Metodologie pro Informační studia a knihovnictví 2

Modul 9: Úvod do induktivní statistiky

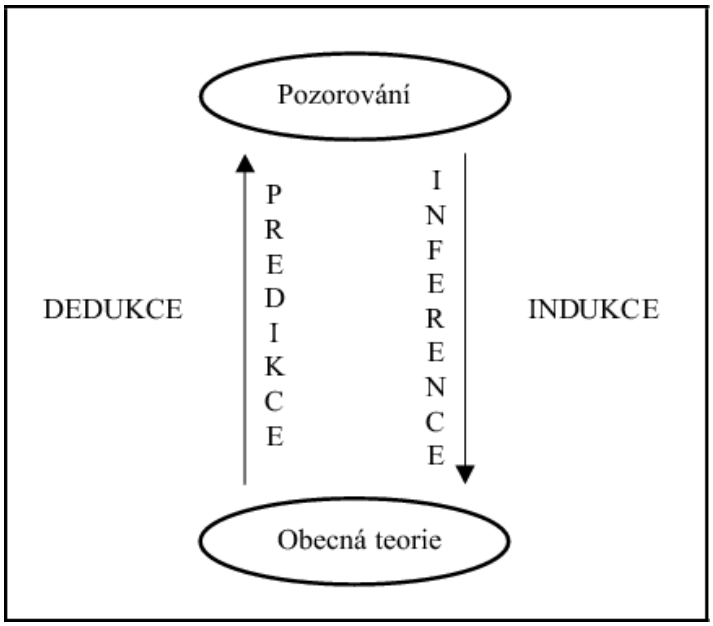
Obsah

Induktivní statistika

Dostáváme se nyní k nové kapitole statistického zpracování dat – k zobecňování na populaci. Dosud naše výpočty vypovídaly vždy jen o našich respondentech – vzorku, který neodpověděl na naše otázky. Cílem výzkumů je ale často vztáhnout výsledky na celou výzkumnou populaci, kterou vzorek zastupuje

Připomeňme si rozdíly mezi deskriptivní a induktivní statistikou:

- **Deskriptivní statistika**: popisuje rozložení četností naměřených proměnných.
- **Statistická indukce**: umožňuje zkoumat vztahy mezi proměnnými a zobecňovat výsledky na základní populaci**.**



Zdroj obrázku: http://new.euromise.org/czech/tajne/ucebnice/html/html/node3.html

Kdy můžeme zobecňovat?

Na úvod je důležité si říci, že zobecňování na populaci si nemůžeme automaticky dovolit v každém výzkumu. **Vzorek totiž musí být reprezentativní vzhledem k populaci**. Toho lze docílit různými způsoby, základním způsobem, se kterým ale počítá statistická indukce je **prostý náhodný výběr**.

Teorie statistické indukce – tedy zobecňování formou zjišťování statistické významnosti - je vyvinuta pro případy velkých reprezentativních náhodných výběrů z velkých základních souborů.

Rabušic a Soukup (2007) říkají:

„Značná část českých sociálních vědců, nemluvě o značné proporci studentů, je posedlá statistickou významností. Testy statistické signifikance v jejich povědomí (neboť tak „pochopili" smysl testováni v kurzech statistiky) slouží jako všemocné zaklinadlo. Jsou přesvědčeni, že bez testů statistických hypotéz není možné získat vědecky relevantní poznatky. Domnívají se, že tyto testy musí aplikovat na všechny

výsledky bez ohledu na to, zdali jejich data pocházejí z pravděpodobnostního (náhodného) výběru, vyčerpávajícího zjišťováni (z cenzu) nebo výběru nenáhodného (kvótního, záměrného, samovýběru). Jsou přesvědčeni, že testy významnosti jim řeknou, co je v datech důležitého, prostřednictvím nalezené statistické signifikance se snaží prokazovat těsnost vztahu dvou proměnných. Nic z toho ovšem statistická významnost neumí."

Logika statistické indukce

Přestože z úvodních řádků vyplývá, že statistickou indukci není možné aplikovat na značnou část výzkumů, které se v praxi realizují, je přesto dobré seznámit se s její logikou.

Základem statistické indukce je **testování statistických hypotéz**, přesněji řečeno zejména testování tzv. nulové hypotézy. Hypotéza je výrok o vztahu proměnných.

- **Nulová hypotéza** předpokládá stav neexistence rozdílu (tj. předpokládá stav shody) mezi proměnnými/skupinami v populaci. (Arbuthnott, 1710)
- **Alternativní hypotéza** předpokládá existenci rozdílu (na základě teorie definujeme předpoklady o rozdílech mezi jednotlivými skupinami v populaci)

Příklady nulových hypotéz:

- H0: Neexistuje rozdíl mezi rozložením proměnných ve vzorku a v populaci.
- H0: Neexistuje vztah mezi časem věnovaným internetu a pohlavím.
- Ho: Neexistuje rozdíl mezi průměrným příjmem mužů a žen zaměstnaných v knihovnách.

Příklady alternativních hypotéz:

- H0: Existuje rozdíl mezi rozložením proměnných ve vzorku a v populaci.
- H1: Neexistuje vztah mezi časem věnovaným internetu a pohlavím.

  H1a: Muži tráví na internetu více času než ženy. (Abychom si mohli dovolit formulovat takto orientovanou hypotézu, měli bychom mít podklady v předchozích výzkumech). NEBO
  H2b: Ženy tráví na internetu více času než muži. (Abychom si mohli dovolit formulovat takto orientovanou hypotézu, měli bychom mít podklady v předchozích výzkumech).

  H0: Neexistuje rozdíl mezi průměrným příjmem mužů a žen zaměstnaných v knihovnách.
  H1a: Muži zaměstnaní v knihovnách mají vyšší příjem než ženy. (Abychom si mohli dovolit formulovat takto orientovanou hypotézu, měli bychom mít podklady v předchozích výzkumech).

Pokud data neodpovídají H0, nulovou hypotézu zamítáme. Zamítnutí nulové hypotézy ovšem samo o sobě většinou nestačí k přijetí hypotézy alternativní.

Pro přijetí či zamítnutí nulové hypotézy je klíčová hladina **statistické významnosti**.

Statistická významnost je pravděpodobnost, s jakou bychom – za předpokladu platnosti nulové hypotézy – mohli obdržet data odporující nulové hypotéze. (*Soukup 2010)*

➔ Je-li statistická významnost nízká, nulová hypotéza nejspíš neplatí.

Zlaté pravidlo pro induktivní statistiku:

- Vysoká hodnota testu statistické významnosti (tj. $\alpha > 0{,}05$) → rozdíl není statisticky významný → **držíme nulovou hypotézu.**
- Nízká hodnota testu statistické významnosti (tj. $\alpha \leq 0{,}05$) → rozdíl je statisticky významný → **zamítáme nulovou hypotézu**.

Princip většiny statistických testů spočívá v tom, že se výsledky naměřených hodnot porovnávají s teoretickým modelem jejich rozložení – z něj jsou odvozeny tzv. kritické hodnoty testu (Reichel 2009). Pro různé druhy hypotéz existuje řada **testovacích kritérií**.

Proč nelze jednoduše zobecnit ze vzorku na populaci aneb zobecňování průměrů

Představte si, že zkoumáme populaci magisterských studentů knihovnictví. Chceme vidět, jak se měnil nějaký konkrétní ukazatel – třeba jejich váhu v kilogramech. Dejme tomu, že je studentů celkem 200. Náš vzorek je 15 studentů (víme už, že takový vzorek by byl velmi malý, ale pro tento příklad si jej ponechme).

Populační průměr sledované vlastnosti je 69,63. Pokaždé, kdy náhodně vybereme nějaký vzorek 15 studentů, dostaneme poněkud odlišné výsledky:

Číslo měření	***Průměr***	***St. odchylka***	***Minimum***	***Medián***	***Maximum***	***Rozpětí***
1.	*66,12*	*9,21*	*47,2*	*65*	*87*	*39,8*
2.	*73,3*	*12,48*	*52,4*	*71,1*	*101,1*	*48,7*
3.	*68,67*	*10,78*	*54*	*69,1*	*85,4*	*31,4*
4.	*69,95*	*10,57*	*54,5*	*68*	*87,8*	*33,3*

Takto bychom mohli pokračovat a při každém výběru bychom dostali poněkud jiné výsledky. Nyní vidíme, že z jednoho měření nelze jednoduše zobecnit průměr – každý výběr je zatížen tzv. **výběrovou chybou**.

Výběrová chyba je chyba, která vyplývá z faktu, že neměříme populaci, ale vzorek. Velikost výběrové chyby vychází především z distribuce vlastnosti v populaci. Pokud je populace homogenní vzhledem k vybranému kritériu, výběrová chyba bude pravděpodobně menší. Výběrová chyba také bude klesat s velikostí vzorku. Vzorek 50 studentů bude mít pravděpodobně nižší výběrovou chybu než vzorek 15 studentů.

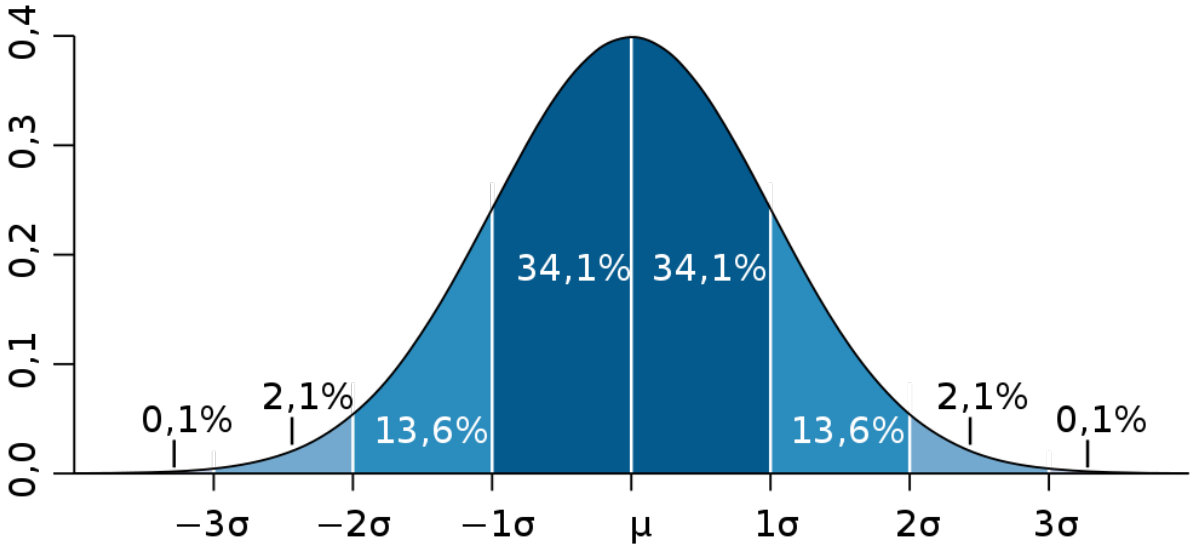
Jak se vypořádat s výběrovou chybou? Musíme pochopit, že ze vzorku nemůžeme se 100%pravděpodobností usuzovat na výsledek (průměr) celé populace. O výsledku tedy můžeme hovořit jen jako o odhadu v rámci určitého intervalu a s určitou mírou jistoty.

Je jasné, že čím nižší míra jistoty, tím menší může být interval, ve kterém se spolehlivě průměr nachází v populaci, a naopak: pokud chceme mít vysokou míru jistoty, interval bude větší.

Nejčastěji volíme **interval spolehlivosti 95 % nebo 99 %**. To znamená, že o naměřeném výsledku můžeme s 95% (respektive 99%) spolehlivostí tvrdit, že se nachází v daném intervalu.

K výpočtu horní a spodní hranice interval spolehlivosti nám pomůže znalost velikosti směrodatné odchylky.

Na obrázku vidíme normální rozložení hodnot v populaci. V intervalu jedné směrodatné odchylky od průměru na obou stranách leží 68,2 % všech naměřených hodnot. V intervalu dvou směrodatných odchylek už leží 95 % a v intervalu tří směrodatných odchylek leží 99 % naměřených hodnot.



Výpočet intervalu spolehlivost v Excelu

V Excelu pro výpočet intervalu spolehlivosti používáme příkaz CONFIDENCE. Podrobný popis použití příkazu najdete zde.

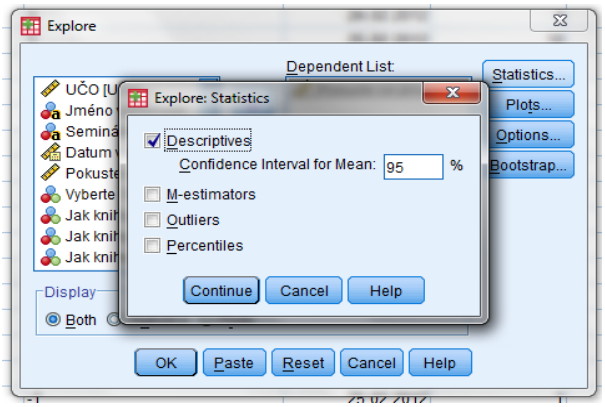
K výpočtu potřebujeme znát:

- **koeficient spolehlivosti** (0,05 pro 95% interval spolehlivosti a 0,01 pro 99% interval spolehlivosti),
- **směrodatnou odchylku v populaci**,
- **velikost výběrového souboru.**

V praxi ale většinou neznáme hodnoty průměru v populaci či výši směrodatné odchylky. Proto byly vyvinuty postupy realizovatelné při využití standardní odchylky naměřeného průměru – tzv. T-rozložení a T-test.

Výpočet intervalu spolehlivost v SPSS

V SPSS používáme záložku Explore, kde si na kartě Statistics upravíme velikost intervalu spolehlivosti:



SPSS vrátí informace o horní a spodní hranici intervalu spolehlivosti.

Hypotézy o shodě dvou populačních průměrů

Pro vyhodnocování hypotézy o shodě dvou průměrů používáme tzv. T-test.

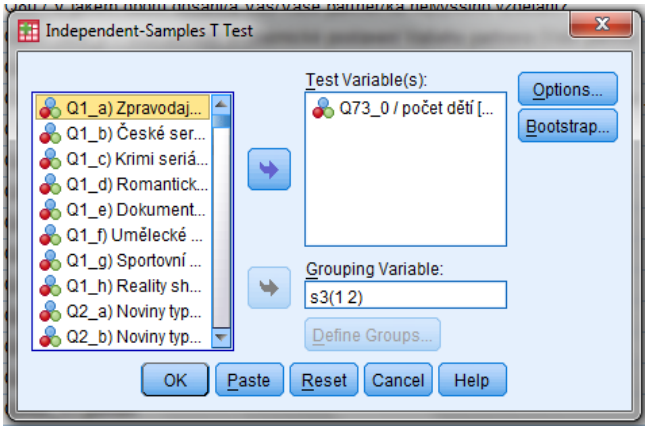
- **Studentův t-test** (William Gosset)
  - *směrodatná odchylka (s), která sama podléhá variabilitě výběru, již nemusí být spolehlivým odhadem populační směrodatné odchylky* (zdroj)
  - Pro nás relevantní: Independent Samples T-test

Např. zkoumáme vztah mezi pohlavím a počtem dětí (v populaci třicátníků) – příklad pracuje s daty z výzkumu Distinkce a hodnoty 2008 (viz Studijní materiály v ISu).

Nulová a alternativní hypotéza:

- *H0: Neexistuje rozdíl mezi počtem dětí u skupin podle pohlaví.*
- *Ha: Existuje rozdíl mezi počtem dětí u skupin podle pohlaví.*

Postup v SPSS: Analyze – Compare Means – Independent Samples T-test



Podíváme se, jaké rozdíly jsme naměřili na vzorku:

Group Statistics

	pohlaví	N	Mean	Std. Deviation	Std. Error Mean
Q73_0 / počet dětí	muž	492	,78	,915	,041
	žena	529	1,13	,939	,041

Existují rozdíly i v populaci?

Interpretujeme test ve dvou krocích:

1. *podíváme se na výsledky F testu o shodě variací*
   - *Signifikance u F > 0,05 → použijeme T-testu pro případ EQUAL VARIANCES ASSUMED*
   - *Signifikance u F < 0,05 → použijeme T-testu pro případ EQUAL VARIANCES NOT ASSUMED*
2. *v příslušném sloupci čteme významnost*

Je-li menší než 0,05, nulovou hypotézu o shodě populačních průměrů lze zamítnout – rozdíl pravděpodobně existuje i v populaci

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Q73_0 / počet dětí	Equal variances assumed	,531	,467	-6,091	1019	,000	-,354	,058	-,468	-,240
	Equal variances not assumed			-6,097	1016,790	,000	-,354	,058	-,468	-,240

Porovnávání více populačních průměrů

Opět si vše ukážeme na příkladu z výzkumu Distinkce a hodnoty 2008.

Např. zkoumáme vztah mezi vzděláním a počtem dětí

Nulová a alternativní hypotéza:

- *H0: Neexistuje rozdíl mezi počtem dětí u jednotlivých vzdělanostních skupin.*
- *Ha: Existuje rozdíl mezi počtem dětí u jednotlivých vzdělanostních skupin.*

*Nejprve si zjistíme rozdíly v **naměřených** průměrech: Analyze – **Compare Means***

Porovnání průměrů ukazuje, že v naměřených hodnotách jsou rozdíly. Jsou však rozdíly i v populaci?

Q73_0 / počet dětí

Q27 / Jaké je Vaše ...	Mean	N	Std. Deviation
základní, bez vyučení	1,19	48	1,214
střední s vyučením	1,07	316	,978
střední bez maturity	,80	104	,885
střední s maturitou	,97	386	,921
vyšší odborné (pomaturitní studium)	,89	36	,854
vysokoškolské bakalářské	,75	36	,806
vysokoškolské magisterské, inženýrské	,71	86	,824
vysokoškolské doktorské	,89	9	1,054
Total	,96	1021	,944

- *1. krok: Analyze – **One way** ANOVA*
- *Options: Descriptives*

Descriptives

Q73_0 / počet dětí

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
základní, bez vyučení	48	1,19	1,214	,175	,83	1,54	0	4
střední s vyučením	316	1,07	,978	,055	,96	1,18	0	4
střední bez maturity	104	,80	,885	,087	,63	,97	0	3
střední s maturitou	386	,97	,921	,047	,87	1,06	0	3
vyšší odborné (pomaturitní studium)	36	,89	,854	,142	,60	1,18	0	3
vysokoškolské bakalářské	36	,75	,806	,134	,48	1,02	0	2
vysokoškolské magisterské, inženýrské	86	,71	,824	,089	,53	,89	0	3
vysokoškolské doktorské	9	,89	1,054	,351	,08	1,70	0	3
Total	1021	,96	,944	,030	,90	1,02	0	4

- *2. krok: **statistika F** a její **signifikance***

ANOVA

Q73_0 / počet dětí

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16,466	7	2,352	2,669	,010
Within Groups	892,887	1013	,881		
Total	909,354	1020			

Podíl variability mezi skupinami (between groups) a variability uvnitř skupin (within groups), konkrétně jejich průměrů součtu druhých mocnin směrodatných odchylek. Pokud platí nulová hypotéza, že rozdíly mezi průměry jsou nulové, musí být obě průměrné hodnoty druhých mocnin podobné a jejich vzájemný poměr (F) tedy musí být blízko 1.

Hodnota je signifikantní (menší než 0,05). Pravděpodobnost podržet nulovou hypotézu je nízká (0,01) → **zamítáme** (tj. průměry v populaci nejsou stejné)

- *3. krok: Chceme vědět, mezi kterými skupinami **statisticky významný rozdíl** existuje*

Post Hoc Tests

Multiple Comparisons

Q73_0 / počet dětí
Bonferroni

(I) Q27 / Jaké je Vaše nejvyšší dosažené vzdělání?	(J) Q27 / Jaké je Vaše nejvyšší dosažené vzdělání?	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
základní, bez vyučení	střední s vyučením	,115	,145	1,000	-,34	,57
	střední bez maturity	,389	,164	,494	-,12	,90
	střední s maturitou	,221	,144	1,000	-,23	,67
	vyšší odborné (pomaturitní studium)	,299	,207	1,000	-,35	,95
	vysokoškolské bakalářské	,438	,207	,974	-,21	1,09
	vysokoškolské magisterské, inženýrské	,478	,169	,134	-,05	1,01
	vysokoškolské doktorské	,299	,341	1,000	-,77	1,37
střední s vyučením	základní, bez vyučení	-,115	,145	1,000	-,57	,34
	střední bez maturity	,275	,106	,274	-,06	,61
	střední s maturitou	,106	,071	1,000	-,12	,33
	vyšší odborné (pomaturitní studium)	,184	,165	1,000	-,33	,70
	vysokoškolské bakalářské	,323	,165	1,000	-,19	,84
	vysokoškolské magisterské, inženýrské	,363*	,114	,042	,01	,72
	vysokoškolské doktorské	,184	,317	1,000	-,81	1,18
střední bez maturity	základní, bez vyučení	-,389	,164	,494	-,90	,12
	střední s vyučením	-,275	,106	,274	-,61	,06

SPSS potvrdilo statisticky významný rozdíl pouze mezi skupinou SŠ s vyučením a VŠ mgr/ing. U ostatních skupin nemůžeme s jistotou říci, že rozdíl existuje i v populaci

Zobecňování výsledků třídění druhého stupně (kontingenčních tabulek)

Druhým příkladem zobecňování z naměřených hodnot na populaci je zobecňování výsledků třídění druhého stupně kategorizovaných dat.

Příklad: Chceme vědět, jak se liší frekvence čtení u skupin podle vzdělání. Formulujeme nulovou a alternativní hypotézu:

- ➔ *H0: Neexistuje rozdíl ve frekvenci čtení mezi skupinami třicátníků s různým vzděláním.*
- ➔ *Ha: Existuje rozdíl ve frekvenci čtení mezi skupinami třicátníků s různým vzděláním.*

Uděláme si kontingenční tabulku (už ji umíme od modulu 7):

Jaké je vaše vzdělání? * Četbě knih (rec) Crosstabulation

			Četbě knih (rec)				
			Několikrát týdně nebo denně	Jednou za měsíc až jednou týdně	Několikrát za rok	Vůbec ne	Total
Jaké je vaše vzdělání?	ZŠ (i nedokončené)	Count	3	11	10	23	47
		% within Jaké je vaše vzdělání?	6,4%	23,4%	21,3%	48,9%	100,0%
	SŠ/VOŠ	Count	172	292	168	200	832
		% within Jaké je vaše vzdělání?	20,7%	35,1%	20,2%	24,0%	100,0%
	VŠ	Count	57	44	17	12	130
		% within Jaké je vaše vzdělání?	43,8%	33,8%	13,1%	9,2%	100,0%
Total		Count	232	347	195	235	1009
		% within Jaké je vaše vzdělání?	23,0%	34,4%	19,3%	23,3%	100,0%

Vidíme poměrně zajímavé rozdíly! Můžeme je zobecnit?

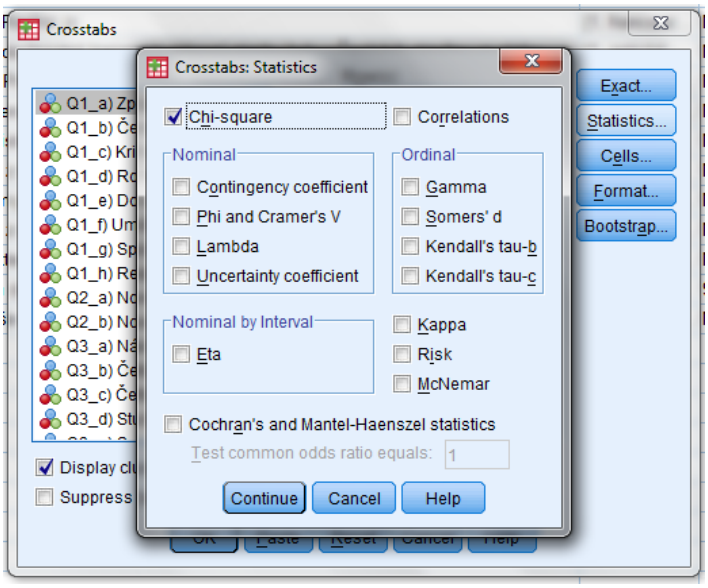
Pro zobecňování rozdílů u kategorizovaných proměnných se používá jako testovací kritérium tzv. test nezávislosti chí kvadrát ($\chi 2$).

Chí-kvadrát je založený na srovnávání naměřených a očekávaných proměnných

- **Očekávaná četnost**: počet jednotek, který by do dané kategorie spadl při náhodném rozložení
- **Naměřená četnost:** počet jednotek, které jsme v dané kategorii ve vzorku naměřili
- **Reziduál**: rozdíl mezi OČ a NČ
- **Adjustované reziduály**: koeficient determinace (AR mají přibližně normální rozložení s průměrem 0 a standardní odchylkou 1)

Chí kvadrát v SPSS

Chí-kvadrát – Analyze – Crosstabs: Statistics



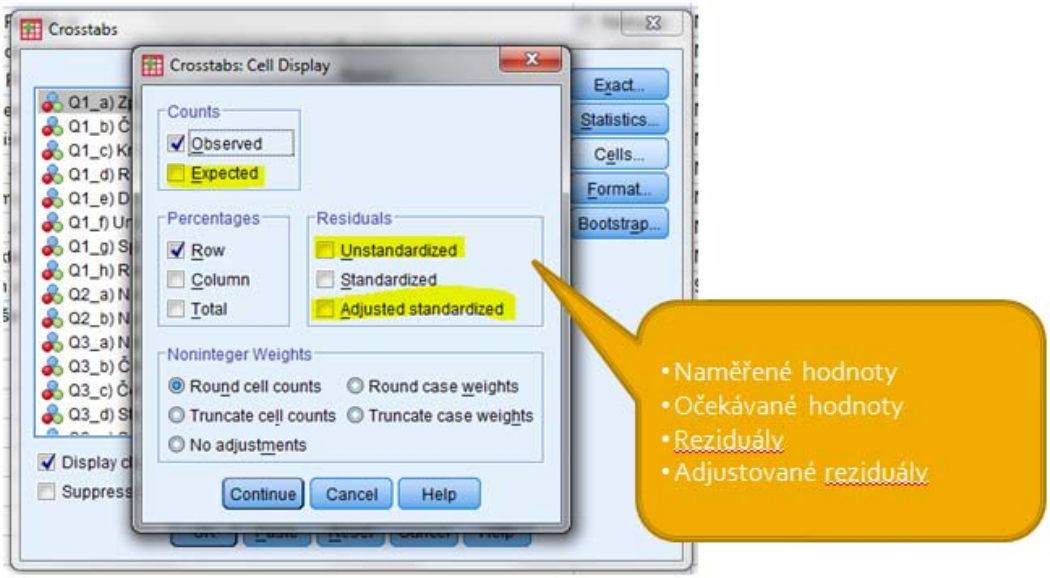
Chi-Square Tests

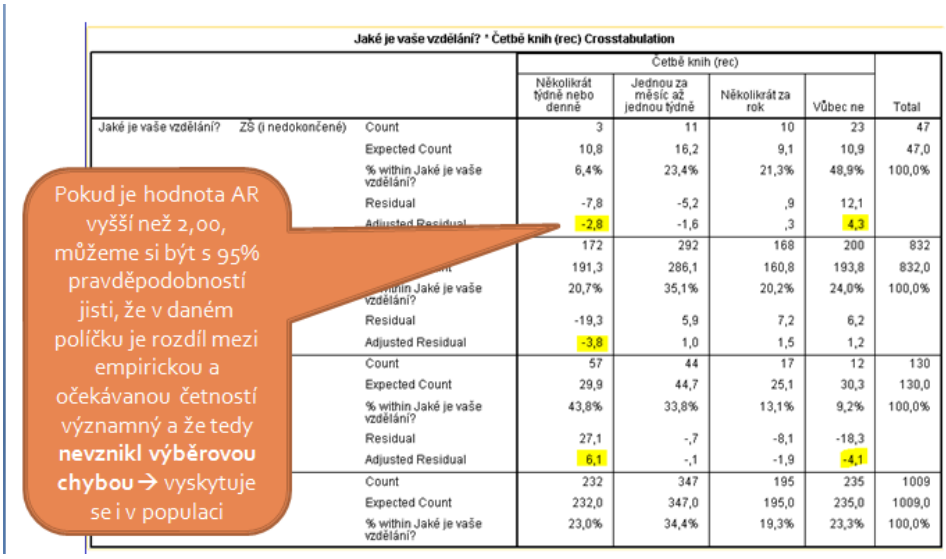
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	61,503[a]	6	,000
Likelihood Ratio	59,263	6	,000
Linear-by-Linear Association	54,887	1	,000
N of Valid Cases	1009		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,08.

Hodnota významnosti o → **zamítáme hypotézu o neexistenci rozdílu v populaci**

Pozor! Chí-kvadrát se dá použít jen pokud více než 20 % políček má očekávanou četnost menší než 5 a minimální očekávaná četnost nesmí být menší než 1





Jaké je vaše vzdělání? * Četbě knih (rec) Crosstabulation

			Četbě knih (rec)				
			Několikrát týdně nebo denně	Jednou za měsíc až jednou týdně	Několikrát za rok	Vůbec ne	Total
Jaké je vaše vzdělání?	ZŠ (i nedokončené)	Count	3	11	10	23	47
		Expected Count	10,8	16,2	9,1	10,9	47,0
		% within Jaké je vaše vzdělání?	6,4%	23,4%	21,3%	48,9%	100,0%
		Residual	-7,8	-5,2	,9	12,1	
		Adjusted Residual	-2,8	-1,6	,3	4,3	
		Count	172	292	168	200	832
		Expected Count	191,3	286,1	160,8	193,8	832,0
		% within Jaké je vaše vzdělání?	20,7%	35,1%	20,2%	24,0%	100,0%
		Residual	-19,3	5,9	7,2	6,2	
		Adjusted Residual	-3,8	1,0	1,5	1,2	
		Count	57	44	17	12	130
		Expected Count	29,9	44,7	25,1	30,3	130,0
		% within Jaké je vaše vzdělání?	43,8%	33,8%	13,1%	9,2%	100,0%
		Residual	27,1	-,7	-8,1	-18,3	
		Adjusted Residual	6,1	-,1	-1,9	-4,1	
		Count	232	347	195	235	1009
		Expected Count	232,0	347,0	195,0	235,0	1009,0
		% within Jaké je vaše vzdělání?	23,0%	34,4%	19,3%	23,3%	100,0%

Pokud je hodnota AR vyšší než 2,00, můžeme si být s 95% pravděpodobností jisti, že v daném políčku je rozdíl mezi empirickou a očekávanou četností významný a že tedy **nevznikl výběrovou chybou** → vyskytuje se i v populaci

Literatura:

Reichel, J. 2009. Kapitoly metodologie sociálních výzkumů. Praha: Grada.

Soukup, P. 2010. „Nesprávné užívání statistické významnosti a jejich možná řešení.“ Data a výzkum – SDA Info 4(2): 77–104.

SOUKUP, Petr - RABUŠIC, Ladislav. Několik poznámek k jedné obsesi českých sociálních věd - statistické významnosti. Sociologický časopis. 2007, roč. 43, č. 2, s. 379-395. ISSN 0038-0288.