# Mesures de dependència basades en rangs: Kendall i Spearman

**Mètodes no paramètrics i de remostratge**
Grau interuniversitari en Estadística UB – UPC

*Prof. Jordi Ocaña Rebull*
*Departament d'Estadística, Universitat de Barcelona*

- Probabilitat de concordança, $\pi_C$, i de discordança, $\pi_D$, entre dues v.a. $X$ i $Y$:

$$\pi_C = \Pr\{X_1 < X_2, Y_1 < Y_2\} + \Pr\{X_1 > X_2, Y_1 > Y_2\}$$
$$= \Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\}$$
$$\pi_D = \Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$$

per dues observacions independents

de $(X,Y)$: $(X_1,Y_1)$ i $(X_2,Y_2)$

# Probabilitats de concordança i discordança

- Pel cas absolutament continu (considerat a partir d'ara):

$\pi_C + \pi_D = 1$, i si $X, Y$ estocàsticament independents: $\pi_C = \pi_D = \dfrac{1}{2}$

- Coeficient $\tau$ de Kendall **poblacional**:

$$\tau = \pi_C - \pi_D$$

o bé:

$$\tau = \pi_C - \left(1 - \pi_C\right) = 2\pi_C - 1$$

## Coeficient $\tau$ ("tau") de Kendall
## Concepte

- Propietats de $\tau$:

1) $-1 \leq \tau \leq +1$, $|\tau| = 1$ *sii* relació funcional monòtona perfecta

2) $X, Y$ independents $\Rightarrow \tau = 0$

($\Leftarrow$ en general falsa)

3) Si $(X, Y) \sim$ normal bivariant:

$$\tau = 0 \Leftrightarrow \rho = 0$$

(raó: $\tau = (2/\pi)\arcsin(\rho)$)

## Coeficient $\tau$ de Kendall. Propietats

$$\binom{n}{2} = \frac{n(n-1)}{2}$$ possibles parelles $X_i, X_j$

(o $Y_i, Y_j$). Per tant:

$$\hat{\pi}_C = \frac{\#\{(X_i - X_j)(Y_i - Y_j) > 0\}}{n(n-1)/2} = \frac{2n_C}{n(n-1)}$$

$$\hat{\pi}_D = \frac{\#\{(X_i - X_j)(Y_i - Y_j) < 0\}}{n(n-1)/2} = \frac{2n_D}{n(n-1)}$$

## Estimació de les probabilitats de concordança i de discordança

$$\hat{\tau}_n = \hat{\pi}_C - \hat{\pi}_D = \frac{2\left(n_C - n_D\right)}{n\left(n-1\right)} = \frac{2S}{n\left(n-1\right)}$$

($S$ = "Estadístic de Kendall",
sovint també designat $K$)
o bé:

$$\hat{\tau}_n = 2\hat{\pi}_C - 1 = \frac{4n_C}{n\left(n-1\right)} - 1$$

## Coeficient $\tau$ de Kendall mostral

## Coeficient $\tau$ de Kendall mostral. Forma alternativa

$$\hat{\tau}_n = \frac{2}{n(n-1)} \overbrace{\sum_{1 \le i < j \le n} \operatorname{sgn}(X_i - X_j)\operatorname{sgn}(Y_i - Y_j)}^{=S \ \text{"Estadístic de Kendall"}}$$

on:

$$\operatorname{sgn}(z) = \begin{cases} -1 \text{ si } z < 0 \\ +1 \text{ si } z > 0 \end{cases}$$

- Propietats de $\hat{\tau}_n$:

1) Estadístic basat en rangs: a totes les expressions anteriors, es podrien substituir els valors $(X_i, Y_i)$ pels seus rangs $(R_i, S_i)$

2) No esbiaixat: $E(\hat{\tau}_n) = \tau$

3) $-1 \leq \hat{\tau}_n \leq +1$

4) $\operatorname{var}(\hat{\tau}_n) \xrightarrow[n \to \infty]{} 0$

5) Consistent: $\hat{\tau}_n \xrightarrow{P} \tau$

# Coeficient $\tau$ de Kendall mostral Propietats

$((X_1, Y_1), \ldots, (X_n, Y_n))$ m.a.s. de $(X, Y)$

$((R_1, S_1), \ldots, (R_n, S_n))$ els seus rangs,

$R_1, \ldots, R_n$ rangs de $X_1, \ldots, X_n$, i per separat,

$S_1, \ldots, S_n$ rangs de $Y_1, \ldots, Y_n$, llavors:

$$\#\{(X_i - X_j)(Y_i - Y_j) > 0\} =$$

$$\#\{(R_i - R_j)(S_i - S_j) > 0\}, \text{ etc.}$$

**Punt 1) anterior: $\tau$ mostral és un estadístic basat en rangs**

$H_0$ : $X, Y$ estocàsticament independents

$H_1 : \tau(X,Y) \neq 0$

(o $H_1^{'} : \tau(X,Y) < 0$, o $H_1^{''} : \tau(X,Y) > 0$)

# Test d'independència basat en el coeficient $\tau$ de Kendall mostral

Si $H_0$ és certa:

1) Distribució de $\hat{\tau}_n$ simètrica i independent de la distribució de $(X,Y)$

2) $E(\hat{\tau}_n) = 0$

3) Per mides mostrals no molt grans està tabulada

# Test d'independència exacte

Si $H_0$ és certa:

1) Per mides mostrals grans, es pot fer servir la següent aproximació a la variància de $\hat{\tau}_n$ :

$$\operatorname{var}(\hat{\tau}_n) \cong \frac{2(2n+5)}{9n(n-1)}$$

2) Distribució asimptòtica (força adequada per $n > 10$):

$$Z = \frac{\hat{\tau}_n - 0}{\sqrt{\operatorname{var}(\hat{\tau}_n)}} = \frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}\hat{\tau}_n \approx N(0,1)$$

# Test d'independència aproximat

- Alternativa bilateral: rebutjarem $H_0$ si $\left|\hat{\tau}_n\right| \geq \tau_\alpha(n)$ (valor crític bilateral pel nivell $\alpha$ i mida mostral $n$)

  (o bé si $\left|S\right| \geq$ valor crític per $S$)
- Alternatives unilaterals $H_1^{'} : \tau(X,Y) < 0$ o $H_1^{''} : \tau(X,Y) > 0$ : rebutjarem $H_0$ si $\left|\hat{\tau}_n\right| \geq \tau_\alpha^*(n)$ (valor crític unilateral pel nivell $\alpha$ i mida mostral $n$) i $\hat{\tau}_n$ té el mateix signe indicat a $H_1$ (o bé si $\left|S\right| \geq \ldots$)

# Test exacte

# Test asimptòtic

| $H_1:\ \tau \neq 0$ | $H_1^{'}:\ \tau < 0$ | $H_1^{''}:\ \tau > 0$ |
| --- | --- | --- |
| es rebutja $H_0$ si: | es rebutja $H_0$ si: | es rebutja $H_0$ si: |
| $\lvert Z \rvert \geq z_\alpha$ | $Z \leq -z_{2\alpha}$ | $Z \geq z_{2\alpha}$ |

$z_p$ valor crític >0 a taula $N(0,1)$ per prova bilateral per nivell de significació $p$

- “tau-a”: estadístic de Kendall sense empats
- “tau-b”: $\hat{\tau}_B = \dfrac{n_C - n_D}{\sqrt{(N-t)(N-u)}}$

$N = n(n-1)/2$

$t = \sum_{i=1}^{s_X} t_i(t_i - 1)/2$, $s_X$ sèries d'empats per $X$, de llargada $t_i, i = 1,\dots,s_X$

$u = \sum_{i=1}^{s_Y} u_i(u_i - 1)/2$, $s_Y$ sèries d'empats per $Y$, de llargada $u_i, i = 1,\dots,s_Y$

# Empats i $\tau$ de Kendall

$$Z_B = \frac{n_C - n_D}{\sqrt{v}} \approx N(0,1)$$

$$v = (v_0 - v_t - v_u)/18 + v_1 + v_2$$

$$v_0 = n(n-1)(2n+5)$$

$$v_t = \sum_{i=1}^{s_X} t_i(t_i - 1)(2t_i + 5)$$

$$v_u = \sum_{i=1}^{s_Y} u_i(u_i - 1)(2u_i + 5)$$

$$v_1 = \sum_{i=1}^{s_X} t_i(t_i - 1)\sum_{i=1}^{s_Y} u_i(u_i - 1) / (2n(n-1))$$

$$v_2 = \frac{\sum_{i=1}^{s_X} t_i(t_i - 1)(t_i - 2)\sum_{i=1}^{s_Y} u_i(u_i - 1)(u_i - 2)}{(9n(n-1)(n-2))}$$

# Distribució asimptòtica de tau-b

- "tau-c", més apropiat per mostres grans:

$$\hat{\tau}_C = \frac{2k\left(n_C - n_D\right)}{n^2\left(k-1\right)}$$

$$k = \min\left\{n - \sum_{i=1}^{s_X} t_i, n - \sum_{i=1}^{s_Y} u_i\right\}$$

mínim nombre d'observacions no empatades a $X$ o a $Y$

- Es pot utilitzar la mateixa estimació de la variància que amb "tau-b"

# Empats i $\tau$ de Kendall

- És el coeficient de correlació usual de Pearson, calculat sobre els rangs ($R_i$,$S_i$)
  - (Per tant, evidentment és un estadístic basat en rangs)

$$r_S = \frac{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)^2 \sum_{i=1}^{n}\left(S_i - \bar{S}\right)^2}}$$

- No és gaire clar quin paràmetre poblacional estima...

## Coeficient de correlació (mostral) per rangs de Spearman

- ...més ben dit, sí que és clar, però no és gaire clara la seva interpretació:

$$\rho_S = 3\{P_C - P_D\}$$

$$P_C = \Pr\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right]$$

$$P_D = \Pr\left[(X_1 - X_2)(Y_1 - Y_3) < 0\right]$$

per tres observacions independents

de $(X,Y)$: $(X_1,Y_1),(X_2,Y_2)$ i $(X_3,Y_3)$

# Coeficient de correlació (poblacional) de Spearman

Atès que:

$$\bar{R} = \bar{S} = \frac{n+1}{2}$$

$$\sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2 = \sum_{i=1}^{n} \left(S_i - \bar{S}\right)^2 = \frac{n\left(n^2 - 1\right)}{12}$$

$$\boxed{r_S = \frac{12}{n\left(n^2 - 1\right)} \sum_{i=1}^{n} R_i S_i - 3\,\frac{n+1}{n-1}}$$

## Coeficient de correlació per rangs de Spearman. Fórmula de càlcul

1) $-1 \leq r_S \leq +1$

$r_S = 1 \; sii \; R_i = S_i$ per tot $i$

$r_S = -1 \; sii \; R_i = n - S_i$ per tot $i$

Si $X, Y$ independents:

2) la distribució exacta de $r_S$ és simètrica

i no depèn de les distribucions de $X$ i $Y$

3) $E(r_S) = 0 \quad \text{var}(r_S) = 1/(n-1)$

4) $\sqrt{n-1}\, r_S \approx N(0,1)$

# Propietats de $r_S$

- Les propietats 2 a 4 anteriors permeten construir un test per rebutjar "$H_0 : X$ i $Y$ independents"
- En general es considera preferible el test basat en el coeficient de Kendall, ja que:
  - El paràmetre (i per tant el concepte) de dependència que posa de manifest $H_1$ no té una interpretació clara
  - La convergència a la normal és més lenta per $r_S$

## Test d'independència basat en el coeficient de Spearman