7.2.6 Spearmanův korelační koeficient pořadí

Anglický psycholog Charles Edward Spearman (1863–1945) navrhl svůj koeficient korelace tak, že koreloval postupem podle Pearsona *pořadí* jednotlivých měření obou proměnných. Význam tohoto kroku spočívá v tom, že jeho koeficient zachycuje monotónní vztahy (ne pouze lineární, ale obecně rostoucí nebo klesající); je rezistentní vůči odlehlým hodnotám.

Spearmanovým korelačním koeficientem, jehož teoretickou hodnotu značíme ρ_s, měříme sílu vztahu X a Y, když nemůžeme předpokládat linearitu očekávaného vztahu nebo normální rozdělení proměnných X a Y. Závislost proměnných může mít obecně vzestupný nebo sestupný charakter. Jestliže $r_s = 1$, resp. $r_s = -1$, párové hodnoty (x_i, y_i) leží na nějaké vzestupné, resp. klesající funkci. Hodnoty r_s nemění jakákoli vzestupná transformace původních dat. Pro malé rozsahy je jeho výpočet méně pracný než výpočet Pearsonova korelačního koeficientu.

Odhadem ρ_s, je výběrový koeficient korelace r_s $(-1 \leq r_s \leq 1)$, který pro daný výběr (x_i, y_i) spočteme podle vzorce

$$r_s = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)},$$

kde D_i jsou rozdíly pořadí R_x a R_y hodnot x_i a y_i vzhledem k ostatním hodnotám seřazeného výběru podle velikosti. Před výpočtem je nutno oběma řadám čísel x_i a y_i tato pořadí přiřadit. Jestliže dvě čísla v řadě hodnot x_i, resp. y_i jsou stejná, přiřadíme jim průměrnou hodnotu příslušných pořadí. Obdobně provedeme tuto úpravu pro více stejných hodnot. V každé řadě nesmí být více než 1/5 pozorování stejných. Pokud se tak stane, musíme celý výpočet upravit.

PŘÍKLAD 7.7

Výpočet Spearmanova korelačního koeficientu

Výpočet r_s si ukážeme pro hodnoty z tabulky 7.10:

$$r_s = 1 - \frac{6 \times 26}{10(100 - 1)} = 0{,}84.$$

Pro posouzení statistické významnosti koeficientu r_s slouží tabulka X z přílohy B. Přesahuje-li hodnota $|r_s|$ tabulkovou hodnotu pro daný počet párů měření n a hladinu významnosti, můžeme vztah považovat za prokázaný. Pro náš příklad, testujeme-li dvoustrannou hypotézu $\rho_s = 0$ na hladině 1 %, je tabulková hodnota 0,746 (tabulka obsahuje kritické hodnoty pro dvoustranné testy). Vztah

Tab. 7.10 Příklad postupu při výpočtu Spearmanova korelačního koeficientu pořadí

x	y	R_x	R_y	$D = R_x - R_y$	$D \times D$
187	72	10,00	6,50	3,50	12,25
170	60	1,00	1,00	0,00	0,00
180	73	6,50	8,00	−1,50	2,25
184	74	8,00	9,00	−1,00	1,00
178	72	5,00	6,50	−1,50	2,25
180	70	6,50	4,50	2,00	4,00
172	62	2,00	2,00	0,00	0,00
176	70	3,00	4,50	−1,50	2,25
186	80	9,00	10,00	−1,00	1,00
177	67	4,00	3,00	1,00	1,00
Součet					26,00

mezi oběma proměnnými z příkladu je tedy prokázán. U větších výběrů $(n \geq 30)$ lze na hladině α použít přibližný z-test hypotézy $\rho_s = 0$:

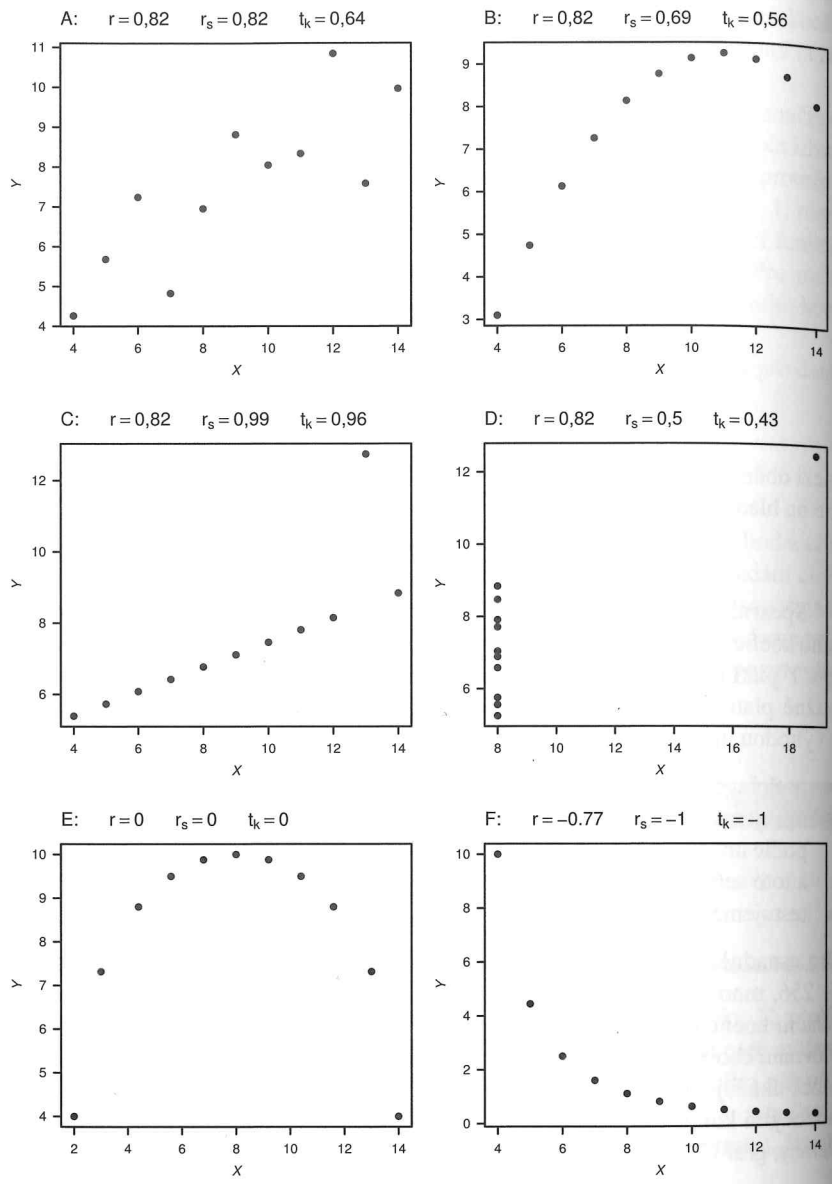
$$z = |r_s| \sqrt{n - 1}.$$

Spearmanův koeficient r_s někdy používáme pro odhad Pearsonova korelačního koeficientu, resp. r, jelikož pro dvojrozměrně normálně rozdělené proměnné X a Y platí přibližný vztah $\rho = 2 \sin(0{,}523\rho_s)$. Tento vzorec je upřesněním přibližně platného vztahu $\rho = \rho_s$. Podle Spearmana lze jeho koeficient korelace s výhodou uplatnit v situacích, kdy:

- potřebujeme rychlý a rezistentní odhad korelačního koeficientu r;
- testujeme schopnost zkoumané osoby správně řadit objekty nebo vlastnosti podle určitých hledisek tak, že ji necháme seřadit tyto objekty nebo vlastnosti a toto seřazení pak srovnáme se standardem;
- testujeme možnost přítomnosti monotónního trendu v časové řadě měření.

Pro usnadnění interpretace jsou na obrázku 7.5 znázorněna data z příkladu 7.3 (s. 256, množina 1 = A, 2 = B, 3 = C, 4 = D) a uvedeny k nim vypočtené korelační koeficienty podle Pearsona, Spearmana a Kendalla, aby bylo umožněno srovnání chování těchto koeficientů (viz odstavec o Pearsonově koeficientu). Obrázek ukazuje, jak Spearmanův koeficient zachytí vztah reprezentovaný různými bodovými konfiguracemi. Graf F dokumentuje jeho schopnost měřit monotónní vztahy, graf C ukazuje jeho rezistenci vůči odlehlým hodnotám.

Obr. 7.5 Zobrazení různých bodových konfigurací a k nim dopočítaného Pearsonova (r), Spearmanova (r_s) a Kendallova (t_k) korelačního koeficientu



7.2.7 Kendallův koeficient pořadové korelace

Korelační koeficient má měřit „sílu vztahu" dvou proměnných. Ale různé korelační koeficienty ho měří různým způsobem. Pearsonův i Spearmanův korelační koeficient mohou mít hodnotu 0,3, ale pokaždé to znamená něco trochu jiného. Kendallův korelační koeficient má na rozdíl od předchozích dvou jednoduchou pravděpodobnostní interpretaci. Jeho teoretickou hodnotu v populaci označujeme τ_k nebo Kendallovo *tau*.

Zatímco Spearman koreloval pořadí, Kendall založil svoji statistiku na inverzích v pořadí. Vycházíme z dat, která se týkají metrického nebo ordinálního hodnocení n objektů ($i = 1, 2, \ldots, n$) podle dvou kritérií X a Y. Ke každému objektu i získáme ohodnocení (x_i, y_i). Nejdříve seřadíme dvojice (x_i, y_i) tak, že hodnoty x_i budou tvořit rostoucí posloupnost. Jestliže mezi kritérii X a Y je kladná asociace, pak také y_i budou mít vzestupnou tendenci. Při záporné asociaci budou mít y_i sestupnou tendenci. Kendall proto rozlišuje vztah $y_j > y_i$, resp. $y_j < y_i$, pokud $j > i$ ($i = 1, 2, \ldots, n-1$). V prvním případě nastává tzv. **konkordance**, jež skóruje pro kladnou asociaci, ve druhém **diskordance**, která skóruje pro negativní asociaci. Počet všech konkordancí, resp. diskordancí označíme P, resp. Q. Rozdíl $S = P - Q$ někdy nazýváme Kendallovo S a je jednoduchou mírou závislosti. Převaha konkordancí, resp. diskordancí vede ke kladné, resp. záporné hodnotě S. Možná škála hodnot S závisí na rozsahu výběru n. Jednoduchá úprava však tento problém vyřeší. S se totiž může pohybovat mezi hodnotami $-0{,}5n(n-1)$ a $0{,}5n(n-1)$. Proto se Kendallův koeficient *tau* t_k počítá podle formule

$$t_k = \frac{S}{D} = \frac{P - Q}{D},$$

kde jmenovatel D je maximální možný počet konkordancí, resp. diskordancí a má hodnotu $n(n-1)/2$.

PŘÍKLAD 7.8

Výpočet konkordancí a Kendallova koeficientu pořadové korelace

Vypočítáme počet diskordancí a konkordancí pro data v tabulce 7.11. Protože počty P a Q jsou přibližně stejné, mezi proměnnou X a Y není pravděpodobně žádná asociace. S má hodnotu −2.

Kendallův koeficient $t_k = -2/36 = -0{,}05$.

Tab. 7.11 Příklad výpočtu Kendallova koeficientu pořadové korelace

Věk (X)	Cholesterol (Y)	Konkordance	Diskordance
41	274	1	7
45	209	4	3
50	194	5	1
51	270	1	4
54	165	4	0
59	234	2	1
62	281	0	2
68	238	0	1
71	208	0	0
Součet		$P = 17$	$Q = 19$

Platí $-1 \leq t_k \leq 1$ a hodnot právě ± 1 nabývá t_k ve stejných situacích jako Spearmanův koeficient. Kritické hodnoty pro rozhodování, kdy je možné zamítnout hypotézu nezávislosti X a Y (H_0: $\tau_k = 0$), nalezneme pomocí speciálních tabulek. Některé programy dokážou spočítat přesnou p-hodnotu pro test nulové hodnoty τ_k. Pro velká n má t_k přibližně normální rozdělení se střední hodnotou 0 a směrodatnou odchylkou s_τ

$$s_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}},$$

pokud proměnné X a Y jsou nezávislé. Rozhodování o nulové hodnotě τ_k vychází z testovací z-statistiky $z = t_k / s_\tau$, kterou porovnáváme s kritickými hodnotami standardizovaného normálního rozdělení.

Interpretace τ_k je přímočařejší než u Spearmanova koeficientu ρ_s. Jestliže $\tau_k = p$, můžeme u dvou náhodně vybraných jedinců očekávat s pravděpodobností p, že jejich seřazení podle kritéria X bude stejné jako seřazení podle kritéria Y. Většinou oba koeficienty mají přibližně stejnou velikost.

V kapitole 8.4 poznáme využití Kendallova korelačního koeficientu při hodnocení závislosti v kontingenčních tabulkách, jež vznikly klasifikací objektů podle dvou ordinálních znaků.

Jestliže v údajích existují shody ($x_j = x_i$, resp. $y_j = y_i$), musíme výpočet modifikovat, protože v tomto případě nemůže koeficient dosáhnout hodnoty -1, resp. 1. Modifikaci uplatňujeme při větším počtu shod a týká se jmenovatele D ve vzorci pro výpočet Kendallova *tau*. Označme

symboly u, resp. v počty shodných pořadí mezi x_i, resp. y_i postupně v jednotlivých skupinách shodných pořadí a symboly U a V součty, které mají tvar:

$$U = 0{,}5 \sum u(u-1),$$
$$V = 0{,}5 \sum v(v-1).$$

Modifikace výpočtu spočívá v nahrazení D číslem $D' = \sqrt{(D-U)(D-V)}$. Takto modifikovaný výpočet Kendallova *tau* nazýváme **Kendallovo *tau-b***, značíme t_b. Kendallovo t_b lze interpretovat jako korelaci mezi hodnotami dx a dy, kde dx se rovná 1, resp. -1, pokud pro $j > i$ je $x_j > x_i$, resp. $x_j < x_i$, a nule v ostatních případech. Hodnoty dy počítáme obdobně. Jak hodnoty dx, tak hodnoty dy spočítáme pro všechna možná srovnání, kterých je $n(n-1)/2$. (Zvára, 2000)

7.2.8 Bodově biseriální korelační koeficient a koeficient ϕ

Vztah mezi spojitou metrickou proměnnou a binární proměnnou se měří biseriálním korelačním koeficientem r_{pb} tak, že n dvojic měření se rozdělí na dvě skupiny podle hodnoty alternativního parametru a spočte se hodnota r_{pb} podle vzorce

$$r_{pb} = \frac{(\bar{x}_1 - \bar{x}_2)}{s} \sqrt{\frac{n_1 n_2}{n(n-1)}},$$

kde n_i, resp. $\bar{x}_i$ jsou počty, resp. průměrná hodnota spojitého parametru v obou skupinách a s je společná směrodatná odchylka. Tento koeficient r_{pb} testujeme podobně jako normální korelační koeficient. Jestliže $r_{pb} > 1$, resp. $r_{bk} < -1$, dosadíme za něj hodnotu 1, resp. -1. Uvedený vzorec se v praxi nepoužívá, protože stejnou hodnotu dostaneme použitím algoritmu pro Pearsonův koeficient korelace pro dvojice hodnot obou proměnných, přičemž binární proměnnou zastupují nuly a jedničky. Jestliže binární proměnná vznikla dichotomizací spojité normálně rozdělené proměnné, můžeme spočítat odhad Pearsonova korelačního koeficientu obou spojitých proměnných pomocí tzv. biseriálního korelačního koeficientu (viz Howell, 1992, s. 270).

Koeficient ϕ je Pearsonův korelační koeficient vypočítaný pro dvě alternativní proměnné, které kódujeme pomocí hodnot 0 a 1. (Existuje i jednodušší výpočet, ale ten nemá v době počítačů opodstatnění.) Platí, že $\phi^2 = \chi^2/n$, kde χ^2 je testovací statistika nezávislosti v čtyřpolní tabulce a n je počet dvojic, z nichž se počítá korelační koeficient. Test nulové hodnoty koeficientu ϕ se provádí stejně jako test nezávislosti pro čtyřpolní tabulku, která je tvořena četnostmi kombinací hodnot obou proměnných (viz kap. 8.3.1).