Nemnormális, parametrizált eloszlású valószínűségi változók*

Kotosz Balázs

PhD, a Budapesti Corvinus Egyetem adjunktusa

E-mail: balazs.kotosz@uni-corvinus.hu

Ferenci Tamás

MSc, a Budapesti Corvinus Egyetem demonstrátora

E-mail: tamas.ferenci@medstat.hu

Szimulációs vizsgálatok során gyakran szükségessé válik adott jellemzőkkel rendelkező eloszlásból származó véletlen számok generálása. Amennyiben valamilyen jellegzetes, közismert eloszlásról van szó, a szükséges műveletek könnyen elvégezhetők, illetve a megfelelő programcsomagok ezeket tartalmazzák. Ha azonban bizonyos paraméternek tekintett tulajdonságokkal, például adott értékű momentumokkal rendelkező eloszlásokra van szükségünk, komoly akadályokba ütközhetünk. A szerzők tanulmányukban bemutatnak és megvizsgálnak néhány megoldási lehetőséget (Pearson-, Johnson-eloszláscsaládok, általánosított λ-eloszlás, Burr XII, Tukey-féle „g-and-h" és Fleishman transzformációs módszer), azok alkalmazhatósági korlátaival együtt, részletesen tárgyalva az illesztéssel kapcsolatos témákat is.

TÁRGYSZÓ:
Statisztikai módszertan.
Valószínűség-eloszlás.
Momentumok.

* Itt szeretnénk köszönetet mondani a lektornak értékes észrevételeiért. Természetesen a tanulmányban előforduló esetleges hibákért kizárólag a szerzőket terheli felelősség.

Eloszlásillesztés alatt első közelítésben azt a statisztikai feladatot értjük, melynek során valamilyen empirikus adatsorhoz (mintához) olyan elméleti eloszlást keresünk, hogy az empirikus adatsor eloszlása és az elméleti eloszlás a leghasonlóbb legyen (a hasonlóság valamilyen mértéke szerint). E dolgozatban kizárólag az egyváltozós statisztika területén fogunk mozogni.

Ahhoz, hogy a feladatot végre tudjuk hajtani, két részfeladat megoldására van szükség: először a mintánkból az eloszlására vonatkozó információkat szükséges kinyerni, majd ezeket kell felhasználni az elméleti eloszlás meghatározásakor.

Ez utóbbi – figyelembe véve, hogy a gyakorlatban paraméterek által befolyásolt eloszlásokkal találkozunk – ismét csak két részfeladatot jelent: a felhasznált eloszlás megválasztását, majd, miután ezt rögzítettünk, az optimális paraméterezés megállapítását. Az első feladatot, az eloszlás mellett történő elköteleződést számos tényező befolyásolhatja (a modellező implicit ismeretei a szóba jövő eloszlások szakmai tartalmáról, előzetes várakozások stb.), ezért kevésbé algoritmizálható. (Valamilyen illeszkedésvizsgáló próbát használva azonban arra is mód van természetesen, hogy a modellező több, önmagában optimálisan paraméterezett eloszlásnak az empirikus adatokkal vett illeszkedése alapján válasszon.)

Az eloszlás kiválasztása után következő feladat az optimális paraméterek meghatározása. Mivel itt a mintából kell következtetni a sokasági paraméterre, jól láthatóan egy becslési feladatot kaptunk, amelyre számos közismert eljárás, például a népszerű maximum likelihood-elv (ML) használható.

Ha azonban történetileg visszatekintünk erre a kérdésre, azt látjuk, hogy a XX. század elején – bár az eloszlásillesztések ekkor szinte fénykorukat élték – még nem volt, legalábbis mai formájában, széles körű használatban az ML-elv. (Noha *Pearson* már a századforduló környékén megsejtette, és bizonyos értelemben használta is e módszert.) A kor statisztikusai tehát más elvre támaszkodtak, az egyik legnépszerűbb a momentumok alapján történő illesztés, az ún. momentumok módszere (MM) volt.

Ennek során meghatározták az empirikus adatsor első néhány (tipikusan négy) momentumát, majd azt tekintették optimális paraméterkombinációnak, mely ugyanilyen momentumokkal rendelkező elméleti eloszlást adott. Amíg az ML-elv minden mintaelemet közvetlenül felhasznál, addig a momentumok alapján történő illesztés 4 számra redukálja az adatbázist – ami nyilvánvalóan rontja az illeszkedést. Hatalmas előnye viszont, hogy a legtöbb eloszlás esetén az elméleti eloszlás paramétereinek függvényében felírt, és az empirikus adatokkal egyenlővé tett momentumok, mint egyenletek alkotta egyenletrendszer analitikusan megoldható volt.

Későbbiekben, az elméleti és számítástechnikai fejlődésnek köszönhetően a momentumok módszerének ilyen alkalmazása kikerült a napi gyakorlatból. Érdekes viszont, hogy az utóbbi időben, egészen más indíttatásokból, ismét előtérbe kerültek e módszerek. Ezen okok egyike[1] a Monte-Carlo-szimulációs módszerek széles körű elterjedése, melyekkel végzett bizonyos vizsgálatoknál szükségessé válik adott értékű momentumokkal rendelkező eloszlásokból származó véletlenszámok generálása. Mivel jelen dolgozatunkat egy ilyen alkalmazás inspirálta, e kérdést röviden bemutatjuk. (Részletesebben lásd például *Ferenci* [2009]-et.)

Tegyük fel, hogy egy statisztikai próba valamilyen eloszlási (tipikusan normalitási) feltevéssel él a sokaságokra vonatkozóan, melyből a mintái származnak (mint a közismert Student-féle *t*-próba), és vizsgálni kívánjuk, hogy a próba mennyire robusztus e feltevés megsértésére nézve. Ennek egyike lehetősége a Monte-Carlo-módszer, melynek során a feltevést irányítottan megsértő (adott mértékben nemnormális) sokaságból származó véletlenszámok tömegét generáljuk, és – ezeken végrehajtva a vizsgált tesztet – megfigyeljük, hogy az empirikus elsőfajú hibaarány konvergál-e a szignifikanciaszinthez. Ehhez szükséges, hogy képesek legyünk adott mértékben nemnormális sokaságból származó véletlenszámok generálására; ez tipikusan adott (nemnormális) ferdeséget/csúcsosságot jelent. Fontos megjegyezni, hogy e feladat jó minőségű megoldása azt is igényli, hogy olyan eloszlást válasszunk, melyből a lehető legtöbb ferdeség/csúcsosság értékhez generálható véletlenszám, tehát a lehető legszélesebben, legtöbb ferdeség/csúcsosság eléréséhez paraméterezhető (hiszen a megoldás során majd a ferdeség/csúcsosság síkon akarunk végigiterálni).

Vegyük észre, hogy ez a probléma eltér a momentumok módszerének alapfeladatától, hiszen itt a momentumok nem egy empirikus adatsorból számolhatók, hanem előre, a modellező által meghatározottak.[2] Ez a tény (tehát hogy az említett két feladatrész közül csak a másodikat szükséges megoldani: az elméleti eloszlást úgy megválasztani és paraméterezni, hogy momentumai meghatározott értékek legyenek) egy lényeges módosulást mégis jelent: a feladat innentől nem statisztikai értelemben vett becslés (így például a becsléselméleti tulajdonságait sem lehet vizsgálni, szemben a szó hagyományos értelmében vett momentumok módszerével, ahol ez központi kérdés). Mi e különbségtétel hangsúlyozása végett használjuk az „eloszlásillesztés” kifejezést.

Ennek vizsgálata arra motivál minket, hogy fellapozzuk a momentum módszer és az eloszlásillesztés klasszikus irodalmát, és újra áttekintsük a korábban még egészen más okból vizsgált feladatot.

[1] Egy másik fontos és aktuális téma, melyre itt csak utalni tudunk, a momentum módszer egy általánosítása, a GMM. Erről lásd például *Hall* [2005]-öt.

[2] Egy másik terület, ahol – teljesen más indíttatásból – de épp ugyanerre szükség lehet, és melyet ismét csak utalás szintjén tudunk megemlíteni, a bayes-i statisztika (*Lee* [2009]). Itt ugyanis a priorok létrehozásához használt külső információ gyakran épp momentumok (vagy épp kvantilisek) formájában áll rendelkezésre.

Annál is inkább szükség van erre, mert a legtöbb jól ismert, alapozó statisztikai kurzuson is oktatott eloszlás (például normális, t, χ^2, F, exponenciális, lognormális) nem alkalmas 4 momentum alapján történő illesztésre (sem); a szóba jövő eloszlások pedig még egyetemi szinten is újszerűek lehetnek. Ezek áttekintését kíséreljük meg most.

Ezt az alapfeladatot kiegészítjük azon kérdés vizsgálatával, hogy hogyan lehetséges egy eloszlást (momentumai helyett) kvantiliseivel illeszteni. (Bár ennek megoldására csak egy, az eloszlásoknak még az előzőnél is szűkebb köre képes.) A momentumok kapcsán eddig elmondott legtöbb megjegyzés változatlanul érvényes kvantilisek alapján történő illesztésre is.

A dolgozat első részében az eloszlásillesztés két módszerét, a momentumokon és a kvantiliseken alapuló illesztést tekintjük át, különös tekintettel a fogalmak és a jelölések egységes definiálására. A második részben a céloknak megfelelő eloszlásokat, illetve eloszláscsaládokat mutatjuk be, így rendre a Pearson-, a Burr-, a Johnson-, az általánosított λ, a g-and-h, végül a Fleishman-eloszlást. Egyes levezetések és eredmények – bonyolultságuk, hosszuk miatt – az internetes Mellékletben kaptak helyet (www.ksh.hu/statszemle).

1. Az eloszlásillesztés két módszere

Ebben a részben definiáljuk pontosan, hogy mit értünk a momentumok, illetve kvantilisek alapján történő illesztésen. A dolgozat egészében alkalmazott eloszlásfüggetlen jelölésrendszert is itt vezetjük be. Ez már csak azért is fontos, mert a forrásmunkák majdnem egy évszázadot fognak át, amely idő alatt igen jelentősen változtak bizonyos statisztikai jelölésekkel kapcsolatos szokások; így most egyúttal arra is kísérletet teszünk, hogy ezeket egységes keretben mutassuk be.

1.1. Momentumok alapján történő illesztés

Egy valószínűségi változó[3] n-edik nyers momentumának a

$$\mu'_n = \int_{-\infty}^{+\infty} x^n \cdot f(x)\,\mathrm{d}x \qquad n = 0,1,...$$

integrált nevezzük, ha az konvergens (*Kendall–Stuart* [1977]). Egy változó nulladik nyers momentuma szükségképp 0, az első nyers momentuma a várható értéke.

[3] A továbbiakban sokszor – némileg hanyagul – erre úgy is fogunk hivatkozni, mint egy „eloszlás momentumra", tudva természetesen, hogy itt precízen egy valószínűségi változó momentumáról van szó.

Az *n*-edik centrális momentumának a

$$\mu_n = \int_{-\infty}^{+\infty} \left(x - \mu_1'\right)^n \cdot f\left(x\right) \mathrm{d}x \qquad\qquad n = 0,1,...$$

integrált nevezzük, ha az konvergens. (Könnyen belátható, hogy ha egy valószínűségi változónak létezik *n*-edik nyers momentuma, akkor létezik *n*-edik centrális momentuma is.) Mint látható, a centrális momentumot a várható érték körül értelmeztük. Ez nem szükségszerű, de mi a mostani tárgyalásunkban ezt fogadjuk el definíciónak. A nulladik centrális momentuma értelemszerűen minden eloszlásnak 1, az első centrális momentum értéke 0, míg a második a szórásnégyzet.

Végül bevezethető a standardizált centrális momentum fogalma is. Mivel ennek a három, és annál magasabb momentumok esetén van igazán értelme, így a jelölés indexe is a harmadik ilyen momentumnál vesz fel 1 értéket:

$$\gamma_{n-2} = \frac{\mu_n}{\left(\mu_2^{1/2}\right)^n}.$$

Így γ_1 a ferdeség, γ_2 pedig a csúcsosság mutatója lesz.[4] Példának okáért, a normális eloszlás ferdesége e mutatókkal $\gamma_1 = 0$, csúcsossága[5] $\gamma_2 = 3$.

Mindezek alapján – a Cauchy–Bunyakovszkij–Schwarz-egyenlőtlenség felhasználásával – belátható, hogy szükségképp minden eloszlásra teljesül a

$$\gamma_2 \geq \gamma_1^2 + 1$$

összefüggés, mely a ferdeség függvényében határoz meg egy minimális csúcsosságot. (Kissé leegyszerűsítve azt mondja ki, hogy nem léteznek nagyon ferde, és mégis lapult eloszlások.) Ebből következik, hogy a ferdeség/csúcsság síkon létezik egy – parabolikus görbe által kijelölt – „lehetetlen tartomány", ahol nem létezhet eloszlás.

A ferdeség és csúcsosság specifikált értékeit rendre g_1-gyel és g_2-vel fogjuk jelölni.

[4] A mutatók jelölésének tekintetében az irodalom megosztott. Az α, a β, a γ és a δ különböző sorszámai egyaránt felbukkannak különböző írásokban, sokszor hasonló, vagy éppen négyzetes tartalommal. Az általunk választott jelölésekben a mutatókat *Pearson* nyomán – bár vele nem teljesen azonosan – definiáltuk. Ezzel kapcsolatban megjegyezzük, hogy a ferdeség/csúcsosság síkot sokszor $\gamma_1^2 - \gamma_2$ tengelyeken ábrázolták, ráadásul a függőleges tengelyt fejjel lefelé fordítva. Mi konzisztensen a szokott állású $\gamma_1 - \gamma_2$ síkot fogjuk használni.

[5] Pontosan ez utóbbi az oka annak, hogy több helyen az általunk definiált mutató 3-mal csökkentett értékét nevezik csúcsosságnak („excess kurtosis", „többlet csúcsosság"), hiszen így a normális eloszlásra mindkét mutató 0 értékű lesz. Mivel a mi álláspontunk szerint e megoldás ad hoc, valószínűségelméletileg nem tiszta, nem követjük ezt a szisztémát, és az említett módon definiált mutatót fogjuk dolgozatunkban használni.

Ezek szerint a momentumok alapján történő illesztés feladata a következőként határozható meg. Adott egy $f\left(\underline{\Theta}\right)$ eloszlás, ahol a $\underline{\Theta}$ paramétervektor az eloszlás jellemzőit határozza meg. Keressük azt a $\underline{\Theta}^*$ paramétervektort, melyre teljesül, hogy

$$\mu_n\left(\underline{\Theta}^*\right) = m_n$$

valamely $m_n, n = 1,2,...H$ számsorozatra. Itt tipikusan $H = 4$.

1.2. Kvantilisek alapján történő illesztés

A feladat igen hasonló az előbbihez, egyedül a p-ed rendű ($p \in (0,1)$) kvantilis

$$\rho_p : \int_{-\infty}^{\rho_p} f(x)\mathrm{d}x = p$$

egyenlettel definiált fogalmát kell bevezetnünk.

Ekkor a kvantilisek alapján történő illesztés feladata így fogalmazható meg. Adott egy $f\left(\underline{\Theta}\right)$ eloszlás, ahol a $\underline{\Theta}$ paramétervektor az eloszlás jellemzőit határozza meg. Keressük azt a $\underline{\Theta}^*$ paramétervektort, melyre teljesül, hogy

$$\rho_{q_n}\left(\underline{\Theta}^*\right) = \rho_n$$

valamely $\{\rho_n, q_n\}$ (n=1,2,...H) párokból álló sorozatra.

2. Eloszláscsaládok

A következőkben bemutatjuk a legfontosabb olyan eloszláscsaládokat, melyek gyakorlati problémák esetén lehetővé teszik a momentumok és/vagy kvantilisek alapján történő illesztést. (Természetesen mindenhol megadjuk ennek korlátait is.) A leírások során bemutatjuk az eloszlásokat, és külön kitérünk az illesztés elvégzésének statisztikai hátterére.

2.1. Pearson-eloszláscsalád

A modern statisztika egyik legnagyobb alakja, *Karl Pearson* a XX. század első évtizedeiben vezette be azt az eloszláscsaládot, mely mai napig az ő nevét viseli. Az

ezzel kapcsolatos ismereteket cikkek egész sorában közölte (*Pearson* [1893], [1895], [1901], [1916]), melyek közül az első 1893-ben, az utolsó 1916-ban jelent meg. A sokszor a saját korát is megelőző közlés 12, római számmal azonosított eloszlás meglehetősen kusza rendszerét eredményezte, melyek számát Pearson folyamatosan növelte a cikkek során, de időközben bizonyos eloszlásokat át is definiált, míg másokról megállapította, hogy az előzők speciális esetei.

Tovább bonyolítja a helyzetet, hogy ezen eloszlások egy része – ahogy a valószínűségelmélet fejlődésével feltárultak az összefüggések – más nevet kapott a későbbiekben. Így fordulhat elő, hogy a Pearson-rendszerben vannak olyan eloszlások, amelyek – bár rájuk csak egy rejtélyes római szám utal – valójában teljesen triviálisak, míg más eloszlásoknak oly kevés a gyakorlati jelentősége, hogy azóta szinte feledésbe mentek.

Pearson eredeti célja az volt, hogy – biostatisztikai indíttatásból – olyan eloszlásrendszert alkosson, mely lehetővé teszi az illesztést a legkülönfélébb ferdeségű és csúcsosságú empirikus adatokra; még pontosabban, hogy olyan eloszlásrendszert adjon meg, mellyel minden esetben elvégezhető az illesztés az első négy momentum alapján – épp, amire nekünk is szükségünk van a korábban már vázolt okokból.

Pearson alapötlete az volt, hogy az eloszlásokat a sűrűségfüggvényükkel adta meg, de nem közvetlenül ($f(x)$ alakban), hanem egy rá vonatkozó differenciálegyenlettel ($\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ alakban):

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{x}{b_0 + b_1 x + b_2 x^2} \cdot f. \qquad /1/$$

Bár első ránézésre igen szokatlan megadása ez egy sűrűségfüggvénynek, bizonyos tulajdonságok mégis kényelmesen leolvashatók. Egyrészt, ha egy ennek megfelelő eloszlás az egész $x \in \mathsf{R}$ számegyenesen értelmezett, akkor egy és csakis egy helyen, az $x = 0$ pontban vesz fel szélsőértéket. (Mivel itt eleve adott a sűrűségfüggvény deriváltja, ez közvetlenül leolvasható.) Az is könnyen belátható (lásd a Mellékletben), hogy ez a szélsőérték maximum lesz, tehát levonhatjuk azt a következtetést, hogy az ilyen (egész számegyenesen értelmezett) Pearson-eloszlások unimodálisak, módusszal a 0 pontban. Természetesen ez nem szükségképp teljesül azokra az eloszlásokra, melyek nem értelmezettek a teljes valós számegyenesen: ezeknél szélsőérték (módusz) lehet továbbá az értelmezési határokban is. Ezen felül az is észrevehető, hogy az $x \to \pm\infty$ határátmenetben a $\frac{\mathrm{d}f}{\mathrm{d}x}$ szintén nullába tart, tehát az eloszlás mindkét végén elenyészik. (Ezek teljesülése Pearsont is motiválta az eloszláscsalád kialakításakor.)

Az /1/ egyenlet átrendezésével és kiintegrálásával megkaphatjuk a Pearson-rendszer nyers momentumaira érvényes következő összefüggést:

$$nb_0\mu'_{n-1} + (n+1)b_1\mu'_n + \left[(n+2)b_2 + 1\right]\mu'_{n+1} = 0.$$

Ez jól láthatóan egy rekurzív összefüggés, mely lehetővé teszi, hogy $\mu'_0 (=1)$ és μ'_1 (tehát valójában csak μ'_1) ismeretében meghatározzuk az összes momentumot. (Ismerve természetesen az eloszlást leíró 3 paramétert, b_0 -t, b_1 -t és b_2 -t.) Belátható, hogy e momentumok közül az első négy egyértelműen meghatározza az eloszlás 3 paraméterét, így annak sincs akadálya, hogy ezeket a (centrális) momentumok függvényében írjuk fel (lásd a Mellékletet).

Mivel a fenti együtthatók még semmit nem mondanak magukról az eloszlásokról, így a problémát tovább kell vizsgálnunk: meg kell oldalunk a bemutatott differenciálegyenletet. Ennek menete terjedelmi okokból a Mellékletben található, mi most a végeredményre koncentrálunk.

2.1.1. A három alapvető Pearson-eloszlás

Bár (amint azt a 2.1. bevezetésében is említettük) Pearson 12 eloszlást definiált, ezek közül csak 3 van, ami nemnulla területet fed le a ferdeség/csúcsosság síkból, a többi – ebből is következően – átmeneti, illetve elfajuló eset. Mi a következőkben – összhangban eredeti céljainkkal – erre a 3 eloszlásra, a Pearson I, IV és VI eloszlásokra fogunk koncentrálni. Ezek az eloszlások a bemutatott általános esetből származtathatók, bizonyos sajátosságokat ($b_0 + b_1 x + b_2 x^2$ -nak van-e valós gyöke, illetve ha igen, akkor azok hogy helyezkednek el) is figyelembe vevő paraméterezéssel. (Ennek oka a Mellékletben közölt differenciálegyenlet-megoldásból válik világossá.)

Pearson IV. Amennyiben a $b_0 + b_1 x + b_2 x^2$ -nek nincs valós gyöke, a következő alakot érdemes (lásd a Mellékletet) használni:

$$f(x) = k\left(1 + \frac{x^2}{\alpha^2}\right)^m \cdot \exp\left[\nu \cdot \operatorname{arctg}\left(\frac{x}{\alpha}\right)\right]. \qquad /2/$$

Mivel a /2/ eloszlásfüggvény minden $x \in \mathsf{R}$ esetében értelmezett és valós, így az eloszlás tartója a teljes számegyenes. Összevetve ezt az áttekintésben mondottakkal, egyből adódik, hogy az eloszlás unimodális és harang[6] alakú.

[6]Az eloszlások alakja kapcsán a harang közismert jelentésű, az U-alakkal arra utalunk, hogy a sűrűségfüggvény egy lokális maximum szélsőértéktől indulva csökken, majd a globális minimum után növekszik, és egy lokális maximum az értelmezési tartomány felső határa. (A két lokális maximum közül bármelyik lehet globális maximum.) Az L- és a J-alakú eloszlások egymás tükörképei, így az L-alakú eloszlásoknál a módusz az értelmezési tartomány alsó, míg a J-alakúaknál a felső határa.

Az integrációs konstansból adódó – és eddig kötetlen – k értéke azon peremfeltétellel határozható meg, hogy a sűrűségfüggvény integrálja a teljes számegyenesen egységnyi. A részletek mellőzésével (lásd például *Heinrich* [2004]-et) ennek értéke:

$$k = \frac{\Gamma(m)}{\sqrt{\pi}\alpha\Gamma(m-1/2)}\left|\frac{\Gamma(m+i\nu/2)}{\Gamma(m)}\right|^2 = \frac{\left|\frac{\Gamma(m+i\nu/2)}{\Gamma(m)}\right|^2}{\alpha B(m-1/2,1/2)}.$$

Ebből a felírásból az utolsó szükséges információ is kiolvasható: az eloszlás akkor normálható, azaz akkor létezik, ha $m > 1/2$.

Pearson I és VI. Ha a $b_0 + b_1 x + b_2 x^2$-nek van valós gyöke, akkor (lásd a Mellékletet):

$$f(x) = k \cdot (x - a_1)^{\frac{\sqrt{b_1^2 - 4b_0b_2} + b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}} \cdot (a_2 - x)^{\frac{\sqrt{b_1^2 - 4b_0b_2} - b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}}.$$

A *Pearson I eloszlás* esetén az így meghatározott sűrűségfüggvény az $x \in (a_1, a_2)$ tartományon vesz fel valós értéket, csak ott értelmezett. Ez az eloszlás tehát mindkét irányból korlátos tartón, egy véges intervallumon értelmezett csak. Az eloszlás harang-, U- és J-alakú is lehet.

Vezessük be az $m_1 = \frac{\sqrt{b_1^2 - 4b_0b_2} + b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}$ és az $m_2 = \frac{\sqrt{b_1^2 - 4b_0b_2} - b_1}{2b_2\sqrt{b_1^2 - 4b_0b_2}}$ jelöléseket, továbbá transzformáljuk a számegyenest úgy, hogy az origót a_1-be toljuk, az egységet pedig $(a_2 - a_1)$-nek választjuk. Ekkor a sűrűségfüggvény így írható:

$$f(x) = kx^{m_1}(1-x)^{m_2}.$$

Ekkor a normalizációs konstans beláthatóan $1/B(m_1+1, m_2+1)$ lesz, így (az immár nyilvánvalóan a $(0;1)$ intervallumon értelmezett) sűrűségfüggvény:

$$f(x) = \frac{1}{B(m_1+1, m_2+1)} x^{m_1}(1-x)^{m_2}.$$

Végül pedig, ebből az eloszlás létezésének feltétele is leolvasható: $m_1, m_2 > -1$.

Pearson VI eloszlásnál az előző eset előjeleinek megfordításával, alkalmas transzformációval a következő sűrűségfüggvényhez jutunk:

$$f(x) = \frac{1}{B(m_1+1, m_2+1)} x^{m_1}(1+x)^{-m_1-m_2-2}. \qquad /3/$$

Ez azért új lényegileg, mert nem egy tartományon (két gyök között), hanem azon kívül értelmezett; a /3/ speciális esetben például az $x \in (0, \infty)$-n. Ezen eloszlás tartója tehát mindig egy félegyenes. Alakja harang vagy J.

A /3/ sűrűségfüggvényből közvetlenül látható, hogy a létezés feltétele, hogy $m_2 > -1, m_1 + m_2 < -1$.

2.1.2. Az alapvető Pearson-eloszlások illesztése momentumok alapján

A következőkben megadjuk az eloszlások illesztéséhez szükséges összefüggéseket.

Pearson IV. Vezessük be az $r = 2m - 2$ jelölést. Ezzel a nyers momentumok számítási módszere:

$$\begin{aligned}
\mu_1' &= -\frac{a\nu}{r}, \\
\mu_2' &= \frac{a^2}{r(r-1)}\left(r + \nu^2\right), \\
\mu_n' &= \frac{a}{r-n+1}\left[(n-1)a\mu_{n-2}' - \nu\mu_{n-1}'\right].
\end{aligned}$$

Ezekből a standardizált momentumok és a ferdeség/csúcsosság ($\gamma_1 - \gamma_2$) mutatói számíthatók. Ha ez utóbbit megtesszük, és az eredményeket egyenlővé tesszük a specifikált g_1 és g_2 értékekkel, majd a kapott egyenletrendszert megoldjuk, akkor a következőket kapjuk:

$$r = 2(m-1) = \frac{6\left(g_2 - g_1^2 - 1\right)}{2g_2 - 3g_1^2 - 6},$$

$$\nu = \frac{r(r-2)g_1}{\sqrt{16(r-1) - g_1(r-2)^2}},$$

$$a = \frac{\sqrt{\mu_2\left[16(r-1) - g_1(r-2)^2\right]}}{4}.$$

Pearson I. Pearson itt is megadta a nyers és standardizált momentumokat, ízelítőül az első két nyers momentum (ehhez legyen $b = a_1 + a_2$):

$$\mu_1' = \frac{b(m_1+1)}{m_1 + m_2 + 2}, \quad \mu_2' = \frac{b^2(m_1+2)(m_1+1)}{(m_1 + m_2 + 3)(m_1 + m_2 + 2)}.$$

(A további tagok számítására rendelkezésre áll egy (igaz meglehetősen összetett) rekurzív formula.)

A paraméterek számítása specifikált momentumok alapján:

$$r = 2\frac{6\left(g_2 - g_1^2 - 1\right)}{3g_1^2 - 2g_2 + 6}, \quad \varepsilon = \frac{r^2}{4 + \frac{1}{4}g_1^2\left(r+1\right)^2 \big/ \left(r+1\right)}.$$

Ezek meghatározása után a két ismeretlen paraméter az

$$\left(m+1\right)^2 - r\left(m+1\right) + \varepsilon = 0$$

másodfokú egyenlet két gyökeként kapható meg.

Pearson VI. A nyers momentumok számítására meglehetősen bonyolult (de explicit alakú) formula áll rendelkezésre; alakjukat tekintve a Pearson I-gyel lesznek analógak. Ebből is következően, a számítás menete egyezik Pearson I-gyel.

2.1.3. A Pearson-eloszláscsalád lefedési tartománya

Az eloszlásokat bemutató részben minden esetben megadtuk az eloszlás létezésének feltételét. (Tipikusan származtatott paraméterek alapján, ám végeredményben az eredeti differenciálegyenlet paramétereit használva.) Nincsen akadálya tehát annak, hogy ezeket a feltételeket átírjuk a minta tulajdonságaira; ez minden esetben megtehető lesz pusztán a ferdeség és csúcsosság felhasználásával.

Pearson IV és VI létezésének a feltétele $2g_2 - 3g_1^2 - 6 > 0 \Rightarrow g_2 > \frac{3}{2}g_1^2 + 3$, míg a Pearson I-é $6 + 3g_1^2 - 2g_2 > 0 \Rightarrow g_2 < \frac{3}{2}g_1^2 + 3$.

A Pearson IV-et és VI-ot a $g_2 = \frac{3\left(2\sqrt{g_1^6 + 12g_1^4 + 48g_1^2 + 64} + 13g_1^2 + 16\right)}{32 - g_1^2}$ egyenletű görbe különíti el egymástól.

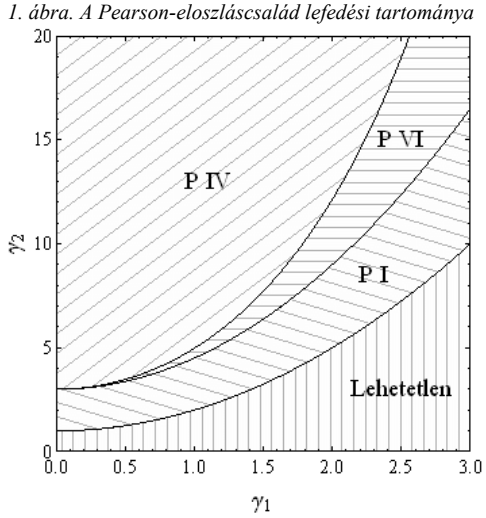
Ezen görbe alatt a Pearson VI, fölötte (ad infinitum) a Pearson IV eloszlás található. (Lásd az 1. ábrát.)

A Pearson-eloszláscsalád legfontosabb, minden más jó tulajdonságánál fontosabb előnye, hogy minden ferdeség/csúcsosság pontra illeszthető; lefedi az egész ferdeség/csúcsosság síkot. Még a most bemutatott igen általános eloszlások közül is csak kevés bír ehhez fogható lefedéssel.

Hátránya viszont, hogy az eloszlás, illetve kvantilisfüggvénye nem adható meg explicit alakban, így nincs egyszerű, általános módszer Pearson-eloszlásból származó

véletlenszámok generálására. Nem is beszélve arról, hogy ha iterálunk a ferdeség/csúcsosság síkon, akkor még az algebrai formák között is váltogatnunk kell, ami szintén impraktikus számítástechnikai szempontból.

1. ábra. A Pearson-eloszláscsalád lefedési tartománya



2.2. Burr-eloszláscsalád

Irwing W. Burr amerikai statisztikus 1942-ben publikált cikke (*Burr* [1942]) tekinthető az első közlésnek a témában. Burr 12 eloszlást adott meg írásában (mindegyiket eloszlásfüggvényével), melyeket gyakorlati szempontból fontosnak nevezett. Ezen eloszlások közül azonban egyetlen, a XII-es kapott nagy figyelmet a későbbiekben.

Már maga Burr is kiemelte ezt az eloszlást az idézett cikkében, és példaként tárgyalta az empirikus adatokhoz való illesztésének módszerét. Helyesen mutatott ugyanis rá, hogy az eloszlás paraméterei révén változatos γ_1 ferdeség és γ_2 csúcsosság mutatókat tud felvenni.

Hatke [1949] már azt is vizsgálta, hogy milyen ferdeség/csúcsosság értékekre végezhető el az illesztés, ám a ma szokásos $\gamma_1 - \gamma_2$ sík helyett a $\gamma_1^2 - \delta$ síkot használta a lefedettség megadásához, amely δ mérték ma már nincs[7] használatban, ráadásul később sikerült igazolni, hogy adatai részben tévesek: az eloszlás nagyobb területen illeszthető, mint a cikk megadta.

[7] $\delta = \dfrac{2\gamma_2 - 3\gamma_1^2 - 6}{\gamma_2 + 3}$ definíció szerint; használatát *Craig* [1936] javasolta, alapvetően azért, mert bevezetésével a Pearson-eloszlások sokkal egyszerűbb formát öltenek bizonyos számításokban.

Hosszú szünet után, 1968-ban Burr új cikkekkel jelentkezett (*Burr–Cislak* [1968]), melyben elsődlegesen a Burr-eloszlású sokaságokból vett minták becsléselméleti tulajdonságaival foglalkozott, ám emellett rámutatott *Hatke* [1949] előbb említett hibájára, és frissítette a lefedettséget mutató ábrát. Nem sokkal később (*Burr* [1973]) rövid közleményben bemutatta azokat az immár elektronikus számítógéppel, nagy pontossággal számított táblázatait, melyekkel finoman lehet illeszteni az eloszlásokat. Az eredményeket még mindig a $\gamma_1^2 - \delta$ síkon adta meg grafikusan, és továbbra sem foglalkozott a határok analitikus felírásának kérdésével.

Ebből a szempontból *Rodriguez* [1977] jelentett óriási előrelépést. A szerző egyrészt a ma szokásos $\gamma_1 - \gamma_2$ síkon adta meg az eloszlás lefedését (megmutatva, hogy számos, gyakorlatban fontos eloszlás illeszthető a Burr XII-vel), másrészt a lefedettséget illetően nem numerikus számításokon alapuló, analitikus eredményeket is elért.

Az utolsó fontos elméleti fejlemény *Tadikamalla* [1980] cikke, melyben tisztázta a kapcsolatot a Burr XII és pár egyéb fontos eloszlás között, egyúttal felhívta a figyelmet a Burr III eloszlás XII-höz hasonló kedvező illesztési tulajdonságaira.

2.2.1. A Burr XII eloszlás származtatása és definíciója

Burr eredeti cikkében (*Burr* [1949]) azt tűzte ki feladatul, hogy a gyakorlatban előforduló adatokhoz történő illesztésre alkalmas eloszlásokat adjon meg eloszlásfüggvénnyel[8]. A korszak legnépszerűbb, empirikus adatok illesztésére szolgáló rendszere, a Pearson-eloszláscsalád nem felel meg ennek a szempontnak, hiszen az eloszlások sűrűségfüggvényét ragadja meg (ahogy azt mi is tárgyaltuk a 2.1. pontban) egy, a sűrűségfüggvényre felírt differenciálegyenlet segítségével.

Burr úgy látott neki a feladatnak, hogy megalkotta Pearson differenciálegyenletének analógiáját eloszlásfüggvényre felírva:

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = F(x)\left[1 - F(x)\right]g(x).$$

Az analógia nyilvánvaló, ha a $g(x) = \frac{1}{a + bx + cx^2}$-et tekintjük, és figyelembe vesszük, hogy a nevezőben itt csak $F(x)$ szerepelhet ($x \cdot F(x)$ nem), hogy az minden $x \in \mathsf{R}$-re nemnegatív legyen. (Ellenkező esetben sérülne az eloszlásfüggvény nemcsökkenő tulajdonsága.)

[8] Ez azért hangsúlyos, mert a korszak korlátozott számítástechnikai lehetőségei mellett komoly előnyökkel bírt empirikus adatok illesztésénél az eloszlásfüggvény használata (hiszen az intervallumok valószínűsége integrálás helyett egyszerű kivonással kapható meg) szemben az egyébként szokásosabb sűrűségfüggvényekkel. Hasonlóképp könnyebben ragadhatók meg a kvantilisértékek is.

Rögtön látható, hogy ez egy szétválasztható változójú differenciálegyenlet, amit szeparálva, majd az integrálást parciális törtekre bontással elvégezve, kapjuk, hogy:

$$F(x) = \frac{1}{e^{-G(x)} + 1}.$$

Burr a cikkében 12, általa fontosnak vélt konkrét $F(x)$ eloszlásfüggvényt ad meg. Ezek közül az utolsó, továbbiakban a Burr XII:

$$F(x) = 1 - \frac{1}{\left(1 + x^c\right)^k},$$

amely csak az $x \in (0, \infty)$ tartományon értelmezett, és $0 < c, k \in \mathbb{R}$. (A Mellékletben megmutatjuk, hogy az általános formulából hogyan kapható meg a Burr XII.)

2.2.2. A Burr XII tulajdonságai

Sűrűségfüggvény. A Burr XII sűrűségfüggvénye egyszerű deriválással adódik:

$$f(x) = F'(x) = \frac{kcx^{c-1}}{\left(1 + x^c\right)^{(k+1)}},$$

ahol továbbra is $x > 0$.

A Mellékletben részletesebben is elemezzük a sűrűségfüggvény jellegét. Ebből ki fog derülni, hogy $c > 1$ esetben az eloszlás unimodális, $\sqrt[c]{\frac{c-1}{kc+1}}$ módusszal, $c \leq 1$ esetben L-alakú.

Kvantilisfüggvény. A Burr XII kvantilisfüggvénye (tehát az eloszlásfüggvényének az inverze) egyszerű algebrai átalakításokkal megkapható az eloszlásfüggvényből:

$$F^{-1}(p) = Q(p) = \left[\frac{1}{(1-p)^{1/k}} - 1\right]^{1/c}.$$

Ezzel kapcsolatban kiemeljük, hogy lehetséges a kvantilisfüggvényt zárt alakban előállítani, ami nagy számítástechnikai egyszerűsítést jelent, ha ilyen eloszlást követő véletlenszámokat kell generálnunk.

2.2.3. A Burr XII momentumai analitikusan

A momentumokon alapuló illesztés kulcsfeladata az elméleti eloszlás momentumainak felírása általánosságban, az eloszlás ismeretlenjeinek segítségével.

Átlag és szórás illesztése. A Burr XII-nek csak két paramétere van, így világos, hogy az első 4 momentumot – célkitűzésünk szerint – bizonyosan nem fogjuk tudni tetszőlegesen megszabni. Egyik megoldás, hogy a ferdeség/csúcsosság beállítása után meghatározzuk az – így már adódó – átlagot és szórást, majd az eloszlás x változójához hozzáadjuk az elméleti és az empirikus átlag különbségét, illetve szorozzuk azt az elméleti és empirikus szórás hányadosával. Ez a művelet könnyen belefoglalható az eloszlásfüggvénybe is, például:

$$F(x) = 1 - \frac{1}{\left[1 + \left(\frac{x-\mu}{\sigma}\right)^{c}\right]^{k}},$$

ahol választható például $\mu = \mu' - \bar{\mu}(c,k)$ és $\sigma = \frac{\sigma'}{\bar{\sigma}(c,k)}$. (Itt $\bar{\mu}(c,k)$, illetve $\bar{\sigma}(c,k)$ jelenti a harmadik és negyedik momentumhoz (az első kettőtől függetlenül) illesztett eloszlás első két momentumát, μ' és σ' pedig a kívánt várható értéket és a szórást.) Ilyen módon a 2 paraméteres eloszlásunk könnyedén 4 paraméteressé alakítható. (Ez a megoldás látható például *Hönschová* [2008]-ban is.)

Amennyiben nem definiált momentumokhoz, hanem empirikus adatokhoz illesztünk, akkor egy másik (kézenfekvő, és az előző által is sugallt) lehetőség, hogy az eloszlásból csak az adódó várható értéket vonjuk ki, illetve szórásával osztjuk le, majd ehhez az empirikus adatok standardizáltját illesztjük. (Ez kézi számításnál praktikus, hiszen táblázatba célszerű volt eleve a standardizált értékeket foglalni.)

Mivel az említettek semmilyen lényegi módosítást nem jelentenek, így a továbbiakban az eredeti eloszlást használjuk, nem törődve a várható értékkel és szórással, tudva, hogy azokat tetszőlegesen beállíthatjuk anélkül, hogy az bármiben módosítaná a most következő tárgyalást.

Ferdeség és csúcsosság. A következőkben az eloszlás momentumait, centrális momentumait és standardizált centrális momentumait származtatjuk, hogy így megkapjuk a ferdeség és csúcsosság már bemutatott γ_1 és γ_2 mutatószámait (a levezetések terjedelmi okokból a Mellékletben kaptak helyet). Az így kapott formulákat használhatjuk később az illesztéshez.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\Gamma^{-3}\left[2\lambda_{c,k}^3(1) - 3\Gamma(k)\lambda_{c,k}(1)\lambda_{c,k}(2) + \Gamma^2(k)\lambda_{c,k}(3)\right]}{\left\{\Gamma^{-2}(k)\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right]\right\}^{3/2}} = \frac{2\lambda_{c,k}^3(1) - 3\Gamma(k)\lambda_{c,k}(1)\lambda_{c,k}(2) + \Gamma^2(k)\lambda_{c,k}(3)}{\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right]^{3/2}}. \quad /4/$$

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2} =$$

$$= \frac{\Gamma^{-4}\left[-3\lambda_{c,k}^4(1) + 6\Gamma(k)\lambda_{c,k}^2(1)\lambda_{c,k}(2) - 4\Gamma^2(k)\lambda_{c,k}(1)\lambda_{c,k}(3) + \Gamma^3(k)\lambda_{c,k}(4)\right]}{\left\{\Gamma^{-2}(k)\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right]\right\}^2} = \quad /5/$$

$$= \frac{-3\lambda_{c,k}^4(1) + 6\Gamma(k)\lambda_{c,k}^2(1)\lambda_{c,k}(2) - 4\Gamma^2(k)\lambda_{c,k}(1)\lambda_{c,k}(3) + \Gamma^3(k)\lambda_{c,k}(4)}{\left(\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right)^2}.$$

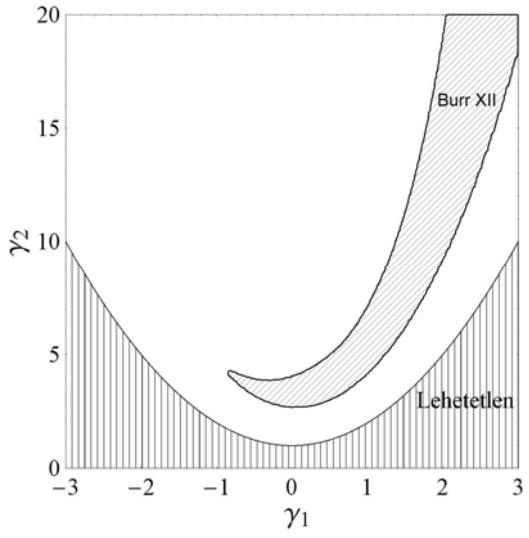
Ezeket a kifejezéseket (melyek c és k függvényei) egyenlővé kell tenni a specifikált g_1 és g_2 értékekkel, majd a kapott egyenletrendszert meg kell oldani c-re és k-ra. Ezt természetesen csak numerikusan tudjuk megtenni, ráadásul a megoldás még így is számos problémát felvet(het): numerikus instabilitás (például kerekítésekből adódó hibák), konvergencia kérdése stb. Egyszóval, bár itt ezt a kérdést egyáltalán nem tárgyaljuk, fontos jelezni, hogy a megoldás ettől még nem feltétlenül triviális.

A várható érték és a szórás illesztésének kérdését már tárgyaltuk, így az illesztés az eddigiek ismeretében teljeskörűen elvégezhető a Burr XII lefedési tartományában.

2.2.4. A Burr XII lefedési tartománya

A /4/ és /5/ egyenletek c és k argumentumait végigfuttatva lehetséges tartományukon, könnyen meghatározhatjuk – legalábbis empirikusan – a lefedési tartományt. (Lásd a 2. ábrát.)

2. ábra. A Burr XII eloszlás lefedési tartománya



A tartományt határoló görbékre *Rodriguez* [1977] analitikus egyenleteket is ad, ezekkel most – részben matematikai bonyolultságuk miatt – nem foglakozunk.

A Burr XII eloszlás, bár első ránézésre a lehetséges ferdeség/csúcsosság sík kis részét fedi, valójában igen praktikus, hiszen e „kis" rész számos, nagy gyakorlati jelentőségű eloszlást tartalmaz (többek között részeket mindhárom alapvető Pearson-típusból, a normális és logisztikus eloszlást, részeket mind a Johnson-féle S_U, mind az S_B eloszlásokból, részeket a Weibull- és a gamma-eloszlásokból stb.). Ennek következtében a kis fedés ellenére igen sok gyakorlati alkalmazásban jön szóba a használata, amit számos publikáció mutat az elmúlt évtizedekből.

Ezzel kapcsolatban az is előnyként jegyzendő meg, hogy a lefedett rész egyetlen algebrai alakú eloszlással érhető el (szemben például a Johnson- vagy Pearson-eloszlásokkal), így nem szükséges tartományonként eltérő eszköztár használata.

A Burr XII további előnye, hogy az eloszlásfüggvénye, illetve – ami ebből a szempontból még fontosabb – annak inverze (a kvantilisfüggvény) is megadható zárt alakban. Ez – figyelembe véve a közismert valószínűségszámítási tételt – azt jelenti, hogy a Burr XII eloszlást követő véletlen számok generálása igen egyszerűen, mindössze egy egyenletes véletlenszám-generátorral megvalósítható. Ez igen komoly előny akkor, ha számítógépes szimulációkhoz van szükség nagy mennyiségű Burr XII eloszlású véletlenszámra.

2.3. A Johnson-eloszláscsalád

Norman L. Johnson (*Karl Pearson* fiának témavezetése alatt készített) PhD-dolgozatában mutatta be a később róla elnevezett eloszláscsaládot. 1949-es munkájában, Pearson megoldásához hasonlóan, *Johnson* [1949] is eloszláscsaládot definiált, vagyis nem egyetlen formula paraméterezésével, hanem a $\gamma_1 - \gamma_2$ sík különböző területeire különböző függvényeket definiálva érte el célját.

Az eloszlásokat sűrűségfüggvényükön keresztül határozta meg, implicite:

$$z = \gamma + \delta \cdot \log f\left(\frac{x-\mu}{\lambda}\right),$$

ahol *z* standard normális eloszlás, míg az *f* függvény háromféle lehet:

– lognormális $S_L : f(u) = u$,
– korlátozatlan $S_U : f(u) = u + \sqrt{1+u^2}$,
– korlátozott $S_B : f(u) = u/(1-u)$.

A háromféle eloszlás együttesen teljesen lefedi a lehetséges $\gamma_1 - \gamma_2$ síkot.

Az S_L eloszlás ennek a síknak egyetlen egyenesét fedi le, így itt a két mutató egymást egyértelműen meghatározza. Az így kapott eloszlások egy oldalon korlátozottak, míg végtelenek a másik oldalon.

Az S_U eloszlások a $\gamma_1 - \gamma_2$ sík S_L vonal feletti területet fedik le, magukban foglalva a Pearson IV, V, VII eloszlásokat, illetve bizonyos VI-os eloszlásokat. Az így kapott eloszlások mindkét oldalukon végtelenek.

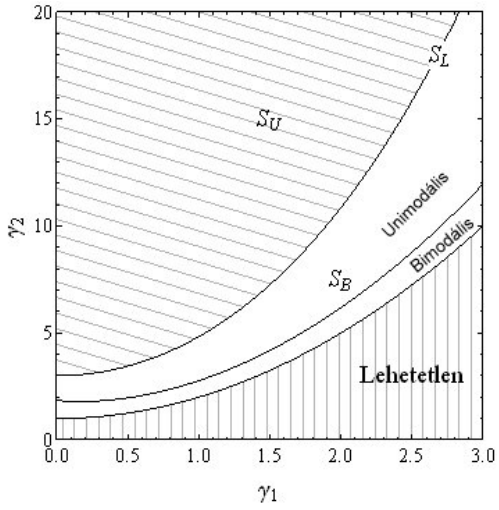
Az S_B eloszlások az S_L vonal és az eloszlások létezésének – korábban ismertetett – alsó határa közti területet fedik le, azaz ide értendők Pearson I, II, III és bizonyos VI-os eloszlásai. Ezek az eloszlások mindkét oldalon korlátozottak (*Draper* [1952]).

Johnson [1949] azt is megmutatta, hogy az S_B eloszlások bimodalitásának szükséges és elégséges feltétele, hogy

$$\delta < 2^{-1/2} \qquad\qquad |\gamma| < \frac{\sqrt{1-2\delta^2} - 2\delta^2 \tanh^{-1}\sqrt{1-2\delta^2}}{\delta}.$$

Ez a feltétel tágabb területet fed le, mint az a terület, ahol minden létező eloszlás szükségszerűen kétmódusú (*Draper* [1952]). Az eloszláscsalád lefedési tartományát a 3. ábra szemlélteti.

3. ábra. A Johnson-eloszláscsalád lefedési tartománya



Az eloszlás kiterjesztéseként *Johnson* [1954], illetve *Tadikamalla* és *Johnson* ([1980], [1982]) a Laplace-, illetve a logisztikus eloszlást (L-eloszláscsalád) használta a normális helyett. Utóbbi az eloszlásfüggvény és az inverz eloszlásfüggvény egyszerűbb kifejezhetősége miatt könnyebb illesztést tesz lehetővé.

2.3.1. A Johnson-eloszláscsalád illesztése momentumok alapján

A három eloszlás illesztését különválasztva kell kezelni. A szétválasztáshoz először a kívánt g_1 érték alapján az $\omega = e^{\delta^{-2}}$ helyettesítéssel a

$$(\omega - 1)(\omega + 2)^2 = g_1$$

egyenletet kell megoldani, majd ebből a

$$\gamma_2 = \omega^4 + 2\omega^3 + 3\omega^2 - 3$$

kifejezést kell értékelni. Ha $\gamma_2 > g_2$, akkor az S_B, egyébként az S_U eloszlás illesztése szükséges (*Hill et al.* [1976]).

Az S_U görbe esetén a momentumok zárt alakban kifejezhetők, az illesztés így ezek alapján numerikusan elvégezhető (az egyes formulák a függelékben megtalálhatók). Az illesztés megkönnyítésére Johnson több alkalommal is (*Johnson* [1965], [1974]) publikált táblázatokat, amelyek a $\gamma_1 - \gamma_2$ értékekhez tartozó γ és δ értékeket tartalmazzák. Helyettesítések sorozatával a kifejezések egyszerűsíthetők, egy negyedfokú, kétismeretlenes egyenletrendszerre, amiből az eredeti paraméterek visszaszámíthatók (*Tuenter* [2001]).

Az S_B görbék momentumai nem fejezhetők ki zárt alakban, így az illesztés még nehezebb. A megfelelő közelítő táblázatokat *Pearson* és *Hartley* [1972] közölte.

A momentumok alapján történő illesztés esetén tehát – hasonlóan Pearson eloszláscsaládjához – először meg kell találni, hogy melyik a megfelelő eloszlás, és a paraméterek becslése csak ezután végezhető el.

Amint az a 3. ábrából is kitűnik, a Johnson-eloszláscsalád a teljes ferdeség/csúcsosság síkot lefedi, ez nagyon fontos elméleti előnye.

2.3.2. A Johnson-eloszláscsalád illesztése kvantilisek alapján

Az illesztéshez több tanulmány szerint is szimmetrikus percentiliseket célszerű választani (*Bukac* [1972], *Mage* [1980], *Slifker–Shapiro* [1980]). Belátható, hogy ebben az esetben helyettesítések sorozatával a probléma egy másodfokú egyenletrendszer megoldásához vezet, ami a jelenlegi számítástechnikai háttér mellett általában nem okoz nehézséget.

A kvantilisen alapuló meghatározás szimulációs vizsgálatok alapján (különösen a korlátozott függvényre) nem csak egyszerűbb, de kisebb négyzetes hibával (MSE) is rendelkezik, mint a momentumokon alapuló (*Wheeler* [1980]).

2.4. Az általánosított λ-eloszlás

Az általánosított λ-eloszlás (GLD) ötlete eredetileg *Tukey-tól* származik (*Tukey* [1960]). Az eloszlásnak mindössze egyetlen szabadon állítható paramétere van, így a

normálistól több szempont szerint adott módon eltérő eloszlások előállítására nem alkalmas (lásd a Mellékletet).

A helyzet és a terjedelem kezelését biztosító technikák ismeretében – szimmetrikus eloszlásokat eredményező – triviális általánosítást adott meg *Ramberg* és *Schmeiser* [1972].

Két évvel később került sor (*Ramberg–Schmeiser* [1974]) a formula további általánosítására, a továbbiakban erre RS-eloszlásként hivatkozunk:

$$Q(u) = \lambda_1 + \lambda_2^{-1}\left[u^{\lambda_3} - (1-u)^{\lambda_4}\right],$$

ahol λ_1 a helyzetért, λ_2 a szóródásért, λ_3 és λ_4 az eloszlás alakjáért felelősek. Összhangban az 1972-es eredményekkel, a $\lambda_3 = \lambda_4$ eset szimmetrikus eloszlásokat ad.

Ramberg és szerzőtársai (*Ramberg et al.* [1979]) megmutatták, hogy bizonyos paraméter-kombinációkra kapott eredmények nem lehetnek az eloszlás kvantilisei (adott λ_2 mellett a λ_3–λ_4 tér bizonyos kombinációi nem érvényesek). A létezés feltétele a sűrűségfüggvény nemnegativitásával, azaz a

$$\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \geq 0$$

feltétellel egyenértékű (*Su* [2005]).

Az elérhető eloszlások túlnyomó része egymóduszú; azonban korlátozott formában, de U-alakú és nyesett (L-alakú) eloszlások is előállíthatók. A $\lambda_3 \geq 1, \lambda_4 \leq 2$ paraméterezés U-alakú, míg a $\lambda_3 = 0$ paraméter L-alakú eloszlásokhoz vezet (*Ramberg et al.* [1979]).

A λ_3–λ_4 sík teljes lefedettségének biztosításra is találtak megoldást (*Freimer et al.* [1988]) a következő paraméterezésen keresztül (FMKL):

$$Q(u) = \lambda_1 + \lambda_2^{-1}\left[\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4}}{\lambda_4}\right].$$

Az FMKL-eloszlás már a teljes λ_3–λ_4 térben definiált, az illesztés egyetlen feltétele, hogy $\lambda_2 > 0$ legyen. Az eloszlás *k*-adik momentuma – hasonlóan az RS-eloszláshoz – akkor véges, ha $\min(\lambda_3, \lambda_4) > -1/k$. Az illesztéshez szükséges alapszámításokra vonatkozó irodalom ugyanakkor meglehetősen hiányos.

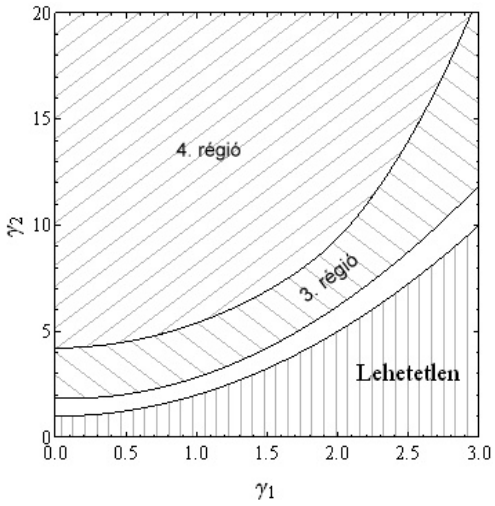
2.4.1. Az általánosított λ-eloszlás illesztése momentumok alapján

Az RS-eloszlás momentumokon alapuló illesztéséhez az eloszlás sűrűségfüggvényéből tudunk kiindulni. A momentumok definíciói alapján a paraméterek függvé-

nyében kifejezhetők a szükséges centrális momentumok, illetve a ferdeség és csúcsosság mutatói is. A Mellékletben található formulákból látható, hogy a γ_1 és γ_2 értékek csak a λ_3 és λ_4 függvényei, így az eloszlás ferdeségének és csúcsosságának meghatározása után a várható érték és a variancia külön paraméterezhető. Problémát okoz, hogy a λ_3- és a λ_4-értékek zárt alakban nem fejezhetők ki, így az egyenletrendszer megoldása erősen számításigényes, különösen a formulákban található béta függvények számítása miatt.

Karian és *Dudewicz* arra is felhívja a figyelmet, hogy az RS-eloszlás csak az $1{,}8\left(\gamma_1^2+1\right)\le\gamma_2$ teret tudja lefedni, így $\gamma_1-\gamma_2$ tér egy szűk sávjában nem lehetséges eloszlások generálása. Ez éppen az a sáv, ahol a lehetséges minimális csúcsosságnál csak kissé nagyobb csúcsosságértékek találhatók. (*Karian–Dudewicz* [2000]), amint a 4. ábra is mutatja. Valóban látható, hogy az általánosított λ-eloszlás egy szűk, közvetlenül a lehetetlen tartomány fölötti sáv kivételével lefedi a ferdeség/csúcsosság síkot. Az ábrán azt is megadtuk, hogy melyik *Karian–Dudewicz* [2000] szerinti régióból kikerülő paraméterekkel végezhető el a lefedés. (További régiók a nagyobb ferdeségeknél kaphatnak szerepet; csak többféle lefedésre adnak módot, a lefedés teljességét nem befolyásolják.)

*4. ábra. Az általánosított λ-eloszlás lefedési tartománya**



* A 3. régióban $\lambda_3, \lambda_4 > 0$, míg a 4. régióban $\lambda_3, \lambda_4 < 0$.

Az egyes paraméter-kombináció intervallumokban javasolt kezdőértékekről jó áttekintést ad (*Karian–Dudewicz* [2000]), míg *Lakhany–Mausser* [2000]-nél további illesztési módszerek értékelését is megtaláljuk.

2.4.2. Az általánosított λ-eloszlás illesztése kvantilisek alapján

A GLD-eloszlás valamennyi változata kvantilis függvényével adott, így kézenfekvőnek látszik a kvantiliseken alapuló illesztés. A 4 paraméteres változatok (RS és FMKL) esetén 4 kvantilis érték megadásával meghatározhatók a paraméterek. A kvantilis függvény formájából adódik, hogy az egyenletrendszerből gyorsan kiejthető a λ_1 és a λ_2 paraméter, így egy kétegyenletes, kétismeretlenes nemlineáris egyenletrendszert kell megoldani. Az egyenletrendszer csak hatványfüggvényeket tartalmaz, megoldása tehát nagyságrendekkel gyorsabban elvégezhető, mint a momentumokon alapuló illesztés (*Su* [2005]).

A problémát ebben az esetben az okozza, hogy 4 kvantilis közvetlenül nem tudja jól leírni az eloszlás alakját. A kvantiliseken alapuló csúcsosságmutatók is általában 4 kvantilist használnak (*Kim–White* [2004]), amik azonban éppen az eloszlás helyzetét nem határozzák meg. A probléma áthidalására *Karian* és *Dudewicz* 4 kvantilisen alapuló mutatót javasolt. Az első mutató, a medián szolgál az eloszlás helyzetének kiváltására (az első nyers momentum párjaként az eloszlás helyzetéért felel). A második mutató valamilyen interpercentilis mutató lehet, az adatok középső tartományának terjedelmét mutatja, $0 < u < 0{,}25$ (a szóródás mutatója). A harmadik mutató a ferdeséget írja le, míg a negyedik a csúcsosság egy lehetséges mérőszáma.

$$\rho_1 = F^{-1}(0{,}5) = \lambda_1 + \frac{\left(\frac{1}{2}\right)^{\lambda_3} - \left(\frac{1}{2}\right)^{\lambda_4}}{\lambda_2}$$

$$\rho_2 = F^{-1}(1-u) - F^{-1}(u) = \frac{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}}{\lambda_2}$$

$$\rho_3 = \frac{F^{-1}(0{,}5) - F^{-1}(u)}{F^{-1}(1-u) - F^{-1}(0{,}5)} = \frac{(1-u)^{\lambda_4} - u^{\lambda_3} + \left(\frac{1}{2}\right)^{\lambda_3} - \left(\frac{1}{2}\right)^{\lambda_4}}{(1-u)^{\lambda_3} - u^{\lambda_4} + \left(\frac{1}{2}\right)^{\lambda_4} - \left(\frac{1}{2}\right)^{\lambda_3}}$$

$$\rho_4 = \frac{F^{-1}(0{,}75) - F^{-1}(0{,}25)}{\rho_2} = \frac{\left(\frac{3}{4}\right)^{\lambda_3} - \left(\frac{1}{4}\right)^{\lambda_4} + \left(\frac{3}{4}\right)^{\lambda_4} - \left(\frac{1}{4}\right)^{\lambda_3}}{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}}.$$

A harmadik és a negyedik mutató csak λ_3 és λ_4 függvénye, így ebben az esetben is alkalmazható a rekurzív megoldás, először λ_3 és λ_4 meghatározása, majd abból λ_2, végül λ_1 kalibrálása.

2.5. A g-and-h-eloszlás

A g-and-h-eloszlás, az eredeti λ-eloszláshoz hasonlóan, *John Wilder Tukey* nevéhez fűződik. Az eloszlás egy 1977-es konferencia-előadásban (*Tukey* [1977]) került ismertetésre, amelyből tanulmány nem készült.

A g-and-h-eloszlást kvantilisfüggvényével (inverz eloszlásfüggvényével) definiáljuk, a standard normális eloszlásból (z) kiindulva, az alábbi transzformációval:

$$q(z) = g^{-1}\left(\mathrm{e}^{gz} - 1\right)\mathrm{e}^{hz^2/2} \quad g \neq 0 \quad h > 0\,.$$

A paraméterek közvetlenül alakítják az eloszlás alakját, így g felel a ferdeségért (irányban és nagyságban), h pedig a csúcsosságért (a kurtózissal pozitívan korrelál).

A két paraméter szerinti határeloszlások ($g \to 0$, illetve $h \to 0$) is meghatározhatók, illetve könnyen belátható, hogy a $g \to 0, h \to 0$ paraméterezés szerinti határeloszlás éppen a standard normális eloszlást adná vissza.

A g-and-h-eloszlás sűrűségfüggvénye az a következő alakban írható fel:

$$f_{q(Z)}\left(q(z)\right) = f_{q(Z)}\left(q(z), \frac{f_Z(z)}{q'(z)}\right).$$

Ahogy *Headrick* (*Headrick et al.* [2008]) megmutatja, a $q(z)$ transzformáció szigorú monotonitása miatt a sűrűségfüggvény unikális, globális maximumponttal rendelkezik, vagyis a kapott eloszlások egymóduszúak. Az eloszlások helyzetének jellemzésére az inverz eloszlásfüggvénnyel való megadásnak megfelelően a medián mutatkozik a legegyszerűbb középértéknek. Belátható, hogy a medián a $q(z=0)=0$ helyen lesz, ahogy a kiindulásul szolgáló standard normális eloszlásra is igaz.

2.5.1. A g-and-h eloszlás illesztése momentumok alapján

A sűrűségfüggvény felhasználásával az eloszlás momentumai definiálhatók:

$$E\left[q(z)^k\right] = \int_{-\infty}^{+\infty} q(z)^k \, f_Z(z)\,\mathrm{d}z\,.$$

A k-ad rendű momentum létezésének feltétele, hogy $0 \le h < 1/k$ teljesüljön. Ebből az első négy nyers momentum viszonylag egyszerűen származtatható. A pontos formulák a Mellékletben találhatók.

A harmadik- és negyedik centrális momentumokon alapuló ferdeség és csúcsosság mutatók (γ_1 és γ_2) a g és h függvényében megadhatók (a formulák a Melléklet-

ben találhatók). Ahogy *Rayner–MacGillivray* [2002] jelzi, valamennyi ferdeség elérhető, ugyanakkor a 3 alatti csúcsosságok (lapult eloszlások) nem képezhetők.

A momentumokon alapuló illesztéshez – még akkor is, ha az csupán a ferdeséghez és a csúcsossághoz való igazodást jelenti – megoldandó egyenletrendszer numerikusan is erősen számításigényes.

Az eloszlásnak két paramétere van, így γ_1 és γ_2 alapján *g* és *h* értékéből az első két momentum egyértelműen következik. Ha tetszőleges – vagy standard – első két momentummal rendelkező eloszlást szeretnénk illeszteni, akkor további általánosítás szükséges, ami az inverz eloszlásfüggvénnyel való megadás miatt analitikusan nehezen kezelhető, ismereteink szerint jelenleg megoldatlan probléma.

2.5.2. A g-and-h-eloszlás illesztése kvantilisek alapján

A kvantilisfüggvénnyel való definiálás sugallja a kvantiliseken alapuló illesztést. Figyelembe véve, hogy az eloszlás mediánja mindig 0, a triviális kvantilisen túl két kvantilis megadása egyértelműen meghatározza az eloszlást. Az 1.2. pontban leírtak alapján ez előny, ha célunk adott empirikus eloszláshoz elméleti eloszlás illesztése, ugyanakkor a normálistól adott mértékben eltérő eloszlás paraméterezése nem oldható meg. A kvantiliseken alapuló illesztés előnyei a többdimenziós eloszlások esetén jelennek meg (*Field–Genton* [2006]).

2.6. Fleishman-eloszlás

Bár „eloszlás” néven említjük, a Fleishman-eloszlás sokkal inkább egy nevezetes eloszlás-transzformációs eljárás. *Allen Fleishman* 1978-ben publikálta (*Fleishman* [1978]); már eredetileg is Monte-Carlo-szimulációkra gondolva.

Annak ellenére, hogy a módszer igen elegáns, és számítástechnikai szempontból is rendkívül jól kezelhető, mintegy 2 évtizedig szinte a feledés homályába merült. Jelentős elméleti fejlemény egészen a 2000-es évekig nem történt, és az alkalmazások többsége (*Headrick–Sheng–Hodis* [2007]) is az 1990-es évekre esik.

2002-ben Headrick kiterjesztette a módszert (*Headrick* [2002]), hogy az 6 momentumig alkalmas legyen illesztésre, majd ugyanő 2007-ben megoldotta (*Headrick–Kovalchuk* [2007]) a legfontosabb nyitott kérdést: megadta analitikus alakban a Fleishman-eloszlás sűrűség- és eloszlásfüggvényét.

2.6.1. A Fleishman-eloszlás áttekintése, illesztés

Fleishman módszere mindössze egy standard normális eloszlású véletlenszámot igényel. Alapötlete: készítsük el az így generált véletlen változó egy transzformáltját.

A transzformáló függvény konkrét alakja, paraméterei nyilván hatással lesznek a transzformált változó eloszlására, így momentumaira is. Ha a transzformáló függvénynek egyetlen paramétere van, akkor azzal nyilván egyszerre állítjuk az összes momentumot (legjobb esetben is), így általánosságban nincs arra remény, hogy el tudjuk érni, hogy a transzformált változó több (például négy) momentuma előírás szerinti legyen. Viszont ahogy növeljük paramétereinek számát, úgy válhat kellően „testreszabhatóvá" a transzformált függvény. A polinomiális transzformáló függvény alkalmas arra, hogy ezt megvalósítsa, hiszen kényelmesen állítható a szabad paraméterek száma. Ha például 4 paraméterre van szükségünk, mert a transzformált változó négy momentumát szeretnénk előírás szerintire állítani, akkor válasszuk az

$$Y = G(Z) = a + bZ + cZ^2 + dZ^3 \qquad /6/$$

transzformáló függvényt, ahol tehát $Z \sim N(0,1)$.

Nincs más dolgunk, mint meghatározni a paramétereket. Ehhez azt kell tudnunk, hogy a transzformált változó momentumai hogyan írhatók fel a transzformációs függvény ismeretében. Például, az első momentum, a várható érték esetén nincs nehéz dolgunk, hiszen az ismert tétel szerint, ha képezzük egy X valószínűségi változó $V = g(X)$ transzformáltját (ahol $g : \mathsf{R} \to \mathsf{R}$), akkor $E(V) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\mathrm{d}x$ folytonos esetben. Azaz pusztán a transzformáló függvény és az eredeti sűrűségeloszlás ismeretében (egyetlen integrálással) megadható a transzformált változó várható értéke. Maradva a /6/ transzformáló függvénynél, azt kapjuk, hogy

$$E(Y) = \int_{-\infty}^{\infty} \left(a + bZ + cZ^2 + dZ^3\right) \cdot \varphi(x)\mathrm{d}x = a + c.$$

Ha tehát azt szeretnénk, hogy generált eloszlásunk várható értéke μ legyen, nincs más dolgunk, mint kielégíteni a $a + c = \mu$ egyenletet.

A további momentumokra hasonló kifejezések kaphatók, bár a számítások, ha nem is bonyolultabbak, de mindenestre jóval munkaigényesebbek lesznek (a hosszas polinomszorzások miatt), ezért csak a végeredményt közöljük. Felírva tehát ugyanezt a következő három momentumra is, három további egyenletet kapunk, így már megoldható lesz az négy ismeretlenes és immár négy egyenletes egyenletrendszerünk. Az eredményül kapott egyenletrendszer tehát, standardizált ($\mu = 0$, $\sigma = 1$) esetben:

$$a + c = 0$$

$$b^2 + 6bd + 15d^2 + 2c^2 = 1$$

$$2c\left(b^2 + 24bd + 105d^2 + 2\right) = g_1$$

$$24\left(bd + c^2\left(1 + b^2 + 28bd\right) + d^2\left(12 + 48bd + 141c^2 + 255d^2\right)\right) = g_2.$$

Ennek megoldásával a keresett transzformációs együtthatókat kapjuk. (Természetesen teljesen nyilvánvaló, hogy a megoldást numerikus úton kell végeznünk.)

Ha megvannak a transzformációs együtthatók, nincs más dolgunk, mint a standard normális eloszlású véletlenszámokat generálni, majd a felparaméterezett polinomiális függvénnyel transzformálni őket.

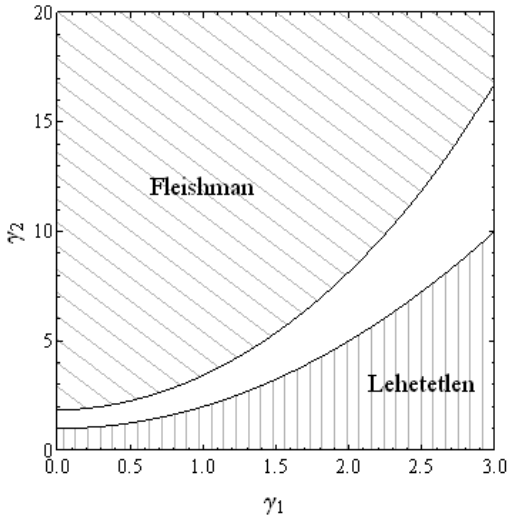
Ahogy említettük, *Headrick* [2002] megadja a szükséges formulákat az első 6 momentumhoz történő illesztésre is.

2.6.2. A Fleishman-eloszlás lefedési tartománya

A Fleishman-eloszlás nagyjából lefedi a teljes ferdeség/csúcsosság síkot, ám a módszer egy, az elméleti minimumnál kicsit magasabban húzódó minimum-csúcsossággal rendelkezik (adott ferdeségre). Ebből következően a lapult eloszlások generálása problémába ütközhet. Példának okáért, *Headrick–Sawilowsky* [2002] megmutatta, hogy szimmetrikus esetben a legkisebb elérhető csúcsosság $\gamma_1 = 1{,}85$ (1 helyett).

Saját számítási eredményeinket[9] e tekintetben az 5. ábra közli, melyen jól látható a nemgenerálható tartomány elhelyezkedése.

5. ábra. A Fleishman-transzformációs módszer lefedési tartománya



A Fleishman-módszer legnagyobb előnye akkor jelentkezik, ha nagy mennyiségű adott eloszlást követő véletlenszám-generálás szükséges. A módszer tervezéséből

[9] Ezen ábránál (és a többinél is) a számításokat és a rajzolást végző .nb Mathematica munkafüzet elérhető a szerzőknél.

adódóan ehhez mindössze egy standard normális eloszlású véletlenszám-generátor szükséges, ennek birtokában néhány szorzással és összeadással, azaz számítástechnikailag igen egyszerűen előállíthatók a kívánt véletlenszámok.

A módszer további jellemzője, hogy a ferdeség/csúcsosság síkon a lefedett terület igen nagy, ám nem a teljes elméletileg lehetséges terület.

*

A tanulmányban áttekintettük azokat az eloszlásokat, illetve eloszláscsaládokat, amelyek alkalmasak lehetnek változatos alakú – lehetőség szerint a ferdeség/csúcsosság síkot minél jobban lefedő – eloszlásból származó minták generálásához. Fontos szempontnak tartottuk, hogy eloszláscsaládok esetén a megfelelő eloszlás kiválasztása minél egyszerűbb legyen, az eloszlások paramétereinek megválasztása a kívánt empirikus eloszláshoz minél egyszerűbben megtörténhessen. Az irodalomban erre a célra fellelhető hat eloszlás(család) legfontosabb jellemzőit a táblázatban foglaljuk össze.

Az eloszlások legfontosabb jellemzői

Eloszlás(család)	Az eloszlás megadása	Momentumokon alapuló illesztés	Kvantiliseken alapuló illesztés	Ferdeség/csúcsosság lefedés
Pearson	sűrűségfüggvény (közvetetten)	lehetséges	nem ajánlott	Teljes; 3 algebrai alakkal
Burr	eloszlásfüggvény	lehetséges	nem ajánlott	Részleges, túl kis és túl nagy minimum feletti csúcsosságok egyaránt lehetetlenek; 1 algebrai alakkal
Johnson	sűrűségfüggvény (közvetetten)	lehetséges	ajánlott	Teljes; 2 algebrai alakkal
GLD	inverz eloszlásfüggvény	lehetséges, de nem ajánlott	ajánlott	Szinte teljes; kis minimum feletti csúcsosságok lehetetlenek; 1 algebrai alakkal
g-and-h	inverz eloszlásfüggvény	lehetséges	nem kivitelezhető	Szinte teljes; kis minimum feletti csúcsosságok lehetetlenek; 1 algebrai alakkal
Fleishman	sűrűségfüggvény	ajánlott	nem megoldott	Szinte teljes; kis minimum feletti csúcsosságok lehetetlenek; 1 algebrai alakkal

Irodalom

BUKAC, J. L. [1972]: Fitting S_B Curves Using Symmetrical Percentile Points. *Biometrika*. 59. évf. 688–690. old.

BURR, I. W. [1942]: Cumulative Frequency Functions. *The Annals of Mathematical Statistics.* 13. évf. 2. sz. 215–232. old.

BURR, I. W – CISLAK, P. J. [1968]: On a General System of Distributions: I. Its Curve-Shape Characteristics; II. The Sample Median.; III. The Sample Range. *Journal of the American Statistical Association*. 63. évf. 322. sz. 627–643. old.

BURR, I. W. [1973]: Parameters for a General System of Distributions to Match a Grid of α_3 and α_4 . *Communications in Statistics*. 2. évf. 1. sz. 1–21. old.

CRAIG, C. C. [1936]: A New Exposition and Chart for the Pearson System of Frequency Curves. *Annals of Mathematical Statistics*. 7. évf. 1. sz. 16–28. old.

DRAPER, J. [1952]: Properties of Distributions Resulting from Certain Simple Transformations of the Normal Distribution. *Biometrika*. 39. évf. 3–4. sz. 290–301. old.

FERENCI, T. [2009]: *Using Massively Parallel Processing in the Testing of the Robustness of Statistical Tests with Monte Carlo Simulation*. Challenges for Analysis of the Economy, the Businesses, and Social Progress International Scientific Conference. November 19–21. Szeged.

FIELD, C. – GENTON, M. G. [2006]: The Multivariate g-and-h Distribution. *Technometrics*. 48. évf. 1. sz. 104–111. old.

FLEISHMAN, A. I. [1978]: A Method for Simulating Non-normal Distributions. *Psychometrika*. 43. évf. 521–532. old.

FREIMER, M. ET AL. [1988]: A Study of the Generalized Tukey Lambda Family. Communications in Statistics. *Theory and Methods*. 17. évf. 10. sz. 3547–3567. old.

HALL, A. R. [2005]: *Generalized Method of Moments*. Oxford University Press. Oxford.

HATKE, M. A. [1949]: A Certain Cumulative Probability Function. *Annals of Mathematical Statistics*. 20. évf. 3. sz. 461–463. old.

HEADRICK, T. C. [2002]: Fast Fifth-Order Polynomial Transforms for Generating Univariate and Multivariate Non-normal Distributions. *Computational Statistics & Data Analysis*. 40. évf. 685–711. old.

HEADRICK, T. C. – SAWILOWSKY, S. S. [2000]: Weighted Simplex Procedures for Determining Boundary Points and Constants for the Univariate and Multivariate Power Methods. *Journal of Educational Behavioral Statistics.* 25. évf. 417–436. old.

HEADRICK, T. C. – SHENG Y. – HODIS F. A. [2007]: Numerical Computing and Graphics for the Power Method Transformation Using Mathematica. *Journal of Statistical Software*. 19. évf. 3. sz. 1–17. old.

HEADRICK, T. C. – KOWALCHUK, R. K. [2007]: The Power Method Transformation: Its Probability

Density Function, Distribution Function, and Its Further Use for Fitting Data. *Journal of Statistical Computation and Simulation*. 77. évf. 229–249. old.

HEADRICK, T. C. – KOWALCHUK, R. K. – SHENG, Y. [2008]: Parametric Probability Densities and Distribution Functions for Tukey g-and-h Transformations and Their Use for Fitting Data. *Applied Mathematical Sciences*. 2. évf. 9. sz. 449–462. old.

HEINRICH, J. [2004]: *A Guide to the Pearson Type IV Distribution*. http://www-cdf.fnal.gov/publications/cdf6820_pearson4.pdf (Elérés dátuma: 2010. május 15.)

HILL, I. D. – HILL, R. – HOLDER, R. L. [1976]: Fitting Johnson Curves by Moments. *Journal of the Royal Statistical Society*. Series C (Applied Statistics). 25. évf. 2. sz. 180–189. old.

HÖNSCHHOVÁ, E. [2008]: *Estimation of the Scale Parameter in Burr Distribution*. ROBUST 2008 Poster Section. Szeptember 8–12. Prbylina.

JEFFREYS, H. [1948]: *Theory of Probability.* Oxford University Press. Oxford.

JOHNSON, N. L. [1949]: Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*. 36. évf. 1–2. sz. 149–176. old.

JOHNSON, N. L. [1954]: Systems of Frequency Curves Derived from the First Law of Laplace. *Trabajos de Estadistica.* 5. évf. 283–291. old.

JOHNSON, N. L. [1965]: Tables to Facilitate Fitting S_U Frequency Curves. *Biometrika*. 52. évf. 3–4. sz. 547–558. old.

JOHNSON, N. L. [1974]: Extensions and Corrections to 'Tables to Facilitate Fitting S_U Frequency Curves'. *Biometrika*. 61. évf. 1. sz. 203–205. old.

KARIAN, Z. – DUDEWICZ, E. [2000]: *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods*. CRC Press. Boca Raton.

KENDALL, M. G. – STUART, A. [1977]: The Advanced Theory of Statistics. *Distribution Theory*. Vol. 1. Charles Friffin & Company. London.

KIM, T-H. – WHITE, H. [2004]: On More Robust Estimation of Skewness and Kurtosis. *Finance Research Letters*. 1. évf. 56–73. old.

LAKHANY, A. – MAUSSER, H. [2000]: Estimating the Parameters of the Generalized Lambda Distribution. *Algo Research Quarterly*. 3. évf. 3. sz. 47–58. old.

LEE, P. M. [2009]: *Bayesian Statistics: An Introduction*. Wiley. New York.

MAGE, D. T. [1980]: An Explicit Solution for S_B Parameters Using Four Percentile Points. *Technometrics*. 22. évf. 247–251. old.

PEARSON, K. [1893]: Contributions to the Mathematical Theory of Evolution. *Proceedings of the Royal Society of London.* 54. köt. 329–333. old.

PEARSON, K. [1895]: Contributions to the Mathematical Theory of Evolution, II: Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London*. 186. köt. 343–414. old.

PEARSON, K. [1901]: Mathematical Contributions to the Theory of Evolution, X: Supplement to a Memoir on Skew Variation. *Philosophical Transactions of the Royal Society of London*. Series A. Containing Papers of a Mathematical or Physical Character. 197. köt. 443–459. old.

PEARSON, K. [1916]: Mathematical Contributions to the Theory of Evolution, XIX: Second Supplement to a Memoir on Skew Variation. *Philosophical Transactions of the Royal Society of London*. Series A. Containing Papers of a Mathematical or Physical Character. 216. köt. 429–457. old.

PEARSON, E. S. – HARTLEY, H. O. [1972]: *Biometrika Tables for Statisticians*. Vol. 2. University Press. Cambridge.

RAMBERG, J. S. ET AL. [1979]: A Probability Distribution and Its Uses in Fitting Data. *Technometrics*. 21. évf. 2. sz. 201–214. old.

RAMBERG, J. S. – SCHMEISER, B. W. [1972]: An Approximate Method for Generating Symmetric Random Variables. *Communications of the ACM*. 15. évf. 11. sz. 987–990. old.

RAMBERG, J. S. – SCHMEISER, B. W. [1974]: An Approximate Method for Generating Asymmetric Random Variables. *Communications of the ACM*. 17. évf. 2. sz. 78–82. old.

RAYNER, G. D. – MACGILLIVRAY, H. L. [2002]: Numerical Maximum Likelihood Estimation for the g-and-k and the Generalized g-and-h Distributions. *Statistics and Computing*. 12. évf. 57–75. old.

RODRIGUEZ, R. N. [1977]: A Guide to the Burr Type XII Distributions. *Biometrika*. 64. évf. 1. sz. 129–134. old.

SLIFKER, B. K. – SHAPIRO, S. S. [1980]: The Johnson System: Selection and Parameter Estimation. *Technometrics*. 22. évf. 239–246. old.

SU, S. [2005]: A Discretized Approach to Flexibly Fit Generalized Lambda Distributions to Data. J*ournal of Modern Applied Statistical Methods*. 4. évf. 2. sz. 408–424. old.

TADIKAMALLA, P. R. – JOHNSON, N. L. [1980]: *Systems of Frequency Curves Generated by Transformations of Logistic Variables.* Kézirat.

TADIKAMALLA, P. R. [1980]: A Look at the Burr and Related Distributions. *International Statistical Review / Revue Internationale de Statistique*. 48. évf. 3. sz. 337–344. old.

TADIKAMALLA, P. R. – JOHNSON, N. L. [1982]: Systems of Frequency Curves Generated by Transformations of Logistic Variables. *Biometrika*. 69. évf. 2. sz. 461–465. old.

TUENTER, H. J. H. [2001]: An Algorithm to Determine the Parameters of S_U-curves in the Johnson System of Probability Distributions by Moment Matching. *Journal of Statistical Computation and Simulation*. 70. évf. 4. sz. 325–347. old.

TUKEY, J. W. [1960]: *The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts.* Technical Report 36. Statistical Techniques Reasearch Group. Princeton University. Princeton.

TUKEY, J. W. [1977]: *Modern Techniques in Data Analysis.* Regional Research Conference. Június 13–17. North Dartmouth, MA.

WHEELER, R. E. [1980]: Quantile Estimators of Johnson Curve Parameters. *Biometrika*. 67. évf. 3. sz. 725–728. old.

Summary

In simulational studies, it is often necessary to generate random numbers coming from distributions that have specified properties. If a well-known, typical distribution is used, the necessary steps can be performed easily, and they are included in statistical program packages. However, if we need distributions that have properties considered to be parameters, such as arbitrarily set moments, we might face problems. Now we present and examine a few solutions for this problem (Pearson-, Johnson-distribution families, Generalized λ-distribution, Burr XII, Tukey "g-and-h" and Fleishman transformation), with their limits of application, and an analysis of the questions that arise when fitting them.

Melléklet

A) A Pearson-féle differenciálegyenlet megoldása

A megadott differenciálegyenlet egy változó együtthatójú, lineáris differenciálegyenlet, így megoldása semmilyen problémát nem okozhat:

$$f(x) = \exp\left(-\int \frac{x}{b_0 + b_1 x + b_2 x^2} \right).$$

Ez az integrál is meghatározható különösebb gond nélkül, először parciális törtekre bontva, majd alkalmasan helyettesítve:

$$\int \frac{x}{b_0 + b_1 x + b_2 x^2} = \frac{\ln\left(b_0 + b_1 x + b_2 x^2\right)}{2b_2} - \frac{b_1}{b_2\sqrt{4b_0 b_2 - b_1^2}} \cdot \operatorname{arctg}\left[\frac{2b_2 x + b_1}{\sqrt{4b_0 b_2 - b_1^2}} \right] + C.$$

Ebből már felírható a sűrűségfüggvény:

$$f(x) = k \cdot \left(b_0 + b_1 x + b_2 x^2\right)^{\frac{1}{2b_2}} \cdot \exp\left(\operatorname{arctg}\left[\frac{2b_2 x + b_1}{\sqrt{4b_0 b_2 - b_1^2}} \right] \right)^{-\frac{b_1}{b_2\sqrt{4b_0 b_2 - b_1^2}}}.$$

Ha $4b_0 b_2 - b_1^2 > 0$ teljesül, akkor ez a kifejezés x-től függetlenül feltétlenül valós, hiszen az arctg argumentuma valós, a $\sqrt{4b_0 b_2 - b_1^2}$ szintén valós, továbbá a $b_0 + b_1 x + b_2 x^2$ polinomnak nem lesz valós gyöke, így felthetjük, hogy értéke végig pozitív. (Azt ugyanis az általánosság korlátozása nélkül feltehetjük, hogy a b_2 főegyüttható pozitív; ez x előjelének esetleges megcserélésével biztosan elérhető.)

Ha ellenben a $b_0 + b_1 x + b_2 x^2$-nek van valós gyöke, a sűrűségfüggvényt célszerű átalakítani, hogy a komplex argumentumok megjelenését kiküszöböljük. Ehhez induljunk ki a következő összefüggésből:

$$\operatorname{arctg}(ix) = i\operatorname{artgh}(x),$$

illetve az artgh definíciójának figyelembevételével:

$$\operatorname{arctg}(ix) = i\operatorname{artgh}(x) = \frac{i}{2}\ln\left(\frac{1+x}{1-x}\right).$$

Ebből következően

$$\operatorname{arctg}(x) = \operatorname{arctg}\left(i\cdot\frac{x}{i}\right) = \frac{i}{2}\ln\left(\frac{1+\frac{x}{i}}{1-\frac{x}{i}}\right).$$

Innen pedig:

$$\exp\left(\operatorname{arctg}(x)\right) = \left(\frac{1+\frac{x}{i}}{1-\frac{x}{i}}\right)^{\frac{i}{2}} = \left(1+\frac{x}{i}\right)^{\frac{i}{2}}\left(1-\frac{x}{i}\right)^{-\frac{i}{2}}.$$

Ennek ismeretében térjünk vissza a sűrűségfüggvény explicit felírására, és alakítsuk át azt:

$$f(x) = k\cdot\left(b_0+b_1x+b_2x^2\right)^{\frac{1}{2b_2}}\cdot\left[\left(1+\frac{2b_2x+b_1}{i\sqrt{4b_0b_2-b_1^2}}\right)^{\frac{i}{2}}\left(1-\frac{2b_2x+b_1}{i\sqrt{4b_0b_2-b_1^2}}\right)^{-\frac{i}{2}}\right]^{-\frac{b_1}{b_2\sqrt{4b_0b_2-b_1^2}}} =$$

$$= k\cdot\left(b_0+b_1x+b_2x^2\right)^{\frac{1}{2b_2}}\cdot\left[\left(1+\frac{2b_2x+b_1}{\sqrt{b_1^2-4b_0b_2}}\right)^{\frac{i}{2}}\left(1-\frac{2b_2x+b_1}{\sqrt{b_1^2-4b_0b_2}}\right)^{-\frac{i}{2}}\right]^{-\frac{b_1}{b_2\sqrt{4b_0b_2-b_1^2}}} =$$

$$= k\cdot\left(b_0+b_1x+b_2x^2\right)^{\frac{1}{2b_2}}\cdot\left[\left(1+\frac{2b_2x+b_1}{\sqrt{b_1^2-4b_0b_2}}\right)\left(1-\frac{2b_2x+b_1}{\sqrt{b_1^2-4b_0b_2}}\right)^{-1}\right]^{\frac{b_1}{2b_2\sqrt{b_1^2-4b_0b_2}}}.$$

Ez a felírás nyilvánvalóan ekvivalens az előzővel, ám épp akkor nem jelennek meg benne komplex számok, ha abban megjelennének (tehát akkor, ha a $b_0 + b_1 x + b_2 x^2$ -nak van valós gyöke.)

Nem szükségszerű, hogy elsőként az arctg -t tartalmazó formulát vezessük le, és abból származtassuk a másikat. A klasszikus levezetési lehetőség, hogy a valós, illetve komplex gyökök feltételezése mellett eltérő helyettesítést választunk az integrálás során, így magával a integrálással jutunk két különböző megoldáshoz. (Az egyik esetben egy $\frac{1}{1+x^2}$ sémájú integrandusból származik az arctg , a másik esetben egy $\frac{1}{1-x^2}$ sémájúból az artgh , ami viszont rögtön logaritmusokra cserélhető.) Az előzőekben bemutatott azonosságok mutatják az átjárást a két felírás között.

Ebből adódik (ez például *Jeffryes* [1948] megközelítése), hogy mindkét formula felírható egyszerű közös alakba a következőképp:

$$f(x) = A(x - c_1)^{m_1} (c_2 - x)^{m_2} .$$

Annak igazolása, hogy a Pearson-eloszlások szélsőértéke maximum

A sűrűségfüggvény második deriváltja

$$\frac{d^2 f}{dx^2} = \frac{d}{dx} \frac{x \cdot f}{\left(b_0 + b_1 x + b_2 x^2\right)} = \frac{f}{\left(b_0 + b_1 x + b_2 x^2\right)} \left(b_0 - b_2 x^2\right),$$

ami az x=0 szélsőértékpontban szükségképp pozitív, hiszen azt az általánosság korlátozása nélkül feltehetjük, hogy a b_0 főegyüttható pozitív. (Ez egy esetleges előjelcserével ugyanis mindenképp elérhető.)

A Pearson-eloszlások paraméterei a momentumokkal kifejezve

Ezek a kifejezések azért kaptak a Mellékletben helyet, mert – ahogy azt a főszövegben is megmutattuk – közvetlenül nincsen jelentőségük a Pearson-eloszlások illesztése során. A keresett együtthatók:

$$\begin{aligned} b_0 &= -\frac{\mu_2\left(4\mu_2\mu_4 - 3\mu_3^2\right)}{A} \\ b_1 &= -\frac{\mu_3\left(\mu_4 + 3\mu_2^2\right)}{A} \\ b_2 &= -\frac{\left(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3\right)}{A}, \end{aligned}$$

ahol $A = 10\mu_4\mu_2 - 18\mu_2^3 - 12\mu_3^2$.

A Pearson IV eloszlás pontos származtatása

Ebben az esetben a levezetés során a sűrűségfüggvényre elsőként kapott formulából érdemes kiindulni, melyet most az egyértelműség kedvéért megismétlünk:

$$f(x) = k \cdot \left(b_0 + b_1 x + b_2 x^2\right)^{\frac{1}{2b_2}} \cdot \exp\left(\operatorname{arctg}\left[\frac{2b_2 x + b_1}{\sqrt{4b_0 b_2 - b_1^2}}\right]\right)^{-\frac{b_1}{b_2\sqrt{4b_0 b_2 - b_1^2}}}.$$

Végezzük el a $\nu = \frac{b_1}{b_2\sqrt{4b_0 b_2 - b_1^2}}$, az $m = \frac{1}{2b_2}$ és az $\alpha = \frac{\sqrt{4b_0 b_2 - b_1^2}}{2b_2}$ egyszerűsítő helyettesítéseket, továbbá helyezzük át a koordinátarendszerünk origóját a (régi x szerinti) $x + \frac{b_1}{2b_2}$ pontba. Ekkor a következő egyszerűbb alakot nyerjük:

$$f(x) = k_1\left[b_2\left(x^2 + \alpha^2\right)\right]^m \cdot \exp\left[\nu \cdot \operatorname{arctg}\left(\frac{x}{\alpha}\right)\right],$$

avagy másik szokásos formájában (a konstans megváltoztatásával):

$$f(x) = k\left(1+\frac{x^2}{\alpha^2}\right)^m \cdot \exp\left[\nu \cdot \operatorname{arctg}\left(\frac{x}{\alpha}\right)\right].$$

A Pearson I és VI eloszlás levezetése során alkalmazott átalakítás

Ha a $b_0 + b_1 x + b_2 x^2$-nek van valós gyöke, úgy célszerűbb a levezetés során a sűrűségfüggvényre másodikként kapott formulát használni, hiszen ez nem fog komplex számokat tartalmazni:

$$f(x) = k \cdot \left(b_0 + b_1 x + b_2 x^2\right)^{\frac{1}{2b_2}} \cdot \left[\left(1+\frac{2b_2 x + b_1}{\sqrt{b_1^2 - 4b_0 b_2}}\right)\left(1-\frac{2b_2 x + b_1}{\sqrt{b_1^2 - 4b_0 b_2}}\right)^{-1}\right]^{\frac{b_1}{2b_2\sqrt{b_1^2-4b_0b_2}}}. \quad /7/$$

Jelölje $b_0 + b_1 x + b_2 x^2$ két gyökét a_1 és a_2 ($a_1 < a_2$):

$$a_{1,2} = \frac{-b_1 \pm \sqrt{b_1^2 - 4b_0 b_2}}{2b_2}.$$

Ezzel a /7/ kifejezés így írható:

$$f(x) = k_1 \cdot \left[b_2 (x-a_1)(x-a_2)\right]^{\frac{1}{2b_2}} \cdot \left\{\left[\frac{2b_2}{\sqrt{b_1^2 - 4b_0 b_2}}(x-a_1)\right]\left[\frac{2b_2}{\sqrt{b_1^2 - 4b_0 b_2}}(a_2-x)\right]^{-1}\right\}^{\frac{b_1}{2b_2\sqrt{b_1^2-4b_0b_2}}}.$$

A konstansok összegyűjtésével:

$$f(x) = k \cdot (x-a_1)^{\frac{\sqrt{b_1^2-4b_0b_2}+b_1}{2b_2\sqrt{b_1^2-4b_0b_2}}} \cdot (a_2-x)^{\frac{\sqrt{b_1^2-4b_0b_2}-b_1}{2b_2\sqrt{b_1^2-4b_0b_2}}}.$$

(Látható, hogy itt a konstans akár komplex tagot is tartalmazhat, de ennek nincs jelentősége, hiszen értékét úgyis később, a normalizációs feltétel alapján határozzuk meg.)

B) A Burr-eloszlás

A Burr XII származtatása az általános formulából

A Burr XII-t akkor kapjuk, ha az általános differenciálegyenletben

$$g(x)=\frac{ck\left(1+x^{c}\right)^{-k-1}}{\left[1-\left(1+x^{c}\right)^{-k}\right]^{2}\cdot\left(\frac{1}{1-\left(1+x^{c}\right)^{-k}}-1\right)},$$

innen ugyanis $G(x)=-\ln\left[\frac{1}{1-\left(1+x^{c}\right)^{-k}}-1\right]$.

A Burr XII sűrűségfüggvényének részletesebb elemzése

A sűrűségfüggvény jellegének megértéséhez írjuk fel az első deriváltját:

$$f'(x)=kcx^{c-2}\left(1+x^{c}\right)^{-k-2}\left(c-x^{c}-kcx^{c}-1\right).$$

Az itt szereplő változókra vonatkozó korlátozások ($x,c,k>0$) miatt az utolsó zárójeles szorzó előtti kifejezés bizonyosan pozitív, így a szélsőérték létezésének szükséges feltétele:

$$\left(c-x^{c}-kcx^{c}-1\right)=0\Rightarrow x=\sqrt[c]{\frac{c-1}{kc+1}}.$$

A kiszámított x pontban a második derivált negatív, így a szélsőérték valóban létezik, és jellegét tekintve maximum. Tartalmilag ez azt jelenti, hogy az x pont az eloszlás módusza lesz.

Az előbbiek azonban csak a $c > 1$ esetben állnak fenn. Ha $c = 1$ akkor formálisan ugyan létezik maximumpont, de a helye épp a 0, ahol a sűrűségfüggvény nem értelmezett, $c < 1$ esetben ráadásul még szélsőérték sem létezik, hiszen az első derivált mindenhol negatív.

Nyers momentumok

Az eloszlás n-edik momentuma:

$$\mu'_n = E\left[X^n\right] = \int_{-\infty}^{+\infty} x^n \cdot f_X(x)\, x,$$

illetve

$$\mu'_n = \int_0^{+\infty} x^n \cdot f_X(x)\, x = \int_0^{+\infty} x^n \cdot kcx^{c-1}\left(1+x^c\right)^{-k-1} x,$$

figyelembe véve, hogy az eloszlás csak a nemnegatív félegyenesen értelmezett.

Az integrálást helyettesítéssel fogjuk elvégezni, mégpedig a

$$t = x^c\left(1+x^c\right)^{-1}$$

helyettesítést alkalmazva. Ebből

$$x = \left(\frac{t}{1-t}\right)^{\frac{1}{c}},$$

ahonnan

$$x = \frac{t^{\frac{1}{c}-1}}{c\left(1-t\right)^{\frac{1}{c}+1}} t.$$

Az integrálás új határai: $t = 0$ (ha $x = 0$) és $t = 1$ (ha $x = \infty$).

Ezek használatával:

$$\begin{aligned}\mu'_n &= \int_0^{+\infty} x^n \cdot kcx^{c-1}\left(1+x^c\right)^{-k-1} x = \\ &= \int_0^1 kc\left(\frac{t}{1-t}\right)^{\frac{n+c-1}{c}}\left[1+\left(\frac{t}{1-t}\right)\right]^{-k-1}\frac{t^{\frac{1}{c}-1}}{c(1-t)^{\frac{1}{c}+1}}t = \\ &= \int_0^1 k\left(\frac{t}{1-t}\right)^{\frac{n+c-1}{c}}\left(\frac{1}{1-t}\right)^{-k-1}\frac{t^{\frac{1}{c}-1}}{(1-t)^{\frac{1}{c}+1}}t = \\ &= k\int_0^1 t^{\frac{n}{c}}(1-t)^{k-\frac{n}{c}-1}t = k\cdot \mathrm{B}\left(\frac{n}{c}+1, k-\frac{n}{c}\right) \equiv k\cdot \mathrm{B}_{c,k}(n),\end{aligned}$$

felhasználva az Euler-féle béta függvényt, és az ez alapján definiált $\mathrm{B}_{c,k}(n)$ rövid jelölést.

Ahogy az ebből is rögtön látható, az eloszlásnak csak akkor létezik véges n-edik momentuma, ha $k-\frac{n}{c}>0 \Rightarrow n<ck$. Például a ferdeség és csúcsosság létezéséhez a $ck>4$ feltétel szükséges.

Centrális momentumok

Induljunk ki a centrális momentum definíciójából (ismét figyelembe véve, hogy $x>0$):

$$\mu_n = \int_0^{\infty}\left(x-\mu'_1\right)^n f_X(x)x.$$

Alkalmazzuk a binomiális tételt a hatványozás kifejtésére:

$$\begin{aligned}\mu_n &= \int_0^{\infty}\left(\sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}x^i\left(\mu'_1\right)^{n-i}\right)\cdot f_X(x)x = \\ &= \sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}\left(\mu'_1\right)^{n-i}\int_0^{\infty}x^i\cdot f_X(x) = \sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}\left(\mu'_1\right)^{n-i}\mu'_i.\end{aligned}$$

Ilyen módon tehát megoldottuk a problémát, hiszen a centrális momentumok számítását visszavezettük a – már ismert – nyers momentumok számítására:

$$\mu_n = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} (\mu_1')^{n-i} \mu_i' = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} \left[k \cdot \mathrm{B}_{c,k}(1)\right]^{n-i} \cdot \left[k \cdot \mathrm{B}_{c,k}(i)\right].$$

Ezt a kifejezést némileg lerövidíthetjük, ha észrevesszük, hogy

$$k \cdot B_{c,k}(i) = k \cdot \frac{\Gamma\left(\frac{i}{c}+1\right)\Gamma\left(k-\frac{i}{c}\right)}{\Gamma(k+1)} = \frac{\Gamma\left(\frac{i}{c}+1\right)\Gamma\left(k-\frac{i}{c}\right)}{\Gamma(k)} \equiv \frac{\lambda_{c,k}(i)}{\Gamma(k)},$$

felhasználva az Euler-féle gamma függvény jól ismert $\Gamma(k+1) = k\Gamma(k)$ rekurzív formuláját (amiből $\frac{k}{\Gamma(k+1)} = \frac{1}{\Gamma(k)}$), illetve bevezetve a

$$\lambda_{c,k}(i) = \Gamma\left(\frac{i}{c}+1\right)\Gamma\left(k-\frac{i}{c}\right) = \mathrm{B}_{c,k}(i) \cdot \Gamma(k+1)$$

rövid jelölést. (Azért nem a $k \cdot B_{c,k}(i)$-ra vezetünk be rövidítést, mert így ki külön tudjuk az i-től nem függő tagot kezelni; $\lambda_{c,k}(i)$ viszont már mindhárom változótól függ.)

Ezekkel a Burr XII általános centrális momentuma:

$$\mu_n = \sum_{i=0}^{n} (-1)^{n-i} \binom{n}{i} \frac{\lambda_{c,k}^{n-i}(1)\lambda_{c,k}(i)}{\Gamma^{n-i+1}(k)}.$$

Speciálisan a második centrális momentum, azaz a szórásnégyzet:

$$\begin{aligned} \mu_2 = \sigma^2 \quad &= \frac{\lambda_{c,k}^2(1)\lambda_{c,k}(0)}{\Gamma^3(k)} - \frac{2\lambda_{c,k}^2(1)}{\Gamma^2(k)} + \frac{\lambda_{c,k}(2)}{\Gamma(k)} = \frac{-\lambda_{c,k}^2(1)}{\Gamma^2(k)} + \frac{\lambda_{c,k}(2)}{\Gamma(k)} = \\ &= \Gamma^{-2}(k)\left[\Gamma(k)\lambda_{c,k}(2) - \lambda_{c,k}^2(1)\right], \end{aligned}$$

figyelembe véve, hogy $\lambda_{c,k}(0) = \Gamma(k)$.

A harmadik centrális momentum (az előző számításokat követve):

$$\begin{aligned}\mu_3 &= -\frac{\lambda_{c,k}^3(1)}{\Gamma^3(k)} + 3\frac{\lambda_{c,k}^3(1)}{\Gamma^3(k)} - 3\frac{\lambda_{c,k}(1)\lambda_{c,k}(2)}{\Gamma^2(k)} + \frac{\lambda_{c,k}(3)}{\Gamma(k)} = \\ &= \Gamma^{-3}\left[2\lambda_{c,k}^3(1) - 3\Gamma(k)\lambda_{c,k}(1)\lambda_{c,k}(2) + \Gamma^2(k)\lambda_{c,k}(3)\right].\end{aligned}$$

Hasonlóképp a negyedik centrális momentum:

$$\begin{aligned}\mu_4 &= \frac{\lambda_{c,k}^4(1)}{\Gamma^4(k)} - 4\frac{\lambda_{c,k}^4(1)}{\Gamma^4(k)} + 6\frac{\lambda_{c,k}^2(1)\lambda_{c,k}(2)}{\Gamma^3(k)} - 4\frac{\lambda_{c,k}(1)\lambda_{c,k}(3)}{\Gamma^2(k)} + \frac{\lambda_{c,k}(4)}{\Gamma(k)} = \\ &= \Gamma^{-4}\left[-3\lambda_{c,k}^4(1) + 6\Gamma(k)\lambda_{c,k}^2(1)\lambda_{c,k}(2) - 4\Gamma^2(k)\lambda_{c,k}(1)\lambda_{c,k}(3) + \Gamma^3(k)\lambda_{c,k}(4)\right].\end{aligned}$$

C) A Johnson-féle S_B eloszlás (centrális) momentumai

A momentumok:

$$\mu_1' = -\omega^{1/2}\sinh\Omega \qquad \mu_2 = \frac{1}{2}(\omega-1)(\omega\cosh 2\Omega + 1)$$

$$\mu_3 = -\frac{1}{4}\omega^{1/2}(\omega-1)^2\left[\omega(\omega+2)\sinh 3\Omega + 3\sinh\Omega\right]$$

$$\mu_4 = \frac{1}{8}(\omega-1)^2\left[\omega^2\left(\omega^4+2\omega^3+3\omega^2-3\right)\cosh 4\Omega + 4\omega^2(\omega+2)\cosh 2\Omega + 3(2\omega+1)\right]$$

ahol $\omega = e^{\delta^{-2}}$, $\Omega = \gamma/\delta$.

Ebből:

$$\gamma_1 = \frac{\omega^{1/2}(\omega-1)^{1/2}\left[\omega(\omega+2)\sinh 3\Omega + 3\sinh\Omega\right]}{\sqrt{2}\left(\omega\cosh 2\Omega + 1\right)^{3/2}}$$

$$\gamma_2 = \frac{\omega^2\left(\omega^4+2\omega^3+3\omega^2-3\right)\cosh 4\Omega + 4\omega^2(\omega+2)\cosh 2\Omega + 3(2\omega+1)}{2\left(\omega\cosh 2\Omega + 1\right)^2}$$

D) A λ-eloszlás

A Tukey-féle λ-eloszlás a kvantilis függvényével (inverz eloszlásfüggvényével) adott:

$$Q(u)=\begin{cases}\dfrac{u^{\lambda}-(1-u)^{\lambda}}{\lambda}, & \text{ha } \lambda\neq 0\\[2ex] \dfrac{\ln(u)}{1-u} & \text{ha } \lambda=0\end{cases}.$$

Az eloszlásnak mindössze egyetlen szabadon állítható paramétere van, így a normálistól több szempont szerint adott módon eltérő eloszlások előállítására nem alkalmas. Ugyanakkor több általánosítás is rendelkezésre áll, amelyek változatosabb eloszlásokat tesznek lehetővé.

A helyzet és a terjedelem kezelését biztosító technikák ismeretében triviális általánosítást adott meg *Ramberg* és *Schmeiser* (*Ramberg–Schmeiser* [1972]):

$$Q(u)=\lambda_1+\lambda_2^{-1}\left[u^{\lambda_3}-(1-u)^{\lambda_3}\right].$$

Az így előálló eloszlások szimmetrikusak lesznek.

E) A Ramberg-Schmeiser-féle általánosított λ-eloszlás centrális momentumai

$$\mu_1 = \lambda_1 + \frac{\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}}{\lambda_2}$$

$$\mu_2 = \frac{\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1} - \left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^2}{\lambda_2^2}$$

$$\mu_3 = \frac{\frac{1}{3\lambda_3+1} - 3\beta(2\lambda_3+1,\lambda_4+1) + 3\beta(\lambda_3+1,2\lambda_4+1) - \frac{1}{3\lambda_4+1} - 3\left(\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1}\right)\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right) + 2\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^3}{\lambda_2^3}$$

$$\mu_4 = \frac{\begin{array}{l}\frac{1}{4\lambda_3+1} - 4\beta(3\lambda_3+1,\lambda_4+1) + 6\beta(2\lambda_3+1,2\lambda_4+1) - 4\beta(\lambda_3+1,3\lambda_4+1) - \frac{1}{4\lambda_4+1} \\ -4\left(\frac{1}{3\lambda_3+1} - 3\beta(2\lambda_3+1,\lambda_4+1) + 3\beta(\lambda_3+1,2\lambda_4+1) + \frac{1}{3\lambda_4+1}\right)\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right) \\ +6\left(\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1}\right)\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^2 - 3\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^4\end{array}}{\lambda_2^4}$$

Ebből a $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$ és a $\gamma_2 = \frac{\mu_4}{\mu_2^2}$ formulák felhasználásával:

$$\gamma_1 = \frac{\frac{1}{3\lambda_3+1} - 3\beta(2\lambda_3+1,\lambda_4+1) + 3\beta(\lambda_3+1,2\lambda_4+1) - \frac{1}{3\lambda_4+1} - 3\left(\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1}\right)\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right) + 2\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^3}{\left[\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1} - \left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^2\right]^{3/2}}$$

és

$$\gamma_2 = \frac{\begin{array}{l}\frac{1}{4\lambda_3+1} - 4\beta(3\lambda_3+1,\lambda_4+1) + 6\beta(2\lambda_3+1,2\lambda_4+1) - 4\beta(\lambda_3+1,3\lambda_4+1) - \frac{1}{4\lambda_4+1} \\ -4\left(\frac{1}{3\lambda_3+1} - 3\beta(2\lambda_3+1,\lambda_4+1) + 3\beta(\lambda_3+1,2\lambda_4+1) + \frac{1}{3\lambda_4+1}\right)\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right) \\ +6\left(\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1}\right)\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^2 - 3\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^4\end{array}}{\left[\frac{1}{2\lambda_3+1} - 2\beta(\lambda_3+1,\lambda_4+1) + \frac{1}{2\lambda_4+1} - \left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right)^2\right]^2}$$

adódik.

F) A g-and-h eloszlás különböző momentumai

$$\mu_1' = \frac{e^{g^2/2-2h} - 1}{g(1-h)^{1/2}}$$

$$\mu_2' = \frac{1 - 2e^{g^2/2-4h} + e^{2g^2/1-2h}}{g^2(1-2h)^{1/2}}$$

$$\mu_3' = \frac{3e^{g^2/2-6h} + e^{9g^2/2-6h} - 3e^{2g^2/1-3h} - 1}{g^3\left(1-3h\right)^{1/2}}$$

$$\mu_4' = \frac{e^{8g^2/1-4h}\left(1 + 6e^{6g^2/4h-1} + e^{8g^2/4h-1} - 4e^{7g^2/8h-2} - 4e^{15g^2/8h-2}\right)}{g^4\left(1-4h\right)^{1/2}}$$

$$\gamma_1 = \frac{\dfrac{3e^{g^2/2-6h} + e^{9g^2/2-6h} - 3e^{2g^2/1-3h} - 1}{\left(1-3h\right)^{1/2}} - 3\dfrac{\left(1 - 2e^{g^2/2-4h} + e^{2g^2/1-2h}\right)\left(e^{g^2/2-2h} - 1\right)}{\left(1-2h\right)^{1/2}\left(1-h\right)^{1/2}} + 2\dfrac{\left(e^{g^2/2-2h} - 1\right)^3}{\left(1-h\right)^{3/2}}}{g^3\left[\dfrac{\left(1 - 2e^{g^2/2-4h}\right) + e^{g^2/1-2h}}{\left(1-2h\right)^{1/2}} + \dfrac{\left(e^{g^2/2-2h} - 1\right)^2}{\dfrac{h-1}{g^2}}\right]^{3/2}}$$

$$\gamma_2 = \frac{\begin{array}{c}\dfrac{e^{8g^2/1-4h}\left(1 + e^{6g^2/4h-1} + e^{8g^2/4h-1} - 4e^{7g^2/8h-2} - 4e^{15g^2/8h-2}\right)}{\left(1-4h\right)^{1/2}} - 4\dfrac{\left(3e^{g^2/2-6h} + e^{9g^2/2-6h} - 3e^{2g^2/1-3h} - 1\right)\left(e^{g^2/2-2h} - 1\right)}{\left(1-3h\right)^{1/2}\left(1-h\right)^{1/2}} - 6\dfrac{\left(e^{g^2/2-2h} - 1\right)^4}{\left(h-1\right)^2} \\ -12\dfrac{\left(1 - 2e^{g^2/4h-2} + e^{2g^2/2h-1}\right)\left(e^{g^2/2-2h} - 1\right)^2}{\left(1-2h\right)^{1/2}\left(h-1\right)} + 3\dfrac{\left(1 - 2e^{g^2/4h-2} + e^{2g^2/2h-1}\right)^2}{2h-1}\end{array}}{\left[\dfrac{\left(1 - 2e^{g^2/2-4h}\right) + e^{2g^2/2h-1}}{\left(2h-1\right)^{1/2}} + \dfrac{\left(e^{g^2/2-2h} - 1\right)^2}{h-1}\right]^2}$$