Kontingenční tabulky

Používají se k testování závislosti veličin, které se neměří kvantitativně, ale nabývají jen několika málo hodnot a zjišťuje se příslušnost či nepříslušnost daných hodnot.

Příklad: pacient: muž - žena

léčil se - neléčil se

zemřel - nezemřel

Do tabulek se zaznamenávají napozorované četnosti (třeba počty pacientů) s danou kombinací znaků.

Nechť veličina Y nabývá hodnot $1, 2, \dots, r$ a veličina Z hodnot $1, \dots, c$.

Kontingenční tabulka vypadá takto:

	Z				
Y	1	2	$\dots$	c	Σ
1	n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1\cdot} = \sum_j n_{1j}$
2	n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2\cdot}$
$\vdots$					$\vdots$
r	n_{r1}	n_{r2}	$\dots$	n_{rc}	$n_{r\cdot}$
Σ	$n_{\cdot 1} = \sum_i n_{i1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot c}$	$n = \sum_i \sum_j n_{ij}$

n_{ij} je počet případů, kdy se ve výběru vyskytla dvojice (i, j), tj. platilo, že $Y_i = i$ a současně $Z = j$

tomu odpovídají (neznámé) pravděpodobnosti $p_{ij} = P(Y = i, Z = j)$

veličiny n_{ij} pak mají sdružené multinomické rozdělení s param. p_{ij}

a platí pro ně, že $\sum_i \sum_j \frac{(n_{ij} - n p_{ij})^2}{n p_{ij}}$ má rozdělení $\chi^2_{(r-1)(c-1)}$

asymptoticky pro $n \to \infty$

Veličiny Y a Z budou nezávislé, bude-li platit

$$p_{ij} = P(Y=i, Z=j) = P(Y=i)\,P(Z=j) = \Big(\sum_j p_{ij}\Big)\Big(\sum_i p_{ij}\Big) = p_{i.}\,p_{.j}$$

Hypotéza nezávislosti Y a Z se tedy formuluje jako $H_0: p_{ij} = p_{i.}\,p_{.j}$

Neznámé parametry se přitom nahradí relativními četnostmi

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, \quad \hat{p}_{.j} = \frac{n_{.j}}{n}$$

Testové kritérium tedy bude

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.}\,n_{.j}}{n}\right)^2}{\frac{n_{i.}\,n_{.j}}{n}} = n \sum_i \sum_j \frac{n_{ij}^2}{n_{i.}\,n_{.j}} - n \geq \chi^2_{(r-1)(c-1)}(\alpha).$$

Protože test je asymptotický, dá se použít, když $\frac{n_{i.}\,n_{.j}}{n} > 5$ pro vš. i, j

Speciálně ve čtyřpolních tabulkách bude

$$\chi^2 = n\,\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}\,n_{2.}\,n_{.1}\,n_{.2}} \sim \chi^2_1$$

Příklad: Údaje o nemocných mužích

	přežili	zemřeli	celkem
léčení	5	6	11
neléčení	3	4	7
celkem	8	10	18

$$\chi^2 = 18 \cdot \frac{(5 \cdot 4 - 6 \cdot 3)^2}{11 \cdot 7 \cdot 8 \cdot 10} = \frac{18 \cdot 4}{11 \cdot 7 \cdot 8 \cdot 10} = \frac{9}{770} < 3{,}84 = \chi^2_1(0{,}05)$$

Nedá se tedy, že by léčení mělo nějaký vliv na přežití či nepřežití pacienta. I když tedy to může být tím, že málo četností pro χ^2-test.

Co s tím?

Pacient může uvažovat takto: bude-li se léčit, má šanci na přežití 5:6

nebude-li se léčit, má šanci 3:4

Která šance je větší?

$$\frac{5:6}{3:4} = \frac{5\cdot 4}{3\cdot 6} = \frac{20}{18} > 1$$

Tedy lepší bude dál se léčit.

Podíl šancí obecně je veličina $b = \dfrac{n_{11}\, n_{22}}{n_{12}\, n_{21}}$ se nazývá interakce

a je odhadem teoretické interakce $\beta = \dfrac{p_{11}\, p_{22}}{p_{12}\, p_{21}}$.

Hypotéza nezávislosti se pak formuluje jako $H_0: \beta = 1$.

V testu je používají logaritmické interakce $d = \ln b$, $\delta = \ln \beta$

a hypotéza pak má tvar $H_0: \delta = 0$.

Platí, že veličina

$$D = \frac{d}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \qquad \left(= \frac{0,10536}{0,974679} = 0,108 \right)$$

má za platnosti hypotézy rozdělení (asymptoticky) $N(0,1)$.

H_0 se zamítá, když $|D| \geq u\left(\frac{\alpha}{2}\right)$. $\quad (= 1,959)$

Další příklad: Údaje o nemocných ženách

	přežily	zemřely	celkem
léčeny	6	3	9
neléčeny	9	5	14
celkem	15	8	23

interakce $b = \dfrac{6:3}{9:5} = \dfrac{30}{27} > 1$, tj. výhodnější je dál se léčit

Údaje o nemocných lidech (muži a ženy celkem)

	přežili	zemřeli	celkem
léčení	11	9	20
neléčení	12	9	21
celkem	23	18	41

interakce $k = \dfrac{11:9}{12:9} = \dfrac{99}{108} < 1$

Závěr: Pro muže i ženu je výhodnější se léčit, ale pro člověka je výhodnější se neléčit.

Čím to je? Podobně jako u korelačních koeficientů často nestačí uvažovat pouze závislost dvojic faktorů, ale všech. Tady máme vlastně 3 veličiny: léčba, přežití, pohlaví a měli bychom analyzovat trojrozměrnou kontingenční tabulku.