ZOBECŇOVÁNÍ VÝBĚROVÝCH VÝSLEDKŮ NA ZÁKLADNÍ SOUBOR

INTERVAL SPOLEHLIVOSTI PRO HODNOTU PRŮMĚRU

Naším cílem při sociologických analýzách, který jsou založeny na práci s daty z výběrového souboru, je zobecňovat výsledky z výběru na celou základní populaci. Při univariační analýze toho docílíme tak, že určujeme intervaly spolehlivosti. Připomínáme, že zobecňování výsledků z výběru na základní soubor (statistická inference) je možné pouze tehdy, pokud je výběr reprezentativní, tedy pokud byl vybrán takovými postupy, které zajišťují, že každý prvek základní souboru měl stejnou šanci dostat se do souboru výběrového.

Příklad P5.3a: Z příkladu **P2.3**, kde jsme počítali, jaká je průměrná hodnota postoje k důležitosti Boha v životě českých respondentů (na desetibodové stupnici byl průměr 3,63) chceme stanovit interval spolehlivosti (*confidence interval*, *CI*) pro základní soubor (jelikož výběrový soubor je reprezentativní pro populaci ČR starší 18 let – k tomu viz článek Jana Řeháka v časopise *Sociální studia* 6 z roku 2001, str. 16 – má toto úsilí smysl), abychom zjistili, jaké hodnoty průměru můžeme očekávat v celé dospělé populaci České republiky. Intervaly spolehlivosti pro hodnotu průměru vypočítá SPSS v proceduře *Analyze —Descriptive Statistics — Explore* pro proměnnou q33.
Výsledek:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Q33 Bůh - důležitost v životě	1846	96,8%	62	3,2%	1908	100,0%

Descriptives

			Statistic	Std. Error
Q33 Bůh - důležitost v životě	Mean		3,63	,07
	95% Confidence Interval for Mean	Lower Bound	3,49	
		Upper Bound	3,77	
	5% Trimmed Mean		3,43	
	Median		2,00	
	Variance		9,345	
	Std. Deviation		3,06	
	Minimum		1	
	Maximum		10	
	Range		9	
	Interquartile Range		5,00	
	Skewness		,858	,057
	Kurtosis		-,614	,114

Vzhledem k tomu, že chceme mít poměrně velkou jistotu o hodnotě průměru základního souboru, je v SPSS nastavena standardně jistota 95 %. V tabulce *Desciptives* vidíme, že dolní hranice intervalu spolehlivosti je 3,49 a jeho horní hranice 3,77. Tato čísla tedy říkají, že s 95% jistotou můžeme očekávat, že průměrná hodnota odpovědí na otázku o důležitosti Boha v našem životě by se v celé české populaci pohybovala mezi 3,49–3,77. (Poznámka: všimněte

si hodnoty směrodatné chyby – *Std. Error* 0,07. Násobte tuto hodnotu dvěma a postupně ji odečtěte a přičtěte k hodnotě průměru. Jaký bude výsledek? No přesně takový, jaký vypočítal SPSS. Interval spolehlivosti se tedy, pokud znáte směrodatnou chybu, dá lehce spočítat i ručně. A jak vypočítáme směrodatnou chybu? I tu lze lehce spočítat a to tak, že podělíme směrodatnou odchylku druhou odmocninou velikosti výběrového souboru (N). Zkontrolujme si: velikost výběrového souboru je, jak vidíme z tabulky *Case Processing Summary*, 1846 – je třeba pracovat pouze s údajem o platných odpovědích, ti, kdo na tuto otázku neodpověděli, nebyli do výpočtu průměru zahrnuti. Druhá odmocnina tohoto čísla je 42,96. Směrodatná odchylka (Std. Deviation v tab. *Desciptives*) je 3,06, pak 3,06/42,96 = 0,07, což je hodnota směrodatné chyby.

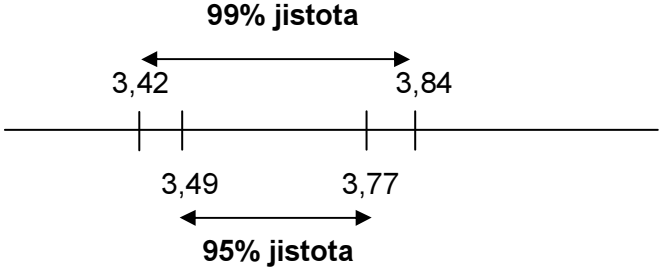
Pokud bychom chtěli mít interval spolehlivosti stanoven s jistotou 99 %, nastavíme v dialogovém okně v proceduře *Explore* – *Statistics* hodnotu intervalu spolehlivosti na 99 %. Lze ji ovšem vypočítat i ručně. Ruční výpočet je nesmírně jednoduchý. Hodnotu směrodatné chyby násobíme třemi a výsledek přičteme a odečteme od hodnoty průměru. Takže v našem případě:
0,07 * 3 = 0,21.
3,63 + 0,21 = 3,84
3,63 – 0,21 = 3,42
99% interval spolehlivosti je tedy 3,42–3,84. Zkontrolujme výsledek z výpočtu v SPSS:

Descriptives

			Statistic	Std. Error
Q33 Bůh - důležitost v životě	Mean		3,63	,07
	99% Confidence Interval for Mean	Lower Bound	3,45	
		Upper Bound	3,82	
	5% Trimmed Mean		3,43	
	Median		2,00	
	Variance		9,345	
	Std. Deviation		3,06	
	Minimum		1	
	Maximum		10	
	Range		9	
	Interquartile Range		5,00	
	Skewness		,858	,057
	Kurtosis		-,614	,114

Výsledek se nepatrně liší. Rozdíl vznikl tím, že náš ruční výpočet je méně přesný, neboť při požadavku 99 % spolehlivosti je třeba násobit směrodatnou chybu ne třemi, nýbrž konstantou 2,85. Pro praktické sociologické účely a ruční výpočty ovšem tato malá nepřesnost není podstatná.

Srovnejme nyní 95% interval spolehlivosti (3,49–3,77) s jeho 99% bratrancem (3,42–3,84).



Vidíme, že daní za větší jistotu je širší interval spolehlivosti, z něhož paradoxně vyplývá určitá vyšší „nevědomost“: mám 99 % jistotu, že průměr české populace v tomto postoji leží někde mezi hodnotou 3,42 až 3,84.

95% spolehlivost je v sociálních vědách obvykle dobrou hranicí jistoty, takže se v SPSS s implicitně zabudovaným vzorcem pro výpočet intervalu spolehlivosti můžeme spokojit (blíže k této problematice viz článek v čítance: Řehák, J. 1974. Poznámky k analýze sociologických dat. *Sociologický časopis*, č. 4: 425–433).

* * *

INTERVAL SPOLEHLIVOSTI PRO HODNOTU PODÍLU (%)

Interval spolehlivosti se stanovuje nejenom pro hodnotu průměru, ale také pro hodnotu nějakého podílu (%). Víme-li např. z výzkumu veřejného mínění, že 75 % respondentů v reprezentativním souboru souhlasí s názorem, že schopní lidé by měli hodně vydělávat, musí nás zajímat otázka, v jakém intervalu se bude tento podíl pohybovat v celé populaci ČR.

V případě, že stanovujeme interval spolehlivosti pro podíl (procento) a ne pro průměr, nemůžeme žel plně využít SPSS, neboť tento software kupodivu nemá tuto proceduru zabudovanou ve svých paměťových vzorcích. Proto si musíme pomoci prostřednictvím drobných triků. Na tomto místě bych rád upřímně poděkoval kolegovi Janu Řehákovi, který nám tyto triky poradil. Záleží přitom na tom, zdali hledáme interval spolehlivosti pro proměnnou, která je dichotomická (má jenom dvě varianty znaku, např. muž–žena, je spokojen–je nespokojen atd.), nebo polytomická (má více variant znaku).

Příklad P5.3b: Interval spolehlivosti pro dichotomické proměnné.

Trik spočívá v tom, že hodnoty dichotomie (ať byly kódovány jako 0 a 1, nebo jako 1 a 2) převedeme (rekódujeme procedurou *Recode*) na hodnoty 0 a 100. Pro takto upravenou proměnnou pak již v proceduře *Explore* spočteme normální průměr (t.j. procento) a jeho interval spolehlivosti, který je v dané situaci hledaným intervalem spolehlivosti pro procenta.

Ukázka výpočtu. Chceme zjistit, jaký je v souboru EVS-ČR1999 interval spolehlivosti pro rozložení odpovědí na otázku q42: *Myslíte si, že žena musí mít děti, aby se splnilo její poslání, nebo to není nutné?* Jelikož je to dichotomická proměnná, můžeme uplatnit Řehákův trik. Nejdříve tedy musíme rekódovat původní hodnoty 1 a 2 na hodnoty 0 a100. Proveďme: Původní proměnná:

Tab. A

Q42 Žena musí mít děti, aby splnila poslání

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 ano	795	41,7	44,1	44,1
	2 není to nutné	1007	52,8	55,9	100,0
	Total	1803	94,5	100,0	

Rekódovaná proměnná:

```
RECODE
  q42 (1=0)(2=100).
EXECUTE.
```

Tab. B

Q42 Žena musí mít děti, aby splnila poslání

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	795	41,7	44,1	44,1
	100	1007	52,8	55,9	100,0
	Total	1803	94,5	100,0	

Tab. C

Descriptives

			Statistic	Std. Error
Q42 Žena musí mít děti, aby splnila poslání	Mean		55,87	1,17
	95% Confidence Interval for Mean	Lower Bound	53,58	
		Upper Bound	58,17	
	5% Trimmed Mean		56,52	
	Median		100,00	
	Variance		2466,89	
	Std. Deviation		49,67	
	Minimum		0	
	Maximum		100	
	Range		100	
	Interquartile Range		100,00	
	Skewness		-,237	,058
	Kurtosis		-1,946	,115

Vidíme, že vypočítaný průměr 55,87 odpovídá podílu respondentů, kteří se domnívají, že není nutné, aby žena měla děti (55,9 ve sloupci *Valid Percent* v tab. A nebo B). Proto pro tento údaj můžeme údaje o horní a dolní hranici 95% intervalu spolehlivosti pro průměr (v tabulce C) chápat jako údaje o horní a dolní hranici intervalu spolehlivosti pro toto procento. Tudíž v základním souboru, to je mezi dospělou populací ČR, se pohybuje podíl lidí, kteří si myslí, že není nutné, aby žena měla děti k naplnění jejího poslání, mezi 53,6 a 58,2 %.

Pro výpočet intervalu spolehlivosti pro podíl lidí, zastávají názor, že žena musí mít děti k naplnění poslání, již musíme použít kalkulačky – stačí ale pouze hodnoty intervalů spolehlivosti odečíst od 100: 100–53,58 = 46,42 a 100 – 58,17 = 41,83. Podíl respondentů s tímto postojem se bude tak v základním souboru pohybovat mezi 41,8 a 46,2 %.

Příklad P5.3c: Interval spolehlivosti pro polytomické proměnné.

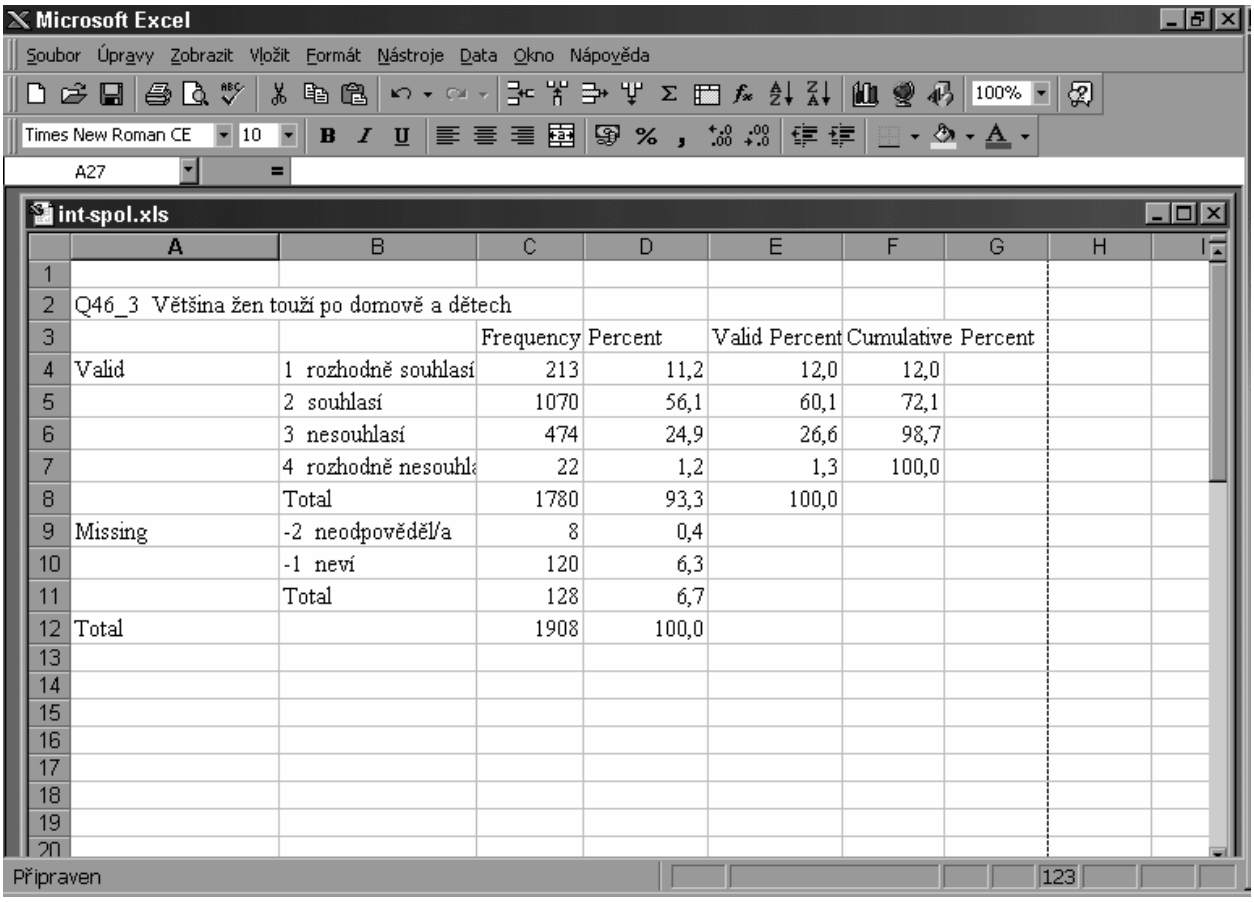
U vícehodnotového znaku se postupuje, pokud nechceme interval spolehlivosti počítat ručně, jinak. Tabulku, kterou v SPSS dostaneme z *Frequencies*, zkopírujme (prostřednictvím příkazu *Copy*) a vložme ji do Excelu. V něm si připravíme příslušný vzorec pro výpočet intervalu spolehlivosti:

$$\sqrt{\frac{p(100-p)}{N}}$$

a pak už jen dosazujeme příslušná data. A pokud si tento excelovský soubor uložíme jako matrici, můžeme se k němu opakovaně vrátit a vypočítat velmi rychle interval spolehlivosti pro jakoukoliv polytomickou proměnnou.

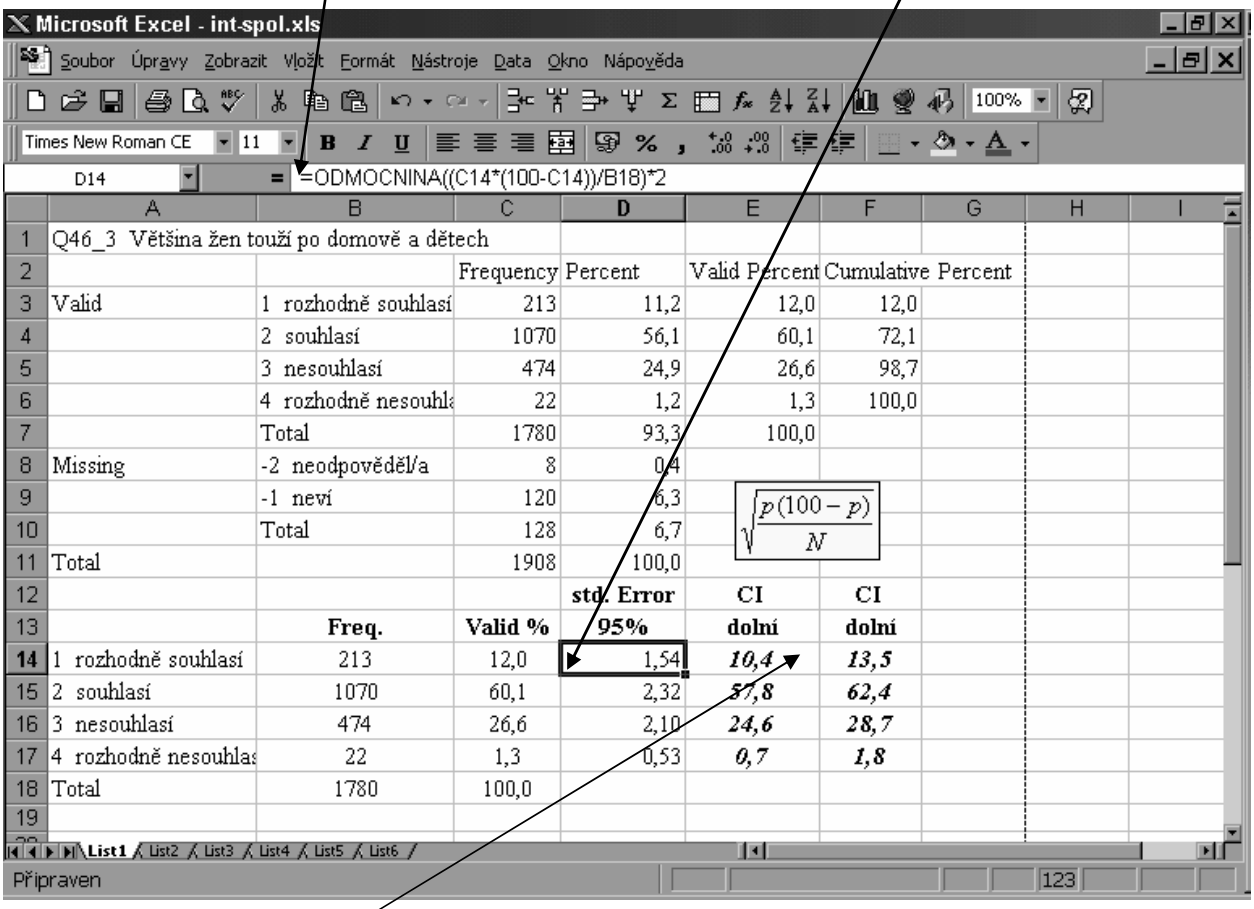
Ukázka:

V příkladu **P2.1** jsme se zajímali o rozložení proměnné q46_3. Vypočítejme pro jednotlivá procenta intervaly spolehlivosti. Nejdříve si tedy v SPSS udělejme znovu třídění prvního stupně této proměnné a použijme k tomu proceduru *Frequencies*. Vypočtený výsledek zkopírujeme a vložíme do tabulkového procesoru Excel. Bude to vypadat takto:



	A	B	C	D	E	F	G	H	I
1									
2	Q46_3 Většina žen touží po domově a dětech								
3			Frequency	Percent	Valid Percent	Cumulative Percent			
4	Valid	1 rozhodně souhlasí	213	11,2	12,0	12,0			
5		2 souhlasí	1070	56,1	60,1	72,1			
6		3 nesouhlasí	474	24,9	26,6	98,7			
7		4 rozhodně nesouhla	22	1,2	1,3	100,0			
8		Total	1780	93,3	100,0				
9	Missing	-2 neodpověděl/a	8	0,4					
10		-1 neví	120	6,3					
11		Total	128	6,7					
12	Total		1908	100,0					
13									
14									
15									
16									
17									
18									
19									
20									

Nyní si vložme příslušné vzorce pro výpočet intervalu spolehlivosti (jeho matematickou podobu jsme jen pro připomenutí přidali do volného prostoru). Vidíte, že jsme si v nové tabulce zkopírovali údaje o absolutních četnostech (*Frequency*), z nichž ovšem pro výpočet použijeme, jak říká vzorec, pouze jeden údaj, jímž je celková velikost souboru (1 780). To je ve vzorci ono *N*. Dále jsme si zkopírovali údaje o platných procentech (*Valid Percent*), neboť to jsou hodnoty *p* ve vzorci. Přidali jsme tři nové sloupce. Do sloupce **Std. error 95** jsme za použití excelovské syntaxe vepsali celý vzorec pro výpočet směrodatné chyby, kterou stanovujeme pro 95% jistotu. Tento vzorec je vepsán do buňky D14 a způsob zápisu je zobrazen v dialogovém okně.



	A	B	C	D	E	F	G	H	I
1	Q46_3 Většina žen touží po domově a dětech								
2			Frequency	Percent	Valid Percent	Cumulative Percent			
3	Valid	1 rozhodně souhlasí	213	11,2	12,0	12,0			
4		2 souhlasí	1070	56,1	60,1	72,1			
5		3 nesouhlasí	474	24,9	26,6	98,7			
6		4 rozhodně nesouhla	22	1,2	1,3	100,0			
7		Total	1780	93,3	100,0				
8	Missing	-2 neodpověděl/a	8	0,4					
9		-1 neví	120	6,3					
10		Total	128	6,7					
11	Total		1908	100,0					
12				**std. Error**	**CI**	**CI**			
13		**Freq.**	**Valid %**	**95%**	**dolní**	**dolní**			
14	1 rozhodně souhlasí	213	12,0	1,54	***10,4***	***13,5***			
15	2 souhlasí	1070	60,1	2,32	***57,8***	***62,4***			
16	3 nesouhlasí	474	26,6	2,10	***24,6***	***28,7***			
17	4 rozhodně nesouhlas	22	1,3	0,53	***0,7***	***1,8***			
18	Total	1780	100,0						
19									

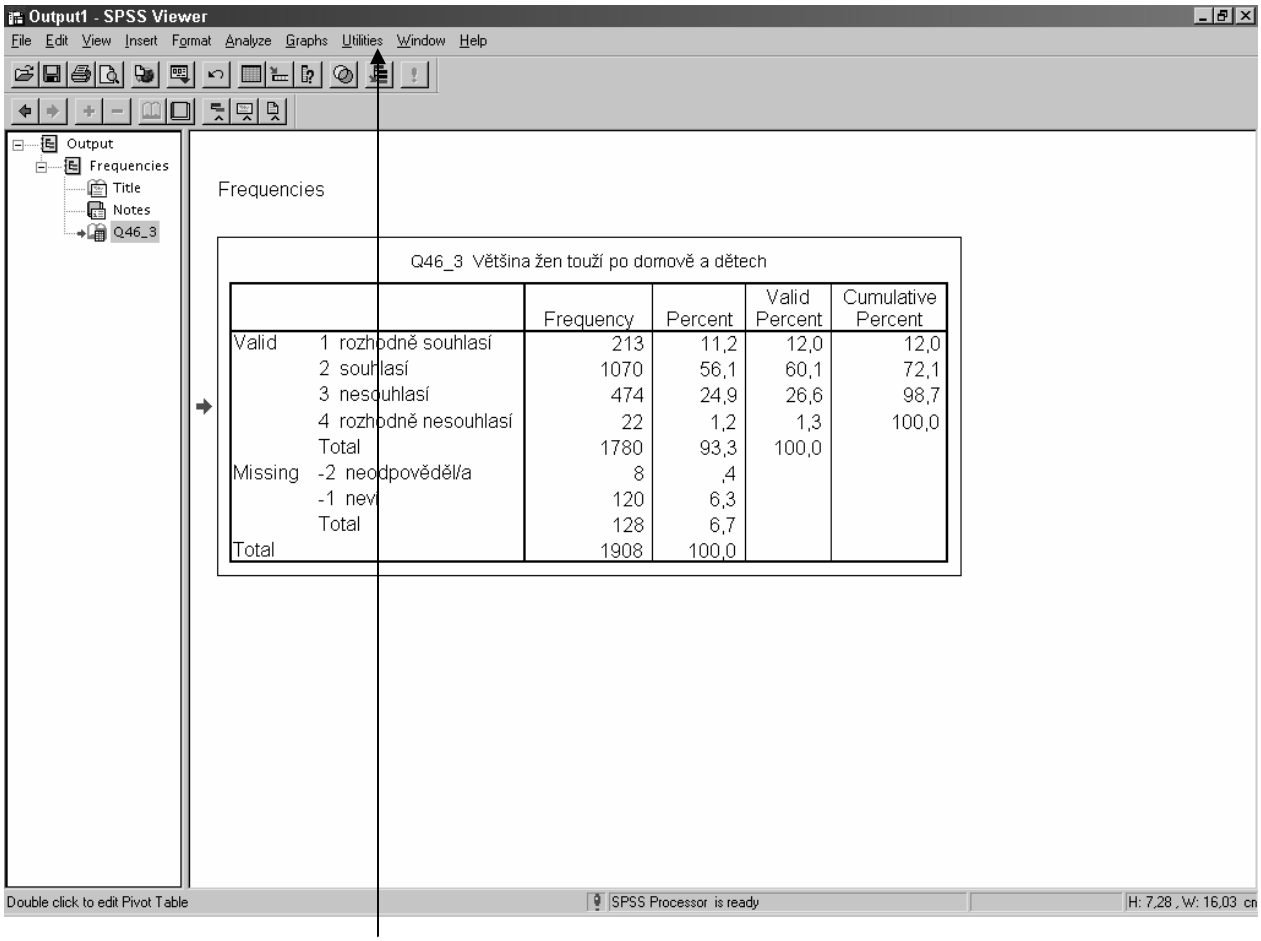
Další dva přidané odstavce jsou již přímo hodnoty dolního a horního intervalu spolehlivosti. Pod údajem ***10,4*** je vzorec = C14-D14 a pod údajem ***13,5*** je vzorec = C14+D14, tedy operace, kdy od 12 % těch, kdo rozhodně souhlasí s výrokem, že *Většina žen touží po domově a dětech*, nejdříve odečítáme velikost směrodatné chyby (1,5) a pak ji k 12 % přičítáme. Tím získáváme interval spolehlivosti (10 – 14 %) pro podíl obyvatel ČR, kteří rozhodně souhlasí s tímto výrokem. Vidíme, že směrodatná chyba je v každém řádku jiná a je to pochopitelné. Pro 1,3 % musí být menší než pro 60,1 %. Proto uvádějí-li někdy agentury pro výzkum veřejného mínění velikost výběrové chyby (což je samo o sobě velmi chvályhodný fakt) a tvrdí, že např. velikost výběrové chyby je 2 %, není to informace správná (proč?).

Máte-li takto připravenou matrici, pak při výpočtu dalších intervalů spolehlivosti z jiných výpočtů SPSS stačí přepsat údaje o velikosti vzorku (buňka B18) a do buněk C14…C17 dosadit příslušná validní procenta. Excel (a v tom je jeho kouzlo), okamžitě přepočítá nově dosazené údaje a vy máte k dispozici nové intervaly spolehlivosti.[1] Pokud bude vaše nová proměnná mít vyšší počet variant než 4, budete si muset přidat příslušný počet řádků, do nichž vepíšete

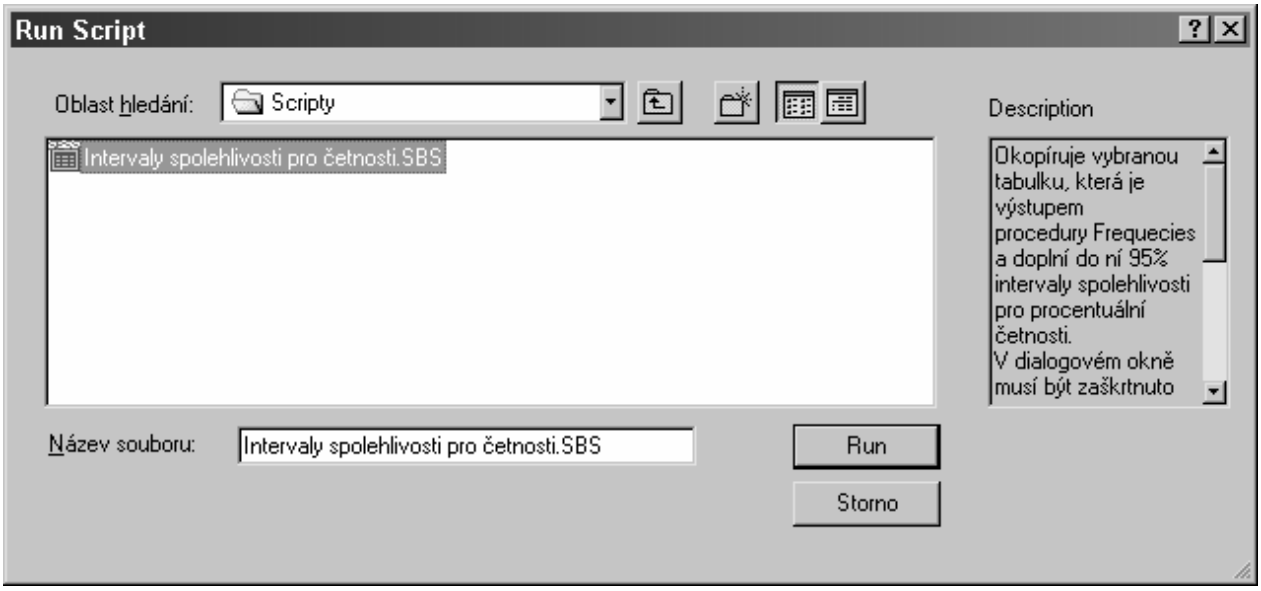
[1] Tuto matrici naleznete na dokumentovém serveru informačního systému MU.

patřičné vzorce (dávejte přitom velký pozor, abyste – pokud budete vzorce kopírovat – v nich měli správně označeny všechny odkazy na buňky).
Tento postup je, jak jistě uznáte, poněkud krkolomný. Proto kolega Řehák sepsal v rámci jazyka SPSS malý prográmek (v jazyce SPSS se mu říká *script*), který intervaly spolehlivosti doplňuje přímo do tabulky z *Frequencies*. Postup je následující:

1. Necháte si spočítat *Frequencies* příslušné proměnné.
2. V Outputu SPSS na tuto tabulku 1x kliknete a tím ji označíte (viz malou šipku u tabulky na obrázku).



3. Klikněte na tlačítko *Utilities* a v něm na příkaz *Run Script*.
4. V dialogovém okně naveďte SPSS tam, kde máte uložen script pro intervaly spolehlivosti.



Když si kliknete na jméno tohoto *scriptu*, vkopíruje se do okénka „Název souboru“ a v okně vpravo nazvaném *Description* se objeví popis procedury. Klikněte na příkaz *Run*. Objeví se nová tabulka, která uvádí příslušné horní a dolní meze intervalu spolehlivosti pro jednotlivé varianty znaku (viz tab. 5.3, kterou jsme editovali tak, že jsme odmazali sloupce pro intervaly spolehlivosti kumulovaných četností).

Tab. 5.3: Intervaly spolehlivosti vypočtené z Řehákova *scriptu*

Četnostní tabulka s intervaly spolehlivosti proměnné Q46_3 Většina žen touží po domově a dětech

:

Hodnoty		Statistiky						
		Četnost	Relativní četnost	Dolní mez[a]	Horní mez[a]	Rel. četnost platných hodnot	Dolní mez[a]	Horní mez[a]
Platné	1 rozhodně souhlasí	213	11,18%	9,77%	12,60%	11,99%	10,48%	13,50%
	2 souhlasí	1070	56,09%	53,86%	58,31%	60,11%	57,84%	62,39%
	3 nesouhlasí	474	24,86%	22,92%	26,80%	26,65%	24,59%	28,70%
	4 rozhodně nesouhlasí	22	1,17%	,69%	1,65%	1,25%	,73%	1,77%
	Celkem	*1780*	*93,30%*	*92,18%*	*94,42%*	*100,00%*		
Vynechané	-2 neodpověděl/a	8	,43%	,14%	,73%			
	-1 neví	120	6,27%	5,18%	7,35%			
	Celkem	*128*	*6,70%*	*5,58%*	*7,82%*			
Celkem		1908	100,00%					

a. 95%ní interval spolehlivosti. K výpočtu je použita asymptotická metoda, která předpokládá, že celkový počet pozorování je větší než 30 a v každé kategorii se vyskytuje alespoň 5 případů.

V této tabulce nás zajímají sloupce pro validní četnosti (sloupce platných hodnot), které jsou vyznačeny žlutě. Když tyto intervaly spolehlivosti zkontrolujete s intervaly, které jsme vypočítali v excelovské tabulce výše, uvidíte, že jsou totožné.