Kontingenční tabulky a testy nezávislosti nominálních veličin

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

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

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

$n_{j.} = n_{j1} + ... + n_{js}$ 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: $W = \left\langle \chi^2_{1-\alpha}\left((r-1)(s-1)\right), \infty\right)$.

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

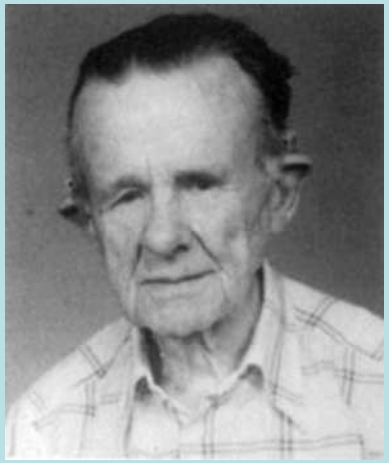
Podmínky dobré aproximace

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

Měření síly závislosti

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

Význam hodnot Cramérova koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.



Carl Harald Cramér (1893 – 1985): Švédský matematik

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 $K \in W$.

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

Statistický software poskytuje p-hodnotu pro Fisherův přesný test. Jestliže vyjde $p \leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α.
Fisherův test se používá při malých rozsazích výběrů (pokud $n \leq 20$ nebo pokud $20 < n \leq 40$ a některá z teoretických četností je menší než 5).

Podíl šancí ve čtyřpolní kontingenční tabulce

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 ku počtu neúspěchů za okolností I je $\frac{a}{c}$ (šance na úspěch za okolností I).

Poměr počtu úspěchů ku počtu neúspěchů za okolností II je $\frac{b}{d}$ (šance na úspěch za okolností II).

Podíl těchto dvou poměrů je podíl šancí: $OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$.

Pokud OR =1, pak okolnosti nemají vliv na výskyt jevu.
Pokud OR > 1, pak za okolností I je vyšší šance na výskyt jevu než za okolností II.
Pokud OR < 1, pak za okolností I je nižší šance na výskyt jevu než za okolností II.

Podíl šancí považujeme za odhad neznámého teoretického podílu šancí $o\rho = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$.

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

Interval spolehlivosti pro podíl šancí

Logaritmus teoretického podílu šancí oρ má přibližně normální rozložení a směrodatná odchylka jeho odhadu, tj. logaritmu podílu šancí OR, je $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$.

Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ln oρ jsou

$$\ln \mathrm{OR} - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2}, \ln \mathrm{OR} + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,u_{1-\alpha/2}.$$

Odlogaritmováním dostaneme meze 100(1-α)% asymptotického intervalu spolehlivosti pro oρ:

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

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. $o\rho = 1$) proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $o\rho \neq 1$).

Testování nezávislosti lze provést pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro podíl šancí $o\rho$:

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

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

Relativní riziko ve čtyřpolní kontingenční tabulce

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 za okolností I ku celkovému počtu úspěchů a neúspěchů je $\frac{a}{a+c}$ (riziko úspěchu za okolností I).

Poměr počtu úspěchů za okolností II ku celkovému počtu úspěchů a neúspěchů je $\frac{b}{b+d}$ (riziko úspěchu za okolností II).

Podíl těchto dvou rizik je relativní riziko: $RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{a(b+d)}{b(a+c)}$.

Pokud RR =1, pak okolnosti nemají vliv na výskyt jevu.
Pokud RR > 1, pak okolnosti I zvyšují četnost výskytu jevu.
Pokud RR < 1, pak okolnosti I snižují četnost výskytu jevu.

Relativní riziko považujeme za odhad neznámého teoretického relativního rizika $r\rho = \frac{\pi_{11}\pi_{\bullet 2}}{\pi_{12}\pi_{\bullet 1}}$.

Interval spolehlivosti pro relativní riziko

Logaritmus teoretického rizika rρ má přibližně normální rozložení a směrodatná odchylka jeho odhadu, tj. logaritmu relativního rizika RR, je $\sqrt{\frac{1}{a}-\frac{1}{a+c}+\frac{1}{b}-\frac{1}{b+d}}=\sqrt{\frac{c}{a(a+c)}+\frac{d}{b(b+d)}}$.

Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ln rρ jsou

$$\ln RR-\sqrt{\frac{c}{a(a+c)}+\frac{d}{b(b+d)}}u_{1-\alpha/2},\ln RR+\sqrt{\frac{c}{a(a+c)}+\frac{d}{b(b+d)}}u_{1-\alpha/2}.$$

Odlogaritmováním dostaneme meze 100(1-α)% asymptotického intervalu spolehlivosti pro rρ:

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

Testování nezávislosti ve čtyřpolních tabulkách pomocí relativního rizika

Na asymptotické hladině významnosti α testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $r\rho = 1$) proti alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $r\rho \neq 1$).

Testování nezávislosti lze provést pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro relativní riziko $r\rho$:

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

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