

Testování exponenciálního a Poissonova rozložení

Test dobré shody

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

- a) Je-li distribuční funkce spojitá, pak data rozdělíme do r třídících intervalů (u_j, u_{j+1}) , $j = 1, \dots, r$. Zjistíme absolutní četnost n_j j -tého třídícího intervalu a vypočteme pravděpodobnost p_j , že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j -tém třídícím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.
- b) Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídících intervalů použijeme varianty $x_{[j]}$, $j = 1, \dots, r$. Pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j a vypočteme pravděpodobnost p_j , že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak
- $$p_j = \Phi(x_{[j]}) - \lim_{x \rightarrow x_{[j]}^-} \Phi(x) = P(X = x_{[j]}).$$

Testová statistika: $K = \sum_{j=1}^r \frac{(n_j - np_j)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-p)$, 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 $K \geq \chi^2_{1-\alpha}(r-1-p)$. Aproximace se považuje za vyhovující, když $np_j \geq 5$, $j = 1, \dots, r$.

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

Jednoduchý test exponenciálního rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr X_1, \dots, X_n pochází z exponenciálního rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim \text{Ex}(\lambda)$ 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, \dots, X_n pochází z Poissonova rozložení. Označme M výběrový průměr a S^2 výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny $X \sim \text{Po}(\lambda)$ 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.: Byla zkoumána doba životnosti 45 součástek (v hodinách). Výsledky jsou uvedeny v tabulce rozložení četností:

Doba životnosti	Počet součástek
(0, 50]	15
(50, 100]	14
(100, 150]	6
(150, 200]	5
(200, 250]	2
(250, 300]	1
(300, 350]	1
(350, 400]	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení. Použijte a) test dobré shody, b) jednoduchý test exponenciálního rozložení (využijte toho, že z původních dat byl vypočten průměr $m = 99,93$ a rozptyl $s^2 = 7328,91$).

Řešení:

ad a)

Zadáme vektor $x_j = [0:50:400]'$ a vektor pozorovaných četností $n_j = [15\ 14\ 6\ 5\ 2\ 1\ 1\ 1]'$.

Celkový rozsah souboru je $n = \sum(n_j)$ a parametr $\lambda = 99,93$.

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

Protože nejsou splněny podmínky dobré aproximace pro $j = 4, 5, 6, 7, 8$, je třeba sloučit třídící intervaly 4 až 8 do jednoho. Dostaneme novou tabulku rozložení četností

Doba životnosti	Počet součástek
(0, 50]	15
(50, 100]	14
(100, 150]	6
(150, 400]	10

Zadáme nový vektor $x_j = [0\ 50\ 100\ 150\ 400]'$ a nový vektor pozorovaných četností $n_j = [15\ 14\ 6\ 10]'$. Znovu vypočteme teoretické četnosti $np_j = n \cdot \text{diff}(\text{expcdf}(x_j, \lambda))$.

Nyní již jsou splněny podmínky dobré aproximace. Vypočítáme testovou statistiku

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

Protože testová statistika $K = 1,5153$ se nerealizuje v kritickém oboru $W = \langle 5,9915, \infty \rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

ad b)

Testovou statistiku K vypočteme podle vzorce $K = \frac{(n-1)S^2}{M^2}$. Kritický obor má tvar:

$W = \langle 0; \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1); \infty \rangle$. V našem případě $K = 32,2924$,

$W = \langle 0; 27,575 \rangle \cup \langle 64,202; \infty \rangle$, H_0 tedy nezamítáme na asymptotické hladině významnosti 0,05.

Samostatný úkol: Vypočtete p-hodnotu pro jednoduchý 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. ($p = 0,1912$).

Příklad 2.: Studujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů:

Počet pacientů	Pozorovaná četnost
0	79
1	188
2	282
3	275
4	196
5	114
6	45
7	10
8	7
9	3
10	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z Poissonova rozložení. Použijte a) test dobré shody, b) jednoduchý test Poissonova rozložení.
Řešení:

ad a)

Postupujeme podobně jako v příkladu 1, ale místo funkce expcdf použijeme funkci poispdf, abychom vypočítali hodnoty pravděpodobnostní funkce Poissonova rozložení v bodech 0 až 10. Odhad parametru lambda získáme jako vážený průměr počtu pacientů ($m = 2,7992$).

Protože nejsou splněny podmínky dobré aproximace, je třeba sloučit poslední tři varianty do jedné.

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

Upozornění: $\pi(8) = 1 - \sum_{x=0}^7 \pi(x)$

$K = 8,502$, $W = \langle 14,067, \infty \rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

ad b)

$K = \frac{(n-1)S^2}{M}$, kde $m = 2,7992$, $s^2 = 2,6594$

$K = 1139,1$, $W = \langle 0; 1104,93 \rangle \cup \langle 1296,86; \infty \rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Samostatný úkol: Vypočtete p-hodnotu pro jednoduchý test Poissonova rozložení.

($p = 0,2187$)

Další možnosti ověřování exponenciálního rozložení – využití funkce probplot (pravděpodobnostně – pravděpodobnostní graf), Kolmogorovův – Smirnovův test (funkce kstest).