LEKCE 8
MĚŘENÍ SÍLY ASOCIACE MEZI DVĚMA PROMĚNNÝMI

V minulé kapitole jsme si ukázali, jak zjistit, zdali jsou dvě proměnné na sobě závislé či nikoliv. Použili jsme k tomu testu chí-kvadrát a adjustovaných reziduí. Chí-kvadrát (χ^2) má ovšem své velké limity, jichž bychom si měli být vědomi. Bryman a Cramer (1997)[1] uvádějí čtyři:

1. Chí-kvadrát nedokáže změřit sílu vztahu (podrobněji pojednáme o tomto tématu níže). I když hodnota chí-kvadrátu bude pro nějaké dvě proměnné vysoká a bude i silná jeho statistická významnost (např. $p < 0{,}001$), tak to ještě neznamená, že souvislost mezi těmito proměnnými je vyšší než u chí-kvadrátu, jehož hodnota je nižší a má menší statistickou významnost (např. $p < 0{,}05$). Hodnota chí-kvadrátu je totiž mimo jiné také závislá na počtu řádků a sloupců, tedy na počtu variant obou znaků, což je víceméně technický parametr, který má jen málo společného s věcnou stránkou analýzy. Jediné, co nám údaje o chí-kvadrátu a jeho významnosti říkají, je, jak mnoho si můžeme být jisti, že mezi proměnnými je skutečný vztah, který nebyl způsoben výběrovou chybou.
2. Chí-kvadrát se vůbec nehodí pro situaci, kdy hledáme vztah mezi dvěma ordinálními nebo dvěma kardinálními znaky, popřípadě kombinaci obou. Norušis (1998) k tomu poznamenává, že chí-kvadrát není v takové situaci dost silný test na to, aby odhalil odchylky od nezávislosti.
3. Chí-kvadrát by se neměl používat pro tabulky o velikosti 2 x 2. Jak jsme již uvedli v pozn. 1 textu příkladů ke kapitole 7, je třeba v takovém případě použít Yatesovy korekce (*Correction for Continuity*). Někteří autoři navrhují, že v takové situaci je lépe použít fi koeficientu (*phi coefficient* - ϕ), jak uvidíme později v této kapitole.
4. Chí-kvadrát je nespolehlivý, pokud více než 20 % políček mají očekávané četnosti menší než 5 nebo pokud minimální očekávaná četnost je menší než 1.

Jak je tedy zřejmé, procedura chí-kvadrát není ideálním postupem pro hledání vztahu mezi dvěma proměnnými. Tím hlavním problémem, opakujeme, je to, že nedokáže odpovědět na otázku, jak silná je zjišťovaná souvislost (asociace nebo korelace) mezi dvěma proměnnými.

Otázka na to, zdali je mezi dvěma proměnnými vztah (asociace, korelace) je jednou ze základních otázek, kterou si při bivariační analýze dat klademe. Vztah mezi dvěma proměnnými existuje, pokud hodnoty jedné proměnné jsou vztaženy k hodnotám druhé proměnné, pokud kovariují. Zajímá nás např., 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. Při těchto otázkách se zajímáme nejen o to, zdali je mezi uvedenými proměnnými souvislost, ale také jakou má tato souvislost sílu, jak je těsná a jakou má povahu, jaký má směr – viz Loether a McTavish (1988) v předchozí kapitole. Pokud např. zjistíme, že mezi vzděláním a mírou rasové intolerance je souvislost, zajímá nás, zdali se míra intolerance se zvyšujícím se vzděláním zvyšuje, nebo snižuje a jak je tato souvislost silná (těsná).

Pro zjištění síly či těsnosti vztahu počítáme tzv. koeficienty asociace nebo korelace.[2] Je to číslo, které nabývá hodnot v intervalu od 0 do 1. Hodnota blízko 0 indikuje nezávislost, čím více se blíží jedné, tím silnější souvislost mezi proměnnými existuje. Hledáme-li souvislost mezi ordinálními či kardinálními znaky, míra korelace se bude pohybovat v intervalu od -1 do +1. I zde platí, že čím blíže je hodnota blízko jedné nebo -1, tím silnější je mezi proměnnými vztah. Pro měření síly vztahu se používá řady nejrůznějších měr asociace a korelace. V SPSS je získáme následovně:

Procedura:

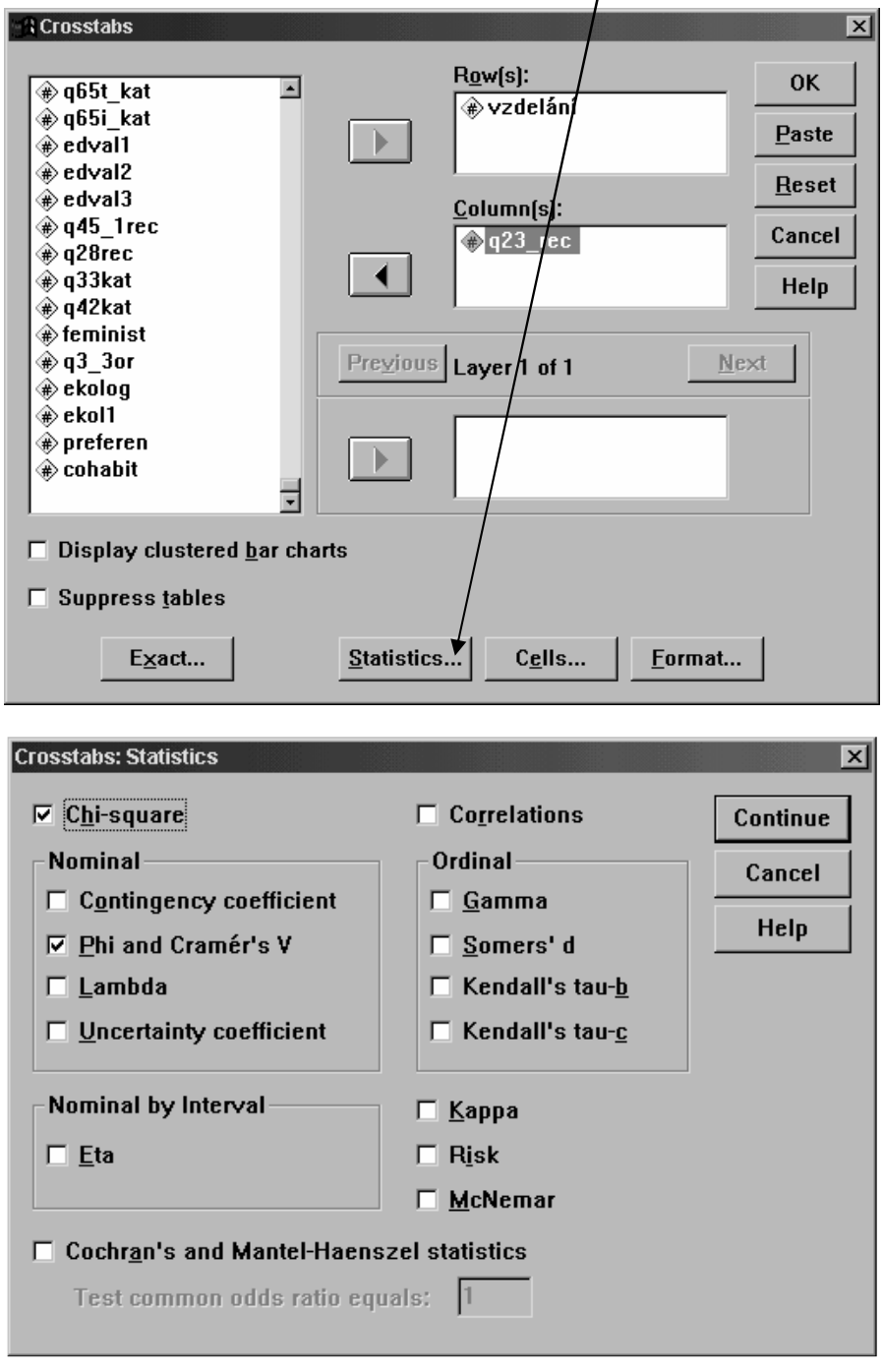
---

[1] Bryman, A., Cramer, D. 1997. *Quantitative Data Analysis with SPSS for Windows*. Routledge.

[2] Existuje úzus, že při měření síly souvislosti mezi nominálními znaky hovoříme o asociaci, při měření síly souvislosti mezi ordinálními a kardinálními znaky hovoříme o korelaci.

ANALYZE – DESCRIPTIVE STATISTICS – CROSSTABS – STATISTICS – volba příslušných koeficientů (viz obr. 8.1)

Obr. 8.1: Postup pro výpočet koeficientů asociace a korelace



To, jaký druh koeficientu kliknutím do příslušného okénka zvolíme, závisí na několika okolnostech, z nichž ta nejdůležitější je povaha proměnných (nominální, ordinální, kardinální či dichotomická), jejichž vztah hledáme.

Míry asociace pro nominální znaky

Při hledání vztahu mezi dvěma nominálními proměnnými nemáme při interpretaci příliš mnoho prostoru. Vzhledem k povaze těchto znaků, to je k faktu, že uspořádání jejich variant v sobě nenese žádné pořadí a může tedy být libovolné, předurčuje možnost interpretace – především nemůžeme nic říci o směru vztahu.

Není totiž možné např. konstatovat, že se zvyšující se barvou vlasů se zvyšuje příklon k náboženským denominacím nebo že se snižujícím se rodinným stavem se zvyšuje zakoupená značka automobilu. U žádné z těchto proměnných totiž neexistuje smysluplné pořadí kategorií, takže samozřejmě výroky, které by chtěly pracovat se směrem vztahu (typu čím více X, tím méně Y apod.), jsou nesmyslné. Jediné, co u vztahu dvou nominálních znaků lze změřit, je jeho těsnost. Je samozřejmé, že hodnoty koeficientů pro nominální znaky se budou pohybovat v intervalu od 0 do 1 (proč?). Koeficient pro nominální znaky je třeba použít i tehdy, když zjišťujeme souvislosti mezi jedním znakem nominálním a jedním ordinálním. Obecně totiž platí, že pro volbu koeficientu je rozhodující ta proměnná, která je v hierarchii měření (nominální–ordinální–intervalová) na nižším stupni.

a) Míry založené na chí-kvadrátu

Koeficient fí (*phi coefficient* - ϕ) se používá pro situaci, kdy kontingenční tabulka má podobu tabulky 2 x 2, to je má dva řádky a dva sloupce. Vypočítá se tak, že hodnota chí-kvadrát se podělí velikostí vzorku a výsledek se odmocní.

V případě, že máme vyšší počet řádků a sloupců než 2, použijeme jako míru asociace tzv. Cramérova V (*Cramér's V*).

SPSS počítá také koeficient kontingence (*coefficient of contingecny*), avšak ten nedoporučujeme používat. Jeho nevýhodou je, že jeho hodnota příliš závisí na počtu řádků a sloupců a že nenabývá nikdy hodnoty 1, i když se jedná o perfektní souvislost. Např. v tabulce 4 x 4 je nejvyšší možná hodnota tohoto koeficientu, jak upozorňuje Norušis, pouze 0,87.

Příklad 8. 1
Testujme nulovou hypotézu, že úroveň dosaženého vzdělání respondenta (*vzdelani*) nemá vliv na to, k jakému druhu náboženského vyznání se hlásí (*q23rec*).

Řešení:
Podle vzoru na obrázku 8.1 necháme spočítat koeficient *Cramerovo V*. Koeficient *fi* nelze v tomto případě použít, neboť naše tabulka bude mít 4 řádky pro vzdělání a 3 řádky pro náboženské vyznání (proměnnou q23 jsme rekódovali do smysluplného počtu kategorií tak, aby některé varianty nového znaku byly vůbec obsazeny). Výsledek výpočtu je uveden v tabulce 8.1. Vidíme v ní, že jisté rozdíly mezi některými stupni vzdělání a náboženského vyznání sice existují (např. u respondentů se základním vzděláním bylo 38 % lidí římskokatolického vyznání, u lidí se vzděláním VŠ jich bylo pouze 19 %), avšak celková souvislost je velmi nízká: V = 0,10. Jelikož statistická významnost této hodnoty je 0,000, musíme zamítnout nulovou hypotézu o neexistenci vztahu rozdílu mezi vzděláním a náboženským vyznáním. Nicméně souvislost je nízká.

Tab. 9. 1: Náboženská víra podle vzdělání respondenta

VZDELÁNÍ kategorizace q94 * Q23_REC Náboženské vyznání Crosstabulation

			Q23_REC Náboženské vyznání			
			1 Římskokatolické	2 Ostatní	3 Nehlásí se	Total
VZDELÁNÍ kategorizace q94	1 základní	Count	137	21	205	363
		Row %	37,7%	5,8%	56,5%	100,0%
	2 vyučen	Count	222	29	527	778
		Row %	28,5%	3,7%	67,7%	100,0%
	3 SŠ	Count	136	39	375	550
		Row %	24,7%	7,1%	68,2%	100,0%
	4 VŠ	Count	37	10	148	195
		Row %	19,0%	5,1%	75,9%	100,0%
Total		Count	532	99	1255	1886
		Row %	28,2%	5,2%	66,5%	100,0%

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,137	,000
	Cramer's V	,097	,000
N of Valid Cases		1886	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Míry souvislosti pro ordinální znaky

a) souvislost založená na měření konkordance (souhlasu) a diskordance (nesouhlasu)

Toto jsou míry založené na srovnávání párů hodnot. Příklad:

	X1	X2
R1	1	2
R2	2	3
R3	3	2

Srovnejme odpovědi prvního respondenta (R1) a druhého respondenta (R2). Hodnoty respondenta R2 jsou v obou případech vyšší než hodnoty R1 (2 je větší než 1 a 3 je větší než 2). Toto je příklad konkordantního páru. Případ konkordance nastává vždy, když hodnoty obou proměnných jsou vyšší (nebo nižší) než obě hodnoty druhého případu.

Případ diskordance nastává tehdy, jestliže hodnota proměnné u jednoho případu je vyšší (nebo nižší) než hodnota téže proměnné u druhého případu a u druhé proměnné je tomu přesně naopak. V naší tabulce jsou respondenti R2 a R3 diskordantní (3:2 a 2:3).

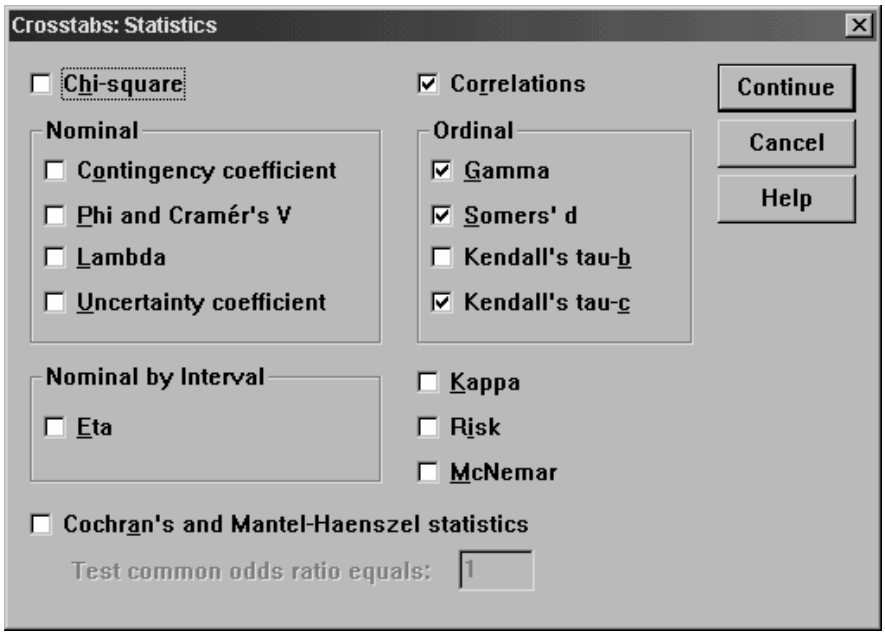
V případě, že dvě pozorování mají stejné hodnoty v jedné nebo obou proměnných, říkáme, že jsou spřaženy. U srovnání dvou případů je možných pět různých výsledků: Mohou být 1. konkordatní, 2. diskordantní, 3. spřaženy v první proměnné, 4. spřaženy v druhé proměnné nebo 5. spřaženy v obou proměnných. Jestliže bude většina párů v našich datech konkordantních, bude asociace mezi příslušnými proměnnými pozitivní, což znamená, že s růstem hodnot (nebo také poklesem) jedné proměnné porostou (nebo budou klesat) hodnoty druhé proměnné. Jestliže většina páru je diskordantních, je asociace záporná, tedy se zvyšující se hodnotou jedné proměnné se bude snižovat hodnota druhé proměnné a naopak. Pokud je počet konkordantních a diskordantních párů stejný, není mezi proměnnými žádná asociace.

Koeficienty:
Goodman-Kruskalovo gamma
Kendallovo tau b
Kendallovo tau c
Somersovo d
Spearmanovo rhó - korelace založená na pořadí

Příklad 8.2:
Zajímá nás, zdali existuje souvislost mezi názorem na to, zdali žena musí mít děti, aby se naplnilo její poslání (proměnná q42 v souboru EVS ČR 1999), a věkem respondenta kategorizovaného do věkových skupin (proměnná vek_kat).
Řešení:

Jelikož obě proměnné jsou proměnné ordinální,[3] zvolíme v příslušném dialogovém okně patřičné koeficienty (viz obr. 8.2)

Obrázek 8. 2: Ukázka zadání výpočtu koeficientů pro ordinální znaky



Tabulka 8.2: Výpočet pro zjištění souvislosti mezi věkem a postojem k poslání ženy

Crosstab

			Q42 Žena musí mít děti, aby splnila poslání		
			1 ano	2 není to nutné	Total
VEK_KAT kategorizace věku	1 18-29	Count	131	273	404
		Row %	32,4%	67,6%	100,0%
	2 30-39	Count	105	184	289
		Row %	36,3%	63,7%	100,0%
	3 40-49	Count	150	204	354
		Row %	42,4%	57,6%	100,0%
	4 50-59	Count	172	154	326
		Row %	52,8%	47,2%	100,0%
	5 60-69	Count	136	123	259
		Row %	52,5%	47,5%	100,0%
	6 70+	Count	101	66	167
		Row %	60,5%	39,5%	100,0%
Total		Count	795	1004	1799
		Row %	44,2%	55,8%	100,0%

[3] Proměnná q42 je de facto proměnná dichotomická, ale dichotomické proměnné mají tu vlastnost, že mohou být považovány jak za ordinální, tak za intervalové; proměnná „věkové skupiny" je rovněž proměnnou ordinální – zapamatujme si, že jakmile intervalovou proměnnou kategorizujeme do skupin (v našem případě věkové skupiny vznikly rekódováním proměnné věk, která byla měřena jako proměnná intervalová), přeměníme ji tímto krokem na ordinální znak.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	-,160	,020	-8,123	,000
		VEK_KAT kategorizace věku Dependent	-,213	,026	-8,123	,000
		Q42 Žena musí mít děti, aby splnila poslání Dependent	-,128	,016	-8,123	,000

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-c	-,211	,026	-8,123	,000
	Gamma	-,257	,031	-8,123	,000
	Spearman Correlation	-,187	,023	-8,055	,000[c]
Interval by Interval	Pearson's R	-,187	,023	-8,049	,000[c]
N of Valid Cases		1799			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Již samotná kontingenční tabulka naznačuje, že jistý vztah mezi sledovanými znaky existuje: názor, že žena musí mít děti, aby se naplnilo její poslání, je zastáván silněji s narůstajícím věkem. A co říkají korelační koeficienty? Především vidíme, že máme dvě skupiny koeficientů: asymetrické (zaměřené – *directional*) a symetrické. Asymetrické dokáží změřit souvislost v situaci, kdy jsme schopni rozlišit nezávisle a závisle proměnnou. V našem případě je závisle proměnná (*dependent*) a42, takže pro analýzu vztahu musíme vzít hodnotu korelace –0,13 (přesně –0,128, ale zaveďme si pravidlo, že hodnotu koeficientů budeme zaokrouhlovat na dvě desetinná místa).

Pokud porovnáme tuto hodnotu s hodnotami koeficientů symetrických,[4] zjistíme, že každý nabývá poněkud jiných hodnot. Není to chyba, je to dáno způsobem výpočtu. Pro který z nich je třeba se rozhodnout? Populární je Spearmanův koeficient nebo Kendallovo tau-c (jeho varianta Kendallovou tau-b je určena pro čtvercovou tabulku). Spearman obecně nabývá nižších hodnot než Kendall. V našem příkladu má souvislost dvou znaků hodnotu –0,187 (tedy –0,19) podle Spearmanova koeficientu a 0,211 (0,21) podle Kendalla. Gamma je ještě vyšší: 0,26. Co s tím? Doporučujeme používat pravidelně pouze jeden koeficient.

Velmi populární je Spearmanův koeficient, ale novější literatura upozorňuje, že v případě, kdy máme malý datový soubor nebo když mnoho hodnot proměnné je stejného pořadí (jsou to tzv. svázaná pořadí, *tied ranks* a tato situace vzniká tehdy, když jedna proměnná má poměrně malý počet kategorií), je lepší používat Kendallova tau (de Vaus 2002).[5] Výhodou tohoto koeficientu je navíc to, že je lepším odhadem korelace, která existuje v populaci (Field 2002). Jak Spearmanův, tak Kendallův koeficient jsou koeficienty neparametrické, což znamená, že nemusíme dbát na předpoklady určené pro použití parametrických charakteristik.

[4] Nenechejme se zmást, že v tabulce je také uveden výpočet koeficientu Pearsonova. Ten se používá, jak uvidíme dále, pro měření souvislosti dvou intervalových znaků. To, že je vytištěn mezi koeficienty pořadovými, je způsobeno nastavením SPSS: Spearmanův koeficient získáte tehdy, když si v dialogovém okně *Statistics* zakliknete požadavek na Correlations. Tato procedura tiskne ovšem koeficienty dva: Spearmanův i Pearsonův.

[5] de Vaus, D. 2002. *Analyzing Social Science Data. 50 Key Problems in Data Analysis*. Sage, London.

Všimněte si, že v každé tabulce jsou v posledním sloupci uvedeny také hodnoty statistické významnosti příslušného koeficientu. Vidíme, že všechny jsou 0,000, takže musíme zamítnout nulovou hypotézu, že mezi znaky bude v základním souboru souvislost. Zjištěnou korelaci musíme proto očekávat také v základním souboru (není dílem výběrové chyby), míra této souvislosti je ovšem nízká.

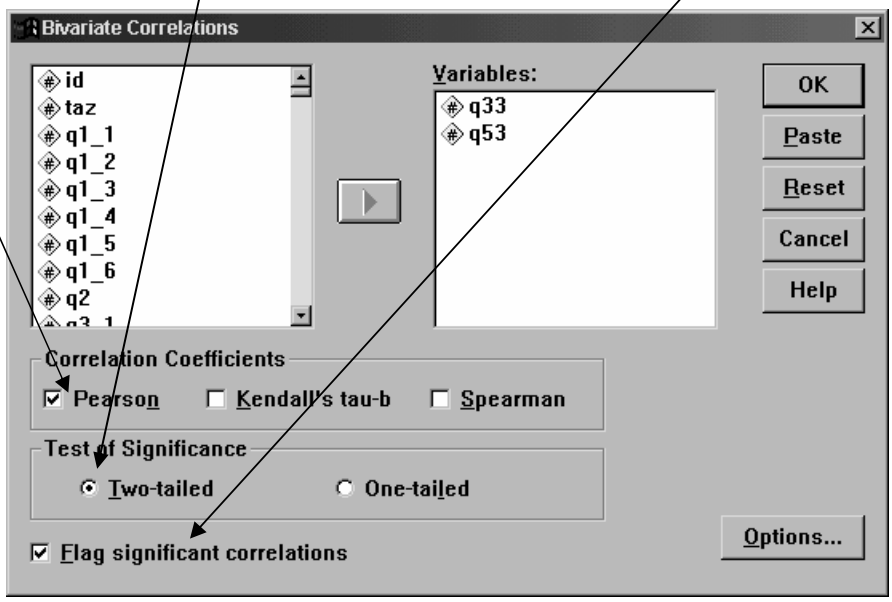
Míra souvislosti pro intervalové znaky

Souvislost mezi dvěma znaky intervalovými se měří prostřednictvím jednoho jediného koeficientu – Pearsonova koeficientu lineární korelace.

Intervalové znaky se mimo jiné vyznačují tím, že mají dlouhé stupnice měření (např. proměnná věk má u dospělých respondentů více než 60 kategorií, příjem může mít desetitisíce kategorií, levo-pravá politická orientace může mít deset kategorií atd.). Bylo by proto nesmyslné nechat vytvářet pro takovéto znaky tabulku třídění II. stupně (*Crosstabs*). Např. kdybychom třídili proměnnou levo-pravá politická orientace měřenou na desetibodové stupnici s proměnnou důležitost Boha v životě jedince měřenou rovněž na desetibodové stupnici, vznikne tabulka o 10 sloupcích a 10 řádcích, která se nedá smysluplně interpretovat. Z tohoto důvodu má SPSS nastavenou možnost vypočítat Pearsona bez tabulky Crossatbs.[6] Je jí procedura *Correlate*, která tiskne jako výstup matici korelací.

Procedura:
ANALYZE – CORRELATE – BIVARIATE – proměnné, jejichž vztahy hledáme – volba koeficientu – volba jedno či dvoustranného testu signifikance – zdůrazni signifikantní korelace

Obr. 8.3: Dialogové okno pro zadání výpočtu matice korelací



Příklad 8.3:
Existuje statistická souvislost mezi politickou orientací měřenou na levopravém kontinuu a názorem na důležitost Boha v životě jedince?

Řešení:

[6] Jistě jste si v tabulce 8.2 všimli, že Pearsonův koeficient je zabudován i v proceduře *Crosstabs*. Tento způsob výpočtu má smysl použít tehdy, když intervalové proměnné mají krátké stupnice měření. Což by např. bylo v případě, kdybychom hledali souvislost mezi počtem dětí (hodnoty této proměnné se pohybují od 0 do 4) a mírou anomie (tato stupnice nabývá hodnot od 0 do 5).

Zadání tohoto výpočtu ukazuje obr. 8.3. Výstup vypadá následovně (viz tab. 8.3):

Tabulka 8.3

Correlations

		Q33 Bůh - důležitost v životě	Q53 Levice - pravice
Q33 Bůh - důležitost v životě	Pearson Correlation	1,000	,147**
	Sig. (2-tailed)	,	,000
	N	1846	1711
Q53 Levice - pravice	Pearson Correlation	,147**	1,000
	Sig. (2-tailed)	,000	,
	N	1711	1758

**. Correlation is significant at the 0.01 level (2-tailed).

Hledaná korelace je 0,15 (0,147) a jelikož jsou u ní dvě hvězdičky, je tato korelace signifikantní na hladině významnosti 0,01 (pokud by se objevila jenom jedna, byla by korelace signifikantní na hladině významnosti 0,05). Kladné znaménko znamená, že se zvyšující se hodnotou proměnné q33 se zvyšuje také hodnota proměnné q53. Pro věcnou interpretaci se musíme pro jistotu vždy podívat, jakým směrem je stupnice u obou proměnných orientována. V našem případě je q33 směrována od nedůležitosti Boha k jeho důležitosti a proměnná q53 od levice k pravici. Což znamená, že čím více lidé zdůrazňují důležitost Boha v jejich životě, tím jsou politicky více orientováni doprava. Těsnost této souvislosti však není příliš velká. To, že je statisticky významná říká, že přibližně tak velkou souvislost můžeme očekávat také v základním souboru.

Příklad 8.4:
Existuje vzájemná souvislost mezi pocitem svobodného rozhodování o svém životě (q9), spokojeností s životem (q10) a politickou orientací (953)?
Řešení:

Tabulka 8.4:

Correlations

		Q9 Kontrola nad životem	Q10 Spokojenost se životem	Q53 Levice - pravice
Q9 Kontrola nad životem	Pearson Correlation	1,000	,423**	,134**
	Sig. (2-tailed)	,	,000	,000
	N	1888	1885	1749
Q10 Spokojenost se životem	Pearson Correlation	,423**	1,000	,163**
	Sig. (2-tailed)	,000	,	,000
	N	1885	1899	1753
Q53 Levice - pravice	Pearson Correlation	,134**	,163**	1,000
	Sig. (2-tailed)	,000	,000	,
	N	1749	1753	1758

**. Correlation is significant at the 0.01 level (2-tailed).

Výsledná matice korelací má vždy podobu čtvercové tabulky obsahující tolik řádků a sloupců, kolik proměnných vstupuje do analýzy. Všimněte si, že korelace proměnných se sebou samými jsou umístěny na diagonále tabulky a jsou vždy rovny 1. Hodnoty jednotlivých bivariačních korelací jsou zobrazeny zrcadlově pod a nad diagonálou, stačí se proto dívat pouze do jedné poloviny matice.

V tabulce 8.4 vidíme, že existuje poměrně silná a signifikantní korelace mezi kontrolou nad životem a spokojeností se životem (0,42): Se zvyšujícím se pocitem kontroly nad svým životem roste také spokojenost se životem. Korelace mezi politickou orientací a kontrolou nad životem je nízká, byť signifikantní (0,13) – spokojeni jsou spíše ti, kdo jsou orientováni pravicově. Podobně nízká je korelace mezi politickou orientací a spokojeností se životem (0,16): pravicově orientovaní respondenti mají tendenci být spokojenější se svým životem.

V souvislosti s tímto příkladem stojí zato upozornit na možnost, jak organizovat tvar korelační matice prostřednictvím příkazu syntaxe. Předpokládejme nyní, že bychom považovali levo-pravou orientaci respondenta za nezávisle proměnnou ovlivňující pocit kontroly nad životem i spokojenost se životem. Nezajímali bychom se tedy o korelaci mezi kontrolou nad životem a spokojeností se životem. Abychom dostali pohodlný výstup pro čtení hledaných souvislostí, lze matici korelací uspořádat do takovéto podoby (viz tab. 8.5):

Tabulka 8.5

Correlations

		Q9 Kontrola nad životem	Q10 Spokojenost se životem
Q53 Levice - pravice	Pearson Correlation	,134**	,163**
	Sig. (2-tailed)	,000	,000
	N	1749	1753

**. Correlation is significant at the 0.01 level (2-tailed).

Vidíme, že ve srovnání s tab. 8.4 je tato tabulka jednodušší na čtení. Abychom tuto tabulku získali, museli jsme použít syntaxe. Zatímco syntax pro tabulku 8.4 vypadá takto:[7]

```
CORRELATIONS
  /VARIABLES=q9 q10 q53
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

pro tabulku 8.5 pak takto:

```
CORRELATIONS
  /VARIABLES=q53 with q9 q10
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Jediný rozdíl spočívá v uspořádání proměnných. Do syntaxe, kterou jsme získali vlepením příkazu z dialogového okna (viz pozn. 6), vepíšeme ručně na první místo nezávisle proměnnou (q53), za ni ručně vepíšeme spojku *with*, za nímž následují závisle proměnné. Když bychom přeložili smysl obou syntaktických zápisů do normální češtiny, tak ten první říká: vypočítej vzájemné korelace proměnných q9, q10 a q53. Ten druhý pak sděluje: vypočítej matici korelací pro q53 s proměnnými q9 a q10.

Příklad 8.5:
Zajímají nás souvislosti mezi vzděláním respondenta (ISCED1) a jeho pocitem kontroly nad svým životem a spokojeností se životem.

[7] Syntax získáme tak, že v dialogovém okně procedury *Bivariate Correlations* (viz obr. 8.3) klikneme na tlačítko *Paste*. Věty syntaxe se objeví v syntaxovém souboru – na dolní liště obrazovky se objeví nové tlačítko s logem SPPS a písmenem S…

Řešení:
Jelikož hledáme korelaci mezi jasně definovanou nezávisle proměnnou proměnnou (vzdělání) a dvěma závislými proměnnými, použijeme způsob zadání výpočtu přes syntax. A jelikož jedna z proměnných je ordinální (vzdělání), musíme při zadávání výpočtu v dialogovém okně *Bivariate Correlations* zakliknout místo požadovaného Pearsonova koeficientu koeficient Spearmanův.

```
NONPAR CORR
  /VARIABLES=isced1 with q9 q10
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Tabulka 8.6

Correlations

			Q9 Kontrola nad životem	Q10 Spokojenost se životem
Spearman's rho	ISCED1	Correlation Coefficient	,125**	,117**
		Sig. (2-tailed)	,000	,000
		N	1866	1874

**. Correlation is significant at the .01 level (2-tailed).

Vzdělání koreluje s oběma proměnnými jen slabě, byť statisticky vysoce významně. Pocit kontroly nad životem a spokojenost se životem není tedy na vzdělání příliš závislý.

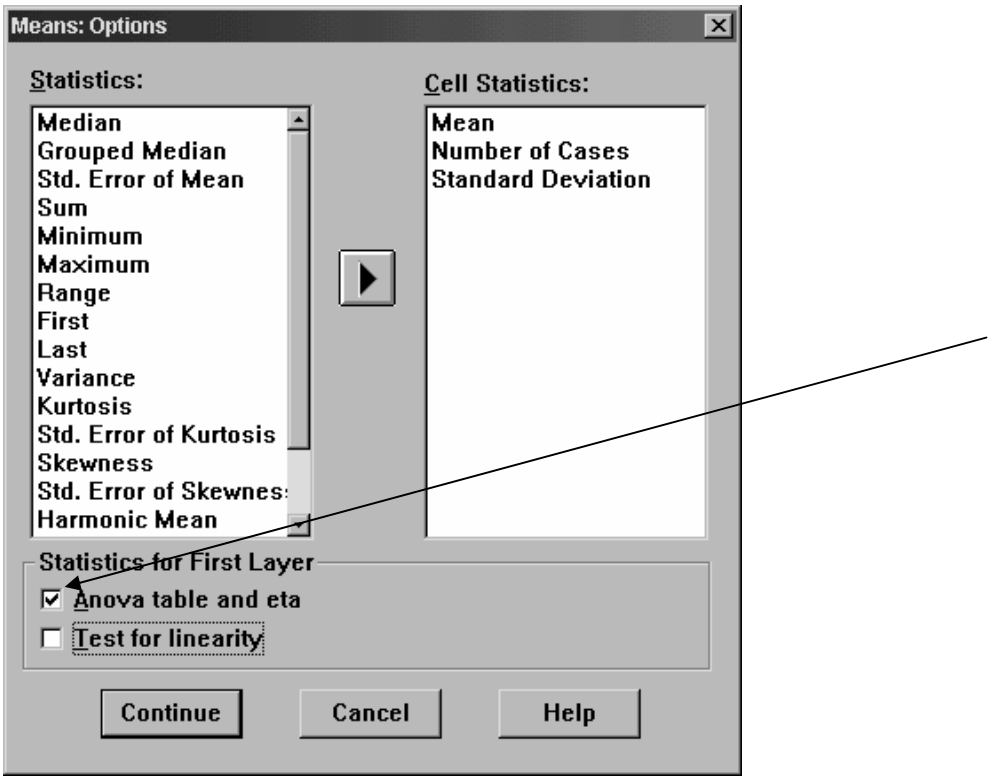
Koeficient pro souvislost nominálního znaku s kardinální (nebo dlouhou ordinální) proměnnou

V analýze dat se setkáváme s případy, kdy nás zajímá souvislost mezi nominální proměnnou a proměnnou kardinální. např. v datovém souboru EVS 1999 nás zajímá, zdali existuje souvislost mezi náboženským vyznáním[8] a spokojeností se životem. Na základě znalostí, které již máme, bychom mohli tuto úlohu řešit prostřednictvím srovnání průměrů, prostřednictvím procedury *Means*. Byla by to dobrá volba, neboť tato procedura v sobě také obsahuje výpočet **koeficientu eta**, který je určen právě pro měření souvislosti mezi nominální proměnnou a proměnnou intervalovou. Ukažme si postup výpočtu.

Přes tlačítka *Analyze – Compare means – Means* si v dialogovém okně *Means* do *Dependent list* nastavíme jako závisle proměnnou q10 (spokojenost s životem) a do *Independent list* q23-rec jako proměnnou nezávislou (a navíc nominální). Pak ve stejném okně ještě klikneme na *Options* a v něm si zvolíme výpočet *Anova table and eta* (viz obr. 8.4).

[8] Náboženské vyznání (proměnná q23) má mnoho variant, které jsou obsazeny jen malým počtem respondentů. Z tohoto důvodu jsme z proměnné q22 (zdali se respondent hlásí nebo nehlásí k nějakému náboženskému vyznání) a z proměnné q23 (k jakému vyznání se hlásí) vytvořili novou proměnnou (prostřednictvím procedury *Compute if*) q23_rec, která má varianty: 1. Římskokatolické, 2. Ostatní, 3. Nehlásí se k vyznání.

Obr. 8.4: Zadání výpočtu koeficientu eta v proceduře Means



Výstupní výpočty jsou následující (viz tabulky 8.7a – 8.7c)

Tabulka 8.7a:

Report

Q10 Spokojenost se životem

Q23_REC Náboženské vyznání	Mean	N	Std. Deviation
1 Římskokatolické	6,96	530	2,04
2 Ostatní	7,06	97	2,28
3 Nehlásí se	7,09	1257	1,88
Total	7,05	1883	1,95

Tento výstup již známe. Rozdíly v průměrech nejsou příliš velké, věřící i nevěřící jsou se svým životem poměrně spokojeni.

Tabulka 8.7b:

ANOVA Table

	Q10 Spokojenost se životem * Q23_REC Náboženské vyznání		
	Between Groups (Combined)	Within Groups	Total
Sum of Squares	6,425	7143,621	7150,05
df	2	1881	1883
Mean Square	3,213	3,798	
F	,846		
Sig.	,429		

Tato tabulka (tab. 9.7b) analýzy rozptylu je výsledkem požadovaného výpočtu *Anova table and eta*. Hodnota statistické signifikance rozdílů v průměrech je 0,43, tedy hodnota, která nám velí podržet nulovou hypotézu o neexistenci rozdílů v populaci.

Tabulka 8.7c:

Measures of Association

	Eta	Eta Squared
Q10 Spokojenost se životem * Q23_REC Náboženské vyznání	**,030**	,001

Tabulka 8.7c zobrazuje hodnotu koeficientu eta. Je velmi nízká, 0,03, takže potvrzuje to, co již naznačovaly průměry: mezi proměnnou náboženské vyznání a spokojeností se životem není žádná souvislost.

* * *

Začínáte mít pocit, že se v množství koeficientů pomalu ztrácíte? Nezoufejte, každodenní analytická praxe Vás velmi rychle naučí se v tomto množství orientovat a používat ty koeficienty, které jsou pro danou situaci adekvátní. Abychom vám v této orientaci napomohli, uvádíme dvě přehledné tabulky (tab. 8.8 a tab. 8.9) všech používaných koeficientů, které má SPSS ve svých výpočetních operacích, dále podmínky pro jejich použití a některé základní charakteristiky.

Tab. 8.8: Přehled měr asociace a jejich charakteristiky

Úroveň měření	**Počet kategorií**	***Vhodná metoda***	***Vhodný koeficient***
1. Nominální / Nominální	2 x 2	Crosstabs	Phi, Lambda
2. Nominální / Nominální	3+ x 2+	Crosstabs	Cramerovo V, Lambda
3. Nominální / Ordinální	3+ x 3+	Crosstabs	Cramerovo V, Lambda
4. Nominální / Intervalová	nominální nezávislá	a) Crosstabs (pokud má intervalová proměnná málo kategorií b) Means, ANOVA	Eta Eta
5. Ordinální / Ordinální	obě proměnné s malým počtem kategorií	Crosstabs	Gamma, Kendalovo tau b (pro čtvercovou tabulku, Sommersovo D, Kendalovo tau c (pro obdélníkovou tabulku)
6. Ordinální / Ordinální	jedna proměnná s mnoha kategoriemi	pořadová korelace	Kendalovo tau c
7. Ordinální / Ordinální	obě proměnné s mnoha kategoriemi	pořadová korelace	Kendalovo tau c Spearmanovo rhó
8. Ordinální / Intervalová	obě proměnné s několika kategoriemi	a) Crosstabs b) Srovnání průměrů pokud je závisle proměnná intervalová	Eta, stejné koeficienty jak v 5. Eta
9. Ordinální / Intervalová	ordinální s několika kategoriemi, intervalová s mnoha	a) Means b) Pořadová korelace	Eta Kendalovo tau
10. Ordinální / Intervalová	obě s mnoha kategoriemi	pořadová korelace	Kendalovo tau Spearmanovo rhó
11. Intervalová / Intervalová		bodový graf	Pearsonovo R, Regrese

Tab. 8.9: Charakteristiky měr asociace

Koeficient	Velikost tabulky	Rozsah hodnot	Směr	Symetrický	Linearita
Phi	2 x 2	<0; 1>	ne	ano	ne
Cramerovo V	větší než 2 x 2	<0; 1>	ne	ano	ne
Lambda	jakákoliv velikost	<0; 1>	ne	obě verze	ne
Gamma	jakákoliv velikost	<-1; 1>	ano	ano	ano
Somersovo d	jakákoliv velikost	<-1; 1>	ano	obě verze	ano
Kendallovo tau b	čtvercové tabulky	<-1; 1>	ano	ano	ano
Kendallovo tau c	jakákoliv velikost	<-1; 1>	ano	ano	ano
Eta	jakákoliv velikost	<0; 1>	ne	ne	ne
Spearmanovo rhó	jakákoliv velikost	<-1; 1>	ano	ano	ano
Pearsonovo r	netabelovat	<-1; 1>	ano	ano	ano

Při práci s korelačními koeficienty je třeba mít na paměti neustále jednu důležitou věc. Koeficienty pro pořadové proměnné i pro proměnné intervalové měří lineární vztah (viz poslední sloupec v tab. 8.9). Z toho ale vyplývá jedno důležité pravidlo: Vychází-li korelace pro ordinální a intervalové znaky nízká, znamená to pouze, že vztah mezi proměnnými nemá lineární povahu. Možná je souvislost mezi znaky velmi těsná, ale má jinou než lineární podobu. Co s tím udělat v praxi? De Vaus (2002) navrhuje:

- Máte-li pochybnosti o linearitě vztahu, použijte pro měření souvislosti koeficient eta (pozor ale, etu má smysl použít tehdy, když nezávisle proměnná má jen nepříliš vysoký počet variant). Eta je v tom případě vynikající koeficient, který je citlivý na zachycení i nelineárního vztahu.
- Použijte koeficientu pro nominální znaky, např. Cramerova V. Pokud tento koeficient vyjde vyšší než ten, který jste použili původně, je to indikace toho, že hledaný vztah není lineární.
- Použijte k analýze grafu nebo tabulky, abyste zjistili, v kterých kategoriích dochází k odchylce od linearity.

A co je nízká a co vysoká korelace? De Vaus navrhuje následující klasifikaci:

Tab. 9.10: Interpretace hodnot korelačního koeficientu

Hodnota korelace	interpretace souvislosti
0,01 – 0,09	triviální, žádná
0,10 – 0,29	nízká až střední
0,30 – 0,49	střední až podstatná
0,50 – 0,69	podstatná až velmi silná
0,70 – 0,89	velmi silná
0,90 – 0,99	téměř perfektní

Naše zkušenost nám ovšem říká, že když v sociologických analýzách zjistíme korelaci v řádu 0,3, máme důvod k radosti.

V souvislosti s výší korelace je třeba upozornit ještě na dva aspekty měření souvislosti dvou vztahů.

1. Ani vysoká míra korelace nemusí znamenat přítomnost kauzálního (příčinného) vztahu.
2. Při analýze nějakého problému je bivariační korelace pouhým vstupním krokem, neboť – jak již dobře víte, společenské jevy jsou velmi složitě multideterminovány. Proto ani např. zjištění korelace na úrovni 0,6 nás nesmí vést k domněnce, že jsme objevili vysvětlující faktor nebo dokonce příčinu. Aby statistici mírnili naše nadšení nad výší koeficientu korelace, zavedli tzv. **koeficient determinace**. Jeho výpočet je velmi jednoduchý: hodnotu zjištěného Pearsonova korelačního koeficientu umocníte na druhou a výsledek vynásobíte stem. Předpokládejme např. že jsme zjistili, že mezi přiřazením se respondenta k levici či pravici (měřené na 10-ti bodové stupnici) a jeho postojem, kdo by měl být odpovědný za život jedince, zdali jedince sám, nebo stát (rovněž měřený na 10-ti bodové stupnici), je korelace 0,63. Umocněním na druhou získáme výsledek 0,40. Ten po vynásobení stem (0,40 * 100 = 40 %) říká, že politická orientace respondenta (na kontinuu levice versus pravice) vysvětluje pouze ze 40 % variabilitu postoje k odpovědnosti za život jedince. Zbylých 60 % variability je třeba připsat působení jiných faktorů – zjistit které to jsou, je právě cílem vaší analýzy. Korelaci na úrovni 0,60 ovšem nacházíme v sociologických datech poměrně zřídka, typičtější je korelace na úrovni 0,25 – 0,30. Pokud by měl zjištěný koeficient korelace např. hodnotu 0,30, z něj vypočtený koeficient determinace je pouhopouhých 9 %!

Korelační koeficienty by při prezentování výsledků měly být doprovázeny také údajem o jejich statistické signifikanci[9], neboť korelační koeficient vypovídá pouze o vztahu dvou proměnných ve výběrovém souboru a neříká nic o korelaci v souboru základním. Naším cílem ovšem je zobecnit (generalizovat) výsledky z výběru na základní soubor. To nám umožňuje právě test signifikance.

Výsledek testu statistické signifikance je velmi závislý na velikosti souboru. Platí totiž, že:

- silné korelace lze získat spíše v malých souborech než ve velkých,
- silné korelace budou často statisticky nevýznamné v malých souborech,
- silná korelace v malém souboru může být statisticky nevýznamná, ale nízká korelace ve velkém souboru může být statisticky významná (de Vaus 2002).

Ilustrace: Korelace mezi vzděláním a příjem a vzděláním a frekvencí návštěv kostela v souborech o různě velkém počtu respondentů (náhodné výběry z téhož souboru)

	Příjem				Frekvence návštěv kostela			
	N=1500	N=60	N=30	N=15	N=1500	N=60	N=30	N=15
Vzdělání	0,38***	0,47**	0,44	0,80	0,08**	-0,06	-0,08	0,23

Pozn. *** = signifikance na $\alpha < 0{,}001$; ** = signifikance na $\alpha < 0{,}01$; * = signifikance na $\alpha < 0.05$,
Pramen: de Vaus (2002:177)

Vidíme např., že i nízká korelace u návštěv kostela (0,08) vychází statisticky signifikantní. Také vidíme, že v malých souborech jsou obvykle korelace mnohem vyšší než ve velkých souborech. Je to způsobeno tím, že výsledky pro malé soubory jsou značně ovlivněny výběrovou chybou, vznikají jako důsledek výběrové procedury. Naopak působení výběrové procedury, chyby výběru je ve velkém souboru méně pravděpodobné.

Testy signifikance umožňují statistickou inferenci, to je zobecnění výsledků z výběru na základní soubor, neboť nám říkají, jaká je pravděpodobnost, že výsledek zjištěný ve výběrovém souboru způsoben výběrovou chybou (*sampling error*). Z toho samozřejmě vyplývá, že nemá smysl používat testy statistické signifikance pro jiné než pravděpodobnostní (a tedy reprezentativní) výběry![10]. A nezapomeňte na to, že testy signifikance mohou být oboustranné (*two-tailed*) a jednostranné (*one-tailed*), což souvisí s tím, jak jsme

[9] O problematice statistické signifikance píše výborně a přehledně již několikrát citovaný de Vaus (2002) na stranách 167–179 a 187–193. Doporučujeme přečíst (kniha je v několika exemplářích v ústřední knihovně FSS MU).

[10] Navzdory tomu najdete výzkumné zprávy (často od psychologů), v nichž výzkumníci uvádějí i u nereprezentativních výběrů údaje a statistické signifikanci. Je to samozřejmě nesmysl.

si již uvedli dříve, zdali máme naši substantivní hypotézu nesměrovanou (*non-directional*), nebo směrovanou (directional). Nesměrovaná hypotéza je např. hypotéza o tom, že mezi muži a ženami bude rozdíl v míře nezaměstnanosti – zde bychom museli použit dvoustranného testu. Pokud bychom ale měli hypotézu, že nezaměstanost bude vyšší u žen než u mužů, máme hypotézu směrovanou, a tudíž použijeme testu jednostranného.

Nezapomeňte také, že stále platí, že statistická signifikance nemá nic společného s významností věcnou, meritorní. I statisticky vysoce signifikantní výsledek může mít z věcného hlediska nulový význam.

A konečně, i nulový výsledek ve vědě má svůj význam. Zjistíme-li mezi nějakými dvěma proměnnými nulovou korelaci, je to výsledek výzkumu, který přidáváme do mozaiky vědeckého poznání. Nalezená nula, zvláště tam, kde očekáváme souvislost, je důležitým badatelským výsledkem. Nehoňte se proto tzv. metodou „výlovu rybníka“ pouze za silnými korelacemi. Na základě precizní operacionalizace vašeho výzkumu testujte vaše substantivní hypotézy a mějte radost z každého nalezeného výsledku – pokud byste např. zjistili, že mezi vzděláním ženy a počtem jejich dětí je pouze nízká korelace (např. 0,10), jste na stopě důležitého faktu, neboť až dosud tato korelace byla střední síly. Obecně platí, že rozhodující pro úspěch vašeho výzkumu je dobrá předchozí příprava výzkumu a dobré (literaturou podepřené) hypotézy.