Téma 1 – Příklad 2

Zadání příkladu

V následující tabulce je uvedeno 30 dvojic hodnot znaku x a y. Roztřiďte tyto hodnoty do tabulky dvourozměrného rozdělení četností a vypočítejte hodnoty podmíněných průměrů a podmíněných rozptylů proměnné y.

Pořadí dvojice	x_i	y_j	Pořadí dvojice	x_i	y_j	Pořadí dvojice	x_i	y_j
1	1	1	11	2	3	21	3	4
2	1	2	12	2	1	22	4	4
3	1	1	13	3	3	23	3	5
4	1	3	14	3	2	24	3	3
5	2	1	15	3	2	25	4	3
6	2	4	16	3	1	26	4	4
7	2	2	17	3	5	27	1	4
8	2	2	18	2	3	28	2	5
9	1	4	19	1	1	29	4	5
10	1	4	20	2	1	30	4	1

Vypracování příkladu

Korelační tabulka dvojrozměrně zobrazuje dvě číselné proměnné a jejich sdružené četnosti (absolutní). Nejprve zobrazíme unikátní hodnoty proměnných do řádků a sloupců – bývá zvykem používat řádky pro proměnnou *x* a sloupce pro proměnnou *y*. Vnitřní buňky tabulky je třeba zkonstruovat tak, aby jednotlivé hodnoty proměnných *x* a *y* měly v křížově odpovídající buňce absolutní četnost výskytu dat. Například pro x = 1 a y = 1 je četnost rovna třem, existují tři takové dvojice hodnot (v zadání s pořadovými čísly 1, 3, 19).

y_j / x_i	1	2	**3**	**4**	5	Součty četností $n_{i.}$	$\bar{y}_i$	s_i^2
1	3	1	1	3	0	8	2,5000	1,7500
2	3	2	2	1	1	9	2,4400	1,8242
3	1	2	2	1	2	8	3,1250	1,8594
4	1	0	1	2	1	5	3,4000	1,8400
Součty četností $n_{.j}$	8	5	6	7	4	**30**	x	x

Jednou z kontrol správnosti je křížový součet okrajových četností, který musí dohromady představovat rozsah souboru[1].

Podmíněnou charakteristikou rozumíme určitou hodnotu deskriptivní statistiky pro proměnnou *y*, která platí za předpokladu určité hodnoty proměnné *x*. V našem případě budeme počítat podmíněný průměr pomocí vzorce:

$$\bar{y}_i = \frac{\sum_{j=1}^{l} y_j n_{ij}}{n_{i\cdot}} ;$$

kde n_{ij} jsou příslušné sdružené absolutní četnosti a $n_{i\cdot}$ je okrajová absolutní četnost, která je součtem počtu hodnot *y* v případě určité hodnoty *x*.

Pro hodnotu x = 1 je podmíněný průměr roven hodnotě

$\bar{y}_i = \frac{1\cdot 3+2\cdot 1+3\cdot 1+4\cdot 3+5\cdot 0}{8} = 2{,}5$. Takto postupujeme i pro další hodnoty x.

Podmíněný rozptyl bude vypočtený dle vzorce:

$$s_i^2 = \frac{\sum_{j=1}^{l} \left(y_j - \bar{y}_i\right)^2 n_{ij}}{n_{i\cdot}} .$$

Pro hodnotu x = 1 je podmíněný rozptyl roven hodnotě

$s_i^2 = \frac{(1-2{,}5)^2\cdot 3+(2-2{,}5)^2\cdot 1+(3-2{,}5)^2\cdot 1+(4-2{,}5)^2\cdot 3+(5-2{,}5)^2\cdot 0}{8} = 1{,}75$. Takto postupujeme i pro další hodnoty x.

1V případě chybějících údajů je třeba postupovat podle některé ze známých metod jejich doplnění, nebo použít pouze kompletní dvojice (v SGP bývá označováno „Complete Cases Only“).

Řešení v SGP

V programu Statgraphics stačí zadat všechny hodnoty dvojic do dvou samostatných sloupců – proměnných x a y. Je třeba pouze dbát na to, aby hodnoty párově správně odpovídaly. Výsledné vektory pak vypadají takto.

	x	y
1	1	1
2	1	2
3	1	1
4	1	3
5	2	1
6	2	4
7	2	2
8	2	2
9	1	4
10	1	4
11	2	3
12	2	1
13	3	3
14	3	2
15	3	2
16	3	1
17	3	5
18	2	3
19	1	1
20	2	1
21	3	4
22	4	4
23	3	5
24	3	3
25	4	3
26	4	4
27	1	4
28	2	5
29	4	5
30	4	1

Procedura v SGP: Describe – Categorical Data – Crosstabulation (Frequency Table)

Tvorbu dvourozměrné tabulky zařídí procedura Crosstabulation. Ve vstupním dialogu vybereme řádkovou proměnnou Row Variable (použijeme tradičně x) a sloupcovou proměnnou Column Variable (y). V okně Frequency Table můžeme přes doplňkový panel Pane Options zobrazit relativní četnosti vzhledem k řádku, sloupci, nebo celému souboru (Table Percentages). Dále jsou zde možnosti, které odkazují na chí-kvadrát test o nezávislosti

dvou kategoriálních proměnných (Expected Frequencies, Deviations, Chi-Square Values). Jejich použití by v tomto případě bylo ovšem chybné!

Frequency Table for x by y

	1	2	3	4	5	Row Total
1	3	1	1	3	0	8
	10,00%	3,33%	3,33%	10,00%	0,00%	26,67%
2	3	2	2	1	1	9
	10,00%	6,67%	6,67%	3,33%	3,33%	30,00%
3	1	2	2	1	2	8
	3,33%	6,67%	6,67%	3,33%	6,67%	26,67%
4	1	0	1	2	1	5
	3,33%	0,00%	3,33%	6,67%	3,33%	16,67%
Column Total	8	5	6	7	4	30
	26,67%	16,67%	20,00%	23,33%	13,33%	100,00%

Cell contents:
Observed frequency
Percentage of table

Procedura v SGP: Describe - Numeric Data – Subset Analysis
(Summary Statistics)

Při vstupním dialogu zadáme jako *Data* proměnnou, z jejíž hodnot budeme počítat výběrové charakteristiky (podmiňovanou), proměnnou *y*, do políčka *Codes* zadáme proměnnou podmiňující, tedy *x*.

V okně Summary Statistics vidíme jednotlivé podmíněné charakteristiky. Jejich zobrazení můžeme upravit v doplňkových možnostech Pane Options. Nezapomeňme, že jde o výběrové charakteristiky, takže v případě, že nás zajímají charakteristiky základní, musíme upravit výrazem (n-1/n). U charakteristik tvaru rozdělení je odlišnost výraznější, jelikož program Statgraphics užívá vzorců odlišných od klasických momentových měr.

Summary Statistics
Data variable: y

x	*Count*	*Average*	*Variance*
1	8	2,5	2,0
2	9	2,44444	2,02778
3	8	3,125	2,125
4	5	3,4	2,3
Total	30	2,8	2,02759

The StatAdvisor
This table shows sample statistics for the 4 levels of x.

Interpratace

V případě, že proměnná x nabývá hodnoty jedna, je průměrná hodnota proměnné y rovna dvěma a půl. Rozptyl hodnot y je v tomto případě roven 1,75.

Řešení v MS Excel

Dle verze MS Excel je nutné využít specifických statistických funkcí pro výpočty podmíněných charakteristik, případně využít odvozených funkcí (např. +IF)

PRŮMĚR – Průměr hodnot.

VAR.P – Rozptyl základního souboru (od Excel 2010).

VAR.S – Rozptyl výběru (od Excel 2010).

VAR – Rozptyl základního souboru.

VAR.VÝBĚR – Rozptyl výběru.