VYUŽITÍ SYSTÉMU g - h ROZDĚLENÍ PŘI ANALÝZE DAT

Jiří Militký, VÚTZ, Dvůr Králové n.L.

1. ÚVOD

Při průzkumové analýze dat je účelem kvantifikovat jejich statistické zvláštnosti, které se týkají zejména tvaru výběrového rozdělení. Běžně se pro tyto účely používá kvantilových charakteristik šikmosti a špičatosti /1/. Další možností je použití vhodného systému empirických frekvenčních funkcí, pomocí kterého se aproximuje výběrové rozdělení. Oba tyto přístupy je možné spojit s využitím takových systémů frekvenčních funkcí, které vycházejí z kvantilové funkce (jako je např. Tukeyův systém lambda rozdělení). Vzhledem k tomu, že při analýze dat je referenční rozdělení normální (a odchylky od normality jsou obyčejně připisovány transformaci původně normálně rozdělených dat), je vhodné, aby empirický systém rozdělení byl generován jako transformace normálně rozdělených náhodných veličin (jako je Johnsonův systém).

Syntézou těchto požadavků je tzv. systém g-h rozdělení, jehož parametry jsou přímo kvantilové míry šikmosti a špičatosti. Lze ho tedy konstruovat velmi jednoduše přímo z experimentálních dat. Navíc je "automaticky" zajištěna robustnost resp. lokální charakterizovatelnost zvláštností rozdělení dat.

Systém g-h rozdělení lze kromě průzkumové analýzy dat uplatnit také pro generaci náhodných veličin z rozdělení s definovanými šikmostmi a špičatostmi (které mohou být závislé na vzdálenosti od mediánu). Tyto náhodné veličiny lze s výhodou použít při různých simulačních studiích resp. ověřování robustních metod.

V tomto příspěvku jsou přehledně uvedeny základní vlastnosti g-h rozdělení, které se dostalo i do statistické encyklopedie /2/.
Je popsána struktura části programu EXGR, která využívá g-h rozdělení pro účely průzkumové analýzy.

2. DEFINICE g-h ROZDĚLENÍ

Systém g-h empirických rozdělení je založen na monotónní transformaci standardizované náhodné veličiny s normálním rozdělením. Jde o jeden z translačních systémů rozdělení, kdy se nejprve provádí standardizace původních náhodných proměnných (kvantilů) x na normalizované kvantily

$$z = (x - \tilde{x}_{0,5}) / R \qquad (1)$$

kde $\tilde{x}_{0,5}$ je parametr (medián) a R je parametr měřítka.
Pro třídu g-rozdělení (pouze sešikmených) se volí obecná transformace typu

$$Q_{g,0}(z) = G(z) \cdot z \qquad (2)$$

kde $G(z)$ je lichá funkce, pro kterou musí platit, že

$$Q_{g,0}(0) = 0 \quad \text{a} \quad \lim_{z \to 0} Q_{g,0}(z) \approx z$$

$G(z)$ je vlastně operátor šikmosti, který závisí na parametru šikmosti g. Je vhodné, aby pro g > 0 vycházela rozdělení sešikmená vpravo, pro g < 0 rozdělení sešikmená vlevo a pro g = 0 normální rozdělení. Všem těmto požadavkům vyhovuje jednoduchá funkce

$$G(z) = [\exp(g \cdot z) - 1] / (g \cdot z) \qquad (3)$$

Lze jednoduše nalézt, že frekvenční funkce g-rozdělení $f_g(x)$ je dána vztahem

$$f_g(x) = 1/[\sqrt{2\pi} \cdot |(x - \tilde{x}_{0,5}) \cdot g + R|] \cdot \exp(-\{\ln[(x - \tilde{x}_{0,5}) \cdot g/R + 1]\}^2 / (2g^2))$$

Jde tedy o lognormální rozdělení (rozdělení typu S_L v Johnsonově systému frekvenčních funkcí).
Pro třídu h-rozdělení (symetrických s různou špičatostí) se volí transformace typu

$$Q_{0,h}(z) = H(z) \cdot z \qquad (4)$$

Operátor špičatosti $H(z)$ musí být rostoucí kladná sudá funkce závislá na parametru špičatosti h. Pro h = 0 jde o normální rozdělení a čím je h > 0 větší, tím má odpovídající rozdělení delší konce. Tomu vyhovuje volba

$$H(z) = \exp(hz^2/2) \tag{5}$$

Z rov. (5) plyne, že parametr h může být i záporný. Podmínka monotónnosti funkce $Q_{g,0}$ je však narušena, pokud je $z^2 > -1/h$. Třída h-rozdělení definované rov. (4) a (5) se v oblasti konců chová jako Paretovo rozdělení. Volba faktoru 1/2 v rov. (5) také zajišťuje, že pro h ≈ 1 je $Q_{0,h}$ rozdělení blízké Cauchyho rozdělení /3/.

Frekvenční funkci h rozdělení lze však získat pouze numericky, protože nelze nalézt analyticky transformaci inverzní k rov. (4).

Pro obecnou třídu g-h rozdělení se volí dvouparametrická transformace

$$Q_{g,h}(z) = G(z) \cdot H(z) \cdot z \tag{6}$$

kde $G(z)$ je voleno podle rov.(3) a $H(z)$ je definováno rov. (5). Také v tomto případě je třeba odpovídající frekvenční funkci počítat numericky.

Pro generaci náhodných čísel z g-h rozdělení se zvolenými parametry g a h postačuje generovat kvantily z_p standardního normální rozdělení (zde p určuje 100 P %ní kvantil) a dosazovat do rov. (6) za $z = z_p$.

3. MOMENTY

I když se ani při generaci náhodných proměnných ani průzkumové analýze dat momenty nepoužívají, je pro účely porovnání s ostatními systémy empirických rozdělení vhodné znát výrazy pro první čtyři momenty. Ty zde budou funkcemi parametrů g, h. Pro standardizované náhodné proměnné $Y = Q_{g,h}(z)$ z g-h rozdělení je střední hodnota definována vztahem

$$E(Y) = \frac{1}{g\sqrt{h-1}}\left[\exp\left(\frac{g^2}{2(1-h)}\right) - 1\right] \qquad 0 \le h \le 1 \tag{7}$$

a pro rozptyl platí

$$D(Y) = \frac{1}{g^2\sqrt{1-2h}}\left[\exp(2w) - 2\exp\left(\frac{w}{2}\right) + 1\right] - \frac{1}{g^2(1-h)}\left[\exp\left(\frac{w}{2}\right) - 1\right]^2 \tag{8}$$

$$0 \le h < 0{,}5$$

kde $w = g^2/(1-2h)$.

Pro případ g-rozdělení je h = 0 a pak z rov. (7)

$$E(Y) = [\exp(g^2/2) - 1] / g \tag{7a}$$

resp. z rov. (8)

$$D(Y) = \exp(g^2) \cdot [\exp(g^2) - 1] / g^2 \tag{8a}$$

Šikmost g_1 je u g-rozdělení vyjádřitelná ve tvaru

$$g_1 = [\exp(3g^2) - 3\exp(g^2) + 2] / \sqrt{(\exp(g^2) - 1)^3} \tag{9}$$

a pro špičatost g_2 platí

$$g_2 = [\exp(6g^2) - 4\exp(3g^2) + 6\exp(g^2) - 3] / (\exp(g^2) - 1)^2 \tag{10}$$

Pro případ h-rozdělení je g = 0, takže $E(Y) = g_1 = 0$. Pro rozptyl zde platí

$$D(Y) = 1 / \sqrt{(1-2h)^3} \qquad 0 \le h < 0{,}5 \tag{8b}$$

a špičatost je ve tvaru

$$g_2 = 3(1-2h)^3 / \sqrt{(1-4h)^6} \tag{10a}$$

Při výpočtu vyšších momentů g-h rozdělení je možné použít Martinezova vztahu

$$E(Y^n)=\frac{1}{g^n\sqrt{1-nh}}\sum_{i=0}^{n}(-1)^i\binom{n}{i}\exp\left\{\frac{[(n-i)g]^2}{2(1-nh)}\right\} \tag{11}$$

který platí pro $g \neq 0$ a $0 \leq h \leq 1/n$ /3/.

4. ODHADY PARAMETRŮ g-h ROZDĚLENÍ

S ohledem na návaznost na průzkumovou analýzu dat se pro odhad parametrů g, h resp. obou používá výběrového mediánu $\tilde{x}_{0,5}$ a dvojic kvantilů $\tilde{x}_p, \tilde{x}_{1-p}$ pro vhodná $0 < p < 0{,}5$. U malých výběrů se pro tyto účely využívá všech pořádkových statistik $x_{(i)}$. U větších výběrů se používá tzv. písmenových hodnot, kde $p = 2^{-i}$ $i = 2,3$ (odpovídají kvartilům, oktilům, sedecilům atd.)

g - Rozdělení

Pro zvolené p lze určit parametr g (a také R) dvou typů

$$\tilde{x}_p = \tilde{x}_{0,5} + R\cdot Y = \tilde{x}_{0,5} + R\left[\frac{\exp(g\cdot z_p)-1}{g}\right] \tag{12a}$$

$$\tilde{x}_{1-p} = \tilde{x}_{0,5} + R\left[\frac{\exp(-g z_p)-1}{g}\right] \tag{12b}$$

kde z_p jsou standardizované kvantily normálního rozdělení (vzhled k symetrii je $z_p = -z_{1-p}$).
Po jednoduchých úpravách vyjde

$$g_p = -\frac{1}{z_p}\cdot \ln\left[\frac{\tilde{x}_{1-p}-\tilde{x}_{0,5}}{\tilde{x}_{0,5}-\tilde{x}_p}\right] \tag{13}$$

Index p zde ukazuje, že parametr šikmosti obecně může záviset na poloze kvantilů vzhledem k mediánu.

Pokud je g_p přibližně konstantní (resp. není funkcí p) postupuje se tak, že se určí medián $\tilde{g}$ ze všech g_p. Jednoduše je pak možné odhadnout parametru R jako směrnici v Q-Q grafu, kde se vynášejí hodnoty $\tilde{x}_p$ vs. $Q_{g,0}(z_p)$. Přímka v tomto grafu indikuje, že lze použít g-rozdělení s konstantní hodnotou g. Pro stejný účel je možné použít také graf symetrie, kdy se vynášejí polosumy $0{,}5(\tilde{x}_p + \tilde{x}_{1-p})$ vs. $z_p^2/2$. S využitím rov. (12a) a (12b) snadno určíme teoretickou závislosti

$$0{,}5(\tilde{x}_p+\tilde{x}_{1-p}) = \tilde{x}_{0,5} + 0{,}5\frac{R}{g}\left[\exp(g z_p)+\exp(-g z_p)-2\right] \approx$$
$$\approx \tilde{x}_{0,5} + R\cdot g\cdot z_p^2/2 \tag{14}$$

Platí-li tedy předpoklad konstantnosti parametru sešikmení g, vyjde v grafu symetrie přibližně lineární závislost.

Pokud není g_p konstantní, předpokládá se obyčejně, že je to jednoduchý polynom vzhledem k normalizovaným kvantilům z_p resp. jejich čtvercům. Jednoduché je zobecnění typu /3/

$$g_p = g_0 + g_1 z_p^2 \tag{15}$$

Rov. (15) lze snadno ověřit na základě grafu šikmosti, kdy se vynáší g_p vs. z_p^2. Pokud vyjde v tomto grafu přibližně lineární závislost, je třeba uvažovat obecnější rozdělení s nekonstantním g_p vyjádřeným rov. (15). Parametry g_0, g_1 v rov. (15) lze snadno určit z úseku a směrnice v grafu šikmosti (doporučuje se použití robustní regrese).

h - Rozdělení

Při znalosti parametru polohy $\tilde{x}_{0,5}$ a parametru měřítka R lze určit pro každé $\tilde{x}_p$ odpovídající hodnotu h_p přímo z definičního vztahu (4). Vyjde

$$h_p = \left[2\ln\left(\frac{(\tilde{x}_p-\tilde{x}_{0,5})/R}{z_p}\right)\right] / z_p^2 \tag{16}$$

v případě, že je výběrové rozdělení symetrické, musí pochopitelně být $h_p = h_{1-p}$. Jako vhodný parametr měřítka se doporučuje interkvartilový odhad směrodatné odchylky mediánu /3/

$$R = 0{,}926\ .(\tilde{x}_{0,75} - \tilde{x}_{0,25})\ /\ \sqrt{n} \tag{17}$$

kde n je rozsah výběru.
Za předpokladu symetrického rozdělení lze přímo z definičních vztahů

$$x_p = \tilde{x}_{0,5} + R \cdot z_p \cdot \exp(h \cdot z_p^2/2) \tag{18a}$$

a

$$x_{1-p} = \tilde{x}_{0,5} - R \cdot z_p \cdot \exp(h \cdot z_p^2/2) \tag{18b}$$

(zde $z_p < 0$ a $z_{1-p} = -z_p$) dospět k lineární funkci vzhledem k h

$$\ln\left[\frac{\tilde{x}_{1-p} - \tilde{x}_p}{-2z_p}\right] = \ln R + h\, z_p^2/2 \tag{19}$$

Z rov. (19) je patrné, že vynesením $\ln[(\tilde{x}_{1-p} - \tilde{x}_p)/(-2z_p)]$ vs. $z_p^2/2$ vyjde v případě platnosti h-rozdělení s konstantním parametrem špičatosti h přibližně lineární závislost. Tato závislost se označuje jako graf pseudo-sigma a umožňuje odhad h i R ze směrnice resp. úseku regresní přímky.

Pro normální rozdělení je graf pseudosigma prakticky horizontální přímka s nulovou směrnicí a úsekem $\ln R$. Pro nesymetrická rozdělení vycházejí grafy pseudosigma nelineární.
Pokud není h_p konstantní (ale $h_p \neq h_{1-p}$) předpokládá se, že jde opět o jednoduchou závislost typu rov. (15). Tedy

$$h_p = h_0 + h_1 z_p^2 \tag{20}$$

Rov. (20) lze jednoduše ověřit na základě grafu špičatosti, kdy se vynáší h_p v závislosti na z_p^2. Pokud vyjde v tomto grafu přibližně lineární závislost, znamená to, že platí obecnější rozdělení s nekonstantním parametrem špičatosti vyjádřeným rov. (20).

g-h rozdělení

To je nejčastější případ, kdy je rozdělení dat nesymetrické a ani po symetrizační transformaci neodpovídá délkou konců normálnímu rozdělení. Vzhledem k volbě $Q_{g,h}(z)$ ve tvaru rov. (6) s funkcemi $G(z), H(z)$, lze snadno určit, že rov. (13) platí nezávisle na velikosti h. To znamená, že lze nejprve nalézt odhad parametru špičatosti g (stejně jako u "čistého" g-rozdělení) a pak provést symetrizační transformaci dat před odhadem h.

Při znalosti g můžeme snadno určit s využitím z rov. (6) polorozpětí

$$\tilde{x}_{1-p} - \tilde{x}_{0,5} = \frac{R}{g}(\exp(-g z_p) - 1) \cdot \exp(h\, z_p^2/2) \tag{21}$$

Po úpravě vyjde lineární závislost vzhledem k h

$$y^* = \ln\left[\frac{g(\tilde{x}_{1-p} - \tilde{x}_{0,5})}{\exp(-g z_p) - 1}\right] = \ln R + h\, z_p^2/2 \tag{22}$$

Vynesením y^* vs. $z_p^2/2$ (modifikovaný graf pseudosigma) vyjde v případě platnosti g-h rozdělení s konstantními g, h přibližně lineární závislost.
Jednoduše lze také postupovat v případě, že g_p je vyjádřeno rov. (15) a h_p je vyjádřeno rov. (20). Protože je odhad g_p nezávislý na h_p, lze stejně jako u "čistého" g-rozdělení nalézt parametry g_0, g_1. Pro odhad parametrů h_0, h_1 však již nelze použít rov. (20), ale je třeba provést symetrizační transformaci (stejně jako u rov. (21)).
Po úpravách vyjde vztah

$$y_p^* = \ln R + \frac{h_0}{2} z_p^2 + \frac{h_1}{2} z_p^4 \tag{23}$$

kde
$$\gamma_p^{*} = \ln\left[\frac{(\tilde{x}_{1-p} - \tilde{x}_{0,5})\,\tilde{g}_p}{\exp(-\tilde{g}_p z_p) - 1}\right]$$

se určuje při znalosti g_0, g_1 pro každé p. Závislost γ_p^{*} vs. $z_p^2/2$ se označuje jako <u>zobecněný graf špičatosti</u>. Vyjde-li parabolický, znamená to, že je nutné použít zobecněné g-h rozdělení s nekonstantními parametry g, h.

Pro ověřování platnosti různých typů g-h rozdělení lze pochopitelně použít formální aparát lineární regrese a testovat významnost směrnic, resp. úseků ve výše uvedených grafech. Pro účely předběžné analýzy dat však běžně postačuje posouzení vlastních grafů.

5. PROGRAM EXGR

Program EXGR byl vytvořen v jazyce HPL pro stolní počítač HP 9825 vybavený plotterem. Skládá se z řady různých grafů pro posouzení statistických zvláštností dat. Kromě řady dalších se pro posouzení <u>šikmosti</u> používá grafů

- symetrie, t.j. závislosti $0.5(\tilde{x}_p + \tilde{x}_{1-p})$ vs. $z_p^2/2$
- šikmosti, t.j. závislosti g_p vs. z_p^2

Pro posouzení <u>špičatosti</u> se používá grafů :

- pseudosigma, t.j. závislosti $\ln[(\tilde{x}_{1-p} - \tilde{x}_p)/(-2z_p)]$ vs. $z_p^2/2$
- špičatosti, t.j. závislosti h_p vs. z_p^2.

V případě, že vyjde výrazná šikmost vyjádřená parametrem g (resp. g_p), provádí se před posuzováním špičatosti symetrizační transformace kvantilů

$$\tilde{x}_p^{*} = g_p\,\tilde{x}_p \,/\, [\exp(g_p z_p) - 1] \tag{24}$$

Vzhledem k účelu použití se pouze kreslí jednotlivé grafy spolu s odpovídajícím robustním (mediánovým) odhadem regresní přímky.

6. ZÁVĚR

Z výše uvedeného je patrné, že manipulace s g-h rozdělením je velmi snadná. Navíc je možné nejdříve určit šikmost (část g-rozdělení) a pak zpracovávat špičatost (část h-rozdělení) po symetrizační transformaci. Pro účely generace náhodných čísel je výhodné, že je g-h rozdělení definované přes kvantilové funkce. Toho se také s výhodou používá při odhadech parametrů. Pro účely předběžné analýzy dat je zase výhodné, že lze jednoduše popsat i případy nekonstantní šikmosti resp. špičatosti v závislosti na vzdálenosti od mediánu.
Také grafy symetrie a pseudosigma lze interpretovat s ohledem na jejich význam vzhledem ke g-h rozdělení.

Při použití g-h rozdělení jako systému empirických rozdělení je však nevýhodou, že nelze obecně analyticky nalézt odpovídající frekvenční funkce. Pokud je to účelem, bude zřejmě výhodnější použití jiného translačního systému empirických rozdělení, jako je Johnsonův, Shapiro - Wilkův atp.

7. LITERATURA

/1/ Militký J.: Zpracování experimentálních dat I, předběžná analýza, Skripta DT Ostrava 1986

/2/ ncyklopedia of Statistical Sciences, vol 4, str. 298-301, J.Wiley New York 1982

/3/ Hoaglin D.C., Mosteler F., Tukey J.W. Eds. : Exploring data tables trends and shapes, J.Wiley New York 1985, kap. 11