Két valószínűségi változó korrelációjának különböző mérőszámai[1]

STEINER FERENC, HAJAGOS BÉLA[2]

A dolgozat kimutatja, hogy az (1) szerint definiált, r_c-vel jelölt klasszikus korrelációs együtthatót katasztrofális mértékben torzítja akár egyetlen, nem túl távol eső értékpár is, még akkor is, ha n = 100 adatpárunk van. A (8) egyenletben megadott rezisztifikált korrelációs együttható Monte-Carlo vizsgálataiból kiderült, hogy a (8) kifejezés egyben robusztus is.

F. STEINER, B. HAJAGOS: Different characteristics of the correlation of two sets of data

It is shown in the present article that the classical definition of calculated correlation coefficient (Eq. 1) is in a high degree not resistant. The resistificated formula (Eq. 8) is, in the contrary, not only resistant, but also robust, too.

1. Példa a korrelációs együtthatónak nem rezisztens voltára

Ha n elemű mintát veszünk mind a ξ-vel, mind az η-val jelölt valószínűségi változóból, az így kapott n db (x_i,y_i) értékpár alapján a klasszikus statisztika a következő formula szerinti becslést ajánlja a korrelációs együttható helyes értékének ($r_{true} \equiv r_t$ -nek) a becslésére:

$$r_c = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \qquad (1)$$

ahol a c index arra utal, hogy adatpárokból való *számítás* eredményét adja az (1) egyenlet ($r_c \equiv r_{calculated}$). Az $\bar{x}$, ill. $\bar{y}$ szokásosan az x_i-k, ill. y_i-k számtani átlagát jelentik.

Az (1) egyenletet számtalan statisztikai könyv közli. Ha ezek közül mégis [CRAMÉR 1945]-re hivatkozunk, azt azzal indokolhatjuk, hogy a szerző egy, a klasszikus statisztikában gyakran idézett esetre, amikor is ξ és η egyaránt Gauss-típusú valószínűségi változók, megadja az r_c valószínűségsűrűség függvényének a képletét is a valódi r_t értékhez és a mintavételkor választott n-hez. A formula a következő:

$$f(r_c) = \frac{n-2}{\pi}\left(1-r_t^2\right)^{\frac{n-1}{2}}\cdot\left(1-r_c^2\right)^{\frac{n-4}{2}}\int_0^1 \frac{x^{n-2}}{\left(1-r_t r_c x\right)^{n-1}\cdot\sqrt{1-x^2}}\,dx \qquad (2)$$

(A fenti képletet [STEINER 1990] a 265. oldalon közli). A ξ és η valószínűségi változók korrelálatlansága esetén, amikor tehát $r_t = 0$, a (2) sűrűségfüggvény alakja a következőre egyszerűsödik:

$$f(r_c) = \frac{1}{\sqrt{\pi}}\frac{\Gamma[(n-1)/2]}{\Gamma[(n-2)/2]}\left(1-r_c^2\right)^{\frac{n-4}{2}} \qquad (3)$$

([STEINER 1990], 266. oldal).

A fenti formulák nélkül is természetesnek fogjuk tartani, hogy amennyiben n-et minél kisebbre választjuk, az (1) formula szerinti r_c érték r_t körüli ingadozása egyre nagyobb lesz. A [STEINER 1990] 265. oldalán közölt, szintén [CRAMÉR 1945] munkájából átvett ábra alapján kimondható, hogy n = 10 db adatpár választása az r_c-becslések megengedhetetlenül nagymértékű ingadozását idézi elő; szinte ijesztő mértékűnek minősíthető ez a bizonytalanság a korrelálatlanság (r_t = 0) esetén. Ezek után az olvasó aligha fogja sokallni, hogy nemcsak ebben a pontban, hanem a dolgozat további pontjaiban szereplő Monte-Carlo vizsgálatok során is *a szerzők következetesen n = 100 adatpár mintavételezéséből indulnak ki.*

Az *1. ábra* erre az esetre (n = 100) mutatja be a (3) formula szerinti, nyilván az origóra szimmetrikus — és immár elfogadható keskenységű — görbét. De ez az a pont, ahol a modern statisztika egyik alapkérdését már nem mulaszthatjuk el feltenni: az (1) formula által szolgáltatott becslések vajon kellően rezisztensek-e? Vagyis: kiugró adatpárok nem torzíthatják-e megengedhetetlen mértékben az r_c értékét, még az outlierek kisarányú fellépte esetén is?

Konkretizáljuk a ξ és η valószínűségi változókat az

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left[-(x-10)^2/2\right]$$

és

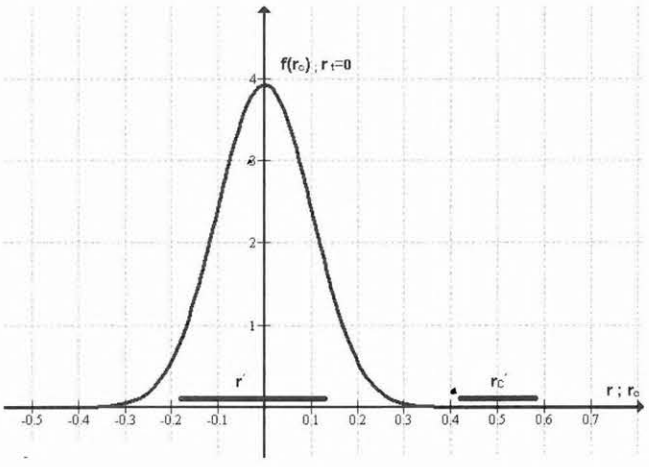
$$f(y) = \frac{1}{\sqrt{2\pi}}\exp\left[-(y-10)^2/2\right]$$

valószínűségsűrűség-függvényeikkel, azaz mindkét esetben legyen egységnyi a szórás (ami itt egyben a skálaparaméter is), a helyparaméterek is azonosak: 10-es értékűek, korreláltságról pedig szó sincs (r_t = 0). Az utóbbira utalva kiugró mintaelemet is tartalmazó 100 db értékpárt könnyen generálhatunk úgy, hogy a jelen esetben minimumra redukáljuk az outlieradatpár-arányt: legyen csak egyetlen belőle, mégpedig a (0,0) adatpár. A választott értékpár, az (x,y) síkon egyetlen pont, mint ahogyan pontokként jelennek meg az imént definiált $f(x)$ és $f(y)$ szerint 99-szer szabályosan generált x_i és y_i Gauss-véletlenszámok (x_i,y_i) értékpárjai is. Az utóbbi 99 adatpár pontfelhőjének közepe nyilván a (10,10) pont, s

[1] Beérkezett: 2008. január 15-én

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

így kiderül, hogy a (0,0) origó, mint outlier, egyáltalában nincs erőltetetten távol a pontfelhő középpontjától: a „3σ-szabályra" gondolva, ami most $S = 1$ miatt a 99 pont 3 sugarú körön belüliségét mondja ki igen nagy (de persze nem pontosan 1) valószínűséggel, egyben azt is jelenti, hogy az outlier-távolság (10,10)-től a pontfelhő középpontjától még ötszöröse sincs a szabályos pontpárok 3σ körsugarának.



1. ábra. A (3) szerinti, azaz az $r_t = 0$ esethez tartozó $f(r_c)$ -görbe (l. (1) egyenlet), $n = 100$ Gauss-eloszlású adatpár esetére Egyetlen nem túl távoli outlier a jobb oldalt feltüntetett szakaszra torzítja a 10 felvett esetben az ilyenkor r_c' -vel jelölt értékeket. Pontosan ugyanazon értékpár-százasok a (8) egyenlettel számolva origó körüli r'-értékekre vezetnek, azaz semmiféle outlier-torzítás nem jelentkezik

Fig. 1. The probability density curve $f(r_c)$ (for the definition of r_c see Eq. 1) in uncorrelated case for random number-pairs of Gaussian type, if $n = 100$ pairs of data are given. (For the analytical expression see Eq. 3.) One single, not very far lying outlier-pair distorts fully the result (see the r_c' interval above the abcissa); in the contrary, the resistifyed Eq. 8 results for the same 10 cases points of the r'-interval which corresponds to the outlier-free $f(r_c)$ curve for the uncorrelated case

Az 1. ábra jobb oldalán az abszcissza fölé húzott egyenes szakasz jelzi annak a 10 db, (1) szerint számított értéknek a terjedelmét, amelyet a szabályosan generált 99 értékpár 10 realizációja eredményezett a (0,0) outlier esetén. (A torzított és ezért megkülönböztetésül r_c' -vel jelölt 10 db, (1) szerint számított érték ui. véletlenszerűen a következőknek adódott: 0,59; 0,42; 0,49; 0,52; 0,50; 0,56; 0,47; 0,50; 0,49; 0,49.) Ezek az r_c' értékek olyan távol vannak az 1. ábra valószínűségsűrűség-görbe gyakorlatilag számításba veendő értéktartományától, hogy az 1. ábra az *(1) formula nem rezisztens voltát szembeszökően bizonyítja.* Megemlítjük, hogy ugyanezt a következtetést vonhatjuk le az $r_t = 0{,}2$-höz és $r_t = 0{,}5$-höz tartozóan a szerzők által kiszámított szabályos $f(r_c)$ görbék és az ugyancsak egyetlen: (0,0) outlierhez tartozó torzított 10 db r_c' és a valódi r_t érték nagy eltérése között, ezért ezen ábrák közlésétől eltekintettünk.

2 A korrelációs együttható klasszikus formulájának rezisztens analogonjai

Az 1. pont végén megfogalmazott következtetés aligha meglepő, hiszen az (1) képletben szereplő $\bar{x}$ és $\bar{y}$ számtani átlagok már önmaguk sem az (x_i,y_i) $(i = 1,\ldots,n)$ minták rezisztens helyparaméter-becslései. Logikus tehát első lépésként az $\bar{x}$ és $\bar{y}$ becsléseket rezisztens helyparaméter-becslésekre cserélni az (1) formulában; a leggyakoribb érték szinte kínálkozik erre a cserére. Figyeljük meg azonban azt is, hogy a mért értékek és az átlagok különbségei súlyozás nélkül szerepelnek (1)-ben, így az új formulának logikus figyelembe vennie a leggyakoribb értékszámításkor az analóg különbségekre alkalmazott súlyokat is.

A leggyakoribb értékeket az alábbiakban nem a megszokott (pl. a [STEINER 1990]-ben) megadott módon írjuk fel, hanem úgy, hogy a fenti analóg lépések a legkönnyebben legyenek követhetők.

A leggyakoribb értékeket ${}^kM$-val jelöljük, mivel számításakor egy k-val jelölt konstanst 1-nek, 2-nek vagy 3-nak választunk, így az ${}^1M$, ${}^2M$, ${}^3M$ jelöléssel élhetünk. (Később tisztázandó okokból k = 2-t választva ezt általában nem tüntetjük fel, azaz ${}^2M \equiv M$.) A leggyakoribb érték index-helyén adjuk meg, hogy az x_i vagy y_i $(i=1,\ldots,n)$ helyparaméteréről van-e szó, így most az analogon-képzés első lépésén hamar túl lehetünk úgy, hogy (1)-ben $\bar{x}$ helyére ${}^kM_x$-et, $\bar{y}$ helyére pedig ${}^kM_y$-t írunk.

A súlyokat illetően aligha kerülhető meg az, hogy megadjuk a leggyakoribb értékek definícióit az alábbiakban, ha most nem is a szokásos formalizmus a legcélravezetőbb. *Az x_i, ill. y_i $(i=1,\ldots,n)$ minták ${}^kM_x$, ill. ${}^kM_y$ leggyakoribb értéke az alábbi egyenleteket teljesíti:*

$$\sum_{i=1}^{n}\left[(k\varepsilon_x)^2+\left(x_i-{}^kM_x\right)^2\right]^{-1}\left(x_i-{}^kM_x\right)=0\,, \qquad (4)$$

valamint

$$\sum_{i=1}^{n}\left[(k\varepsilon_y)^2+\left(y_i-{}^kM_y\right)^2\right]^{-1}\left(y_i-{}^kM_y\right)=0\,. \qquad (5)$$

A (4)-ben szereplő, ε_x-szel jelölt, és az x_i adatok többségének tömörödését jellemző dihézió, valamint az (5)-ben ε_y-nal jelölt analóg mennyiség az alábbi egyenleteknek tesz (természetszerűleg a (4)-gyel és (5)-tel egyidejűleg) eleget:

$$\varepsilon_x^2=3\left\{\sum_{i=1}^{n}\left[\left(x_i-{}^kM_x\right)^2\right]^2\cdot\left[\varepsilon_x^2+\left(x_i-{}^kM_x\right)^2\right]^{-2}\right\}\cdot \left\{\sum_{i=1}^{n}\left[\varepsilon_x^2+\left(x_i-{}^kM_x\right)^2\right]^{-2}\right\}^{-1}, \qquad (6)$$

valamint

$$\varepsilon_y^2=3\left\{\sum_{i=1}^{n}\left[\left(y_i-{}^kM_y\right)^2\right]^2\cdot\left[\varepsilon_y^2+\left(y_i-{}^kM_y\right)^2\right]^{-2}\right\}\cdot \left\{\sum_{i=1}^{n}\left[\varepsilon_y^2+\left(x_i-{}^kM_x\right)^2\right]^{-2}\right\}^{-1}. \qquad (7)$$

A (4) és (5) egyenletekből már leolvashatjuk, hogy az új korrelációs együttható-formulákban a kétféle analóg (az

(1)-ben $(x_i - \bar{x})$ és $(y_i - \bar{y})$ helyére kerülő) $(x_i - {}^kM_x)$, ill. $(y_i - {}^kM_y)$ különbségek milyen súlyokkal szorzandók, hogy *a rezisztifikált, ^kr-val jelölt korrelációs együtthatót kapjuk eredményül:*

$$ {}^k r = \frac{\sum_{i=1}^{n}\left[(k\varepsilon_x)^2 + (x_i - {}^kM_x)^2\right]^{-1}\cdot(x_i - {}^kM_x)}{\left\{\sum_{i=1}^{n}\left[(k\varepsilon_x)^2 + (x_i - {}^kM_x)^2\right]^{-2}\cdot(x_i - {}^kM_x)^2\right\}^{1/2}} \cdot \frac{\left[(k\varepsilon_y)^2 + (y_i - {}^kM_y)^2\right]^{-1}\cdot(y_i - {}^kM_y)}{\left\{\sum_{i=1}^{n}\left[(k\varepsilon_y)^2 + (y_i - {}^kM_y)^2\right]^{-2}\cdot(y_i - {}^kM_y)^2\right\}^{1/2}} \quad (8) $$

Amennyiben $k = 2$, az r bal felső indexét nem szoktuk jelölni (ahogyan ezt az ^kM leggyakoribb értékeknél is elhagyjuk).

Az 1. ábra jobb oldalán levő intervallumba eső, torzított r_c' értékekre vezető (mivel egyetlen outlier-adatpárt is tartalmazó) adatpár-százasokkal kiszámítottuk a (8) szerinti r' értékeket is. Úgy látszik, hogy a fentiekben végrehajtott rezisztifikálás eredményre vezetett: ugyanazokhoz a százelemű adatpár-halmazokhoz immár nem a klasszikus (1) formula szerinti, a függetlenség esetén teljesen irreális r_c'-értékek adódtak, hanem a valódi r_t korrelációs együttható zérus voltára, azaz a korrelálatlanságra utalva az origó körül jelennek meg (l. az origó feletti egyenes szakaszt) persze olyan ingadozással, amelyet $n = 100$ esetén az 1. ábra görbéje szerint az elmélet is megkövetel: –0,16 és +0,13 között. (Talán nem felesleges ugyanolyan sorrendben azt a tíz r'-értéket is felsorolni, amelyek ugyanazon adatpár-százasokhoz adódtak, amelyek az 1. pont végén felsorolt r_c'-értékeket eredményezték az (1) formula használatakor: +0,13; –0,16; –0,11; +0,01; –0,05; +0,04; –0,14; +0,06; –0,06; 0,11.)

3. A rezisztifikált korrelációs együtthatót: ^kr-t szolgáltató formulák Monte-Carlo vizsgálata

3.1. A vizsgálat módszere

3.1.1. A vizsgálatokkor alkalmazott véletlenszám-típusok

Talán szokatlan egész könyvek alapvető szemléleteire, ill. az előbbiek által egyértelműen sugallt, ezért immár tényként elfogadható eredményre mintegy globális jelleggel hivatkozni, de [STEINER 1990] és [STEINER 1997] (amelyek mögött egy egész statisztikai munkacsoport több évtizedes munkája áll) egyértelműen az

$$ f_a(x) = n(a)\cdot\left(1 + x^2\right)^{-a/2} \qquad (a > 1) \qquad (9) $$

típuscsalád elemeivel való sűrűségfüggvény-modellezés kiemelkedő előnyeit mutatja. (A (9)-beli $n(a)$ normálási faktor gammafüggvényekkel $\Gamma[a/2]\cdot\left\{\Gamma[(a-1)/2]\cdot\sqrt{\pi}\right\}^{-1}$ szerint számítható).

A (9)-ben az $a \to \infty$ határesetben a Gauss-típus adódik; a csökkentésével azután a szárnyak tartománya egyre nagyobb valószínűségi súlyt kap, míg az $a = 1$ érték közelébe nem érünk, ahol a szárnyak súlya már olyan nagy, hogy a gyakorlatban előforduló esetek ritkán modellezhetők ezekkel az $f_a(x)$ sűrűségfüggvényekkel.

A statisztika gyakorlatában előforduló a hibaingadozástípusok valószínűségsűrűségét legáttekinthetőbben úgy írhatjuk fel, hogy bevezetjük a $t = 1/(a - 1)$ mennyiséget, ekkor ugyanis ez a g-vel jelölt típus-előfordulási valószínűségfüggvény egyszerűen

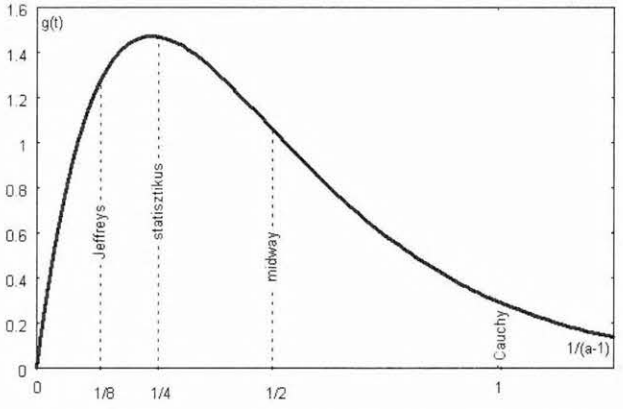
$$ g(t) = 16\cdot t\cdot e^{-4t} \qquad (10) $$

alakú.

A *2. ábra* ([STEINER 1980] 233. oldaláról bizonyos átjelölésekkel átvéve) grafikusan mutatja be ezt a függvényt, amelyik határozott maximumot mutat a = 5-nél (t = 1/4-nél), megfelelően a [DUTTER 1986/87]-ben a geostatisztikára vonatkozóan közöltekkel. Ennél lényegesen kisebb, de nem elhanyagolható valószínűségsűrűség jellemzi az a = 2 esetét, azaz a (9) alapján a

$$ f_C(x) = \frac{1}{\pi}\frac{1}{1 + x^2} \qquad (11) $$

sűrűségfüggvényű ún. Cauchy-eloszlást. Hogy ilyen eloszlással is foglalkoznunk kell (azaz $g(1)$ nem tekinthető elhanyagolható valószínűségsűrűség értéknek), arra vonatkozóan [LANDY, LANTOS 1991]-re, annak ábrájára hivatkozhatunk, de talán még inkább [TARANTOLA 1987] javaslatára, amely szerint, ha joggal kell feltételeznünk aktuális esetünkben nem elhanyagolható mértékben outliereket, akkor az $f_C(x)$ eloszlást feltételező statisztikai algoritmussal célszerű dolgoznunk.



2. ábra. Az $f_a(x)$ eloszlástípusok $g(t)$ valószínűségsűrűséggörbéje: a = 5, statisztikus (t = 1/4); a = 3, midway (t = 1/2); a = 2, Cauchy (t = 1). Az ábra a fentieket kiegészítve az a = 9 (Jeffreys, t = 1/8) helyét is kijelöli

Fig. 2. In the present article the Monte-Carlo calculations were carried out for the $f_a(x)$ distributions a = 5 (statistical, t = 1/4); a = 3, (midway, t = 1/2); a = 2 (Cauchy-type, t = 1). The figure shows also the case belonging to a = 9 (Jeffreys-type, t = 1/8) and the probability density-curve $g(t)$ of the type occurrences

Ami a (10)-ből következő $g(0) = 0$ állítást illeti, azaz azt a klasszikus szemlélet számára meghökkentő kijelentést, hogy a statisztika gyakorlatában a Gauss-típus előfordulása zérus valószínűségsűrűségű, hangsúlyozottan kell utalni arra, hogy a modern statisztika irodalmából számos utalás felsorolásával támaszthatnánk ezt alá. Talán nagyobb súlya van azonban annak a múlt század közepéről származó megállapításnak, amelyet Sir Harald JEFFREYS tett (nagyon sok geofizikai mérési adatrendszer értelmezése után), hogy

ti. kis szárnyú eloszlásokat keresve legfeljebb 6 és 10 közé eső a-értékekkel jellemzett $f_a(x)$ sűrűségfüggvényekkel modellezhető mérési adatrendszereket talált, de Gauss típusút soha (idézi [KERÉKFI 1978]). A 2. ábra görbéje mutatja, hogy az origó (vagyis a Gauss-típus) közelében gyorsan nőnek nagy értékekre a valószínűségsűrűségek, ezért a fentiekben megismert $6 < a < 10$, ún. Jeffreys-intervallum egyik, (Gauss-hoz közeli,) $a = 9$ típusparaméterű $f_a(x)$ sűrűségfüggvényét önkényesen ugyan, de nem minden alap nélkül nevezhetjük Jeffreys-eloszlásnak és alkalmazhatjuk azokban a gyakorlati eseteinkben, amikor a szokásosnál szignifikánsan rövidebbek mért adataink szárnytartományai.

Végül megemlítjük (különösebb indoklás mellőzésével, de a 2. ábrára pillantva logikusan), hogy az $a = 3$-hoz tartozó, „midway"-nek nevezett típus vizsgálatát is szükségesnek tartjuk, Így tehát a következő véletlenszám-típusok közül választjuk vizsgálatunk tárgyait:

— Jeffreys-eloszlástípus, $a = 9$, sűrűségfüggvénye

$$f_J(x) = \frac{35}{32}\left(1+x^2\right)^{-9/2}; \qquad (12)$$

— statisztikus-eloszlástípus, $a = 5$, sűrűségfüggvénye

$$f_{st}(x) = \frac{3}{4}\left(1+x^2\right)^{-5/2}; \qquad (13)$$

— midway-eloszlástípus, $a = 3$, sűrűségfüggvénye

$$f_{mw}(x) = 0{,}5\left(1+x^2\right)^{-3/2}; \qquad (14)$$

— Cauchy-eloszlástípus, $a = 2$, sűrűségfüggvénye

$$f_C(x) = \frac{1}{\pi}\cdot\frac{1}{\left(1+x^2\right)}. \qquad (15)$$

3.1.2. Adott r_t korrelációs együtthatójú adatpárok előállítása

Azonos típusú ξ és η valószínűségi változók r_t valódi korrelációs együtthatójának megfelelő, n = 100 db (x_i,y_i) értékpár gépi számítását kell megoldanunk, mivel a Monte-Carlo számítások a 3.1.1. végén felsorolt típusoknál nem mellőzhetők, azaz már nem választhatjuk az analitikus levezetés kényelmes útját, amely Gauss-esetben követhető volt és a (2) és (3) sűrűségfüggvények görbe-formuláihoz vezetett. Ezek után igen nagyra kell az N ismétlési számot választani ahhoz, hogy sűrűségfüggvény-görbékként elfogadható ${}^3r,r$ és ${}^1r$ hisztogramokat kapjunk eredményül.

Mielőtt a 3.1.1. végén felsorolt típusokkal foglalkoznánk, kénytelenek vagyunk egy kis kitérőt tenni.

Ismeretes, hogy a szimmetrikus stabilis típuscsalád sűrűségformulája

$$f_\alpha(x) = \frac{1}{\pi}\int_0^\infty e^{-t^\alpha/\alpha}\cdot\cos(tx)\,dt \qquad (0 < \alpha \le 2) \qquad (16)$$

(ld. pl. [STEINER 1990] 34. old.). Az $\alpha = 1$ típusparaméter Cauchy-típust eredményez; ez az egyetlen eset, amikor közvetlen kapcsolat, mi több, egybeesés van a 3.1.1. végén felsorolt típusokkal. (Az $\alpha = 2$ típusparaméter Gauss-típust szolgáltat, de ez most számunkra érdektelen.)

Az $f_\alpha(x)$-ek sajátságaira az imént alkalmazott jelzőt, hogy ti. egy eloszlás *stabilis*, még definiálnunk kell. Ez annyit jelent, hogy amennyiben ξ és η azonos α típusparaméterű valószínűségi változók, akkor összegük, a $(\xi+\eta)$ valószínűségi változó is ugyanazzal az α-val lesz jellemezhető.

Elméletileg legegyszerűbben erre a típuscsaládra definiálható az r_t korrelációs együttható. A ξ és η korrelációját r_t jellemzi, ha fennáll a következő egyenlet:

$$\eta = r_t\cdot\xi + \left(1+r_t^\alpha\right)^{1/\alpha}\cdot\zeta \qquad (17)$$

(l. [STEINER 1990] 250. old. 7-10.), ahol a ζ-t ugyanaz az α jellemzi, mint ξ-t, és ξ-től független valószínűségi változó. Ha x_i, y_i és z_i a ξ, η és ζ valószínűségi változóhoz tartozó, gép által generált véletlen számok, akkor triviális, hogy a Monte-Carlo vizsgálatokhoz szükséges (x_i,y_i) adatpár

$$y_i = r_t\cdot x_i + \left(1+r_t^\alpha\right)^{1/\alpha}\cdot z_i \qquad (18)$$

szerint adódik.

A továbbiakban a egész dolgozatban a (18) egyenlet szerint végezzük számításainkat, — de milyen jogcímen? Hiszen (16) a stabilis eloszlásokra érvényes, és különben is: a 3.1.1. végén felsorolt $f_a(x)$ típusoknál milyen α alkalmazandó?

A szerzők nagy örömmel közlik olvasóikkal, hogy az $f_a(x)$-család és az $f_a(x)$-szupermodell annyira szoros rokonságban állnak egymással, hogy (18)-ban igen jó közelítéssel akár statisztikus véletlenszámokat is jelenthetnek az x_i,y_i és z_i mennyiségek.

Ami az $f_a(x)$ és $f_a(x)$ típusok hasonlóságát illeti, azzal kapcsolatban először [STEINER 1990] 247. oldali ábrájára utalunk. A sűrűséggörbék együttfutásának szoros mértéke azonban lehetne csak egyetlen szerencsésen választott példa is, a [STEINER (*ed.*) 1997] App. VIII. azonban a típustávolságok egzakt definíciójára támaszkodva kvantitatív módon jellemzi az $f_a(x)$ és $f_a(x)$ közeli rokonságát széles típustartományra vonatkozóan. Ugyanitt találhatók azok az α-értékek is, amelyeket a 3.1.1. végén felsorolt esetekben alkalmaznunk kell:

Jeffreys: α = 1,842; statisztikus: α = 1,677;
midway: α = 1,387; Cauchy: α = 1. (19)

Maguk az egymástól független x_i és z_i véletlenszámok generálása abban az esetben igényli a minimális gépidőt, ha analitikus alakban adottak az aktuális esetre az eloszlásfüggvények inverzei is, hiszen ekkor a gép által a (0,1) intervallumban azonos valószínűségsűrűséggel szolgáltatott, azaz egyenletes eloszlású, u-val jelölt számokból az éppen szolgáltatott u-értékkel egyszerűen $F^{-1}(u)$ lesz az F eloszlásfüggvényű típusnak megfelelő véletlenszám.

Monte-Carlo vizsgálatainkba nem vonjuk be a Jeffreys-eloszlást (ez nem eredményezne plusz következtetéseket), ezért két eloszlástípusra adjuk csak meg az inverz eloszlásfüggvényeket:

— midway-eloszlástípus, $a = 3$, inverz eloszlásfüggvénye

$$F_{mw}^{-1}(u) = \frac{u-0{,}5}{\sqrt{u(1-u)}}; \qquad (20)$$

— Cauchy-eloszlástípus, $a = 2$, inverz eloszlásfüggvénye

$$F_C^{-1}(u) = \mathrm{tg}\pi(u-0{,}5). \qquad (21)$$

(A fentiek érvényessége kontrollálható az $a = 3$-hoz és $a = 2$-höz tartozó eloszlásfüggvényekkel, l. [STEINER 1990] 50. oldal.)

Sajnos a leggyakrabban előforduló statisztikus eloszlástípusra (a = 5) nem célszerű megadni explicit alakban az inverz eloszlásfüggvényt, de [HAJAGOS 2005] matematikai meggondolásainak köszönhetően ekkor is könnyen számítható u-ból a statisztikus eloszlású x véletlenszám:

$$\varphi = 2\mathrm{arctg}\sqrt{u}\ ; \quad \kappa = 2\cos\left[\frac{\varphi+\pi}{3}\right]; \quad x = \frac{\kappa}{\sqrt{1-\kappa^2}}. \qquad (22)$$

A Monte-Carlo számításokkal már sűrűségfüggvénynek tekinthető hisztogramokat akarunk nyerni és a kvantilisértékek hibáját is minél kisebbre szeretnénk leszorítani. Egy-egy típushoz és r_t-hez ezért N = 100 000 db adatpárszázasból számított korrelációs együttható görbéje fogja alábbi ábráinkon Monte-Carlo vizsgálataink eredményeit bemutatni.

3.2. A Monte-Carlo vizsgálatok eredményei

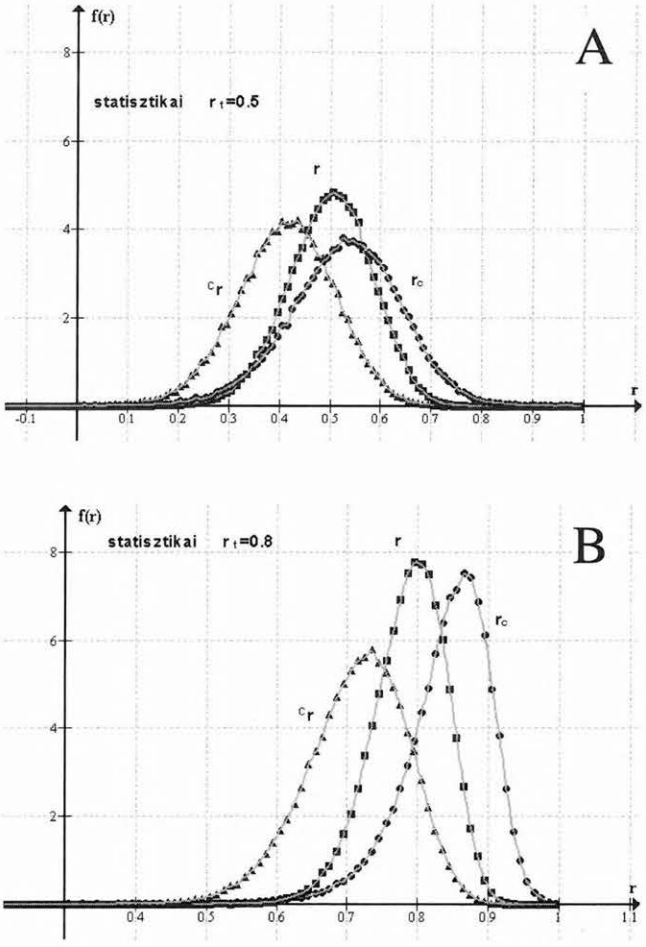
3.2.1. A korrelációs együtthatók görbéi statisztikus eloszlás esetén

Ne felejtsük el, hogy az r, ${}^{C}r$, ${}^{J}r$ (${}^{2}r$, ${}^{1}r$, ${}^{3}r$) formuláit (ld. a (8) képletet) a klasszikus (1) r_c definícióból pusztán analógiára támaszkodva, és nem egzakt levezetéssel nyertük, így a 3.1. módszerével dolgozva szükséges kontrollálni, hogy a rezisztifikált korrelációs együtthatók görbéi megfelelnek-e elvárásainknak. Áttekinthetetlenségre vezetne, ha a 3.1.1. végén felsorolt eloszlások és a rezisztifikált korrelációk mindegyikét (3-féle) vizsgálat tárgyává tennénk, különösen akkor, ha az r_t értékére is sok értéket vennénk fel. Éppen ezért a további vizsgálatainkban *az elméleti (helyes) korrelációs együtthatóra csak két értéket veszünk fel: az r_t = 0,5 és 0,8 értékeket, a Jeffreys-eloszlást és az ${}^{J}r$ (≡${}^{3}r$) korrelációs együtthatót pedig nem vonjuk be vizsgálatainkba.*

A 2. ábra alapján nem igényel külön magyarázatot, hogy (az N = 100 000 ismétlési számot alkalmazó) Monte-Carlo számításainkat legelőször a statisztikus eloszlásra végezzük el. Az eredményeket 0,01 hosszúságú részintervallumonként számláljuk meg, így ezek a hisztogrammok (N nagy volta miatt) már akár valószínűségsűrűség-függvényeknek is tekinthetők. Ugyanezt az eljárást követjük majd a 3.2.2. és 3.2.3. pontban is.

A *3a. ábra* az r, ${}^{C}r$ és a klasszikus r_c görbéket r_t = 0,5 esetére, a *3b. ábra* pedig az r_t = 0,8-ra mutatja be. Mivel a statisztikus eloszlásra vonatkozó leggyakoribb érték is, valamint r is k = 2-vel számítandó, nem lehet csodálkozni, hogy az aktuális r_t-knek az r-görbék felelnek meg a leginkább, de a megfelelés mértékét akár meglepőnek is minősíthetjük. — Még az a kérdés is felvethető ennek alapján, hogy *a leggyakoribb értékek eseteire analitikusan gyakran nehézkesen végezhető (vagy el sem végezhető) számítások helyett vajon más esetekben is nem alkalmazható-e az egyszerűbb klasszikus eredményekből kiinduló ugyanolyan (vagy hasonló) analogonképzés, mint ami az 1. pontban az (1) képletből a (8) formulára vezetett?*

Az ${}^{C}r$-görbék balra tolódását (0,08-dal, ill. 0,07-del) akár nagynak is minősíthetjük, bár az eltolódások kisebbek a görbék félérték-szélességeinél. Végül a hagyományos r_c-vel számolva a helyes r-görbétől jobbra tolódik a görbe, a bizonytalanságot jellemző félértékszélesség pedig a 3a. ábrán szignifikánsan nagyobb.



3. ábra. A k = 1-hez és 2-höz a (8) szerint tartozó ${}^{C}r$ és r sűrűségfüggvény-görbe, valamint az (1) szerint klasszikusan számított r_c-görbe statisztikus eloszlású eloszláspárok, valamint az elméleti korrelációs együttható r_t = 0,5 (A) és r_t = 0,8 (B) értékeire

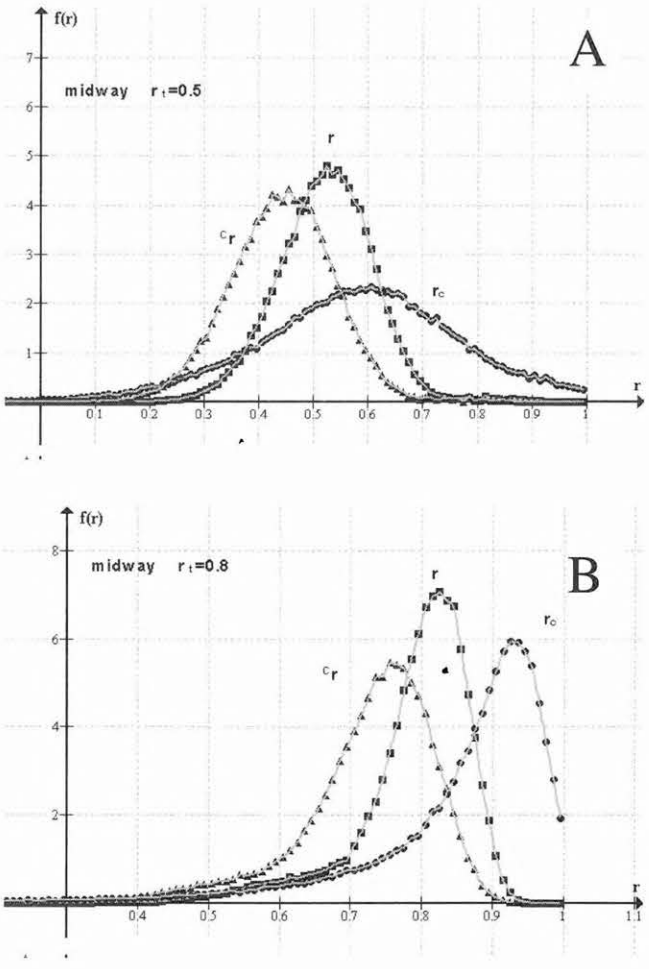
Fig. 3. The curves ${}^{C}r$, r (which correspond to k = 1 and 2 in Eq. 8) and the classical r_c-curve (see Eq. 1) for statistically distributed random number-pairs, if the theoretical correlation coefficient is r_t = 0,5 (A) and r_t = 0,8 (B), respectively

3.2.2. A korrelációs együtthatók görbéi midway-eloszlás esetén

Az $f_a(x)$ eloszláscsalád a = 3-hoz tartozó típusához nem tartozik erre az esetre levezett korrelációs együttható, így r_t = 0,5 és r_t = 0,8-hoz szorosan kötődő görbéket aligha várhatunk a midway-eredményeket bemutató *4a.* és *4b. ábrák*on. Indokolt viszont most a robusztusság kérdésének a felvetése az r, ${}^{C}r$ és r_c korrelációs együtthatók görbéit összehasonlítva.

A 4a. ábra szerint az r-görbe móduszának az eltérése a legkisebb az r_t -től: kb. 1/3-ot tesz ki a növekedés irányában. A másik irányban már nagyobb: 1/2 körüli az r_t = 0,5-től az ${}^{C}r$-görbe móduszának eltérése, de a legnagyobb abszolút értékű eltérést a klasszikus r_c-görbe mutatja a maga 0,6 körüli móduszával.

A robusztusságot illetően hasonló a sorrend az r_t = 0,8 esetén is, de nem árt kihangsúlyozni, hogy a klasszikus r_c-görbe módusza itt már 0,1-nél is nagyobb mértékben tér el r_t-től: kb. 1 1/3 ez a különbség.



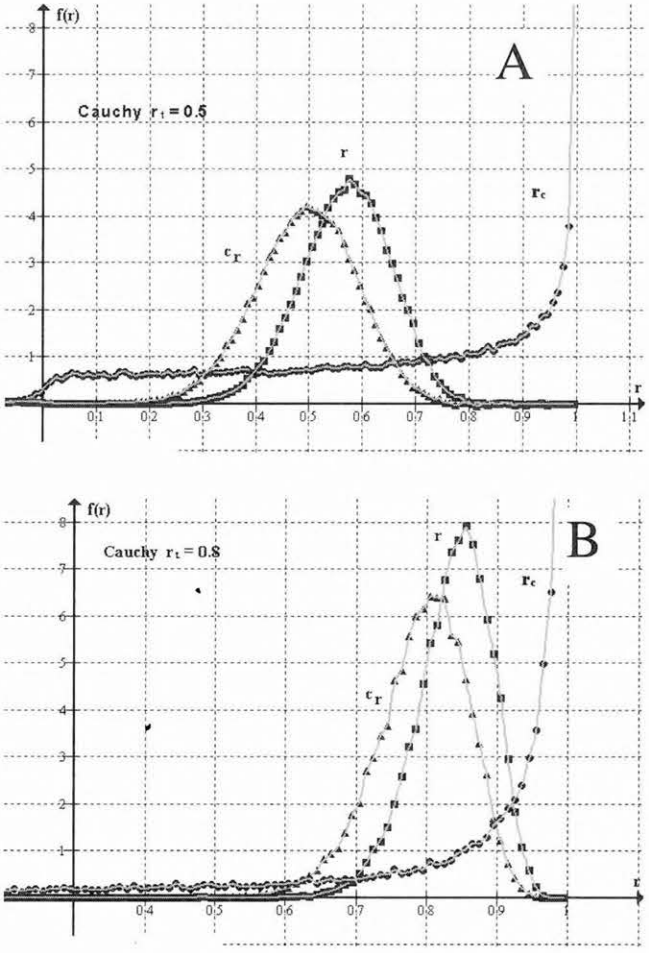
4. ábra. A k = 1-hez és 2-höz a (8) szerint tartozó ${}^{C}r$ és r sűrűségfüggvény-görbe, valamint az (1) szerint klasszikusan számított r_c-görbe midway eloszláspár, valamint az elméleti korrelációs együttható r_t = 0,5 (A) és r_t = 0,8 (B) értékeire

Fig. 4. The curves ${}^{C}r$, r (which correspond to k = 1 and 2 in Eq. 8) and the classical r_c-curve (see Eq. 1) for midway distributed random number-pairs, if the theoretical correlation coefficient is r_t = 0,5 (A) and r_t = 0,8 (B), respectively

3.2.3. A korrelációs együtthatók görbéi a Cauchy-eloszlása esetén

Az eredményül kapott sűrűségfüggvényeket az *5a.* és *5b. ábra* mutatja be. Mivel ${}^{C}r$ a (8) k = 1-gyel számított értékeit jelenti, esetünkben ezeket a görbéket várjuk a legszabályosabban elhelyezkedőknek. Valóban: aligha lehet jobban az r_t = 0,5-höz tartozóan jobban illeszkedő görbét elképzelni, mint amelyet az 5a. ábra ${}^{C}r$ valószínűségsűrűség-eloszlásként erre az esetre eredményezett. — Az r-görbéket illetően talán itt sem túlzott a robusztusság megkívánható mértékének a teljesüléséről beszélni, hiszen a móduszok eltérése r_t-től kisebb 0,1-nél, s ez sokkal kisebb érték, mint akár az r, akár az ${}^{C}r$ esetében a meghatározási bizonytalanságot jellemző félértékszélesség.

Nos, és mi a helyzet a hagyományos r_c-görbékkel? Az 5a. és 5b. ábra r_c-görbéi az aktuális r_t-től teljesen független, lapos, az $r_c \to 1$-nél azonban hirtelen igen nagymértékben növekedő sűrűségfüggvényt szolgáltatnak. *A hagyományos, r_c-vel jelölt ((1) szerint számolt) korrelációs együtthatók a Cauchy-típusú valószínűségi változók esetén már teljesen használhatatlanok.* Ennek fényében a (8) formula által szolgáltatott értékeket már teljes joggal nevezhetjük nemcsak rezisztenseknek, hanem robusztusoknak is.



5. ábra. A k = 1-hez és 2-höz a (8) szerint tartozó ${}^{C}r$ és r sűrűségfüggvény-görbe, valamint az (1) szerint klasszikusan számított r_c-görbe Cauchy eloszláspár, valamint az elméleti korrelációs együttható r_t = 0,5 (A) és r_t = 0,8 (B) értékeire

Fig. 5. The curves ${}^{C}r$, r (which correspond to k = 1 and 2 in Eq. 8) and the classical r_c-curve (see Eq. 1) for Cauchy distributed random number-pairs, if the theoretical correlation coefficient is r_t = 0,5 (A) and r_t = 0, (B), respectively

3.2.4. A számítási eredmények kvantilisei

Amennyiben az olvasó bizonyos további kvantitatív következtetésekre is kíváncsi, ebben az utolsó alfejezetben táblázatosan közöljük a p = 0,1 és p = 0,9 valószínűségekhez tartozó $q(0,9)$ kvantilisek értékeit, a jól ismert alsó és felső kvartilis, valamint szextilis értékeket (q_a, q_f, Q_a és Q_f), a mediánnal és módusszal együtt. A megadott 3 tizedesjegy közül az utolsó már *bizonytalannak* tekintendő.

Köszönetnyilvánítás

A jelen dolgozat a T 049852 számú OTKA kutatások részeként készült el.

			$q(0,1)$	Q_a	q_a	medián	q_f	Q_f	$q(0,9)$	modusz
Statisztikus eloszláspár	r_t=0,5	${}^{C}r$	0,282	0,314	0,344	0,409	0,473	0,500	0,528	0,423
		r	0,386	0,413	0,439	0,497	0,551	0,575	0,599	0,506
		r_c	0,383	0,421	0,454	0,528	0,599	0,629	0,661	0,542
	r_t=0,8	${}^{C}r$	0,612	0,639	0,663	0,714	0,759	0,777	0,796	0,733
		r	0,713	0,732	0,749	0,786	0,819	0,832	0,846	0,799
		r_c	0,755	0,781	0,803	0,845	0,879	0,846	0,905	0,868
Midway eloszláspár	r_t=0,5	${}^{C}r$	0,320	0,351	0,380	0,450	0,507	0,533	0,561	0,453
		r	0,410	0,439	0,466	0,524	0,580	0,603	0,627	0,531
		r_c	0,337	0,403	0,462	0,585	0,700	0,752	0,809	0,603
	r_t=0,8	${}^{C}r$	0,601	0,648	0,682	0,742	0,789	0,807	0,826	0,846
		r	0,688	0,733	0,761	0,808	0,844	0,857	0,870	0,822
		r_c	0,681	0,759	0,815	0,894	0,938	0,953	0,966	0,930
Cauchy eloszláspár	r_t=0,5	${}^{C}r$	0,368	0,430	0,454	0,496	0,537	0,559	0,613	0,500
		r	0,455	0,511	0,533	0,571	0,607	0,627	0,675	0,577
		r_c	0,151	0,381	0,504	0,710	0,868	0,925	0,988	-
	r_t=0,8	${}^{C}r$	0,707	0,753	0,771	0,799	0,825	0,839	0,871	0,810
		r	0,766	0,805	0,819	0,843	0,864	0,876	0,902	0,848
		r_c	0,537	0,857	0,947	0,970	0,990	0,993	0,998	-

HIVATKOZÁSOK

CRAMÉR H. 1945: Mathematical Methods of Statistics. Almquist & Wiksells, Uppsala, 575 p.

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

HAJAGOS B. 2005: Geostatisztikai eloszlású véletlenszámok gyors előállítása. Magyar Geofizika **46**, 2, 64-65. o.

KERÉKFI P. 1978: Robusztus becslések. Alkalmazott Matematikai Lapok 4, 327–357. o.

LANDY I., LANTOS M. 1991: A practical example for the Cauchy distribution. (App. 1: STEINER (*Ed.*): The Most Frequent Value, Akadémiai Kiadó, Budapest, 221–223. o.)

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

STEINER F. (*Ed.*) 1997: Optimum Methods in Statistics. Akadémiai Kiadó, Budapest, 370 p.

TARANTOLA A. 1987: Inverse Problem Theory. Elsevier, Amsterdam, 613 p.