

## Cvičení 5 s návodem

### Příklady na testování exponenciálního a Poissonova rozložení

#### Teoretická část

##### I. Test dobré shody

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z rozložení s distribuční funkcí  $\Phi(x)$ .

Testová statistika  $K = \sum_{j=1}^r \frac{(n_j - np_j)^2}{np_j}$  se za platnosti nulové hypotézy asymptoticky řídí

rozložením  $\chi^2(r-p-1)$ , kde  $p$  je počet odhadovaných parametrů daného rozložení.

Přitom

$n_j$  je absolutní četnost  $j$ -tého třídícího intervalu pro veličinu  $X$  resp.  $j$ -té varianty veličiny  $X$ ,  
 $np_j$  je teoretická četnost  $j$ -tého třídícího intervalu pro veličinu  $X$  resp.  $j$ -té varianty veličiny  $X$ .  
Platí-li nulová hypotéza, pak  $p_j = \Phi(u_{j+1}) - \Phi(u_j)$  resp.  $p_j = \Phi(x_{[j]}) - \lim_{x \rightarrow x_{[j]}^-} \Phi(x) = P(X = x_{[j]})$ .

Kritický obor:  $W = \langle \chi^2_{1-\alpha}(r-p-1), \infty \rangle$ .

Jestliže  $K \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ .

Aproximace se považuje za vyhovující, když  $np_j \geq 5$ ,  $j = 1, \dots, r$ .

Při nesplnění podmínky  $np_j \geq 5$ ,  $j = 1, \dots, r$  je třeba některé intervaly resp. varianty slučovat.

##### II. Jednoduchý test exponenciálního rozložení (Darlingův test)

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z exponenciálního rozložení.

Testová statistika  $K = \frac{(n-1)S^2}{M^2}$ , která se v případě platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(n-1)$ .

Přitom  $M$  je výběrový průměr a  $S^2$  je výběrový rozptyl daného náhodného výběru.

Kritický obor:  $W = \langle 0, \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty \rangle$ .

Jestliže  $K \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ .

##### III. Jednoduchý test Poissonova rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z Poissonova rozložení.

Testová statistika  $K = \frac{(n-1)S^2}{M}$ , která se v případě platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(n-1)$ .

Přitom  $M$  je výběrový průměr a  $S^2$  je výběrový rozptyl daného náhodného výběru.

Kritický obor:  $W = \langle 0, \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty \rangle$ .

Jestliže  $K \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ .

**Příklad 1.:** V systému hromadné obsluhy byla sledována doba obsluhy 70 zákazníků (v min). Výsledky jsou uvedeny v tabulce rozložení četností:

Doba obsluhy	Počet zákazníků
(0, 3]	14
(3,6]	16
(6,9]	10
(9,12]	9
(12,15]	8
(15,18]	5
(18,21]	3
(21,24]	5

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení. Použijte:

- test dobré shody,
- Darlingův test exponenciálního rozložení

**Řešení:**

Testujeme  $H_0$ : náhodný výběr  $X_1, \dots, X_{70}$  pochází z  $Ex(\lambda)$  proti  $H_1$ : non  $H_0$ .

Ad a) Nejprve odhadneme parametr  $\lambda$  exponenciálního rozložení:

$$\hat{\lambda} = \frac{1}{m} = \frac{1}{\frac{1}{n} \sum_{j=0}^r n_j x_{[j]}} = \frac{1}{70} (14 \cdot 1,5 + 16 \cdot 4,5 + \dots + 5 \cdot 22,5) = 0,1122$$

Pravděpodobnost, že náhodná veličina s rozložením  $Ex(\lambda)$ , kde  $\lambda = 0,1122$  se bude realizovat v intervalu  $(u_j, u_{j+1})$  je

$$p_j = \Phi(u_{j+1}) - \Phi(u_j), j = 1, \dots, r, \text{ kde } \Phi(x) = 1 - e^{-\lambda x}.$$

Výpočty potřebné pro stanovení testové statistiky  $K$  uspořádáme do tabulky.

$(u_j, u_{j+1})$	$x_{[j]}$	$n_j$	$p_j$	$np_j$
(0, 3]	1,5	14	0,2858	20,0033
(3,6]	4,5	16	0,2041	14,2871
(6,9]	7,5	10	0,1458	10,2044
(9,12]	10,5	9	0,1041	7,2884
(12,15]	13,5	8	0,0744	5,2056
(15,18]	16,5	5	0,0531	3,7181
(18,21]	19,5	3	0,0378	2,6556
(21,24]	22,5	5	0,0271	1,8967

Podmínky dobré aproximace nejsou splněny, sloučíme tedy intervaly (15,18], (18,21] a (21,24].

$(u_j, u_{j+1})$	$x_{[j]}$	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2 / np_j$
(0, 3]	1,5	14	0,2858	20,0033	1,8017
(3,6]	4,5	16	0,2041	14,2871	0,2054
(6,9]	7,5	10	0,1458	10,2044	0,0041
(9,12]	10,5	9	0,1041	7,2884	0,4020
(12,15]	13,5	8	0,0744	5,2056	1,5000
(15,24]	19,5	13	0,1181	8,2704	2,7047

Testová statistika  $K = 1,8017 + \dots + 2,7047 = 6,6178$ ,  $r = 6$ ,  $p = 1$ ,  $r - p - 1 = 4$ ,  $\chi^2_{0,95}(4) = 9,4877$ .

Testová statistika se nerealizuje v kritickém oboru  $W = \langle 9,4877, \infty \rangle$ , na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že doba obsluhy se řídí exponenciálním rozložením.

Ad b)

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$m = \frac{1}{70} (14 \cdot 1,5 + 16 \cdot 4,5 + \dots + 5 \cdot 22,5) = 8,9143$$

$$s^2 = \frac{1}{69} [19 \cdot (1,5 - 8,9143)^2 + 16 \cdot (4,5 - 8,9143)^2 + \dots + 5 \cdot (22,5 - 8,9143)^2] = 41,1447$$

$$\text{Testová statistika: } K = \frac{(n-1)S^2}{M^2} = \frac{69 \cdot 41,1447}{8,9143^2} = 35,7265.$$

$$\text{Kritický obor: } W = \langle 0, \chi^2_{0,025}(69) \rangle \cup \langle \chi^2_{0,975}(69), \infty \rangle = \langle 0; 47,9242 \rangle \cup \langle 93,8565, \infty \rangle.$$

$H_0$  zamítáme na asymptotické hladině významnosti 0,05.

### Řešení pomocí MATLABu:

Ad a)

Zadáme vektor mezí  $uj = [0:3:24]'$ , vektor středů  $xj = [1.5:3:22.5]'$ , vektor pozorovaných četností  $nj = [14 16 10 9 8 5 3 5]'$ . Celkový rozsah souboru je  $n = \text{sum}(nj)$  ( $n=70$ ) a parametr  $\lambda = 1/\text{sum}(nj \cdot xj)/n$  ( $\lambda=0,1121$ ).

Vypočteme teoretické četnosti  $npj = n \cdot \text{diff}(\text{expcdf}(uj, \lambda))$

Protože nejsou splněny podmínky dobré aproximace pro poslední tři intervaly, je třeba je sloučit do jednoho.

Zadáme nový vektor  $uj = [0 3 6 9 12 15 24]'$  a nový vektor pozorovaných četností  $nj = [14 16 10 9 8 13]'$ .

Znovu vypočteme teoretické četnosti  $npj = n \cdot \text{diff}(\text{expcdf}(xj, \lambda))$ .

Nyní již jsou splněny podmínky dobré aproximace.

Vypočítáme testovou statistiku  $K = \text{sum}((nj - npj).^2 ./ npj)$  a kvantil  $\chi^2_{1-\alpha}(r - p - 1) = \chi^2_{0,95}(4)$  pomocí funkce `chi2inv(0.95,4)`.

Protože testová statistika  $K = 6,6178$  se nerealizuje v kritickém oboru  $W = \langle 9,4877, \infty \rangle$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

Ad b)

Použijeme funkci darling.m.

Zadáme vstupní vektor středů třídících intervalů společně s absolutními četnostmi třídících intervalů:

$X = [1.5 \ 14; 4.5 \ 16; 7.5 \ 10; 10.5 \ 9; 13.5 \ 8; 16.5 \ 5; 19.5 \ 3; 22.5 \ 5]'$

Zavoláme funkci darling:

$[zमितnuti, K, p, lambda] = \text{darling}(X)$

Dostaneme výsledek:

zमितnuti =

1

K =

35.7265

p =

6.1430e-004

lambda =

8.9143

**Příklad 2.:** Na jistém nádraží byl sledován počet přijíždějících vlaků za 1 h. Pozorování bylo prováděno celkem 15 dnů (tj. 360 h) a výsledky jsou uvedeny v tabulce:

Počet vlaků za 1 hodinu	0	1	2	3	4	5	6	7 a víc
četnost	27	93	103	58	50	21	6	2

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že počet přijíždějících vlaků za 1 h se řídí Poissonovým rozložením, a to a) testem dobré shody, b) jednoduchým testem Poissonova rozložení.

**Řešení:**

Testujeme  $H_0$ : náhodný výběr  $X_1, \dots, X_{360}$  pochází z  $Po(\lambda)$  proti  $H_1$ : non  $H_0$ .

Ad a) Nejprve odhadneme parametr  $\lambda$  Poissonova rozložení:

$$\hat{\lambda} = m = \frac{1}{n} \sum_{j=0}^r n_j x_{[j]} = \frac{1}{360} (27 \cdot 0 + 93 \cdot 1 + \dots + 2 \cdot 7) = 2,3$$

Pravděpodobnost, že náhodná veličina s rozložením  $Po(\lambda)$ , kde  $\lambda = 2,3$  bude nabývat hodnot

$$0, 1, \dots, 7 \text{ a víc je } p_j = \frac{\lambda^j}{j!} e^{-\lambda} = \frac{2,3^j}{j!} e^{-2,3}, j = 0, 1, \dots, 6, p_7 = 1 - (p_0 + p_1 + \dots + p_6).$$

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	$n_j$	$p_j$	$n p_j$
0	27	0,1003	36,0932
1	93	0,2306	83,0143
2	103	0,2652	95,4665
3	58	0,2033	73,1910
4	50	0,1169	43,0848
5	21	0,0538	19,3590
6	6	0,0216	7,4210
7 a víc	2	0,0094	3,3703

Podmínky dobré aproximace nejsou splněny, sloučíme tedy varianty 6 a 7 a víc.

j	n <sub>j</sub>	p <sub>j</sub>	np <sub>j</sub>	(n <sub>j</sub> - np <sub>j</sub> ) <sup>2</sup> / np <sub>j</sub>
0	27	0,1003	36,0932	2,2909
1	93	0,2306	83,0143	1,2012
2	103	0,2652	95,4665	0,5945
3	58	0,2033	73,1910	3,1529
4	50	0,1169	43,0848	1,4887
5	21	0,0538	19,3590	0,1391
6 a víc	8	0,0300	10,7912	0,7220

$K = 2,2909 + 1,2012 + \dots + 0,7220 = 9,5892$ ,  $r = 7$ ,  $p = 1$ ,  $r - p - 1 = 5$ ,  $\chi^2_{0,95}(5) = 11,0705$ .  
Protože  $9,5892 < 11,0705$ , nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05. Nepodařilo se tedy prokázat, že počty přijíždějících vlaků za 1 h se neřídí Poissonovým rozložením.

Ad b)

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$m = \frac{1}{360} (27 \cdot 0 + 93 \cdot 1 + \dots + 2 \cdot 7) = 2,3$$

$$s^2 = \frac{1}{359} [27 \cdot (0 - 2,3)^2 + 93 \cdot (1 - 2,3)^2 + \dots + 2 \cdot (7 - 2,3)^2] = 2,121448$$

$$\text{Testová statistika: } K = \frac{(n-1)S^2}{M} = \frac{359 \cdot 2,121448}{2,3} = 331,1304,$$

$$\text{Kritický obor: } W = \langle 0, \chi^2_{0,025}(359) \rangle \cup \langle \chi^2_{0,975}(359), \infty \rangle = \langle 0,308,4 \rangle \cup \langle 413,4, \infty \rangle$$

$H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

### Řešení pomocí MATLABu:

Ad a) Zadáme vektor  $x_j = [0:7]'$  a vektor pozorovaných četností  $n_j = [27 \ 93 \ 103 \ 58 \ 50 \ 21 \ 6 \ 2]'$ .

Celkový rozsah souboru je  $n = \text{sum}(n_j)$  a odhad parametru lambda:  $\text{lambda} = \text{sum}(n_j \cdot x_j) / n$ .

Vektor pravděpodobností:  $p_j = \text{poisspdf}(x_j, \text{lambda})$ .

Poslední pravděpodobnost musíme nahradit doplňkem do jedné:  $p_j(8) = 1 - \text{sum}(p_j(1:7))$

Vypočteme teoretické četnosti  $np_j = n \cdot p_j$

Protože nejsou splněny podmínky dobré aproximace pro variantu 7 a víc, je třeba sloučit varianty 6 a 7 a víc do jedné.

Zadáme nový vektor  $x_j = [0:6]'$  a nový vektor pozorovaných četností  $n_j = [27 \ 93 \ 103 \ 58 \ 50 \ 21 \ 8]'$ .

Znovu vypočteme vektor pravděpodobností:  $p_j = \text{poisspdf}(x_j, \text{lambda})$ .

Poslední pravděpodobnost musíme nahradit doplňkem do jedné:  $p_j(7) = 1 - \text{sum}(p_j(1:6))$

Vypočteme teoretické četnosti  $np_j = n \cdot p_j$

Nyní již jsou splněny podmínky dobré aproximace.

Vypočítáme testovou statistiku  $K = \text{sum}((n_j - np_j).^2 ./ np_j)$  a kvantil  $\chi^2_{1-\alpha}(r - p - 1) = \chi^2_{0,95}(5)$  pomocí funkce `chi2inv(0,95,5)`.

Protože testová statistika  $K = 9,582$  se nerealizuje v kritickém oboru  $W = \langle 11,0705, \infty \rangle$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

Ad b)

Použijeme funkci `darling.m`.

Zadáme vstupní vektor variant společně s absolutními četnostmi těchto variant:

`X = [1.5 14; 4.5 16; 7.5 10; 10.5 9; 13.5 8; 16.5 5; 19.5 3; 22.5 5]'`

Zavoláme funkci `darling`:

`[zamitnuti, K, p, lambda] = darling(X, 'poiss')`

Dostaneme výsledek:

`zamitnuti =`

`0`

`K =`

`331.1304`

`p =`

`0.2968`

`lam =`

`2.3000`

### **Další možnosti ověřování exponenciálního rozložení:**

využití funkce `probplot` (pravděpodobnostně – pravděpodobnostní graf),

Kolmogorovův – Smirnovův test (funkce `kstest`).

Použití K-S testu

Vygenerujeme 100 hodnot z exponenciálního rozložení s parametrem 2:

`x = exprnd(2, 100,);`

Provedeme porovnání výběrové distribuční funkce s distribuční funkcí exponenciálního rozložení  $Ex(2)$ :

`[h, p, ksstat] = kstest(x, [x, expcdf(x, 2)])`

Význam výstupních parametrů:

$h = 0$ , když nezamítáme hypotézu o exponenciálním rozložení  $Ex(2)$  na hladině významnosti 0,05,  $h = 1$ , když tuto hypotézu zamítáme.

$p$  je odpovídající  $p$ -hodnota

$ksstat$  je hodnota testové statistiky.