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]

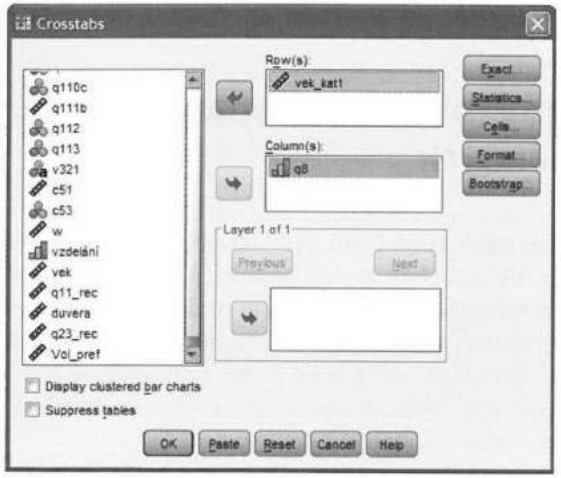
[191] Vztahy mezi proměnnými nehledáme samozřejmě pouze u kategorizovaných proměnných, ale také u proměnných spojitých, kardinálních. U dvou kardinálních proměnných ovšem sledujeme, zdali kovariují, tedy zdali se odchylky od průměru v jedné proměnné podobají odchylkám od průměru v druhé proměnné. O tom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

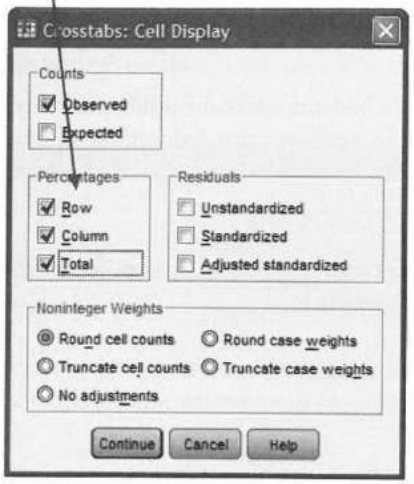
$$(199 / 784) \times 100 = 25{,}4\ \%.$$

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 % $(199 / 444) \times 100 = 44{,}8\ \%$). 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: $(199 / 1\,864) \times 100 = 10{,}7\ \%$. 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

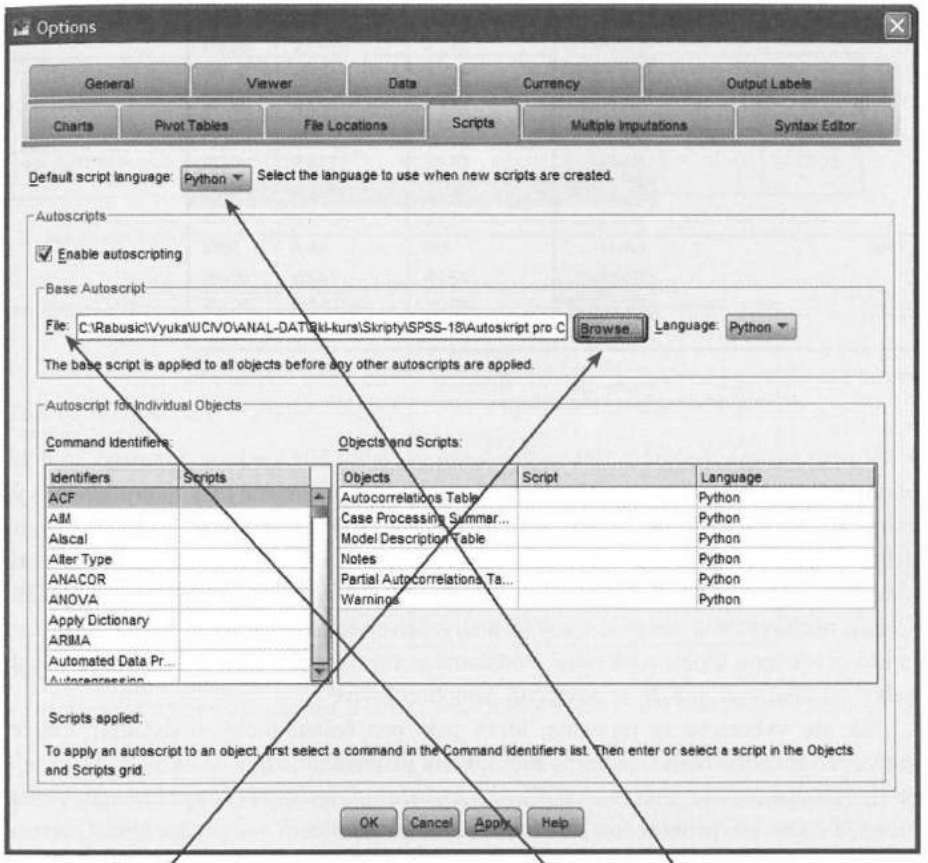
Po spuštění příkazu – to je nejdříve se kliknutím na *Continue* vrátíme do dialogu *Crosstabs*, v němž klikneme na *OK* – získáme výstup 8.2a.

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.