Testování exponenciálního a Poissonova rozložení

Test dobré shody

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

Distribuční funkce $\Phi(x)$ je spojitá:
data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+} \rangle$, $j = 1, ..., r$;
zjistíme absolutní četnost n_j j-tého třídicího intervalu;
vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$

Distribuční funkce $\Phi(x)$ má nejvýše spočetně mnoho bodů nespojitosti:
místo třídicích intervalů použijeme varianty $x_{[j]}$, $j = 1, \ldots, r$;
pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j;
vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak $p_j = \Phi(\quad) - \lim_{x \to \quad} \Phi(\quad) = P(X = x_{[\quad]})$.

Testová statistika: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$.

Platí-li nulová hypotéza, pak $K \approx \chi^2(r-p-1)$, kde p je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení p = 2, protože z dat odhadujeme střední hodnotu a rozptyl.)
Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-p-1)$.
Aproximace se považuje za vyhovující, když $np_j \geq 5$, $j = 1, ..., r$.

Upozornění: Hodnota testové statistiky K je silně závislá na volbě třídicích intervalů. Navíc při nesplnění podmínky $np_j \geq 5$, $j = 1, ..., r$ je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace.

Jednoduchý test exponenciálního rozložení (Darlingův test)

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: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle$. Jestliže $K \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

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

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:
a) test dobré shody,
b) Darlingův test exponenciálního rozložení

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

Ad a) Nejprve odhadneme parametr λ exponenciálního rozložení: $\lambda = \frac{1}{m} = \frac{1}{\frac{1}{n}\sum_{j=} \; \cdot} = \frac{1}{\frac{1}{70}(4 \cdot 1{,}5 + 6 \cdot 4{,}5 + \ldots + \cdot 22{,}5)} = ,1122$

Pravděpodobnost, že náhodná veličina s rozložením $Ex(\lambda)$, kde $\lambda = 0{,}1122$ se bude realizovat v intervalu $\langle u_j, u_{j+} \rangle$ je
$p_j = \Phi(u_{j+1}) - \Phi(u_j)$, $j = 1, \ldots, r$, kde $\Phi = -$.
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

$\langle u_j, u_{j+} \rangle$	$x_{[j]}$	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].

$\langle u_j, u_{j+} \rangle$	$x_{[j]}$	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 $K = 1{,}8017 + \ldots + 2{,}7047 = 6{,}6178$, $r = 6$, $p = 1$, $r - p - 1 = 4$, $\chi^2_{0,95}(4) = 9{,}4877$.
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) Darlingův test exponenciálního rozložení je založen 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: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle$.

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$m = \frac{1}{70}(4 \cdot 1{,}5 + 6 \cdot 4{,}5 + \ldots + \cdot 22{,}5) = ,9143$

$s^2 = \frac{1}{69}(9 \cdot (,5 - ,9143)^2 + 6 \cdot (4,5 - ,9143)^2 + \ldots + \cdot (22,5 - ,9143)^2) = 1,1447$

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

Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle = \langle 0, \chi^2_{0{,}025}(69)\rangle \cup \langle \chi^2_{0{,}975}(69), \infty\rangle = \langle 0; 47{,}9242\rangle \cup \langle 93{,}8565, \infty\rangle$.

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 = sum(nj) (n=70) a parametr lambda = n/sum(nj'*xj) (lambda=0,1121).
Vypočteme teoretické četnosti npj=n*diff(expcdf(uj,1/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*diff(expcdf(uj,1/lambda)).
Nyní již jsou splněny podmínky dobré aproximace.
Vypočítáme testovou statistiku K=sum((nj-npj).^2./npj) a kvantil $\chi_{1-\alpha}(k-r-1) = \chi_{0,95}(4)$ 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)

Zadáme vektor mezí uj = [0:3:24]', vektor středů xj =[1.5:3:22.5]' a vektor pozorovaných četností
nj = [14 16 10 9 8 5 3 5]'. Celkový rozsah souboru je n = sum(nj)
Vypočteme průměr m=sum(nj'*xj)/n. Dostaneme 8,9143.
Vypočteme rozptyl skv=sum(nj'*(xj-m).^2)/(n-1). Dostaneme 41,1447.
Vypočteme testovou statistiku K=(n-1)*skv/m^2. Dostaneme 35,7265.
Nakonec stanovíme kvantily $\chi_{\alpha/2}(k-1) = \chi_{0,025}(69)$ a $\chi_{1-\alpha/2}(k-1) = \chi_{0,975}(69)$: chi2inv(0.025,69) a chi2inv(0.975,69).
Dostaneme 47,9242 a 93,8565. Protože testová statistika patří do kritického oboru, na asymptotické hladině významnosti 0,05 zamítáme nulovou hypotézu.

Samostatný úkol: Vypočtěte p-hodnotu pro Darlingův test exponenciálního rozložení.
Pro oboustrannou alternativu se počítá podle vzorce $p = 2\min\{\Phi(K), 1 - \Phi(K)\}$, kde Φ je distribuční funkce rozložení, kterým se řídí testová statistika, když H_0 platí. (p = 0,0006).

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 $X_1, \ldots, X_{360}$ pochází z Po(λ) proti H_1: non H_0.

Ad a) Nejprve odhadneme parametr λ Poissonova rozložení: $\lambda = n = \frac{1}{n}\sum_{j=} x \square = \frac{1}{360}(7 \cdot 0 + 3 \cdot 1 + \ldots + 2 \cdot 7) = 2{,}3$

Pravděpodobnost, že náhodná veličina s rozložením Po(λ), kde $\lambda = 2{,}3$ bude nabývat hodnot 0, 1, ..., 7 a víc je

$p_j = \frac{\lambda}{j!} e^{-} = \frac{2{,}3^j}{j!} e^{-,3}$, $j =),1,...,6$, $p_7 = - (0 +)_1 \ldots +)_6)$.

Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_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_j	p_j	np_j	$(n_j - np_j)^2 / np_j$
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

$K = 2,2909 + 1,2012 + \ldots + 0,7220 = 9,5892$, $r = 7$, $p = 1$, $r - p - 1 = 5$, $\chi^2_{0,95}(5) = 11,0705$.
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) Test je založen 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: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle$.

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$m = \frac{1}{360}(27 \cdot 0 + 93 \cdot 1 + \ldots + 8 \cdot 7) = 2,3$

$s^2 = \frac{1}{359}\left[27 \cdot (0 - 2,3)^2 + 93 \cdot (1 - 2,3)^2 + \ldots + 8 \cdot (7 - 2,3)^2\right] = 2,121448$

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

Kritický obor: $W = \langle 0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty\rangle = \langle 0, \chi^2_{0,025}(359)\rangle \cup \langle \chi^2_{0,975}(359), \infty\rangle = \langle 0; 308,4\rangle \cup \langle 413,4, \infty\rangle$

H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Řešení pomocí MATLABu:
Ad a) Zadáme vektor xj = [0:7]' a vektor pozorovaných četností nj = [27 93 103 58 50 21 6 2]'.
Celkový rozsah souboru je n = sum(nj) a odhad parametru lambda: lambda=sum(nj'*xj)/n.
Vektor pravděpodobností: pj=poisspdf(xj,lambda).
Poslední pravděpodobnost musíme nahradit doplňkem do jedné: pj(8)=1-sum(pj(1:7))
Vypočteme teoretické četnosti npj=n*pj
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 xj = [0:6]' a nový vektor pozorovaných četností nj = [27 93 103 58 50 21 8]'.
Znovu vypočteme vektor pravděpodobností: pj=poisspdf(xj,lambda).
Poslední pravděpodobnost musíme nahradit doplňkem do jedné: pj(7)=1-sum(pj(1:6))
Vypočteme teoretické četnosti npj=n*pj
Nyní již jsou splněny podmínky dobré aproximace.
Vypočítáme testovou statistiku K=sum((nj-npj).^2./npj) a kvantil $\chi_{1-\alpha}(k-1-r) = \chi_{0,95}(5)$ 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) Zadáme vektor xj = [0:7]' a vektor pozorovaných četností nj = [27 93 103 58 50 21 6 2]'.
Vypočteme průměr m=sum(nj'*xj)/n. Dostaneme 2,3.
Vypočteme rozptyl skv=sum(nj'*(xj-m).^2)/(n-1). Dostaneme 2,1215.
Vypočteme testovou statistiku K=(n-1)*skv/m. Dostaneme 331,1353.
Nakonec stanovíme kvantily $\chi_{\alpha/2}(k-1-r) = \chi_{0,025}(359)$ a $\chi_{1-\alpha/2}(k-1-r) = \chi_{0,975}(359)$: chi2inv(0.025,359) a chi2inv(0.975,359).
Dostaneme 308,4 a 413,4. Protože testová statistika nepatří do kritického oboru, nelze na asymptotické hladině významnosti 0,05 zamítnout nulovou hypotézu.

Samostatný úkol: Vypočtěte p-hodnotu a) pro test dobré shody (p = 0,0877), b) pro jednoduchý test Poissonova rozložení (p = 0,2969).

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.