Analýza závislosti dvou nominálních náhodných veličin

Motivace

Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny nominálního typu jsou stochasticky nezávislé. Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá.
Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány 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.

Kontingenční tabulky

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

Nyní pořídíme dvourozměrný náhodný výběr 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

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

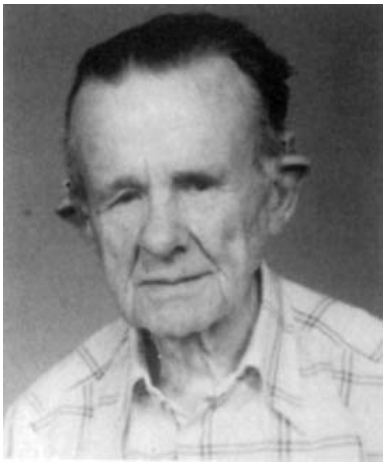
Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti $\frac{n_{j.} n_{.k}}{n}$ aspoň v 80% případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20% neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

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 1, tím je těsnější závislost mezi X a Y, čím blíže je 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:

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

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

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

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

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

Nyní dosadíme do vzorce pro testovou statistiku K:

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

Dále stanovíme kritický obor:

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

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

Vypočteme Cramérův koeficient: $V=\sqrt{\frac{76{,}4}{360\cdot 2}}=0{,}3267\,.$

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 četnost
1	I	univerzitní	50
2	I	technický	30
3	I	ekonomický	10
4	II	univerzitní	30
5	II	technický	50
6	II	ekonomický	20
7	III	univerzitní	10
8	III	technický	20
9	III	ekonomický	30
10	IV	univerzitní	50
11	IV	technický	10
12	IV	ekonomický	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 (typ skoly) Četnost označených buněk > 10 Pearsonův chí-kv. : 76,8359, sv=6, p=,000000				
X	Y univerzitní	Y technický	Y ekonomický	Řádk. součty
I	35,0000	27,5000	27,5000	90,0000
II	38,8889	30,5556	30,5556	100,0000
III	23,3333	18,3333	18,3333	60,0000
IV	42,7778	33,6111	33,6111	110,0000
Vš.skup.	140,0000	110,0000	110,0000	360,0000

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 záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát, Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

	Statist. : X(4) x Y(3) (typ skoly)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	76,83589	df=6	p=,00000
M-V chí-kvadr.	84,53528	df=6	p=,00000
Fí	,4619881		
Kontingenční koeficient	,4193947		
Cramér. V	,3266749		
Koeficient nejistoty	X=,0861227	Y=,1075343	X|Y=,09564

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, …, 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.
(Jde o podobný problém jako v analýze rozptylu jednoduchého třídění, kde porovnáváme shodu středních hodnot intervalové či poměrové proměnné. V tomto případě však porovnáváme shodu pravděpodobnostního rozložení nominální proměnné.)
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 četnost
1	Eskdale	A	33
2	Eskdale	B	6
3	Eskdale	O	56
4	Eskdale	AB	5
5	Annandale	A	54
6	Annandale	B	14
7	Annandale	O	52
8	Annandale	AB	5
9	Nithsdale	A	98
10	Nithsdale	B	35
11	Nithsdale	O	115
12	Nithsdale	AB	5

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

	Kontingenční tabulka (krevni skupiny) Četnost označených buněk > 10 (Marginální součty nejsou označeny)					
	X	Y A	Y B	Y O	Y AB	Řádk. součty
Četnost	Eskdale	33	6	56	5	100
Řádk. četn.		33,00%	6,00%	56,00%	5,00%	
Četnost	Annandale	54	14	52	5	125
Řádk. četn.		43,20%	11,20%	41,60%	4,00%	
Četnost	Nithsdale	98	35	115	5	253
Řádk. četn.		38,74%	13,83%	45,45%	1,98%	
Četnost	Vš.skup.	185	55	223	15	478

Ověříme podmínky dobré aproximace:

Souhrnná tab.: Očekávané četnosti (krevni skupiny) Četnost označených buněk > 10 Pearsonův chí-kv. : 10,4537, sv=6, p=,106812					
X	Y A	Y B	Y O	Y AB	Řádk. součty
Eskdale	38,7029	11,50628	46,6527	3,13808	100,0000
Annandale	48,3787	14,38285	58,3159	3,92259	125,0000
Nithsdale	97,9184	29,11088	118,0314	7,93933	253,0000
Vš.skup.	185,0000	55,00000	223,0000	15,00000	478,0000

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.

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:

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

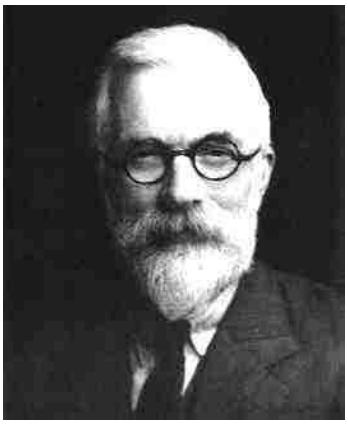
Platí-li hypotéza o nezávislosti veličin X, Y, pak K se asymptoticky řídí rozložením $\chi^2(1)$.

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 tuto 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 $p \leq \alpha$, 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 četnost
1	obézní	obézní	15
2	obézní	neobézní	9
3	neobézní	obézní	7
4	neobézní	neobézní	19

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. Dostaneme výstupní tabulku:

	Statist. : X(2) x Y(2) (obezita rodicu)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	6,410777	df=1	p=,01134
M-V chí-kvadr.	6,548348	df=1	p=,01050
Yatesův chí-kv.	5,048207	df=1	p=,02465
Fisherův přesný, 1-str.			p=,01188
2-stranný			p=,02163
McNemarův chí-kv. (A/D)	,2647059	df=1	p=,60691
(B/C)	,0625000	df=1	p=,80259

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. Na problém lze pohlížet tak, že máme dva nezávislé výběry z alternativních rozložení, první má rozsah $n_1 = a+c$ a pochází z rozložení $A(\vartheta_1)$, druhý má rozsah $n_2 = b+d$ a pochází z rozložení $A(\vartheta_2)$. Testujeme hypotézu H_0: $\vartheta_1 - \vartheta_2 = 0$ proti oboustranné alternativě. V kapitole o hodnocení dvou nezávislých náhodných výběrů z alternativních rozložení jsme použili testovou statistiku

$$T_0 = \frac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$ která se za platnosti nulové hypotézy asymptoticky řídí rozložením N(0,1). (M_* je vážený průměr výběrových průměrů.) Nyní použijeme testovou statistiku $K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$, stejně 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	$\frac{20}{240} = 8{,}3\%$	$\frac{80}{220} = 36{,}4\%$
ne	$\frac{220}{240} = 91{,}7\%$	$\frac{140}{220} = 63{,}6\%$

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:

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

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

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny.
Realizace testové statistiky:

$$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{460(20 \cdot 140 - 80 \cdot 220)^2}{240 \cdot 220 \cdot 100 \cdot 360} = 53{,}01.$$

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.

Nyní ještě provedeme test hypotézy, že výskyt chřipky v očkované a kontrolní skupině je shodný, pomocí testové statistiky $T_0 = \dfrac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$, která se za platnosti nulové hypotézy asymptoticky řídí rozložením N(0,1).
V tomto případě

$$m_1 = \frac{20}{240},\ m_2 = \frac{80}{220},\ m_* = \frac{20+80}{460} = \frac{100}{460},\ n_1 = 240,\ n_2 = 220$$

Realizace testové statistiky: $t_0 = \dfrac{\dfrac{20}{240} - \dfrac{80}{220}}{\sqrt{\dfrac{100}{460}\cdot\dfrac{360}{460}\cdot\left(\dfrac{1}{240}+\dfrac{1}{220}\right)}} = -7{,}28069$

Kritický obor:

$$W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = (-\infty, -u_{0{,}995}\rangle \cup \langle u_{0{,}995}, \infty) = (-\infty, -2{,}5758\rangle \cup \langle 2{,}5758, \infty)$$

Testová statistika se realizuj ev kritickém oboru, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu, že výskyt chřipky v očkované a kontrolní skupině se neliší.
(Povšimněte si, že kvadrát testové statistiky T_0 má stejnou hodnotu jako testová statistika K.)

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o třech proměnných X (varianty 1 – chřipka ano, 0 – chřipka ne), Y (varianty 1 – očkování ano, 0 – očkování ne) a četnost a čtyřech případech:

	1 X	2 Y	3 cetnost
1	chripka ano	ockovani ano	20
2	chripka ano	ockovani ne	80
3	chripka ne	ockovani ano	220
4	chripka ne	ockovani ne	140

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. Dostaneme výstupní tabulku:

Statist.	Statist. : X(2) x Y(2) (chripka.sta)		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	53,00842	df=1	p=,00000
M-V chí-kvadr.	55,60618	df=1	p=,00000
Yatesův chí-kv.	51,37366	df=1	p=,00000
Fisherův přesný, 1-str.			p=,00000
2-stranný			p=,00000
McNemarův chí-kv. (A/D)	88,50625	df=1	p=0,0000
(B/C)	64,40334	df=1	p=,00000

Vidíme, že p-hodnota pro Fisherův exaktní oboustranný test je blízká 0, tedy na hladině významnosti 0,01 zamítáme hypotézu, že výskyt chřipky v očkované a kontrolní skupině se neliší.
(Testová statistika K je uvedena v 1., řádku výstupní tabulky.)
Pokud chceme posoudit sílu vlivu očkování na výskyt chřipky, vypočteme pomocí systému STATISTICA Cramérův koeficient: na záložce Možnosti zaškrtneme Fí (tabulky 2x2) & Cramérovo V & C:

Statist.	Statist. : X(2) x Y(2) (chripka.sta)		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	53,00842	df=1	p=,00000
M-V chí-kvadr.	55,60618	df=1	p=,00000
Fí pro tabulky 2 x 2	-,339464		
Tetrachorická korelace	-,576894		
Kontingenční koeficient	,3214476		

Cramérův koeficient je roven absolutní hodnotě koeficientu Fí, tedy 0,3395. Vidíme, že se jedná o středně silnou závislost.

Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá podíl šancí (odds ratio). 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 $OR = \frac{ad}{bc}$.

Pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro podíl šancí lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický 100(1-α)% interval spolehlivosti pro skutečný podíl šancí má meze:

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

Jestliže interval spolehlivosti neobsahuje 1, 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:

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

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

Po odlogaritmování dostaneme:

$d = e^{-0{,}028} = 0{,}972,\ h = e^{1{,}692} = 5{,}433$

Protože interval (0,972; 5,433) obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

b) Testování pomocí statistiky K:

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_{1.}n_{.1}}{n}=\frac{28\cdot 56}{125}=12{,}544\,,\quad \frac{n_{1.}n_{.2}}{n}=\frac{28\cdot 69}{125}=15{,}456\,,$$

$$\frac{n_{2.}n_{.1}}{n}=\frac{97\cdot 56}{125}=43{,}456\,,\quad \frac{n_{2.}n_{.2}}{n}=\frac{97\cdot 69}{125}=53{,}544$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}=\frac{125\cdot(17\cdot 58-11\cdot 39)^2}{28\cdot 97\cdot 56\cdot 69}=3{,}6953$$

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: $V=\sqrt{\frac{K}{n(m-1)}}=\sqrt{\frac{3{,}6953}{125(2-1)}}=0{,}1719$

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

Výpočet pomocí systému STATISTICA:

a) Testování pomocí podílu šancí:

Vytvoříme datový soubor s proměnnými dm (dolní mez) a hm (horní mez). Do Dlouhého jména proměnné dm napíšeme:
=exp(log((17*58)/(11*39))-sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))
a do Dlouhého jména proměnné hm napíšeme:
=exp(log((17*58)/(11*39))+sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))
Získáme výsledek:

	1 dm	2 hm
1	0,9724	5,4324

Protože interval (0,9724; 5,4324) obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

b) Testování pomocí statistiky K:

Vytvoříme datový soubor s proměnnými prijeti (varianta 1 – přijat, 0 – nepřijat), dojem (varianta 1 – dobrý, 0 – špatný) a cetnost a čtyřmi případy:

	1 prijeti	2 dojem	3 cetnost
1	prijat	dobry	17
2	prijat	spatny	11
3	neprijat	dobry	39
4	neprijat	spatny	58

Ověříme splnění podmínek dobré aproximace:
Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 prijeti, List 2 dojem – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (prijeti_na_fakultu.sta)
Četnost označených buněk > 10
Pearsonův chí-kv. : 3,69533, sv=1, p=,054568

prijeti	dojem spatny	dojem dobry	Řádk. součty
neprijat	53,54400	43,45600	97,0000
prijat	15,45600	12,54400	28,0000
Vš.skup.	69,00000	56,00000	125,0000

Vidíme, že podmínky dobré aproximace jsou splněny. V záhlaví této tabulky je uvedena hodnota testové statistiky K = 3,69533 a odpovídající asymptotická p-hodnota = 0,054568. Nulovou hypotézu tedy nelze zamítnout na asymptotické hladině významnosti 0,05.

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-α)% empirický asymptotický levostranný interval spolehlivosti pro OR neobsahuje číslo 1.

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-α)% empirický asymptotický pravostranný interval spolehlivosti pro OR neobsahuje číslo 1.

Pokud jsou šance na úspěch stejné za prvních i druhých okolností, pak proti nulové hypotéze postavíme oboustrannou alternativu

H_1: OR ≠ 1.
Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když 100(1-α)% empirický asymptotický oboustranný interval spolehlivosti pro OR neobsahuje číslo 1.

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) levostrannou 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: $\mathrm{OR} = \frac{ac}{bd} = \frac{6 \cdot 13}{4 \cdot 1} = \frac{39}{2} = 19{,}5$. 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:
=exp(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:

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

	1 DM	2 HM
1	1,777296	213,9486

Vidíme, že 1,7773 < OR < 213,9486 s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 1, 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:
=exp(log(19,5)-sqrt(1/6+1/4+1/1+1/13)*VNormal(0,95;0;1))

	1 DM
1	2,612213

Protože interval (2,612213; ∞) neobsahuje 1, 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:
=exp(log(19,5)+sqrt(1/6+1/4+1/1+1/13)*VNormal(0,95;0;1))

	1 HM
1	145,5663

Protože interval (-∞; 145,5663) obsahuje 1, 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.