11 Kontingenční tabulky

11.1 Test nezávislosti

Nechť náhodný vektor $\boldsymbol{X} = (Y, Z)'$ má diskrétní rozdělení, přičemž veličina Y nabývá hodnot $1, \ldots, r$ a veličina Z hodnot $1, \ldots, c$. Označme

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

Předpokládejme, že se uskutečnil výběr o rozsahu n z tohoto rozdělení. Nechť n_{ij} je počet těch případů, kdy se ve výběru vyskytla dvojice (i, j). Náhodné veličiny n_{ij} mají pak sdružené *multinomické rozdělení* s parametrem n a s pravděpodobnostmi p_{ij}. Proti odstavci 10.1 máme tu nepodstatnou odlišnost, že pravděpodobnosti p_{ij} i empirické četnosti n_{ij} píšeme ve tvaru matic místo ve tvaru vektorů. Matici (n_{ij}) se říká *kontingenční tabulka*. Podobně jako v dřívějších analogických situacích píšeme

$$n_{i.} = \sum_j n_{ij}, \qquad n_{.j} = \sum_i n_{ij}.$$

Samozřejmě platí

$$n = \sum_i n_{i.} = \sum_j n_{.j} = \sum_i \sum_j n_{ij}.$$

Číslům $p_{i.}$ a $p_{.j}$ se říká *marginální pravděpodobnosti* a hodnotám $n_{i.}$ a $n_{.j}$ *marginální četnosti*. Matice pravděpodobností (p_{ij}) a kontingenční tabulka (n_{ij}) jsou uvedeny v tab. 11.1.

V praxi vzniká kontingenční tabulka tak, že se na statistických jednotkách sledují dva znaky. Tyto znaky mohou být svou povahou diskrétní a nabývat jen konečně mnoha hodnot (např. muž — žena, krevní skupina 0 — A — B — AB), nebo sami záměrně vytvoříme jen konečně mnoho kategorií (místo uvedení přesného reakčního času zkoumaných osob na nějaký podnět se zaznamená jako výsledek jen to, zda reakce byla rychlá či pomalá). Někdy může mít znak objektivně danou číselnou hodnotu (např. počet dětí), jindy přiřazená číselná hodnota zachovává uspořádání, ale nevypovídá o skutečné velikosti znaku (např. barvy uspořádáme do skupin podle rostoucí vlnové délky; přiřazeno však je nakonec jen pořadí barev,

Tabulka 11.1: Matice pravděpodobností a kontingenční tabulka

Matice pravděpodobností

	Z			
Y	1	$\cdots$	c	Σ
1	p_{11}	$\cdots$	p_{1c}	$p_{1.}$
$\cdots$	$\cdots\cdots$			$\cdots$
r	p_{r1}	$\cdots$	p_{rc}	$p_{r.}$
Σ	$p_{.1}$	$\cdots$	$p_{.c}$	1

Kontingenční tabulka

	Z			
Y	1	$\cdots$	c	Σ
1	n_{11}	$\cdots$	n_{1c}	$n_{1.}$
$\cdots$	$\cdots\cdots$			$\cdots$
r	n_{r1}	$\cdots$	n_{rc}	$n_{r.}$
Σ	$n_{.1}$	$\cdots$	$n_{.c}$	n

ne vlnová délka). V mnoha případech přiřazujeme čísla $1, 2, \ldots$ jen jako označení, ale nejde ve skutečnosti ani o přesné hodnoty ani o objektivní pořadí (např. uvedená čísla slouží jako kódy pro různé dělnické profese). Zde se omezíme na základní popis obecných kontingenčních tabulek a nebudeme se zabývat speciálními metodami, jimiž lze vytěžit další informaci v případě uspořádaných hodnot znaku nebo v případě znaků přímo vyjádřených číselně.

Nejčastější úlohou při rozboru kontingenčních tabulek je provedení testu hypotézy, že veličiny Y a Z jsou na sobě nezávislé. Než přistoupíme k odvození tohoto *testu nezávislosti*, dokážeme jedno přípravné tvrzení.

Věta 11.1 *Veličiny Y a Z jsou nezávislé tehdy a jen tehdy, platí-li $p_{ij} = p_{i.}p_{.j}$ pro všechny dvojice (i, j).*

Důkaz. Podle definice uváděné v teorii pravděpodobnosti jsou Y a Z nezávislé právě tehdy, platí-li $\mathsf{P}(Y \in A, Z \in B) = \mathsf{P}(Y \in A)\mathsf{P}(Z \in B)$ pro kterékoli množiny $A \subset \{1, \ldots, r\}$, $B \subset \{1, \ldots, c\}$. Zvolíme-li $A = \{i\}$, $B = \{j\}$, pak odtud plyne, že v případě nezávislosti veličin Y a Z musí platit $p_{ij} = p_{i.}p_{.j}$. Obrácenou implikaci dokážeme pro $A = \{1, 2\}$, $B = \{1, 2, 3\}$. Obecný případ se dokazuje úplně stejně, jen zápis je o něco složitější. Máme

$$\begin{aligned}\mathsf{P}(Y \in A, Z \in B) &= \sum_{i=1}^{2}\sum_{j=1}^{3} p_{ij} = \sum_{i=1}^{2}\sum_{j=1}^{3} p_{i.}p_{.j} \\ &= \left(\sum_{i=1}^{2} p_{i.}\right)\left(\sum_{j=1}^{3} p_{.j}\right) = \mathsf{P}(Y \in A)P(Z \in B). \qquad \square\end{aligned}$$

Proto hypotézu nezávislosti H_0 můžeme psát ve tvaru

$$H_0 : p_{ij} = p_{i.}p_{.j}, \qquad i = 1, \ldots, r;\ j = 1, \ldots, c.$$

To znamená, že pravděpodobnosti p_{ij} multinomického rozdělení jsou funkcemi menšího počtu neznámých parametrů, jimiž jsou marginální pravděpodobnosti $p_{i.}$ a $p_{.j}$. Ale tyto marginální pravděpodobnosti nejsou nezávislé, neboť

$$\sum_i p_{i.} = \sum_j p_{.j} = 1.$$

Máme-li splnit podmínku (4) věty 10.4, pak do neznámých parametrů nemůžeme počítat pravděpodobnosti $p_{r.}$ a $p_{.c}$, neboť ty lze už z ostatních marginálních pravděpodobností vypočítat. Počet neznámých parametrů je tudíž $m = r-1+c-1 = r+c-2$. Přitom se samozřejmě omezujeme jen na ty situace, kdy všechny marginální pravděpodobnosti jsou kladné. Kdyby tomu tak nebylo, prostě by se některé řádky nebo některé sloupce vynechaly.

K odhadu neznámých parametrů $p_{1.}, \dots, p_{r-1,.}$ a $p_{.1}, \dots, p_{.,c-1}$ použijeme soustavu rovnic (10.14). Místo veličin X_i zde samozřejmě máme veličiny n_{ij}. Dostáváme

$$\sum_{j=1}^{c} \left(\frac{n_{ij}}{p_{i.}} - \frac{n_{rj}}{p_{r.}} \right) = 0, \qquad i = 1, \dots, r-1 \tag{11.1}$$

a

$$\sum_{i=1}^{r} \left(\frac{n_{ij}}{p_{.j}} - \frac{n_{ic}}{p_{.c}} \right) = 0, \qquad j = 1, \dots, c-1, \tag{11.2}$$

protože pro $h = 1, \dots, r-1$ platí

$$\frac{\partial p_{i.} p_{.j}}{\partial p_{h.}} = \begin{cases} p_{.j} & \text{pro } i = h, \\ -p_{.j} & \text{pro } i = r, \\ 0 & \text{v ostatních případech.} \end{cases}$$

Podobný výsledek dostaneme pro parciální derivace podle $p_{.l}$. Vzorec (11.1) však platí i pro $i = r$ a (11.2) platí i pro $j = c$. Proto místo (11.1) můžeme psát

$$\frac{n_{i.}}{p_{i.}} - \frac{n_{r.}}{p_{r.}} = 0, \qquad i = 1, \dots, r. \tag{11.3}$$

Odtud

$$n_{i.} = \frac{n_{r.}}{p_{r.}} p_{i.}, \qquad i = 1, \dots, r.$$

Sečtením přes všechna i dostaneme $n = n_{r.}/p_{r.}$, takže odtud máme pro $p_{r.}$ odhad $\hat{p}_{r.} = n_{r.}/n$. Dosazením do (11.3) nakonec získáme odhady

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, \qquad i = 1, \dots, r.$$

Podobně se vyřeší i (11.2) a dostane se

$$\hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}, \qquad j = 1, \ldots, c.$$

Podle věty 10.4 má pak veličina

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} \tag{11.4}$$

asymptoticky rozdělení χ^2, jehož počet stupňů volnosti je roven

$$rc - (r + c - 2) - 1 = (r - 1)(c - 1).$$

Vzhledem k (10.8) můžeme χ^2 také počítat podle vzorce

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - n. \tag{11.5}$$

Když vyjde $\chi^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$, zamítáme hypotézu H_0 o nezávislosti veličin Y a Z. Ke shodě s limitním rozdělením se opět vyžaduje, aby všechny teoretické četnosti $n_{i\cdot} n_{\cdot j}/n$ byly větší než 5. Není-li tato podmínka splněna, spojují se některé řádky nebo některé sloupce.

Příklad 11.2 U 6800 mužů byla zjišťována barva očí a barva vlasů (viz Yule a Kendall 1950). Výsledky jsou uvedeny v tab. 11.2.

Tabulka 11.2: Barva očí a vlasů

	Barva vlasů				
Barva očí	Světlá	Kaštanová	Černá	Zrzavá	Celkem
Světle modrá	1 768	807	189	47	2 811
Šedá nebo zelená	946	1 387	746	53	3 132
Tmavohnědá	115	438	288	16	857
Celkem	2 829	2 632	1 223	116	6 800

Podle vzorce (11.4) nebo (11.5) vypočteme $\chi^2 = 1\,073{,}51$. Vzhledem k tomu, že $\chi^2 \geq \chi^2_6(0{,}05) = 12{,}59$, zamítneme hypotézu, že barva očí a barva vlasů u mužů jsou nezávislé znaky. $\diamondsuit$

11.2 Test homogenity multinomických rozdělení

Někdy se stává, že marginální řádkové četnosti $n_{i\cdot}$ jsou předem stanoveny. Pak ale i-tý řádek kontingenční tabulky $n_{i1} \quad \dots \quad n_{ic}$ má *multinomické rozdělení* s parametry $n_{i\cdot}$, $q_{i1}, \dots, q_{ic}$, kde $q_{i1}, \dots, q_{ic}$ jsou nějaké pravděpodobnosti splňující podmínku $q_{i1} + \cdots + q_{ic} = 1$. Většinou je pak třeba testovat *hypotézu homogenity*, která říká, že pravděpodobnosti $q_{i1}, \dots, q_{ic}$ nezávisejí na řádkovém indexu i [takže všechny řádky matice (q_{ij}) jsou stejné]. Na základě jistého zobecnění věty 10.4 lze dokázat (viz Cramér 1946), že i v tomto případě za platnosti hypotézy homogenity má veličina χ^2 počítaná podle vzorce (11.4) nebo (11.5) asymptoticky rozdělení $\chi^2_{(r-1)(c-1)}$. Hypotéza homogenity se zamítá v případě $\chi^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$.

Příklad 11.3 V severozápadním Skotsku byla provedena studie, která měla prokázat, zda je percentuelní zastoupení krevních skupin na celém území homogenní či nikoli. V oblasti Eskdale bylo náhodně vybráno 100 osob, v oblasti Annandale 125 osob a v oblasti Nithsdale 253 osob. Výsledky převzaté ze sbírky Osborn (1979) jsou uvedeny v tab. 11.3.

Tabulka 11.3: Krevní skupiny

Oblast	A	B	0	AB	Celkem
Eskdale	33	6	56	5	100
Annandale	54	14	52	5	125
Nithsdale	98	35	115	5	253
Celkem	185	55	223	15	478

Podle (11.5) dostaneme $\chi^2 = 10{,}45$. Toto číslo je menší než kritická hodnota $\chi^2_6(0{,}05) = 12{,}59$. Nemůžeme tudíž zamítnout hypotézu, že rozdělení krevních skupin je ve všech třech oblastech stejné. ♢

Všimněme si podrobněji testu homogenity v případě $c = 2$.

Věta 11.4 *Je-li $c = 2$, pak*

$$\chi^2 = \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \sum_{i=1}^{r} \frac{n_{i1}^2}{n_{i\cdot}} - n \frac{n_{\cdot 1}}{n_{\cdot 2}}. \tag{11.6}$$