

Téma 2: Výpočet číselných charakteristik jednorozměrného a dvourozměrného datového souboru

Úkol 1.: U 100 náhodně vybraných domácností byl zjišťován způsob zásobování bramborami (znak X, varianty 1 = vlastní sklep, 2 = jinde, 3 = nákup) a bydliště (znak Y, varianty 1 = velké město, 2 = malé město, 3 = vesnice).

způsob zásobování	bydliště		
	velké město	malé město	vesnice
vlastní sklep	13	15	14
jinde	11	7	2
nákup	19	9	10

a) Pro oba znaky určíme modus.

b) Vypočteme Cramérův koeficient znaků X, Y.

Návod: Otevřeme nový datový soubor se třemi proměnnými X, Y, četnost a devíti případy. Do proměnné X napíšeme 3 jedničky, 3 dvojky a 3 trojky, do proměnné Y napíšeme 3 krát pod sebe 1, 2, 3 a do proměnné četnost napíšeme odpovídající simultánní absolutní četnosti dvojic variant (X, Y), tj. 13, 15, 14, 11, 7, 2, 19, 9, 10. Proměnným vytvoříme návěští a popíšeme význam jednotlivých variant.

ad a) Výpočet modu: Statistika – Základní statistiky/tabulky – Popisné statistiky – OK – klikneme na tlačítko se závažím – zaškrtneme Stav zapnuto, vybereme proměnnou vah četnost – OK - Proměnné X, Y – OK – Detailní výsledky – zaškrtneme Modus.

Proměnná	Popisné statistiky (brambory)	
	Modus	Četnost modu
X	1,000000	42
Y	1,000000	43

Proměnná X má modus 1, tj. nejvíce domácností skladuje brambory ve vlastním sklepě a proměnná Y má také modus 1, tj. nejvíce domácností bydlí ve velkém městě.

ad b) Výpočet Cramérova koeficientu: Statistika – Základní statistiky/tabulky – Kontingenční tabulky – OK – Specif. tabulky - List 1 X, List 2 Y - OK – na záložce Možnosti ve Statistikách 2 rozměrných tabulek zaškrtneme Fí (tabulky 2x2) & Cramérovo V & C – přejdeme na záložku Detailní výsledky – Detailní 2-rozm. tabulky.

Statist.	Statist. : X(3) x Y(3) (brambory)		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	6,420286	df=4	p=,16989
M-V chí-kvadr.	7,075760	df=4	p=,13195
Fí	,2533828		
Kontingenční koeficient	,2456207		
Cramér. V	,1791687		

Na posledním řádku najdeme, že Cramérův koeficient nabývá hodnoty 0,179, tedy mezi způsobem zásobování bramborami a bydlištěm domácnosti existuje jen slabá závislost – viz následující tabulka:

Cramérův koeficient	interpretace
mezi 0 až 0,1	zanedbatelná závislost
mezi 0,1 až 0,3	slabá závislost
mezi 0,3 až 0,7	střední závislost
mezi 0,7 až 1	silná závislost

Úkol 2.: Otevřeme datový soubor znamky.sta.

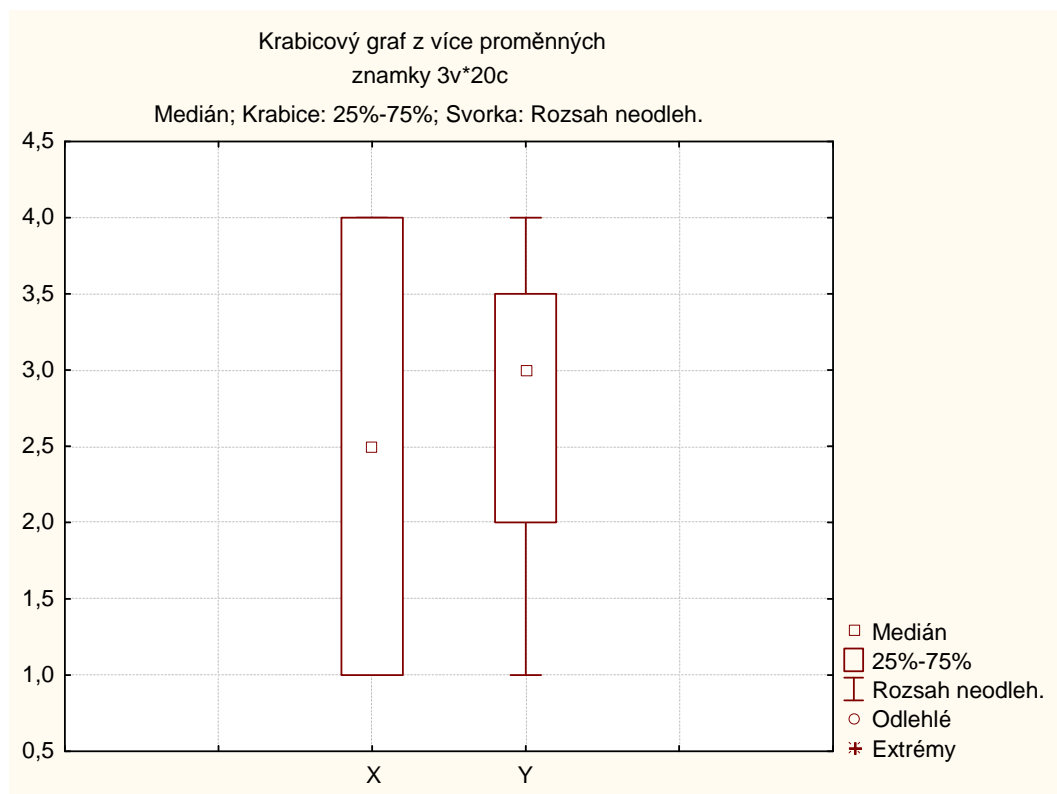
- a) Pro známky z matematiky a angličtiny vypočteme medián, dolní a horní kvartil, kvartilovou odchylku a vytvoříme krabicový diagram.
- b) Vypočteme Spearmanův koeficient korelace známek z matematiky a angličtiny pro všechny studenty, pak zvlášť pro muže a zvlášť pro ženy. Získané výsledky budeme interpretovat.

Návod:

ad a) Statistika – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X, Y – OK – Detailní výsledky - zaškrtneme Medián, Dolní & horní kvartily, Kvartil. rozpětí – Výpočet.

Proměnná	Popisné statistiky (znamky)			
	Medián	Spodní kvartil	Horní kvartil	Kvartilové rozpětí
X	2,500000	1,000000	4,000000	3,000000
Y	3,000000	2,000000	3,500000	1,500000

Vytvoření krabicového diagramu: Grafy – 2D Grafy – Krabicové grafy – vybereme Vícenásobný – Proměnné X, Y – OK.



ad b) Statistika – Neparametrická statistika – Korelace – OK – Proměnné X, Y – OK – Spearman R.

Pro všechny:

	Spearmanovy korelace (znamky) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000	
Proměnná	X	Y
X	1,000000	0,688442
Y	0,688442	1,000000

Počítáme-li Spearmanův korelační koeficient pro ženy (resp. pro muže), použijeme filtr: tlačítko Select Cases – Zapnout filtr – včetně případů – některé, vybrané pomocí výrazu Z=0 (resp. Z=1).

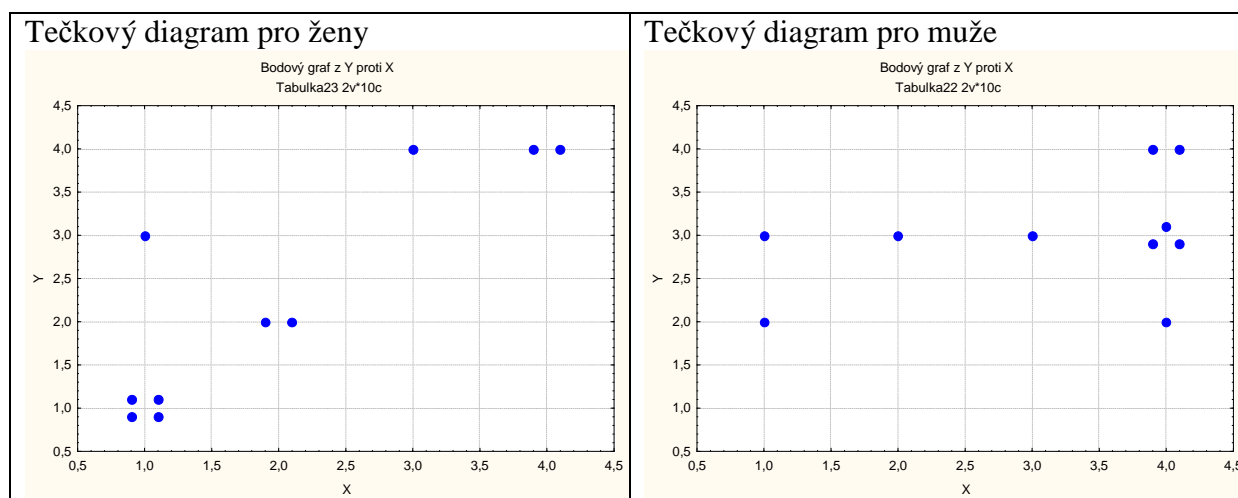
Pro ženy:

	Spearmanovy korelace (znamky) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000 Zhrnout podmínku: Z=0	
Proměnná	X	Y
X	1,000000	0,860314
Y	0,860314	1,000000

Pro muže:

	Spearmanovy korelace (znamky) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000 Zhrnout podmínku: Z=1	
Proměnná	X	Y
X	1,000000	0,373544
Y	0,373544	1,000000

Vidíme, že nejsilnější přímá pořadová závislost mezi známkami z matematiky a angličtiny je u žen, $r_s = 0,86$. U mužů je tato závislost mnohem slabší, $r_s = 0,37$. U žen tedy dochází k tomu, že se sdružují podobné známky z obou předmětů, zatímco u mužů se projevuje spíše tendence k různým známkám. Je to zřetelně vidět na dvourozměrných tečkových diagramech.



Význam hodnot Spearmanova (i Pearsonova) koeficientu korelace je popsán v tabulce:

Absolutní hodnota korelačního koeficientu	Interpretace hodnoty
0	lineární nezávislost
(0, 0,1)	velmi nízký stupeň závislosti
[0,1, 0,3)	nízký stupeň závislosti
[0,30, 0,50)	mírný stupeň závislosti
[0,50, 0,70)	význačný stupeň závislosti
[0,70, 0,90)	vysoký stupeň závislosti
[0,90, 1)	velmi vysoký stupeň závislosti
1	úplná lineární závislost

Úkol 3.: Otevřeme datový soubor ocel.sta.

a) Pro mez plasticity a mez pevnosti vypočteme aritmetický průměr, směrodatnou odchylku, rozptyl, koeficient variace, šikmost a špičatost. Výsledky porovnáme s údaji ve skriptech Popisná statistika (viz str. 30).

b) Vypočteme Pearsonův koeficient korelace meze plasticity a meze pevnosti. Dále vypočteme také kovarianci a výsledek porovnáme s výsledkem ve skriptech Popisná statistika (str. 30).

Návod:

ad a) Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X, Y – OK – Detailní výsledky - zaškrtneme Průměr, Směrodat. odchylka, Rozptyl, Variační koeficient, Šikmost, Špičatost – Výsledky.

Proměnná	Popisné statistiky (ocel)					
	Průměr	Rozptyl	Sm.odch.	Koef.prom.	Šikmost	Špičatost
X	95,8833	1070,240	32,71453	34,11910	-0,046758	-0,605826
Y	114,4000	1075,125	32,78911	28,66181	0,297889	-0,592621

Vysvětlení: Rozptyl a směrodatná odchylka vyjdou ve STATISTICE jinak než ve skriptech, protože STATISTICA ve vzorci pro výpočet rozptylu nepoužívá $1/n$, ale $1/(n-1)$. Koeficient variace (v tabulce označený jako Koef. Prom.) je udán v procentech.

ad b) Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměnných – X, Y – OK, na záložce Možnosti zrušíme volbu Včetně průměrů a sm. odch. – Výpočet.

Proměnná	Korelace (ocel)	
	X	Y
X	1,00	0,93
Y	0,93	1,00

Označ. korelace jsou významné na hlad. $p < ,05000$
N=60 (Celé případy vynechány u ChD)

Vidíme, že mezi X a Y existuje silná přímá lineární závislost.

Kovariance se počítá složitěji. Statistiky – Vícenásobná regrese - Proměnné Nezávislá X, Závislá Y – OK – OK – Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky - Kovariance.

Proměnná	Kovariance (ocel)	
	X	Y
X	1070,240	1002,471
Y	1002,471	1075,125

Vysvětlení: Na hlavní diagonále jsou rozptyly proměnných X, Y, mimo hlavní diagonálu je kovariance. Kovariance vyjde ve STATISTICĚ jinak než ve skriptech, protože ve STATISTICĚ se ve vzorci pro výpočet kovariance nepoužívá $1/n$, ale $1/(n-1)$.

Úkol k samostatnému řešení

Máme k dispozici údaje o rozměrech lebek staroegyptské populace. Jedná se o 216 mužů a 109 žen.

Proměnná X ... největší délka mozkovny v mm (tj. přímá vzdálenost kranio-metrických bodů glabella a opisthocranion)

Proměnná Y ... největší šířka mozkovny v mm (tj. přímá vzdálenost kranio-metrických bodů euryon dx a euryon sin)

Proměnná Z ... pohlaví osoby (1 ... muž, 0 ... žena)

(Data jsou uložena v souboru lebky.sta.)

a) Pro největší délku a největší šířku mozkovny mužů vypočítejte aritmetický průměr, směrodatnou odchylku, rozptyl, koeficient variace, šikmost a špičatost.

b) Vypočítejte Pearsonův koeficient korelace největší délky a největší šířky mozkovny mužů.

Dále vypočítejte také kovarianci těchto dvou znaků a nakreslete dvourozměrný tečkový diagram.

Výsledky pro bod (a):

Proměnná	Popisné statistiky (lebky.sta)					
	Zhrnout podmínku: z=1					
	Průměr	Rozptyl	Sm.odch.	Var.koef.	Šikmost	Špičatost
X	182,0324	40,76639	6,384856	3,507538	-0,055067	-0,451137
Y	137,1852	23,27717	4,824642	3,516883	0,085290	-0,248534

Výsledky pro bod (b), koeficient korelace:

Proměnná	Korelace (lebky.sta)	
	Zhrnout podmínku: z=1	
	X	Y
X	1,000000	0,168157
Y	0,168157	1,000000

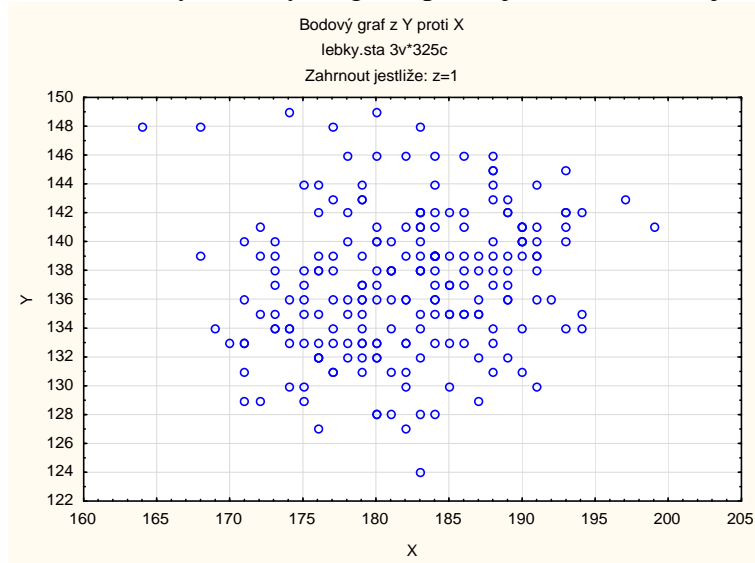
Vidíme, že mezi X a Y existuje nízký stupeň přímé lineární závislosti.

Výsledky pro bod (b), kovariance:

Proměnná	Kovariance (lebky.sta)	
	Zhrnout podmínku: z=1	
	X	Y
X	40,76639	5,18002
Y	5,18002	23,27717

Vysvětlení: Na hlavní diagonále jsou rozptýly proměnných X, Y, mimo hlavní diagonálu je kovariance.

Dvourozměrný tečkový diagram pro největší délku a největší šířku mozkovny mužů:



Ze vzhledu diagramu je patrné, že mezi oběma znaky existuje slabá přímá lineární závislost.