5. Testování exponenciálního a Poissonova rozložení

5.1. Věta (test dobré shody – viz přednáška 2)
H_0: náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$
H_1: non H_0

Testová statistika: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j} \approx \chi^2(r-p-1)$, když H_0 platí

r … počet třídicích intervalů $\langle u_j, u_{j+1} \rangle$ ve spojitém případě resp. počet variant $x_{[j]}$ v diskrétním případě.
n_j … absolutní četnost j-tého třídicího intervalu resp. j-té varianty.
p_j … pravděpodobnost, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu resp. j-tou variantou.
p … počet odhadovaných parametrů testovaného rozložení.

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

$K \in W \Rightarrow$ H_0 zamítáme na asymptotické hladině významnosti α.
Podmínky dobré aproximace: $np_j \geq 5$, j = 1, ..., r. Při nesplnění těchto podmínek se doporučuje slučování některých třídicích intervalů resp. variant.

5.2. Příklad: Byla zjišťována doba životnosti 45 součástek (v hodinách). Ze získaných údajů byl vypočten výběrový průměr m = 99,93 h a výběrový rozptyl s^2 = 7328,9 h^2. Máme k dispozici roztříděné údaje:

Doba životnosti	Počet součástek
$\langle 0,50 \rangle$	15
$\langle 50,100 \rangle$	14
$\langle 100,150 \rangle$	6
$\langle 150,200 \rangle$	5
$\langle 200,250 \rangle$	2
$\langle 250,300 \rangle$	1
$\langle 300,350 \rangle$	1
$\langle 350,400 \rangle$	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že doba životnosti se řídí exponenciálním rozložením.

Řešení: $\hat{\lambda} = \frac{1}{m} = \frac{1}{99,93}$, testujeme H_0: $X_1, \ldots, X_{45} \sim Ex\left(\frac{1}{99,93}\right)$ proti H_1: non H_0. Počítáme pravděpodobnosti $p_j = \int_{u_j}^{u_{j+1}} \frac{1}{99,93} e^{-\frac{x}{99,93}} dx$, $j = 1, 2, \ldots, 8$

j	$\langle u_j, u_{j+1} \rangle$	n_j	p_j	$np_j = 45p_j$
1	$\langle 0,50 \rangle$	15	0,3937	17,72
2	$\langle 50,100 \rangle$	14	0,2387	10,74
3	$\langle 100,150 \rangle$	6	0,1447	6,51
4	$\langle 150,200 \rangle$	5	0,0878	3,95
5	$\langle 200,250 \rangle$	2	0,0532	2,39
6	$\langle 250,300 \rangle$	1	0,0323	1,45
7	$\langle 300,350 \rangle$	1	0,0196	0,88
8	$\langle 350,400 \rangle$	1	0,0119	0,53

Vidíme, že pro $j = 4, \ldots, 8$ nejsou splněny podmínky dobré aproximace. Posledních 5 intervalů tedy sloučíme do jednoho.

Dostaneme novou tabulku

j	$\langle u_j, u_{j+1} \rangle$	n_j	p_j	$np_j = 45p_j$
1	$\langle 0,50 \rangle$	15	0,3937	17,7157
2	$\langle 50,100 \rangle$	14	0,2387	10,7413
3	$\langle 100,150 \rangle$	6	0,1447	6,5127
4	$\langle 150,400 \rangle$	10	0,2046	9,2084

Testová statistika:

$$K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j} = \frac{(15-17,7157)^2}{17,7157} + \frac{(14-10,7413)^2}{10,7413} + \frac{(6-6,5127)^2}{6,5127} + \frac{(10-9,2084)^2}{9,2084} =$$

$$= 1,5133$$

Kritický obor: $W = \langle \chi^2_{1-\alpha}(r-p-1), \infty) = \langle \chi^2_{0,95}(4-1-1), \infty) = \langle 5,9915, \infty)$

$K \notin W \Rightarrow$ H_0 nezamítáme na asymptotické hladině významnosti 0,05.

5.3. Poznámka: V MATLABu se test dobré shody pro exponenciální rozložení provádí pomocí funkce tds_exp.m.

```
function [zamitnuti,K,p,lambda]=tds_exp(uj,nj,alfa)
% test dobre shody k overeni exponencialniho rozlozeni
% syntaxe: [zamitnuti,K,p,lambda]=tds_exp(uj,nj,alfa)
% vstupni parametry:
% uj ... sloupcovy vektor s mezemi tridicich intervalu
% nj ... sloupcovy vektor absolutnich cetnosti tridicich intervalu
% alfa ... hladina vyznamnosti testu
% vystupni parametry:
% zamitnuti ... =0, kdyz H0 nezamitame
%                   =1, kdyz H0 zamitame
% K ... hodnota testove statistiky
% p ... p-hodnota testu
% lambda ... odhad parametru exponencialniho rozlozeni
delka=size(uj);
delka=delka(:,1);
dti=diff(uj/2);
xj=[uj(1:delka-1)+dti];
n=sum(nj);
lambda=n/(nj'*xj);
npj=[n*diff(expcdf(uj,1/lambda))];
%test podminek dobre aproximace....hodnota 1 pro poruseni
if sum(npj<5)>0
   poruchy_podminek=(npj<5)'
   error('Nejsou splneny podminky dobre aproximace.')
end;
K=sum((nj-npj).^2./npj);
kvantil=chi2inv(1-alfa,size(nj,1)-2);
p=1-chi2cdf(K,size(nj,1)-2);
zamitnuti=(p<alfa);
```

Při řešení pomocí funkce tds_exp.m zohledníme, že při původním třídění do 8 intervalů nebyly splněny podmínky dobré aproximace a budeme pracovat se 4 intervaly.

Zadáme vektor mezí:
uj=[0;50;100;150;400];
vektor pozorovaných četností:
nj=[15;14;6;10];
a hladinu významnosti:
alfa=0.05;

Zavoláme funkci tds_exp:
[zamitnuti,K,p,lambda]=tds_exp(uj,nj,alfa)

Dostaneme výsledek:
zamitnuti=0, K=1.4068, p=0.4949, lambda=0.0091
Protože p-hodnota je větší než hladina významnosti 0,05, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

5.4. Příklad: Sledujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na zubní pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů (máme 16 x 75 = 1200 intervalů).

Počet pacientů	0	1	2	3	4	5	6	7	8 a víc
četnost	79	188	282	275	196	114	45	10	11

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že počet příchozích pacientů během půl hodiny se řídí Poissonovým rozložením.

Řešení:

$$\hat{\lambda} = m = \frac{1}{1200}\left(79\cdot 0 + 188\cdot 1 + 282\cdot 2 + 275\cdot 3 + 196\cdot 4 + 114\cdot 5 + 45\cdot 6 + 10\cdot 7 + 11\cdot 8\right) = 2{,}7992,$$

testujeme H_0: $X_1, \ldots, X_{1200} \sim Po(2{,}7992)$ proti H_1: non H_0. Počítáme pravděpodobnosti

$$p_j = \frac{2{,}7992^j}{j!} e^{-2{,}7992},\ j = 0, 1, \ldots, 7,\ p_8 = 1 - \sum_{j=0}^{7} p_j .$$

j	0	1	2	3	4	5	6	7	8
n_j	79	188	282	275	196	114	45	10	11
np_j	73,0329	204,4313	286,1186	266,9646	186,8195	104,5878	48,7931	19,5114	9,7406

Podmínky dobré aproximace jsou splněny.
Testová statistika:

$$K = \sum_{j=1}^{r} \frac{\left(n_j - np_j\right)^2}{np_j} = \frac{\left(79 - 73{,}0329\right)^2}{73{,}0329} + \frac{\left(188 - 204{,}4313\right)^2}{204{,}4313} + \ldots + \frac{\left(11 - 9{,}7406\right)^2}{9{,}7406} = 8{,}5019$$

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}\left(r - p - 1\right), \infty\right) = \left\langle \chi^2_{0{,}95}\left(9 - 1 - 1\right), \infty\right) = \left\langle 14{,}067, \infty\right)$

$K \notin W \Rightarrow$ H_0 nezamítáme na asymptotické hladině významnosti 0,05.

5.5. Poznámka: V MATLABu se test dobré shody pro Poissonovo rozložení provádí pomocí funkce tds_poiss.m.

```
function [zamitnuti,K,p,lambda]=tds_poiss(xj,nj,alfa)
% test dobre shody k overeni Poissonova rozlozeni
% syntaxe: [zamitnuti,K,p,lambda]=tds_poiss(xj,nj,alfa)
% vstupni parametry:
% xj ... sloupcovy vektor variant sledovane veliciny
% nj ... sloupcovy vektor absolutnich cetnosti variant
% alfa ... hladina vyznamnosti testu
% vystupni parametry>
% zamitnuti ... =0, kdyz H0 nezamitame
%                   =1, kdyz H0 zamitame
% K ... hodnota testove statistiky
% p ... p-hodnota testu
% lambda ... odhad parametru Poissonova rozlozeni
n=sum(nj);
r=size(xj,1);
lambda=sum(nj'*xj)/n;
pj=poisspdf(xj(1:r-1),lambda);
pj=[pj;1-sum(pj)];
npj=n*pj;
%test podminek dobre aproximace....hodnota 1 pro poruseni
if sum(npj<5)>0
   poruchy_podminek=(npj<5)'
   error('Nejsou splneny podminky dobre aproximace.')
end;
K=sum((nj-npj).^2./npj);
kvantil=chi2inv(1-alfa,size(nj,1)-2);
p=1-chi2cdf(K,size(nj,1)-2);
zamitnuti=(p<alfa);
```

Příklad vyřešíme pomocí funkce tds_poiss.m.

Zadáme vektor variant:
xj=[0:8]';
vektor pozorovaných četností:
nj=[79;188;282;275;196;114;45;10;11];
a hladinu významnosti:
alfa=0.05;

Zavoláme funkci tds_poiss:
[zamitnuti,K,p,lambda]=tds_poiss(xj,nj,alfa)

Dostaneme výsledek:
zamitnuti=0, K=8.5019, p=0.2904, lambda=2.7992

Protože p-hodnota je větší než hladina významnosti 0,05, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

5.6. Věta: Darlingův (jednoduchý) test exponenciálního rozložení
Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z exponenciálního rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Ex(\lambda)$ je $E(X) = 1/\lambda$ a rozptyl je $D(X) = 1/\lambda^2$.

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

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

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

5.7. Příklad: Pro data z příkladu 5.2. proveďte na hladině významnosti 0,05 Darlingův test.
Řešení: $n = 45$, $m = 99{,}93$ h, $s^2 = 7328{,}9$ h^2

Testová statistika: $K = \frac{(n-1)S^2}{M^2} = \frac{44 \cdot 7328{,}91}{99{,}93^2} = 32{,}2924$

Kritický obor:

$W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty) = \langle 0, \chi^2_{0{,}025}(44)\rangle \cup \langle \chi^2_{0{,}975}(44), \infty) = \langle 0{,}27{,}575\rangle \cup \langle 64{,}202,$

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o exponenciálním rozložení nezamítáme na asymptotické hladině významnosti 0,05.

5.8. Věta: Jednoduchý test Poissonova rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z Poissonova rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim Po(\lambda)$ je $E(X) = \lambda$ a rozptyl je $D(X) = \lambda$.

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

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

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

5.9. Příklad: Pro data z příkladu 5.4. proveďte na hladině významnosti 0,05 jednoduchý test Poissonova rozložení.

Řešení: Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$\hat{\lambda} = \mathrm{m} = \frac{1}{1200}\left(79\cdot 0 + 188\cdot 1 + 282\cdot 2 + 275\cdot 3 + 196\cdot 4 + 114\cdot 5 + 45\cdot 6 + 10\cdot 7 + 11\cdot 8\right) = 2{,}7992$$

$$\mathrm{s}^2 = \frac{1}{1199}\left[79\cdot\left(0-2{,}7992\right)^2 + 188\cdot\left(1-2{,}7992\right)^2 + \ldots + 11\cdot\left(8-2{,}7992\right)^2\right] = 2{,}6594$$

$$\mathrm{K} = \frac{\left(\mathrm{n}-1\right)\mathrm{S}^2}{\mathrm{M}} = \frac{1199\cdot 2{,}6594}{2{,}7882} = 1139{,}1$$

Kritický obor: $\mathrm{W} = \left\langle 0, \chi^2{}_{\alpha/2}\left(\mathrm{n}-1\right)\right\rangle \cup \left\langle \chi^2{}_{1-\alpha/2}\left(\mathrm{n}-1\right), \infty\right) = \left\langle 0; 1104{,}93\right\rangle \cup \left\langle 1296{,}86; \infty\right)$

H_0 nezamítáme na asymptotické hladině významnosti 0,05.

5.10. Poznámka: Darlingův test i jednoduchý test Poissonova rozložení můžeme v MATLABu provést pomocí funkce darling.m.

```
function [zamitnuti,K,p,lambda]=darling(X,distrib,alfa)
% TEST K OVERENI EXPONENCIALNIHO A POISSONOVA ROZLOZENI
%  function [zamitnuti,K,p,lambda]=darling(X,ROZLOZENI,ALFA)
%    X muze byt n-vektor pozorovanych velicin, jejichz rozdeleni overujeme;
%        - pro souhrnne zadana data je X tvaru (r x 2), kde prvni sloupec
%          obsahuje jednotlive varianty a druhy sloupec cetnosti;
%        - pro vypoctene statistiky je X=[n,m,s2], kde n=pocet pozorovani,
%                          m=vyberovy prumer a s2=vyberovy rozptyl
%    ROZLOZENI je 'exp' pro overeni exponencialniho rozlozeni (implicitni)
%          nebo 'poiss' pro overeni Poissonova rozlozeni
%    ALFA je hladina vyznamnosti testu (implicitne 0.05)
%
%   vystup: zamitnuti=1 => ZAMITAME hypotezu o shode rozdeleni
%          zamitnuti=0 => hypotezu o shode rozdeleni NEZAMITAME
%          K = hodnota testoveho kriteria
%          p = p-hodnota testu
%          lambda = odhadnuty parametr rozdeleni

% (c) Ondrej Petrik, 10.03.2010

if (nargin==1) distrib='exp'; end
if (nargin<3) alfa=0.05; end

[a,b]=size(X);
if(a<b) X=X';[a,b]=size(X); end
if(a==3&&b==1) n=X(1); m=X(2); s2=X(3);
elseif(b==1&&a~=3) m=mean(X); n=a; s2=var(X);
else vaha=X(:,2); n=sum(vaha);
    s2=var(X(:,1),vaha)*n/(n-1); m=X(:,1)'*vaha/n;
end

lambda=m;
if strcmp(distrib,'exp')lambda=1/m; m=m^2;  disp('Darlinguv test exponencialniho rozlozeni');
else disp('Jednoduchy test Poissonova rozlozeni');
end
K=(n-1)*s2/m;
p=2*min(chi2cdf(K,n-1),1-chi2cdf(K,n-1));
zamitnuti=(p<alfa);
```

Příklady 5.7 a 5.9 vyřešíme pomocí funkce darling.m.

Řešení příkladu 5.7: Známe údaje o době životnosti 45 součástek, údaje jsou roztříděny do 8 třídicích intervalů: $\langle 0,50\rangle : 15$, $\langle 50,100\rangle : 14$, $\langle 100,150\rangle : 6$, $\langle 150,200\rangle : 5$, $\langle 200,250\rangle : 2$, $\langle 250,300\rangle : 1$, $\langle 300,350\rangle : 1$, $\langle 350,400\rangle : 1$

Zadáme vstupní vektor středů třídicích intervalů společně s absolutními četnostmi třídicích intervalů:
X= [25 15;75 14;125 6;175 5;225 2;275 1;325 1;375 1];

Zavoláme funkci darling:
[zamitnuti,K,p,lambda]=darling(X)

Dostaneme výsledek:
zamitnuti=0, K=31.6619, p=0.1644, lambda=0.0101

Darlingův test nezamítá hypotézu o exponenciálním rozložení na asymptotické hladině významnosti 0,05.

Řešení příkladu 5.9: Sledujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na zubní pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů (máme 16 x 75 = 1200 intervalů).

Počet pacientů	0	1	2	3	4	5	6	7	8 a víc
četnost	79	188	282	275	196	114	45	10	11

Zadáme vektor variant xj = [0:8]' a vektor pozorovaných četností nj = [79 188 282 275 196 114 45 10 11]' a utvoříme matici X: X = [xj nj];

Zavoláme funkci darling:
[zamitnuti,K,p,lambda]=darling(X,'poiss')

Dostaneme výsledek:
zamitnuti=0, K=331.1304, p=0.2187, lambda=2.7992

Jednoduchý test Poissonova rozložení nezamítá hypotézu o Poissonově rozložení na asymptotické hladině významnosti 0,05.

5.11. Poznámka: Pro výpočet kvantilů Pearsonova chí-kvadrát rozložení pro počet stupňů volnosti nad 30 můžeme použít aproximační vzorec: $\chi^2{}_\alpha(n) \approx \frac{1}{2}\left(u_\alpha + \sqrt{2n-1}\right)^2$. Pro kvantily z příkladu 5.9. dostáváme:

$$\chi^2{}_{0,025}(1199) \approx \frac{1}{2}\left(u_{0,025} + \sqrt{2\cdot 1199 - 1}\right)^2 = \frac{1}{2}\left(-1{,}96 + \sqrt{2397}\right)^2 = 1104{,}46$$

$$\chi^2{}_{0,975}(1199) \approx \frac{1}{2}\left(u_{0,975} + \sqrt{2\cdot 1199 - 1}\right)^2 = \frac{1}{2}\left(1{,}96 + \sqrt{2397}\right)^2 = 1296{,}42$$

5.12. Poznámka: Pro vizuální posouzení, zda naše data pocházejí z exponenciálního rozložení, lze také použít P-P graf.

Způsob konstrukce: spočteme standardizované hodnoty $z_i = \frac{x_i - m}{s}$, i = 1, ..., n a uspořádáme je podle velikosti $z_{(1)} \le \ldots \le z_{(n)}$. Na vodorovnou osu vyneseme hodnoty distribuční funkce exponenciálního rozložení $\Phi\left(z_{(i)}\right) = 1 - e^{-\lambda z_{(i)}}$, i = 1, …, n a na svislou osu hodnoty empirické distribuční funkce $F_n\left(z_{(i)}\right) = \frac{i}{n}$, i = 1, …, n. Pokud se body $(\Phi(z_{(i)}), F(z_{(i)}))$ řadí kolem hlavní diagonály čtverce [0,1] x [0,1], lze soudit, že data pocházejí z exponenciálního rozložení.