6. Vícerozměrný statistický soubor pokračování

6.1 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 6.2.1 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

6.1.1 Korelační koeficient

V dalším textu se budeme snažit osvětlit pojmy uvedené v předchozí tabulce 6.2.1

Jestliže jsou obě veličiny povahově kvantitativního charakteru používáme k měření 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** vylučujeme lineární vztah mezi znaky **x** a **y** , ale to neznamená, že mezi nimi nemůže vztah být viz např. Obr. 6.8 nebo Obr. 6.7. 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 o statistické regresy

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

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{(n-1).s_x.s_y} \qquad (6.1),$$

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

$$r = \frac{\overline{x.y} - \bar{x}.\bar{y}}{\sigma_X.\sigma_Y} \qquad (6.2),$$

kde výše uvedené hodnoty jsou počítány podle těchto principů:

$$\overline{x.y} = \frac{\sum_{i=1}^{n} X_i Y_i}{n} \qquad (6.3),$$

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{a} \quad \bar{y} = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad (6.4),$$

$$\sigma_X = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad (6.5),$$

$$\sigma_Y = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (6.6).$$

Příklad 6.2.1

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 6.2.2

	Čas		x_i^2	y_i^2	$x_i.y_i$
	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
	$\sum_{i=1}^{n} x_i = 50$	$\sum_{i=1}^{n} y_i = 30$	$\sum_{i=1}^{n} x_i^2 = 600$	$\sum_{i=1}^{n} y_i^2 = 200$	$\sum_{i=1}^{n} x_i y_i = 340$
	$\bar{x} = 10$	$\bar{y} = 6$	$\overline{x^2} = 120$	$\overline{y^2} = 40$	$\overline{x.y} = 68$

Řešení:

Nejdříve určíme hodnotu **r** podle vztahu (6.2), potřebujeme ještě určit hodnoty s_x a s_y.

Tabulka 6.2.3

	Čas		$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}).(y_i - \bar{y})$
	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
	$\sum_{i=1}^{n} x_i = 50$	$\sum_{i=1}^{n} y_i = 30$	$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right) = 0$	$\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y}) = 40$
	$\bar{x} = 10$	$\bar{y} = 6$			

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

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

Počítejme nyní podle vzorce (6.5) . Musíme nejdříve určit hodnoty σ_x a σ_y . Tyto hodnoty jsou $\sigma_x = \sqrt{20} = 2.\sqrt{5}$ a $\sigma_y = 2$.

$$r = \frac{68 - 6.10}{2.\sqrt{5}.2} = \frac{8}{4.\sqrt{5}} = \frac{2.\sqrt{5}}{5} \cong 0,89 .$$

Příklad 6.2.2
Vypočítejte hodnotu korelačního koeficientu z údajů uvedených v následující korelační tabulce:

Tabulka 6.2.4

x y	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
$n_{.j}$	10	15	25	16	12	10	88	645	5 005
$y_j n_{.j}$	30	60	125	96	84	80	475		
$y_j^2 n_{.j}$	90	240	625	576	588	640	2 759		

Řešení:
Počítáme hodnotu korelačního koeficientu **r** podle (6.2). Potřebné hodnoty určíme z předchozí tabulky. Tedy

$\bar{x} = \frac{645}{88} = 7,329; \bar{y} = \frac{475}{88} = 5,398; \sigma_X = 1,776; \sigma_Y = 1,489$. Z těchto hodnot již můžeme spočítat korelační koeficient **r** :

r = -0,144

Výsledná hodnota je velmi nízká, nemůžeme tedy hovořit o lineárním vztahu mezi X a Y. V kapitole o statistických testech se seznámíme s možností testovat pomocí této spočtené hodnoty výrok o nezávislosti náhodných veličin X a Y.

Platí totiž tvrzení , které za předpokladu , že náhodné veličiny X a Y jsou typu normální rozdělení a korelační koeficient těchto náhodných veličin je roven nule . Nechť dále je $n \geq 3$. Potom je náhodná veličina

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

typu studentova rozdělení s (n – 2) stupni volnosti .

Pomocí tohoto tvrzení je možno sestrojit parametrický test na nezávislost náhodných veličin X a Y . Podstatné v tomto tvrzení je předpoklad o normalitě obou náhodných veličin. Abychom nemuseli složitě zjišťovat hodnotu testových statistik , jsou většinou kritické hodnoty pro tento test uváděny v tabulkách jako například v tab. korelačních koeficientů.

Pokud bychom pokračovali v našem příkladu , nalezneme kritickou hodnotu pro n = 88 jako přibližně 0,208 , protože naše naměřená hodnota je absolutně menší než kritická hodnota nemůžeme nezávislost obou statistických znaků vyloučit.

6.1.2 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

$$\mathbf{R} = n - \mathbf{Q+1}. \qquad (6.7)$$

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 $d_i = \mathbf{Q_i} - \mathbf{R_i}$ takto

$$r_S = 1 - \frac{6.\sum_{i=1}^{n} d_i^{\,2}}{n.\left(n^2 - 1\right)} \qquad (6.8).$$

Při shodném pořadí jsou hodnoty všech $d_i = 0$, tedy $\mathbf{r_S} = 0$. Je – li pořadí opačné, použijeme výraz (6.7) a $d_i = Q - n + Q - 1 = 2 . Q - n - 1$, hodnota Q postupně nabývá všech hodnot od 1 do n. Zjistěme jaké hodnoty potom nabývá výraz (6.11) . Dosaďme přímo do (6.7) a získáváme

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

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

$$-1 < r_S < 1 \qquad (6.9)$$

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

$$Z = r_S.\sqrt{n-1} \qquad (6.10).$$

Tento vztah využijeme v části věnované statistickým hypotézám. Abychom nemuseli složitě zjišťovat hodnotu testových statistik , jsou většinou kritické hodnoty pro tento test uváděny v tabulkách jako například v tab. spearmanových korelačních koeficientů.

Příklad 6.2.3

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 6.2.5

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 diference	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 (6.8). Dostáváme

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

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

Podobně jako v předchozí části budeme pomocí statistického testu ověřovat nezávislost hodnocení pro hodnotu n = 10 nalezneme kritickou hodnotu jako 0,6364 pro hladinu významnosti 0,05 . Tedy na této hladině významnosti bychom museli připustit závislost mezi oběma hodnoceními , pokud bychom ale zvolili hladinu významnosti rovnu 0,01 nemohli bychom nezávislost obou šetření vyvrátit!

6.1.3 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 6.2.3.1

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:

a) Případ, kdy všechny jednotky výběru , které mají určitou variantu jedné náhodné veličiny , mají i určitou variantu druhé náhodné veličiny. Tomuto říkáme **úplná asociační závislost**.

b) 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 6.2.6

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**

$$r_A = \frac{n.n_{11} - n_{1.}.n_{.1}}{\sqrt{n_{1.}.n_{2.}.n_{.1}.n_{.2}}} \qquad (6.11).$$

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 $D = n_{11} . n_{22} – n_{12} . n_{21} = 0$, potom je hodnota $r_A = 0$ a dané náhodné veličiny jsou asociačně nezávislé.

d)

Příklad 6.2.4

Byly vyšetřovány vztahy mezi vlastnictvím automobilu a ochotou jezdit hromadnou dopravou do zaměstnání. Výsledné hodnoty šetření jsou uvedeny v následující tabulce 6.2.7. Zjistěte míru asociace těchto veličin.

Tabulka 6.2.7

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

Řešení:

Použijeme vzorec (6.14). Po dosazení příslušných hodnot získáme r_A=**-0,73586.** Uvedené vztahy ukazují na střední míru závislosti mezi vlastnictvím automobilu a ochotou jezdit hromadnou dopravou.

6.1.4 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 měr. Předpokládejme, že náhodná veličina X nabývá s variant a náhodná veličina Y nabývá r variant . K nejznámějším patří **Pearsonův koeficient kontingence**

$$c = \sqrt{\frac{\chi^2}{n + \chi^2}}, \qquad (6.12)$$

kde

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n.\left(n_{ij} - \frac{n_{i.}.n_{.j}}{n} \right)^2}{n_{i.}.n_{.j}} \qquad (6.13)$$

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**

$$K = \frac{\chi^2}{n.\sqrt{(r-1).(s-1)}}, \qquad (6.14)$$

v němž má χ^2 stejný význam jako v (6.13).

Příklad 6.2.5
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 6.2.8.

Tabulka 6.2.8

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 (6.13) převedeme tabulku 6.2.8 na tvar, kdy v jednotlivých buňkách tabulky budou hodnoty vypočtené ze vztahu (6.13).

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 χ^2 = 14,8386. Tuto hodnotu dosadíme nejdříve do vztahu (6.15) 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 = 0,0326. Podle něho můžeme zamítnout vztah mezi jednotlivými odpověďmi.