

STATISTICKÉ
ZPRACOVÁNÍ DAT –
FISHERŮV EXAKTNÍ TEST

Metoda chí kvadrát x Fisherův test

- Pro zjištění závislosti – metoda chí kvadrát
- V některých případech metodu chí kvadrát nelze použít
 - rozsah souboru menší než 20
 - očekávané četnosti jsou malé
- Lze použít Fisherův test - založen na jiném principu
- Fisherův exaktní test je založen na výpočtu přesné (exaktní) pravděpodobnosti, se kterou bychom za platnosti nulové hypotézy o nezávislosti veličin získali naši konkrétní realizaci kontingenční tabulky

Fisherův test

- Zjišťujeme závislost dvou kvalitativních veličin na prvcích téhož výběru
- Máme náhodný výběr rozsahu n rozdělený do dvou skupin (skupina 1, skupina 2)
- Skupiny mohou nabývat hodnotu jednoho ze dvou znaků (znak 1, znak 2)
- Příkladem - skupina ženy, muži, znak kouří, nekouří
- Úkolem testu je rozhodnout, zda znaky jsou na sobě závislé nebo nezávislé (zda znak 1 má vliv na znak 2)
- Fisherův exaktní test odvozen pro kontingenční tabulku 2x2 tzv. čtyřpolní tabulku, ale existuje i jeho zobecnění pro libovolnou kontingenční tabulku

ZÁKLADNÍ PRINCIP FISHEROVA TESTU

- Testujeme nulovou hypotézu proti alternativní hypotéze.
- Nulová hypotéza H_0 : znaky 1 a 2 jsou nezávislé (Pozorované četnosti by měly odpovídat očekávaným četnostem)
- Alternativní hypotéza H_1 : Mezi znaky 1, 2 je závislost

- Nepředpokládá se, že teoretické rozdělení četností je známé, ale počítá se přímo pravděpodobnost odchylky od nulové hypotézy
- Při testování se generují varianty pozorované tabulky četností a určuje se pravděpodobnost výskytu všech obměn, které mají stejné součty okrajových četností
- Hlavní myšlenkou testu je výpočet pravděpodobnosti, se kterou bychom získali čtyřpolní tabulky stejně nebo více vzdálené od nulové hypotézy při zachování marginálních četností

VÝPOČET TESTOVÉ STATISTIKY

	Znak 1	Znak 2	Součet
Skupina 1	a	b	$a+b$
Skupina 2	c	d	$c+d$
Součet	$a+c$	$b+d$	n

Čtyřpolní tabulka

a, b, c, d četnosti
 $a+b, c+d, a+c, b+d$ okrajové četnosti tzv. marginální četnosti.

- Z hodnot a, b, c, d se vybere hodnota a a od té se postupně odečítá a po té přičítá hodnota 1, aby součet okrajových četností zůstal stejný a byly vyčerpány všechny možné případy. Např. pokud se od hodnoty a odečte 1, musí se k hodnotě b přičíst 1, k hodnotě c přičíst 1 a od hodnoty d odečíst 1, aby okrajové četnosti zůstaly stejné
- Generují se všechny možné varianty tabulky četností
- Pro původní a každou vygenerovanou tabulku se vypočítá pravděpodobnost

Vzorec pro výpočet pravděpodobnosti

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

p_i pravděpodobnost vypočtená z tabulky i

a, b, c, d četnosti uvnitř tabulky i

n rozsah souboru

Hodnota testového kritéria

- Hodnotou testového kritéria testové statistiky je součet všech vypočtených pravděpodobností menších nebo stejných jako hodnota pravděpodobnosti, která přísluší čtyřpolní tabulce sestrojené na základě pozorovaných hodnot
- Hodnotou testového kritéria se porovnává s hladinou významnosti α
- V případě oboustranného testu se sčítají hodnoty všech vypočtených pravděpodobností u tabulek, které jsou menší nebo rovny než skutečná zjištěná četnost
- Pokud je $\sum p_i < \alpha$, potom nulovou hypotézu o nezávislosti zamítáme a přijímáme hypotézu, že určitá závislost existuje

Příklad 1

- Skupina 1 a 2, znak 1 a 2, zkoumáme závislost mezi skupinami a znaky
- Hladina významnosti 5 %
- Ze získaných dat vytvoříme čtyřpolní tabulku

	Znak 1	Znak 2	Součet
Skupina 1	2	5	7
Znak 2	3	2	5
Součet	5	7	12

- Z této tabulky vybereme hodnotu 2 (skupina 1, znak 1) a od hodnoty 2 postupně odečítáme 1 a po té přičítáme hodnotu 1
- Ostatní hodnoty doplňujeme tak, aby součet okrajových četností zůstal stejný
- Dostaneme následující tabulky:

	Znak 1	Znak 2	Součet
Skupina 1	0	7	7
Znak 2	5	0	5
Součet	5	7	12

	Znak 1	Znak 2	Součet
Skupina 1	1	6	7
Znak 2	4	1	5
Součet	5	7	12

	Znak 1	Znak 2	Součet
Skupina 1	3	4	7
Znak 2	2	3	5
Součet	5	7	12

	Znak 1	Znak 2	Součet
Skupina 1	4	3	7
Znak 2	1	4	5
Součet	5	7	12

	Znak 1	Znak 2	Součet
Skupina 1	5	2	7
Znak 2	0	5	5
Součet	5	7	12

- Pravděpodobnost pro čtyřpolní tabulku sestavenou na základě pozorovaných hodnot je 0,265152
- Menší nebo stejné hodnoty nabývají pravděpodobnosti p_1 , p_2 , p_3 , p_5
- Hodnota testové statistiky (p -hodnota) je součet všech vypočtených pravděpodobností menších nebo stejných jako hodnota pravděpodobnosti pro čtyřpolní tabulku sestavenou na základě pozorovaných hodnot, tzn., že je

$$\sum p_i = p_1 + p_2 + p + p_3 + p_5 = 0,001263 + 0,044192 + 0,265152 + 0,220960 + 0,026515 = 0,558082$$
- Není splněna podmínka $\sum p_i < \alpha$, platí $\sum p_i > 0,05$, nulovou hypotézu o nezávislosti přijímáme a lze konstatovat, že mezi skupinami 1 a 2 a znaky 1 a 2 není závislost

- Generování všech možných variant tabulky četností je poměrně pracné, ale existuje řada programů, kde stačí zadat hodnoty zjištěných četností do tabulky a výsledkem je hodnota testové statistiky
- Příkladem vhodného programu je odkaz na <http://www.langsrud.com/fisher.htm>
- Aplikaci, která umožňuje zobecnění na kontingenční tabulku max 2x5 <https://quantitativeskills.com/sisa/statistics/fiveby2.htm>

Příklad 2

- <http://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickyh-a-biologickyh-dat--analyza-a-management-dat-pro-zdravotnicke-obory--testovani-hypotez-o-kvalitativnich-promennych--fisheruv-exaktni-test>