A hibameghatározás bizonytalanságai

CSERNYÁK LÁSZLÓ*–HAJAGOS BÉLA**–STEINER FERENC**

A hiba becsült értékei a mintaelemszám (n) növekedésével egyre jobban közelítenek egy elvi értéket; aktuális esetben, ha n elég nagy, legkönnyebben a hibabecslés aszimptótikus szórása alapján tájékozódhatunk a bizonytalanság mértékéről. Matematikai statisztikai kézikönyvek (ld. pl. Cramér 1958) megadják a hiba empirikus szórására (σ_{emp}) és a valószínű hibára (Q) vonatkozó általános formulákat, de ha a fejlődés a gyakran előforduló esetekben új hiba definíció alkalmazását igényli (ld. a P definícióját illetően Ferenczy et al. 1988), akkor a gyakorlati szakembernek az erre vonatkozó bizonytalanságról is tájékozottnak kell lennie. A dolgozat megadja a P aszimptótikus szórásának formuláját és (ábrákon) az értékeit is ($f_a(x)$-re optimális k-kra és konstans k = 2-re egyaránt) és az $f_a(x)$ szupermodellen hasonlítja össze az eredményeket egyéb módon számított hibák (elsősorban σ_{emp}) aszimptótikus szórásával.

A dolgozat végül megadja a leggyakoribb értékek aszimptótikus szórására vonatkozó becslésnek a hibáját is.

С увеличением количества элементов проб (n) оцениваемое значение приближается к определенному теоретическому значению; в действительности, если число n достаточно велико, то степень достоверности проще всего может быть оценена на основе асимптотической дисперсии оценки погрешности. В спарвочниках по математической статистике (см., например, Крамер, 1958) приводятся общие формулы для определения эмпирической дисперсии и вероятной погрешности (Q), однако часто возникает необходимость применения нового метода определения погрешности (в связи с определением погрешности P см. работу Ференци и др., 1988); причем специалиста, применяющего этот метод на практике, необходимо информировать о достоверности метода.

В работе приводится формула асимптотической дисперсии P, ее значения (см. рисунки) для f_a (x) при оптимальных значениях k и постоянном езначении k = 2, а также с помощью супермодели f_a (x) результаты сравниваются с асимптотической дисперсией (в первую очередь $\sigma_{эпм}$) погрешностей, определенных другими способами.

В заключение в работе приводится погрешность оценки асимптотической дисперсии методом наиболее частых значений.

The estimations of error approximate in general more and more a theoretical value, if the sample size (n) increases; the uncertainty is measured by the asymptotic variance of the error estimation. Handbooks of statistics (e. g. Cramér 1958) contain general formulae for the asymptotic variance of the samle standard deviation (σ_{emp}) and that of the probable error (Q), but new error definitions (as e. g. the definition of P in Ferenczy et al. 1988) need corresponding informations about the uncertainties of the error estimations in question. They are given in the present paper for the dihesion ε, for P and for the asymptotic variance of the most frequent value (A_M), mainly on ground of the supermodel $f_a(x)$. Comparisons are made with the asymptotic variance of the sample standard deviation and with that of the semi-intersextile range. – Preliminary Monte Carlo investigations show that in the domain $5 \leq n \leq 160$ the uncertainties of the dihesions can be calculated already according to the given asymptotic rule.

1. Bevezetés

Még nem is olyan túl régen felesleges elméletieskedésnek számított gyakorlati szakemberek körében a hiba hibájával foglalkozni. A számítástechnika fejlődése azonban lehetővé tette nagy hatásfokú, de sok műveletet igénylő statisztikai algoritmusok alkalmazását, s ez különböző módszerek gyakorlati összehasonlítása esetén (a relatív hatásfok becslésekor) elengedhetetlenné teszi a hiba kellően pontos ismeretét. A 20% körüli hatásfok-különbségnek például már komoly költségkihatásai lehetnek, – de ezt a különbséget nyilván még indikálni sem tud-

* ELGI, Budapest
** NME Geofizikai Tanszék

juk kellő biztonsággal, ha a hibabecslések szórása mondjuk 30%. A hiba hibájának a vizsgálata tehát ma már közvetlen gyakorlati fontossággal bír, és bár továbbra is szigorúbbak a követelményeink a százalékos hibát illetően magára a primer módon mért mennyiségekre, mint a hiba hibájára vonatkozóan, az utóbbi ma már semmiképpen sem fogadható el oly nagy értékűnek, hogy az alig legyen több nagyságrendi tájékoztatásnál. Sajnos a σ_{emp} értéke bizonyos körülmények között az adatok túlnyomó többségére jellemző eltéréseket még nagyságrendileg sem jellemzi helyesen, így más módon például a P hibával történő jellemzés ilyenkor egyenesen elengedhetetlennek tűnik, de a *Ferenczy et al. 1988*-ban definiált P egyébként is sokkal megbízhatóbb (robusztusabb, rezisztensebb) hibajellemző. A P hibájára vonatkozó vizsgálat ennek a jellemzőnek a behatóbb ismeretét, adekvát használatát és ezeken keresztül a földtudományi adatrendszerek gazdaságos interpretációját egyaránt elősegítheti.

A P definíciója adott $f(x)$ sűrűségfüggvény esetén (ld. *Ferenczy et al. 1988*)

$$P = \varepsilon \cdot \exp\left\{\frac{1}{2}\int_{-\infty}^{\infty} \ln\left[1+\left(\frac{x}{k\varepsilon}\right)^2\right]\cdot f(x)\,dx\right\} \tag{1}$$

ahol x a mért adatnak a k faktorral végrehajtott, általános leggyakoribb érték szerinti kiegyenlítés (ld. *Steiner 1985*) eredményétől való eltérése, $f(x)$ ezen eltérések valószínűségsűrűség-függvénye és ε az $f(x)$ dihéziója. (Szimmetrikus hibaeloszlás esetén sok esetben felesleges a fenti, bonyolultan hangzó specifikáció, hiszen a kiegyenlítések ugyanazt a hiperfelületet, illetőleg pontot – a szimmetriapontot – definiálják; utóbbi esetben – bármely kiegyenlítésnél – az x a mért érték távolsága a szimmetriaponttól.)

Foglalkozzunk először a

$$P = P/\varepsilon \tag{1a}$$

hibájának a meghatározásával.

2. A $\overline{P}$-meghatározás hibája

A robusztus statisztika egyik középponti jelentőségű fogalma az $IC(x)$-szel jelölt hatásfüggvény, amelynek definícióját közvetlenül $\overline{P}$-ra írjuk fel:

$$IC(x) = \lim_{\Delta\to 0} \frac{\overline{P}[(1-\Delta)f(x)+\Delta\cdot\delta(x)] - \overline{P}[f(x)]}{\Delta} \tag{2}$$

ahol $\delta(x)$ a Dirac-δ (ld. pl. *Steiner 1985*). Látjuk (kicsiny Δ-kat feltételezve), hogy a hatásfüggvény Δ-szorosa minden x-re azt adja meg, hogy egy Δ valószínűséggel jelentkező járulékos x-értékű adat mennyivel változtatja meg P értékét.

Első lépésként konstansnak tekintjük ε-t. Ekkor írhatjuk (behelyettesítve (1)-ből $\overline{P} = P/\varepsilon$-t a (2) kifejezésbe):

$$IC(x) = \lim_{\Delta\to 0}\frac{1}{\Delta}\left\{\exp\left[(1-\Delta)\,\frac{1}{2}\int_{-\infty}^{\infty}\ln\left[1+\left(\frac{x}{k\varepsilon}\right)^2\right]\cdot f(x)dx + \right.\right.$$

$$+\frac{\Delta}{2}\ln\left[1+\left(\frac{x}{k\varepsilon}\right)^2\right]\Bigg]-\exp\left[\frac{1}{2}\int_{-\infty}^{\infty}\ln\left[1+\left(\frac{x}{k\varepsilon}\right)^2\right]\cdot f(x)\,dx\right]\Bigg\}=$$

$$=\lim_{\Delta\to 0}\frac{\bar{P}}{\Delta}\left\{\left[\frac{1+\left(\frac{x}{k\varepsilon}\right)^2}{\bar{P}}\right]^{\Delta}-1\right\}=\bar{P}\ln\frac{\sqrt{1+\left(\frac{x}{k\varepsilon}\right)^2}}{\bar{P}}. \tag{3}$$

(Az utolsó lépésben felhasználtuk, hogy bármely $u>0$ mennyiségre, a l'Hospital szabály szerint

$$\lim_{\Delta\to 0}\frac{u^{\Delta}-1}{\Delta}=\lim_{\Delta\to 0}\frac{u^{\Delta}\cdot\ln u}{1}=\ln u.$$

Az aszimptótikus szórásnégyzet általános alakja (ld. pl. *Steiner 1985*):

$$A^2=\int_{-\infty}^{\infty}IC^2(x)f(x)\,dx, \tag{4}$$

így konstans ε esetén az aszimptótikus szórás (3) alapján

$$A_{\bar{P}}=\bar{P}\cdot\sqrt{\int_{-\infty}^{\infty}\left[\frac{\ln\sqrt{1+\left(\frac{x}{k\varepsilon}\right)^2}}{\bar{P}}\right]^2\cdot f(x)dx}. \tag{5}$$

(Furcsa, hogy az integrandusznak zérushelye van $\frac{x}{k\varepsilon}=\sqrt{\bar{P}^2-1}$-nél; persze, ha valamilyen x-re valóban elfogadható, hogy az nem számít P hibája szempontjából, annak közel kell lennie P-hez; Cauchy-nál ez az $x=\sqrt{3}=1{,}73$ valóban nincs messze $P=2$-től.)

3. A dihézió meghatározásának hibája

Valamely S skálaparaméter meghatározásának aszimptótikus szórásnégyzete (ld. pl. *Steiner 1985*)

$$A^2=S^2\frac{\int_{-\infty}^{\infty}\chi^2\left(\frac{x}{S}\right)f(x)\,dx}{\left[\int_{-\infty}^{\infty}\chi'\left(\frac{x}{S}\right)\cdot\frac{x}{S}f(x)\,dx\right]^2}, \tag{6}$$

ha S-et az

$$\int_{-\infty}^{\infty}\chi\left(\frac{x}{S}\right)\cdot f(x)dx=0 \tag{7}$$

követelés definiálja és $f(x)$ az origóra szimmetrikus. A minta alapján való S-meghatározás (7) összeg-megfelelőjével történik.

A $\varkappa$-függvény analitikus alakja az ε dihézió meghatározásakor

$$\chi(x) = \frac{3x^2 - 1}{(x^2+1)^2}. \tag{8}$$

Nincs tehát semmi akadálya, hogy $f(x) = f_a(x)$-szel az a típusparaméter függvényeként számítsuk ki a (6) szerinti A_ε aszimptótikus szórást, mint a függvényét, ahol $f_a(x)$ a *Csernyák* és *Steiner 1982* által modellezési célokra bevezetett eloszlásmodell-család. (Az $f_a(x)$-et definiáló formulát Ferenczy et al. 1988 is közli ebben a folyóiratszámban.)

Az A_ε-ra (6) -tal, illetve (8)-cal kapott numerikus eredményeket kitünően közelíti $a > 1{,}8$, azaz $0 < \frac{1}{a-1} < 1{,}25$ esetén a következő empírikus formula:

$$A_\varepsilon = 2\varepsilon\left[1 - \frac{a-2}{\pi \cdot (a-1)}\right]. \tag{9}$$

Az $A_\varepsilon/\varepsilon$ pontos értékeinek függését az eloszlástípustól a $\left(0 < \frac{1}{a-1} < 2{,}5\right)$ tartományra az *1. ábra* vastagon kihúzott görbéje mutatja. (A (9) egyszerű empírikus formula $a < 1{,}8$ is használható, csak pontatlanabb: $\frac{1}{a-1} = 2{,}5$-nél, azaz $a = 1{,}4$-hél 5% körüli eltéréssel adja A_ε helyes értékét; $\frac{1}{a-1} = 1{,}67$-nél, azaz $a = 1{,}6$-nál már csak 2% az eltérés.)

4. A skálaparaméter-becslés további két módszerének a hibája

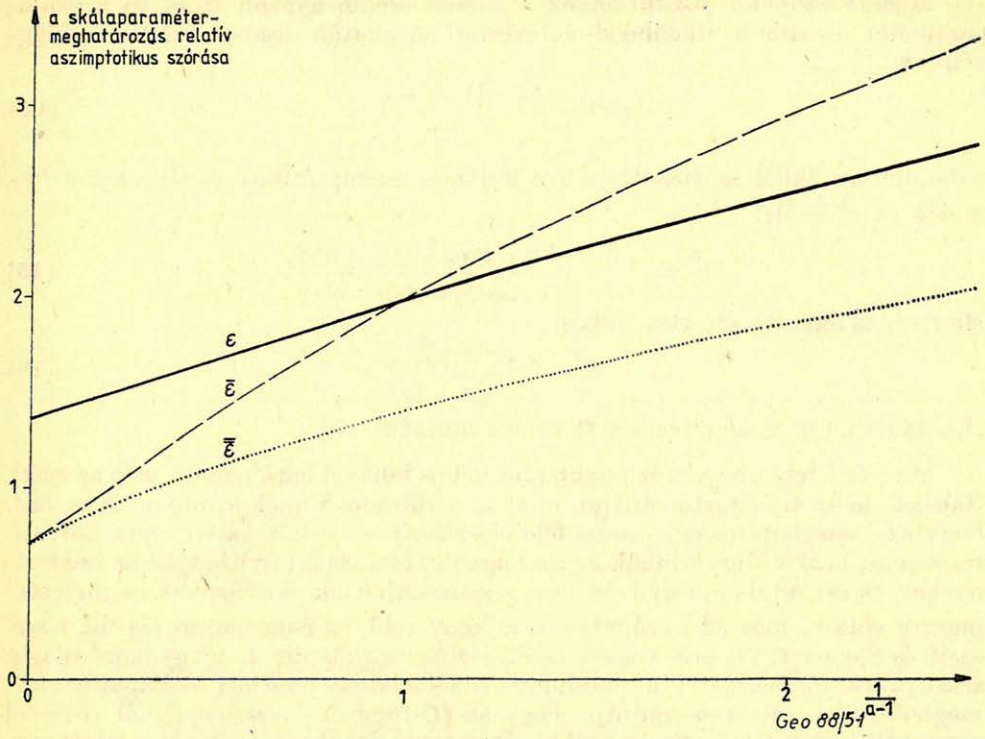
Az általános leggyakoribb érték-számítás második variánsánál, amely *Hajagos 1985* eredményeire épül, skálaparaméterként (röviden megfogalmazva) közvetlenül k meghatározása történik az alábbi $\varkappa$-függvény szerint:

$$\chi(x) = \frac{(a+1)x^2 - 1}{(1+x^2)^2}. \tag{10}$$

(A fenti, kissé nagyvonalú, ui. általános esetben csak közelítőleg érvényes megfogalmazás az általános leggyakoribb érték számítására vonatkozóan teljesen korrekt az egész $f_a(x)$ típus-családra, azaz a (10) alapján kapott $\bar{\varepsilon}$ ekkor pontosan $k\varepsilon$-nal egyenlő, ahol k az általános leggyakoribb értékszámítás szokásos variánsában használt paraméter, (ld. pl. a $k(a)$-ra vonatkozó empírikus formulát a *Ferenczy et al. 1988*-ban).

Az A_ε^2 aszimptótikus szórásnégyzet egzakt formuláját nem túl bonyolult analitikus alakban sikerül megadnunk, ha bevezetjük a következő jelölést:

$$S_j = \int_{-\infty}^{\infty} \left(\frac{S^2}{S^2+x^2}\right)^j f(x)dx. \tag{11}$$



1. ábra. A skálaparaméterbecslés relatív bizonytalansága háromféle meghatározási módszerre vonatkozóan, az $f_a(x)$ szupermodell különböző eloszlástípusaira

Рис. 1. Относительная достоверность оценки параметров шкалы для трех различных методов определения, для различных типов распределения супермодели $f_a(x)$

Fig. 1. Uncertainties of scale parameter estimations of different kind, in cas e of model distributions from the supermodel $f_a(x)$.

Így $S = \bar{\varepsilon}$-ra (6) alapján a következő eredmény adódik:

$$A_{\bar{\varepsilon}}^2 = \bar{\varepsilon}^2 \frac{S_2 - 2\dfrac{a+2}{a+1} S_3 + \left(\dfrac{a+2}{a+1}\right)^2 \cdot S_4}{4\left[S_1 - \dfrac{3a+5}{a+1} S_2 + \dfrac{2(a+2)}{a+1} S_3\right]^2}. \tag{12}$$

Ha az aktuális $f(x)$ azonos valamelyik $f_a(x)$-szel, akkor figyelembe vesszük az S_j-kre, illetve ezek összefüggéseire vonatkozó ismert képleteket (ld. *Hajagos 1985*), s így az $f_a(x)$-ekre vonatkozóan a következő alakra egyszerűsödik $A_{\bar{\varepsilon}}^2$ kifejezése:

$$A_{\bar{\varepsilon}}^2 = \bar{\varepsilon}^2 \cdot \frac{(a+2)\cdot(a+4)\cdot(a^2+3a+6)}{2a(a-1)\cdot(a+1)\cdot(a+6)}. \tag{13}$$

Az $A_{\bar{\varepsilon}}/\bar{\varepsilon}$ értékeit szaggatott vonallal rajzolt görbe mutatja az *1. ábrán.*

Mielőtt azonban diszkutálnánk a kapott eredményeket, írjuk fel a skálaparaméter maximum likelihood-elv szerinti meghatározásához tartozó χ-függvényt is:

$$\chi(x) = \frac{(a-1)\cdot x^2 - 1}{1+x^2}, \tag{14}$$

valamint az ebből származtatható általános aszimptótikus szórásnégyzet-formulát ($S = \bar{\bar{\varepsilon}}$-sal):

$$A^2_{\bar{\bar{\varepsilon}}} = \bar{\bar{\varepsilon}}^2 \frac{(a-1)^2 - 2a(a-1)S_1 + a^2 S_2}{4a^2(S_1 - S_2)^2}. \tag{15}$$

Ha $f_a(x)$ az aktuális eloszlás, akkor

$$A^2_{\bar{\bar{\varepsilon}}} = \bar{\bar{\varepsilon}}^2 \frac{a+2}{2(a-1)}; \tag{16}$$

$A_{\bar{\varepsilon}}/\bar{\varepsilon}$ görbéjét az *1. ábra* pontozott vonala mutatja.

Az $\bar{\varepsilon}$ és $\bar{\bar{\varepsilon}}$ tehát egyaránt kisebb százalékos hibával határozható meg az egész Gauss-Cauchy típustartományon, mint az ε dihézió. Ennek azonban az az ára, hogy az $\bar{\varepsilon}$ meghatározás a Gauss-féle eloszlástípushoz közeledve egyre kevésbé rezisztens, azaz a durva hibák egyre nagyobb értékben torzíthatják az eredményeket, és ezt általában nyilván nem kockáztathatjuk. A szárnyak szennyezésmentes voltára még az $\bar{\varepsilon}$-számításnál is kényesebb az $\bar{\bar{\varepsilon}}$-meghatározás (ld. *Csernyák* és *Steiner 1985b* erre vonatkozó részletes vizsgálatait; hogy gyakorlatilag a szárnyakra támaszkodik a maximum likelihood-elv szerinti skálaparaméter-meghatározás, azt az is mutatja, hogy az *IC*-függvény a szárnyaknál veszi fel maximális értékeit). — Ez az utóbbi alternatíva azonban egyéb okok miatt sem ajánlható gyakorlati alkalmazásra: mint *Csernyák* és *Steiner 1985a* kimutatta, még szimmetrikus eloszlásnál is előfordulhat ennél a skálaparaméter-becslésnél az, hogy a kettős iteráció másik ágának eredményeként kapott helyparaméter-becslésnek (azaz a statisztikai algoritmus általában leglényegesebb eredményének) végtelen nagy lesz az aszimptótikus szórása. —

Megemlítjük még, hogy az (11)-ben definiált S_j mennyiségek bevezetésével természetesen a számunkra legérdekesebb eset: a dihézió-meghatározás A^2_ε aszimptótikus szórásnégyzete is kifejezhető, ha tetszőleges $f(x)$ esetét akarjuk vizsgálni. A (6)-ból és (10)-ből adódó formula a következő ($S = \varepsilon$-ra):

$$A^2_\varepsilon = \varepsilon^2 \cdot \frac{9S_2 - 24S_3 + 16S_4}{(14S_2 - 16S_3)^2}. \tag{17}$$

5. A P hiba empirikus értékének a bizonytalanságai

Aszimptótikus értelemben jogos a P-re az *IC*-függvényt az alábbiak szerint felírni:

$$\frac{IC(P;\, x)}{P} = \frac{IC(\varepsilon;\, x)}{\varepsilon} + \frac{IC(\bar{P};\, x)}{\bar{P}}. \tag{18}$$

Ezzel P relatív hibájára vonatkozóan a következő kifejezés adódik általános esetben:

$$A_p/P = \sqrt{(A_{\bar{P}}/\bar{P})^2 + (A_\varepsilon/\varepsilon)^2 + C}, \tag{19}$$

ahol

$$C = \frac{2}{P} \int_{-\infty}^{\infty} IC(\varepsilon; x) \cdot IC(\overline{P}; x) f(x) dx. \tag{19a}$$

A (19)-ben a gyök alatti első tagként elfogadjuk a konstans ε-ra kapott (5)-ből ismert kifejezést; a számításhoz szükséges $\overline{P}$-t ekkor (1), illetve (1a) szerint $f_a(x)$ családra

$$\overline{P} = \mathrm{ezp}\left\{\int_0^{\infty} \ln\left[1 + x^2\right] \cdot f_a(x) dx\right\}$$

definiálja. (Ha az $f_a(x)$-családon kívüli $f(x)$ valószínűségeloszlásra akarunk $A_{\overline{P}}/\overline{P}$-t számolni, a (19) egyenletbe nyilván az általános (1), illetve (1a), valamint az (5) formulák alapján meghatározott $\overline{P}$ és $A_{\overline{P}}$ kerül. Ugyanekkor persze a (17) fogja a második tagot szolgáltatni.)

A C számításához szükségünk van az $IC(\varepsilon; x)$ függvény formulájára:

$$IC_\varepsilon(x) = \varepsilon \frac{3\left(\frac{x}{\varepsilon}\right)^2 - 1}{(14S_2 - 16S_3) \cdot \left[1 + \left(\frac{x}{\varepsilon}\right)^2\right]^2} \tag{20}$$

(a (11) szerint és $S = \varepsilon$-nal számított S_2-vel, illetve S_3-mal); a C integranduszában szereplő másik függvényt: $IC(\overline{P}; x)$-et már (3)-ból ismerjük.

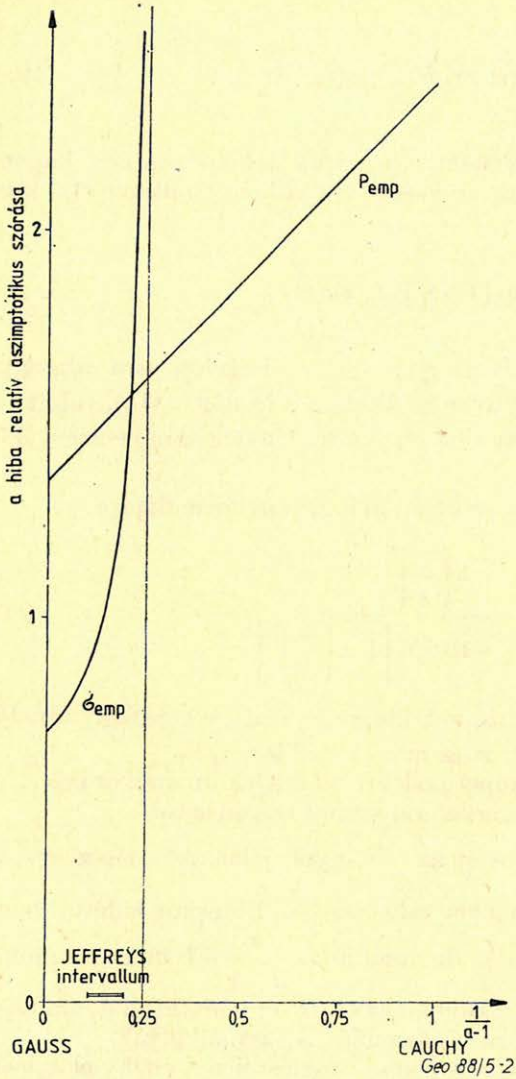
Mivel az A_p/P értékeit az $f_a(x)$ szupermodellre számítjuk ki, amikor is a (19)-ben a gyökjel alatti második tag egyszerűen (9) szerint számítható.

A *2. ábra* mutatja az A_P/P görbéjét az $\frac{1}{a-1}$-gyel jellemzett eloszlástípus függvényében. Az *1. ábra* A_P/ε görbéjével való összehasonlításkor kiderül, hogy A_p/P-ben a dihéziómeghatározás hibája dominál a $0 < \frac{1}{a-1} < 1$ intervallumban olyannyira, hogy ebben a típustartományban az A_P/P-re vonatkozó gyors tájékozódásként akár az $A_\varepsilon/\varepsilon$-ra vonatkozó (9) formulát is használhatjuk.

A P relatív aszimptótikus hibája a Gauss-eloszlástól a Cauchy-eloszlásig kb. *1,8*-szeresére növekszik a *2. ábra* szerint. Az ebben az irányban tapasztalt növekedést az eloszlás-szárnyak súlyosságának növekedése miatt természetesnek találjuk. – Sőt, a növekedés mértéke inkább mérsékeltnek mondható, különösen, ha az A_P/P görbéjét az $A\sigma_{\mathrm{emp}}/\sigma_{\mathrm{emp}}$ görbéjével hasonlítjuk össze.

6. A szórás empirikus értékének a meghatározási hibái

$A_{\sigma\,\mathrm{emp}}$ vizsgálatakor az x alatt továbbra is a kiegyenlítés eredményétől mért távolságot értjük; ez ebben az esetben persze a legkisebb négyzetek elve szerinti kiegyenlítést jelenti. (Legegyszerűbb eleve szimmetrikus hibaeloszlásra gondolni, amikor a különbségtétel felesleges.) A matematikai statisztika kézikönyvei (ld.



2. ábra. A hibabecslés relatív bizonytalansága kétféle hibadefinícióra vonatkozóan, az $f_a(x)$ szupermodell különböző eloszlástípusaira

Рис. 2. Относительная достоверность оценки погрешностей для двух различных методов определения погрешностей, для различных типов распределения супермодели $f_a(x)$

Fig. 2. Uncertainties of error estimations of different kind, in case of model distributions from the supermodel $f_a(x)$.

pl. *Cramér 1958*) közlik a σ_{emp}-re vonatkozó aszimptótikus szórás ($A_{\sigma\,emp}$) formuláját:

$$A_{\sigma\,emp} = \frac{1}{2\sigma}\sqrt{\int_{-\infty}^{\infty} x^4 f(x)\,dx - \sigma^4}\,; \tag{21}$$

ennek a kifejezésnek a létezéséhez $f(x)$ negyedik momentumának véges voltát persze fel kell tételeznünk.

Ez az utóbbi feltétel az $f_a(x)$ eloszláscsaládnál csak $a>5$ esetén teljesül. Így a $3>a\leq 5$ tartományban a nagy számok törvényének teljesülési üteméről nincs közelebbi információnk, csak azt tudjuk, hogy nem áll fenn a meghatáro-

zási pontosság növekedésére vonatkozóan az $1/\sqrt{n}$-nel való arányosság megnyugtató sajátsága; a pontosságnövekedés üteme éppen nagy n-eknél lassul le, amikor pedig éppen lehetőleg pontos értékekre törekednénk. – Ha $a \leq 3$, már σ sem létezik, így ezeknél az eloszlásoknál a nagy számok törvénye már semmilyen formájában sem teljesül: nem növekszik a pontosság n növekedésével. (Hogy a nagy számok törvénye fordítva is teljesülhet, arra nézve ld. *Csernyák* és *Steiner 1982* vizsgálatait.)

Kimutatható (ld. *Hajagos* és *Steiner 1988*), hogy az $f_a(x)$ eloszláscsaládra a következő egyszerű formulával számíthatjuk σ_{emp} esetén a relatív hibát:

$$A_{\sigma\,emp}/\sigma_{emp} = \sqrt{\frac{a-2}{2(a-5)}}. \qquad (22)$$

Az aszimptótikus szórás relatív értékeinek ezt a görbéjét szintén a *2. ábra* mutatja be.

Következtetéseink az alábbi pontokba fogalalhatók, beleértve az eddigi megállapításokat is:

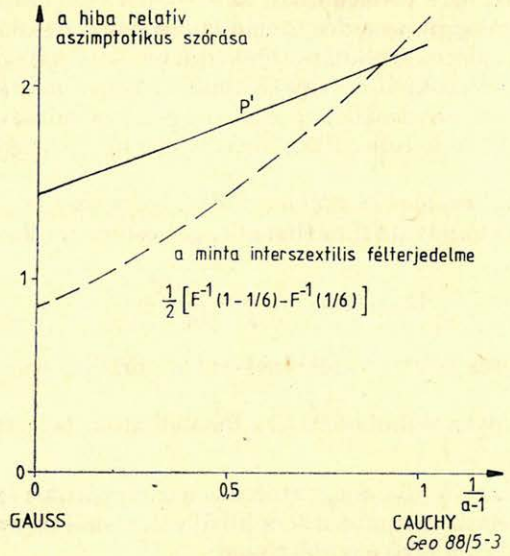
1. $a \leq 5$-re végtelen nagy a σ_{emp}-meghatározás aszimptótikus szórása;
2. a σ_{emp} és a P_{emp} relatív aszimptótikus hibája $a \approx 6$-nál egyezik meg (ez a Jeffreys-intervallum szélén levő eloszlástípus);
3. $a > 6$ esetén ugyan kisebb σ_{emp} relatív hibája, mint P_{emp}-é, a rezisztencia teljes hiánya miatt azonban σ_{emp} alkalmazása ebben a típustartományban sem jasolható a geofizikai és geológiai vizsgálatok túlnyomó többségénél.

A fentiek lényegében σ_{emp} diszkvalifikálását jelentik minden olyan esetben (pl. hatásfokbecsléseknél), amikor a hiba hibája nem lehet túlságosan nagy. (Ha szinte csak hiba nagyságrend meghatározására korlátozódnak igényeink, vagy valóban a Gauss-eloszlás közvetlen közelében levő szűk típustartomány jelentkezése várható, ráadásul garantáltan durvahiba-mentesen – mint pl. egyes geodéziai méréssorozatoknál, – akkor a σ_{emp} továbbra is elfogadható hibajellemzőnek.)

7. A P'-meghatározás hibái

A P hibát bevezető dolgozat (*Ferenczy et al. 1988*) $k = 2$ alkalmazását javasolja arra az esetre, ha nincsenek a hibaeloszlás típusára vonatkozó előzetes ismereteink (ekkor, ha szükségesnek ítéljük a megkülönböztetést, a P' jelölést alkalmazzuk). Indokolt tehát, hogy vizsgálataink erre az egyszerűbben használható hibadefinícióra is kiterjedjenek.

A *3. ábra* görbéje mutatja az erre az esetre vonatkozó relatív aszimptótikus szórás-görbét. (A számítások – mutatis mutandis – most is (19) alapján történtek.) Talán meglepőnek találjuk, összehasonlítva a *3. ábrát* a *2. ábrával*, hogy a Gauss-Cauchy tartományon valamivel kisebb ingadozással adódnak a hiba hibái P'-nél, mint P-nél (sőt könnyű tájékozódásul akár praktikus is elfogadni $A_P/P' = 1{,}8$-at jellemző értéknek, amitől az eltérés maximálisan 24% alatt marad a teljes Cauchy–Gauss tartományon). Nem szabad elfeledkeznünk azonban arról, hogy a hiba hibájának praktikusabb viselkedésével szemben a $k = 2$-vel végzett helyparaméter-meghatározások valamennyivel kisebb pontossága áll a $0 \leq \frac{1}{a-1} \leq 1$ intervallum két végéhez közeledve.



3. ábra. Az interszextilis félterjedelem és a P' becslési bizonytalanságainak összehasonlítása a $f_a(x)$ szupermodell különböző eloszlástípusaira

Рис. 3. Сравнение достоверности интерсекстильного полусъема и оценки P для различных типов распределения супермодели $f_a(x)$

Fig. 3. Comparison of the estimation uncertainties of the semi-intersextile range and those of the P' in case of model distributions from the supermodel $f_a(x)$.

8. Az interszextilis félterjedelem empirikus értékének bizonytalanságai

Tanulságos megvizsgálni, hogy a *2:1* valószínűségarányt pontosan megvalósító interkvantilis félterjedelmek, mint hibajellemzők, milyen hibával határozhatók meg. – Az $f_a(x)$-családra vonatkozóan az általános formulákat *Hajagos* és *Steiner 1988* közli tetszőleges kvantilisre, így semmi akadálya ezek alapján egy az eddigiekkel analóg görbe szerkesztésének.

A *3. ábrán* bemutatott, szaggatott vonalú görbe kedvező képet mutat az interkvantilis félterjedelem-meghatározás hibájára a *2:1* valószínűségaránynál.

(Legyen szabad e helyen felvetni az *interszextilis félterjedelem* terminus technicus használatát az *interkvartilis félterjedelem* mintájára a szinte használhatatlanul nehézkes, és még így sem ennyire egyértelmű következő kifejezés helyett: *a 2:1 valószínűségarányhoz tartozó interkvantilis félterjedelem.* – Az interszextilis félterjedelem nyilván a következőképpen írható fel az F eloszlásfüggvény inverzével:

$$\frac{1}{2}[F^{-1}(1-1/6)-F^{-1}(1/6)],$$

míg az interkvartilis félterjedelem kifejezése

$$\frac{1}{2}[F^{-1}(1-1/4)-F^{-1}(1/4)].$$

A továbbiakban élni fogunk a javasolt kifejezéssel.)

Az interszextilis félterjedelem (mint hibajellemző) relatív aszimptótikus szórása csaknem ugyanolyan kedvező értékről indul $\frac{1}{a-1} = 0$-nál, mint a σ_{emp}-re vonatkozó görbe a *2. ábrán*. Nagy különbség, hogy a σ_{emp} – a másik hibajellemzővel ellentétben – nem robusztus: az *5*-höz közeli (de annál nagyobb) a-k esetén teljesül ugyan még a nagy számok törvénye, de igen nagy aszimptótikus szórással (ld. a (22) formulát), *5*-nél kisebb a-knál pedig még ennél is kedvezőtlenebb sajátságok lépnek fel. (Erről a korábbiak során már említést tettünk; pl. a Cauchy-eloszláshoz közeledve n-nel egyre inkább növekvő σ_{emp}-értékeket kapunk (!), ami a hiba megítélését persze teljesen tévútra vezetheti.)

A fentiekben csak a robusztusság szempontjait említettük, a *3. ábra* két görbéjének összehasonlításakor azonban, amikor is két robusztus hibameghatározási eljárást kell egybevetnünk, újra gondolnunk kell a durva hibájú adatokra (outlier-ekre) való érzéketlenség, azaz a rezisztencia szempontjaira is. – Nem maradhatnak figyelmen kívül számítástechnikai vonatkozások sem.

Az utóbbiakkal kezdve: az interszextilis félterjedelem empirikus értékének meghatározásához szükség van olyan műveletekre (pl. nagyság szerinti sorba rendezés), amelyek plusz gépóraigényt és memóriakapacitás-többletet egyaránt jelentenek (magához a kiegyenlítéshez viszonyítva). Meggondolandó, hogy megéri-e ez azt a többletet, amit a hibameghatározás pontosságában nyerünk növekvő a-val a Jeffreys-intervallum felé, majd azon túl haladva a Gauss-eloszlásig. – Ha nincs információnk az eloszlástípusról, amely Cauchy-félének is adódhat (amikor pedig már fordított a helyzet a pontosságot illetően), akkor lehet, hogy fáradságosabban jutunk pontatlanabb hibajellemzőhöz.

Elgondolkodtató mindenesetre az interszextilis félterjedelem *3. ábrán* látható pontossági fölénye a Gauss-Cauchy típusintervallum nagyobbik részén: ha érdemesnek ítéljük, vállalni fogjuk a fentiekben említett számítástechnikai többletet, – annál is inkább, mert így nem csak közelítőleg, hanem pontosan *2:1* valószínűségarányt megadó hibához fogunk jutni. (Ami a pontosságok maximális eltérését illeti, a Gauss-eloszlásnál *32%*-kal kisebb az interszextilis félterjedelem relatív aszimptótikus szórása, mint a P'-é.)

A fentiek azonban természetesen csak arra az esetre vonatkozhatnak, amikor az adatok eloszlása *tiszta* $f_a(x)$ eloszlással modellezhető. A rezisztencia problémáinak teljes, vagy legalább valamennyire részletes elemzése nagyobb terjedelmet igényelne, ezért ezt mellőzni vagyunk kénytelenek. Annyi azonban azonnal belátható, hogy ha adataink több mint egy hatoda minősül durva hibájúnak a reális adattömörülés egyik oldalán, akkor az interszextilis félterjedelemnek semmi köze sem lesz az anyaeloszláshoz, míg a P továbbra is döntően az anyaeloszlás jellemzője marad a legtöbb ilyen extrém esetben is. (Ráadásul még arra is lehetőségünk van, hogy könnyen megszabaduljunk a távoli adatoktól súly szerinti vágással, ld. pl. *Steiner 1988;* ilyen megoldásra, mint erre *Ferenczy et al. 1988* már utalt, akkor lehet szükség, amikor a durva hibájú adatok *igen távoliak*, ugyanakkor a százalékos arányuk is *jelentős*. Egy igen egyszerű példa bemutatása bizonyára hasznos lesz: ha összesen *15* adatunk közül *3* db *+50*-es értékű durva hibájú adatunk van, egyébként *12* adat a (*−12, +12*) intervallumon egyenletes eloszlásból származó ún. *ideális minta* (±1, ±3, ±5, ±7, ±9 és ±11), akkor az interszextilis félterjedelem empírikus értékét *28,5*-nek találjuk, – azaz több, mint három és félszer akkorának, mint magából az ideális mintából nyerhető *8*-as értéket. – A P' is megérzi a nagy százalékban jelenlevő durva hibákat, de ezek

P' értékét viszonylag csak kis mértékben: 23%-kal növelik (az interszextilis félterjedelem-meghatározás ebben az esetben tehát *15,5*-szer nagyobb hibával van terhelve). Így – ha némi torzulással is, – P'_{emp} továbbra is az adattömörödésre magára vonatkozóan informál bennünket a hiba nagyságáról (s itt nem éltünk még a súly szerinti eliminálás imént említett tartaléklehetőségével sem). Rezisztencia-okokból tehát (egyéb itt nem tárgyalt esetekben is) a P hiba alkalmazását fogjuk általában kedvezőbbnek ítélni, hiszen mint láttuk, durva hibájú adatok miatt könnyen elveszíthetjük azt a pontossági előnyt a hibameghatározásban, amit a *3. ábra* görbepárja a vizsgált típusintervallum Gauss-eloszlás felé eső részén mutat. Ilyen, viszonylag mérsékelt előnyöknek a sokkal nagyobb, esetleg katasztrófális mértékű hátrányok kockázatának megszüntetése érdekében történő feladásakor szokás a robusztus statisztika irodalmában, *Anscombe 1960* nyomán, a *biztosítási díj* hasonlatával élni. Az ε, $\bar{\varepsilon}$ és $\bar{\bar{\varepsilon}}$ meghatározási hibájára vonatkozó, az *1. ábrán* látható görbék egybevetésekor ugyanúgy idézhető lett volna ez a találó analógia.

9. A leggyakoribb érték aszimptótikus szórásának meghatározási bizonytalanságai

Az eddigiekben tárgyalt hibadefiníciók az anyaeloszlásra vonatkoztak, azaz az egyes adatok hibáit jellemzik. Ezek értékei nem függhetnek n-től (pontosabban csak az empírikus értékek statisztikus ingadozása áll kapcsolatban n-nel). Magának a leggyakoribb értékek szerinti kiegyenlítés eredményének, pl. az M leggyakoribb értéknek a hibája persze annál pontosabb, minél nagyobb az n értéke. Pontosabban: bebizonyítható (ld. *Csernyák* és *Steiner 1983*), hogy szimmetrikus eloszlásokra az aszimptótikus szórás mindig véges, azaz a nagy számok törvénye a legelőnyösebb alakban: $\sqrt{n}$-nel arányos pontosságnövekedést biztosítva teljesül.

Szimmetrikus eloszlásokra és $k = 1$-re az M aszimptótikus hibája

$$A_M = \frac{\varepsilon}{\sqrt{n(\varepsilon)}} \tag{23}$$

(Csernyák és Steiner 1983), ahol $n(\varepsilon)$ azonos S_1-gyel. Az $n(.)$ jelölés arra szeretne figyelmeztetni, hogy ennek a mennyiségnek heurisztikusan könnyen értelmezhető (és gyakorlatilag is használható) jelentése van: $n.n(.)$ azonos a kiegyenlítés által figyelembe vett effektív adatszámmal. – A mintákra a kapott dihézió osztva a súlyösszeg gyökével nyilván az M-ek szórására ad becslést.

A leggyakoribb érték általánosítására Hajagos 1985 adott elvi alapot. Erre az általános leggyakoribb értékre az aszimptótikus szórást

$$A_M = \frac{\sqrt{a+2}}{a} \cdot \frac{k\varepsilon}{\sqrt{n(k\varepsilon)}} \tag{24}$$

adja (a k faktor az a típusparaméternél eredményez optimális hatásfokot; a $k(a)$-függvény képletét illetően ld. *Ferenczy et al. 1988*).

Jelöljük ((la) mintájára) $\bar{A}$-sal az A_M/ε hányadost; ezt nagy n-nél ε ingadozásai már csak jelentéktelenül befolyásolják, gondolatmenetünk tehát teljesen analóg lehet a *2.* és *5.* pontban követetthez. Így az A_M mennyiség relatív aszimptótikus szórása (A_{A_M}/A_M) felírható a következőképpen (v. ö. a (19) formulával):

$$A_{A_M}/A_M = \sqrt{(A_\varepsilon/\varepsilon)^2 + (A_{\overline{A}}/\overline{A})^2 + K}, \tag{25}$$

ahol

$$K = \frac{2}{A_M} \int_{-\infty}^{\infty} IC(\varepsilon; x) \cdot IC(\overline{A}; x) f(x) dx. \tag{25a}$$

Foglalkozzunk először az $\overline{A}$ becslésének aszimptótikus szórásával, $A_{\overline{A}}$-sal. – Kiindulásunk most is a hatásfüggvény:

$$IC(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} \{\overline{A}[(1-\Delta) f(x) + \Delta \cdot \delta(x)] - \overline{A}[f(x)]\} =$$

$$= \lim_{\Delta \to 0} \frac{k\sqrt{a+2}}{a \cdot \Delta} \left[\frac{1}{(1-\Delta) \cdot n(k\varepsilon) + \Delta \dfrac{(k\varepsilon)^2}{(k\varepsilon)^2 + x^2}} - \frac{1}{\sqrt{n(k\varepsilon)}} \right] =$$

$$= \lim_{\Delta \to 0} \frac{1}{\Delta} \frac{k\sqrt{a+2}}{a \cdot \sqrt{n(k\varepsilon)}} \left[\frac{1}{\sqrt{(1-\Delta)\left[1 - \dfrac{(k\varepsilon)^2}{n(k\varepsilon)\left((k\varepsilon)^2 + x^2\right)}\right]}} - 1 \right] =$$

$$= \overline{A} \frac{1}{2} \left[1 - \frac{(k\varepsilon)^2}{n(k\varepsilon) \cdot \left((k\varepsilon)^2 + x^2\right)} \right] \tag{26}$$

(az utolsó lépésben, a határátmenet előtt, felhasználtuk a binominális sorfejtést). Ebből az aszimptótikus szórásnégyzet (ld. újra a (4) általános formulát,) a következő:

$$A_{\overline{A}}^2 = \overline{A}^2 \frac{1}{4} \int_{-\infty}^{\infty} \left[1 - \frac{(k\varepsilon)^2}{n(k\varepsilon) \cdot \left((k\varepsilon)^2 + x^2\right)} \right]^2 f(x) dx = \overline{A}^2 \frac{1}{4} \left[1 - \frac{S_2}{S_1^2} \right] \tag{27}$$

(az S_j-k definícióját illetően ld. a (11) formulát). Felhasználva az $f_a(x)$ családra érvényes

$$S_2 = \frac{a+1}{a+2} S_1 \quad \text{és} \quad S_1 = \frac{a-1}{a} \tag{28}$$

összefüggéseket (mindkettőre nézve ld. *Hajagos 1985*), az $\overline{A}$ relatív aszimptótikus szórása az $f_a(x)$ szupermodellre vonatkozóan végül az

$$A_{\overline{A}}/\overline{A} = \sqrt{\frac{1}{2(a+2) \cdot (a-1)}} \tag{29}$$

alakban írható fel. A (25)-öt (és (9)-et) figyelembe véve tehát végül az f_a-családra vonatkozóan a következőt írhatjuk:

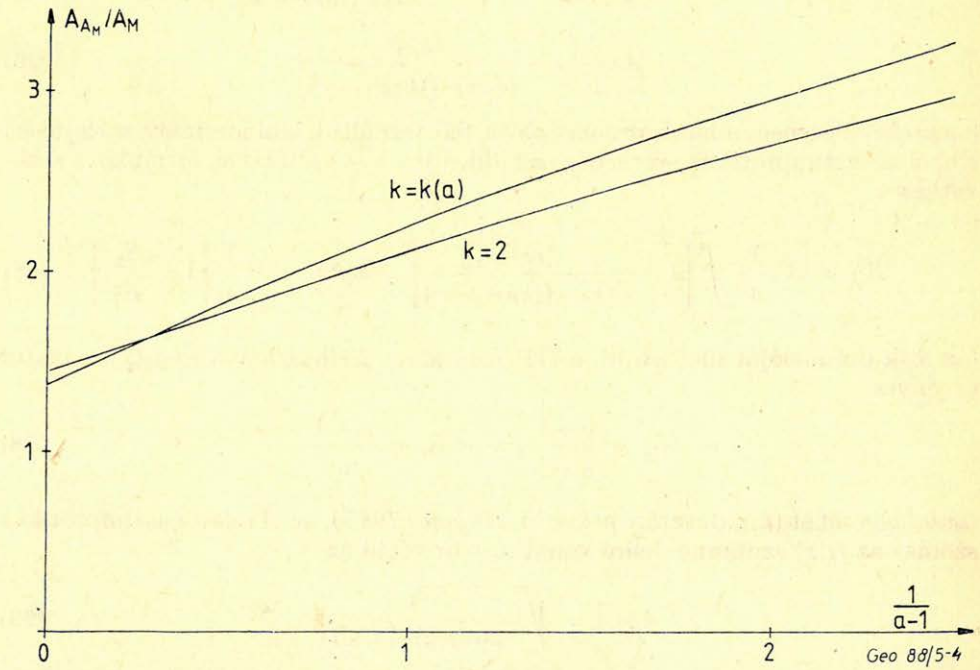
$$A_{A_M}/A_M = \sqrt{4\left[1 - \frac{a-2}{\pi \cdot (a-1)}\right]^2 + \frac{1}{2(a+2) \cdot (a-1)} + K} \tag{30}$$

(a K értékeit illetően ld. a *Táblázatot*.

Táblázat

a	ε	$A_\varepsilon/\varepsilon$	k=k(a)		k=2			
			K	A_{A_M}/A_M	$n(2\varepsilon)$	$A_{\overline{A}}/\overline{A}$	K	A_{A_M}/A_M
1,4	1,503	2,797	2,492	3,269	0,480	0,418	0,812	2,969
1,6	1,273	2,379	1,665	2,749	0,571	0,333	0,607	2,526
2	1,000	2,000	1,000	2,263	0,666	0,250	0,444	2,123
2,5	0,812	1,793	0,666	1,989	0,721	0,202	0,363	1,902
3	0,697	1,686	0,500	1,842	0,750	0,176	0,323	1,788
4	0,561	1,576	0,333	1,687	0,778	0,151	0,283	1,671
6	0,428	1,489	0,200	1,559	0,799	0,131	0,252	1,577
10	0,314	1,432	0,111	1,472	0,811	0,118	0,232	1,515

Nagy a-értékekre a gyök alatti második és harmadik tag elhanyagolhatóan kicsiny lesz, — de még a Cauchy-eloszlásnál ($a = 2$) is csak *1/8* a második tag értéke, s bár ekkor már $K = 1$, — az első összeadandó *4*-es értéke mellett, — az A_{A_M}/A_M értékét az A ingadozásainak a figyelembevétele még a Cauchy-eloszlásnál is csak kb. *13*%-kal emeli meg, $A_\varepsilon/\varepsilon$ értékével összehasonlítva. Ez annyit jelent, hogy a Gauss-Cauchy típustartományon nyugodtan használható az



4. ábra. A leggyakoribb érték aszimptótikus szórásának becslési bizonytalanságai kétféle k-választás esetén, az $f_a(x)$ szupermodell különböző eloszlástípusaira

Рис. 4. Достоверность оценки асимптотической дисперсии методом наиболее частого значения для двух различных значений k, для, различных типов распределений супермодели $f_a(x)$

Fig. 4. Uncertainties of the estimations for the asymptotic standard deviations of most frequent values, at alternative choice of the value k, in case of model distributions from the supermodel $f_a(x)$

$$A_{A_M}/A_M \approx 2\left[1 - \frac{a-2}{\pi \cdot (a-1)}\right]$$

közelítés.

Hasonló következtetésre jutunk abban az esetben, amikor az aktuális eloszlástípustól függetlenül, pontosabban: a típusra vonatkozó ismeret hiányában $k = 2$-vel végezzük számításainkat. Az a különböző értékeihez a (25) gyök alatt szereplő mennyiségeit, valamint a $k = 2$ esetre vonatkozó A_{A_M}/A_M értékeket a táblázat tartalmazza. (Maga A_M ekkor, azaz $k = 2$ esetén, $1{,}1 \cdot \varepsilon/\sqrt{n(2\varepsilon)}$-ként számítandó.) – A *4. ábra* bemutatja mind a $k = k(a)$, mind a $k = 2$ esetre vonatkozóan az A_{A_M}/A_M görbét.

Az ε konstans voltát (pl. $\overline{P}$ bizonytalanságának vizsgálatakor) az áttekinthetőség megőrzésére törekedve tételeztük fel. Szigorúan, véve a P-nek és A_M-nek a bizonytalanságára közölt eredmények felső korlátnak tekintendők, azonban mind a részletesebb (itt nem közölt) analitikus vizsgálatok, mind pedig Monte-Carlo-eredményeink azt mutatják, hogy a valóságos értékek olyan közel vannak a felső korláthoz, hogy az eltérés a gyakorlat számára egyelőre érdektelen. – Végül megemlítjük előzetes Monte Carlo vizsgálatainknak azt az érdekes eredményét, hogy a dihézió meghatározásának hibájára vonatkozóan már kicsiny mintaelemszámnál is igen jó közelítéssel teljesül a dolgozatban megadott aszimptótikus törvényszerűség.

IRODALOM

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W., Tukey, J. W. (1982): Robust Estimates of Location. – Princeton University Press, Princeton, N. J.

Anscombe, F. J. (1960): Rejection of Outliers. Technometrics, *2* (1)

Cramér, H. (1958): Mathematical Methods of Statistics. Princeton University Press, Princeton.

Csernyák, L., Steiner, F. (1983): Limit distribution of the most frequent values of samples from symmetrical distributions. Acta Geodaet., Geophys. et Mont. Acad. Sci. Hung. *18* (1–2) 1983.

Csernyák, L., Steiner, F. (1982): Untersuchungen über das Erfüllungstempo des Gesetzes der großen Zahlen. Publ. Techn. Univ. Miskolc, Ser. A. Mining *37* (1–2) pp. 47–64.

Csernyák, L., Steiner, F. (1985a): Bemerkungen zu der sogenannten "Cauchy Maximum Likelihood"-Abschätzung. Publ. Techn. Univ. Miskolc. Ser. A. Mining *38* (3–4) pp. 203–209.

Csernyák, L., Steiner, F. (1985b): Die Suche nach einer geeigneten Abschätzungsmethode für die Geophysik. Publ. Techn. Univ. Miskolc, Ser. A. Mining *40* (1–4) pp. 183–223.

Ferenczy, L., Hajagos, B., Steiner, F. (1988): A hagyományos hibadefiníció fogyatékosságai. Javaslat új hibadefiníció alkalmazására.

Hajagos, B. (1985): Die verallgemeinerten Student-schen t-Verteilungen und die häufigsten Werte. Publ. Techn. Univ. Miskolc, Ser. A. Mining *40* (1–4) pp. 225–238.

Hajagos, B., Steiner, F. (1988): Asymptotic behaviour of error estimations. Need for a practice in error estimation on new bases. Acta Geod., Geophys. et Mont. Acad. Sci. Hung. *23* (3–4)

Steiner, F. (1985): Robusztus becslések. Egyetemi jegyzet, Tankönyvkiadó, Budapest.

Steiner, F. (1988): Most frequent value procedures. Geophysical Transactions, *34*.