Petr Mareš
Ladislav Rabušic
Petr Soukup

Analýza sociálněvědních dat (nejen) v SPSS

muni
PRESS

v populaci rozdíl či souvislost existuje. V současnosti se v literatuře ještě hodně hovoří o pravděpodobnosti $\mathbf{1 - \beta}$, tzv. **síle testu**. Technicky jde o pravděpodobnost, že správně zamítneme nulovou hypotézu, která neplatí.[190] Samozřejmě že by tato pravděpodobnost měla být co největší (doporučení je minimálně 0,8), ale její výpočet není zcela snadný, resp. je k němu třeba používat speciální procedury. Dodejme, že pokud používáme běžné statistické postupy představené v této učebnici a máme výběrové soubory v řádu minimálně stovek výběrových jednotek, je naše síla testu vždy dostačující.

Literarura

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Science* (2nd ed.). Hillsdale, NJ: Erlbaum.

Field, A. (2009). *Disvovering Statistics Using SPSS* (3rd ed.). London: Sage.

Hendl, J. (2004). *Přehled statistických metod zpracování dat*. Praha: Portál.

Wonnacot, T. H., & Wonnacot, R. J. (1993). *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing.

[190] Jde o analogii pravděpodobnosti odsouzení opravdového viníka.

Kapitola 8

Základy dvourozměrné (bivariační) analýzy kategoriálních proměnných

Až dosud jsme se převážně zabývali analýzami, které byly založeny na srovnávání průměrů a rozptylů (variancí), tedy úlohami, kdy závisle proměnná 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. vesnice, 2. město, 3. velkoměsto), případně že jde o proměnné dichotomické.

V dvourozměrné analýze zkoumáme vztahy mezi dvěma proměnnými. Znamená to, že se ptáme, do jaké míry jedna proměnná ovlivňuje druhou proměnnou. Například 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 výraz, že „jedna proměnná ovlivňuje druhou"? Mezi proměnnými existuje vztah, pokud rozložení (distribuce) hodnot jedné kategorizované proměnné je asociováno s rozložením hodnot druhé kategorizované proměnné.[191] Řečeno jinak: hodnoty jedné proměnné jsou rozloženy takovým způsobem, že jsou vzorovány v závislosti na rozložení hodnot druhé proměnné.

Procedura, která nám pomůže vztah (asociaci) mezi dvěma proměnnými odhalit, se nazývá **třídění druhého stupně**: třídíme totiž rozložení variant znaku jedné proměnné podle rozložení variant znaku druhé proměnné. V jazyce SPSS je pojmenována jako *crosstabulation* neboli křížová tabulace – česky ovšem raději hovoříme o vytváření a analýze kontingenčních tabulek.[192]

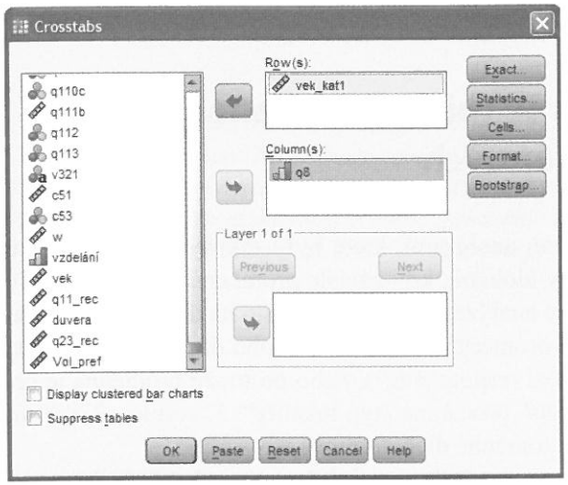
[191] Vztahy mezi proměnnými nehledáme samozřejmě pouze u kategorizovaných proměnných, ale také u proměnných spojitých, kardinálních. U dvou kardinálních proměnných ovšem sledujeme, zdali kovariují, tedy zdali se odchylky od průměru v jedné proměnné podobají odchylkám od průměru v druhé proměnné. O tom ale až v následující kapitole.

[192] Jak uvidíme dále, kontingenční tabulky má smysl vytvářet pouze pro kategorizované proměnné s relativně nevelkým počtem kategorií.

Příklad 8.1

Na základě údajů v datovém souboru „EVS99-cvicny" zjistěme, jak se liší názor na to, zdali je možné lidem důvěřovat (proměnná *q8*) v závislosti na věkových kategoriích (nezávisle proměnná *vek_kat1*). Platí náš předpoklad, že s rostoucím věkem narůstá nedůvěra vůči lidem?

Řešení: Procedura *Analyze – Descriptive Statistics – Crosstabs – Rows* (vek_kat1) – *Columns* (q8) – viz obr. 8.1.



Obr. 8.1 Zadání pro kontingenční tabulku (*Crosstabs*)

Na základě tohoto zadání, kdy jsme do řádků umístili nezávisle proměnnou a do sloupců proměnnou závislou (*q8*), získáme výstup 8.1.[193]

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

Výstup 8.1 Kontingenční tabulka pro proměnné *věk* a *názor na důvěru*

Z výstupu 8.1 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.

[193] Kontingenční tabulky si v SPSS obvykle organizujeme následujícím způsobem: nezávisle proměnnou umisťujeme do řádků tabulky, závisle proměnnou do sloupců.

I když by se na první pohled zdálo, že starší respondenti tento názor zastávali častěji než respondenti mladší (199 : 87), 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, což znemožňuje přímé 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 namísto absolutních četností (počtů) budeme srovnávat relativní četnosti, procenta.

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

Avšak předtím 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 celkovým počtem případů příslušného řádku. Ten nalezneme ve sloupci *Total*. Tak např. řádkové procento pro 87 respondentů ze skupiny 50+ (50 a více let), kteří si myslí, že lidem je možné důvěřovat, je:

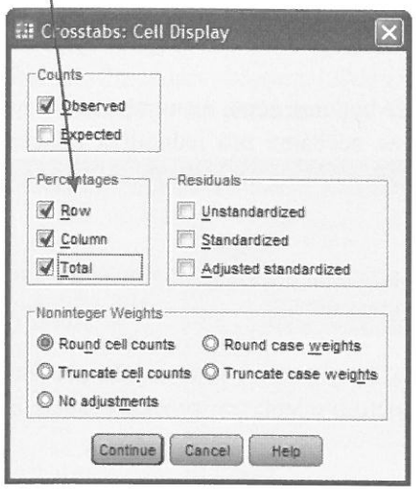
$$(199 / 784) \times 100 = 25{,}4\ \%.$$

Tento údaj čteme následovně: z respondentů ve věku 50+ let si 25 % myslí, že lidem se dá důvěřovat. Naopak 75 % (585/784 nebo v tomto případě i 100–25) z této věkové skupiny 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 celkovým počtem případů ve sloupcové kategorii, který 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+ let.

Celková procenta (*Total %*) pak získáme tak, že absolutní četnost v políčku dělíme celkovým počtem případů v souboru (resp. jen těch, u nichž máme platné odpovědi na obě analyzované otázky). Ten je uveden v křížovém součtu celkových počtů četností sloupců a řádků. Našich 199 respondentů ve věku 50+ let, kteří si myslí, že lidem lze důvěřovat, tedy tvoří: (199 / 1 864) × 100 = 10,7 %. Čteme: ze všech respondentů našeho souboru bylo 11 % těch, kteří měli 50+ let a kteří byli současně přesvědčeni, že lidem lze důvěřovat.

Všechny tři druhy procent za nás vypočítá SPSS. Kliknutím v dialogovém okně na tlačítko *Cells* se rozbalí tato nabídka, v níž ve čtverci *Percentages* zvolíme *Row*, *Column* a *Total* (viz obr. 8.2).



Obr. 8.2 Zadání pro výpočet řádkových, sloupcových a celkových procent

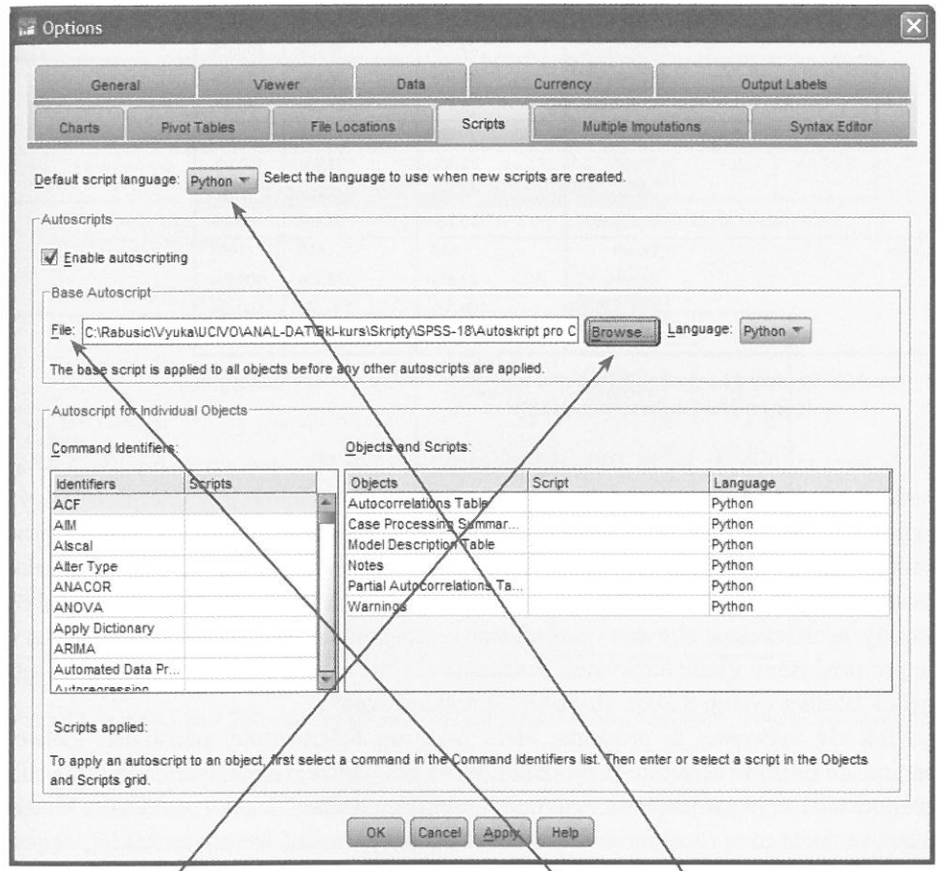
Po spuštění příkazu – to je nejdříve se kliknutím na *Continue* vrátíme do dialogu *Crosstabs*, v němž klikneme na *OK* – získáme výstup 8.2a.

vek_kat1 tri vekove skupiny * q8 Důvěra v lidi Crosstabulation

			q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
vek_kat1 tri vekove skupiny	1 18-29	Count	87	326	413
		% within vek_kat1 tri vekove skupiny	21,1%	78,9%	100,0%
		% within q8 Důvěra v lidi	19,6%	23,0%	22,2%
		% of Total	4,7%	17,5%	22,2%
	2 30-49	Count	158	508	666
		% within vek_kat1 tri vekove skupiny	23,7%	76,3%	100,0%
		% within q8 Důvěra v lidi	35,6%	35,8%	35,7%
		% of Total	8,5%	27,3%	35,7%
	3 50+	Count	199	585	784
		% within vek_kat1 tri vekove skupiny	25,4%	74,6%	100,0%
		% within q8 Důvěra v lidi	44,8%	41,2%	42,1%
		% of Total	10,7%	31,4%	42,1%
Total		Count	444	1419	1863
		% within vek_kat1 tri vekove skupiny	23,8%	76,2%	100,0%
		% within q8 Důvěra v lidi	100,0%	100,0%	100,0%
		% of Total	23,8%	76,2%	100,0%

Výstup 8.2a Kontingenční tabulka souvislosti důvěry a věku s řádkovými, sloupcovými a celkovými procenty

Výstup 8.2a je poněkud nepřehledný v tom, že řádková a sloupcová procenta pojmenovává výrazem *% within...* (a následuje jméno řádkové a sloupcové proměnné). Abychom si výstup zpřehlednili, navrhujeme nastavit si další skript, který popisek v tabulkách pojmenuje jednodušším způsobem. Příslušný autoskript nastavíme následovně. V základní obrazovce SPSS na horní liště najdeme tlačítko *Edit* a rozklikneme jeho roletku. Úplně dole pak klikneme na tlačítko *Options*. V rozbaleném dialogovém okně najdeme tlačítko *Scripts* a kliknutím jej otevřeme. Ukáže se tato obrazovka (viz obr. 8.3):



Obr. 8.3 Způsob nastavení skriptu pro zjednodušení popisku tabulky *Crosstabs*

V ní si nejdříve nastavíme jako default jazyk skriptu na *Python*. Poté klinutím na *Browse* si najdeme v našem adresáři skript s názvem „Autoskript pro Crosstabs.py" (soubor je na přiloženém CD) a vložíme jej do políčka *File*. Kliknutím na *OK* vše potvrdíme.

Když si nyní necháme znovu udělat kontingenční tabulku pro asociaci mezi kategorizovaným věkem a názorem na důvěru k lidem, získáme výstup 8.2b.

vek_kat1 tri vekove skupiny * q8 Důvěra v lidi Crosstabulation

			q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
vek_kat1 tri vekove skupiny	1 18-29	Count	87	326	413
		Row %	21,1%	78,9%	100,0%
		Column %	19,6%	23,0%	22,2%
		% of 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%
		% of 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%
		% of Total	10,7%	31,4%	42,1%
Total		Count	444	1419	1863
		Row %	23,8%	76,2%	100,0%
		Column %	100,0%	100,0%	100,0%
		% of Total	23,8%	76,2%	100,0%

Výstup 8.2b Upravená kontingenční tabulka souvislosti důvěry a věku s řádkovými, sloupcovými a celkovými procenty

V něm vidíme, že počet 199 respondentů (v řádku 50+) jednou znamená 25,4 %, podruhé 44,8 % a potřetí 10,7 %. Každý podíl má samozřejmě jiný interpretační význam a my si musíme v analýzách tohoto druhu dát dobrý pozor na to, jaká procenta vlastně chceme interpretovat. Jelikož ve vědě jako v každé jiné činnosti také platí princip efektivity, tedy snaha dosahovat maximálních výsledků s minimálními vstupy, necháváme si obvykle v našich analýzách spočítat jen ten druh procenta, který je pro příslušnou úlohu adekvátní. Podstatně si tím i zjednodušíme analytický život, neboť tabulka výstup 8.2b je zbytečně „mnohomluvná".

Jak ale 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. Ostatně na tom je založena i naše hypotéza, že s narůstajícím věkem bude slábnout důvěra v ostatní lidi.

Jestliže víme, která proměnná je nezávislá, podíváme se, kam jsme ji v kontingenční tabulce umístili. Pokud je v **řádcích** tabulky, počítáme **řádková** procenta. Tím dosáhneme toho, že všechny počty v kategoriích nezávisle proměnné vyrovnáme (položíme je za základ, tj. sto procent), což umožní smysluplné srovnání. Pokud je nezávisle proměnná

ve **sloupci**, počítáme **sloupcová** procenta. A co je důležité, o umístění proměnných do řádků či sloupců rozhodujeme při práci v SPSS sami při zadávání příkazu.[194]

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

Podívejme se tedy, jak by měla vypadat tabulka, s jejíž pomocí odpovíme na naši otázku (viz výstup 8.2c).

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
vek_kat1 tri vekove skupiny * q8 Důvěra v lidi	1863,997[a]	97,7%	44,003	2,3%	1908	100,0%

a. Number of valid cases is different from the total count in the crosstabulation table because the cell counts have been rounded.

vek_kat1 tri vekove skupiny * q8 Důvěra v lidi Crosstabulation

			q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
vek_kat1 tri vekove skupiny	1 18-29	Count	87	326	413
		Row %	**21,1%**	78,9%	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	1419	1863
		Row %	23,8%	76,2%	100,0%

Výstup 8.2c

Pozn. Rámeček 8.1 na s. 250 ukazuje, jak má vypadat formát tabulky, když naše výsledky publikujeme.

Z první části výstupu vidíme, že z celkového počtu respondentů na tuto otázku neodpovědělo 44 dotázaných neboli 2,3 %. Pozor, do kontingenční tabulky jsou vždycky zahrnuti pouze ti respondenti, kteří mají platné údaje u obou proměnných – celkem jich bylo 1 864 (to je 97,7 %).[195]

[194] Dodejme, že pokud tabulku použijeme například v diplomové práci či článku, musíme buď do názvu či pod ni do poznámky uvést, jaký typ procent obsahuje, abychom usnadnili její čtení (viz dále).

[195] Nejste překvapeni, že se v políčku u platného N (Valid N) objevuje údaj s desetinnými místy (1863,997)? Je to důsledek vážení souboru.

Výsledek třídění je poněkud překvapující. S narůstajícím věkem sice poněkud narůstá podíl osob, které 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 %. Rozdíly mezi procenty v políčkách se nazývají epsilon (a značí se řeckým písmem ε). Například hodnota epsilon pro respondenty ve věku (50+) a (18–29) je 25,4 − 21,0 = 4,4 %. Jelikož v analýze dat platí hrubé pravidlo, že teprve rozdíl (epsilon), který se blíží 10 %, indikuje i věcně podstatný rozdíl (to je takový, který nevznikl v důsledku výběrové chyby), vyslovujeme závěr, že v otázce důvěry k lidem se čeští respondenti nelišili v závislosti na věku. Zamítáme tak naši výzkumnou hypotézu, že s narůstajícím věkem bude také narůstat nedůvěra v ostatní lidi.

Při publikaci výsledků ovšem tabulku v takové podobě, jako je ve výstupu 8.2c, nikdy nezveřejňujeme, není totiž pro čtenáře přehledná. Musíme ji proto upravit podle následujících zásad:

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 umisťujeme do sloupců, takže počítáme sloupcová procenta. Tento požadavek ale není striktní, umístění proměnných také závisí na tom, jak dlouhé názvy mají jednotlivé kategorie.
5. U nezávisle proměnné uvádíme i procenta „celkem" (obvykle tedy 100 %) a současně i absolutní počty případů.
6. V poznámce pod tabulkou se uvádí zdroj dat a velikost souboru.

Tabulka z výstupu 8.2c by tedy podle těchto zásad měla být pro případnou publikaci upravena takto:

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 % (413)	100 % (666)	100 % (784)

Zdroj: EVS ČR 1999, N = 1864.

Tab. 8.1 Důvěra k lidem podle věku (sloupcová %)

Rámeček 8.1 Náležitosti tabulek

Tento příklad je dobrou ukázkou toho, že i „nula" ve vědě je důležitým poznatkem. My jsme zjistili, na rozdíl od našeho předpokladu, že mezi věkovými skupinami není v zásadě rozdíl v postoji „důvěra v ostatní lidi". Tato zjištěná „nula" v sobě ovšem obsahuje podstatný fakt, na jehož základě nyní víme, že v roce 1999 nebyly starší osoby vůči ostatním lidem méně důvěřivé než ty mladší.

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

V našem příkladu jsme hledali vztah mezi kategorizovaným věkem a postojem k jiným lidem z hlediska důvěry. Tuto úlohu jsme mohli řešit i jinak. Jelikož náš datový soubor obsahuje také údaje o věku v jeho nekategorizované podobě (je to 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, lze použít t-test. Výsledek je na výstupu 8.3.

Group Statistics

	Q8 Důvěra vlidi	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

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

Výstup 8.3 T-test pro průměrný věk u kategorií „důvěry v lidi"

Rozdíl v průměrném věku není příliš velký, věkový průměr je v obou kategoriích podobný (46,95 : 45,41). Proto také test nulové hypotézy, že rozdíl se v základním souboru (populaci) mezi těmito kategoriemi nebude odlišovat, vychází statisticky nevýznamný, takže nulovou hypotézu nelze zamítnout. Jinou technikou jsme zde tak dospěli ke stejnému výsledku. Máme tudíž jistotu, 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 (ani věcně ani statisticky) významná.

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

8.1 Test nezávislosti chí-kvadrát (χ^2)

Příklad 8.2

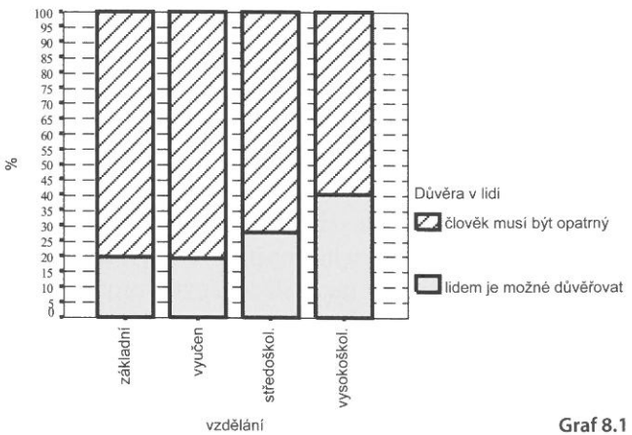
Hledejme v datech „EVS99-cvicny" odpověď na otázku, zdali je důvěra v lidi ovlivněna vzděláním respondentů. Naše výzkumná hypotéza bude znít, že se zvyšujícím se vzděláním bude narůstat podíl těch, kteří si myslí, že lidem je možné důvěřovat a že tento vztah je statisticky významný.

Řešení: Řádková procenta ve výstupu 8.4, jakož i graf na výstupu 8.4 ukazují, že mezi jednotlivými vzdělanostními kategoriemi existují rozdíly v názorech na důvěru k jiným lidem, přičemž lidé se středoškolským a vysokoškolským vzděláním mají tendenci lidem více důvěřovat 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 osob z různých vzdělanostních skupin stejný.

vzdelání kategorizace q94 * q8 Důvěra v lidi Crosstabulation

			q8 Důvěra v lidi		
			1 lidem je možné důvěřovat	2 člověk musí být opatrný	Total
vzdelání 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%

Výstup 8.4 Důvěra v lidi podle vzdělání



Graf 8.1 Důvěra v lidi podle vzdělání (v %)

Test provedeme na základě výpočtu statistiky **chí-kvadrát** χ^2 (*chi-square*). Chí-kvadrát je založen na srovnání napozorovaných (empirických) a očekávaných četností. Vychází z jednoduché myšlenky, že existuje model rozložení dat, které by vzniklo tak, že mezi sledovanými hodnotami dvou proměnných není žádná asociace – vzniklo by tedy působením náhody (což jsou tzv. očekávané četnosti). Srovnáme-li tento model se skutečným, empirickým rozložením reálných dat, zjistíme, zdali model náhodného rozložení odpovídá empirickým datům, nebo ne. Takže:

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

Podívejme se nyní opět do výstupu 8.4. 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 (okrajovou) četnost příslušného sloupce a marginální (okrajovou) 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.[196] Očekávaná četnost je vyšší než ta, kterou jsme my zjistili empiricky (71 případů). Zjištěný rozdíl 15,6 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 ještě musíme vypočítat rozdíl mezi empirickou a očekávanou četností, ten umocnit na druhou, podělit hodnotou očekávané četnosti a jednotlivé výsledky sečíst. Tím získáme hodnotu testové statistiky χ^2 chí-kvadrát. Tu pak – jako při každém testování nulové hypotézy – porovnáme s matematickým modelem rozložení, v tomto případě s modelem chí-kvadrát,[197] čímž zjistíme statistickou významnost.

Všechny výše naznačené operace za nás samozřejmě provede SPSS, ale pokud chceme, můžeme výpočet kontrolovat. V *Crosstabs* 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 pak tento výstup (viz výstup 8.5).

[196] Pro zvídavé dodejme, že tento postup vychází z definice nezávislosti, protože při výpočtu zohledňujeme odděleně rozdělení obou proměnných (okraje tabulky), a nikoliv jejich kombinace (tzv. sdružené četnosti uvnitř tabulky).

[197] Abychom byli zcela přesní, musíme vzít ještě v úvahu jeden prvek, a tím je počet stupňů volnosti, které odpovídají součinu $(k - 1) \times (l - 1)$, kde k a l jsou počty řádkových, resp. sloupcových kategorií.

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%

Výstup 8.5 Očekávané četnosti a rezidua v proceduře *Crosstabs*

Řádek *Residual* (česky rezidua) udává numerický rozdíl mezi napozorovanou (*Count*) a očekávanou (*Expected Count*) četností. Má-li znaménko +, znamená to, že napozorovaná četnost je vyšší, než bychom očekávali, kdyby platila nulová hypotéza nezávislosti. Záporné znaménko vyjadřuje pravý opak, tedy že napozorovaná četnost je nižší, než jaká by měla být, kdyby platila nulová hypotéza. Zkontrolujme, že rozdíl mezi empirickými a očekávanými četnostmi pro 71 respondentů je skutečně 15,6, jak jsme zjistili ručním výpočtem výše. V rutinní analytické praxi informace tohoto druhu nepotřebujeme, a proto takto detailní tabulku de facto nevyžadujeme. **Zadání testu chí-kvadrát** pro naši úlohu je následující: *Analyze – Desriptive Statistics – Crosstabs* – v dialogovém okně klikneme na lištu *Statistics* a v objevivším se novém dialogovém okně zaškrtneme políčko *Chi-square*. Výsledkem je výstup 8.6.

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.

Výstup 8.6 Tabulka hodnot chí-kvadrátu

Hodnota testového kritéria Pearsonova chí-kvadrát testu je 46,479 a její vypočtená pravděpodobnost chyby prvního druhu (hovorově též hladina významnosti) je 0,000. Musíme proto zamítnout nulovou hypotézu o nezávislosti vztahu mezi vzděláním a názorem na důvěru v ostatní lidi. Naopak očekáváme, že i v základním souboru se lidé budou ve své důvěře v ostatní odlišovat v závislosti na tom, jakého vzdělání dosáhli.

Další údaje ve výstupu 8.6 nejsou v dané situaci zajímavé.[198] Pozornost bychom ale 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.

Vraťme se ještě jedním příkladem k testu chí-kvadrát **o nezávislosti**, kdy testujeme, zdali jedna proměnná závisí na druhé.[199] Můžeme např. testovat hypotézu, zdali v populaci existuje 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.

---

[198] *Continuity Correction* je Yatesovou korekcí (opravou) Pearsonova chí-kvadrátu pro tabulky 2×2, tedy tabulky, v nichž obě proměnné jsou dichotomické, takže mají každá jen dvě varianty; mnozí totiž tvrdí, že v tabulce 2×2 dochází při standardním výpočtu chí-kvadrátu k přeceně-ní jeho hodnot, proto musí být výpočet upraven; *Likelihood Ratio* je statistika velmi podobná chí-kvadrátu a pro velké výběry dosahuje velmi podobných hodnot (namísto odečítání četností se četnosti dělí); Fisherův exaktní test (*Fisher's Exact Test*) můžeme jako sociologové směle ignorovat, je totiž určen pro malé výběry. *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.

[199] Pozor ale, pořád mějme na paměti, že se jedná pouze o test, zdali v populaci budou dvě sledované proměnné statisticky nezávislé. I tento test, jako každý test statistické inference, předpokládá, že výběrový soubor, z něhož pocházejí naše data, má vzhledem k zobecňované populaci reprezentativní charakter.

Test chí-kvadrát ve výstupu 8.7[200] říká, že zamítáme hypotézu o nezávislosti těchto dvou proměnných, neboť statistická signifikance je menší než 0,05. Což tedy znamená, že volební preference jsou rodinným stavem respondenta nějakým způsobem ovlivněny – v níže uvedeném výstupu 8.8 například vidíme, že podporu komunistům vyjadřovali především vdovci a vdovy (23 %), naopak svobodní a svobodné by je téměř nevolili (4 %).[201]

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	75,033[a]	18	,000
Likelihood Ratio	71,519	18	,000
Linear-by-Linear Association	5,062	1	,024
N of Valid Cases	1586		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,23.

Výstup 8.7 Volební preference podle rodinného stavu respondenta a test chí-kvadrát

Pozn. Proměnná „volební preference" (*vol_pref*) vznikla transformací proměnné *q72* takto:

```
RECODE q72 (6=1) (1=2) (3=3) (13=4) (9=5) (2=6) (14=6) (4 thru 5=6)
(7 thru 8=6)  (10 thru 12=6) (96 thru 97=7) (ELSE=SYSMIS) INTO
Vol_pref.
VARIABLE LABELS Vol_pref 'Rekodovaná q72'.
VALUE LABELS Vol_pref 1 'KSČM' 2 'ČSSD' 3 'KDU' 4 'US' 5 'ODS'
6 'jiná' 7 'nevolil by'.
```

[200] 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". Syntax pro transformaci proměnné je: RECODE rod_stav (1=1) (2=2) (5=4) (3 thru 4=3) (ELSE=SYSMIS) INTO rod_stav1. VARIABLE LABELS rod_stav1 ‚rodinný stav (sloučeno rozvedený + odloučení)'. EXECUTE. VALUE LABELS rod_stav1 1 'ženatý/vdaná' 2 'vdovec/vdova' 3 'rozvedení/odloučení' 4 'svobodný/á'. EXECUTE.

[201] Aniž bychom chtěli předbíhat dosavadní znalosti, je zde třeba vzít v úvahu možný vliv třetí proměnné, která do tohoto vztahu může intervenovat a kterou bychom měli ověřit (zde to však neuděláme). Touto skrytou proměnnou je věk. Svobodní jsou věkově mladí lidé, vdovci a vdovy naopak spíše staršího věku.

Vol_pref Rekodovaná q72 * rod_stav1 1 'ženatý Crosstabulation

			rod_stav1 1 'ženatý				Total
			1 ženatý/vdaná	2 vdovec/vdova	3 rozvedení/odloučení	4 svobodný/á	
Vol_pref Rekodovaná q72	1 KSČM	Row %	62,7%	22,4%	8,1%	6,8%	100,0%
		Adjusted Residual	-,4	5,5	,4	-4,0	
	2 ČSSD	Row %	69,0%	8,8%	4,9%	17,3%	100,0%
		Adjusted Residual	1,9	-,8	-1,7	-,5	
	3 KDU	Row %	59,2%	15,3%	4,1%	21,4%	100,0%
		Adjusted Residual	-1,1	1,8	-1,3	,8	
	4 US	Row %	59,9%	5,6%	6,8%	27,8%	100,0%
		Adjusted Residual	-1,2	-2,0	-,3	3,3	
	5 ODS	Row %	67,5%	7,7%	7,2%	17,6%	100,0%
		Adjusted Residual	1,5	-1,7	-,2	-,4	
	6 jiná	Row %	64,2%	6,1%	5,6%	24,0%	100,0%
		Adjusted Residual	,0	-1,9	-1,0	2,1	
	7 nevolil by	Row %	60,6%	10,7%	11,6%	17,1%	100,0%
		Adjusted Residual	-1,5	,4	3,3	-,6	
Total		Row %	64,2%	10,1%	7,4%	18,3%	100,0%

Výstup 8.8 Třídění volebních preferencí podle rodinného stavu respondenta

Pozn. Ve výstupu jsme z důvodů přehlednosti vynechali údaj o počtech případů (*count*) v jednotlivých políčkách.

Výsledky třídění ve výstupu 8.8 lze však dále specifikovat. Poslouží nám k tomu údaje tzv. **adjustovaných reziduí** (*Adjusted Residual*), které jsme si nechali do tabulky výstupu 8.8 vypočítat. Adjustované reziduum je založeno na rozdílu mezi napozorovanou a očekávanou četností (jak jsme si ukázali ve výstupu 8.5). Řečeno jazykem statistiky, je to rozdíl mezi frekvencí očekávanou (f_e) a frekvencí napozorovanou (f_o).[202] Tento rozdíl se jmenuje *delta*, značí se odpovídajícím řeckým písmenem delta (Δ) a adjustované reziduum standardizuje rezidua podělením jejich směrodatnou odchylkou.

Díky tomu lze adjustovaná rezidua testovat z hlediska statistické významnosti,[203] přičemž platí, že pokud je jeho hodnota vyšší v absolutní hodnotě než 2,00 (tj. menší než −2, nebo větší než 2), můžeme si být s 95% pravděpodobností jisti, že v daném políčku je rozdíl mezi napozorovanou (empirickou) a očekávanou četností statisticky významný a že tedy nevznikl jako důsledek výběrové chyby. Interpretačně má tato informace obrovský význam, neboť nám umožňuje detailní vhled do vztahu mezi proměnnými (do struktury závislosti). Například vidíme, že v řádku těch, kdo preferují KSČM, máme dvě statisticky významná adjustovaná rezidua (pro lepší orientaci jsou zvýrazněna). U vdovců/vdov je hodnota rezidua +5,5. To znamená, že vdovci a vdovy by volili komunisty významně častěji, než by odpovídalo předpokladu nezávislosti. Naopak

[202] Indexy *e* a *o* vycházejí z anglických ekvivalentů, tj. *expected* (očekávané) a *observed* (napozorované).

[203] Rezidua nemají žádné běžně známé teoretické rozdělení, adjustovaná pak mají přibližně standardizované normální rozdělení.

svobodní respondenti by komunisty volili mnohem méně často (*Adj. res.* = −4,0), než by odpovídalo hypotéze nezávislosti. Statisticky významně častěji by svobodní volili Unii svobody (US).

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 skript SPSS nazvaný *Znaménkové schéma.*[204] Tento programek udělá to, že namísto adjustovaných reziduí vloží do příslušných políček znaménkové schéma, které nám ukáže, kde lze očekávat statisticky významné souvislosti. Podmínkou je, že musíme nechat SPSS vypočítat **tabulku adjustovaných reziduí**. Získáme ji takto: výpočet volíme v proceduře *Crosstabs*, kliknutím na políčko *Cells* získáme obr. 8.2 (viz s. 246) a v něm zaškrtneme pouze *Adjusted Standardized*. Když na takto vzniklou tabulku pustíme skript (*Utilities – Run script – Znaménkové_schéma.py*), získáme výstup 8.9.

Vol_pref Rekodovaná q72 * rod_stav1 1 'ženatý Crosstabulation

Sign Scheme

		rod_stav1 1 'ženatý			
		1 ženatý/vdaná	2 vdovec/vdova	3 rozvedení/odloučení	4 svobodný/á
Vol_pref Rekodovaná q72	1 KSČM	o	+++	o	---
	2 ČSSD	o	o	o	o
	3 KDU	o	o	o	o
	4 US	o	-	o	++
	5 ODS	o	o	o	o
	6 jiná	o	o	o	+
	7 nevolil by	o	o	+++	o

Výstup 8.9 Znaménkové schéma pro vztah mezi rodinným stavem respondenta a volebními preferencemi

Znaménkové schéma ukazuje, v kterých polích jsou statisticky významné rozdíly mezi napozorovanými a očekávanými četnostmi. Počet symbolů indikuje totéž jako v případě znamének u t-testu (viz kapitolu 7); symbol + pak znamená, že napozorované četnosti jsou vyšší než očekávané, symbol − situaci opačnou, že napozorované četnosti jsou nižší než očekávané. A samozřejmě záleží na počtu znamének, neboť ten indikuje pravděpodobnost chyby:

\+ alfa = 0,05 (napozorované četnosti vyšší než očekávané a signifikantní s 95% pravděpodobností – riziko chyby max. 5 %)[205]

\+ + alfa = 0,01 (riziko chyby max. 1 %, signifikace 99 %)

[204] Jde o postup původně navržený Linhartem a Šafářem (1967, s. 437), detailně rozpracovaný manželi Řehákovými (1978). Oba články stojí dodnes za přečtení a lze je nalézt online.

[205] Rizikem zde máme na mysli pravděpodobnost chyby prvního druhu (viz kapitolu 7). Plně si uvědomujeme, že ve statistice se riziko (relativní) užívá v jiném, přesně stanoveném, významu.

\+ + + alfa = 0,001 (riziko chyby max. 0,1 %, signifikace 99,9 %)

\− alfa = 0,05 (napozorované četnosti nižší než očekávané, riziko chyby do 5 %)

\− − alfa = 0,01

\− − − alfa = 0,001

A jak výsledek interpretujeme? Svobodní respondenti statisticky významně (tj. nejspíše i v celé populaci) častěji než náhodně (na hladině významnosti $\alpha = 0{,}01$) odpověděli, že by ve volbách (výzkum se konal v roce 1999) volili stranu Unie svobody a statisticky méně často by volili komunisty. Komunisty by naopak statisticky významně často volili respondenti, jejichž rodinný stav byl vdovec/vdova. Rozvedení respondenti by statisticky významně častěji nešli volit. V ostatních polích tabulky nejsou mezi náhodným rozložením a empirickým rozložením žádné zobecnitelné rozdíly (nejsou statisticky významné), nelze je tedy v populaci očekávat.

Na tomto místě je třeba znovu vyslovit upozornění, že ve velkých souborech vykazují i malé rozdíly mezi empirickými a očekávanými četnostmi vysokou statistickou signifikaci. Proto je vždy nutné doplnit výsledky testů statistické signifikace reziduí rozborem procentuálního rozložení v kontingenční tabulce a uplatnit pravidlo hrubé „tesařské tužky" o desetiprocentním rozdílu, o němž jsme se zmínili na počátku této kapitoly.

Shrňme nyní **postup pro analýzu kontingenčních tabulek** do základních kroků:

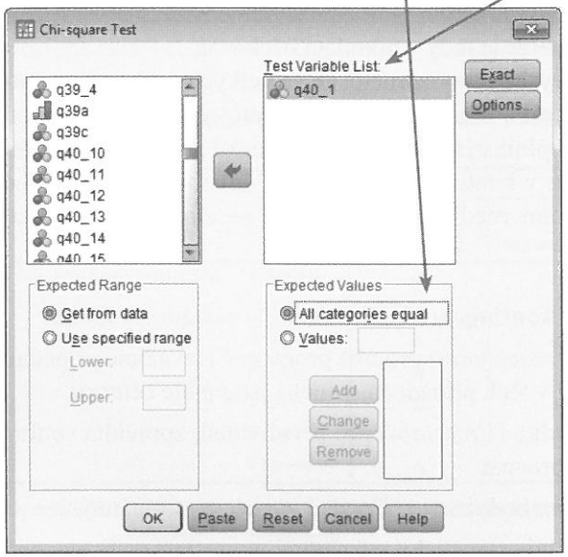
1. Vytvoříme četnostní tabulky (*Frequencies*) pro obě proměnné a zvážíme případné sloučení či vynechání kategorií v těch případech, v nichž jsou malé četnosti.
2. Zobrazíme si kontingenční tabulku (*Crosstabs*) pro první vhled, zpravidla ve formě sloupcových či řádkových procent.
3. Vypočítáme chí-kvadrát test a rozhodneme o případné závislosti proměnných.
4. V případě, že chí-kvadrát test bude statisticky významný, vypočteme adjustovaná rezidua a znaménkové schéma k posouzení struktury závislosti.
5. Interpretujeme nalezenou závislost (díky výstupům z kroků 2 a 4).
6. Popíšeme sílu souvislosti koeficientem asociace (viz následující kapitolu) a tím posoudíme věcnou významnost nalezené souvislosti.

8.1.1 Použití testu chí-kvadrát v jednorozměrné analýze

V této kapitole jsme si ukázali, jak odhalit a testovat asociaci mezi proměnnými prostřednictvím testu chí-kvadrát. Obdobného testu (s podobným názvem) 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 jednorozměrné, neboť nás bude zajímat, zdali se empirické rozložení kategorií jedné proměnné odlišuje od předpokládané distribuce této proměnné.

Jako nulovou hypotézu můžeme například 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 EVS-ČR1999, proměnná *q40_1*. Použijeme testu chí-kvadrát, který je ve skupině procedur pod názvem *Nonparametric tests*: *Analyze – Nonparametric tests – Legacy Dialogs – Chi-Square.*

V dialogovém okně vložíme zvolenou proměnnou do *Test Variable List* a ponecháme zaškrtnutý způsob výpočtu *All categories equal* (tj. je očekávána shoda podílu jednotlivých kategorií odpovědí) – viz obr. 8.4.



Obr. 8.4 Ukázka dialogového okna zadání pro chí-kvadrát pro jeden výběr

Výsledky:

q40_1 Věrnost v manželství

	Observed N	Expected N	Residual
1 velmi důležité	1427	649,3	777,7
2 spíše důležité	493	649,3	-156,3
3 nepříliš důležité	28	649,3	-621,3
Total	1948		

Test Statistics

	q40_1 Věrnost v manželství
Chi-Square[a]	1563,543
df	2
Asymp. Sig.	,000

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 649,3.

Výstup 8.10 *Chi-Square Test* pro jeden výběr

V první části tabulky výstupů 8.10 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 (*Expected N*) by měly být rovny 649,3. Proč? Protože v souboru bylo celkem 1 948 osob, a má-li být tento počet rozdělen do tří stejně velkých skupin, musíme 1 948 podělit 3, což se rovná 649,3. Statistická významnost testu chí-kvadrát vyšla, jak ukazuje druhá část výstupu 8.10, 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.

To ale není u tohoto druhu analýzy všechno, co lze vytěžit. Můžeme získat ještě detailnější informaci. Aplikací skriptu *test dobré shody* lze zjistit, zdali se empirické četnosti (*Observed N*) jednotlivých kategorií statisticky signifikantně odlišují od příslušných četností očekávaných (jde o obdobu znaménkového schématu u kontingenčních tabulek, v tomto případě ale pracujeme s tříděním prvního stupně). Jak to provedeme? Tabulku, kterou jsme získali výše (to je v proceduře *Analyze – Nonparametric tests – Legacy Dialogs – Chi-Square*), označíme v outputu SPSS kliknutím a pustíme na ni skript *test dobré shody.py*. Tabulka se promění následovně (viz výstup 8.11):

q40_1 Věrnost v manželství

	Observed N	Expected N	Sign Scheme
1 velmi důležité	1427	649,3	+++
2 spíše důležité	493	649,3	---
3 nepříliš důležité	28	649,3	---
Total	1948		

Výstup 8.11 Použití skriptu *test dobré shody.py*

Skript nahradil v původní tabulce rezidua znaménkovým schématem, které graficky zvýrazňuje buňky, jejichž četnost se statisticky významně liší od očekávané četnosti. Typ znaménka opět reprezentuje směr odchylky, jak jsme uvedli výše, znaménka opět ukazují statistickou významnost výsledku.

U testu chí-kvadrát dobré shody pro rozložení četností jedné proměnné je zapotřebí poznamenat, že v sociologii a příbuzných disciplínách lze tento test použít zejména k otestování reprezentativity našeho výběrového souboru. Jako očekávané rozdělení bereme údaje o populaci a zjišťujeme, zda náš výběr je od ní jen náhodně odlišný. Pokud tedy nezamítáme nulovou hypotézu, je náš výběr nejspíše reprezentativní a vice versa.

A ještě jeden typ úloh se u tohoto druhu testování objevuje, a to o shodě četností mezi zvolenými dvojicemi kategorií v tabulce frekvencí. To zjistíme prostřednictvím skriptu *test shody četností.py*. Funguje tak, že jej nasadíme na tabulku *Frequencies*.

Příklad 8.3

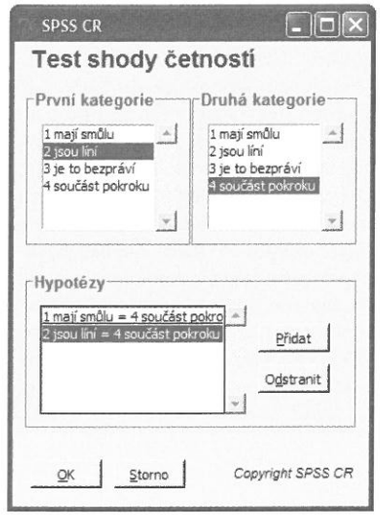
Chceme zjistit, zdali se od sebe statisticky významně odlišují počty respondentů, kteří na otázku, proč jsou u nás lidé, kteří žijí v nouzi, odpověděli, že to je proto, že mají smůlu, nebo proto, že jsou líní, od respondentů, kteří chudobu připisují společenskému pokroku. Opět použijeme data ze souboru „EVS99-cvicny". Tabulka rozložení četností této proměnné vypadá následovně (viz výstup 8.12):

q11_rec Proc lidé zijí v nouzi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 mají smůlu	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 součást pokroku	322	16,9	18,5	100,0
	Total	**1734**	**90,9**	**100,0**	
Missing	System	174	9,1		
Total		**1908**	**100,0**		

Výstup 8.12 Proč jsou u nás lidé, kteří žijí v nouzi

Chceme zjistit, zdali počet 285 respondentů, resp. podíl osob, které si myslí, že lidé u nás žijí v nouzi proto, že mají smůlu (16,4 %), je statisticky významně (v populaci) odlišný od počtu 322 osob (18,5 %), které si myslí, že nouze je prostě součástí pokroku. Podobně zdali podíl 786 respondentů (41,2 %), kteří si myslí, že chudí jsou chudými, protože jsou líní, se liší od podílu osob 322 respondentů, kteří se domnívají, že nouze je součást pokroku. Na tuto tabulku nasadíme v outputu SPSS skript *test shody četností.py*. Po spuštění skriptu (*Utilities – Run Script*) ještě musíme SPSS sdělit, které dvojice kategorií chceme srovnávat: v našem případě to je kategorie 1 s kategorií 4 a kategorie 2 s kategorií 4 – viz obr. 8.5.



Obr. 8.5 Zadání pro test shody četností

Testované dvojice se volí z okének první a druhé kategorie. V okně „první kategorie" zaklikneme příslušnou kategorii („jsou líní") a v okně druhá kategorie tu, s níž chceme srovnávat („součást pokroku"). Pak kliknutím na *Přidat* je umístíme do okna hypotézy. Totéž provedeme pro dvojici „mají smůlu" a „součást pokroku". Výsledek je na výstupu 8.13.

Výsledky testu shody četností

Testované kategorie	Rozdíl četností	Rozdíl procent	Podíl četností	Testová statistika Chí-kvadrát (1)	Dosažená hladina významnosti
1 mají smůlu : 4 součást pokroku	-37	-2,1%	,886	2,203	,138
2 jsou líní : 4 součást pokroku	465	26,8%	2,446	194,990	,000

Výstup 8.13 Výsledek testu shody četností

Co výsledek říká? Rozdíl v počtech nebo v procentech mezi 1. a 4. skupinou není statisticky signifikantní, neboť „dosažená" hladina významnosti (viz poslední sloupec tabulky) je 0,138, takže mohl vzniknout v důsledku výběrové chyby. Na druhé straně rozdíl mezi 2. a 4. skupinou signifikantní je. Můžeme tedy s jistotou očekávat, že vysoký rozdíl mezi těmi, kteří zastávají názor, že chudí jsou chudými proto, že jsou líní, a těmi, kdo si myslí, že chudí žijí v nouzi, neboť je to součást pokroku, bude existovat i v souboru základním (české dospělé populaci). Pro výraz „rozdíly procent" je ještě třeba si vysvětlit pojem „procentní body", což činíme v rámečku 8.2.

Procenta a procentní body

Kromě „normálních" procent používáme ještě tzv. procentní body. Jaký je mezi procenty a procentními body rozdíl? Představte si, že prohlížeč Firefox před dvěma lety používalo 20 % lidí, ale tento rok ho používá již 30 % lidí. Jaký je nárůst mezi těmito dvěma lety? Chybou by bylo říci, že nárůst je desetiprocentní. Pokud Firefox používalo nejdříve 20 % lidí a poté 30 %, tak to znamená, že ho nyní používá o polovinu více lidí než předtím, tj. o 50 % více (základem, tj. 100 % byla pětina populace). Jenomže je trochu neobratné říkat, že nárůst je 50 %, když by pro všechny bylo srozumitelnější sdělení, že se počet procent zvedl o deset. A přesně k tomu máme procentní body. V tomto případě můžeme říci, že nárůst využívání prohlížeče Firefox činí deset procentních bodů. Procentní body tedy použijeme v případě, když chceme vyjádřit rozdíl mezi dvěma procentuálními údaji (procento je tedy vztaženo k absolutnímu základu, procentní body k relativnímu základu, tj. k procentům). Není překvapením, že běžně se procentní body používají v ekonomii (rozdíl v úrokových mírách, míře inflace), ale čím dál více pronikají i do ostatních věd, proto si na ně zvykněme. Dodejme, že zatímco procenta značíme známou značkou „%", pro procentní body se buď užívá plného výrazu, nebo zkratky „p.b.".

Rámeček 8.2

Literatura

Linhart, Z., & Šafář, J. (1967). Programování třídění a statistických výpočtů pomocí samočinných počítačů. *Sociologický časopis*, *3*(4), 435–443.

Řehák, J., & Řeháková, B. (1978). Analýza kontingenčních tabulek: rozlišení dvou základních typů úloh a znaménkové schéma. *Sociologický časopis*, *14*(6), 619–631.

Kapitola 9

Měření vztahů mezi dvěma proměnnými (korelační analýza)

9.1 Asociace a korelace

V minulé kapitole jsme si ukázali, jak hledat asociaci mezi proměnnými v kontingenční tabulce na základě rozložení případů v jejich políčkách, a naučili jsme se také používat test chí-kvadrát pro potvrzení existence takové asociace v základním souboru. Řekli jsme si ale také, že chí-kvadrát nám sice dovolí zamítnout nulovou hypotézu o neexistenci asociace mezi sledovanými proměnnými, ale neposkytne nám žádnou informaci o síle této asociace. Vysvětlení, co asociace vlastně jsou a jaké jsou jejich charakteristiky, jsme v předchozí kapitole tak trochu odbyli, takže to nyní napravíme.

To, zdali je mezi dvěma proměnnými vztah (asociace, kontingence, korelace), je jedna ze základních otázek, kterou si při dvojrozměrné analýze dat klademe. Při hledání vztahu mezi dvěma proměnnými sledujeme, zda změny v jedné proměnné jsou doprovázeny i změnami v druhé proměnné.

Zajímá nás například, zdali existuje vztah mezi vzděláním a průměrným věkem v době prvního sňatku, zdali školní prospěch dětí souvisí s majetkovou úrovní jejich rodičů, zdali míra anomie souvisí s postojem k systému českého sociálního zabezpečení atd.

Hledání vztahu mezi proměnnými (hledání asociace) je z analytického hlediska operace, která má čtyři kroky. Při zkoumání párové asociace bychom si totiž, jak říkají Loether a McTavish (1988), měli klást čtyři následující otázky:

1. **Zdali asociace existuje, či nikoliv.**
2. **Jak je asociace silná (těsná)** – to je, jak silně rozložení variant jedné proměnné určují rozložení variant druhé proměnné.