

Čtyřpolní tabulky – v R: **table()**, **prop.table()**, **ctable()**, **matrix()**, **xtabs()**

Definice čtyřpolní tabulky je zřejmá – je to nejjednodušší možná kontingenční tabulka, kdy obě sledované náhodné veličiny mají pouze dvě varianty, kterých mohou nabývat. Stejně jako v případě obecné kontingenční tabulky můžeme pomocí statistických metod rozhodovat o statistické závislosti dvou sledovaných veličin.

Příklad čtyřpolní tabulky:

X/Y	y1	y2	Celkem
x1	a	b	a+b
x2	c	d	c+d
Celkem	a+c	b+d	a+b+c+d

Při rozhodování o nezávislosti ve čtyřpolní tabulce můžeme samozřejmě použít Pearsonův chí-kvadrát test, neboť tento test lze použít na jakoukoliv kontingenční tabulku, nicméně u tohoto testu je nutné hlídat jeho předpoklady: 80 % očekávaných četností, větších než 5. V případě čtyřpolní tabulky teda ve všech buňkách.

Nedodržení předpokladů pro Pearsonův chí-kvadrát test může vést k nesmyslným závěrům. Situace s malými pozorovanými, a tedy i očekávanými četnostmi jsou ale bohužel relativně časté, a to samé platí i pro čtyřpolní tabulky. Zlatým standardem pro hodnocení čtyřpolních tabulek se proto stal jiný test, tzv. **Fisherův exaktní test (Fisher exact test)**, který je založen na výpočtu přesné (exaktní) pravděpodobnosti, se kterou bychom za platnosti nulové hypotézy o nezávislosti veličin X a Y získali naši konkrétní realizaci čtyřpolní tabulky.

Fisherův exaktní test – v R: **fisher.test()**

K pochopení Fisherova exaktního testu vycházíme ze čtyřpolní kontingenční tabulky:

Nulovou hypotézou je v případě Fisherova testu *nezávislost sledovaných veličin X a Y*. Hlavní myšlenkou Fisherova exaktního testu je výpočet pravděpodobnosti, se kterou bychom získali čtyřpolní tabulky stejně nebo více vzdálené od nulové hypotézy při zachování pozorovaných marginálních četností. Zachování marginálních četností znamená, že se soustředíme pouze na situace, které odpovídají stejným četnostem jednotlivých variant náhodných veličin.

Pravděpodobnost získání konkrétního výsledku čtyřpolní tabulky s danými marginálními četnostmi lze vypočítat pomocí vzorce

$$p = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)! (a+c)! (c+d)! (b+d)!}{n! a! b! c! d!}$$

Výpočet testové statistiky potom probíhá následovně: spočítáme pravděpodobnosti p^* , příslušné všem možným tabulkám, které lze získat při zachování marginálních četností. Výsledná **testová statistika, respektive p-hodnota**, Fisherova exaktního testu *je součtem pravděpodobností p^* menších nebo stejných jako hodnota p , která přísluší čtyřpolní tabulce sestavené na základě pozorovaných hodnot*. Sčítáme tak pravděpodobnosti možností, které jsou více nebo stejně vzdáleny od nulové hypotézy,

jinými slovy tedy představují extrémnější nebo stejně extrémní variantu výsledku. Z výpočetního postupu je vidět, že Fisherův exaktní test není úplně standardním testem, neboť roli testové statistiky zde, na rozdíl od všech předchozích testů, hraje přímo p-hodnota. Tu potom pro rozhodnutí o platnosti nulové hypotézy srovnáme se zvolenou hladinou významnosti testu α , je-li p-hodnota testu menší než zvolené α , zamítáme nulovou hypotézu o nezávislosti veličin X a Y .

Příklad:

Uvažujeme 60 pozorování s tím, že tentokrát budeme zjišťovat, zda veličina X souvisí s veličinou Y . Pomocí Fisherova exaktního testu chceme testovat nulovou hypotézu o nezávislosti těchto veličin. Pozorovaná data, respektive pozorovanou čtyřpolní tabulku představuje následující tabulka:

X/Y	y1	y2	Celkem
x1	a = 11	b = 31	a+b = 42
x2	c = 6	d = 12	c+d = 18
Celkem	a+c = 17	b+d = 43	a+b+c+d = 60

Pravděpodobnost příslušná pozorované čtyřpolní tabulce p je následující

$$p = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)! (a+c)! (c+d)! (b+d)!}{n! a! b! c! d!} = \frac{42! 17! 18! 43!}{60! 11! 31! 6! 12!} = 0,205$$

Dále **vypočítáme pravděpodobnosti p^*** , pro jednotlivé možnosti kontingenční tabulky se zachováním marginálních četností, tedy se zachováním řádkových a sloupcových součtů. Výsledek zobrazuje následující tabulka:

Pravděpodobnosti p^* příslušné jednotlivým možnostem kontingenční tabulky

Možnost	a	b	c	d	p^*
1.	0	42	17	1	$4,6 \times 10^{-14}$
2.	1	41	16	2	$1,7 \times 10^{-11}$
3.	2	40	15	3	$1,8 \times 10^{-9}$
4.	3	39	14	4	$9,1 \times 10^{-8}$
5.	4	38	13	5	$2,5 \times 10^{-6}$
6.	5	37	12	6	$4,1 \times 10^{-5}$
7.	6	36	11	7	$4,3 \times 10^{-4}$
8.	7	35	10	8	0,003
9.	8	34	9	9	0,015
10.	9	33	8	10	0,050
11.	10	32	7	11	0,121
12.	11	31	6	12	0,205
13.	12	30	5	13	0,245
14.	13	29	4	14	0,202
15.	14	28	3	15	0,111
16.	15	27	2	16	0,039
17.	16	26	1	17	0,008
18.	17	25	0	18	$6,6 \times 10^{-4}$

Pro všechny řádky tabulky kromě řádku 13 tedy platí $p^* \leq p$ ($p = 0,205$). Výsledná p-hodnota Fisherova exaktního testu je dána součtem p^* všech řádků kromě řádku 13.

P-hodnotu testu tedy spočítáme jako $1 - 0,245 = 0,755$ a vzhledem k tomu, že platí $0,755 > 0,05$, nezamítáme na hladině významnosti $\alpha = 0,05$ nulovou hypotézu o nezávislosti veličin X a Y .

V Rku:

```
> ctyrpolni <- matrix(c(11, 31, 6, 12),
+                       ncol = 2,
+                       dimnames = list("X" = c("x1", "x2"),
+                                           "Y" = c("y1", "y2")))
> ctyrpolni
      Y
X     y1 y2
x1    11  6
x2    31 12
> fisher.test(x = ctyrpolni)
```

Fisher's Exact Test for Count Data

```
data:  ctyrpolni
p-value = 0.7553
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1868689 2.9024467
sample estimates:
odds ratio
 0.7138493
```

Pokud je x = matice, bere se jako dvourozměrná kontingenční tabulka a y můžeme ignorovat. Její položky měly být nezáporná celá čísla. Jinak musí být jak x , tak y faktory stejné délky. Neúplné případy jsou odstraněny.

Z výstupu získáváme p-hodnotu Fisherova exaktního testu (p -value) = 0,755 (stejný výsledek jako je uvedený počítáním).

Pro 2×2 tabulky je H_0 o nezávislosti ekvivalentní hypotéze, že poměr šancí je roven jedné.

Fisherův exaktní test byl odvozen primárně pro čtyřpolní tabulky, nicméně **existuje i jeho zobecnění na libovolnou kontingenční tabulku.**

Pagano M, Halvorsen KT. An Algorithm for Finding the Exact Significance Levels of $r \times c$ Contingency Tables. Journal of the American Statistical Association, 1981; 76: 931-934.

POZOR!!

Fisherův exaktní test v R může házet chybu pro výpočet z větších tabulek. Náročný na výpočet. Objem potřebných výpočtů roste s rozměrem tabulky, a proto se často využívá metoda Monte Carlo. V Rku zadáním argumentu `simulate.p.value = TRUE` -> počítá p -hodnoty pomocí simulace Monte Carlo.