Lâm sàng thống kê

Ước tính khoảng tin cậy 95% cho một biến số đã hoán chuyển sang đơn vị logarít

Hỏi*: "Nhiều biến số lâm sàng không tuân theo luật phân phối Gaussian, do đó cách tính khoảng tin cậy 95% theo phương pháp thông thường không thể áp dụng. Nếu phải biến đổi biến số sang logarít thì cách tính khoảng tin cậy 95% sẽ như thế nào?"*

Rất nhiều biến số lâm sàng (và trong sinh học nói chung) như lượng đường trong máu, độ cholesterol trong máu, và nhiều chỉ số sinh học khác không tuân theo luật phân phối chuẩn. Trong trường hợp này, phương pháp mô tả biến số thường là số trung vị (median), và các điểm tứ phân vị 25% và 75% (tức là 25th quartile và 75th quartile). Nhưng cũng có trường hợp phân tích, chúng ta cần phải hoán chuyển các biến số này sang một đơn vị khác sao cho tuân theo luật phân phối chuẩn. Một trong những hàm số hoán chuyển là logarít. Khi một biến số đã hoán chuyển sang một đơn vị khác thì tất cả các số trung bình và độ lệch chuẩn cũng thay đổi, cách diễn dịch cũng thay đổi. Bài viết này sẽ trình bày một cách tính rất đơn giản để duy trì ý nghĩa sinh học ban đầu của biến số.

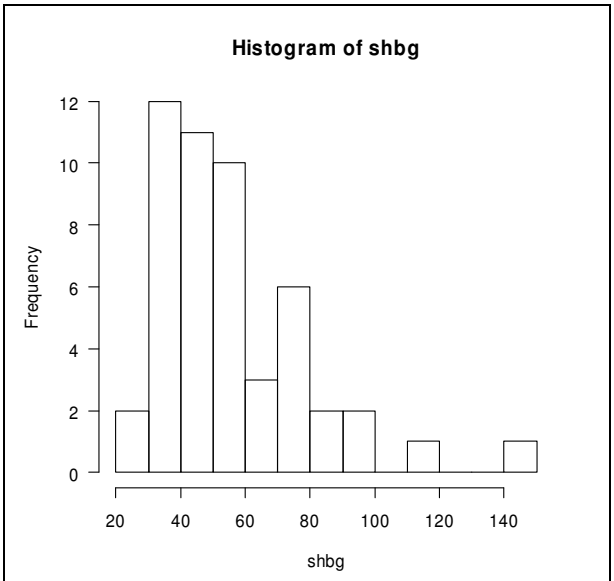
Hãy lấy một ví dụ cụ thể. Chúng ta đo lường độ SHBG ở 50 bệnh nhân nam tuổi 60 trở lên, và kết quả như sau:

```
53.6,  87.1,  35.2, 40.7, 74.5, 35.6, 82.9, 50.2, 33.8, 40.6,
110.5, 147.6, 35.8, 52.5, 72.5, 90.5, 37.8, 76.0, 48.5, 44.7,
53.2,  32.6,  39.3, 49.4, 34.6, 99.3, 46.4, 73.2, 57.7, 24.9,
45.5,  46.7,  45.9, 50.8, 69.2, 57.2, 30.0, 31.5, 50.8, 46.6,
70.8,  64.4,  34.2, 51.9, 49.8, 78.3, 52.1, 33.4, 35.5, 67.4
```

Một vài chỉ số thống kê cho biến số SHBG có thể ước tính như sau:

Số trung bình	55.46
Độ lệch chuẩn	23.42
Trung vị	50

Nếu tính theo luật phân phối chuẩn, khoảng tin cậy 95% của SHBG là: 55.46–1.96×23.42 = 9.55 và 101.37 mmol/L. Nhưng trước khi chấp nhận sự hợp lí của khoảng tin cậy này, chúng ta phải xem qua phân phối của biến SHBG (**Biểu đồ 1**) dưới đây.



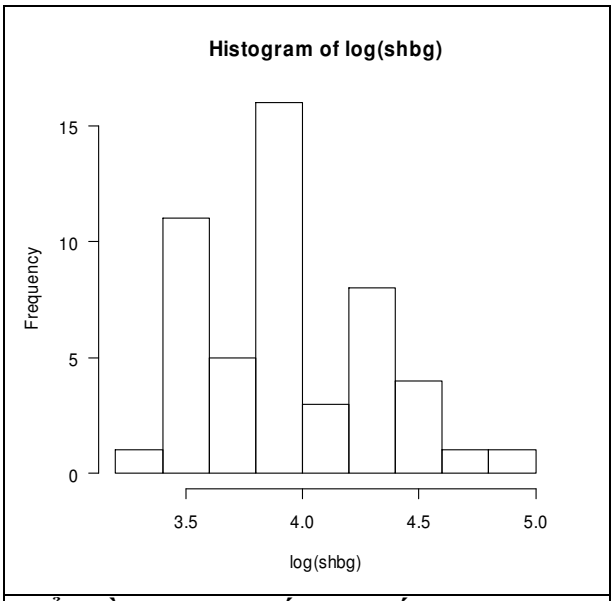
Biểu đồ 1: Phân phối của biến SHBG

Như có thể thấy, phần lớn bệnh nhân có độ SHBG thấp hơn 80 mmol/L, và rất ít bệnh nhân với SHBG cao hơn 80 mmol/L. Nói cách khác, phân phối của SHBG xiên lệch về những giá trị thấp, không cân đối, tức là không tuân theo luật phân phối chuẩn (Normal distribution). Do đó, khoảng tin cậy 95% và số trung bình vừa ước tính trên không có ý nghĩa vì đã vi phạm một qui luật thống kê học.

Cách "khắc phục" cho tình trạng này là hoán chuyển SHBG sang một đơn vị sao cho tuân theo luật phân phối chuẩn. Vì độ lệch về một phía (phía trái) chúng ta có thể áp dụng hàm số logarít để hoán chuyển. Chẳng hạn như thay vì 53.6, chúng ta chuyển thành log(53.6) = 3.98. Tiếp tục hoán chuyển như thế, chúng ta sẽ có một dãy số mới như sau:

```
3.982 4.467 3.561 3.706 4.311 3.572 4.418 3.916 3.520 3.704 4.705 4.995 3.578 3.961
4.284 4.505 3.632 4.331 3.882 3.800 3.974 3.484 3.671 3.900 3.544 4.598 3.837 4.293
4.055 3.215 3.818 3.844 3.826 3.928 4.237 4.047 3.401 3.450 3.928 3.842 4.260 4.165
3.532 3.949 3.908 4.361 3.953 3.509 3.570 4.211
```

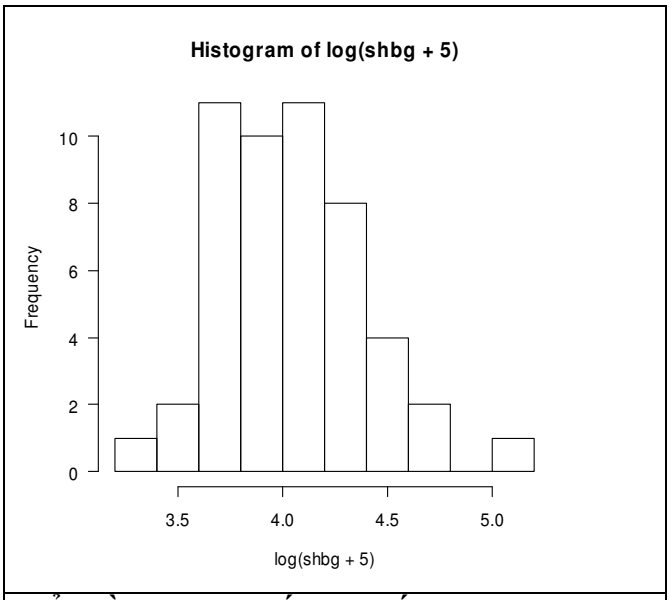
Bây giờ chúng ta thử xem phân phối của log(SHBG) (**Biểu đồ 2**):



Biểu đồ 2: Phân phối của biến log(SHBG)

Phân phối này vẫn chưa thoả đáng, vì vẫn còn xiên lệch. Chúng ta để ý thấy giá trị cao nhất của log(SHBG) là khoảng 5, cho nên chúng ta có thể áp dụng một hàm số hoán chuyển mới: log(SHBG + 5). Chẳng hạn như nếu SHBG = 53.6, thì log(SHBG+5) = log(53.6 + 5) = 4.07. Giá trị mới này cho 50 bệnh nhân và biểu đồ phân phối như sau:

```
4.071 4.523 3.694 3.822 4.376 3.704 4.476 4.011 3.658 3.820 4.749 5.028 3.709 4.052
4.350 4.559 3.757 4.394 3.980 3.906 4.064 3.627 3.791 3.996 3.679 4.647 3.940 4.359
4.138 3.398 3.922 3.945 3.930 4.022 4.307 4.130 3.555 3.597 4.022 3.944 4.328 4.240
3.669 4.041 4.004 4.422 4.045 3.648 3.701 4.282
```



Biểu đồ 3: Phân phối của biến log(SHBG+5)

Bây giờ thì chúng ta đã thành công hoán chuyển SHBG sang phân phối chuẩn. Do đó, chúng ta có thể ước tính số trung bình và độ lệch chuẩn cho biến số mới:

Trung bình log(SHBG+5): 4.041
Độ lệch chuẩn (SD) của log(SHBG+5): 0.3427

Như vậy, khoảng tin cậy 95% của biến số mới là: 4.041-1.96×0.3427 = 3.369 đến 4.041+1.96×0.3427 = 4.712.

Vấn đề đặt ra là chúng ta cần phải hoán chuyển ngược lại đơn vị mmol/L, vì một đơn vị logarít rất khó hiểu và khó diễn dịch. Để hoán chuyển ngược lại, chúng ta tạm gọi log(SHBG+5) = y, và mục tiêu là chúng ta tìm SHBG:

$$\text{Log}(\text{SHBG} + 5) = y$$

Do đó,

$$\text{SHBG} + 5 = e^{y}$$

Hay, cụ thể hơn:

$$\text{SHBG} = e^{y} - 5$$

Do đó, số trung bình và khoảng tin cậy 95% SHBG có thể ước tính như sau:

Trung bình SHBG: $e^{4.041} - 5 = 51.86$ mmol/L
Và khoảng tin cậy 95%: $e^{3.369} - 5 = 24.05$ đến $e^{4.712} - 5 = 106.3$ mmol/L.

Chúng ta có thể so sánh kết quả ước tính "sai" và kết quả ước tính "đúng" như sau:

	Ước tính không hoán chuyển	Ước tính dựa vào hoán chuyển logarít
Số trung bình	55.46	51.86
Khoảng tin cậy 95% CI	23.42 – 101.37	24.05 – 106.3

Nhìn vào Biểu đồ 1, chúng ta sẽ thấy ngay rằng các ước số dựa vào hoán chuyển logarít hợp lí hơn là những ước số không hoán chuyển, vì chúng phản ảnh đầy đủ hơn sự phân phối của SHBG.

Ví dụ trên đây cho thấy trước khi phân tích bằng bất cứ mô hình nào, chúng ta cần phải xem xét cẩn thận phân phối của biến số. Bởi vì phần lớn các phương pháp phân tích thống kê dựa vào giả định luật phân phối chuẩn, vi phạm giả định này cũng có nghĩa là kết quả không có ý nghĩa khoa học cao.

Ghi chú:

Các tính toán trên đây có thể thực hiện bằng máy tính cầm tay hay Excel. Nhưng đối với bạn đọc quen sử dụng ngôn ngữ thống kê R, thì các tính toán và biểu đồ trên được thực hiện bằng các mã sau đây. (Bạn đọc có thể cắt tất cả mã và dán vào R để tự mình kiểm nghiệm).

```
# nhập số liệu 50 bệnh nhân vào biến có tên là shbg

shbg <- c(53.6, 87.1, 35.2, 40.7, 74.5, 35.6, 82.9, 50.2, 33.8, 40.6,
          110.5, 147.6, 35.8, 52.5, 72.5, 90.5, 37.8, 76.0, 48.5,
          44.7, 53.2, 32.6, 39.3, 49.4, 34.6, 99.3, 46.4, 73.2, 57.7,
          24.9, 45.5, 46.7, 45.9, 50.8, 69.2, 57.2, 30.0, 31.5, 50.8,
          46.6,  70.8, 64.4, 34.2, 51.9, 49.8, 78.3, 52.1, 33.4,
          35.5, 67.4)

# ước tính số trung bình, độ lệch chuẩn và 95% CI

mean(shbg)
sd(shbg)
lower95 <- mean(shbg) - 1.96*sd(shbg)
upper95 <- mean(shbg) + 1.96*sd(shbg)

# vẽ biểu đồ 1

hist(shbg, breaks=15)

# hoán chuyển sang log(shbg+5)

logshbg <- log(shbg +5)

# vẽ biểu đồ 3

hist(logshbg)

# tính số trung bình, sd, 95% CI
m <- mean(logshbg)
stdev <- sd(logshbg)
lower95 <- mean(logshbg) - 1.96*sd(logshbg)
upper95 <- mean(logshbg) + 1.96*sd(logshbg)

# hoán chuyển ngược về shbg

exp(m) - 5
exp(lower95) - 5
exp(upper95) - 5
```

Muốn biết thêm cách sử dụng R cho phân tích thống kê, các bạn có thể tham khảo cuốn sách "**Phân tích số liệu và tạo biểu đồ bằng R**" của tôi, do Nhà xuất bản Khoa học Kỹ thuật phát hành đầu năm 2007.

Nguyễn Văn Tuấn