Földtani Közlöny, Bull. of the Hungarian Geol. Soc. (1988) 118. 175—183

A geostatisztikai számítások hatásfoknövelésének szükségessége és új lehetőségei*

*Steiner Ferenc***

(5 ábrával, 3 táblázattal)

Összefoglalás: A dolgozat rövid elméleti összefoglalóval és néhány példával hívja fel módszerfejlesztők és alkalmazók szíves figyelmét a statisztikai értelemben vett hatásfoknövelés gazdaságos *lehetőségeire*, valamint arra, hogy ezekkel a lehetőségekkel már a közeljövőben *szükségszerűen* élnünk is kell, ha valóban a hatékonyság optimumára törekszünk. — A leggyakoribb értékek szerinti kiegyenlítés minden olyan feladatnál azonnal bevezethető, ahol a legkisebb négyzetek elve közvetlenül nyert eddig felhasználást; bonyolultabb esetek esetleg járulékos elméleti adaptálást is szükségessé tehetnek.***

A geostatisztikát lehet nagyon szűk értelemben definiálni, de a matematikai geológia már csaknem két évtizede megjelenő nemzetközi folyóiratában (Journal of the International Association for Mathematical Geology) egyre inkább találhatók általánosabb — sőt nagyon általános — definíciók erre a tudományágra. Értelmezésemben a geostatisztika szintén általánosan értendő (másképpen talán nem is volna indokolt külön diszciplínának tekinteni), amelyet *célkitűzése* és *módszere* definiál, amennyiben a bányászat és a további geológiai-geofizikai kutatás számára törekszik a jelentős költségkihatású döntéseknél közvetlenül felhasználható, minél nagyobb megbízhatóságú információk (geoinformációk) statisztikai kinyerésére a rendelkezésre álló adatrendszerből. A célkitűzés annyiban teszi specifikussá az alkalmazott matematikai statisztikai eljárások összességét metodikailag is, hogy azoknak a földtani és geofizikai kutatás mérési eredményrendszereire szabottaknak kell lenniök. Ez egyrészt nyilván adaptálások sorát jelenti, másrészt azonban azt, hogy *az eljárásoknak a geofizikai és földtani kutatások mérési adatrendszereire, ezek eloszlástípusaira kell előnyösen alkalmazhatóaknak lenniök.*

A kutatási költségek nagysága fokozottan teszi szükségessé annak az egyébként is természetes követelménynek a maradéktalan érvényre juttatását, hogy a *hatékonyság optimuma* valósuljon meg. Nyilvánvaló, hogy az azonos megbízhatóságú információhoz szükséges adatbeszerzési (mérési) és számítási (gép-

* Előadásként elhangzott 1987. május 21-én, Miskolcon, a Borsodi Műszaki Hetek keretében, a Magyarhoni Földtani Társulat és a Magyar Geofizikusok Egyesülete által együttesen szervezett előadóülésen.

** Nehézipari Műszaki Egyetem Geofizikai Tanszék 3515 Miskolc-Egyetemváros.

*** A dolgozat egyszerű megfogalmazásokra törekszik, az I. táblázatban pl. röviden foglalja össze az alapelveket és az abból következő algoritmusokat. Amennyiben a tisztelt olvasó részletesebb információkat igényel, ill. fel akarja frissíteni valószínűségszámítási ismereteit, PRÉKOPA (1962) kitűnő, lényegre törő könyvének regiszt ere alapján gyorsan tájékozódhat a maximum likelihood-elvtől a sűrűségfüggvény fogalmáig a dolgozatban szereplő fogalmak nagyobb részéről. Az I-divergencia, a robusztusság, a rezisztencia fogalmaira, az iterative számolt súlyozott kiegyenlítések végrehajtási technikájára nézve STEINER (1985) ad felvilágosítást.

óra-) költségeket *együttesen* kell figyelembe venni, ill. ezen együttes költségek minimumára kell törekedni.

Hogy a matematikai statisztikai módszerek számításigényessége milyen nagy mértékben térhet el egymástól, azt legyen szabad egy táblázatba való sűrítéssel bemutatnom. (Az itt és a továbbiakban szereplő táblázatok célja, hogy bizonyos egyszerűsítések és bizonyos szempontok kiemelése révén szemléletesen álljanak előttünk lényeges összefüggések.) — Tekintsük tehát most főleg ebből a szempontból az *I. táblázatot.*

Kiindulásunk a matematikai statisztika legfontosabbnak és legkorszerűbbnek tekinthető két alapelve, a maximum likelihood-elv és az I-divergencia minimalizálása, mint alapelv (az utóbbira nézve ld. pl. Hajagos, 1982). Az egyik esetben, a maximum likelihood-elvnél, az eloszlástípus pontos ismeretét tételezzük fel, — a másik esetben *modellezzük* az a priori pontosan szinte sohasem ismert eloszlástípust. — Talán felesleges hangsúlyoznom, hogy mennyivel modernebb ez az utóbbi szemlélet általánosságban is, de szakterületeink speciális esetében különösen. — Az ún. I-divergenciával mért információveszteség minimumára törekedve, azonnal megkapjuk azt a még nagyon általános alakú formulát, amelynek megoldása megadja az n db mérési eredmény alapján a legindokoltabban valóságos értéknek elfogadható, T-vel jelölt mennyiséget. (A jelölés a vonatkozó nemzetközi statisztikai szakirodalomban általánosan elfogadott; szimmetrikus eloszlásoknál pl. a szimmetriapontot jelenti.)

Ha hibáink eloszlását harang alakúnak tételezzük fel, mégpedig igen általános értelemben véve ezt a fogalmat, — akkor a g sűrűségfüggvény az $\frac{x-T}{S}$, ún. standardizált változó *négyzetének* a függvénye, tehát így írható: $g\left(\left[\frac{x-T}{S}\right]^2\right)$. Ebben az esetben viszont könnyen ellenőrizhetjük, hogy a fenti általános formula iteratíve ismételt súlyozott átlagszámításra redukálódik, ahol a φ súlyok g'/g-ként számítandók (a g' alatt a g sűrűségfüggvény deriváltja értendő).

Látjuk, hogy a két elv azonos iterációs algoritmusra vezet ugyan, de ennek számtalan realizációja van aszerint, hogy a valóságos eloszlásokat milyen eloszlástípusokkal lehet adekvát módon modellezni.

Amikor még csak mechanikus számológépek álltak rendelkezésre, amelyekkel 400 szorzás ill. osztás, vagy 1200 összeadás ill. kivonás volt a műszakonkénti norma (6 jegyre, ellenőrzéssel), akkor a számítástechnikailag legegyszerűbb algoritmus volt csak kivitelezhető: a legkisebb négyzetes elv szerinti, amely nem igényel iterációt és súlyszámítást sem. Ez — látjuk a táblázatból —, gyakorlatilag a hibák Gauss-eloszlásának a feltételezésével egyértelmű.

A számításigényesség szempontja az egy műveletre eső költségek már több évtizede és jelenleg is változatlanul tartó meredek csökkenése következtében — a gyakorlati feladatok természetétől függően — másod-, harmadrendűvé, esetleg tized-huszadrendűvé vált. Át kell tehát tekintenünk azt, hogy ha a g modelleloszlásra különböző feltevésekkel élünk, az milyen következményekkel jár.

Ha g tetszőleges, akkor az *I. táblázatban* feltett konkrét kérdésekre nem adhatunk ugyan választ, de azt hangsúlyoznunk kell, hogy egy egészen tetszőlegesen felvett g eloszlástípus általában *feleslegesen sok* számítást tehet szükségessé, amelynek végrehajtása esetleg még ma is problémákat okozhat.

Nagyon egyszerűvé válnak iteratív súlyszámításaink, ha a hibákat az

A táblázat megmutatja, hogy a matematikai statisztika két nevezetes alapelve (szimmetrikus hibaeloszlás esetén) olyan iterációs eljárásokra vezet, amelyeknek lépései súlyozott átlagképzések (általános esetben súlyozott kiegyenlítések). A modelleloszlások helyes megválasztásával jó statisztikai sajátságokkal rendelkező, gazdaságosan kivitelezhető eljárásokhoz jutunk

The table shows that both well-known basic principles of mathematical statistics (in case of a symmetrical distribution of errors) will lead to such iteration procedures the steps of which are weighted average calculations (or, in a general case, weighted adjustments). By a proper choice of model distributions methods of good statistical characteristics and economically feasible will be obtained

I. táblázat — Table I.

Matematikai statisztikai elvek és meghatározási módszerek kapcsolata; hasonlóságok és különbségek

	A MAXIMUM LIKELIHOOD-ELV	AZ I-DIVERGENCIA MINIMALIZÁLÁSA
Elvi kiindulás (a módszerek alapgondolata):	*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*	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 el helyes értéknek, amellyel az *információveszteség minimális*
	(néhány ismert logikai lépés)	(differenciálás, mintára való adaptálás)
Az alapelv szerint az a helyes T-érték, amely kielégíti a következő egyenletet:	$\sum_{i=1}^{n}\left(\frac{\partial f(x_i;T)}{\partial T}\Big/ f(x_i;T)\right)=0$	$\sum_{i=1}^{n}\left(\frac{\partial g(x_i;T)}{\partial T}\Big/ g(x_i;T)\right)=0$

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 voltát 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

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

jelöléssel nyilván

Az alapelvek gyakorlatilag súlyozott átlagképzés iteratív végrehatását írják elő:

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

alakú lesz, ezt iterációs algoritmusként kell értelmeznünk.

A számításigényesség mértéke a φ analitikus alakjától, azaz a g modelleloszlás megválasztásától függ. További egyszerűsítésként legyen $S = 1$

A g modelleloszlás megválasztásának lehetőségei (a konkrét esetek $T = 0$-ra felírva):	Tetszőleges	$f_a(x)=\frac{1}{c(a)\cdot(1+x^2)^{a+\frac{1}{2}}}$	$f_G(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
A súlyfüggvény és számításának gépóraigénye	$\varphi(x)$ általában bonyolult kifejezés, számítása viszonylag nagy gépóraigényű	$\varphi(x_i-T)=\frac{1}{1+(x_i-T)^2}$; számítása a valódi (nem elfajult) súlyfüggvények közül *minimális számú művelet* végrehajtását igényli	$\varphi(x_i-T)=1$; a súlyozott átlag közönséges *számtani középértéke*be megy át, iteráció sem szükséges
KÖVETKEZMÉNYEK: A T-meghatározás milyen aktuális elosztástípusokra maximális hatásfokú?	(Az alábbi kérdésekre nyilván csak specifikált esetben adható válasz.)	Az $f_a(x)$-szel jellemzett eloszlásokra, amelyek a különböző értékeinél egymástól jelentősen eltérő gyakorlati eseteket képesek modellezni	*Csak egyetlen* eloszlástípusra (az f_G-vel jellemzett GAUSS-eloszlásra)
Mennyire érzéketlen a hatásfok az eloszlás típusának változásaira?		Az a típusparaméter tág tartományában a hatásfok az eloszlástípusra *nagymértékben érzéketlen* (azaz *robusztus)*	*A hatásfok* az aktuális eloszlásnak a GAUSS-félétől való eltérése esetén *meredeken csökken* (nem robusztus)
Mennyire érzéketlen az eredmény durva hibájú adatokra, azaz az eljárás rezisztens-e?		Az a típusparaméter tág tartományában *az eljárás rezisztens*	*Nem rezisztens,* az eredményt néhány durva hibájú adat jelentősen módosíthatja, vagy teljesen tönkreteheti
A megfelelő kiegyenlítési módszer, amikor tehát az eljárástól nemcsak egyetlen T-adat meghatározását várjuk:		*A leggyakoribb értékek szerinti kiegyenlítés* (M-fitting) Az eredményül adódó hiperfelületet a tömörödési tendenciát mutató pontok határozzák meg (tekintet nélkül a kieső adatokra)	*A legkisebb négyzetes kiegyenlítés* (az M-fitting határesete, ha $a \to \infty$). Az eredményeket szemléltető hiperfelület úgy igyekszik elhelyezkedni, hogy lehetőleg a pontok egyikétől se legyen túlságosan távol (akkor is, ha ezzel eltávolodik a pontok tömörödési tartományától)

Idődiagram a statisztika néhány alapelvére, és a GAUSS-eloszlás feltételezésének sokáig gazdaságos, ma azonban már idejétmúlt voltára vonatkozóan

Time diagram concerning basic principles of the statistics, showing that the postulation of a GAUSSian distribution, approach that was for a long time quite rentable, is now fully out of date

II. táblázat — Table II.

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)

1700 — 1778 — 1800 — 1812 — 1886 — 1900 — 1912 — 1938 — 1960 — 1977 — 2000

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:

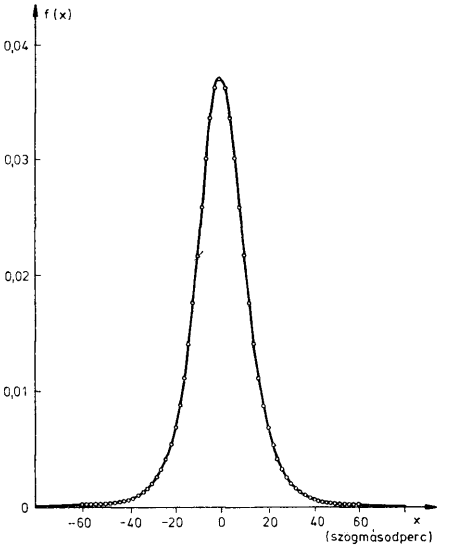
kb 1960-tól

Robusztus és rezisztens módszerek kifejlesztése, a kártékonnyá vált dogmák felszámolásának megkezdése (TUKEY, HUBER)

*I. táblázat*ban definiált $f_a(x)$ eloszlások valamelyikével modellezzük. Ez jelenti az ún. leggyakoribb értékek számítását, vagy általánosabban: a leggyakoribb értékek szerinti kiegyenlítést (l. STEINER 1973, CSERNYÁK 1973 és STEINER 1985). Ez a legkisebb négyzetes elv alkalmazásához viszonyítva kb. két nagyságrenddel nagyobb számításigényt jelent, így még nem is olyan túl régen az esetek zömében a GAUSS-eloszlás feltételezése jelentette a hatékonyság maximumát akkor is, ha így a megkívánt pontosságot esetleg csak jelentősen több mérési adat biztosíthatta (l. erre vonatkozóan a *II. táblázat*ot).

Tekintsük meg egy percre közelebbről is a *II. táblázat*ot, amely néhány kiemelt szempontra vonatkozóan egyrészt időbeli változásokat, másrészt kontinuitásokat hangsúlyoz.

Most kezdenek túlsúlyba jutni azok az esetek, amikor a nagyobb számú mérési adat követelménye (a végeredmény megbízhatóságának növeléséhez) csak lényegesen nagyobb költséggel teljesíthető, mint statisztikai értelemben nagy hatásfokú algoritmusok alkalmazásának bevezetése. — Legyen szabad itt felidéznem a hatásfok definícióját: azt az arányt adja meg a hatásfok valamely algoritmusra és hibaeloszlás-típusra, hogy *adataink hány százaléka lenne elegendő a végeredmény ugyanakkora megbízhatóságához, ha az adott eloszlástípusra*



1. ábra. Példa gyakorlati valószínűségeloszlásoknak valamely $f_a(x)$-modelleloszlással való jó közelíthetőségére. Folytonos vonal: NEWCOMB által tapasztalt hibaeloszlás; nullkörök: $f_a(x)$-értékek a = 5,3, S = 21,1 és T = 0 esetén

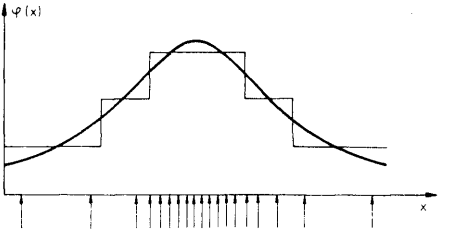
Fig. 1. Example of the possibility for a good approximation of practical probability distributions by a model distribution $f_a(x)$. Continuous line: distribution of errors observed by NEWCOMB; zero-circles: $f_a(x)$ values for a = 5.3, S = 21.1 and T = 0

optimális algoritmust alkalmaznánk. Ha pl. hibáink eloszlása

$$f_4(x) = \frac{2}{\pi(1 + x^2)^2}$$

szerinti, de a GAUSS-eloszlás feltételezésével, azaz a legkisebb négyzetes elv szerint számolunk, 50%-os csak a hatásfok: *ekkor tehát költségesen beszerzett adataink felét eltékozoljuk.* A számítási költségeknek az utóbbi évtizedekben bekövetkezett és változatlanul tartó meredek csökkenése új helyzetet teremt a hatékonyság optimumát illetően: véget ért az a mintegy másfél évszázados (kényelmes) időszak, amikor a GAUSS-eloszlás feltevéséből következő legkisebb négyzetes algoritmusnak statisztikai értelemben vett kicsiny hatásfoka esetén is optimális hatékonyságot lehetett elérni a mérési + számítási költségek együttes figyelembevételekor (l. újra a *II. táblázatot).* Ma már majdnem minden matematikai statisztikai számításnak, így *a geostatisztikai számítások zömének is okvetlenül szükséges nagy hatásfokra törekednie, mert másképpen a hatékonyság jelentősen el fog maradni a jelenlegi lehetőségektől.*

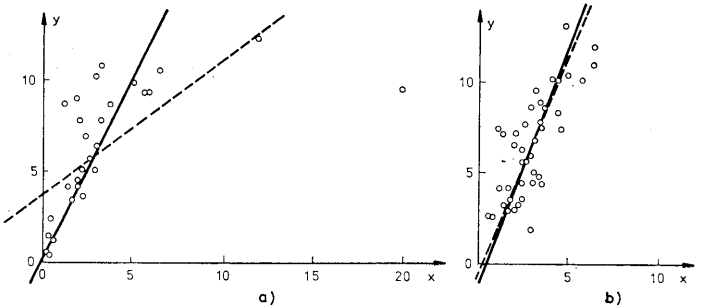
A másfél száz esztendő még a történelemben is hosszú idő, nemhogy a matematikai statisztika tudománytörténetében, így nem csodálkozhatunk azon, hogy a GAUSS-eloszlásnak hibaeloszlásként való hosszú ideig állandó (és mint láttuk, a hatékonyság optimuma szemszögéből *indokolt)* feltevése az elméleti és gyakorlati szakemberek között egyaránt úgy rögződött, hogy a hibaeloszlás a *valóságban is* GAUSS-eloszlás. Ez a dogmává merevedett (és ma már az esetek többségében káros) nézet annyira része a köztudatnak, hogy az ennek ellentmondó megállapításokat, amelyek a szakirodalomban egyre gyakrabban olvashatók, bizonyos kétkedés fogadja. Megnyugtatásul legyen szabad felhívni arra a figyelmet, hogy azok a szakemberek, akik valamely szakterület gyakorlati adatrendszereivel *és* a matematikai statisztika elméletével *egyaránt* foglalkoztak, nem a GAUSS-eloszlás túlnyomó előfordulásának nézetét vallották *régebben sem* (l. a *II. táblázat* néhány idevágó idézetét). Külön figyelmet érde-



2. ábra. A leggyakoribb értékek számítási módszerének legrégebbi előzménye eddigi ismereteink szerint SHORT 1761-es közleményében található. SHORT módszere a lépcsős függvény szerinti súlyfüggvénynek felel meg, a leggyakoribb értékek számítása a törésmentes görbével ábrázolt súlyfüggvénnyel történik. — A nyilak $f_4(x)$ eloszlásból származó 20 elemű ideális minta elemeinek felelnek meg

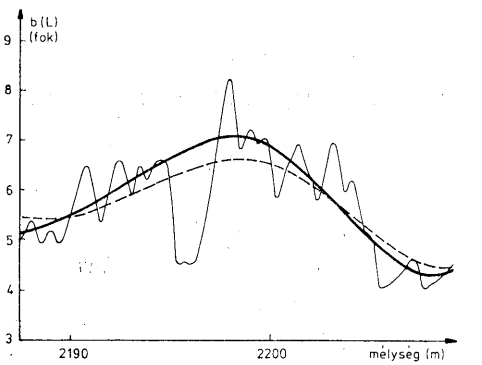
Fig. 2. The earliest records concerning the method of calculating the most frequent values are to be found, as far as our knowledge goes, in a paper by SHORT published in 1761. SHORT's method corresponds to the weight-function relative to a step function, the most frequent values being calculated by a weight-function represented by a curve with no break. — The arrows correspond to the elements of an idealized 20-element sample deriving from a $f_4(x)$ distribution

mel az, hogy a *csillagász* NEWCOMB és a *geofizikus* JEFFREYS szolgáltak az eloszlásokra vonatkozóan konkrét utalásokkal: az *1. ábrán* látjuk, hogy NEWCOMB (1886) által a hibaeloszlásra megadott analitikus alak gyakorlatilag az $a = 5{,}3$-hoz tartozó $f_a(x)$-eloszlásnak felel meg, JEFFREYS (1961) szerint pe-



3. ábra. Kétféle fémtartalom (x-szel és y-nal jelölve) szomszédos mélységszakaszokra (a és b). Folytonos egyenesek mutatják a leggyakoribb érték szerinti kiegyenlítés eredményeit, a szaggatottak a legkisebb négyzetek elve alapján kapott eredményeket. Az utóbbiakat nagymértékben befolyásolhatják kieső adatok (outliers): az a) mutatja, hogy két adatpár elég ahhoz, hogy teljesen eltorzítsa a legkisebb négyzetes eredményeket (a folytonos vonalak mindkét esetben arányosságra utalnak)

Fig. 3. Two different metal contents (x and y, respectively) for adjacent depth intervals (a and b). Solid straight lines indicate the results of adjustment according to the most frequent value, the dashed lines represent the results obtained on the basis of the least squares. The latter may be largely influenced by the outliers: a) indicates that two pairs of data are sufficient to distort the least square results (the continuous lines refer to proportionality in both cases)



4. ábra. Lyukferdeség-szelvény simítása spline-függvény számításával, irreálisan gyors változások kiküszöbölése céljából. Vékony vonal: mért ferdeség-szelvény; szaggatott vonal: spline-kiegyenlítés a legkisebb négyzetek elve szerint; vastag folytonos vonal: spline-kiegyenlítés a leggyakoribb értékek szerint

Fig. 4. Smoothing of borehole inclination by calculating a spline function, in order to eliminate irreally rapid changes. Thin solid line: measured inclination profile; dashed line: spline adjustment according to the least squares principle; thick solid line: spline adjustment according to the most frequent values

dig a legjobb esetben is általában csak a 6 és 10 közötti intervallumban levőnek adódik az $f_a(x)$-eloszlásnak tekintett hibaeloszlás a típusparamétere (azaz nem több ennél; a $\rightarrow \infty$-re adódna GAUSS-eloszlás). De nem csodálkozhatunk azon, hogy csillagász és geofizikus vette magának azt a fáradságot, hogy az eloszlástípust meghatározza: akár egy üstökös áthaladására, akár egy földrengésre vonatkozzanak is az adatok, teljes képtelenség az esemény „megismétlésével" több adathoz, és ezáltal pontosabb eredményhez jutni, így a valóságos eloszlásra nézve nagyobb hatásfokú algoritmus alkalmazása lehetett egyedül a járható út a nagyobb pontossághoz az ilyen jellegű vizsgálatoknál, — még ha ez akkoriban nagyon sok munkát igényelt is. (JEFFREYS (1932) arról számol be, hogy az akkori mechanikus számológépekkel egyetlen iterációs lépés végrehajtásához 6 óra volt szükséges.)

Amikor a hatásfoknövelés új lehetőségeiről beszélünk, az „új" jelző elsősorban relatíve értendő a statisztikában, így speciálisan a geostatisztikában is, a jelenleg legelterjedtebben alkalmazott, (kimondva vagy kimondatlanul) a legkisebb négyzetes elvre épülő módszerekhez viszonyítva. De jogosan látszik alkalmazhatónak az „új" jelző a Nehézipari Műszaki Egyetem Geofizikai Tanszékén tömörülő, de számos kültagot számláló team által geostatisztikai alkalmazásra is megfontolásra ajánlott koncepció *egészének* vonatkozásában Közvetlen gyakorlati értéke ugyanis csak olyan koncepciónak ill. eljárásnak lehet, amelynél bizonyos kritériumok egyidejű teljesülése biztosított (l. erre nézve a *III. táblázatot)*, amely tehát nem csak ad hoc felvetett ötlet, hanem mintegy „teljes vértezetében" jelentkezik. (A szakirodalomban nagyon sok ad hoc gondolattal találkozunk, amelyek távol állnak attól, hogy azokra a *III. táblázat* kritériumai egyidejűleg teljesüljenek.)

A leggyakoribb értékek szerinti kiegyenlítés koncepciója nagyon sok építőelemében természetes, kézenfekvő elgondolás, így az egyes mozaik-kockák analógiáit megtalálhatjuk — éppen az elgondolások természetességének az iga-

A statisztikai eljárásokkal szemben támasztott követelmények
Requirements regarding statistical procedures

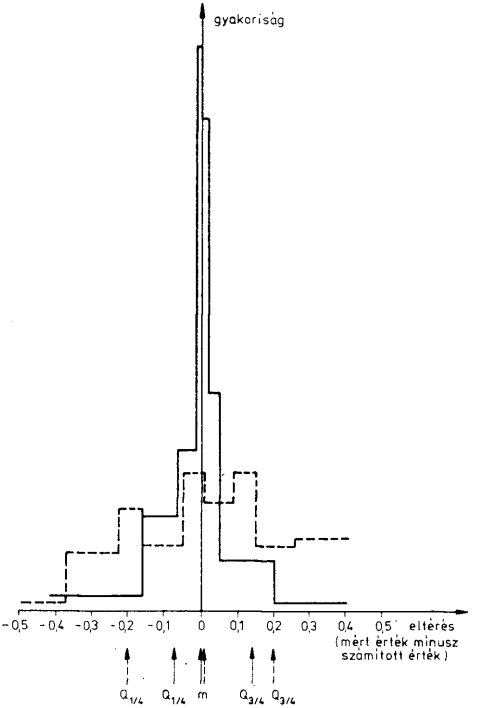
III. táblázat—Table III.

Matematikai statisztikai elvek és meghatározási módszerek kapcsolata; hasonlóságok és különbségek
Relationship of mathematical statistical principles and methods of determination; similarities and dissimilarities

As eredményesség és az általános gyakorlati alkalmazhatóság kritériumai. (Minek kell *együtt* adottnak lennie az alkalmazó szemszögéből egy matematikai statisztikai eljárásnál ahhoz, hogy az eredményesen és általánosan legyen alkalmazható?).
Criteria of efficiency and overall practical applicability. (What ought to be given as prerequisites from the aspects of the user of a mathematical statistical procedure in order to use it efficiently and universally?).

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 esetére
ELOSZLÁSMODELL-CSALÁD	álljon rendelkezésre a valóságban előforduló valószínűségeloszlás-típusok minél adekvátabb modellezése céljából egy kellően általános, de lehetőleg egyszerűen kezelhető modelleloszlás-család
NAGY HATÁSPOK	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 gépidőráfordítással 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.

zolására — akár a távolabbi múltban is. Legyen erre példa a *2. ábra*, ahol alul egy $f_4(x)$ anyaeloszlásból származó, 20 elemű ún. ideális mintát mutatunk be, fölötte a leggyakoribb értékszámítás súlyfüggvényének a görbéjével. — Legelőször azt figyeljük meg, hogy a minta szélén a súlyok már lényegesen kisebbek a maximális súlyértéknél, hiszen e szélső adatok ingadozása a legnagyobb. A valószínűségelmélet egyes teoretikusainak feltevésével ellentétben ui. — amely szerint a statisztikus ingadozásokat végtelen sok végtelen kicsiny hatás szuperpozíciója hozza létre, — nagyon jól tudjuk, hogy bizonyos ritkán előforduló, és éppen ezért a gyakorlati kezelhetőség kedvéért elhanyagolt körül-



5. ábra. A mért értékektől való eltérések gyakorisági diagramjai, többváltozós kiegyenlítés után. Folytonos vonal: leggyakoribb érték szerinti kiegyenlítés, szaggatott vonal: hagyományos (legkisebb négyzetes) kiegyenlítés. A bejelölt alsó és felső kvartilisok mutatják, hogy a valószínű hiba lényegesen kisebb a leggyakoribb érték szerinti kiegyenlítésnél

Fig. 5. Diagrams of frequency of deviations from the measured values, after multivariable adjustment. Solid line: adjustment according to the most frequent value, dashed line: conventional (least square) adjustment. The lower and upper quartiles marked in show that the probable error is substantially smaller in case of the adjustment according to the most frequent value

mény igenis véges — sőt, esetleg durva hibát is előidéző — hatást fejthet ki, és geostatisztikai algoritmusainknak ekkor sem szabad zavarba jönniök, azaz jelentős hibájú, vagy teljesen használhatatlan eredményt adniok. Ezt biztosítja a súlyfüggvény bemutatott görbéje a leggyakoribb értékek szerinti kiegyenlítéseknél, — de mennyire közel van ehhez az a lépcsős függvény, amit SHORT (1763) már két és negyed századdal ezelőtt alkalmazott! (SHORT egyébként szintén csillagász volt.) De nem kell csodálkoznunk: a valóságos adatrendszerek adekvát kezelése a hatásfok maximumára törekedve nyilván a jelenleg javasolthoz hasonló algoritmust kellett, hogy kézenfekvővé tegyen bármikor, így két és negyed évszázaddal ezelőtt is.

Befejezésül még legyen szabad három példát bemutatnom. A *3a. ábra* a leggyakoribb értékek szerinti kiegyenlítés (l. pl. STEINER, 1985) legelső alkalmazása: kétféle fémtartalom összefüggése egy előfordulás valamely mélységszintjére vonatkozóan. A legkisebb négyzetes kiegyenlítés eredményét (szaggatott vonal) „elhúzzák" a kieső pontok, míg a leggyakoribb érték szerinti kiegyenlítést ezek nem zavarják (folytonos vonal). — Az utóbbi helyességét támasztják egyébként alá a szomszédos mélységszint adatpárjai is (l. a *3b. ábrát).*

A leggyakoribb értékek szerinti kiegyenlítés ugyanúgy figyelembe tudja venni az eredmények között esetleg szigorúan megkövetelendő összefüggéseket, mint a legkisebb négyzetes módszer. (A legközeletűbb példa erre geodéziai jellegű: a háromszög szögeinek 180°-nak kell lenniök.) — Ennek egyik gyakorlatilag fontos következménye az, hogy a mereven előírt analitikus alak kényszerétől mentes, ún. spline-kiegyenlítés is végrehajtható az új módszerrel. A HURSÁN-TAKÁCS (1986) alapján rajzolt *4. ábrán* ferdeség-szelvény egy szakaszát látjuk, ahol a kis mélységszelvény-különbségekhez tartozó nagy változások semmiképpen nem lehetnek reálisak (pontosabban a mérőeszköz saját folyamatait tükrözik), így a spline-kisimítás feltétlenül indokolt. A leggyakoribb értékek szerinti kiegyenlítés olyan görbét eredményez, amit egy előítéletmentes értékelő kézzel is berajzolna —, a legkisebb négyzetes esetben azonban az eredménygörbét elhúzza egy különösen nagy amplitudójú ingadozás. A két eredmény közötti különbség a változás teljes tartományának mintegy 10%-a, tehát egyáltalában nem hanyagolható el.

Az utolsó példa komplex mélyfúrási adatrendszerekre alkalmazott kétféle (legkisebb négyzetes, ill. leggyakoribb értékek szerinti), hatváltozós másodfokú kiegyenlítés kétféle eltérésrendszerének két gyakorisági diagramja az *5. ábrán* (FERENCZY-TAKÁCS [1986] egyik ábrája alapján). Látjuk, hogy a *legkisebb négyzetes kiegyenlítés* olyan együtthatórendszert szolgáltatott, amelylyel a meghatározandó mennyiség *valószínű hibája csaknem kétszer akkora,* mint ha a kiegyenlítést a javasolt új módszer szerint hajtjuk végre, és a tárolóparaméter helyes értékét ennek megfelelően számítjuk a hatféle szelvényadatból.

Irodalom — References

CSERNYÁK L. (1973): On the most frequent value and cohesion of probability distributions — Acta Geodaet., Geophy. et Mont. Acad. Sci. Hung. 8. (3—4).

FERENCZY L.—TAKÁCS E. (1986): Valószínűségelméleti alapokon nyugvó kvantitatív karotázs interpretációs rendszer hatékonyságának és megbízhatóságának növelése — Miskolc. Jelentés. (Kézirat).

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

HURSÁN L.—TAKÁCS E. (1986): A lyukferdeségmérések kiértékelésének új lehetőségei — Miskolc. Jelentés. (Kézirat)

JEFFREYS, H. (1932): An alternative to the rejection of observations — Proceedings of the Royal Soc. of London Ser. A. 137.

JEFFREYS, H. (1961): Theory of Probability. Oxford. Clarendon Press.

NEWCOMB, S. (1886): A generalized theory of the combination of observations so as to obtain the best result — American Journal of Mathematics, 8.
PRÉKOPA A. (1962): Valószínűségelmélet műszaki alkalmazásokkal. Budapest. Műszaki Könyvkiadó.
SHORT, J. (1763): Second paper concerning the parallax of the suc etc. — Philos. Trans. Roy. Soc. London, 53.
STEINER F. (1973): Most frequent value and cohesion of probability distributions — Acta Geodaet., Geophys. et Mont. Acad. Sci. Hung. 8. (3—4).
STEINER F. (1985): Robusztus becslések. Budapest. Tankönyvkiadó.

A kézirat beérkezett: 1987. VI. 15.

Need and possibilities for increasing the efficiency of geostatistical calculations

*F. Steiner**

From the aspects of the probability theory, the geological and geophysical data may vary heavily in type of distribution. Their statistical processing must be optimal (unlike it is the case with the conventional least squares principle), for, e.g. using an algorithm of merely 50% efficiency would equal discarding the half of one's data acquired at a high cost. Although modern methods require more calculations, but, given the steady reduction of price per one operation, the data acquisition + calculation complex cannot be economically rentable unless up-to-date statistical methods are used.

Using simplifications (that are unavoidable), the author presents the basic statistical principles, the nature of the calculations based thereon and their specific features. A timetabulation is used to visualize the trends of development, with simultaneous presentation of the opinions on distributions met in the practice, opinions (not changing in time) held by researchers who dealt with both practical data systems and the theory of mathematical statistics. — The members of model distribution family $f_a(x)$ provide good approximations to distributions occurring in geology and geophysics. For this reason by using the most-frequent-values-calculations based on these distributions and/or adjustments based upon such a principle, geostatistical techniques of high efficiency can be developed.

Manuscript received: 15th June, 1987.

Необходимость увеличения эффективности и новые возможности статистических расчетов в геологических науках

Ф. Штейнер

Геологические и геофизические данные с точки зрения теории вероятностей могут характеризоваться различными типами распределений. Необходимо стремиться к их обработке статистически оптимальным способом, отклоняющимся от традиционного принципа наименьших квадратов, ибо, например, применение альгоритма с эффективностью в 50% равноценно отказу от половины данных, приобретенных в результате существенных затрат. Хотя в современных методах требуется применение большого объема расчетов, при все снижающейся стоимости отдельно взятой операции приобретение данных в комплексе с расчетами все чаще становится экономичным и достаточно эффективным лишь при использовании современных статистических методов.

В статье с неизбежными упрощениями представлены принципы статистики, а также характер основывающихся на них расчетов и их особенности. Хронологической таблицей иллюстрируются направления усовершенствования; в ней представлены точки зрения исследователей, в равной мере занимавшихся теорией математической статистики и системами эмпирических данных, по распределениям, встречающимся на практике. Члены семейства распределений по модели $f_a(x)$ дают хорошее приближение распределениям, встречающимся в геологии и геофизике, поэтому путем расчета наиболее частых значений в соответствии с таким распределением или же путем выравниваний, основывающихся на этом принципе, могут быть разработаны высокоэффективные геостатистические способы.

* Technical University for Heavy Industry, Department of Geophysics, H-3515 Miskolc-Egyetemváros, Hungary