

# APLIKÁCIA ŠTATISTICKÝCH TESTOV DOBREJ ZHODY

**PhDr. Eva Ostertagová, PhD.**

Technická univerzita v Košiciach  
Fakulta elektrotechniky a informatiky  
Katedra matematiky a teoretickej informatiky  
Němcovej 32, 042 00 Košice  
e-mail: [eva.ostertagova@tuke.sk](mailto:eva.ostertagova@tuke.sk)

## Abstract

The aim of article is an application of basic statistical tests of normality. The importance of normal distribution is undeniable since it is an underlying assumption of many statistical procedures. When the normality assumption is violated, interpretation and inferences may not be reliable or valid.

**Key words:** Pearson test, Kolmogorov-Smirnov test, Liliefors test, Shapiro-Wilk normality test, Jarque-Bera normality test

## ÚVOD

Testy dobrej zhody predstavujú kategóriu štatistických testov, ktoré umožňujú na vopred zvolenej hladine významnosti  $\alpha$  testovať nulovú hypotézu  $H_0$ , že daný náhodný výber bol realizovaný z rozdelenia stanoveného typu, ale prípadne s neznámymi parametrami proti alternatívnej hypotéze  $H_1$ , že nepochádza z tohto rozdelenia. Je teda napríklad možné testovať hypotézu, že príslušné rozdelenie je normálne so známymi alebo neznámymi parametrami  $\mu$  a  $\sigma$ .

Medzi najznámejšie testy dobrej zhody patria Pearsonov test, Kolmogorovov-Smirnovov test a Lilieforsov test. Tieto testy sú do istej miery univerzálne. Za túto univerzálnosť sa však platí zníženou účinnosťou testov. Táto skutočnosť viedla k vypracovaniu špeciálnych testov založených na charakteristických vlastnostiach predpokladaného modelu. Špeciálnym testom dobrej zhody je napr. Shapiro-Wilkov test normality a Jarque-Bera test normality.

## 1. PEARSONOV TEST

Pearsonov test ( $\chi^2$ -test) patrí medzi najznámejšie testy dobrej zhody. Je to test založený na rozdieloch medzi pozorovanými (empirickými) a očakávanými (teoretickými) početnosťami. Tento test sa odporúča používať pre náhodný výber veľkého rozsahu.

Daný je náhodný výber  $x_1, x_2, \dots, x_n$  z rozdelenia pravdepodobnosti s neznámou distribučnou funkciou  $F(x)$ , daná je známa distribučná funkcia  $F_0(x)$  teoretického rozdelenia pravdepodobnosti, daný je rozklad reálnej osi

na disjunktné triedy (intervaly)  $I_1, I_2, \dots, I_k$ . Na vopred zvolenej hladine významnosti  $\alpha$  testujeme nulovú hypotézu  $H_0 : F(x) = F_0(x)$  proti alternatívnej hypotéze  $H_1 : F(x) \neq F_0(x)$ .

Testovacia štatistika u tohto testu je

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}, \text{ ktorá má (pre } n \geq 50 \text{)}$$

pri platnosti  $H_0$  približne  $\chi^2$ -kvadrát rozdelenie pravdepodobnosti s  $k - r - 1$  stupňami voľnosti, kde  $n_i$  je počet zložiek náhodného výberu o rozsahu  $n$  v  $I_i$ ,  $p_i$  je pravdepodobnostná miera  $I_i$  určená distribučnou funkciou  $F_0$ ,  $n \cdot p_i$  je teoretická početnosť v intervale  $I_i$ ,  $r$  je počet odhadovaných parametrov pri určovaní  $F_0$ ,  $k$  je počet tried.

Hypotézu  $H_0$  zamietame na hladine významnosti  $\alpha$ , ak  $\chi^2 > \chi^2_{1-\alpha, k-r-1}$ , kde  $\chi^2_{1-\alpha, k-r-1}$  je tabelovaná kritická hodnota  $\chi^2$ -rozdelenia.

Pri Pearsonovom teste je potrebné dôsledne dodržiavať tzv. Cochranovo pravidlo. Podmienka použitia daného testu je, aby  $n \cdot p_i \geq 5$  pre všetky  $i = 1, 2, \dots, k$ .

Bolo overené, že pre počet stupňov voľnosti  $k - r - 1 \geq 3$  stačí, aby bolo  $n \cdot p_i \geq 4$  a pre  $k - r - 1 \geq 6$  stačí, aby bolo  $n \cdot p_i \geq 1$ . Ak nie sú podmienky splnené, odporúča sa zlúčenie susedných intervalov s malou početnosťou [2].

V MATLABe môžeme použiť štandardnú funkciu *chi2gof*.

## 2. KOLMOGOROVOV-SMIRNOVOV TEST

Kolmogorovov-Smirnovov (KS) test je založený na porovnávaní distribučnej funkcie predpokladaného (teoretického) rozdelenia spojitého typu s empirickou (výberovou) distribučnou funkciou. Má sa používať len k overovaniu hypotéz, ktoré určujú teoretickú distribučnú funkciu jednoznačne, t.j. nielen čo do tvaru, ale aj čo do hodnôt parametrov – úplne špecifikovaný model. Na rozdiel od Pearsonovho testu je možné tento test realizovať aj pre náhodné výbery pomerne malých rozsahov.

Na vopred zvolenej hladine významnosti  $\alpha$  testujeme nulovú hypotézu  $H_0 : F(x) = F_0(x)$  proti alternatívnej hypotéze  $H_1 : F(x) \neq F_0(x)$ , kde  $F$  je empirická distribučná funkcia výberového súboru a  $F_0$  je uvažovaná teoretická distribučná funkcia.

Pre namerané hodnoty  $x_1, x_2, \dots, x_n$  (za predpokladu, že všetky sú navzájom rôzne) má empirická distribučná funkcia tvar (uvažujeme usporiadané hodnoty  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ ):

$$F(x) = \begin{cases} 0 & \text{pre } -\infty < x < x_{(1)} \\ \frac{i}{n} & \text{pre } x_{(i)} \leq x < x_{(i+1)}, i = 1, \dots, n-1 \\ 1 & \text{pre } x_{(n)} < x < \infty \end{cases}$$

Testovacia štatistika  $D$  je definovaná ako najväčšia vzdialenosť medzi hodnotami  $F$  a  $F_0$ :  $D = \sup |F_0(x) - F(x)|$ , pre  $x \in R$ . Vzhľadom na takto definovanú empirickú distribučnú funkciu sa dá testovacia štatistika  $D$  vypočítať zo vzťahu:  $D = \max \left\{ \max(d_i^-), \max(d_i^+) \right\}$ , pre  $1 \leq i \leq n$ , kde  $d_i^- = |F_0(x_{(i)}) - F(x_{(i)})|$ ,  $d_i^+ = |F_0(x_{(i)}) - F(x_{(i)}^+)|$ .

Hypotézu  $H_0$  zamietame na hladine významnosti  $\alpha$ , ak vypočítaná hodnota  $D$  prekročí tabelovanú kritickú hodnotu  $D_\alpha(n)$ , t.j.  $D > D_\alpha(n)$ . Pre malé rozsahy výberu  $n \leq 30$  použijeme presnú hodnotu  $D_\alpha(n)$  z tabuľky kritických hodnôt KS testu. Pre veľké rozsahy výberu používame asymptotickú hodnotu

$$\sqrt{-\frac{\ln \frac{\alpha}{2}}{2n}}, \text{ teda napr. pre } \alpha = 0,05 \text{ hodnotu } \frac{1,3581}{\sqrt{n}}$$

a pre  $\alpha = 0,01$  hodnotu  $\frac{1,6276}{\sqrt{n}}$  [3].

V MATLABe môžeme použiť štandardnú funkciu *kstest*.

### 3. LILIEFORSOV TEST

Lilieforsov (LF) test je modifikáciou KS testu. Ak nie sú parametre teoretického rozdelenia známe – neúplne špecifikovaný model, tak sa nahradia bodovými odhadmi.

Tak ako v prípade KS testu pre malé  $n$  používame presnú kritickú hodnotu z tabuľky Lilieforsovho testu, pre veľké  $n$  je možné použiť približnú asymptotickú hodnotu  $\frac{0,89}{\sqrt{n}}$

pre  $\alpha = 0,05$  a  $\frac{1,04}{\sqrt{n}}$  pre  $\alpha = 0,01$  [3].

V MATLABe môžeme použiť štandardnú funkciu *lilietest*.

### 4. SHAPIRO-WILKOV TEST NORMALITY

Shapiro-Wilkov (SW) test je založený na zisťovaní skutočnosti či sa body zostrojeného kvantil-kvantilového grafu (Q-Q plotu) významne líšia od regresnej priamky prelozenej týmito bodmi.

Na vopred zvolenej hladine významnosti  $\alpha$  testujeme nulovú hypotézu  $H_0: F(x) = F_0(x)$  proti alternatívnej hypotéze  $H_1: F(x) \neq F_0(x)$ , kde  $F$  je empirická distribučná funkcia výberového súboru o rozsahu  $n$  a  $F_0$  je uvažovaná teoretická distribučná

funkcia normálneho rozdelenia  $norm(\mu, \sigma)$  s bližšie nešpecifikovanými parametrami  $\mu$  a  $\sigma$ .

SW test sa používa predovšetkým pre výbery menších rozsahov  $n \leq 50$ . Uvažujme usporiadané hodnoty  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  náhodného výberu  $x_1, x_2, \dots, x_n$ . Testovacia štatistika má tvar

$$W = \frac{\left( \sum_{i=1}^m a_{i,n} \cdot (x_{(n-i+1)} - x_{(i)}) \right)^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2}, \text{ kde } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ je}$$

výberový priemer,  $a_{i,n}$  sú tabelované váhy

a  $m = \frac{n}{2}$ , ak je  $n$  párne, resp.  $m = \frac{n-1}{2}$ , ak je  $n$  nepárne [1].

Čím viac sa hodnota testovacej štatistiky  $W$  blíži k 1, tým je zhoda medzi teoretickým a empirickým rozdelením lepšia. Hypotézu  $H_0$  o normálnom rozdelení základného súboru, z ktorého bol realizovaný náhodný výber, zamietame na hladine významnosti  $\alpha$ , ak platí  $W \leq W_\alpha(n)$ , kde  $W_\alpha(n)$  je tabelovaná kritická hodnota.

SW test nie je do programu MATLAB implementovaný, ale je ho možné stiahnuť ako súbor *swtest.m* na stránkach MATLAB Central.

### 5. JARQUE-BERA TEST NORMALITY

Jarque-Bera (JB) test je založený na meraní rozdielu koeficienta šikmosti  $\gamma_3$  a špicatosti  $\gamma_4$  nameraných dát s koeficientmi šikmosti a špicatosti normálneho rozdelenia.

Na vopred zvolenej hladine významnosti  $\alpha$  testujeme nulovú hypotézu  $H_0: F(x) = F_0(x)$  proti alternatívnej hypotéze  $H_1: F(x) \neq F_0(x)$ , kde  $F$  je empirická distribučná funkcia výberového súboru o rozsahu  $n$  a  $F_0$  je uvažovaná teoretická distribučná funkcia normálneho rozdelenia  $norm(\mu, \sigma)$  s neznámymi parametrami.

$$\text{Testovacia štatistika je } JB = n \cdot \left( \frac{\gamma_3^2}{6} + \frac{\gamma_4^2}{24} \right),$$

ktorá má pri platnosti  $H_0$  pre veľké  $n$  približne  $\chi^2$ -kvadrát rozdelenie s dvoma stupňami voľnosti (jeden pre  $\gamma_3$  a jeden pre  $\gamma_4$ ).

Tento test je platný len asymptoticky, a teda je potrebné mať veľký rozsah náhodného výberu. Pri rozsahu výberu  $n < 100$  je potrebná ostrážitosť pri používaní tohto testu.

Hypotézu  $H_0$  zamietame na hladine významnosti  $\alpha$ , ak vypočítaná hodnota  $JB$  prekročí kritickú hodnotu  $\chi_{1-\alpha, 2}^2$ , t.j.  $JB > \chi_{1-\alpha, 2}^2$  [3].

V MATLABe môžeme použiť štandardnú funkciu *jbtest*.

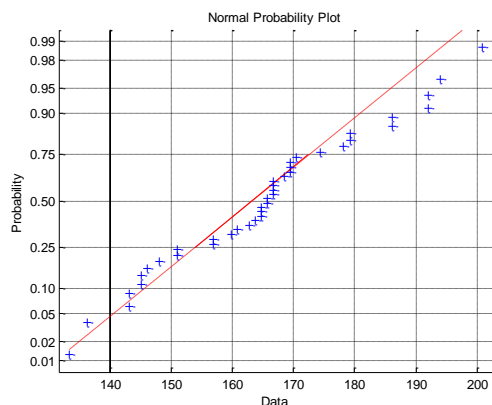
## 6. APLIKÁCIA TESTOV NORMALITY V TECHNICKEJ PRAXI

V istej slovenskej firme, ktorú nebudeme menovať, prechádza hotový produkt mnohými fázami od predpríprav až po konečné úpravy. Firma dostane určitú zákazku, množstvo kvalifikovaných inžinierov ju spracuje, ďalej prichádza na rad schválenie a samotná továrenská zlievarenská časť. V nej sa vytvorí model danej zákazky, z ktorého sa vytvorí odliatok a zase do odliatku sa zlieva zmes určitých prísad (nebudeme uvádzať). Použitá zmes prísad musí mať prísne stanovené technologické vlastnosti. Denne sú odoberané vzorky a technologické vlastnosti sú vyhodnocované v laboratóriu. Pri našom štatistickom skúmaní pre túto firmu sme sa zamerali na jednu dôležitú technologickú vlastnosť, a to na väznosť v tlaku. Väznosť v tlaku môžeme považovať za spojitú náhodnú premennú. Túto technologickú vlastnosť by chceli v tejto firme optimalizovať na hodnotu 160 kPa [4].

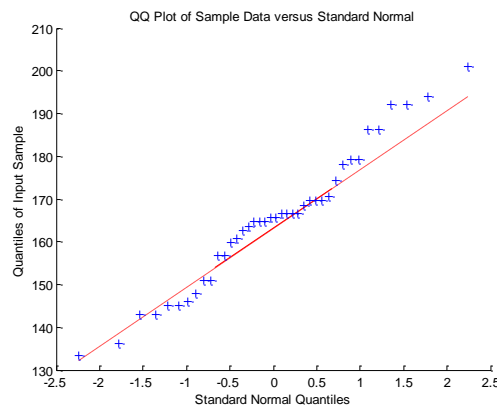
Pre realizáciu štatistickej analýzy bolo potrebné vykonať testovanie normality. V tomto príspevku budeme testovať hypotézu, že výberový súbor o rozsahu  $n = 40$  pochádza z normálneho rozdelenia. Normalitu otestujeme pomocou grafických aj vybraných štatistických testov s použitím MATLABu.

Na overenie normality použijeme grafický test normality, tzv. normal probability plot (N-P plot) a tiež quantile-quantile plot (Q-Q plot). Ak sú dáta z normálneho rozdelenia, tak body ležia na priamke, čo je v našom prípade približne splnené. Uvedené grafy sú zobrazené na obr. 1–2.

```
x=[164.64,166.60,146.02,194.04,186.20,...
192.08,178.06,160.72,145.04,166.60,...
186.20,150.92,162.68,163.66,143.08,...
143.08,174.44,156.80,192.08,169.54,...
165.62,156.80,166.60,169.54,136.22,...
159.74,179.34,168.56,145.06,200.90,...
164.64,166.60,170.52,165.62,164.64,...
147.98,179.34,133.28,150.92,169.54];
normplot(x),figure,qqplot(x)
```



Obr. 1 N-P plot



Obr. 2 Q-Q plot

Ďalej si zvolíme hladinu významnosti  $\alpha = 0,05$  a budeme testovať nulovú hypotézu  $H_0$ , že daný výber je skutočne z normálneho rozdelenia oproti alternatívnej hypotéze  $H_1$ , že pochádza z iného než normálneho rozdelenia.

Pre LF test, SW test a JB test použijeme tieto príkazy MATLABu:

```
[h,p,lfstat,cval]=lillietest(x,0.05,'norm')
[h,p,swstat]=swtest(x,0.05)
[h,p,jbstat,cval]=jbtest(x,0.05)
```

Výsledky použitých testov sú uvedené v tab. 1.

Tab. 1 Sumárna tabuľka pre vybrané testy

	LF test	SW test	JB test
Hodnota testovacej štatistiky	0,1182	0,9726	0,3731
Kritická hodnota	0,1385	-	4,7481
P-hodnota	0,1650	0,4335	0,5000
Prijímaná hypotéza	$H_0$	$H_0$	$H_0$

## ZÁVER

Na základe vybraných grafických testov a štatistických testov dobrej zhody sa potvrdila testovaná hypotéza, že sa daný výberový súbor riadi normálnym rozdelením pravdepodobnosti.

## Literatúra

- [1] Markechová, D., Tirpáková, A., Stehlíková, B.: Základy štatistiky pre pedagógov. UKF v Nitre, 2011. 405 s. ISBN 978-80-8094-899-3.
- [2] Ostertagová, E.: Aplikovaná štatistika. Elfa, Košice, 2011. 161 s. ISBN 978-80-8086-171-1.
- [3] Ostertagová, E.: Aplikovaná štatistika v počítačovom prostredí MATLABu. Equilibria, Košice, 2012. 193 s. ISBN 978-80-8143-006-0.
- [4] Ostertagová, E., Kalina, P.: Použitie Matlabu pri štatistickej kontrole kvality. Transfer inovácií, 16/2010, s. 202-204. ISSN 1337-7094.

**Príspevok bol spracovaný v rámci riešenia grantovej úlohy VEGA č. 1/0102/11 Metódy a techniky experimentálneho modelovania vnútropodnikových výrobných a nevýrobných procesov.**