Návod na statistický software PSPP
část 2. – Kontingenční tabulky

Jiří Šafr
FHS UK

poslední revize 31. srpna 2010

Logika kontingenčních tabulek

Postup a logika **kontingenčních tabulek** je popsán v Prezentaci 2 (spss2_tabulky.ppt).

Platí základní princip:

Uspořádání tabulky

sloupcová procenta:

V kategoriích nezávislé proměnné ukazujeme kompletní (100 %) distribuci závislé proměnné.

	NEZÁVISLÁ - vysvětlující			
ZÁVISLÁ - vysvětlovaná		Pohlaví		
	Spokojenost	**Muž**	**Žena**	Celkový součet
	1 (nespokojen)	41 % (5)	22 % (2)	7
	2	41 % (5)	11 % (1)	6
	3 (spokojen)	16 % (2)	66 % (6)	8
	Celkový součet	100 % (12)	100 % (9)	**21**

Nejčastěji bývá **závislá** proměnná nalevo v řádcích a **nezávislá** (vysvětlující) ve sloupcích.

Chceme následující **tabulku ukazující závislost příjmu na vzdělání** (příklad z AD_100511.doc): příjem (kvartily) a vzdělání (4 kategorie).

prijem4 Příjem-osobní – kvartily (s24) * vzd4 Vzdělání Crosstabulation

% within vzd4 Vzdělání

		vzd4 Vzdělání				
		1 ZŠ	2 VYUČ	3 SŠ	4 VŠ	Total
prijem4 Příjem-osobní – kvartily (s24)	1 I. do 7 tis.	62,6%	26,4%	28,9%	19,6%	31,6%
	2 II. 7–9 tis.	26,2%	31,9%	15,7%	17,6%	24,1%
	3 III. 9–15 tis.	7,5%	29,7%	29,2%	21,6%	26,2%
	4 IV. nad 15 tis.	3,7%	11,9%	6,2%	41,2%	18,1%
Total		100,0%	100,0%	100,0%	100,0%	100,0%

*V kategoriích nezávislé proměnné (zde Vzdělání) ukazujeme kompletní (100 %) distribuci závislé proměnné (prijem4).

*Pozor! **Směr kauzality je vždy věcí teorie, nelze ji určit z dat samotných.**

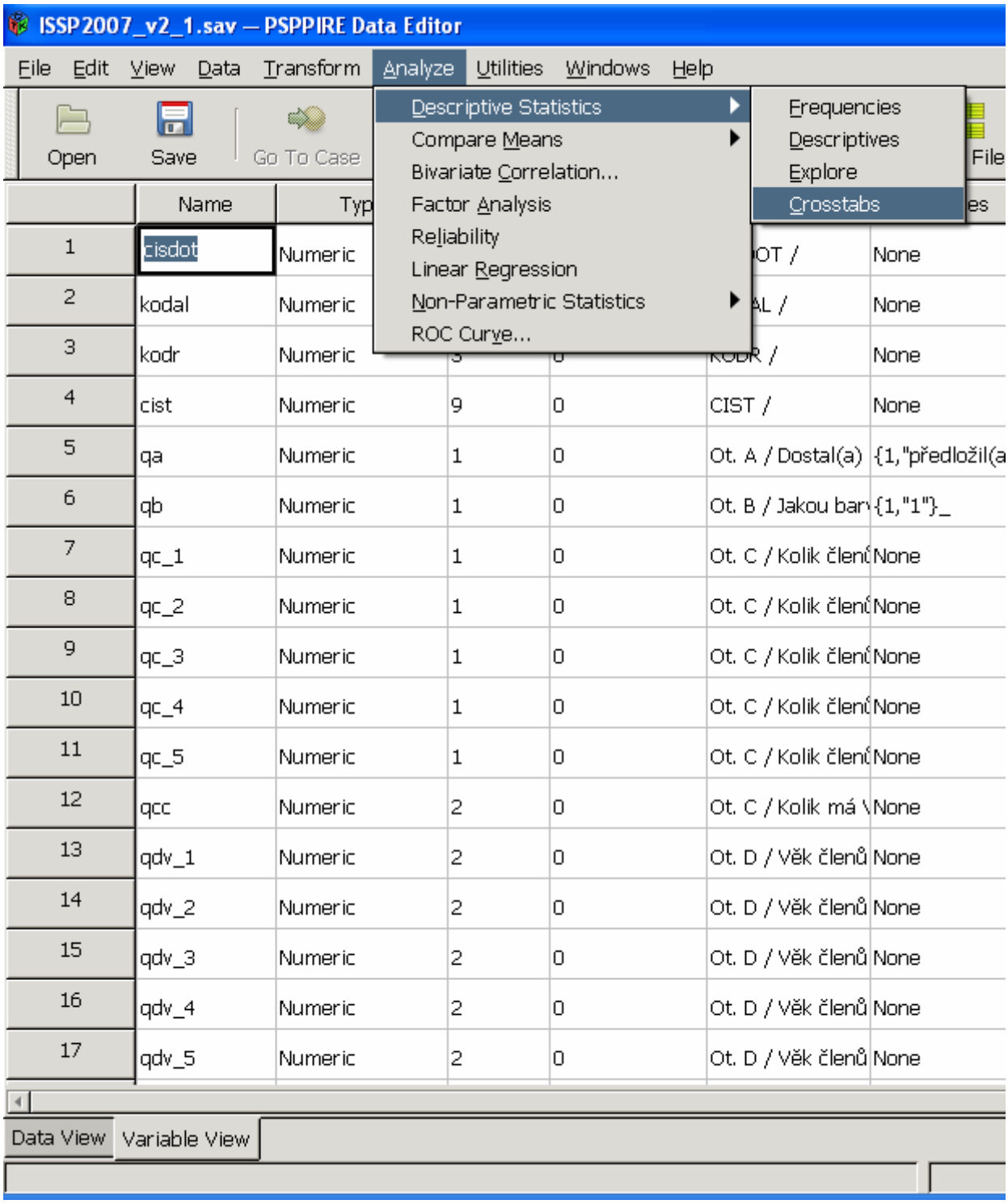
*Informace z tabulky interpretujeme tak, že porovnáme navzájem podskupiny nezávislé proměnné (zde stupně vzdělání) podle vlastností závislé proměnné (zde kvartily příjmu)

→ tabulku čteme „po řádcích“ (pokud máme nezávislý znak ve sloupcích, závislý v řádcích a sloupcová procenta jako zde, což je nejobvyklejší, ale lze to cele otočit o 90st).

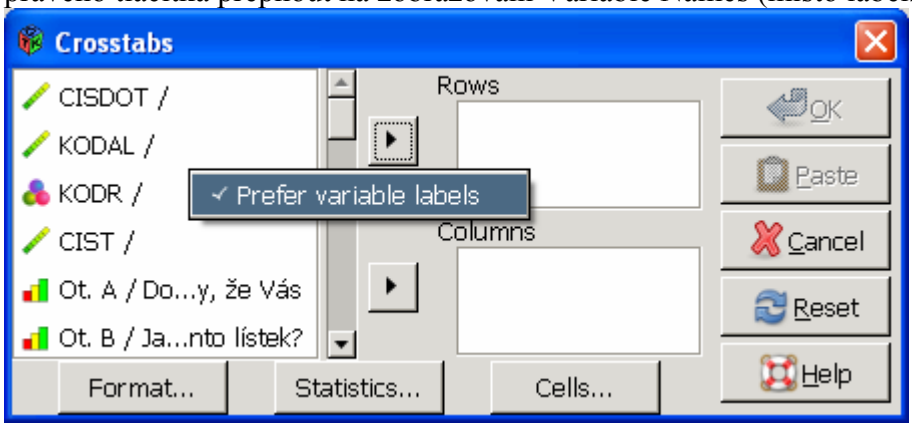
Postup vytváření kontingenčních tabulek v PSPP (SPSS)

Nejprve si otevřeme datový soubor, zde je to ISSP2007_v2_1.sav.

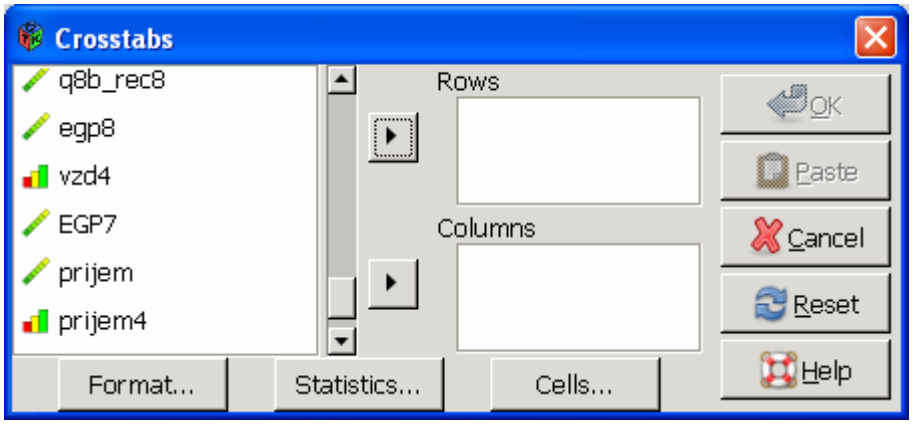
V PSPP přes menu zadáme:
Analyze → Descriptive Statistics → Crosstabs.

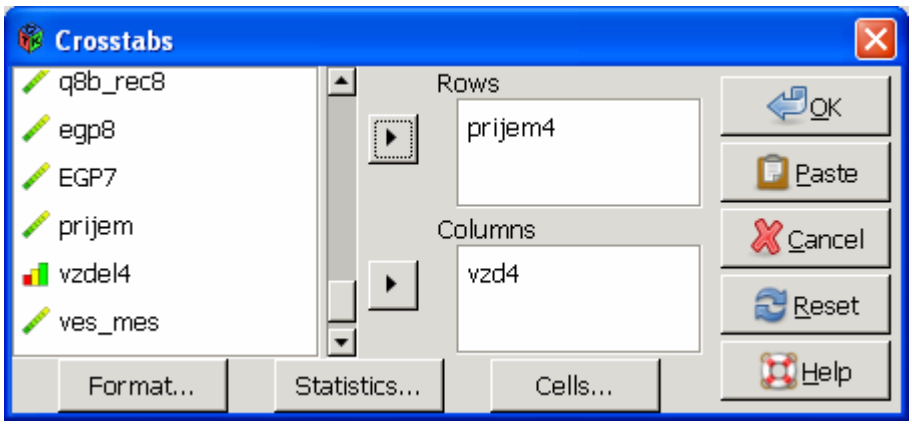


Pro jednoduší orientaci si můžeme pomocí myši – najedeme na seznam proměnných – a jejího pravého tlačítka přepnout na zobrazování Variable Names (místo labels).

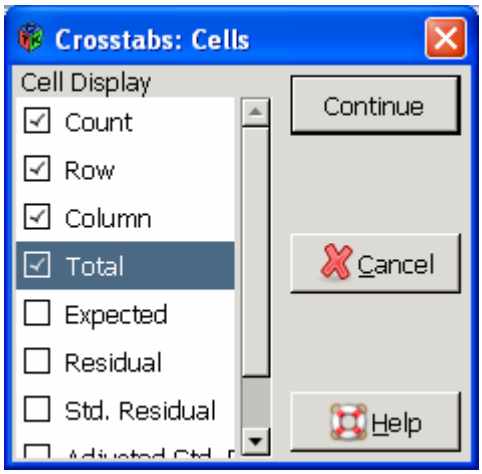


Najdeme si proměnné, které chceme do tabulky a do řádků = Rows dáme **závislou (zde prijem4**) a do sloupců = **Rows (vzdel4)**

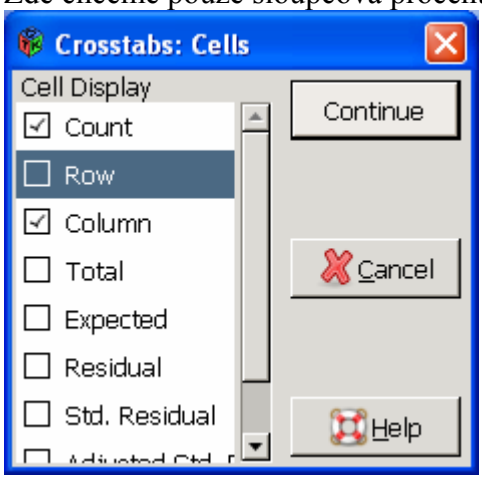




Dále musíme nastavit, co bude v políčkách tabulky v sekci **Cells**:
PSPP má přednastaveny všechny možnosti (četnosti, řádková, sloupcová i celková %), vhodné je vybrat pouze to, co budeme potřebovat, aby tabulka byla přehledná.

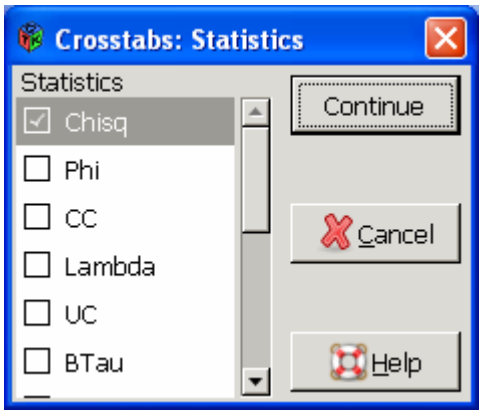


Zde chceme pouze sloupcová procenta (**Colum**) a ponecháme též absolutní četnosti (**Count**).



Statistics: Statistické testy a míry asociace

Pokud ponecháme nastavení – zakliknuto je **Chisq**, tak se nám ve výstupu rovněž objeví základní test asociací v tabulce tzv. Test dobré shody Chisquare (Chíkvadrát), který testuje nulovou hypotézu, že kategorie proměnných jsou na sobě nezávislé. Můžeme ho vypnout, ale doporučuji ponechat, je to základní test, interpretace je jednoduchá (navíc je to test neparametrický, takže je vhodný i na malé výběrové soubory). Hlídat ale musíme dostaečné četnosti v jednotlivých polích tabulky (viz dále).



Výstup (output) a jak ho čteme:

Nejprve dostaneme tabulku Summary informující nás o počtu platných případů (průnik za obě proměnné v kontingenční tabulce).

Summary.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Příjem-osobní – kvartily (s24) * Vzdělání	843	69.0%	379	31.0%	1222	100.0%

Zde vidíme, že z celkového počtu případů (respondentů ve výzkumu) 1222 odpovědělo na obě otázky pouze 843, takže máme 31% missingů – chybějících hodnot, což je hodně. Nepsané pravidlo pro velké výběrové soubory (cca nad 100 případů) říká, že je-li chybějících hodnot do 5%, pak to můžeme ignorovat, jinak bychom měli zjistit, „kdo jsou ti, kdo neodpovídají“. Zde je to způsobeno odpovědí na otázku po příjmu (to by nám řekla tabulka třídění prvního stupně pomocí FREQUENCIES – viz návod na PSPP část 1).

Hlavní tabulka, která nás zajímá a kterou později upravíme do zprávy/ diplomky je samotná **kontingenční tabulka**:

prijem4 * vzd4 [count, column %].

prijem4	*vzd4*				Total
	ZŠ	VYUČ	SŠ	VŠ	
I. do 7 tis.	67,0 62,6%	95,0 26,4%	94,0 28,9%	10,0 19,6%	266,0 31,6%
II. 7–9 tis.	28,0 26,2%	115,0 31,9%	51,0 15,7%	9,0 17,6%	203,0 24,1%
III. 9–15 tis.	8,0 7,5%	107,0 29,7%	95,0 29,2%	11,0 21,6%	221,0 26,2%
IV. nad 15 tis.	4,0 3,7%	43,0 11,9%	85,0 26,2%	21,0 41,2%	153,0 18,1%
Total	107,0 100,0%	360,0 100,0%	325,0 100,0%	51,0 100,0%	843,0 100,0%

Tabulka v sobě obsahuje jak relativní podíly (%) tak absolutní četnosti. Můžeme je s výhodou rozdělit při zadávání na dvě – provedeme analýzu 2x za sebou jednou s Count, podruhé s Column.

Samotné absolutní četnosti vypadají takto (zde pozor na nesmyslné desetinné místo)

prijem4 * vzd4 [count].

prijem4	*vzd4* ZŠ	VYUČ	SŠ	VŠ	Total
I. do 7 tis.	67,0	95,0	94,0	10,0	266,0
II. 7–9 tis.	28,0	115,0	51,0	9,0	203,0
III. 9–15 tis.	8,0	107,0	95,0	11,0	221,0
IV. nad 15 tis.	4,0	43,0	85,0	21,0	153,0
Total	107,0	360,0	325,0	51,0	843,0

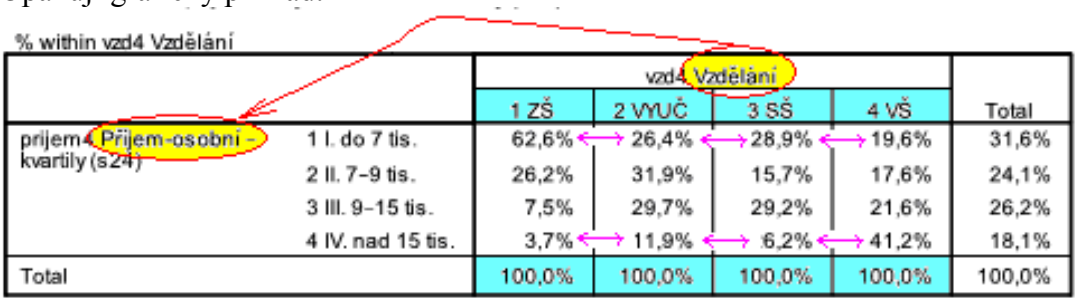
Tuto tabulku absolutních četností nemůžeme interpretovat, protože četnosti ve spoluvýskytu kategorií jsou závislé na tzv. marginálních četností, tj. rozložení kategorií obou znaků.
Ale nezapomeňte zde kontrolovat, že v každém políčku tabulky máte alespoň cca 5 případů (v absolutní četnosti). Pokud ne, tak můžete některé kategorie sloučit (minimum je 2 x 2 tabulka), například pomocí příkazu RECODE (o tom v 3 dílu návodu na PSPP).

Samotné sloupcové četnosti vypadají takto:

prijem4 * vzd4 [column %].

prijem4	*vzd4* ZŠ	VYUČ	SŠ	VŠ	Total
I. do 7 tis.	62,6%	26,4%	28,9%	19,6%	31,6%
II. 7–9 tis.	26,2%	31,9%	15,7%	17,6%	24,1%
III. 9–15 tis.	7,5%	29,7%	29,2%	21,6%	26,2%
IV. nad 15 tis.	3,7%	11,9%	26,2%	41,2%	18,1%
Total	100,0%	100,0%	100,0%	100,0%	100,0%

Interpretace provádíme takto: nejprve zkontrolujeme absolutní četnosti (v Count), zda máme dostatečné množství případů v políčkách tabulky. Pak čteme sloupcová procenta a to pořádcích porovnáváme, zda se lišší % zastoupení v příjmové kategorii podle vzdělání.
Opakuji grafický příklad:



Také můžeme sledovat kupení četností na diagonále, které indikuje lineární závislost.

		vzd4 Vzdělání				
		1 ZŠ	2 VYUČ	3 SŠ	4 VŠ	Total
prijem4 Příjem-osobní - kvartily (s24)	1 I. do 7 tis.	**62,6%**	26,4%	28,9%	19,6%	31,6%
	2 II. 7-9 tis.	26,2%	**31,9%**	15,7%	17,6%	24,1%
	3 III. 9-15 tis.	7,5%	29,7%	**29,2%**	21,6%	26,2%
	4 IV. nad 15 tis.	3,7%	11,9%	26,2%	**41,2%**	18,1%
Total		100,0%	100,0%	100,0%	100,0%	100,0%

Zde se zdá, že vyšší příjmové kategorie nacházíme u vzdělanějších respondentů, tedy že příjem roste se vzděláním. Ale podrobněji viz dále znaménkové schéma.

Tabulku lze pochopitelně otočit o 90st., aniž by se změnil její význam, ale při zadání musíme **prohodit proměnné** (vzd4 a prijem4) **z řádků na sloupce** a zadat, že chceme **řádková procenta** (**Row**). Pak čteme tabulku „po sloupcích“.

vzd4 * prijem4 [row %].

	prijem4				
vzd4	I. do 7 tis.	II. 7–9 tis.	III. 9–15 tis.	IV. nad 15 tis.	Total
ZŠ	62,6%	26,2%	7,5%	3,7%	100,0%
VYUČ	26,4%	31,9%	29,7%	11,9%	100,0%
SŠ	28,9%	15,7%	29,2%	26,2%	100,0%
VŠ	19,6%	17,6%	21,6%	41,2%	100,0%
Total	31,6%	24,1%	26,2%	18,1%	100,0%

V principu to tedy vypadá takto:

% within vzd4 Vzdělání

		prijem4 Příjem-osobní - kvartily (s24)				
		1 I. do 7 tis.	2 II. 7-9 tis.	3 III. 9-15 tis.	4 IV. nad 15 tis.	Total
vzd4 Vzdělání	1 ZŠ	62,6%	26,2%	7,5%	3,7%	100,0%
	2 VYUČ	26,4%	31,9%	29,7%	11,9%	100,0%
	3 SŠ	28,9%	15,7%	29,2%	26,2%	100,0%
	4 VŠ	19,6%	17,6%	21,6%	41,2%	100,0%
Total		31,6%	24,1%	26,2%	18,1%	100,0%

Udělat správně zorientovanou tabulku nemusí stačit. Vystačíme si tím například tam, kde máme data o úplné populaci, ale pokud chceme zobecnit výsledky našeho výzkumu, tedy chceme-li tvrdit, že rozdíly mezi kategoriemi naměřené v našem výběrovém vzorku platí i pro celou populaci, pak musíme provést ještě statistický test této závislosti.

Test dobré shody Chí kvadrát – (ne)závislost proměnných v kontingenční tabulce

Ale zpět do outputu. Ve výstupu máme ještě **tabulku s výsledkem statistického testu nezávislosti**.
Statistická tzv. nulová hypotéza zde říká, že příjem je na vzdělání nezávislý.
Hodnota testového kritéria Chíkvadrát je 122,42 při 9 stupních volnosti (df) a *p*-hodnota (Asympt.signif.) je 0,00.
Nulovou hypotézu zde nemůžeme přijmout neboť **dosažená hladina významnosti *p* je menší než 0,05** (což odpovídá zvolené chybě 5%). Znamená to, že i pro celou populaci ČR platí, že alespoň jedna kategorie příjmu je ovlivněna alespoň jednou kategorií vzdělání. Tento test nám ale neříká, které kategorie to jsou, ani kolik jich je (zda platí zde kvazi-lineární závislost, nebo zda je závislost jen mezi dvěma kategoriemi). Intuitivně na to můžeme usoudit z rozdílů mezi sloupcovými procenty, čteme-li tabulku po řádcích. Můžeme použít statistický test, který by nám dále ukázal kde je souvislost a jak je silná, když si do políček tabulky přidáme tzv. Adjustované residua (viz dále).

Zdůrazňuji, že **smyslem testování hypotéz je usuzovat na to, zda vztahy ve vašem výběrovém souboru** (vzorku, zde 1222 respondentů ve výzkumu ISSP) **s přijatelnou mírou chyby** (nejčastěji volíme 5 %) **platí i pro celou populaci**. Nic jiného! Neříká, jak moc je vztah mezi proměnnými těsný.
Splněno k tomu musí být mnoho podmínek, mj.: výběr vzorku musí být náhodný, dostatečně velký (minimálně 30-50 případů, ale záleží na velikosti populace), z dostatečně velké populace, nesmí však být příliš velký (protože pak jsou všechny souvislosti statisticky významné) a v polích tabulky musí být dostatečný počet případů (nesmí být prázdná).
K principu testování hypotéz viz více článek P. Soukupa Soukup, P. 2007. "Statisticky významný neznamená důležitý." na http://www.socioweb.cz/index.php?disp=teorie&shw=298&lst=108

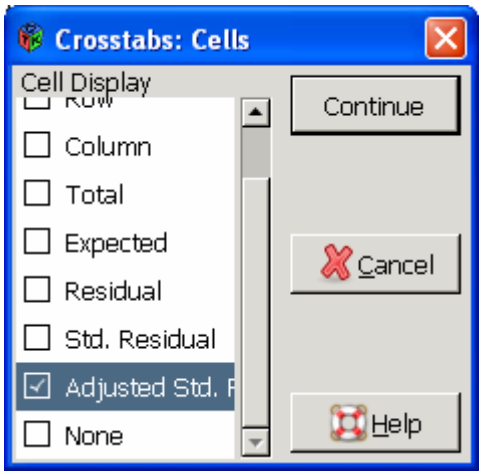
Chi-square tests.

Statistic	*Value*	*df*	*Asymp. Sig. (2-sided)*
Pearson Chi-Square	122,42	9	,00
Likelihood Ratio	123,44	9	,00
Linear-by-Linear Association	70,15	1	,00
N of Valid Cases	843		

Znaménkové schéma – v kterých polích tabulky je jak silná souvislost?

Při interpretaci souvislosti v tabulce používáme nejčastěji sloupcová procenta (viz výše), s výhodou lze využít tzv. Adjustovaných residuí, na jejichž základě se utvoří znaménkové schéma odchylek v polích. Jde o jednoduchý způsob k zčitelnění % odchylek od tzv. teoretického rozdělení četností v málo přehledné kontingenční tabulce tak, aby v ní byla rychlá optická orientace. Procentní odchylky převádíme na znaménka „+“ a „—“, podle vypočtené veličiny „z“ [Šafář 1969: 566; podrobně viz Řehák, Řeháková 1986].

V CROSSTABS si zadáme Adjustovaná standardizovaná residua (**Adjusted Std. Residuals**)



Výsledkem může být buď samostatná tabulka jako zde nebo další řádek v kontingenční tabulce:

prijem4 * vzd4 [adj. resid.].

prijem4	*vzd4* ZŠ	VYUČ	SŠ	VŠ	Total
I. do 7 tis.	7,4	-2,8	-1,3	-1,9	,0
II. 7–9 tis.	,5	4,6	-4,5	-1,1	,0
III. 9–15 tis.	-4,7	2,0	1,6	-,8	,0
IV. nad 15 tis.	-4,1	-4,0	4,8	4,4	,0
Total					

Platí, že čím vyšší hodnoty (uvažujeme ale **pouze vyšší než 1,96**), tím větší souvislost mezi kategoriemi nezávislé a závislé proměnné. Buď jsou kategorie v tom kterém políčku nadreprezentovány (kladné hodnoty) nebo podreprezentovány (záporné hodnoty) oproti teoretickému očekávání (to zohledňuje odlišně zastoupení kategorií uvažovaných znaků).

Graficky to lze vyjádřit pomocí tzv. **znaménkového schématu**, které slouží ke zčitelnění odchylek od tzv. teoretického rozdělení četností v málo přehledné kontingenční tabulce tak, aby ní byla rychlá optická orientace.
% odchylky převádíme na znaménka „+“ a „–“

- kde abs(z) >= **3.29** nahradíme +++ resp. – – –,
- kde abs(z) >= **2.58** nahradíme ++ resp. – –,
- kde abs(z) >= **1.96** nahradíme + resp. –.
- pro abs(z) < **1.96** dáme 0

což odpovídá klíči:
0 = statisticky nevýznamné
+, – = významná odchylka na 5 % hladině statistické významnosti
++, – – = významnost od 0,1 % do 1 %,
+++, – – –, = pravděpodobnost náhodného výskytu odchylky menší než 0,1 %

V našem případě to pak vypadá takto (vytvořeno v SPSS, ale znaménka lze dodatečně udělat v Excelu či Wordu – stačí přepsat Adjustované residua podle výše uvedeného pravidla):

		vzd4 Vzdělání			
		1 ZŠ	2 VYUČ	3 SŠ	4 VŠ
prijem4 Příjem-osobní – kvartily (s24)	1 I. do 7 tis.	+++	--	o	o
	2 II. 7–9 tis.	o	+++	---	o
	3 III. 9–15 tis.	---	+	o	o
	4 IV. nad 15 tis.	---	---	+++	+++

Interpretace znamének je ve smyslu „nacházíme výrazně více lidí s určitou vlastností (zde kategorií vzdělání) v daných polích tabulky oproti tzv. očekávané četnosti“. Nejde tedy o žádný průměr.

Nezapomeňte, než uděláte znaménkové schéma, nejprve byste měli provést Chíkvadrát test pro celou tabulku. Zjistit, zda můžete zamítnout nulovou hypotézu o tom, že žádná kategorie v tabulce se „statisticky významně“ neodlišuje od ostatních (tedy můžeme-li přijmout tzv. alternativní hypotézu, že alespoň v jednom poli se zjištěná četnost odlišuje). Připomínám, že tento test říká, že tento rozdíl platí nejen pro data v našem vzorku, ale že ho nalezneme i v celé populaci.

Zadání kontingenční tabulky pomocí syntaxu

Následující text lze zkopírovat do Syntax editoru v PSPP/SPSS (v menu: File / New / Syntax) a spustit (pomocí Run nebo Ctrl+R). Nejprve je třeba otevřít datový soubor ISSP2007_v2_1.sav.

```
*zadání kontingenční tabulky v SPSS / PSPP:
pro COL = sloupcová procenta.

CROSSTAB prijem4 BY vzd4 /CELL=COL COUNT.

*pro test nezávislosti Chí kvadrát přidáme.
CROSSTAB prijem4 BY vzd4 /CELL=COL COUNT /STATISTICS=CHISQ

*pro přidání adjustovaných residuí ještě.
CROSSTAB prijem4 BY vzd4 /CELL=COL COUNT ASRESID
/STATISTICS=CHISQ.

*tabulky % a N samostatně dostaneme takto.
CROSSTAB prijem4 BY vzd4 /CELL= COUNT.
CROSSTAB prijem4 BY vzd4 /CELL= COL.
CROSSTAB prijem4 BY vzd4 /CELL= ASRESID.
```

*nebo otočeno o 90st, tedy pro **ROW = řádková procenta** a **prohodíme proměnné**!.

```
CROSSTAB vzd4 BY prijem4  /CELL=ROW COUNT.
CROSSTAB vzd4 BY prijem4  /CELL=ROW COUNT ASRESID.
```