ODHADY PARAMETRU TUKEYOVA A ZOBECNĚNÉHO TUKEYOVA SYSTÉMU ROZDĚLENÍ

Marie Hušková, MFF UK

0. ÚVOD

Tukey (1960) zavedl systém jednoparametických rozdělení s kvantilovou funkcí

(1) $$F^{-1}(u;\lambda) = (u^{\lambda} - (1-u)^{\lambda})/\lambda \qquad u\in(0,1), \quad \lambda\in R_1.$$

V literatuře se pro něj ujal název Tukeyův systém rozdělení.

Ramberg a Schmeiser (1974) zavedli systém rozdělení závisející na čtyřech parametrech $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ s kvantilovou funkcí

(2) $$F^{-1}(u;\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \lambda_1 + (u^{\lambda_2} - (1-u)^{\lambda_2})\lambda_4 \qquad u\in(0,1),$$

kde $\lambda_1\in R_1$ je parametr polohy, $|\lambda_4|$ je parametr měřítka a λ_2, λ_3 jsou parametry charakterizující chování chvostů, $(\lambda_2,\lambda_3,\lambda_4)\in\Lambda = \{\lambda_i > 0,\ i=2,3,4\}\cup\{\lambda_i<0,\ i=2,3,4\}\cup\{\lambda_2<-1,\ \lambda_3>0,\ \lambda_4<0\}\cup\{\lambda_2>0,\ \lambda_3<-1,\ \lambda_4<0\}\cup\{\lambda_2=0,\ \lambda_3\lambda_4>0\}\cup\{\lambda_3=0,\ \lambda_2\lambda_4>0\}$.

Systém rozdělení tvořený rozděleními s kvantilovou funkcí (2) s $\lambda_1\in R_1$ a $(\lambda_2,\lambda_3,\lambda_4)\in\Lambda$ se nazývají zobecněný Tukeyův systém rozdělení.

Tyto systémy zahrnují řadu běžně používaných rozdělení popř. jejich aproximace. V práci bude pojednáno o vlastnostech těchto systémů a odhadech jejich parametrů.

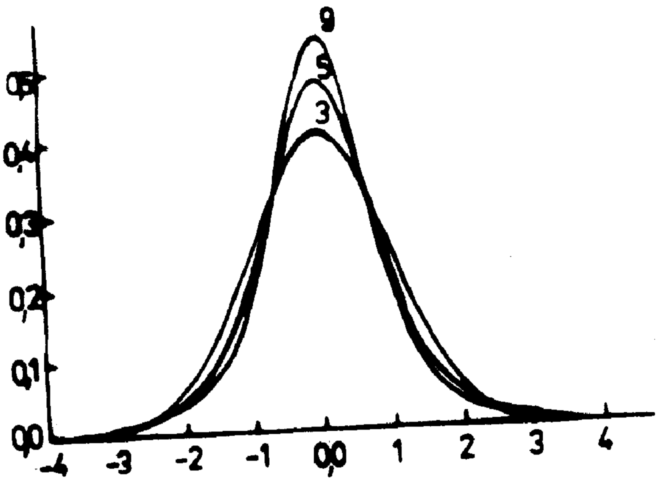
1. NĚKTERÉ VLASTNOSTI SYSTÉMU ROZDĚLENÍ

Tukeyův systém rozdělení zahrnuje symetrické rozdělení s tzv. těžkými chvosty $(\lambda<0)$, i lenkými chvosty $(\lambda>0)$. Při $\lambda<1$ nebo $\lambda>2$ jde o rozdělení unimodální jednovrcholové a při $1<\lambda<2$ má rozdělení tzv. $\hat{U}$ - tvar, Cauchyovo rozdělení, pro $\lambda=0$ logistické a při $\lambda=1$ nebo 2 rozdělení rovnoměrné. Při $\lambda=0{,}08$ obdržíme rozumnou aproximaci pro dvojitě exponenciální rozdělení, při $\lambda=0{,}14$ pro normální. Tento systém poskytuje také dobrou aproximaci pro t-rozdělení.

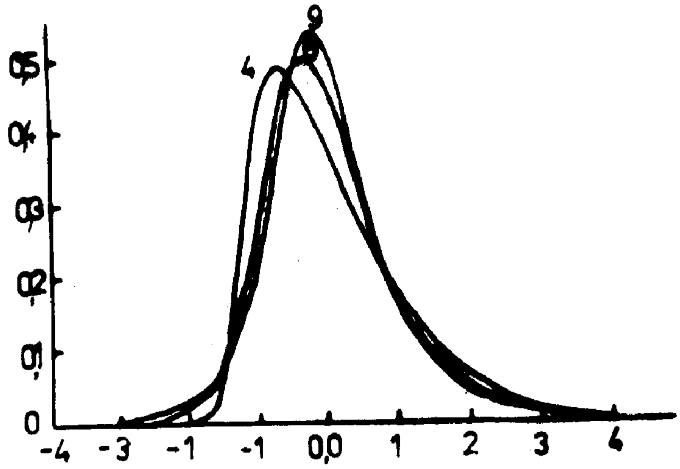
Zobecněný Tukeyův systém rozdělení zahrnuje některé asymetrické rozdělení. Poskytuje dobrou aproximaci např. pro exponenciální rozdělení.

Na obrázcích 1,2,3,4 jsou uvedeny grafy hustot několika rozdělení náležejících do těchto systémů. ($\alpha_i = E(X-EX)^i(E(X-EX)^2)^{-1/2}$ $\quad i=3,4$).

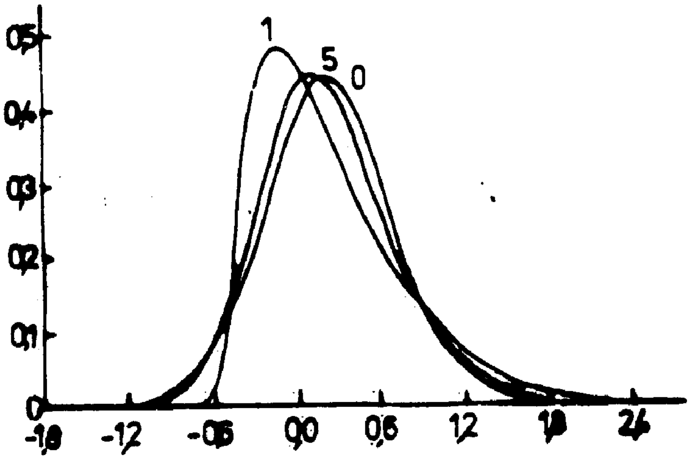
Uvědomíme-li si fakt, že má-li náhodná veličina U rovnoměrné rozdělení na $(0,1)$ a F je nějaká distribuční funkce, pak $F^{-1}(U)$ má rozdělení s distribuční funkcí F, lze uvedené systémy rozdělení výhodně využít při různých simulačních studiích. Navíc se oba systémy ukázaly velmi užitečné při reprezentaci dat, jestliže model není znám.



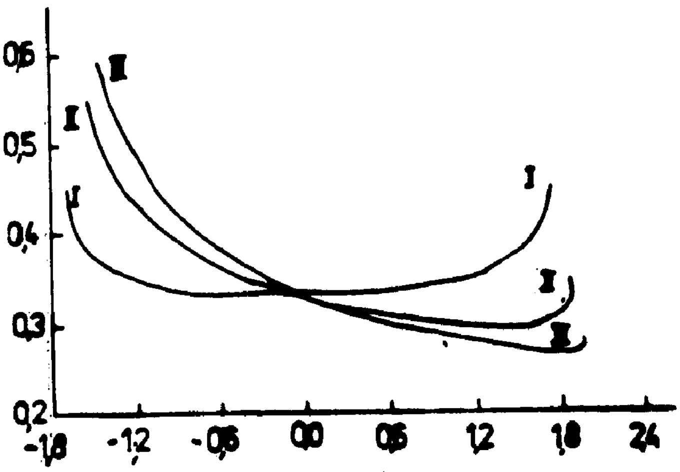
Obr. č.1 Grafy hustot pro
$\alpha_3 = 0$, $\alpha_4 = 3,5,9$



Obr. č.2 Grafy hustot pro
$\alpha_3 = 1$, $\alpha_4 = 4,6,9$



Obr. č.3 Grafy hustot pro
$\alpha_3 = 0,\ 0.5, 1;$ $\alpha_4 = 4$



Obr. č.4 Grafy hustot
s tzv. U - tvarem

Obecně neexistuje explicitní vyjádření distribuční funkce nebo hustoty. Momenty lze vyjádřit následovně (pro systém (2)):

(3) $$EX = \lambda_1 + \lambda_4 \left((\lambda_2 + 1)^{-1} - (\lambda_3 + 1)^{-1} \right) \quad \text{pro} \quad \lambda_2 > -1, \ \lambda_3 > -1,$$

(4) $$E(X - \lambda_1)^k = \lambda_4^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j B(\lambda_2(k - j) + 1, \ \lambda_3 j + 1)$$

$$\text{pro} \quad \lambda_2 k > -1, \quad \lambda_3 k > -1,$$

kde $B(\alpha,\beta)$ je beta funkce.

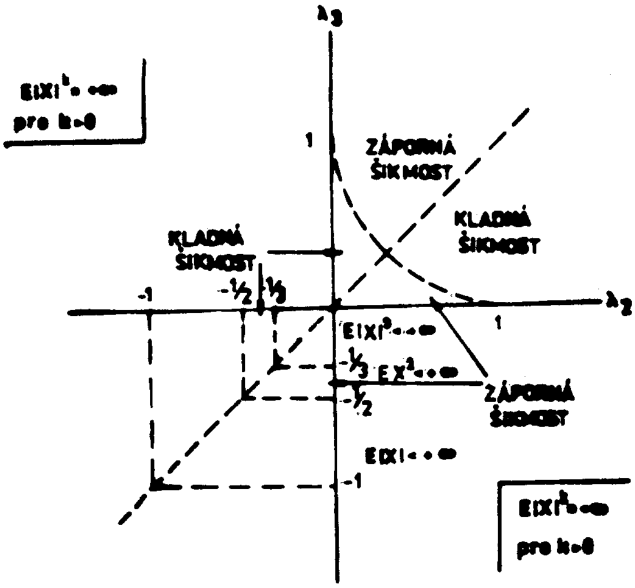
Pro momenty náhodné veličiny X s kvantilovou funkcí (1) platí

(6) $$EX = 0$$

(7) $$EX^k = \lambda^{-k} \sum_{j=0}^{k} \binom{k}{j} (-1)^j B(\lambda(k - j) + 1, \ \lambda j + 1) \quad \text{pro} \quad \lambda k > -1,$$

V článku Ramberg a kol. (1979) jsou tabelovány hodnoty $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ pro $EX = 0$, $EX^2 = 1$, $\alpha_3 = 0(0.05)2$, $\alpha_4 = 1.8(0.2)9$, $\alpha_3 = E(X-EX)^3$, $\alpha_4 = E(X-EX)^4$.

Tabulka č.5 obsahuje přehled o konečnosti momentů a kladné či záporné šikmosti ($E(X-EX)^3 > 0$ resp. < 0).



Obr. 5 Některé vlastnosti rozdělení v závislosti na λ_2, λ_3.

Poznamenejme, že Fisherova míra informace vzhledem k parametru posunutí je konečná pro $\lambda_2, \lambda_3, \lambda \notin \langle 1/2, 2 \rangle$.

2. ODHADY PARAMETRŮ $\lambda, \lambda_1, \lambda_2, \lambda_3, \lambda_4$

Pozornost soustředíme hlavně na systém (2), neboť (1) je speciálním případem (2).

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení s kvantilovou funkcí (2) a $X_{(1)}, \ldots X_{(n)}$, je příslušný uspořádaný výběr. V dalším bude $X_{(a)}$ značit [a]-tou pořádkovou statistiku, kde [a] je celá část a.

Za předpokladu konečnosti 4. momentu můžeme odhadnout parametry $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ metodou momentů, t.j. za odhady vezmeme řešení rovnic:

$$EX^k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \qquad k = 1,2,3,4. \tag{8}$$

Explicitní řešení obecně neexistuje. Můžeme však využít tabulky Ramberga a kol. (79) zmíněné výše. Podobně za odhad parametru λ můžeme vzít řešení rovnice (dle 7):

$$EX^3 = \frac{1}{n} \sum_{i=1}^{n} X_i^3 , \tag{9}$$

při jejichž řešení můžeme využít tabulky Ramberga a kol. (79).

Můžeme však též použít velice jednoduché odhady založené na pořádkových statistikách. Je-li $\lambda_2 < 1$, pak můžeme pro λ_2 použít odhad

$$\hat{\lambda}_2(M,a,b,s) = \frac{1}{\log s^{-1}} \log \frac{X_{(aM)} - X_{(bM)}}{X_{(asM)} - X_{(bsM)}} , \tag{10}$$

kde $M,a,b,s > 0$, $s \neq 1$, $a \neq b$, $\max(a,b,as,bs) < n/2)$. Podobně, je-li $\lambda_3 < 1$, lze λ_3 odhadnout

$$(11) \qquad \hat{\lambda}_3(M,a,b,s) = \frac{1}{\log s^{-1}} \log \frac{X_{(n-aM)} - X_{(n-bM)}}{X_{(n-asM)} - X_{(n-bsM)}} .$$

Oba tyto odhady jsou konsistentní, pro $n \to \infty$, $M \to \infty$, $M/n \to 0$ (a,b,s jsou pevné) platí

$$(12) \qquad \hat{\lambda}_i(M,a,b,s) = \lambda_i + O_p(\max(M^{-1/2}, (M/n)^{1-\lambda_i})) \qquad \text{pro } \lambda_i < 1$$

$$= 1 + O_p(1) \qquad \text{pro } \lambda_i \geq 1,$$

$i = 2,3$. Pro praktické účely je vhodná např. volba $s = 1/2$, $a = 4$, $b = 2$, dostaneme pak

$$\hat{\lambda}_2(M,4,2,1/2) = \frac{1}{\log 2} \log \frac{X_{(4M)} - X_{(2M)}}{X_{(2M)} - X_{(M)}}$$

a M volíme podle rozsahu výběru $0{,}02n - 0{,}05n$.

Další typ odhadů parametrů (λ_2, λ_3) získáme řešením následujících transcendentních rovnic:

$$(13) \qquad \frac{X_{(a_i n)} - X_{(b_i n)}}{X_{(c_i n)} - X_{(d_i n)}} = \frac{a_i^{\lambda_2} - (1-a_i)^{\lambda_3} - b_i^{\lambda_2} + (1-b_i)^{\lambda_3}}{c_i^{\lambda_2} - (1-c_i)^{\lambda_3} - d_i^{\lambda_2} + (1-d_i)^{\lambda_3}} \qquad i = 1,2$$

kde $1 > a_i, b_i, c_i, d_i > 0$, $(a_i,b_i) \neq (c_i,d_i), (d_i,c_i)$, $i = 1,2$, $(a_1,b_1,c_1,d_1) \neq (a_2,b_2,c_2,d_2), (c_2,d_2,a_2,b_2)$, $a_i \neq b_i$, $c_i \neq d_i$.

Získané odhady (ozn. $(\tilde{\lambda}_2, \tilde{\lambda}_3)$) jsou konsistentní; pro $n \to \infty$ a a_i, b_i, c_i, d_i, $i = 1,2$, pevné platí

$$(14) \qquad \tilde{\lambda}_i = \lambda_i + O_p(n^{-1/2}) \qquad i = 2,3.$$

Parametr měřítka λ_4 můžeme odhadnout buď

$$(15) \qquad \hat{\lambda}_4(M,N) = (X_{(n-M)} - X_{(N)})/2, \quad \text{je-li } \lambda_2 > 0, \quad \lambda_3 > 0,$$

nebo

$$(16) \qquad \tilde{\lambda}_4(a,b) = \frac{X_{(an)} - X_{(bn)}}{a^{\tilde{\lambda}_2} - (1-a)^{\tilde{\lambda}_3} - b^{\tilde{\lambda}_2} + (1-b)^{\tilde{\lambda}_3}} ,$$

kde $M,N < n/2$, $a \neq b \in (0,1)$, $\tilde{\lambda}_2$, $\tilde{\lambda}_3$ jsou odhady získané řešením rovnic (13). Oba odhady jsou za jistých předpokladů konsistentní.
Pro $n \to \infty$ a,b pevné platí

$$(17) \qquad \tilde{\lambda}_4(a,b) = \lambda_4 + O_p(n^{-1/2}).$$

Jestliže navíc $M/n \to 0$, $N/n \to 0$ M,N mohou být pevné nebo konvergovat k nekonečnu, $\lambda_i > 0$ $i = 2,3$, pak lze dokázat platnost

(18) $$\hat{\lambda}_4(M,N) = \lambda_4 + O_p(\max(\frac{M}{n},(\frac{M}{n})^{\lambda_2}, \frac{N}{n}, (\frac{N}{n})^{\lambda_3})).$$

Parametr λ_1 můžeme odhadnout buď

(19) $$\hat{\lambda}_1(M,N) = (X_{(n-M)} + X_{(N)})/2 , \qquad \text{je-li} \quad \lambda_2 > 0, \ \lambda_3 > 0,$$

nebo

(20) $$\tilde{\lambda}_1(a,b) = (X_{(an)} + X_{(bn)})/2 - \tilde{\lambda}_4(a^{\tilde{\lambda}_2} - (1-a)^{\tilde{\lambda}_3} + b^{\tilde{\lambda}_2} - (1-b)^{\tilde{\lambda}_3})/2$$

kde $a \neq b \in (0,1)$ $M,N < n/2$. Odhady mají vlastnosti analogické $\hat{\lambda}_4(M,N)$ respective, $\tilde{\lambda}_4(a,b)$.

Podívejme se nyní na odhad parametru λ Tukeyova systému rozdělení. Nabízejí se dva typy (ve shodě s předchozím) jednak

(21) $$\hat{\lambda}(M,a,b,s) = (\hat{\lambda}_2(M,a,b,s) + \hat{\lambda}_3(M,a,b,s))/2, \qquad \text{pro} \quad \lambda < 1$$

kde $\hat{\lambda}_i(M,a,b,s)$ $i=2,3$ jsou definovány (10) respective, (11) jednak řešení rovnice

$$\frac{X_{(an)} - X_{((1-a)n)}}{X_{(cn)} - X_{((1-c)x)}} = \frac{a^{\lambda} - (1-a)^{\lambda}}{c^{\lambda} - (1-a)^{\lambda}},$$

kde $a \neq c \in (0,1)$.

Odhady parametrů λ_2, λ_3 popř. λ můžeme velmi výhodně využít např. při konstrukci asymptoticky optimálních odhadů v následujících dvou typických úlohách neparametrických metod.

a) Nechť $X_1,\dots,X_n$ jsou nezávislé náhodné veličiny, X_i má distribuční funkci $F(x-\theta c_i)$, kde $(c_1,\dots,c_n)$ jsou známé regresní konstanty, θ je neznámý parametr, který chceme odhadnout. Asymptoticky optimální R-odhad je generován skórovou funkcí

$$\varphi_f(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \qquad u \in (0,1)$$

kde f a f' jsou hustota a její derivace příslušné F, F^{-1} je kvantilová funkce. Jestliže distribuční funkce F náleží do zobecněného Tukeyova systému pak platí (pro λ_2, λ_3 $\notin$ $\langle 1/2, 2\rangle$):

$$\varphi_f(u) = -\frac{\lambda_2(\lambda_2-1)\,u^{\lambda_2-2} - \lambda_3(\lambda_3-1)(1-u)^{\lambda_3-2}}{\lambda_4(\lambda_2 u^{\lambda_2-1} + \lambda_3(1-u)^{\lambda_3-1})^2}, \qquad u \in (0,1).$$

Tedy optimální skórovou funkcí známe až na parametry λ_2, λ_3, λ_4, nahradíme-li je odhady navrženými výše, pak R-odhad generovaný takto vzniklou skórovou funkcí je asymptoticky optimální (ovšem za předpokladu, že F náleží do zobecněného Tukeyova systému).

b) Nechť $(X_1,\dots,X_n)$ je náhodný výběr z rozdělení s distribuční $F(X-\theta)$, kde θ je neznámý parametr, který chceme odhadnout a F je symetrická distribuční funkce. Asymptoticky optimální L-odhad je generován funkcí

$$J_f(u) = \varphi'_f(u)\ f(F^{-1}(u))\ (\int_0^1 \varphi'_f(y)\ f(F^{-1}(y))\ dy)^{-1}, \qquad u \in (0,1).$$

Jestliže F náleží do Tukeyova systému, známe J_f až na parametr λ . Nahradíme-li λ některým z odhadů navržených výše, pak L-odhad, generovaný takto vzniklou funkcí J_f, je asymptoticky optimální (za předpokladu, že F náleží do Tukeyova systému). Další podrobnosti může čtenář nalézt v citované literatuře.

LITERATURA

Chan, L. K. and Rhodin, L. S. (1980): Robust estimation of location using optimally chosen sample quantiles. Technometrics 22, 225-237.

Filliben, J. J. (1969). Simple and robust estimation of the location parameter of a symmetric distribution. Ph. D. dissertation, Princeton University, Princeton, N. Y.

Hušková, M. (1986). Simple estimators of the parameters of generalized Tukey's family, vyjde CMUC.

Hušková, M. (1986). Adaptive estimators of parameters generalized Tukey's family, zasláno k otištění.

Joiner, B. L. and Rosenblatt, J. R. (1971). Some properties of the range in samples from Tukey's symmetric λ -distributions. J. Amer. Statist. Assoc, 66, 394-399.

Jones, D. H. (1979). An efficient adaptive distribution-free test for location, J. Amer. Statist, Assoc, 74, 822-828.

Ramberg, J. S. and Schmeiser, B. W. (1972). An approximate method for generating symmetric random variables, Comm. of the ACM, 15, 987-990.

Ramberg, J. S. and Schmeiser, B. W. (1974). An approximative method for generating asymmetric random variables. Comm. of the ACM, 17, 78-82.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J. and Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. Technometrics 21, 201-214.

Tukey, J. W. (1960). The practical relationship between the common transformations of percentages or counts and of amounts. Technical Report 36, Statistical Research Group, Princeton University, Princeton, N. Y.