

Crosstabs – kontingenční tabulky

Kontingenční tabulky

Kontingenční tabulky užíváme ke zjišťování vztahu dvou kategorizovaných proměnných. Výstupní tabulka obsahuje jedno pole pro každou kombinaci hodnot těchto proměnných. Tabelovat můžeme:

- Četnosti (pozorované i očekávané)
- Procenta (řádková, sloupcová nebo vzhledem k celé tabulce)
- Rezidua (nestandardizovaná, standardizovaná, adjustovaná standardizovaná)

Tabulku lze rovněž doplnit sloupcovým grafem četností. Pro zjišťování závislosti řádkové a sloupcové proměnné jsou k dispozici různé typy statistických testů, z nichž nejčastěji je užíván Pearsonův test chí-kvadrát.

Volání procedury v IBM SPSS Statistics

Analyze → Descriptive Statistics → Crosstabs

Nastavení dialogu

The screenshot shows the 'Crosstabs' dialog box in IBM SPSS. On the left is a list of variables. In the center, 'Priority [rec11]' is in the 'Row(s):' field and 'Věkové kategorie [katvek]' is in the 'Column(s):' field. Below these, 'Pohlaví [ot.14]' is in a layer box labeled 'Layer 1 of 1'. On the right, buttons for 'Exact...', 'Statistics...', 'Cells...', and 'Format...' are visible. At the bottom, there are checkboxes for 'Display clustered bar charts' and 'Suppress tables', and buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'. Arrows point from text labels to specific parts of the dialog: 'řádky' points to the 'Row(s):' field, 'sloupce' points to the 'Column(s):' field, 'vrstvy (třídění vyšších stupňů)' points to the layer box, 'zobrazit sloupcový graf' points to the 'Display clustered bar charts' checkbox, and 'nevytvářet tabulku' points to the 'Suppress tables' checkbox.

- Do políček *Row(s)* resp. *Column(s)* přeneseme řádkovou resp. sloupcovou proměnnou. Jestliže do některého z polí vložíme více proměnných, vytvoří se pro každou z nich samostatná tabulka.

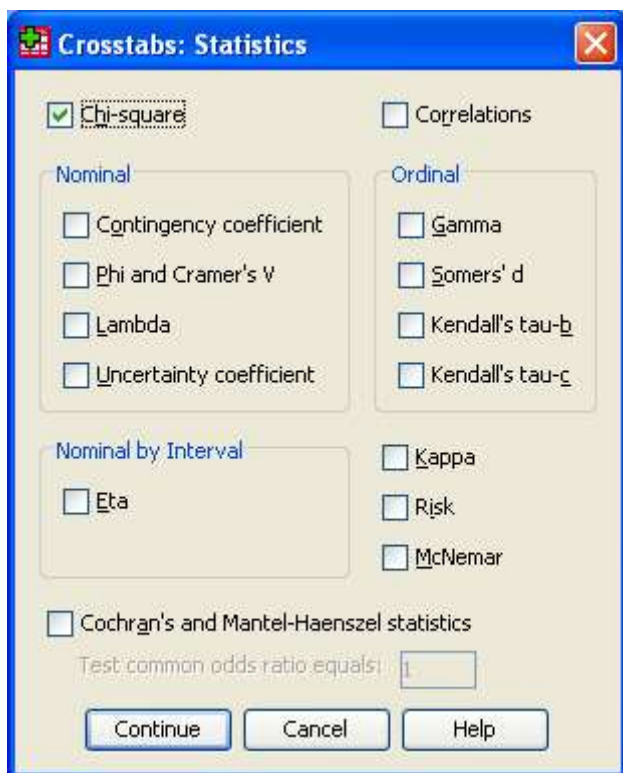
Listy procedur IBM SPSS Statistics

- Pokud potřebujeme získat tabulku vyššího stupně třídění, využijeme políčko *Layer* a pomocí tlačítka *Next* přecházíme mezi dalšími vrstvami třídění.
- Zaškrtneme-li tlačítko *Display clustered bar charts*, zobrazí se ve výstupu také skupinkový sloupcový graf.
- Při označení *Suppress tables* se ve výstupu nevytvoří kontingenční tabulka, ale jenom zvolené statistiky nebo grafy.

Tlačítko *Exact*

Tlačítko je dostupné jen při nainstalování modulu *Exact Tests*. Umožňuje provádět spolehlivé statistické testy nezávislosti i pro tabulky, které vykazují malé očekávané četnosti v některých buňkách.

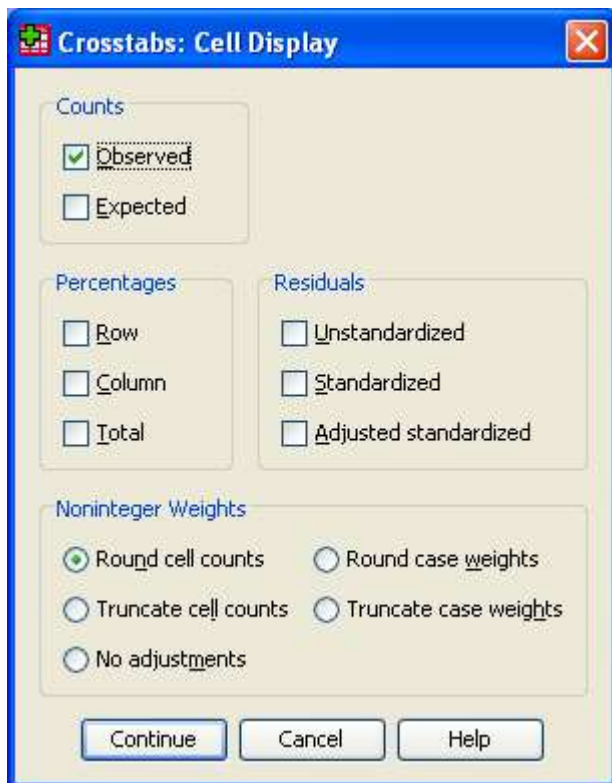
Tlačítko *Statistics*



Pomocí tlačítka *Statistics* zaškrtneme požadované statistiky a testy. Procedura obsahuje celkem 15 různých měr závislosti dvou proměnných. Statistiky jsou rozděleny do několika skupin podle typu proměnných. Podrobnější informace jsou k dispozici v příloze.

Nejčastěji užívaným testem je Pearsonův chí-kvadrát (*Chi-square*). Pro ordinální proměnné je dále k dispozici například Spearmanův koeficient pořadové korelace rho (*Correlations*).

Tlačítko Cells



Tlačítkem *Cells* nastavíme, co se má zobrazit v buňkách kontingenční tabulky.

Counts (četnosti)

Observed (pozorované četnosti) – kolik záznamů obsahuje danou kombinaci hodnot.

Expected (očekávané četnosti) – kolik záznamů s danou kombinací hodnot očekáváme, za předpokladu, že zkoumané proměnné jsou statisticky nezávislé.

Percentages (procenta)

Row (řádková procenta),

Column (sloupcová procenta),

Total (procenta vzhledem k celé tabulce).

Residuals (rezidua)

Unstandardized (nestandardizovaná) – rozdíl četnosti a očekávané četnosti.

Standardized (standardizovaná) – nestandardizovaná rezidua vydělená odhadem své směrodatné odchylky.

Adjusted standardized (adjustovaná standardizovaná) – nestandardizovaná rezidua vydělená odhadem své standardní chyby, porovnávají se vzhledem k hodnotám standardizovaného normálního rozdělení.

Noninteger Weights (neceločíselné váhy)

Hodnoty, udávající četnosti v kontingenční tabulce, jsou obvykle celočíselné – jedná se o počty. Jsou-li však případy váženy neceločíselnými vahami, mohou vycházet také tyto četnosti neceločíselné. Z toho důvodu je k dispozici několik možností, jak se s touto situací vyrovnat:

Round cell counts – váhy jednotlivých případů jsou užity k výpočtu četností, avšak před výpočtem statistik jsou výsledné hodnoty zaokrouhleny.

Truncate cell counts – váhy jednotlivých případů jsou užity k výpočtu četností, avšak před výpočtem statistik jsou výsledné hodnoty oříznuty.

Round case weights – váhy jednotlivých případů jsou před výpočtem zaokrouhleny.

Truncate case weights – váhy jednotlivých případů jsou před výpočtem oříznuty.

No adjustments – pracujeme s neceločíselnými vahami i četnostmi. Jestliže však využíváme tlačítko *Exact* (je k dispozici pouze máme-li nainstalovaný modul *Exact Tests*), hodnoty četností v kontingenční tabulce musí být zaokrouhleny nebo oříznuty dříve, než dojde k výpočtu exaktních statistik.

Tlačítko Format



Kategorie řádkové proměnné můžeme setřídít vzestupně nebo sestupně podle kódů.

Výstupy

Přehled o počtu platných a chybějících případů

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
PS 1998 * Vzdělání respondenta	1434.298 ^a	35.7%	2583.011	64.3%	4017.309	100.0%

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

Informace o počtu platných a chybějících hodnot. V tomto případě je pod tabulkou poznámka, která upozorňuje na to, že případy jsou váženy neceločíselnými vahami a v kontingenční tabulce budou hodnoty zaokrouhleny.

Kontingenční tabulka

PS 1998 * Vzdělání respondenta Crosstabulation

Count		Vzdělání respondenta				Total
		ZŠ	SŠ bez maturity	SŠ s maturitou	VŠ	
PS 1998	ČSSD	67	195	154	53	469
	KDU-ČSL	27	52	33	18	130
	DŽJ	13	12	3	0	28
	KSČM	34	57	26	7	124
	ODS	47	162	168	62	439
	SPR-RSČ	7	6	9	1	23
	US	15	32	62	30	139
	jiná strana	16	26	33	7	82
Total		226	542	488	178	1434

Kontingenční tabulka ukazuje, jak jsou rozloženy preference pro jednotlivé strany mezi kategorie vzdělání. Řádky tabulky znázorňují strany volené v parlamentních volbách v roce 1998, sloupce kategorie vzdělání. Buňky vyjadřují počty respondentů, z dané věkové kategorie, kteří volili uvedenou stranu. Poslední řádek resp. sloupec, je součtem předchozích řádků resp. sloupců.

Řádková procenta

PS 1998 * Vzdělání respondenta Crosstabulation

% within PS 1998

		Vzdělání respondenta				Total
		ZŠ	SŠ bez maturity	SŠ s maturitou	VŠ	
PS 1998	ČSSD	14.3%	41.6%	32.8%	11.3%	100.0%
	KDU-ČSL	20.8%	40.0%	25.4%	13.8%	100.0%
	DŽJ	46.4%	42.9%	10.7%	.0%	100.0%
	KSČM	27.4%	46.0%	21.0%	5.6%	100.0%
	ODS	10.7%	36.9%	38.3%	14.1%	100.0%
	SPR-RSČ	30.4%	26.1%	39.1%	4.3%	100.0%
	US	10.8%	23.0%	44.6%	21.6%	100.0%
	jiná strana	19.5%	31.7%	40.2%	8.5%	100.0%
Total		15.8%	37.8%	34.0%	12.4%	100.0%

Tabulka udává řádková procenta. V tomto případě můžeme porovnat, jak se liší procentuální zastoupení věkových kategorií pro jednotlivé politické strany. Z tabulky například poznáme, že mezi voliči US a ODS je více vysokoškolsky vzdělaných osob, než je tomu u jiných stran.

Poslední sloupec je součtem předchozích, je však spíše kontrolní – hodnota je vždy rovna 100 %. Poslední řádek udává celkové procentuální zastoupení kategorií vzdělání.

Test Chí-kvadrát

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	100.486 ^a	21	.000
Likelihood Ratio	100.072	21	.000
Linear-by-Linear Association	9.649	1	.002
N of Valid Cases	1434		

a. 4 cells (12.5%) have expected count less than 5. The minimum expected count is 2.85.

Pearsonův test Chí-kvadrát (*Pearson Chi-Square*) na prvním řádku umožňuje testovat nezávislost řádkové a sloupcové proměnné. Rozhodující jsou hodnoty asymptotické signifikance. Porovnáváme je s předem danou hodnotou hladiny spolehlivosti α , zpravidla 0,05. V našem případě je signifikance menší než 0,05, a na této hladině významnosti tedy lze prohlásit, že zkoumané proměnné jsou na sobě závislé. Druhý řádek obsahuje alternativní test téhož. Třetí řádek zjišťuje existenci lineárního vztahu a má tedy smysl pouze v případě, že jsou obě proměnné ordinální.

Poznámka pod čarou nás informuje o málo početných kombinacích – prázdné či málo zaplněné buňky bývají zdrojem nespolehlivosti asymptotických statistik. Pro test Chí-kvadrát by neměla mít žádná buňka očekávanou četnost menší než 1 a více než

Listy procedur IBM SPSS Statistics

20 % buněk by nemělo mít očekávanou četnost menší než 5. V našem případě lze tedy považovat výsledek testu za adekvátní.

Adjustovaná standardizovaná rezidua

PS 1998 * Vzdělání respondenta Crosstabulation

Adjusted Residual		Vzdělání respondenta			
		ZŠ	SŠ bez maturity	SŠ s maturitou	VŠ
PS 1998	ČSSD	-1.1	2.1	-.7	-.9
	KDU-ČSL	1.6	.5	-2.2	.5
	DŽJ	4.5	.6	-2.6	-2.0
	KSČM	3.7	2.0	-3.2	-2.4
	ODS	-3.5	-.5	2.2	1.3
	SPR-RSČ	1.9	-1.2	.5	-1.2
	US	-1.7	-3.8	2.8	3.5
	jiná strana	1.0	-1.2	1.2	-1.1



PS 1998 * Vzdělání respondenta Crosstabulation

Adjusted Residual		Vzdělání respondenta			
		ZŠ	SŠ bez maturity	SŠ s maturitou	VŠ
PS 1998	ČSSD	o	+	o	o
	KDU-ČSL	o	o	-	o
	DŽJ	+++	o	--	-
	KSČM	+++	+	--	-
	ODS	---	o	+	o
	SPR-RSČ	o	o	o	o
	US	o	---	++	+++
	jiná strana	o	o	o	o

Pro specifikování, u kterých kategorií nastal významný rozdíl, můžeme použít adjustovaná standardizovaná rezidua. Tyto hodnoty porovnáváme s kvantily standardizovaného normálního rozložení pro naši zvolenou hladinu spolehlivosti.

Výsledek této analýzy lze graficky znázornit znaménkovým schématem (viz obrázek). Tabulku s adjustovanými residui snadno upravíme do uvedené podoby pomocí skriptu *Znaménkové schéma*, který je volně k dispozici na stránkách www.acrea.cz. Skript porovnává adjustovaná rezidua s 95%, 99% a 99,9% kvantily standardizovaného normálního rozdělení (tj. zaokrouhleně $\pm 1,96$ / $\pm 2,58$ / $\pm 3,29$) a hodnoty vně těchto intervalů indikují (na příslušné hladině spolehlivosti) narušení nezávislosti v dané buňce. V každé buňce je potom znaménko rezidua uvedeno tolikrát, kolik z uvedených mezí bylo překročeno.

Listy procedur IBM SPSS Statistics

Z naší tabulky tedy například zjistíme, že mezi respondenty, kteří volili DŽJ a KSČM je výrazně více osob se základním vzděláním, než bychom očekávali a naopak mezi voliči ODS je takových respondentů méně. Mezi voliči US je výrazně více vysokoškoláků.

Ukázka alternativního způsobu zadávání tabulky

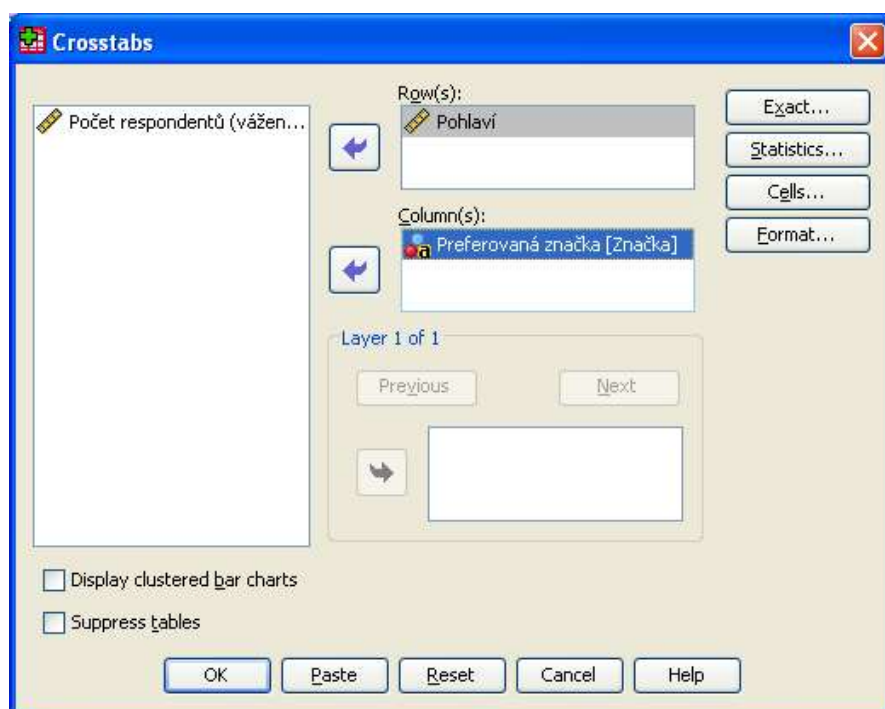
Data mohou do procedury *Crosstabs* vstupovat nejen ve tvaru, kdy jeden případ představuje například jednoho respondenta, ale také agregovaně jako četnosti. Potom každý řádek představuje jednu kombinaci řádkové a sloupcové proměnné. Kromě toho musí datový soubor obsahovat další sloupec vyjadřující četnosti. Pomocí tohoto sloupce jsou potom případy váženy (nabídka *Data, Weight Cases*).

Proměnná vyjadřující počet nemusí nutně být celočíselná. Četnosti mohou být odvozeny ze souboru, který již byl vážený – například pomocí designových vah (u některých jednotek byla větší pravděpodobnost zahrnutí do souboru).

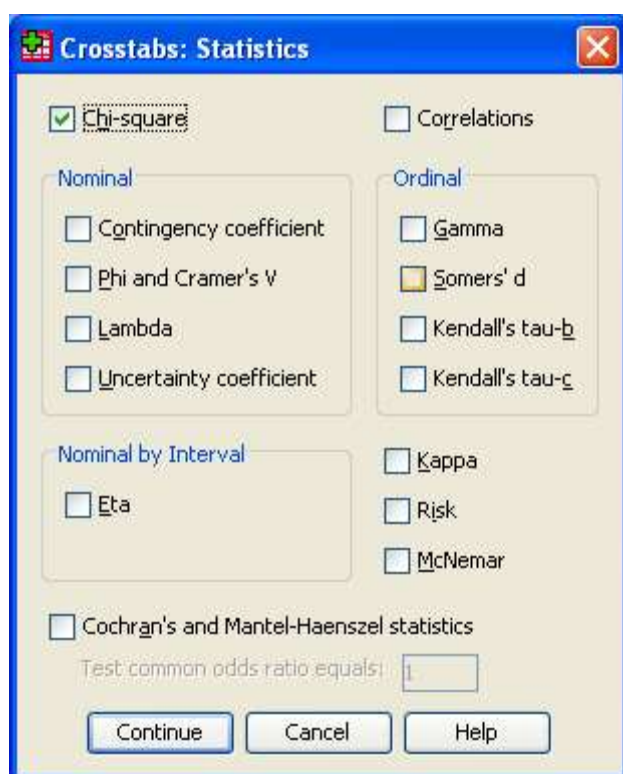
	Pohlaví	Značka	Počet	var	var	var
1	muž	A	540,09			
2	muž	B	436,96			
3	muž	C	588,72			
4	muž	D	637,64			
5	muž	E	470,14			
6	muž	F	581,23			
7	žena	A	573,68			
8	žena	B	497,63			
9	žena	C	600,81			
10	žena	D	578,68			
11	žena	E	437,66			
12	žena	F	560,84			
13						
14						

Datový soubor na předchozím obrázku představuje počty mužů a žen, kteří preferují určitou značku. Z tohoto souboru vytvoříme kontingenční tabulku a pokusíme se otestovat, zda muži a ženy mají stejné preference.

Listy procedur IBM SPSS Statistics

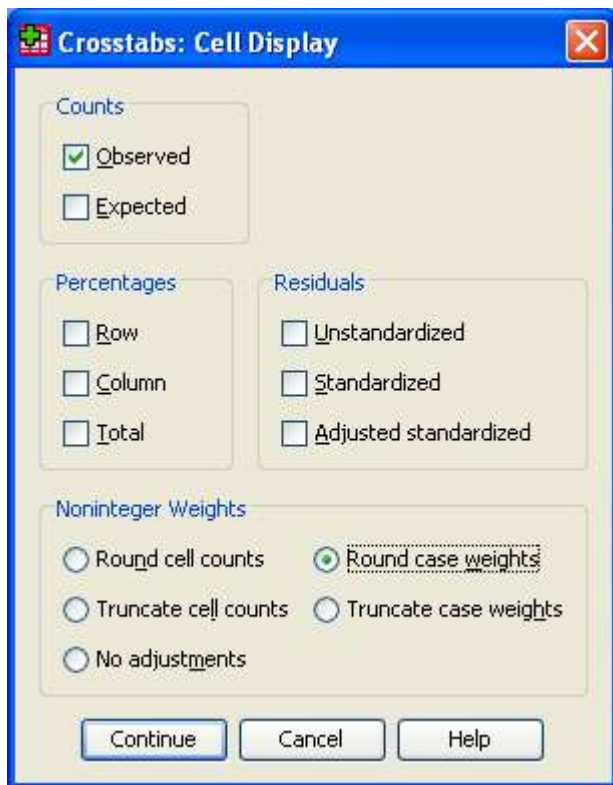


V dialogovém okně procedury *Crosstabs* přeneseme do pole *Row(s)* proměnnou *Pohlaví*, do pole *Column(s)* proměnnou *Preferovaná značka*.



Pomocí tlačítka *Statistics* označíme výpočet testu Chí-kvadrát (Chi-square).

Listy procedur IBM SPSS Statistics



Dále poklepeme myší na tlačítko Cells a označíme vhodnou volbu pro práci s neceločíselnými vahami.

Pohlaví * Preferovaná značka Crosstabulation

Count		Preferovaná značka						Total
		A	B	C	D	E	F	
Pohlaví	muž	540	437	589	638	470	581	3255
	žena	574	498	601	579	438	561	3251
Total		1114	935	1190	1217	908	1142	6506

Ve výstupu dostáváme jednak přehlednou kontingenční tabulku, jednak výsledek Pearsonova testu chí-kvadrát. Na základě hodnoty signifikance v prvním řádku (0.092), která je větší než 5 % nezamítáme na 95% hladině spolehlivosti nulovou hypotézu, že preference mužů a žen jsou stejné.

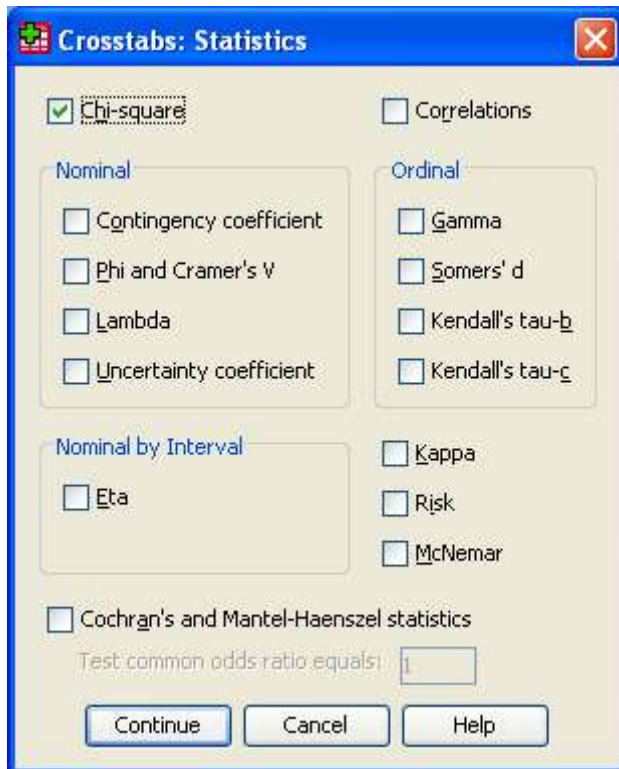
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.474 ^a	5	.092
Likelihood Ratio	9.479	5	.091
N of Valid Cases	6506		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 453.72.

Příloha 1

Tlačítko Statistics



Tlačítko *Statistics* nabízí výpočet několika testů nezávislosti a poměrně velké množství měr síly vztahu dvou či více kategorizovaných proměnných. Volba *Chi-square* zahrnuje Pearsonovu chí-kvadrát statistiku, test poměrem věrohodností, Fisherův exaktní test a Mantel-Haenszelovu statistiku. Další míry jsou v dialogu rozděleny do skupin podle typu proměnných, pro které jsou navrženy.

Pro kardinální a ordinální proměnné¹ je k dispozici Pearsonův a Spearmanův korelační koeficient (*Correlations*). Pro nominální proměnné jsou určeny: koeficient kontingence (*Contingency coefficient*), ϕ a Cramérovo V (*Phi a Cramer's V*), lambda a Goodman-Kruskalovo tau (*Lambda*) a koeficient nejistoty (*Uncertainty coefficient*). Pro ordinální proměnné lze užít: koeficient gama (*Gamma*), Somersovo d (*Somers' d*), Kendallovo tau-b (*Kendall's tau-b*) a Kendallovo tau-c (*Kendall's tau-c*). Pro kombinaci nominální a kardinální proměnné je vhodný koeficient eta (*Eta*). Pro čtvercové tabulky jsou k dispozici: koeficient kappa (*Kappa*), poměr šancí a relativní riziko (*Risk*) a McNemar-Bowkerův test (*McNemar*). Volba *Cochran's and Mantel-Haenszel statistics* zahrnuje Cochranovu, Mantel-Haenszelovu, Breslow-Dayovu a Taronovu statistiku.

¹ U kardinálních proměnných je dáno uspořádání kategorií a jejich vzájemné vzdálenosti. U ordinálních proměnných se bere v úvahu pouze uspořádání kategorií, vzdálenost dvou po sobě jdoucích kategorií se předpokládá vždy stejná.

Chí-kvadrát testy (*Chi-square*)

Pearsonův chí-kvadrát test se používá k testování hypotézy nezávislosti dvou kategorizovaných proměnných tvořících kontingenční tabulku. Označíme-li řádkovou proměnnou A a její kategorie A_1, A_2, \dots, A_R a sloupcovou proměnnou B a její kategorie B_1, B_2, \dots, B_S , pak za platnosti hypotézy nezávislosti platí:

$$P(A=A_i, B=B_j) = P(A=A_i) * P(B=B_j) \text{ pro } i = 1, \dots, R \text{ a } j = 1, \dots, S.$$

Jinými slovy, pravděpodobnost současného výskytu A_i a B_j se rovná součinu pravděpodobnosti výskytu A_i a pravděpodobnosti výskytu B_j . Označíme-li dále n_{ij} četnosti v jednotlivých polích tabulky, n_{i+} a n_{+j} příslušné marginálie a n celkový počet případů, pak četnosti očekávané za platnosti hypotézy nezávislosti lze vyjádřit takto:

$$n * P(A=A_i, B=B_j) = n * P(A=A_i) * P(B=B_j),$$

což odhadneme:

$$n * (n_{i+} / n) * (n_{+j} / n) = n_{i+} * n_{+j} / n.$$

Výraz $n_{i+} * n_{+j} / n$ označíme E_{ij} .

Pearsonova chí-kvadrát statistika χ_P^2 je rovna součtu čtverců rozdílů pozorovaných a očekávaných četností normovaných očekávanými četnostmi a má za platnosti hypotézy nezávislosti asymptoticky chí-kvadrát rozdělení s $(R-1)*(S-1)$ stupni volnosti:

$$\chi_P^2 = \sum \sum \frac{(n_{ij} - E_{ij})^2}{E_{ij}}.$$

Aproximace rozložení Pearsonovy chí-kvadrát statistiky pomocí chí-kvadrát rozdělení je dobrá, pokud jsou všechny očekávané četnosti E_{ij} větší než 1 a alespoň 80 % z nich je větších než 5 (Cochranovo doporučení).

Pro čtyřpolní tabulky, tj. tabulky s rozměry 2x2, se počítá navíc Pearsonova chí-kvadrát statistika s Yatesovou korekcí na spojitost:

$$\chi_C^2 = \sum \sum \frac{(\max(0, |n_{ij} - E_{ij}| - 0,5))^2}{E_{ij}}.$$

Jedná se o korekci neshody mezi diskretním rozložením případů do polí kontingenční tabulky a spojitým chí-kvadrát rozdělením. Hodnota Pearsonovy chí-kvadrát statistiky s Yatesovou korekcí je vždy menší nebo rovna Pearsonově chí-kvadrát statistice bez korekce. Signifikance testu s korekcí je proto vyšší než u testu bez korekce. Test s korekcí tedy zamítá hypotézu nezávislosti méně často než test bez korekce – jedná se o konzervativnější test.

Test poměrem věrohodností se také používá k testování hypotézy nezávislosti dvou kategorizovaných proměnných. Na rozdíl od Pearsonovy statistiky vychází z podílu pozorovaných a očekávaných četností. Za platnosti hypotézy nezávislosti má statistika poměrem věrohodností χ_{LR}^2 asymptoticky chí-kvadrát rozdělení s $(R-1)*(S-1)$ stupni volnosti a počítá se pomocí následujícího vzorce:

$$\chi_{LR}^2 = 2 \sum \sum n_{ij} \ln \left(\frac{n_{ij}}{E_{ij}} \right).$$

Fisherův exaktní test se v modulu *IBM SPSS Statistics Base* počítá pouze pro tabulky s rozměry 2x2, v nichž je alespoň jedna očekávaná četnost menší než 5 (pro tabulky RxS je tento test k dispozici v modulu *IBM SPSS Exact Tests*). Fisherův test je (podobně jako výše zmíněné testy) testem hypotézy nezávislosti, nicméně nevychází z aproximace, ale je to test přesný. V případě tabulek s nízkými četnostmi, kdy je aproximace chí-kvadrát rozdělením méně přesná, mu dáváme přednost před Pearsonovým chí-kvadrát testem i před testem poměrem věrohodností. Ve výstupu se zobrazuje signifikance oboustranného i jednostranného testu, protože zde neplatí, že signifikance jednostranného testu je polovinou signifikance testu oboustranného.

Mantel-Haenszelův test se používá k testování hypotézy nezávislosti mezi dvěma kardinálními proměnnými proti alternativě lineárního trendu. Jeho výpočet vychází z Pearsonova korelačního koeficientu r . Zamítnutí hypotézy nezávislosti pomocí Mantel-Haenszelova testu má tentýž význam jako prokázání nenulovosti Pearsonova korelačního koeficientu – více o korelacích lze nalézt v listu procedury *Bivariate Correlations*. Mantel-Haenszelova statistika má za platnosti hypotézy nezávislosti asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti a vypočte se následovně:

$$\chi^2_{MH} = (n - 1)r^2.$$

Tento test je ve výstupu uveden pod názvem *Linear-by-Linear Association*.

Korelační koeficienty (*Correlations*)

Korelace je mírou lineární závislosti dvou proměnných, ať už je počítaná přímo z hodnot (kódů kategorií), které jsou uloženy v datové matici (Pearsonův korelační koeficient), nebo z pořadí hodnot (Spearmanův koeficient pořadové korelace). Oba tyto koeficienty lze spočítat také pomocí procedury *Bivariate Correlations (Analyze, Correlate, Bivariate)*. Více o těchto statistikách a jejich vlastnostech lze nalézt v listu procedury *Bivariate Correlations*.

Míry pro nominální proměnné (*Nominal*)

Míry pro nominální proměnné lze rozdělit na základě principu, z něhož vycházejí, do dvou skupin: na míry asociace založené na Pearsonově chí-kvadrát statistice a na míry vycházející z rozkladu variability. Do první skupiny patří koeficient kontingence, ϕ a Cramérovo V , do druhé skupiny se řadí koeficient λ , Goodman-Kruskalovo τ a koeficient nejistoty.

Samotnou Pearsonovu chí-kvadrát statistiku není vhodné používat pro porovnání síly vztahu u tabulek, které se liší počtem řádků či sloupců nebo celkovým počtem případů, protože její hodnota roste se zvětšujícími se rozměry tabulky a s rostoucím počtem případů. Koeficient kontingence, ϕ a Cramérovo V penalizují hodnotu Pearsonovy chí-kvadrát statistiky počtem případů, případně také počtem kategorií řádkové či sloupcové proměnné.

Koeficient kontingence CC penalizuje Pearsonovu chí-kvadrát statistiku pouze počtem případů, a to následujícím způsobem:

$$CC = \sqrt{\frac{\chi_P^2}{\chi_P^2 + n}}.$$

Koeficient kontingence nabývá hodnot z intervalu $[0, \sqrt{(q-1)/q}]$, kde q je minimum z počtu řádků a sloupců tabulky – nula indikuje nezávislost proměnných, hodnoty blízké horní mezi intervalu ukazují na silnou závislost. Maximální možná hodnota tohoto koeficientu však závisí na rozměrech tabulky, což znesnadňuje jeho použití pro porovnání síly vztahů v tabulkách odlišných rozměrů.

Koeficient ϕ upravuje Pearsonovu chí-kvadrát statistiku pouze počtem případů a počítá se následovně:

$$\phi = \sqrt{\frac{\chi_P^2}{n}}.$$

Koeficient ϕ může nabýt nezáporných hodnot, menších nebo rovných výrazu $\sqrt{(q-1)}$, kde q je minimum z počtu řádků a sloupců tabulky. Čím vyšší je hodnota ϕ , tím silnější je vztah mezi proměnnými, nula odpovídá nezávislosti proměnných. Závislost maximální možné hodnoty tohoto koeficientu na rozměrech tabulky, znesnadňuje jeho použití pro porovnání síly vztahů v tabulkách odlišných rozměrů.

Pro čtyřpolní tabulky, tj. tabulky s rozměry 2x2, je ϕ rovno Pearsonovu korelačnímu koeficientu. V tomto případě nabývá hodnot z intervalu $[-1, 1]$, znaménko přebírá od Pearsonova korelačního koeficientu.

Druhá mocnina koeficientu ϕ se vyskytuje v korespondenční analýze pod názvem inercie.

Cramérovo V penalizuje Pearsonovu chí-kvadrát statistiku nejen počtem případů, ale i rozměry tabulky. Získá se následovně:

$$V = \sqrt{\frac{\chi_P^2}{n(q-1)}},$$

kde q je minimum z počtu řádků a sloupců tabulky.

Cramérovo V nabývá hodnot z intervalu $[0, 1]$ – nula ukazuje na absenci asociace, hodnoty blízké jedné na silnou asociaci mezi řádkovou a sloupcovou proměnnou. Horní mez intervalu již není (na rozdíl od dvou předešlých měr) závislá na rozměrech tabulky. Tuto mez koeficient dosahuje v případě úplné asociace, což je například lineární závislost řádkové a sloupcové proměnné a jiné typy závislosti.

Jestliže zkoumáme symetrický vztah dvou nominálních proměnných, lze doporučit použití Cramérova V.

Druhá skupina měr pro nominální proměnné vychází z rozkladu variability jedné proměnné na variabilitu vysvětlenou znalostí druhé proměnné a zbylou variabilitu. Jedná se tedy o míry asymetrického vztahu, kdy je jedna proměnná považovaná za

závislou a druhá za nezávislou. Míry lambda, Goodman-Kruskalovo tau a koeficient nejistoty vyjadřují podíl variability závislé proměnné vysvětlené znalostí nezávislé proměnné. Jedná se tedy o koeficienty determinace a liší se pouze volbou míry variability. Koeficient lambda je založen na variačním poměru, Goodman-Kruskalovo tau na Giniho míře variability a koeficient nejistoty na entropii.

Při výpočtu těchto měr lze jednou uvažovat rozklad variability sloupcové proměnné a podruhé rozklad variability řádkové proměnné. Tímto způsobem získáme dvě varianty asymetrického koeficientu. Koeficient lambda a koeficient nejistoty navíc existují i v symetrické podobě, kdy se hodnoty dvou asymetrických měr kombinují (přesněji viz níže). Všechny tyto míry nabývají hodnoty z intervalu $[0,1]$, kde nula indikuje absenci vztahu mezi proměnnými a hodnoty blízké jedné silný vzájemný vztah proměnných. Hodnotu 1 nabývají pouze v případě, pokud je v každém řádku tabulky maximálně jedna buňka s nenulovou četností.

Koeficient lambda je založen na míře variability zvané variační poměr. Variační poměr se počítá jako jedna mínus relativní četnost modální kategorie, což je zároveň pravděpodobnost chybné predikce při modálním rozhodování². Označíme-li indexem m příslušnou modální kategorii, pak lze $\lambda_{B|A}$, tj. koeficient lambda pro sloupcovou proměnnou B při znalosti řádkové proměnné A, vyjádřit následujícím způsobem:

$$\lambda_{B|A} = \frac{1 - \frac{n+m}{n} - \sum_{i=1}^R \frac{n_{i+}}{n} \left(1 - \frac{n_{im}}{n_{i+}}\right)}{1 - \frac{n+m}{n}}.$$

Koeficient $\lambda_{A|B}$ pro řádkovou proměnnou A při znalosti sloupcové proměnné B získáme analogicky.

Koeficient lambda lze interpretovat jako relativní redukci chyby predikce při modálním rozhodování. Pravděpodobnost chybné predikce sloupcové proměnné B bez znalosti řádkové proměnné A je dána výrazem $\left(1 - \frac{n+m}{n}\right)$, který tvoří jmenovatel vzorce pro výpočet $\lambda_{B|A}$. Pravděpodobnost chybné predikce sloupcové proměnné B při znalosti proměnné A je $\sum_{i=1}^R \frac{n_{i+}}{n} \left(1 - \frac{n_{im}}{n_{i+}}\right)$, což je výraz vyskytující se v čitateli vzorce pro výpočet $\lambda_{B|A}$.

Asymetrický koeficient lambda je vhodný, pokud klasifikace dle nezávislé proměnné předchází klasifikaci dle závislé proměnné. Symetrickou variantu koeficientu lambda lze použít pro symetrické vztahy. Symetrický koeficient lambda vyjadřuje relativní redukci chyby predikce v situaci, kdy je v polovině případů za závislou proměnnou brána sloupcová proměnná a v polovině případů řádková proměnná:

$$\lambda = \frac{\frac{1}{2} \left[1 - \frac{n+m}{n} - \sum_{i=1}^R \frac{n_{i+}}{n} \left(1 - \frac{n_{im}}{n_{i+}}\right) + 1 - \frac{n+m}{n} - \sum_{j=1}^S \frac{n_{+j}}{n} \left(1 - \frac{n_{mj}}{n_{+j}}\right) \right]}{1 - \frac{1}{2} \left(\frac{n+m}{n} + \frac{n+m}{n} \right)}.$$

Goodman-Kruskalovo tau vychází z Giniho míry variability, která se počítá jako jedna mínus součet čtverců relativních četností jednotlivých kategorií, což je zároveň

² Při modálním rozhodování je predikovaná kategorie s nejvyšší četností, tj. modální kategorie.

pravděpodobnost chybné predikce při proporcionálním rozhodování³. Goodman-Kruskalovo tau pro sloupcovou proměnnou B při znalosti řádkové proměnné A se značí $\tau_{B|A}$ a spočítá se následovně:

$$\tau_{B|A} = \frac{1 - \sum_{j=1}^S \left(\frac{n_{+j}}{n}\right)^2 - \sum_{i=1}^R \frac{n_{i+}}{n} \left(1 - \sum_{j=1}^S \left(\frac{n_{ij}}{n_{i+}}\right)^2\right)}{1 - \sum_{j=1}^S \left(\frac{n_{+j}}{n}\right)^2}.$$

Goodman-Kruskalovo $\tau_{A|B}$ pro řádkovou proměnnou A při znalosti sloupcové proměnné B získáme analogicky.

Goodman-Kruskalovo tau lze interpretovat jako relativní redukci chyby predikce při proporcionálním rozhodování.

Koeficient nejistoty obdržíme použitím entropie jako míry variability. Entropie sloupcové proměnné B se spočítá následovně:

$$U_B = - \sum_{j=1}^S \frac{n_{+j}}{n} \ln \left(\frac{n_{+j}}{n} \right).$$

Koeficient nejistoty pro sloupcovou proměnnou B při znalosti řádkové proměnné A značený $U_{B|A}$ lze pak vyjádřit takto:

$$U_{B|A} = \frac{- \sum_{j=1}^S \frac{n_{+j}}{n} \ln \left(\frac{n_{+j}}{n} \right) - \sum_{i=1}^R \frac{n_{i+}}{n} \left(- \sum_{j=1}^S \frac{n_{ij}}{n_{i+}} \ln \left(\frac{n_{ij}}{n_{i+}} \right) \right)}{- \sum_{j=1}^S \frac{n_{+j}}{n} \ln \left(\frac{n_{+j}}{n} \right)}.$$

Koeficient nejistoty $U_{A|B}$ pro řádkovou proměnnou A při znalosti sloupcové proměnné B získáme analogicky. Symetrický koeficient nejistoty se počítá jako harmonický průměr obou variant asymetrického koeficientu:

$$U = \frac{2}{\frac{1}{U_{B|A}} + \frac{1}{U_{A|B}}} = 2 \frac{- \sum_{j=1}^S \frac{n_{+j}}{n} \ln \left(\frac{n_{+j}}{n} \right) - \sum_{i=1}^R \frac{n_{i+}}{n} \left(- \sum_{j=1}^S \frac{n_{ij}}{n_{i+}} \ln \left(\frac{n_{ij}}{n_{i+}} \right) \right)}{- \sum_{j=1}^S \frac{n_{+j}}{n} \ln \left(\frac{n_{+j}}{n} \right) - \sum_{i=1}^R \frac{n_{i+}}{n} \ln \left(\frac{n_{i+}}{n} \right)}.$$

Jeho hodnota leží vždy mezi hodnotami asymetrických variant koeficientu nejistoty.

Míry pro ordinální proměnné (*Ordinal*)

Koeficienty gama, Somersovo d, Kendallovo tau-b a tau-c jsou míry určené pro dvě ordinální proměnné. Kromě Somersova d jsou to míry symetrické, Somersovo d je primárně asymetrická míra, ale existuje i jeho symetrická varianta.

Všechny tyto koeficienty mají vlastnosti koeficientů pořadové korelace: mohou nabývat hodnot od -1 do 1, v případě nezávislosti proměnných jsou rovny nule.

Všechny dále vycházejí z počtu konkordantních a diskordantních dvojic. Za konkordantní dvojici se považuje taková dvojice případů, kde má případ s vyšší hodnotou proměnné A i vyšší hodnotu proměnné B. Diskordantní dvojicí rozumíme takovou dvojici případů, kde případ s vyšší hodnotou proměnné A má nižší hodnotu proměnné B. Za shodu označíme situaci, kdy se hodnoty u proměnné A nebo B shodují. Dále zavedme toto značení:

³ Při proporcionálním rozhodování se kategorie predikují náhodně s pravděpodobnostmi určenými jejich relativními četnostmi.

Listy procedur IBM SPSS Statistics

P ... počet konkordantních dvojic,

Q ... počet diskordantních dvojic,

T_A ... počet dvojic, které mají stejnou hodnotu proměnné A, ale odlišnou hodnotu proměnné B,

T_B ... počet dvojic, které mají stejnou hodnotu proměnné B, ale různou hodnotu proměnné A.

Koeficient gama je symetrickou mírou založenou pouze na počtu konkordantních a diskordantních dvojic, shody nebere v úvahu. Značíme ji γ a vypočítáme takto:

$$\gamma = \frac{P-Q}{P+Q}.$$

Jedná se tedy o rozdíl podílu konkordantních a diskordantních dvojic ze všech dvojic kromě shod.

Krajních hodnot -1 a 1 nabývá nejen v případě perfektní lineární závislosti, tj. kdy se nenulové četnosti vyskytují pouze na hlavní či vedlejší diagonále tabulky, ale i když jsou buňky s nenulovou četností rozmístěny podél hlavní či vedlejší diagonály tabulky (všechny však musí být na stejné straně diagonály).

V případě vícerozměrných tabulek, kdy zadáme jednu nebo více proměnných do vrstev tabulky, bude spočítán koeficient gama pro každou jednotlivou vrstvu a také parciální koeficient gama.

Somersovo d je míra vycházející nejen z počtu konkordantních a diskordantních dvojic, ale i z počtu shod. Pro sloupcovou proměnnou B při znalosti řádkové proměnné A se Somersovo $d_{B|A}$ počítá následovně:

$$d_{B|A} = \frac{P-Q}{P+Q+T_A}.$$

Jedná se tedy o rozdíl podílu konkordantních a diskordantních dvojic ze všech dvojic kromě shod v proměnné B.

Somersovo $d_{A|B}$ pro řádkovou proměnnou A při znalosti sloupcové proměnné B se vypočte analogicky. Symetrický koeficient se získá jako harmonický průměr obou jeho asymetrických variant:

$$d = \frac{2}{\frac{1}{d_{B|A}} + \frac{1}{d_{A|B}}} = \frac{2(P-Q)}{2(P+Q)+T_A+T_B}.$$

Jeho hodnota je vždy mezi hodnotami asymetrických variant Somersova d.

Kendallovo tau-b je symetrická míra vycházející z počtu konkordantních a diskordantních dvojic a z počtu shod. Spočítá se následovně:

$$\tau_b = \frac{P-Q}{\sqrt{(P+Q+T_A)(P+Q+T_B)}}.$$

Pro čtyřpolní tabulky je shodná s Pearsonovým lineárním korelačním koeficientem.

Hraniční hodnoty 1, resp. -1 může tato míra dosáhnout pouze pro čtvercové tabulky, a to v případě perfektní lineární závislosti (nenulové četnosti pouze na hlavní, resp.

Listy procedur IBM SPSS Statistics

vedlejší diagonále tabulky). Hodnota Kendallova tau-b je v absolutní hodnotě vždy menší než hodnota koeficientu gama, ale větší než hodnota symetrické varianty Somersova d. Platí tedy:

$$|\text{symetrické Somersovo d}| < |\text{Kendallovo tau-b}| < |\text{gama}|.$$

Kendallovo tau-b lze spočítat také pomocí procedury *Bivariate Correlations (Analyze, Correlate, Bivariate)*.

Kendallovo tau-c je symetrická míra navržená jako varianta Kendallova tau-b pro obdélníkové tabulky. Spočítá se následovně:

$$\tau_c = \frac{2q(P-Q)}{n^2(q-1)},$$

kde q je minimum z počtu řádků a sloupců tabulky. Hraniční hodnoty 1, resp. -1 může dosáhnout pouze pro čtvercové tabulky, a to v případě perfektní lineární závislosti, nicméně pro obdélníkové tabulky se hraniční hodnotě 1, resp. -1 blíží více než Kendallovo tau-b.

Pokud zkoumáme symetrický vztah dvou ordinálních proměnných lze doporučit Kendallovo tau-b pro čtvercové tabulky a Kendallovo tau-c pro obdélníkové tabulky, pro asymetrické vztahy použijeme Somersovo d.

Nominální a kardinální proměnná (*Nominal by Interval*)

Koeficient eta je totožný s koeficientem eta počítaným v proceduře *Means (Analyze, Compare Means, Means)*. Jedná se o míru vycházející z rozkladu celkové variability kardinální proměnné na variabilitu uvnitř skupin daných nominální proměnnou a na variabilitu mezi skupinami. Variabilita je měřená pomocí rozptylu počítaného z hodnot uložených v datové matici (kódů jednotlivých kategorií). Koeficient je v proceduře *Crosstabs* počítán oběma možnými způsoby, tj. za kardinální proměnnou se jednou považuje sloupcová proměnná a podruhé řádková proměnná kontingenční tabulky.

Míry pro čtvercové tabulky

Cohenovo kappa je mírou shody dvou hodnotitelů, která bere v potaz i možnost náhodné shody. Tento koeficient je založen na porovnání pravděpodobnosti shody vypočtené z dané tabulky (vyjádřeno jako součet celkových procent diagonálních prvků) a očekávané pravděpodobnosti shody za předpokladu nezávislosti daných dvou hodnotitelů, tj. pravděpodobnosti náhodné shody. Hodnota Cohenova kappa se spočte následovně:

$$\kappa = \frac{\sum \frac{n_{ii}}{n} - \sum \frac{n_{i+} n_{+i}}{n^2}}{1 - \sum \frac{n_{i+} n_{+i}}{n^2}}.$$

Maximální možná hodnota je 1 a vyjadřuje perfektní shodu. Nula odpovídá situaci, kdy pravděpodobnost shody odhadnutá z tabulky je stejná jako pravděpodobnost náhodné shody. Kappa může nabývat i záporných hodnot v případech, kdy pravděpodobnost shody odhadnutá z tabulky je menší než očekávaná pravděpodobnost shody, což však nastává zřídka.

Poměr (podíl) šancí (odds ratio) a **relativní riziko** (relative risk) jsou určeny pouze pro čtyřpolní tabulky (tj. čtvercové tabulky s rozměry 2x2).

Poměr šancí je mírou asociace mezi dvěma dichotomickými proměnnými. Na rozdíl od relativního rizika, přistupuje k oběma proměnným symetricky. Jeho hodnota se získá:

$$R_0 = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Poměr šancí vyjadřuje podíl šancí⁴ sledovaného jevu v první a ve druhé skupině pro situaci, kdy jedna z proměnných charakterizuje výskyt jevu a druhá reprezentuje skupiny. Poměr šancí roven jedné odpovídá nezávislosti proměnných, skupiny se vzhledem k výskytu jevu neliší. Hodnota větší než jedna charakterizuje častější výskyt jevu v první skupině, hodnota menší než jedna častější výskyt jevu ve druhé skupině.

Procedura dále počítá relativní riziko pro skupiny určené kategoriemi sloupcové proměnné. Jedná se o porovnání šance pro jednotlivé kategorie sloupcové proměnné a šance vypočtené z celého souboru. Relativní rizika pro obě sloupcové kategorie se vypočítají následovně:

$$R_1 = \frac{n_{11}/n_{21}}{(n_{11}+n_{12})/(n_{21}+n_{22})} = \frac{n_{11}/(n_{11}+n_{12})}{n_{21}/(n_{21}+n_{22})},$$
$$R_2 = \frac{n_{12}/n_{22}}{(n_{11}+n_{12})/(n_{21}+n_{22})} = \frac{n_{12}/(n_{11}+n_{12})}{n_{22}/(n_{21}+n_{22})}.$$

Relativní riziko rovněž vyjadřuje pro každou sloupcovou kategorii podíl řádkových relativních četností.

Poměr šancí i relativní riziko nabývají nezáporných hodnot. Hodnota jedna odpovídá nezávislosti proměnných. Pro obě míry jsou rovněž k dispozici intervaly spolehlivosti (vycházející z normálního rozdělení), které lze užít k testování nezávislosti – neobsahuje-li interval spolehlivosti hodnotu jedna, pak lze na příslušné hladině spolehlivosti zamítnout hypotézu nezávislosti.

McNemar-Bowkerův test je určen k testování hypotézy symetrie ve čtvercových kontingenčních tabulkách. Hypotézu symetrie lze vyjádřit takto: $p_{ij} = p_{ji}$ pro všechny prvky kontingenční tabulky mimo diagonálu. Testová statistika se spočte následovně:

$$\chi^2 = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

Za platnosti hypotézy symetrie má tato statistika asymptoticky chí-kvadrát rozložení s $R(R-1)/2$ stupni volnosti. Procedura provede test pouze v případě, že proměnné tvořící tabulku, jsou kódovány stejně, např. obě mají kódy 1, 2, 3.

Speciálním případem McNemar-Bowkerova testu je McNemarův test pro tabulky s rozměry 2x2. Tady hypotéza symetrie splývá s hypotézou marginální homogenity (zastoupení kategorií řádkové a sloupcové proměnné je totožné). Signifikance McNemarova testu se počítá přesně pomocí binomického rozdělení, jedná se tedy

⁴ Šance (odds) vyjadřuje podíl pravděpodobnosti, že jev nastane, vzhledem k pravděpodobnosti, že jev nenastane.

Listy procedur IBM SPSS Statistics

o exaktní test. McNemarův test pro dvě dichotomické proměnné lze spočítat také pomocí procedury *2 Related Samples Tests (Analyze, Nonparametric Tests, Legacy Dialogs, 2 Related Samples ...)* a pomocí procedury *Nonparametric Tests: Two and More Related Samples (Analyze, Nonparametric Tests, Related Samples)*. V těchto procedurách se však signifikance testu počítá přesně pomocí binomického rozdělení pouze v případě, že počet případů ve dvou polích mimo diagonálu tabulky není větší než 25. Jinak se používá aproximace pomocí chí-kvadrát rozdělení s jedním stupněm volnosti a při výpočtu statistiky se zahrnuje korekce na spojitost.

Statistiky pro vícerozměrné tabulky

Statistiky popsané v tomto oddíle se týkají vícevrstvých tabulek s rozměry $2 \times 2 \times K$ – tabulky jsou tedy tvořeny dvěma dichotomickými proměnnými a dalšími kategorizovanými proměnnými určujícími vrstvy tabulky. Vrstvám tabulky se v tomto případě říká také strata. Dvěma základními hypotézami pro tento typ tabulek je homogenita poměrů šancí a podmíněná nezávislost dvou dichotomických proměnných tvořících tabulku. Homogenita poměrů šancí znamená, že poměry šancí v jednotlivých stratech jsou stejné. Hypotéza podmíněné nezávislosti vyjadřuje, že poměry šancí v jednotlivých stratech jsou všechny rovny jedné – jedná se tedy o striktnější hypotézu. Neplatí-li hypotéza homogenity poměrů šancí, nemá již smysl uvažovat hypotézu podmíněné nezávislosti.

Breslow-Dayova a Taronova statistika slouží k testování hypotézy o homogenitě poměrů šancí. Za platnosti této hypotézy mají obě statistiky asymptoticky chí-kvadrát rozdělení s $(K-1)$ stupni volnosti, kde K je počet strat, tedy vrstev tabulky. Breslow-Dayova statistika je založena na Mantel-Haenszelově odhadu společného poměru šancí (viz níže). Taronova statistika je upravená Breslow-Dayova statistika. Její hodnota je vždy menší než hodnota Breslow-Dayovy statistiky.

Cochranova a Mantel-Haenszelova statistika jsou určeny k testování hypotézy o podmíněné nezávislosti dvou dichotomických proměnných. Mantel-Haenszelova statistika vychází z předpokladu, že data pocházejí z hypergeometrického rozdělení. Cochranova statistika předpokládá, že řádky či sloupce v jednotlivých stratech odpovídají dvěma nezávislým binomickým rozdělením. Oba testy jsou založeny na porovnání pozorované četnosti n_{11k} a očekávané četnosti v příslušné buňce tabulky. Odlišné předpoklady však vedou k různým odhadům rozptylu n_{11k} . Mantel-Haenszelova statistika navíc obsahuje korekci na spojitost. Za platnosti testované hypotézy má Mantel-Haenszelova statistika asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti. Cochranova statistika má za platnosti testované hypotézy asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti pouze v případě, že počet strat je fixní.

Mantel-Haenszelův odhad společného poměru šancí (common odds ratio) je mírou síly asociace dvou dichotomických proměnných v tabulce s rozměry $2 \times 2 \times K$. Platí-li předpoklad, že poměr šancí je ve všech stratech (vrstvách tabulky) stejný, pak lze společný poměr šancí odhadnout následovně:

Listy procedur IBM SPSS Statistics

$$\theta_{MH} = \frac{\sum \frac{n_{11k}n_{22k}}{n_{++k}}}{\sum \frac{n_{12}n_{21k}}{n_{++k}}}.$$

Mantel-Haenszelův odhad nabývá nezáporných hodnot, hodnota jedna odpovídá podmíněné nezávislosti proměnných. Společný poměr šancí se vyjadřuje také zlogaritmován – podmíněné nezávislosti pak odpovídá hodnota nula. Za platnosti předpokladu o shodě poměrů šancí ve všech vrstvách tabulky, mají odhad společného poměru šancí i jeho logaritmus asymptoticky normální rozdělení. Z tohoto faktu vychází výpočet intervalů spolehlivosti, které jsou také součástí výstupu. Výstupní tabulka dále obsahuje i signifikanci testu shody mezi společným poměrem šancí a hodnotou zadanou v poli *Test common odds ratio equals*.