Tên học viên : Đỗ Văn Nhân, K24 - xác suất thống kê

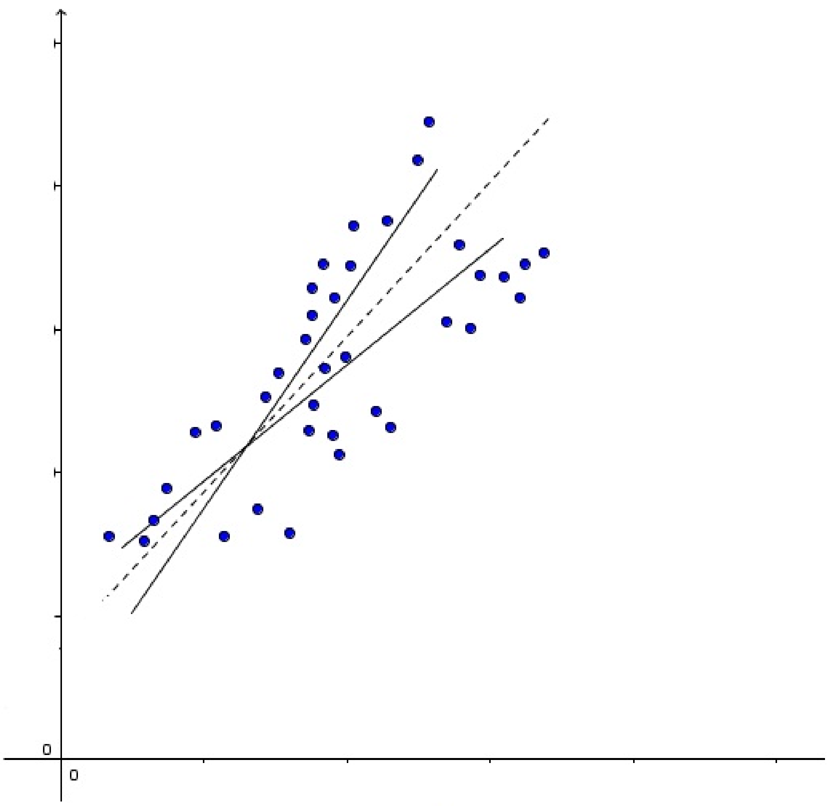
Môn : Thống kê toán nâng cao

Giảng Viên : GS - TS Nguyễn Bác Văn

Bài tập. Hình dưới đây biểu diễn phân phối của giá khoai tây vào tháng chạp năm 1936 (x) và tháng chạp năm 1937 (y) tại 46 vùng khác nhau ở Thụy Điển theo số liệu thống kê chính thức.

Những đặc trưng thông thường của mẫu là

$$\overline{x} = 660,57;\ \overline{y} = 120,91;\ s_1 = 106,86;\ s_2 = 120,91$$
$$r = 0,7928;\ b_{1\ 2} = 0,7007;\ b_{2\ 1} = 0,8971;\ n = 46$$



Giá khoai tây tại 46 vùng ở Thụy Điển tháng chạp năm 1936 (x) và tháng chạp 1937 (y).
Các đường hồi quy và đường hồi quy trực giao (đường nét đứt).

Giả sử các giá trị (x, y) lập thành một mẫu từ họ chuẩn và ta muốn thu được các tin tức về giá trị chưa biết của hệ số hồi quy β_{21} và hệ số tương quan ρ của họ này.

Ta có

$\overline{x}, \overline{y}$ lần lượt là trung bình giá khoai tây tại 46 vùng vào tháng 12/1936 và 12/1937.

$s_1,\ s_2$ lần lượt là độ lệch tiêu chuẩn giá khoai vào tháng 12/1936 và 12/1937.

$r = \dfrac{m_{1\ 1}}{s_1 s_2} = \dfrac{m_{1\ 1}}{\sqrt{m_{2\ 0}}\sqrt{m_{0\ 2}}}$ là hệ số tương quan của mẫu với

$$m_{1\ 1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\ ;\ m_{2\ 0} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ và } m_{0\ 2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

$b_{12} = \dfrac{m_{1\ 1}}{m_{2\ 0}} = \dfrac{rs_2}{s_1}$ và $b_{21} = \dfrac{m_{1\ 1}}{m_{0\ 2}} = \dfrac{rs_1}{s_2}$ tương ứng là các hệ số hồi quy của mẫu.

Đặt

$$t = \frac{s_1\sqrt{n-2}}{s_2\sqrt{1-r^2}}\left(b_{2\ 1} - \beta_{2\ 1}\right)$$

thì t sẽ có phân phối Student với $n-2$ bậc tự do

(theo kết quả đã được chứng minh trong Bartlett [54]) .

Tiếp theo, với độ tin cậy $1-\alpha$ cho trước, ta có

$$\begin{aligned}
P\left(|t| \le t^{(n-2)}_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\
\Rightarrow P\left(-t^{(n-2)}_{1-\frac{\alpha}{2}} \le \frac{s_1\sqrt{n-2}}{s_2\sqrt{1-r^2}}\left(b_{2\ 1} - \beta_{2\ 1}\right) \le t^{(n-2)}_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\
\Rightarrow P\left(-t^{(n-2)}_{1-\frac{\alpha}{2}}\frac{s_2\sqrt{1-r^2}}{s_1\sqrt{n-2}} \le b_{2\ 1} - \beta_{2\ 1} \le t^{(n-2)}_{1-\frac{\alpha}{2}}\frac{s_2\sqrt{1-r^2}}{s_1\sqrt{n-2}}\right) &= 1-\alpha \\
\Rightarrow P\left(b_{2\ 1} - t^{(n-2)}_{1-\frac{\alpha}{2}}\frac{s_2\sqrt{1-r^2}}{s_1\sqrt{n-2}} \le \beta_{2\ 1} \le b_{2\ 1} + t^{(n-2)}_{1-\frac{\alpha}{2}}\frac{s_2\sqrt{1-r^2}}{s_1\sqrt{n-2}}\right) &= 1-\alpha
\end{aligned}$$

hay β_{21} có khoảng tin cậy tương tứng với độ tin cậy $1-\alpha$ là

$$\left[b_{2\ 1} - \frac{s_2\sqrt{1-r^2}}{s_1\sqrt{n-2}}t^{(n-2)}_{1-\frac{\alpha}{2}};\ b_{2\ 1} + \frac{s_2\sqrt{1-r^2}}{s_1\sqrt{n-2}}t^{(n-2)}_{1-\frac{\alpha}{2}}\right]$$

trong đó $t^{(n-2)}_{1-\frac{\alpha}{2}}$ là phân vị mức $1-\dfrac{\alpha}{2}$ của phân phối Student với $n-2$ bậc tự do.
Thay giá trị $b_{2\ 1};\ s_2;\ s_1;\ r$ và n ta được

$$\beta_{2\ 1} \in \left[0,8971 - 0,1039.t^{(n-2)}_{1-\frac{\alpha}{2}},\ 0,8971 + 0,1039.t^{(n-2)}_{1-\frac{\alpha}{2}}\right].$$

Cụ thể với các giá trị của $\alpha = \{0,05;\ 0,01;\ 0,001\}$ ta lần lượt có được khoảng tin cậy cho $\beta_{2\ 1}$ như sau :

$$\begin{aligned}
\alpha &= 5\%\ ,\ t^{(n-2)}_{1-\frac{\alpha}{2}} = 2,0208 \quad \text{và} \quad \beta_{2\ 1} \in [0,687;\ 1,107] \\
\alpha &= 1\%\ ,\ t^{(n-2)}_{1-\frac{\alpha}{2}} = 2,6942 \quad \text{và} \quad \beta_{2\ 1} \in [0,617;\ 1,177] \\
\alpha &= 0,1\%,\ t^{(n-2)}_{1-\frac{\alpha}{2}} = 3,5311 \quad \text{và} \quad \beta_{2\ 1} \in [0,530;\ 1,264]\,.
\end{aligned}$$

Tiếp theo, xét hệ số tương quan mẫu r. Theo công thức (27.8.1)

$$\begin{aligned}\mathrm{Var}\,(r) &= \mathrm{Var}\left(\frac{m_{1\ 1}}{\sqrt{m_{2\ 0}}\sqrt{m_{0\ 2}}}\right)\\ &= \frac{\rho^2}{4n}\left(\frac{\mu_{4\ 0}}{\mu_{2\ 0}^2}+\frac{\mu_{0\ 4}}{\mu_{02}^2}+\frac{2\mu_{2\ 2}}{\mu_{2\ 0}.\mu_{0\ 2}}+\frac{4\mu_{2\ 2}}{\mu_{1\ 1}^2}-\frac{4\mu_{3\ 1}}{\mu_{1\ 1}.\mu_{2\ 0}}-\frac{4\mu_{1\ 3}}{\mu_{1\ 1}.\mu_{0\ 2}}\right)\end{aligned}$$

trong trường hợp họ chuẩn, biểu thức trở thành

$$\mathrm{Var}\,(r) = \frac{\left(1-\rho^2\right)^2}{n}$$

ta được giá trị xấp xỉ của ρ và sai số tiêu chuẩn

$$\sqrt{\mathrm{Var}\,(r)} = \frac{1-r^2}{\sqrt{n}} = \frac{1-0,7928^2}{\sqrt{46}} \approx 0,0548.$$

Nếu phân phối mẫu của r khá gần phân phối chuẩn thì phân phối này có thể sử dụng để làm giả thiết

$$H:\rho = c \qquad \text{với } c \text{ là hằng số đã cho nào đó.}$$

Từ công thức (29.7.2) xét hàm MĐXS của r là

$$f_n\,(r) = \frac{n-2}{\pi}\left(1-\rho^2\right)^{\frac{n-1}{2}}\left(1-r^2\right)^{\frac{n-4}{2}}\int_0^1 \frac{x^{n-2}}{\left(1-\rho r x\right)^{n-1}}\frac{dx}{\sqrt{1-x^2}}$$

Khi $\rho \neq 0$ thì tốc độ hội tụ hàm phân phối mẫu của r đến phân phối chuẩn khá chậm. Khi $n = 46$, vì các kết quả nhận được nhờ sai số tiêu chuẩn không chính xác. Do đó chúng ta sử dụng phép biến đổi của Fisher (29.7.3) với

$$U = \frac{1}{2}\log\frac{1+r}{1-r} \text{ và } \zeta = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$$

Theo công thức (29.7.4), ta được

$$E\,(U) = E\left(\frac{1}{2}\log\frac{1+r}{1-r}\right) = \zeta + \frac{\rho}{2\,(n-1)} = \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2\,(n-1)} \text{ và } \mathrm{Var}\,(U) = \frac{1}{n-3}$$

Đặt

$$Z = \frac{U - E\,(U)}{\sqrt{\mathrm{Var}\,(U)}}$$

thì Z sẽ có phân phối chuẩn $\mathcal{N}\,(0,1)$. Ta có

$$Z = \frac{\frac{1}{2}\log\frac{1+r}{1-r} - \left(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}\right)}{\sqrt{\frac{1}{n-3}}} = \sqrt{n-3}\left[\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) - \left(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2\,(n-1)}\right)\right]$$

Vậy với độ tin cậy $1-\alpha$, ta xem số liệu là phù hợp với giá trị giả thiết nào đó của ρ, ta có

$$\begin{aligned}
&P\left(|Z| \le z_{1-\frac{\alpha}{2}}\right) &&= 1-\alpha\\
\Rightarrow\ &P\left\{-z_{1-\frac{\alpha}{2}} \le \sqrt{n-3}\left[\frac{1}{2}\log\frac{1+r}{1-r} - \left(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}\right)\right] \le z_{1-\frac{\alpha}{2}}\right\} &&= 1-\alpha\\
\Rightarrow\ &P\left\{\frac{-z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \le \frac{1}{2}\log\frac{1+r}{1-r} - \left(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}\right) \le \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right\} &&= 1-\alpha\\
\Rightarrow\ &P\left\{-\frac{1}{2}\log\frac{1+r}{1-r} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \le -\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}\right) \le -\frac{1}{2}\log\frac{1+r}{1-r} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right\} &&= 1-\alpha\\
\Rightarrow\ &P\left\{\frac{1}{2}\log\frac{1+r}{1-r} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \le \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \le \frac{1}{2}\log\frac{1+r}{1-r} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right\} &&= 1-\alpha
\end{aligned}$$

trong đó $z_{1-\frac{\alpha}{2}}$ là phân vị mức $1-\dfrac{\alpha}{2}$ của phân phối chuẩn $\mathcal{N}(0,1)$.

Vậy với mỗi độ tin cậy $1-\alpha$ cho trước thì khoảng tin cậy cho

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

sẽ là

$$\left[\frac{1}{2}\log\frac{1+r}{1-r} - \frac{z^{(n-2)}_{1-\frac{\alpha}{2}}}{\sqrt{n-3}};\ \frac{1}{2}\log\frac{1+r}{1-r} + \frac{z^{(n-2)}_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right]$$

thay $r = 0,7928$ và $n = 46$; ta được

$$\left[1,07892 - 0,15249.z_{1-\frac{\alpha}{2}};\ 1,07892 + 0,15249.z_{1-\frac{\alpha}{2}}\right] \qquad (\diamond)$$

Khoảng tin cậy cho ρ tương ứng với mỗi

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

sẽ được giải từ phương trình

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} = k$$

với k là hệ số xác định từ 2 biên của khoảng tin cậy $(\diamond)$.

Cụ thể nếu $\alpha = 5\%$ ta có $z_{1-\frac{\alpha}{2}} = 1,96$ và khi đó

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \in [0,78003;\ 1,37782]$$

Giải bất phương trình

$$0,78003 \leq \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \leq 1,37782$$

$$\Leftrightarrow \begin{cases} \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \geq 0,78003 \;\Rightarrow \rho \geq 0,6486 \\ \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \leq 1,37782 \;\Rightarrow \rho \leq 0,8783 \end{cases}$$

ta được khoảng tin cậy cho ρ với $\alpha = 0,05$ là

$$[0,6486;\; 0,8783]$$

Tương tự với $\alpha = 1\%$ thì $z_{1-\frac{\alpha}{2}} = 2,58$; khi đó

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \in [0,68548;\; 1,47237]$$

Giải bất phương trình và ta được

$$\rho \in [0,5913;\; 0,8980]$$

Cuối cùng với $\alpha = 0,1\%$, thì $z_{1-\frac{\alpha}{2}} = 3,29$ và

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \in [0,5772;\; 1,58064]$$

Giải bất phương trình; ta được

$$\rho \in [0,5164;\; 0,9171]$$

Vậy

$$\begin{aligned} \alpha = \;\; 5\%\,, &\quad \text{và} \quad \rho \in [0,6486;\; 0,8783] \\ \alpha = \;\; 1\%\,, &\quad \text{và} \quad \rho \in [0,5913;\; 0,8980] \\ \alpha = 0,1\%, &\quad \text{và} \quad \rho \in [0,5164;\; 0,9171]\,. \end{aligned}$$