

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 hypotézu, 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.

Kontingenční tabulky

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

Označme:

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

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

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

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

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

Pořídíme dvourozměrný náhodný výběr $(X_1, Y_1), \dots, (X_n, Y_n)$ rozsahu n z rozložení, kterým se řídí dvourozměrný diskretní 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

$n_{j.} = n_{j1} + \dots + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$

$n_{.k} = n_{1k} + \dots + n_{rk}$ 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

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

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

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

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

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 $\frac{n_{j \cdot} \cdot n_{\cdot k}}{n}$ 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.

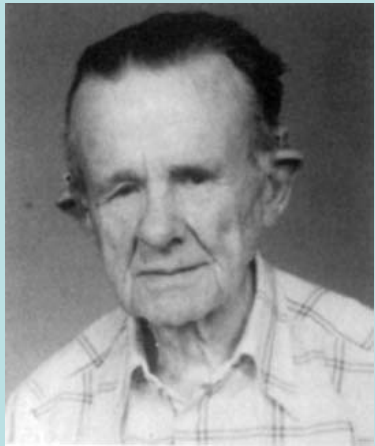
Význam hodnot Cramérova koeficientu:

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 Cramérovův 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

$$\begin{aligned} \frac{n_{1.}n_{.1}}{n} &= \frac{90 \cdot 140}{360} = 35, & \frac{n_{1.}n_{.2}}{n} &= \frac{90 \cdot 110}{360} = 27,5, & \frac{n_{1.}n_{.3}}{n} &= \frac{90 \cdot 110}{360} = 27,5, \\ \frac{n_{2.}n_{.1}}{n} &= \frac{100 \cdot 140}{360} = 38,9, & \frac{n_{2.}n_{.2}}{n} &= \frac{100 \cdot 110}{360} = 30,6, & \frac{n_{2.}n_{.3}}{n} &= \frac{100 \cdot 110}{360} = 30,6, \\ \frac{n_{3.}n_{.1}}{n} &= \frac{60 \cdot 140}{360} = 23,3, & \frac{n_{3.}n_{.2}}{n} &= \frac{60 \cdot 110}{360} = 18,3, & \frac{n_{3.}n_{.3}}{n} &= \frac{60 \cdot 110}{360} = 18,3, \\ \frac{n_{4.}n_{.1}}{n} &= \frac{110 \cdot 140}{360} = 42,8, & \frac{n_{4.}n_{.2}}{n} &= \frac{110 \cdot 110}{360} = 33,6, & \frac{n_{4.}n_{.3}}{n} &= \frac{110 \cdot 110}{360} = 33,6 \end{aligned}$$

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

Dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{(90 - 35)^2}{35} + \frac{(90 - 27,5)^2}{27,5} + \dots + \frac{(110 - 33,6)^2}{33,6} = 76,84.$$

Dále stanovíme kritický obor:

$$W = \left\{ \chi^2_{1-0,05} \leq \chi^2_{0,95} \leq \chi^2_{1-0,05} \right\} = \left\{ \chi^2_{0,95} \leq \chi^2_{0,95} \leq \chi^2_{0,95} \right\} = \left\{ 12,6, \infty \right\}$$

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

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

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 (ty p školy)				
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 a Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

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		

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: $w = \left(\chi^2_{1-\alpha}, \infty \right)$

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 odchyľují 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., Yates, McNemar (2x2). Dostaneme výstupní tabulku:

Statist.	Statist. : X(2) x Y(2) (obezita rodicu)		
	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.

Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá výběrový **podíl šancí** (odds ratio). Považujeme ho za odhad neznámého teoretického podílu šancí $op = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$. Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j\cdot}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{\cdot k}$	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 $\pi_{jk} = \pi_{j\cdot}\pi_{\cdot k}$, tudíž teoretický podíl šancí $op = 1$. Závislost veličin X, Y bude tím silnější, čím více se op bude lišit od 1. Avšak $op \in (0, \infty)$, tedy hodnoty op 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. $\ln \varphi = 0$) proti alternativě H_1 : X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\ln \varphi \neq 0$).

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

Kritický obor: $W = (-\infty, -u_{1-\alpha/2}] \cup [u_{1-\alpha/2}, \infty)$.

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

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

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

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:

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

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:

$$\frac{n_{1,n_1}}{n} = \frac{28 \cdot 56}{125} = 12,544, \quad \frac{n_{1,n_2}}{n} = \frac{28 \cdot 69}{125} = 15,456,$$

$$\frac{n_{2,n_1}}{n} = \frac{97 \cdot 56}{125} = 43,456, \quad \frac{n_{2,n_2}}{n} = \frac{97 \cdot 69}{125} = 53,544$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n \sum_{j=1}^m \sum_{k=1}^m \frac{(n_{jk} - \frac{n_{j.} \cdot n_{.k}}{n})^2}{\frac{n_{j.} \cdot n_{.k}}{n}}}{\sum_{j=1}^m \sum_{k=1}^m \frac{n_{jk} \cdot (n_{jk} - 1)}{n_{j.} \cdot n_{.k}}} = \frac{125 \cdot (17 \cdot 58 - 1 \cdot 39)^2}{28 \cdot 97 \cdot 56 \cdot 69} = 3,6953$$

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

Protože testová statistika se nerealizuje k kritickému 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.

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 jednostranný 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 jednostrannou alternativu

$H_1: OR < 1$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch jednostranné alternativy, když $100(1-\alpha)\%$ empirický asymptotický jednostranný 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

- oboustrannou alternativu, tj. tvrzení, úspěch v matematice a hra na hudební nástroj spolu souvisí,
- pravostrannou alternativu, tj. tvrzení, že šance na úspěch v matematice jsou vyšší pro žáky, kteří hrají na nějaký hudební nástroj,
- 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: $OR = \frac{ac}{bd} = \frac{6 \cdot 13}{4 \cdot 1} = \frac{39}{2} = 19,5$. 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 sestojíme 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 případě. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

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

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

$$=\log(19,5)+\sqrt{1/6+1/4+1/1+1/13}*\text{VNormal}(0,975;0;1)$$

	1 DM	2 HM
1	0,575093	5,365736

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	0,960198

Protože interval $(0,960198; \infty)$ 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	4,980631

Protože interval $(-\infty; 4,980631)$ 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.