Hodnocení kontingenčních tabulek

Motivace

Při zpracování dat se velmi často setkáváme s veličinami nominálního typu, např.:

rodinný stav ženicha a nevěsty – svobodný/á, rozvedený/á, vdovec/vdova,

barva očí – modrá, hnědá, zelená, šedá,

barva vlasů – černá, hnědá, blond, rezavá,

krevní skupina – A, B, AB, 0,

atd.

Máme-li k dispozici n objektů, na nichž zjišťujeme hodnoty dvou nominálních veličin X a Y, můžeme testovat některé hypotézy. Nejdůležitější jsou:

hypotéza o nezávislosti

hypotéza o homogenitě (o shodnosti struktury)

hypotéza o symetrii

Hypotéza o nezávislosti: nulová hypotéza tvrdí, že znaky X a Y jsou nezávislé.

Např. nás zajímá, zda barva očí a barva vlasů jsou ve sledované populaci jedinců nezávislé.

Intenzitu případné závislosti měří různé koeficienty, které nabývají hodnot od 0 do 1. Čím je takový koeficient bližší 1, tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

Hypotéza o homogenitě (o shodnosti struktury): nulová hypotézy tvrdí, že rozložení pravděpodobností náhodné veličiny Y je stejné za různých podmínek, které vyjadřují varianty náhodné veličiny X.

Např. nás může zajímat, zda věková struktura hospitalizovaných pacientů – veličina Y, nabývá variant: do 25 let, 26-45 let, 46-65 let, 66 let a více, je stejná ve dvou nemocnicích – veličina X, nabývá variant 1 a 2.

Hypotéza o symetrii: nulová hypotéza tvrdí, že pokus neovlivní pravděpodobnos výskytu sledovaného znaku. Tato hypotéza se používá v situacích, kdy na n objektech sledujeme pravděpodobnost výskytu nějakého znaku před pokusem a po pokusu (obdoba párového t-testu).

Např. nás zajímá, zda požití alkoholu ovlivní schopnost řidiče úspěšně projet náročnou uzavřenou trať. V tomto případě veličina X nabývá variant 1 – projel trať bez chyb, 2 – projel trať s chybami a veličina Y nabývá variant 1 – před požitím alkoholu, 2 – po požití alkoholu.

Kontingenční tabulky

Nechť X,Y jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť X nabývá variant $x_{[1]}, \ldots, x_{[r]}$ a Y nabývá variant $y_{[1]}, \ldots, y_{[s]}$.

Označme:

$\pi_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]})$ … simultánní pravděpodobnost dvojice variant $(x_{[j]}, y_{[k]})$

$\pi_{j.} = P(X = x_{[j]})$ … marginální pravděpodobnost varianty $x_{[j]}$

$\pi_{.k} = P(Y = y_{[k]})$ … marginální pravděpodobnost varianty $y_{[k]}$

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

	y	$y_{[1]}$	…	$y_{[s]}$	$\pi_{j.}$
x	π_{jk}				
$x_{[1]}$		π_{11}	…	π_{1s}	$\pi_{1.}$
…		…	…	…	…
$x_{[r]}$		π_{r1}	…	π_{rs}	$\pi_{r.}$
$\pi_{.k}$		$\pi_{.1}$	…	$\pi_{.s}$	1

Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y). Zjištěné absolutní simultánní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})$ uspořádáme do kontingenční tabulky:

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

$n_{j.} = n_{j1} + \ldots + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$
$n_{.k} = n_{1k} + \ldots + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$

Simultánní pravděpodobnost π_{jk} odhadneme pomocí simultánní relativní četnosti $p_{jk} = \frac{n_{jk}}{n}$, marginální pravděpodobnosti $\pi_{j.}$ a $\pi_{.k}$ odhadneme pomocí marginálních relativních četností $p_{j.} = \frac{n_{j.}}{n}$ a $p_{.k} = \frac{n_{.k}}{n}$.

Testování hypotézy o nezávislosti

Testujeme nulovou hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny.

Kdyby náhodné veličiny X, Y byly stochasticky nezávislé, pak by platil multiplikativní vztah

$\forall j = 1, \ldots, r$, $\forall k = 1, \ldots, s$: $\pi_{jk} = \pi_{j.}\, \pi_{.k}$ neboli $\frac{n_{jk}}{n} = \frac{n_{j.}}{n} \cdot \frac{n_{.k}}{n}$, tj. $n_{jk} = \frac{n_{j.} n_{.k}}{n}$. Číslo $m_{jk} = \frac{n_{j.} n_{.k}}{n}$ se nazývá teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})$.

Testová statistika: $K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.} n_{.k}}{n}\right)^2}{\frac{n_{j.} n_{.k}}{n}}$.

Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty)$.

Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.

Podmínky dobré aproximace

Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti $\frac{n_{j.} n_{.k}}{n}$ aspoň v 80% případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20% neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

Měření síly závislosti

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde m = min{r,s}. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je k 1, tím je závislost mezi X a Y těsnější, čím blíže je k 0, tím je tato závislost volnější.

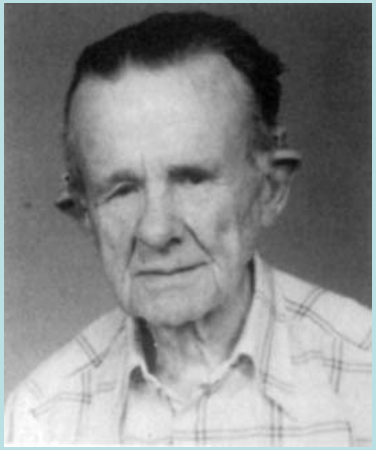
Význam hodnot Cramérova koeficientu:

mezi 0 až 0,1 … zanedbatelná závislost,

mezi 0,1 až 0,3 … slabá závislost,

mezi 0,3 až 0,7 … střední závislost,

mezi 0,7 až 1 … silná závislost.



Carl Harald Cramér (1893 – 1985): Švédský matematik

Příklad

V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází (veličina X) a typ školy, na kterou se hlásí (veličina Y). Výsledky jsou zaznamenány v kontingenční tabulce:

Sociální skupina	Typ školy			$n_{j.}$
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
$n_{.k}$	140	110	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení:

Nejprve vypočteme všech 12 teoretických četností:

Sociální skupina	Typ školy			$n_{j.}$
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
$n_{.k}$	140	110	110	360

$$\frac{n_{1.}n_{.1}}{n} = \frac{90 \cdot 140}{360} = 35, \quad \frac{n_{1.}n_{.2}}{n} = \frac{90 \cdot 110}{360} = 27{,}5, \quad \frac{n_{1.}n_{.3}}{n} = \frac{90 \cdot 110}{360} = 27{,}5,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{100 \cdot 140}{360} = 38{,}9, \quad \frac{n_{2.}n_{.2}}{n} = \frac{100 \cdot 110}{360} = 30{,}6, \quad \frac{n_{2.}n_{.3}}{n} = \frac{100 \cdot 110}{360} = 30{,}6,$$

$$\frac{n_{3.}n_{.1}}{n} = \frac{60 \cdot 140}{360} = 23{,}3, \quad \frac{n_{3.}n_{.2}}{n} = \frac{60 \cdot 110}{360} = 18{,}3, \quad \frac{n_{3.}n_{.3}}{n} = \frac{60 \cdot 110}{360} = 18{,}3,$$

$$\frac{n_{4.}n_{.1}}{n} = \frac{110 \cdot 140}{360} = 42{,}8, \quad \frac{n_{4.}n_{.2}}{n} = \frac{110 \cdot 110}{360} = 33{,}6, \quad \frac{n_{4.}n_{.3}}{n} = \frac{110 \cdot 110}{360} = 33{,}6$$

Vidíme, že podmínky dobré aproximace jsou splněny, všechny teoretické četnosti převyšují číslo 5.

Dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{(50-35)^2}{35} + \frac{(30-27{,}5)^2}{27{,}5} + \dots + \frac{(50-33{,}6)^2}{33{,}6} = 76{,}84.$$

Dále stanovíme kritický obor:

$$W = \left\langle \chi^2_{1-\alpha}((r-1)(s-1)); \infty \right) = \left\langle \chi^2_{0{,}95}((4-1)(3-1)); \infty \right) = \left\langle \chi^2_{0{,}95}(6); \infty \right) = \left\langle 12{,}6; \infty \right)$$

Protože $K \in W$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05.

Vypočteme Cramérův koeficient: $V = \sqrt{\frac{76{,}84}{360 \cdot 2}} = 0{,}326.$

Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných (X - sociální skupina, Y – typ školy, četnost) a 12 případech:

	1 X	2 Y	3 četno
1	I	univerzi	50
2	I	technick	30
3	I	ekonom	10
4	II	univerzi	30
5	II	technick	50
6	II	ekonom	20
7	III	univerzi	10
8	III	technick	20
9	III	ekonom	30
10	IV	univerzi	50
11	IV	technick	10
12	IV	ekonom	50

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (
Četnost označených buněk > 10
Pearsonův chí-kv. : 76,8359, sv=6, p

X	Y univerzi	Y technic	Y ekonomi	Řádk součt
I	35,00	27,50	27,50	90,00
II	38,88	30,55	30,55	100,0
III	23,33	18,33	18,33	60,00
IV	42,77	33,61	33,61	110,0
Vš.sku	140,0	110,0	110,0	360,0

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny. V záhlaví tabulky je uvedena hodnota testové statistiky K = 76,8359, počet stupňů volnosti 6 a odpovídající p-hodnota. Je velmi blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny.
Hodnotu testové statistiky a Cramérův koeficient dostaneme také tak, že na na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát a Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

Statist.	Chi-kv	sv	p
Pearsonův chí-k	76,83	df=	p=,00
M-V chí-kvadr.	84,53	df=	p=,00
Fí	,4619		
Kontingenční ko	,4193		
Cramér. V	,3266		

Testování hypotézy o homogenitě (o shodnosti struktury)

Na asymptotické hladině významnosti α testujeme hypotézu

H_0: $\pi_{1k} = \pi_{2k} = \ldots = \pi_{rk}$, $k = 1, 2, \ldots, s$ proti alternativě H_1: aspoň jedna dvojice pravděpodobností se liší.

Nulová hypotéza tvrdí, že rozložení pravděpodobností náhodné veličiny Y je stejné za různých podmínek, které vyjadřují varianty náhodné veličiny X.

Testová statistika i kritický obor jsou stejné jako při testování hypotézy o nezávislosti.

Příklad: V severozápadním Skotsku byla provedena studie, která měla prokázat, zda je procentuální zastoupení krevních skupin na celém území homogenní či nikoliv. V oblasti Eskdale bylo náhodně vybráno 100 osob, v oblasti Annandale 125 osob a v oblasti Nithsdale 253 osob. Výsledky jsou uvedeny v tabulce:

oblast	Krevní skupina				$n_{j.}$
	A	B	0	AB	
Eskdale	33	6	56	5	100
Annandale	54	14	52	5	125
Nithsdale	98	35	115	5	253
$n_{.k}$	185	55	223	15	478

Na asymptotické hladině významnosti 0,05 proveďte test homogenity.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných (X - oblast, Y – krevní skupina, četnost) a 12 případech:

	1 X	2 Y	3 četno
1	Eskdal	A	3
2	Eskdal	B	6
3	Eskdal	O	5
4	Eskdal	AB	5
5	Annan	A	5
6	Annan	B	1
7	Annan	O	5
8	Annan	AB	5
9	Nithsda	A	9
1	Nithsda	B	3
1	Nithsda	O	11
1	Nithsda	AB	5

Nejprve vytvoříme kontingenční tabulku řádkově podmíněných relativních četností, abychom získali představu o procentuálním zastoupení krevních skupin ve sledovaných třech oblastech:

	Kontingenční tabulka (krevní skup Četnost označených buněk > 10 (Marginální součty nejsou označe					
	X	Y A	Y B	Y O	Y AB	Řád souč
Četnost	Eskdal	3	6	5	5	10
Řádk. č		33,0	6,0	56,0	5,0	
Četnost	Annan	5	1	5	5	12
Řádk. č		43,2	11,2	41,6	4,0	
Četnost	Nithsda	9	3	11	5	25
Řádk. č		38,7	13,8	45,4	1,9	
Četnost	Vš.sku	18	5	22	1	47

Ověříme podmínky dobré aproximace:

Souhrnná tab.: Očekávané četnosti (kre Četnost označených buněk > 10 Pearsonův chí-kv. : 10,4537, sv=6, p=,1					
X	Y A	Y B	Y O	Y AB	Řádk součt
Eskdal	38,7(	11,50	46,65	3,138	100,0
Annand	48,3/	14,38	58,3'	3,922	125,0
Nithsda	97,9'	29,11	118,0	7,939	253,0
Vš.sku	185,0	55,00	223,0	15,00	478,0

Podmínky dobré aproximace jsou splněny.
Testová statistika nabývá hodnoty 10,45372, p-hodnota je 0,10681, což znamená, že na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že procentuální zastoupení krevních skupin ve sledovaných třech oblastech Skotska je shodné.

Testování hypotézy o symetrii

Má-li kontingenční tabulka stejný počet řádků jako sloupců (tj. r = s), nazývá se čtvercová. Pokud veličiny X a Y mají stejné varianty, můžeme testovat hypotézu symetrie H_0: $\pi_{jk} = \pi_{kj}$ pro všechny dvojice (j,k).

Testová statistika: $K = \sum_{j=1}^{r} \sum_{k=j+1}^{r} \frac{(n_{jk} - n_{kj})^2}{n_{jk} + n_{kj}}$. Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2(r(r-1)/2)$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}(r(r-1)/2), \infty)$.

H_0 zamítáme na asymptotické hladině významnosti α, když $K \in W$.

Příklad: Na souboru 45 náhodně vybraných žáků byla zjišťována obtížnost dvou úloh. Každý žák řešil obě úlohy v náhodném pořadí (polovina žáků v jednom pořadí, polovina v opačném). Výsledky řešení byly klasifikovány do tří kategorií:

Řešení 1. úlohy	Řešení 2. úlohy			$n_{j.}$
	správně	částečně správně	chybně	
správně	8	5	1	14
částečně správně	2	15	1	18
chybně	4	3	6	13
$n_{.k}$	14	23	8	45

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že obě úlohy jsou stejně obtížné.

Řešení:

Vypočítáme realizaci testové statistiky:

$$K = \frac{(5-2)^2}{5+2} + \frac{(1-4)^2}{1+4} + \frac{(1-3)^2}{1+3} = 4{,}085$$

Stanovíme kritický obor: $W = \langle \chi^2_{1-\alpha}(r(r-1)/2), \infty) = \langle \chi^2_{1-\alpha}(3), \infty) = \langle 7{,}815, \infty)$.

Testová statistika se nerealizuje v kritickém oboru, tedy na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že obě úlohy jsou stejně obtížné.

Čtyřpolní tabulky

Nechť $r = s = 2$. Pak hovoříme o čtyřpolní kontingenční tabulce a používáme označení: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$.

X	Y		$n_{j.}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	a+b
$x_{[2]}$	c	d	c+d
$n_{.k}$	a+c	b+d	n

Test nezávislosti ve čtyřpolní tabulce

Testovou statistiku pro čtyřpolní kontingenční tabulku lze zjednodušit do tvaru:

$$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Platí-li hypotéza o nezávislosti veličin X, Y, pak K se asymptoticky řídí rozložením $\chi^2(1)$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty)$

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \in W$.

Povšimněte si, že za platnosti hypotézy o nezávislosti $ad = bc$.

Pro čtyřpolní tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako Fisherův faktoriálový test.

Sir Ronald Aylmer Fisher (1890 – 1962): Britský statistik a genetik.
(Fisherův přesný test je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998. Princip spočívá v tom, že pomocí kombinatorických úvah se vypočítají pravděpodobnosti toho, že při daných marginálních četnostech dostaneme tabulky, které se od nulové hypotézy odchylují aspoň tak, jako daná tabulka.)

Upozornění: STATISTICA poskytuje p-hodnotu pro Fisherův přesný test. Jestliže vyjde $p \le \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.

Příklad: V náhodném výběru 50 obézních dětí ve věku 6 – 14 let byla zjišťována obezita rodičů. Veličina X – obezita matky, veličina Y – obezita otce. Výsledky průzkumu jsou uvedeny v kontingenční tabulce:

X	Y		$n_{j.}$
	ano	ne	
ano	15	9	24
ne	7	19	26
$n_{.k}$	22	28	50

Pomocí Fisherova exaktního testu ověřte, zda lze na hladině významnosti 0,05 zamítnout hypotézu o nezávislosti náhodných veličin X a Y.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o třech proměnných X, Y (varianty 0 – neobézní, 1 – obézní) a četnost a čtyřech případech:

	1 X	2 Y	3 četno
1	obézn	obézn	1
2	obézn	neobé	9
3	neobé	obézn	7
4	neobé	neobé	1

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Fisher exakt., Yates, McNemar (2x2). Dostaneme výstupní tabulku:

Statist.	Statist. : X(2) x Y(2) (d Chi-kva	sv	p
Pearsonův chi-kv	6,410	df=	p=,01
M-V chi-kvadr.	6,548	df=	p=,01
Yatesův chi-kv.	5,048	df=	p=,02
Fisherův přesný,			p=,01
2-stranný			p=,02
McNemarův chi-k	,2647	df=	p=,60
(B/C)	,0625	df=	p=,80

Vidíme, že p-hodnota pro Fisherův exaktní oboustranný test je 0,02163, tedy na hladině významnosti 0,05 zamítáme hypotézu, že obezita matky a otce spolu nesouvisí.

Test homogenity ve čtyřpolní tabulce

Na asymptotické hladině významnosti α testujeme hypotézu H_0: $\pi_{1k} = \pi_{2k}$, k = 1, 2 proti alternativě H_1: aspoň jedna dvojice pravděpodobností se liší.

Testová statistika $K = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ je stejná jako u testu nezávislosti. Tato statistika se v případě platnosti nulové hypotézy asymptoticky řídí rozložením $\chi^2(1)$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty)$. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \in W$.

Příklad: 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. Na asymptotické hladině významnosti 0,01 testujte hypotézu, že výskyt chřipky v očkované a kontrolní skupině je shodný.

Řešení:

Údaje uspořádáme do čtyřpolní kontingenční tabulky, kde roli veličiny X hraje onemocnění chřipkou a roli veličiny Y existence očkování.

X onemocnění chřipkou	Y existence očkování		$n_{j.}$
	ano	ne	
ano	20	80	100
ne	220	140	360
$n_{.k}$	240	220	460

Vypočteme sloupcově podmíněné relativní četnosti:

X onemocnění chřipkou	Y existence očkování	
	ano	ne
ano	8,3%	36,4%
ne	91,7%	63,6%

Vidíme, že v očkované skupině onemocnělo chřipkou 8,3% lidí, v kontrolní skupině však 36,4%. Zjistíme, zda takto velký rozdíl je způsoben pouze náhodnými vlivy.

Ověříme splnění podmínek dobré aproximace, tedy nejprve vypočteme teoretické četnosti:

$$\frac{n_{1.}n_{.1}}{n} = \frac{100 \cdot 240}{460} = 52,17,\quad \frac{n_{1.}n_{.2}}{n} = \frac{100 \cdot 220}{460} = 47,83,\quad \frac{n_{2.}n_{.1}}{n} = \frac{360 \cdot 240}{460} = 187,83,\quad \frac{n_{2.}n_{.2}}{n} = \frac{360 \cdot 220}{460} = 172,17$$

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny.

Realizace testové statistiky:

$$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{460(20 \cdot 140 - 80 \cdot 220)^2}{240 \cdot 220 \cdot 100 \cdot 360} = 53,0$$, Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty) = \langle \chi^2_{0,99}(1), \infty) = \langle 6,635, \infty)$.

Protože $K \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,01. S rizikem omylu nejvýše 0,01 jsme tedy prokázali, že výskyt chřipky v očkované a kontrolní skupině se liší.

Test symetrie ve čtyřpolní tabulce (McNemarův test)

Jde o obdobu párového testu pro ordinální, intervalové či poměrové proměnné. Na asymptotické hladině významnosti α testujeme nulovou hypotézu

H_0: $\pi_{12} = \pi_{21}$ proti alternativě H_1: $\pi_{12} \neq \pi_{21}$. (Jedná se o shodu pravděpodobností v políčkách na vedlejší diagonále.)

Například sledujeme, zda určité ošetření ovlivňuje pravděpodobnost výskytu jistého znaku na objektech základního souboru.

Testová statistika $K = \frac{(b-c)^2}{b+c}$ se v případě platnosti H_0 asymptoticky řídí rozložením $\chi^2(1)$.

Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty)$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \in W$.

Podmínky dobré aproximace: $b + c > 8$.

Příklad: Máme zjistit, zda požití alkoholu ovlivňuje schopnost řidičů správně projet určitou trasu. Bylo náhodně vybráno 100 řidičů, kteří měli za úkol projet uvedenou trasu. Po požití alkoholu dostali stejný úkol. Výsledky experimentu máme v tabulce:

Před požitím alkoholu	Po požití alkoholu		$n_{j.}$
	bezchybně	s chybami	
bezchybně	45	35	80
s chybami	15	5	20
$n_{.k}$	60	40	100

Test proveďte na asymptotické hladině významnosti 0,05.

Řešení:

Před požitím alkoholu správně projelo trať 60 řidičů, po požití pouze 40. Ověříme, zda rozdíl mezi těmito počty řidičů je způsoben pouze náhodnými vlivy.

Vidíme, že podmínky dobré aproximace jsou splněny, b + c = 50 > 8.

Vypočteme testovou statistiku: $K = \frac{(b-c)^2}{b+c} = \frac{(35-15)^2}{35+15} = 8$. Kritický obor: $W = \langle \chi^2_{1-\alpha}(1), \infty) = \langle \chi^2_{0,95}(1), \infty) = \langle 3{,}841, \infty)$. Protože $K \in W$, H_0 zamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 0,05 jsme tedy prokázali, že požití alkoholu ovlivňuje schopnost řidiče správně projet určitou trasu.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o třech proměnných X (1 – před požitím alkoholu, 2 – po požití alkoholu), Y (1 – bezchybně, 2 – s chybami) a četnost a čtyřech případech:

	1 X	2 Y	3 četno
1	před požitím a	bezchy	4
2	před požitím a	s chyba	3
3	po požití alkol	bezchy	1
4	po požití alkol	s chyba	5

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Fisher exakt., Yates, McNemar (2x2). Dostaneme výstupní tabulku:

Statist.	Statist. : X(2) x Y(2		
	Chi-kva	sv	p
Pearsonův chí-kv	2,343	df=	p=,12
M-V chí-kvadr.	2,458	df=	p=,11
Yatesův chí-kv.	1,627	df=	p=,20
Fisherův přesný,			p=,09
2-stranný			p=,20
McNemarův chí-k	30,42	df=	p=,00
(B/C)	7,220	df=	p=,00

Zajímá nás poslední řádek označený McNemarův chí-kv. (B/C). Testová statistika nabývá hodnoty 7,22 (systém STATISTICA používá při výpočtu testové statistiky tzv. opravu na nespojitost: $K = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}$). Odpovídající p-hodnota je 0,00721, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu, že požití alkoholu neovlivňuje schopnost řidiče správně projet určitou trasu.

Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá výběrový podíl šancí (odds ratio). Považuje-me ho za odhad neznámého teoretického podílu šancí $o\rho$. Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

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

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je tedy

$$OR = \frac{ad}{bc}.$$

Jsou-li veličiny X, Y nezávislé, pak $\pi = $, tudíž teoretický podíl šancí $o\rho$. Závislost veličin X, Y bude tím silnější, čím více se $o\rho$ bude lišit od 1. Avšak $o\rho \in (0, \infty)$, tedy hodnoty $o\rho$ jsou kolem 1 rozmístěny nesymetricky. Z tohoto důvodu raději používáme logaritmus teoretického či výběrového podílu šancí.

Testování nezávislosti ve čtyřpolních tabulkách pomocí podílu šancí

Na asymptotické hladině významnosti α testujeme hypotézu H$_0$: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\ln OR = 0$) proti alternativě H$_1$: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\ln OR \neq 0$).

Testová statistika $T_0 = \dfrac{\ln OR}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$ se asymptoticky řídí rozložením $N(0,1)$, když nulová hypotéza platí.

Kritický obor: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.

Nulovou hypotézu tedy zamítáme na asymtotické hladině významnosti α, když se testová statistika realizuje v kritickém oboru W.

Testování nezávislosti lze provést též pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro logaritmus podílu šancí $\ln OR$, který je dán vzorcem:

$$(d, h) = \left(\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2},\ \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2}\right)$$

Jestliže interval spolehlivosti neobsahuje 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

Příklad (testování nezávislosti pomocí podílu šancí a pomocí statistiky K):

U 135 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Řešení:

a) Testování pomocí podílu šancí:

$OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2,298$. Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojmem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojmem.

Provedeme další pomocné výpočty:

$$\ln OR = 0,832,$$

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}} = 0,439,\ u_{0,975} = 1,96$$

Dosadíme do vzorců pro meze asymptotického intervalu spolehlivosti pro podíl šancí:

$$d = \ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0,832 - 0,439 \cdot 1,96 = -0,028,\quad h = \ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} = 0,832 + 0,439 \cdot 1,96 = 1,692$$

Protože interval (-0,028; 1,692) obsahuje číslo 0, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

b) Testování pomocí statistiky K:

přijetí	dojem		$n_{j.}$
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
$n_{.k}$	56	69	125

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_{1.}n_{.1}}{n} = \frac{28 \cdot 56}{125} = 12{,}54,\quad \frac{n_{1.}n_{.2}}{n} = \frac{28 \cdot 69}{125} = 15{,}45,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{97 \cdot 56}{125} = 43{,}45,\quad \frac{n_{2.}n_{.2}}{n} = \frac{97 \cdot 69}{125} = 53{,}54$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{125(17 \cdot 58 - 11 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 3{,}695$$

Kritický obor: $W = \langle \chi^2_{0,95}(1), \infty) = \langle 3{,}841, \infty)$.

Protože testová statistika se nerealizuje k kritickém oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Vypočteme ještě Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}} = \sqrt{\frac{3{,}6953}{125(2-1)}} = 0{,}171$

Vidíme, že mezi dojmem u přijímací zkoušky a přijetím na fakultu je pouze slabá závislost.

Poznámka k jednostranným alternativám:
Nulová hypotéza tvrdí, že podíl šancí je roven 1, tj. H_0: OR = 1.

Pokud víme, že za prvních okolností je šance na úspěch vyšší než za druhých okolností, pak proti nulové hypotéze postavíme pravostrannou alternativu
H_1: OR > 1.
Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch pravostranné alternativy, když $100(1-\alpha)\%$ empirický asymptotický levostranný interval spolehlivosti pro ln OR neobsahuje číslo 0.

Pokud víme, že za prvních okolností je šance na úspěch nižší než za druhých okolností, pak proti nulové hypotéze postavíme levostrannou alternativu
H_1: OR < 1.
Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch levostranné alternativy, když $100(1-\alpha)\%$ empirický asymptotický pravostranný interval spolehlivosti pro ln OR neobsahuje číslo 0.

Pokud jsou šance na úspěch stejné za prvních i druhých okolností, pak proti nulové hypotéze postavíme oboustrannou alternativu
H_1: OR $\neq$ 1.
Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když $100(1-\alpha)\%$ empirický asymptotický oboustranný interval spolehlivosti pro ln OR neobsahuje číslo 0.

Příklad: U 24 žáků 6. třídy základní školy bylo zjišťováno, zda jsou úspěšní v matematice (tj. mají na posledním vysvědčení známku 1 nebo 2 z matematiky) a zda hrají na nějaký hudební nástroj. Z 10 úspěšných matematiků 6 hrálo na nějaký hudební nástroj, kdežto ve skupině neúspěšných matematiků hrál pouze 1 žák na hudební nástroj. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že úspěch v matematice a hra na hudební nástroj jsou nezávislé veličiny. Proti nulové hypotéze postavte

a) oboustrannou alternativu, tj. tvrzení, úspěch v matematice a hra na hudební nástroj spolu souvisí,
b) pravostrannou alternativu, tj. tvrzení, že šance na úspěch v matematice jsou vyšší pro žáky, kteří hrají na nějaký hudební nástroj,
c) pravostrannou alternativu, tj. tvrzení, že šance na úspěch v matematice jsou nižší pro žáky, kteří hrají na nějaký hudební nástroj.

Řešení:

Máme kontingenční tabulku

úspěch v M	hra na hudební nástroj		$n_{j.}$
	ano	ne	
ano	6	4	10
ne	1	13	14
$n_{.k}$	7	17	24

Vypočteme podíl šancí: $OR = \frac{ad}{bc} = \frac{6\cdot 13}{4\cdot 1} = \frac{39}{2} = 19{,}5$. Podíl šancí nám říká, že žák, který hraje na nějaký hudební nástroj, má 19,5 x větší šanci na úspěch v matematice než žák, který nehraje na žádný hudební nástroj.

Ad a)

Pro testování nulové hypotézy proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti:

Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a jednom případu. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

=log(19,5)-sqrt(1/6+1/4+1/1+1/13)*VNormal(0,975;0;1)

a analogicky do Do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:

=log(19,5)+sqrt(1/6+1/4+1/1+1/13)*VNormal(0,975;0;1)

	1 DM	2 HM
1	0,575	5,365

Vidíme, že 0,575093 < ln OR < 5,365736 s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05 ve prospěch oboustranné alternativy. S rizikem omylu nejvýše 5% se tedy prokázalo, že úspěch v matematice souvisí s hrou na hudební nástroj.

Ad b)
Pro testování nulové hypotézy proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti:
Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:
=log(19,5)-sqrt(1/6+1/4+1/1+1/13)*VNormal(0,95;0;1)

	1 DM
1	0,960

Protože interval (0,960198; ∞) neobsahuje 0, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05 ve prospěch pravostranné alternativy. S rizikem omylu nejvýše 5% se tedy prokázalo, že žáci, kteří hrají na nějaký hudební nástroj, mají vyšší šance na úspěch v matematice.

Ad c)
Pro testování nulové hypotézy proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti:
Do Dlouhého jména proměnné HM napíšeme vzorec pro dolní mez:
=log(19,5)+sqrt(1/6+1/4+1/1+1/13)*VNormal(0,95;0;1)

	1 HM
1	4,980

Protože interval (-∞; 4,980631) obsahuje 0, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05 ve prospěch levostranné alternativy. Neprokázalo se tedy, že žáci, kteří hrají na nějaký hudební nástroj, mají nižší šance na úspěch v matematice.