Xác Suất Thống Kê

Ngày 19 tháng 5 năm 2014

1. Hệ số tương quan giữa X và Y

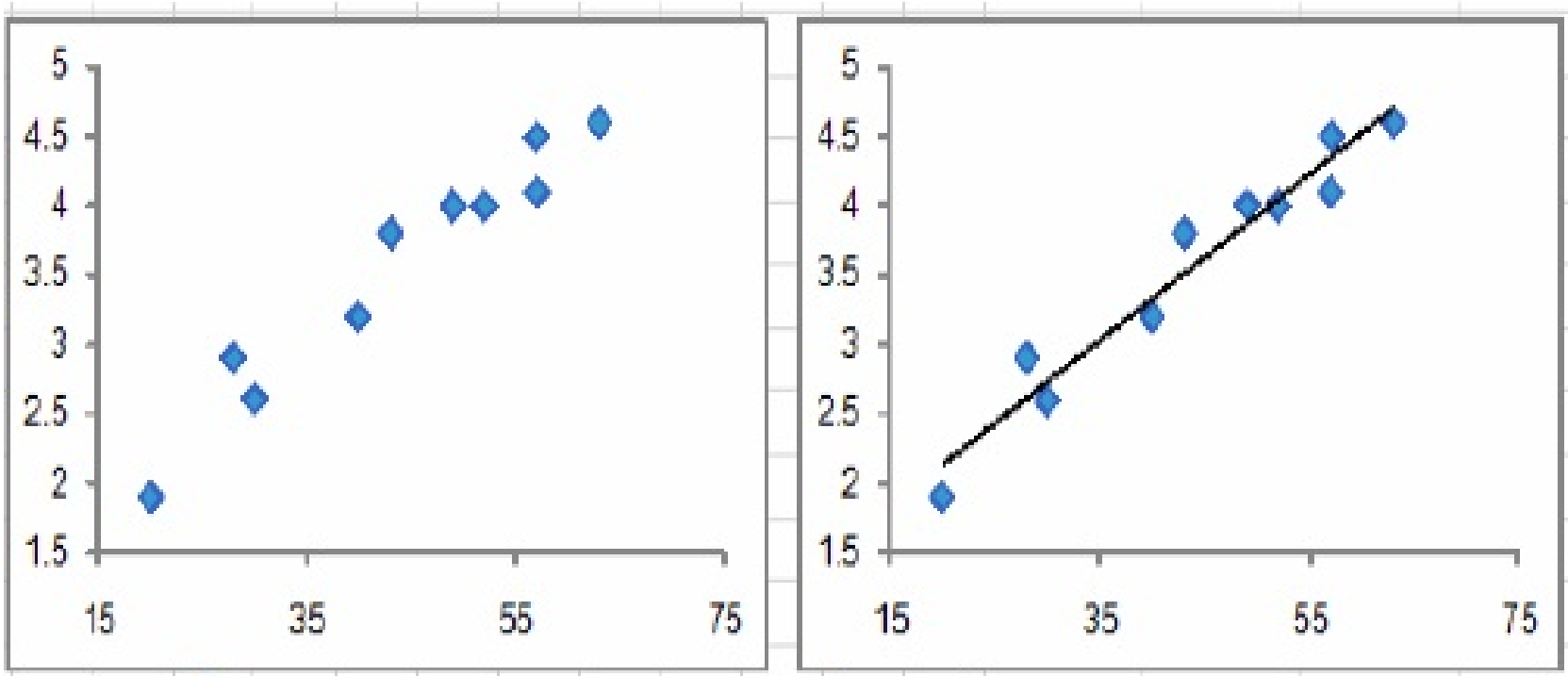
• Để minh họa cho vấn đề, chúng ta thử xem xét nghiên cứu sau đây mà trong đó nhà nghiên cứu đo lường độ cholesterol (Y) trong máu của 10 đối tượng nam ở độ tuổi (X).

Kết quả đo lường như sau:

X	20	52	30	57	28
Y	1,9	4,0	2,6	4,5	2,9

X	43	57	63	40	49
Y	3,8	4,1	4,6	3,2	4,0

Biểu đồ liên hệ giữa độ tuổi và độ cholesterol:



Biểu đồ trên đây gợi ý cho thấy mối liên hệ giữa độ tuổi (X) và cholesterol (Y) là một đường thẳng (tuyến tính).

• Để "đo lường" mối liên hệ này, chúng ta có thể sử dụng hệ số tương quan:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\widehat{s}_x^2.\widehat{s}_y^2}}.$$

Trong đó $\overline{xy} = \frac{1}{n}\sum_{\substack{i=1\\j=1}} n_{ij}x_iy_i$, $n = \sum n_{ij}$.

Chú ý. $\sqrt{\widehat{s}_x^2.\widehat{s}_y^2}$ có sai số bé hơn $\widehat{s}_x.\widehat{s}_y$.

Ý nghĩa

• Hệ số tương quan đo mối *quan hệ tuyến tính* giữa x, y.

1) $-1 \leq r_{xy} \leq 1$.

2) Nếu $r_{xy} = 0$ thì hai biến số không có quan hệ tuyến tính; nếu $r_{xy} = \pm 1$ thì hai biến số có quan hệ tuyến tính tuyệt đối.

3) Nếu $r_{xy} < 0$ thì quan hệ giữa x, y là giảm biến (có nghĩa là khi x tăng thì y giảm).

4) Nếu $r_{xy} > 0$ thì quan hệ giữa x, y là đồng biến (có nghĩa là khi x tăng thì y cũng tăng).

VD 1. Tính hệ số tương quan giữa độ tuổi và cholesterol cho ở bảng trên. Ta có:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 43,9\,;\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 3,56\,;$$

$$\overline{xy} = \frac{1}{n}\sum_{\substack{i=1\\ j=1}} n_{ij} x_i y_i = 167,26\,;$$

$$\widehat{s}_x^2 = 183,29\,;\ \widehat{s}_y^2 = 0,6944\,.$$

$$\text{Vậy } r_{xy} = \frac{\overline{xy} - \bar{x}.\bar{y}}{\sqrt{\widehat{s}_x^2 . \widehat{s}_y^2}} = 0,9729\,.$$

2. Đường thẳng hồi qui

• Để tiện việc theo dõi và mô tả mô hình, gọi độ tuổi cho cá nhân i là x_i và cholesterol là y_i, $i = \overline{1,10}$.

– Các điểm có tọa độ $(x_i; y_i)$ tạo thành đường gấp khúc và gần với đường thẳng có dạng $y = ax + b$. Người ta dùng đường thẳng $y = ax + b$ để tính xấp xỉ các giá trị y_i theo x_i: $y_i = ax_i + b + \varepsilon_i$ với một sai số ε_i, đường thẳng này được gọi là đường thẳng hồi quy.