VIII. KONTINGENČNÉ TABUĽKY

Anotácia

Analýza kontingenčných tabuliek umožňuje analyzovať väzbu medzi dvoma kategoriálnymi premennými. Základným spôsobom testovania je tzv. chi-square test, ktorý porovnáva pozorované četnosti kombinácií kategórií oproti očakávaným četnostiam, ktoré vychádzajú z teoretickej situácie, kedy je vzťah medzi premennými náhodný.

Test dobrej zhody je využívaný tiež na porovnanie pozorovaných četností oproti očakávaným četnostiam daných určitým pravidlom (typickým príkladom je Hardy-Weinbergova rovnováha v genetike)

Špecifickým typom výstupov odvodených z kontingenčných tabuliek sú tzv. odds ratia a relatívne riziká, využívané často v medicíne na identifikáciu a popis rizikových skupín pacientov.

Test dobrej zhody - základná teória

$$\chi^2_{(s.v.)} = \sum \frac{\left[\text{pozorovaná četnosť} - \text{očakávaná četnosť}\right]^2}{\text{očakávaná četnosť}}$$

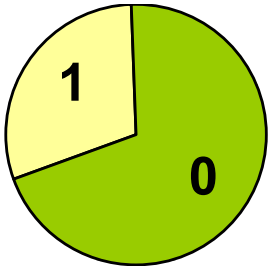
$$\chi^2_{(s.v.)} = \underbrace{\frac{\left[\text{pozorovaná četnosť} - \text{očakávaná četnosť}\right]^2}{\text{očakávaná četnosť}}}_{\text{1. jav}} + \underbrace{\frac{\left[\text{pozorovaná četnosť} - \text{očakávaná četnosť}\right]^2}{\text{očakávaná četnosť}}}_{\text{2. jav}} + \ldots$$

Test dobrej zhody - základná teória

Binomické javy (1/0)

$$\chi^2_{(1)} = \frac{\left[\text{pozorovaná četnosť} - \text{očakávaná četnosť}\right]^2}{\text{očakávaná četnosť}} + \frac{\left[\text{pozorovaná četnosť} - \text{očakávaná četnosť}\right]^2}{\text{očakávaná četnosť}}$$

I. jav 1 | II. jav 2



Príklad ✓ 10 000 ľudí hádže mincou → rub: 4 000 prípadov (R); líce: 6 000 prípadov (L)

? Výsledok je možné považovať za štatisticky významne odlišný (alebo neodlišný) od očakávaného pomeru R : L = 1 : 1 ?

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

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

⇒ **Rozdiel je vysoko štatisticky významný (p << 0,001]**

Kontingenčné tabuľky
H0 :Nezávislosť dvoch jevov A a B

Kontingenčná tabuľka 2 x 2

B \ A	+	-	Podiel (+)
+	a	b	$\frac{a}{(a+b)}$ p_1
-	c	d	$\frac{c}{(c+d)}$ p_2
Podiel (+)	$\frac{a}{(a+c)}$	$\frac{b}{(b+d)}$	

N = a + b + c + d

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

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

Očakávané četnosti:

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

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

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

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

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

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

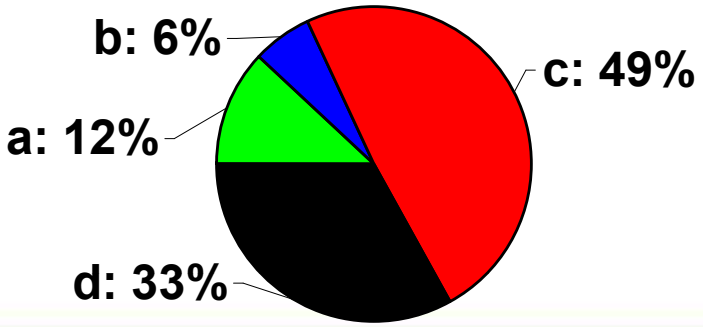
Kontingenčné tabuľky: príklad

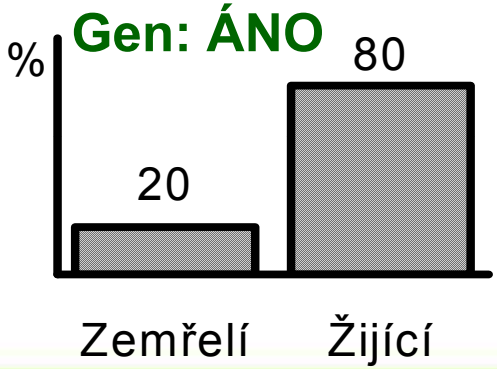
gén / †	Áno	Nie	Σ
Áno	20	82	102
Nie	10	54	64
Σ	30	136	166

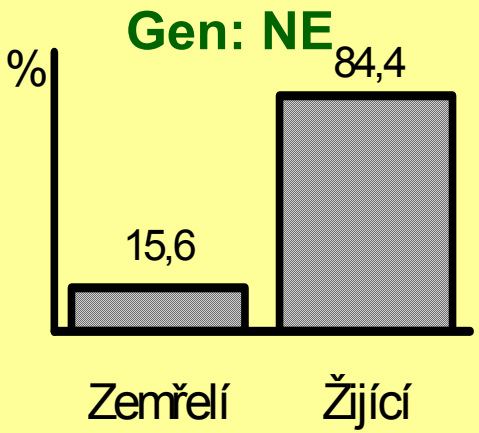
$F_A = 102 * 30 / 166 = 18,43$
$F_B = 102 * 136 / 166 = 83,57$
$F_C = 11,57$
$F_D = 52,43$

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

Kontingenčná tabuľka v obrázku







R x C kontingenčná tabuľka

Výber: N ľudí zo sociologického prieskumu (delikventi)
Jav **A**: Pôvod z rozvrátených rodín
Jav **B**: Stupeň zločinnosti I < II < III < IV

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

Stupne voľnosti:
(R-1) * (C-1) = 1 * 3 = 3

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

Tabuľky: $\chi^2_{(1-\alpha)}{}^{(\nu)}$

Očakávané četnosti:

$$p_a = \frac{a}{a+e} \qquad p_b = \frac{b}{b+f}$$

$$p_c = \frac{c}{c+g} \qquad p_d = \frac{d}{d+h}$$

Test dobrej zhody: príklad I

? **Ověřte na datech z pokusu se 100 květinkami 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		
	Žlutá	Červená	n
$f_{poz.}$	84	16	100
$f_{oček.}$	75	25	

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

St. voľnosti = n = k - 1 = 1

Zamietame hypotézu zhody porovnávaných četností

Pri testovaní H_0 sme použili matematický zápis (0,025 < P < 0,05). Z tabuliek $\chi 2$ rozložení vidíme, že pravdepodobnosť prekročenia hranice 2,706 je 0,1 (10 %), čo môže byť stručne zapísané ako P ($\chi 2 \geq 2{,}706$) = 0,10.
Ďalej je možné zistiť pre P ($\chi 2 \geq 3{,}841$) = 0,05. V riešenej úlohe sme dospeli k hodnote testovej štatistiky $\chi 2$ = 4,320. Pre tento prípad môžme teda písať 0,025 < P ($\chi 2 \geq 4{,}320$) < 0,05; a jednoduchšie 0,025 < P < 0,05. Ide v podstate o približné určenie hraníc chyby 1. druhu.

Test dobrej zhody: príklad II

Tento príklad je rozšírením problému z príkladu 1 na porovnanie pozorovaných a očakávaných frekvencí pre viac kategórií sledovaného znaku:

✔ Celkem bylo zkoumáno 250 semen určitého druhu rostliny a roztříděno 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
$f_{poz.}$	152	39	53	6	250
$f_{oček.}$	140,6250	46,8750	46,8750	15,6250	

v = 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 dobrej zhody: príklad III

Zložitejšie príklady riešené porovnávaním frekvencií je možné rozdeliť na testovanie dielčích hypotéz:

✓ Predpokladajme, že chceme pre dáta z predchádzajúcej úlohy testovať hypotézu existencie štiepneho pomeru 9 : 3 : 3 pre prvé tri kategórie semien:

	žluté/hladké	žluté/vrásčité	zelené/hladké	n
$f_{poz.}$	152	39	53	244
$f_{ček.}$	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.

✓ Teraz otestujeme hypotézu štiepneho pomeru kategórií zelené/vrásčité: ostatné typy = 1:15

	zelené/vrásčité	ostatní	n
$f_{poz.}$	6	244	25
$f_{ček}$	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.

Test dobrej zhody: príklad IV - využitie aditivity testu

✓ U 193 párů dvojčat byly zjištěny následující poměry pohlaví: **56 Ch - Ch**
72 Ch – H
65 H - H

? ***Za predpokladu, že narodenie chlapčeka má rovnakú pravdepodobnosť ako narodenie dievčatka, môžme očakávať pomery pre výššie uvedené skupiny = 0,25 : 0,5 : 0,25. Overte tento predpoklad na uvedenom vzorku populácie.***

Σ 193 párov 1/4 : 1/2 : 1/4
očakávané četnosti = 48,25 : 96,50 : 48,25

$$\chi^2_{(2)} = 13,28$$

Prečo môžeme v predchádzajúcom prípade očakávať zamietnutie H_0?

Testujte nasledujúce hypotézy:
1) Sú relatívne počty párov so zhodným pohlavím v zhode s očakávanými četnosťami? (ignorujte Ch – H páry)
2) Je relatívna četnosť kombinácie Ch - Ch a H - H párov oproti párom s rozdielnym pohlavím v zhode s očakávanými četnosťami?

Σ 121 párů 1 : 1
očekávané četnosti = 60,5 : 60,5

$$\chi^2_{(1)} = 0,669$$

$$\frac{H - H}{Ch - Ch}$$

Σ 193 párů 1 : 1
očekávané četnosti = 96,5 : 96,5

$$\chi^2_{(1)} = 12,44$$

Test dobrej zhody: príklad V

Města - zatížení exhalacemi - třídy (A > B > C > D)

Svět: A : B : C : D = 2 : 3 : 6 : 4

Konkrétní země (n = 184 měst): A : B : C : D = 32 : 151 : 182 : 116

H_0: shoda f_i a F_i **a = 0,05**

F_A: 64,13 F_C: 192,39

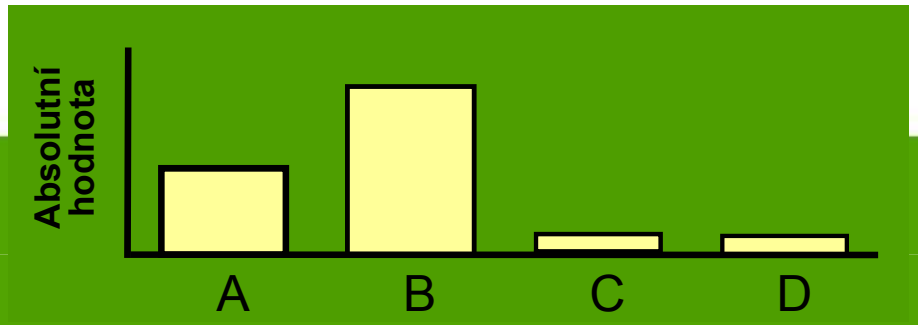
F_B: 96,19 F_D: 128,27

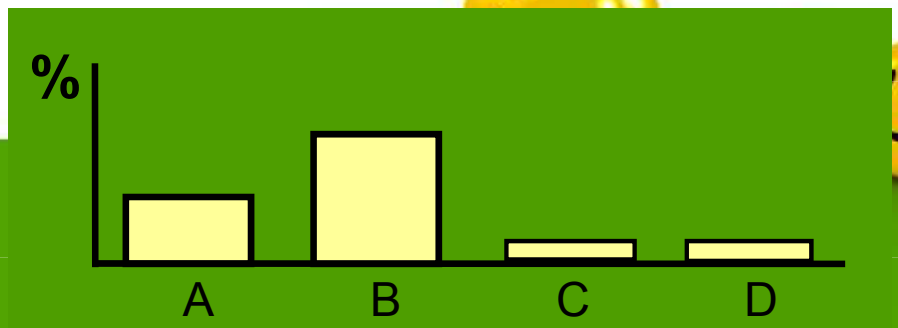
$$\chi^2_{(3)} = \frac{(32 - 64,13)^2}{64,13} + \ldots\ldots\ldots + \frac{(116 - 128,27)^2}{128,27} = \underline{\underline{49,06}}$$

$$Tabulky : \chi^{2\,(\nu)}_{1-\alpha} = \chi^{2\,(3)}_{0,95} = 7,81$$

Zamítáme hypotézu shody pozorovaných četností s očekávanými.

Příspěvek kategorií A, B, C, D k celkové hodnotě $\chi 2$





c2 test - příklad složitější kontingenční tabulky I

Caffeine consumption and marital status in antenatal patiens (from Martin and Bracken, 1987)

	Caffeine consumption (mg/day)				
Marital status	**0**	**1 - 150**	**151 - 300**	**> 300**	**Total**
Married	652	1537	598	242	3029
Divorced, separed or widowed	36	46	38	21	141
Single	218	327	106	67	718
Total	906	1910	742	330	3888

Caffeine consumption and marital status data

	Caffeine consumption (mg/day)				
Marital status	**0**	**1 - 150**	**151 - 300**	**> 300**	**Total**
Married	22 %	51 %	20 %	8 %	3029 (100 %)
Divorced, separed or widowed	26 %	33 %	27 %	15 %	141 (100 %)
Single	30 %	46 %	15 %	9 %	718 (100 %)
Total	23 %	49 %	19 %	8 %	3888 (100 %)

c2 test - příklad složitější kontingenční tabulky II

Expected frequencies

	Caffeine consumption (mg/day)				
Marital status	**0**	**1 - 150**	**151 - 300**	**> 300**	**Total**
Married	705,8	1488	578,1	257,1	3029
Divorced, separed or widowed	32,9	69,3	26,9	12,0	141
Single	167,3	352,7	137	60,9	718
Total	906	1910	742	330	3888

Contributions of each cell

	Caffeine consumption (mg/day)				
Marital status	**0**	**1 - 150**	**151 - 300**	**> 300**	**Total**
Married	4,11	1,61	0,69	0,89	7,30
Divorced, separed or widowed	0,30	7,82	4,57	6,82	19,51
Single	15,36	1,88	7,02	0,60	24,86
Total	19,77	11,31	12,28	8,31	51,66

c2 test - příklad frakcionace složitější kontingenční tabulky I

Cílem rozsáhlejšího průzkumu populace bylo prozkoumat vztah mezi dvěma typy chorob a krevními skupinami u lidí. Konkrétní data jsou uvedena v tabulce:

Krevní skupina	Žaludeční vředy	Rakovina žaludku	Kontrola	Celkem
0	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
Celkem	1796	883	6087	8766

Vypočítejte testovou charakteristiku pro tuto kontingenční tabulku a otestujte nulovou hypotézu nezávislosti jevů ($\chi 2 = 40{,}54$; 4 st. volnosti)

c2 test - příklad frakcionace složitější kontingenční tabulky II

K podrobnějšímu průzkumu složitějších tabulek výrazně napomáhá přepis původní tabulky do podoby procentického zastoupení kategorií:

Krevní skupina	Žaludeční vředy	Rakovina žaludku	Kontrola
0	983	383	2892
A	679	416	2625
B	134	84	570
Celkem	1796	883	6087

Z této tabulky je patrné:

1. **Jsou jenom malé rozdíly v distribuci krevních skupin u kontroly a u skupiny nemocných rakovinou žaludku.**

2. **Pacienti s vředy mají mnohem častěji krevní skupinu 0.**

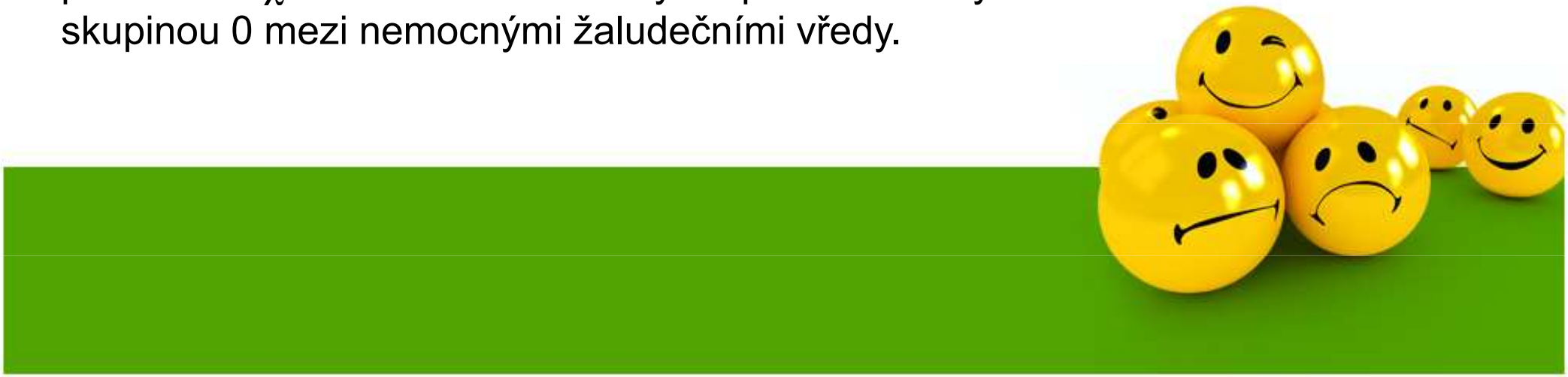
Na základě těchto poznatků je možné sestrojit menší kontingenční tabulku, která otestuje hypotézu o shodné distribuci krevních skupin pro nemocné rakovinou a pro zdravé lidi.

Sestavte tuto tabulku a otestujte nulovou hypotézu.
($\chi 2$ = 5,64 (2 st. v.), P je přibližně rovna 0,06)

c2 test - příklad frakcionace složitější kontingenční tabulky III

- Z tohoto dílčího testu vyplývá možnost sloučení skupiny nemocných rakovinou a zdravých lidí neboť se vzhledem k distribuci krevních skupin chovají jako homogenní populace. Dalším logickým krokem v podrobné analýze je testování shody relativních četností výskytu krevních skupin A a B mezi kombinovaným vzorkem (sloučená skupina s rakovinou a kontrola) a mezi vzorkem lidí nemocných žaludečními vředy - tzn. nyní neuvažujeme krevní skupinu 0. Výsledkem tohoto testu je $\chi 2 = 0{,}68$ (1 st. vol.); $P > 0{,}7$. Vzorky pro krevní skupiny A a B lze tedy sloučit do směsného vzorku A + B.

- Nyní otestujeme shodu relativních četností výskytu skupiny 0 oproti A + B, a to mezi kombinovanou populací (kontrola + nemocní rakovinou) a mezi vzorkem nemocných vředařů ($\chi 2 = 34{,}29$; 1 st. vol.). Lze tedy shrnout, že vysoká hodnota původního $\chi 2$ se 4 st. volnosti byla způsobena zvýšenou četností lidí s krevní skupinou 0 mezi nemocnými žaludečními vředy.



c2 test - příklad frakcionace složitější kontingenční tabulky IV

Průběh hodnocení lze shrnout do tabulky:

Srovnání	St. volnosti	χ^2
0, A, B skupina u pacientů s rakovinou (**r**) x kontrola (**k**)	2	5,64
A, B skupina u pacientů s vředy x kombinovaný vzorek (**r + k**)	1	0,68
0, A, B skupina u pacientů s **s** vředy x kombinovaný vzorek (**r + k**)	1	34,29
Celkem	**4**	**40,61**

Celkový součet testových statistik χ^2 (40,61) odpovídá přibližně původní hodnotě χ^2 (40,54). Což platí i o stupních volnosti (4). Tato skutečnost potvrzuje, že jsme detailním rozborem vyčerpali informační obsah původní kontingenční tabulky a kromě popsané závislosti (zvýšený výskyt krevní skupiny 0 u lidí s žaludečními vředy) jsou jednotlivé kategorie zkoumaných jevů zcela nezávislé.

Kontingenčná tabuľka 2 x 2: Riešenie pri nedostatočnej veľkosti vzorky

Yates' corection | **Fisher's exact test**

✓ H_0: Nezávislost jevů

Test analyzuje všetky možné 2 x 2 tabuľky, ktoré dávajú rovnakú sumu riadkov a stĺpcov ako tabuľka zdrojová.
Algoritmus každej tabuľke priraďuje pravdepodobnosť, že taká situácia nastane, ak je H_0 pravdivá.

Spectacle wearing among juvenile delinquents and non-delinquents who failed a vision test (Weindling et al., 1986)

		Juvenile delinquents	Non- deliquents	Total
Spectacle wearers	**Yes**	1	5	6
	No	8	2	10
	Total	9	7	16

Kontingenčná tabuľka 2 x 2: Riešenie pri nedostatočnej veľkosti vzorky

Všetkny možné varianty tabuľky s danou sumou riadkov a stĺpcov

(I)

0	6
9	1

(II)

1	5
8	2

(III)

2	4
7	3

(IV)

3	3
6	4

(V)

4	2
5	5

(VI)

5	1
4	6

(VII)

6	0
3	7

Pravdepodobnosť náhodného vzniku variant tabuľky

	a	b	c	d	P
(I)	0	6	9	1	0,00087
(II)	1	5	8	2	0,02360
(III)	2	4	7	3	0,15734
(IV)	3	3	6	4	0,36713
(V)	4	2	5	5	0,33042
(VI)	5	1	4	6	0,11014
(VII)	6	0	3	7	0,01049
				Total	0,99999

2 x 2 frekvenčná tabuľka pre párové usporiadanie: Mc Nemar's test

Príklad: Porovnanie 2 metód stanovenia antigénu v krvi (antigén vždy prítomný)

✓ **H_0: metoda 1 = metoda 2**

Metoda 1	Metoda 2	Frekvence
úspěch	úspěch	202
úspěch	neúspěch	60
neúspěch	úspěch	42
neúspěch	neúspěch	10

} $\sum = 102$

$$\chi^2_{(c)} = \frac{(|60 - 42| - 1)^2}{102} = 2,83$$

H_0 nezamietnutá

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