A variancia-meghatározás hibái különböző típusú valószínűségeloszlásoknál[1]

STEINER FERENC[2]

A variancia (az L_2-norma minimális értékének négyzete, azaz a szórásnégyzet) a klasszikus statisztika egyik alapmennyisége, ezért meghatározási hibáinak az ismerete minél több hibaeloszlás-típusra vonatkozóan alapvető fontosságú. Különösen igaz ez a klasszikus geostatisztikára, ahol a varianciával definiált (általában $\gamma(h)$-val jelölt, szemivariogramnak is nevezett) variogram fundamentális szerepet játszik.

Ha a szórást (σ_{VAR}-t) választjuk a VAR-ral jelölt variancia meghatározási hibájának a jellemzésére, nagy n mintaelemszámoknál ezt egyszerűen számíthatjuk $\sigma_{VAR} = A_{VAR}/\sqrt{n}$ -ként, mivel a varianciák A_{VAR} aszimptotikus szórását a klasszikus statisztika (a negyedik momentum segítségével) szintén egyszerű formulával adja meg. E dolgozatban azonban fény derül arra, hogy A_{VAR} értéke a földtudományok hibaeloszlásainak kb. csak egynegyedére véges, és végtelen értékű az irodalom szerint a geostatisztikában leggyakrabban várható eloszlástípusra is. Eszerint a (szerencsére csak látszólagos) ellentmondás szerint a $\gamma(h)$ variogramgörbe pontjainak számításánál az esetek többségében olyan nagy hibával szembesülünk, amely már a számítási eredmény realitásának a kérdését is felveti.

A jelen dolgozat kimutatja, hogy ha a szórás helyett a varianciák hibáit interszextilis félterjedelmükkel mérjük, minden olyan típusra teljesülőnek találjuk a nagy számok törvényét (azaz a hiba az n mintaelemszám növelésekor csökken), amelyre a varianciának magának véges az értéke.

F. STEINER: Errors of the variance-determinations for different parent distribution types

As the variance (the square of the minimum L_2-norm, i.e., the square of the scatter) is one of the basic characteristics of the conventional statistics, it is of practical importance to know the errors of its determination for different parent distribution types. This statement is outstandingly valid for the geostatistics because the $\gamma(h)$ variogram (called also as semi-variogram) is defined as the half variance of some quantity difference (e.g. difference of ore concentrations in function of the h distance of the measuring points) and this $\gamma(h)$-curve plays a basic role in the classical geostatistics.

If the scatter (σ_{VAR}) is chosen to characterise the determination uncertainties of the variance (denoted the latter by VAR), this can be easily calculated as the quotient $A_{VAR}/\sqrt{n}$ (if the number n of the elements in the sample is large enough); for the so-called asymptotic scatter A_{VAR} a simple formula is known (containing the fourth moment). The present paper shows that the A_{VAR} has finite value unfortunately only for about a quarter of distribution types occurring in the earth sciences, it must be especially accentuated that A_{VAR} has infinite value for that distribution type which most frequently occurs in the geostatistics.

It is proven by the present paper that the law of large numbers is always fulfilled (i.e., the error always decreases if n increases) for the error-determinations if the semi-intersextile range is accepted (instead of the scatter); the single (quite natural) condition is the existence of the theoretical variance for the parent distribution.

1. Bevezetés

A múlt század közepén dél-afrikai aranybányák készletszámítására KRIGE szakmai (elsősorban bányász-) körökben nagy visszhangot kiváltó, szellemes eljárást vezetett le valószínűségelméleti alapon a szükségképpen csak diszkrét pontokban ismert érckoncentráció-értékek interpolálására. Az új módszert elméletileg is tanulmányozták (ill. megalapozták), valamint gyakorlatilag is továbbfejlesztették (előbbire MATHERON [1965], utóbbira JOURNAL, HUIJBREGTS [1978] szolgálhat példaként); a létrejött eredmény a klasszikus geostatisztika gerincének tekinthető.

Az ezen a szakterületen dolgozók aligha igénylik a módszer ismertetését; többi olvasómnak legyen szabad tankönyvemet [STEINER 1990] ajánlanom, ahol magyar nyelven az eljárás ismertetése (valószínűségelméleti levezetéssel és egy példa részletes bemutatásával együtt) megtalálható. Ebből kiderül, hogy a krigelés kulcsfogalma a (szemivariogramnak is nevezett) variogram, amelynek definíciója:

$$\gamma(h) = \frac{1}{2} VAR\left[Z(\vec{x}) - Z(\vec{x}+\vec{h})\right], \qquad (1)$$

ahol Z a (pl. térbelileg) interpolálandó jellemző mennyiség (érctartalom vagy egyéb), az $\vec{x}$-tartomány a vizsgált térrészt definiálja, a szögletes zárójelben pedig azon pontpárokra vonatkozó jellemző mennyiségek értékkülönbsége szerepel, amely pontok h távolságra vannak egymástól. *VAR* nyilván a varianciát jelenti, ami köztudomásúan a szórásnégyzettel, azaz a minimális értékű L_2-norma négyzetével azonos. A $\gamma(h)$ kulcsszerepe a krigelés végrehajtásakor abban áll, hogy a $\gamma(h)$ maximális értékéből az aktuális h ponttávolsághoz tartozó $\gamma(h)$-t levonva, az ilyen távolságban mérhető Z értékpárok (c_{ik}-val jelölt) kovarianciáit kapjuk (ld. pl. STEINER [1990] 291. oldalán a 8.6 ábrát), márpedig ezek a c_{ik} értékek annak a *Krige-mátrixnak* az elemei, amelynek (ill. inverzének) alapján az interpoláció végeredménye: a vizsgált térrész bármely, be nem mért pontjára vonatkozó Z érték már elemi úton számítható.

[1] Beérkezett: 2000. október 16-án

[2] Miskolci Egyetem Geofizikai Tanszék,
H-3515 Miskolc, Egyetemváros

A fentiekből következik, hogy akár speciálisan krigeléskor, akár a címben megjelölt, sokkal általánosabb témakörben is a klasszikus statisztika kellős közepén vagyunk: az L_2 eltérésnorma iménti felmerülése már erre figyelmeztethetett. Márpedig e sorok írója az elmúlt években–évtizedekben nem a klasszikus, azaz az $f_G(x) = (2\pi)^{-1/2}\sigma^{-1} \cdot \exp\left[-x^2/2\sigma^2\right]$ valószínűség-sűrűségfüggvényű hibaeloszlások gyakori (vagy pláne kizárólagos) előfordulását feltételező statisztika területén tevékenykedett (f_G jelölésében a G index arra utal, hogy a Gauss-féle típusról, az ún. *normál*eloszlás sűrűségfüggvényéről van szó, amelynek fenti kifejezésében σ a variancia gyökét, a szórást, azaz az L_2-norma minimális értékét jelenti), hanem a statisztikát a gyakorlat igényeinek megfelelő sokkal nagyobb általánossággal művelte, hibaeloszlásként túlnyomóan az

$$f_a(x) = \Gamma(a/2) \cdot \pi^{-1/2} \cdot \Gamma^{-1}[(a-1)/2] \cdot \cdot S^{-1} \cdot \left[1 + (x/S)^2\right]^{-a/2} \qquad a > 1 \qquad (2)$$

sűrűségfüggvényekkel definiált szupermodell valamelyik típusát feltételezve. (Már a (2) kínálta sokféle $f_a(x)$-típus is jelzi az általánosság mértékét, egyben azt is, hogy ez az $f_G(x)$ kizárólagos előfordulását feltételező klasszikus szemlélet általánosítása is egyben, hiszen bizonyítható, hogy $a \to \infty$ esetén $f_a \to f_G$, azaz a Gauss-típus is az $f_a(x)$ szupermodell elemének tekinthető.) Miért e *„pálfordulás"* a szerző részéről a klasszikus statisztika irányában?

E kérdés megválaszolása előtt legyen szabad a szerzőnek egy teljesen szubjektív, ezért apró betűs bekezdést ideiktatnia, — talán kismértékben a SZILASI [2000]-ből vett következő megjegyzés hatására is: *„...én tényleg, soha, sehol nem olvastam olyan utasítást, hogy a tudományos szövegnek egyben ... feltétlenül unalmasnak is kell lennie."* — Abban a szerencsében volt részem, hogy egyetemi tanulmányaim során a valószínűség elméletét SZŐKEFALVI-NAGY Béla professzor előadásában hallgathattam annak következményeként, hogy RIESZ professzorral közösen írt könyvük [RIESZ, SZŐKEFALVI-NAGY 1952] megjelenésével egy csapásra lett fiatalon világhírűvé, így a JATE Matematikai Intézete nem tehette meg, hogy bizonyos tárgyak előadását ne őrá bízza és az viszont nyilván azonnal nem volt megoldható, hogy ezek mind abból a témakörből valók legyenek, amelyek világhíressé tették. Nem tudhatom, hogy a valószínűségelméletet mennyire szívesen adta elő — feltehetően szívesebben, mint az ábrázoló geometriát, amelynek ezt megelőzően kényszerűen szintén előadója volt, — de nagyon jól jártam, hogy olyan zsenitől (és nem a klaszszikus valószínűségelmélet lezártságában bigott módon hívő középkádertől) ismertem meg a valószínűségelmélet alapjait, aki bizonyos távolságtartással és az őrá oly jellemző csípős megjegyzésekkel sem fukarkodva adta elő ezt a tárgyat. Nehéz lenne megítélni, hogy a fentieknek milyen mértékben volt szerepe abban (valami tudat alatti csatornán keresztül), hogy egy zsenge tanársegédként írt cikkem [STEINER 1959] szakmai tartalmának statisztikai komponense az L_2-norma alkalmazásától való elszakadást jelentette, — és ráadásul ezzel a lépéssel egy Egyed-féle jó alapgondolat [EGYED 1955] megszabadult a gyakorlati alkalmazhatóságot is már időnként megkérdőjelező nagy bizonytalanságtól, amelyet az 1955-ös cikkben az L_2-alapú statisztika alkalmazása okozott. Sietek hozzátenni, hogy az idézett cikkem írásakor csak a *józan paraszti ész* logikáját követtem, és ez teljesen ösztönösen vezetett engem a legkisebb négyzetek klasszikus elvétől való eltérésre. Több évtized távolából is hálával tartozom EGYED professzornak, aki ezt a klasszikus statisztikától való ösztönös elszakadásomat egy beszélgetésünk során tudatossá tette bennem, bizonyára ezzel is előkészítve a talajt ahhoz, hogy már 1965-ben (gravitációs témakörben) írt kandidátusi értekezésemben bátorkodtam kimondani a *legnagyobb reciprokok* elvét (ld. pl. STEINER (*ed.*) [1997], 367. oldal), amelynek alapján robusztus és rezisztens eljárások származtathatók akárhány ismeretlen paraméter meghatározására általános esetben is. Alig titkolható büszkeséggel teszem hozzá, hogy egyetlen évvel korábban jelent meg az a dolgozat [HUBER 1964], amelyet startként szokás elfogadni a robusztus statisztika fejlődéstörténetében, de ez a Hubercikk még nem szolgált általános esetekre alkalmazható algoritmusokkal: belül marad a helyparaméter-meghatározások problémakörén.

A fenti apró betűs bekezdés előtt feltett kérdést az tette látszólag indokolttá, hogy a klasszikus statisztika egy alapmennyisége e dolgozat vizsgálódásának a tárgya. A vizsgálati módszer azonban a klasszikus statisztikáénál tágabb horizontú lesz, nemcsak az eredményesség érdekében, hanem azért is, mivel egyre több geofizikus kolléga bír már a klasszikusnál általánosabb statisztikai szemlélettel (ld. pl. az L_1-normán alapuló módszerek egyre nagyobb térhódítását), ennek megfelelően a *„normális eloszlás széles körű felléptét"* (PRÉKOPA [1962], 289. oldal 5. szövegsor) is feltehetően egyre kevesebben tételezik fel, a modern statisztika eszköztára pedig — látszólagos ellentmondásokat feltárva — esetleg olyan esetekben is elbizonytalanítható a gyakorlati szakembert, amikor az nem indokolt. Amennyire lehet, ilyen eseteknek elejét kell venni (úgy érzem, hogy ennek elősegítése kötelessége a gyakorlati statisztika általános szemléletű elméletével foglalkozó szakembereknek). A variogram-, vagy általánosabban a variancia-meghatározások vonatkozásában a jelen dolgozat ezt a célt kívánja elérni.

2. A variancia becslése

Az (1) egyenletben szereplő $\gamma(h)$-t, vagy általánosan: valamilyen ξ valószínűségi változó pontos variancia-értékét nyilván csak becsülni tudjuk, ha (mint a gyakorlatban szinte mindig) csak n db x_i mérési adatra támaszkodhatunk. A pontos érték persze egyszerűen meghatározható az $f(x)$ valószínűség-sűrűségfüggvény ismeretében a következő integrállal:

$$VAR(\xi) = \int_{-\infty}^{\infty} (x-E)^2 f(x)\mathrm{d}x, \qquad (3)$$

ahol E a ξ várható értéke, azaz $E = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x$. Ha x statisztikai ingadozást jelent, az esetek túlnyomó többségében indokolt a zérus értékre szimmetrikus $f(x)$-et várni (a (2)-ben definiált szupermodell mindegyik hibatípusmodellje ilyen, de az (1)-ben szereplő Z differenciákra is ez áll), s ekkor egyszerűen

$$VAR = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x \qquad (4)$$

a variancia pontos értékét definiáló formula. Minták esetében (azaz n db, az $f(x)$-ből véletlenszerűen kapott x_i adatra) a (4) formulát a statisztikában megszokott szabályokkal írjuk át összegformulává:

$$VAR = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \ , \qquad (5)$$

tudva persze azt, hogy a bal oldal most már nem pontos értéket jelent, csupán annak becslését. (A $\gamma(h)$ (1) kifejezésében minden rögzített h távolsághoz az (5)-beli x_i-k az összes lehetséges módon képzett Z különbségeket jelentik — tehát csak közvetve van közük az (1)-beli, rádiuszvektor értelmű $\underline{x}$ vektorhoz, — és ha a bemért pontok között $n(h)$ db, egymástól h távolságra levő pontpár van, a VAR-becslés (5) kifejezésébe n helyett nyilván $n(h)$-t kell írnunk mindkét helyen. Így számítják a gyakorlatban az összes lehetséges h-hoz $\gamma(h)$ értékeit — pl. valamely adott, mondjuk térbeli pontrács sarokpontjaiban mért Z értékek alapján. A becslésjelleg miatt nem várható sima görbe; az ingadozó pontokat ezért célszerűen analitikusan adott formula szerinti görbét feltételezve egyenlítik ki, talán leggyakrabban a *szférikus modellnek* nevezett $\gamma(h) = C\left[1{,}5 \cdot h/H - 0{,}5(h/H)^3\right]$ függvényt alkalmazva a $0 \le h \le H$ tartományon, ahol H hatástávolságot jelent: ez az érték az a legkisebb h távolsága azon pontpároknak, amelyeknél a mért Z értékek kovarianciája már zérus. Ha $h > H$, $\gamma(h)=C$, ld. újra a STEINER [1990] 291. oldalán a 8.6 ábrát. Hogy egy ilyen megnyugtatóan sima görbe milyen nagymértékben ingadozó pontok kiegyenlítéséből adódott, azt talán néha jobb nem is tudni; mindenesetre az imént idézett könyv 296. oldalának 8.9 ábráján első pillanatban ijesztő pontfelhőt látunk. A pontok ilyen mértékű diszperziója azonban nem tekinthető tipikusnak: a krigelés részletes, ugyanakkor áttekinthető bemutatása szükségképpen összességében is kisszámú Z adatot igényelt, így azután az $n(h)$ értékek nagyon kicsinyek lettek, következésképpen heurisztikusan is indokoltnak fogadhatjuk el az (5) szerint számított pontok nagy statisztikai ingadozását. Emellett nagy, a vizsgált térrész méretével összemérhető h értékeknél az (5) szerinti számításoknál torzítás is felléphet az $n(h)$-k extrém kicsiny volta miatt, mégpedig a $\gamma(h)$ pontbecslések csökkenését eredményezve. A klasszikus statisztikából azonban régóta ismert a torzításmentes varianciabecslés formulája is, amely általános esetben

$$VAR = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (6)$$

alakban írható, ahol $\bar{x}$ az x_i-k számtani átlagát jelenti. Az alább ismertetendő Monte Carlo-számítások mindegyikében így történt a *VAR*-becslések számítása.

3. A varianciabecslések aszimptotikus szórása

Ha valamely statisztikai mennyiséget elég sok adatból számítunk ki, akkor ennek a b-vel jelölt becslésnek a σ_b-vel jelölt szórását szerencsés esetben $\sigma_b = A/\sqrt{n}$ -ként számíthatjuk (ahol A-t a b aszimptotikus szórásának nevezzük és n az adatok száma).

A (4)-ben definiált *VAR* kifejezést a klasszikus statisztikában második centrális momentumnak is nevezik és ekkor m_2-vel jelölik. Az r-edik centrális momentum (m_r) kifejezés per analogiam

$$m_r = \int_{-\infty}^{\infty} x^r f(x)\mathrm{d}x, \qquad (7)$$

de ezek a momentumok origóra szimmetrikus $f(x)$ sűrűségfüggvényeknél, amelyekkel a hibaeloszlásokat is modellezzük, nyilván csak páros r-eknél szolgáltatnak zérustól eltérő értékeket.

A variancia A_{VAR}-ral jelölt aszimptotikus szórásának formulája a következő (ld. pl. CRAMÉR [1945]):

$$A_{VAR} = \sqrt{m_4 - m_2^2}\ , \qquad (8)$$

tehát csak az r=4-hez és r=2-höz tartozó momentumok ismerete szükséges A_{VAR} meghatározásához.

Vegyük alapul hibaeloszlásként a (2)-ben definiált $f_a(x)$ szupermodell valamely a-hoz tartozó típusát standardizált alakban, azaz S=1-et helyettesítve. Az utóbbi lényegtelen, azaz eredményekre hatással nem levő, pusztán egyszerűbb, így egyben áttekinthetőbb írásmódot lehetővé tevő művelet interpretálható úgy, hogy, mivel x egységét tetszőlegesen választhatom meg, így standardizáláskor csupán az történik, hogy történetesen az S skálaparamétert választom x egységének. (Ha vizsgálódásaink után az eredeti egységrendszerre akarunk visszatérni, akkor a standardizált esetre vonatkozó *VAR*-t, A_{VAR}-t, vagy bármilyen, x^2 dimenziójú jellemzőt S^2-tel kell szoroznunk.)

A (2) szupermodell alapulvétele általában is, így most is nemcsak azért előnyös, mert a típusok rendkívüli sokféleségét kínálja (a gyakorlat számára túlnyomóan már túl rövid szárnyú modellnek bizonyuló Gauss-típustól egészen az extrém nagy súlyú szárnytartományokkal rendelkező típusokig), hanem azért is, mert $f_a(x)$ egyszerű analitikus alakja gyakran vezet elemien egyszerű formulákra. Utóbbira talán az egyik legmeggyőzőbb példa éppen a (7) szerint számított m_r-eknek az $f_a(x)$ szupermodellre adódó formulája:

$$m_r = \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (r-1)}{(a-3)\cdot(a-5)\cdot\ldots\cdot[a-(r+1)]} \qquad (9)$$

(ld. pl. STEINER (ed.) [1997] 52. oldalán a (2-5) formulát).

Azonnal látható (9)-ből, hogy magának a varianciának (azaz m_2-nek) az értéke

$$VAR = \frac{1}{a-3}, \qquad (10)$$

s mivel (szintén (9)-ből)

$$m_4 = \frac{3}{(a-3)\cdot(a-5)}\ , \qquad (11)$$

következésképpen a variancia aszimptotikus szórása (ld. (8)-at)

$$A_{VAR} = \frac{\sqrt{2}}{a-3}\cdot\sqrt{1+\frac{3}{a-5}} \qquad (12)$$

alakban írható fel. Ha A_{VAR} értéke a *VAR*-hoz viszonyítottan érdekel bennünket, nyilván az

$$A_{VAR}/VAR = \sqrt{2}\cdot\sqrt{1+\frac{3}{a-5}} \qquad (13)$$

formulát fogjuk alkalmazni.

4. A geostatisztikában előforduló hibatípusok

DUTTER [1986/87] szerint a geostatisztikában leggyakrabban az a=5-höz tartozó $f_a(x)$ valószínűségeloszlás-típus előfordulása várható, amelynek sűrűségfüggvénye (2) szerint standard esetben

$$f_{st}(x) = \frac{3}{4} \cdot \frac{1}{\left(\sqrt{1+x^2}\right)^5} . \qquad (14)$$

(Az „st" index arra utal, hogy néhány csillagászati, geodéziai stb. adatrendszer típusvizsgálata is a=5 körüli értékre vezetett, így ez a típus *statisztikus eloszlásnak* nevezhető.) Mivel nincs okunk feltételezni, hogy a $\gamma(h)$ variogrampontok számításának alapjául szolgáló Z differenciák eloszlása nem áll közel a (14) szerinti statisztikushoz (hiszen utóbbinak a definiálásakor kiindulásként DUTTER fentebb idézett, éppen a geostatisztikára vonatkozó megállapítása szolgált), első pillanatban megdöbbenéssel állapítjuk meg, hogy (12)-ből ekkor a variancia aszimptotikus szórásának végtelen voltára kell következtetnünk. Ha ezt a körülményt úgy interpretálnánk, hogy hiába sűrítem (akár szinte minden ésszerű határon túl is) a mérési pontok számát pl. a $\gamma(h)$-meghatározáshoz, hogy annak $A_{VAR}/\sqrt{n(h)}$ szórása az $n(h)$ növelésével csökkenjen, igyekezetem hiábavaló: maga az A_{VAR} végtelen lévén, a szórás változatlanul végtelen marad már a (14) szerinti statisztikai eloszlásnál is (persze az $a<5$ típusparaméterekkel jellemzett eloszlásoknál ugyanúgy), amely f_{st}-ről pedig a (14) előtti sorok Dutter-idézete azt állította, hogy a geostatisztikában leggyakrabban ez a hibaeloszlás fordul elő.

Azokat az olvasóimat, akik eddig DUTTER nevével esetleg nem találkoztak, — pedig őt a fentiekben mértékadó tekintélyként idéztem, — legyen szabad arról tájékoztatnom, hogy ez a bécsi professzor a robusztus statisztika irodalmában ismert név, HUBER társszerzőségével is jelentetett meg cikkeket, — márpedig HUBERt a nemcsak fundamentális jelentőségű, már idézett HUBER [1964] tette kiemelkedően elismertté, hanem az is, hogy elsőként jelentetett meg monográfiát a robusztus statisztikáról [HUBER 1981]. A fentiek — azazhogy DUTTER kiváló elméleti szakember a robusztus statisztika általános témakörében, — még nem nyugtatnák meg e sorok szerzőjét afelől, hogy valóban kompetens bármely gyakorlati diszciplínára vonatkozóan típuseloszlások előfordulási valószínűségére bármilyen kijelentést tenni. Gondoljunk azoknak a klasszikus valószínűségelméletet művelő kiváló matematikusoknak az extrém nagy számára, akik a Gauss-típus igen gyakori — vagy akár kizárólagos — előfordulását hirdetik a hibaeloszlásokra vonatkozóan, sőt, a Ljapunov-tételre (ld. pl. PRÉKOPA [1962]) hivatkozva ezt bizonyítottnak is hiszik, megfeledkezve CRAMÉR nevezetes tételéről, amely szerint akárhány apró hiba szuperponálódása alakítja is ki a gyakorlatban tapasztalt hibaeloszlást, ez utóbbi csak akkor lehet Gauss típusú, ha minden egyes összetevője már maga is eleve Gauss típusú volt. — Azt hiszem, joggal mondhatjuk ki, hogy típuseloszlás-előfordulás valamilyen diszciplínán belüli gyakoriságára vonatkozó kijelentést csak olyan matematikustól fogadhatunk el, aki az adott diszciplínát is kellő mélységben ismeri. Nos, DUTTER a leobeni egyetemen geostatisztikát ad elő (ennek jegyzetét idéztem DUTTER [1986/87]-ként), így ez a feltétel teljesül.

A geofizikusok nagy szerencséjére talán JEFFREYS személye teljesíti a fent megfogalmazott kettős követelményt a legpregnásabban: kiváló matematikusként monográfiát írt a valószínűségelméletről [JEFFREYS 1961], ugyanakkor (mint gyakorlati statisztikus) a geofizika több szakterületén (pl. a szeizmológiában) mért adatrendszerekből hámozta ki a lényeges információkat, a lehető legnagyobb pontosságra törekedve. Utóbbit a klasszikus statisztika alkalmazásával nem érhette el, mert az adatrendszerek hibaeloszlásait sohasem találta Gauss típusúaknak, így a mért adatrendszerek optimális értékelésére iterációs eljárásokat fejlesztett ki és alkalmazott már a XX. század harmincas éveiben, — bár az akkori *„számítástechnika"* egész munkacsoportjának egy műszaknyi munkáját igényelte egyetlen iterációs lépés végrehajtásához. JEFFREYS a hibaeloszlások még legrövidebbnek talált szárnyait is a $6<a<10$ tartományba eső a típusparaméterű $f_a(x)$ eloszlásokkal találta modellezhetőeknek (idézi KERÉKFI [1978]). A $6<a<10$ típusintervallumot ezért *Jeffreys-intervallumnak* nevezzük; nos, ezen intervallum bármely típusára véges értéket ad (12) a varianciák A_{VAR} aszimptotikus szórására, de ez nem a leggyakrabban előforduló típusok tartománya. Itt nem részletezendő okokból célszerű volt az a=9 esetet mint ezen intervallum reprezentánsát kiemelni; ennek a Jeffreys-eloszlásnak nevezett típusnak a standard esetre vonatkozó, $f_J(x)$ -szel jelölt sűrűségfüggvénye (2) szerint a következő:

$$f_J(x) = \frac{35}{32} \cdot \frac{1}{\left(\sqrt{1+x^2}\right)^9} . \qquad (15)$$

Az f_J az f_{st}-hez viszonyítva nyilván az f_a-típusok rövid szárnyú tartományát képviseli (az f_G-vel jelölt, $a \to \infty$ -hez tartozó Gauss-eloszlás persze az f_a-típusok közül a legrövidebb szárnyú, de ilyen hibaeloszlással nemcsak JEFFREYS nem találkozott gyakorlati adatrendszereinél, de modern szerzők egész sora sem — legyen szabad itt elhagynom az erre vonatkozó idézetlistát). Nézzük az f_{st} *másik oldalát*: milyen nagy súlyú szárnyak fordulhatnak elő a geofizikában? STEINER (ed.) [1991] elektromágneses példát mutat be arra, hogy a típusmodellezéshez az

$$f_C(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} \qquad (16)$$

sűrűségfüggvényű Cauchy-eloszlásra is szükségünk lehet (azonnal látható (2)-ből, hogy ez az a=2 típusparaméterhez tartozó $f_a(x)$ eloszlás). Sőt, STEINER (ed.) [1997] egy példában bemutatja, hogy még az $a<2$ eset is előfordulhat adatrendszereinknél, de ezek csak sporadikus esetek lehetnek; már a Cauchy körüli eloszlások előfordulása sem gyakori, de semmiképpen nem elhanyagolható valószínűségű.

Jó lenne egy egyszerű analitikus függvénybe sűríteni mindazt, amit szavakkal a fentiekben kissé hosszadalmasan tudtam csak elmondani. Az áttekinthetőség fokozása érdekében vezessük be $t \equiv \frac{1}{a-1}$-et új típusparaméterként; ezzel egyben azt is elérjük (a STEINER (ed.) [1997]-ben bizonyítottan), hogy új típusparaméterünk *„kézzelfogható"* értelemmel bír: a Gauss-tól mért típustávolság konstansszorosával egyenlő. A Cauchy-eloszlásra (a=2

miatt) t=1, az a=5-höz tartozó statisztikai típusra t=0,25, az a=9-cel jellemzett Jeffreys-félére t=0,125. Vizsgálódásaink szemszögéből a t=0,5 esetet is kiemelten fontosnak kell ítélnünk, mert (10) szerint ennél az (a=3-mal is jellemezhető) típusnál kezdődik az a típusintervallum, amelytől indulva (azaz $t \geq 0{,}5$ -re) már nincs értelmezve a variancia, amit úgy is szokás rövidebben megfogalmazni, hogy $VAR = \infty$ teljesül (a (4) szerinti integrál divergens).

A fentiekben megismert t alkalmazásával a hibatípus-eloszlások $g(t)$ -vel jelölt sűrűségfüggvénye

$$g(t) = 16 \cdot t \cdot \mathrm{e}^{-4t} \tag{17}$$

szerint írható fel (görbeként e függvényt STEINER [1990] a 233. oldalon a 6.19 ábrán mutatja be). Ez az egyszerű formula maximumát valóban a DUTTER által leggyakrabban előfordulóként megadott hibatípusnál éri el (amely típust most a t=0,25 típusparaméter definiálja), a t=0-hoz tartozó Gauss-típus (17) szerint valóban zérus valószínűségsűrűségű, a Jeffreys-intervallum típusait viszont a maximális $g(t)$ érték felénél nagyobb valószínűségsűrűségek jellemzik, — míg a t=1-hez tartozó Cauchy-eloszlás valószínűségsűrűsége már kb. csak ötöde a maximális $g(t)$ értéknek: így teljesül itt az a korábbi megállapítás, hogy szakterületünkön ugyan nem várható gyakorinak a Cauchy környéki típusok előfordulása, de ez utóbbi esemény egyáltalában nem elhanyagolható valószínűségű. A t-nek 1-nél nagyobb, növekvő értékeihez a $g(t)$ értékek gyors zérushoz tartása tartozik, kifejezve azt, hogy $a < 2$ típusparaméterű eloszlások fellépte már szinte elhanyagolhatóan ritka.

A (17) ugyan elegánsan sűrítve adja vissza a típuseloszlások előfordulási gyakoriságairól az irodalomban található utalásokat, de hiba lenne a földtudományok minden szegmensére szinte természettörvényként elfogadni. A jelen sorok szerzője azonban nem ismer egyéb javaslatokat a szakirodalomból, ezért a (17) alapján válaszolja meg azt a kérdést, hogy az előforduló típusok hány százalékában véges a varianciák (egyben persze az (1) szerinti $\gamma(h)$ variogrampontok) aszimptotikus szórása. Mivel $t \geq 0{,}25$ -nél (12) szerint A_{VAR} már végtelen, a varianciák hibáit szórással, azaz $\sigma_{VAR} = A_{VAR} / \sqrt{n}$ -ként a várható hibatípusok 26,4%-ánál tudjuk csak jellemezni, mivel a $P\{t < 0{,}25\}$ valószínűség (17) szerint

$$P\{t < 0{,}25\} = \int_0^{0{,}25} g(t)\mathrm{d}t = 0{,}264 \tag{18}$$

értékűnek adódik.

5. A σ_{VAR} varianciahibák Monte Carlo-meghatározásai és kapcsolatuk az A_{VAR} aszimptotikus szórással

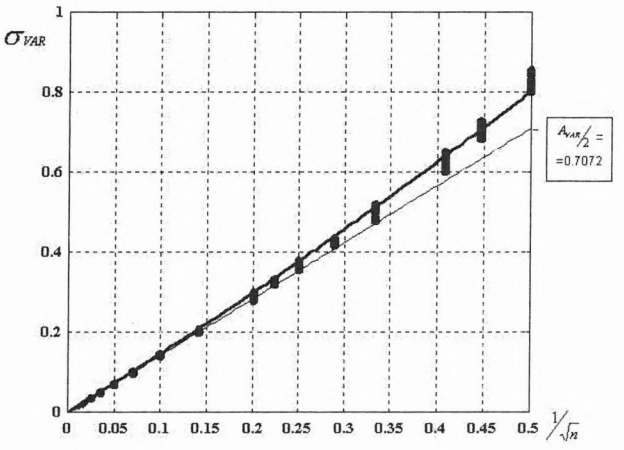
Ha a statisztika irodalmában megadják bármely jellemzőhöz és becslésmódjához az aszimptotikus szórás értékét vagy formuláját, azt általában nem felejtik el hozzátenni, hogy a becslések σ szórása $A/\sqrt{n}$ -ként számítható, ha n elég nagy. Annak a közlésével általában már adós szokás maradni, hogy az adott esetben milyen n értéket tekinthetünk *elég nagy*nak? 25-öt, 50-et vagy 100-at? Esetleg még ennél is sokkal nagyobb értéket, mondjuk n=10 000-et? HAJAGOS, STEINER [2000]-ből látható, hogy amennyiben az aszimptotikus szórással kezelhető típustartománynak a határához közeledünk, a $\sigma = A/\sqrt{n}$ szabály csak minden határon túl növő n-ekre ad pontos értéket.

Jelenleg nyilván a

$$\sigma_{VAR} = A_{VAR} / \sqrt{n} \tag{19}$$

érvényességi tartományai érdekelhetnek bennünket, de mivel a (12) szerinti A_{VAR} a (18)-ból láthatóan az összes előforduló hibatípusoknak csak 26,4%-ára véges, elégedjünk meg három hibatípus: a Gauss-, Jeffreys- és az a=6-hoz tartozó f_a típus Monte Carlo-vizsgálatával. Emlékezünk, hogy a harmadikként említett eloszlás a Jeffreys-intervallum felső, a kritikus (a Gauss-típustól távolodva elsőként $A_{VAR} = \infty$-t eredményező) a=5-höz közel álló határa, így az a=5,9;5,8;…;5,1 sorozat időigényes vizsgálatával mutathatnánk csak be azt, amit per analogiam kézenfekvő a varianciákra vonatkozóan is várni a HAJAGOS, STEINER [2000] eredményei alapján: az n, mint a (19) alsó érvényességi határa, minden határon túl nő, ha $a \to 5$.

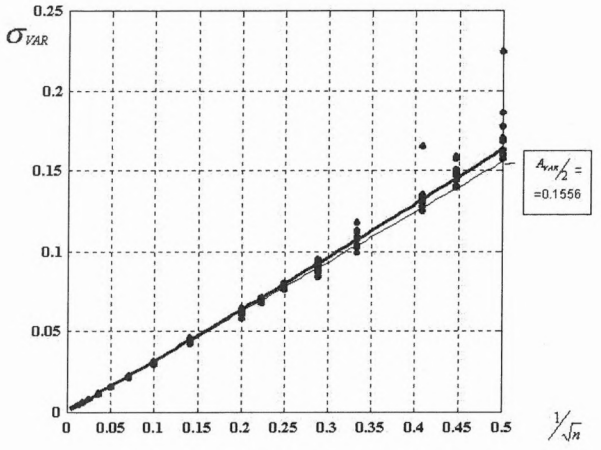
A Monte Carlo-vizsgálatok N ismétlési száma mindvégig 2400 volt; ezek eredményeit az *1–3. ábrán* pontok jelzik. Az ábrákon látható, hogy azonos n-hez több pont is tartozik, mert a gép az egész fenti procedúrát azonos n-hez minden esetben 11-szer ismételte meg.



1. ábra. A varianciák szórása az $1/\sqrt{n}$ függvényében, ha az anyaeloszlás Gauss típusú. A (20) összefüggésbeli c faktort a Monte Carlo-eredményekből az eltérések P_C-normájának minimalizálásával határoztuk meg

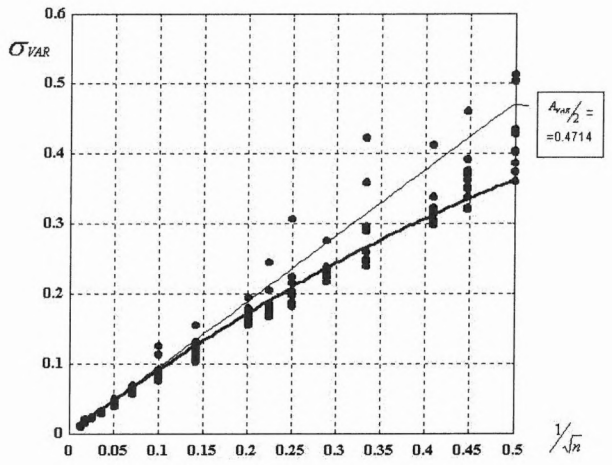
Fig. 1. The scatter of the variances *vs.* $1/\sqrt{n}$ in case of the standard Gaussian parent distribution. The c coefficient in Eq. 20 was determined on the basis of the Monte Carlo results using P_C - norm minimisation; the thin straight line corresponds to the asymptotic rule

A $(1/\sqrt{n}\,;\ \sigma_{VAR})$ pontok elhelyezkedését vizuálisan megítélve az 1., 2. és 3. ábrán (minimálisan n=4 elemű mintát tartva reálisnak σ_{VAR} becsléséhez, az $1/\sqrt{n}$ jelentésű abszcisszát csak a 0,5 értékig vettem fel az ábrákon), egyszerű parabolikus kiegyenlítést látszik célszerűnek végrehajtani



2. ábra. A varianciák szórása az $1/\sqrt{n}$ függvényében, ha az anyaeloszlás Jeffreys típusú. A (20) összefüggésbeli c faktort a Monte Carlo-eredményekből az eltérések P_C -normájának minimalizálásával határoztuk meg

Fig. 2. The scatter of the variances vs. $1/\sqrt{n}$ in case of the Jeffreys (see Eq. 15) parent distribution. The c coefficient in Eq. 20 was determined on the basis of the Monte Carlo results using P_C-norm minimisation; the thin straight line corresponds to the asymptotic rule



3. ábra. A varianciák szórása az $1/\sqrt{n}$ függvényében, az anyaeloszlás a=6-hoz tartozó standard $f_a(x)$-eloszlására. A (20) összefüggésbeli c faktort a Monte Carlo-eredményekből az eltérések P_C-normájának minimalizálásával határoztuk meg

Fig. 3. The scatter of the variances vs. $1/\sqrt{n}$ in case of the standard $f_a(x)$ for a=6 as parent distribution. The c coefficient in Eq. 20 was determined on the basis of the Monte Carlo results using P_C-norm minimisation; the thin straight line corresponds to the asymptotic rule

a hibaeloszlás-típusok neve	***a*** **típusparaméter**	A_{VAR} **a varianciák aszimptotikus szórása**	***c*** **az 1/*n* szorzója a (20) formulában**
Gauss	∞	1,4145	0,3645
Jeffreys	9	0,3118	0,0304
a Jeffreys-intervallum Gausstól távolabbi határa	6	0,9428	-0,4383

1. táblázat. A variancia szórásának n-től való függését $\sigma_{VAR} = A_{VAR}/\sqrt{n} + c/n$ alakban megadó (20) egyenlet paraméterei három, standard alakban megadott anyaeloszlástípusra

Table 1. The values of the parameters figuring in Eq. 20 for three parent distribution types (given in standard form)

$$\sigma_{VAR} = A_{VAR}/\sqrt{n} + c/n \qquad (20)$$

alakban. Tanulságos megemlíteni, hogy A_{VAR}-t és c-t egyaránt ismeretlennek tekintve, a klasszikus, azaz az L_2-normát minimalizáló kiegyenlítés A_{VAR} értékeire a (12)-ből (ill. (13)-ból) pontosan számítható A_{VAR} értékektől szignifikánsan eltérő adatokat is szolgáltatott, ezzel szemben a P_C (ld. pl. STEINER (ed.) [1997] 20. oldalán levő P.2 táblázatot) minimumhelye helyes értékeket eredményezett A_{VAR}-ra mindhárom esetben. (A (13)-at a standard Gauss-esetnél alkalmaztuk: az egységnyi szórás miatt *VAR*=1 is teljesül, s így (13)-ból közvetlenül leolvasható, hogy erre az ($a = \infty$ -nel is jellemezhető) típusra $A_{VAR} = \sqrt{2}$.) Így a P_C szerint egyenlítettem ugyan ki mindhárom ponthalmazt, de felesleges „sportteljesítménynek" ítélve A_{VAR} kiegyenlítéssel való meghatározását, kiegyenlítéskor a pontosan ismert A_{VAR}-okat helyettesítve (20)-ba, kiegyenlítéssel csak a c meghatározása történt. A (20) paramétereit a vizsgált három esetre az *1. táblázat* tartalmazza.

Az 1., 2. és 3. ábrán vékony vonallal az origóbeli érintőt is feltüntettük; ez adná meg a σ_{VAR} függését az összes n mintaméretre vonatkozóan, ha (19) nemcsak aszimptotikusan teljesülne. Ha a kiegyenlítés eredménye szerinti vastag vonal már grafikusan nem különíthető el a vékony vonalú egyenestől, akkor már az egyszerű (19) formulát alkalmazhatjuk. A fentiekben már kifogásolt „ha n elegendően nagy" feltétel helyébe számértékek lépnek: a Jeffreys-eloszlásnál $n \geq 25$ -nek és az a=6-tal jellemzett f_a hibatípusnál $n \geq 100$ -nak kell teljesülnie ahhoz, hogy (19) a gyakorlat számára pontos eredményt adjon.

6. A $Q_{VAR,n}$ varianciahibák Monte Carlo-meghatározásai és a belőlük levonható következtetések véges A_{VAR} esetén

Ebben a pontban tulajdonképpen a 7. pontot szándékozom előkészíteni, amennyiben már most öntsünk tiszta vizet a pohárba: az $A_{VAR} = \infty$ és következésképpen a $\sigma_{VAR} = \infty$ esetét sem kell szerencsére tragikusan felfognunk, mivel a *VAR*-becslések egyéb hibajellemzői nem okvetlenül végtelenek, ha σ_{VAR} végtelennek adódik is. Magyarul: nem túl szerencsés a választásunk, ha a szórást választjuk hibajellemzőnek. Milyen más alternatívánk van?

Többet is felsorolhatnék, de elégedjünk meg a Q-val jelölt interszextilis félterjedelemmel (grafikus definícióját standard $f_4(x)$ -re a STEINER (ed.) [1997] P.2 ábrája mutatja

be a 21. oldalon), amelynek néhány praktikus sajátságát már STEINER [1990] is megfogalmazta: sohasem végtelen értékű, aszimptotikus viselkedését még f_a-nál általánosabb esetekre is egyszerű formula írja le, — és hogy egy klaszszikus statisztikát művelő szakember fülének is bizonyára jól hangzó tulajdonságáról is beszámoljak: $Q \approx \sigma$ teljesül a *normáleloszlásra* (azaz a Gauss-félére); a pontos összefüggés erre az eloszlásra a következő:

$$Q = 0{,}9674 \cdot \sigma \,. \tag{21}$$

A Q interszextilis félterjedelem általános definíciója előtt definiálnunk kell a Q_f felső szextilist az

$$\int_{Q_f}^{\infty} f(x)\mathrm{d}x = 1/6 \tag{22a}$$

integrállal, az alsó szextilist pedig a következőképpen:

$$\int_{-\infty}^{Q_a} f(x)\mathrm{d}x = 1/6 \,. \tag{22b}$$

A (Q_a, Q_f) intervallumra nyilván 2/3 valószínűséggel esnek x értékek (Q_a-nál kisebb és Q_f -nél nagyobb x értékek előfordulása egyaránt 1/6 valószínűségű); kézenfekvő tehát a (Q_a, Q_f) intervallum felét hibajellemzőnek elfogadni és ezt Q-val jelölve interszextilis félterjedelemnek nevezni:

$$Q = \frac{1}{2}(Q_f - Q_a) \,. \tag{22c}$$

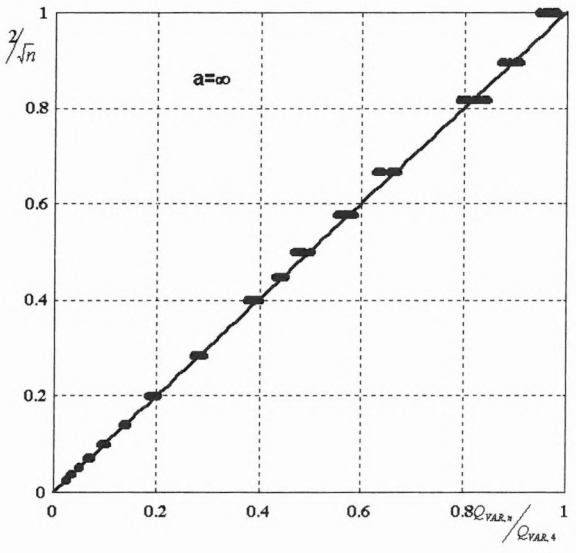
Monte Carlo-számításunk során N=2400-szor generálunk n elemű mintát az éppen vizsgált anyaeloszlásból, minden mintára meghatározva a (6) kifejezést és nagyság szerint rendezve az így kapott *VAR*-becsléseket, már csak azt a két értéket kell megadnunk, amelynél $N/6$ db *VAR* kisebb (vagy egyenlő), ez a $Q_{VAR,n,a}$, valamint amelynél $N/6$ db *VAR*-becslés nagyobb (vagy egyenlő), ez $Q_{VAR,n,f}$ lévén, (22c) szerint már számíthatjuk a kézenfekvően $Q_{VAR,n}$-nel jelölt hibajellemzőt. Maga *VAR* is szoros kapcsolatban van az anyaeloszlás egyik primer hibajellemzőjével, a szórással ($VAR = \sigma^2$ miatt), ezért ha a $Q_{VAR,n}$ hiba hibáját (akár elméleti úton, vagy másképpen) tüzetes vizsgálat tárgyává tennénk, az már a *hiba hibájának a hibája* vizsgálati célkitűzések kategóriába esne, amely nem érdektelen kérdéskör ugyan, de túlzásnak ítélném ennek részletes vizsgálatát. Ehelyett minden n-re mindig 11-szer meghatározva (persze egymástól teljesen függetlenül) a $Q_{VAR,n}$ értékeket, ezek (normálás után) ugyanúgy külön pontként kerülnek az ábrákra, ahogyan (az N elemű *VAR* minta sorba rendezését és a normálást leszámítva) hasonló előkészületi munka után számítottuk az N db *VAR* becslésből a σ_{VAR} értékeket és így minden n-hez annyi pont került (a σ_{VAR}-ok kis ingadozásának kedvező esetében esetleg vizuálisan már nem elkülöníthetően) az 1., 2. és 3. ábrára, ahányszor az egész procedúrát a gép végrehajtotta, tehát mindig 11 darab.

A fent már szóba hozott normálás végrehajtása abból a célból történik, hogy az eredmények függetlenek legyenek a skálaparamétertől (mivel mindig standard alakú anyaeloszlásból indulunk ki, ahol az S=1 teljesül, minden $Q_{VAR,n}$ és a korábbiakban kapott σ_{VAR} eredményünk is S^2-tel szorzandó, ha az S skálaparaméter értékét az $S \neq 1$ reláció jellemzi). Számítástechnikailag nyilván a legkevésbé munkaigényes $Q_{VAR,4}$ meghatározása, így ezt a meghatározást (11 helyett) ezerszer is megismételhetjük, amely értékhalmaz mediánja már kellően pontos ahhoz, hogy normálásra alkalmazhassuk; a további ábráinkon szereplő pontok egyik koordinátája ezért $Q_{VAR,n}/Q_{VAR,4}$ lesz. Mivel nyilván 1 ennek az immár skálaparaméter-független hányadosnak a maximális értéke, a pontok másik koordinátájára is könnyen megvalósíthatjuk az 1 értékű maximumot, ha az eddigi $1/\sqrt{n}$ helyett $2/\sqrt{n}$ -et alkalmazunk. Hogy görbéink sohasem indulhassanak végtelen iránytangenssel az origóból, ezért (végül is tehát az analitikus kezelhetőség megkönnyítésére) $Q_{VAR,n}/Q_{VAR,4}$ szerepel a továbbiakban abszcisszaként és $2/\sqrt{n}$ ordinátaként; az origóban a végtelen aszimptotikus szórások eseteiben így adódó zérus iránytangens megjelenése analitikus szemszögből megítélve nemcsak hogy nem probléma, sőt inkább egyszerűsödést eredményez (vö. a (24) és (27) formulákat).

A *4., 5.* és *6. ábra* a fenti meggondolással adódó eredmények kiegyenlítő görbéit mutatja be ugyanazokra a véges aszimptotikus szórású esetekre (Gauss, Jeffreys, a=6), amelyek σ_{VAR}-görbéit az 1., 2. és 3. ábrán már megismertük. A 4. ábrán egyetlen egyenes adódik; az 5. és 6. ábra görbéi pedig nem elhanyagolható hosszúságú (nyilván az origóbeli érintővel egybeeső) egyenes szakasszal indulnak, arra utalva, hogy a $Q_{VAR,n}$ mennyiségekre a (19)-cel analóg

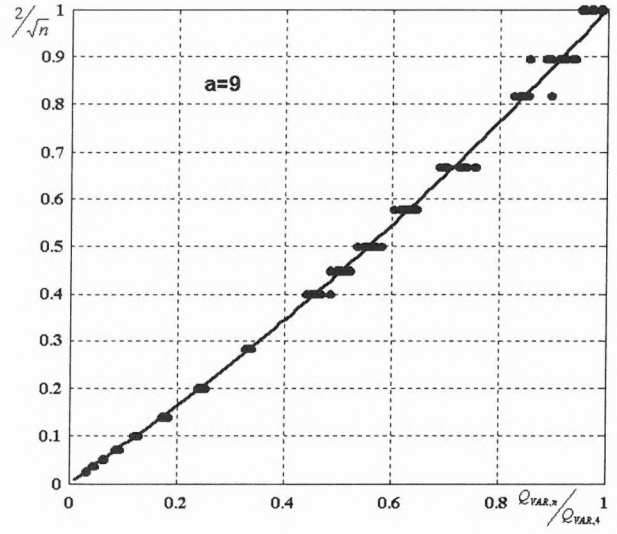
$$Q_{VAR,n} = Q_{VAR} / \sqrt{n} \tag{23}$$

aszimptotikus törvényszerűség teljesül véges aszimptotikus szórású esetekben, csak egyelőre még nem ismerjük a Q_{VAR} *aszimptotikus interszextilis félterjedelem* számszerű értékeit.



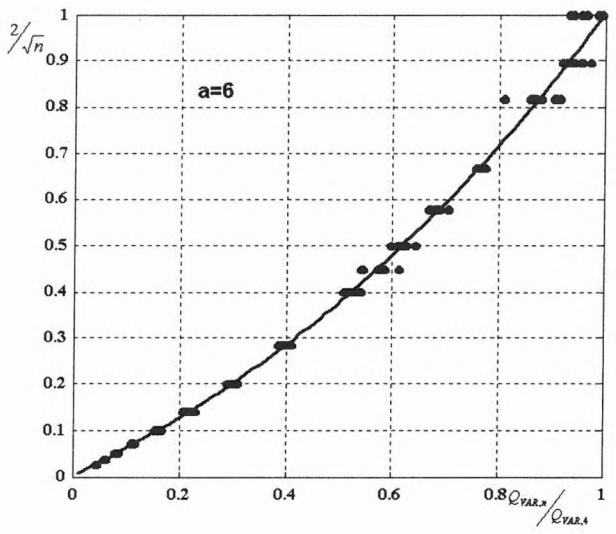
4. ábra. $2/\sqrt{n}$ görbéje a $Q_{VAR,n}/Q_{VAR,4}$ függvényében Gauss típusú hibaeloszlásra, a Monte Carlo-eredményeket a (24) egyenlet szerint L_2-minimalizálással meghatározva (a görbeparaméterek számértékeire vonatkozóan ld. a 2. táblázatot)

Fig. 4. The $2/\sqrt{n}$ curve vs $Q_{VAR,n}/Q_{VAR,4}$ in case of the Gaussian parent distribution, fitting the Monte Carlo results according to Eq. 24. (The numerical values of the parameters are given in Table 2)



5. ábra. $2/\sqrt{n}$ görbéje a $Q_{VAR,n}/Q_{VAR,4}$ függvényében Jeffreys típusú hibaeloszlásra, a Monte Carlo-eredményeket a (24) egyenlet szerint L_2-minimalizálással meghatározva (a görbeparaméterek számértékeire vonatkozóan ld. a 2. táblázatot)

Fig. 5. The $2/\sqrt{n}$ curve vs $Q_{VAR,n}/Q_{VAR,4}$ in case of the Jeffreys-type parent distribution, fitting the Monte Carlo results according to Eq. 24. (The numerical values of the parameters are given in Table 2)



6. ábra. $2/\sqrt{n}$ görbéje a $Q_{VAR,n}/Q_{VAR,4}$ függvényében a hibák a=6-hoz tartozó $f_a(x)$-eloszlására, a Monte Carlo-eredményeket a (24) egyenlet szerint L_2-minimalizálással meghatározva (a görbeparaméterek számértékeire vonatkozóan ld. a 2. táblázatot)

Fig. 6. The $2/\sqrt{n}$ curve vs $Q_{VAR,n}/Q_{VAR,4}$ in case of the a=6 parent distribution, fitting the Monte Carlo results according to Eq. 24. (The numerical values of the parameters are given in Table 2.)

a a típus-paraméter	$Q_{VAR,4}$ a négyelemű minták varianciabecsléseinek interszextilis félterjedelme	b	u	v	w a $2/\sqrt{n}$ és a $Q_{VAR,n}/Q_{VAR,4}$ kapcsolatát leíró (27) egyenlet négy paramétere, Monte Carlo-eredmények alapján	Q_{VAR} a varianciák aszimptotikus interszextilis félterjedelme; alkalmazását a (23) egyenlet mutatja	0,9674·A_{VAR} a varianciák aszimptotikus szórása a Gauss-eloszlásra érvényes Q/σ hányados értékével szorozva	a Q_{VAR} és a 0,9674·A_{VAR} kicsiny százalékos eltérései (ld. a (26) definíciót)
∞	0,7000	1,0000	-	-	-	1,4000	1,3681	2,28%
9	0,1185	0,4749	1,629	1,061	0,1969	0,3160	0,3012	4,68%
6	0,2321	0,4776	1,414	1,885	-0,6055	0,9719	0,9121	6,15%
5	0,3355	0	1,246	2,132	-0,6312	∞	∞	-
4	0,5815	0	1,672	3,029	-1,2596	∞	∞	-
3,5	0,8733	0	1,995	3,545	-2,8765	∞	∞	-

2. táblázat. A 4–7. ábrák görbéit leíró (24), ill. (27) egyenlet szerinti *b, u, v, w* paraméterek (3., 4., 5. és 6. oszlop); a második oszlopbeli $Q_{VAR,4}$-ből és *b*-ből (25) szerint meghatározott Q_{VAR} aszimptotikus interszextilis félterjedelem értékei a 7. oszlopban találhatók. Az utolsó oszlop a nagy *n*-ekhez számított empirikus varianciák típusának Gauss-féléhez közeli voltára utal (esetleg aszimptotikusan az is)

Table 2. The curves in Figs. 4–7 were fitted according to Eq. 24 and Eq. 27; the parameters (*b, u, v, w*) are given in the third, fourth, fifth and sixth columns for six parent distribution types. The Q_{VAR} values were calculated on the basis of the $Q_{VAR,4}$ and *b* values according to Eq. (25). The percentual difference-values given in the last column hint that the distribution-type of the variances calculated for large *n*-s is near to the Gaussian

A 4., 5. és 6. ábrák pontjainak kiegyenlítése azzal az analitikus formulával történik, amely az egységnégyzetbeli, hasonló lefutású változások eseteire HAJAGOS, STEINER [2000]-ben már olyan jól bevált. Jelen esetünkben ez

$$2/\sqrt{n} = b \cdot x + (1-b) \cdot (x^u - w \cdot x^v \cdot \ln x) \qquad (24)$$

alakú összefüggést jelent, ahol az x egyszerűsítő jelölés a $Q_{VAR,n}/Q_{VAR,4}$ hányadost jelenti. A 4., 5. és 6. ábra eseteire a *2. táblázat* első három sora adja meg egyrészt $Q_{VAR,4}$-nek, másrészt a (24) egyenletben szereplő paraméterek (azaz a *b, u, v* és *w*) értékeit.

Könnyen belátható, hogy a $Q_{VAR,4}$ és a *b* origóbeli iránytangens birtokában a (23)-beli aszimptotikus interszextilis félterjedelem

$$Q_{VAR} = 2 \cdot Q_{VAR,4} / b \qquad (25)$$

szerint számítható. Ezeket az értékeket is feltüntettük a 2. táblázat első három sorában, s mivel ezek közel álltak az

A_{VAR} aszimptotikus szórások 1. táblázatbeli értékeihez, a nagy n-ekhez tartozó varianciák Gauss-hoz közeli típusát feltételeztük. Valóban, mivel a (21) szerint az ilyen típusra $Q_{VAR}=0{,}9674\cdot A_{VAR}$ teljesül, a táblázat a $0{,}9674\cdot A_{VAR}$ szorzatokat is tartalmazza, és ezek Δ_{VAR}-ral jelölt százalékos eltéréseit a Monte Carlo-eredményekből nyert Q_{VAR} értékektől

$$\Delta_{VAR}=\left(Q_{VAR}-0{,}9674\cdot A_{VAR}\right)/Q_{VAR}\cdot 100 \qquad (26)$$

szerint számítva, kicsiny (és legalább részben a teljes Monte Carlo-procedúra + kiegyenlítés) számlájára írható százalékos eltérések adódtak mindössze. Ezek szerint $Q_{VAR,n}$-et az $5<a<\infty$ tartomány bármely a paraméterrel jellemzett $f_a(x)$ anyaeloszlás-típusára (23) szerint számíthatjuk, úgy, hogy Q_{VAR} helyébe a (12) szerint számított A_{VAR} 0,9674-szeresét helyettesítjük, feltételezve persze, hogy n elég nagy.

7. A $Q_{VAR,n}$ varianciahibák értékei a varianciák végtelen aszimptotikus szórása esetén

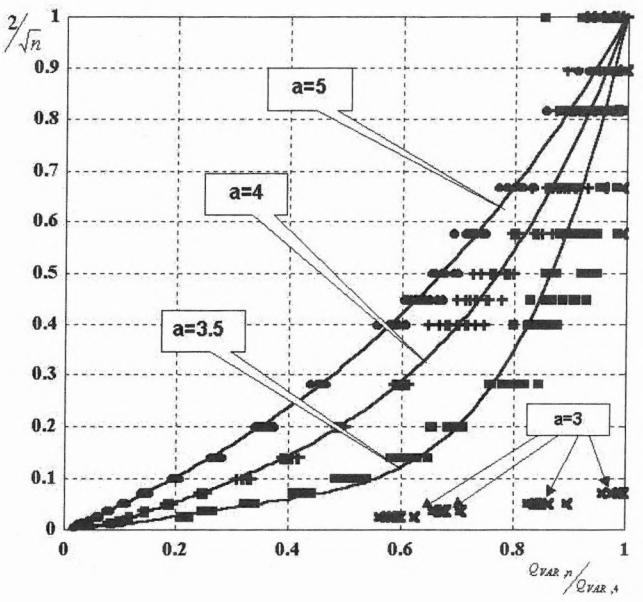
A *7. ábra* a 6. pontban részletesen leírt Monte Carlo-módszerrel nyert ($Q_{VAR,n}/Q_{VAR,4}$; $2/\sqrt{n}$) pontokat mutatja be az egységnégyzetben, amelyek az a típusparaméter 5; 4 és 3,5 értékeihez tartozó $f_a(x)$ anyaeloszlásokhoz tartoznak; (12)-ből láthatóan mindhárom esetre $A_{VAR}=\infty$.

Minél közelebb van az a 3-hoz, az origótól zérus iránytangenssel indulóan annál hosszabb $Q_{VAR,n}/Q_{VAR,4}$ intervallumon esnek a pontok az abszcisszatengelyhez nagyon közel (a 7. ábrán az „a=3,5" jelű görbe mutatja legpregnánsabban ezt a tendenciát,) ami azt jelenti, hogy a nagy számok törvénye teljesül ugyan (azaz nagyobb n mintaelemszámnál csökken a varianciák $Q_{VAR,n}$-nel mért hibája), de ez a=3,5-nél a $Q_{VAR,4}$ hibának már akár csak felére csökkentéséhez is igen nagy n mintaelemszámot igényel. Az alkalmazott Monte Carlo-eljárás kissé kétségtelenül túlfeszített teszteléseként az a=3-hoz is számítottunk pontokat (ezek közül az abszcisszához közelieket nyilakkal jelöltük, a többi gyakorlatilag az abszcissza 1-es értékénél emelt függőleges egyenesre esett). Ezek feltüntetése talán nem volt felesleges, mert a görbék origóból való *vízszintes* indulása és kisebb vagy nagyobb hosszon az abszcisszatengely közelében maradása mindegyik $5\geq a>3$ relációnak eleget tevő $f_a(x)$ anyaeloszlástípusra jellemző, ha pontjainkból most (a $Q_{VAR,n}/Q_{VAR,4}$ hányadost továbbra is x-ként jelölve,) a

$$\frac{2}{\sqrt{n}}=x^u-w\cdot x^v\cdot\ln x \qquad (27)$$

analitikus kifejezés szerint végezzük el a pontok kiegyenlítését. Ekkor ui. a (24)-ben szereplő b értéke természetesen zérus; a többi görbeparamétert (tehát a (27) összes görbeparaméterét) a 7. ábra három görbéjére a 2. táblázat számszerűen adja meg.

A görbék közelítőleg *vízszintes* szakaszainak hossza a=5-höz közeledve egyre rövidül, sőt, nullához tart (a 7. ábrán az 5-tel jelölt görbén ezért nem látszik vízszintes szakasz, de grafikusan az sem, hogy a görbe — mint minden (27) alakú függvény — az origóban zérus iránytangensű). Másrészt viszont a-t minden határon túl 3 közelébe csökkentve, a számított görbék egyre közelebb vannak ahhoz a *görbéhez*, amely az egységnyi hosszúságú abszcisszavonalból és az abszcissza 1-es értékénél arra merőlegesen emelt, szintén egységnyi hosszúságú egyenesből áll. Az az előbbiekben jogosan túlfeszítettnek minősített tesztelés, amely a=3-ra végzett számításokból állt, valóban a fentiekben leírt, két egymásra merőleges egyenes szakaszból álló *határgörbéhez* közeleső pontokat eredményezett. (A túlfeszítettség ismételt hangsúlyozását azért tartom indokoltnak, mert (10) szerint már elvi értéke sem létezik a varianciának a=3 esetén.) HAJAGOS, STEINER [2000]-ben az $\bar{x}$ számtani átlagokra kapott teljesen analóg eredményeket a végtelen aszimptotikus szórások típustartományára, amely ott a $3\geq a>2$ relációt teljesítő a-kat jelentette; az a=2-nél már nem létezik várható érték, így ez a határ az $\bar{x}$-okra ugyanolyan szerepű volt, mint most a varianciáknál az a=3 határ.



7. ábra. $2/\sqrt{n}$ görbéi a $Q_{VAR,n}/Q_{VAR,4}$ függvényében, végtelen aszimptotikus szórású varianciák néhány esetére: az a=5; 4 és a=3,5 típusparaméter-értékekhez tartozó $f_a(x)$ hibaeloszlásokra. A (27) egyenlet szerinti L_2-minimalizálás a Monte Carlo-eredmények alapján, az u, v és w paramétereknek a meghatározására történt b=0 feltétellel; a numerikus eredményeket a 2. táblázat tartalmazza. Az a=3 típusparaméter esetén a variancia már definiálatlan, így a Monte Carlo-eredmények $\left[Q_{VAR,n}/Q_{VAR,4};\ 2/\sqrt{n}\right]$ pontjai érthetően részben az abszcissza, részben a $Q_{VAR,n}/Q_{VAR,4}$=1 függőleges egyenes közelében adódtak (ezek közül az ábra csak az abszcissza-közelieket jelöli meg)

Fig. 7. The $2/\sqrt{n}$ curves vs $Q_{VAR,n}/Q_{VAR,4}$ for three cases characterized by infinite A_{VAR} asymptotic scatter: the parent distributions are of $f_a(x)$ type to the type parameters a=5; 4 and a=3.5. The u, v and w parameters of the three curves were determined by fitting of the Monte Carlo results according to Eq. 27 (the numerical values are given in Table 2). In case of a=3 the variance itself is already undefined, consequently the points $\left[Q_{VAR,n}/Q_{VAR,4};\ 2/\sqrt{n}\right]$ representing the Monte Carlo results and denoted in the figure by x are very near to the abscissa (see the arrows), on one hand, and on the other hand, very near (or even coinciding with) the vertical straight line defined by $Q_{VAR,n}/Q_{VAR,4}$=1.

A varianciák végtelen aszimptotikus szórásával jellemzett $5\geq a>3$ típusparaméterű $f_a(x)$ eloszlások másik határát: az a=5-höz tartozót azonban legnagyobb örömünkre a

nagy számok törvényét gazdaságosan teljesítő viselkedés jellemzi (ld. a 7. ábrán az 5 jelű görbét), hiszen emlékezünk a korábbiakból arra, hogy ez tekinthető a geostatisztikában (DUTTER 1986/87), de általánosabban is a leggyakrabban előforduló hibaeloszlás-típusnak.

Nem szeretném a dolgozat végén az olvasónak afeletti örömét elrontani, hogy a nagy számok törvénye (már ti. hogy növekvő n-nel csökken a hiba), mint láttuk, minden véges varianciájú esetben teljesül, azaz feloldódott az a pusztán látszólagosnak bizonyult ellentmondás, amelyet a $\gamma(h)$ variogrampont- vagy egyéb variancia-meghatározás túlnyomóan problémamentesnek talált volta és a között a tény között feszült, hogy a variancia aszimptotikus szórása a geostatisztikában előforduló eloszlástípusoknak csak kb. egynegyedére véges, és az irodalomban leggyakrabban előfordulónak mondott típusra is már végtelen. A megoldást (Kolumbusz tojásaként) az hozta meg, hogy a hibát a korábbiakban általánosan elfogadott σ szórás helyett interszextilis félterjedelmével (Q-val) jellemeztük; ezzel sikerült reális képet kapnunk a hibaviszonyokról. Örömünket esetleg kissé beárnyékolhatja, de persze nem hallgatható el, hogy a hibacsökkenés mértéke nagymértékben függ az anyaeloszlás típusától, s így pl. ha n=4-ről n=100 mintaelemszámra térünk át, a 4.–7. ábrákról azonnal leolvasható, hogy az n=4-re jellemző hiba n=100 alkalmazásakor (azaz 0,2 értékű $2/\sqrt{n}$ -re) a Gauss-típusnál 20%-ára, a Jeffreys-típusnál 25%-ára, a=6-nál 30%-ára, a=5-nél 35%-ára, a=4-nél 50%-ára végül a=3,5-nél már csak 70%-ára csökken. A felsorolás utolsó két adata arra utal, hogy a nagy számok törvénye — a gyakorlati szakember bánatára — eléggé gazdaságtalanul is teljesülhet, ha az a típusparaméter értéke túl közel van 3-hoz, az elméleti variancia létezésének határához. Ilyen esetekben ijesztően (és túlnyomóan teljesíthetetlenül) nagy n-ek is adódhatnak, ha egy szignifikáns hibacsökkenési mértéket eleve kikötünk, amelyet a négyelemű minimális mintaméretről egy alkalmas n-re váltva akarunk elérni. Amennyiben a minimális mintaméretet jellemző hibát egyötödére (20%-ára) akarjuk lecsökkenteni, akkor a fentiekben láttuk, hogy ez a Gauss-anyaeloszlásnál már n=100-zal elérhető, de az 5., 6. és 7. ábrákról leolvashatóan a Jeffreys-típusnál (a=9-nél) már n=156, az a=6 esetében n=237, a leggyakrabban előforduló a=5-re pedig n=400 mintaelemszámra van szükségünk (megítélés kérdése, hogy ezt már túl soknak ítéljük-e, hiszen csak négyszerese a Gauss-típust jellemző értéknek). A 7. ábráról leolvashatóan ugyanilyen hibacsökkenéshez azonban az a=4 típusnál 1600, míg az a=3,5 típusnál már kereken ötezer mintaelemszám szükséges.

* * *

A szerző megragadja az alkalmat, hogy köszönetét fejezhesse ki az FKFP 0914/1997 projektnek az 1997 óta többnyire team-tevékenység keretében végzett munkájának támogatásáért. Ebbe nemcsak oktatásfejlesztést és kutatómunkát, de az ezen időszakban közlésre leadott cikkek írását is beleértem, amelyeket túlnyomóan HAJAGOS Béla ny. főiskolai tanár társszerzővel, legnagyobb számban az Acta Geodaetica et Geophysica Acad. Sci. Hung. folyóirat jelentetett meg. Ugyanitt fog megjelenni a szerzőnek (HAJAGOS Béla társszerzőségével) az a dolgozata is, amely a jelen cikk angol nyelvű pandantjaként tekinthető, hiszen az ábrák és táblázatok azonosak lesznek, ezek hordozván a számítások eredményeit, amelyeknek programozási és futtatási munkáiért HAJAGOS Béla barátomnak hálás köszönetemet fejezem ki, mint a jelen dolgozat szerzője. Ez az angol nyelvű dolgozat a jelenleginél rövidebb lesz (lényegében megmaradva az ábrák kommentálásánál), ezért ezt a cikket már csak kétszerzős változatban lesz korrekt megjelentetni (a rövidséget az ugyanabban az Actában megjelentetett cikkekre való nagyobb mérvű támaszkodás lehetősége teszi indokolttá). — Az azonos eredmények két dolgozatban történő, de más szövegű (terjedelemben is eltérő) megjelentetését persze előzetesen egyeztettem a két folyóirat főszerkesztőjével.

HIVATKOZÁSOK

CRAMÉR H. 1945: Mathematical methods of statistics. Almqvist & Wiksells, Uppsala, 575 p.

DUTTER R. 1986/87: Mathematische Methoden in der Montangeologie. Vorlesungsnotizen. Manuscript, Leoben

EGYED L. 1955: Új módszer az átlagsűrűség meghatározására. Geofizikai Közlemények **4**, 2, 31–36

HAJAGOS B., STEINER F. 2000: The fulfilment of the law of large numbers for arithmetic means in case of infinite asymptotic scatter. Acta Geod., Geoph. Acad. Sci. Hung. **35**, 4

HUBER P. J. 1964: Robust estimation of a location parameter. Ann. Math. Statist., **35**, 73–101

HUBER P. J. 1981: Robust statistics. Wiley, New York, 308 p.

JEFFREYS H. 1961: Theory of probability. Clarendon Press, Oxford

JOURNAL A. G., HUIJBREGTS Ch. J. 1978: Mining geostatistics. Academic Press, New York, 600 p.

KERÉKFI P. 1978: A robusztus becslésekről. Alkalmazott Matematikai Lapok, **4**, 327–357

MATHERON G. 1965: Les variables regionalisées et leur estimation. Masson & Cie, Paris, 305 p.

PRÉKOPA A. 1962: Valószínűségelmélet műszaki alkalmazásokkal. Műszaki Könyvkiadó, Budapest, 440 p.

RIESZ F., SZŐKEFALVI-NAGY B. 1952: Leçons d'Analyse Fonctionnelle. Akadémiai Kiadó, 448 p.

STEINER F. 1959: Zur Ermittlung des Koeffizienten der gravimetrischen Höhenreduktion. Gerlands Beiträge zur Geophysik **68**, 1, 15–20

STEINER F. 1990: A geostatisztika alapjai. Tankönyvkiadó, Budapest, 363 p.

STEINER F. (ed.) 1991: The most frequent value. Akadémiai Kiadó, Budapest, 315 p.

STEINER F. (ed.) 1997: Optimum methods in statistics. Akadémiai Kiadó, Budapest, 370 p.

SZILASI L. 2000: A Kopereczky-effektus. Jelenkor Kiadó, Budapest, 248 p.