Adalékok a kívülállóérték fogalmához

Dr. Monhor Davaadorzsín egyetemi docens,
Nyugat-Magyarországi Egyetem
Geoinformatikai Főiskolai Kar

1. Motiváció és bevezetés

Napjainkban széleskörű intenzív kutatás folyik a *kívülállóérték* elmélete és gyakorlati alkalmazása terén egyaránt. Ezzel szemben a szakirodalomban még nincs egyértelmű, elfogadott, egységes definíció a kívülállóértékre. Ez azt jelenti, hogy még nincs egyszerű és egyértelmű válasz arra kérdésre: „Mi a kívülállóérték?"

E megállapítást támasztja alá az alábbi két idézet is. 1995-ben *U. Gather* [8] a következőket írja: „The problem of *outliers* in random data sets is a very interesting, important and common one. Nevertheless there is no formal and generally accepted definition of what is meant by an *outlier*. Terms like outlier, spurious observation, contaminant, gross error and others are used with different and overlapping meanings." („A véletlen adatokban rejlő kívülállóérték problémája nagyon érdekes, fontos és gyakran előforduló kérdés. Nincs azonban formális és általánosan elfogadott definíció arra, hogy egy kívülállóérték alatt mit értünk. Kívülállóértéket, helytelen mérési értéket, szennyező értéket, durva hibát és egyebeket eltérő és átfedő értelmezésekkel használják.")

A fenti idézet egy matematikai statisztika elméletével foglalkozó cikkből származik, s így a kívülállóérték definíciójának hiányát a matematikai statisztika oldaláról hangsúlyozza. A geodézia oldaláról is megerősíti ezt, ahogy *K. R. Koch* [12], neves geodéta megjegyzi: „In statistical literature, the word 'outlier' has never been defined precisely..." („Statisztikai irodalomban az „outlier" szót soha nem definiálták precízen...") E sorok után világos, hogy nagy jelentősége van a kívülállóérték mibenléte tisztázásának és annak definíciójával való foglalkozásnak.

A jelen dolgozat célja az, hogy először történelmileg röviden áttekintsük a kívülállóértékkel kapcsolatos kutatások fejlődésének fő irányvonalát, s utána, annak elemzésére támaszkodva, megpróbálkozzunk a szóban forgó fogalom kategorizáló és leíró definíciójával, végezetül rátérünk a kívülállóérték bizonyos kategóriája valószínűségelméleti, ill. matematikai statisztikai elméleti hátterére is.

2. Rövid történelmi áttekintés

Valószínűleg mérési adatok kidolgozásánál először csillagászok használták a kívülállóérték kezelését, amely mai terminológiával nem más, mint a kívülállóérték elvetése (rejecting outliers).

1757-ben, azaz 245 évvel ezelőtt, *Boscovich* [3] 10 mérési eredmény közül két értéket nagyon eltérőnek talált, s a megmaradt 8 érték átlagát használta a Föld elliptikus alakjának meghatározására. Tehát itt két figyelemreméltó ténnyel találkozunk: az első a kívülállóérték elvetéséről, a második geodéziai alkalmazásáról szól. Így elmondható, hogy a kívülállóérték elméletének csírája a geodéziával kapcsolatos (ui. a csillagászati helymeghatározás tág értelemben a geodéziához kapcsolódik). Később *Legendre* [17, 18] is a kívülállóérték elvetés mellett foglalt állást. *Helmert* 1877. évi egyik munkája [13] címéből valószínűsíthető, hogy az kapcsolódhat a kívülállóértékhez, továbbá, a folyóirat neve révén geodéziával is foglalkozhatott. Így geodéziai mérések matematikai feldolgozásának talaján gyökerezett a kívülállóérték.

A kívülállóértékkel foglalkozó régi munkák közül *Daniel Bernoulli* [2] munkájára érdemes figyelni. *Bernoulli* megjegyzi, hogy a csillagászok körében elterjedt módszernek tekinthető a kívülállóérték elvetése, továbbá a következőket írja: „Nem látok módot arra, hogy válaszvonalat húzzak azok között, amelyeket feltétlenül el kell távolítani, és azok között, amelyeket fenn kell tartani mindenképpen; sőt az is előfordulhat, hogy éppen az eltávolított mérési érték az, ami a többiekre a legjobb korrekcióval szolgálhatott volna. Mindez mellett, minden alkalommal nem utasítom el (ítélem el) egyik vagy másik mérési eredmény elvetési elvét, valójában helyeslem az elvet, mivel megfigyelés (mérés) közben adódhat olyan ok, amely önmagában a mérő, ill. megfigyelő számára kételyt ébreszt."

E sorokból belátható, hogy *Bernoulli* a kívülállóérték elvetésének helyeslése mellett alternatív lehetőségét, azaz a kívülállóérték fenntartását is javasolja. E javaslatban a robusztus módszer csíráját látja több szerző, így a robusztus módszer eredete Bernoulli korához vezet vissza. Azonban – e sorok írója szerint – itt egy nagyon lényeges dologról esett szó: *Bernoulli* nem helyettesíti az egyiket (pl. elvetést) a másikkal (fenntartással), hanem két önálló módszerként vagy inkább hozzáállásként említi, továbbá hangsúlyozza a kettő közötti választásnak nehézségét is. Ilyen választás ma sem könnyű, s meggondolandó a kívülállóérték elvetési elmélet és robusztus módszerek közötti relációban.

Boscovich és *Bernoulli* munkáját követő kb. 170–180 éven keresztül, a mérési adatok feldolgozásánál hol elvetettek, hol fenntartottak olyan adatot, illetve adatokat, amely, ill. amelyek kapcsán kétség merült fel. Az elvetés, illetve fenntartás megítélése általában szubjektív jellegű volt. Az elvetést jórészt alkalmazták, s ez az irányzat az uralkodó vonalat képezte. Ad hoc jellegű elvetési szabályokat, ill. „kritériumokat" találtak, és alkalmaztak. Azonban csak az 1930-as években látott napvilágot néhány olyan kritérium, amely a matematikai statisztika szemszögéből nézve viszonylag megalapozottnak tekinthető. Ezek a dolgozatok lendületet adtak a kívülállóértékkel kapcsolatos további kutatáshoz, s így a 60-as évekig a kívülállóérték statisztikai kezelésében történt számottevő előrehaladás. Mindemellett, nem született meg egy egységes kívülállóérték-elmélet, s egy rövid megtorpanási vagy stagnálási periódus következett. Ez az időszak a robusztus módszerek elméletének megjelenésével esik egybe. Az 1960-as években *Tukey* [28] és munkatársai azt mutatták be, hogy statisztikai becslésekben a mintaközép nagyméretű ingadozást előidézhet. Ez irányú kutatások talaján, 1964-ben *Huber* [14] a maximum-likelihood becslést általánosította, s úgynevezett M-becsléseket (Maximum-likelihood típusú becslések) vezetett be, és megállapította azok kedvező becslési tulajdonságait. Ezzel *Huber* a robusztus becslések elméletének alapját rakta le, melyet a robusztus módszerek intenzív fejlődésének és széleskörű elterjedésének korszaka követett. Ám félresiklások, ill. tévhitek is akadtak. Sokan például azt hitték, hogy a kívülállóérték mindennemű problémáját megoldják e módszerek. Így, a kívülállóérték elméleti kutatása megtorpant. Az utóbbi időben viszont a kívülállóérték újra kutatási területté válik.

3. Gondolatok a kívülállóérték fogalmáról

Mindannyian tudjuk, hogy mit jelent a statisztikai minta kifejezés. Mindamellett először pontosítsuk e fontos és alapvető fogalmat, a továbbiak egyértelmű megfogalmazása végett.

A matematikai statisztikában az egymástól független, azonos eloszlású n számú valószínűségi változók $\xi_1, \xi_2, ..., \xi_n$ összességét n elemű statisztikai mintának (röviden statisztikai mintának vagy még rövidebben mintának) szokták nevezni. Az egyes valószínűségi változók a minta elemei; a minta elemeinek száma a minta nagysága. E definíció a matematikai statisztikai fogalmak és módszerek elméleti tárgyalásában, illetve egyértelmű és általános megfogalmazásokban elkerülhetetlen.

Alapdefiníció: *A kívülállóérték a statisztikai minta egy olyan eleme, amely a minta megmaradt részétől meglehetősen távol esik, illetve a minta túlnyomó elemeinek eloszlási viselkedésétől eltérő jelleget mutat.*

A fenti alapdefiníció igen általános leírást ad a kívülállóérték alapvető jellegére, azonban a konkrét jellemzés hiányzik, s ezt a hiányt az alábbi kategorizálást eredményező, pontosabb definíciókkal pótoljuk.

Finomító definíciók:

(a) Nem kívánatos kívülállóérték: Ez a mérés, megfigyelés és számítógépes adat be- és átvitel, illetve továbbítás és egyéb hasonló műveletek során fellépő, a megmaradt részek statisztikai elemzését torzító elem.

(b) Váratlan vagy modell-módosító kívülállóérték: Ez olyan váratlan kívülállóérték, amelynek megjelenése egy új hozzáállást eredményez a tanulmányozás alatt lévő jelenségre, következésképp az alkalmazandó matematikai modell változtatása szükségessé válik.

(c) Természetes ingadozással kapcsolatos kívülállóérték: Ez olyan kívülállóérték, amely az adatokban rejlő változékonyságot markánsan demonstrálja.

Az alábbiakban részletezzük a fenti definíciókat.

A kívülállóérték *(a) kategóriája* a mérési adatok kidolgozásában a *durva hibá*t jelenti. Effajta kívülállóértékre a külföldi irodalomban többnyire a gross error, blunder elnevezéseket használják. A kívülállóérték fogalmának eredete ilyen hibás mérési eredményekkel kapcsolatos, s így az (a) típusú kívülállóértékek a „legősibb" kívülállóértékek. Magától értetődő, hogy ilyen fajta kívülállóértékeket ki kell deríteni, és el kell távolítani a mérési adatokból, amennyiben erre mód és lehetőség van.

Éppen ezért – ahogyan az előbb vázolt történelmi áttekintés is mutatja – hosszú ideig ez a gondolat domináns volt. Ez azonban korántsem olyan egyszerű, még manapság sem. A nehézség több forrásból fakad. Az egyik lényeges probléma az, hogy nincs olyan egyértelmű érték, amelyen túl a mérési eredmény durva hiba, s azon belül viszont természetes változékonyságként elfogadható lenne. *Masson D'Autumn*ra hivatkozva [16] arra figyelmeztet, hogy „egyre növekvő veszéllyel állunk szemben: kompetens szakemberek ellenőrzése nélkül, számítógéppel automatikusan dolgoznak fel egyre több adatot. Valójában, a legkisebb négyzettel történő kiegyenlítést felhasználók már elismerték a veszélyt. A számításokban rejlő nem kívánt kívülállóértékek elfojtásának elhalaszthatatlan gyakorlati igényét érezve, durva hibák ellen saját intuitív próba és hiba eljárások használatával a programokat robusztizálják". Ez az idézet igen tanulságos, mert az időszerű, fontos megállapítás mellett, egy elterjedt tévhit „markáns megjelenése" is megtalálható benne. Az idézet figyelmeztet arra, hogy a számítógépes adatfeldolgozás során is keletkezhet durva hiba. Valójában a billentyű, tizedesvessző hibás leütése, az adatmásolás és továbbítás során esetleg fellépő adatsérülések, a berendezés észrevétlen meghibásodásából adódó adatsérülések és egyebek mind-mind durva hibát idézhetnek elő, mely sokszor „láthatatlan" és nehezen deríthető ki. A szóban forgó idézet a robusztus módszerrel kapcsolatban eléggé elterjedt tévhitet is „jócskán" tartalmaz. A „robusztizálást", ill. a robusztus módszert, mint a durva hibát elfojtó, elnyomó (suppressing) statisztikai eszközként említi. Ha egy statisztikai módszer robusztus, akkor ez azt jelenti, hogy a módszer bizonyos (természetes) ingadozással szemben stabil kell legyen. Ez fontos dolog, viszont a durva hiba nem egy „természetes" ingadozás, így az azzal szemben érzéketlen becslés vagy statisztikai módszer nem lenne más, mint a durva hiba elrejtésére szolgáló eszköz, ami nem dicséretre méltó. Lényeg az, hogy a durva hibát és az egyéb hasonló, nagy méretű kívülállóértékeket ki kell deríteni, azonosítani kell, és el kell távolítani. Bár ez nem olyan egyszerű, ahogyan azt előbb is említettük. Erre hivatott a kívülállóérték kiderítési elmélet (outlier detection theory, outlier identification theory stb.), amely létezni létezik, azonban teljes egészében még nincs elfogadható módon egységesen feldolgozva. *Ádám J.* et al [1] τ-statisztikával történő kívülállóérték statisztikai kiderítési módszer alkalmazásáról számol be. A kívülállóérték kiderítésének az adatok matematikai-statisztikai feldolgozásába történő bevonása növeli a matematikai elemzés pontosságát.

A *(b) kategóriá*hoz tartozó kívülállóérték nem kapcsolódik közvetlenül a mérési hibákhoz, hanem a modell illeszkedési vizsgálatok különböző típusaihoz. Az utóbbi időben egyre több dolgozat foglalkozik a regressziós modell és a kívülállóérték kapcsolatával. A kívülállóérték fontos szerepet játszik regressziós modellek reziduáljainak elemzésében. Itt mind a kívülállóérték kiderítését, mind pedig a robusztus módszert alkalmazzák.

A váratlan kívülállóérték nagyon érdekes jelenség a matematikai modellezéssel kapcsolatos statisztikai elemzésben. Például egy új gyógyszer hatásának vizsgálatában vagy elemi részecskék statisztikai elemzésében a kívülállóérték nem hibára, hanem teljesen új jelenségre utalhat. Itt a kívülállóérték kiderítésére különösen nagy szükség van. Ilyen esetben új modellek alkotása válik szükségessé.

A *(c) kategóriá*hoz tartozó kívülállóértékeket véletlennel együtt járó természetes jelenségként fogjuk fel. E megközelítésből adódóan a kívülállóérték eltávolítása nem célszerű, hanem annak létezéséhez illeszkedő statisztikai módszerek alkalmazása a teendő. Ebben az esetben a robusztus statisztikai módszer egy hatékony eszköz. *Carosio* [4] kezdeményezte a robusztus módszerek geodéziai alkalmazását, s utána ezzel többen foglalkoztak. Magyarországon robusztus módszer geodéziai alkalmazásával *Detrekői* [5, 6], *Kalmár, Somogyi* és *Závoti* [25, 26] foglalkozott. A hibaelmélet részletes tárgyalása a [6]-ban található.

4. A folytonos eloszlás egy sajátossága és a kívülállóérték

Ahogyan az előző pontban tisztáztuk, a kívülállóérték egyik kategóriája a valószínűségi változó változékonyságából eredő ingadozásként fogható fel. Következésképpen felmerül az a kérdés, hogy egy valószínűségi változó mennyire képes változékonyságát „megmutatni". Egy folytonos valószínűségi változó minden egyes lehetséges értékét nulla valószínűséggel vesz fel. Amennyiben a lehetséges értékek halmaza felülről korlátlan intervallum, akkor „elméletileg" bármilyen nagy értéket is felvehet a szóban forgó folytonos valószínűségi változó. Ez a lehetőség a folytonos valószínűségi változó természetében rejlik. Alulról korlátlan folytonos valószínűségi változó esetén, hasonlóképpen bármilyen kis értéket is felvehet a valószínűségi változó. Ez az egyszerű gondolat ráébreszt bennünket arra, hogy a kívülállóérték egyik kategóriája egy olyan jelenség, amely a folytonos valószínűségi változó markáns változékonyságának megjelenése. Ez egy deduktív és kvalitatív motiváció, amelyet a „valószínűségelméleti nagy eltérések“ fejezetre támaszkodva kvantitatívan is megközelíthetünk.

A nagy eltérések elmélete kapcsolódik a központi határeloszlás tételekhez. Központi határeloszlás tételnek szokás nevezni minden olyan tételt, ami bizonyos ξ_1, ξ_2, ... valószínűségi változók megfelelően normált és központosított összegeinek sorozata eloszlásának normális eloszláshoz való konvergenciáját állítja. A központi határeloszlás tételeknél

$$P\left(\varsigma_n < x\right) \tag{1}$$

ahol ς_n-nel jelöljük az összegeket.

Itt a valószínűségek konvergenciáját $n \to \infty$ és tetszőlegesen választott, de rögzített x valós szám esetén vizsgálják.

Azonban a nagy x érték esetén, azaz n növekedésével szintén növekedő $x=x_n$ esetén

$$P\left(\varsigma_n > x\right) \tag{2}$$

valószínűség aszimptotikus viselkedésének tanulmányozása egy sor elméleti, ill. gyakorlati kérdésben bizonyult fontosnak. Ezt a problémakört nagy eltéréseknek szokták nevezni [19]. A nagy eltérések elméletéből [19, 30] ismert a következő reláció:

$$P\left(\varsigma_n > x_n\right) \sim \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{x_n} e^{-\frac{1}{2}x_n} \tag{3}$$

Itt például x_n=4 esetén a (3) formulával nagy eltérések valószínűségét kiszámítva a 0.013497741 valószínűséget kapjuk. Így, standard normális eloszlást nagy eltéréssel megközelítve, körülbelül száz mérésnél egy esetben 4-nél nagyobb értékre „juthatunk“. Viszont ha a közelítés helyett az „ideális határeloszlás“, vagyis a standard normális eloszlás lenne, akkor nulla valószínűséget kapnánk a szokásos standard normális eloszlás táblázatából. Ebben a megközelítésben, standard normális eloszlásnál is az igen nagy érték (= kívülállóérték) előfordulásának lehetősége „elméletileg“ is megalapozott.

5. Statisztikai minta és kívülállóérték

Előzetesen vegyünk egy m-elemű statisztikai mintát egy folytonos valószínűségi változóra. Itt az „előzetesen“ határozó arra utal, hogy a mintavételt tovább fogjuk folytatni, azaz a szóban forgó minta nem egy végleges minta, hanem csak egy előzetes m-elemű minta. Legyen x a legnagyobb elem ebben az előzetes mintában. Most tovább folytatjuk a mintavételt addig, amíg x-nél nagyobb mintaelemet kapunk. Jelöljük n-nel e további mintaelemek számát. Ekkor ennek az n számnak (nyilvánvaló, hogy n egy diszkrét valószínűségi változó) valószínűségi eloszlása

$$p(n) = \frac{m}{(m+n)(m+n-1)}, \quad n = 1, 2, .. \tag{4}$$

formulával adható meg.

A (4) formula, valamint a kérdés felvetése *Wilks*-től [29] származik. Itt *Wilks* a valószínűségi változó folytonosságát feltételezte, ami lényeges, mert nélküle a kérdés felvetése sem jöhet szóba. Nem tett említést azonban a valószínűségi változó korlátosságáról, illetve korlátlanságáról. Így korlátos és folytonos valószínűségi változóra is érvényes a (4) formula. Ez valóban lehetséges, azonban hozzá kell tennünk, hogy ez a lehetőség csak azért van meg, mert korlátos, nyílt intervallum és számegyenes között „semmi különbség“ nincs az általános topológiai értelemben (itt az általános topológia szó a modern matematika egyik ágát jelenti). Viszont, ha zárt intervallum lenne a valószínűségi változó lehetséges értékeinek halmaza, akkor már az állítás veszíti el az érvényességét. Másfelől a gyakorlati alkalmazást, közelebbről a kívülállóértéket szem előtt tartva, a *Wilks*-formulát folytonos felülről korlátlan valószínűségi változóra vonatkoztatom.

Rényi Alfréd [23] egy ξ_k mintaelemet kiemelkedőnek nevezett, ha $\xi_k > \xi_j$, midőn $j < k$. Más szóval, egy mintaelem kiemelkedő, ha az összes megelőző mintaelemeknél nagyobb. E definíció szerint ξ_1 mindig (triviálisan) kiemelkedő. Legyen υ_o =1; jelölje υ_1 az első 1-nél nagyobb sorszámú (tehát az első nem triviális) kiemelkedő mintaelem sorszámát, υ_2 az első υ_1-nél nagyobb (tehát a második nem triviális) kiemelkedő elem sorszámát (n=2, 3,...) . Ebben a jelölésben, a υ_n valószínűségi változó határeloszlását *Rényi* a következő tételben határozta meg [23].

Tétel. Ha mN jelenti egy folytonos sokaságból vett végtelen minta első N eleme között a kiemelkedő elemek számát, akkor 1 valószínűséggel fennáll a

$$\lim_{N\to\infty} \frac{\mu_N}{\log N} = 1 \tag{5}$$

reláció.

Ezt a tételt is a felülről korlátlansági feltétellel kiegészítve, azt az állítást nyerhetjük, hogy a „kívülállóérték" előfordulása a mintaelem számával logaritmikusan arányos. Itt jegyezzük meg, hogy a *Wilks* és *Rényi* által definiált mintaelemek nem valódi kívülállóértékek, hanem azok között előfordul a kívülállóérték, s azok közül némelyik igen közel állhat a valódi kívülállóértékhez.

Mind a *Wilks*-formula, mind pedig a Rényi-tétel használatával azt állíthatjuk, hogy felülről korlátlan, folytonos valószínűségi változóra vonatkozó, elég nagy statisztikai minta esetén bármilyen nagy mintaelem fordulhat elő. Ez a tény azt mutatja, hogy a kívülállóérték statisztikai mintában történő előfordulása bizonyos esetekben „természetes jelenség", továbbá elméletileg is alátámasztja a 3. pontbeli kívülállóértékek kategorizálását.

6. Kiugróérték vagy kívülállóérték?

A magyar geodéziai irodalomban az angol „outlier", „outlying" observation szavak helyett a kiugróérték kifejezést használják, a „kiugróérték" kifejezés elterjedőben van. Tekintettel arra, hogy a magyar geodéziai irodalom – tudomásom szerint – „outlier"-ral nemigen foglalkozott, így a szóban forgó kifejezés még nem teljesen honosodott meg. Így változtatásra is van lehetőség. Ha az „outlier" szó szemantikájára jobban figyelünk, akkor a kívülállóérték, kívülhelyezkedő érték szó jobban tükrözi a fordítást, valamint magát a fogalmat is, mint a kiugróérték szó, hiszen a „lie" szónak idevágó értelmezése az Oxford Advanced Learner's Dictionary, Encyclopedic Edition, Oxford, London szótár szerint 'be situated', ami nem más, mint helyezkedik, vagy áll valahol (pl. The town lies on the coast. Sri Lanka lies on the south of India etc.)

Egy másik indokként említem, hogy az „outlier" fogalmára vonatkozólag egy angol matematikai statisztikai szakirodalomban volt szó „inlier"-ról is. Tehát a beugró-érték és a kiugróérték kifejezésnél jobban illeszkedne a belülálló (belülhelyezkedő), ill. kívülálló (kívülhelyezkedő) kifejezés.

Nemcsak az angol, hanem pusztán a magyar nyelv szemantikáján belül gondolkodva is a „kiugróérték" terminológia kifogásolható. A Magyar Értelmező Kéziszótár (szerkesztette: *Juhász József, Szőke István, O. Nagy Gábor* és *Kovalszky Miklós*, Akadémiai Kiadó, Budapest, 1975) szerint a kiugrik igének több mint tíz jelentése van. Ezek közül a 9. jelentése a következő: szembetűnik, feltűnően látszik. A vastag betűs cím jól kiugrik (a szövegből). <Rejtett hiba> előbukkan, megkerül. A szótárból történő kiválasztási értelemben e jelentés a fogalomhoz legközelebb áll, azonban mégsem annyira jó, mert a „kiugróérték" kifejezés mindjárt azt a benyomást, ill. érzést kelti, hogy az az érték, amiről a fogalom szól, önmagában feltűnően látszik, rögtön „kiugrik", ez viszont ellentmond az előbb kifejtett kiderítési és egyéb nehézségeknek.

Pár évvel ezelőtt egy magyar napilapban azt olvastam, hogy „Ez egy kívülálló számára érthetetlen." Ez is azt sugallja, hogy a kívülálló szó a mindennapi életben egy természetes szó. Összefoglalva, azt a javaslatot szeretném tenni, hogy az „outlier" magyar megfelelőjeként a továbbiakban a „kívülállóérték" szót használjuk.

7. Következtetések

A kívülállóérték egy olyan fogalom, amely régi keletű, de még nem teljesen tisztázódott, ill. alakult ki annak elmélete és módszertana. A dolgozatban kifejtett gondolatok alapján az alábbi következtetések vonhatók le:

- a fogalom sokszínűsége és egymástól merőben eltérő feladatokban való megjelenése miatt a kategorizálás jobban járható út és kiindulópont lehet a további alapos tanulmányozáshoz;
- a jelen dolgozatban megfogalmazott kategorizálás abban az értelemben integráló hozzáállás és definíció, hogy bármelyik kívülállóérték a felsorolt három kategória valamelyikéhez tartozik;
- a kategorizálás egy másik előnye az, hogy a kívülállóérték fajtájához megfelelő módszer kiválasztását előtérbe helyezzük, s ezáltal a különféle tévhitek, ill. félreértelmezések helye szűkül;
- a dolgozat 4. és 5. pontbeli valószínűségelméleti-matematikai statisztikai eredményeket – tudomásom szerint – sem maguk a szerzők, sem más kutatók nem hozták összefüggésbe a kívülállóértékekkel. Ezzel szemben e sorok írója az említett eredményeket a kívülállóérték elmélet kvantitatív alapozására alkalmazza. Itt külön kiemelnénk, hogy ezek az eredmények eloszlásfüggetlen relációk;
- a kívülállóértéknek a nagy eltérések elméletével történő kapcsolása új elméleti meggondolás.

8. Köszönetnyilvánítás

A szerző köszönetet mond *dr. Alpár Gyulá*nak, a műszaki tudomány doktorának, akinek a dolgozat bírálata során tett konstruktív kritikája hasznos volt.

Végezetül megjegyezzük, hogy e tanulmány OTKA TO35242 nyilvántartási számú kutatási támogatással készült el.

IRODALOM

1. *Ádám J.–Bányai L.–Csapó G.–Szűcs L.*: Szélső pontosságú geodéziai mérések a sóskúti mikrohálózatban mozgásvizsgálati célból, Geodézia és Kartográfia, 2001/4., 16–22.

2. *Bernoulli, D.*: The most probable choice between several discrepant observations and the formation therefrom of the most likely induction, reprinted in. Biometrika, 48 (1961), 1–18.

3. *Boscovich, R. J.*: De litteraria expeditione per pontificiam ditionem, et synopsis am plioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii, 4 (1757), 353–396.

4. *Carosio, A.:* Robuste Ausgleichung, Vermessung, Photogrammetrie, Kulturtechnik, 11 (1979), 293–297.

5. *Detrekői, Á.*: A durva hibák figyelembevétele a mérési eredmények feldolgozásakor, Geodézia és Kartográfia, 1986/3., 155–160.

6. *Detrekői Á.*: Kiegyenlítő számítások, Tankönyvkiadó, Budapest, 1991

7. *Ferguson, T. S.:* On the rejection of outliers, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume I, 1961, 253–287.

8. G*ather, U.*: Outlier Models and Some Related Inferential Issues, in The Exponential Distribution, ed. by N. Balakrishnan and A. P. Basu, University of Missouri-Columbia, Gordon and Breach Publishers, 2000, 221–239.

9. *Grubbs, F. E.*: Sample criteria for testing outlying observations, Annals of Mathematical Statistics, 21 (1950), 27–58.

10. *Grubbs, F. E.*: Procedures for detecting outlying observations in samples, Technometrics, 11 (1961), 1–21.

11. *Hawkins, D. M.:* Identification of outliers, Chapman and Hall, London, 1980.

12. *Hekimoglu, S–Koch, K. R.*: How can reliability of the robust methods be measured? Third Turkish-German Joint Geodetic Day, Istanbul, Turkey, June 1–4, 1999, 179–196.

13. *Helmert, R.*: Ueber den Maximalfehler einer Beobachtung, Zeitschrift f. Vermessungswesen, 6 (1877), 131–147.

14. H*uber, P. J.*: Robust estimation of a location parameter, Annals of Mathematical Statistics, 35 (1964), 73–101.

15. *Koch, K. R.*: Parameter Estimation and Hypothesis Testing in Linear Models, Springer, 1999

16. *Kubik, K.–Wenig, W.–Frederiksen, P.*: Oh, Gross errors!, Australian Journal of Geodesy, Photogrammetry and Surveying, 42 (1985), 1–18.

17. *Legendre, A. M.*: Nouvelles méthodes pour la détermination des orbites des cometes, Courcier, Paris, 1805

18. *Legendre, A. M.*: Méthode des moindres carrés, pour trouver le milien le plus probable entre les résultats de différentes observations, Mem. Inst. de France, 1810. 149–154.

19. *Linnik, Yu. V.*: On the probability of large deviations for the sums of independent variables, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume II, 1961, 289–306.

20. *Masson D'Autumn, G.*: Manuscript presented at the Symp. Comm. III of ISP, London 1971, 6–8.

21. *Monhor, D.:* Valószínűségelmélet, jegyzet, Székesfehérvár, 2002

22. *Monhor, D.*: Mérési hibák, központi határeloszlás tételek, Hagen-féle hipotézisek és normális eloszlás, Geodézia és Kartográfia, 2001/1. 11–16.

23. *Rényi, A.*: Egy megfigyeléssorozat kiemelkedő elemeiről, A Magyar Tudományos Akadémia Matematikai és Fizikai Osztályának Közleményei, 12 (1962), 105–121.

24. *Schröder, W.*: Why research into history of geosciences? Acta Geod. Geoph. Hung. 36 (2001) 225–229.

25. *Somogyi, J.–Kalmár, J.:* The use of robust estimation in geodetic data processing, Acta Geod. Geoph. Hung., 26 (1991), 57–68.

26. *Somogyi, J.–Závoti, J.*: Robust estimation with iteratively reweighted least squares method, Acta Geod. Geoph. Hung., 28 (1993), 413–420.

27. *Stefansky, W.*: Rejecting outliers by maximum normed residuals, Annals of Mathematical Statistics, 42 (1971), 35–45.

28. *Tukey, J. W.*: A survey of sampling from contaminated distributions, Contributions to Probability and statistics, I. Olkin, S. G. Ghunje, W. Hoffding (eds.), Stanford University Press, 1960, 448–485.

29. *Wilks, S.*: Recurrence of extreme observations, Journal of Australian Mathematical Society, 1 (1959), 106–112.

30. *Yurinski, V. V.*: Exponential Inequalities for Sums of Random Vectors, Journal of Multivariate Analysis, 6 (1976), 473–499.

Clarifications of and complements to the concept of outlier

Davaadorjin Monhor
Summary

In the paper the concept of outlier is examined from the standpoint of possible causes and ways of dealing with them. As a result of the examination, a general and descriptive definition of outlier was introduced. The definition gives rise to three categories of all types of outliers. These categories are (a) unwanted outliers, i.e., gross-errors and similar kinds of outliers, (b) outliers arising from model-fitting analysis of residuals and unexpected outliers, (c) outliers that manifests intrinsic variability of randomness.

Based on the results in [23, 29], quantitative probabilistic reasoning is given to explain the presence of outliers in statistical sample. The paper also contains a brief outline of outliers and ways of dealing with outliers in historical setting.

❑