Ověřování normality

David Hampel

Ústav statistiky a operačního výzkumu,
Mendelova univerzita v Brně

evropský sociální fond v ČR · EVROPSKÁ UNIE · MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY · OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Kurz pokročilých statistických metod
Global Change Research Centre AS CR, 5.–7. 8. 2015

Tato akce se koná v rámci projektu: Vybudování vědeckého týmu environmentální metabolomiky a ekofyziologie a jeho zapojení do mezinárodních sítí (ENVIMET; r.č. **CZ.1.07/2.3.00/20.0246**) realizovaného v rámci Operačního programu Vzdělávání pro konkurenceschopnost

Obsah

Motivace

- Téměř každý statistický test vyžaduje splnění určitých předpokladů.
- Normalita je častým předpokladem použitelnosti celé řady testů.
- Usuzovat na (ne)normalitu můžeme již na základě povahy zpracovávaných dat.

Grafické ověřování normality

Normalitu je možno přibližně ověřovat s pomocí diagnostických grafů. Nejčastěji se používá

- histogram (viz Popisná statistika),
- „Normal Probability plot“ (N-P plot)
- „Quantile – Quantile plot“ (Q-Q plot) a
- „Probability – Probability plot“ (P-P plot).

Všechny tyto grafy jsou v současnosti standartní výbavou statistických programů.

Normal Probability plot

N-P plot konstruujeme tak, že na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na svislou osu kvantily u_{α_j}, kde

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

Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince. Pocházejí-li data z normálního rozdělení, pak budou všechny dvojice $(x_{(j)}, u_{\alpha_j})$ ležet na přímce.

Quantile – Quantile plot

Q-Q plot konstruujeme tak, že na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozdělení, kde

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

kde r_{adj} a n_{adj} jsou korigující faktory oba menší než 0.5. Často se klade $r_{adj} = 0.375$ a $n_{adj} = 0.25$. Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince. Pocházejí-li data z testovaného rozdělení, pak budou všechny dvojice $(K_{\alpha_j}(X), x_{(j)})$ ležet na přímce.

N-P plot a Q-Q plot – příklad

Desetkrát nezávisle na sobě byla změřena jistá konstanta.

2 1.8 2.1 2.4 1.9 2.1 2 1.8 2.3 2.2

Je třeba otestovat normalitu.

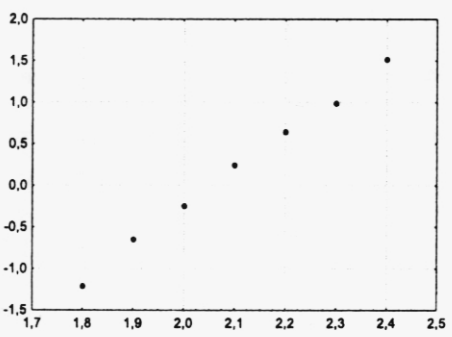
usp. hodnoty	1.8	1.8	1.9	2	2	2.1	2.1	2.2	2.3	2.4
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	1.5	1.5	3	4.5	4.5	6.5	6.5	8	9	10

N-P plot

$j = (1.5, 3, 4.5, 6.58910)$

$\alpha_j = \frac{3j-1}{3n+1} = (0.1129, 0.2581, 0.4032, 0.5968, 0.7419, 0.8387, 0.9355)$

$u_{\alpha_j} = (-1.2112, -0.6493, -0.245, 0.245, 0.6493, 0.9892, 1.5179)$

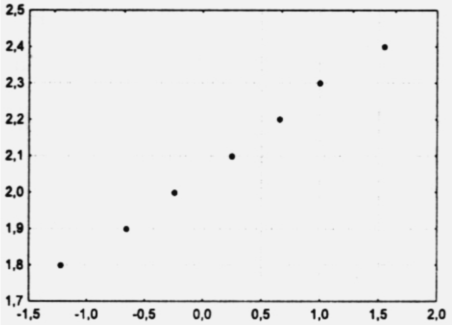


Q-Q plot

$j = (1.5, 3, 4.5, 6.58910)$

$\alpha_j = \frac{j-0.375}{n+0.25} = (0.1098, 0.2561, 0.4024, 0.5976, 0.7439, 0.8415, 0.939)$

$K_{\alpha_i}(X) = u_{\alpha_i} = (-1.2278, -0.6554, -0.247, 0.247, 0.6554, 1.0005, 1.566)$



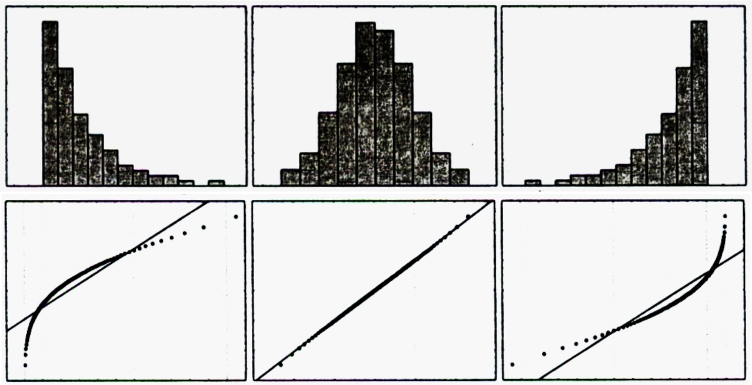
Spočtou se standardizované hodnoty

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

Na vodorovnou osu se vznesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = \frac{j}{n}$. Jsou-li některé hodnoty $x_{(1)} \leq \cdots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince. Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce $[0,1] \times [0.1]$ lze usuzovat na dobrou shodu empirického a teoretického rozdělení.

Histogram a N-P plot

Rozdělení s kladnou šikmostí	Normální rozdělení	Rozdělení se zápornou šikmostí



Testy normality

V dalším si uvedeme několik testů, které umožňují testovat (nejen) normalitu více exaktněji než vizuálním zhodnocením grafu.

Kolmogorov-Smirnov test

Testujeme hypotézu, že náhodný výběr pochází z rozdělení s distribuční funkcí $\Phi(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)|.$$

Nulovou hypotézu zamítneme na hladině významnosti α když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem

$$\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}.$$

Kolmogorov-Smirnov test – příklad

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K-S testu zjistěte na hladině významnosti 0.05, zda tato data pocházejí z normálního rozdělení.

Odhadem střední hodnoty je výběrový průměr $m = 11$, odhadem rozptylu je výběrový rozptyl $s^2 = 10$.

Hodnoty výběrové distribuční funkce $F(x)$:

1. $x < 8 : F(x) = 0$
2. $8 \leq x < 9 : F(x) = \frac{1}{5} = 0.2$
3. $9 \leq x < 10 : F(x) = \frac{2}{5} = 0.4$
4. $10 \leq x < 12 : F(x) = \frac{3}{5} = 0.6$
5. $12 \leq x < 16 : F(x) = \frac{4}{5} = 0.8$
6. $x \geq 16 : F(x) = 1$

Kolmogorov-Smirnov test – příklad

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K-S testu zjistěte na hladině významnosti 0.05, zda tato data pocházejí z normálního rozdělení.

Odhadem střední hodnoty je výběrový průměr $m = 11$, odhadem rozptylu je výběrový rozptyl $s^2 = 10$.

Hodnoty teoretické distribuční funkce $\Phi_T(x)$:

1. $\Phi_T(8) = \Phi(\frac{8-11}{\sqrt{10}}) = 0.17106$
2. $\Phi_T(9) = \Phi(\frac{9-11}{\sqrt{10}}) = 0.26435$
3. $\Phi_T(10) = \Phi(\frac{10-11}{\sqrt{10}}) = 0.37448$
4. $\Phi_T(12) = \Phi(\frac{12-11}{\sqrt{10}}) = 0.62552$
5. $\Phi_T(16) = \Phi(\frac{16-11}{\sqrt{10}}) = 0.94295$

(Φ je distribuční funkce $N(0,1)$.)

Kolmogorov-Smirnov test – příklad

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K-S testu zjistěte na hladině významnosti 0.05, zda tato data pocházejí z normálního rozdělení.

Odhadem střední hodnoty je výběrový průměr $m = 11$, odhadem rozptylu je výběrový rozptyl $s^2 = 10$.

Rozdíly mezi $\Phi_T(x)$ a $F(x)$:

1. $d_1 = 0.2 - 0.17106 = 0.02894$
2. $d_2 = 0.4 - 0.26435 = 0.13565$
3. $d_3 = 0.6 - 0.37448 = 0.22552$
4. $d_4 = 0.8 - 0.62552 = 0.17448$
5. $d_5 = 1 - 0.94295 = 0.05705$

Kolmogorov-Smirnov test – příklad

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K-S testu zjistěte na hladině významnosti 0.05, zda tato data pocházejí z normálního rozdělení.

Odhadem střední hodnoty je výběrový průměr $m = 11$, odhadem rozptylu je výběrový rozptyl $s^2 = 10$.

Testová statistika

$$D_5 = 0.22552$$

je menší než kritická hodnota 0.343, hypotézu o normalitě na hladině významnosti 0.05 nezamítáme.

Shapiro-Wilk test

Tento test je určen pro testování normality. Je založen na zjištění, zda body v Q-Q plotu jsou výrazně odlišné od regresní přímky proložené těmito body. Používá se především pro výběry menších rozsahů, $n < 50$. Testové kriterium je dáno jako

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

kde

$$(a_1, \ldots, a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}},$$

$m = (m_1, \ldots, m_n)^\top$ je střední hodnota a V kovarianční matice pořadových statistik. Hypotézu o normalitě zamítáme, když bude W příliš malé.

Testy normality založené na šikmosti a špičatosti

- Pokud je statistika

$$U_3 = \frac{a_3}{\sqrt{\frac{6(n-2)}{(n+1)(n+3)}}},$$

kde a_3 je **šikmost**, větší než kvantil $u_{\alpha/2}$, normalitu zamítáme ve prospěch asymetrie.

- Pokud je statistika

$$U_4 = \frac{a_4 - \left(3 - \frac{6}{n-1}\right)}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}},$$

kde a_4 je **špičatost**, větší než kvantil $u_{\alpha/2}$, normalitu zamítáme ve prospěch více či méně špičatého rozdělení.

Oba testy jsou vhodné pro rozsáhlé výběry ($n > 200$ popř. $n > 500$).

Testy dobré shody

Testujeme hypotézu, že náhodný výběr pochází z rozdělení s distribuční funkcí $\Phi(x)$.

- Určíme r třídících intervalů a spočteme, kolik hodnot je v každém intervalu - n_i.
- Spočteme teoretické pravděpodobnosti p_i, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v těchto intervalech.
- Spočteme hodnotu testové statistiky

$$K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}.$$

- Pokud vyjde $K \geq \chi^2_{1-\alpha}(r-1-p)$ (p je počet odhadovaných parametrů zkoumaného rozdělení), hypotézu o normalitě zamítáme. Výsledky testu se považují za věrohodné, pokud $np_j \geq 5 \ \forall j$.