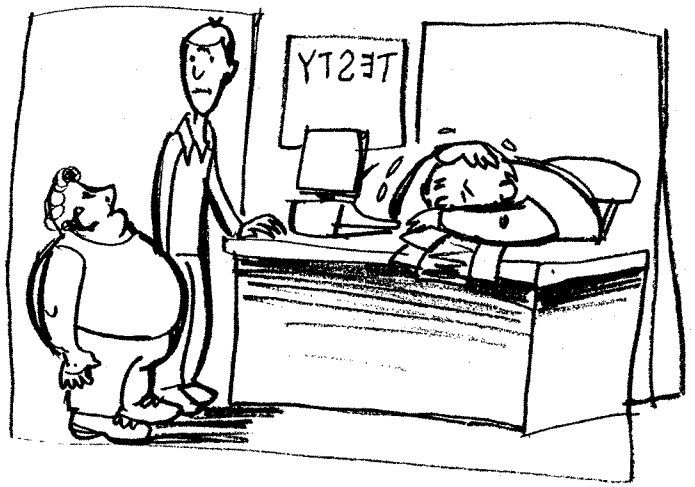


„Prý jsme mu zkazili jeho pozitivní korelaci mezi výškou a váhou."

Cílem regresní a korelační analýzy je popis statistických vlastností vztahu dvou nebo více proměnných. Dvojrozměrný bodový graf nebo korelační tabulka dávají první představu o rozdělení sledovaných proměnných. Graf často indikuje překvapivé vlastnosti dat jako nelinearitu vztahu, nehomogenitu nebo přítomnost odlehlých hodnot. Na obrázku 7.1 je rovněž vynesena přímka, která byla proložena body metodou nejmenších čtverců. Vliv třetí proměnné na rozložení bodů můžeme zachytit různým tvarem nebo barvou bodů v závislosti na hodnotě této proměnné (např. u dat o výšce a váze bychom mohli použít různé značky pro body odpovídající chlapcům a dívkám, pokud bychom tuto informaci o proměnné *pohlaví* měli k dispozici). Některé možné konfigurace dat v grafu popíšeme v následujícím odstavci.

7.2 Korelační analýza

V nejobecnějším smyslu, slovo „korelace" označuje míru stupně asociace dvou proměnných. Říká se, že dvě proměnné jsou korelované (resp. asociované), jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé proměnné. Míra této tendence může sahat od neexistence korelace (všechny hodnoty proměnné Y se vyskytují stejně pravděpodobně s každou hodnotou proměnné X) až po absolutní korelaci (s danou hodnotou

proměnné X se vyskytuje právě jedna hodnota proměnné Y). Pro měření korelace byla navržena řada koeficientů. Liší se podle typů proměnných, pro které se využívají. Statistické usuzování o korelačních koeficientech se opírá o teorii pravděpodobnosti pro společné rozdělení dvou nebo více náhodných proměnných.

Při zkoumání korelačních vztahů má rozhodující význam kvalitativní rozbor příslušného materiálu. Nemá smysl měřit závislost tam, kde na základě logické úvahy nemůže existovat. Často je zbytečné měřit závislosti i z jiných důvodů. Je to zejména tehdy, když je korelace způsobena: a) formálními vztahy mezi proměnnými; b) nehomogenitou studovaného základního materiálu; c) působením společné příčiny.

Formální korelace vzniká např. tehdy, když se zjišťuje korelace procentuálních charakteristik, jež se navzájem doplňují do 100 % (např. korelace procentního zastoupení bílkovin a tuku v potravinách).

Jestliže populace, kterou studujeme, obsahuje subpopulace, pro něž se průměrné hodnoty proměnných X a Y liší, vypočtené korelační vztahy jsou touto **nehomogenitou** silně ovlivněny a jejich hodnoty nepopisují skutečný vztah mezi uvažovanými proměnnými. Nehomogenita materiálu se projeví na bodovém grafu tak, že shluky bodů pro subpopulace se budou nacházet v různých oblastech souřadnicového systému. Na obrázku 7.2 je modelově ukázáno působení nehomogenity. Ta má za důsledek, že korelačním koeficientem hodnotíme bez diferenciace najednou dva shluky bodů, které přísluší k různým populacím. Na obrázku A to vede k nenulovém korelačními koeficientu i přesto, že v obou shlucích jsou proměnné nekorelované, naopak proměnné na obrázku B jsou v obou shlucích proměnné korelované, ale celková korelace je nulová.

Příkladem **korelací způsobených společnou příčinou** jsou vztahy mezi některými mírami těla, např. mezi délkou pravé a levé ruky. Jiným známým příkladem jsou zdánlivé korelace způsobené časovým faktorem nebo faktorem modernizace u dvou řad údajů.

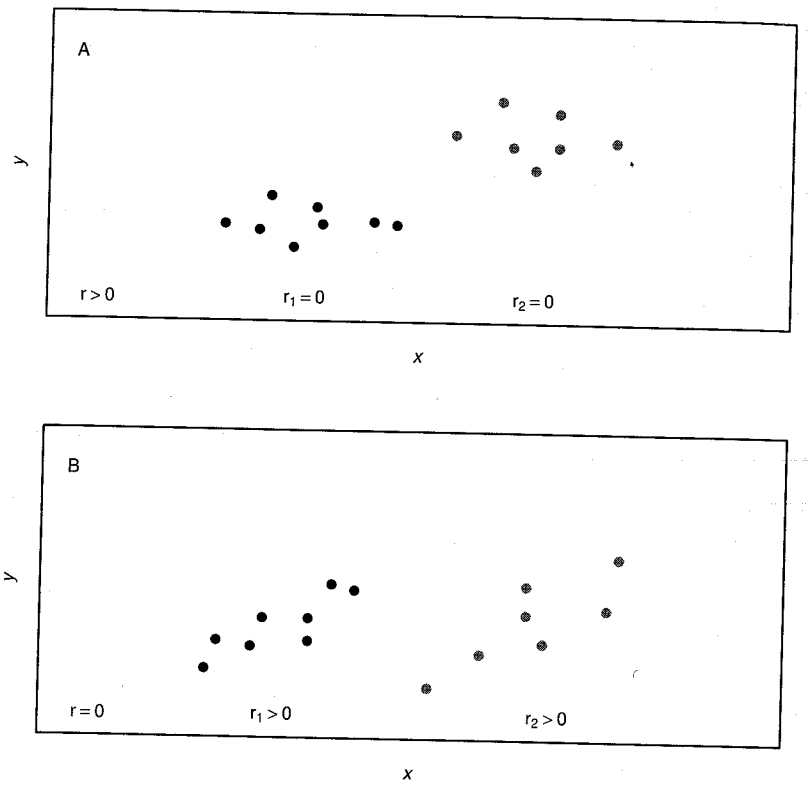
PŘÍKLAD 7.1

Zdánlivé korelace

Počet televizních přístrojů na osobu koreluje s očekávanou délkou života. Ve státech, kde je mnoho televizních přístrojů, dosahují obyvatelé vysokého věku. Je možné změnou počtu televizních přístrojů dosáhnout prodloužení věku v oblastech světa, kde je nižší očekávaná délka života?

Podobným korelacím se někdy říká „nesmyslné" korelace. Hodnota korelace je vysoká. Nesmyslný by byl závěr o příčinném působení. Korelační závislost

Obr. 7.2 Příklad kladné (A) a nulové (B) korelace, které jsou způsobené nehomogenitou dat

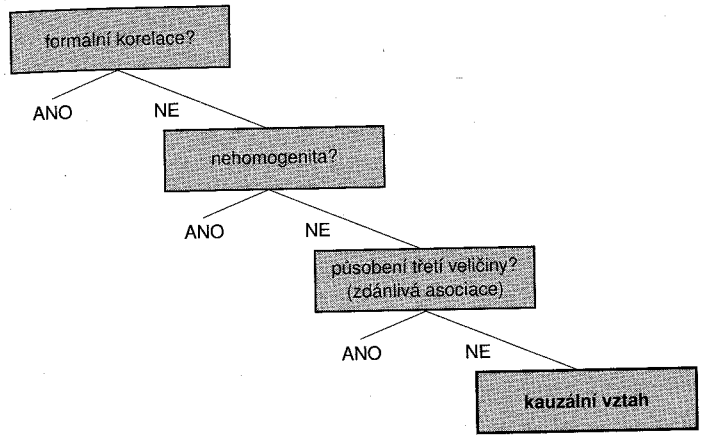


Korelační koeficient r je vypočtený pro všechny body, koeficienty r_1 a r_2 odděleně pro každý shluk zvlášť

je zdůvodněna proměnnou „národní důchod", jež je společnou příčinou obou proměnných.

Kromě tohoto působení proměnné jako „společné příčiny" mohou působit matoucí (rušivé) proměnné, které korelují jak s cílovou proměnnou, tak s proměnnou ovlivňující. Proměnná v tomto případě znesnadňuje interpretaci, protože nelze rozlišit vliv matoucí a sledované ovlivňující proměnné na cílovou proměnnou. Uvádíme pořadí, v němž máme vylučovat nezajímavé korelace, než se dostaneme do fáze, kdy by velká korelace mohla indikovat kauzální vztah (obr. 7.3).

Obr. 7.3 Postup pro ověření kauzálního vztahu



7.2.1 Pearsonův korelační koeficient

Přes některé své nedostatky zůstává Pearsonův korelační koeficient r nejdůležitější mírou síly vztahu dvou náhodných spojitých proměnných X a Y. Počítáme jej z n párových hodnot $\{(x_i, y_i)\}$ změřených na n jednotkách náhodně vybraných z populace. Korelační koeficient r nabývá hodnot z intervalu $[-1; 1]$. Jestliže má hodnotu 1 nebo -1, pak y-souřadnici bodu lze přesně spočítat pomocí lineárního vztahu z jeho x-souřadnice. Korelační koeficient r počítáme pomocí tzv. kovariance s_{xy} a směrodatných odchylek s_x a s_y obou proměnných:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Vzorec s kovariancí pomáhá porozumět tomu, že r má kladnou hodnotu, pokud asociace proměnných je pozitivní. Dejme tomu, že studujeme korelaci výšky a váhy studentů. Jedinci, kteří mají hodnotu výšky nad průměrem, mívají nadprůměrnou i hodnotu váhy. Oba rozdíly od průměru, jež spolu násobíme při výpočtu kovariance, budou mít u vyšších a těžších jedinců kladnou hodnotu. Jedinci, kteří mají menší výšku, mají obvykle i menší váhu. U nich jsou oba

rozdíly od průměrů záporné, a proto je součin rozdílů od průměru rovněž kladný. Protože je většina sčítanců kladných, musí být kladná i výsledná hodnota kovariance, a tedy i korelačního koeficientu. Tuto interpretaci lze ještě lépe pochopit při výpočtu r pomocí standardizovaných hodnot. Platí totiž vzorec

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n} x'_i y'_i}{n-1},$$

kde x' a y' označují standardizované hodnoty.

Důležité vlastnosti Pearsonova korelačního koeficientu r shrneme pomocí několika tvrzení:

1. Platí $-1 \le r \le 1$.
2. Jestliže $|r| = 1$, leží všechny body na nějaké přímce.
3. Jestliže $r = 0$, nazýváme X a Y nekorelované proměnné. Dvě náhodné proměnné jsou tím více korelovány, čím blíže je hodnota r k číslům 1 nebo -1. V tom případě lze vztah obou proměnných dobře vyjádřit přímkou.
4. Jestliže $r < 0$, resp. $r > 0$, tak se Y v průměru zmenšuje, resp. zvětšuje při zvětšování proměnné X. Říkáme, že asociace je záporná, resp. kladná.
5. Pearsonův korelační koeficient vyjadřuje pouze sílu *lineárního* vztahu. Špatně měří jiné vztahy, ať jsou jakkoli silné.
6. Korelační koeficient se nezmění, když změníme jednotky měření proměnných X a Y.
7. Podobně jako průměr nebo směrodatná odchylka je korelační koeficient r velmi ovlivněn odlehlými hodnotami.
8. Korelační koeficient r nerozlišuje mezi závisle a nezávisle proměnnou.
9. Korelační koeficient r není úplným popisem dat i při velmi silném lineárním vztahu. Pro úplnější popis potřebujeme znát rovnici přímky, která vyjadřuje tvar vztahu.
10. Pokud jedna z proměnných nemá náhodný charakter (její hodnoty jsou pevně určeny), není vhodné korelační koeficient použít.
11. Korelace, ať je jakkoli silná, *neznamená* sama o sobě průkaz příčinného vztahu, tedy toho, že změny proměnné X skutečně působí změny proměnné Y.

Mezi proměnnými mohou existovat nejrůznější vztahy a máme i různé způsoby, jak je měřit. Některé z nich popíšeme v dalších odstavcích. Ačkoli korelační koeficient se používá velmi často, je nutné mít na paměti jeho omezení.

PŘÍKLAD 7.2

Výpočet korelačního koeficientu

Budeme hodnotit závislost výšky a váhy, jejichž hodnoty jsme naměřili u 10 studentů. Vypočítáme korelační koeficient pro párové hodnoty, které jsou uvedeny spolu s potřebnými dopočítanými hodnotami v tabulce 7.3. Hodnoty jsou zobrazeny na obrázku 7.1 (s. 249).

Součet v posledním sloupci je základem pro výpočet kovariance

$$\text{Cov}(x\ y) = s_{xy} = 259\ (10\quad 1) = 28\ 8$$

Dále jsme zjistili: $\bar{x} = 1790\ 10 = 179$; $\bar{y} = 700\ 10 = 10$; $s_x = 5\ 61$; $s_y = 5\ 83$. Korelační koeficient má tedy hodnotu $r = 28\ 8\ (5\ 61\quad 5\ 83) = 0\ 88$.

Tab. 7.3 Příklad postupu výpočtu korelačního koeficientu

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
	187	72	8	2	16
	170	60	9	10	90
	180	73	1	3	3
	184	74	5	4	20
	178	72	1	2	2
	180	70	1	0	0
	172	62	7	8	56
	176	70	3	0	0
	186	80	7	10	70
	177	67	2	3	6
Součet	1790	700	0	0	259

Někdy se zařazují hodnoty korelace do pásem podle síly asociace. V tabulce 7.4 uvádíme jeden z návrhů. Interpretace hodnot korelačního koeficientu není tak přímočará, jako je tomu u většiny jednorozměrných charakteristik. Proto se doporučuje dopočítat další charakteristiky, jako jsou parametry proložené přímky nebo směrodatná chyba odhadu při regresi (viz další kapitola).

Tab. 7.4 Pásma síly asociace podle velikosti korelačního koeficientu r

Síla asociace	$\|r\|$
malá	0,1–0,3
střední	0,3–0,7
velká	0,7–1,0

Hodnota korelačního koeficientu je bohužel silně ovlivňována odlehlými hodnotami ve výběru. Zkreslení také nastane, když se při výběru objektů omezíme pouze na ty, jejichž hodnota proměnné X nebo Y musí ležet v určitém intervalu. Korelační koeficient r má pak tendenci být menší než korelace r' vypočítaná bez omezení kladeného na data. Pro úpravu zkresleného korelačního koeficientu vlivem omezení rozsahu měření proměnné X použijeme vzorec

$$r' = \frac{Ur}{\sqrt{(U^2 - 1)r^2 + 1}},$$

kde $U = s/s'$ je poměr směrodatné odchylky s měření X ve studii a směrodatné odchylky s' v populaci bez restrikce.

Korelační koeficient je také ovlivněn nepřesností metod, kterými měříme obě proměnné. Jestliže známe r_{yy} a r_{xx} koeficienty spolehlivosti měření obou proměnných (jedná se o korelace opakovaných měření), lze se přiblížit hodnotě korelačního koeficientu bezchybně změřených proměnných $r_{x'y'}$ pomocí úpravy

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}.$$

PŘÍKLAD 7.3

Význam exploračního zobrazení dvojrozměrných dat

Jednoduchým příkladem toho, jakou důležitou roli hraje explorační zobrazení dat, je zkoumání čtyř sérií modelových dat podle Anscomba (1973), které uvádí tabulka 7.5. Základní statistické charakteristiky proměnných X a Y a jejich korelační koeficient mají pro první sérii dat hodnoty $\bar{x} = 9\,0$; $s_x = 3\,31$; $\bar{y} = 7\,5$; $s_y = 2\,03$ a $r = 0\,816$. Pokud spočteme tyto charakteristiky pro ostatní série, zjistíme, že jsou stejné. Pokud však všechny čtyři série zobrazíme graficky (viz obr. 7.5A-D, s. 270), výsledek je dost překvapivý.

Tab. 7.5 Série modelových dat se stejnými základními statistickými charakteristikami a korelačními koeficienty

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8 04	10	9 14	10	7 46	8	6 58
8	6 95	8	8 14	8	6 77	8	5 76
13	7 58	13	8 74	13	12 74	8	7 71
9	8 81	9	8 77	9	7 11	8	8 84
11	8 33	11	9 26	11	7 81	8	8 47
14	9 96	14	8 1	14	8 84	8	7 04
6	7 24	6	6 13	6	6 08	8	5 25
4	4 26	4	3 1	4	5 39	19	12 5
12	10 84	12	9 13	12	8 15	8	5 56
7	4 82	7	7 26	7	6 42	8	7 91
5	5 68	5	4 74	5	5 73	8	6 89

7.2.2 Pravděpodobnostní rozdělení dvou náhodných proměnných

Teorie pravděpodobnosti popisuje nejen rozdělení jedné náhodné proměnné, ale i společná pravděpodobnostní rozdělení dvou nebo více náhodných proměnných. Této teorie je zapotřebí tehdy, když chceme navrhovat pravděpodobnostní modely vztahu proměnných a zdůvodnit procedury pro statistické usuzování v korelační a regresní analýze. V našem jednoduše pojatém výkladu budeme postupovat tak, abychom mohli získané výsledky využít i v kapitole o analýze závislosti kategoriálních proměnných.

Zatím jsme se seznámili s jednou dvojrozměrnou charakteristikou, s Pearsonovým korelačním koeficientem r. Teoretickou hodnotu Pearsonova korelačního koeficientu v populaci označujeme ρ. Získali bychom ji výpočtem z údajů o všech prvcích populace. Výběrový koeficient r je bodovým odhadem této hodnoty. S rostoucím rozsahem výběru n se hodnota výběrového korelačního koeficientu r_n blíží ke své teoretické hodnotě ρ.

Teoretickou hodnotu ρ lze přímo odvodit podobně jako teoretickou střední hodnotu μ, když známe společné pravděpodobnostní rozdělení náhodných proměnných, pro které korelační koeficient počítáme. Koncept dvojrozměrného

pravděpodobnostního rozdělení a techniku výpočtu teoretických hodnot ozřejmíme pomocí jednoduchého příkladu. Postupujeme podobně jako v jednorozměrném případě (viz kap. 4.2).

Představme si, že v daném pokusu můžeme získat pro hodnoty proměnných X a Y pouze tři různé hodnoty: $x \in (7; 15; 2)$, $y \in (3; 6; 9)$. Společné pravděpodobnostní p_{xy} rozdělení proměnných X a Z je popsáno tabulkou pravděpodobností všech možných kombinací uvedených hodnot (tab. 7.6). Poslední sloupec, resp. poslední řádek tabulky 7.6 a) obsahuje jednorozměrná rozdělení p_x a p_y náhodných proměnných X a Y. Tyto hodnoty jsme dostali jako součet pravděpodobností v daném řádku, resp. sloupci. Nazýváme je **marginální rozdělení**. Z tabulky je vidět, že dvojici hodnot $(x, y) = (6; 7)$ lze dostat v náhodném pokusu s pravděpodobností 0,1, avšak pravděpodobnost výskytu jevu $y = 7$ je 0,2. Pomocí marginálních rozdělení spočítáme očekávané hodnoty pro proměnné X a Y a také dopočítáme teoretické hodnoty rozptylů obou proměnných pro proměnnou X a Y – (tab. 7.6b, c).

Tab. 7.6 Příklad dvojrozměrného pravděpodobnostního rozdělení a výpočet jeho charakteristik

a) Dvojrozměrné rozdělení

	y			
x	7	15	2	p_x
3	0,1	0,2	0,0	0,3
6	0,1	0,0	0,3	0,4
9	0,0	0,1	0,2	0,3
p_y	0,2	0,3	0,5	1,0

b) Výpočet průměrných hodnot pro proměnnou X

x	p_x	xp_x	x^2p_x
3	0,3	0,9	2,7
6	0,4	2,9	14,4
9	0,3	2,7	24,3
Součet	1,0	6,0	41,4
		$= E(x)$	$= E\ x^2$

c) Výpočet průměrných hodnot pro proměnnou Y

y	p_y	yp_y	y^2p_y
7	0,2	1,4	9,8
15	0,3	4,5	67,5
2	0,5	1,0	2,0
Součet	1,0	6,9	79,3
		$= E(y)$	$= E\ y^2$

Teoretické hodnoty průměru a rozptylu náhodných proměnných X a Y jsou tedy:

$$\mu_x = E(X) = \sum_i x_i p_i(x) = 6{,}0 \qquad \sigma_x^2 = E(x^2) - \mu_x^2 = 41{,}4 - 6^2 = 5{,}4$$

$$\mu_y = E(Y) = \sum_j y_j p_j(y) = 6{,}9 \qquad \sigma_y^2 = E(y^2) - \mu_y^2 = 79{,}3 - 6{,}9^2 = 31{,}7$$

To znamená, že $\sigma_x = \sqrt{5{,}4} = 2{,}32$ a $\sigma_y = \sqrt{31{,}7} = 5{,}63$.

Teoretickou kovarianci σ_{xy} (také ji značíme $\mathrm{Cov}(x, y)$) vypočteme modifikací vzorce pro výběrovou kovarianci:

$$\mathrm{Cov}(x, y) = \sigma_{xy} = E\left[(X - \mu_x)(Y - \mu_y)\right]$$
$$= E(XY) - \mu_x\mu_y = \sum_i \sum_j x_i y_j\, p_{ij}(x, y) - \mu_x\mu_y$$

Nejdříve spočítáme hodnoty $E(XY)$:

$$E(XY) = \sum_i \sum_j x_i y_j \cdot p_{ij}(x, y)$$
$$= 0{,}1 \times 3 \times 7 + 0{,}2 \times 3 \times 15 + 0{,}0 \times 3 \times 2$$
$$+ 0{,}1 \times 6 \times 7 + 0{,}0 \times 6 \times 15 + 0{,}3 \times 6 \times 2$$
$$+ 0{,}0 \times 9 \times 7 + 0{,}1 \times 9 \times 15 + 0{,}2 \times 9 \times 2$$
$$= 36{,}0$$

Takže $\sigma_{xy} = E(XY) - \mu_x\mu_y = 36{,}0 - 6{,}0 \times 6{,}9 = -5{,}4$. Teoretickou hodnotu koeficientu korelace pak dostaneme dosazením teoretických hodnot do vzorce pro výběrový koeficient korelace

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-5{,}4}{2{,}32 \times 5{,}63} = -0{,}41.$$

Teoretická korelace $-0{,}41$ indikuje sílu závislosti mezi oběma proměnnými.

V případě spojitých náhodných proměnných jsou tyto výpočty sice komplikovanější, ale stejně jako u jednorozměrných charakteristik se koncepčně moc neliší.

Dále připomeneme pojem nezávislosti náhodných proměnných a ukážeme, že pokud náhodné proměnné jsou nezávislé, pak se jejich korelační koeficient rovná nule. V dřívějším výkladu jsme **nezávislost náhodných proměnných** vymezili požadavkem, že realizace jedné náhodné proměnné neovlivňuje chování druhé

náhodné proměnné (např. hodnota jedné proměnné u určité osoby neovlivňuje hodnotu měření jiné proměnné u téže ani jiné osoby).

Definice nezávislosti dvou náhodných proměnných vychází z počítání pravděpodobností podmnožin dvojrozměrného prostoru $R \times R$. Nechť množiny A_x, resp. A_y mají pravděpodobnosti $P_x(A_x)$, resp. $P_y(A_y)$ vzhledem k rozdělení proměnné X, resp. Y. Pak X a Y jsou **stochasticky nezávislé**, nebo prostě nezávislé, pokud pravděpodobnost množiny $A_x \times A_y$ vzhledem k uvažovanému dvojrozměrnému rozdělení lze vypočítat vynásobením pravděpodobností obou množin $P(A_x \times A_y) = P_x(A_x)P_y(A_y)$. Tato podmínka musí platit pro všechny podmnožiny A_x a A_y. Je patrné, že se jedná o převedení pojmu nezávislosti náhodných jevů na chování náhodných proměnných.

Pro náš příklad popíšeme dvojrozměrné rozdělení tabulkou 7.6a. Pravděpodobnosti p_{xi}, resp. p_{yj} vznikly součtem pravděpodobností v řádku, resp. v sloupci tabulky. Nazýváme je **marginální pravděpodobnosti**. Definují marginální rozdělení, které popisuje náhodné chování izolovaných proměnných X a Y. Jestliže proměnné X a Y jsou nezávislé, pak z definice plyne, že pravděpodobnosti v tabulce jsou součiny marginálních pravděpodobností (viz tab. 7.7b).

Pojem stochastické nezávislosti dále ilustruje výpočet podmíněné pravděpodobnosti $p(x = i \mid y = j)$, tedy pravděpodobnosti, že náhodná proměnná X

Tab. 7.7 Obecné dvojrozměrné rozdělení a nezávislost proměnných

a) Dvojrozměrné rozdělení proměnných X a Y

x \ y	7	15	2	p_x
3	p_{11}	p_{12}	p_{13}	p_{x1}
6	p_{21}	p_{22}	p_{23}	p_{x2}
9	p_{31}	p_{32}	p_{33}	p_{x2}
p_y	p_{y1}	p_{y2}	p_{y3}	1

b) Podmínka pro nezávislost X a Y

x \ y	7	15	2	p_x
3	$p_{x1}p_{y1}$	$p_{x1}p_{y2}$	$p_{x1}p_{y3}$	p_{x1}
6	$p_{x2}p_{y1}$	$p_{x2}p_{y2}$	$p_{x2}p_{y3}$	p_{x2}
9	$p_{x3}p_{y1}$	$p_{x3}p_{y2}$	$p_{x3}p_{y2}$	p_{x2}
p_y	p_{y1}	p_{y2}	p_{y3}	1

c) Podmíněná rozdělení proměnné X za podmínky $y = y_i$

x \ y	7	15	2	p_x
3	p_{x1}	p_{x1}	p_{x1}	p_{x1}
6	p_{x2}	p_{x2}	p_{x2}	p_{x2}
9	p_{x3}	p_{x3}	p_{x3}	p_{x2}
p_y	p_{y1}	p_{y2}	p_{y3}	1

Tab. 7.8 Rozdělení pravděpodobnosti pro dvě nezávislé proměnné

Proměnná x \ Proměnná y	7	15	2	p_x
3	0,06	0,09	0,15	0,3
6	0,08	0,12	0,2	0,4
9	0,06	0,09	0,15	0,3
p_y	0,2	0,3	0,5	1

bude mít hodnotu i, za předpokladu, že náhodná proměnná Y má hodnotu j. Protože platí $p(x = i \mid y = j) = p_{xiyj}/p_{yj}$, má tabulka hledaných podmíněných pravděpodobností tvar jako tabulka 7.7c. Hodnoty v ní vyjadřují, že fixujeme-li proměnnou Y, je podmíněné rozdělení náhodné proměnné X stejné pro všechny hodnoty proměnné Y a toto rozdělení se shoduje s příslušným marginálním rozdělením proměnné X. Pojmenujme očekávanou hodnotu náhodné proměnné X při fixované hodnotě náhodné proměnné Y „podmíněná očekávaná hodnota". Z tabulky 7.7c je patrné, že podmíněné očekávané hodnoty náhodné proměnné X jsou stejné pro všechny hodnoty náhodné proměnné Y.

Pro ilustraci, jak se nezávislost projevuje na hodnotě korelačního koeficientu, vytvoříme z původního dvojrozměrného rozdělení proměnných X a Y v naší tabulce 7.7a nové rozdělení, aby proměnné byly nezávislé. Postupujeme tak, že zachováme podobu jednorozměrných marginálních pravděpodobnostních rozdělení proměnných X a Y a dopočítáme ostatní pravděpodobnosti podle předpisu pro nezávislost. Nové dvojrozměrné pravděpodobnostní rozdělení popisuje tabulka 7.8. Ukážeme, že se v tomto případě – to je u nezávislých náhodných proměnných – očekávaná hodnota jejich součinu rovná součinu jejich očekávaných hodnot, tedy $E(XY) = E(X)E(X)$. Protože $p_{ij} = p_i p_j$, platí:

$$E(XY) = \sum_i \sum_j x_i y_j p_{ij} = \sum_i \sum_j x_i y_j p_i p_j = \sum_i x_i p_i \sum_j y_j p_j = E(X)E(Y).$$

Důležitá je okolnost, že uvedený vztah lze zobecnit a dokázat i pro spojité náhodné proměnné.

Označili jsme $E(X) = \mu_x$ a $E(Y) = \mu_y$. Protože pro nezávislé proměnné platí:

$$E[(X - \mu_x)(Y - \mu_y)] = E(XY) - \mu_x\mu_y = E(X)E(Y) - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0,$$

plyne z toho, že kovariance σ_{xy}, a tedy i (teoretický) korelační koeficient ρ dvou stochasticky nezávislých náhodných proměnných jsou vždy rovné nule (čtenář se může přesvědčit přímým výpočtem pro hodnoty v poslední tabulce). Neplatí

to však obráceně. Nulová hodnota korelačního koeficientu neznamená vždy, že proměnné jsou stochasticky nezávislé. Pro jednu významnou třídu rozdělení však i toto obrácené tvrzení platí. Jedná se o tzv. dvojrozměrné normální rozdělení náhodných proměnných (X, Y). Jde o rozšíření pojmu normálního rozdělení, které jsme poznali v kap. 4.5.3, na systém dvou proměnných. Dvojrozměrné normální rozdělení je jednoznačně určeno průměry a rozptyly obou proměnných a jejich korelačním koeficientem ρ_{xy}. Zobecnění pro vícerozměrné normální rozdělení se provádí analogicky.

Uvedeme základní vzorce pro kovarianci a rozptyl náhodných proměnných, které se často používají při výkladu metody korelační a regresní analýzy. Symbol *Var* označuje rozptyl proměnné a Cov kovarianci dvou proměnných. Nechť a, b, c, d jsou konstanty a X, Y, Z, U náhodné proměnné, pak platí:

1. $\mathrm{Cov}(X, X) = Var(X)$,
2. $\mathrm{Cov}(aX + bY, cZ + dU) =$
   $ac\,\mathrm{Cov}(X, Z) + ad\,\mathrm{Cov}(X, U) + bc\,\mathrm{Cov}(Y, Z) + bd\,\mathrm{Cov}(Z, U)$,
3. $Var(aX+bY) = \mathrm{Cov}(aX+bY, aX+bY) = a^2Var(X)+b^2Var(Y)+2ab\,\mathrm{Cov}(X, Y)$,
4. $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$ pro nekorelované proměnné X a Y,
5. $\mathrm{Cov}(X, Y) = 0{,}25(Var(X + Y) - Var(X - Y))$.

7.2.3 Odhad a testování korelačního koeficientu

Popíšeme testy a intervaly spolehlivosti pro Pearsonův korelační koeficient. Tyto metody lze použít za předpokladu, že společné rozdělení obou proměnných lze modelovat dvojrozměrným normálním rozdělením nebo – jinak vyjádřeno – rozdělení obou proměnných je normální a jejich vztah je přibližně lineární.

Při posuzování, zda se vypočítaná hodnota korelačního koeficientu významně liší od nuly, použijeme tabulku IX z přílohy B, kde jsou hodnoty kritických mezí pro výběrový korelační koeficient v závislosti na rozsahu výběru. Jestliže bylo k dispozici n párových hodnot, má vypočtený korelační koeficient $n - 2$ stupňů volnosti. Přesahuje-li v absolutní hodnotě hodnotu v tabulce pro požadovanou hladinu významnosti, můžeme vztah považovat za prokázaný na dané hladině významnosti. Snadno nahlédneme, že s rostoucím počtem pozorování prokážeme statistickou významnost i velmi malého korelačního koeficientu.

Jestliže chceme testovat obecnější hypotézu $H_0: \rho_{xy} = \rho_0$ proti alternativě $H_1: \rho_{xy} \quad \rho_0$, kde $\rho_0 \quad 0$, musíme použít tzv. Fisherovu z-transformaci (arctanh – „arkustangens hyperbolický“):

$$z = \dot{z}(r) = \operatorname{arctanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

kde $\dot{z}$ označujeme Fisherovu transformaci. Touto transformací jsme rozšířili interval hodnot $-1 \le r \le +1$ na interval $-\infty \le z \le +\infty$. Nová proměnná má přibližně průměr

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

a směrodatnou odchylku

$$s_z = \sqrt{\frac{1}{n-3}},$$

takže pro test nulové hypotézy lze použít interval spolehlivosti ve tvaru

$$z - ts_z \le \mu_z \le z + ts_z,$$

kde t je kritická hodnota pro dvoustranný test zjištěná pomocí t-rozdělení o $n - 2$ stupních volnosti na odpovídající hladině významnosti.

Zpět do měřítka korelačního koeficientu převedeme oba krajní body intervalu pomocí inverzní transformace $\dot{z}^{-1}$:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Získáme tak interval spolehlivosti pro ρ_{xy}.

PŘÍKLAD 7.4

Testování hodnoty korelačního koeficientu

Test hypotézy H_0: $_{xy}$ = 0 5 proti H_1: $_{xy}$ 0 5 pro náš případ, kdy n = 10, r = 0 88, provedeme pomocí intervalu spolehlivosti s hladinou 0 95. Vypočteme nejdříve Fisherovu z-transformaci (protože $_0$ se nerovná nule)

$$z = \frac{1}{2}\ln\frac{1+r}{1\quad r} = \frac{1}{2}\ln\frac{1+0\,88}{1\quad 0\,88} = \frac{1}{2}\ln\frac{1\,88}{0\,12} = 1\,375$$

a směrodatnou odchylku

$$s_z = \frac{\overline{1}}{n\quad 3} = \frac{\overline{1}}{10\quad 3} = \frac{\overline{1}}{7} = 0\,37796 \quad 0\,378$$

Kritická hodnota t-rozdělení s 8 stupni volnosti má pro zvolenou hladinu spolehlivosti hodnotu 2,306. Interval spolehlivosti má tedy tvar

$$1\,375 \quad 2\,306 \quad 0\,378 \quad _z \quad 1\,375 + 2\,306 \quad 0\,378 \quad (0\,504; 2\,247)$$

Pomocí zpětné transformace $\dot{z}^{-1}$ převedeme tento interval do měřítka pro r a dostáváme (0 465; 0 977). Protože hodnota 0,5 leží v tomto intervalu, nemůžeme nulovou hypotézu zamítnout.

Pokud chceme testovat významnost rozdílu dvou korelačních koeficientů r_1 a r_2, získaných změřením dvojic proměnných ve dvou rozdílných skupinách r_1 a r_2, transformujeme oba korelační koeficienty Fisherovou transformací na hodnoty $\dot{z}_1$ a $\dot{z}_2$. Přibližně platný 95% interval spolehlivosti pro rozdíl Δ_z má pak tvar

$$\dot{z}_1 - \dot{z}_2 - 1{,}96\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}} \le \Delta_z \le \dot{z}_1 - \dot{z}_2 + 1{,}96\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}.$$

Na meze tohoto intervalu následně uplatníme zpětnou transformaci $\dot{z}^{-1}(\Delta_z)$, abychom získali interval spolehlivosti pro hodnotu $\Delta = \rho_1 - \rho_2$.

7.2.4 Problém třetí proměnné v korelační analýze

Korelace mezi dvěma proměnnými je někdy zavádějící a obtížně se interpretuje. Musíme zohlednit, že korelace dvou proměnných může být ovlivněna několika dalšími proměnnými. Mnoho atributů – jako např. výška, váha, síla, mentální schopnost, slovní zásoba, dovednost číst – roste v rozmezí 6 až 18 let s věkem. Korelace těchto proměnných budou určitě pozitivní. Když z nich však vyloučíme působení věku, pravděpodobně klesnou k nule. Vliv rušivého faktoru „věk" kontrolujeme dvěma způsoby. Buď měříme vztah proměnných pouze pro vybranou věkovou kategorii, nebo použijeme tzv. parciální korelační koeficient. Podívejme se na tuto druhou možnost podrobněji. Budeme uvažovat rušivý vliv pouze jedné proměnné, ačkoli postup výpočtu parciálního korelačního koeficientu lze zobecnit pro libovolné množství rušivých parametrů. Pro jeho užití platí stejné předpoklady a omezení jako v případě normálního korelačního koeficientu.

Předpokládáme lineární asociace mezi proměnnými X, Y a Z zachycené korelačními koeficienty ρ_{xy}, ρ_{xz}, ρ_{yz}. **Parciální korelační koeficient** $\rho_{xy.z}$ měřící sílu vztahu proměnných X, Y po vyloučení vlivu parametru Z vypočítáme podle vzorce:

$$\rho_{xy.z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}.$$

Jeho odhad pomocí naměřených hodnot (x_i, y_i, z_i) získáme dosazením výběrových hodnot korelačních koeficientů za teoretické. Výpočty a odhady parciálních korelačních koeficientů $\rho_{yz.x}$ a $\rho_{xz.y}$ dostaneme příslušnou cyklickou záměnou korelačních koeficientů.

Při testování nulové hodnoty parciálního korelačního koeficientu postupujeme stejně jako v případě jednoduchého korelačního koeficientu. Abychom však nalezli správnou kritickou mez, použijeme počet stupňů volnosti $n - 3$, kde n je počet trojic dat ve výběru.

PŘÍKLAD 7.5

Problém třetí proměnné v korelační analýze

V rámci screeningové akce bylo vyšetřeno 142 starších žen, u kterých byly také zaznamenávány parametry věk (v), krevní tlak (t) a koncentrace cholesterolu (c) v krvi. Pro ně se vypočítaly korelační koeficienty r_{vt} = 0 33; r_{vc} = 0 5; r_{tc} = 0 25. Protože zvýšené hodnoty krevního tlaku by mohly souviset se zvýšeným množstvím cholesterolu na stěnách cév, byla tato otázka důkladněji statisticky zkoumána. Parametry t a c s věkem rostou, tážeme se proto, zda jejich poměrně slabší korelace není způsobena efektem parametru věk. Vliv věku jako rušivého parametru se eliminuje zjištěním parciálního korelačního koeficientu $r_{tc\,v}$:

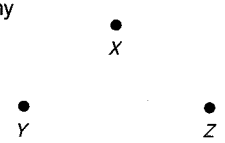
$$r_{tc\,v} = \frac{0\ 25 \quad 0\ 33 \quad 0\ 50}{(1 \quad 0\ 33^2)(1 \quad 0\ 50^2)} = 0\ 1$$

Pro 139 = (142 3) stupňů volnosti se nedá na hladině významnosti 5 % prokázat významnost tohoto korelačního koeficientu. Tímto statistickým zkoumáním jsme neukázali, že pro každou věkovou kategorii je krevní tlak pozitivně korelován s hladinou cholesterolu v krvi.

Výpočet parciálního korelačního koeficientu provádíme ve studiích, v nichž nás zajímá hlubší analýza vztahu mezi proměnnými a ověřování hypotéz o příčinných vztazích. Uvedeme v přehledu různé konfigurace korelačních vztahů proměnných X, Y, Z, přičemž budeme uvažovat i směr možné kauzality (obr. 7.4). Uvedená kauzální schémata implikují hodnoty korelačních koeficientů (v praktické analýze ovšem předpokládáme rovnost nule pouze přibližnou). Naopak to jednoznačně neplatí. Například $X \to Z \to Y$ má stejné koeficienty jako $Y \to Z \to X$. Stejně tak situace c) a d) jsou empiricky neodlišitelné. V těchto případech interpretujeme vztahy na základě dosavadních teoretických poznatků a pomocí základních kritérií pro ověřování kauzálního vztahu: a) silná asociace mezi proměnnými; b) prokázání této asociace v různých podmínkách (konzistence asociace); c) prokázaní změny hodnoty jedné proměnné při změně hodnoty druhé proměnné; d) působení proměnné klasifikované jako příčina předchází efektu v čase; e) existence věrohodného teoretického modelu působení.

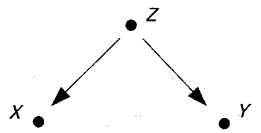
Obr. 7.4 Různé konfigurace korelačních vztahů

a) X, Y, Z jsou nekorelovány
$r_{xy} = 0$
$r_{yz} = 0$
$r_{xz} = 0$

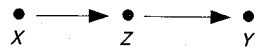


b) X a Y jsou dvě nekorelované příčiny pro proměnnou Z
$r_{xy} = 0$
$r_{yz} \neq 0$
$r_{xz} \neq 0$

c) Z je společná příčina X a Y
$r_{xy} \neq 0$
$r_{yz} \neq 0$
$r_{xz} \neq 0$
ale $r_{xy.z} = 0$



d) vztah X a Y je zprostředkován Z
$r_{xz} \neq 0$
$r_{yz} \neq 0$
$r_{xy} = r_{xz}r_{yz}$
ale $r_{xy.z} = 0$



7.2.5 Vliv dvou nezávisle proměnných na závisle proměnnou

Mnohonásobný koeficient korelace se používá v situacích, kdy chceme zjistit celkovou sílu vztahu mezi zvolenou proměnnou na jedné straně a několika dalšími (predikujícími) proměnnými $X_2, X_3, \ldots, X_k$ na straně druhé. Hodnotí se jím význam kumulativního vlivu více proměnných na zvolenou cílovou proměnnou. Mnohonásobný korelační koeficient, který pro tři proměnné značíme $\rho_{x.yz}$, je číselnou mírou možnosti predikce cílové proměnné X pomocí proměnných Y a Z:

$$\rho_{x.yz} = \sqrt{\frac{\rho_{xy}^2 + \rho_{xz}^2 - 2\rho_{xz}\rho_{xy}\rho_{yz}}{1 - \rho_{yz}^2}}.$$

Jeho odhad získáme dosazením příslušných výběrových korelačních koeficientů do tohoto vzorce. Nulovou hypotézu, že $\rho_{x.yz} = 0$, testujeme pomocí F-testu provedeného transformovanou hodnotou $r_{x.yz}$:

$$F = \frac{r_{x.yz}^2(n-3)}{2(1-r_{x.yz}^2)}$$

V tomto statistickém testu zjišťujeme, zda je hodnota F větší než kritická mez F-rozdělení se stupni volnosti 2 a $n - 3$. (V kapitole o mnohonásobné lineární regresní analýze se budeme tímto problémem zabývat podrobněji.)

PŘÍKLAD 7.6

Výpočet mnohonásobného korelačního koeficientu

Výzkum vycházel ze zkušenosti sportovní praxe, že osvojení motorické dovedností závisí komplexně na různých znacích jedince. Na závěr základního lyžařského kurzu pro šestnáctileté účastníky se změřil čas ve slalomu u 36 dívek. Také se u nich zjišťovaly další charakteristiky. V tabulce 7.9 uvádíme korelace dosaženého času ve slalomu a dvou vybraných parametrů z této studie, abychom mohli spočítat, jak silně dosažený čas na těchto parametrech závisí.

Mnohonásobný korelační koeficient mezi dosaženým časem ve slalomu jako cílovou proměnnou a prediktory Y a Z má hodnotu:

$$r_{X\,ZY} = \frac{\overline{0\ 34^2 + 0\ 46^2 \quad 2(\ 0\ 34)(0\ 46)(0\ 45)}}{1 \quad 0\ 45^2} = 0\ 77$$

Tab. 7.9 Korelační matice pro tři proměnné charakterizující skupinu účastníků lyžařského kurzu

	X	Y	Z
Čas ve slalomu (X)	1 00	0 34	0 46
Test rovnováhy (Y)	0 34	1 00	0 45
Test sociální úzkosti (Z)	0 46	0 45	1 00

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:

$$r_s = 1 \quad \frac{6 \quad 26}{10(100 \quad 1)} = 0\ 84$$

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 $\rho_s = 0$:

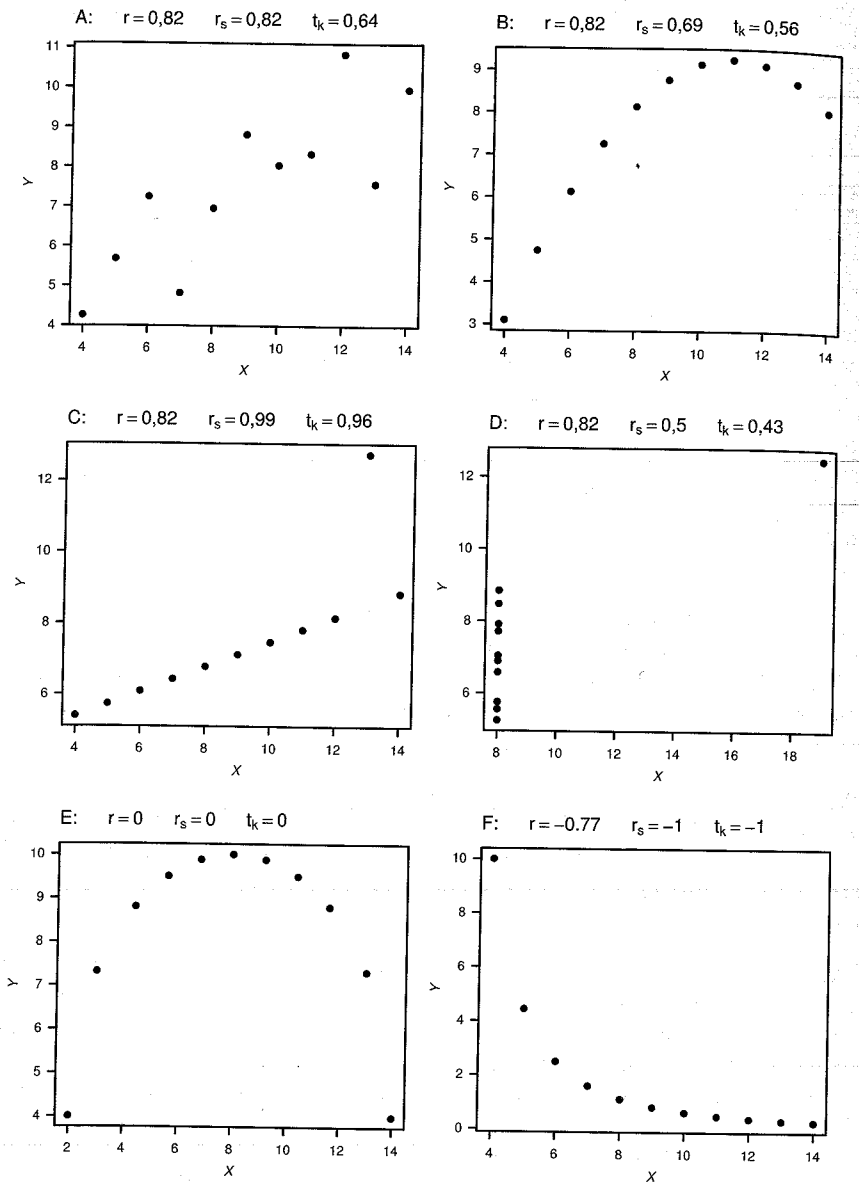
$$z = |r_s| \sqrt{n - 1}.$$

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).