OVÌØENÍ NORMÁLNÍHO ROZDÌLENÍ - MÉNÌ BÌŽNÉ TESTY

Jana Kubanová
Ústav matematiky, FES, Univerzita Pardubice

Abstract: *The normal distribution of population is presumption of many statistical tests. The acceptance of this presumption is most frequently tested by chi-square test and by one-sample Kolmogorov–Smirnov´s test. Some other tests are described in this article, serving for confirmation of this hypothesis, such as the tests based on skewness and curtosess, the Jarque – Berra´s test, the Shapiro-Wilk´s normality test and the D'Agostin test.*

Pøi statistických analýzách se velmi èasto používají testy, které vycházejí z pøedpokladu normálního rozložení základního souboru. V nìkterých pøípadech mùžeme být o splnìní tohoto pøedpokladu pøesvìdèeni na základì døívìjších zkušeností nebo z výzkumù realizovaných na stejných nebo podobných souborech. Pokud si nejsme jisti splnìním pøedpokladu normality, je tøeba ho ovìøit nìkterým ze statistických testù. Nejèastìji používané jsou χ^2 test a jednovýbìrový Kolmogorovùv- Smirnovùv test (Andìl,1993).

Nyní vzpomeneme struènì ještì nìkolik dalších testù, na jejichž základì lze ovìøit normalitu rozložení základního souboru.

a) Testy založené na šikmosti a špièatosti

Pomocí tìchto testù testujeme nulovou hypotézu H_0, že náhodný výbìr pochází ze základního souboru X s normálním rozložením pravdìpodobností $N(\mu, \sigma)$.

O normálním rozdìlení je známo, že jeho koeficienty šikmosti a špièatosti (asymetrie a excesu) jsou rovny nule. Toho se využije v následujícím testu.

Je známo, že výbìrové koeficienty šikmosti a špièatosti (asymetrie a excesu) jsou definovány následujícím zpùsobem:

Výbìrový koeficient šikmosti S_k

$$S_k = \frac{M_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2}\right)^3}$$

Výbìrový koeficient špièatosti E_k

$$E_k = \frac{M_4}{s^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2\right)^2} - 3$$

Koeficient šikmosti i špièatosti normálního rozdìlení je roven nule, proto by mìly i tyto odhady být blízké nule.

Lze dokázat, že pøi náhodném výbìru dostateèného rozsahu z normálního rozdìlení pravdìpodobností $N(\mu, \sigma)$ má náhodná velièina S_k pøibližnì $N(ES_k, \sqrt{DS_k})$ rozdìlení a náhodná velièina E_k pøibližnì $N(EE_k, \sqrt{DE_k})$ rozdìlení pravdìpodobností, kde pro støední hodnotu a disperzi uvedených koeficientù platí (Andìl,1978):

$ES_k = 0$ $\qquad DS_k = \dfrac{6(n-2)}{(n+1)(n+3)}$

$EE_k = \dfrac{-6}{n+1}$ $\qquad DE_k = \dfrac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$

Při testu založeném na šikmosti zamítneme hypotézu o normálním rozložení pravděpodobností základního souboru tehdy, když

$$\frac{|S_k|}{\sqrt{DS_k}} \geq z_\alpha \ ,$$

kde $z_\alpha = \Phi^{-1}\left(\dfrac{2-\alpha}{2}\right)$. Tuto hodnotu hledáme v tabulkách N(0, 1) rozdělení pravděpodobností.

Při testu založeném na špičatosti hypotézu o normalitě rozložení zamítneme, je-li

$$\frac{|E_k - EE_k|}{\sqrt{DE_k}} \geq z_\alpha$$

V některých případech jsou testy založené na šikmosti a špičatosti citlivější na porušení normality než uvedený χ^2 test.

b) Test kombinace výběrové šikmosti a špičatosti (Jarque – Berra)

Testujeme opět nulovou hypotézu H_0, že náhodný výběr pochází ze základního souboru X s normálním rozložením pravděpodobností $N(\mu, \sigma)$.

K testování nulové hypotézy používáme testovací kritérium tvaru

$$\chi = \frac{S_k^2}{DS_k} + \frac{(E_k - EE_k)^2}{DE_k},$$

kde S_k je výběrový koeficient šikmosti a E_k výběrový koeficient špičatosti. Za předpokladu platnosti H_0 má náhodná veličina χ asymptoticky χ^2 rozložení pravděpodobností se 2 stupni volnosti. Hypotézu H_0 zamítáme v případě, kdy $\chi > \chi^2_{\alpha,2}$.

c) Shapiro-Wilkův test normality

Dalším z testů, který může sloužit k ověření normality je test odvozený Shapiro a Wilkem. Test je vhodný pro výběry rozsahu $7 \leq n \leq 30$. V testu se používá realizace $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ uspořádaného náhodného výběru $X_{(1)}, \ldots, X_{(n)}$.
Testujeme nulovou hypotézu

H_0: X má $N(\mu, \sigma)$ rozdělení pravděpodobností proti alternativní hypotéze

H_1: X nemá $N(\mu, \sigma)$ rozdělení pravděpodobností.

Testovací kritérium má tvar: $SW = \dfrac{\left(\sum_{i=1}^{m} a_i(n)(X_{(n-i+1)} - X_{(i)})\right)^2}{\sum_{i=1}^{n}(X_i - \bar{x})^2}$,

kde $a_i(n)$ jsou tabelované konstanty, $m = \frac{n}{2}$, je-li n sudé èíslo a $m = \frac{n-1}{2}$, je-li n liché èíslo, $X_{(n-i+1)}$ a $X_{(i)}$ jsou poøádkové statistiky vytvoøené z náhodného výbìru $X_1, ..., X_n$ jeho uspoøádáním do neklesající posloupnosti.

Konstanty $a_i(n)$ souvisí s oèekávanými poøádkovými statistikami normálního rozložení, tzn., že má-li náhodná velièina X rozdìlení $N(\mu, \sigma)$, budou body $(a_i(n), X_i)$ seskupeny až na náhodné odchylky kolem regresní pøímky se smìrnicí σ.

Princip testu je v tom, že se odhadne parametr σ náhodnou velièinou $s^* = \sum_{i=1}^{n} a_i X_i$ a jeho odhad se porovná s odhadem založeným na náhodné velièinì $\sum_{i=1}^{n} (X_i - \bar{x})^2$.

Hypotézu o normalitì rozložení zamítáme, je-li $SW < SW^*(n, \alpha)$. Hodnoty $SW^*(n, \alpha)$ hledáme ve statistických tabulkách (Palenèár a kol.,2001).

V následující tabulce 1 jsou uvedeny koeficienty $a_i(n)$ Shapiro-Wilkova testu a 95%-ní kvantily pro èetnosti $n = 5$ až $n = 10$ náhodného výbìru.
Vzhledem k symetrii kvantilù normálního rozdìlení platí: $a_i = - a_{n-i+1}$.

Tab.1: Koeficienty $a_i(n)$ Shapiro-Wilkova testu a 95%-ní kvantily pro $n = 5$ až $n = 10$ náhodného výbìru.

i	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
1	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739
2	0,2413	0,2816	0,3031	0,3164	0,3244	0,3291
3	0	0,0875	0,1401	0,1743	0,1976	0,2141
4	-	-	0	0,0561	0,0947	0,1224
$SW^*_{0,95}$	0,762	0,788	0,803	0,818	0,829	0,842

Pro vìtší èetnosti náhodného výbìru se hodnoty koeficientù a_i i hodnoty testovacího kritéria $W^*_{1-\alpha}$ poèítají podle Roystonova postupu (Royston,1982).
Konstanty a_i jsou tabelované napø v (Owen,1962), (Sarhan,Greenberg,1962) ale i v ÈSN 01 0225.

Praktické pokusy ukázaly, že tento test je vzhledem k odchylkám od normality citlivìjší, než ostatní známe testy dobré shody.

d) D´Agostinùv test

Tento test se používá k testování hypotézy o normalitì pro výbìry o rozsahu $30 \le n \le 100$, tzn. že opìt testujeme nulovou hypotézu

H_0: X má $N(\mu, \sigma)$ rozdìlení pravdìpodobností proti alternativní hypotéze

H_1: X nemá $N(\mu, \sigma)$ rozdìlení pravdìpodobností.

Pøi testování postupujeme tak, že hodnoty realizace náhodného výbìru uspoøádáme do neklesající posloupnosti $x_{(1)} £ x_{(2)} £ \ldots £ x_{(n)}$ a definujeme náhodnou velièinu tvaru

$$D = \frac{\sqrt{n}(A - 0{,}2820948)}{0{,}0299860}$$

kde $A = \dfrac{\sum_{i=1}^{n}\left(\frac{n+1}{2} - i\right)\left(x_{(n-i+1)} - x_{(i)}\right)}{\sqrt{n^3 \sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2}}$

a $x_{(n-i+1)}$, $x_{(i)}$ jsou hodnoty poøádkových statistik z náhodného výbìru $X_1, \ldots, X_n$.

Testovanou hypotézu zamítneme, platí-li pro hodnotu náhodné velièiny D:

$$D < D\left(n, \frac{\alpha}{2}\right) \text{ nebo } D > D\left(n, 1 - \frac{\alpha}{2}\right).$$

Pokud se snažíme hodnotit simultánnì výsledky rùzných testù normality, vyvstávají problémy s urèením chyby prvního druhu. Proto je vhodné se na tyto výsledky testù dívat s urèitým nadhledem a považovat je pouze za orientaèní kriteria pøi vytváøení koneèného názoru.

V nìkterých pøípadech mùže k posouzení normality dobøe posloužit grafické znázornìní hodnot statistického znaku. Uvádíme pro orientaci nìkteré možnosti:

1. *Histogram rozdìlení èetností statistického znaku.*
Histogram nám umožní posoudit, zda hodnoty statistického znaku pøipomínají prùbìh Gaussovy køivky.

2. *Graf empirické distribuèní funkce.*
Sledujeme, zda je graf empirické distribuèní funkce podobný esovité køivce distribuèní funkce normálního rozložení. Pøed sestrojením grafu je vhodné data normovat.
3. *Q-Q diagram.*

Na svislé ose y nanášíme výbìrové kvantily, na vodorovné ose x jim odpovídající kvantily normálního rozložení pravdìpodobností. Pokud výbìr odpovídá normálnímu rozložení pravdìpodobností, pak body, odpovídající jednotlivým pozorováním, budou ležet kolem pøímky $y = x$.

Literatura:

1. Andìl,J.: *Statistické metody*. Matfyzpress, Praha 1993
2. Kubanová,J.: *Statistické metody pro ekonomickou a technickou praxi*, 2003 (v tisku)
3. Owen,D.B.: *Handbook of Statistical Tables*. Massachusetts, Adison-Wesley, 1962
4. Palenèár,R.,Ruiz,J.M.,Janiga,I.,Horníková,A.: *Štatistické metódy v metrologických a skúšobných laboratóriach*. Bratislava 2001, ISBN 80-968449-3-8
5. Royston,J.P.: *Appl.Statist*. 31,115,1982
6. Sarhan,A.E.,Greenberg,B.G.: *Contribution to Order Statistics*. New York, J. Wiley 1962

Kontaktní adresa:
Paed.Dr. Jana Kubanová, CSc., Ústav matematiky, FES, Univerzita Pardubice
Studentská 84, 532 10 Pardubice
tel.: 466 036 046
e-mail: Jana.Kubanova@upce.cz

Recenzoval: doc. RNDr. Bohdan Linda, CSc., vedoucí Ústavu matematiky, FES, Univerzita Pardubice