

14. Korelace

14.1 Způsoby měření závislostí pro různé typy dat

Při práci se statistickými údaji se velmi často setkáváme s daty, která jsou tvořena dvojicemi, trojicemi hodnot. Složky takovýchto dat je možno vyšetřovat samostatně metodami předchozích kapitol, ztrácíme však vazbu mezi nimi. Právě metodami zobrazování dat, jejich účelnému seskupování do tabulek či jiných pomocných nástrojů se bude zabývat tato kapitola. Navíc si vymežíme i metody pro analýzu vazeb mezi složkami vybraných dat.

14.2 Dvourozměrná rozdělení četností

Předpokládejme, že v provedeném výběru je obsaženo n jednotek, u nichž sledujeme dva kvantitativní znaky x a y . Jejich konkrétním naplněním jsou tedy dvojice hodnot (x_i, y_j) , kde $i=1, \dots, p$ a $j=1, \dots, r$. Obecně není možné předpokládat, že čísla p a r jsou totožná, neboť každý ze znaků x a y může mít různý počet variant. Dohromady však tvoří n členný výběr. Označme absolutní četnosti jednotlivých kombinací variant x_i a y_j symbolem n_{ij} .

Seřaďme uspořádané dvojice (x_i, y_j) uvedené ve výběru podle nějakého pravidla (např. podle velikosti x_i a poté podle velikosti y_j). Potom v případě dvou znaků x a y budeme rozumět pod pojmem **dvourozměrného rozdělení absolutních četností** zobrazení, které přiřazuje jednotlivým uspořádaným dvojicím (x_i, y_j) , pro $i=1, \dots, p$ a $j=1, \dots, r$, hodnoty absolutních četností n_{ij} . Podobně bychom mohli provést takovéto přiřazení u hodnot relativních četností, kdy jednotlivé hodnoty n_{ij} budeme dělit součtem všech absolutních četností n .

Velmi podobně můžeme uvést i pojem **dvourozměrných intervalových rozdělení četností**. Nejdříve data vhodně uspořádáme (půjde o vhodné uspořádání kartézského součinu intervalů) a potom provedeme stejný postup jako v předchozím odstavci s tím, že místo uspořádaných dvojic budeme pracovat s uspořádanými dvojicemi intervalů.

Podobně můžeme zobecnit předchozí definice na případ, kdy u náhodného výběru sledujeme k kvantitativních znaků $z_1, \dots, z_k$. Potom hovoříme o k rozměrném rozdělení četností (absolutních nebo relativních) resp. o k rozměrném rozdělení intervalových četností.

Dvourozměrné rozdělení četností dvou kvantitativních znaků x a y lze velmi přehledně uvést v tabulce, která se obecně nazývá **korelační tabulka**. Popíšme si dále tuto tabulku: V prvním řádku, v němž je uveden postupně obsah jednotlivých sloupců se uvedou seřazené vzestupně všechny varianty jednoho ze znaků nalezené ve výběru; v prvním sloupci se postupně uvedou seřazené vzestupně všechny varianty druhého znaku nalezené ve výběru. Na průsečících takto popsanych sloupců a řádků se uvedou postupně absolutní (relativní) četnosti daných dvojic hodnot. Poslední sloupec (resp. poslední řádek) korelační tabulky obsahuje tzv. **podmíněné rozdělení absolutních (relativní) četností** druhého znaku (v případě posledního řádku prvního znaku).

Jestliže se vyšetřuje závislost dvou znaků, které nabývají velké množství hodnot, nebo předpokládáme, že jsou spojité, potom jednotlivé varianty znaků účelně seskupujeme do intervalů. Výsledná korelační tabulka je potom tabulkou dvourozměrného intervalového rozdělení absolutních (relativních) četností.

Tabulka 14.1 **Korelační tabulka absolutních četností**

x_i	y_j				n_i
	y_1	y_2	...	y_r	
x_1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
...					
x_p	n_{p1}	n_{p2}		n_{pr}	$n_{p.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

V této tabulce tedy vždy platí:

$$n_{i.} = \sum_{j=1}^r n_{ij} \quad (14.1)$$

$$n_{.j} = \sum_{i=1}^p n_{ij} \quad (14.2)$$

$$n = \sum_{i=1}^p n_{i.} = \sum_{j=1}^r n_{.j} = \sum_{i=1}^p \sum_{j=1}^r n_{ij} \quad (14.3)$$

Hodnoty n_{ij} se někdy nazývají také nepodmíněné četnosti a hodnoty $n_{i.}$ a $n_{.j}$ se nazývají nepodmíněné marginální četnosti.

Graficky můžeme zobrazovat takováto dvourozměrná data například sloupcovým grafem, v případě, že budeme dodržovat pravidla pro tvorbu histogramů, můžeme vytvořit také histogram pro dva znaky (někdy se nazývá stereogram). Existují i jiné způsoby zobrazení dvourozměrných dat, ale ty naráží především na komplikace s jejich konstrukcí, proto je uvádět nebudeme.

Jestliže chceme zobrazit dvourozměrné (nebo i vícerozměrné) rozdělení četností v případě kvalitativních znaků, můžeme využít pro zaznamenání údajů tabulku, která se shoduje s korelační tabulkou. Sledujeme – li vztahy mezi dvěma alternativními znaky nazýváme korelační tabulku v tomto případě **asociační tabulkou**. Jestliže alespoň jeden kvalitativní znak nabývá více než dvou variant nazýváme korelační tabulku **kontingenční**.

Uveďme dále několik příkladů korelačních tabulek.

Příklad 14.1

Výsledky písemných prací z matematiky a fyziky studentů jedné třídy jsou uvedeny dále v tabulce:

žák	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
M	3	1	3	2	4	3	1	4	2	1	2	3	2	3	1	3	4	3	2	5	4	2	3	4	3	4	4
F	2	2	4	2	5	1	1	4	5	2	1	4	3	2	2	3	3	2	4	4	3	1	1	3	5	3	4

Uveďte odpovídající korelační tabulku!

Řešení:

		MATEMATIKA					
		1	2	3	4	5	$n_{i.}$
FYZIK A	1	1	2	2	0	0	5
	2	3	1	3	0	0	7
	3	0	1	1	4	0	6
	4	0	1	2	2	1	6
	5	0	1	1	1	0	3
	$n_{.j}$	4	6	9	7	1	27

Uved'me pro úplnost stejnou tabulku pro relativní četnosti:

		MATEMATIKA					
		1	2	3	4	5	$n_{i.}$
FYZIKA	1	0,0 38	0,0 74	0,0 74	0 0	0 0	0,1 86
	2	0,1 11	0,0 38	0,1 11	0 0	0 0	0,2 60
	3	0 0	0,0 38	0,0 38	0,1 49	0 0	0,2 22
	4	0 0	0,0 38	0,0 74	0,0 74	0,0 38	0,2 22
	5	0 0	0,0 38	0,0 38	0,0 38	0 0	0,1 11
	$n_{.j}$	0,1 49	0,2 22	0,3 33	0,2 60	0,0 38	1

Nepřesnosti v dané tabulce jsou zaviněny zaokrouhlováním jednotlivých hodnot.

Příklad 14.2

U náhodně vybraných součástek se proměřovala jejich délka a výška. Pro každý z těchto rozměrů se určilo, zda je vyhovující (v normě) či ne . 239 součástek mělo oba rozměry správné, 14 mělo správně výšku a nesprávně délku, 60 mělo správně délku a nesprávně výšku a 7 mělo oba rozměry nesprávně. Převed'te tyto hodnoty do korelační tabulky.

Řešení:

Délka		výška		celkem
		správná	nesprávná	
	správná	239	60	299
	celkem	253	67	320

Jde o případ asociační tabulky.

Příklad 14.3

Sledovala se souvislost mezi rodinným stavem ženichů a nevěst v souboru československých snoubeneckých párů v roce 1966. Upravte tuto tabulku na tabulku s relativními četnostmi.

		rodinný stav nevěst			celkem
		svobodné	ovdovělé	rozvedené	
rodinný stav ženichů	svobodný	95 145	902	4 385	100 432
	ovdovělý	919	1 092	1 050	3061
	rozvedený	5 876	1 090	5 265	12 231
celkem		101 940	3 084	10 700	115 724

Řešení:

		rodinný stav nevěst			celkem
		svobodné	ovdovělé	rozvedené	
rodinný stav ženichů	svobodný	82,217%	0,779%	3,789%	86,786 %
	ovdovělý	0,794%	0,944%	0,907%	2,645%
	rozvedený	5,078%	0,942%	4,550%	10,569%
celkem		88,089%	2,665%	9,246%	100,000%

Jak jsme viděli v předchozích případech je možno zobrazovat pomocí různých variant korelační tabulky i data , která nejsou kvantitativního charakteru. Je možno provádět srovnávání dat ordinálních (záleží nám na pořadí) např. pořadí při vyhodnocení otázky , pořadí subjektivních pocitů atd. Dále je možno zobrazovat také nominálně měřené znaky v kontingenčních tabulkách.

Slovo **asociace** znamená v našich pojmech sloučení, složení. Slovo **kontingence** znamená vazba.

Kromě popisu dat jen prostou korelační tabulkou , je naší snahou většinou získat doplňující informace o vazbách mezi jednotlivými znaky (korelační analýza) a o způsobu vyjádření těchto vazeb (regresní analýza). Těmito pojmy se budeme nyní u dvou znaků dále zabývat.

Z pohledu vztahu dvou znaků již známe z prostředí např. středoškolské matematiky funkční závislost dvou veličin. Známe – li stranu čtverce, je možno určit jednoznačně jeho obvod, známe – li vklad , počet období, úrokovou míru , je možno určit výši vkladu včetně úroků. V prostředí statistiky se však setkáváme a vyšetřujeme jiné typy vztahů.

Definice 14.1

Řekneme, že znak y (náhodná veličina Y) je statisticky závislý na znaku x (náhodné veličině X), jestliže změna hodnoty znaku x má za následek změnu podmíněného rozdělení znaku y (náhodné veličiny Y).

V praxi není většinou podmíněné rozdělení znaku y známo, proto se snažíme získat o něm představu pomocí nástrojů regresní analýzy.

Definice 14.2

Řekneme, že znak y (náhodná veličina Y) je korelačně závislý na znaku x (náhodná veličina X), jestliže změna hodnoty znaku x má za následek změnu podmíněné střední hodnoty rozdělení znaku y (náhodné veličiny Y).

Zůstávají – li podmíněné průměry závisle proměnné konstantní , i když se hodnoty nezávisle proměnné mění jakkoli , považuje se závisle proměnná za korelačně nezávislou na příslušných nezávisle proměnných.

Podle příslušných definic je zřejmé, že pojem korelační závislosti je slabší než pojem statistické závislosti. Korelační závislost je vždy i statistickou závislostí, protože změna podmíněného průměru znamená i změnu podmíněného rozdělení. Na druhou stranu korelační nezávislost ještě nemusí znamenat statistickou nezávislost. To že se nemění podmíněný průměr nemusí znamenat, že se nemění podmíněné rozdělení. Statistická závislost může být i korelační nezávislostí, jestliže se změny podmíněného rozdělení projeví ve změnách jiných popisných hodnot než v podmíněném průměru. Statistická nezávislost je v každém případě také korelační nezávislostí, neboť neměnnost podmíněného rozdělení znamená současně konstantnost podmíněného průměru.

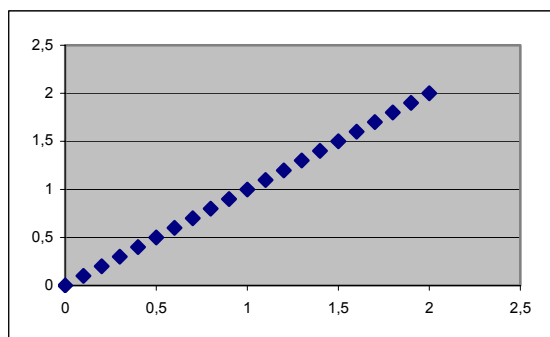
Korelaci náhodných veličin chápeme vzájemnou korelační závislost uvažovaných náhodných veličin. Přitom rozlišujeme :

1. Jednoduchou korelaci – korelaci dvou náhodných veličin;
2. Mnohonásobnou regresi – korelaci jedné náhodné veličiny a skupiny dvou a více náhodných veličin.

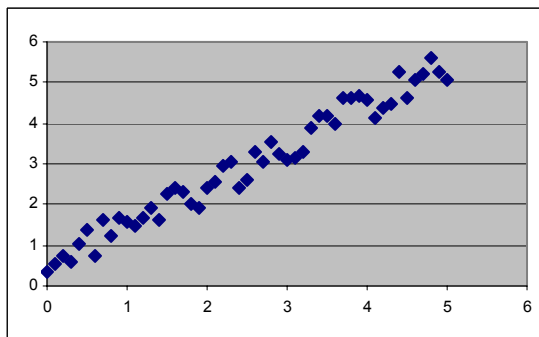
Důležité je pochopit jaký je vztah mezi korelační závislostí na jedné straně a kauzální závislostí na straně druhé. Jestliže jsou dvě náhodné veličiny korelačně závislé, pak to znamená, že mezi těmito náhodnými veličinami může existovat kauzální závislost. Nelze ale rozlišit zda jde o kauzální závislost bezprostřední , kdy změny jedné veličiny podmiňují změny druhé, nebo o kauzální závislost zprostředkovanou . Existence korelační závislosti dvou náhodných veličin nemůže být důkazem toho, že mezi nimi existuje kauzální závislost.

Na dalších obrázcích jsou uvedeny různé modelové situace korelace dvou náhodných veličin (dvou znaků).

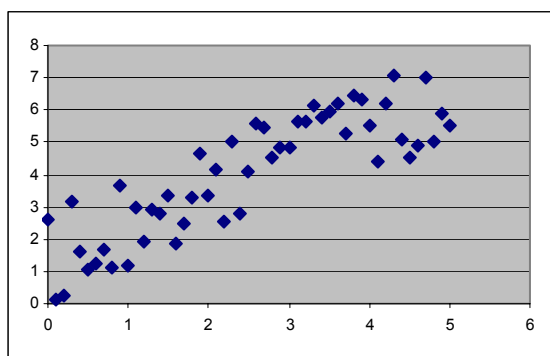
Obr. 14.1 – Extrémně kladná korelace



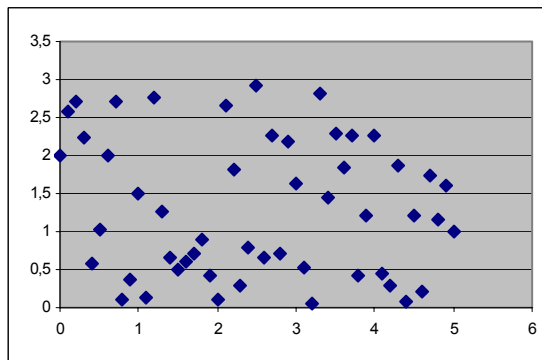
Obr. 14.2 – Silná kladná korelace



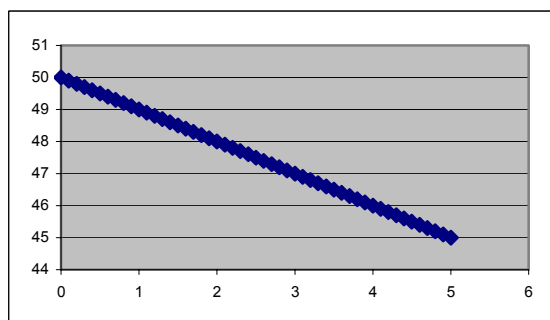
Obr 14.3. – Slabá kladná korelace



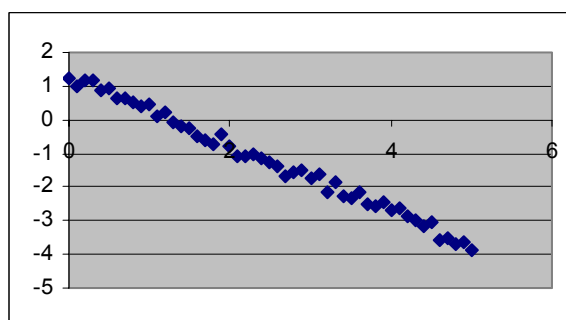
Obr 14.4 – Korelace blízká nule



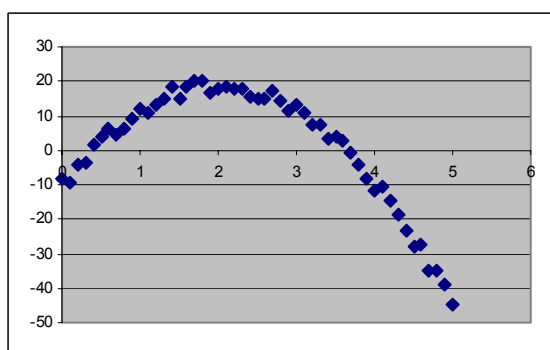
Obr 14.5 – Extrémně záporná korelace



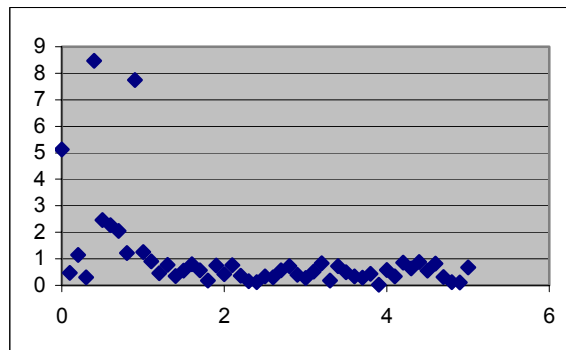
Obr 14.6 – Silná záporná korelace



Obr 14.7 – Nelineární korelace



Obr 14.8. – Nelineární korelace



Z jednotlivých obrázků je patrné, že některá vyjádření vztahu mezi náhodnými veličinami X a Y (resp. znaky x a y) mohou mít i podobu skoro funkční , dokonce i tehdy když dané hodnoty spolu příliš nekorelují (např. Obr 14.7) , ale mohou mít podobu lineárního vztahu (Obr. 14.1), když spolu korelují velmi.

14.3 Míry korelace dvourozměrných statistických souborů

Jak jsme již uvedli v předchozí části je jednou z nejdůležitějších charakteristik vícerozměrných statistických souborů charakteristika vazby mezi jednotlivými znaky. Podstatné pro naše další šetření bude to , zda jsou data kvantitativní povahy, nebo jsou ordinální nebo nominální. V následující tabulce jsou uvedeny jednotlivé typy charakteristik míry korelace pro různé druhy dat.

Tabulka 14.2 Míry korelace

		Proměnná Y		
		kvantitativní	ordinální	nominální
Proměnná X	kvantitativní	korelační koeficient r		
	ordinální		koeficient pořadové korelace r_s	
	nominální	biseriální koeficient r_{bis}		koeficient Φ asociační koeficient Q kontingenční koeficient C

14.3.1 Výběrový korelační koeficient

V dalším textu se budeme snažit osvětlit pojmy uvedené v předchozí tabulce 14.2. Jestliže jsou obě veličiny povahově kvantitativního charakteru používáme k měření výběrový korelační Pearsonův koeficient r . Jde o bezrozměrnou veličinu, která může nabývat hodnot mezi -1 a 1. Pomocí toho koeficientu měříme většinou sílu lineární vztahu mezi znaky x a y . Jestliže nabývá krajních mezních hodnot -1 nebo 1 můžeme vztah mezi znaky vyjádřit pomocí funkčního lineárního vztahu. Pro hodnotu $r=-1$ při rostoucích hodnotách x hodnoty y klesá a při hodnotě $r=1$ při rostoucích hodnotách x hodnota y klesá. Pro hodnotu $r=0$ můžeme vyloučit lineární vztah mezi znaky x a y , ale to neznamená, že mezi nimi nemůže vztah být viz např. Obr. 14.8 nebo Obr. 14.7. Při užívání tohoto koeficientu je zapotřebí mít na zřeteli klasické předpoklady lineárního modelu, tedy především normalitu a to, že pro libovolnou hodnotu x_i má náhodná hodnota Y střední hodnotu, která je dána příslušnou hodnotou na regresní přímce. Více o těchto předpokladech se dozvíme v kapitole 13. o statistické regresi. Mějme dále na paměti, že skutečnou hodnotu korelačního koeficientu ρ neznáme a můžeme ji odhadovat právě pomocí r .

Při výpočtu korelačního koeficientu r se používají hodnoty odhadu směrodatné odchylky s_x a s_y ,

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y} \quad (14.4),$$

tento způsob výpočtu využívá hodnoty počítané přímo z korelační tabulky. Pro některé výpočty je ale jednodušší používat následující vzorec

$$r = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \quad (14.5),$$

kde výše uvedené hodnoty jsou počítány podle těchto principů:

$$\overline{x \cdot y} = \frac{\sum_{i=1}^n X_i Y_i}{n} \quad (14.6),$$

$$\bar{x} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{a} \quad \bar{y} = \frac{\sum_{i=1}^n Y_i}{n} \quad (14.7),$$

$$\sigma_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})^2} \quad (14.8),$$

$$\sigma_y = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (14.9).$$

Příklad 14.3

Byly zjišťovány reakční doby řidičů v určité situaci před stresem a po stresu. Zjištěné údaje jsou uvedeny v tabulce níže. Zjistěte hodnotu korelačního koeficientu!

Tabulka 14.3

	Čas		x_i^2	y_i^2	$x_i \cdot y_i$
	Po stresu	Před stresem			
A	3	3	9	9	9
B	7	5	49	25	35
C	11	7	121	49	77
D	14	6	196	36	84
E	15	9	225	81	135
	$\sum_{i=1}^n x_i = 50$	$\sum_{i=1}^n y_i = 30$	$\sum_{i=1}^n x_i^2 = 600$	$\sum_{i=1}^n y_i^2 = 200$	$\sum_{i=1}^n x_i y_i = 340$
	$\bar{x} = 10$	$\bar{y} = 6$	$\overline{x^2} = 120$	$\overline{y^2} = 40$	$\overline{x \cdot y} = 68$

Řešení:

Nejdříve určíme hodnotu r podle vztahu (5.4), potřebujeme ještě určit hodnoty s_x a s_y .

Tabulka 14.4

	Čas		$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
	Po stresu	Před stresem			
A	3	3	-7	-3	21
B	7	5	-3	-1	3
C	11	7	1	1	1
D	14	6	4	0	0
E	15	9	5	3	15
	$\sum_{i=1}^n x_i = 50$	$\sum_{i=1}^n y_i = 30$	$\sum_{i=1}^n (x_i - \bar{x}) = 0$	$\sum_{i=1}^n (y_i - \bar{y}) = 0$	$\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 40$
	$\bar{x} = 10$	$\bar{y} = 6$			

$$s_x = 5 \text{ a } s_y = \sqrt{5}$$

$$r = \frac{40}{4.5 \cdot \sqrt{5}} = \frac{2}{\sqrt{5}} = \frac{2 \cdot \sqrt{5}}{5} \cong 0,89$$

Počítejme nyní podle vzorce (14.5). Musíme nejdříve určit hodnoty σ_x a σ_y . Tyto hodnoty jsou $\sigma_x = \sqrt{20} = 2 \cdot \sqrt{5}$ a $\sigma_y = 2$.

$$r = \frac{68 - 6 \cdot 10}{2 \cdot \sqrt{5} \cdot 2} = \frac{8}{4 \cdot \sqrt{5}} = \frac{2 \cdot \sqrt{5}}{5} \cong 0,89.$$

Příklad 14.4

Vypočítejte hodnotu korelačního koeficientu z údajů uvedených v následující korelační tabulce:

Tabulka 14.4

$\begin{matrix} x \\ y \end{matrix}$	3	4	5	6	7	8	$n_{i.}$	$x_i n_{i.}$	$x_i^2 n_{i.}$
4	2				1	2	5	20	80
5	6		1		1	4	12	60	300
6	2	2	1	1	3	3	12	72	432
7		3	3	5	5	1	17	119	833
8		5	5	5	2		17	136	1088
9		3	6	3			12	108	972
10		2	9	2			13	130	1300
$n_{.j}$	0	1	5	2	6	1	2	10	88
$y_j n_{.j}$	30	60	125	96	84	80	475		
$y_j^2 n_{.j}$	90	240	625	576	588	640	2 759		

Řešení:

Počítáme hodnotu výběrového korelačního koeficientu r podle (5.5). Potřebné hodnoty určíme z předchozí tabulky. Tedy

$$\bar{x} = \frac{645}{88} = 7,329; \bar{y} = \frac{475}{88} = 5,398; \sigma_x = 1,776; \sigma_y = 1,489. \quad \text{Z těchto hodnot již}$$

můžeme spočítat výběrový korelační koeficient r :

$$r = -0,144$$

Výsledná hodnota je velmi nízká, nemůžeme tedy hovořit o lineárním vztahu mezi X a Y.

Pro vyšetřování statistických hypotéz o nulovosti korelačního koeficientu je důležitá následující věta.

Věta 14.1

Necheť $(X_1, Y_1), \dots, (X_n, Y_n)$ je výběr z dvojrozměrného normálního rozdělení, které má kladné rozptyly a korelační koeficient $\rho = 0$. Potom náhodná veličina

$$T = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$$

je typu studentovo rozdělení s $n-2$ stupni volnosti t_{n-2} .

Důkaz : Viz [1].

Pomocí tohoto tvrzení můžeme testovat hypotézu $H_0 : \rho = 0$ proti alternativní hypotéze $H_1 : \rho \neq 0$. Vypočteme totiž nejdříve hodnotu výběrového korelačního koeficientu r , dále stanovíme výše uvedenou hodnotu T , v případě, kdy $|T| \geq t_{n-2}(\alpha)$, zamítáme H_0 na hladině významnosti α .

Jestliže použijeme toto tvrzení na naše data získáme hodnotu $T = -1,349466$, na hladině významnosti 0,05 je roven kvantit t_{86} číslu 1,98, protože $|T| < 1,98$, nemůžeme zamítnout tvrzení o nezávislosti hodnot X a Y.

14.3.2 Koeficient determinace

V případě lineární regrese (ale nejen v tomto případě) můžeme vyjádřit těsnost vztahu pomocí mezi náhodnou veličinou Y a veličinou x také pomocí poměru variability podmíněné lineární regrese a variabilitou způsobenou náhodnými vlivy.

$$\text{Podle vztahu (13.23) je } S_r = \frac{S_{xx} \cdot S_{YY} - S_{xY}^2}{S_{xx}}.$$

Měříme – li odhad variability Y máme

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n ((Y_i - a - b \cdot x_i) - (-a - b \cdot x_i + \bar{Y}))^2 = S_r + \sum_{i=1}^n (b \cdot (x_i - \bar{x}))^2 - 2 \cdot \sum_{i=1}^n (Y_i - a - b \cdot x_i) \cdot (-a - b \cdot x_i + \bar{Y}) = S_r + b^2 \cdot S_{xx} - 0 = S_r + \left(\frac{S_{xY}}{S_{xx}} \right)^2 \cdot S_{xx}.$$

Tedy

$$S_r = S_{YY} - \frac{S_{xY}^2}{S_{xx}} \Rightarrow R^2 = 1 - \frac{S_r}{S_{YY}} = \frac{S_{xY}^2}{S_{xx} \cdot S_{YY}}.$$

Hodnota R^2 se nazývá **koeficient determinace**, jeho hodnota se pohybuje mezi 0 a hodnotou 1. Koeficient determinace vyjadřuje poměr mezi variabilitou skutečnou (S_{YY}) a variabilitou vypočítávanou (S_r). Je zřejmé, že v lineárním případě je koeficient determinace roven druhé mocnině korelačního koeficientu.

14.3.3 Index determinace

V případě jiných než lineárních regresí nebo mnohonásobných regresí je zapotřebí zabezpečit také míru těsnosti vazby mezi náhodnou veličinou Y a hodnotami x. Stejně jako v předchozí části je celková variabilita náhodné veličiny Y rovna součtu variability vyjádřené regresním vztahem a variability náhodného chování. Definujeme proto podobně jako v předchozím článku

$$I_{xY}^2 = \frac{S_{\hat{Y}}}{S_{YY}},$$

kde $S_{\hat{Y}}$ znamená hodnotu rozptylu empirické regresní funkce a S_{YY} hodnotu rozptylu závisle proměnné Y.

Častěji se k vystižení závislosti používá odmocnina z indexu korelace, tato hodnota se nazývá **index determinace**.

Z obou předchozích hodnot vyplývá, že nejsou symetrické, platí totiž

$$I_{xY} \neq I_{Yx}.$$

14.3.4 Spearmanův korelační koeficient

V případě dvourozměrného souboru kvalitativních údajů, které jsou po složkách ordinálního typu, je možno zjistit stupeň závislosti těchto dvou znaků. K měření takovýchto závislostí se používá Spearmanův korelační koeficient (někdy nazývaný též koeficientem pořadové korelace), který je založený na pořadích jedinců uspořádaných podle velikosti vzhledem k oběma vyšetřovaným znakům. Každému jedinci tedy přiřadíme dvojici pořadí – Q (pořadí podle prvního znaku X) a R (pořadí podle druhého znaku Y).

Jestliže budou hodnoty pořadí Y vzrůstat stejně jako hodnoty X, budou pořadí R a Q stejná. Jestliže bude hodnota pořadí znaku Y s rostoucím pořadím znaku X klesat, budou pořadí obou znaků opačná, tedy

$$R = n - Q + 1. \quad (14.10)$$

Jestliže však budou hodnoty pořadí **R** a **Q** obou znaků libovolná potom očekáváme, že oba znaky budou nezávislé.

Pro n pozorovaných dvojic ve výběru určíme **Spearmanův** korelační koeficient pomocí diferencí pořadí $d_i = Q_i - R_i$ takto

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)} \quad (14.11).$$

Při shodném pořadí jsou hodnoty všech $d_i = 0$, tedy $r_s = 0$. Je-li pořadí opačné, použijeme výraz (5.10) a $d_i = Q - n + Q - 1 = 2 \cdot Q - n - 1$, hodnota Q postupně nabývá všech hodnot od 1 do n . Zjistíme jaké hodnoty potom nabývá výraz (14.11). Dosadíme přímo do (14.11) a získáváme

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n (2i - n - 1)^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot \left(\frac{n^3}{3} - \frac{n}{3} \right)}{n \cdot (n^2 - 1)} = 1 - 2 = -1.$$

Při ostatních případech nabývá r_s hodnoty ležící mezi těmito mezními hodnotami tedy

$$-1 < r_s < 1 \quad (14.12)$$

Pro hodnoty r_s blízké nule můžeme usuzovat, že pořadí **R** a **Q** jsou náhodně zpřeházena a mezi znaky **X** a **Y** není závislost. Pro hodnoty $n \geq 31$ je možné využít vztahu mezi rozdělením $N(0,1)$ a Spearmanovým koeficientem

$$Z = r_s \cdot \sqrt{n-1} \quad (14.13).$$

Tento vztah využijeme v části věnované statistickým hypotézám.

Příklad 14.5

Při přijímacím řízení se provádělo hodnocení komisí a hodnocení speciálním programem. Na základě údajů o deseti studentech rozhodněte o tom, zda jsou obě hodnocení závislá.

Tabulka 14.5

Student	A	B	C	D	E	F	G	H	I	J
Hodnocení komisí	4	6	1	5	10	2	7	3	9	8
Hodnocení programem	1	3	5	7	8	4	6	2	10	9
Diference pořadí	3	3	-4	-2	2	-2	1	1	-1	-1
Čtverec difference	9	9	16	4	4	4	1	1	1	1

Řešení:

Sečteme hodnoty čtverců diferencí $\sum_{i=1}^{10} d_i^2 = 50$ a dosadíme do vztahu (5.11).

Dostáváme

$$r_s = 1 - \frac{6.50}{10 \cdot (100 - 1)} = 1 - \frac{300}{990} = \frac{23}{33} \doteq 0,6970.$$

Hodnota korelačního koeficientu r_s ukazuje na určitou míru závislosti mezi hodnocením komisí a hodnotitelským programem.

14.3.5 Míra asociace

V tomto článku budeme vyšetřovat závislosti dvou náhodných veličin X a Y , které jsou nabývají alternativní (dvouhodnotové). Předpokládáme, že ba znaky X a Y jsou kvalitativní povahy.

Definice 14.3

Řekneme, že náhodné veličiny X a Y jsou asociačně závislé, jestliže v části výběru, který se skládá z jednotek s určitou hodnotou jedné náhodné veličiny, je relativně více nebo méně jednotek s určitou variantou druhé náhodné veličiny.

Z této definice je možno odvodit dva mezní případy:

- Případ, kdy všechny jednotky výběru, které mají určitou variantu jedné náhodné veličiny, mají i určitou variantu druhé náhodné veličiny. Tomuto říkáme **úplná asociační závislost**.
- Jestliže v části výběru, která se skládá z jednotek s určitou variantou jedné náhodné veličiny, je relativně stejný počet jednotek s určitou druhé náhodné veličiny jako v části výběru, který se skládá z jednotek nemajících uvažovanou variantu první náhodné veličiny pak nazveme obě veličiny jako **asociačně nezávislé**.

Asociací tedy rozumíme oboustrannou závislost mezi alternativními náhodnými veličinami kvalitativní povahy. Z hlediska obecného se zkoumá asociace jen dvou znaků (párová) nebo asociace více znaků (mnohonásobná). V tomto textu se budeme zabývat jen jednoduchou asociací. V následující tabulce si uvedeme obecný příklad tzv. asociační tabulky:

Tabulka 14.6

Hodnoty X	Hodnoty Y		Celkem
	+	-	
+	n_{11}	n_{12}	$n_{1.}$
-	n_{21}	n_{22}	$n_{2.}$
Celkem	$n_{.1}$	$n_{.2}$	n

Základní mírou závislosti jednoduché asociace je **koeficient asociace**

$$r_A = \frac{n_{11} \cdot n_{22} - n_{1.} \cdot n_{2.}}{\sqrt{n_{1.} \cdot n_{2.} \cdot n_{.1} \cdot n_{.2}}} \quad (14.14).$$

Podobně jako u ostatních koeficientů, kterými se snažíme měřit závislost náhodných veličin nabývá hodnot z intervalu od -1 ; 1 . Vyšetřeme si krajní případy.

- a) Jestliže jsou hodnoty $n_{12} = n_{21} = 0$, potom $r_A = 1$. jde o případ úplné asociační závislosti.
- b) Jestliže jsou hodnoty $n_{11} = n_{22} = 0$, potom $r_A = -1$. V tomto případě jde opět o úplnou asociační závislost
- c) Jestliže výraz $D = n_{11} \cdot n_{22} - n_{12} \cdot n_{21} = 0$, potom je hodnota $r_A = 0$ a dané náhodné veličiny jsou asociačně nezávislé.

Příklad 14.6

Byly vyšetřovány vztahy mezi vlastnictvím automobilu a ochotou jezdit hromadnou dopravou do zaměstnání. Výsledné hodnoty šetření jsou uvedeny v následující tabulce 14.7. Zjistěte míru asociace těchto veličin.

Tabulka 14.7

Ochota jezdit hromadnou dopravou	Vlastnictví automobilu		Celkem
	ano	ne	
ano	56	283	339
ne	312	35	347
Celkem	368	318	686

Řešení:

Použijeme vzorec (14.14). Po dosazení příslušných hodnot získáme $r_A = -0,73585$. Uvedené vztahy ukazují na střední míru závislosti mezi vlastnictvím automobilu a ochotou jezdit hromadnou dopravou.

14.3.6 Míra kontingence

Pokud vyšetřujeme dvě náhodné veličiny kvalitativního typu, které nejsou alternativní (nenabývají jen dvou hodnot), nemůžeme samozřejmě použít předchozí míru asociace. Pro oboustrannou kvalitativní závislost náhodných veličin, které mají více než dvě varianty, se užívá pojem **kontingence**.

Pro měření těsnosti kontingence dvou náhodných veličin se používá celá řada měř. Předpokládejme, že náhodná veličina X nabývá s variant a náhodná veličina Y nabývá r variant. K nejznámějším patří **Pearsonův koeficient kontingence**

$$c = \sqrt{\frac{\chi^2}{n + \chi^2}}, \quad (14.15)$$

kde

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{n \cdot \left(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{n} \right)^2}{n_{i.} \cdot n_{.j}} \quad (14.16)$$

Výše uvedené symboly jsou definovány stejně jako v předchozí části. Při úplné nezávislosti je hodnota Pearsonova koeficientu kontingence nulová, určitým nedostatkem je to, že ani při úplné závislosti sledovaných náhodných veličin nenabývá hodnoty 1.

Užíváme proto někdy tzv. **Čuprovův koeficient kontingence**

$$K = \frac{\chi^2}{n \cdot \sqrt{(r-1) \cdot (s-1)}}, \quad (14.17)$$

v němž má χ^2 stejný význam jako v (5.15).

Příklad 14.7

V jisté anketě odpovídali respondenti na dvě otázky, každá z nich měla celkem tři možnosti odpovědí a, b, c. Zjistěte, zda mezi oběma otázkami existuje souvislost. Zjištěná data jsou uvedena v tabulce 5.2.8.

Tabulka 14.8

Odpověď na otázku č.1	Odpověď na otázku č.2			Celkem
	a	b	c	
a	15	53	20	88
b	18	59	32	109
c	14	10	7	31
Celkem	47	122	59	228

Řešení:

Podle vzorce (14.16) převedeme tabulku 14.8 na tvar, kdy v jednotlivých buňkách tabulky budou hodnoty vypočtené ze vztahu (14.16).

0,543639	0,742339	0,337415
0,888974	0,007822	0,510292
9,061593	2,616276	0,130186

Takovéto hodnoty sečteme a získáme hodnotu výrazu $\chi^2 = 14,8385$. Tuto hodnotu dosadíme nejdříve do vztahu (5.15) a získáváme $c = 0,2472$. Hodnota Pearsonova koeficientu je velmi malá, můžeme proto předpokládat, že se dané odpovědi velmi odlišují.

Čuprovův koeficient kontingence dává hodnotu $K = 0,0325$. Podle něho můžeme zamítnout vztah mezi jednotlivými odpověďmi.