Inference jako statistický proces 2

Chi^2, asociace a korelace

Proč se zabývat testem nezávislosti?

- Pouhé třídění dvou proměnných a výpočet příslušných procent, byť se jedná o velmi mocnou analytickou proceduru, nestačí k tomu, abychom hledanému vztahu mezi dvěma proměnnými dobře rozuměli. Odhalíme-li totiž, že mezi sledovanými proměnnými je vztah, musíme se dále zajímat o to, zdali jednak tento vztah vydrží i test nezávislosti, jednak jakou má tento vztah sílu. (Mareš, Rabušic, 2002)

Test nezávislosti chí-kvadrát ($\chi 2$)

Jak zjistit asociace?

- Test provedeme na základě výpočtu statistiky chí-kvadrát $\chi 2$ (*chi-square*). Ten je založen na srovnání empirických a očekávaných četnostech.
  - Empirická četnost (*observed count*) pozorovaná hodnota v políčku tabulky
  - Očekávaná četnost (*expected count*) četnost, která by se v políčku objevila, kdyby platila nulová hypotéza

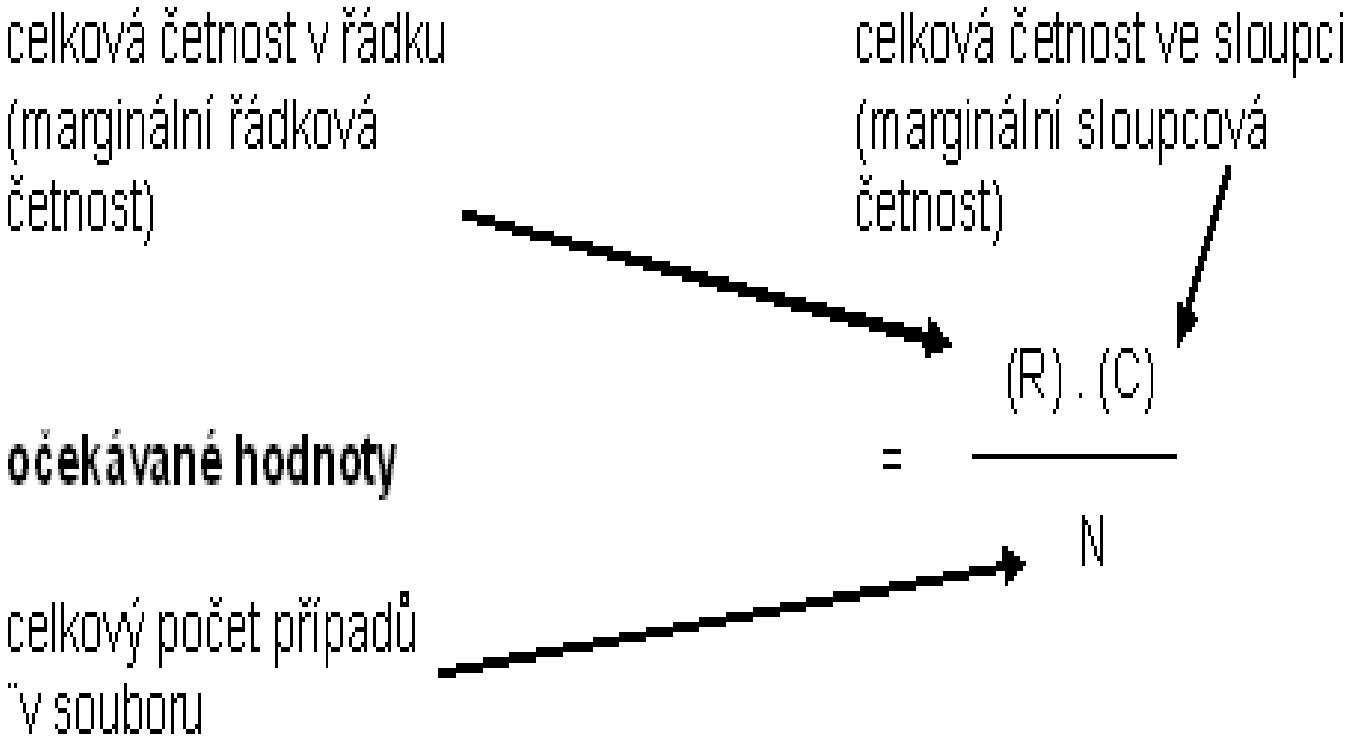
KONTINGENČNÍ TABULKY (Cross tabs)

- Nové pojmy:
  - **EXPECTED COUNT = očekávané četnosti, počet jednotek, který by byl v tomto políčku při nezávislosti obou znaků (náhodné rozložení).**
  - **RESIDUAL = rozdíl mezi pozorovaným počtem jednotek, které mají příslušnou empirickou kombinaci hodnot obou znaků a očekávanou četností. Residuály se dále standardizují a používají se v adjustované (na velikost tabulky) podobě.**
  - **STD. RESIDUAL = Standardizované chi^2 residuály, neboli residuály vydělené druhou odmocninou očekávaných hodnot.**
  - **ADJUSTED RESIDUAL = Adjustované residuály (tak, aby měly přibližně normální rozložení s průměrem = 0 a standardní odchylkou rovnou 1).**

Analýza asociací

- Řádek *Residual*: Má-li znaménko +, znamená to, že empirická četnost je vyšší, než bychom očekávali, kdyby platila nulová hypotéza, záporné znaménko vyjadřuje pravý opak, tedy že empirická četnost je nižší, než jaká by měla být, kdyby platila nulová hypotéza. V rutinní analytické praxi informace tohoto druhu nepotřebujeme, a proto takto detailní tabulku nevyžadujeme.

OČEKÁVANÉ HODNOTY
(PŘEDPOKLAD NEZÁVISLOSTI PROMĚNNÝCH)



Základy bivariační inferenční statistiky

- Test chí-kvadrát:
  - *ANALYZE – DESRIPTIVE STATISTICS* □ *CROSSTABS* □ v dialogovém okně klikneme na lištu *Statistics* a v objevivším se novém dialogovém oknu zaškrtneme políčko *Chi-square*
  - *Test chí-kvadrát je možno také chápat také jako test nezávislosti, kdy testujeme, zdali jedna proměnná závisí na druhé. Můžeme např. testovat hypotézu, zdali existuje nějaká souvislost mezi rodinným stavem respondenta a volebními preferencemi. Je to opět úloha na Crosstabs, ale v jejím rámci si ukážeme, jak je možné v rutinní analytické práci postupovat.*

Užití adjustovaného reziduálu

- Adjustovaný reziduál je založen na rozdílu mezi empirickou a očekávanou četností (jak jsme si ukázali v tab. 8.5). řečeno jazykem statistiky, je to rozdíl mezi frekvencí očekávanou (fo) a frekvencí empirickou (fe). Tomuto rozdílu se říká delta a značí se odpovídajícím řeckým písmenem (Δ).
- V adjustovaném reziduálu je pak tento rozdíl testován z hlediska statistické významnosti, přičemž platí, že pokud je jeho hodnota 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 náhodou. Interpretačně má tato informace obrovský význam, neboť nám umožňuje detailní vhled do vztahu mezi proměnnými.

Příklad (Rabušic, Mareš, 2002)

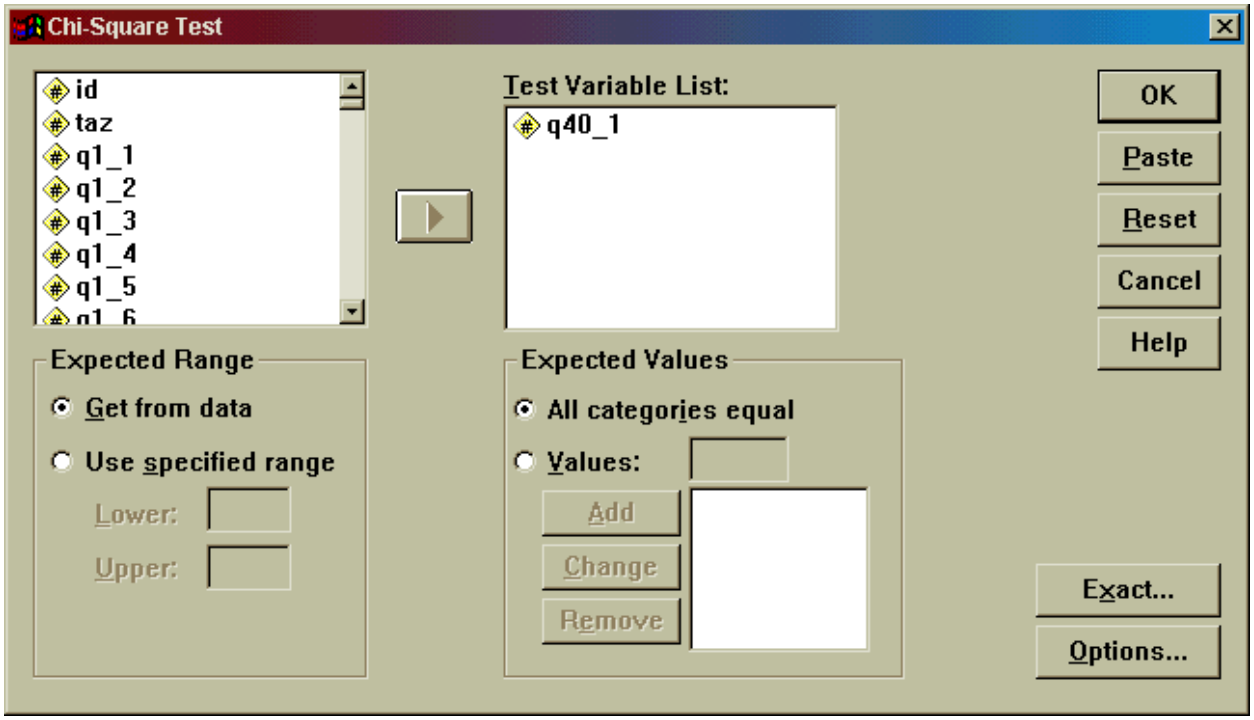
ý stav (sloučeno rozvedený + odloučení) * PREFEREN volební preferenc

			REFEREN volební preference 9						Total
			1 KSČM	2 ČSSD	3 KDU	4 US	5 ODS	9 nedělí by	
ROD_ST1 rodinný s (sloučen rozveden odloučen	1 ženatý/vd	Count	101	196	58	97	253	198	903
		Row %	,2%	1,7%	6,4%	0,7%	8,0%	1,9%	00,0%
		Adjusted Res	-,4	1,9	-1,1	-1,2	1,6	-1,6	
	2 vdovec/v	Count	36	25	15	9	29	35	149
		Row %	,2%	6,8%	0,1%	6,0%	9,5%	3,5%	00,0%
		Adjusted Res	**5,2**	-1,1	1,6	**-2,2**	**-2,1**	,1	
	3 rozveden	Count	13	14	4	11	27	38	107
		Row %	2,1%	3,1%	3,7%	0,3%	5,2%	5,5%	00,0%
		Adjusted Res	,2	-1,9	-1,4	-,4	-,3	**3,1**	
	4 nikdy neoženěn/r vdána	Count	11	49	21	45	66	56	248
		Row %	,4%	9,8%	8,5%	8,1%	6,6%	2,6%	00,0%
		Adjusted Res	**-3,8**	-,2	1,0	**3,6**	,0	-,3	
Total		Count	161	284	98	162	375	327	1407
		Row %	1,4%	0,2%	7,0%	1,5%	6,7%	3,2%	00,0%

Neparametrické užití *chi*^{2}

- testování hypotéz o rozložení hodnot jediné proměnné
- *Analyze - – Nonparametric tests – Chi-Square*

PŘÍKLAD (Mareš, Rabušic, 2002)



Výsledky

Q40_1 Věrnost v manželství

	Observed N	Expected N	Residual
1 velmi důleží	1418	646,7	771,3
2 spíše důlež	498	646,7	-148,7
3 nepříliš důle	24	646,7	-622,7
Total	1940		

Test Statistics

	Q40_1 Věrnost manželství
Chi-Square[a]	1553,769
df	2
Asymp. Sig.	,000

a. 0 cells (,0%) have expected freq
5. The minimum expected cell fr

Významnost test chí-kvadrát vyšla velmi nízká (0,000), takže nulovou hypotézu o tom, že počet osob bude ve třech zmíněných kategoriích postoje k důležitosti věrnosti pro manželství stejný, musíme zamítnout.

KOEFICIENTY

- Každý koeficient má rovněž významnost.
- OPAKOVÁNÍ:
  - nominální a ordinální – asociace – přes crosstabs a statistics,
  - kardinální – korelace (correlate)
- Nutno sledovat koeficienty významnosti – jestliže < 0,05, pak zamítáme H_0 – vztah je „významný - předpokládáme, že existuje i v základním souboru!

Obecná poznámka nakonec

(Mareš, Rabušic, 2002)

Při publikaci výsledků ovšem tabulku v takového podobě, jako jsou tab. 8.4 nebo 8.5, nikdy nezveřejňujeme. Nejsou totiž přehledné. Proto je musíme upravit. Zásady jsou následující:

.Každá tabulka musí mít číslo a název.
.Všechny popisky tabulky musí být česky.
.Názvy proměnných jsou ve sloupcích a řádcích jasně vyjádřeny.
.Nezávisle proměnnou obvykle umísťujeme do sloupců, takže počítáme sloupcová procenta.
.Závisle proměnná, která je v řádku, by měla mít varianty uspořádány od nejvyšší po nejnižší (pokud je měřená na ordinální nebo intervalové úrovni). Tento požadavek se nedodržuje příliš striktně.
.Poslední řádek uvádí celková procenta (obvykle tedy 100 %) a současně i absolutní počty případů.
.V poznámce pod tabulkou se uvádí zdroj dat.