LEKCE 7
ZÁKLADY BIVARIAČNÍ ANALÝZY

Až dosud jsme se převážně zabývali analýzami, které byly založeny na srovnávání průměrů a rozptylů, tedy úlohami, kdy jedna z proměnných byla intervalové (kardinální) povahy. V sociologické analýze ovšem velmi často hledáme vztahy mezi proměnnými, u nichž nemá smysl průměry počítat. Buď z toho důvodu, že se jedná o znaky nominální (např. národnost respondenta) nebo proto, že proměnná je ordinální s malým počtem variant (např. proměnná typ lokality: 1. venkov, 2. město, 3. velkoměsto nebo např. příjmové kategorie) nebo že jde o proměnné dichotomické.

V bivariační analýze hledáme vztahy mezi dvěma proměnnými. Co to znamená? Nic jiného, než že se ptáme, do jaké míry jedna proměnná ovlivňuje druhou proměnnou. Např. při hledání vztahu mezi pohlavím respondenta a tím, zdali respondent preferuje hodnotu svobody či rovnosti se ptáme, zdali se muži a ženy budou lišit v názoru na to, je-li důležitější svoboda, nebo rovnost. A co znamená, že jedna proměnná ovlivňuje druhou? Mezi proměnnými existuje vztah, pokud rozložení (distribuce) hodnot jedné proměnné je asociováno (spojeno) s rozložením hodnot druhé proměnné. Řečeno jinak: hodnoty jedné proměnné jsou rozloženy (distribuovány) takovým způsobem, že nejsou rozloženy náhodně, ale jsou vzorovány v závislosti na rozložení hodnot druhé proměnné. Této situaci také říkáme, že mezi dvěma proměnnými existuje asociace.

Procedura, která nám pomůže otázky tohoto typu odpovědět, se nazývá třídění druhého stupně (třídíme totiž rozložení variant znaku jedné proměnné podle rozložení variant znaku druhé proměnné), v jazyce SPSS pak křížová tabulace (*crosstabulation*) – česky ovšem raději hovoříme o vytváření kontingenčních tabulek.

P7.1:

Na základě údajů v datovém souboru EVS-ČR1999 zjistěme, jak se liší názor na to, zdali je možné lidem důvěřovat (q8) v závislosti na věkových kategoriích (vek_kat1). Platí náš předpoklad, že s rostoucím věkem narůstá nedůvěra vůči lidem?

Procedura:

ANALYZE – DESCRIPTIVE STATISTICS — CROSSTABS – ROWS (vek_kat1) – *COLUMNS* (q8)

Tab. 7.1:

VEK_KAT1 Vekové kategorie * Q8 Důvěra v lidi Crosstabulation

Count

		Q8 Důvěra v lidi		
		1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
VEK_KAT1 Věkové kategorie	1 18-29	87	327	414
	2 30-49	158	508	666
	3 50+	199	585	784
Total		444	1420	1864

Z tabulky 7.1 vyčteme, že např. 87 respondentů ve věku 18–29 let si myslelo, že lidem je možné důvěřovat. Ve věkové skupině 50+ zastávalo tento názor 199 respondentů. I když by se zdálo, že starší respondenti tento názor zastávali častěji než respondenti mladší, *nemůžeme* z těchto údajů takovýto závěr učinit. Srovnáváme zde totiž nesrovnatelné. Jak je vidět v součtech řádků a sloupců (označené slovy *Total*), počty osob v jednotlivých kategoriích jsou různé a to znemožňuje srovnání. Abychom mohli naši úlohu vyřešit, musíme jednotlivé kategorie vyrovnat, neboli standardizovat.

Vyrovnat kategorie samozřejmě neznamená, že budeme nějak manipulovat s daty. Vyrovnání jednotlivých počtů provedeme tak, že necháme pro jednotlivá políčka tabulky vypočítat příslušná procenta a místo absolutních četností budeme srovnávat relativní četnosti, procenta.

Pravidlo 1: Při proceduře *Crosstabs* nemá smysl pracovat s absolutními četnostmi (*count*). Musíme je doplnit o výpočet příslušných procent

Před tím ale, než příslušný výpočet zadáme, musíme rozhodnout, jaká procenta budeme počítat. Máme totiž tři možnosti výpočtu procent: tzv. procenta řádková, sloupcová a celková.

Řádková procenta (*Row %)* se počítají tak, že absolutní četnost v políčku tabulky se dělí příslušným celkovým počtem případů v kategorii. Ten nalezneme ve sloupci *Total*. Tak např. řádkové procento pro 199 respondentů ze skupiny 50+let, kteří si myslí, že lidem je možné důvěřovat, je 199/784) * 100 = 25,4 %. Tento údaj čteme následovně: z respondentů ve věku 50 let a starším si 25 % myslí, že lidem se dá důvěřovat. Naopak 75 % (585 / 784 nebo v tomto případě i 100–25) je přesvědčeno, že člověk musí být ve styku s ostatními lidmi velmi opatrný.

Sloupcová procenta (*Column %*) se počítají analogicky, jen s tím rozdílem, že absolutní četnost v políčku se dělí příslušným celkovým počtem případů v kategorii, kterou ovšem nyní nalezneme v řádku označeném *Total*. Sloupcové procento pro 199 respondentů ve věku 50+ let, kteří si myslí, že lidem lze důvěřovat, je 44,8 % (199/444) * 100 = 44,8 %). Čteme: Ze všech respondentů, kteří si myslí, že lidem je možné důvěřovat, bylo 45 % ve věku 50 více let.

Celková procenta (*Total %*) pak získáme tak, že absolutní četnost v políčku dělíme celkovým počtem případů v souboru. Ten je uveden v křížovém součtu celkových počtů sloupců a řádků. Našich 199 respondentů ve věku 50 a více let, kteří si myslí, že lze důvěřovat, tedy tvoří: (199 / 1864) * 100 = 10,7 %. Čteme: ze všech respondentů našeho souboru bylo 11% těch, kdo měli 50 a více let a kdo jsou současně přesvědčeni, že lidem lze důvěřovat.

Vidíme tedy, že počet 199 respondentů jednou znamenal 25 %, podruhé 45 % a potřetí 11 %. Jelikož ve vědě jako v každé jiné činnosti platí také princip efektivity, tedy snaha dosahovat maximálních výsledků s minimálními vstupy, v analýze počítáme jen ta procenta, která jsou pro příslušnou úlohu adekvátní. V analýze dat proto necháme SPSS vypočítat jen ten druh procenta, který je pro řešení úlohy podstatný. Podstatně si tím i zjednodušíme náš analytický život. Jen posuďte, jak by vypadala tabulka, do níž byste nechali vypočítat všechna procenta (viz tab. 7.2)

Jak vybereme ta procenta, která jsou pro řešení úlohy podstatná? Lehce. Jediné, co musíme učinit, je rozhodnout, která proměnná je nezávislá – tedy ta, o níž předpokládáme, že je příčinou ovlivňující rozložení druhé (závisle) proměnné. V naší úloze je nezávisle proměnnou věk (věkové skupiny), neboť lze předpokládat, že postoj k jiným lidem z hlediska důvěry či nedůvěry bude ovlivňován právě věkem respondenta. Předpokládáme, že s narůstajícím věkem bude slábnout důvěra v ostatní lidi.

Víme-li, že naše nezávisle proměnná je umístěna v řádcích tabulky, necháme vypočítat řádková procenta. Tím dosáhneme toho, že všechny počty v kategoriích nezávisle proměnné vyrovnáme („položíme je za sto", jak říkají statistikové), což umožní smysluplné srovnání. O umístění proměnných do řádků či sloupců rozhodujeme sami v dialogovém okně.

Pravidlo 2: Umístíme-li nezávisle proměnnou do řádků kontingenční tabulky (Rows), použijeme v analýze údaje z řádkových četností. Umístíme-li ji do sloupců (Columns), pracujeme s četnostmi sloupcovými.

Tab. 7.2

VEK_KAT1 Vekové kategorie * Q8 Důvěra v lidi Crosstabulation

			Q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
VEK_KAT1 Vekové kategorie	1 18-29	Count	87	327	414
		Row %	21,0%	79,0%	100,0%
		Column %	19,6%	23,0%	22,2%
		Total %	4,7%	17,5%	22,2%
	2 30-49	Count	158	508	666
		Row %	23,7%	76,3%	100,0%
		Column %	35,6%	35,8%	35,7%
		Total %	8,5%	27,3%	35,7%
	3 50+	Count	199	585	784
		Row %	**25,4%**	74,6%	**100,0%**
		Column %	**44,8%**	41,2%	42,1%
		Total %	**10,7%**	31,4%	42,1%
Total		Count	444	1420	1864
		Row %	23,8%	76,2%	100,0%
		Column %	**100,0%**	100,0%	100,0%
		Total %	23,8%	76,2%	**100,0%**

Podívejme se tedy, jak by měla vypadat tabulka, s jejíž pomocí odpovíme na naši otázku (viz tab. 8.3).

Tab. 7.3

Case Processing Summary

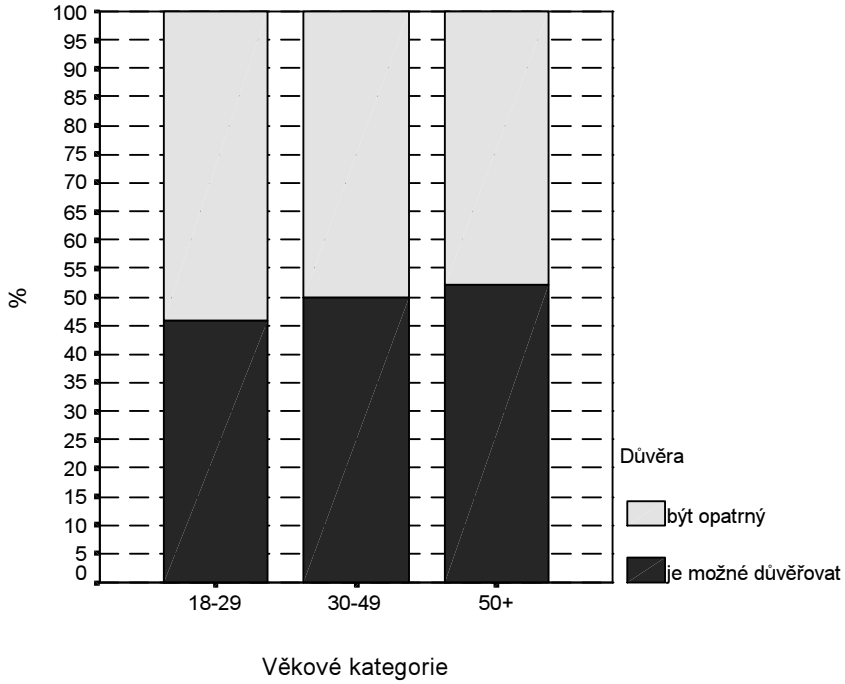
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
VEK_KAT1 Vekové kategorie * Q8 Důvěra v lidi	1864	97,7%	44	2,3%	1908	100,0%

VEK_KAT1 Vekové kategorie * Q8 Důvěra v lidi Crosstabulation

			Q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
VEK_KAT1 Vekové kategorie	1 18-29	Count	87	327	414
		Row %	**21,0%**	79,0%	100,0%
	2 30-49	Count	158	508	666
		Row %	**23,7%**	76,3%	100,0%
	3 50+	Count	199	585	784
		Row %	**25,4%**	74,6%	100,0%
Total		Count	444	1420	1864
		Row %	23,8%	76,2%	100,0%

Vidíme, že z celkového počtu respondentů na tuto otázku neodpovědělo 44 dotázaných. Do kontingenční tabulky jsou vždycky zahrnuti pouze ti, kdo mají údaje u obou proměnných – celkem jich bylo 1 864 (97,7 %). Výsledek třídění je poněkud překvapující. S narůstajícím věkem poněkud narůstá podíl osob, kteří si myslí, že lidem lze důvěřovat (a naopak klesá podíl těch, kdo si myslí, že člověk musí být ve styku s ostatními lidmi velmi opatrný). Rozdíly však nejsou nijak velké: 21 % : 24 % : 25 %, což ilustruje i graf 7.1.[1] Rozdíly mezi procenty v políčkách se nazývají epsilon (a značí se řeckým písmem ε). Hodnota epsilon pro respondenty ve věku 50+ a 18-29 je 25,4-21,0 = 4,4 %.

Obr. 7.1: Důvěra k lidem podle věkových kategorií



Pozn. Tento graf vznikl jako *Bar* graf s tím, že při jeho editaci kliknete myší na tlačítko *Chart*, pak na *Options* a v nich kliknutím zaškrtne *Change scale to 100 %*. Jiný způsob je, že při zadávání třídění si v proceduře Crosstabs zakliknete požadavek *Display clustered bar charts*. Obrázek potom editujete stejně jako v předchozím postupu přes Options atd. (viz výše).

[1] Opakujeme znovu, že grafy, které produkuje SPSS nejsou vůbec hezké a pro publikaci vašich výsledků je třeba grafy vyrobit v excelu. Jejich ukázku naleznete v souboru grafy-bivar.xls. Připomínám, že tabulku z SPSS lze vkopírovat do excelu (pomocí Ctrl C v SPSS a Ctrl V do excelu) a že je dobré si nechat pro tento účel vypočítat tabulku pouze s příslušnými procenty.

Rámeček 7.1

Při publikaci výsledků ovšem tabulku v takového podobě, jako jsou tab. 7.2 nebo 7.3, nikdy nezveřejňujeme. Nejsou totiž přehledné. Proto je musíme upravit. Zásady jsou následující:

1. Každá tabulka musí mít číslo a název.
2. Všechny popisky tabulky musí být česky.
3. Názvy proměnných jsou ve sloupcích a řádcích jasně vyjádřeny.
4. Nezávisle proměnnou obvykle umísťujeme do sloupců, takže počítáme sloupcová procenta.
5. Závisle proměnná, která je v řádku, by měla mít varianty uspořádány od nejvyšší po nejnižší (pokud je měřená na ordinální nebo intervalové úrovni). Tento požadavek se ale nemusí dodržovat příliš striktně.
6. Poslední řádek uvádí celková procenta (obvykle tedy 100 %) a současně i absolutní počty případů.
7. V poznámce pod tabulkou se uvádí zdroj dat.

Tabulka 7.3 by tedy podle těchto zásad musel být pro případnou publikaci upravena takto:

Tabulka 7.3: Důvěra k lidem podle věku (sloupcová %)

Důvěra k lidem	Věkové kategorie		
	18-29	30-49	50+
Lidem je možné důvěřovat	21	24	25
Člověk musí být ve styku s ostatními lidmi opatrný	79	76	75
Celkem	100 % (414)	100 % (666)	100 % (784)

Pramen: EVS ČR 1999

Jelikož v analýze dat zhruba platí pravidlo, že teprve rozdíl (epsilon), který se blíží 10 %, indikuje analyticky podstatný rozdíl (to je takový, který nevznikl náhodou), vyslovujeme závěr, že v otázce důvěry k lidem se čeští respondenti nelišili v závislosti na jejich věku. Zamítáme ovšem naši výzkumnou hypotézu, že s narůstajícím věkem bude také narůstat nedůvěra v ostatní lidi.

Tento příklad je dobrou ukázkou toho, že i „nula“ ve vědě je důležitým poznatkem. My jsme zjistili, že mezi věkovými skupinami není v zásadě rozdíl v postoji k důvěře v ostatní. Tento nulový rozdíl (tato zjištěná „nula“) v sobě ovšem obsahuje podstatný fakt a nový poznatek. Na základě tohoto výsledku jsme si totiž museli opravit naši domněnku, že starší lidé budou vůči ostatním lidem nedůvěřivější než ti mladší.

Pravidlo 3: I nula (nulový rozdíl, nulový výsledek) znamená ve vědě podstatný poznatek.

* * *

V našem příkladě jsme hledali vztah mezi kategorizovaným věkem a postojem k jiným lidem z hlediska důvěry. Tuto úlohu jsme mohli řešit i jinak. Jelikož naše data obsahují údaje o věku v jeho nekategorizované podobě (proměnná *vek*), lze srovnat, zdali se liší průměrný věk osob u lidí, kteří si myslí, že lidem lze důvěřovat, a u lidí, kteří se domnívají, že ve styku s jinými lidmi musí být člověk opatrný. Jelikož zde máme pouze dvě kategorie, musíme použít t-testu.

Výsledek:

Group Statistics

	Q8 Důvěra v lidi	N	Mean	Std. Deviation	Std. Error Mean
VEK	1 lidem je možné důvěřovat	445	**46,95**	16,26	,77
	2 člověk musí být opatrný	1419	**45,41**	16,97	,45

Rozdíl v průměrném věku není příliš velký, neboť věkový průměr je v obou kategoriích podobný. Proto také, jak ukazuje následující výstup, je rozdíl statisticky nevýznamný, jenž nám velí podržet nulovou hypotézu o tom, že věkový rozdíl nebude statisticky signifikantní. Jinou úlohou jsme dospěli ke stejnému výsledku, takže si můžeme být dost jisti, že mezi věkem (ať v jeho hrubé kategorizaci do tří skupin respondentů mladšího, středního a staršího věku, nebo v jeho „přirozené“, nekategorizované podobě) a názorem na důvěru k lidem není souvislost.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
VEK	Equal variances assumed	3,494	**,062**	1,685	1862	**,092**	1,54	,91	-,25	3,33
	Equal variances not assumed			1,722	770,843	,085	1,54	,89	-,21	3,29

* * *

Pouhé třídění dvou proměnných a výpočet příslušných procent, byť se jedná o velmi mocnou analytickou proceduru, nestačí k tomu, abychom hledanému vztahu mezi dvěma proměnnými dobře rozuměli. Odhalíme-li totiž, že mezi sledovanými proměnnými je vztah, musíme se dále zajímat o to, zdali jednak tento vztah vydrží i test nezávislosti, jednak jakou má tento vztah sílu.

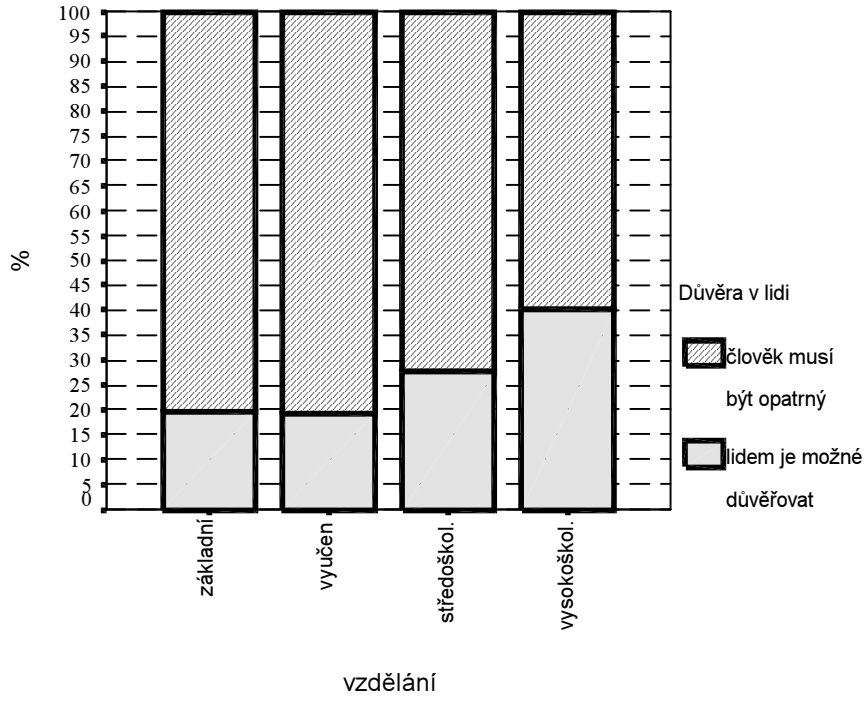
Test nezávislosti chí-kvadrát (χ^2)

P7.2: Hledejme v datech EVS-ČR1999 odpověď na otázku, zdali je důvěra v lidi ovlivněna vzděláním respondentů. Naše výzkumná hypotéza bude znít, že se zvyšujícím se vzděláním bude narůstat podíl těch, kteří si myslí, že lidem je možné důvěřovat a že tento vztah statisticky významný.

Řešení:

Obr. 7.2, jakož i tabulka 7.4 ukazují, že mezi jednotlivými vzdělanostními kategoriemi existují rozdíly v názorech na důvěru k jiným lidem, přičemž lidé se středoškolským a vysokoškolským vzděláním mají tendenci více důvěřovat ostatním, než lidé se vzděláním základním a vyučen. Nás samozřejmě zajímá, zdali tento rozdíl nebyl způsoben náhodou, to je výběrovou chybou, anebo zda máme dostatek evidence k tomu, abychom mohli zamítnout nulovou hypotézu, že v základním souboru bude tento podíl souhlasících mužů a žen stejný.

Obr. 7. 2: Důvěra v lidi podle úrovně dosaženého vzdělání



Tab. 7.4:

VZDELÁNÍ kategorizace q94 * Q8 Důvěra v lidi Crosstabulation

			Q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
VZDĚLÁNÍ kategorizace q94	1 základní	Count	71	292	363
		Row %	19,6%	80,4%	100,0%
	2 vyučen	Count	145	618	763
		Row %	19,0%	81,0%	100,0%
	3 SŠ	Count	151	395	546
		Row %	27,7%	72,3%	100,0%
	4 VŠ	Count	78	116	194
		Row %	40,2%	59,8%	100,0%
Total		Count	445	1421	1866
		Row %	23,8%	76,2%	100,0%

Test provedeme na základě výpočtu statistiky chí-kvadrát χ^2 (*chi-square*). Ten je založen na srovnání empirických a očekávaných četnostech.

Empirická četnost (*observed count*) — pozorovaná hodnota v políčku tabulky
Očekávaná četnost (*expected count*) — četnost, která by se v políčku objevila, kdyby platila nulová hypotéza

Podívejme se nyní do výstupu 7. 4. Vidíme, že v prvním políčku máme respondenty se základním vzděláním, kteří si myslí, že lidem je možné důvěřovat. Bylo jich celkem 71, což je empirická četnost. Očekávaná četnost pro toto políčko se vypočítá velmi snadno: násobíme marginální četnost

příslušného sloupce a marginální četnost příslušného řádku a tento součin podělíme celkovým součtem případů v tabulce. Konkrétně tedy: 363 (celkový počet případů v řádku tohoto políčka) * 445 (celkový počet případů ve sloupci tohoto políčka) / 1866 (celkový počet případů v tabulce) = 86,6. Toto je vyšší hodnota než ta, kterou jsme my zjistili empiricky (71 případů), tedy v našem výzkumu. Tento rozdíl nás však ještě neopravňuje k žádnému závěru. Musíme provést další početní operace, to je vypočítat tímto způsobem očekávané četnosti pro všechna pole tabulky. V každém poli tabulky pak musíme vypočítat rozdíl mezi empirickou a očekávanou četností, ten umocnit na druhou, podělit hodnotou očekávané četnosti a jednotlivé výsledky sečíst. Tím získáme hodnotu chí-kvadrát. Tu pak — jako při každém testování nulové hypotézy — porovnáme s matematickým modelem rozložení, v tomto případě s modelem chí-kvadrát a zjistíme statistickou významnost.

Všechny tyto operace za nás samozřejmě provede SPSS a pokud bychom chtěli, můžeme tento výpočet kontrolovat. V *crosstabsu* si totiž můžeme navolit všechny požadované informace tak, že v dialogovém okně *Cells* zaškrtneme v boxu *Counts* také políčko *Expected* a v boxu *Residuals* políčko *Unstandardized*. Dostaneme tento výstup (viz tab. 7.5)

Tab. 7. 5: Očekávané četnosti a rezidua v proceduře Crosstabs

VZDĚLÁNÍ * Q8 Důvěra v lidi Crosstabulation

			Q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
VZDĚLÁNÍ kategorizace q94	1 základní	Count	71	292	363
		Expected Count	86,6	276,4	363,0
		Row %	19,6%	80,4%	100,0%
		Residual	-15,6	15,6	
	2 vyučen	Count	145	618	763
		Expected Count	182,0	581,0	763,0
		Row %	19,0%	81,0%	100,0%
		Residual	-37,0	37,0	
	3 SŠ	Count	151	395	546
		Expected Count	130,2	415,8	546,0
		Row %	27,7%	72,3%	100,0%
		Residual	20,8	-20,8	
	4 VŠ	Count	78	116	194
		Expected Count	46,3	147,7	194,0
		Row %	40,2%	59,8%	100,0%
		Residual	31,7	-31,7	
Total		Count	445	1421	1866
		Expected Count	445,0	1421,0	1866,0
		Row %	23,8%	76,2%	100,0%

Řádek *Residual* udává numerický rozdíl mezi empirickou (*Count*) a očekávanou (*Expected Count*) četností. Má-li znaménko +, znamená to, že empirická četnost je vyšší, než bychom očekávali, kdyby platila nulová hypotéza, záporné znaménko vyjadřuje pravý opak, tedy že empirická četnost je nižší, než jaká by měla být, kdyby platila nulová hypotéza. V rutinní analytické praxi informace tohoto druhu nepotřebujeme, a proto takto detailní tabulku nevyžadujeme.

Test chí-kvadrát pro naši úlohu zadáme následovně:

ANALYZE – DESRIPTIVE STATISTICS — CROSSTABS — v dialogovém okně klikneme na lištu *Statistics* a v objevivším se novém dialogovém oknu zaškrtneme políčko *Chi-square*

Výsledkem je tato tabulka (tab. 7.6):

Tab. 7.6:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	**46,479[a]**	3	**,000**
Likelihood Ratio	43,813	3	,000
Linear-by-Linear Association	36,316	1	,000
N of Valid Cases	1866		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 46,26.

Hodnota Pearsonova chí-kvadrátu je 46,479 a její dvoustranná hladina významnosti 0,000. Musíme proto zamítnout nulovou hypotézu o nezávislosti vztahu mezi vzděláním a názorem na důvěru v ostatní lidi a očekáváme, že i v základním souboru se lidé budou ve své důvěře v ostatní odlišovat v závislosti na tom, jakého vzdělání dosáhli.

Další údaje v tabulce 7.6 nejsou v dané situaci zajímavé[2]. Pozornost ale bychom vždy měli věnovat poznámce pod tabulkou. Pokud totiž data poruší jeden z důležitých předpokladů chí-kvadrátu, totiž, že ne více než 20 % políček má očekávanou četnost menší než 5 a že minimální očekávaná četnost nesmí být menší než 1, je použití chí-kvadrátu (a koeficientů asociace, které jsou na něm založeny, jak uvidíme později) nekorektní. V případě, že dojde k porušení těchto předpokladů, v poznámce pod tabulkou se objeví varování.

Test chí-kvadrát je možno také chápat také jako test nezávislosti, kdy testujeme, zdali jedna proměnná závisí na druhé. Můžeme např. testovat hypotézu, zdali existuje nějaká souvislost mezi rodinným stavem respondenta a volebními preferencemi. Je to opět úloha na *Crosstabs*, ale v jejím rámci si ukážeme, jak je možné v rutinní analytické práci postupovat.

Test chí-kvadrát v tabulce 7.7 (proměnná q89 „rodinný stav“ byla rekódována, neboť obsahovala kategorii „odloučeni“, v níž bylo pouze 7 osob – tak málo obsazená kategorie způsobuje při třídění potíže, proto jsme ji sloučili s obsahově blízkou kategorií „rozveden/a“)[3] říká, že nemůžeme přijmout hypotézu o nezávislosti těchto dvou proměnných, neboť statistická signifikance je menší než 0,05.

[2] *Continuity Correction* je Yatesovou korekcí (opravou) Pearsonova chí-kvadrátu pro tabulky 2x2, tedy tabulky, v nichž obě proměnné jsou dichotomické, takže mají každá jen dvě varianty; mnozí totiž tvrdí, že v tabulce 2x2 dochází při standardním výpočtu chí kvadrátu k přecenění jeho hodnot, proto musí být výpočet upraven; *Likelihood Ratio* je statistika velmi podobná chí-kvadrátu a pro velké vzorky dosahuje velmi podobných hodnot (viz); Fisherův exaktní test (*Fisher's Exact Test*) můžeme jako sociologové směle ignorovat; *Linear-by-Linear Association* je míra lineárního vztahu mezi proměnnými. Má smysl jen v tom případě, kdy kategorie obou proměnných jsou uspořádány od nejnižší k nejvyšší. Může se tedy použít jako test linearity avšak obě proměnné musejí být minimálně ordinální.

[3] Syntax pro transformaci proměnné je:

```
RECODE q89
(1=1)  (2=2)  (5=4)  (3 thru 4=3)  (ELSE=SYSMIS)  INTO  rod_stav .
VARIABLE LABELS rod_stav 'rodinný stav (sloučeno rozvedený + odloučení)'.
EXECUTE .
```

Tab. 7.7: Volební preference podle rodinného stavu respondenta a test chí-kvadrát

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	66,448[a]	15	**,000**
Likelihood Ratio	63,751	15	,000
Linear-by-Linear Association	1,516	1	,218
N of Valid Cases	1407		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,45.

Což tedy znamená, že volební preference byly rodinným stavem respondenta nějakým způsobem ovlivněny – vidíme např., že podporu komunistům vyjadřovali především vdovci a vdovy (24 %), naopak svobodní a svobodné by je téměř nevolili (4 %).[4]

ROD_STAV rodinný stav (sloučeno rozvedený + odloučení) * PREFEREN volební preference 99 Crosstabulation

			PREFEREN volební preference 99						Total
			1 KSČM	2 ČSSD	3 KDU	4 US	5 ODS	90 nevolil by	
ROD_STAV rodinný stav (sloučeno rozvedený + odloučení)	1 ženatý/vdaná	Count	101	196	58	97	253	198	903
		Row %	11,2%	21,7%	6,4%	10,7%	28,0%	21,9%	100,0%
		Adjusted Residual	-,4	1,9	-1,1	-1,2	1,6	-1,6	
	2 vdovec/vdova	Count	36	25	15	9	29	35	149
		Row %	24,2%	16,8%	10,1%	6,0%	19,5%	23,5%	100,0%
		Adjusted Residual	**5,2**	-1,1	1,6	**-2,2**	**-2,1**	,1	
	3 rozvedený/á	Count	13	14	4	11	27	38	107
		Row %	12,1%	13,1%	3,7%	10,3%	25,2%	35,5%	100,0%
		Adjusted Residual	,2	-1,9	-1,4	-,4	-,3	**3,1**	
	4 nikdy neoženěn/nepro vdána	Count	11	49	21	45	66	56	248
		Row %	4,4%	19,8%	8,5%	18,1%	26,6%	22,6%	100,0%
		Adjusted Residual	**-3,8**	-,2	1,0	**3,6**	,0	-,3	
Total		Count	161	284	98	162	375	327	1407
		Row %	11,4%	20,2%	7,0%	11,5%	26,7%	23,2%	100,0%

Výsledky třídění v tab. 7.7 lze ale ještě dále specifikovat. Poslouží nám k tomu údaje tzv. **adjustovaných reziduí** (*Adjusted Residual*), které jsme si nechali do tabulky 7.7 vypočítat.

Adjustované reziduum je založeno na rozdílu mezi empirickou a očekávanou četností (jak jsme si ukázali v tab. 7.5). Řečeno jazykem statistiky, je to rozdíl mezi frekvencí očekávanou (f_o) a frekvencí empirickou (f_e). Tomuto rozdílu se říká delta a značí se odpovídajícím řeckým písmenem (Δ).
V adjustovaném reziduálu je pak tento rozdíl testován z hlediska statistické významnosti, přičemž platí, že pokud je jeho hodnota vyšší než 2,00, můžeme si být s 95% pravděpodobností jisti, že v daném políčku je rozdíl mezi empirickou a očekávanou četností statisticky významný a že tedy nevznikl náhodou. Interpretačně má tato informace obrovský význam, neboť nám umožňuje detailní

[4] Aniž bychom chtěli předbíhat vaše znalosti, je třeba učinit poznámku o možném vlivu třetí proměnné, která do tohoto vztahu může intervenovat.

vhled do vztahu mezi proměnnými. Tak např. vidíme, že v řádku těch, kdo preferují KSČM, máme dvě statisticky významná adjustovaná rezidua (pro lepší orientaci jsou zvýrazněny). U vdovců/vdov je hodnota rezidua 5,2. To znamená, že vdovci a vdovy by volili komunisty významně častěji, než by odpovídalo předpokladu nezávislosti. Naopak svobodní respondenti by komunisty volili mnohem méně častěji (Adj. res. = -3,8), než by odpovídalo hypotéze nezávislosti. Statisticky významně častěji by svobodní volili Unii svobody.

Tento statistický vhled do dat nám pomáhá detailněji prozkoumat, do jaké míry je možné výsledky třídění (frekvenci určitého políčka tabulky) očekávat i v základním souboru. Celou analýzu je možné ještě zjednodušit, když použijeme Řehákova skriptu *Znaménkové schéma.SBS*. Tento prográmek udělá to, že namísto reziduí vloží do příslušných políček znaménkové schéma, které nám ukáže, jak velký je rozdíl mezi očekávanou a empirickou četností. Podmínkou je, že musíte nechat SPSS vypočítat tabulku adjustovaných reziduí. Ukažme si celou proceduru.

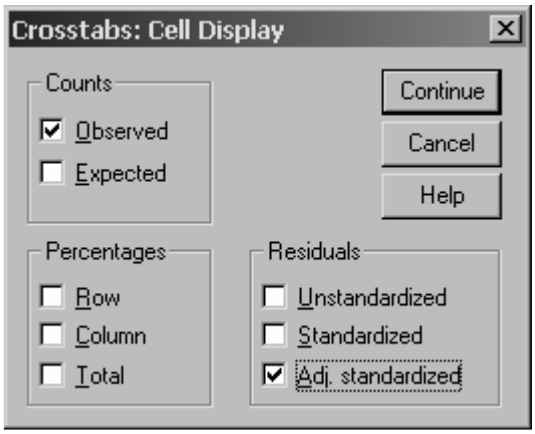
Zajímá nás, zdali rozdíly v názoru na to, proč jsou u nás lidé, kteří žijí v nouzi, jsou podmíněny věkem. Rozložení ukazuje tab.7.8.

Tab. 7.8: Názory na příčinu chudoby podle věkových kategorií

Q11_REC Proc lidé zijí v nouzi * VEK_KAT3 tri vekove skupiny Crosstabulation

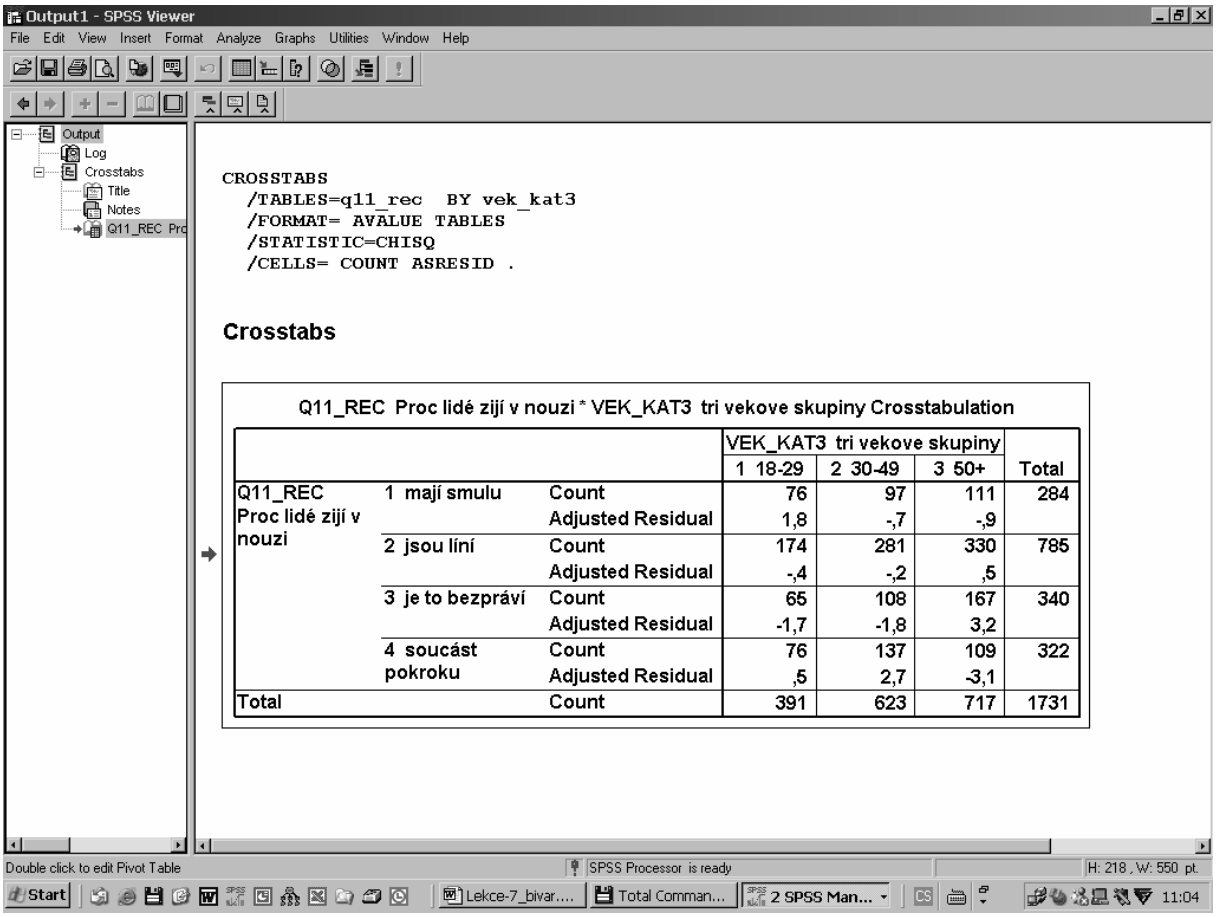
			VEK_KAT3 tri vekove skupiny			
			1 18-29	2 30-49	3 50+	Total
Q11_REC Proc lidé zijí v nouzi	1 mají smulu	Count	76	97	111	**284**
		Column %	19,4%	15,6%	15,5%	**16,4%**
	2 jsou líní	Count	174	281	330	**785**
		Column %	44,5%	45,1%	46,0%	**45,3%**
	3 je to bezpráví	Count	65	108	167	**340**
		Column %	16,6%	17,3%	23,3%	**19,6%**
	4 soucást pokroku	Count	76	137	109	**322**
		Column %	19,4%	22,0%	15,2%	**18,6%**
Total		Count	**391**	**623**	**717**	**1731**
		Column %	**100,0%**	**100,0%**	**100,0%**	**100,0%**

Vidíme, že rozdíly mezi věkovými kategoriemi nejsou příliš velké, např. u příčiny „mají smůlu“ se podíly pohybují mezi 16 a 19 %. Nasaďme na tuto tabulku Řehákův skript. Musíme, si ale nejdříve, jak již řečeno výše, nechat vypočítat tabulku reziduí. Zadání vypadá takto:



Výsledkem je výstup – obr. 7.3.

Obr. 7.3:



Když na tuto tabulku, kterou navíc označíme červenou šipkou (klikneme na ni jedenkrát myší) spustíme proceduru *Utilities – Run script – Znaménkové schéma*, získáme výstup 7. 10.

Tab. 7.10: Znaménkové schéma pro vztah mezi věkem a názory na příčiny chudoby

Q11_REC Proc lidé zijí v nouzi * VEK_KAT3 tri vekove skupiny Crosstabulation

			VEK_KAT3 tri vekove skupiny			
			1 18-29	2 30-49	3 50+	Total
Count	Q11_REC Proc lidé zijí v nouzi	1 mají smulu	76	97	111	284
		2 jsou líní	174	281	330	785
		3 je to bezpráví	65	108	167	340
		4 soucást pokroku	76	137	109	322
	Total		391	623	717	1731
Adjusted Residual	Q11_REC Proc lidé zijí v nouzi	1 mají smulu	o	o	o	
		2 jsou líní	o	o	o	
		3 je to bezpráví	o	o	++	
		4 soucást pokroku	o	++	--	

Znaménkové schéma v dolní polovině tabulky ukazuje, kde jsou statisticky významné rozdíly mezi empirickými a očekávanými četnostmi. Počet symbolů indikuje totéž, jako v případě znamének u t-testu (viz lekce 6); symbol + pak znamená, že empirické četnosti jsou vyšší než očekávané, symbol – pak, že empirické četnosti jsou nižší než očekávané:

\+ ..... alfa = 0,05 (empirické četnosti vyšší než očekávané)
\+ + ... alfa = 0,01
\+ + + ... alfa = 0,001

– ..... alfa = 0,05 (empirické četnosti nižší než očekávané)
– – ..... alfa = 0,01
– – – ..... alfa = 0,001

Respondenti ve věkové skupině 30–49 let statisticky významně častěji než náhodně (na hladině významnosti 0,01) odpověděli, že příčina chudoby je to, že je to prostě součást pokroku. Respondenti starší 50 let pak významně častěji (alfa = 0,01) říkali, že příčinou chudoby je bezpráví ve společnosti a naopak statisticky významně méně častěji (alfa = 0,01) tvrdili, že chudoba je nezbytnou součástí pokroku.

* * *

V této kapitole jsme si ukázali, jak hledat asociaci mezi proměnnými. Na závěr si připomeňme, co o asociaci říkají Loether a McTavish (1988), jejichž text máte ve vaší čítance ke kursu (viz kapitolu 8).

Při zkoumání bivariační asociace bychom měli hledat čtyři následující charakteristiky:

1. Zdali asociace *existuje*, či nikoliv,
2. jak je asociace *silná* (těsná) – to je do jaké rozložení variant jedné proměnné určují rozložení variant druhé proměnné,
3. jaký má asociace *směr* – to je, zdali se jedná o asociaci kladnou, nebo zápornou,
4. jakou má povahu – zdali je monotónní (lineární) či jiného druhu

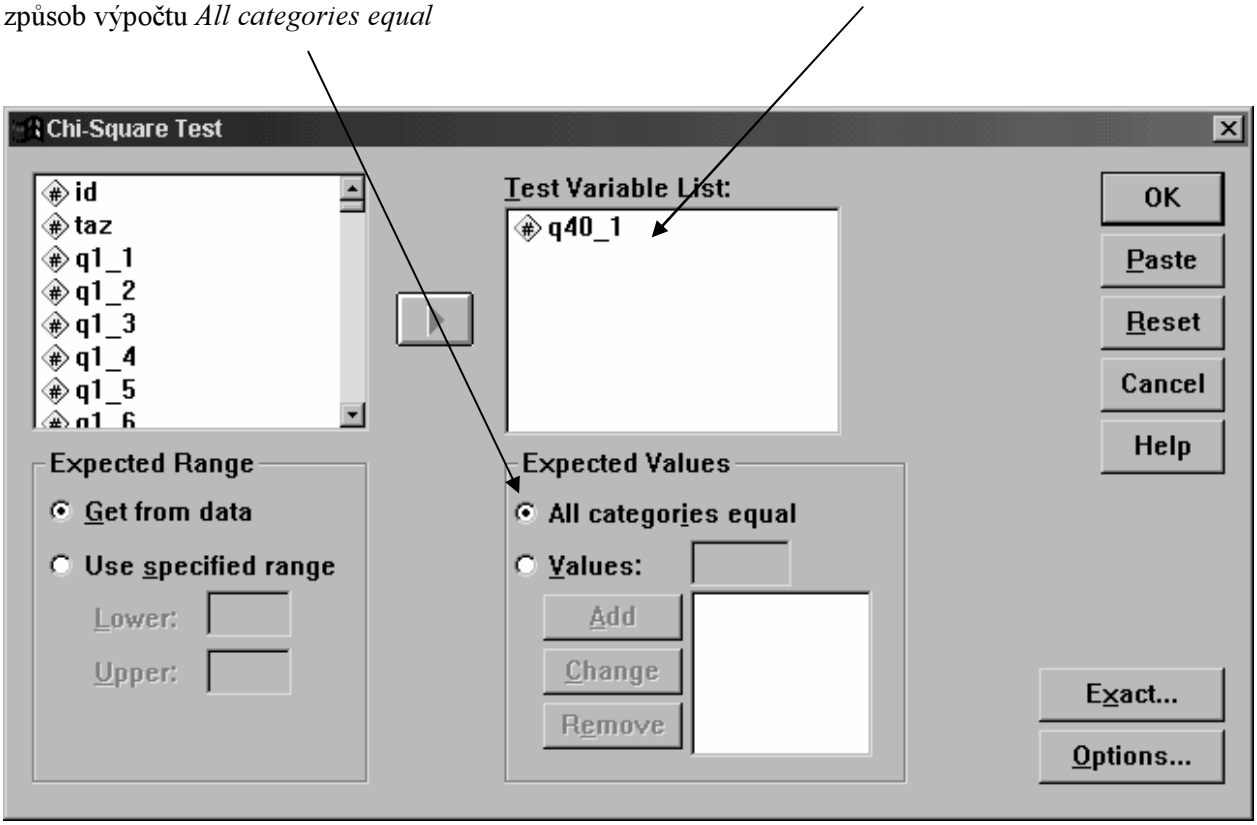
O tom, jak tyto charakteristiky zjišťujeme (měříme), hovoří problematika lekce 8 – měření síly asociace.

Zde jsme se naučili používat test chí-kvadrát pro odhalení asociace. Tohoto testu lze ale použít ještě pro jeden účel, totiž pro testování hypotéz o rozložení hodnot jediné proměnné. Touto úlohou se na chvíli opět vracíme do analýzy univariační, neboť nás bude zajímat, zdali se empirické rozložení kategorií jedné proměnné odlišuje od předpokládané distribuce této proměnné.

Jako nulovou hypotézu můžete např. stanovit, že je pravděpodobné, že rozložení osob, které budou zastávat názor, že věrnost je pro úspěšné manželství velmi důležitá, spíše důležitá a nepříliš důležitá, bude rovnoměrně stejné. Proveďme si tento test, kterému se v SPSS říká THE ONE SAMPLE CHI-SQUARE TEST (Chí-kvadrát pro jeden výběr) na příslušných datech – máme je v souboru EVS99_cvicny, proměnná q40_1. Použijeme testu chí-kvadrát, který je ve skupině procedur pod názvem *Nonparametric tests*.

Analyze – Nonparametric tests – Chi-Square

V dialogovém okně vložíme zvolenou proměnnou do *Test Variable List* a ponecháme zaškrtnutý způsob výpočtu *All categories equal*



Výsledky:

Chi-Square Test

Q40_1 Věrnost v manželství

	Observed N	Expected N	Residual
1 velmi důležité	1418	646,7	771,3
2 spíše důležité	498	646,7	-148,7
3 nepříliš důležité	24	646,7	-622,7
Total	1940		

Test Statistics

	Q40_1 Věrnost v manželství
Chi-Square[a]	1553,769
df	2
Asymp. Sig.	,000

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 646,7.

V první části tabulky výstupů vidíme, že jsme skutečně testovali hypotézu, že počet osob zastávajících názor, že věrnost je pro úspěšné manželství velmi důležitá, spíše důležitá a nepříliš důležitá, bude stejný (očekávané četnosti by měly být rovny 646,7. Proč? No vzhledem k tomu, že celkem bylo v souboru 1940 osob a má-li být tento počet rozdělen do tří stejně velkých skupin, musíme 1940

podělit 3, což se rovná 647,7). Významnost test chí-kvadrát vyšla blízká nule (0,000), takže nulovou hypotézu o tom, že počet osob bude ve třech zmíněných kategoriích postoje k důležitosti věrnosti pro manželství stejný, musíme zamítnout.

V dalším kroku bychom pak mohli testovat, zdali jsou rozdíly v počtech osob u jednotlivých variant statisticky významné. Tuto proceduru ovšem SPSS nemá zabudovanou přímo. Ale můžeme si pomoci dalším z Řehákových skriptů, tentokráte skriptem *Shoda četností.SBS*. Funguje tak, že jej nasadíme na tabulku *Frequencies*, kterou jsme opět označili kliknutím myši.

Příklad: Chceme zjistit, zdali ze od sebe statisticky významně odlišují některé z kategorií odpovědi na otázku, proč jsou u nás lidé, kteří žijí v nouzi. Tabulka rozložení četností vypadá následovně:

Q11_REC Proc lidé zijí v nouzi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 mají smùlu	285	14,9	16,4	16,4
	2 jsou líní	786	41,2	45,3	61,8
	3 je to bezpráví	341	17,9	19,7	81,5
	4 souèást pokroku	322	16,9	18,5	100,0
	Total	1734	90,9	100,0	
Missing	System	174	9,1		
Total		1908	100,0		

My chceme zjistit, zdali počet 285 respondentů (16,4 %), kteří si myslí, že lidé u nás žijí v nouzi proto, že mají smůlu lidé, je statisticky významně odlišný od počtu 322 osob (18.5 %), kteří si myslí, nouze je prostě součást pokroku. Na tuto tabulku nasadíme v outputu SPSS skript *Shoda četností.SBS*. Po spuštění skriptu (*Utilities – Run Script*) ještě musíme SPSS sdělit, které dvě kategorie má srovnávat, my chceme srovnat kategorii 1 s kategorií 4:

A výsledek?

Test shody četností dvou kategorií

:

Kategorie	Statistiky					
	Procenta 1 mají smůlu	Procenta 4 součást pokroku	Rozdíl procent	Z-skór	Signifikance	Znaménkové schéma
1 mají smůlu x 4 součást pokroku	16,4%	18,5%	-2,1%	-1,48	,138	o

Rozdíl mezi 16,4 % a 18,5 % (tedy 2,1 0) není statisticky signifikantní (vypočtená signifikance 0,138 je větší než 0,05), což znamená, že musíme očekávat, že v základním souboru bude podíl těchto dvou skupin shodný. Námi zjištěná dvouprocentní diference vznikla náhodou v důsledku výběrové chyby (všimněte si, že znaménkové schéma tiskne symbol **o**).