

Lâm sàng thống kê

Phân tích tương quan

Nguyễn Văn Tuấn

Một trong những mục tiêu của nghiên cứu y học (và khoa học nói chung) là tìm hiểu những mối tương quan giữa các yếu tố lâm sàng hay các yếu tố cận lâm sàng, và qua đó có thể tiên lượng một yếu tố phụ thuộc từ các yếu tố độc lập. “Mối tương quan” ở đây bao gồm các đặc điểm như mức độ tương quan (degree of correlation) và xây dựng một mô hình tiên đoán. Mô hình ở đây chính là hàm số nối kết hai biến với nhau, và hàm số này phải có độ tin cậy nhất định và có ý nghĩa sinh học để có thể ứng dụng trong lâm sàng.

Chẳng hạn như tìm hiểu mối liên hệ giữa độ tuổi và mật độ xương (bone mineral density, hay BMD) có nghĩa là chúng ta muốn biết mối tương quan giữa hai biến này ra sao và có thể sử dụng độ tuổi để tiên lượng mật độ xương cho một cá nhân hay không. Trong mối liên hệ này, chúng ta xác định BMD là *biến phụ thuộc* (dependent variable) và độ tuổi là *biến độc lập* (independent variable). Nói theo ngôn ngữ toán, gọi BMD là Y và độ tuổi là X , chúng ta muốn tìm hiểu độ tương quan giữa X và Y , và hàm số để mô tả mối liên hệ đơn giản này. Một trong những hàm số đơn giản nhất và có lẽ thông dụng nhất là hồi qui tuyến tính đơn biến (simple linear regression). Trong bài này, tôi sẽ lần lượt giải thích phương pháp phân tích để đạt được hai mục tiêu này.

1. Tóm lược lí thuyết

Để mô tả độ tương quan giữa hai biến, chúng ta cần phải ước tính *hệ số tương quan* (coefficient of correlation). Và, để hiểu “cơ chế” của hệ số tương quan, chúng ta cần làm quen với khái niệm *hiệp biến* (covariance). Chúng ta biết rằng với một biến X hay Y , có ba thông số thống kê mô tả: số cỡ mẫu, số trung (mean), và phương sai (variance), mà tôi đã bàn qua trong bài [Lâm sàng thống kê thứ nhất](#). Nhưng để mô tả mối tương quan giữa hai biến X và Y , chúng ta cần đến hiệp biến.

Có thể hiểu hiệp biến qua hình học lượng giác như sau. Chúng ta biết rằng cho một tam giác vuông, nếu gọi cạnh huyền là c và hai cạnh còn lại là a và b , Định lí Pythagoras cho biết bình phương cạnh huyền bằng tổng bình phương hai cạnh kia:

$$c^2 = a^2 + b^2$$

Nhưng cho một tam giác thường, thì mối liên hệ giữa c và hai cạnh a và b phức tạp hơn với mối liên hệ được định lượng bằng hàm cosine của góc C như sau:

$$c^2 = a^2 + b^2 - 2ab \cdot \cos C$$

Tương tự như vậy, cho hai biến X và Y , và nếu hai biến này hoàn toàn độc lập với nhau, chúng ta có thể phát biểu rằng phương sai của biến $X + Y$ bằng phương sai của X cộng với phương sai của Y :

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

trong đó, “var” là viết tắt của phương sai (tức variance). Chú ý rằng $X+Y$ là một biến mới. Chúng ta cũng chú ý rằng công thức này tương đương với Định lý Pythagoras cho tam giác vuông.

Nếu hai biến X và Y có tương quan nhau, thì công thức trên được thay thế bằng một công thức khác với hiệp biến:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \times \text{Cov}(X, Y)$$

trong đó, “Cov” là viết tắt của hiệp biến (tức covariance). Chúng ta chú ý rằng công thức này tương đương với công thức của tam giác thường, và cũng chú ý rằng công thức trên giống như nhị thức $(x + y)^2 = x^2 + y^2 + 2xy$.

Trên đây là khái niệm. Bây giờ để đi vào chi tiết toán, chúng ta cần một số kí hiệu để viết tắt các chỉ số trên. Gọi x_i và y_i là hai biến quan sát được của X và Y cho cá nhân i . Giả sử chúng ta có n đối tượng thì $i = 1, 2, 3, \dots, n$. Gọi $\bar{x}$ và $\bar{y}$ là hai số trung bình của biến quan sát được x và y ; s_x^2 và s_y^2 lần lượt là phương sai của hai biến, được định nghĩa như sau:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

và

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

Do đó, nếu X và Y độc lập, chúng ta có thể viết:

$$s_{x+y}^2 = s_x^2 + s_y^2 \quad [1]$$

Nhưng nếu X và Y có liên hệ với nhau, công thức trên không đáp ứng được vấn đề mô tả. Chúng ta cần tìm một chỉ số khác mô tả mối liên hệ giữa hai biến, bằng cách nhân độ lệch của biến x từ số trung bình, $(x_i - \bar{x})$, cho độ lệch của biến y , $(y_i - \bar{y})$, thay vì bình phương độ lệch từng biến riêng lẻ như công thức [1]. Nói cách khác, tích số hai độ lệch chính là hiệp biến. Đối với mỗi cá nhân, hiệp biến là:

$$\text{cov}(x_i, y_i) = (x_i - \bar{x})(y_i - \bar{y})$$

Nhưng ở đây chúng ta có n đối tượng, cho nên cần phải cộng tất cả lại và chia cho số đối tượng:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad [2]$$

Công thức [2] chính là định nghĩa của hiệp biến. Từ hai công thức trên, chúng ta có thể rút ra vài nhận xét sơ khởi:

- Phương sai lúc nào cũng là số dương, bởi vì chúng được tính toán từ bình phương, nhưng hiệp biến có thể âm mà cũng có thể dương vì được ước tính từ tích của hai độ lệch.
- Một hiệp biến là số dương có nghĩa là độ lệch từ số trung bình của x tuân theo chiều hướng thuận với y .
- Một hiệp biến là số âm có nghĩa là độ lệch từ số trung bình của x tuân theo chiều hướng nghịch với y .
- Nếu hiệp biến là 0, thì hai biến x và y độc lập nhau, tức không có tương quan gì với nhau.

Một cách để “chuẩn hóa” hiệp biến và phương sai là lấy tỉ số của hai chỉ số này, và đó chính là định nghĩa của *hệ số tương quan*. Hệ số tương quan thường được kí hiệu bằng r :

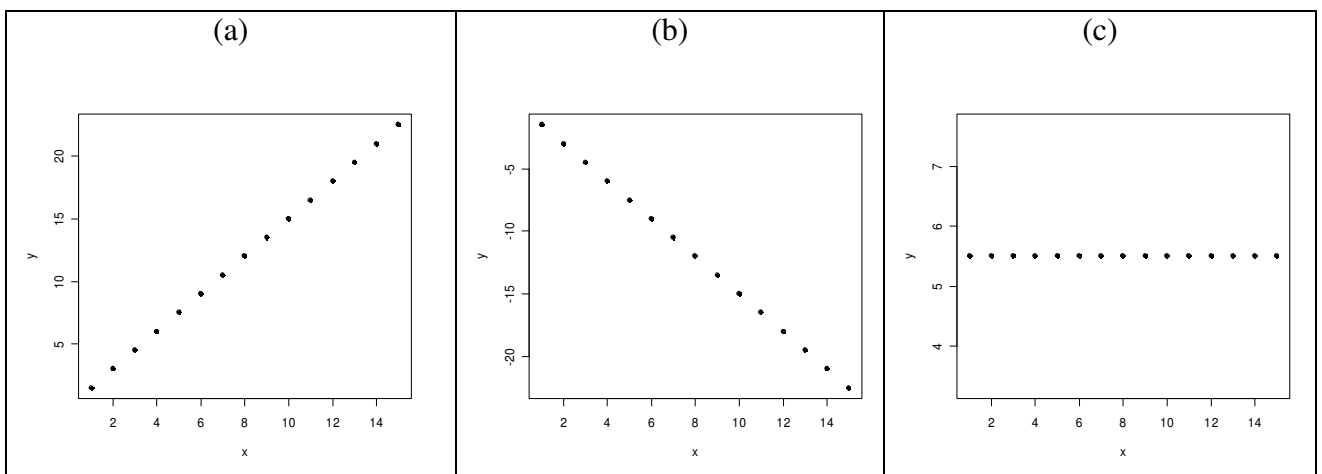
$$r = \frac{Cov(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} = \frac{Cov(x, y)}{s_x \times s_y} \quad [3]$$

(Chú ý rằng căn số bậc hai của phương sai là độ lệch chuẩn, tức là: $s_x = \sqrt{s_x^2}$ và $s_y = \sqrt{s_y^2}$, cho nên công thức trên được mô tả bằng độ lệch chuẩn, thay vì phương sai). Với vài thao tác đại số, có thể viết lại công thức [3] như sau:

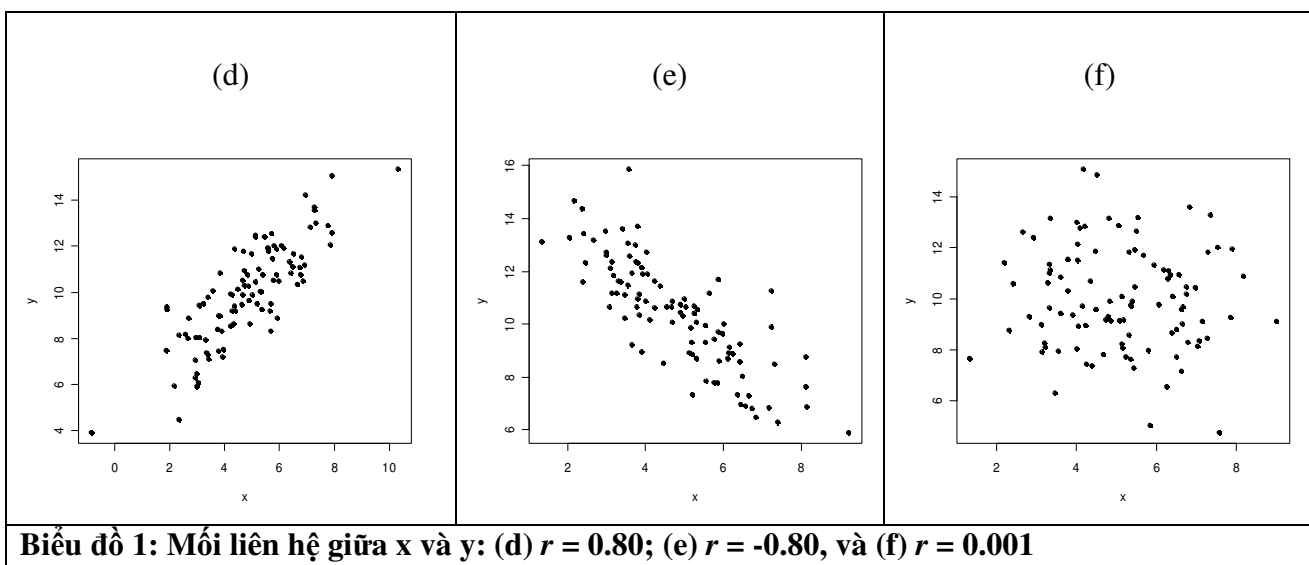
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{(n-1) s_x s_y} \quad [4]$$

Công thức còn được biết đến như là hệ số Pearson (Pearson's correlation coefficient) để ghi nhận cống hiến của nhà thống kê học nổi tiếng Karl Pearson, người đầu tiên phát triển lý thuyết về tương quan vào đầu thế kỉ 20.

Nếu giá trị của r là dương, hai biến x và y cùng biến thiên theo một hướng; nếu giá trị của r là âm, x và y liên hệ đảo ngược: tức khi x tăng thì y giảm, và ngược lại. Nếu $r = 1$ hay $r = -1$ (Biểu đồ 1a và 1b), mối liên hệ của y và x được hoàn toàn xác định; có nghĩa là cho bất cứ giá trị nào của x , chúng ta có thể xác định giá trị của y . Nếu $r = 0$ (Biểu đồ 1c), hai biến x và y hoàn toàn độc lập, tức không có liên hệ với nhau.



Biểu đồ 1: Mối liên hệ giữa x và y : (a) $r = 1$, (b) $r = -1$, và (c) $r = 0$ (độc lập).



Tất nhiên, trong thực tế khoa học thực nghiệm, ít khi nào chúng ta có những mối liên hệ xác định như vừa trình bày. Vì sai số trong đo lường, vì các lí do dao động sinh học, mối liên hệ giữa x và y thường dao động cao hơn -1 và thấp hơn 1 , như Biểu đồ 1d, 1e và 1f.

Vấn đề đặt ra là diễn dịch ý nghĩa của hệ số tương quan như thế nào? Có thể xem hệ số tương quan như là một “hệ số ảnh hưởng” (effect size). Nếu hệ số ảnh hưởng càng cao, thì mối liên hệ có ý nghĩa lâm sàng thực tế. Tuy nhiên, vì ý nghĩa lâm sàng còn tùy thuộc vào bộ môn khoa học. Chẳng hạn như đối với các bộ môn khoa học đòi hỏi độ chính xác cao, hệ số tương quan phải trên 0.8 mới có thể xem là “có ý nghĩa”; nhưng đối với các bộ môn khoa học lâm sàng và y tế công cộng, một hệ số tương quan 0.6 cũng có thể là có ý nghĩa. **Bảng 2** sau đây là những qui ước chung về cách diễn dịch hệ số tương quan trong lâm sàng và y tế công cộng.

Bảng 2. Ý nghĩa của hệ số tương quan

Hệ số tương quan	Ý nghĩa
± 0.01 đến ± 0.1	Mối tương quan quá thấp, không đáng kể
± 0.2 đến ± 0.3	Mối tương quan thấp
± 0.4 đến ± 0.5	Mối tương quan trung bình
± 0.6 đến ± 0.7	Mối tương quan cao
± 0.8 trở lên	Mối tương quan rất cao

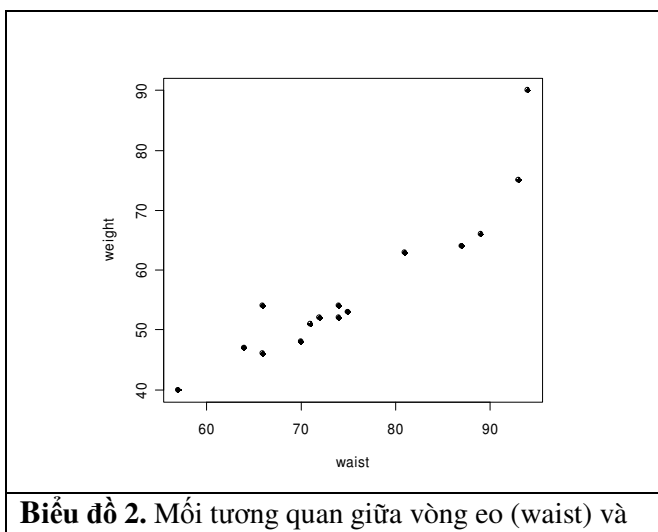
Cần nhấn mạnh một lần nữa, đây chỉ là những giá trị tham chiếu, nó không có nghĩa là những “tiêu chuẩn vàng” để ứng dụng.

2. Một ví dụ cụ thể

Ví dụ 1 – Cân nặng và vòng eo. Số liệu sau đây được trích ra từ một nghiên cứu qui mô (trên 3000 người) ở Việt Nam về mối liên hệ giữa các chỉ số nhân trắc và bệnh tiểu đường. Trọng lượng và vòng eo của 15 đối tượng được đo lường và kết quả như sau:

Trọng lượng (weight; kg)	Vòng eo (waist; cm)
51.0	71.0
66.0	89.0
47.0	64.0
54.0	74.0
64.0	87.0
75.0	93.0
54.0	66.0
52.0	74.0
53.0	75.0
52.0	72.0
48.0	70.0
46.0	66.0
63.0	81.0
40.0	57.0
90.0	94.0

Chú ý rằng cân nặng được tính bằng kg và vòng eo bằng cm. **Biểu đồ 2** sau đây thể hiện mối liên hệ giữa hai biến:



cân nặng (weight) ở 15 đối tượng người Việt được chọn ngẫu nhiên

Áp dụng công thức [1] trên chúng ta có thể mô tả hai biến này qua các chỉ số thống kê như sau:

$$\text{Trọng lượng trung bình: } \bar{x} = \frac{51+66+47+\dots+90}{15} = 57.0 \text{ kg}$$

$$\text{Phương sai của trọng lượng: } s_x^2 = \frac{(51-57)^2 + (66-57)^2 + (47-57)^2 + \dots + (90-57)^2}{15-1} = 163.6$$

$$\text{Vòng eo trung bình: } \bar{y} = \frac{71+89+64+\dots+94}{15} = 75.5 \text{ cm}$$

$$\text{Phương sai vòng eo: } s_y^2 = \frac{(71-75.5)^2 + (89-75.5)^2 + (64-75.5)^2 + \dots + (94-75.5)^2}{15-1} = 122.6$$

Và hiệp biến:

$$\text{Cov}(x, y) = \frac{(51-57) \times (71-75.5) + (66-57) \times (64-75.5) + \dots + (90-57) \times (94-75.5)}{14} = 71.2$$

Tóm lại:

- $n = 15$
- Trọng lượng trung bình: $\bar{x} = 57.0$, độ lệch chuẩn: $s_x = 12.8 \text{ kg}$
- Vòng eo trung bình: $\bar{y} = 75.5$, độ lệch chuẩn: $s_y = 11.1 \text{ cm}$
- Hiệp biến của hai trọng lượng và vòng eo: $\text{Cov}(x, y) = 71.2$

Do đó, hệ số tương quan giữa trọng lượng và vòng eo (theo công thức [3]) là:

$$r = \frac{\text{Cov}(x, y)}{s_x \times s_y} = \frac{71.2}{12.8 \times 11.1} = 0.92$$

Dựa vào qui ước vừa đề cập trong phần trên, chúng ta có thể nói trong nhóm đối tượng này, mối tương quan giữa cân nặng và vòng eo rất cao. Nếu mối tương quan này được lặp lại ở một hay nhiều nhóm đối tượng khác, có thể sử dụng vòng eo để tiên đoán trọng lượng.

3. Khoảng tin cậy 95% của hệ số tương quan

Cũng như các thông số thống kê khác như số trung bình và độ lệch chuẩn, hệ số tương quan cũng chịu ảnh hưởng của dao động giữa các mẫu. Do đó, chúng ta cần phải ước tính khoảng tin cậy 95% của hệ số tương quan. Xin nhắc lại rằng, chúng ta không

biết hệ số tương quan thật (tức là hệ số trong quần thể, và hãy gọi hệ số này là ρ) là bao nhiêu, nên phải sử dụng hệ số r để ước tính ρ .

Muốn ước tính khoảng tin cậy 95% của ρ , chúng ta cần phải ước tính độ lệch chuẩn của r . Li thuyết thống kê cho biết độ lệch chuẩn của r là $s_r = \frac{\sqrt{1-r^2}}{\sqrt{n-2}}$. Khó khăn ở đây, như công thức này cho thấy, là độ lệch chuẩn của r tùy thuộc vào r , tức là mất tính độc lập. Do đó, cần phải tìm một phương pháp khác sao cho khách quan hơn. Nhà thống kê học (và cũng là cha đẻ của khoa học thống kê hiện đại và cha đẻ của lý thuyết di truyền hiện đại) Ronald A. Fisher chứng minh rằng thay vì tính độ lệch chuẩn của r , có thể tính độ lệch chuẩn của một hàm số của r và sẽ đạt được mục tiêu khách quan.

Theo phương pháp của Fisher, trước hết chúng ta cần phải hoán chuyển r sang một chỉ số mới z , qua công thức sau đây:

$$z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \quad [5]$$

Và, có thể chứng minh rằng độ lệch chuẩn của z là:

$$s_z = \frac{1}{\sqrt{n-3}} \quad [6]$$

Do đó, khoảng tin cậy 95% của z là: $z \pm 1.96 \times s_z$. Tất nhiên, sau khi đã ước tính được khoảng tin cậy 95% của z , chúng ta có thể hoán chuyển ngược lại cho khoảng tin cậy 95% của ρ .

Ví dụ 1 (tiếp theo) – Cân nặng và vòng eo: Ở phần trên, chúng ta đã ước tính $r=0.92$; do đó, theo [5] chỉ số z là:

$$z = \frac{1}{2} \log \left(\frac{1+0.92}{1-0.92} \right) = 1.59$$

và với số cỡ mẫu $n = 15$, chúng ta có thể ước tính độ lệch chuẩn của z qua [6] như sau:

$$s_z = \frac{1}{\sqrt{15-3}} = 0.267$$

Khoảng tin cậy 95% của z dao động từ 1.067 đến 2.113 ($1.59 - 1.96 \times 0.267 = 1.067$ đến $1.59 + 1.96 \times 0.267 = 2.113$). Bây giờ, chúng ta có thể hoán chuyển khoảng tin cậy 95% sang hệ số r qua công thức (suy ra từ công thức [5])

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \quad [7]$$

Khi $z = 1.067$, $r = 0.79$. Khi $z = 2.113$, $r = 0.97$. Nói cách khác, hệ số tương quan giữa cân nặng và vòng eo là 0.92 với khoảng tin cậy 95% dao động từ 0.79 đến 0.97.

4. Kiểm định hai hệ số tương quan

Giả sử chúng ta có hai hệ số tương quan r_1 và r_2 , là ước số của hai hệ số ρ_1 và ρ_2 trong một quần thể. Hai hệ số r_1 và r_2 được ước tính từ hai mẫu độc lập n_1 và n_2 đối tượng. Để kiểm định giả định rằng $\rho_1 = \rho_2$ và giả định $\rho_1 \neq \rho_2$, chúng ta trước hết cần phải hoán chuyển r thành chỉ số z :

$$z_1 = \frac{1}{2} \log \left(\frac{1+r_1}{1-r_1} \right) \quad \text{và} \quad z_2 = \frac{1}{2} \log \left(\frac{1+r_2}{1-r_2} \right)$$

Gọi $d = z_1 - z_2$, chúng ta có thể chứng minh rằng phương sai của d là:

$$s_d^2 = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \quad [8]$$

Hay, nói cách khác, độ lệch chuẩn của d là:

$$s_d = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \quad [9]$$

Và kiểm định cho giả thuyết $\rho_1 = \rho_2$ có thể tính toán chỉ số t như sau:

$$t = \frac{d}{s_d} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \quad [10]$$

Có thể chứng minh rằng nếu giả thuyết $\rho_1 = \rho_2$ là đúng thì t tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Điều này có nghĩa là nếu giá trị của t thấp hơn -2 hay cao hơn +2, chúng ta có thể nói hai hệ số tương quan khác nhau có ý nghĩa thống kê.

Ví dụ 2 – So sánh hai hệ số tương quan. Giả sử chúng ta tìm thấy trong y văn ở Mỹ, hệ số tương quan giữa cân nặng và vòng eo là 0.7, và hệ số này được ước tính ở 1000 đối tượng. Những với 15 đối tượng, chúng ta có hệ số 0.92. Câu hỏi đặt ra là độ khác biệt giữa hai hệ số này có ý nghĩa thống kê hay chỉ do các yếu tố ngẫu nhiên gây nên.

- Với hệ số $r_1 = 0.7$, chúng ta có $z = \frac{1}{2} \log \left(\frac{1.7}{0.3} \right) = 0.867$, và phương sai

$$s_1 = \frac{1}{n-3} = \frac{1}{1000-3} 0.001.$$

- Với hệ số $r_2 = 0.92$, chúng ta có $z = 1.59$, và phương sai $s_1 = \frac{1}{15-3} = 0.0833$.

- Do đó, độ khác biệt là: $d = 0.867 - 1.59 = -0.723$. Phương sai của d là: $s_d^2 = 0.001 + 0.0833 = 0.0843$. Do đó, chỉ số t là:

$$t = \frac{-0.723}{\sqrt{0.0843}} = -2.49$$

Vì chỉ số t thấp hơn -2.0, chúng ta có bằng chứng để phát biểu rằng hệ số tương quan ở người Việt cao hơn ở người Mỹ, và độ khác biệt đó có ý nghĩa thống kê.

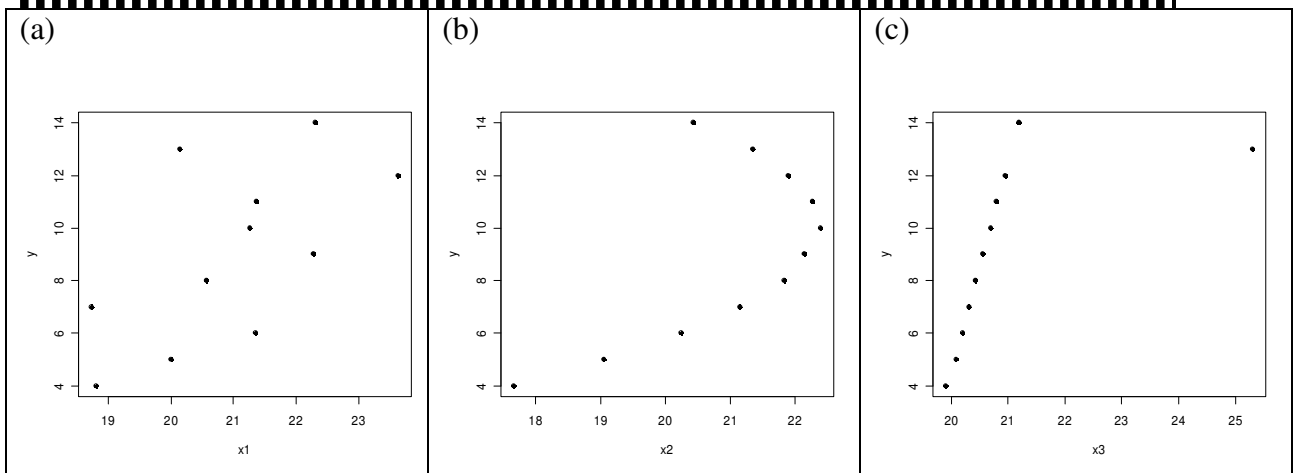
5. Tóm lược

Qua trình bày trên, chúng ta đã biết qua lí thuyết và phương pháp tính cũng như kiểm định một hệ số tương quan (và so sánh hai hệ số tương quan). Cần phải nói thêm rằng các kết quả ước tính vừa trình bày trên chỉ có ý nghĩa nếu hai biến x và y tuân theo luật phân phối chuẩn (tức normal distribution). Nếu giả định phân phối chuẩn không đáp ứng, các kết quả so sánh có thể (nhấn mạnh: “có thể”) không còn giá trị. Tuy nhiên, trong thực tế, với số lượng cỡ mẫu lớn, ngay cả lệch từ luật phân phối chuẩn cũng không có ảnh hưởng lớn đến kết quả.

.....

Cần lưu ý rằng khi nói đến tương quan ở đây, chúng ta nói đến tương quan tuyến tính – tương quan theo luật đường thẳng. Do đó, kết quả phân tích sẽ không có ý nghĩa khi mỗi tương quan không tuân theo luật tuyến tính. Điều này quan trọng, vì trước khi tính toán hệ số tương quan, chúng ta nên vẽ biểu đồ xem mối liên hệ giữa hai biến có tuân theo luật tương quan tuyến tính hay không. Nếu không, có lẽ chúng ta cần phải hoán chuyển số liệu trước khi phân tích. Không có phân tích nào có hữu hiệu hơn là biểu đồ. Đơn thuần dựa vào con số rất có thể sai lầm.

Biểu đồ sau đây cho thấy ba mối tương quan đều có hệ số tương quan như nhau: $r = 0.62$. Nhưng chỉ có Biểu đồ (a) là có ý nghĩa bởi vì mối liên hệ giữa x và y là tuyến tính, còn mối liên hệ (b) và (c) thì rất khác.



Một điều cực kì quan trọng cần nằm lòng khi diễn dịch kết quả phân tích tương quan là hệ số tương quan **không hẳn** phản ánh mối quan hệ nhân quả (cause-and-effect relationship). Nếu không có lí do sinh học, không thể và không nên diễn dịch hệ số tương quan theo định hướng nguyên nhân và hệ quả. Chẳng hạn như mối tương quan giữa trọng lượng và vòng eo như vừa phân tích trên rất cao (0.92), nhưng điều này không có nghĩa là trọng lượng là nguyên nhân làm cho người ta có vòng eo rộng, hay vòng eo rộng là nguyên nhân làm cho người ta cân nặng. Mối liên hệ mà chúng ta quan sát chỉ đơn thuần là tương quan. Việc sử dụng và diễn dịch mối tương quan đó trong bối cảnh lâm sàng ra sao còn tùy thuộc vào kinh nghiệm và kiến thức của từng bộ môn khoa học.

Thuật ngữ sử dụng trong bài viết

Tiếng Việt	Tiếng Anh
Hệ số tương quan	Coefficient of correlation
Hiệp biến	Covariance
Khoảng tin cậy 95%	95% confidence interval
Phương sai	Variance
Độ lệch chuẩn	Standard deviation
Biến phụ thuộc	Dependent variable
Biến độc lập	Independent variable

Chú thích kĩ thuật: Các mã R sau đây đã được sử dụng cho phân tích vừa trình bày.

Mô phỏng cho biểu đồ 1d

```
# trước hết tạo ra 2 dãy số với 100 đối tượng ảo
# tuân theo luật phân phối chuẩn bằng cách sử dụng hàm rnorm
zn1 <- rnorm(100)
zn2 <- rnorm(100)
# muốn có hệ số tương quan = 0.8
r = 0.80
# mô phỏng x với trung bình 5 và độ lệch chuẩn 1.5
x <- 5 + 1.5*zn1
# mô phỏng y với trung bình 10 và độ lệch chuẩn 2 và r=0.80
y <- 10 + r*2.0*zn1 + 2.0*sqrt(1-r^2)*zn2
# vẽ biểu đồ
plot(y ~ x, pch=16)
```

Mô phỏng cho biểu đồ 1e, lần này hệ số tương quan = -0.8

```
zn1 <- rnorm(100)
zn2 <- rnorm(100)
r = -0.80
x <- 5 + 1.5*zn1
y <- 10 + r*2.0*zn1 + 2.0*sqrt(1-r^2)*zn2
plot(y ~ x, pch=16)
```

Mô phỏng cho biểu đồ 1f - hệ số tương quan = 0.001

```
zn1 <- rnorm(100)
zn2 <- rnorm(100)
r = 0.001
x <- 5 + 1.5*zn1
y <- 10 + r*2.0*zn1 + 2.0*sqrt(1-r^2)*zn2
plot(y ~ x, pch=16)
```

nhập số liệu trọng lượng và vòng eo

```
weight <- c(51,66,47,54,64,75,54,52,53,52,48,46,63,40,90)
waist <- c(71,89,64,74,87,93,66,74,75,72,70,66,81,57,94)
```

vẽ biểu đồ 2

```
plot(weight ~ waist, pch=16)
```

tính số trung bình, dùng hàm mean

```

mean(weight)
mean(waist)

# tính độ lệch chuẩn, dùng hàm sd
sd(weight)
sd(waist)

# tính hiệp biến và r
cov(weight, waist)
cov(weight, waist) / (sd(weight)*sd(waist))

# hay có thể sử dụng hàm cor trong R
cor(weight, waist)

# kiểm định r và khoảng tin cậy 95%
cor.test(weight, waist)

# vẽ biểu đồ trong phần "tóm lược"
y <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
x1 <- c(21.26, 20.57, 20.15, 22.28, 21.36, 22.30, 21.35, 18.81, 23.63,
        18.73, 20.01)
x2 <- c(22.40, 21.84, 21.35, 22.15, 22.27, 20.43, 20.25, 17.66, 21.90,
        21.15, 19.05)
x3 <- c(20.70, 20.43, 25.31, 20.56, 20.80, 21.19, 20.21, 19.90, 20.95,
        20.31, 20.09)

plot(y ~ x1, pch=16)
plot(y ~ x2, pch=16)
plot(y ~ x3, pch=16)

```