọn mẫu thứ nhất là: 140, 160, 200, 140, 145, và:

Số trung bình của mẫu: $\bar{x}$ = 157.0 cm
Độ lệch chuẩn: s = 25.4 cm
Sai số chuẩn: SE = 25.4/$\sqrt{5}$ = 11.36 cm

Như vậy, theo lí thuyết xác suất, chúng ta có thể nói rằng xác suất 95% là số trung bình của toàn quần thể dao động từ 157─1.96×11.36 = **139 cm** đến 157+1.96×11.36 = **179 cm**. (Trong thực tế, chúng ta biết rằng số trung bình của toàn quần thể là 160.6 cm).

Tóm tắt

Cần phải nói ngay rằng không một biến số lâm sàng nào có thể được mô tả chỉ bằng một ước số. Để có một "bức tranh" chung về một biến số lâm sàng, chúng ta nên sử dụng ba ước số chính: số cỡ mẫu, số trung bình, và độ lệch chuẩn. Sai số chuẩn không cung cấp thông tin về độ biến thiên của một quần thể, cho nên ước số này không nên sử dụng cho việc mô tả một chỉ số lâm sàng.

Nhưng trong thực tế, vì hiểu sai hay nhập nhằng về độ lệch chuẩn và sai số chuẩn nên các bài báo y học được trình bày thiếu thống nhất. Lúc thì các tác giả trình bày độ lệch chuẩn, lại có khi cung cấp sai số chuẩn. Đây không phải là vấn đề gian lận khoa học, mà chỉ đơn giản là thiếu hiểu biết. Chính vì thế mà ban biên tập các tập san y học quốc tế ra chỉ dẫn khuyến cáo tác giả chỉ nên trình bày độ lệch chuẩn kèm theo số trung bình và cỡ mẫu.

Bởi vì mẫu số của sai số chuẩn là số cỡ mẫu, cho nên sai số chuẩn thường thấp hơn độ lệch chuẩn. Chính vì thế mà có khi tác giả có lẽ ngại trình bày độ lệch chuẩn quá cao (ngại người bình duyệt chất vấn và có thể bài báo bị từ chối) nên họ cố tình trình bày bằng độ lệch chuẩn mà không ghi chú thích! Tình trạng nhập nhằng này mới là gian lận khoa học – nhưng là một gian lận ở trình độ thấp.

Hi vọng rằng những giải thích trên đây của tôi đã cung cấp cho bạn đọc một cách hiểu sâu hơn và rõ ràng hơn về khác biệt giữa độ lệch chuẩn và sai số chuẩn.

Chú thích: *Bài viết này thực chất là dựa vào một bài giảng về phương pháp dịch tễ học mà người viết đã thực hiện ở Bộ môn nội tiết (Đại học Y dược, Thành phố Hồ Chí Minh) vào tháng 7 năm 2006, và buổi tập huấn về nghiên cứu khoa học tại Bệnh viện Đa khoa Kiên Giang vào tháng 2 năm 2007. Thành thật cám ơn các bác sĩ, học viên và bạn đọc ykhoa.net đã đặt nhiều câu hỏi làm cảm hứng cho bài viết.*

Thuật ngữ sử dụng trong bài viết

Tiếng Việt	**Tiếng Anh**
Số trung bình	Mean
Độ lệch chuẩn	Standard deviation (SD)
Sai số chuẩn	Standard error (SE)
Khoảng tin cậy 95%	95% confidence interval
Số trung vị	Median
Phân phối chuẩn	Normal distribution (Gaussian distribution)
Biến thiên	Variation
Phương sai	Variance
Hệ số biến thiên	Coefficient of variation (CV)
Quần thể	Population
Sample	Mẫu
Thông số	Parameter
Estimate	Ước số