Kapitola 7.: Analýza kontingenčních tabulek

Cíl kapitoly

Po prostudování této kapitoly budete umět
- provádět test nezávislosti v kontingenční tabulce
- hodnotit intenzitu závislosti dvou náhodných veličin nominálního typu pomocí Cramérova koeficientu
- provádět Fisherův přesný test ve čtyřpolní kontingenční tabulce a počítat podíl šancí na úspěch za dvojích různých podmínek

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 9 hodin studia.

7.1. 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.

7.2. Testování hypotézy o nezávislosti

7.2.1. Popis testu

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 $y_{[1]}, ..., y_{[s]}$. Pořídíme dvourozměrný náhodný výběr rozsahu n z rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor (X, Y). Zjištěné absolutní 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

Testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$K = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$. Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$.

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

7.2.2. Podmínky dobré aproximace

Výraz $\frac{n_{j.}n_{.k}}{n}$ se nazývá teoretická četnost. Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti 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.

7.2.3. 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.

7.2.4. 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 a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

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

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

Řešení:

Nejprve vypočteme podle vzorce $\frac{n_{j.}n_{.k}}{n}$ všech 12 teoretických četností:

$$\frac{n_{1.}n_{.1}}{n} = \frac{140\cdot 90}{360} = 35,\ \frac{n_{1.}n_{.2}}{n} = \frac{140\cdot 100}{360} = 38{,}9,\ \frac{n_{1.}n_{.3}}{n} = \frac{140\cdot 60}{360} = 23{,}3,\ \frac{n_{1.}n_{.4}}{n} = \frac{140\cdot 110}{360} = 42{,}8,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{110\cdot 90}{360} = 27{,}5,\ \frac{n_{2.}n_{.2}}{n} = \frac{110\cdot 100}{360} = 30{,}6,\ \frac{n_{2.}n_{.3}}{n} = \frac{110\cdot 60}{360} = 18{,}3,\ \frac{n_{2.}n_{.4}}{n} = \frac{110\cdot 110}{360} = 33{,}6,$$

$$\frac{n_{3.}n_{.1}}{n}=\frac{110\cdot 90}{360}=27{,}5,\ \frac{n_{3.}n_{.2}}{n}=\frac{110\cdot 100}{360}=30{,}6,\ \frac{n_{3.}n_{.3}}{n}=\frac{110\cdot 60}{360}=18{,}3,\ \frac{n_{3.}n_{.4}}{n}=\frac{110\cdot 110}{360}=33{,}6$$

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:

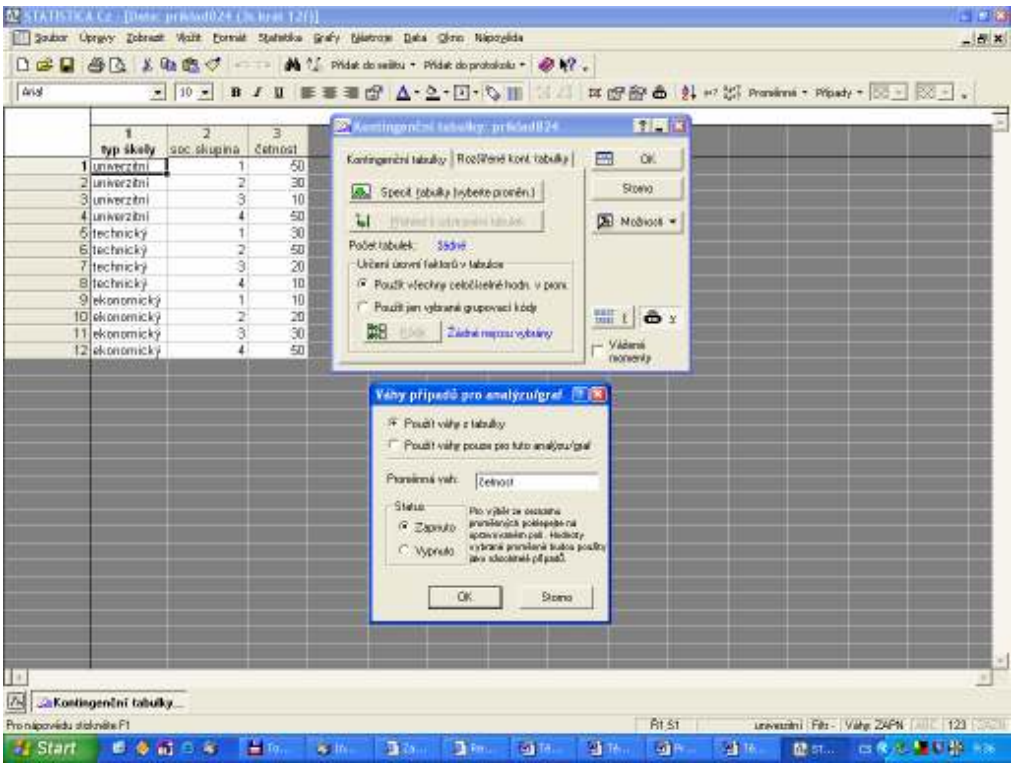
$$W=\left\langle \chi^2_{1-\alpha}((r-1)(s-1)),\infty\right)=\left\langle \chi^2_{0{,}95}((4-1)(3-1)),\infty\right)=\left\langle \chi^2_{0{,}95}(6),\infty\right)=\langle 12{,}6,\infty)$$

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.

Řešení pomocí systému STATISTICA:
Otevřeme nový datový soubor o 12 případech a třech proměnných TYP ŠKOLY, SOC. SKUPINA, ČETNOST). Do proměnné TYP ŠKOLY napíšeme varianty typu školy $x_{[1]} = 1$ (univerzitní), $x_{[2]} = 2$ (technický), $x_{[3]} = 3$ (ekonomický), přičemž každá varianta se objeví čtyřikrát pod sebou. Do proměnné SOC. SKUPINA napíšeme třikrát pod sebe všechny varianty $y_{[1]} = 1$, $y_{[2]} = 2$, $y_{[3]} = 3$, $y_{[4]} = 4$. Do proměnné ČETNOST napíšeme absolutní četnosti jednotlivých dvojic variant $(x_{[j]}, y_{[k]})$.
Statistika – Základní statistiky/tabulky – Kontingenční tabulky – OK – klikneme myší na tlačítko s obrázkem závaží – Status zapnuto – Proměnná vah ČETNOST – OK – Specif. tabulky – List 1 TYP ŠKOLY - List 2 SOC: SKUPINA – OK.



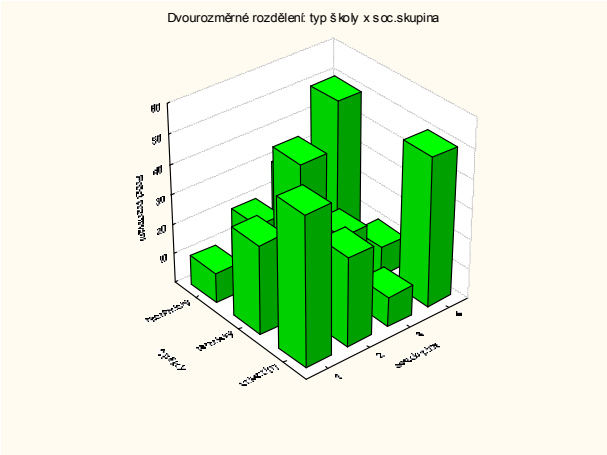
Přesvědčíme se o splnění podmínek dobré aproximace. Na záložce Možnosti zaškrtneme Očekávané četnosti, zvolíme Výpočet. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (priklad824) Četnost označených buněk > 10 Pearsonův chí-kv. : 76,8359, sv=6, p=,000000					
typ školy	soc.skupina 1	soc.skupina 2	soc.skupina 3	soc.skupina 4	Řádk. součty
univerzitní	35,00000	38,8889	23,33333	42,7778	140,0000
technický	27,50000	30,5556	18,33333	33,6111	110,0000
ekonomický	27,50000	30,5556	18,33333	33,6111	110,0000
Vš.skup.	90,00000	100,0000	60,00000	110,0000	360,0000

Vidíme, že všechny teoretické četnosti jsou dostatečně velké, větší než 5. V tabulce je rovněž uvedena realizace testové statistiky K = 76,8359, počet stupňů volnosti = 6. Odpovídající p-hodnota je blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny, z níž uchazeč pochází.
Dále vypočteme Cramérův koeficient. Na záložce Možnosti zaškrtneme Fí & Cramérovo C& V. Přejdeme na záložku Detailní výsledky a vybereme Detailní 2-rozm. tabulky.

	Statist. : typ školy(3) x soc.skupina(4) (priklad824)		
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		

V této tabulce najdeme Cramérův koeficient V = 0,3266749 a také hodnotu testové statistika K s počtem stupňů volnosti 6 a odpovídající p-hodnotou blízkou 0.
Výpočet ještě doplníme grafickým znázorněním simultánních absolutních četností proměnných TYP ŠKOLY a SOC. SKUPINA. Na záložce Detailní výsledky zvolíme 3D histogramy.



Poznámka: Graf lze různě natáčet, stačí v menu vybrat Formát – Vš. možnosti – Zorný bod.

7.3. 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

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 $K \in W$.

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

7.3.1. Příklad

Očkování proti chřipce se zúčastnilo 460 dospělých osob, 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:

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

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

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.

Výpočet pomocí systému STATISTICA:
Vytvoříme datový soubor o třech proměnných X (varianty 0 – neonemocněl chřipkou, 1 – onemocněl), Y (varianty 0 – neočkován, 1 – očkován) a četnost a čtyřech případech:

	1 X	2 Y	3 cetnost
1	neonemocněl	neočkován	140
2	neonemocněl	očkován	220
3	onemocněl	neočkován	80
4	onemocněl	očkován	20

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:

	Statistics: X(2) x Y(2) (chripka.sta		
Statistic	Chi-square	df	p
Pearson Chi-square	53,00842	df=1	p=,00000
M-L Chi-square	55,60618	df=1	p=,00000
Yates Chi-square	51,37366	df=1	p=,00000
Fisher exact, one-tailed			p=,00000
two-tailed			p=,00000
McNemar Chi-square (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 velmi blízká 0, tedy na hladině významnosti 0,01 zamítáme hypotézu, že onemocnění a očkování spolu nesouvisí.

7.3.2. Podíl šancí

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:

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

Jestliže interval spolehlivosti neobsahuje 1, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

7.3.3. Příklad

U 125 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 a dojem u přijímací zkoušky jsou nezávislé veličiny.

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

Řešení:

$OR = \frac{ad}{bc} = \frac{17 \cdot 58}{11 \cdot 39} = 2{,}298$. 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.

Řešení pomocí systému STATISTICA:

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(2,298)-sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))

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

=exp(log(2,298)+sqrt(1/17+1/11+1/39+1/58)*VNormal(0,975;0;1))

	1 DM	2 HM
1	0,972244	5,431562

Vidíme, že 95% asymptotický interval spolehlivosti (0,972; 5,433) pro podíl šancí obsahuje číslo 1, tedy nelze na asymptotické hladině významnosti 0,05 zamítnout hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

7.3.4. 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.

7.3.5. 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: $OR = \frac{ac}{bd} = \frac{6 \cdot 13}{4 \cdot 1} = \frac{39}{2} = 19{,}5$. 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.

Shrnutí

Při testování hypotézy o nezávislosti dvou náhodných veličin nominálního typu vycházíme z *kontingenční tabulky* sestrojené na základě znalosti náhodného výběru rozsahu n z dvourozměrného diskrétního rozložení. Používáme testovou statistiku, která se za splnění *podmínek dobré aproximace* asymptoticky řídí Pearsonovým χ^2 – rozložením. Intenzitu závislosti daných dvou veličin hodnotíme pomocí *Cramérova koeficientu*.

Speciální postavení mezi kontingenčními tabulkami mají *čtyřpolní kontingenční tabulky*, které dostaneme v tom případě, kdy obě sledované náhodné veličiny mají alternativní rozložení. Pro testování nezávislosti v těchto tabulkách existuje *exaktní Fisherův test*. Zajímavou informací, kterou můžeme vyčíst ze čtyřpolních tabulek, je *podíl šancí*. Pokud by šance na úspěch za jedněch i druhých okolností byly přibližně stejné, pak by podíl šancí byl blízký 1. Pomocí asymptotického intervalu spolehlivosti pro podíl šancí lze rovněž testovat hypotézu o nezávislosti veličin X a Y.

Kontrolní otázky

1. Jak testujeme nezávislost nominálních veličin? Jaké podmínky musí být splněny?
2. K čemu slouží Cramérův koeficient? Jaký je význam jeho hodnot?
3. Jak je definována teoretická četnost?
4. Jaká je struktura čtyřpolní kontingenční tabulky?
5. K čemu slouží Fisherův přesný test?
6. Jak lze interpretovat podíl šancí?

Autokorekční test

1. Ve čtyřpolní kontingenční tabulce jsou uvedeny tyto absolutní četnosti: a = 5, b = 3, c = 6, d = 4. Podíl šancí je
a) 1,11
b) 0,625
c) 0,9

2. Náhodná veličina X nabývá tří variant, náhodná veličina Y pak čtyř variant. Na základě náhodného výběru rozsahu 120 z dvourozměrného diskrétního rozložení, jímž se řídí náhodný vektor (X, Y), byla vypočítána realizace testové statistiky K = 12,737. Jak vypadá kritický obor pro test hypotézy o nezávislosti veličin X a Y, pokud volíme asymptotickou hladinu významnosti 0,01?
a) $W = \langle 16{,}8119; \infty)$
b) $W = \langle 0; 16{,}8119 \rangle$
c) $W = \langle 12{,}5916, \infty)$

3. Cramérův koeficient pro zadání z otázky 2 je
a) 0,0531
b) 0,2304
c) 0,1881

4. Dolní mez 95% asymptotického intervalu spolehlivosti pro podíl šancí z otázky 1 je
a) 0,1332
b) -2,0157
c) 6,0798

5. Je dána kontingenční tabulka:

	y	$y_{[1]}$	$y_{[2]}$	$y_{[3]}$	$y_{[4]}$
x	n_{jk}				
$x_{[1]}$		8	8	5	9
$x_{[2]}$		17	5	9	6

Jaká je teoretická četnost pro dvojici variant ($x_{[2]}$, $y_{[2]}$)?
a) 7,7313
b) 7,1791
c) 5,8209

Správné odpovědi: 1c) 2a) 3b) 4a) 5b)

Příklady

1. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví a vypočtěte Cramérův koeficient, jsou-li k dispozici následující údaje:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

Výsledek:
Podmínky dobré aproximace jsou splněny, pouze jedna teoretická četnost klesne pod 5. Testová statistika se realizuje hodnotou 3,5 , počet stupňů volnosti = 2, kritický obor je $W = \langle 5{,}991; \infty \rangle$. Hypotézu o nezávislosti pohlaví a pedagogické hodnosti tedy nezamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient V = 0,187.

2. 100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

pohlaví	nápoj	
	A	B
muž	20	30
žena	30	20

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.
Výsledek:
Vypočtená p-hodnota je 0,07134. Protože p-hodnota je větší než 0,05, nezamítáme na hladině významnosti hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

3. Z rakouské statistické ročenky z roku 1968 jsou převzaty údaje o rodinném stavu ženicha a nevěsty před sňatkem:

ženich	nevěsta		
	svobodná	ovdovělá	rozvedená
svobodný	44833	391	2343
ovdovělý	978	581	560
rozvedený	3405	375	2535

Na asymptotické hladině významnosti 0,01 testujte hypotézu, že rodinný stav ženicha a nevěsty jsou nezávislé. Vypočtěte rovněž Cramérův koeficient.
Výsledek:
Podmínky dobré aproximace jsou splněny. Testová statistika se realizuje hodnotou 15557,67, počet stupňů volnosti = 4, kritický obor je $W = \langle 13{,}2767; \infty \rangle$. Hypotézu o nezávislosti rodinného stavu ženicha a nevěsty tedy zamítáme na asymptotické hladině významnosti 0,01. Cramérův koeficient V = 0,377.

4. Zkouškovou písemku ze statistiky psalo 37 studentů. Z 16 mužů uspělo 9, z 21 žen uspělo 12. Vypočtěte podíl šancí na úspěch a pomocí asymptotického intervalu spolehlivosti pro podíl šancí testujte na asymptotické hladině významnosti 0,05 hypotézu, že úspěch u zkoušky ze statistiky nezávisí na pohlaví studenta.
Výsledek:
Podíl šancí na úspěch pro muže a pro ženy je 0,96, tedy muži mají nepatrně nižší šanci na úspěch než ženy. 95% asymptotický interval spolehlivosti pro podíl šancí je (0,2595; 3,5827). Jelikož tento interval obsahuje 1, hypotézu o nezávislosti úspěchu a pohlaví nezamítáme na asymptotické hladině významnosti 0,05.