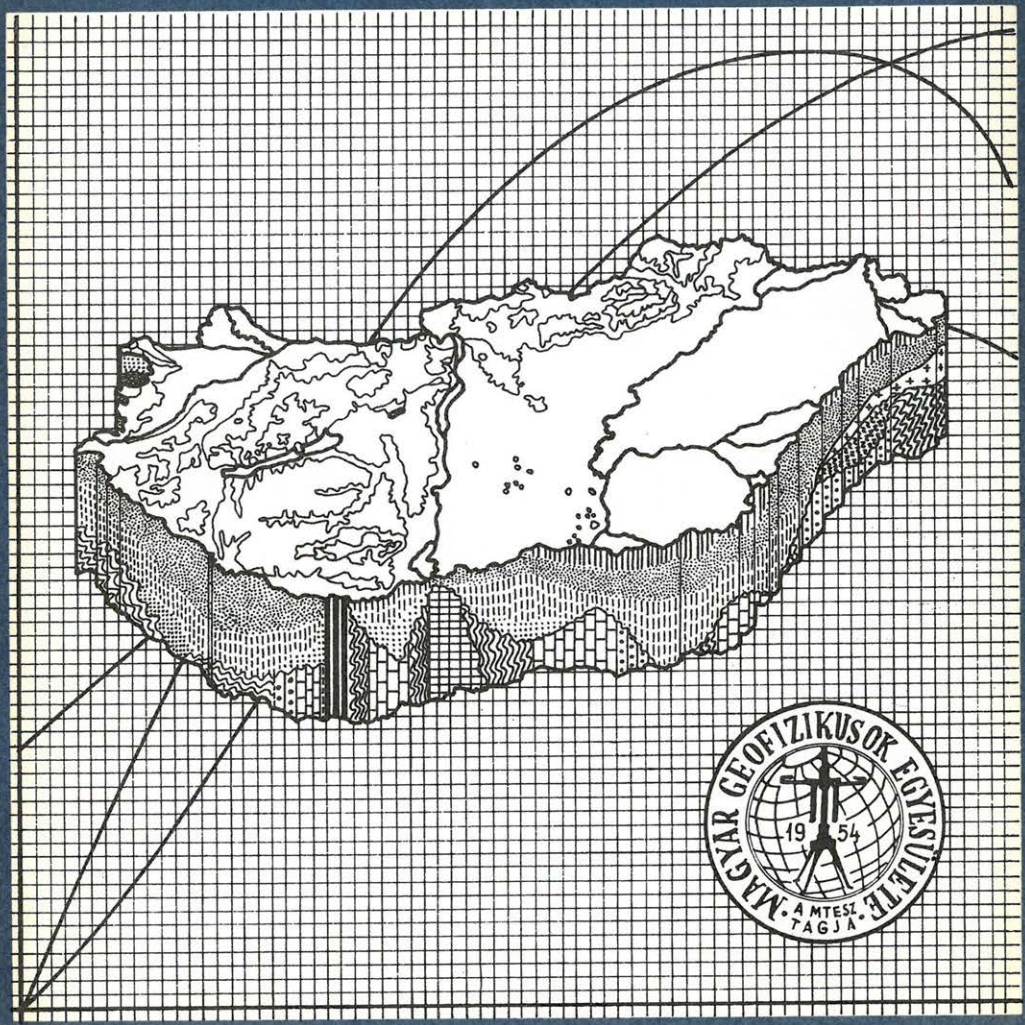
MAGYAR GEOFIZIKA

3



A MAGYAR GEOFIZIKUSOK EGYESÜLETE FOLYÓIRATA
BUDAPEST, 1988. XXIX. ÉVFOLYAM, 3. SZÁM

Szeidovitz

MAGYAR GEOFIZIKA
a MAGYAR GEOFIZIKUSOK EGYESÜLETE folyóirata
XXIX. évfolyam 3. szám

HU ISSN 0025–0120

TARATALOMJEGYZÉK

Főszerkesztő: Zelei András

Szerkesztőbizottság: Deres János, Kilényi Éva, Meskó Attila, Rádler Béla, Verő József

Szerkesztőség címe: 1368 Budapest VI., Anker köz 1. félemelet 17. Postafiók 240.

MAGYAR GEOFIZIKA

Kiadja: Delta Szaklapkiadó és Műszaki Szolgáltató Leányvállalat Budapest, Közraktár u. 4. 1093 Telefon: 175-200 levélcím: Budapest, Pf. 97. 1442

Felelős kiadó BUDAI FERENC főigazgató

Terjeszti a MAGYAR POSTA

Előfizethető a Hírlapkézbesítő Hivataloknál és a Posta Hírlapelőfizetési és Lapellátási irodáján. 1900 Budapest, V., József Nádor tér 1., vagy átutalással a 215–96162 pénzforgalmi jelzőszámra. Egy szám ára 32,50 Ft. Előfizetés fél évre 97,50 Ft, egy évre 195– Ft. Külföldön terjeszti a Kultúra, 1389 Budapest Pf. 149. és a Magyar Média 1392 Budapest, Pf. 279. 86–253

88.790. Állami Nyomda, Budapest – Felelős vezető: Mihalek Sándor igazgató

Egyesületi tagoknak tagdíj ellenében

Megjelenik évente hatszor

Index: 26507

СОДЕРЖАНИЕ

CONTENTS

Főszerkesztő: Zelei András

Szerkesztő bizottság: Deres János, Kilényi Éva, Meskó Attila, Rádler Béla, Verő József

Szerkesztőség címe: 1368 Budapest VI., Anker köz 1. félemelet 17. Postafiók 240.

Bevezetés

Matematikai statisztikai módszerek nélkülözhetetlenek a geofizika gyakorlatában; hatékonyságuk elsősorban attól függ, hogy kellő mértékben robusztusak és rezisztensek-e. *Robusztusság* alatt azt értjük, hogy a statisztikai módszer használhatósága *nem függ* túlzottan az aktuális valószínűségeloszlás *típusától*, a *rezisztencia a durva hibák*ra való *érzéketlenség*et jelenti.

A leggyakoribb értékek szerinti algoritmusok kifejlesztésében (több vagy kevesebb, különböző jellegű, de mindenképpen értékes munkával) sok magyar geofizikus kolléga, valamint két matematikus vett részt az elmúlt másfél évtizedben. Ennek köszönhetően egy 1985-ben tartott ötnapos mérnöktovábbképző tanfolyam keretében már a robusztus statisztika elméletébe ágyazott általános verzió ismertetésére kerülhetett sor; a továbbképző anyaga egyetemi jegyzetként is megjelent 1985-ben, „Robusztus becslések" címmel. Ezzel megnyílt az út a kiterjedtebb gyakorlati alkalmazások előtt, noha az elmélet még korántsem tekinthető lezártnak.

E füzet első cikke a fent említett továbbképzésen részt nem vett, illetve a jegyzet anyagát részleteiben nem ismerő geofizikus kollégák számára ad rövid áttekintést az alapfogalmakról és a legfontosabb eljárásokról. A második dolgozat a leggyakoribb értékek szerinti számítások néhány karotázs alkalmazását mutatja be, megadva egyben e módszernek a matematikai statisztika elméletén belül elfoglalt helyét. A harmadik és negyedik dolgozat új hibadefiníciót vezet be és tesz vizsgálat tárgyává, mivel a szokásos alapfeltevésektől való elszakadás minden téren, így a hibaszámítás terén is kötelességünkké teszi a dolgoknak az alapoktól induló, szisztematikus újragondolását. Ennek gyakorlati fontossága aligha lehet kérdéses, hiszen például az általánosan alkalmazott empirikus szórás megengedhetetlenül bizonytalan hibaértékeket adhat már a Gauss-típusútól nem túl távoli eloszlástípusok esetén is.

Steiner Ferenc

Введение

Методы математической статистики являются незаменимыми в геофизической практике; эффективность этих методов зависит в первую очередь от того, в достаточной ли степени они устойчивы и резистентны. Под устойчивостью имеется в виду то, что применяемость данного статистического метода практически не зависит от типа действительного распределения вероятности, а резистентность означает независимость от влияния грубых ошибок.

Венгерскими геофизиками и математиками был разработаны методы наиболее частых значений. Эти методы являются устойчивыми и резистентными, кроме этого, их простота обеспечивает возможность их рационального применения.

В данном номере журнала «Венгерская геофизика» опубликовано четыре статьи, связанные с этой темой. В первой статье кратко рассматриваются методы наиболее частых значений, во второй описаны возможности при-

менения данных методов при геофизическом каротаже (здесь же рассматривается связь методов наиболее частых значений с основными принципами математической статистики).

В третьей и четвертой статьях рассматривается новое понятие погрешности, которое является устойчивым, резистентным и полностью соответствует основным принципам метода наиболее частых значений.

Ф. Штейнер

Introduction

Statistical tools are unavoidable in the practice of geophysics; the efficiency of their use depends first of all upon their robustness and resistency. *Robustness* means that the usefulness of a statistical method *doesn't depend* in an undesirable degree *on the type* of the actual probability distribution, *resistance* means the *insensitivity to outliers*.

Hungarian geophysicists and mathematicians developped the so-called *most frequent value procedures*. The latter ones are robust and resistant and in addition, *simple enough* to apply them economically.

The present number of the journal "Magyar Geofizika" contains four contributions to this topic. The first paper gives a brief outline of the most frequent value procedures, the second one presents applications to well logging interpretation (and in the same time interconnections to general statistics). The third and fourth paper deal with a new error definition, being robust, resistant and fully adequate with the conception of the most frequent value calculations.

Prof. F. Steiner

Rövid bevezetés a leggyakoribb érték módszeréhez

FERENCZY LÁSZLÓ*

A dolgozat rövid áttekintést nyújt a leggyakoribb értékek szerinti számítások alapfogalmairól és alapeljárásairól. Részletesebb információkat a „Robusztus becslések" c. egyetemi jegyzet tartalmaz (lásd Steiner F., Tankönyvkiadó, Budapest, 1985).

В работе кратко рассматриваются основные понятия и способы вычислений методом наиболее частых значений. Более детальная информация приведена в университетских конспектах «Устойчивые оценки» (См. Штейнер Ф., издательство «Танкеньвкиадо», Будапешт, 1985 г.)

The paper presents the main concepts and procedures of most frequent value calculations in brief outline. A more detailed treatment is presented in the lecture notes "Robust estimations" (in Hungarian; Tankönyvkiadó, Budapest, 1985).

A leggyakoribb értékek elméleti megalapozása és a leggyakoribb értékek szerinti algoritmusok kidolgozása mintegy másfél évtizedes múltra tekint vissza. E témában számos publikáció és egy egyetemi jegyzet jelent meg, amelyek részletesen tartalmazzák az elméleti és részben gyakorlati kérdéseket, matematikai levezetéseket és bizonyításokat. E cikkben elsősorban a *Robusztus becslések* című jegyzetre *(Steiner 1985)* támaszkodva foglalom össze azokat a legfontosabb alapfogalmakat és definíciókat, valamint eljárásokat, amelyek ezen füzet további dolgozatainak könnyebb megértését segíthetik elő.

1. Leggyakoribb érték fix skálaparaméterű súlyfüggvénnyel

A helyparaméter az adatok helyére, a *skálaparaméter* az adatok tömörülésének mértékére jellemző. Ha S jelöli a skálaparamétert, T a helyparamétert, akkor a valószínűség-sűrűségfüggvény általában

$$f(x) = f(T; S; x) \tag{1}$$

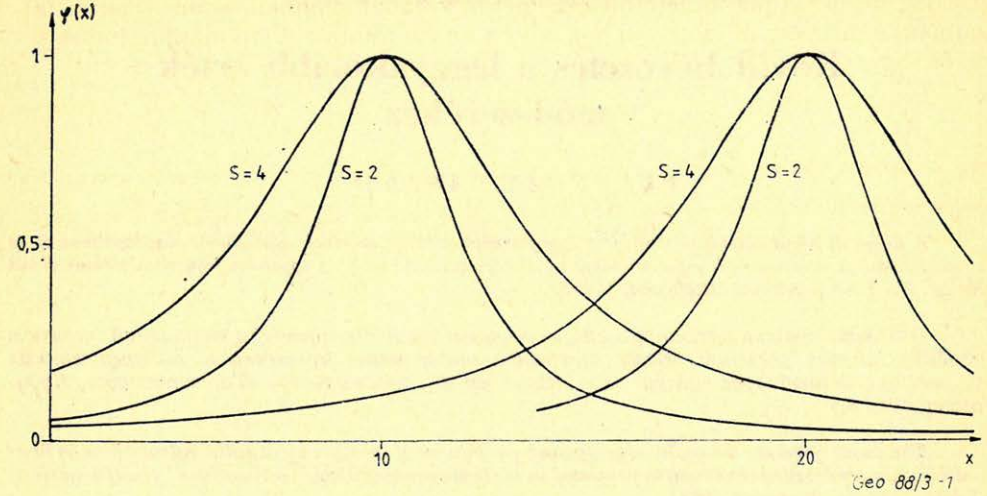
alakú. Például a Cauchy-eloszlás esetében a fenti kifejezés

$$f_C(x) = \frac{1}{\pi} \cdot \frac{S}{S^2 + (x-T)^2}. \tag{2}$$

Az *1. ábrán* mutatjuk be az $S = 2$ és $S = 4$-nek valamint a $T = 10$ és $T = 20$-nak megfelelő $f_C(x)$ görbéket. A T helyparaméter a szimmetriapont helyét, az S skálaparaméter pedig – az előzőtől függetlenül – azt definiálja, hogy az adatok a T környezetében sűrűbb vagy ritkább elhelyezkedéssel várhatók.

Cauchy-eloszlás esetén (2) alapján integrálással meggyőződhetünk arról, hogy S az interkvartilis félterjedelemmel $\left(Q = \frac{Q_{3/4} - Q_{1/4}}{2}\text{-vel}\right)$ azonos. Tulaj-

* NME Geofizikai Tanszék



1. ábra. Négy Cauchy-típusú sűrűségfüggvény, különböző T- és S-értékekhez (a jobb áttekinthetőség kedvéért a maximális értékekhez történt a normálás)

Рис. 1. Четыре различных функции плотности Коши для различных значений T и S/в целях лучшей обзорности нормирование выполнено для максимальных значений)

Fig. 1. Four density functions of Cauchy type (the normalization was made for the maximum value)

donképpen minden eloszlástípus felírható volna Q-val, mint skálaparaméterrel. A gyakorlatban azonban inkább eloszlástípusként más és más skálaparamétert szokás alkalmazni azért, hogy minél egyszerűbben írhassuk fel a sűrűségfüggvény analitikus alakját.

A durva hibákról. Ha a számegyenesen pontonként ábrázoljuk mérési eredményeinket, azonnal szembetűnik, ha egy vagy több *kieső pontunk* van; ekkor durva hibáról szokás beszélni. A durva hibával terhelt mérési eredményre jellemző, hogy az a többi mérési eredménytől jelentősen eltér. Ezeket az adatokat nem szoktuk a *kézi kivitelezésű* számításainkban szerepeltetni (egyszerűen elhagyjuk azokat).

Hogyan lehetne algoritmizálni a durva hibával terhelt értékeknek önkényesnek tűnő elhagyását?

Súlyozott átlagok. A legtöbb valószínűség-elméleti kézikönyv bebizonyítja, hogy az $x_1, x_2, \ldots, x_i, \ldots, x_n$ minta alapján a

$$\frac{\sum_{i=1}^{n} q_i x_i}{\sum_{i=1}^{n} q_i} \tag{3}$$

szerint képzett, ún. súlyozott átlag akkor adja a legkisebb szórású becslést, ha minden q_i súly azonos. Ha azonos mérési körülmények között nyertük adatainkat, akkor a matematikai statisztikának ez az eredménye teljesen kézenfekvő, sőt triviális.

Durva hibákkal terhelt mérési eredményeknél azonban olyan súlyozás alkalmazása az előnyös, melynél a q_i súly a minta adateloszlása alapján kapja értékét, azaz

$$q_i = q_i(x_1, x_2, \ldots, x_n). \tag{3a}$$

Ez gyakorlatilag annyit jelent, hogy a durva hibával terhelt adatok zérus vagy ahhoz közeli súlyt kapnak. Vagyis az $f(x)$ anyaeloszlás erre az x_k adatra már $f(x_k) = 0$, vagy $f(x_k) \approx 0$, tehát az ilyen érték *valószínűtlen*. Ha a (3) helyett a

$$\frac{\sum_{i=1}^{n} f(x_i) \cdot x_i}{\sum_{i=1}^{n} f(x_i)} \tag{4}$$

súlyozott átlagot alkalmazzuk, akkor a durva hibák ellen biztosan védve vagyunk, minden szubjektívnek minősülő, *önkényes* elhagyás nélkül.

A (4) gyakorlati alkalmazásának azonban súlyos akadálya, hogy nem ismerjük sem az $f(x)$ anyaeloszlás típusát, sem paramétereit (T és S). Persze, ha ismernénk, sem mintavételre, sem (4) szerinti helyparaméter becslésre nem volna szükség.

A durva hibájú adatok elvetése, valamint a legsűrűbben elhelyezkedő középső pontok kiemelt *megbecsülése* akkor is megörténik, ha nem az anyaeloszlással, hanem valamely ahhoz hasonló $\varphi(x)$ súlyfüggvénnyel súlyozunk. Ha számítástechnikailag a legegyszerűbb megoldásra törekszünk, akkor a

$$\varphi(x) = \frac{\varepsilon^2}{\varepsilon^2 + (x - M)^2} \tag{5}$$

súlyfüggvény választása a legkedvezőbb.

Leggyakoribb érték fix skálaparaméterű súlyfüggvénnyel. Tekintsük egyelőre ismertnek a $\varphi(x)$ súlyfüggvény ε-nal jelölt skálaparaméterét (meghatározásával később foglalkozunk). Az (5)-ben M-mel jelöltük a súlyfüggvény ismeretlen helyparaméterét. Nagy n mintaelemszámnál az (5) súlyokkal felírt (3) súlyozott átlagtól joggal remélhetjük (szimmetrikus eloszlásoknál teljesen nyilvánvalóan), hogy a valódi M-et szerepeltetve $\varphi(x)$-ben, magát M-et adja vissza:

$$M = \frac{\sum_{i=1}^{n} \varphi(x_i) \cdot x_i}{\sum_{i=1}^{n} \varphi(x_i)}, \tag{6}$$

amelybe behelyettesítve (5)-t, a következő összefüggéshez jutunk:

$$M = \frac{\sum_{i=1}^{n} \dfrac{\varepsilon^2 \cdot x_i}{\varepsilon^2 + (x_i - M)^2}}{\sum_{i=1}^{n} \dfrac{\varepsilon^2}{\varepsilon^2 + (x_i - M)^2}}. \tag{7}$$

Magára az $f(x)$ anya eloszlásra vonatkozóan ennek az összefüggésnek az alábbi integrál alakja alkalmazandó:

$$M = \frac{\int_{-\infty}^{\infty} \frac{\varepsilon^2 \cdot x}{\varepsilon^2 + (x-M)^2} f(x)dx}{\int_{-\infty}^{\infty} \frac{\varepsilon^2}{\varepsilon^2 + (x-M)^2} f(x)dx}. \tag{8}$$

Akár (7), akár (8) egyenleteket tekintjük, a bal oldalon álló M a jobb oldalon is szerepel. Ezek az egyenletek így egy iterációs algoritmus alapformulái. A számítástechnika gyakorlatának megfelelően, a jobb oldalon az M mindenkori értékét tároló rekesz tartalmát kell a törtkifejezésbe behelyettesíteni. Az így számított érték lesz M új értéke, és így tovább, míg az új M nem, vagy csak alig különbözik az előzőtől. Részletesen felírva tehát: az iterációs algoritmus programmal való végrehajtásának tartamára a

$$M_{j+1} = \frac{\sum_{i=1}^{n} \frac{\varepsilon^2 \cdot x_i}{\varepsilon^2 + (x_i - M_j)^2}}{\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M_j)^2}} \tag{7a}$$

összefüggés érvényes. Az iteráció indításához (M kezdőértékére) megfelel pl. a számtani középérték vagy a mintamedián.

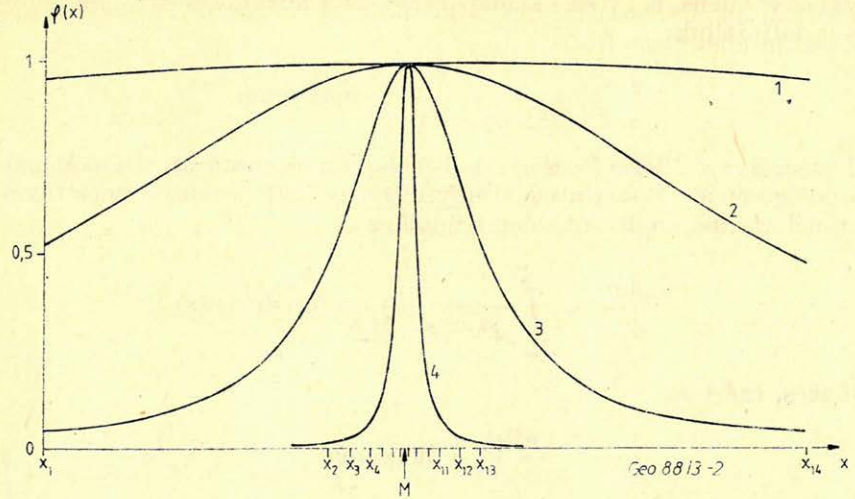
Matematikailag elég annyit mondanunk, hogy azt az M-et tekintjük az $f(x)$ helyparaméterének, amelyik kielégíti (8)-at, s ezt az M értéket az eloszlás leggyakoribb értékének nevezzük. A (7)-et kielégítő M-et pedig a minta leggyakoribb értékének nevezzük és az eloszlás leggyakoribb értékének becslésére használjuk (az M jelölés a most frequent value = leggyakoribb érték elnevezésből származik).

Nyilvánvaló (8)-ból, hogy szimmetrikus $f(x)$-eknél az M azonos a szimmetriaponttal (ami egyben a várható érték). Nem szimmetrikus esetben azonban M-et új helyparaméternek kell tekinteni, melyet nem befolyásolnak a pontok zömétől nagyon távol eső értékek.

2. A súlyfüggvény skálaparaméterének (ε) meghatározása

Maximális effektív adatszám fogalma. Ahogyan az (5) szerinti φ súlyfüggvényben szereplő ε-t az előzőekben az egyszerűség kedvéért ismertnek tételeztük fel, hasonlóan célszerű ezúttal abból kiindulni, hogy ismeretes a súlyfüggvényben szereplő M. A feladat viszont az ismeretlen skálaparaméter meghatározása.

Legyen adva olyan mintánk, ahol a mintaelemek – véletlenül – éppen szimmetrikus elhelyezkedésűek valamely érték körül *(2. ábra)*. Ebben az esetben az (7) egyértelműen adja a szimmetriapontot, bármilyen skálaparaméter-értéket válasszunk is. Ha ε túl nagy, akkor a *2. ábra 1.* görbéjét kapjuk és minden adathoz csaknem ugyanakkora $\varphi(x_i)$ súly tartozik. Láthatjuk, hogy még a nagyon távoli pontok is (x_1, x_{14}) a maximálishoz közeli súlyt kapnak. A pontok aszimmetrikus elhelyezkedésekor túl nagy ε-nál tehát a kieső pontok ugyanúgy



2. ábra. Adalék a (10) feltétel heurisztikus indokolásához

Рис. 2. Добавка для эвритического обоснования условия (10)

Fig. 2. Contribution to a heuristic train of thought leading to the demand Eq. 10

tönkretehetik a (7) szerinti becslést, mint a számtani átlag $\left(\varepsilon_n = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$ szerintit.

Ha túl kicsi az ε, akkor viszont nemcsak a durva hibával terhelt adatok maradnak figyelmen kívül, hanem a centrumban tömörülő pontcsoport szélén is jó néhány (ld. *2. ábra 4.* görbéjét). Ez gyakorlatilag annyit jelent, mintha kisebb lenne az n mintaelemszám. Márpedig a legtöbb becslésnél n csökkenése $1/\sqrt{n}$ szerint növeli a becslés bizonytalanságát.

Valami kompromisszumra tehát mindenképpen szükség van. Ha a két fenti szempontot egymáshoz viszonyítva súlyozni kellene, a *jó* pontok figyelembevételének szempontja kapná a nagyobb súlyt. Milyen kifejezéssel célszerű a (7) szerinti becslés végrehajtásakor effektíve szerepet játszó pontszámot figyelembe venni? A (7)-ben szereplő $\varphi(x)$ súlyok az M közelében közelítően *1* értékűek, az M-től legtávolabb levő, durva hibával terhelt adatok közelítően zérus értékű súlyt kapnak. Az M-től távolabbi, de még jól láthatóan a tömörülő pontcsoporthoz tartozó adatok pedig az M-től való távolsággal az (5) súly szerint egyre csökkenő mértékben járulnak hozzá M értékének kialakításához. Kézenfekvő tehát a súlyok összegét effektív adatszámnak tekinteni:

$$n_{\mathrm{eff}}(\varepsilon) = \sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M)^2}. \tag{9}$$

Az effektív adatszám maximalizálása – az ε-nak ehhez viszonyítva másodlagos, de határozottan egyidejű minimalizálásával – akkor valósul meg, ha a következő kifejezést maximalizáljuk:

$$\frac{n_{\mathrm{eff}}^2(\varepsilon)}{\varepsilon}. \tag{10a}$$

Ez azzal ekvivalens, hogy az ε skálaparamétert a következő extrémum-követelménnyel definiáljuk:

$$\frac{1}{\varepsilon}\sum_{i=1}^{n}\frac{\varepsilon^2}{\varepsilon^2+(x-M)^2}=\text{maximum}. \tag{10}$$

A kohézió és a dihézió fogalma. A (10) alapján meghatározott ε skálaparamétert elsődlegesen az (5) szerinti $\varphi(x)$ súlyfüggvény skálaparaméterének tekintjük. Nagy n-nél azonban nyilvánvalóan fennáll az

$$\frac{n_{\text{eff}}(\varepsilon)}{n}\approx\int_{-\infty}^{\infty}\frac{\varepsilon^2}{\varepsilon^2+(x-M)^2}f(x)\,dx\equiv n(\varepsilon) \tag{11}$$

összefüggés, ezért az

$$\frac{n^2(\varepsilon)}{\varepsilon}=\text{maximum} \tag{12}$$

követelést teljesítő ε-t magára az $f(x)$ eloszlásra jellemző skálaparaméterként is felfoghatjuk. Mivel a (12) kifejezés négyzetgyökének is azonos helyen van a maximuma, ezért a (12) felírható a következőképpen is:

$$\int_{-\infty}^{\infty}\frac{\varepsilon^{3/2}}{\varepsilon^2+(x_i-M)^2}f(x)dx=\text{maximum}. \tag{13}$$

A fenti követelést teljesítő ε a pontok zömének kohéziós tendenciáját jellemzi. Nagy ε kis kohéziót jellemez és fordítva. Célszerű tehát a kohéziót (χ) a következő módon definiálni (ld. *Steiner 1973*):

$$\varkappa=\frac{1}{\varepsilon}. \tag{14}$$

Az ε skálaparaméter tehát reciprok kohéziónak is nevezhető, de az elnevezés nehézkessége miatt az (önkényesen képzett) dihézió megnevezést használjuk.

Az ε gyakorlai számítása. Az M-et továbbra is ismertnek feltételezve, az ε meghatározása a feladat. Elméletileg kimutatható, hogy ε nem lehet nagyobb a mintaterjedelem $\sqrt{3}/2$-szeresénél (ld. *Csernyák, Steiner 1980*):

$$\varepsilon\geq\frac{\sqrt{3}}{2}[\max(x_i)-\min(x_i)]. \tag{15}$$

Tehát ε meghatározható úgy, hogy a fenti kifejezés jobb oldalából kiindulva addig csökkentjük ε értékét, amíg a (13)-mal analóg,

$$\sum_{i=1}^{n}\frac{\varepsilon^{3/2}}{\varepsilon^2+(x_i-M)^2} \tag{13a}$$

kifejezés maximumát el nem érjük. Ezt a maximumhelyet használjuk a $\varphi(x)$ súlyfüggvény skálaparamétereként és egyben ez az ε ad becslést az eloszlás dihéziójára is.

Az előbbi eljárás nemcsak elvi lehetőség, gyakorlatilag is alkalmazható, de – mint az extrémumhelykereső algoritmusok általában – eléggé lassú. Gyorsabban kapjuk meg a súlyfüggvény skálaparaméterét, ha a következő iterációs algoritmust alkalmazzuk:

$$\varepsilon_{k+1}^2 = \frac{3\sum_{i=1}^{n} \frac{\varepsilon_k^4 (x_i - M)^2}{[\varepsilon_k^2 + (x_i - M)^2]^2}}{\sum_{i=1}^{n} \frac{\varepsilon_k^4}{[\varepsilon_k^2 + (x_i - M)^2]^2}} \tag{13b}$$

A fenti formulával meghatározott dihézió a valószínűségi változó ingadozásának egyik mérőszáma. Amennyiben műszereink, a mérési körülmények stb. hibaviszonyai ismertek, a kettős iteráció ε ágát elhagyhatjuk és csak az M ágat futtatjuk konstans – az ismert fizikai körülményekből meghatározott – ε értékkel.

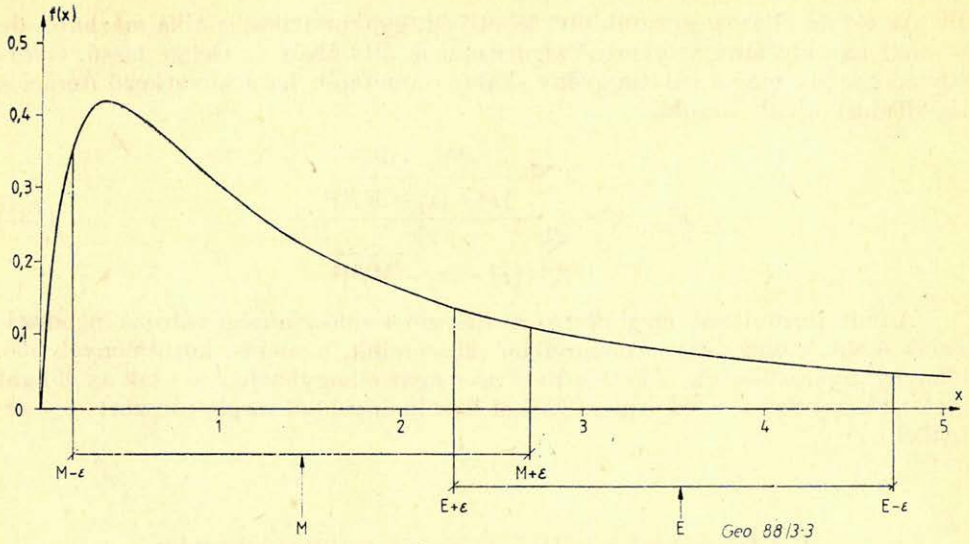
3. A leggyakoribb érték és a dihézió együttes számítása

Anyaeloszlások leggyakoribb értékét a (8) formula csak abban az esetben szolgáltatja egyértelműen, ha az abban szereplő ε is egyértelműen rögzített. Az előző *2.* pont után nyilvánvaló, hogy a (13)-at kielégítő ε-t fogjuk a leggyakoribb érték meghatározásánál alkalmazni és viszont (mindkét számításnál most már elvetjük a gyakorlatban szinte sohasem teljesülő, de a fogalmak bevezetéséhez elengedhetetlen *eleve ismert* ε *eleve ismert* M feltételezéseket). Az M meghatározásához tehát nélkülözhetetlen az ε meghatározása, akár érdekel bennünket, akár nem. Mivel csak az anyaeloszlásra jellemző értékpár határozható meg, a leggyakoribb érték végleges definíciója tartalmazza a dihézió definícióját is. Tehát *$f(x)$ sűrűségfüggvényükkel adott valószínűség-eloszlások leggyakoribb értékének (M) és dihéziójának (ε) azt az M-et és ε-t nevezzük, amelyek a (8) és a (13) feltételt egyidejűleg teljesítik.*

A fenti definíció egyben meghatározási algoritmusnak is tekinthető, mivel az M és ε a (8)-cal, illetve a (13)-mal definiált kettős iterációval számítható, ε-t $3{,}1 \cdot Q$-ról indítva. (Ui. a dihézió és az interkvartilis félterjedelem között fennáll a következő reláció: $\varepsilon \leq 3{,}092 \cdot Q$).

Mérési adatok leggyakoribb értékére és dihéziójára vonatkozó becsléseket – az előzőekkel analóg módon – a (7a) és a (13b) összefüggések (összeg formulák) kettős iterációs algoritmusával határozzuk meg. M induló értékeként a mintaelemek számtani közepét választhatjuk, míg ε-t célszerű a (15) jobb oldala szerint indítani. A kettős iteráció M, illetve ε ágán felváltva végezzük a számításokat.

Megjegyezzük, hogy az anyaeloszláshoz tartozó – az előzőekben definiált – M helyparaméter különbözik a valószínűségszámítás elméletéből ismert és leggyakrabban használt helyparaméterektől (várható érték, medián). M ugyanis az adatok koncentrálódási helyét hivatott megadni. A koncentrálódás mértékére a dihézió a jellemző. Az (M, ε) értékpár nagyon szemléletesen tájékoztat bennünket az anyaeloszlásról: a valószínűségi változó viszonylag nagy valószínűséggel vesz fel értékeket az ($M - \varepsilon$, $M + \varepsilon$) intervallumban. Nem mondható ez el az E-vel jelölt várható értékről aszimmetrikus eloszlások esetében, amikor is a *várható* jelző vagy megkérdőjelezhető, vagy teljesen jogosulatlan (ld. *3. ábra*). (Az E =



3. ábra. Az x értékeknek az $(M-\varepsilon, M+\varepsilon)$ intervallumba esése sokkal valószínűbb aszimmetrikus eloszlásoknál, mint az E „várható" érték körüli, ugyanolyan hosszúságú intervallumbeli előfordulása

Рис. 3. Попадание значений x в интервалы $(M-\varepsilon, M\%\varepsilon)$ более вероятно при асимметричных распределениях, чем встречаемость в пределах интервалов такой же длины, находящихся около «ожидаемого» значения E.

Fig. 3. The occurrence of x in the interval $(M-\varepsilon, M+\varepsilon)$ has much greater probability than the occurrence in $(E-\varepsilon, E+\varepsilon)$

$= \int_{-\infty}^{\infty} xf(x)dx$ helyparaméter becslésére, mint ismeretes, a számtani átlagot használjuk.)

4. A leggyakoribb érték szerinti kiegyenlítés (M-kiegyenlítés)

Ha a leggyakoribb értékek alapgondolatát megtartva akarunk többváltozós kiegyenlítésre vonatkozó algoritmust megadni – mégpedig olyat, melyhez kész könyvtári programokat tudunk felhasználni, – akkor célszerű közvetlenül az M meghatározásának (7a) formulájából kiindulni. Ez az egyenlet ugyanis iterációs lépésenként eleget tesz a következő minimumfeltevésnek (mely differenciálással igazolható):

$$\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + (x_i - M_{j-1})^2} (x_i - M_j)^2 = \min, \tag{16}$$

ahol M_{j-1} az előző iterációs lépésben meghatározott érték, ε pedig a másik iterációs ágon kapott eredmény. Látható, hogy a (16) kifejezés csak a leggyakoribb érték szerinti súllyal különbözik a hagyományos, legkisebb négyzetes kiegyenlítés jól ismert minimumfeltételétől.

Az M helyébe adott analitikus alakú $T(\mathbf{p}, \mathbf{y}_i)$ függvényt helyettesítve (ahol $\mathbf{p}$ a T függvény paraméter vektorát, $\mathbf{y}$ a T függvény független-változó vektorát

jelenti), az M szerinti kiegyenlítés minden iterációs lépése $\mathbf{p}_k$-ra az alábbi minimumfeltétel kielégítését kívánja meg:

$$\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + [x_i - T(\mathbf{p}_{k-1};\, \mathbf{y}_i)]^2} [x_i - T(\mathbf{p}_k;\, \mathbf{y}_i)]^2 = \min, \tag{17}$$

ahol $\mathbf{p}_k$ az *új* paramétervektor, a súlyban szereplő $\mathbf{p}_{k-1}$ pedig az előző iterációs lépés eredményeként kapott, tehát ismert érték vektora.

A T analitikus alakjának természetesen itt is ugyanúgy adottnak kell lenni, mint a hagyományos kiegyenlítésnél. Ha ez a T-függvény ($J = 1$) fokszámú polinom, vagy általánosabban:

$$T(\mathbf{p};\, \mathbf{y}) = p_1 T_1(\mathbf{y}) + p_2 T_2(\mathbf{y}) + \ldots p_J T_J(\mathbf{y}), \tag{18}$$

ismert analitikus alakú T_j-kel, akkor a (17) feltétel teljesítése, azaz a leggyakoribb érték szerinti kiegyenlítés egy-egy iterációs lépése csak lineáris algebrai egyenletrendszer megoldását igényli:

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + [x_i - T(\mathbf{p}_{k-1};\, \mathbf{y}_i)]^2} T_1(\mathbf{y}_i)[p_{k1} T_1(\mathbf{y}_i) + p_{k2} T_2(\mathbf{y}_i) + \ldots + p_{kJ} T_J(\mathbf{y}_i) - x_i] = 0 \\
&\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + [x_i - T(\mathbf{p}_{k-1};\, \mathbf{y}_i)]^2} T_2(\mathbf{y}_i)[p_{k1} T_1(\mathbf{y}_i) + p_{k2} T_2(\mathbf{y}_i) + \ldots + p_{kJ} T_J(\mathbf{y}_i) - x_i] = 0 \\
&\qquad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\sum_{i=1}^{n} \frac{\varepsilon^2}{\varepsilon^2 + [x_i - T(\mathbf{p}_{k-1};\, \mathbf{y}_i)]^2} T_J(\mathbf{y}_i)[p_{k1} T_1(\mathbf{y}_i) + p_{k2} T_2(\mathbf{y}_i) + \ldots + p_{kJ} T_J(\mathbf{y}_i) - x_i] = 0
\end{aligned} \tag{19}$$

A súlyfüggvény skálaparaméterét szintén iterációs formulával nyerjük:

$$\varepsilon_r = \frac{3 \sum_{i=1}^{n} \frac{\varepsilon_{r-1}^4 [x_i - T(\mathbf{p};\, \mathbf{y}_i)]^2}{\{\varepsilon_{r-1}^2 + [x_i - T(\mathbf{p};\, \mathbf{y}_i)]^2\}^2}}{\sum_{i=1}^{n} \frac{\varepsilon_{r-1}^4}{\{\varepsilon_{r-1}^2 + [x_i - T(\mathbf{p};\, \mathbf{y}_i)]^2\}^2}}, \tag{20}$$

ahol $\mathbf{p}$ az utolsó iteráció eredményeként ismert paramétervektor.

A két iterációs ágban (hasonlóan a *3*-ban leírtakhoz) itt is felváltva végezzük a számításokat.

Azonos x_i adatrendszer alapján, azonos analitikus alakú függvény $\mathbf{p}$ paramétervektorát kereshetjük a legkisebb négyzetes elv:

$$\sum_{i=1}^{n} [x_i - T(\mathbf{p};\, \mathbf{y}_i)]^2 = \min \tag{21}$$

szerint is, valamint az előzőekben vázolt leggyakoribb érték szerint is. Általában nem kapunk azonos eredményt és a jelentős eltérések sem ritkák. Ezért az M-szerinti kiegyenlítéssel kapott eredményt megkülönböztetésül $\mathbf{p}_M$-el, míg a (21)-ből nyerhetőt $\mathbf{p}_E$-vel jelöljük. Ehhez csatlakozva az utóbbi módon definiált kiegyenlítést várható érték szerinti kiegyenlítésnek, vagy röviden E-kiegyenlítésnek nevezzük.

5. Az $f_a(x)$ modelleloszlás-család

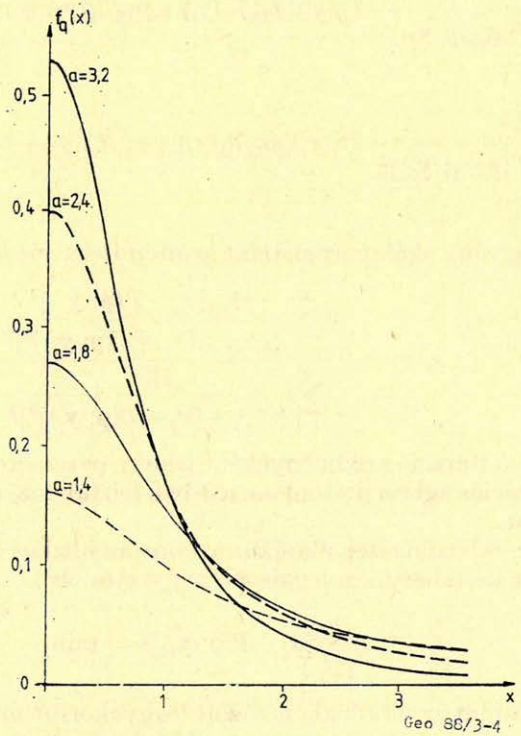
Definiáljuk az a paraméterrel jellemzett eloszlástípus-családot a következőképpen:

$$f_a(x) = \frac{1}{c(a)[\sqrt{a^2+1}]^a} \quad a>1, \tag{22}$$

ahol

$$c(a) = \frac{\sqrt{\pi}\Gamma\left(\frac{a-1}{2}\right)}{\Gamma\left(\frac{a}{2}\right)}, \tag{22a}$$

melyben Γ a jól ismert gamma-függvény. Az origóra szimmetrikus (22) eloszlások



4. ábra. Négy modelleloszlás sűrűségfüggvénye az $f_a(x)$ modellcsaládból

Рис. 4. Четыре функции плотности распределения группы моделей $f_a(x)$

Fig. 4. Four density functions from the supermodel $f_a(x)$

közül néhányat (pozitív x-ekre) a *4. ábrán* mutatunk be. Az a (>1) típusparaméter egész értékeire az eloszlásfüggvények az alábbiak:

$$F_2(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arc\,tg} x;$$

$$F_3(x) = \frac{1}{2} + \frac{x}{2\sqrt{1+x^2}};$$

$$F_4(x) = F_2(x) + \frac{1}{\pi} \frac{x}{1+x^2}; \tag{22b}$$

$$F_a(x) = F_{a-2}(x) + \frac{1 \cdot 3 \cdot \ldots \cdot (a-4)}{2 \cdot 4 \cdot \ldots \cdot (a-3)} \frac{1}{2} \frac{x}{(1-x^2)^{\frac{a-2}{2}}}$$

(ha $a>5$ és páratlan);

$$F_a(x) = F_{a-2}(x) + \frac{2 \cdot 4 \cdot \ldots \cdot (a-4)}{1 \cdot 3 \cdot \ldots \cdot (a-3)} \frac{1}{\pi} \frac{x}{(1-x^2)^{\frac{a-2}{2}}}$$

(ha $a>6$ és páros).

A (22) kifejezés az eloszlástípusok sűrűségfüggvényeit a lehető legegyszerűbb alakban adja. Ha a helyparaméter (jelen esetben szimmetriapont) és a skálaparaméter 1. részben közölt általános jelöléseire, T-re, illetve S-re visszatérünk, akkor $f_a(x)$ a következő általános alakot veszi fel:

$$g_a(x) = \frac{\Gamma\left(\frac{a}{2}\right)}{\Gamma\left(\frac{a-1}{2}\right)\sqrt{\pi}} \cdot \frac{S^{a-1}}{[\sqrt{S^2+(x-T)^2}]^a} \tag{22c}$$

Látjuk, hogy a (22) a $T = 0$ és $S = 1$-nek megfelelő standardizált alaknak felel meg.

Az $f_a(x)$-eloszláscsalád és becslések pontosságára döntően kiható szárnyakat igen széles spektrumban tudja modellezni (lásd *4. ábra*). Ha a értékével felülről közeledünk *1*-hez, akkor egyre közelebb jutunk az olyan lassú (*konstans*/$|x|$ szerinti) aszimptótikus csökkenéshez, amely már nem jellemezhet (sűrűségfüggvény-értelemben) valószínűségi változót. A nagy szárnyak tartományában tehát elmegyünk a legszélső határig. A másik irányban haladva, azaz $a \to \infty$-re a (22) eloszlások egyre jobban közelítik a Gauss-eloszlást, amelyet a gyakorlatban előforduló anyaeloszlások megközelíthetnek, kivételesen el is érhetnek. Az $f_a(x)$-család tehát széles lehetőséget nyújt számunkra, hogy a valóságban előforduló, különböző eloszlástípus-tartományt modellezni tudjuk.

6. Általános leggyakoribb érték

Az általános leggyakoribb érték és jelölése (M_k) onnan származik, hogy azt a

$$M_k = \frac{\sum_{i=1}^{n} \frac{(k\varepsilon)^2 \cdot x_i}{(k\varepsilon)^2 + (x_i - M_k)^2}}{\sum_{i=1}^{n} \frac{(k\varepsilon)^2}{(k\varepsilon)^2 + (x_i - M_k)^2}} \tag{23}$$

kifejezésnek eleget tevő értékkel becsüljük, ahol k értéke az a ismeretében kiszámítható. A $k(a)$ függvény empirikus alakja a következő:

$$k(a) = \sqrt{a-1} + \frac{1}{2\pi} \cdot \frac{\sqrt{a} - \sqrt{2}}{\sqrt{3}}; \qquad \begin{matrix} a \geq 2 \\ k \geq 1. \end{matrix} \tag{24}$$

Ha tehát tudjuk, hogy aktuális eloszlásunk típusa milyen a-jú $f_a(x)$ közelében várható, akkor a $k(a)$ függvény szerinti k-t használva, (23)-at és az

$$\varepsilon^2 = \frac{3 \sum_{i=1}^{n} \frac{(x_i - M_k)^2}{[\varepsilon^2 + (x_i - M_k)^2]^2}}{\sum_{i=1}^{n} \frac{1}{[\varepsilon^2 + (x_i - M_k)^2]^2}} \tag{25}$$

összefüggést használjuk a kettős iterációs algoritmushoz. Az általános leggyakoribb érték számítási programja tehát egyetlen utasításban különbözik csak a szűkebb értelemben vett leggyakoribb értékszámítás programjától. Azaz k-val kell csak megszorozni az ε-ágon kapott eredményt, mielőtt a kettős iteráció M-ágára lépnénk.

Kimutatható (ld. *Hajagos 1985*), hogy ha az x-ek $f_a(x)$ eloszlást követnek és a k értékét a (24) kifejezés szerint választjuk meg, akkor a becslés hatásfoka *100%*-os. (Hatásfok alatt a matematikai statisztikában a minimális aszimptótikus szórásnégyzet és az aktuálisan alkalmazott becsléshez tartozó aszimptótikus szórásnégyzet hányadosát értjük.) Általános leggyakoribb érték számítása esetében és origóra szimmetrikus $f(x)$-ekre az aszimptótikus szórásnégyzetet a következő kifejezés szolgáltatja:

$$A^2(M_k, \varepsilon) = \frac{\int_{-\infty}^{\infty} \frac{x^2}{[(k\varepsilon)^2 + x^2]^2}}{\left\{ \int_{-\infty}^{\infty} \frac{(k\varepsilon)^2 - x^2}{(k\varepsilon)^2 + x^2} f(x)dx \right\}^2}. \tag{26}$$

A leggyakoribb érték szerinti kiegyenlítést is általánosíthatjuk úgy, hogy iterációs lépésenként (19)-ben a

$$\frac{(k\varepsilon)^2}{(k\varepsilon)^2 + [x_i - T(\mathbf{p}_{k-1}; \mathbf{y}_i]^2} \tag{27}$$

súlyokat alkalmazunk az ott szereplő

$$\frac{\varepsilon^2}{\varepsilon^2 + [x_i - T(\mathbf{p}_{k-1}; \mathbf{y}_i)]^2}$$

súlyok helyett, míg ε-t változatlanul a (20) formulából határozzuk meg.

IRODALOM

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

Csernyák, L., Steiner, F. (1980): Practical computation of the most frequent value of data systems. Acta Geodaet., Geophys. et Mont. Acad. Sci. Hung. *15* (1)

Hajagos, B. (1985): Die verallgemeinerten Student-schen t-Verteilungen und die häufigsten Werte. Publications of the Technical University for Heavy Industry, Series A, Mining, *40* (1–4) pp. 225–238.

A leggyakoribb értékek módszere és alkalmazása a karotázs-interpretációban

FERENCZY LÁSZLÓ–STEINER FERENC**

A mélyfúrási geofizikai értelmezés különböző munkafázisaiban alkalmazott matematikai statisztikai módszerek szokásos alapfeltevése, hogy valamely függvénykapcsolattal meghatározott tároló- vagy kőzetparaméter értékének a jellemző értéktől való eltérése Gauss-eloszlást követ. Gyakorlati példák sora mutatja, hogy még az azonos „zónába" tartozó, homogénnek tekintett térrész paraméterértékeinek eloszlása is az esetek túlnyomó többségében lényegesen eltér ettől. Ezért célszerű olyan értelmező algoritmus használata, amely bármilyen eloszlással jellemzett valószínűségi változót kezelni tud.

A tanulmány első részében a leggyakoribb értékek módszerét ismertetjük, bemutatva a módszer fontos, általános jellemzőit. Az elvi előnyök mellett a számítástechnikaiakra is kitérünk. A tanulmány második részében a módszer mélyfúrási geofizikai alkalmazását mutatjuk be.

Основным принципом методов математической статистики, применяемых на различных стадиях интерпретации данных геофизических хисследований глубоких скважин является то, что отличия между определяемыми параметрами коллектора и породы и характерными значениями подчиняются распределению Гаусса. Однако, как показывает практика, распределение значений параметров части пространства, относящегося к одной и той же «зоне» и принятого гомогенным, в подавляющем большинстве случаев в значительной степени отличаются от этого распределения. Поэтому целесообразным является использование такого интерпретирующего алгоритма, который может быть применен в случае случайных величин с любым распределением.

В первой части работы рассматривается метод наиболее частых значений и приводятся наиболее важные характеристики этого метода. Во второй части работы описаны возможности применения метода при геофизическом каротаже глубоких скважин.

It is conventional in different stages of the well-log interpretation to assume Gaussian distribution for any kind of errors or fluctuations. Adequate investigations on practical examples contradict this assumption and therefore one should prefer statistical procedures which can manage advantageously the real distributions. — In the first part of the paper is the essence of the method of the most frequent values presented, the second part recommends its use in the well-log interpretation emphasizing some points of view. — Some possibilities of the applications: (1) determination of "zone-parameters" on ground of data coming from logs and/or core samples; (2) statistical investigation of probability frequency curves and crossplots; (3) specifications of functional dependencies (determination of core-core, log-core and log-log connections); (4) determination of the reservoir characteristics and of the rock composition; (5) qualification of the results.

Szokásos a karotázs-interpretáció különböző fázisaiban bármely hiba vagy fluktuáció Gauss-eloszlását feltételezni. Gyakorlati példák sora bizonyítja azonban, hogy valamely kőzetparaméter értékei, még ha azokat valamely térrész homogénnak minősíthető kőzetanyagán mértük is, a legtöbb esetben ettől lényegesen eltérő valószínűségeloszlást mutatnak. Előnyös ezért olyan értelmező eljárásokat alkalmazni, amelyek a valóságban előforduló eloszlásokat nagy hatásfokkal tudják kezelni.

A matematikai statisztika tudománytörténetében (lásd I. vázlat) hosszú ideig állandó volt az a feltevés (mely az adott számítástechnikai lehetőségek mellett a hatékonyság optimuma szempontjából indokolt volt), hogy a hibaeloszlások Gauss-eloszlást követnek. Ez a nézet megmerevedett és sajnos úgy él a köz-

* NME Geofizikai Tanszék

I. vázlat

. Időtáblázat
a matematikai statisztika egyes fontosabb eseményeiről, folyamatairól és megállapításairól

A matematikai statisztika fejlődésének néhány, a hatékony algoritmusok szempontjából fontos időpontja ill. szakasza:

Az ezerhétszázas években a maximum likelihood-elv első megjelenése (D. Bernoulli).
Ezt az elvet később Gauss is alkalmazta, de csak másfél évszázad múlva vált nyilvánvalóvá a sokirányú használhatóság

A XVIII-XIX. század fordulójától, Legendre és Gauss munkásságával indul a legkisebb négyzetes elv gyakorlati alkalmazása. Másfél évszázadon át számos kutató fejlesztette tovább speciális szakterületek problémáinak megoldására

Fisher névadó cikkében (újra) kimondja a maximum likelihood-elvet (1912-ben). Cramér és Dugué csatlakozó munkáiból derül ki, hogy ezzel az elvvel adott f(x) sűrűségfüggvényű eloszláshoz képezhető nagyon általános esetben is 100 %-os hatásfokú algoritmus

Az információveszteség (I-divergencia) minimalizálásának módszere a valóság legjobb modellezéséhez vezet; a gyakorlati algoritmusok részben megegyeznek azokkal, amelyek a maximum likelihood-elvből következnek (Kullback, Perez, Rényi)

1778 1912 1960

1700 1800 1900 2000

1812 1886 1938 1977

A valóságban milyen hibaeloszlások várhatók és milyenek nem? A választ illetően csak a gyakorlati problémákkal és a matematikai statisztikával egyaránt foglalkozók tekinthetők kompetensnek

Gauss (saját maga által is fordítottnak minősített) gondolatmenete: ha a számtani átlag a legjobb algoritmus, akkor a hibák $f_G(x)$-eloszlásúak (Gauss-eloszlás, ún. "normális eloszlás")

Newcomb a Gauss-eloszlásról: "egészen kivételesek azok az esetek, amikor úgy találjuk, hogy a hibák valóban követik ezt a törvényt"

Jeffreys: a leggondosabb, azonos körülmények között végzett mérések is általában legfeljebb 5-9 szabadságfokkal jellemzett Student-eloszlású adatsort eredményeznek (és nem Gauss-eloszlást)

Mosteller és Tukey: vannak, akik félreértik a "normális" szót, "rendszerint előfordulót" értve alatta, de, amennyire mi tudjuk, olyan eloszlások, amelyek pontosan követik ezt a formulát, sohasem fordulnak elő a gyakorlatban

A mérés+számítás komplexumra vonatkozó optimális hatékonyság szemszögéből mennyire tekinthető jogosnak a Gauss-eloszlás feltevése? A helyes válasz változása az idő függvényében:

GAUSS GAUSS GAUSS GAUSS GAUSS A VALÓSÁG ADEKVÁT MODELLJE

Az esetek többségében a Gauss-eloszlás feltevése (a legkisebb négyzetes elv) szolgáltatja az optimális hatékonyságot, annak ellenére, hogy a valóságban leggyakrabban előforduló eloszlásokra a legkisebb négyzetes algoritmusok matematikai statisztikai értelemben gyakran kis hatásfokúak

Az esetek egyre növekvő százalékában a valóságos eloszlásnak megfelelő, nagy hatásfokú algoritmusokkal érhető el a hatékonyság optimuma; a Gauss-eloszlás feltevése egyre gyakrabban jelent megengedhetetlenül alacsony hatékonyságot

Új szempontok, fogalmak és tennivalók a matematikai statisztikában:

kb 1960-tól

Robusztus és rezisztens módszerek kifejlesztése, a kártékonnyá vált dogmák felszámolásának megkezdése (Tukey, Huber)

tudatban, hogy az ennek ellentmondó megállapításokat, melyek a szakirodalomban egyre gyakrabban olvashatók, bizonyos kétkedéssel fogadja. Az *I. vázlat* idézetei azonban egyértelműen mutatják, hogy valóságos adatrendszerekkel és a matematikai statisztikával egyaránt foglalkozó szakemberek már régóta ismerik a gyakorlatban előforduló eloszlások szignifikáns eltérését a Gauss félétől.

Egy általános módszertani áttekintéshez elengedhetetlen az, hogy visszanyúljunk a fő statisztikai alapelvekhez, ú.m. a jól ismert (és *1912*-ben kimondott) maximum likelihood elvhez, valamint az ún. I-divergencia minimalizálásának elvéhez (amely egy fél évszázaddal későbbi). Ez a viszonylag új elv számunkra sokkal inkább elfogadható, hiszen általában nem ismerjük pontosan az eloszlás típusát, s így csak modellezhetjük azt, — és bizonyára elfogadjuk az információveszteségként felfogott I-divergencia minimalizálására vonatkozó követelést.

A *II/a. vázlat* szemléletesen mutatja, hogy harang alakú hibaeloszlás esetén (a Gauss-görbe csak egyetlen a végtelen sok alternatíva közül,) mindkét elv a súlyozott átlagok iteratív számítását írja elő az $x_1, \ldots, x_i, \ldots, x_n$ minta alapján legindokoltabban elfogadható T meghatározására.

Matematikai statisztikai elvek és meghatározási módszerek kapcsolata; hasonlóságok és különbségek (I)

II/a. vázlat

Elvi kiindulás (a módszerek alapgondolata):

A MAXIMUM LIKELIHOOD-ELV

Tudjuk, hogy f(x) az aktuális eloszlás sűrűségfüggvénye; az $x_1, \ldots, x_i, \ldots, x_n$ mért értékek (azaz a minta) alapján azt fogadjuk el helyes értéknek, amellyel számolva a minta maximális valószínűségű

(néhány ismert logikai lépés)

Az alapelv szerint az a helyes T-érték, amely kielégíti a következő egyenletet:

$$\sum_{i=1}^{n} \left(\frac{\frac{\partial f(x_i;T)}{\partial T}}{f(x_i;T)} \right) = 0$$

AZ I-DIVERGENCIA MINIMALIZÁLÁSA

Az ismeretlen f(x) sűrűségfüggvényű eloszlást egy adott analitikus alakú g(x) eloszlással helyettesítjük (modellezzük); az információveszteséget az ún. I-divergenciával mérve, a minta alapján azt fogadjuk helyes értéknek, amellyel az információveszteség minimális

(egyetlen differenciálás)

$$\sum_{i=1}^{n} \left(\frac{\frac{\partial g(x_i;T)}{\partial T}}{g(x_i;T)} \right) = 0$$

Ha a helyettesítő eloszlás típusa azonos az aktuális eloszlás típusával (azaz g=f), a két alapelv a T meghatározására azonos számítási algoritmust ír elő. (Az információveszteség minimalizálásának követelménye az S skálaparaméter meghatározására már általában eltérő algoritmusra vezet. A gyakorlatban T-t és S-et együtt határozzuk meg, így a teljes eljárás nem azonos a két esetben: a maximum likelihood-elv nem mindig minimalizálja az információveszteséget. - Egyszerűség kedvéért a továbbiakban S ismert volták tételezzük fel.)

Ha a modelleloszlás sűrűségfüggvényét így írhatjuk:

$$g\left(\left[\frac{x-T}{S}\right]^2\right)$$

(amivel nyilván szimmetrikusnak feltételeztük a hibák eloszlását, - ez, speciális esetektől eltekintve, megtehető), - akkor a T-t definiáló fenti egyenlet a

$$\varphi(x_i-T) = g'\left(\left[\frac{x-T}{S}\right]^2\right)/g\left(\left[\frac{x-T}{S}\right]^2\right)$$

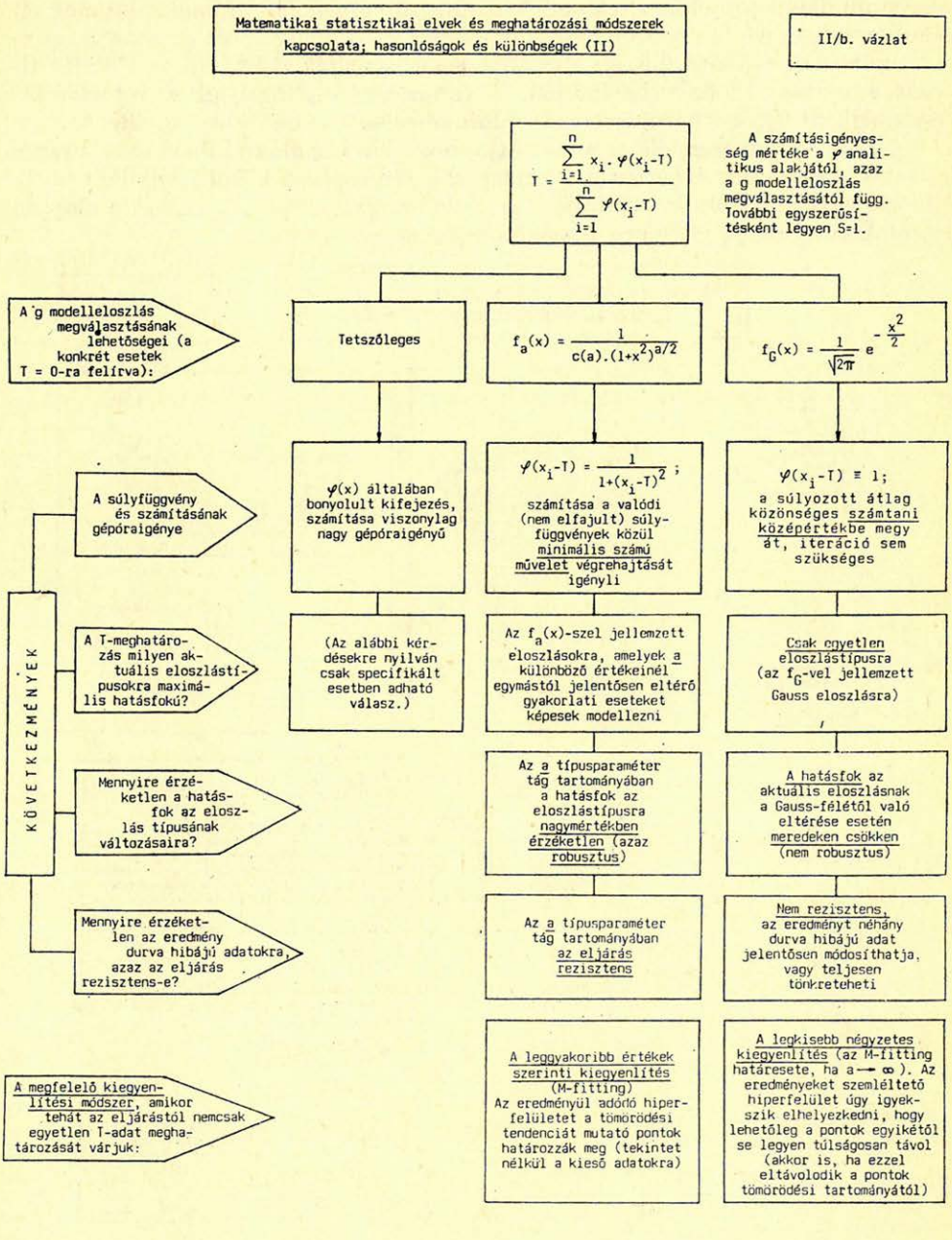
jelolessel nyilván

$$T = \frac{\sum_{i=1}^{n} x_i \cdot \varphi(x_i-T)}{\sum_{i=1}^{n} \varphi(x_i-T)}$$

alakú lesz, amit iterációs algoritmusként kell értelmeznünk.

Az alapelvek gyakorlatilag súlyozott átlagképzés iteratív végrehajtását írják elő:

A g helyettesítő eloszlás választásától függ, hogy az a gépóra-igény túl sok, elfogadható vagy nagyon csekély lesz-e, amely a fenti algoritmus végrehajtásához szükséges: a *II/a. vázlat* mutatja, hogy a súlyok (a φ-értékek) g'/g-ként számítandók. Ha g tetszőleges, ez arra vezethet, hogy túl sok műveletet kell végrehajtanunk. – Ha nem adjuk fel azt a törekvésünket, hogy a T meghatározásának (általános esetben: a többismeretlenes kiegyenlítésnek) nagy legyen a hatás-



foka a valóságban előforduló hibaeloszlások széles spektrumára, akkor a leggyakoribb értékek szerinti kiegyenlítésnek a legkisebb a gépidőigénye (ez a módszer a $g(x) = c \cdot (1+x^2)^{-a/2}$ választásnak felel meg; ld. a *II/b. vázlat*). — De még ez a gépidő is kb. két nagyságrenddel nagyobb, mint a Gauss-eloszlás feltételezésével adódó módszernél (amikor tehát $g(x) = (2\pi)^{-1/2} \cdot \exp(-x^2/2)$-t választjuk helyettesítő eloszlásnak), hiszen ekkor a súlyok számítása és így az iteráció végrehajtása is feleslegessé válik: ebben az esetben, amint az jól ismert, csak a legkisebb négyzetes kiegyenlítés végrehajtása szükséges (azaz egyszerű számtani átlagot kell csak képezni, ha egyetlen ismeretlenünk van).

A számítások tovább már nem fokozható egyszerűsége (azaz a lehető legrövidebb gépidő) természetesen nem lehet egyetlen figyelembe veendő szempont. Ahogyan az egyre inkább ismertté válik, a legkisebb négyzetes módszernek sok hátránya van: egyrészt néhány durva hibájú adat jelentősen torzíthatja (vagy teljesen tönkre is teheti) az eredményeket, másrészt sok, a valóságban előforduló hibaeloszlásra ennek a hagyományos eljárásnak (statisztikai értelemben véve) nagyon kicsiny a hatásfoka (ld. újra a *II/b. vázlatot*). Egy, mondjuk *50%*-os hatásfok pedig jól ismerten nem jelent kevesebbet, mint azt, hogy kidobtuk drágán mért adataink felét.

Régóta ismert, hogy a valóságban előforduló eloszlások nagyon sokfélék lehetnek. Ennek ellenére a Gauss-eloszlás feltételezése másfél évszázadon keresztül indokolt volt abban az értelemben, hogy a mérések + számítások együttes költsége ekkor adódott minimálisnak azokban az esetekben is, amikor a statisztikai értelemben vett hatásfok csak *50%*, *25%*, vagy akár még ennél is kevesebb volt. (Nota bene: néhány évtizeddel ezelőtt a mechanikus gépekkel végzett műveletek normája műszakonként *400* szorzás osztás vagy *1200* összeadás/kivonás

AZ EREDMÉNYESSÉG ÉS ÁLTALÁNOS GYAKORLATI ALKALMAZHATÓSÁG KRITÉRIUMAI
(Minek kell együtt adottnak lennie az alkalmazó szemszögéből egy matematikai statisztikai eljárásnál ahhoz, hogy az eredményesen és általános legyen alkalmazható?)

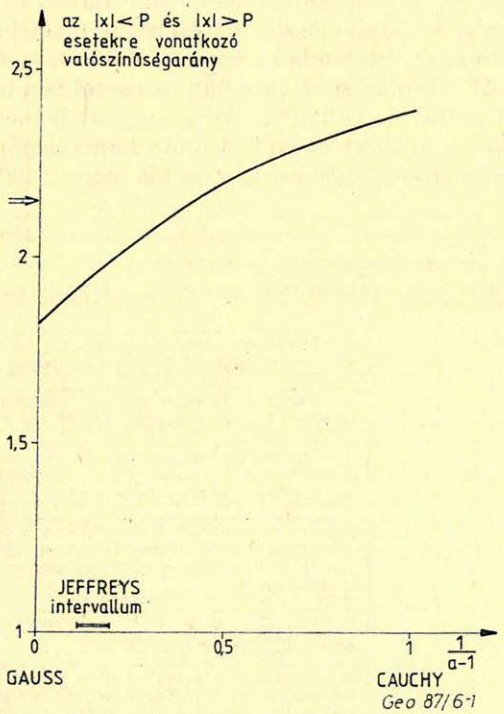
III. vázlat

ÁTTEKINTHETŐSÉG	a statisztikai algoritmus működése heurisztikusan közvetlenül is értelmezhető és követhető legyen;
ELMÉLETI MEGALAPOZOTTSÁG	az algoritmus feleljen meg a matematikai statisztika korszerű elméleti eredményeinek, legyen azokból levezethető;
ÁLTALÁNOSÍTHATÓSÁG	a helyparaméter-meghatározásként definiált statisztikai algoritmus minden további nélkül általánosítható legyen a többváltozós kiegyenlítések eseteire;
ELOSZLÁSMODELL-CSALÁD	álljon rendelkezésre a valóságban előforduló valószínűségeloszlástípusok minél adekvátabb modellezése céljából egy kellően általános, de lehetőleg egyszerűen kezelhető modelleloszlás-család;
NAGY HATÁSFOK	a statisztikai algoritmus legyen minél nagyobb hatásfokú az eloszlásmodell-család tagjaira;
KIS GÉPÓRA-IGÉNY	az algoritmus számítástechnikai szempontból legyen lehetőleg egyszerű, hogy a legfontosabb (pl. a hatásfokra vonatkozó) követelményeket minél kisebb gépidő-ráfordítással elégíthessük ki;
ROBUSZTUSSÁG	az algoritmus hatásfoka legyen elegendően érzéketlen a hibák eloszlástípusának változásaira;
REZISZTENCIA	az algoritmus legyen nagymértékben érzéketlen a kiütő, azaz durva hibával terhelt adatokra, hiszen ezek esetenként előfordulhatnak, és többváltozós kiegyenlítések esetén a szokásos vizuális eliminálás (pontelhagyás) módszere nemcsak gazdaságtalan és szubjektív, de ilyenkor alkalmazhatatlan is.

volt, azaz nagyon drágák voltak az iteratív számítások.) A számítási költségek meredek csökkenése teljesen új szituációt teremtett: a leggyakoribb érték szerinti kiegyenlítés egyrészt minden további nélkül alkalmazható, másrészt a sajátságai megfelelnek azoknak a kritériumoknak, amelyek modern statisztikai módszereknél elengedhetetlenek (ld. a *III. vázlatot*, ahol az alkalmazás szemszögéből lényeges kritériumokat foglaltuk össze).

Ennek a statisztikai módszernek az alkalmazását a karotázs-értelmezésben a következő szempontok kiemelésével ajánljuk:

1. a karotázs-értelmezés adatrendszerei túlnyomó többségükben nem Gauss-eloszlások;
2. a durva hibájú adatok automatikus kiküszöbölésével és a pontatlanabb adatok kisebb súlyú figyelembevételével az eredmények elsősorban a pontos adatokra támaszkodnak és így az eredmények is megbízhatóbbak;
3. a módszer még kevesebb adat esetén is eredményesebb lehet, mint a hagyományos.



1. ábra. Lyukferdeség szelvény feldolgozása a legkisebb négyzetek elve és a leggyakoribb érték szerint

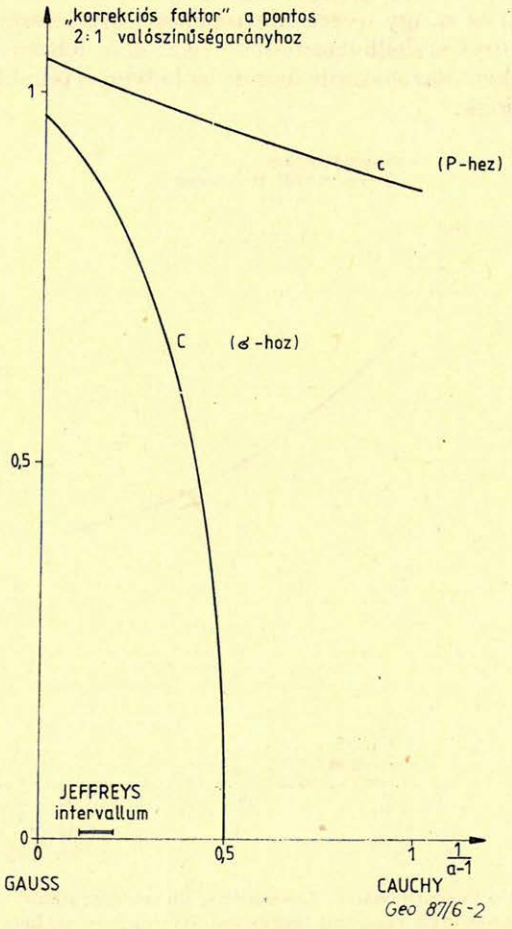
Рис. 1. Обработка данных разреза клинометрии скважины методом наименьших квадратов и наиболее частых значений

Fig. 1. Deviation log smoothed by the least squares technique is heavily influenced by outlying values being in a short depth interval, whilst the result of the smoothing carried out according to the most frequent value technique is insensitive to the outliers

Néhány lehetőség az alkalmazásra a karotázs-értelmezésben:

- a *zónaparaméterek* meghatározása szelvény és/vagy magadatok alapján,
- gyakorisági görbék és crossplot-ok statisztikai vizsgálata;
- függvénykapcsolatok meghatározása (mag-mag, szelvény-mag és szelvény-szelvény relációban);
- tárolóparaméterek és kőzetösszetétel meghatározása;
- az eredmények minősítése.

Végül a leggyakoribb érték szerinti és legkisebb négyzetes kiegyenlítés összehasonlítására három példát mutatunk be.



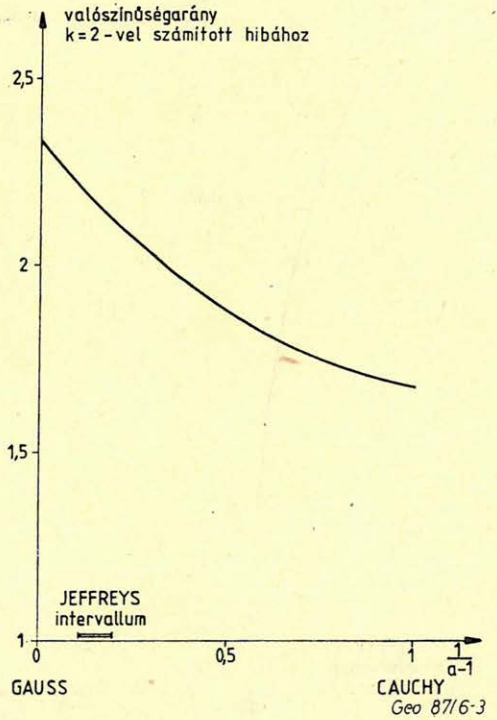
2. ábra. Szimulált szelvényadatok kvantitatív értelmezése a legkisebb négyzetek elve és a leggyakoribb érték szerint

Рис. 2. Количественная интерпретация данных модели разреза методом наименьших квадратов и наиболее частых значений

Fig. 2. Quantitative interpretation of simulated logs carried out according to the least squares principle (dotted lines) and to the most frequent value technique (continuous lines), respectively

Az *1. ábra* ferdeség-szelvényt mutat be egy mélységszakaszra; ilyen gyors változások nem fordulhatnának elő az alkalmazott fúrószár hosszából következtethetően. A javasolt módszerrel kapott spline-függvény gyakorlatilag ugyanaz a görbe, amit egy előítéletmentes értelmező kézzel is berajzolna. A legkisebb négyzetes számítások eredményét azonban kedvezőtlenül befolyásolja egy nagy amplitúdójú fluktuáció. — Az eredmények közötti eltérés feltétlenül szignifikáns.

A karotázs-alkalmazások túlnyomó többsége azonban nem követhető olyan egyszerűen nyomon, mint az iménti spline-példában. A *2. ábra* a két kiegyenlítéssel kapott eredményeket mutatja be arra az esetre, ha (egyszerűség kedvéért) *25%* porozitású (és agyagmentes) víztároló homokra számítunk ki *6*-féle karotázs-szelvényt, hibát szuperponálunk az adatokra (*1%*-os valószínű hibával és *12,7%*-os maximális hibával) és az így nyert adatrendszert értelmezzük (a törmelékes tárolókra szokásos kőzetmodellt használva). A *2. ábra* lehetővé teszi, hogy az így kapott eredményeket összehasonlíthassuk a helyes értékekkel; az új módszer előnyei nyilvánvalóak.



3. ábra. A permeabilitás logaritmusának és szelvény, illetve magadatok kapcsolatának vizsgálata legkisebb négyzetek elve, valamint leggyakoribb érték szerinti hatváltozós másodfokú kiegyenlítéssel

Рис. 3. Исследование зависимости между логарифмом проницаемости и данными разреза или керна методом наименьших квадратов, а также с вторичным выравниванием шести переменных методом наиболее частых значений

Fig. 3. The different degree of the dependence of the logarithm of the permeability upon log and core data of six kind, investigated by two different fitting technique and supposing quadratic dependence

A harmadik példa: a permeabilitás logaritmusát kifejeztük hatféle szelvény-, illetve magadat hatváltozós másodfokú függvényeként. A *3. ábra* mutatja, hogy a valószínű hiba csaknem a felére csökken, ha az ebben az előadásban javasolt módszert használjuk a hagyományos legkisebb négyzetes eljárás helyett.

IRODALOM

Csernyák, L.: On the most frequent value and cohesion of probability distributions. Acta Geodaet., Geophys, et Mont. Acad. Sci. Hung. *8* (3 – 4) 1973.

Ferenczy, L., Takács, E.: Valószínűségelméleti alapokon nyugvó kvantitatív karotázs interpretációs rendszer hatékonyságának és megbízhatóságának növelése. Jelentés. Kézirat. – Miskolc, 1986.

Hajagos, B.: Der häufigste Wert, als eine Abschätzung von minimalem Informationsverlust etc. Publications of the Technical University for Heavy Industry, Series A, Mining, *37* (1 – 2) 1982.

Hursán, L., Takács, E.: A lyukferdeségmérések kiértékelésének új lehetőségei. Jelentés. Kézirat. – Miskolc, 1986.

Jeffreys, H.: An alternative to the rejection of observations. Proceedings of the Royal Soc. of London Ser. A. *137* 1932.

Jeffreys, H.: Theory of Probability. Clarendon Press, Oxford, 1961.

Newcomb, S.: A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics, *8*, 1986.

Short, J.: Second paper concerning the parallax of the sun etc. Philos. Trans. Roy. Soc. London, *53* 1763.

Steiner, F.: Most frequent value and cohesion of probability distributions. Acta Geodaet., Geophys. et Mont. Acad. Sci. Hung. *8* (3 – 4) 1973.

Steiner, F.: Robusztus becslések. Bp., Tankönyvkiadó, 1985.

A hagyományos hibadefiníció fogyatékosságai. Javaslat új hibadefiníció alkalmazására

FERENCZY LÁSZLÓ–HAJAGOS BÉLA*–STEINER FERENC**

A geofizikai adatrendszerek feldolgozása és értelmezése gyakran hagyományos matematikai statisztikai eljárások alkalmazásával történik, amelyek gyakorlatilag az adatok Gauss-eloszlását tételezik fel. Geofizikai, geológiai és csillagászati adathalmazokon végzett vizsgálataink során azonban azt tapasztaltuk, hogy az eloszlások az esetek túlnyomó többségében a Gausstól szignifikánsan eltérnek. Ezek hatékony statisztikai kezelése csak új, a robusztusság kívánalmainak megfelelő statisztikai algoritmusok használatával lehetséges.

A robusztus eljárások gyakorlati alkalmazásának egyik kulcskérdése, hogy a hibát adekvát módon definiáljuk. Ehhez elengedhetetlen az eddigi hibadefiníciók revíziója is.

A tanulmány első részében a hagyományos hibadefiníciót (σ-t) és annak fogyatékosságait, második részében az általunk javasolt új hibadefiníciót tárgyaljuk. Végezetül táblázatban foglaljuk össze a hagyományos statisztika (legkisebb négyzetek elve) és az általános leggyakoribb értékek koncepciója szerinti hibadefiníciók leglényegesebb vonásait.

Обработка и интерпретация данных геофизических исслздований часто выполняется с применением традиционных методов математической статистики, которые предполагают, что данные подчиняются распределению Гаусса. Однако опыт исследования геофизических, геологических и асторномических данных показывает, что распределение этих данных в большинстве случаев отличается от распределения Гаусса. Поэтому эффективность статистической обработки этих данных может быть обеспечена только при применении новых статистических алгоритмов, подчиняющихся требованиям устойчивости.

Одним из основных вопросов практического применения устойчивых методов является однозначеность определения погрешностей. Для решения этого вопроса необходимым является пересмотр применявшихся ранее методов определения погрешностей.

В первой части статьи рассматривается традиционный метод определения погрешностей (σ–т) и его недостатки, а во второй части обсуждается предлагаемый нами новый метод определения погрешностей. В заключение в таблице сведены наиболее важные характеристики традиционного статистического метода определения погрешностей (метод наименьших квадратов) и определения погрешностей методом наиболее частых значений.

It is both principally and practically inconsistent, dubious and disadvantageous to calculate errors in traditional way, if we use robust statistical methods. Disregarding for a moment this point of view, however, the usual definition of error shows many shortcomings also for itself, if it isn't guaranteed that the distribution of deviations is of Gaussian type. The error definition proposed in the present paper realizes a very similar characterization of the error for a great variety of distribution types, as the standard deviation does it but the latter one only for the Gaussian type and for its very neighbourhood.

Bevezető meggondolások

A geofizikai adatrendszerekben rejlő információ minél hatékonyabb kinyerésének jogos igénye modern statisztikai algoritmusok alkalmazását teszi elengedhetetlenné. Az új módszerek használatához viszont az eddigi hibadefiníciók revíziója és új hibadefiníciók megadása válik szükségessé.

Pontosabban fogalmazva az a tulajdonképpeni kérdés, hogy egy olyan esetben, amikor valamely egy, vagy többváltozós, valamilyen elv szerint végrehajtott kiegyenlítés után adott a mért és számított értékek különbségének d_i el-

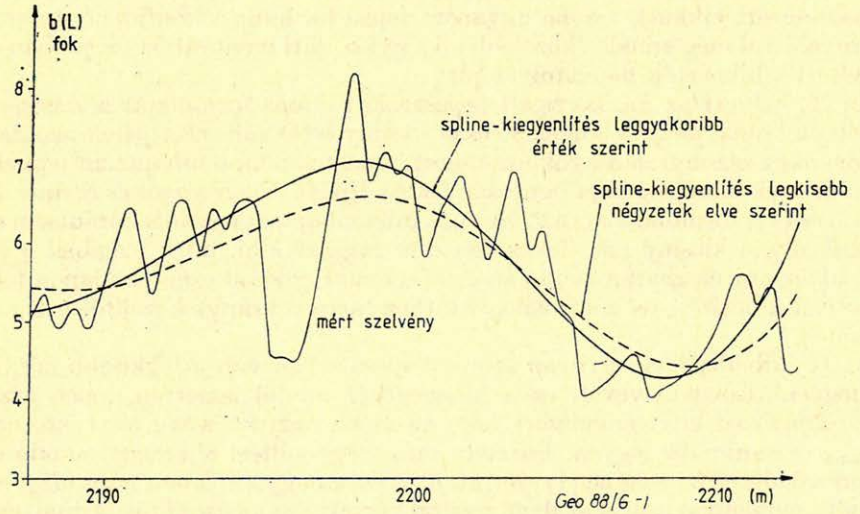
* NME Geofizikai Tanszék

térésrendszere (n db adat), akkor hogyan számítandó a hiba. Vagyis az az egyetlen érték, amelyik a gyakorlati szakember számára kézzelfoghatóan jellemzi a mért értékek megbízhatóságát. Meggondolásainkban ugyan a d_i eltérések fognak szerepelni, de az ezek alapján számított hiba természetesen magát az adatrendszert jellemzi.

(Magának a kiegyenlítés eredményének a megbízhatósága kellően kézben tartható a becslések aszimptótikus szórásainak a számításával, ld. pl. *Huber 1981* vagy *Steiner 1985*, így ezzel a problémával e helyen nem foglalkozunk. A kiegyenlítés eredménye n növekedésével egyre megbízhatóbb lesz, míg a jelen dolgozat magát a hibaeloszlást elemzi, és arra ad meg – n-től nyilván független – jellemzőt.)

Matematikai szempontból nézve az az első lépés, hogy a d_i eltéréseket valamilyen x-szel jelölt és $f(x)$ sűrűségfüggvénnyel jellemzett valószínűségi változó mintavételezett értékeinek fogjuk fel. A második lépésben egy skálaparamétert választunk ki, és ezt használjuk fel az x-ek diszperziójának a jellemzésére. (A skálaparaméter definíciójára nézve újra az imént idézett két munkára hivatkozunk.)

E második lépésnek matematikailag nincs nagy súlya: ismert eloszlástípus esetén ugyanis egy tetszőleges kiválasztott skálaparaméter alapján bármelyik másik kiszámítható. (A legismertebb példa erre: a σ szórásból, amennyiben az x



1. ábra. Az $|x| < P$ esetek valószínűségének aránya az $|x| > P$ esetek valószínűségéhez viszonyítva különböző eloszlástípusoknál. (Az x-ek a kiegyenlítés utáni eltérések, a az eloszlástípust megadó paraméter, ld. a (4) definíciós formulát.) A nyíl az ordinátán a Gauss-eloszlásra és a σ-ra vonatkozó, analóg módon képzett valószínűségarányt mutatja

Рис. 1. Доля вероятности случаев $|x| < P$ по отношению к вероятхости случаев $|x| < P$ для различных типов распределений x – разница после выравнивания, a – параметр, задающий тип распределения, см. формулу определения (4). Стрелка на оси координат показывает долю вероятности, полученную аналогичным образом для распределения Гаусса и σ

Fig. 1. Probability ratio of the events $|x| < P$ and $|x| > P$, respectively, for various distribution types. (Eq. 4 defines the meaning of the type parameter a.) The value at the arrow corresponds to the analogous ratio of the events $|x| < \sigma$ and $|x| > \sigma$, respectively, in case of a Gaussian distribution

valóban Gauss-eloszlású, $0{,}6745\sigma$ szerint számítható a *Bessel (1815)* által definiált valószínű hiba, azaz a q interkvartilis félterjedelem.) A hibadefiníció kiválasztásakor tehát annak kell mérvadónak lennie, hogy a gyakorlati szakember tulajdonképpen miről akar gyors tájékoztatást kapni a hiba értéke alapján.

A hagyományos hibadefiníció

Induljunk ki a

$$\sigma_{\text{emp}} = \frac{1}{n} \sum_{i=1}^{n} d_i^2 \tag{1}$$

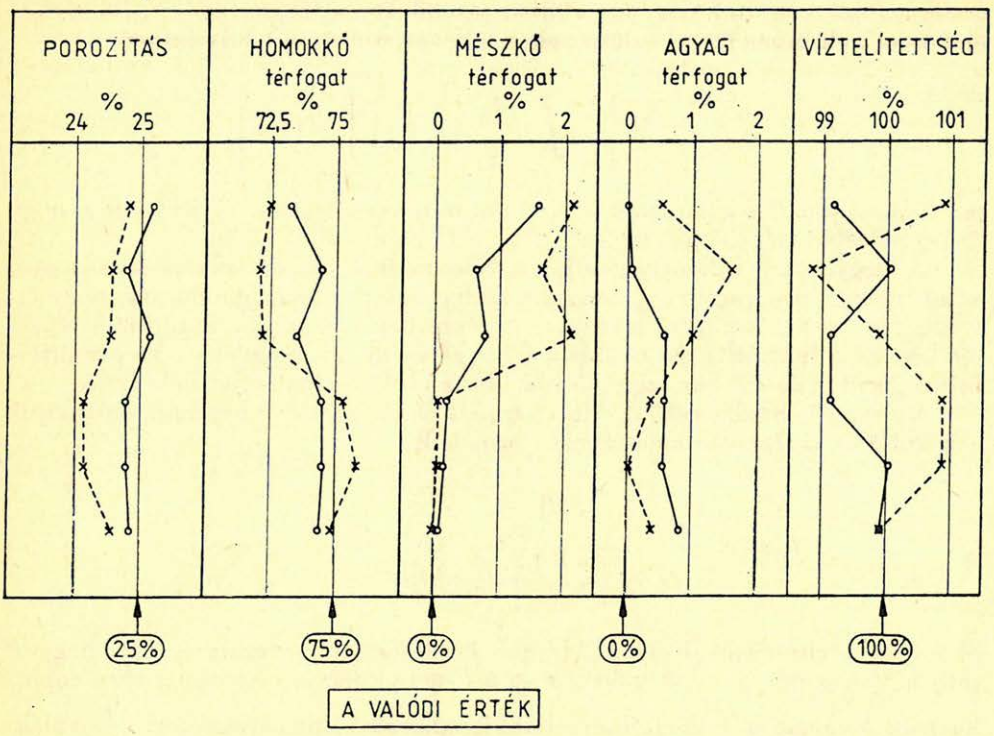
jól ismert hibajellemzőből (ún. tapasztalati szórásból), amely $n \to \infty$ esetén a σ szórást közelíti (ha ez utóbbi valóban létezik). Gauss-típusú hibaeloszlásnál közismert, hogy az eltérések abszolút értékére $|x| < \sigma$ az esetek *68,26*%-ában teljesül. Az (1) kifejezés birtokában tehát a következőre lehet számítani: *kb. kétszer akkora annak a valószínűsége*, hogy az (1) szerint számított hibánál *kisebb* abszolút értékű d_i eltéréssel találkozunk, mint annak, hogy *nagyobbal* (a valószínűségarány közel *2:1*).

Néhány megjegyzés az előzőekhez:

- A valószínűségaránnyal való jellemzés gondolata Besseltől származik (ld. a már idézett cikket), s noha ugyanazt fejezi ki, mint a konfidenciaszint, egy árnyalattal még annál is közelebb áll a gyakorlati mentalitáshoz, jobban érzékelteti a hibaérték mondanivalóját;
- Az (1) helyett az ún. korrigált tapasztalati szórás formuláját is szerepeltethettük volna, de ez a korrekció csak kicsiny értékváltozást jelent azokhoz az igen nagy bizonytalanságokhoz képest, amelyek a hiba hibájaként lépnek fel, ha az x-ek eloszlása nem pontosan Gauss-típusú (ld. *Hajagos és Steiner 1988*, (13) és (14) formula, illetve *2. ábra*). A hiba hibájánál ugyanis korántsem szoktunk olyan kicsiny százalékos értékhez ragaszkodni, mint magánál a tulajdonképpeni meghatározandó adatnál. (Imént például ezen az alapon tekinthettük a *68,26*%-os konfidenciaszinthez tartozó arányt közelítőleg *2:1* értékűnek.)
- Az (1) hibadefiníció nyilván szoros kapcsolatban van a legkisebb négyzetes kiegyenlítés alapelvével: ez a kiegyenlítés ui. jól ismerten éppen aszerint *alakítja ki* a d_i eltérésrendszert, hogy az eltérés négyzetösszeg, azaz ezzel együtt σ_{emp} is minimális legyen. Bármely más kiegyenlítési eljárásnál adódó d_i' eltérésrendszerre tehát az (1) szerinti σ_{emp} csak nagyobb lehet, mint σ'_{emp} – ami már önmagában is jelzi az ilyen módon végzett összehasonlítás semmitmondó (sőt félrevezető) voltát.

Az új hibadefiníció

Nyilvánvaló, hogy új, a robusztusság és rezisztencia követelményeit is kellő mértékben kielégítő hibadefinícióra van szükség. – A σ_{emp} típusfüggő (azaz nem robusztus) voltára már utaltunk. A durva hibákra való nagyfokú érzékenység pedig azonnal látható az (1) formulából: egy, vagy néhány olyan d_i érték, amelyik lényegesen nagyobb a többinél, gyakorlatilag teljesen meghatározhatja σ_{emp} értékét, függetlenül a zömmel előforduló d_i-k értékeitől. Az esetek túlnyomó több-



2. ábra. A pontos 2:1 valószínűségarányhoz tartozó „korrekciós-görbék". A görbéket összehasonlítva megállapítható, hogy a P hibadefiníció *robusztus* (az eloszlástípustól csak lényegtelen mértékben függ), a σ pedig *nem robusztus* (csak egy szűk típustartományra korlátozódik a használhatósága)

Рис. 2. Коррекционные кривые точной доли вероятности 2 : 1. Сравнение кривых показывает, что определение погрешностей P является устойчивым (т. е. лишь в очень незначительной степени зависит от типа распределения, а σ не является устойчивой (т. е. возможность ее использования ограничивается только на определенные типы распределения)

Fig. 2. "Correction curves" to get the exact 2:1 value for the ratio of probabilities. The comparison of the curves shows that the error P is *robust* (i. e., its meaning depends only insignificantly upon the actual type of probability distribution), on the contrary, the error σ *isn't robust* (i. e., the practical usefulness of σ is limited to a narrow type domain)

ségét jellemző megbízhatósági mértékről a σ_{emp}-érték alapján így semmiféle támpontunk nem lesz.

Az eltérésrendszerekre a következő hibadefiníció elfogadását javasoljuk:

$$P_{emp} = \varepsilon \cdot \left\{ \prod_{i=1}^{n} \left[1 + \left(\frac{d_i}{k\varepsilon} \right)^2 \right] \right\}^{\frac{1}{2n}}, \qquad (2)$$

ahol a d_i-k a k paraméterrel végrehajtott általános leggyakoribb érték szerinti kiegyenlítés eredményétől számított eltérések és ε ezek dihéziója (ld. pl. *Steiner*

1985; a P jel arra utal, hogy ezt a hibát produktumból számítjuk). A P elméleti értéke az x eltérések $f(x)$ eloszlása esetén a következőképpen számítandó:

$$P = \varepsilon \cdot \exp\left\{\frac{1}{2}\int_{-\infty}^{\infty} \ln\left[1+\left(\frac{x}{k\varepsilon}\right)^2\right]\cdot f(x)\,dx\right\} \qquad (3)$$

(ennek alapján P-t logaritmikus hibának is nevezhetnénk). Nyilvánvaló, hogy $f(x)$ eloszlásból származó d_i-k esetén $P_{emp} \to P$, ha $n \to \infty$.

A kiegyenlítés valamely módjához való kötődés (azaz k szereplése különösnek tűnhet P-ben vagy P_{emp}-ben mindaddig, míg meg nem gondoljuk, hogy (1) is kiegyenlítéshez kötött (a legegyszerűbb esetben a számtani középérték előzetes képzése jelenti ezt). Ez ráadásul σ_{emp} esetén mereven egyetlen kiegyenlítésfajtát jelent, míg a P-ben szereplő k az aktuális eloszlásnak megfelelő érték.

Ha modelleloszlás-családként, *Csernyák* és *Steiner 1982* nyomán, elfogadjuk a következő, a típusparaméterű szupermodellt:

$$f_a(x) = \frac{\Gamma\left(\frac{a}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{a-1}{2}\right)\cdot(1+x^2)^{a/2}} \qquad (1<a<\infty), \qquad (4)$$

akkor már feltehetjük legplauzibilisebb kérdésünket, nevezetesen azt, hogy P milyen valószínűségarányt valósít meg a P-nél kisebb és nagyobb x-ekre vonatkozóan. A válasz az *1. ábráról* olvasható le, ahol az a típusparamétert $\frac{1}{a-1}$ alakban szerepeltettük az abszcisszán, így az *1*-es értéknél a Cauchy-eloszlás szerepel, az origóban pedig a Gauss-féle eloszlástípus. (Könnyű igazolni ui., ld. pl. *Hajagos 1985*, hogy $a \to \infty$ esetén Gauss-típusú eloszláshoz jutunk.) Az *1. ábra* ordinátáján kis nyíl jelzi a Gauss-eloszlás esetén a σ által megvalósított valószínűségarány pontos értékét.

Mivel *2:1* körüli arányok szerepelnek az *1. ábrán*, azt is felhordtuk (ld. a *2. ábra c* jelű görbéjét), hogy milyen c-kel adódna $c \cdot P$-nél *pontosan 2:1* arány különböző eloszlástípusoknál (ilyen értelemben c korrekciós faktornak tekinthető, amelynek csak elméleti vizsgálatainkban van jelentősége, a gyakorlatban azonban az alkalmazása felesleges). Látjuk, hogy c nem mutat nagymértékű ingadozást az *1*-es érték körül, így a P-t elfogadhatjuk eloszlástípusok nagyon különböző fajtáira is a *2:1* valószínűségarányt közelítőleg megadó hibadefinícióként.

Ezt abból a szempontból is örömmel veszzük tudomásul, hogy a P által közelítőleg megvalósított valószínűségarány azonos a σ által a Gauss-eloszlásra közelítőleg megvalósítottal. A σ nem robusztus jellege miatt azonban ezek a σ-ra vonatkozó arányok gyorsan megváltoznak, ha az aktuális eloszlástípus távolodik a Gauss-félétől. Ezt a c-vel analóg módon definiált C értékek görbéjével szemléltetjük a *2. ábrán:* most $C \cdot \sigma$ adja a pontos *2:1* valószínűségaránynak megfelelő hibaértéket. A C-görbe meredeken csökken a *típus-intervallum* felezőpontjához közeledve, hiszen a ≤ 3 esetén $\sigma = \infty$. Az $\frac{1}{a-1} = 0{,}5$-től kezdve tehát semmiféle C-vel nem kompenzálhatjuk a-nak az eloszlás szárnyainak *legkülső* szakaszaira való, egészségtelenül nagy érzékenységét.

A P optimum sajátsága

A P-nek, mint határozatlansági jellemzőnek egy fontos tulajdonságára információelméleti vizsgálat világít rá a következőképpen.

Tekintsük a legegyszerűbb esetet: egyetlen érték meghatározása tehát a feladat. Ha az M_k leggyakoribb értéket határozzuk meg valamely aktuális $f(x)$ eloszlásra, az azzal egyértelmű, hogy valamelyik a-értékhez tartozó $f_a(x)$ eloszlást használjuk helyettesítő (modell-) eloszlásként az I-divergencia

$$I = \int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)} dx \tag{5}$$

kifejezésében $g(x)$ helyén, és azt az M_k-értéket fogadjuk el helyesnek, amelyiknél az információ-veszteségként értelmezett I minimális. (A k- és a-értékek között kölcsönösen egyértelmű kapcsolat áll fenn, ld. pl. *Steiner 1985.*) A levezetést nem részletezzük, de ennek első lépéseként $f_a(x)$-nek a (4)-ben adott standard kifejezése helyett az az általánosabb alak kerül (5)-be $g(x)$ helyére, amelyiket (4)-ből úgy kapunk, hogy $\frac{x - M_k}{\varepsilon k}$-t helyettesítünk x helyett, és normálási okokból még $(k\varepsilon)$-nal osztunk. Így azonnal felismerhető az I kifejezésében a P-ben (ld. (3)) szereplő integrál. Ha az egyenlet két oldalának exp-függvényét vesszük, kiderül, hogy P e^I-vel arányos úgy, hogy a szorzótényező M_k-t már nem tartalmazza, így P ugyanott lesz minimális, ahol I. Ez azt jelenti, hogy a leggyakoribb értékeket számítva, minimális hibájú eredményeket kapunk. (Ha $f(x)$ valamely f_a-val azonos típusú. akkor nincs információveszteség, optimális az algoritmus.)

A fentiekkel szoros kapcsolatban áll az alábbi, a praktikus mentalitáshoz közelebb álló gondolatmenet.

A leggyakoribb értékek szerinti kiegyenlítés a

$$\prod_{i=1}^{n} [(k\varepsilon)^2 + d_i^2] = \min \tag{6}$$

feltételt teljesíti. (Logaritmálás és differenciálás után azt kapjuk, hogy a $\frac{(k\varepsilon)^2}{(k\varepsilon)^2 + d_i^2}$ súlyokkal szorzott eltérésnégyzeteket kell minimalizálnunk, ami gyakorlatilag arra az ismert iterációs algoritmusra vezet, amelynek minden lépése súlyozott legkisebb négyzetes kiegyenlítés.) Kiemelve a (6) kifejezésből $(k\varepsilon)^{2n}$-t, (azaz szorzórényezőként $(k\varepsilon)^2$-et), $2n$-edik gyököt vonva és végül k-val osztva, a P_{emp} (2) kifejezését kapjuk, ami nyilván ugyanott minimális, ahol az eredeti (6) kifejezés.

P-számítás

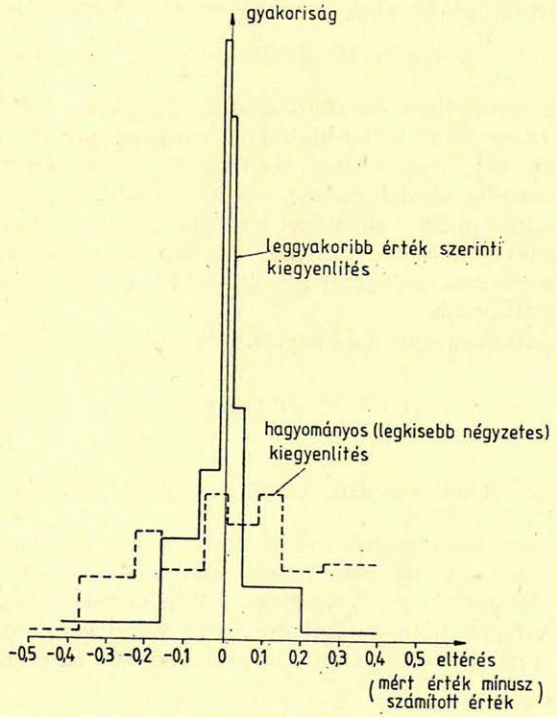
Ha gépben van az adatrendszerünk, ennek σ_{emp} hibáját minden járulékos adat nélkül, *gombnyomásra* kapjuk meg. Persze ennek az adatnak valószínűségaránnyal, vagy konfidencia-szinttel megadott *értelme* csak akkor lesz, ha a hibaeloszlás típusát ismerjük. A *68,26%*-os konfidencia-szint jelentést feltételezni pl. gyakorlatilag annyi, hogy a hibaeloszlást Gauss-típusúnak hisszük.

Ilyen értelemben a σ_{emp}-számítás implicit módon tartalmaz típusra vonatkozó megszorítást, – persze lehet, hogy a felhasználó a számítás során erre nem

is gondol. A P (2) vagy (3) szerinti számításánál azonban a típusról nyilatkozni kell, hiszen a k értéke az a típusparamétertől függ. – A következő empírikus függvénnyel számíthatjuk kielégítő pontossággal k-t adott a-ból (ld. *Steiner 1985*):

$$k = \sqrt{a-1} + \frac{1}{2\pi} \cdot \frac{\sqrt{a} - \sqrt{2}}{\sqrt{3}}. \qquad (7)$$

A típusra vonatkozó ismeret igénye elbizonytalaníthatja első pillanatban a felhasználót, de az alábbiakban kiderül, hogy a P robusztussága miatt módunk van olyan egyszerűsítésre, amely járulékos információ nélkül, szintén *gombnyomásra* szolgáltat hibát, mégpedig olyant, amelyik széles típustartományon jó közelítéssel valósítja meg a *2:1* valószínűségarányt (nem úgy, mint a hibaként csak nagyon szűk típustartományra, gyakorlatilag a Gauss-eloszlás közvetlen közelére, s ott is csak garantáltan durvahiba-mentes esetre használható σ_{emp}).



A helyes értéktől való eltérések gyakorisági diagramjai.
Geo 88/6-3

3. ábra. Valószínűségarányok az „automatikusan számítható" P hibához (ekkor a hibaeloszlás-típus előzetes közelítő ismerete nem szükséges)

Рис. 3. Доли вероятности для погрешности P, вычисляемой автоматически (в этом случае нет необходимости в примерном определении типа распределения)

Fig. 3. Probability ratios to the error P which is "automatically" computable (i. e., even a rough preliminary estimation of the distribution type is superfluous)

Közelítőleg 2:1 valószínűségarányt adó hibadefiníciók a kiegyenlítés utáni d_i eltérésre vonatkozóan
(A 2:1 arány annyit jelent, hogy d_i kétszer akkora valószínűséggel *kisebb* a hiba értékénél, mint amilyen valószínűséggel *nagyobb* nála)

Определения погрешностей, дающие соотношения вероятностей примерно 2 : 1, по отношению к разнице $д_i$ после выравнивания
(соотношение 2 : 1 означает, что вероятность того, что значение $д_i$ меньше значения погрешности, в два раза больше, чем вероятностьтого, что значение $д_i$ больше значения погрешности

Error definitions giving approximately a probability ratio of 2:1 concerning the deviations d_i got by fitting.
(The ratio 2:1 means that $|d_i|$ less than the error is twice so much probable than a greater $|d_i|$)

	A hibaformula	Határérték, amit a hiba közelít $n \to \infty$ esetén (x az eltérés, $f(x)$ ennek sűrűségfüggvénye)	A hiba rezisztens-e?	A hiba robusztus-e?	Optimum-sajátság
Hagyományos statisztika szerint	$\sigma_{\text{emp}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$	$\sigma = \sqrt{\int_{-\infty}^{\infty} x^2 \cdot f(x)dx}$ (gyakran végtelen, ekkor nem tudni, mit fejez ki)	*nem*, mert σ_{emp}-et teljesen meghatározhatja egy vagy néhány kieső adat, s ekkor semmi köze σ-hoz. (Ha valamilyen algoritmus szerint hagyjuk el a kieső adatokat, σ_{emp} az elfogadott algoritmus szerint adódhat jelentősen különböző értékűnek.)	*nem*, mert a kb. 2:1 arány csak a Gauss-félénél és a hozzá igen közeli eloszlásoknál teljesül, egyébként nincs információnk a konfideciaszintről	legkisebb négyzetes kiegyenlítés utáni d_i adatrendszerre σ_{emp} minimális
Az általános leggyakoribb értékek koncepciója szerint	$P_{\text{emp}} = \varepsilon \left\{ \prod_{i=1}^{n} \left[1 + \frac{d_i^2}{(k\varepsilon)^2} \right] \right\}^{\frac{1}{2n}}$ (Ha az eloszlástípus ismeretlen, k = 2-t használjuk.)	$P = \varepsilon \cdot \exp \cdot \left\{ \frac{1}{2} \int_{-\infty}^{\infty} \ln \left[1 + \left(\frac{x}{k\varepsilon} \right)^2 \right] \cdot f(x)dx \right\}$ (az egész $f_a(x)$-családra véges érték)	*igen*, kieső adatok csak lényegtelen hatást gyakorolnak. (Jelentősebbé ezek hatása csak a dihéziónál nagyságrendekkel nagyobb és/vagy nagy %-os arányban előforduló kieső adatok esetén válhat.)	*igen*, a 2:1 arány az $f_a(x)$ modellcsalád Gauss-tól Cauchy-ig terjedő szakaszán közelítőleg teljesül	a k faktorral végrehajtott leggyakoribb érték szerinti kiegyenlítés utáni d_i adatrendszerre P_{emp} minimális

Ha nincs az eloszlástípusra vonatkozóan előzetes információnk, *javasoljuk a* $k = 2$ *használatát:* ezzel (2) egyértelműen szolgáltat egy P_{emp} hibát. (A vesszővel való megkülönböztetés csak összehasonlító elméleti vizsgálatoknál indokolt, gyakorlati esetekben konstans k alkalmazásakor is a P_{emp} vagy P jelölést használjuk.)

A *3. ábra* az $f_a(x)$ eloszláscsaládra mutatja a P'-re vonatkozó valószínűségarányokat. Megállapítható, hogy a P' valóban széles típustartományon valósít meg közelítőleg *2:1* valószínűségarányt (az ettől az aránytól való eltérések a Gauss-Cauchy típusintervallumon alig valamivel nagyobbak, mint a P-nél, v.ö. a *3.* és *1. ábrát*). – Megjegyezzük még, hogy $k = 2$-vel dolgozva, magának a kiegyenlítési procedúrának a hatásfoka is *90% fölött van* a Gauss-eloszlástól csaknem egészen a Cauchy-féléig (utóbbinál *89%*); *100%*-hoz nagyon közeliek a hatásfokok $k = 2$ esetén a gyakran előforduló, *3* és *10* közötti a-kal (*0,1* és *0,5* közötti $\frac{1}{a-1}$-gyel) jellemzett eloszlástípusoknál. Figyelemre méltó, hogy ebben a gyakorlatilag fontos tartományban a valószínűségarányok a *2:1* értéktől *10%*-nál kisebb mértékben térnek csak el (ld. újra a *3. ábrát*).

Összefoglalás

A dolgozat lezárásaként a hagyományos és a jelen cikkben javasolt hiba összehasonlításának a megkönnyítésére táblázatba rendezve közöljük a dolgozatban szereplő hibadefiníciókat, valamint néhány főbb megállapítást, illetve megjegyzést (lásd *1. táblázat*).

IRODALOM

Huber, P. J.: Robust Statistics, Wiley, New York, 1981.

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

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)

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.

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.

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

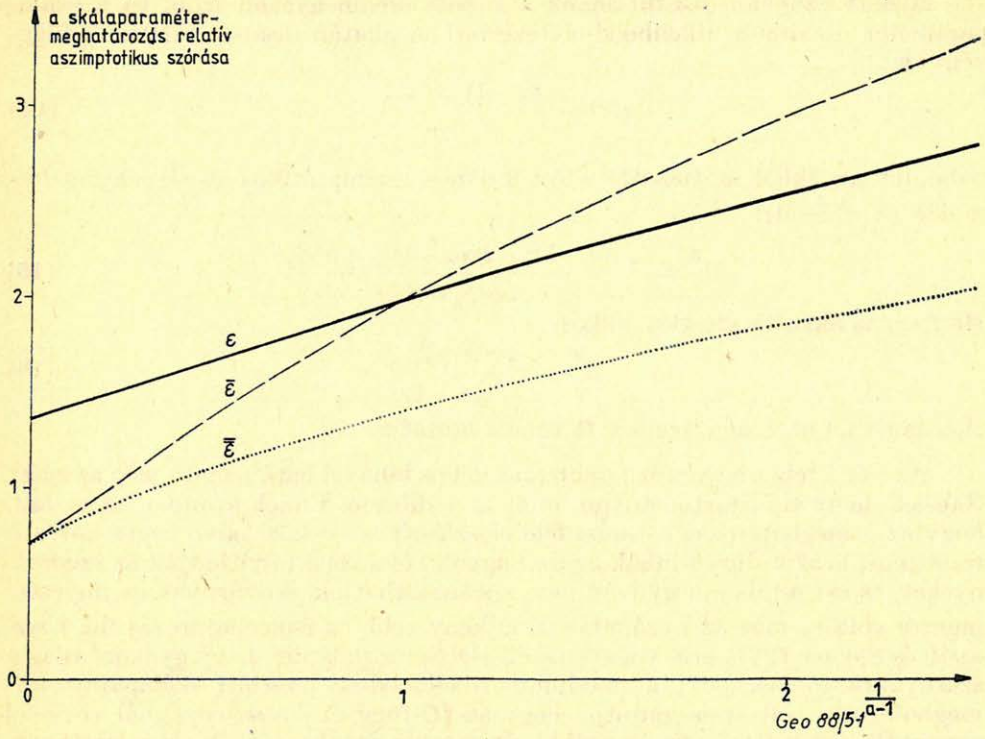
Az általános leggyakoribb érték-számítás második variánsánál, amely *Hajagos 1985* eredményeire épül, skálaparaméterként (röviden megfogalmazva) közvetlenül k meghatározása történik az alábbi $\varkappa$-függvény szerint:

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

(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.

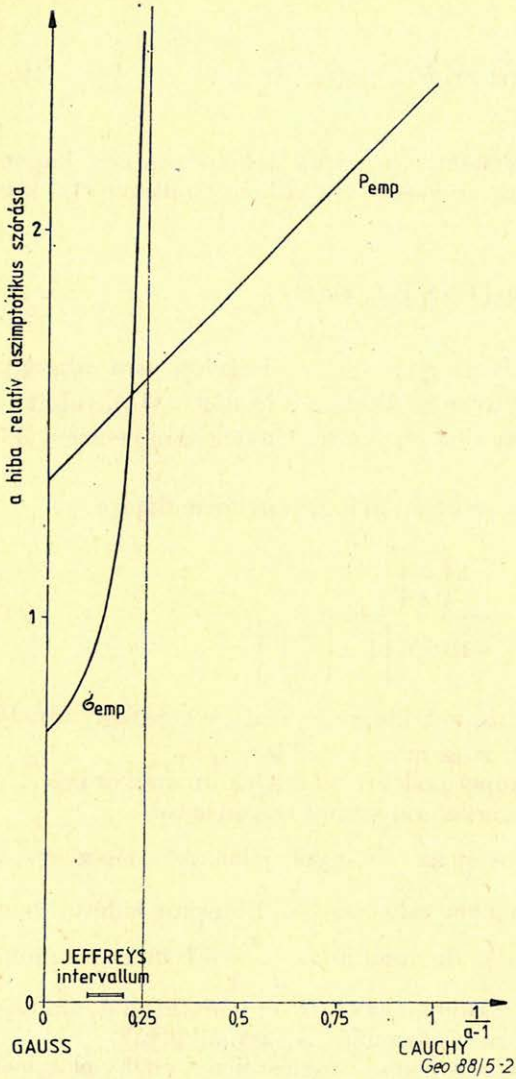
Mivel az A_p/P értékeit az $f_a(x)$ szupermodellre számítjuk ki, amikor is a (19)-ben a gyökjel alatti második tag egyszerűen (9) szerint számítható.

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:

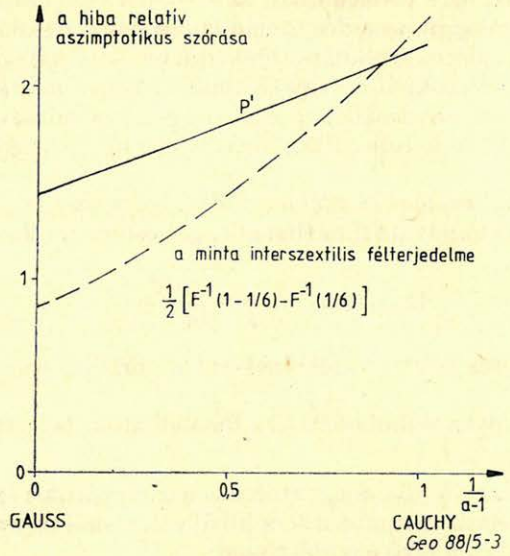
1. $a \leq 5$-re végtelen nagy a σ_{emp}-meghatározás aszimptótikus szórása;
2. a σ_{emp} és a P_{emp} relatív aszimptótikus hibája $a \approx 6$-nál egyezik meg (ez a Jeffreys-intervallum szélén levő eloszlástípus);
3. $a > 6$ esetén ugyan kisebb σ_{emp} relatív hibája, mint P_{emp}-é, a rezisztencia 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:

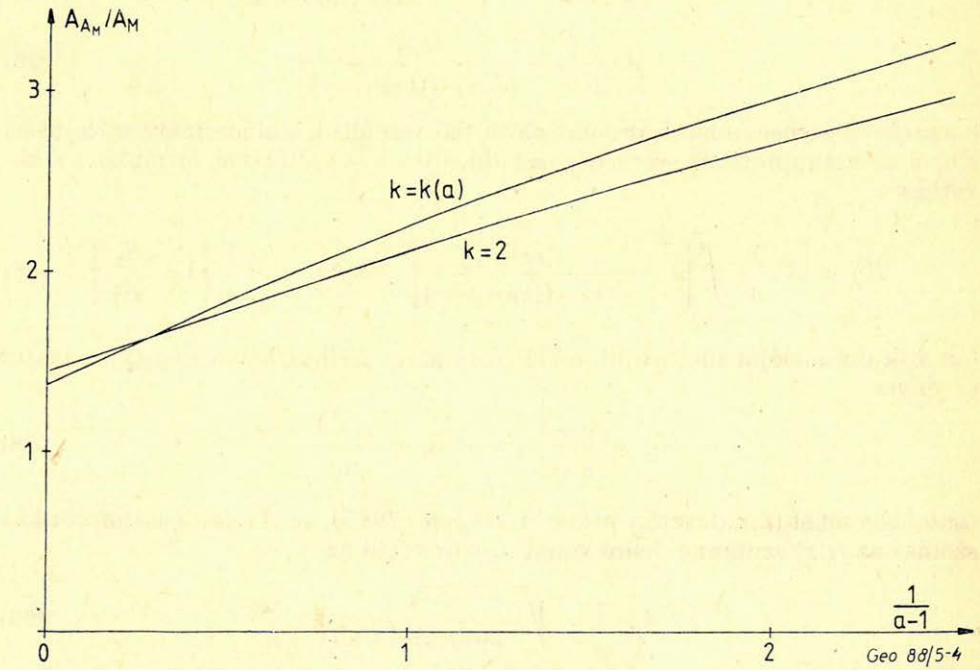
$$A_{A_M}/A_M = \sqrt{4\left[1 - \frac{a-2}{\pi \cdot (a-1)}\right]^2 + \frac{1}{2(a+2) \cdot (a-1)} + K} \tag{30}$$

(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*.

KREMSZNER MIKLÓS

1932—1988

Alkotó erejének teljében, 1988 január 18-án váratlanul hunyt el, családjában, valamint baráti és szakmai közösségében pótolhatatlan űrt hagyva hátra.

Fiatalsága Sopronhoz kötötte és oda is tért meg. Munkája a fővároshoz kapcsolta az egész országra kiterjedően.

A Nehézipari Műszaki Egyetemen szerzett geofizikusmérnöki oklevelet 1954-ben, megszerezve később a mérnök-közgazdász képesítést is.

A munkát a Magyar Állami Eötvös Lóránd Geofizikai Intézetben kezdte, tevékenyen közreműködve a mélyfurási geofizika szénkutatásban történő meghonosításánál.

A vízkutatással 1958-ban jegyezte el magát és ez maradt a működési területe az utolsó pillanatig. Elsők között volt, aki ilyen feladatokat vállalt Mongoliában.

1959-től a Vízkutató és Fúró Vállalat dolgozója. Kezdetben észlelő és kiértékelő, 1969-től a Geofizikai Osztály vezetője.

Egyesületünknek alapító tagja, a Mélyfúrási Geofizikai Szakosztály, az Oktatási Bizottság és az Országos Elnökség tevékeny tagja volt.

Emlékét szeretettel őrizzük!

L. S.

Ára: 32,50 Ft