Круглов А.И. 605-202 ST 12

$\boxed{1}$ $X=(X_1,\dots,X_n)$, $Y=(Y_1,\dots,Y_n)$, $f: \mathbb{R}^2 \to \mathbb{R}$, $f(x,y)=-f(y,x)$, $\forall\, 1 \le i < j \le n$ $C_{ij}(X)=f(X_i, X_j)$,

обобщ. к-т корр. $\hat{r} = \dfrac{\sum_{i<j} C_{ij}(X)\, C_{ij}(Y)}{\sqrt{\sum_{i<j} C_{ij}(X)^2 \sum_{i<j} C_{ij}(Y)^2}}$. Выразить через $\hat{r}$ к-ты Пирсона, Спирмэна, Кэндалла.

1) Пирсон: $\hat{\rho} = \dfrac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2 \sum_{i=1}^n (Y_i-\bar{Y})^2}}$. Распишем $\sum_{i=1}^n (X_i-\bar{X})^2 = \sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)^2 = \sum_{i=1}^n \Big(X_i^2 - 2X_i \cdot \frac{1}{n} \cdot$

$\cdot \sum_{j=1}^n X_j + \frac{1}{n^2}\big(\sum_{j=1}^n X_j\big)^2\Big) = \sum_{i=1}^n X_i^2 - \frac{2}{n}\underbrace{\big(\sum_{i=1}^n X_i\big)}_{A}\underbrace{\big(\sum_{j=1}^n X_j\big)}_{A} + \frac{1}{n}\underbrace{\big(\sum_{j=1}^n X_j\big)^2}_{A} = \sum_{i=1}^n X_i^2 - \frac{1}{n}\big(\sum_{i=1}^n X_i\big)^2 = \frac{1}{n}\sum_{i<j}(X_i - X_j)^2 = \sum_{i<j} C_{ij}(X)^2$

Таким образом, $C_{ij}(X) = \dfrac{X_i - X_j}{\sqrt{n}}$, $C_{ij}(Y) = \dfrac{Y_i - Y_j}{\sqrt{n}}$. Ясно, что числитель тоже согласуется с

таким определением C_{ij}: $\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^n \big(X_i - \frac{1}{n}\sum_{j=1}^n X_j\big)\big(Y_i - \frac{1}{n}\sum_{j=1}^n Y_j\big) = \sum_{i=1}^n \Big(X_i Y_i - Y_i \cdot \frac{1}{n}\sum_{j=1}^n X_j -$

$- X_i \cdot \frac{1}{n}\sum_{j=1}^n Y_j + \frac{1}{n^2}\big(\sum_{j=1}^n X_j\big)\big(\sum_{j=1}^n Y_j\big)\Big) = \sum_{i=1}^n X_i Y_i - \frac{2}{n}\big(\sum_{i=1}^n Y_i\big)\big(\sum_{j=1}^n X_j\big) + \frac{1}{n}\big(\sum_{j=1}^n X_j\big)\big(\sum_{j=1}^n Y_j\big) = \sum_{i=1}^n X_i Y_i - \frac{1}{n}\big(\sum_{j=1}^n X_j\big)\big(\sum_{j=1}^n Y_j\big) =$

$= \frac{1}{n}\sum_{i<j}(X_i - X_j)(Y_i - Y_j) = \sum_{i<j} C_{ij}(X)\, C_{ij}(Y)$. Итог: $\hat{\rho} = \hat{r}$ при $C_{ij}(X) = \dfrac{X_i - X_j}{\sqrt{n}}$

2) Спирмен: $\rho_S = \dfrac{\sum_{i=1}^n (R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^n (R_i-\bar{R})^2 \sum_{i=1}^n (S_i-\bar{S})^2}}$, где R и S – ранги наблюдений в X и Y. Заметим, что

ρ_S – это $\hat{\rho}$, но вместо самих наблюдений подставлены ранги: $X_i \mapsto \operatorname{rank}(X_i) = j$, $X_{(j)} = X_i$

Значит, $C_{ij}(X) = \dfrac{\operatorname{rank}(X_i) - \operatorname{rank}(X_j)}{\sqrt{n}}$. Кстати, в этом и предыдущем пунктах можно было бы

взять C_{ij} без $\sqrt{n}$ в знаменателе, т.к. он сокращается при подсчёте к-та корреляции и ни

на что не влияет. Итог: $\rho_S = \hat{r}$ при $C_{ij}(X) = \dfrac{\operatorname{rank}(X_i) - \operatorname{rank}(X_j)}{\sqrt{n}}$

3) Кендалл: $\tau = \dfrac{S-R}{S+R}$, где S – # согласованных пар, R – # несогласованных пар, причём пары согласованы,

когда $\operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j) = 1$. Значит, $S = \sum_{i<j} \dfrac{\operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j) + 1}{2}$, $R = \sum_{i<j} \dfrac{1 - \operatorname{sign} \cdot \operatorname{sign}}{2}$

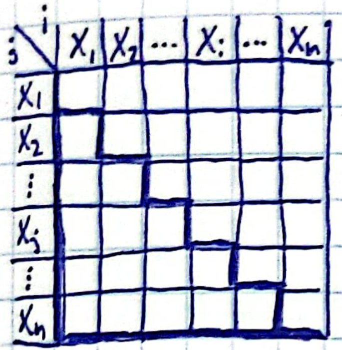
(тут без огр. общн. считаем, что $X_i \ne X_j$, $Y_i \ne Y_j$, т.е. $\operatorname{sign} \in \{-1, 1\}$, а иначе можно, например, относить

пары к числу согл./несогл., в зависимости от задачи). Тогда $\tau = \dfrac{S-R}{S+R} = \dfrac{\sum_{i<j} \operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j)}{1}$

Значит, $C_{ij}(X) = \operatorname{sign}(X_i - X_j)$. Итог: $\tau = \hat{r}$ при $C_{ij}(X) = \operatorname{sign}(X_i - X_j)$

2 Док-ть, что $\hat{r}$ - действительно выб. к-т корр., т.е. что $|\hat{r}| \le 1$ и $\hat{r} = \pm 1$ достигается, а также

что при независ. выборках $E\hat{r} = 0$.



Рассмотрим A - матрицу значений $C_{ij}(X) = f(X_i, X_j)$. Треугольник ниже диагонали задействован в суммах в определении $\hat{r}$ $\left(\sum_{i<j} \dots\right)$. Поскольку $f(x,y) = -f(y,x)$, то $A^T = -A$ (в частности, на диагонали нули: $f(x,x) = -f(x,x) = 0$).

Рассмотрим $\mathrm{trace}\left(\frac{1}{2} A^T A\right) = \mathrm{trace}\left(-\frac{1}{2} AA\right) = \frac{1}{2} \sum_{i=1}^{n} (-AA)_{ii} =$

$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}^2(X) = \sum_{i<j} C_{ij}^2(X)$. Пусть B - матрица значений $C_{ij}(Y) = f(Y_i, Y_j)$, тогда ясно, что

$\mathrm{trace}\left(\frac{1}{2} B^T B\right) = \sum_{i<j} C_{ij}^2(Y)$, $\mathrm{trace}\left(\frac{1}{2} A^T B\right) = \sum_{i<j} C_{ij}(X)\, C_{ij}(Y)$, то есть $(\hat{r})^2 = \left(\dfrac{\mathrm{trace}\left(\frac{1}{2} A^T B\right)}{\sqrt{\mathrm{trace}\left(\frac{1}{2} A^T A\right) \cdot \mathrm{trace}\left(\frac{1}{2} B^T B\right)}}\right)^2 =$

$= \left(\dfrac{\mathrm{trace}(A^T B)}{\sqrt{\mathrm{trace}(A^T A) \cdot \mathrm{trace}(B^T B)}}\right)^2 = \dfrac{\mathrm{trace}^2(A^T B)}{\mathrm{trace}(A^T A) \cdot \mathrm{trace}(B^T B)} = \dfrac{\langle A, B\rangle^2}{\langle A, A\rangle \cdot \langle B, B\rangle}$

По нер-ву КБШ, $\langle A, B\rangle^2 \le \langle A, A\rangle \cdot \langle B, B\rangle \Rightarrow (\hat{r})^2 \le 1 \Rightarrow \boxed{|\hat{r}| \le 1}$. В КБШ равенство

достигается при $A = tB$, $t \in \mathbb{R}$ или один из них $= 0$ (например, $A = 0$). Случай $A = 0$ / $B = 0$

нас не интересует, т.к. он зануляет знаменатель и $\hat{r}$ будет не определен. При $A = tB$ имеем

$C_{ij}(X) = t \cdot C_{ij}(Y)$, т.е. $f(X_i, X_j) = t \cdot f(Y_i, Y_j)$ $\forall i,j$. Ясно, что это выполняется, например,

для $f(x_i, x_j) = x_i - x_j$ и $Y = \frac{1}{t} \cdot X$: $(x_i - x_j) = t \cdot \left(\frac{x_i}{t} - \frac{x_j}{t}\right) \Rightarrow \boxed{\hat{r} = \pm 1 \text{ достигается}}$.

Пусть X, Y независимы, тогда A, B - тоже независимы. Значит, $\mathrm{cov}(A,B) = E\langle A, B\rangle = 0$,

т.е. числитель у $E\hat{r}$ зануляется $\Rightarrow E\hat{r} = \dfrac{E\langle A,B\rangle^2}{E(\langle A,A\rangle \langle B,B\rangle)} = \dfrac{(E\langle A,B\rangle)^2}{E(\langle A,A\rangle \langle B,B\rangle)} = 0 \Rightarrow \boxed{E\hat{r} = 0}$