5. Vícerozměrný statistický soubor

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 vymezíme i metody pro analýzu vazeb mezi složkami vybraných dat.

5.1 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,…,p a j=1,…,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,…,p a j=1,…,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 $\mathbf{z_1,\ldots,z_k}$. 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**. Popišme 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 popsaných 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 5.1 **Korelační tabulka absolutních četností**

	y_j				$n_{i.}$
x_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:

$$n_{i.} = \sum_{j=1}^{r} n_{ij} \qquad (5.1)$$

$$n_{.j} = \sum_{i=1}^{p} n_{ij} \qquad (5.2)$$

$$n = \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{r} n_{.j} = \sum_{i=1}^{p} \sum_{j=1}^{r} n_{ij} \qquad (5.3)$$

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 5.1.1
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	$n_{i.}$
FYZIKA	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

Uveďme pro úplnost stejnou tabulku pro relativní četnosti:

		MATEMATIKA					
		1	2	3	4	5	$n_{i.}$
FYZIKA	1	0,038	0,074	0,074	0	0	0,186
	2	0,111	0,038	0,111	0	0	0,260
	3	0	0,038	0,038	0,149	0	0,222
	4	0	0,038	0,074	0,074	0,038	0,222
	5	0	0,038	0,038	0,038	0	0,111
	$n_{.j}$	0,149	0,222	0,333	0,260	0,038	1

Nepřesnosti v dané tabulce jsou zaviněny zaokrouhlováním jednotlivých hodnot.

Příklad 5.1.2

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. Převeďte tyto hodnoty do korelační tabulky.

Řešení:

		výška		
		správná	nesprávná	celkem
Délka	správná	239	60	299
	nesprávná	14	7	21
	celkem	253	67	320

Jde o případ asociační tabulky.

Příklad 5.1.3

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 naši 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 5.1.1

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 5.1.2

Řekneme, že znak **y** (náhodná veličina **Y**) je korelačně závislý na znaku **x** (náhodné veličině **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.

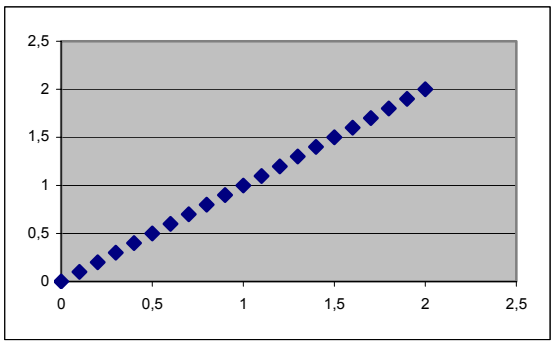
Korelací náhodných veličin chápeme vzájemnou korelační závislost uvažovaných náhodných veličin. Přitom rozlišujeme :

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

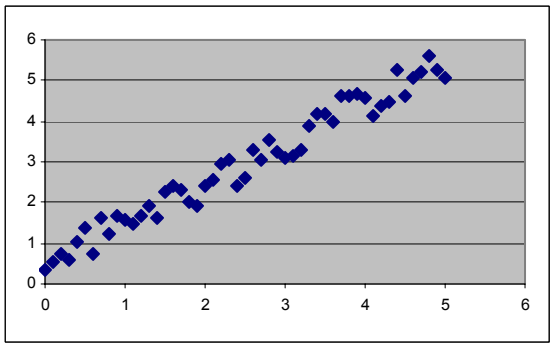
Důležité je pochopit jaký je vztah mezi korelační závislostí na jedné straně a kauzální závislostí na straně druhé. Jestliže jsou dvě náhodné veličiny korelačně závislé, pak to znamená, že mezi těmito náhodnými veličinami může existovat kauzální závislost. Nelze ale rozlišit zda jde o kauzální závislost 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).

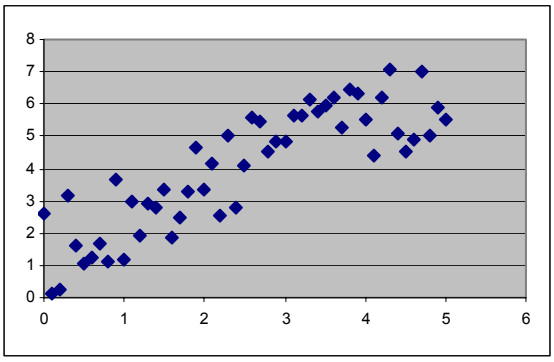
Obr. 5.1 –Extrémně kladná korelace



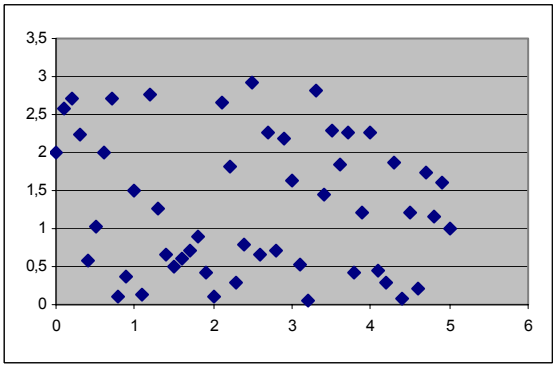
Obr. 5.2 – Silná kladná korelace



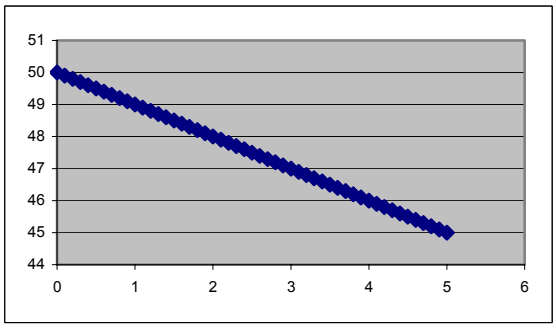
Obr 5.3. – Slabá kladná korelace



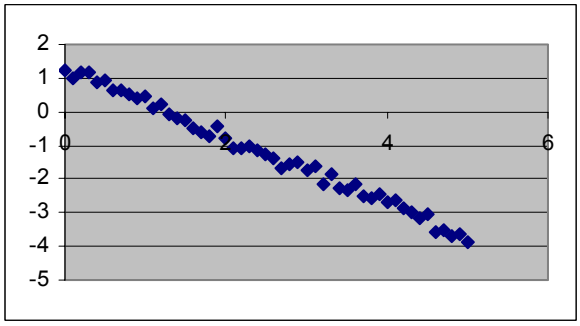
Obr 5.4 – Korelace blízká nule



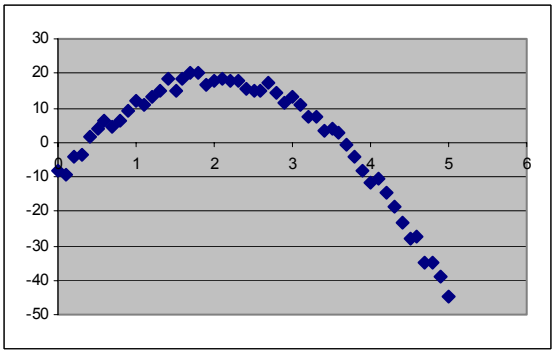
Obr 5.5 – Extrémně záporná korelace



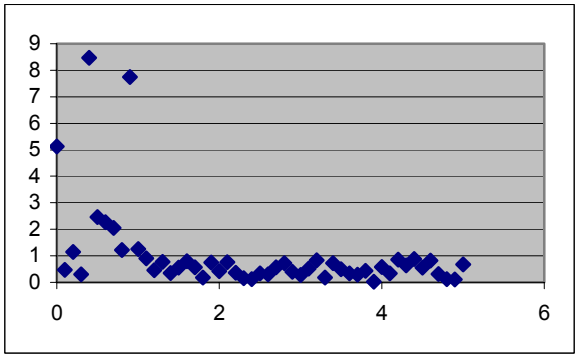
Obr 5.6 – Silná záporná korelace



Obr 5.7 – Nelineární korelace



Obr 5.8. – Nelineární korelace



Z jednotlivých obrázků je patrno, 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 5.7) , ale mohou mít podobu lineárního vztahu (Obr. 1), když spolu korelují velmi.