Ověřování normálního rozdělení

Marika Dienová

Seminář Vybrané partie z aplikované matematiky

Ústav matematiky a statistiky
Janáčkovo nám. 2a
Brno

27.3.2008

Osnova

- Testy dobré shody
- Test pomocí výběrové šikmosti
- Test pomocí výběrové špičatosti
- Posouzení normality graficky
- Kolmogorovův-Smirnovův test
- Shapirův-Wilkův test

Testy dobré shody

Nechť $X_1, ..., X_n$ je náhodný výběr.
Testujeme hypotézu H_0, že se jedná o výběr z $N(\mu, \sigma^2)$

- Vytvoříme $k \geq 4$ třídících intervalů $(-\infty, b_1), [b_1, b_2), \ldots, [b_{k-1}, \infty)$, které označíme $J_i, i = 1, \ldots, k$. Pravděpodobnost p_i, že daná veličina padne do J_i je rovna:

$$p_i = \int_{J_i} f(x)dx \qquad \text{kde} \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- Parametry μ a σ odhadneme metodou minimálního χ^2.
  Po úpravě dostaneme soustavu:

$$\mu = \frac{1}{n}\sum_{i=1}^{k}\frac{X_i}{p_i}\int_{J_i} xf(x)dx \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{k}\frac{x_i}{p_i}\int_{J_i}(x-\mu)^2 f(x)dx$$

  Řešení označme $\widehat{\mu}$ a $\widehat{\sigma}$

Testy dobré shody

- Testovací statistika má tvar:

$$\chi^2 = \sum_{i=1}^{k} \frac{[X_i - np_i(\widehat{\mu}, \widehat{\sigma})]^2}{np_i(\widehat{\mu}, \widehat{\sigma})} \tag{1}$$

- Pokud $\chi^2 > \chi^2_{k-3}(\alpha)$, zamítneme na hladině významnosti α hypotézu H_0.
- Hodnota testovací statistiky je silně závislá na volbě třídících intervalů. Navíc při nesplnění podmínky $np_i \geq 5$ je třeba některé intervaly slučovat, což ovšem vede ke ztrátě informace.
- Test je silný pouze v případě velkého počtu dat $n \geq 50$.

Výběrová šikmost a špičatost

Výběrová šikmost

$$a_3 = \frac{n \sum_{i=1}^{n} (X_i - \overline{X})^3}{[\sum_{i=1}^{n} (X_i - \overline{X})^2]^3} \tag{2}$$

Výběrová špičatost

$$a_4 = \frac{n \sum_{i=1}^{n} (X_i - \overline{X})^4}{[\sum_{i=1}^{n} (X_i - \overline{X})^2]^2} \tag{3}$$

Za předpokladu, že výběr pochází z normálního rozdělení, má šikmost i špičatost asymptoticky normální rozdělení s parametry:

$$E(a_3) = 0 \qquad D(a_3) = \frac{6(n-2)}{(n+1)(n+3)}$$

$$E(a_4) = 3 - \frac{6}{(n+1)} \qquad D(a_4) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Testování pomocí výběrové šikmosti

- Testujeme nulovou hypotézu o normalitě výběru oproti alternativě, že výběr pochází z nějakého nesymetrického rozdělení.
- Protože šikmost normálního rozdělení je rovna nule, měla by být veličina a_3 blízká této hodnotě.
- Testovací statistika má tvar:

$$K_3 = \frac{a_3}{\sqrt{D(a_3)}} \tag{4}$$

- Jestliže $n \geq 200$, lze využít asymptotické normality. Pokud $|K_3| \geq u(\frac{\alpha}{2})$, zamítáme nulovou hypotézu.
- Tento test však vyjde neprůkazně, pokud se data liší od normality v něčem jiném, než je šikmost.

Testování pomocí výběrové šikmosti

Vylepšený postup (D'Agostino)

$$b = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} \qquad W^2 = \sqrt{2(b-1)} - 1$$

$$\delta = \frac{1}{\sqrt{\ln W}} \qquad a = \sqrt{\frac{2}{W^2 - 1}}$$

$$Z_3 = \delta \ln\left[\frac{K_3}{a} + \sqrt{\left(\frac{K_3}{a} + 1\right)}\right]$$

Veličina Z_3 má přibližně rozdělení $N(0,1)$.
Pokud bude $|Z_3| \geq u(\frac{\alpha}{2})$, zamítneme nulovou hypotézu.
Tato aproximace může být použita už pro $n > 8$

Testování pomocí výběrové špičatosti

- Protože špičatost normálního rozdělení je rovna 3, měla by být veličina a_4 blízká této hodnotě.
- Testovací statistika má tvar:

$$K_4 = \frac{a_4 - E(a_4)}{\sqrt{D(a_4)}} \tag{5}$$

- Jestliže $n \geq 500$, lze využít asymptotické normality. Pokud $|K_4| \geq u(\frac{\alpha}{2})$, zamítáme nulovou hypotézu.
- Tento test však vyjde neprůkazně, pokud se data liší od normality v něčem jiném, než je špičatost.

Testování pomocí výběrové špičatosti

Vylepšený postup (D'Agostino)

$$B = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}} \qquad A = 6 + \frac{8}{B}\left(\frac{2}{B} + \sqrt{1 + \frac{4}{B^2}}\right)$$

$$Z_4 = \frac{1 - \frac{2}{9A} - \sqrt[3]{\frac{1-\frac{2}{A}}{1+K_4\sqrt{\frac{2}{A-4}}}}}{\sqrt{\frac{2}{9A}}}$$

Veličina Z_4 má přibližně rozdělení $N(0,1)$.
Pokud bude $|Z_4| \geq u(\frac{\alpha}{2})$, zamítneme nulovou hypotézu.
Tato aproximace může být použita už pro $n \geq 20$

Test kombinace šikmosti a špičatosti

- Tento test je založen na veličině $K_3^2 + K_4^2$.
  Hypotézu o normalitě zamítáme, pokud $K_3^2 + K_4^2 \geq \chi_2^2(\alpha)$.
  Tento postup se ovšem doporučuje pouze pro výběry o rozsahu $n > 200$.
- Pro $n \geq 20$ můžeme použít test založený na $Z_3^2 + Z_4^2$.
  Hypotézu o normalitě zamítáme, pokud $Z_3^2 + Z_4^2 \geq \chi_2^2(\alpha)$.

Normal probability plot (N-P plot)

- N-P plot se konstruuje tak, že na vodorovnou osu nanášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ a na svislou osu kvantili u_{α_j}, kde

$$\alpha_j = \frac{3j - 1}{3n + 1}$$

- Pokud jsou některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající dané skupině.
- Pocházejí-li data z normálního rozdělení, pak budou všechny dvojice (x_j, u_{α_j}) ležet na přímce.

Quantile-Quantile plot (Q-Q plot)

- Pomocí Q-Q plotu můžeme graficky posoudit, zda data pocházejí z nějakého známého rozdělení.
- Q-Q plot se konstruuje tak, že na svislou osu nanášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ a na vodorovnou osu kvantili u_{α_j}, kde

$$\alpha_j = \frac{j - r_{adj}}{n + n_{adj}}$$

  n_{adj} a r_{adj} jsou korigující faktory. Implicitně se klade $n_{adj} = 0,25$ a $r_{adj} = 0,375$.
- Pokud jsou některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající dané skupině.
- Body (u_{α_j}, x_j) metodou nejmenších čtverců proložíme přímku. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a normálním rozdělením.

Probability-Probability plot (P-P plot)

- Spočteme standardizované hodnoty

$$z_{(j)} = \frac{x_{(j)} - m}{s} \qquad j = 1, 2, \ldots, n.$$

  kde m je aritmetický průměr hodnot $x_1, \ldots, x_n$ a s je jejich směrodatná odchylka.

- Na vodorovnou osu vyneseme hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = \frac{1}{n}$.

- Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce $\langle 0, 1\rangle \times \langle 0, 1\rangle$, lze usuzovat na dobrou shodu empirického a teoretického rozložení.

- Pokud jsou některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající dané skupině.

Kolmogorovův-Smirnovův test (K-S test)

- K-S test testuje nulovou hypotézu H_0 říkající, že výběr $X_1, \ldots, X_n$ pochází z rozdělení s distribuční funkcí $\Phi(x)$.
- Nechť $F_n(x)$ je výběrová distribuční funkci.
- Testovací statistika: $D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)|$
- V případě, že $D_n \geq D_n(\alpha)$, zamítneme nulovou hypotézu na hladině významnosti α, kde $D_n(\alpha)$ je tabelovaná kritická hodnota.
- Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem

$$D_n(\alpha) \approx \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$$

- Nulová hypotéza ovšem musí specifikovat distribuční funkci zcela přesně, včetně všech jejích případných parametrů.

Shapirův-Wilkův test normality (S-W test)

- S-W test testuje hypotézu, že náhodný výběr $X_1, \ldots, X_n$ pochází z normálního rozdělení s parametry (μ, σ^2).
- Test je založen na základě zjištění, zda body Q-Q grafu se významně odlišují od regresní přímky proložené těmito body.
- Používá se především pro $n < 50$.
- Testovací statistika má tvar:

$$W = \frac{[\sum_{i=1}^{n} a_i x_{(i)}]^2}{\sum_{i=1}^{n} (x_i - \overline{x})}$$

kde koeficienty a_i byly speciálně odvozeny pro tento test a jsou tabelovány (např. v ČSN 010225)

Reference

[1] Jiří Anděl: *Základy matematické statistiky*,1. vydání. Praha : MATFY-ZPRESS, 2005.

[2] Jiří Anděl: *Matematická statistika*, 2. vyd. Praha : SNTL - Nakladatelství technické literatury, 1985

[3] Budíková, Marie - Lerch, Tomáš - Mikoláš, Štěpán: *Základní statistické metody*, 1. vyd. Brno: Masarykova univerzita, 2005.

[4] Dominik Grůza: *Ověřování normality*, Diplomová práce, Brno 2007