Statistická analýza kompozičních tabulek

Kamila Fačevicová, Karel Hron

Katedra matematické analýzy a aplikací matematiky,
Univerzita Palackého v Olomouci

Od kontingenčních ke kompozičních tabulkám

- **Kontingenční tabulky:** vztah dvou faktorů zachycený v tabulce s diskrétními vstupy. Test nezávislosti

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}.$$

- **Kompoziční tabulky:** spojitá analogie kompozičních tabulek. Umožňují analýzu vztahu mezi faktory na základě výběru n tabulek.

$I \times J$ kompoziční tabulky

- Speciální případ $I \cdot J$-složkových kompozičních dat $\mathbf{x} = \mathcal{C}(x_{11}, \ldots, x_{1J}, \ldots, x_{I1}, \ldots, x_{IJ})$, které jsou přeuspořádány do tabulky o rozměru $I \times J$

$$\mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}.$$

- Nesou informaci o vztahu mezi řádkovým a sloupcovým faktorem.
- Příklad:

věk\ BMI	podváha	normální v.	nadváha	obezita
$25-44$	0.0144	0.2196	0.1410	0.0554
$45-64$	0.0022	0.1014	0.1792	0.0988
$65-84$	0.0014	0.0473	0.0900	0.0493

Tabulka : Rozdělení populace České republiky v roce 2008 podle věku a hodnoty BMI indexu.

$I \times J$ kompoziční tabulky

- Operace uzávěru:

$$\mathcal{C}(\mathbf{x}) = \begin{pmatrix} \frac{\kappa \cdot x_{11}}{\sum_{i,j=1}^{I,J} x_{ij}} & \cdots & \frac{\kappa \cdot x_{1J}}{\sum_{i,j=1}^{I,J} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa \cdot x_{I1}}{\sum_{i,j=1}^{I,J} x_{ij}} & \cdots & \frac{\kappa \cdot x_{IJ}}{\sum_{i,j=1}^{I,J} x_{ij}} \end{pmatrix}.$$

- Výběrovým prostorem je $I \cdot J$-složkový **simplex** dimenze $I \cdot J - 1$

$$\mathcal{S}^{IJ} = \{\mathbf{x} = (x_{11}, \ldots, x_{1J}, \ldots, x_{IJ}) | x_{ij} > 0,$$

$$i = 1, 2, \ldots, I, j = 1, 2, \ldots, J; \sum_{i,j=1}^{I,J} x_{ij} = \kappa\}.$$

Základní operace s kompozičními tabulkami - Aitchisonova geometrie

- **Perturbace:**

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C} \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix}.$$

- **Mocninná transformace:**

$$\alpha \odot \mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix}.$$

- **Skalární součin:**

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}}.$$

- **Aitchisonova délka a vzdálenost:**

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} \quad \text{a} \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a .$$

Rozklad kompozičních tabulek

- Pomocí projekcí mohou být kompoziční tabulky rozděleny na nezávislou a interakční část.

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int}.$$

- **Nezávislá tabulka:**

$$\mathbf{x}_{ind} = \left(x_{ij}^{ind} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} x_{kj} x_{il} \right)^{\frac{1}{IJ}} \right)_{i,j=1}^{I,J}.$$

Popisují jen vztah mezi řádky nebo sloupci. Analogie s případem nezávislosti v kontingenčních tabulkách.

- **Interakční tabulka:**

$$\mathbf{x}_{int} = \left(x_{ij}^{int} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} \frac{x_{ij}}{x_{kj} x_{il}} \right)^{\frac{1}{IJ}} \right)_{i,j=1}^{I,J}.$$

Popisují vztah mezi různými řádky a sloupci.

Analýza vztahu mezi faktory

Pokud jsou řádkový a sloupcový faktor nezávislé, pak

- veškerá informace o tabulce $\mathbf{x}$ je obsažena v její nezávislé části ($\mathbf{x} = \mathbf{x}_{ind}$),
- interakční tabulka je neutrální prvek (všechny prvky jsou si rovny),
- všechny ilr souřadnice tabulky $\mathbf{x}_{int}$ jsou nulové.

Vyjádření kompoziční tabulky v souřadnicích

Souřadnice D-složkových kompozičních dat představují $D-1$-rozměrný reálný vektor

$$\mathbf{z} = h(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a) = (z_1, z_2, \ldots, z_{D-1}),$$

kde $\mathbf{e}_i = \mathcal{C}(e_{i1}, \ldots, e_{i,D})$, $i = 1, \ldots, D-1$ tvoří ortonormální bázi D-složkového simplexu.

Pro tyto souřadnice platí, že

$$h(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathbf{z}_1 + \beta \cdot \mathbf{z}_2, \quad \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle.$$

Mezi různými systémy souřadnic existuje ortogonální vztah.

Vyjádření kompoziční tabulky v souřadnicích

Interakční tabulku lze převést do souřadnic pomocí:

- Postupného binárního dělení $\Rightarrow I \cdot J - 1$ obecně nenulových souřadnic.
- Vztahu

$$\frac{1}{\sqrt{r \cdot s \cdot (r-1) \cdot (s-1)}} \ln \prod_{i=1}^{r-1} \prod_{j=1}^{s-1} \frac{x_{ij} x_{rs}}{x_{is} x_{rj}} \qquad (1)$$

pro $r = 2, 3, \ldots, I$ a $s = 2, 3, \ldots, J \Rightarrow (I-1)(J-1)$ obecně nenulových souřadnic, zbývající nulové.

- Permutací řádků nebo sloupců tabulky a použitím vztahu (1) $\Rightarrow (I-1)(J-1)$ obecně nenulových souřadnic s novou interpretací.

Srovnání metod pro převod interakční tabulky do souřadnic

	Postupné binární dělení		Nová metoda
+	Interpretace pomocí bilancí	+	Souvislost s poměry šancí v tabulce $\mathbf{x}$.
−	Zdlouhavý výpočet o několika krocích.	+	Rychlý výpočet přímo z tabulky $\mathbf{x}$.
−	Více nenulových souřadnic, než je dimenze prostoru $\mathcal{S}^{I\times J}(\mathbf{x}_{int})$.	+	Počet nenulových souřadnic je roven dimenzi prostoru $\mathcal{S}^{I\times J}(\mathbf{x}_{int})$.
−	Vede k problémům se singularitou souřadnic.	+	Problém se singularitou je eliminován.

Tabulka : Srovnání metod pro vyjádření tabulky $\mathbf{x}_{int}$ v souřadnicích.

Příklad - vztah mezi věkem a indexem BMI

- Zabýváme se vztahem mezi věkem a BMI.
- K dispozici máme výběr kompozičních tabulek typu 3×4 popisujících rozdělení populace 18-ti evropských zemí podle věku a BMI v roce 2008.
- Faktor věk má 3 úrovně: $25 - 44$, $45 - 64$ a $65 - 84$ let.
- Faktor BMI má 4 úrovně: podváha, normální váha, nadváha a obezita.

věk\ BMI	podváha	normální v.	nadváha	obezita
$25 - 44$	0.0144	0.2196	0.1410	0.0554
$45 - 64$	0.0022	0.1014	0.1792	0.0988
$65 - 84$	0.0014	0.0473	0.0900	0.0493

Příklad - rozklad na nezávislou a interakční tabulku

- **Nezávislá tabulka pro Českou republiku:**

$$\mathbf{x}_{ind} = \begin{pmatrix} 0.0061 & 0.1716 & 0.2218 & 0.1090 \\ 0.0039 & 0.1090 & 0.1409 & 0.0692 \\ 0.0020 & 0.0569 & 0.0736 & 0.0361 \end{pmatrix}$$

- **Interakční tabulka pro Českou republiku:**

$$\mathbf{x}_{int} = \begin{pmatrix} 0.1813 & 0.0973 & 0.0483 & 0.0387 \\ 0.0444 & 0.0707 & 0.0967 & 0.1085 \\ 0.0541 & 0.0632 & 0.0930 & 0.1037 \end{pmatrix}$$

- **Průměrná tabulka** (ve smyslu Aitchisonovy geometrie):

$$\overline{\mathbf{x}}_{int} = \begin{pmatrix} 0.1472 & 0.0959 & 0.0585 & 0.0462 \\ 0.0550 & 0.0747 & 0.0907 & 0.1023 \\ 0.0599 & 0.0677 & 0.0915 & 0.1027 \end{pmatrix}$$

Prvky si nejsou rovny $\Rightarrow$ svědčí proti nezávislosti faktorů.

Příklad - převod interakční tabulky do souřadnic

$$z_{22}^{int} = \frac{1}{2} \ln \frac{x_{11}x_{22}}{x_{12}x_{21}}$$

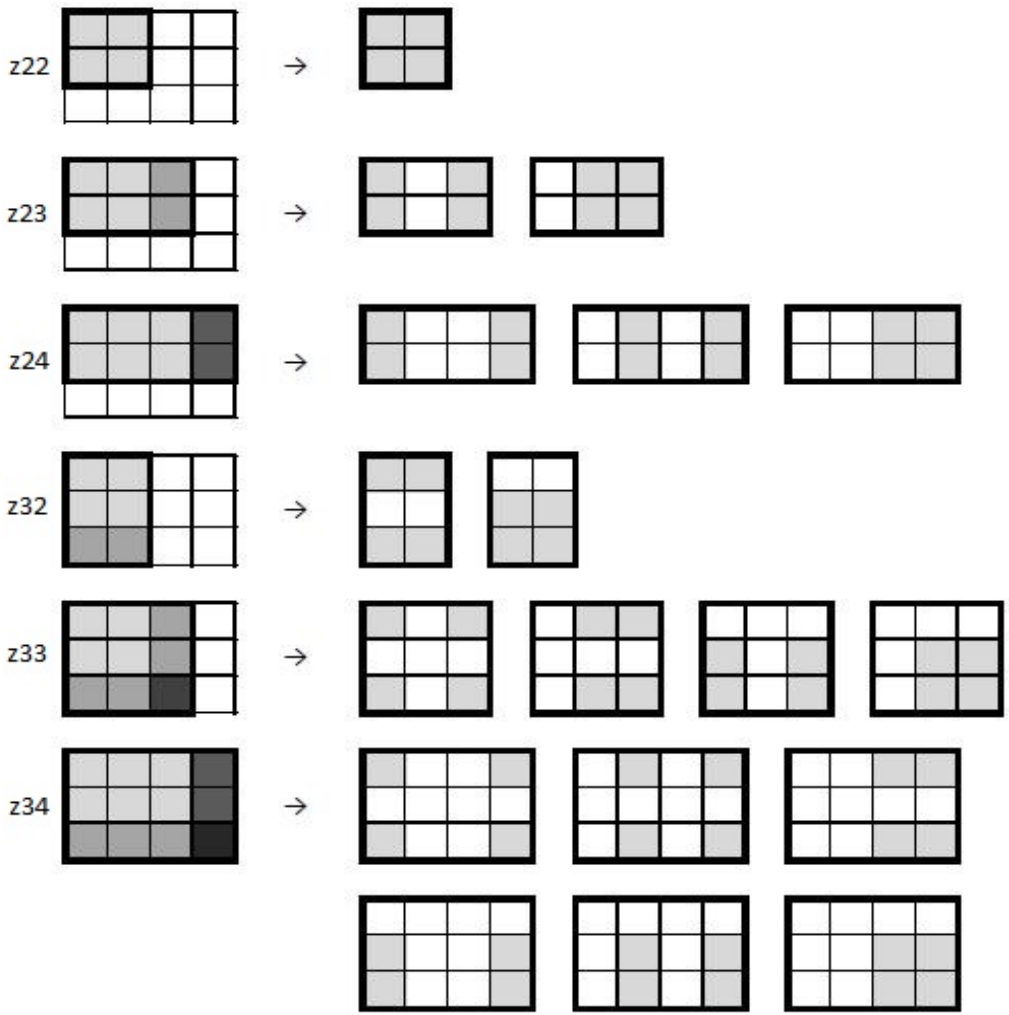
$$z_{23}^{int} = \frac{1}{\sqrt{12}} \ln \frac{x_{11}x_{12}x_{23}^2}{x_{21}x_{22}x_{13}^2}$$

$$z_{24}^{int} = \frac{1}{\sqrt{24}} \ln \frac{x_{11}x_{12}x_{13}x_{24}^3}{x_{21}x_{22}x_{23}x_{14}^3}$$

$$z_{32}^{int} = \frac{1}{\sqrt{12}} \ln \frac{x_{11}x_{21}x_{32}^2}{x_{31}^2x_{12}x_{22}}$$

$$z_{33}^{int} = \frac{1}{\sqrt{36}} \ln \frac{x_{11}x_{12}x_{21}x_{22}x_{33}^4}{x_{31}^2x_{32}^2x_{13}^2x_{23}^2}$$

$$z_{34}^{int} = \frac{1}{\sqrt{72}} \ln \frac{x_{11}x_{12}x_{13}x_{21}x_{22}x_{23}x_{34}^6}{x_{31}^2x_{32}^2x_{33}^2x_{14}^3x_{24}^3}$$

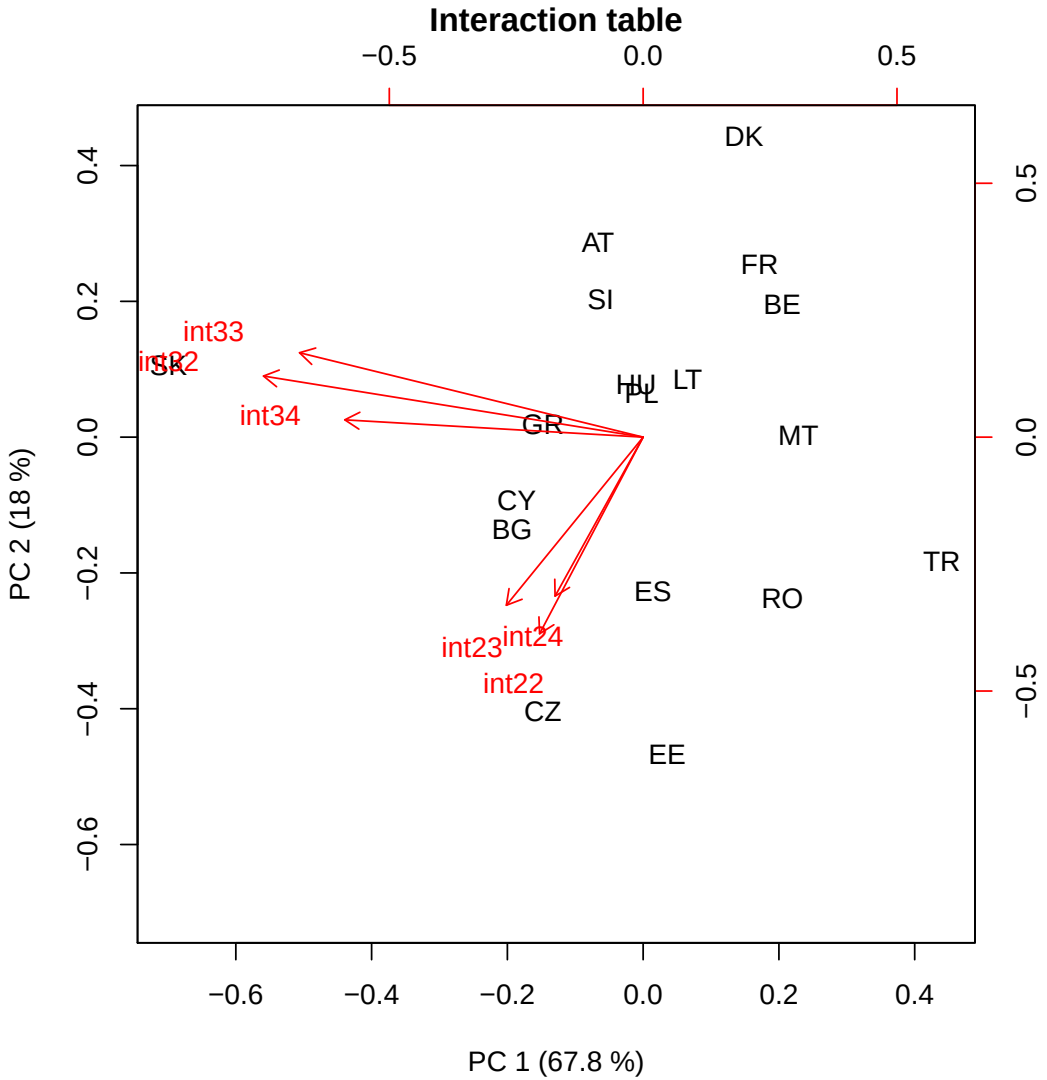


Příklad - výběrové charakteristiky

$\mathbf{z}_{int}$	z_{22}	z_{23}	z_{24}	z_{32}	z_{33}	z_{34}
V. průměr	0.3674	**0.6096**	**0.6494**	0.1057	0.3624	0.3783
V. s. odchylka.	0.1488	**0.1357**	**0.1426**	0.2412	0.2175	0.1945

- Opět svědčí proti nezávislosti věku a BMI.
- Nejvíce jsou od nuly vzdáleny souřadnice z_{23} a z_{24} (poměry šancí agregující podváhu a normální váhu, resp. podváhu, normální váhu a nadváhu).
- V případě náhodného výběru (nezávislosti kompozičních tabulek) lze nezávislost dále ověřit pomocí testování nulovosti souřadnic $\mathbf{z}_{int}$.

Příklad - grafická reprezentace souřadnic



Co najdete na posteru

- Definici kompozičních tabulek jako spojité analogie kontingenčních tabulek.
- Metodu analýzy vztahu mezi dvěma faktory s využitím výběru n tabulek.
- Rozklad tabulek na interakční a nezávislou část.
- Vztah pro výpočet souřadnic interakční tabulky.
- Interpretaci souřadnic ve smyslu poměru šancí.
- Příklad použití metody pro analýzu vztahu mezi věkem a BMI.

Reference

- J. Aitchison (1986) The Statistical Analysis of Compositional Data. Chapman and Hall, London.
- J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barceló-Vidal (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3), 279–300.
- J.J. Egozcue, J.L. Díaz-Barrero, V. Pawlowsky-Glahn (2008) Compositional analysis of bivariate discrete probabilities. In: J. Daunis-i-Estadella, J.A. Martín-Fernández, eds, Proceedings of CODAWORK'08, University of Girona, Spain.
- K. Fačevicová, K. Hron, V. Todorov, D. Guo, M. Templ (2013) Logratio approach to statistical analysis of 2×2 compositional tables. Journal of Applied Statistics, DOI: 10.1080/02664763.2013.856871.
- K. Fačevicová, K. Hron, V. Todorov (2014) Compositional tables analysis in coordinates. V přípravě.