

4 Mnohorozměrná data

Průvodce studiem

Tato kapitola má posloužit pro orientaci v problematice statistické analýzy vícerozměrných dat. Jsou zde uvedeny důležité výběrové charakteristiky, které lze z dat vyhodnotit. Také jsou zmíněny techniky ověřování předpokladů o rozdělení, zejména o normálním rozdělení a některé transformace, které lze užít v analýze dat. Počítejte se třemi hodinami studia s tím, že se k probírané látce budete ještě podle potřeby vracet.



Prozatím jsme se zabývali otázkami abstraktního popisu vztahů mezi náhodnými veličinami, především náhodným vektorem. Nyní obrátíme pozornost k praktičtějším problémům mnohorozměrných dat. Připomeňme, že veličiny, které zjišťujeme na sledovaných objektech, jsou různého typu. Podle oboru jejich hodnot rozlišujeme:

- veličiny spojité - mohou nabývat nespočetně mnoho hodnot, např. čas, délka atd.
- veličiny nespojité (diskrétní) - nabývají jen spočetně mnoho hodnot, v praxi jen konečného počtu a často několika málo možných hodnot, např. kategoriální veličiny, vyjadřující příslušnost k nějaké skupině (kategorií) objektů
- alternativní (dichotomické, binární) veličiny, patří mezi nespojité, ale tím, že mohou nabývat jen dvou možných hodnot, často interpretovaných jako ANO/NE, TRUE/FALSE nebo 1/0, bývá někdy užitečné nahlížet na ně jako na zvláštní typ veličin

Dále veličiny můžeme rozlišovat podle škály, ve které měříme:

- nominální (kategoriální)
- ordinální (pořadové)
- rozdílové (intervalové) kvantitativní (metrické, je definována vzdálenost dvou hodnot)
- poměrové kvantitativní (metrické, je definována vzdálenost dvou hodnot)

Kromě těchto hledisek, která klasifikují veličiny, je důležité mít i na paměti, jak vlastně data vznikla, co zobrazují a jaké jsou vztahy mezi pozorovanými objekty. Podstatné je, zda objekty můžeme považovat za nezávislé nebo zda vznikla jako řada pozorování téhož objektu v různých obdobích. Různé situace rozlišuje následující tabulka.

Úlohy řešené analýzou dat, časový prvek v datech:

počet objektů	počet veličin	počet období	typ úlohy
1	p	1	kasuistika, případová studie
n	1	1	jednorozměrná analýza
1	1	T	jednorozměrná časová řada
n	1	T	$T = 2$ párové srovnání, $T > 2$ opakovaná měření
n	p	1	vícerozměrná analýza dat
1	p	T	vícerozměrná časová řada
n	p	T	longitudinální studie

Poznámka: $n, p, T > 1$

Analýza vícerozměrných dat se většinou zabývá daty, kdy máme p veličin pozorovaných na n objektech. Rozlišit můžeme následující situace:

- všechny veličiny metrické – n bodů v \mathbf{R}^p
- všechny veličiny kategoriální – mnohorozměrné kontingenční tabulky
- smíšená data – datová matice se rozdělí na podsoubory – analogie s dvou/vícevýběrovými úlohami

Pro charakterizaci vícerozměrných dat potřebujeme odhadnout charakteristiky p -rozměrného vektoru náhodných veličin, tj. vektor středních hodnot a kovarianční matici (je symetrická), tedy určit následující počet výběrových charakteristik:

$$2p + \frac{p(p-1)}{2} = 2p + \frac{p^2 - p}{2} = \frac{p^2 + 3p}{2}$$

Počet odhadovaných parametrů tedy roste *kvadraticky* s počtem veličin, např. pro $p = 10$ je to 65 charakteristik, pro $p = 40$ je to už 860 charakteristik.

Pokud jsou veličiny kategoriální, potřebujeme odhadovat i sdruženou pravděpodobnostní funkci. Počet políček v tabulce roste exponenciálně s počtem veličin. Tudíž pokud mají být tyto odhady důvěryhodné, potřebujeme, aby vícerozměrná data byla měla dostatečný rozsah, tzn. n bylo velké.

4.1 Výběrové charakteristiky

Vícerozměrná dat jsou reprezentována datovou maticí $(n \times p)$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

tzn., že i –tý řádek datové matice je řádkový vektor pozorování p veličin na i –tém objektu.

Vektor průměrů jednotlivých veličin

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

tj.

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad j = 1, 2, \dots, p$$

Wishartova matice typu $(p \times p)$ má prvky

$$w_{jj'} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}), \quad j, j' = 1, 2, \dots, p$$

a můžeme ji zapsat také jako

$$\mathbf{W} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Výběrová kovarianční matice je opět typu $p \times p$

$$\mathbf{S} = \frac{1}{n-1} \mathbf{W},$$

její prvky jsou výběrové kovariance tj.

$$s_{jj'} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Korelační matice má prvky $r_{jj'} = s_{jj'}/(s_j s_{j'})$ tj. výběrové korelační koeficienty a má tvar

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$



4.2 Lineární transformace proměnných

Při analýze vícerozměrných dat je často výhodné pracovat s odvozenými veličinami, které vzniknou z původních lineární transformací, např. s centrovanými proměnnými

$$v_{ij} = x_{ij} - \bar{x}_j,$$

pro které vektor průměrů je nulový, $\bar{\mathbf{v}} = \mathbf{0}$, a kovarianční a korelační matice se nezmění, tzn.

$$\mathbf{S}_v = \mathbf{S}_x, \quad \mathbf{R}_v = \mathbf{R}_x.$$

Další často užívanou transformací je normování. Normované hodnoty proměnných vzniknou vycentrováním a vydělením směrodatnou odchylkou původních proměnných, tj.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, p.$$

Pak vektor průměrů je nulový $\bar{\mathbf{z}} = \mathbf{0}$, všechny rozptyly a směrodatné odchylky normovaných proměnných jsou rovny jedné a kovarianční i korelační matice normovaných proměnných jsou si rovny, tj. $\mathbf{S}_z = \mathbf{R}_z = \mathbf{R}_x$ a jsou rovny korelační matici původních netrasformovaných proměnných.

Pro veličinu, která vznikne lineární kombinací původních proměnných

$$u_i = \mathbf{c}^T \mathbf{x}_i = \sum_{j=1}^p c_j x_{ij}$$

platí

$$\bar{u} = \mathbf{c}^T \bar{\mathbf{x}}, \quad s_u^2 = \mathbf{c}^T \mathbf{S}_x \mathbf{c}.$$

4.3 Vzdálenost dvou objektů

Řádek datové matice, tj. vektor \mathbf{x}^T můžeme považovat za souřadnice bodu v p -rozměrném prostoru. Pak je užitečné zabývat se vzdáleností dvou objektů. Jedno-rozměrná vzdálenost (když $p = 1$) dvou objektů i, i' je absolutní hodnota rozdílu pozorovaných hodnot,

$$d(i, i') = |x_i - x_{i'}|.$$

Pro vícerozměrná data můžeme definovat různé vzdálenosti. Eukleidovská vzdálenost dvou objektů i, i' je

$$D_E(i, i') = \sqrt{\sum_{j=1}^p (x_{ij} - x_{i'j})^2} = \sqrt{\sum_{j=1}^p d_j^2(i, i')}$$

Normovaná vzdálenost

$$D_N(i, i') = \sqrt{\sum_{j=1}^p (z_{ij} - z_{i'j})^2} = \sqrt{\sum_{j=1}^p \frac{d_j^2(i, i')}{s_j^2}}$$

Mahalanobisova vzdálenost respektuje jak rozdílnost ve variabilitě, tak korelační strukturu. Její čtverec je pak



$$D_M^2(i, i') = \mathbf{d}^T \mathbf{S}^{-1} \mathbf{d} = (\mathbf{x}_i - \mathbf{x}_{i'})^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_{i'}),$$

kde $\mathbf{d} = \mathbf{x}_i - \mathbf{x}_{i'}$.

Pokud $\mathbf{S} = \sigma^2 \mathbf{I}$ (všechny rozptyly jsou shodné, veličiny nekorelované), pak $D_N = D_M$.

Pro vyhledávání odlehlých pozorování je užitečná výběrová Mahalanobisova vzdálenost od těžiště našich pozorování (od vektoru průměrů)

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

Pro p -rozměrné normální rozdělení populace je

$$\frac{(n-p)n}{(n^2-1)p} D_i^2 \sim F(p, n-p),$$

podle toho tedy můžeme posoudit, zda je pozorování odlehlé.

4.4 Chybějící hodnoty v datech

Zpracování mnohorozměrných dat v reálných úlohách je někdy komplikováno tím, že data nejsou úplná, hodnoty některých prvků datové nejsou k dispozici (slangově označovány jako missingy, missings).

Obvyklý jednoduchý postup je vypustit veličiny s mnoha missingy a vypustit případy (objekty) s mnoha missingy a úsudku o těchto veličinách a objektech se zříci.

Nejběžnější postup užívaný ve většině statistických paketů je automatické vypouštění případů (objektů) s jedním nebo více missingy, tzv. CASEWISE strategie. Tato strategie je z hlediska dalšího zpracování nejbezpečnější, ale může někdy vést k příliš velké ztrátě informace, kdy mnoho objektů je vyřazeno jen kvůli jedné chybějící hodnotě. Ale při této strategii výběrová kovarianční matice zůstane pozitivně definitivní (když hodnota $(\mathbf{X}) = p$), neboť hodnota $(\mathbf{X}^T \mathbf{X}) = p$.

Pro odhad kovarianční matice lze užít i strategii PAIRWISE, kdy každá kovariance se počítá ze všech možných dvojic a průměry v ní jen z hodnot užitých pro výpo-

čet kovariance nebo strategii ALLVALUE, kdy pro průměry se užijí všechny možné (dostupné) hodnoty. Při tomto postupu se využije dat důkladněji, ale výběrová kovarianční matice nemusí být pozitivně definitivní.

Jiný přístup k práci s missingy je tak zvaná imputace, čili doplnění chybějících hodnot nějakými vhodnými, obvykle náhodnými hodnotami z rozdělení, které je shodné nebo podobné s rozdělením pozorovaných hodnot v datové matici. Cílem imputace je zabránit ztrátě informace nevyužitím všech pozorovaných hodnot v datové matici za cenu rizika, že informaci obsaženou v datech trochu zkreslíme. Běžnými metodami imputace, které jsou implementovány ve standardním statistickém software, jsou následující postupy doplnění chybějících hodnot:

- průměrem
- náhodně z předpokládaného rozdělení s využitím odhadu jeho parametrů, nejčastěji je chybějící prvek x_{ij} nahrazen hodnotami z $N(\mu, \sigma^2)$, kde hodnoty parametrů odhadneme z dostupných dat, $\hat{\mu} = \bar{x}_j$, $\hat{\sigma}^2 = s_j^2$
- regresním modelem, jehož parametry odhadneme ze zbývajících $(n - 1)$ objektů, chybějící hodnota se nahradí hodnotou predikovanou modelem, případně ještě modifikovanou náhodným kolísáním.

4.5 Ověřování normality

Mnoho metod vícerozměrné analýzy dat vychází z předpokladu, že náhodná složka v datech má normální rozdělení. Někdy je vyžadována vícerozměrná normalita, tj. p -rozměrné normální rozdělení, někdy stačí jen normalita některých veličin. Jak víme, p -rozměrné sdružené normální rozdělení má marginální rozdělení normální, avšak neplatí to naopak, tzn. marginální normalita nezaručuje sdruženou normalitu. Uvedeme stručně některé metody, kterými lze normalitu (většinou jen marginální) testovat.

Jednorozměrná normalita

Testy dobré shody, ve kterých se empirické rozdělení porovnává s normálním rozdělením. Rozpětí pozorovaných hodnot se rozdělí na r intervalů a porovnají se četnosti pozorovaných hodnot v jednotlivých intervalech s teoretickými četnostmi, které bychom očekávali při výběru stejného rozsahu z normálního rozdělení. Hranice intervalů se volí

- buď ekvidistantně (intervaly jsou stejně široké) tak, aby teoretické (očekávané) četnosti $\Psi_i > 1$ byly pro všechny intervaly a $\Psi_i > 5$ pro 80% hodnot (tzv. Cochranovo pravidlo).
- nebo hranice r intervalů se volí tak, aby teoretické četnosti Ψ_i byly konstantní, nejčastěji se volí $\Psi_1 = \dots = \Psi_r = 5$. Pokud se rozhodneme pro tento způsob

volby hranic intervalů, volbou požadované hodnoty teoretické četnosti je určen při daném rozsahu výběru i počet intervalů r .

Testová statistika je pak

$$\sum_{i=1}^r \frac{(n_i - \Psi_i)^2}{\Psi_i^2} \sim \chi_{(r-1)}^2$$

Kolmogorovův test – porovnává se výběrová (F_n a teoretická (F) distribuční funkce. Výběrová distribuční funkce definována jako

$$F_n(0) = 0, \quad F_n(i) = \frac{i}{n}$$

a testovou statistikou je maximum absolutní hodnoty rozdílu porovnávaných distribučních funkcí

$$k_2 = \max(|F - F_n|).$$

Pro malé rozsahy výběru jsou kritické hodnoty této statistiky tabelovány, pro $n > 50$ se může užít asymptotická aproximace

$$k_2(0, 95) \doteq \frac{1,36}{\sqrt{n}} \quad k_2(0, 99) \doteq \frac{1,63}{\sqrt{n}}$$

Testy šikmosti a špičatosti

Při výpočtu šikmosti a špičatosti se užívají centrální výběrové momenty M_k .

$$M_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Šikmost je definována jako

$$g_1 = \frac{M_3}{M_2^{3/2}}$$

a normální rozdělení má šikmost rovnou nule (je symetrické).

Špičatost je rovna

$$g_2 = \frac{M_4}{M_2^2} - 3$$

a i ta je pro normální rozdělení nulová, neboť u normálního rozdělení je poměr $M_4/M_2^2 = 3$.

K testování normality lze užít statistiky

$$K^{(3)} = \sqrt{\frac{g_1^2(n+1)(n+3)}{6(n-2)}}$$

a

$$K^{(4)} = \sqrt{\frac{(n+1)^2(n+3)(n+5)}{24n(n-2)(n-3)}} \left(g_2 + \frac{6}{n+1} \right),$$

které obě mají přibližně normované normální rozdělení $N(0, 1)$.

4.6 Grafické metody ověřování normality

Tyto grafické metody umožňují rychlé vizuální posouzení shody empirického rozdělení s normálním (případně i jiným) rozdělením a proto jsou velmi často využívány a také jsou implementovány v běžně užívaném statistickém software. Kromě histogramů, do kterých se proloží i teoretické rozdělení a grafů porovnávajících empirickou distribuční funkci s teoretickou se užívá tzv. QQ-graf, kvantilový graf. QQ je zkratka pro kvantil (angl. quantile).

Pro sestrojení kvantilového grafu nejdříve uspořádáme výběr, tj.

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Hodnoty výběrové distribuční funkce se spočtou jako

$$VDF(x_{(i)}) = \frac{i - \frac{1}{2}}{n} \quad \text{nebo} \quad VDF(x_{(i)}) = \frac{i}{n+1}$$

a kvantily $x_{(i)}$ se vynesou do grafu proti odpovídajícím kvantilům normovaného normálního rozdělení, (tj. např. proti hodnotám $u(i/(n+1))$, kde $u(p)$ je p -kvantil normovaného normálního rozdělení). Pokud je výběrové rozdělení normální, grafem je přibližně přímka.

4.7 Transformace dat

Pokud zjistíme, že naměřená data nejsou z normálního rozdělení, je někdy užitečné použít vhodnou transformaci, aby transformací původní veličiny vznikla odvozená veličina, která normální rozdělení má. Potom lze aplikovat metody vyžadující normální rozdělení na transformované veličiny.

Transformací rozumíme takový přepočet, $y_i = f(x_i)$, aby se y_i , $i = 1, \dots, n$ přiblížilo výběru z normálnímu rozdělení.

Ze zkušenosti lze doporučit tyto transformace:

- odmocninová transformace $y = \sqrt{x}$, když x jsou četnosti
- logitová transformace $y = \ln\left(\frac{x}{1-x}\right)$, když x jsou relativní četnosti

- logaritmická transformace $y = \ln x$ pokud měřená veličina má log-normální rozdělení - např. výdaje na domácnost, náklady na výrobek atd.

Jinou možností je *transformace odvozená z dat*, kterou navrhli Box a Cox v roce 1964. Tato transformace poskytne hodnoty odvozené veličiny y , které se nejvíce přibližují normálnímu rozdělení.

$$y = \begin{cases} \frac{x^2-1}{\lambda} & \text{pro } \lambda \neq 0 \\ \ln x & \text{pro } \lambda = 0 \end{cases}$$

λ se odhadne jako $\lambda = \lambda_{max}$, které maximalizuje věrohodnostní funkci

$$\ln L(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 + (\lambda - 1) \sum_{i=1}^n \ln x_i \right]$$

Asymptotický interval $100(1 - \alpha)\%$ - ní spolehlivosti pro λ :

$$2 [\ln L(\lambda_{max}) - \ln L(\lambda)] \leq \chi_{1-\alpha}^2(1),$$

čili v tomto intervalu jsou všechna x , pro která platí:

$$\ln L(x) \geq \ln L(\lambda_{max}) - \frac{1}{2} \chi_{1-\alpha}^2(1)$$

Charakteristiky pro data, která nejsou z normálního rozdělení

Takovými charakteristikami jsou ty, jejichž hodnoty nejsou ovlivněny odhledlými hodnotami v datech. Jako příklad uvedeme uřezávaný průměr (trimmed mean):

$$\bar{x}(\alpha) = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} x_{(i)} \quad m = \text{int} \left(\frac{\alpha n}{100} \right)$$

α je % uříznutých pořádkových statistik na každém konci.

Podobně lze zavést i uřezávaný odhad rozptylu atd.



Shrnutí

- *vektor výběrových průměrů, výběrová kovarianční matice, výběrová korelační matice*
- *ověřování normality, QQ graf*



Kontrolní otázky

1. *Vygenerujte si (např. v Excelu) náhodný výběr z rovnoměrného rozdělení o rozsahu 100 a zkonstruujte QQ graf*
2. *Vygenerujte náhodný výběr stejného rozsahu z normálního rozdělení, zkonstruujte QQ graf a porovnejte s QQ grafem z předchozí otázky*