

Hodnocení kontingenčních tabulek

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 $x_{[1]}, \dots, x_{[r]}$ a Y nabývá variant $y_{[1]}, \dots, y_{[s]}$.

Označme:

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

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

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

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

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

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

$n_{\cdot k} = n_{1k} + \dots + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$

Simultánní pravděpodobnost π_{jk} odhadneme pomocí simultánní relativní četnosti $p_{jk} = \frac{n_{jk}}{n}$, marginální pravděpodobnosti $\pi_{j\cdot}$

a $\pi_{\cdot k}$ odhadneme pomocí marginálních relativních četností $p_{j\cdot} = \frac{n_{j\cdot}}{n}$ a $p_{\cdot k} = \frac{n_{\cdot k}}{n}$.

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 multiplikační vztah

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

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

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

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

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 $\frac{n_j \cdot n_k}{n}$ 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: $V = \frac{\sqrt{K}}{\sqrt{m}}$, 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.

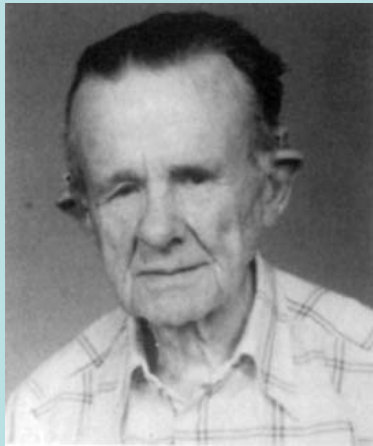
Význam hodnot Cramérova koeficientu:

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

$$\begin{aligned} \frac{n_{11}n_1}{n} &= \frac{90 \cdot 40}{360} = 10, & \frac{n_{12}n_2}{n} &= \frac{90 \cdot 10}{360} = 2,5, & \frac{n_{13}n_3}{n} &= \frac{90 \cdot 10}{360} = 2,5 \\ \frac{n_{21}n_1}{n} &= \frac{10 \cdot 40}{360} = 1,1, & \frac{n_{22}n_2}{n} &= \frac{10 \cdot 10}{360} = 0,28, & \frac{n_{23}n_3}{n} &= \frac{10 \cdot 10}{360} = 0,28 \\ \frac{n_{31}n_1}{n} &= \frac{60 \cdot 40}{360} = 6,7, & \frac{n_{32}n_2}{n} &= \frac{60 \cdot 10}{360} = 1,7, & \frac{n_{33}n_3}{n} &= \frac{60 \cdot 10}{360} = 1,7 \\ \frac{n_{41}n_1}{n} &= \frac{11 \cdot 40}{360} = 1,1, & \frac{n_{42}n_2}{n} &= \frac{11 \cdot 10}{360} = 0,3, & \frac{n_{43}n_3}{n} &= \frac{11 \cdot 10}{360} = 0,3 \end{aligned}$$

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

Dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{50^2}{90} + \frac{30^2}{90} + \dots + \frac{50^2}{110} = 34.$$

Dále stanovíme kritický obor:

$$W = \chi_{1-0,05}^2(4-1) = \chi_{0,95}^2(3) = 7,88$$

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

Vypočteme Cramérův koeficient: $V = \sqrt{\frac{7,88}{36}} = 0,46$

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 četn
	1	I univerzi	5
	2	I technick	3
	3	I ekonom	1
	4	II univerzi	3
	5	II technick	5
	6	II ekonom	2
	7	III univerzi	1
	8	III technick	2
	9	III ekonom	3
	10	IV univerzi	5
	11	IV technick	1
	12	IV ekonom	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 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	38,89	30,56	30,56	100,00
III	23,33	18,33	18,33	60,00
IV	42,78	33,64	33,64	110,00
vs.skl	140,00	110,00	110,00	360,00

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
F1	,4619		
Kontingenční ko	,4193		
Cramér. V	,3266		

Test homogenity v tabulce typu 2 x s

Máme kontingenční tabulku, v níž veličina X má jen dvě varianty a veličina Y s variant:

	y	Y _[1]	...	Y _[s]	$\pi_{j.}$
X	π_{jk}				
X _[1]		π_{11}	...	π_{1s}	$\pi_{1.}$
X _[2]		π_{21}	...	π_{2s}	$\pi_{2.}$
$\pi_{.k}$		$\pi_{.1}$...	$\pi_{.s}$	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	π_{jk}				
X _[1]		n_{11}	...	n_{1s}	$n_{1.}$
X _[2]		n_{21}	...	n_{2s}	$n_{2.}$
$\pi_{.k}$		$n_{.1}$...	$n_{.s}$	n

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

Na problém lze pohlížet tak, že máme s nezávislých náhodných výběrů z alternativních rozložení, přičemž první má rozsah $n_1 = n_{11} + n_{21}$ a pochází z rozložení $A(\underline{q})$, ..., s-tý má rozsah $n_s = n_{1s} + n_{2s}$ a pochází z rozložení $A(\underline{q})$. Testujeme hypotézu $H_0: \underline{q} \dots =$ proti alternativě H_1 : non H_0 .

V kapitole o hodnocení náhodných výběrů z alternativních rozložení jsme použili testovou statistiku:

$$Q = \frac{1}{n} \sum_{j=1}^s \frac{(n_j - \frac{n \cdot p_j}{s})^2}{\frac{n \cdot p_j}{s}}, \text{ když } H_0 \text{ platí.}$$

Kritický obor: $W = \chi^2_{s-1, \alpha}$

H_0 tedy zamítáme na asymptotické hladině významnosti α , když $Q > W$. Přitom $M = \frac{1}{n} \sum_{j=1}^s n_j \bar{x}_j$ je vážený průměr výběrových průměrů.

Nyní použijeme testovou statistiku $K = \sum_{j=1}^s \sum_{k=1}^s \left(\frac{n_{jk} - \frac{n_j \cdot n_k}{n}}{\frac{n_j \cdot n_k}{n}} \right)^2$, stejně jako u testu nezávislosti. Lze dokázat, že při výše

uvedeném označení jsou statistiky Q a K totožné. Tedy test homogenity lze provést stejně jako test nezávislosti. Tato statistika se v případě platnosti nulové hypotézy asymptoticky řídí rozložením $\chi^2(s-1)$. Kritický obor: $W = \chi^2_{s-1, \alpha}$. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α , když $K > W$.

Příklad: 104 náhodně vybraných matek bylo dotázáno, zda jejich kojeneček dostává dudlík. Zjišťoval se též nejvyšší stupeň dosaženého vzdělání matky.

Vzdělání matky	Počet matek	Počet dětí s dudlíkem
ZŠ	39	27
SŠ	47	34
VŠ	18	15

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že používání dudlíku nezávisí na vzdělání matky. (Jedná se o příklad 8.6.2. ze skript Základní statistické metody. Zde je uvedeno, že testová statistika Q se realizuje hodnotou 1,267, kritický obor je $W_{=5992}_{\infty}$, tedy nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.)

Řešení: Data zapíšeme do kontingenční tabulky 2 x 3.

	Matka ZŠ	Matka SŠ	Matka VŠ	$n_{.i}$
Dudlík ano	27	34	15	76
Dudlík ne	12	13	3	28
$n_{.k}$	39	47	18	104

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_{11}n_{1.}}{n} = \frac{76 \cdot 39}{104} = 28,5, \frac{n_{12}n_{1.}}{n} = \frac{76 \cdot 47}{104} = 34,5, \frac{n_{13}n_{1.}}{n} = \frac{76 \cdot 18}{104} = 13,5, \frac{n_{21}n_{2.}}{n} = \frac{28 \cdot 39}{104} = 10,5, \frac{n_{22}n_{2.}}{n} = \frac{28 \cdot 47}{104} = 12,5, \frac{n_{23}n_{2.}}{n} = \frac{28 \cdot 18}{104} = 4,8$$

Podmínky dobré aproximace jsou splněny, pouze v 1 případě ze 6 je teoretická četnost menší než 5.

Dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{27^2}{29 \cdot 39} + \frac{34^2}{13 \cdot 47} + \frac{3^2}{3 \cdot 18} = 1,268$$

Kritický obor: $W_{=2,1} s_{=1, \infty} \chi^2_{0,05} 2, \infty = 5992_{\infty}$

Na asymptotické hladině významnosti 0,05 se tedy neprokázalo, že používání dudlíku závisí na vzdělání matky.

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:

$$K = \frac{n \cdot \text{nad}^2}{(a+b)(c+d)(a+c)(b+d)}$$

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

Kritický obor: $W = [1, \infty)$

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 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 četno
1	obezn	obezn	1
2	obezn	neobe	9
3	neobe	obezn	7
4	neobe	neobe	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-kv	sv	p
Pearsonuv chi-kv	6,410	df=	p=,01
M-V chi-kvadr.	6,548	df=	p=,01
Yatesuv chi-kv.	5,048	df=	p=,02
Fisheruv přesny, 2-stranny			p=,01 p=,02
McNemaruv chi-k (B/C)	,2647 ,0625	df=	p=,60 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. Na problém lze pohlížet tak, že máme dva nezávislé výběry z alternativních rozložení, první má rozsah $n_1 = a+c$ a pochází z rozložení $A(\underline{q})$, druhý má rozsah $n_2 = b+d$ a pochází z rozložení $A(\underline{q})$. Testujeme hypotézu $H_0: \underline{q} = 0$ proti oboustranné alternativě. V kapitole o hodnocení náhodných výběrů z alternativních rozložení jsme použili

testovou statistiku $T_0 = \frac{M_1 - M_2}{\sqrt{M_1(1-M_1) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$, která se za platnosti nulové hypotézy asymptoticky řídí rozložením $N(0,1)$.

(M_* je vážený průměr výběrových průměrů.) Nyní použijeme testovou statistiku $K = \frac{n_1 a^2 + n_2 b^2}{n_1 + n_2}$, stejně jako u testu nezávislosti. Tato statistika se v případě platnosti nulové hypotézy asymptoticky řídí rozložením $\chi^2(1)$. Kritický obor: $W = [1, \infty)$. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α , když $K \geq W$.

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:

X onemocnění chřipkou	Y existence očkování		n _{j.}
	ano	ne	
ano	20	80	100
ne	220	140	360
n _k	240	220	460

$$\frac{n_1 n_1}{n} = \frac{100 \cdot 240}{460} = 51,7, \quad \frac{n_1 n_2}{n} = \frac{100 \cdot 220}{460} = 45,83$$

$$\frac{n_2 n_1}{n} = \frac{360 \cdot 240}{460} = 187,83, \quad \frac{n_2 n_2}{n} = \frac{360 \cdot 220}{460} = 172,17$$

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

Realizace testové statistiky:

$$K = \sum_{j=1}^k \frac{(n_{kj} - \frac{n_{k.} n_{.j}}{n})^2}{\frac{n_{k.} n_{.j}}{n}} = \frac{460(20 - 51,7)^2}{240} + \frac{460(80 - 45,83)^2}{220} = 30,7$$

Kritický obor: $W = \chi_{1-\alpha}^2(1) = \chi_{0,99}^2(1) = 3,84$

Protože $K > W$, 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 provedeme výpočet pomocí statistiky $T_0 = \frac{M_1 - M_2}{\sqrt{M_1(1-M_1) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$, která se v případě platnosti nulové hypotézy

asymptoticky řídí rozložením $N(0,1)$.

Přitom očkovaných bylo 240, z nich onemocnělo 20, neočkovaných bylo 220, z nich onemocnělo 80.

V našem případě tedy $n_1 = 240$, $n_2 = 220$, $m_1 = \frac{20}{240}$, $m_2 = \frac{80}{220}$, $m = \frac{20+80}{240+220} = \frac{100}{460} = \frac{5}{23}$.

Ověření podmínek $n_1 \hat{q} (1 - \hat{q}) > 9$ a $n_2 \hat{q} (1 - \hat{q}) > 9$: Parametry q_1 a q_2 neznáme, nahradíme je odhady m_1 a m_2 , tedy $20 \cdot (1 - 20/240) = 18,333 > 9$, $80 \cdot (1 - 80/220) = 50,909 > 9$.

Realizace testového kritéria:

$$t_0 = \frac{m_1 - m_2}{\sqrt{m \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{\frac{20}{240} - \frac{80}{220}}{\sqrt{\frac{5}{23} \left(\frac{1}{240} + \frac{1}{220} \right)}} = -2,80$$

Kritický obor je $W = (-\infty, u_{1-0,995}] \cup [u_{0,995}, \infty) = (-\infty, -2,5758] \cup [2,5758, \infty)$. Protože testové kritérium patří do kritického oboru, H_0 zamítáme na asymptotické hladině významnosti 0,05.

Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{a}{c} \cdot \frac{d}{b}$, která se nazývá výběrový **podíl šancí** (odds ratio). Považujeme ho za odhad neznámého teoretického podílu šancí $\frac{a}{c} : \frac{b}{d}$. 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

$$OR = \frac{a}{c} \cdot \frac{d}{b}$$

Jsou-li veličiny X, Y nezávislé, pak $\pi_{11} = \pi_{10} \cdot \pi_{01}$, tudíž teoretický podíl šancí $\frac{a}{c} = \frac{b}{d}$. Závislost veličin X, Y bude tím silnější, čím více se $\frac{a}{c}$ bude lišit od 1. Avšak $\frac{a}{c} \cdot \frac{d}{b} = 1$, tedy hodnoty $\frac{a}{c}$ 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. $I_{\alpha, \alpha}$) proti alternativě $H_1: X, Y$ nejsou stochasticky nezávislé náhodné veličiny (tj. $I_{\alpha, \alpha}$).

Testová statistika $T_0 = \frac{\ln OR}{\sqrt{\frac{1}{a+b+c+d}}}$ 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 $\ln OR$, který je dán vzorcem:

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

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:

$OR = \frac{17}{39} \cdot \frac{58}{11} = 2,29$. 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:

$$\ln OR = \ln 2,29 = 0,832,$$

$$\sqrt{\frac{1}{a+b+c} + \frac{1}{d+e+f} + \frac{1}{g+h+i}} = \sqrt{\frac{1}{17+11+39} + \frac{1}{11+58+97}} = 0,139,$$

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

$$\ln d = \ln OR - \frac{1}{\sqrt{\frac{1}{a+b+c} + \frac{1}{d+e+f} + \frac{1}{g+h+i}}} = 0,832 - \frac{1}{0,139} = -0,28,$$

$$\ln h = \ln OR + \frac{1}{\sqrt{\frac{1}{a+b+c} + \frac{1}{d+e+f} + \frac{1}{g+h+i}}} = 0,832 + \frac{1}{0,139} = 1,692.$$

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:

$$\frac{n_1 n_{11}}{n} = \frac{28 \cdot 17}{125} = 3,808, \quad \frac{n_1 n_{12}}{n} = \frac{28 \cdot 11}{125} = 2,464,$$

$$\frac{n_2 n_{21}}{n} = \frac{97 \cdot 17}{125} = 13,216, \quad \frac{n_2 n_{22}}{n} = \frac{97 \cdot 11}{125} = 8,5424$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n \sum_{j=1}^2 \sum_{k=1}^2 \frac{f_{kj}^2}{n_{kj}} - \sum_{k=1}^2 n_{k.}^2}{n \sum_{k=1}^2 n_{k.} - \sum_{k=1}^2 n_{k.}^2} = \frac{12 \cdot 1758 - 39^2}{28 \cdot 125 - 97^2} = 0,595$$

Kritický obor: $W = \chi_{0,95,1, \infty}^2 = 3,841$.

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

Vypočteme ještě Cramérův koeficient: $V = \sqrt{\frac{K}{n-1}} = \sqrt{\frac{3,6953}{124}} = 0,171$

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. $H_0: o_p = 1$.

Pokud víme, že za prvních okolností je šance na úspěch vyšší než za druhých okolností, pak proti nulové hypotéze postavíme pravostrannou alternativu

$H_1: o_p > 1$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch pravostranné alternativy, když 100(1- α)% empirický asymptotický jednostranný interval spolehlivosti pro $\ln o_p$ 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

$H_1: o_p < 1$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch levostranné alternativy, když 100(1- α)% empirický asymptotický jednostranný interval spolehlivosti pro $\ln o_p$ neobsahuje číslo 0.

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

$H_1: o_p \neq 1$.

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α ve prospěch oboustranné alternativy, když 100(1- α)% empirický asymptotický oboustranný interval spolehlivosti pro $\ln o_p$ 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,
- 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: $OR = \frac{6 \cdot 13}{1 \cdot 39} = 2$. 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 θ 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	2
	DM	HM
1	0,575	5,365

Vidíme, že $0,575093 < \ln \theta < 5,365736$ 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:

$$=\log(19,5)-\text{sqrt}(1/6+1/4+1/1+1/13)*\text{VNormal}(0,95;0;1)$$

	1 DM
1	0,960

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:

$$=\log(19,5)+\text{sqrt}(1/6+1/4+1/1+1/13)*\text{VNormal}(0,95;0;1)$$

	1 HM
1	4,980

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.