LEKCE03
PŘÍKLAD

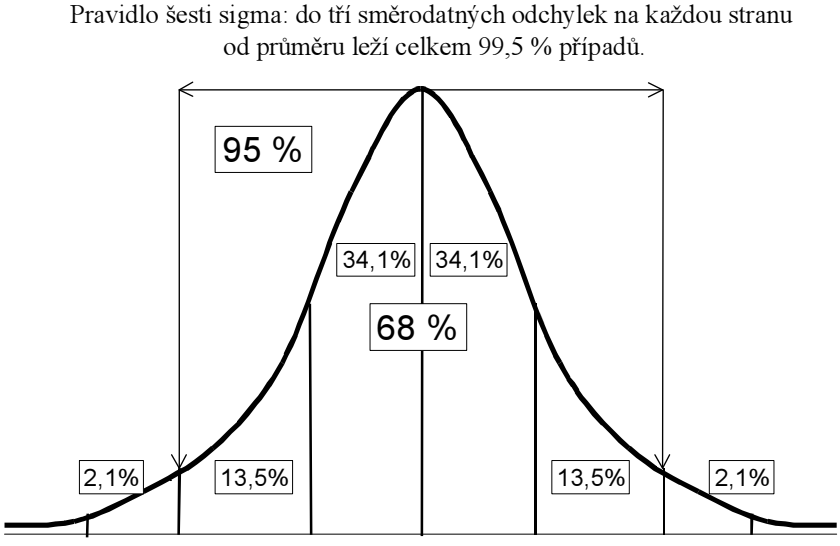
NORMALITA ROZLOŽENÍ A Z SKÓRY; ZOBECŇOVÁNÍ VÝBĚROVÝCH VÝSLEDKŮ NA ZÁKLADNÍ SOUBOR

V předchozích lekcích jsme si ukázali, že před tím, než začneme analyzovat data, je u proměnných měřených na intervalové úrovni vždy dobré se přesvědčit, jaký tvar má rozložení jednotlivých znaků. Zajímá nás především, zdali má distribuce četností tvar rozložení normálního. Tato informace je v analýze dat velmi důležitá, neboť spousta statistických procedur má jako jeden z předpokladů to, že hodnoty proměnných jsou normálně rozloženy.

Spousta biologických, psychických ale i sociálních vlastností má tu charakteristiku, že jsou rozloženy zvláštním způsobem kolem střední hodnoty – totiž že jsou rozloženy normálně. Toto rozložení má podobu zvonovité křivky – nazývá se tak v angličtině (*Bell Curve*), ve francouzštině se hovoří o „křivce policejního klobouku". Ve vědeckém jazyce se hovoří o Gaussově křivce nebo také o křivce normálního rozložení.

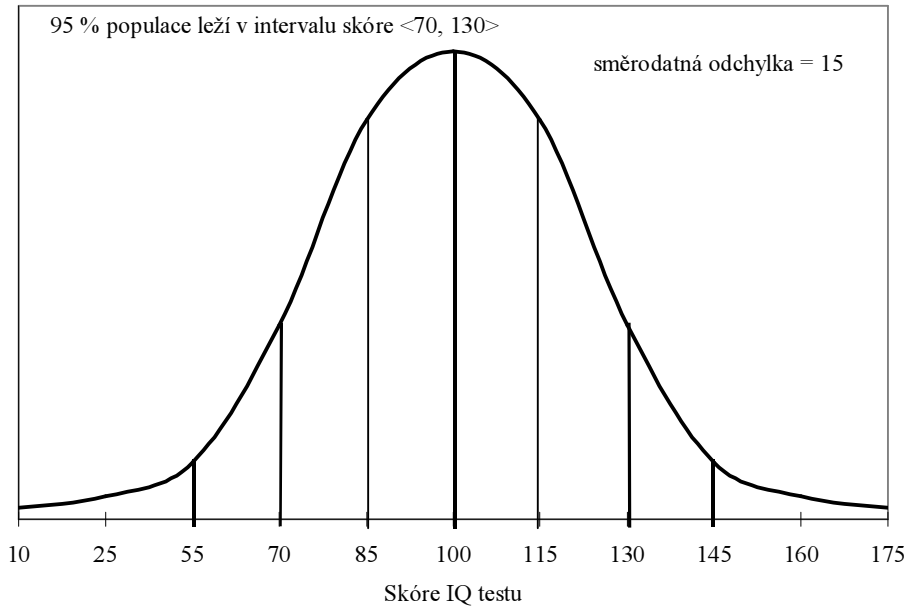
Normální rozložení má – mimo nádherného a ladného symetrického tvaru – několik pěkných vlastností: předně, je přesně určena střední hodnotou a směrodatnou odchylkou. V normálním rozložení má průměr, medián i modus stejnou hodnotu. Většina hodnot se soustřeďuje kolem průměru. Navíc platí, že do čtyř sigma (σ = sigma je symbol pro směrodatnou odchylku), tedy dvě směrodatné odchylky na každou stranu od průměru spadne většina pozorovaných hodnot, přesně 95,34 %. Do šesti sigma pak padne přesně 99,7 % pozorovaných hodnot (tedy v rozsahu +3 a -3 směrodatných odchylek). Do jedné směrodatné odchylky na každou stranu spadne 68,26 % případů (viz obr. 1).

Obr. 1: Křivka normálního rozložení a její základní charakteristiky (σ)



Převedeme-li tento fakt do empirické roviny, tak to znamená, že např. v IQ testech, kde se předpokládá, že průměr je 100 a směrodatná odchylka (σ) je 15, spadne 68 % populace mezi hodnoty 85 a 115 (tedy jednu σ na každou stranu od průměru 100) a 95 % populace se pohybuje mezi hodnotami 70 a 130 (viz obr. 2).

Obr. 2: Rozložení skóre v IQ testu

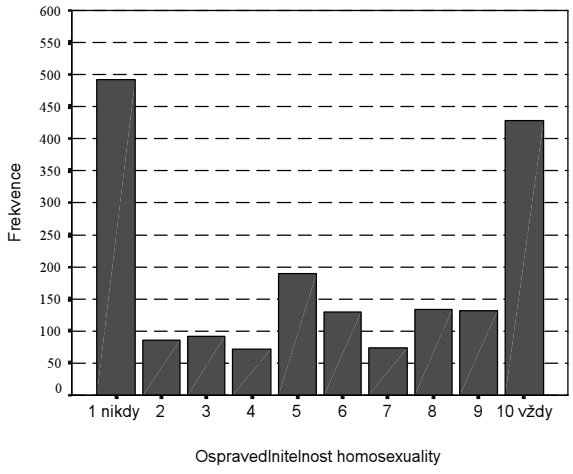


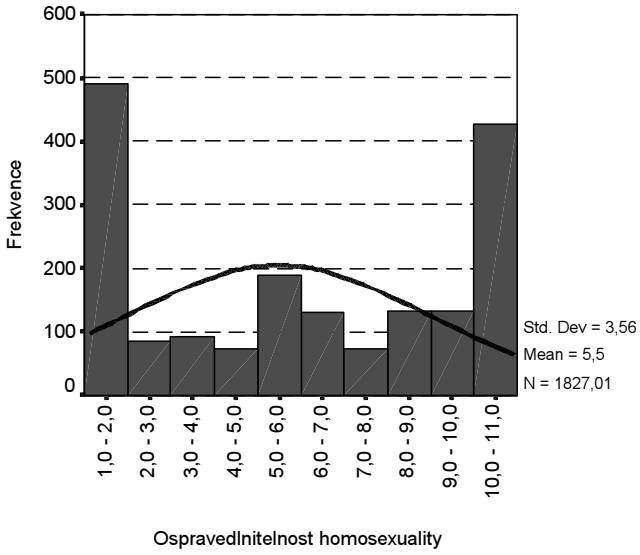
Jelikož v sociologii často pracujeme s daty výběrového souboru, musíme se zajímat nejenom o to, zdali jsou normálně rozloženy charakteristiky výběrového souboru, ale také, zdali toto normální rozložení můžeme očekávat i v základním souboru.

Příklad P5.1: Ve výzkumu EVS ČR1999 byla respondentům položena baterie otázek, zdali jsou různá lidská jednání (např. sebevražda, užívání drog, rozvod apod.) ospravedlnitelná či nikoliv. Postoj k těmto činnostem byl zjišťován na desetibodové stupnici, kdy 1 znamenala, že příslušné jednání není nikdy ospravedlnitelné a 10 znamenala, že takové jednání je vždy ospravedlnitelné. Naším úkolem je zjistit, zdali rozložení názorů na homosexualitu má tvar normálního rozložení a zdali můžeme předpokládat, že tato normální distribuce bude nalezena i v základním souboru.

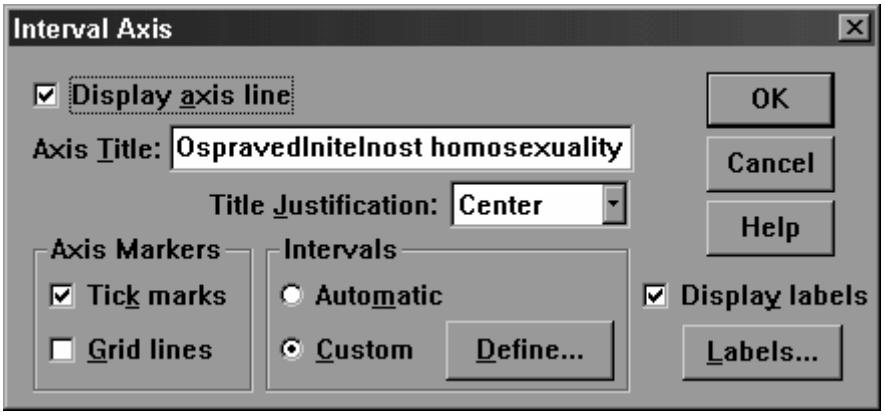
Řešení:

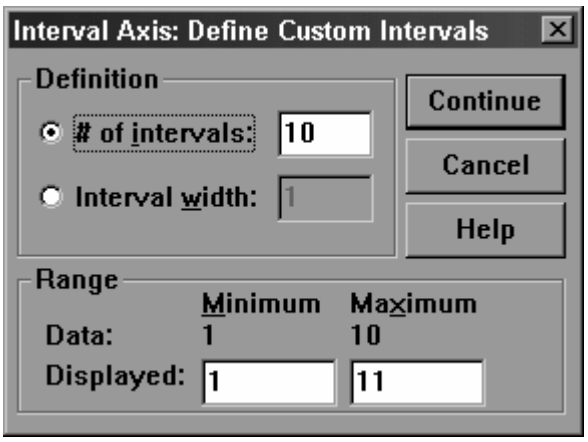
Nejdříve si nechejme nakreslit grafy rozložení této proměnné:





Poznámka: Pravý graf s vloženou křivkou normálního rozložení získáme editací grafu, který automaticky vytvoří SPSS. Podle níže uvedeného návodu upravíme nastavení osy X (*Interval Axis*), že zaškrtneme políčko *Interval Customs* a klikneme na *Define*. V objevivším se dialogovém okně budeme požadovat 10 intervalů a *Maximum Displayed Range* nastavíme na 11.





Oba grafy již pouhou okometrickou analýzou naznačují, že postoje respondentů ve výběrovém souboru mají daleko k normalitě (distribuce de facto připomíná U rozložení). K testování normality máme několik prostředků.

1. Testujeme, zdali se rozložení podstatně odlišuje od normality z hlediska šikmosti a špičatosti. K tomu potřebujeme výpočet šikmosti a špičatosti a jeho směrodatnou chybu. Tyto údaje nám poskytne procedura *Explore*, kterou již známe (*Analyze – Descriptive Statistics – Explore*). Když podělíme hodnoty šikmosti (*skewness*) a špičatosti (*Kurtosis*) jejich směrodatnými chybami (*Std. Error*) – nezáleží zde na znaménku – a výsledek vyjde vyšší než 2,5, je normalita tvaru rozložení z hlediska šikmosti či špičatosti porušena. V našem případě, jak vidíme z tabulky 4.1, je šikmost –0,017 a její směrodatná chyba 0,057. Poměr šikmosti k její směrodatné chybě je: 0,017 / 0,057 = 0,30. Špičatost je –1,565 a její směrodatná chyba 0,114, poměr je tedy 1,565 / 0,114 = 13,7. Obě charakteristiky tak ukazují, že rozložení nemá tvar rozložení normálního.

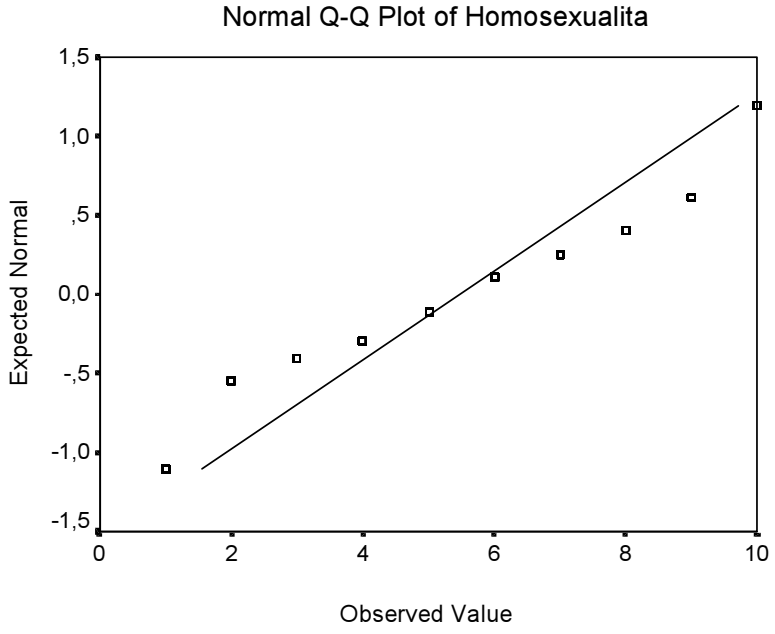
Tab. 5.1: Vypočtené charakteristiky proměnné Q65A_8, postoj k ospravedlnitelnosti či neospravedlnitelnosti homosexuality

Descriptives

			Statistic	Std. Error
Q65A_8 Homosexualita	Mean		5,47	,08
	95% Confidence Interval for Mean	Lower Bound	5,31	
		Upper Bound	5,64	
	5% Trimmed Mean		5,47	
	Median		5,00	
	Variance		12,641	
	Std. Deviation		3,56	
	Minimum		1	
	Maximum		10	
	Range		9	
	Interquartile Range		8,00	
	Skewness		-,017	,057
	Kurtosis		-1,565	,114

2. Prostřednictvím **formálního statistického testu** testujeme nulovou hypotézu, že data pocházejí z normálního rozložení. Výpočet získáme současně se zadáním grafického testu normality, tedy při proceduře
   Analyze – Descriptive Statistics – Explore — Plots — Normality Plots with Tests

Výstup:
a) grafický (tzv. Q-Q graf)



Pokud by se zobrazené body shlukovaly kolem přímky, naznačovalo by to, že data pocházejí z populace, v níž jsou normálně rozložena. V našem případě se rozložení postojů k homosexualitě od přímky odchyluje, což je indikací toho, že ani v populaci není tento postoj normálně rozložen.

b) statistický

Součástí výše uvedené procedury je i následující tabulka Kolmogorova-Smirnova testu normality:

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Q65A_8 Homosexualita	,165	1827	,000

a. Lilliefors Significance Correction

Pokud je vypočtená signifikance (ve sloupci *Sig.*) velmi nízká, to je menší než 0,05, máme dobré důvody pochybovat o tom, že předpoklad normality rozložení v základním souboru je správný.

* * *

Příklad P5.2: Naznačme si platnost centrální limitní věty (*central limit theorem*) která říká: ať je rozdělení základního souboru jakékoliv, rozdělení střední hodnoty výběrového souboru bude vždy normální, jestliže rozsah výběrového souboru dosáhne alespoň jisté minimální velikosti, jíž je velikost alespoň 30 (viz Swoboda, str. 153).

Jak známo, z populace je možné teoreticky udělat nekonečné množství výběrových souborů. Představme si nyní, že soubor 1 908 respondentů, kteří odpovídali na naše otázky ve výzkumu EVS ČR1999, je základním souborem, že jsme tedy provedli vyčerpávající zjišťování (de facto census) v nějakém malém státečku, který má 1908 obyvatel. Z tohoto základního souboru můžeme prostřednictvím SPSS udělat celou řadu náhodných výběrových souborů o velikosti, řekněme 20 % z celého souboru. Procedura k tomu je následující:

Data—Select Cases—Random Sample of Cases—Sample—Approximately… % of all cases

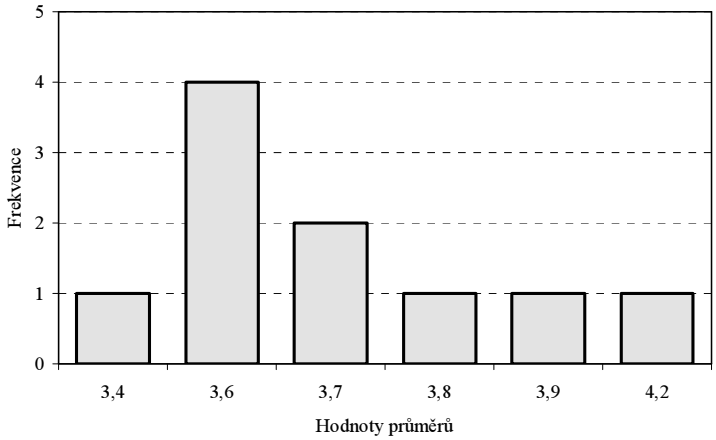
My těch náhodných výběrů uděláme pouze 10 a budeme sledovat, jak se mění hodnota průměru proměnné q33 *Jak důležitý je Bůh ve Vašem životě?* Výsledky uvádí tabulka 4.2. Připomeňme si, jak již víme z lekce druhé, že hodnota „skutečného“ průměru (to je průměru populace našeho imaginárního ministátu) byla 3,63 a jeho směrodatná odchylka 3,06.

Tab. 5.2: Různé náhodné výběry a měnící se hodnoty průměrů proměnné q33

Výběr	Průměr	Směrod. odchylka	N
1.	4,15	3,28	371
2.	3,58	3,05	375
3.	3,75	3,02	361
4.	3,85	3,12	406
5.	3,41	2,92	368
6.	3,74	3,11	371
7.	3,64	3,08	373
8.	3,58	2,97	388
9.	3,71	3,01	362
10.	3,56	3,05	367

Když z hodnot průměrů a jejich frekvence uděláme příslušný graf (do grafu jsme zanesli hodnoty průměrů zaokrouhlené na jedno desetinné místo), vidíme, že ač náš počet výběrů zdaleka nedosáhl třiceti, jak zní podmínka centrální limitní věty, nabývá rozložení průměrů tvaru, které připomíná normální rozložení (viz obr. 5.1).

Obr. 5.1: Rozložení hodnot průměrů proměnné q33 z deseti náhodně vybraných vzorků



Když navíc vypočteme z průměrů jednotlivých výběrů celkový průměr, dostaneme hodnotu 3,70, která není příliš vzdálena od průměru 3,63.

* * *

Příklad P5.3: Z skóry (standardizovaná směrodatná odchylka)

V některých úlohách potřebujeme porovnat, jak jsou vzdáleny jednotlivé hodnoty od průměru. Předpokládejme, že v testu ze statistické analýzy dat někdo získal 77 bodů a jiný 66 bodů. Když víme, že průměrný výsledek v testu byl 70 bodů, můžeme vypočítat, jaká je pozice těchto dvou výsledků vzhledem k celkovému rozložení hodnot výsledků testu. Nástrojem k tomu jsou tzv. Z-skóry. Potřebujeme k tomu znát kromě průměru navíc směrodatnou odchylku, neboť vzorec pro výpočet této charakteristiky je:

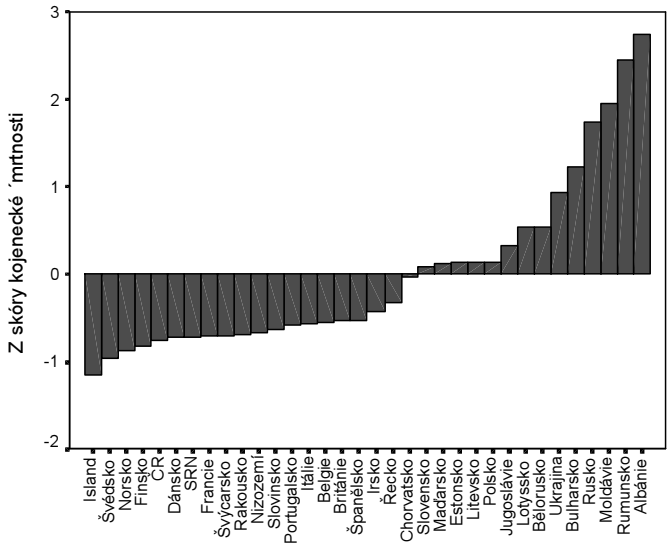
Z-skór = (hodnota znaku – průměr) / směrodatná odchylka.

Víme-li, že směrodatná odchylka od průměrného bodového skóre v testu z analýzy dat byla 5, pak výsledek 66 bodů umisťuje tohoto studenta do vzdálenosti -0,8 směrodatné odchylky od průměru (66–70/5 = -0,8) a výsledek 77 bodů do +1,4 směrodatné odchylky od průměru (77-70/5 = 1,4). Z-skór tedy říká, kolik standardních odchylek je určitý případ pod nebo nad průměrem. Je-li Z-skór roven 0, je na průměru, je-li 1, je jednu směrodatnou odchylku nad průměrem. Z-skóry lze nejen vypočítat, ale také uložit jako novou proměnnou a dále s ní pracovat.

Z-skóry např. umožňují srovnat relativní pozici respondenta z hlediska různých proměnných. Např. zjistíme, že respondent má v příjmu Z-skór 2,1 a ve vzdělání má -1. Znamená to tedy, že tato osoba je v příjmové kategorii více než dvě směrodatné odchylky nad průměrem (a když se podíváte na obrázek normálního rozložení, a uvědomíte si, že do plochy nad dvě směrodatné odchylky spadá jen 2,14 % případů s nejvyššími hodnotami nad průměrem, je to příjem velmi vysoký). Ve vzdělání je však pod průměrem. Je to tedy člověk, který ač má nízké vzdělání, patří mezi osoby s nejvyššími příjmy (kdopak to asi je?). Bez standardizace prostřednictvím Z-skóru by takovéto srovnání nebylo možné, neboť každá proměnná má jiné jednotky měření, odlišné průměry a odlišné směrodatné odchylky.

A nyní příklad z reálných dat. Z demografické statistiky máme údaje o kojenecké úmrtnosti (viz soubor *dmg-data.sav*, proměnná *kojen_um*). Když si prostřednictvím procedury *Descriptives* necháme uložit z-skóry této proměnné, uloží se nám na konec matice nová proměnná nazvaná *zkojen_u*. Pak si necháme celý soubor utřídit pomocí procedury *Data – Sort cases – Sort by (zkojen_u)*, čímž se pořadí matice změní tak, že v prvním řádku se objeví země s nejnižší hodnotou z-skóre kojenecké úmrtnosti (Island) a na posledním (34.) místě Albánie. Když si pak tuto novou proměnnou necháme zpracovat do grafu, získáme obrázek 5.2.

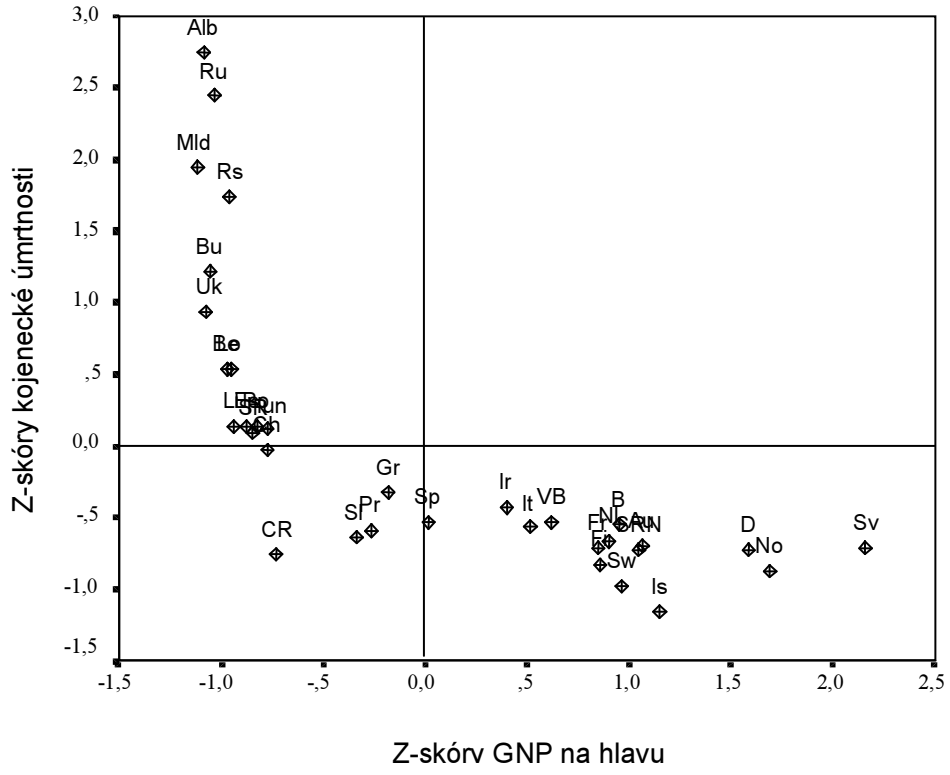
Obr. 5.2: Pořadí evropských zemí podle z-skórů kojenecké úmrtnosti v roce 1999.



Z dat víme, že průměr kojenecké úmrtnosti byl v Evropě v roce 1999 8,34 zemřelých dětí do jednoho roku na 1000 živě narozených a směrodatná odchylka byla 4,97. Z obrázku je pak patrné, jak mnoho se jednotlivé evropské země v tomto ukazateli odlišují. Na průměrné hodnotě je Chorvatsko, hodnoty nižší než průměr, mají všechny západoevropské země, k nimž se z bývalých komunistických zemí řadí pouze ČR (a zdůrazněme, že naše kojenecká úmrtnost je jedna z nejnižších na světě) a Slovinsko.

Nechejme si vypočítat z-skóry ještě pro další proměnnou tohoto souboru, a to pro proměnnou hrubý národní produkt na hlavu (*gnp_head*), který je uváděn v US dolarech na hlavu (tato data zachycují situaci v roce 1998). Když tuto novou proměnnou (*zgnp_hea*) umístíme do *scatter* grafu spolu se z-skóry kojenecké úmrtnosti, získáme velmi zajímavý obrázek (viz obr. 5.3). Říká nám, že země s nadprůměrným hrubým národním produktem – v Evropě byl průměr GNP na hlavu 13 899 US dolarů a směrodatná odchylka byla 12 086 – mají také obvykle podprůměrnou kojeneckou úmrtnost a naopak země s podprůměrným GNP (chudé země) mají obvykle vysokou kojeneckou úmrtnost.

Obr. 5.3: Evropské země podle hrubého národního příjmu na hlavu (GNP) a kojenecké úmrtnosti v roce 1999



ZOBECŇOVÁNÍ VÝBĚROVÝCH VÝSLEDKŮ NA ZÁKLADNÍ SOUBOR

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.

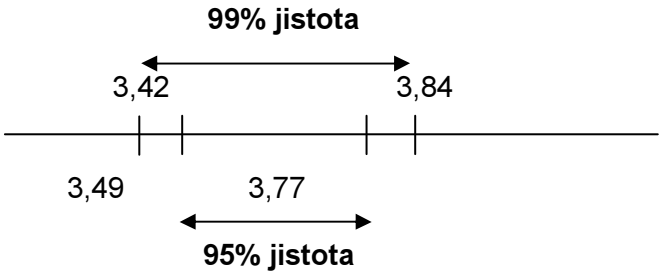
Pokud bychom chtěli mít interval spolehlivosti stanoven s jistotou 99 %, nastavíme v dialogovém okně v proceduře *Explore – Statistics* hodnotu intervalu spolehlivosti na 99 %. Lze ji ovšem vypočítat i ručně. Ruční výpočet je nesmírně jednoduchý. Hodnotu směrodatné chyby násobíme třemi a výsledek přičteme 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 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.

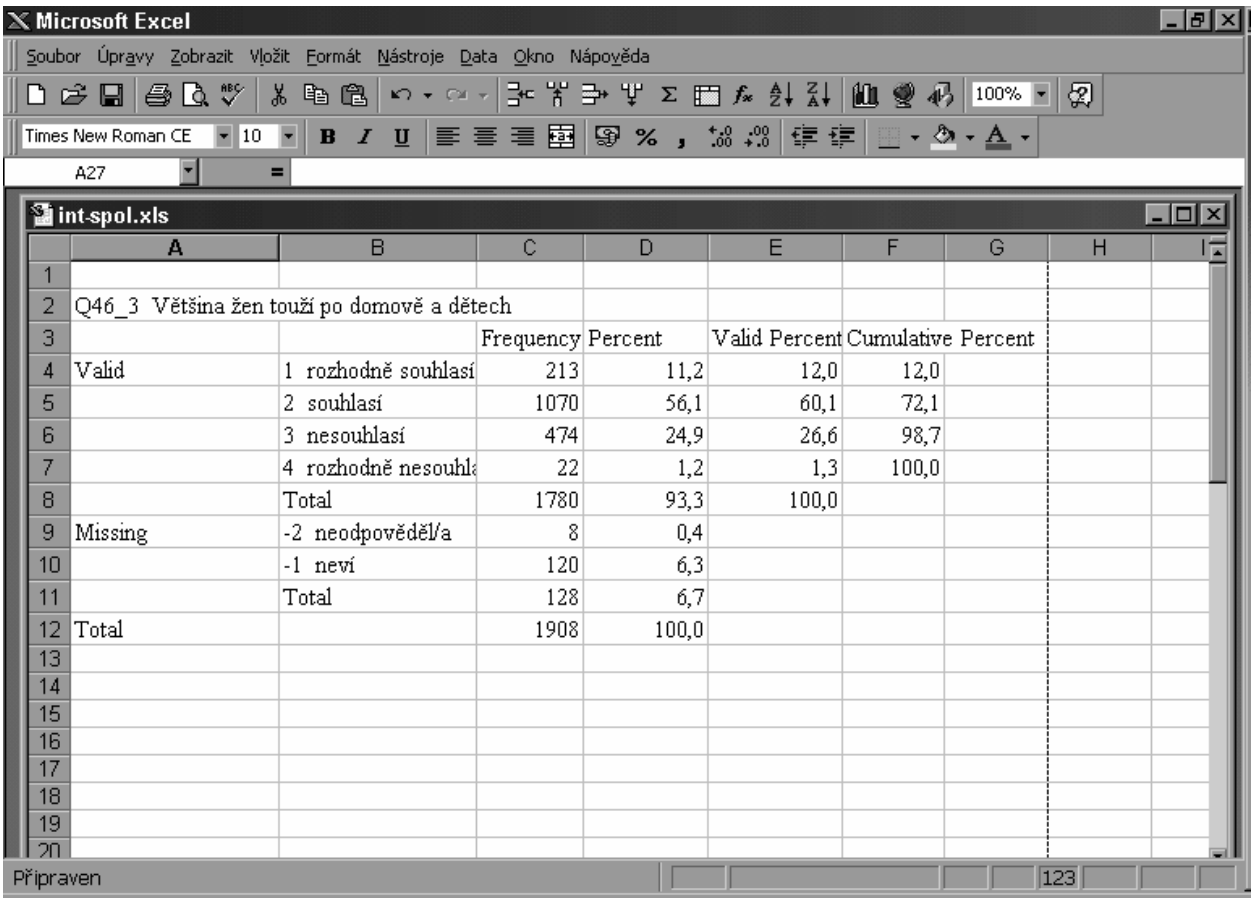
U vícehodnotového znaku se postupuje, pokud nechceme interval spolehlivosti počítat ručně, jinak. Tabulku, kterou v SPSS dostaneme z *Frequencies*, zkopírujme (prostřednictvím příkazu *Copy*) a vložme ji do Excelu. V něm si připravíme příslušný vzorec pro výpočet intervalu spolehlivosti:

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

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.

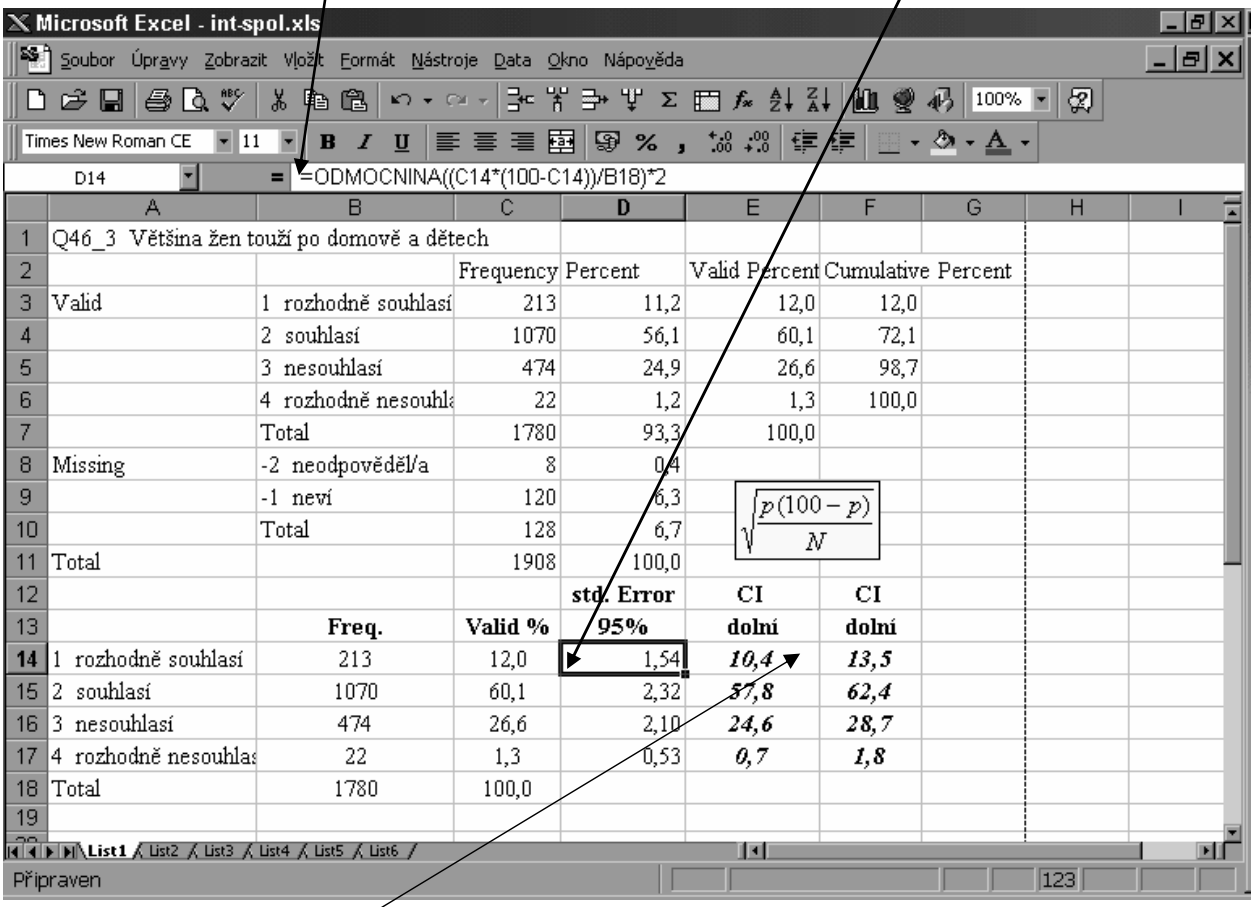
Ukázka:

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
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 nesouhl	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									

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

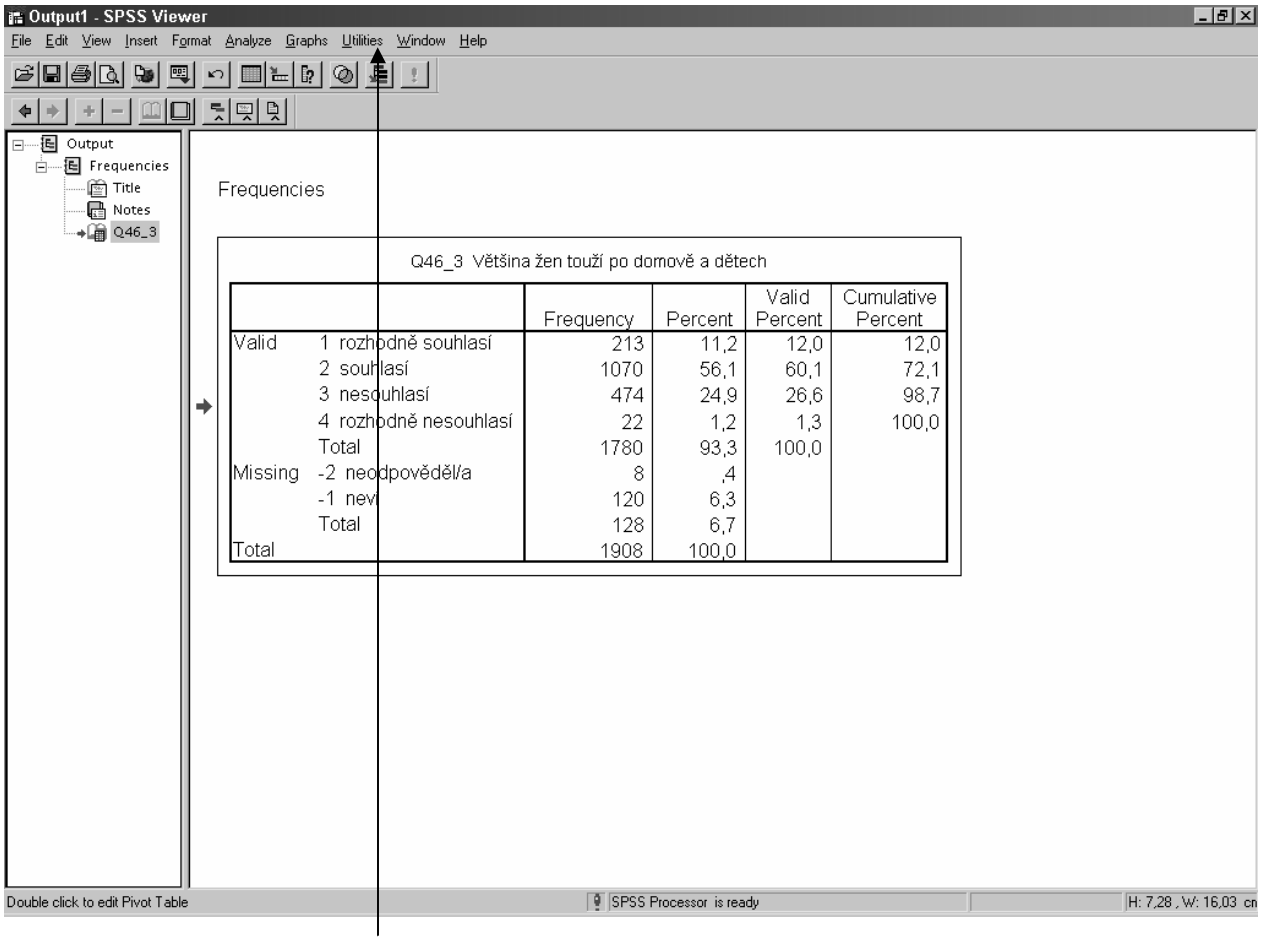
Další dva přidané odstavce jsou již přímo hodnoty dolního a horního intervalu spolehlivosti. Pod údajem ***10,4*** je vzorec = C14-D14 a pod údajem ***13,5*** je vzorec = C14+D14, tedy operace, kdy od 12 % těch, kdo rozhodně souhlasí s výrokem, že *Většina žen touží po domově a dětech*, nejdříve odečítáme 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

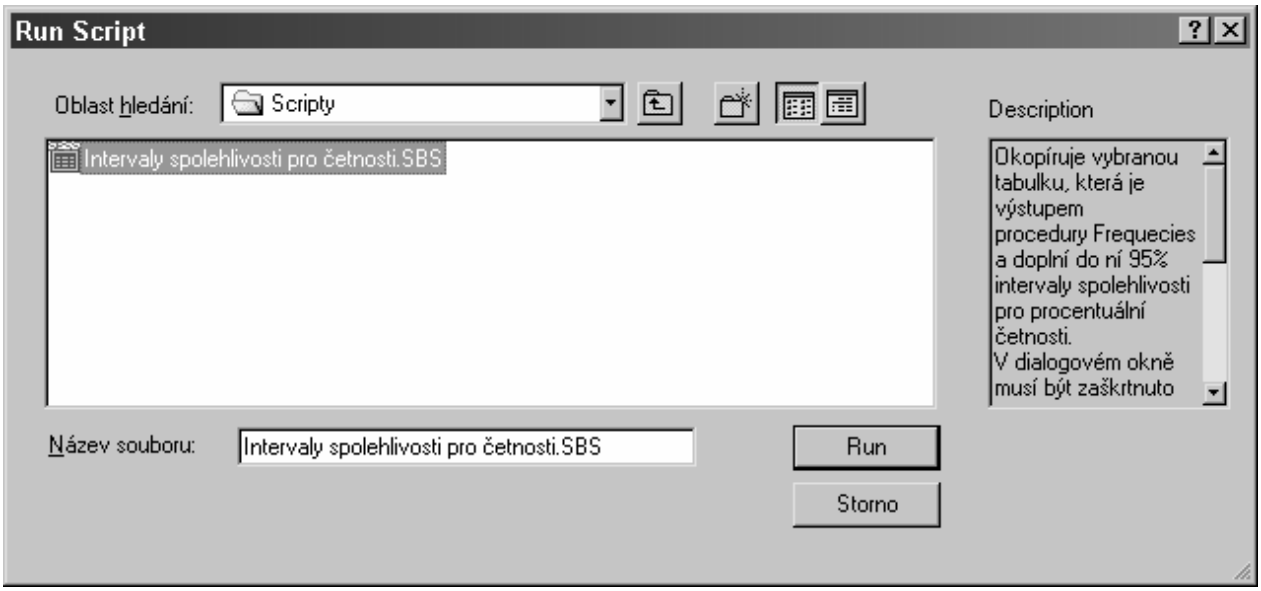
[1] Tuto matrici naleznete na dokumentovém serveru informačního systému MU.

počet variant než 4, budete si muset přidat příslušný počet řádků, do nichž vepíšete 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.