APLIKÁCIA PEARSONOVHO TESTU DOBREJ ZHODY V TECHNICKEJ PRAXI

PhDr. Eva Ostertagová, PhD.
Technická univerzita v Košiciach
Fakuta elektrotechniky a informatiky
Katedra matematiky a teoretickej informatiky
Němcovej 32, 042 00 Košice
eva.ostertagova@tuke.sk

Abstract

The aim of article is an application of Pearson's chi-squared test by means of MATLAB software. This test is used to test if a sample of data came from a population with a specific distribution.

Key words: null and alternative hypothesis, Poisson distribution, observed and theoretical frequency, Pearson's test statistic

ÚVOD

Testy dobrej zhody umožňujú testovať hypotézy o celom tvare rozdelenia pravdepodobnosti, ktoré je reprezentované distribučnou funkciou. Tieto testy umožňujú na vopred zvolenej hladine významnosti α testovať nulovú hypotézu H_0, že daný náhodný výber bol realizovaný z rozdelenia stanoveného typu, ale prípadne s neznámymi parametrami.

Medzi testy dobrej zhody patria predovšetkým Pearsonov test, Kolmogorovov test, a Kolmogorovov-Smirnovov test. Tieto testy sú do istej miery univerzálne. Možno ich použiť k overeniu zhody empirického rozdelenia s akýmkoľvek modelom. Za túto univerzálnosť sa platí zníženou účinnosťou testov. Táto skutočnosť viedla k vypracovaniu špeciálnych testov založených na charakteristických vlastnostiach predpokladaného modelu. Špeciálnym testom dobrej zhody je napr. test normality pomocou momentových charakteristík.

PEARSONOV TEST

Pearsonov test (χ^2-test) patrí medzi najznámejšie testy dobrej zhody. Je to jednoduchý test založený na rozdiele medzi pozorovanými (empirickými) a očakávanými (teoretickými) početnosťami.

Daný je náhodný výber $x_1, x_2, \ldots, x_n$ 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) $I_1, I_2, \ldots, I_k$. Na vopred zvolenej hladine významnosti α testujeme nulovú hypotézu $H_0 : F(x) = F_0(x)$ proti alternatívnej hypotéze $H_1 : F(x) \neq F_0(x)$. Testovacia štatistika (charakteristika) je $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$, ktorá má pre $n \geq 50$ pri platnosti H_0 približne *chí*-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. Kritická oblasť je $K_\alpha = \left(\chi^2_{1-\alpha,\, k-r-1}, \infty \right)$.

Pri Pearsonovom teste je potrebné dôsledne dodržiavať tzv. Cochranovo pravidlo. Podmienka použitia daného testu je $n \cdot p_i \geq 5$ pre všetky $i = 1, 2, \ldots, k$. Táto podmienka sa niekedy v praxi ťažko dodržiava. Jej prísne dodržiavanie je však potrebné len pri malom počte stupňov voľnosti testovacej štatistiky χ^2. 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 platilo $n \cdot p_i \geq 1$. Ak nie sú tieto podmienky splnené, odporúča sa zlúčenie susedných intervalov s malou početnosťou.

VÝPOČET V PROSTREDÍ MATLABU

Budeme riešiť úlohu: U 500 televízorov určitého typu bol sledovaný počet porúch počas 100 hodín skúšobnej prevádzky. Výsledky pozorovania sú zaznamenané v tab. 1. Na hladine významnosti $\alpha = 0{,}05$ budeme testovať hypotézu o tom, že počet porúch má Poissonovo rozdelenie.

Tab.1 Tabuľka početností

x_i	0	1	2	3	4	5	6	7
n_i	199	169	87	31	9	3	1	1

Danú úlohu budeme riešiť v niekoľkých krokoch, pričom pri výpočte budeme vhodne aplikovať MATLAB.

Postavíme proti sebe hypotézy $H_0 : F(x) = F_0(x)$ a $H_1 : F(x) \neq F_0(x)$. Testujeme či sa distribučná funkcia F rozdelenia, z ktorého pochádza náhodný výber, rovná distribučnej funkcii F_0 Poissonovho rozdelenia *poiss*(λ), pričom parameter rozdelenia λ je neznámy. Použijeme hladinu významnosti $\alpha = 0{,}05$.

Keďže parameter Poissonovho rozdelenia $E(X) = \lambda$ nepoznáme, aproximujeme ho bodovým odhadom, ktorým je výberový priemer, t.j. $\lambda \approx \bar{x}$. Aplikácia MATLABu:

```
x=[0*ones(1,199),1*ones(1,169),2*ones(1,87),...
3*ones(1,31),4*ones(1,9),5*ones(1,3),6,7];
n=length(x);vp=mean(x)
```

Výsledok z MATLABu je:

```
vp = 1
```

Pre výpočet testovacej štatistiky vytvoríme pomocnú tabuľku, kde v 1. stĺpci budú hodnoty x_i, pričom prvá hodnota 0 zostane v tomto prípade nezmenená a poslednú hodnotu 7 nahradíme $+\infty$

(viď definičný obor funkcie pravdepodobnosti Poissonovho rozdelenia). V 2. stĺpci budú teoretické pravdepodobnosti p_i vypočítané podľa definície funkcie pravdepodobnosti Poissonovho rozdelenia $f(x)=\begin{cases}\dfrac{\lambda^x\cdot e^{-\lambda}}{x!}, & ak \quad x=0,\,1,\,2,\,\ldots\\ 0, & inak\end{cases}$.

V MATLABe použijeme funkciu *poisspdf*(x,λ). Do 3. stĺpca uložíme teoretické početnosti $n{\cdot}p_i$, ktoré vypočítame vynásobením teoretických pravdepodobností p_i číslom n. Do 4. stĺpca dáme empirické početnosti n_i. Ilustrujeme použitie MATLABu:

```
xi=[0:6,inf];ni=[199,169,87,31,9,3,1,1];
p1=poisspdf(0:6,vp);p2=1-sum(p1);
pi=[p1,p2];npi=n*pi;tab1=[xi;pi;npi;ni]'
```

Dostaneme takýto výstup z MATLABu:

```
tab1 =
         0    0.3679  183.9397  199.0000
    1.0000    0.3679  183.9397  169.0000
    2.0000    0.1839   91.9699   87.0000
    3.0000    0.0613   30.6566   31.0000
    4.0000    0.0153    7.6642    9.0000
    5.0000    0.0031    1.5328    3.0000
    6.0000    0.0005    0.2555    1.0000
       Inf    0.0001    0.0416    1.0000
```

Ďalej použijeme Cochranovo pravidlo. Počet tried $k = 8$, počet odhadovaných parametrov rozdelenia $r = 1$, teda $k-r-1 = 8-1-1 = 6 \geq 6$, kde stačí, aby $n\cdot p_i \geq 1$. V treťom stĺpci tabuľky vidíme, že táto podmienka nie je splnená na konci tabuľky. Ľahko sa presvedčíme, že nestačí zlúčiť posledné tri triedy. Je potrebné zlúčiť až štyri susedné triedy z konca tabuľky. Potom bude platiť $k-r-1 = 5-1-1 = 3 \geq 3$. Tu ale stačí $n\cdot p_i \geq 4$, čo už bude splnené. Tabuľku doplníme o stĺpec hodnôt $\dfrac{(n_i - n\cdot p_i)^2}{n\cdot p_i}$, po sčítaní ktorých dostaneme testovaciu charakteristiku $\chi^2 = 4{,}8574$.

Ilustrujeme MATLAB napr. v takejto forme:

```
npi=[npi(1:4),sum(npi(5:8))];
ni=[ni(1:4),sum(ni(5:8))];w=(ni-npi).^2./npi;
tab2=[npi;ni;w]',chi2=sum(w)
```

Výstup z MATLABu:

```
tab2 =
  183.9397  199.0000    1.2331
  183.9397  169.0000    1.2134
   91.9699   87.0000    0.2686
   30.6566   31.0000    0.0038
    9.4941   14.0000    2.1385
chi2 = 4.8574
```

Hypotézu H_0 nezamietame, keď platí $\chi^2 \leq \chi^2_{1-\alpha,\,k-r-1}$. Kritickou hodnotou testu je kvantil $c = \chi^2_{0{,}95;\,3} = 7{,}8147$ (viď štatistické tabuľky). Výpočet kvantilu pomocou MATLABu:

```
c=chi2inv(.95,3)
```

Výstup z MATLABu je:

```
c = 7.8147
```

Keďže $\chi^2 = 4{,}8574 \notin K_\alpha = (\,7{,}8147\,;\,\infty\,)$, hypotézu H_0 nezamietame. Pearsonov test so spoľahlivosťou 95 % preukázal, že ide o Poissonovo rozdelenie.

Budeme ilustrovať ešte aplikáciu štandardnej funkcie MATLABu *chi2gof*, ktorej použitie vedie k tomu istému výsledku:

```
bins=[0:6,inf];
obsCounts=[199,169,87,31,9,3,1,1];
n=sum(obsCounts);
x=[0*ones(1,199),1*ones(1,169),2*ones(1,87),...
3*ones(1,31),4*ones(1,9),5*ones(1,3),6,7];
vp=mean(x);p1=poisspdf(0:6,vp);p2=1-sum(p1);
lambdaHat=[p1,p2];
expCounts =n*lambdaHat;
[h,p,st]=chi2gof((0:7),'ctrs',(0:7),'frequency',...
obsCounts,'expected',expCounts,'nparams',1)
c=chi2inv(0.95,3)
```

Časť výstupu z MATLABu je:

```
h = 0
p = 0.1825
chi2stat: 4.8574
df: 3
O: [199 169 87 31 14]
E: [183.9397 183.9397 91.9699 30.6566 9.4941]
c = 7.8147
```

Výsledok h = 0 znamená, že hypotézu H_0 nezamietame na hladine významnosti $\alpha = 0{,}05$. To isté rozhodnutie je možné uskutočniť aj na základe p-hodnoty, pretože $p = 0{,}1825 > \alpha = 0{,}05$. Ostatné výsledky sú zrejmé, ak porovnáme výsledky oboch spôsobov riešenia.

ZÁVER

Personov test dobrej zhody je významnou metódou matematickej štatistiky, ktorá umožňuje overenie štatistickej zhody empirického rozdelenia s niektorým z teoretických rozdelení. Je univerzálnym testom pre diskrétne aj spojité rozdelenia, ktorý ale vyžaduje dostatočne veľký rozsah výberového súboru.

Literatúra

[1] OSTERTAGOVÁ, E.: Aplikovaná štatistika. Elfa, Košice, 2011.

[2] OSTERTAGOVÁ, E.: Pravdepodobnosť a matematická štatistika v príkladoch. Elfa, Košice, 2005.

Príspevok bol spracovaný v rámci riešenia grantovej úlohy KEGA 3/7353/09.