

Kontingenční tabulky



Test dobré shody
Fisherův přesný test
McNemar test

Anotace



- Analýza kontingenčních tabulek umožňuje analyzovat vazbu mezi dvěma kategoriálními proměnnými. Základním způsobem testování je tzv. chi-square test, který srovnává pozorované četnosti kombinací kategorií oproti očekávaným četnostem, které vychází z teoretické situace, kdy je vztah mezi proměnnými náhodný.
- Test dobré shody je využíván také pro srovnání pozorovaných četností proti očekávaným četnostem daným určitým pravidlem (typickým příkladem je Hardy-Weinbergova rovnováha v genetice)
- Specifickým typem výstupů odvozených z kontingenčních tabulek jsou tzv. odds ratio a relativní rizika, využívaná často v medicíně pro identifikaci a popis rizikových skupin pacientů.

Test dobré shody - základní teorie



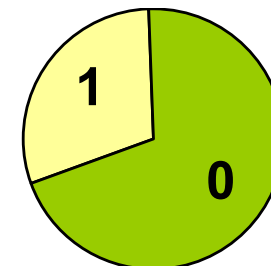
$$\chi^2_{(s.v.)} = \sum \frac{\left[\begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\text{očekávaná četnost}}$$

$$\chi^2_{(s.v.)} = \underbrace{\frac{\left[\begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\text{očekávaná četnost}}}_{\text{1. jev}} + \underbrace{\frac{\left[\begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\text{očekávaná četnost}}}_{\text{2. jev}} + \dots$$

Test dobré shody - základní teorie

Binomické jevy (1/0)

$$\chi^2_{(1)} = \frac{\left[\begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\underbrace{\text{očekávaná četnost}}_{\text{I. jev 1}}} + \frac{\left[\begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\underbrace{\text{očekávaná četnost}}_{\text{II. jev 2}}}$$



I. jev 1

II. jev 2

Příklad



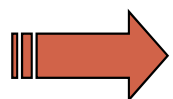
10 000 lidí hází mincí → rub: 4 000 případů (R)
líc: 6 000 případů (L)



Lze výsledek považovat za statisticky významně odlišný (nebo neodlišný) od očekávaného poměru R : L = 1 : 1 ?

$$\chi^2_{(1)} = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400$$

Tabulková hodnota: $\chi^2_{(0,95)} (\nu = 1) = \underline{\underline{3,84}} \quad (0,95 = 1 - \alpha)$



Rozdíl je vysoce statisticky významný ($p \ll 0,001$)

Test dobré shody: příklad I



? Ověřte na datech z pokusu se 100 květinami určitého druhu, že barva květů se geneticky štěpí v poměru žlutá : červená = 3 : 1.

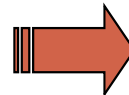
✓ H_0 : Pozorovaná frekvence pro jednotlivé barvy květů jsou vzorkem populace mající poměr mezi žlutými a červenými květy 3 : 1.

Součet frekvencí u obou barev květů (f_i) se rovná 100 a pozorované frekvence u kategorií barvy budou srovnány s očekávanými frekvencemi (uvedeny v závorkách):

	Kategorie barvy		n
	Žlutá	Červená	
$f_{\text{poz.}}$	84	16	100
$f_{\text{oček.}}$	75	25	

$$\chi^2 = \sum \frac{(f_{\text{poz.}} - f_{\text{oček.}})^2}{f_{\text{oček.}}} = \frac{(84 - 75)^2}{75} + \frac{(16 - 25)^2}{25} = 4,320$$

St. volnosti = $n = k - 1 = 1$



Zamítáme hypotézu shody srovnávaných četností

Při testování H_0 jsme použili matematický zápis ($0,025 < P < 0,05$). Z tabulek χ^2 rozložení vidíme, že pravděpodobnost překročení hranice 2,706 je 0,1 (10 %), což může být stručně zapsáno jako $P(\chi^2 \geq 2,706) = 0,10$.

Dále lze zjistit pro $P(\chi^2 \geq 3,841) = 0,05$. V řešené úloze jsme dospěli k hodnotě testové statistiky $\chi^2 = 4,320$. Pro tento případ lze tedy psát $0,025 < P(\chi^2 \geq 4,320) < 0,05$; a jednodušeji $0,025 < P < 0,05$. Jde v podstatě o přibližné určení hranic chyby 1. druhu.

Test dobré shody: příklad II



Tento příklad je rozšířením problému z příkladu 1 na srovnání pozorovaných a očekávaných frekvencí pro více kategorií sledovaného znaku:

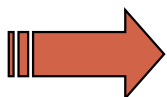


Celkem bylo zkoumáno 250 semen určitého druhu rostliny a roztríděno do následujících kategorií: žluté/hladké; žluté/vrásčité; zelené/hladké; zelené/vrásčité. Předpokládaný poměr výskytu těchto kategorií v populaci je 9 : 3 : 3 : 1. Následující tabulka obsahuje původní data z pozorování a dále postup při testování H_0 .

	žluté/hladké	žluté/vrásčité	zelené/hladké	zelené/vrásčité	n
$f_{\text{poz.}}$	152	39	53	6	250
$f_{\text{oček.}}$	140,6250	46,8750	46,8750	15,6250	

$$\nu = k - 1 = 3$$

$$\chi^2 = \frac{11,3750^2}{140,6250} + \frac{7,8750^2}{46,8750} + \frac{6,1250^2}{46,8750} + \frac{9,6250^2}{15,6250} = 8,972$$



Zamítáme hypotézu shody pozorovaných četností s očekávanými

Test dobré shody: příklad III

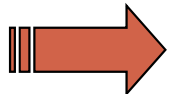
Složitější příklady řešené srovnáváním frekvencí je možné rozdělit na testování dílčích hypotéz:

- ✓ Předpokládejme, že chceme pro data z předchozí úlohy testovat hypotézu existence štěpného poměru 9 : 3 : 3 pro první tři kategorie semen:

	žluté/hladké	žluté/vrásčité	zelené/hladké	n
$f_{\text{poz.}}$	152	39	53	244
$f_{\text{oček.}}$	146,400	48,800	48,800	

$$n = k - 1 = 2$$

$$\chi^2 = \frac{5,600^2}{146,40} + \frac{9,800^2}{48,80} + \frac{4,200^2}{48,80} = 2,544$$



Nezamítáme hypotézu shody pozorovaných četností s očekávanými.

- ✓ Nyní otestujeme hypotézu štěpného poměru kategorií zelené/vrásčité:ostatní typy = 1:15

	zelené/vrásčité	ostatní	n
$f_{\text{poz.}}$	6	244	25
$f_{\text{oček.}}$	15,625	234,375	

$$n = k - 1 = 1$$

$$\chi^2 = \frac{9,625^2}{15,625} + \frac{9,625^2}{234,375} = 6,324$$



Zamítáme hypotézu shody pozorovaných četností s očekávanými.

Kontingenční tabulka I



- Máme dvě nominální veličiny, X (má r variant) a Y (má s variant)
- Kontingenční tabulka typu $r \times s$

$x_{[j]}$ \ $y_{[k]}$	$y_{[1]}$	$y_{[s]}$	$n_{j.}$
$x_{[1]}$	n_{11}	n_{1s}	$n_{1.}$
.
.
$x_{[r]}$	n_{r1}	n_{rs}	$n_{r.}$
$n_{.k}$	$n_{.1}$.	.	$n_{.s}$	n

- Označení:
 n_{jk} - simultánní absolutní četnost,
 $n_{j.}$ - marginální absolutní četnost

Kontingenční tabulka II



- Kontingenční tabulka umožňuje testování následujících hypotéz:
 1. **Hypotézu o nezávislosti,**
 2. **Hypotézu o shodnosti struktury (test homogeneity)**
 3. **Hypotézu o symetrii**

Testování nezávislosti I



- Motivace: Souvisí spolu výskyt dvou nominálních znaků měřených na jediném výběru?
- Příklad: Barva očí (modrá, zelená, hnědá) a barva vlasů (hnědá, černá, blond) u vybraných 30 studentů jsou nezávislé.
- Nulová hypotéza: Znaky X a Y jsou nezávislé náhodné veličiny.
- Alternativní hypotéza: Znaky X a Y jsou závislé náhodné veličiny.
- Test: **Pearsonův chí-kvadrát**

$$K = \sum_{j=1}^r \sum_{k=1}^s \frac{(n_{jk} - e_{jk})^2}{e_{jk}} \stackrel{\text{H}_0 \text{ platí}}{\approx} \chi^2((r-1)(s-1))$$

Očekávané (teoretické) četnosti e_{jk} : $e_{jk} = \frac{n_{j.} \cdot n_{.k}}{n}$

- H_0 zamítáme na hladině významnosti α , pokud $K \geq \chi_{1-\alpha}^2((r-1)(s-1))$
- **Předpoklady testu ?**

Testování nezávislosti II



- **Předpoklady Pearsonova chí-kvadrát testu:**

1. Jednotlivá pozorování shrnutá v kontingenční tabulce jsou nezávislá, tj. každý prvek patří jen do jedné buňky kont. tabulky, nemůže zároveň patřit do dvou.
2. **Podmínky dobré aproximace:** Očekávané (teoretické) četnosti jsou aspoň v 80 % případů větší nebo rovné 5 a ve 100 % případů nesmí být pod 2. (Pokud není tento předpoklad splněn, je vhodné sloučit kategorie s nízkými četnostmi).

- **Měření síly závislosti:**

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r, s\}$, V je z intervalu (0,1)

Význam hodnot: 0-0,1....zanedbatelná závislost

0,1-0,3...slabá závislost

0,3-0,7...střední závislost

0,7-1 silná závislost

Kontingenční tabulky

H0 :Nezávislost dvou jevů A a B



**Kontingenční
tabulka
2 x 2**

<div style="display: flex; align-items: center; justify-content: center;"> <div style="border-right: 1px solid black; padding-right: 5px;"> <div style="border-bottom: 1px solid black; padding-bottom: 5px;"> <div style="display: flex; align-items: center;"> <div style="border-right: 1px solid black; padding-right: 5px;">↓ B</div> <div style="padding-left: 5px;">→ A</div> </div> </div> </div> </div>	+	-	Podíl (+)
+	a	b	$\frac{a}{(a + b)}$ p₁
-	c	d	$\frac{c}{(c + d)}$ p₂
Podíl (+)	$\frac{a}{(a + c)}$	$\frac{b}{(b + d)}$	

$$N = a + b + c + d$$

$$P(B^+) = \frac{(a + b)}{N}$$

$$P(B^-) = \frac{(c + d)}{N}$$

Očekávané četnosti:

$$F_{(A)} = \frac{(a+b)(a+c)}{N}$$

$$F_{(C)} = \frac{(a+c)(d+c)}{N}$$

$$\chi^2_{\nu=1} = \sum_{i=1}^4 \frac{(f_i - F_i)^2}{F_i}$$

$$F_{(B)} = \frac{(a+b)(b+d)}{N}$$

$$F_{(D)} = \frac{(b+d)(c+d)}{N}$$

$$\nu = 1 = (r-1) * (c-1)$$

$$\chi^2_c = \sum \sum \frac{(|f_{ij} - F_{ij}| - 0,5)^2}{F_{ij}}$$

$$P_{(A)}; P_{(B)}$$

Kontingenční tabulky: příklad

gen \ †	Ano	Ne	Σ
Ano	20	82	102
Ne	10	54	64
Σ	30	136	166

$$F_A = 102 * 30 / 166 = 18,43$$

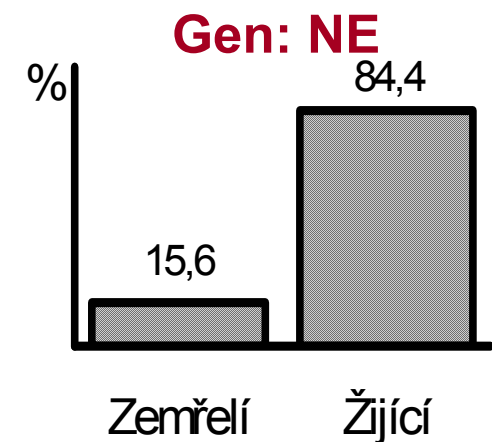
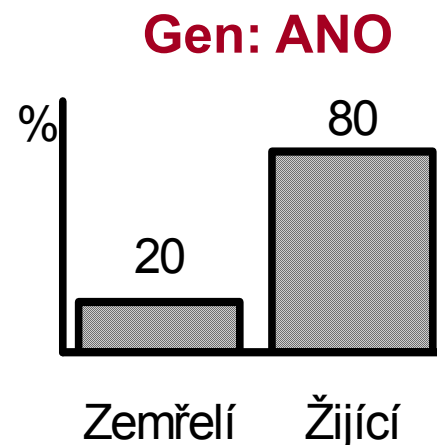
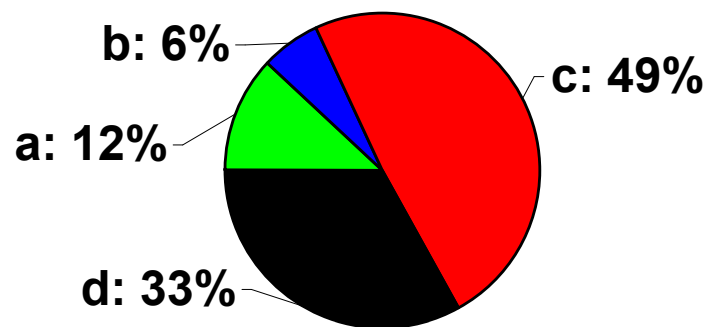
$$F_B = 102 * 136 / 166 = 83,57$$

$$F_C = 11,57$$

$$F_D = 52,43$$

$$\chi^2_{(1)} = \frac{(20-18,43)^2}{18,43} + \frac{(82-83,57)^2}{83,57} + \frac{(10-11,57)^2}{11,57} + \frac{(54-52,43)^2}{52,43} = 0,423 \quad 0,423 < \chi^2_{0,95}^{(1)} = 3,84$$

Kontingenční tabulka v obrázku



Příklad 1: Řešení v softwaru Statistica I



- Na hladině významnosti 0,05 testujte hypotézu o nezávislosti genu a stavu pacienta. Simultánní četnosti znázorněte graficky.

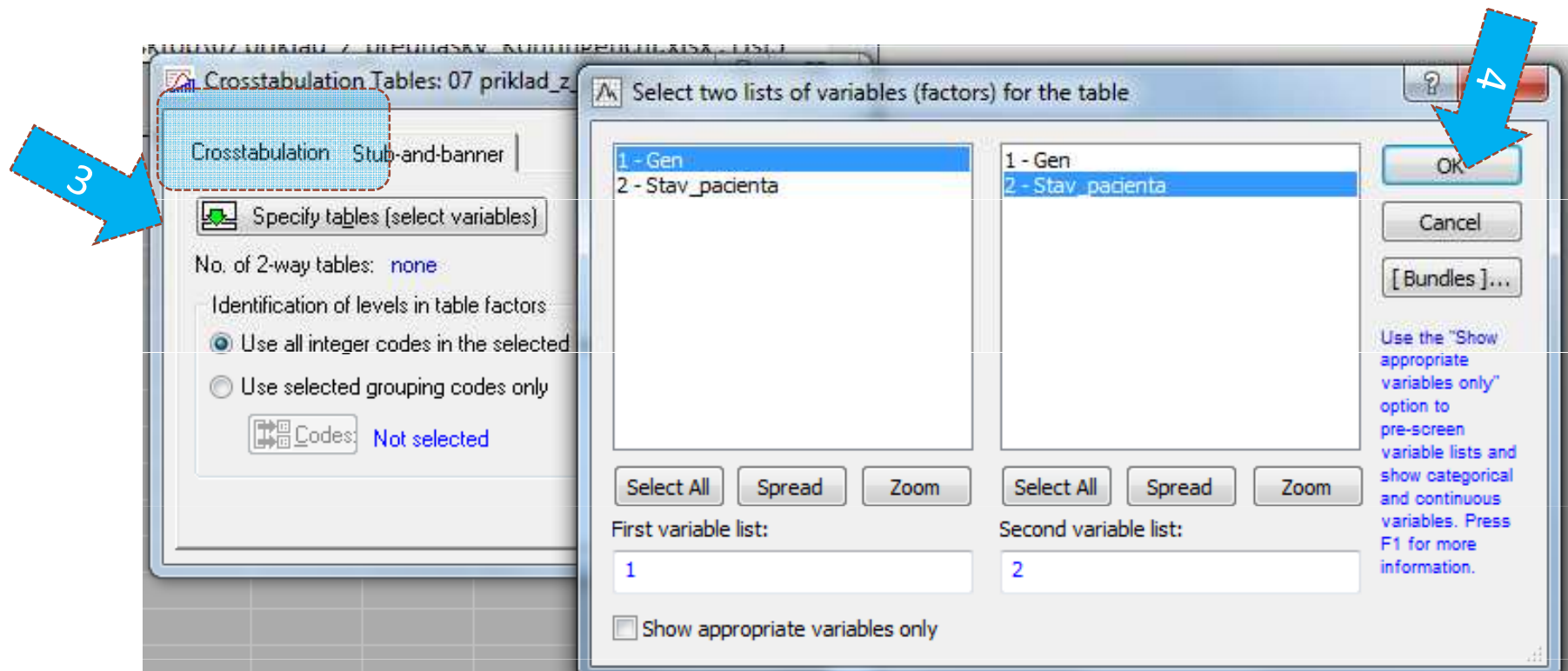
- V menu **Statistics** zvolíme **Basic statistics**,
Vybereme **Tables and banners**

The screenshot shows the Statistica software interface. The 'Statistics' menu is open, and the 'Basic statistics' option is highlighted with a red dashed box. A blue arrow points to the 'Basic statistics' icon. Below the menu, the 'Basic Statistics and Tables' dialog box is open, and the 'Tables and banners' option is selected, also highlighted with a red dashed box and a blue arrow. The background shows a data table with columns 'Gen' and 'Stav_p'.

	1 Gen	2 Stav_p
1	přítomer	úmrtí
2	přítomer	úmrtí
3	přítomer	úmrtí
4	přítomer	úmrtí
5	přítomer	úmrtí
6	přítomer	úmrtí
7	přítomer	úmrtí
8	přítomer	úmrtí
9	přítomer	úmrtí
10	přítomer	úmrtí
11	přítomer	úmrtí
12	přítomer	úmrtí

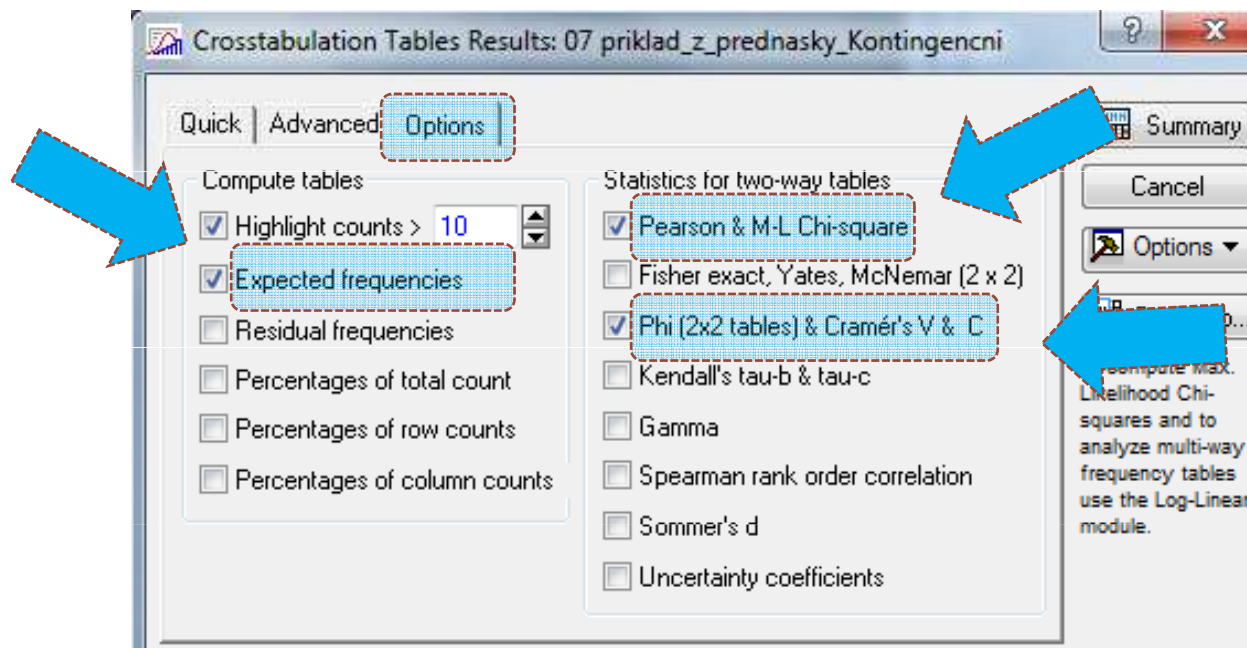
Příklad 1: Řešení v softwaru Statistica II

- Vybereme proměnné, které chceme testovat



Příklad 1: Řešení v softwaru Statistica III

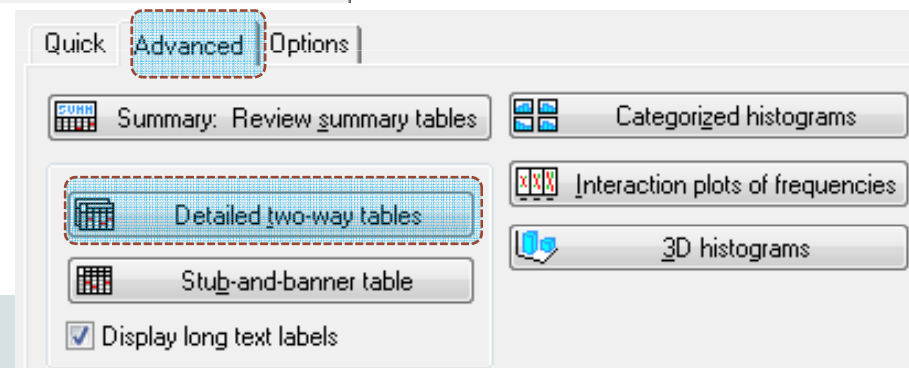
- Na záložce **Options** zaškrtneme **Expected frequencies (Očekávané četnosti)** (k ověření podmínek dobré aproximace)



- Zaškrtneme Pearsonův chí-kvadrát

- Pokud chceme vypočítat i Cramérův koeficient zaškrtneme Phi & Cramer's V

- Poté se vrátíme na záložku **Advanced**, kde a zvolíme **Detailed two-way tables**



Příklad 1: Řešení v softwaru Statistica IV

Tab.1: Pozorované četnosti

Summary Frequency Table (07 priklad_z_prednasky_K
Marked cells have counts > 10
(Marginal summaries are not marked)

Gen	Stav_pacienta úmrtí	Stav_pacienta žijící	Row Totals
přítomen	20	82	102
nepřítomen	10	54	64
All Grps	30	136	166

Tab. 2: Očekávané četnosti

Summary Table: Expected Frequencies (07 priklad_z_pre
Marked cells have counts > 10
Pearson Chi-square: ,421322, df=1, p=,516278

Gen	Stav_pacienta úmrtí	Stav_pacienta žijící	Row Totals
přítomen	18,43373	83,5663	102,0000
nepřítomen	11,56627	52,4337	64,0000
All Grps	30,0000	136,0000	166,0000



Jsou splněny podmínky dobré aproximace? ANO

Tab. 3: Paersonův chí-kvadrát

Hodnota testové statistiky Počet stupňů volnosti p- hodnota

Statistic	Chi-square	df	p
Pearson Chi-square	,4213223	df=1	p=,51628
M-L Chi-square	,4277117	df=1	p=,51311
Phi for 2 x 2 tables	,0503794		
Tetrachoric correlation	,0949754		
Contingency coefficient	,0503156		

R x C kontingenční tabulka



Výběr: N lidí ze sociologického průzkumu (delikventi)

Jev **A**: Původ z rozvrácených rodin

Jev **B**: Stupeň zločinnosti I < II < III < IV

A \ B	I.	II.	III.	IV.	Σ
ANO	a	b	c	d	číslo 1
NE	e	f	g	h	
Σ	číslo2				

Stupně volnosti:

$$(R-1) * (C-1) = 1 * 3 = 3$$

$$F_a = \frac{\text{číslo 1} \cdot \text{číslo 2}}{N}$$

Tabulky: $\chi^2_{(1-\alpha)}^{(v)}$

Očekávané četnosti:

$$p_a = \frac{a}{a + e}$$

$$p_b = \frac{b}{b + f}$$

$$p_c = \frac{c}{c + g}$$

$$p_d = \frac{d}{d + h}$$

Čtyřpolní tabulky



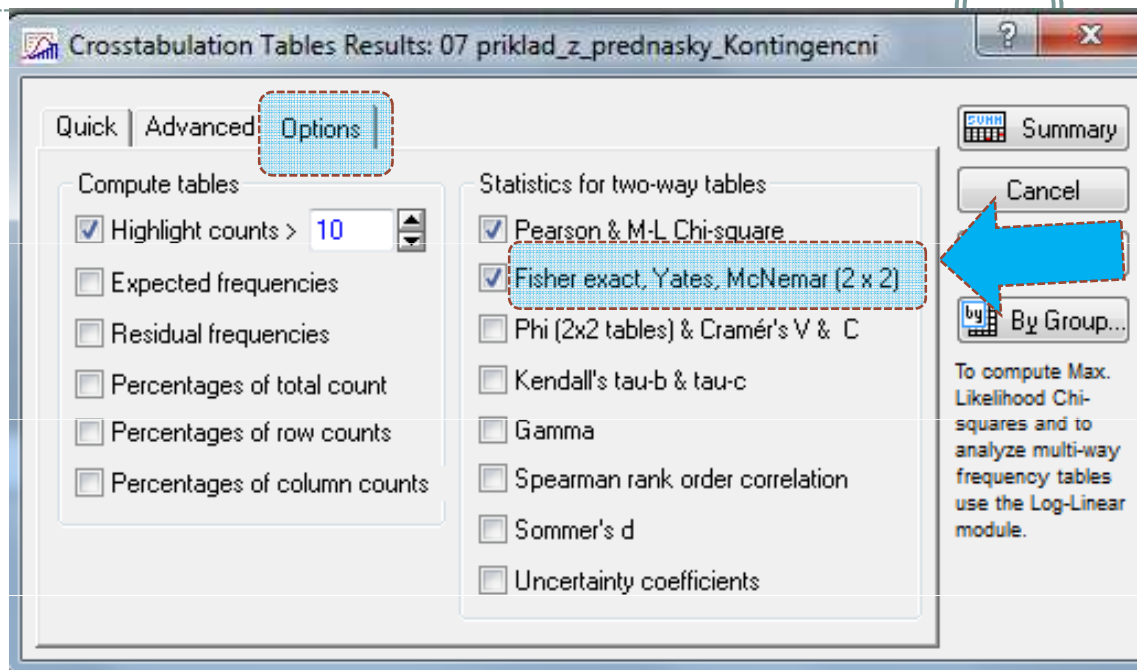
- Máme dvě nominální veličiny, X (má dvě varianty) a Y (má dvě varianty)
- Kontingenční tabulka typu **2 x 2**

	Y _[1]	Y _[2]	n _{j.}
X _[1]	a	b	a+b
X _[2]	c	d	c+d
n _{.k}	a+c	b+d	n

Výsledek pokusu	okolnosti		n _{j.}
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
n _{.k}	a+c	b+d	n

- Definice: **podíl šancí (odds ratio)** $OR = \frac{ad}{bc}$
 Jestliže asymptotický 100(1-α)% interval spolehlivosti $\ln OR \pm \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{1-\alpha/2}$ neobsahuje 1, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.
- Test: **Fisherův přesný (exaktní) test (slouží též k testování v tabulce r x s, když nemáme splněny podmínky dobré aproximace)**

Řešení v softwaru Statistica: Fisherův přesný test



- Na záložce **Options** zaškrtneme **Fisher exact**

- Výstupní tabulka

Statistic	Statistics: Gen(2) x Stav_paci		
	Chi-square	df	p
Pearson Chi-square	,4213223	df=1	p=,51628
M-L Chi-square	,4277117	df=1	p=,51311
Yates Chi-square	,1952605	df=1	p=,65857
Fisher exact, one-tailed			p=,33259
two-tailed			p=,54314
McNemar Chi-square (A/D)	14,71622	df=1	p=,00012
(B/C)	54,79348	df=1	p=,00000

Pro jednostranný test

Pro oboustranný test



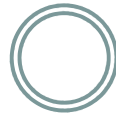
Testování homogenity (testování shody struktury)

- Motivace: Zajímá nás výskyt nominálního znaku u r nezávislých výběrů z r různých populací.
- Příklad: Je zájem o sport stejný u děvčat jako u chlapců?
- Nulová hypotéza: pravděpodobnostní rozdělení kategoriální proměnné je stejné v různých populací
- Test: **Pearsonův chí-kvadrát**

		Dívky	Chlapci	
Zájem o sport	Ano	a	b	$a+b$
	Ne	c	d	$c+d$
		$a+c$	$b+d$	n

Některé marginální četnosti (buď sloupcové nebo řádkové) jsou předem pevně stanoveny

Testování homogenity: příklad I



Očkování proti chřipce se zúčastnilo 460 dospělých, z nichž 240 dostalo očkovací látku proti chřipce a 220 dostalo placebo. Na konci experimentu onemocnělo 100 lidí chřipkou, 20 z nich bylo z očkované skupiny a 80 z kontrolní skupiny. Je to dostatečný důkaz, že očkovací látka byla účinná?

Nulová hypotéza: Procento výskytu chřipky je v očkované a kontrolní skupině stejné.

Test hypotézy o symetrii (McNemarův test pro čtyřpolní tabulku)



- Motivace: Na osobách sledujeme binární proměnnou před pokusem a po něm, cílem je zjistit, zda došlo ke změně v rozdělení této proměnné.
- **Analýza párových dichotomických proměnných**

Četnostní tabulka

		po		$n_{j.}$
		+	-	
před	+	a	b	$a+b$
	-	c	d	$c+d$
$n_{.k}$		$a+c$	$b+d$	n

Tabulka teoretických pravděpodobností

		po		
		+	-	
před	+	p_{11}	p_{12}	$p_{1.}$
	-	p_{21}	p_{22}	$p_{2.}$
		$p_{.1}$	$p_{.2}$	

- Nulová hypotéza: $p_{ij} = p_{ji}$, pokus nemá vliv na výskyt daného znaku
- Testová statistika: $\chi^2 = \frac{(b - c)^2}{b + c}$ pokud je větší než kritická hodnota χ^2 rozdělení o jednom stupni volnosti (vhodné pro počty údajů $b+c > 8$), pak nulovou hypotézu zamítáme

Mc Nemarrův test: příklad I



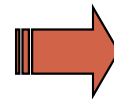
Zjistěte, zda výuka o pozitivním působení sportu na zdraví vede ke změně postojů žáků ke sportování.

Nulová hypotéza: Počet žáků, kteří změní svůj postoj pozitivním směrem, je pouze náhodně odlišný od počtu žáků, kteří změní svůj postoj negativním směrem.

		Postoj po výuce		
		+	-	
Postoj před výukou	+	5	3	8
	-	16	2	18
		21	5	26

$$\chi^2 = \frac{(3 - 16)^2}{3 + 16} = 8,89$$

Tabulky: $\chi^2_{1-\alpha}(v = 1) = 3,84$



H₀ zamítnuta

Závěr: Výuka má pozitivní vliv na postoj žáků vzhledem k provozování sportu.