Eloszláscsaládokhoz való illeszkedés vizsgálata
Ph.D. értekezés

Osztényiné Krauczi Éva
Témavezető:
Dr. Csörgő Sándor
Konzulensek:
Dr. Pap Gyula és Dr. Szűcs Gábor

Matematika- és Számítástudományi Doktori Iskola
Szegedi Tudományegyetem, Bolyai Intézet
Szeged, 2016

Tartalomjegyzék

1. fejezet

Bevezetés

A hipotézisvizsgálat, és ezen belül az illeszkedésvizsgálat fontos területe a matematikai statisztikának. Arra a kérdésre, hogy mikor merült fel az első ilyen típusú probléma az emberiség történetében, a teljes ismeret hiányában nem tudunk teljes bizonyossággal válaszolni. Annyi ismert, hogy 1812-ben Laplace csillagászati vizsgálataiban statisztikai módszert használt annak a hipotézisnek az eldöntésére, hogy a naprendszer üstökösei szerves részei a naprendszernek, vagy csak külső behatolók. Ha csak külső behatolók az üstökösök, akkor pályasíkjuk és az ekliptika közötti szög egyenletes eloszlású kell legyen a $(0,2\pi)$ intervallumon, vagyis egy illeszkedésvizsgálatot kellett elvégeznie.

Az illeszkedésvizsgálat igazi úttörői K. Pearson, E. S. Pearson, A. Fisher és J. Neymann voltak, akik az első eljárásokat dolgozták ki annak a hipotézisnek az eldöntésére, hogy egy véletlen mennyiség eloszlása a minta gyakoriságeloszlása alapján tekinthető-e egy megadott F eloszlással megegyezőnek. Ezt nevezzük egyszerű illeszkedésvizsgálatnak. Később szükség lett olyan eljárásokra is, melyekkel arról a hipotézisről tudtak döntést hozni, hogy a minta egy megadott eloszláscsaládból származik-e. Ezeket az eljárásokat nevezzük összetett illeszkedésvizsgálatnak.

A 2. fejezetben a disszertáció szempontjából fontos történeti előzményeket gyűjtötük össze. Ehhez del Barrio, Cuesta-Albertos és Matrán [33] cikkét használtuk, melyben egy jó összefoglalás található. Mivel a 4. és 5. fejezetekben tárgyalt illeszkedésvizsgálati módszerek, valamint a 3. fejezetben bevezetésre kerülő egyik módszer eloszláscsaládokhoz való illeszkedés ellenőrzésére alkalmasak, illetve alkalmas, így ebben a fejezetben az ezzel kapcsolatos fontosabb eddigi eredmények bemutatása a cél. Az eredmények bemutatása alatt egyrészt a pontos módszer, a tesztstatisztika, másrészt a tesztstatisztika határeloszlásának megadását értjük. Ezen eljárások két nagy osztályát tárgyaljuk részletesen, az egyik a minta eloszlásának és az eloszláscsalád eloszlásainak távolságán alapuló tesztek, a másik a regresszió-, illetve korrelációtesztek. Ennek az az oka, hogy a 4. és 5. fejezetekben lévő tesztek ezekhez az osztályokhoz tartoznak.

A 3. fejezetben egy eljárást javaslunk egyenletes eloszlás esetén egyszerű, illetve összetett illeszkedésvizsgálatra. Az ötlet a következő. Legyenek $U_1, U_2, \ldots, U_n$ független, [0,1] intervallumon egyenletes eloszlású véletlen változók, egy minta. Emellett adott egy determinisztikus $d_n \in (0,1)$ távolságszint minden mintamérethez. A [0,1] intervallumon húzzuk végig ezt a távolságszintet, és figyeljük meg, hogy a rendezett minta elemei hány osztályba esnek. Egy klaszterbe azok az elemei tartoznak a rendezett mintának, amelyekre teljesül az, hogy az egymást követő elemek távolsága nem nagyobb, mint d_n. Egy adott mintához

és távolságszinthez tartozó osztályok számát nevezzük klaszterszámnak. Csörgő S. és Wu [23] három különböző rátával nullához tartó távolságszint sorozat mellett bebizonyították a klaszterek számának aszimptotikus normalitását. Ennek a tételnek bizonyítjuk a többdimenziós változatait különböző intervallumon egyenletes eloszlások esetében, majd használjuk egyenletesség tesztelésére ismert és ismeretlen intervallumon. Bebizonyítjuk a Csörgő–Wu-féle, különböző távolságszintekhez tartozó klaszterszámok együttes aszimptotikus normalitását három esetben: ha a minta a [0,1], ha az ismert $[a, b]$ illetve ha egy ismeretlen intervallumon egyenletes eloszlásból származik. Így ebből adódóan aszimptotikus χ^2-tesztet kapunk egyszerű, illetve összetett nullhipotézis ellenőrzésére. Meghatározzuk a tesztek erejét különböző [0,1] intervallumon folytonos alternatívákkal szemben szimulációval, valamint összehasonlítjuk az új tesztek erejét az Inglot és Ledwina [48] által bevezetett „data driven smooth" teszttel. Ez a fejezet tartalmazza a Krauczi [59] cikk eredményeit.

A 4. fejezetben az L^2-Wasserstein távolságot használó del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] által bevezetett normalitás tesztet vizsgáljuk. Egy eltolás- és skálamentes tesztstatisztikát kaptak, amely egyrészt úgy tesztel normális eloszláscsaládhoz való tartozást, hogy minimális távolságot keres kvantilis-függvények távolságának segítségével; másrészt a tesztstatisztikából látható, hogy korrelációtesztet határoz meg. Ebből a kétféle megközelítésből származik a teszt későbbi elnevezése is, kvantilis korreláció teszt, amely elnevezést Csörgő Sándortól hallottam először. Ennek a normalitástesztnek számos alternatívával szembeni erővizsgálatát végezzük el szimuláció segítségével, valamint összehasonlítjuk más normalitástesztek viselkedésével. Mivel a Wilk–Shapiro-teszttel aszimptotikusan ekvivalens a „spanyolok"[34] tesztje, nem meglepő az erővizsgálat eredménye. Ez a fejezet tartalmazza a Krauczi [52] cikk eredményeit.

Az utolsó, 5. fejezetben Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34], valamint del Barrio, Cuesta-Albertos és Matrán [33] által bevezetett kvantilis korreláció teszt súlyozott változatát vezetjük be logisztikus eloszláscsalád esetében. A súlyfüggvény használatát a tesztstatisztikában egymástól függetlenül de Wet [28, 29] és Csörgő S. [19, 20] különböző motivációból javasolta. Csörgő a súlyfüggvény bevezetésével a tesztstatisztika határeloszlásának létezését remélte több eloszláscsalád esetében, de Wet pedig a normális eloszláscsalád esetében használt tesztstatisztika határeloszlásának végtelen soros előállításában tapasztalt „szabadságifok vesztést" remélte előidézni más eloszláscsaládok esetében is. „Szabadságifok vesztés" alatt azt értjük, hogy a határeloszlás soros előállításában az első kettő tag hiányzik. Mi a Csörgő-féle [20] eredményt a de Wet által, eltolás eloszláscsalád esetére javasolt konkrét súlyfüggvénnyel bizonyítjuk logisztikus eltolás-skála eloszláscsalád esetében. Del Barrio, Cuesta-Albertos és Matrán [33] a tesztstatisztika határeloszlását megadták súlyozott Brown-hidak Karhunen–Loève-sorfejtéseként. Ugyanezen technikával meghatározzuk az általunk kapott határeloszlás soros alakját. Majd ugyancsak egy szimulációs erővizsgálat következik, valamint összehasonlítjuk az új teszt erejét az empirikus karakterisztikus függvényre és az empirikus momentum-generáló függvényre alapozott Meintanis [58] tesztekkel. Ez a fejezet tartalmazza a Balogh és Krauczi [6] cikk eredményeit.

2. fejezet

Történeti előzmények

Ebben a fejezetben áttekintést szeretnénk adni arról, hogy honnan indult az illeszkedésvizsgálat, és milyen fontosabb eljárások ismertek. Ehhez del Barrio, Cuesta-Albertos és Matrán [33] cikkét használjuk, melyben egy jó összefoglalás található.

A következőkben bevezetjük az általunk használt jelöléseket. A nemnegatív egészek halmazát $\mathbb{N}$, a valós számok halmazát $\mathbb{R}$ és a komplex számok halmazát $\mathbb{C}$ jelöli. Minden véletlen változó ugyanazon $(\Omega, \mathcal{A}, P)$ valószínűségi mezőn van definiálva. Jelölje $\mathbf{I}_A$ az A esemény indikátor változóját. Legyenek $X_1, \dots, X_n$ független azonos eloszlású véletlen változók, azaz egy statisztikai minta. Jelölje $F(x),\ x\in\mathbb{R}$, a változók közös eloszlásfüggvényét, és

$$Q_F(t) = F^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\}, \qquad t \in (0,1),$$

az F eloszlásfüggvény kvantilisfüggvényét. Legyen

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2, \quad \text{illetve} \quad m_i = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^i$$

a minta átlaga, szórásnégyzete, illetve i-edik centrális momentuma. Jelölje

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{I}_{\{X_k \leq x\}}, \qquad \text{illetve} \qquad \alpha_{F,n}(x) = \sqrt{n}\big(F_n(x) - F(x)\big), \qquad x \in \mathbb{R},$$

az empirikus eloszlásfüggvényt, illetve az empirikus folyamatot. A rendezett mintára az $X_{1,n}, \dots, X_{n,n}$, a minta kvantilisfüggvényére pedig a $Q_n(t),\ t \in [0,1]$, jelölést használjuk. Vegyük észre, hogy tetszőleges $k = 1,2,\dots,n$ és $t \in ((k-1)/n, k/n]$ esetén $Q_n(t) = X_{k,n}$.

Ha a minta a $[0,1]$ intervallumon egyenletes eloszlásból származik, akkor speciálisan jelölje G_n az empirikus eloszlásfüggvényét. Az egyenletes empirikus folyamatot

$$\alpha_n(t) = \sqrt{n}\big(G_n(t) - t\big), \qquad t \in [0,1],$$

a Brown-hidat $B(t),\ t \in [0,1]$, jelöli. Ez utóbbi egy mintafolytonos, $E(B(t)) = 0$ várható értékű és $\mathrm{Cov}(B(s), B(t)) = \min(s,t) - st,\ s,t \in [0,1]$, kovarianciafüggvényű Gauss-folyamat.

Jelölje Φ a standard normális eloszlásfüggvényt, φ a hozzá tartozó sűrűségfüggvényt jelöli. Legyen minden $\sigma > 0$ és minden $\mu \in \mathbb{R}$ esetén $N_\sigma^\mu(x) = \Phi((x-\mu)/\sigma),\ x \in \mathbb{R}$, a μ várható értékű és σ szórású normális eloszlás eloszlásfüggvénye, valamint használjuk az

$\mathbf{N} = \{N_\sigma^\mu \colon \sigma > 0,\, \mu \in \mathbb{R}\}$ jelölést a normális eloszláscsaládra, vagyis az összes normális eloszlás osztályára. Továbbá jelölje az n-dimenziós, $m \in \mathbb{R}^n$ várható érték vektorú és Σ kovarianciamátrixú normális eloszlást $\mathcal{N}_n(m, \Sigma)$ minden $n \in \mathbb{N}$ esetén.

Két metrikus térre lesz szükségünk. Az egyik a $\mathcal{C}[0,1]$ tér, amely az összes [0,1] intervallumon értelmezett, valós értékű, folytonos függvények halmaza. A $\mathcal{C}[0,1]$ tér az

$$\|x\|_\infty := \sup_{0 \le t \le 1} |x(t)|, \qquad x \in \mathcal{C}[0,1],$$

a szuprémum normával van ellátva, mellyel ez a tér teljes, szeparábilis metrikus tér lesz. A másik a $\mathcal{D}[0,1]$ tér, mely azon [0,1] intervallumon értelmezett, valós értékű függvények halmaza, amelyek jobbról folytonosak és van baloldali határértékük. Ez a tér egy olyan távolsággal van ellátva, melyet Szkorohod vezetett be, és amivel ez is teljes, szeparábilis metrikus tér. Részletes bemutatása megtalálható Billingsley [8] könyvében. A Brown-híd a $\mathcal{C}[0,1]$, az egyenletes empirikus folyamat a $\mathcal{D}[0,1]$ tér véletlen elemének tekinthető.

Az értekezésben minden konvergencia úgy értendő, amint $n \to \infty$. A $\to_{\mathcal{D}}$ az eloszlásban való, a $\to_{\mathbf{P}}$ pedig a sztochasztikus konvergenciát jelöli. Az eloszlásbeli egyenlőséget az $=_{\mathcal{D}}$ jelöli.

2.1. Illeszkedésvizsgálat rögzített eloszlás esetén

Az egyszerű illeszkedésvizsgálat azt jelenti, hogy a minta egy adott, rögzített $F_0(x),\ x \in \mathbb{R}$, eloszlásfüggvényhez való illeszkedését vizsgáljuk. Adott egy $X_1, \ldots, X_n$ véletlen minta egy ismeretlen $F(x),\ x \in \mathbb{R}$, eloszlásfüggvényű véletlen változóból. Teszteljük azt az egyszerű nullhipotézis, hogy

$$\mathcal{H}_0 : F = F_0\,.$$

A *Pearson-féle* χ^2*-teszt*et tekinthetjük az első ilyen illeszkedésvizsgálatnak [61]. Az ötlet a következő: osszuk fel a valós egyenest k db páronként diszjunkt cellára, melyek együtt lefedik az egész valós egyenest. Jelölje $C_1, \ldots, C_k$ ezeket a cellákat, és legyen rendre $p_1, \ldots, p_k$ annak a valószínűsége, hogy a nullhipotézis mellett az X véletlen változó beleesik az egyes cellákba. Vagyis, ha $F{=}F_0$, akkor $P(X_1{\in}C_i){=}p_i,\ i{=}1, \ldots, k$. Legyen $O_i^{(n)}$ az i-edik cellába eső megfigyelések száma. Ekkor $O_i^{(n)}$ binomiális eloszlású n és p_i paraméterekkel. Így a Moivre–Laplace-tétel szerint

$$\frac{O_i^{(n)} - np_i}{\sqrt{np_i(1-p_i)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

A többváltozós centrális határeloszlás-tételből következik, hogy ha $l \le k$, akkor a

$$B_l^{(n)} = \frac{1}{\sqrt{n}} \left(O_1^{(n)} - np_1, \ldots, O_l^{(n)} - np_l \right)^\top$$

véletlen vektornak van határeloszlása. A határeloszlás a nulla várható értékű és $\Sigma_l = {=}(\sigma_{i,j})_{i,j=1,\ldots,l}$ kovarianciamátrixú normális eloszlás, ahol a kovarianciamátrix elemei $\sigma_{i,j} = {=} -p_i p_j\,, i \neq j$ esetén, és $\sigma_{i,i} = p_i(1-p_i)$. Sőt, ha $p_i > 0$ minden $i = 1, \ldots, k$ esetén, akkor

a Σ_{k-1} kovarianciamátrixnak létezik inverze, $\Sigma_{k-1}^{-1} = (\nu_{i,j})_{i,j=1,\dots,k-1}$, melynek elemei $\nu_{i,j} = p_k^{-1}$, $i \neq j$ esetén, és $\nu_{i,i} = p_i^{-1} + p_k^{-1}$. Ekkor könnyen látható, hogy

$$\chi^2(n) := \sum_{j=1}^{k} \frac{(O_j^{(n)} - np_j)^2}{np_j} = {B_{k-1}^{(n)}}^{\top} \Sigma_{k-1}^{-1} B_{k-1}^{(n)} \xrightarrow{\mathcal{D}} \chi^2_{k-1},$$

így kapjuk meg a következő jól ismert aszimptotikus eredményt.

2.1. Tétel. *A nullhipotézis teljesülése mellett $\chi^2(n)$ aszimptotikus eloszlása χ^2_{k-1}.*

A teszt hátránya, hogy nagy szabadságot enged a cellák méretének, helyének és számának megválasztásában. Például nem tud különbséget tenni két különböző eloszlás között, melyek a kiválasztott cellákhoz azonos valószínűséget rendelnek.

Az illeszkedésvizsgálati eljárások következő nagy osztálya az *EDF* (Empirical Distribution Function)-*tesztek*. Ezen tesztek alapötlete az, hogy mérjük meg az F_0 hipotetikus eloszlásfüggvény és a mintából számolt F_n empirikus eloszlásfüggvény távolságát, és ezen eltérés nagysága alapján döntsünk a megegyezésről, illetve különbözőségről. Az egyes tesztek abban különböznek egymástól, hogy hogyan mérjük meg a két függvény távolságát.

Az első ilyen teszt 1928-ból Cramér [14], ennek általánosított változata pedig 1931-ből von Mises [75] névéhez fűződik. A von Mises-féle tesztstatisztika

$$\omega_n^2 := n \int_{-\infty}^{\infty} \big(F_n(x) - F_0(x)\big)^2 w(x) \mathrm{d}x$$

alakban van definiálva, tehát súlyozott L^2-normában méri a két függvény távolságát, ahol w a különbözőséget alkalmasan mérő súlyfüggvény. Speciálisan a Cramér-teszt a $w \equiv 1$ választással adódik. Kolmogorov [51] a szuprémum normát használja, a kétoldali tesztstatisztikája

$$D_n := \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

1933-ból, Szmirnov [69, 70] egyoldali tesztstatisztikái az 1930-as évek végéről

$$D_n^+ := \sqrt{n} \sup_{x \in \mathbb{R}} \big(F_n(x) - F_0(x)\big), \qquad D_n^- := \sqrt{n} \sup_{x \in \mathbb{R}} \big(F_0(x) - F_n(x)\big),$$

melyekre $D_n = \max(D_n^+, D_n^-)$. A három statisztikát együtt *Kolmogorov–Szmirnov-statisztikák*nak nevezik. Ezen statisztikák előnye, hogy eloszlásmentes statisztikák, ugyanis minden folytonos F_0 eloszlásfüggvény esetén, a nullhipotézis mellett

$$D_n \overset{\mathcal{D}}{=} \sup_{0 \le t \le 1} |\alpha_n(t)|, \qquad D_n^+ \overset{\mathcal{D}}{=} \sup_{0 \le t \le 1} \alpha_n(t), \qquad \text{és} \qquad D_n^- \overset{\mathcal{D}}{=} \sup_{0 \le t \le 1} (-1)\alpha_n(t).$$

Így minden folytonos eloszlásfüggvényű eloszlás esetén, adott szignifikanciaszinthez és mintaméretheztugyanaz a kritikus érték tartozik. Ez a tulajdonság nem teljesül az ω_n^2 statisztikára, de a Szmirnov [67, 68] 1936-ban javasolt

$$W_n^2(\Psi) := n \int_{-\infty}^{\infty} \Psi\big(F_0(x)\big) \big(F_n(x) - F_0(x)\big)^2 \mathrm{d}F_0(x)$$

változatára már igen, ahol $\Psi(t)$, $0 \leq t \leq 1$, nemnegatív súlyfüggvény. Az összes ilyen statisztikát, amit Ψ változtatásával kapunk, *Cramér–von Mises-típusú statisztikának* nevezünk. A különböző súlyfüggvények használata lehetőséget ad különböző alternatívák felismerésére, éppen ezért a Kolmogorov-statisztikának is bevezették a súlyozott változatát:

$$K_n(\Psi) := \sqrt{n} \sup_{x \in \mathbb{R}} \frac{|F_n(x) - F_0(x)|}{\Psi\big(F_0(x)\big)}.$$

Bár ez se tudta kompenzálni azt a hiányát a szuprémum normának, hogy csak a legnagyobb elterést érzékeli F_n és F_0 között, amíg az L^2-norma ezen két függvény súlyozott átlagos távolságát méri. Ezen heurisztikus megfigyelést a szimuláció is alátámasztja (lásd 4. fejezet, ahol azt tapasztaltuk a normális eloszláscsaládhoz való illeszkedésvizsgálat esetében, hogy a Kolmogorov-tesztnek a legtöbb alternatívával szembeni ereje jóval kisebb, mint más próbák ereje).

Két statisztika különös figyelmet kapott az irodalomban. A $\Psi \equiv 1$ esetben,

$$W_n^2 := n \int_{-\infty}^{\infty} \big(F_n(x) - F_0(x)\big)^2 \mathrm{d}F_0(x)$$

a *Cramér–von Mises-statisztika*; valamint a $\Psi(t) = (t(1-t))^{-1}$, $t \in (0,1)$, mellett

$$A_n^2 := n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(t)(1 - F_0(t))} \mathrm{d}F_0(x)$$

az *Anderson–Darling-statisztika* [4], mely utóbbi a szimulációs vizsgálatok alapján a legerősebb ilyen típusú tesztnek tűnik (lásd például Stephens [71] cikkben, valamint a 4.5. táblázatban a 4.2.2. fejezetben).

Ahhoz, hogy használni tudjuk a gyakorlatban ezeket a teszteket, ismernünk kell az eloszlásfüggvényüket tetszőleges $n \in \mathbb{N}$ esetén, vagy legalább az aszimptotikus eloszlásukat. 1941-ben Szmirnov [70] explicit formában meg tudta adni D_n^+ eloszlásfüggvényét tetszőleges n esetén, Kolmogorov [51] pedig megadott egy rekurzív kifejezést 1933-ban, amivel kiszámítható a $P(D_n < x)$ valószínűség tetszőleges $n \in \mathbb{N}$ és $x \in \mathbb{R}$ esetén. A Cramér–von Mises-típusú statisztikák eloszlásfüggvényének a meghatározása már nagyobb nehézséget okozott. Akkoriban Monte-Carlo szimuláció hiányában fontos kérdés volt, hogy ki tudják-e számolni a kritikus értékeket rögzített $n \in \mathbb{N}$ esetén. Emellett a határeloszlás kérdése elméleti, de gyakorlati szempontból is érdekes volt. Az első aszimptotikus eredményt is a Kolmogorov–Szmirnov-típusú statisztikákra sikerült megkapni:

2.2. Tétel. *Minden $x > 0$ esetén*
(Kolmogorov 1933, [51])

$$\lim_{n \to \infty} P(D_n \leq x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2},$$

(Szmirnov 1941, [70])

$$\lim_{n \to \infty} P(D_n^+ > x) = \lim_{n \to \infty} P(D_n^- < x) = e^{-2x^2}.$$

1948-ban Feller [39] megjegyezte, hogy Kolmogorov és Szmirnov teljesen különböző módszerrel bizonyították állításaikat, és megpróbálta egységesíteni a bizonyításukat. Mivel a D_n, D_n^+ és W_n^2 statisztikák az F_n empirikus és az F_0 elméleti eloszlásfüggvények eltérését mérik, vagyis az $\alpha_{F,n}$ empirikus folyamat funkcionáljai, ezért ezen statisztikák $\mathcal{H}_0$ melletti határeloszlásait valamimilyen közös technikával lehetne származtatni. Így Feller cikke fontos lépés az empirikus folyamatra épített illeszkedésvizsgálat aszimptotikus elméletének egységesítésében. Bár ekkor még magát az empirikus folyamatot és annak a határeloszlását nem vizsgálták.

1949-ben Doob [36] a véges dimenziós eloszlásokat vizsgálva sejtette meg az egyenletes empirikus folyamatnak a Brown-hídhoz való konvergenciáját, de bizonyítani nem tudta. Viszont bizonyította, hogy minden $x > 0$ esetén

$$P\left(\sup_{0\le t\le 1} |B(t)| \le x\right) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}$$

és

$$P\left(\sup_{0\le t\le 1} B(t) > x\right) = e^{-2x^2},$$

vagyis az egyenletes empirikus folyamat abszolút szuprémum és szuprémum funkcionáljainak határeloszlása megegyezik a Brown-híd ugyanezen funkcionáljainak eloszlásával. Ez azt jelenti, hogy ha Doob sejtése igaz, akkor Kolmogorov és Szmirnov eredményeire talán egyszerűbb bizonyítás is adható. 1951-ben Donsker [35] invariancia elve által nyert bizonyítást a sejtés. Az invariancia elv a következőt jelenti. A részletösszeg folyamat minden folytonos funkcionáljának eloszlása konvergál a Brown-mozgás megfelelő funkcionáljának eloszlásához, illetve az egyenletes empirikus folyamat minden folytonos funkcionáljának eloszlása konvergál a Brown-híd megfelelő funkcionáljának eloszlásához.

Ezen eredmények hatására fejlődött ki a metrikus terekben való gyenge konvergencia elmélete többek között Kolmogorovnak, Prohorovnak és Szkorohodnak köszönhetően, amely elmélet segített jobban megérteni az invariancia elvet. Erről szól Billingsley [8] 1968-as könyve. Fontos lépés volt, hogy kidolgozták az elméletet a $\mathcal{C}[0,1]$ és a $\mathcal{D}[0,1]$ tereken. Először a részletösszeg és az empirikus folyamatokat lineáris interpolációval kapott folytonos folyamatokkal közelítették, hogy ne kelljen a $\mathcal{C}[0,1]$ térből kilépniünk. Ezen új folyamat sorozatokra bizonyították a véges dimenziós eloszlások konvergenciáját és a sorozat feszességét, amely kettő tulajdonság együtt a folyamatok eloszlásbeli konvergenciáját adja. A folytonos folyamatokkal való közelítés valahogy mesterkélt. Ahhoz, hogy ezt el tudjuk kerülni, egy gazdagabb téren kell dolgoznunk. Ez a gazdagabb tér a $\mathcal{D}[0,1]$ tér, amelynek már maga az empirikus folyamat is eleme.

2.3. Tétel. *Az* $\alpha_n \xrightarrow{\mathcal{D}} B$ *konvergencia teljesül a* $\mathcal{D}[0,1]$ *téren.*

A 2.3. Tétel lehetővé teszi a 2.2. Tétel természetesebb bizonyítását. Be lehet látni, hogy az $x \mapsto \|x\|_\infty$ leképezés folytonos a Szkorohod-topológiára nézve egy B eloszlása szerint nulla mértékű halmazt kivéve, és mivel $D_n = \|\alpha_n\|_\infty$, ekkor $D_n \xrightarrow{\mathcal{D}} \|B\|_\infty$. Hasonló konvergencia teljesül a D_n^+ és a D_n^- statisztikák esetében.

A 2.3. Tétel teszi lehetővé a Cramér–von Mises-statisztika határeloszlásának meghatározását is. Az $x \mapsto \int_0^1 x^2(t)\mathrm{d}t$ funkcionál szintén folytonos a Szkorohod-topológiára

nézve egy B eloszlása szerint nulla mértékű halmazt kivéve. Így a fenti érvelés ismételt alkalmazásával kapjuk, hogy

$$W_n^2 \xrightarrow{\mathcal{D}} \int_0^1 \big(B(t)\big)^2 \,\mathrm{d}t\,.$$

Innen pedig egy lépés a Cramér–von Mises-típusú statisztikák határeloszlása. Mint a Brown-hidakra vonatkozó iterált logaritmus tétel következményeként Anderson és Darling [4] 1952-ben megmutatta, hogy feltéve az

$$\int_0^\delta \Psi(t)t\log\log\frac{1}{t}\mathrm{d}t \qquad \text{és} \qquad \int_\delta^1 \Psi(t)(1-t)\log\log\frac{1}{1-t}\mathrm{d}t$$

integrálok végességét valamilyen $\delta\in(0,1)$ esetén teljesül a

$$W_n^2(\Psi) \xrightarrow{\mathcal{D}} \int_0^1 \Psi(t)\big(B(t)\big)^2 \,\mathrm{d}t \tag{2.1}$$

konvergencia. Ez az állítás az invarienciaelv alkalmazásával is bizonyítható, ugyanis az $x\mapsto\int_0^1\Psi(t)x^2(t)\,\mathrm{d}t$ funkcionál folytonos a Szkorohod-topológiára nézve egy B eloszlása szerint nulla mértékű halmazt kivéve. A (2.1) konvergencia az Anderson–Darling-féle súlyfüggvény esetén is teljesül, tehát

$$A_n^2 \xrightarrow{\mathcal{D}} \int_0^1 \frac{\big(B(t)\big)^2}{t(1-t)}\,\mathrm{d}t.$$

2.2. Illeszkedésvizsgálat eloszláscsalád esetén

Ebben a fejezetben azokat a teszteket tekintjük, ahol a kérdés az, hogy a minta egy adott eloszláscsaládból származik-e. Itt legyen $\mathcal{F}$ eloszlásfüggvények egy parametrikus eloszláscsaládja, azaz

$$\mathcal{F}=\{F(\cdot,\theta):\theta\in\Theta\}\,,$$

ahol Θ valamilyen nyitott paraméterhalmaz $\mathbb{R}^d$-ben.

Az első vizsgálatok az 1930-as években a normális eloszláscsalád esetében történtek. Fisher [41], Pearson [61] és Williams [79] voltak az elsők, akik a $\sqrt{\beta_1(n)}=m_3(n)/m_2^{2/3}(n)$ és $\beta_2(n)=m_4(n)/m_2^2(n)$ standardizált harmadik és negyedik momentumok segítségével mérték meg a normalitástól való eltérést. 1977-ben Pearson, D'Agostino és Bowman [60] a $\sqrt{\beta_1(n)}$ és $\beta_2(n)$ két alkalmas függvényét használta erre. Ezekkel a tesztekkel az a probléma, hogy a lapultsági és a ferdeségi mutató kevés, hogy karakterizálja a normális eloszlást, emiatt ezen tesztek ereje kicsi bizonyos alternatívákkal szemben. Ezek a tesztek akkor is elfogadják a nullhipotézist, ha a minta ugyan nemnormális eloszlásból származik, de szimmetrikus és a lapultsági mutatója szintén 3, mint normális eloszlásé. Másrészt a gyakorlati alkalmazások szempontjából az is fontos lenne, hogy ha egy eloszlás csak nagyon kicsit különbözik a normális eloszlástól, akkor a teszt azt ne vesse el. Ugyancsak 1977-ben Ali [3] adott eloszlásoknak egy olyan sorozatát, amely ugyan eloszlásban tart a standard normális eloszláshoz, de a lapultsági mutatója felrobban. Vagyis, ha a sorozat elég nagy indexű tagjából származik a mintánk, akkor nagy eséllyel ezek a tesztek elutasítják, pedig valójában közel normális eloszlásról van szó.

Más típusú normalitásteszt például 1954-ből az

$$u_n := \frac{X_{n,n} - X_{1,n}}{\left(\frac{n}{n-1}\right)^{\frac{1}{2}} m_2^{\frac{1}{2}}(n)}$$

statisztika (David, Hartley és Pearson [27]), ami a terjedelem és a szórás, valamint 1947-ből az

$$a_n := \frac{\sum_{j=1}^{n} |X_j - \bar{X}_n|}{n \cdot m_2^{\frac{1}{2}}(n)}$$

statisztika (Geary [43]), ami a mintaátlagtól való átlagos abszolút eltérés és a szórás hányadosából származtatott teszt. Ezek a tesztek csak egyes alternatívákkal szemben viselkednek jól, de kicsi erővel bírnak alternatívák széles skálájával szemben.

A következő alfejezetben azokat a teszteket mutatjuk be, amelyeket rögzített eloszláshoz való illeszkedéstesztek átdolgozásaként kapunk.

2.2.1. Eloszláscsalád tesztelése rögzített eloszláshoz való illeszkedésvizsgálat segítségével

A 2.1. fejezetben rögzített eloszláshoz való illeszkedés teszteket tekintettünk. Egy lehetőség, hogy eloszláscsaládhoz való illeszkedést teszteljünk ezekkel a tesztekkel, ha a θ paraméternek a $\mathcal{H}_0$ mellett egy $\hat{\theta}_n$ becslését véve azt ellenőrizzük, hogy a minta $F(x, \hat{\theta}_n)$, $x \in \mathbb{R}$, eloszlásfüggvényű-e. Ezt javasolta Pearson a χ^2-tesztje esetében. Legyen

$$\hat{\chi}^2(n) := \sum_{j=1}^{k} \frac{(O_j^{(n)} - np_j(\hat{\theta}_n))^2}{np_j(\hat{\theta}_n)},$$

ahol $p_j(\theta)$ annak a valószínűsége, hogy X_1 a j-edik cellába esik $F(x,\theta)$, $x \in \mathbb{R}$, mellett. Pearson nem tudta megadni $\hat{\chi}^2(n)$ aszimptotikus eloszlását. Fisher volt az, aki rámutatott arra, hogy a határeloszlás függ a paraméter becslésének módszerétől, és megmutatta, hogy a szokásos feltételek mellett, ha a θ maximum likelihood becslését vesszük a csoportosított $(O_1^{(n)}, \ldots, O_k^{(n)})$ adatokon, akkor a $\hat{\chi}^2(n)$ statisztikának χ^2_{k-d-1} a határeloszlása (lásd Cochran [13] 1952-ből).

Fisher azt is megfigyelte, hogy a csoportosított $(O_1^{(n)}, \ldots, O_k^{(n)})$ mintából származó $\hat{\theta}_n$ becslésből adódó információvesztés erőcsökkenést eredményez. Ezért Fisher abban az esetben is megvizsgálta $\hat{\chi}^2(n)$ határeloszlását, amikor a θ paraméter egydimenziós, és a teljes mintából vesszük a θ paraméter maximum likelihood becslését. Az eredményét 1954-ben Chernoff és Lehmann [12] d-dimenziós paraméterre általánosította, nevezetesen, hogy megfelelő feltételek mellett

$$\hat{\chi}^2(n) \xrightarrow{\mathcal{D}} \sum_{j=1}^{k-d-1} Z_j^2 + \sum_{j=k-d}^{k-1} \lambda_j Z_j^2, \tag{2.2}$$

ahol Z_j független standard normális változók, és $\lambda_j \in [0,1]$, $j = k-d, \ldots, k-1$, olyan konstansok, amelyek függhetnek a θ paraméter igazi értékétől. Ez a függés mutatja az egyik nagy hátrányát a $\hat{\chi}^2$-teszt használatának eloszláscsalád esetében.

A másik nehézség a $\hat{\chi}^2$ tipusú teszt használatában a cellák választása. Az $O_i^{(n)}$ cellagyakoriságok aszimptotikus normalitásának a következménye a Pearson-féle statisztika aszimptotikus χ^2_{k-1}-eloszlása. Viszont egy kicsi várható gyakorisággal rendelkező cella esetében az $O_i^{(n)}$ változó nagyon lassan konvergál a normális eloszláshoz, ami azt eredményezi, hogy a (2.2) konvergencia lassú. Vagyis az asszimptotikus kritikus értékek használatának létjogosultsága sérülne ebben az esetben. A gyakorlatban ezt úgy próbálják meg elkerülni, hogy „olyan cellákat használnak, amelyekbe legalább 10 megfigyelés esik" (lásd Cochran [13]).

A cellák jó választására nézve 1940-es években Mann és Wald [57] valamint Gumbel [45] azt javasolták rögzített eloszlás esetén, hogy a nullhipotézis mellett azonos valószínűségű cellákat használjunk, ezáltal csökkentve a cellák választásának esetlegességét. Ez a gondolat paraméteres eloszláscsalád esetére úgy vihető át, hogy először vegyük valamilyen alkalmas becslését θ-nak, majd $F(x,\hat{\theta}_n)$, $x\in\mathbb{R}$, mellett azonos valószínűségű cellákat használjunk. Vagyis megint véletlenül fogunk cellákat választani! Ugyanúgy a minta határozza meg, hogy melyik cellákat használjuk, mint amikor olyan cellákat választunk, amelyekbe legalább 10 megfigyelés esik. 1957-ben Watson [76, 77] megmutatta, ha $\hat{\theta}_n$ a teljes mintából származó maximum likelihood becslése θ-nak, valamint a j-edik cella végpontjai $F^{-1}((j-1)/k,\hat{\theta}_n)$ és $F^{-1}(j/k,\hat{\theta}_n)$, akkor (2.2) teljesül. Továbbá, ha $\mathcal{F}$ eltolás-skála család, akkor a λ_j együtthatók nem függnek a θ paramétertől, csak az eloszláscsaládtól.

Az EDF-tesztek adaptációja eloszláscsaládok esetére könnyen kivitelezhető, és hasonlóan a rögzített eloszlás esetére, ezek a tesztek jobb erővel bírnak, mint a $\hat{\chi}^2$-tesztek. Legyen $\hat{\theta}_n$ valamilyen becslése θ-nak. Ekkor a megfelelő becsléses statisztikák

$$\widehat{W}_n^2(\Psi) := n\int_{-\infty}^{\infty}\Psi\left(F(x,\hat{\theta}_n)\right)\left(F_n(x)-F(x,\hat{\theta}_n)\right)^2\mathrm{d}F(x,\hat{\theta}_n)$$

és

$$\hat{K}_n(\Psi) := \sqrt{n}\sup_{x\in\mathbb{R}}\frac{|F_n(x)-F(x,\hat{\theta}_n)|}{\Psi\left(F(x,\hat{\theta}_n)\right)}.$$

A $\Psi\equiv 1$ esetben a két statisztikát a $\widehat{W}_n^2$ és $\hat{K}_n$ jelöli. A kívánatos eloszlásmentesség, ami a rögzített esetben teljesült, itt sajnos nem igaz. Legyen $Z_i^{(n)}=F(X_i,\hat{\theta}_n),\ i=1,\dots,n$, és $\hat{G}_n(t),\ t\in[0,1]$, jelölje a $Z_1^{(n)},\dots,Z_n^{(n)}$ változókhoz tartozó empirikus eloszlásfüggvényt. Ekkor

$$\widehat{W}_n^2(\Psi)=n\int_0^1\Psi(t)(\hat{G}_n(t)-t)^2\mathrm{d}t$$

és

$$\hat{K}_n(\Psi)=\sqrt{n}\sup_{0<t<1}\frac{|\hat{G}_n(t)-t|}{\Psi(t)}.$$

Tehát a két statisztika értéke csak a $\hat{G}_n$ függvénytől függ. Viszont $Z_1^{(n)},\dots,Z_n^{(n)}$ nem független, azonosan egyenletes eloszlású véletlen változók, ami azt eredményezi, hogy a $\hat{G}_n$ függvény funkcionáljainak eloszlására nem alkalmazhatók az eddigiek. Éppen ezért $\hat{G}_n$ nem olyan, amivel klasszikus értelemben tudunk dolgozni. Számos fontos esetben $Z_1^{(n)},\dots,Z_n^{(n)}$ eloszlása nem függ a θ paramétertől, csak az eloszláscsaládtól, vagyis ekkor $\widehat{W}_n^2(\Psi)$ és $\hat{K}_n^2(\Psi)$ paramétermentes. Ez történik az eltolás-skála családok esetében,

amikor olyan $\hat{\theta}_n$ becslést használunk, amiben a becslés felcserélhető a skálázással, illetve az eltolással (lásd David és Johnson [26] 1948-ból). 1967-ben Lilliefors [56] ezt használta fel és készítette el a népszerű táblázatát a normális eloszláscsalád esetére a Kolmogorov–Szmirnov-statisztikához.

A becsléses $\widehat{W}_n^2(\Psi)$ és $\hat{K}_n^2(\Psi)$ típusú statisztikák határeloszlásának a meghatározására tett első kísérlet Darling [25] nevéhez fűződik 1955-ből. A becsléses Cramér–von Mises-statisztika aszimptotikus eloszlását tudta meghatározni abban az esetben, amikor a θ paraméter egydimenziós. 1972-ben Sukhatme [72] kiterjesztette Darling eredményét többdimenziós paraméterekre. Ezekben a cikkekben egy segédfolyamaton keresztül találták meg $\widehat{W}_n^2$ határeloszlását.

1955-ben viszont Kac, Kiefer és Wolfowitz [49] közvetlenül az

$$\hat{\alpha}_n(t) = \sqrt{n}(\hat{G}_n(t) - t), \qquad t \in [0,1],$$

becsléses empirikus folyamatot tanulmányozva kapták meg $\widehat{W}_n^2$ határeloszlását normális eloszláscsalád esetén a maximum likelihood paraméterbecslésekkel: $\hat{\theta}_n = (\hat{X}_n, S_n^2)$. Ugyan a becsléses empirikus folyamatnak a gyenge konvergenciáját nem bizonyították, de megmutatták, hogy

$$\hat{W}_n^2 \xrightarrow{\mathcal{D}} \int_0^1 (Z(t))^2 \, \mathrm{d}t \,,$$

ahol $Z(t)$, $t \in (0,1)$, egy 0 várható értékű és

$$K(s,t) = \min(s,t) - st - \varphi\big(\Phi^{-1}(s)\big)\varphi\big(\Phi^{-1}(t)\big) - \frac{1}{2}\Phi^{-1}(s)\varphi\big(\Phi^{-1}(s)\big)\Phi^{-1}(t)\varphi\big(\Phi^{-1}(t)\big)$$

kovarianciafüggvényű Gauss-folyamat.

A becsléses empirikus folyamat gyenge konvergenciájának általános vizsgálata Durbin [37] nevéhez fűződik 1973-ból. Az eloszláscsaládra és a paraméterre tett megfelelő regularitási feltételek mellett az $\hat{\alpha}_n$ empirikus folymat gyengén konvergál a 0 várható értékű és $K(s,t)$, $s,t \in [0,1]$, kovarianciafüggvényű Gauss folyamathoz. Durbin cikkében explicit formulát adott a $K(s,t)$ kovarianciafüggvényre, és standard számolással megmutatható, hogy ennek speciális esete a Kac, Kiefer és Wolfowitz által megadott kovariancia.

Megjegyezzük, hogy Burke, Csörgő M., Csörgő S. és Révész [10] 1979-es cikkéből következik Durbin eredménye. Ebben a cikkben a becsléses empirikus folyamatot Gauss folyamatok sorozatával közelítik. Azon túl, hogy Durbin tételéből következik a $\hat{W}_n^2(\Psi)$ és $\hat{K}_n^2(\Psi)$ típusú statisztikák nullhipotézis melletti eloszlásbeli konvergenciája, a [10] cikk eredménye az aszimptotikus erők tanulmányozásának is eszköze lehet.

Az empirikus folyamatot tanulmányozó elmélet fejlődésének következményeként további illeszkedést vizsgáló technikák jelentek meg az 1980-as években. Például Feuerverger és Mureika [40], valamint Csörgő S. [15] az empirikus karakterisztikus függvény aszimptotikus eloszlását vizsgálták. A Durbin-tétel analóg változatát empirikus karakterisztikus és kvantilis függvényekre Csörgő S. [16] és LaRiccia és Mason [53] dolgozták ki. Ezen eredmények segítségével új normalitástesztek születtek, melyek közül Murota és Takeuchi Hall és Wels [47], Epps és Pulley [38] valamint Csörgő S. [17, 18] eredményeit említjük meg.

Egy másik ötlet, hogy hogyan tudjuk a rögzített eloszlás esetében használt tesztelési eljárást parametrikus eloszláscsalád esetében használni, a *minimum távolság módszere*.

Legyen δ egy metrika az eloszlásfüggvények halmazán. Ekkor $\Delta(F_n,\mathcal{F})=\inf_\theta \delta(F_n,F(\cdot\,,\theta))$ egy lehetséges mértéke az empirikus eloszlásfüggvény $\mathcal{F}$ parametrikus eloszláscsaládtól való távolságának. Pollard [62] 1980-ban használta ezt először és meghatározta $\Delta(F_n,\mathcal{F})$ határeloszlását, tetszőleges normált lineáris tér értékű véletlen változók esetében.

2.2.2. Regresszió- és korrelációtesztek

Ebben a fejezetben tegyük fel, hogy $\mathcal{F}$ eltolás-skála család, vagyis adott egy H_0 standardizált (0 várható értékű és 1 szórású) eloszlásfüggvény, és az eloszláscsalád többi tagja lineáris transzformációval kapható belőle.

Az ötlet a következő. Legyen $X_1,\dots,X_n$ az $\mathcal{F}$ eloszláscsaládból származó μ várható értékű és σ^2 szórásnégyzetű minta. A korábbi jelöléseknek megfelelően legyen $\mathbf{X}_n^\top = (X_{1,n},\dots,X_{n,n})$ a mintához tartozó rendezett minta. Tekintsünk továbbá egy n elemű mintát H_0 eloszlásfüggvénnyel, és legyen $\mathbf{Z}_n^\top=(Z_{1,n},\dots,Z_{n,n})$ a kapcsolatos rendezett minta. Jelölje $\mathbf{m}_n^\top=(m_{1,n},\dots,m_{n,n})$ illetve $\mathbf{V}_n$ a $\mathbf{Z}_n$ vektor várható érték vektorát illetve kovarianciamátrixát. Könnyen látszik, hogy

$$X_{i,n}\stackrel{\mathcal{D}}{=}\mu+\sigma Z_{i,n},\qquad i=1,\dots,n\,. \tag{2.3}$$

Ha kétdimenziós koordinátarendszerben ábrázoljuk az $(m_{i,n},X_{i,n})$, $i=1,\dots,n$ pontokat, akkor ezeknek közelítőleg egy egyenesre kell esniük, és a linearitás hiánya azt sugallja, hogy X_1 eloszlásfüggvénye nem $\mathcal{F}$-beli. Gyakran ezt csak „szemre" ellenőrzik, de vannak analitikus eljárások is ennek az ellenőrzésére. Két nagy osztálya van ezeknek az eljárásoknak: az egyik a *regresszió*-, a másik a *korrelációtesztek*, mely különböző eljárások valójában ekvivalens tesztekre vezetnek.

Az első esetben a (2.3) lineáris model segítségével adunk egy $\hat\sigma_n^2$ becslést a σ^2 szórásnégyzetre, és ezt hasonlítjuk össze az S_n^2 becsléssel. Ekkor a nullhipotézis mellett a $\hat\sigma_n^2/S_n^2$ tesztstatisztika értéke közel kell legyen 1-hez, ellenkező esetben elvetjük a nullhipotézist. Ezeket az eljárásokat nevezik *regressziótesztek*nek. A másik osztálya ezen teszteknek a ρ korrelációs együttható segítségével ellenőrzi, van-e lineáris kapcsolat az $\mathbf{X}_n$ véletlen vektor és az $\mathbf{m}_n$ determinisztikus vektor között a következőképpen:

$$\rho^2(\mathbf{m}_n,\mathbf{X}_n)=\frac{\left(n\cdot\mathbf{m}_n{}^\top\mathbf{X}_n-1^\top\mathbf{m}_n\cdot 1^\top\mathbf{X}_n\right)^2}{\left(n\cdot\mathbf{m}_n{}^\top\mathbf{m}_n-(1^\top\mathbf{m}_n)^2\right)\left(n\cdot\mathbf{X}_n^\top\mathbf{X}_n-(1^\top\mathbf{X}_n)^2\right)},$$

ahol $1^\top=(1,\dots,1)\in\mathbb{R}^n$. Ekkor a nullhipotézis mellett a $\rho^2(\mathbf{m}_n,\mathbf{X}_n)$ tesztstatisztika értéke közel kell legyen 1-hez, ellenkező esetben elvetjük a nullhipotézist. Ezeket az eljárásokat nevezik *korrelációtesztek*nek.

A regressziótesztek első változata 1965-ből *Wilk és Shapiro* [65] *W normalitásteszt*je. A μ és σ paraméterek legjobb lineáris torzítatlan becslése a (2.3) model alapján az általánosított legkisebb négyzetek módszerével, illetve a szimmetrikus eloszlásokra teljesülő $1^\top\mathbf{V}_n^{-1}\mathbf{m}_n=0$ összefüggés alkalmazásával

$$\hat\mu_n=\bar X_n\qquad\text{és}\qquad\hat\sigma_n=\frac{\mathbf{m}_n{}^\top\mathbf{V}_n^{-1}\mathbf{X}_n}{\mathbf{m}_n^\top\mathbf{V}_n^{-1}\mathbf{m}_n}.$$

Wilk és Shapiro a W tesztstatisztikát a $\hat{\sigma}_n^2/S_n^2$ tesztstatisztika normalizált változataként definiálta

$$W_n := \frac{(\mathbf{m}_n{}^\top \mathbf{V}_n^{-1}\mathbf{X}_n)^2}{\mathbf{m}_n{}^\top \mathbf{V}_n^{-1}\mathbf{V}_n^{-1}\mathbf{m}_n \sum_i (X_i - \bar{X})^2} \tag{2.4}$$

alakban. Ezzel egy regressziótesztet kaptak. Másrészt ez egy korrelációteszt is, ami a normalizációból következik, ugyanis $W_n = \rho^2(\mathbf{V}_n^{-1}\mathbf{m}_n, \mathbf{X}_n)$. Shapiro, Wilk és Chen [63] szimulációs vizsgálatából kiderült, hogy a W-teszt egyike a legerősebb normalitásteszteknek alternatívák széles skálájával szemben. Ezért népszerű módszer a mai napig, annak ellenére, hogy rejteget egy-két nehézséget a használata.

Egyik probléma, hogy magát a W_n tesztstatisztikát bonyolult kiszámítani. Ahhoz, hogy W_n-t meg tudjuk határozni, előzetesen ki kell számolnunk az $\mathbf{m}_n$ vektort és a $\mathbf{V}_n^{-1}$ mátrixot. Ez a mintaméret növekedésével egyre nehezebb feladat, és valójában amikor W_n-et bevezették, legfeljebb 20 elemű minta esetén tudták megadni a $\mathbf{V}_n^{-1}$ mátrix elemeit pontosan. Ezért már Wilk és Shapiro is numerikus közelítéssel számolta W_n értékeit 50-es mintaméretig. Egy másik probléma, hogy az $n = 3$ esetet kivéve nem ismerjük W_n eloszlásfüggvényét. Mivel az $n = 3$ esetben a W-teszt megegyezik az u_n-teszttel, ekkor W_n pontos eloszlása is ismert. Wilk és Shapiro $n = 50$ mintaméretig szimulációval adták meg a kritikus értékeket. A határeloszlás viszont 1986-ig ismeretlen volt, amikor is Leslie, Stephens és Fotopoulos [55] megmutatták a W-teszt aszimptotikus ekvivalenciáját egy másik korrelációteszttel, amely teszt határeloszlása akkor már ismert volt.

Ezek a problémák a W-teszt módosításaihoz vezettek. Az első példányai ezeknek a próbálkozásoknak a *D'Agostino* [24] 1971-ből és a *Shapiro–Francia-korrelációtesztek* [64] 1972-ből, melyek használatát 50-nél nagyobb elemű minták esetén javasolták. A D'Agostino-tesztstatisztika a

$$D_n := \frac{\sum_{i=1}^n (i - \frac{n+1}{2}) X_{i,n}}{n^2 S_n},$$

és a Shapiro–Francia-tesztstatisztika pedig a

$$W'_n := \frac{(\mathbf{m}_n^\top \mathbf{X}_n)^2}{\mathbf{m}_n^\top \mathbf{m}_n \sum_i (X_i - \bar{X})^2}$$

formulával van definiálva. Mindkét cikk szimulációs tanulmánya azt sugallta, hogy ezen tesztek aszimptotikusan ekvivalensek a W-teszttel.

A W'_n további egyszerűsítését javasolta Weisberg és Bingham [78] 1975-ben. Az $\mathbf{m}_n$ vektort helyettesítsük az $\tilde{\mathbf{m}}_n = (\tilde{m}_{1,n}, \ldots, \tilde{m}_{n,n})$ vektorral, ahol

$$\tilde{m}_{i,n} = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right), \quad i = 1, \ldots, n.$$

Ez a statisztika még könnyebben számolható, mint W'_n, valamint Weisberg és Bingham empirikus vizsgálata szerint aszimptotikusan ekvivalens a W_n statisztikával.

A következő fontos változata a W-tesztnek *de Wet és Venter [30] korrelációtesztje* 1972-ből. Az ő tesztstatisztikájuk

$$W_n^* := \sum_{i=1}^n \left(\frac{X_{i,n} - \bar{X}_n}{S_n} - \Phi^{-1}\left(\frac{i}{n+1}\right)\right)^2.$$

Azon túl, hogy ők vezették be a korrelációteszt fogalmát, ez volt az első olyan típusú normalitásteszt, amely határeloszlását is sikerült meghatározni. De Wet és Venter megmutatták, hogy ha $Z_1, Z_2, \ldots$ független, standard normális véletlen változók sorozata, akkor

$$2n(1-W_n^{*1/2}) - \frac{1}{n+1}\sum_{i=1}^{n}\frac{i}{n+1}\left(1-\frac{i}{n+1}\right)\left(\varphi\left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right)\right)^{-2} + \frac{3}{2} \xrightarrow{\mathcal{D}} \sum_{i=3}^{\infty}\frac{Z_i^2-1}{i}\,.$$

Ezzel a tétellel megnyílt a lehetőség arra, hogy más korreláció normalitástesztek határeloszlását megkaphatjuk a W^*-teszttel való aszimptotikus ekvivalencia által. Fontos lépés volt ebben a programban 1987-ből Verril és Johnson [74] eredménye, ahol megmutatták a korrelációtesztek bizonyos általános feltételek melletti aszimptotikus ekvivalenciáját. Így vált világossá, hogy a Shapiro–Francia- és a Weisberg–Bingham-tesztek határeloszlása megegyezik a de Wet–Venter-teszt határeloszlásával. Továbbá a Wilk–Shapiro- és Shapiro–Francia-tesztek aszimptotikus ekvivalenciájából következett a kiindulási W-teszt határeloszlásának ismerete.

3. fejezet

Illeszkedésvizsgálat egyenletes eloszlás esetében

3.1. Együttes klaszterszámok aszimptotikus viselkedése

Legyenek $U_1, U_2 \ldots$ független, a [0,1] intervallumon egyenletes eloszlású véletlen változók, valamint bármely $n \in \mathbb{N}$ esetén legyen $U_{1,n}, \ldots, U_{n,n}$ az $U_1, \ldots, U_n$ mintához tartozó rendezett minta. A minta elemei majdnem biztosan különböznek egymástól, így az $U_{1,n} < < \cdots < U_{n,n}$ reláció majdnem biztosan érvényes. Adott, determinisztikus $d_n \in (0,1)$ távolságszint mellett definiálható egy $\mathcal{G}_n = \mathcal{G}(U_1, \ldots, U_n; d_n)$ véletlen intervallumgráf. A $\mathcal{G}_n$ gráf csúcshalmaza az $U_1, \ldots, U_n$ elemeket reprezentáló $\{1, \ldots, n\}$ halmaz. Két különböző i és j csúcs között akkor és csak akkor van él, ha $|U_i - U_j| < d_n$, ahol $i, j \in \{1, \ldots, n\}$. A mintához tartozó klasztereket úgy definiáljuk, mint ezen mintához tartozó gráf összefüggő komponensei. A K_n klaszterszám a gráf összefüggő komponenseinek a számát jelöli.

Godehardt és Jaworski [44] tanulmányozta az előbb definiált véletlen intervallumgráfot, és sikerült meghatározniuk a K_n eloszlását minden n-re. A klaszterek számának pontos eloszlása mellett természetesen vetődött fel a kérdés, hogy van-e határeloszlása a K_n sorozatnak. Ahhoz, hogy ne degenerált eloszlást kapjunk, a továbbiakban tegyük fel, hogy $d_n \to 0$. Godehardt ésJaworski [44] megmutatták, ha $n^2 d_n \to 0$, akkor $n - K_n \to 0$ majdnem biztosan, vagyis, ha d_n elég gyorsan konvergál nullához, akkor 1 valószínűséggel létezik olyan n_0 (véletlentől függő) küszöbszám, hogy bármely $n \leq n_0$ esetén nincs él a $\mathcal{G}_n$ gráfban. További d_n sorozatok esetében tanulmányozták az adott méretű klaszterek számának az aszimptotikus eloszlását és az $U_1, \ldots, U_n$ minta egy adott elemét tartalmazó klaszter méretének határeloszlását. Sajnos általánosságban nem mondtak semmit K_n viselkedéséről. Csörgő és Wu [23] nem a véletlen gráfos reprezentációt használva három különböző aszimptotikus viselkedésű távolságszint sorozat mellett bebizonyították a klaszterek számának aszimptotikus normalitását. A módszerükkel, amit mi is alkalmazni fogunk, még rátát is adtak az eloszlásfüggvények konvergenciájának sebességére. A következő tételben az ő eredményüket fogalmazzuk meg.

3.1. Tétel (Csörgő és Wu [23]). *(i) Ha* $nd_n \to 0$ *és* $n^2 d_n \to \infty$, *akkor*

$$\Delta_n := \sup_{x\in\mathbb{R}} \left| P\left(\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}(1-e^{-nd_n})}} \le x \right) - \Phi(x) \right| = O\left(\sqrt{\left(nd_n + \sqrt{\frac{4\log n}{n}} \right) \log\frac{1}{nd_n} + \frac{\log\left(n\sqrt{d_n}\right)}{n\sqrt{d_n}}} \right).$$

Ennélfogva

$$\frac{K_n - ne^{-nd_n}}{n\sqrt{d_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

(ii) Ha $0 < \liminf_n nd_n \le \limsup_n nd_n < \infty$, *akkor*

$$\sup_{x\in\mathbb{R}} \left| P\left(\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-2nd_n}(e^{nd_n} - 1 - n^2 d_n^2)}} \le x \right) - \Phi(x) \right| = O\left(\frac{\log^{3/4} n}{n^{1/4}} \right).$$

Ebből következik, hogy ha $nd_n \to c \in (0,\infty)$, *akkor*

$$\frac{K_n - ne^{-nd_n}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, e^{-2c}[e^c - 1 - c^2]).$$

(iii) Ha $nd_n \to \infty$ *és* $ne^{-nd_n} \to \infty$, *akkor*

$$\Delta_n = O\left(\frac{(nd_n)^{3/2}}{\sqrt{e^{nd_n}}} + \sqrt{\varepsilon_n nd_n \log(ne^{-nd_n})} + \sqrt{\frac{e^{nd_n}}{n}} \log(ne^{-nd_n}) \right),$$

ahol Δ_n *ugyanazt a szuprémumot jelöli, mint az (i) esetben, valamint* $\varepsilon_n = \sqrt{(4\log n)/n}$. *És így*

$$\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

A következőkben ennek a tételnek bizonyítjuk a többdimenziós változatait különböző intervallumokon egyenletes eloszlások esetében, majd használjuk egyenletesség tesztelésére ismert és ismeretlen intervallumon.

3.2. Elméleti eredmények

3.2.1. A [0,1] intervallumon egyenletes eloszlásból származó klaszterszámok együttes aszimptotikus viselkedése

Csörgő és Wu [23] megmutatták K_n aszimptotikus normalitását három különböző aszimptotikus viselkedésű távolságszint sorozat mellett. Célunk, hogy ugyanezen távolságszintekhez tartozó klaszterszámok együttes viselkedését megvizsgáljuk.

Tekintsünk $J \geq 1$ darab $d_{n1} \leq d_{n2} \leq \ldots \leq d_{nJ}$, $n \in \mathbb{N}$, távolságszint sorozatot. A $K_{nj}(d_{nj})$ jelölje a d_{nj} távolságszinthez tartozó klaszterek számát minden n és j esetén. Tekintsük a

$$\mathbf{K}_n = \frac{1}{\sqrt{n}} \left(\frac{K_{n1}(d_{n1}) - m_{n1}}{\sigma_{n1}}, \ldots, \frac{K_{nJ}(d_{nJ}) - m_{nJ}}{\sigma_{nJ}} \right)^\top \tag{3.1}$$

a véletlen vektorváltozót az $m_{nj} = ne^{-nd_{nj}}$ és

$$\sigma_{nj} = \sqrt{e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)}, \qquad n \in \mathbb{N}, \qquad j = 1, \ldots, J, \tag{3.2}$$

centralizáló és normalizáló sorozattal. Ekkor a következő határeloszlástételt állíthatjuk.

3.2. Tétel. *Tegyük fel, hogy a $d_{n1} \leq d_{n2} \leq \ldots \leq d_{nJ}$, $n \in \mathbb{N}$, távolságszint sorozatok mindegyike kielégíti az alábbi feltételek valamelyikét:*

(T1) $nd_{nj} \to 0$, $n^2 d_{nj} \to \infty$;

(T2) $0 < \liminf_n nd_{nj} \leq \limsup_n nd_{nj} < \infty$;

(T3) $nd_{nj} \to \infty$, $ne^{-nd_{nj}} \to \infty$.

Továbbá, tegyük fel, hogy

$$s_{ij} := \lim_{n\to\infty} \frac{e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}} - 1 - n^2 d_{ni} d_{nj})}{\sigma_{ni}\sigma_{nj}} \in \mathbb{R}, \qquad 1 \leq i < j \leq J, \tag{3.3}$$

és legyen $s_{jj} := 1$ és $s_{ji} := s_{ij}$. Ekkor

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma), \tag{3.4}$$

a $\Sigma = (s_{ij})_{i,j=1,\ldots,J}$ kovarianciamátrixszal.

Megjegyezzük, hogy a Σ kovarianciamátrix lehet szinguláris is. Ebben az esetben a normális határeloszlás az $\mathbb{R}^J$ térnek egy lineáris alterére koncentrált.

A 3.2. Tétel bizonyítása előtt kimondunk egy állítást, melyet használni fogunk a 3.2. Tétel bizonyításában.

3.3. Állítás. *Legyen $J \geq 1$ természetes szám és $g_{nj} : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, J$, $n \in \mathbb{N}$, mérhető függvényeknek egy rendszere. Tegyük fel, hogy Y_r, $r \in \mathbb{N}$, független azonos eloszlású véletlen változóknak egy olyan sorozata, hogy $E(g_{nj}(Y_r)) = 0$, $s_{jj} := E\left(g_{nj}^2(Y_r)\right) = 1$ minden n, j és r esetén. Továbbá, tegyük fel, hogy minden $i \neq j$ és r esetén*

$$s_{ij} := \lim_{n\to\infty} E\big(g_{ni}(Y_r) g_{nj}(Y_r)\big) \in \mathbb{R}, \tag{3.5}$$

és

$$E\left(|g_{nj}(Y_r)|^3\right) = o(\sqrt{n}). \tag{3.6}$$

Ekkor az $\mathbb{R}^J$ értékű $Z_{nr} = (g_{n1}(Y_r), \ldots, g_{nJ}(Y_r))$, $r = 1, \ldots, n$, $n \in \mathbb{N}$, véletlen vektorokból álló szériasorozatra teljesül az, hogy

$$\frac{Z_{n1} + \cdots + Z_{nn}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma),$$

ahol $\Sigma = (s_{ij})_{i,j=1,\ldots,J}$.

Bizonyítás. Ezt a többdimenziós határeloszlástételt a Cramér–Wold-lemma segítségével bizonyítjuk. Ehhez legyen $c = (c_1, \dots, c_J)^\top \in \mathbb{R}^J$ rögzített, tetszőleges vektor. Ekkor be kell látnunk, hogy

$$c^\top \frac{Z_{n1} + \dots + Z_{nn}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^\top \Sigma c). \tag{3.7}$$

Legyen Σ_n a Z_n vektorváltozó kovarianciamátrixa, ami egy pozitív szemidefinit mátrix. A feltevések szerint $\Sigma = \lim_{n\to\infty} \Sigma_n$, amiből következik, hogy Σ is pozitív szemidefinit. Ez azt jelenti, hogy $c^\top \Sigma c \geq 0$ minden $c \in \mathbb{R}^J$ esetén. Vegyük észre továbbá, hogy

$$D^2\left(c^\top \frac{Z_{n1} + \dots + Z_{nn}}{\sqrt{n}}\right) = \frac{D^2\big(c^\top Z_{n1}\big) + \dots + D^2\big(c^\top Z_{nn}\big)}{n} = \frac{n c^\top \Sigma_n c}{n} = c^\top \Sigma_n c$$

Tekintsük először azt az esetet, amikor $c^\top \Sigma c = 0$. Ekkor a Csebisev-egyenlőtlenség alkalmazásával tetszőleges $\varepsilon > 0$ esetén.

$$P\left(\left|c^\top \frac{Z_{n1} + \dots + Z_{nn}}{\sqrt{n}}\right| > \varepsilon\right) \leq \frac{c^\top \Sigma_n c}{\varepsilon^2} \to \frac{c^\top \Sigma c}{\varepsilon^2} = 0.$$

Ez azt jelenti, hogy

$$c^\top \frac{Z_{n1} + \dots + Z_{nn}}{\sqrt{n}} \xrightarrow{\mathbf{P}} 0,$$

amiből következik, hogy a konvergencia eloszlásban is teljesül. Mivel $c^\top \Sigma c = 0$ esetén $\mathcal{N}(0, c^\top \Sigma c) = 0$ majdnem biztosan, a (3.7) konvergencia ebben az esetben bizonyított.

A (3.7) konvergenciát a $c^\top \Sigma c > 0$ esetben a Ljapunov-tétel segítségével mutatjuk meg. Jegyezzük meg, hogy

$$c^\top \Sigma_n c \to c^\top \Sigma c = \sum_{j=1}^{J} c_j^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{J} c_i c_j s_{ij} > 0,$$

továbbá a jobb oldali kvadratikus alak folytonos az s_{ij} komponensekben. Ebből következik, hogy létezik n_0 küszöbszám és $\varepsilon > 0$, hogy $n \geq n_0$ esetén

$$c^\top \Sigma_n c \geq \sum_{j=1}^{J} c_j^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{J} c_i c_j (s_{ij} - \varepsilon) > 0.$$

Jelölje K az egyenlőtlenségrendszer középső kifejezését. Ekkor

$$s_n^2 = \sum_{r=1}^{n} D^2\big(c^\top Z_{nr}\big) = n c^\top \Sigma_n c \geq nK > 0.$$

Másrészt az L^3-normára vonatkozó háromszög-egyenlőtlenség miatt

$$\sqrt[3]{E\left(|c^\top Z_{nr}|^3\right)} = \|c^\top Z_{nr}\|_{L^3} \leq |c_1| \|g_{n1}(Y_r)\|_{L^3} + \dots + |c_J| \|g_{nJ}(Y_r)\|_{L^3} = o(n^{\frac{1}{6}}) \sum_{j=1}^{J} |c_j|.$$

Mivel $Y_1, Y_2, \ldots$ azonos eloszlásúak, ezért

$$\sup_{1\le r\le n} E\left(|c^\top Z_{nr}|^3\right) \le o(\sqrt{n}) \left(\sum_{j=1}^{J} |c_j|\right)^3 .$$

Ekkor a $c^\top Z_{nr}$, $r=1,\ldots,n$, $n\in\mathbb{N}$, szériasorozat kielégíti a Ljapunov-feltételt $\delta=1$ választással, ugyanis

$$\frac{\sum_{r=1}^n E\left(\left|c^\top Z_{nr} - E(c^\top Z_{nr})\right|^{2+\delta}\right)}{s_n^{2+\delta}} = \frac{\sum_{r=1}^n E\left(\left|c^\top Z_{nr}\right|^3\right)}{s_n^3} \le \frac{n\cdot o(\sqrt{n})\left(\sum_{j=1}^J |c_j|\right)^3}{n^{\frac{3}{2}} K^{\frac{3}{2}}} \to 0.$$

Ennélfogva a Ljapunov-féle centrális határeloszlástételből következik, hogy

$$c^\top \frac{Z_{n1}+\cdots+Z_{nn}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^\top \Sigma c).$$

Így a Cramér–Wold-lemmából következik a bizonyítandó állítás. □

A 3.2. Tétel bizonyítása. A 3.2. Tétel bizonyítása a 3.3. Állításból és a Csörgő és Wu [23] cikk 2.2. fejezetében bemutatott érvelésből jön. Legyenek $Y_1, Y_2, \ldots$ független exponenciális eloszlású véletlen változók $\lambda=1$ paraméterrel, és jelölje $S_m := Y_1+\cdots+Y_m$, $m\in\mathbb{N}$, a kapcsolatos részletösszegeket. Legyen továbbá $F(x)=1-e^{-x}$, $x>0$, a változók közös eloszlásfüggvénye, és jelölje $F_n(x)=\frac{1}{n}\sum_{m=1}^n \mathbf{I}_{\{Y_m\le x\}}$, $x\in\mathbb{R}$, az empirikus eloszlásfüggvényt. Az ismert

$$(U_{1,n},\ldots,U_{n,n}) \stackrel{\mathcal{D}}{=} \left(\frac{S_1}{S_{n+1}},\ldots,\frac{S_n}{S_{n+1}}\right) \tag{3.8}$$

eloszlásbeli egyenlőségből következik, hogy az $U_1,\ldots,U_n$ mintához és d_{nj} távolságszint sorozathoz tartozó $\mathcal{G}_n$ véletlen intervallumgráf azonos eloszlású a

$$\mathcal{G}\left(\frac{S_1}{S_{n+1}},\ldots,\frac{S_n}{S_{n+1}}; d_{nj}\right) = \mathcal{G}(S_1,\ldots,S_n; d_{nj}S_{n+1})$$

véletlen intervallumgráffal. Ez azt jelenti, hogy az összefüggő komponensek száma számolható azon helyek számából, ahol az egymás utáni $S_1,\ldots,S_n$ részletösszegek az adott távolságszintnél nagyobb értékkel különböznek egymástól. Tehát

$$\begin{aligned} K_{nj}(d_{nj}) &\stackrel{\mathcal{D}}{=} 1+\sum_{m=1}^{n-1} \mathbf{I}_{\{Y_{m+1}>d_{nj}S_{n+1}\}} = 1+\left[(n-1)-\sum_{m=1}^{n-1}\mathbf{I}_{\{Y_{m+1}\le d_{nj}S_{n+1}\}}\right] \\ &\stackrel{\mathcal{D}}{=} n-\sum_{m=1}^{n-1}\mathbf{I}_{\{Y_m\le d_{nj}S_{n+1}\}} = n-(n-1)F_{n-1}(d_{nj}S_{n+1}), \end{aligned}$$

tetszőleges $j=1,\ldots,J$ és $n=2,3,\ldots$ esetén. Mivel $F(d_{nj}S_{n+1})=1-e^{-d_{nj}S_{n+1}}$, azt kapjuk, hogy minden j és n esetén

$$\begin{aligned} K_{nj}(d_{nj}) - ne^{-nd_{nj}} &\stackrel{\mathcal{D}}{=} n(1-e^{-nd_{nj}}) - (n-1)F_{n-1}(d_{nj}S_{n+1}) + n[F(d_{nj}S_{n+1}) - (1-e^{-d_{nj}S_{n+1}})] \\ &= n[e^{-d_{nj}S_{n+1}} - e^{-nd_{nj}}] - (n-1)[F_{n-1}(d_{nj}S_{n+1}) - F(d_{nj}S_{n+1})] + F(d_{nj}S_{n+1}). \end{aligned}$$

Mint a Csörgő és Wu [23] cikkben lévő (2.8) és (2.15) felbontásokban, itt is felbonthatók a normált klaszterszámok a konvergencia szempontjából egy fő- és három maradéktagra. Legyen most $j = 1, \dots, J$ és $n = 2, 3, \dots$ esetén a főtag és a maradéktagok

$$
\begin{aligned}
M_{nj} &:= \frac{nd_{nj}e^{-nd_{nj}}(n - S_n) - n[F_n(nd_{nj}) - F(nd_{nj})]}{\sqrt{n}\sigma_{nj}}, \\
R_{nj}^{(1)} &:= \frac{ne^{-nd_{nj}}[e^{nd_{nj} - d_{nj}S_{n+1}} - 1 - (nd_{nj} - d_{nj}S_n)]}{\sqrt{n}\sigma_{nj}}, \\
R_{nj}^{(2)} &:= \frac{(n-1)\left([F_{n-1}(nd_{nj}) - F(nd_{nj})] - [F_{n-1}(d_{nj}S_{n+1}) - F(d_{nj}S_{n+1})]\right)}{\sqrt{n}\sigma_{nj}}, \\
R_{nj}^{(3)} &:= \frac{F(d_{nj}S_{n+1}) - F(nd_{nj}) + \mathbf{I}_{\{Y_n \le nd_{nj}\}}}{\sqrt{n}\sigma_{nj}}.
\end{aligned}
$$

Ekkor az

$$
nF_n(nd_{nj}) = (n-1)F_{n-1}(nd_{nj}) + \mathbf{I}_{\{Y_n \le nd_{nj}\}}
$$

azonosság alkalmazásával algebrailag ellenőrizhető, hogy

$$
\frac{K_{nj}(d_{nj}) - ne^{-nd_{nj}}}{\sqrt{n}\sigma_{nj}} = M_{nj} + R_{nj}^{(1)} + R_{nj}^{(2)} + R_{nj}^{(3)}. \tag{3.9}
$$

A továbbiakban megmutatjuk, hogy az $R_{nj}^{(1)}, R_{nj}^{(2)}, R_{nj}^{(3)}$ maradéktagok sztochasztikusan konvergálnak nullához, majd ezek után meghatározzuk az M_{nj} határeloszlását. A $J = 1$ esetben Csörgő és Wu [23] adott a (3.9) felbontáshoz hasonló reprezentációt, és a távolságszint sorozatra vonatkozó különböző feltételek mellett megmutatták a maradéktagok sztochasztikus konvergenciáját. Szerencsére az általunk bevezetett $R_{nj}^{(1)}, R_{nj}^{(2)}$ és $R_{nj}^{(3)}$ tagok algebrailag kifejezhetők a Csörgő és Wu által definiált maradéktagokból, és ilyen módon az $R_{nj}^{(1)}, R_{nj}^{(2)}$ és $R_{nj}^{(3)}$ maradéktagok konvergenciáját bizonyítani tudjuk.

3.4. Állítás. *A 3.2. Tétel feltételei mellett az $R_{nj}^{(1)}$, $R_{nj}^{(2)}$ és $R_{nj}^{(3)}$ maradéktagok sztochasztikusan nullához konvergálnak.*

Bizonyítás. Rögzítsük j értékét, és legyen $\tilde{\sigma}_{nj} = \sqrt{e^{-nd_{nj}}(1 - e^{-nd_{nj}})}$. Csörgő és Wu [23] a 2.1. Tétel bizonyításában megmutatta, hogy ha a d_{nj} távolságszint sorozat teljesíti a (T1) vagy (T3) feltételt, akkor

$$
\begin{aligned}
\tilde{R}_{nj}^{(1)} &:= \frac{ne^{-nd_{nj}}[e^{nd_{nj} - d_{nj}S_{n+1}} - 1]}{\sqrt{n}\tilde{\sigma}_{nj}} \xrightarrow{\mathbf{P}} 0, \\
\tilde{R}_{nj}^{(2)} &:= \frac{(n-1)\left([F_{n-1}(nd_{nj}) - F(nd_{nj}] - [F_{n-1}(d_{nj}S_{n+1}) - F(d_{nj}S_{n+1})]\right)}{\sqrt{n}\tilde{\sigma}_{nj}} \xrightarrow{\mathbf{P}} 0, \\
\tilde{R}_{nj}^{(3)} &:= \frac{F(d_{nj}S_{n+1})}{\sqrt{n}\tilde{\sigma}_{nj}} \xrightarrow{\mathbf{P}} 0.
\end{aligned}
$$

Algebrailag ellenőrizhető, hogy

$$R_{nj}^{(1)} = \tilde{R}_{nj}^{(1)} \frac{\tilde{\sigma}_{nj}}{\sigma_{nj}} - \frac{nd_{nj}}{\sqrt{n}\sigma_{nj}} \left(1 - \frac{S_n}{n}\right),$$
$$R_{nj}^{(2)} = \tilde{R}_{nj}^{(2)} \frac{\tilde{\sigma}_{nj}}{\sigma_{nj}},$$
$$R_{nj}^{(3)} = \tilde{R}_{nj}^{(3)} \frac{\tilde{\sigma}_{nj}}{\sigma_{nj}} - \frac{F(nd_{nj}) - \mathbf{I}_{\{Y_n \le nd_{nj}\}}}{\sqrt{n}\sigma_{nj}}.$$

Először azt mutatjuk meg, hogy $\tilde{\sigma}_{nj}/\sigma_{nj} \to 1$ és $\sigma_{nj}/d_{nj} \to \infty$. Abban az esetben, ha a távolságszint sorozat a (T1) feltételt elégíti ki, akkor

$$\frac{\tilde{\sigma}_{nj}^2}{\sigma_{nj}^2} = \frac{e^{-nd_{nj}}(1 - e^{-nd_{nj}})}{e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)} = \frac{e^{nd_{nj}} - 1}{e^{nd_{nj}} - 1 - n^2 d_{nj}^2} = h_1(nd_{nj}),$$

ahol $h_1(x) = (e^x - 1)/(e^x - 1 - x^2),\ x > 0$. A L'Hospital-szabály alkalmazásával látható, hogy

$$\lim_{x \to 0} h_1(x) = \lim_{x \to 0} \frac{e^x - 1}{e^x - 1 - x^2} = \lim_{x \to 0} \frac{e^x}{e^x - 2x} = 1,$$

vagyis ebben az esetben $\tilde{\sigma}_{nj}/\sigma_{nj} \to 1$. Hasonló módon

$$\frac{\sigma_{nj}^2}{d_{nj}^2} = \frac{e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)}{d_{nj}^2} = \frac{e^{-nd_{nj}} - e^{-2nd_{nj}} - n^2 d_{nj}^2 e^{-2nd_{nj}}}{nd_{nj}} \frac{n}{d_{nj}} = h_2(nd_{nj}) \frac{n}{d_{nj}},$$

ahol $h_2(x) = (e^{-x} - e^{-2x} - x^2 e^{-2x})/x,\ x > 0$. Szintén a L'Hospital-szabály alkalmazásával látható, hogy

$$\lim_{x \to 0} h_2(x) = \lim_{x \to 0} \frac{e^{-x} - e^{-2x} - x^2 e^{-2x}}{x} = \lim_{x \to 0} e^{-x}(-1) - e^{-2x}(-2) - 2xe^{-2x} - x^2 e^{-2x}(-2) = 1,$$

valamint $n/d_{nj} \to \infty$, ami együtt mutatja, hogy a σ_{nj}/d_{nj} sorozat divergens.

Amennyiben a távolságszint sorozat a (T3) feltételt elégíti ki, akkor a tényezők megfelelő egyszerűsítésével illetve csoportosításával

$$\frac{\tilde{\sigma}_{nj}}{\sigma_{nj}} = \frac{\sqrt{e^{-nd_{nj}}(1 - e^{-nd_{nj}})}}{\sqrt{e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)}} = \sqrt{\frac{e^{nd_{nj}} - 1}{e^{nd_{nj}} - 1 - n^2 d_{nj}^2}} = \sqrt{\frac{1 - e^{-nd_{nj}}}{1 - e^{-nd_{nj}} - n^2 d_{nj}^2 e^{-nd_{nj}}}} \to 1,$$

illetve

$$\frac{\sigma_{nj}}{d_{nj}} = \frac{\sqrt{e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)}}{d_{nj}} = ne^{-nd_{nj}} \sqrt{\frac{e^{nd_{nj}} - 1 - n^2 d_{nj}^2}{n^2 {d_{nj}}^2}} \to \infty.$$

Az eddigi eredményekből azonnal következik $R_{nj}^{(2)} \xrightarrow{\mathbf{P}} 0$.

Az $R_{nj}^{(1)}$ maradéktagban lévő $\sqrt{n}d_{nj}\left(1 - S_n/n\right)/\sigma_{nj}$ tag sztochasztikus nullához tartása a Berry–Esseen-tétel segítségével látható mindkét típusú távolságszint sorozat esetén.

Tetszőleges $\varepsilon > 0$ esetén

$$\begin{aligned} P\left(\left|\frac{d_{nj}}{\sigma_{nj}}\sqrt{n}\left(1-\frac{S_n}{n}\right)\right| > \varepsilon\right) &= P\left(\left|\sqrt{n}\left(1-\frac{S_n}{n}\right)\right| > \varepsilon\frac{\sigma_{nj}}{d_{nj}}\right) \\ &= G_n\Big(-\varepsilon\frac{\sigma_{nj}}{d_{nj}}\Big) + 1 - G_n\Big(\varepsilon\frac{\sigma_{nj}}{d_{nj}}\Big) \\ &= \Phi\Big(-\varepsilon\frac{\sigma_{nj}}{d_{nj}}\Big) + O\left(\frac{1}{\sqrt{n}}\right) + 1 - \Phi\Big(\varepsilon\frac{\sigma_{nj}}{d_{nj}}\Big) + O\left(\frac{1}{\sqrt{n}}\right) \\ &= 2\left(1-\Phi\Big(\varepsilon\frac{\sigma_{nj}}{d_{nj}}\Big)\right) + O\left(\frac{1}{\sqrt{n}}\right) \to 0, \end{aligned}$$

ahol G_n jelölje $\sqrt{n}(1-S_n/n)$ eloszlásfüggvényét. Ebből következik, hogy $R_{nj}^{(1)}$ konvergál nullához.

A $R_{nj}^{(3)}$ maradéktagban található $(F(nd_{nj}) - \mathbf{I}_{\{Y_n \le nd_{nj}\}})/(\sqrt{n}\sigma_{nj})$ mennyiség sztochasztikus viselkedése a Csebisev-egyenlőtlenségből következik. Vegyük észre, hogy most $E(\mathbf{I}_{\{Y_n \le nd_{nj}\}}) = F(nd_{nj})$ és

$$\begin{aligned} D^2(\mathbf{I}_{\{Y_n \le nd_{nj}\}}) &= E(\mathbf{I}_{\{Y_n \le nd_{nj}\}}) - E^2(\mathbf{I}_{\{Y_n \le nd_{nj}\}}) = F(nd_{nj}) - F^2(nd_{nj}) \\ &= (1-e^{-nd_{nj}})e^{-nd_{nj}} = \tilde{\sigma}_{nj}^2. \end{aligned} \tag{3.10}$$

Azt kapjuk, hogy

$$P\left(\left|\frac{F(nd_{nj}) - \mathbf{I}_{\{Y_n \le nd_{nj}\}}}{\sqrt{n}\sigma_{nj}}\right| > \varepsilon\right) \le \frac{D^2\big(\mathbf{I}_{\{Y_n \le nd_{nj}\}}\big)}{\varepsilon^2 n\sigma_{nj}^2} = \frac{\tilde{\sigma}_{nj}^2}{\varepsilon^2 n\sigma_{nj}^2} = \frac{1}{\varepsilon^2 n}\left(\frac{\tilde{\sigma}_{nj}}{\sigma_{nj}}\right)^2 \to 0$$

mindkét típusú távolságszint sorozat és minden $\varepsilon > 0$ esetén. Ezzel sikerült megmutatnunk, hogy $R_{nj}^{(3)}$ is konvergál nullához.

Amennyiben a távolságszint sorozat a (T2) feltételt teljesíti, akkor Csörgő és Wu azt is megmutatta, hogy

$$\begin{aligned} \overline{R}_{nj}^{(1)} &:= \frac{ne^{-nd_{nj}}[e^{nd_{nj}-d_{nj}S_{n+1}} - 1 - (nd_{nj} - d_{nj}S_{n+1})]}{\sqrt{n}\tilde{\sigma}_{nj}} \xrightarrow{\mathbf{P}} 0, \\ \overline{R}_{nj}^{(2)} &:= \frac{(n-1)\left([F_{n-1}(nd_{nj}) - F(nd_{nj}] - [F_{n-1}(d_{nj}S_{n+1}) - F(d_{nj}S_{n+1})]\right)}{\sqrt{n}\tilde{\sigma}_{nj}} \xrightarrow{\mathbf{P}} 0, \\ \overline{R}_{nj}^{(3)} &:= \frac{F(d_{nj}S_{n+1}) - nd_{nj}e^{-nd_{nj}} + \sum_{j=n}^{n+1}[\mathbf{I}_{\{Y_n \le nd_{nj}\}} - F(nd_{nj})]}{\sqrt{n}\tilde{\sigma}_{nj}} \xrightarrow{\mathbf{P}} 0. \end{aligned}$$

Algebrailag ellenőrizhető, hogy ebben az esetben

$$\begin{aligned} R_{nj}^{(1)} &= \overline{R}_{nj}^{(1)} - \frac{nd_{nj}e^{-nd_{nj}}Y_{n+1}}{\sqrt{n}\sigma_{nj}}, \\ R_{nj}^{(2)} &= \overline{R}_{nj}^{(2)}, \\ R_{nj}^{(3)} &= \overline{R}_{nj}^{(3)} - \frac{\mathbf{I}_{\{Y_{n+1} \le nd_{nj}\}} - F(nd_{nj}) - nd_{nj}e^{-nd_{nj}}}{\sqrt{n}\sigma_{nj}}. \end{aligned}$$

A (T2) feltétel mellett

$$0 < \liminf_{n\to\infty} e^{-nd_{nj}} \le \limsup_{n\to\infty} e^{-nd_{nj}} < 1.$$

Ebből következik, hogy $0 < \liminf_{n\to\infty} \sigma_{nj}$, tehát az $R_{nj}^{(3)}$ maradéktagban

$$\frac{nd_{nj}e^{-nd_{nj}}}{\sqrt{n}\sigma_{nj}} \to 0.$$

Mivel az $Y_1, Y_2, \ldots$ sorozat sztochasztikusan korlátos, ezért

$$\frac{nd_{nj}e^{-nd_{nj}}Y_{n+1}}{\sqrt{n}\sigma_{nj}} \xrightarrow{\mathbf{P}} 0.$$

Végül a Csebisev-egyenlőtlenség és a (3.10) azonosság alkalmazásával

$$P\left(\left|\frac{\mathbf{I}_{\{Y_{n+1}\le nd_{nj}\}} - F(nd_{nj})}{\sqrt{n}\sigma_{nj}}\right| > \varepsilon\right) \le \frac{D^2\big(\mathbf{I}_{\{Y_n\le nd_{nj}\}}\big)}{\varepsilon^2 n\sigma_{nj}^2} = \frac{1}{\varepsilon^2 n}\left(\frac{\tilde{\sigma}_{nj}}{\sigma_{nj}}\right)^2 \to 0,$$

minden $\varepsilon > 0$ esetén, hiszen $\tilde{\sigma}_{nj} \le 1$. Tehát sikerült megmutatnunk, hogy tetszőleges távolságszint sorozat esetén $R_{nj}^{(1)}, R_{nj}^{(2)}$ és $R_{nj}^{(3)}$ sztochasztikusan tart nullához (T2) feltétel esetén is. □

A 3.2. Tétel bizonyításának folytatása. Ahhoz, hogy a (3.9) formulában szereplő főtagokat vizsgálni tudjuk, tekintsük a

$$g_{nj}(x) := \frac{nd_{nj}e^{-nd_{nj}}(1-x) - [\mathbf{I}_{\{x\le nd_{nj}\}} - F(nd_{nj})]}{\sigma_{nj}}, \qquad x \in \mathbb{R},$$

$j = 1, \ldots, J$, $n = 1,2, \ldots$, mérhető függvényeket. Ekkor a főtag az alábbi alakban áll elő:

$$M_{nj} = \frac{\sum_{r=1}^{n} g_{nj}(Y_r)}{\sqrt{n}}.$$

Az a célunk, hogy a 3.3. Állítást alkalmazzuk az M_{nj} főtagra. Az így kapott $g_{nj}(Y_r)$ véletlen változó várható értékére teljesül a 3.3. Állítás feltétele, mivel

$$\begin{aligned} E\Big(g_{nj}(Y_r)\Big) &= E\left(\frac{nd_{nj}e^{-nd_{nj}}(1-Y_r) - [\mathbf{I}_{\{Y_r\le nd_{nj}\}} - F(nd_{nj})]}{\sigma_{nj}}\right) \\ &= \frac{nd_{nj}e^{-nd_{nj}}(1-1) - [P(Y_r \le nd_{nj}) - F(nd_{nj})]}{\sigma_{nj}} = 0. \end{aligned}$$

Továbbá vegyük észre, hogy minden $1 \le i \le j \le J$ esetén

$$E\big(\mathbf{I}_{\{Y_r\le nd_{ni}\}}\mathbf{I}_{\{Y_r\le nd_{nj}\}}\big) = P\big(Y_r \le \min(nd_{ni}, nd_{nj})\big) = F(nd_{ni}),$$

amiből következik, hogy

$$\begin{aligned} &E\left([\mathbf{I}_{\{Y_r\le nd_{ni}\}} - F(nd_{ni})][\mathbf{I}_{\{Y_r\le nd_{nj}\}} - F(nd_{nj})]\right) = E\big(\mathbf{I}_{\{Y_r\le nd_{ni}\}}\mathbf{I}_{\{Y_r\le nd_{nj}\}}\big) \\ &- E(\mathbf{I}_{\{Y_r\le nd_{ni}\}})F(nd_{nj}) - F(nd_{ni})E(\mathbf{I}_{\{Y_r\le nd_{nj}\}}) + F(nd_{ni})F(nd_{nj}) \\ &= F(nd_{ni}) - F(nd_{ni})F(nd_{nj}). \end{aligned}$$

Valamint

$$E\big((1-Y_r)\mathbf{I}_{\{Y_r\le nd_{nj}\}}\big)=\int_0^{nd_{nj}}(1-y)e^{-y}\,\mathrm{d}y=\big[(y-1)e^{-y}\big]_0^{nd_{nj}}-\int_0^{nd_{nj}}e^{-y}\,\mathrm{d}y=nd_{nj}e^{-nd_{nj}},$$

és $E((1-Y_r)^2)=D^2(Y_r)=1$. Ekkor

$$\begin{aligned}
E\Big(g_{ni}(Y_r)g_{nj}(Y_r)\Big)=&E\left(\frac{nd_{ni}e^{-nd_{ni}}(1-Y_r)-[\mathbf{I}_{\{Y_r\le nd_{ni}\}}-F(nd_{ni})]}{\sigma_{ni}}\right.\\
&\left.\cdot\frac{nd_{nj}e^{-nd_{nj}}(1-Y_r)-[\mathbf{I}_{\{Y_r\le nd_{nj}\}}-F(nd_{nj})]}{\sigma_{nj}}\right)\\
=&\frac{1}{\sigma_{ni}\sigma_{nj}}\left(nd_{ni}e^{-nd_{ni}}nd_{nj}e^{-nd_{nj}}E\left((1-Y_r)^2\right)+F(nd_{ni})-F(nd_{ni})F(nd_{nj})\right.\\
&\left.-nd_{ni}e^{-nd_{ni}}E\big((1-Y_r)\mathbf{I}_{\{Y_r\le nd_{nj}\}}\big)-nd_{nj}e^{-nd_{nj}}E\left((1-Y_r)\mathbf{I}_{\{Y_r\le nd_{ni}\}}\right)\right)\\
=&\frac{1}{\sigma_{ni}\sigma_{nj}}\left(nd_{ni}e^{-nd_{ni}}nd_{nj}e^{-nd_{nj}}+1-e^{-nd_{ni}}-(1-e^{-nd_{ni}})(1-e^{-nd_{nj}})\right.\\
&\left.-nd_{ni}e^{-nd_{ni}}nd_{nj}e^{-nd_{nj}}-nd_{nj}e^{-nd_{nj}}nd_{ni}e^{-nd_{ni}}\right)\\
=&\frac{e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}}-1-n^2d_{ni}d_{nj})}{\sigma_{ni}\sigma_{nj}}.
\end{aligned}$$

Ha $i<j$, akkor a (3.3) feltétel szerint teljesül a 3.3. Állítás (3.5) feltétele, míg $i=j$ esetben σ_{nj} definíciójából következik, hogy $E(g_{nj}^2(Y_r))=1$. Tehát megmutattuk, hogy teljesülnek a 3.3. Állítás kovarianciákra vonatkozó feltételei. Már csak az maradt hátra, hogy bebizonyítsuk a (3.6) feltevés érvényességét. Háromszor parciálisan integrálva

$$E|Y_r-1|^3=\int_0^1(1-y)^3e^{-y}\,\mathrm{d}y+\int_1^\infty(y-1)^3e^{-y}\,\mathrm{d}y=\frac{12-2e}{e},$$

illetve

$$\begin{aligned}
E[\mathbf{I}_{\{Y_r\le nd_{nj}\}}-F(nd_{nj})]^3&=(1-F(nd_{nj}))^3P\big(Y_r\le nd_{nj}\big)+(0-F(nd_{nj}))^3P\big(Y_r>nd_{nj}\big)\\
&=e^{-3nd_{nj}}(1-e^{-nd_{nj}})+(e^{-nd_{nj}}-1)^3e^{-nd_{nj}}=e^{-nd_{nj}}(1-e^{-nd_{nj}})(2e^{-nd_{nj}}-1).
\end{aligned}$$

Ekkor az L^3-normára vonatkozó háromszög-egyenlőtlenséget használva

$$\begin{aligned}
&\Big(E|g_{nj}(Y_r)|^3\Big)^{1/3}\le\frac{nd_{nj}}{\sigma_{nj}}e^{-nd_{nj}}\left(E|Y_r-1|^3\right)^{1/3}+\frac{1}{\sigma_{nj}}\left(E[\mathbf{I}_{\{Y_r\le nd_{nj}\}}-F(nd_{nj})]^3\right)^{1/3}\\
&=\frac{nd_{nj}}{\sigma_{nj}}e^{-nd_{nj}}\left(\frac{12-2e}{e}\right)^{1/3}+\frac{1}{\sigma_{nj}}\Big[e^{-nd_{nj}}(1-e^{-nd_{nj}})(2e^{-nd_{nj}}-1)\Big]^{1/3}\\
&\le\frac{nd_{nj}}{\sigma_{nj}}e^{-nd_{nj}}\left(\frac{12-2e}{e}\right)^{1/3}+\frac{[e^{-nd_{nj}}(1-e^{-nd_{nj}})]^{1/3}}{\sigma_{nj}}=f(nd_{nj}),
\end{aligned}$$

ahol

$$f(x)=\frac{x\left(\frac{12-2e}{e}\right)^{1/3}+e^{\frac{2}{3}x}(1-e^{-x})^{\frac{1}{3}}}{\sqrt{e^x-1-x^2}},\qquad x>0.$$

Továbbá a három különböző távolságszint sorozat mellett meg kell vizsgálnunk, hogy az $E(|g_{nj}(Y_r)|^3) = o(\sqrt{n})$ aszimptotika teljesül-e. Rögzített j mellett a továbbiakban a $\sim$ aszimptotikus ekvivalencia alatt azt értjük, hogy a két oldal hányadosa egyhez tart, és legyen $x_n = nd_{nj}$. Amennyiben a d_{nj} távolságszint sorozat kielégíti a (T1) feltétel, azaz $x_n \to 0$ és $nx_n \to \infty$, akkor az $e^x \sim 1+x,\ x \to 0$, aszimptotika mutatja, hogy

$$f(x_n) \sim \frac{x_n \left(\frac{12-2e}{e}\right)^{1/3} + x_n^{1/3}}{\sqrt{x_n - x_n^2}} \sim \frac{x_n^{1/3}}{\sqrt{x_n}} = \frac{1}{x_n^{1/6}}, \qquad n \to \infty.$$

Vagyis

$$E|g_{nj}(Y_r)|^3 \sim \frac{1}{{x_n}^{1/2}} = \frac{\sqrt{n}}{(nx_n)^{1/2}} = o(\sqrt{n}).$$

Ha a távolságszint sorozat a (T2) feltételt teljesíti, azaz

$$0 < \liminf_{n\to\infty} x_n \le \limsup_{n\to\infty} x_n < \infty,$$

akkor léteznek olyan $0<a<b<\infty$ valós számok, hogy $x_n \in [a,b]$ minden $n \in \mathbb{N}$ esetén. Mivel f folytonos a pozitív félegyenesen, ezért korlátos az $[a,b]$ intervallumon. Ebből következik, hogy

$$0 < \limsup_{n\to\infty} f(x_n) \le \sup_{x\in[a,b]} f(x) < \infty.$$

Tehát $E|g_{nj}(Y_r)|^3 = f^3(x_n)$ egy korlátos sorozat, amiből következik, hogy $E|g_{nj}(Y_r)|^3 = o(\sqrt{n})$. Ha a távolságszint sorozat a (T3) feltételt teljesíti, azaz $x_n \to \infty$ és $e^{x_n}/n \to 0$, akkor

$$f(x_n) = \frac{x_n e^{-\frac{x_n}{2}} \left(\frac{12-2e}{e}\right)^{1/3} + e^{\frac{x_n}{6}} (1-e^{-x_n})^{\frac{1}{3}}}{\sqrt{1-e^{-x_n}-x_n^2 e^{-x_n}}} \sim e^{\frac{x_n}{6}}$$

aszimptotikus viselkedés látható, vagyis

$$E|g_{nj}(Y_r)|^3 = f^3(x_n) \sim e^{\frac{x_n}{2}} = \left(\frac{e^{x_n}}{n}\right)^{1/2} \sqrt{n} = o(\sqrt{n}).$$

Így mindhárom esetben igaz az $E\left(|g_{nj}(Y_r)|^3\right) = o(\sqrt{n})$ nagyságrend.

Sikerült megmutatnunk, hogy a $\big(g_{n1}(Y_r), \ldots, g_{nn}(Y_r)\big)$ szériasorozatra teljesül a 3.3. Állítás összes feltevése. Ekkor az állításból következik, hogy

$$\big(M_{n1}, \ldots, M_{nn}\big) = \frac{\big(g_{n1}(Y_1), \ldots, g_{nJ}(Y_1)\big) + \cdots + \big(g_{n1}(Y_n), \ldots, g_{nJ}(Y_n)\big)}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma).$$

Mivel $R_{nj}^{(1)}, R_{nj}^{(2)}$ és $R_{nj}^{(3)}$ maradéktagok sztochasztikusan nullához konvergálnak minden $j = 1, \ldots, J$ esetén, a (3.9) felbontásból azonnal következik a tétel állítása. Ezzel a 3.2. Tételt bebizonyítottuk. □

3.5. Következmény. *Speciálisan tegyük fel, hogy $J \ge 2$ és $0 \le J_1 \le J_2 \le J$ olyanok, hogy minden $j \le J_1$ esetén a d_{nj} távolságszintek (T1) típusúak, és minden $j > J_2$ esetén pedig (T3) típusúak. Továbbá tegyük fel, hogy teljesülnek az alábbi feltételek:*

(i) Minden $i<j\le J_1$ *esetén* $s_{ij}:=\lim_{n\to\infty}\sqrt{d_{ni}/d_{nj}}\in\mathbb{R}$ *létezik.*
(ii) Minden $J_1<j\le J_2$ *esetén* $c_j:=\lim_{n\to\infty} nd_{nj}\in\mathbb{R}$ *szintén létezik. Ekkor* $J_1<i<j\le J_2$ *esetén*

$$s_{ij}:=\frac{(e^{c_i}-1-c_ic_j)}{\sqrt{(e^{c_i}-1-c_i^2)(e^{c_j}-1-c_j^2)}}.$$

(iii) Minden $J_2<i<j$ *esetén pedig* $s_{ij}:=\lim_{n\to\infty}e^{-n(d_{nj}-d_{ni})/2}\in\mathbb{R}$ *is létezik. Legyen továbbá* $s_{ji}:=s_{ij}$ *és* $s_{jj}:=1$. *Ekkor a (3.4) konvergencia érvényes, a*

$$\Sigma=\begin{pmatrix}\Sigma_1 & 0 & 0\\ 0 & \Sigma_2 & 0\\ 0 & 0 & \Sigma_3\end{pmatrix}$$

blokkdiagonális kovarianciamátrixszal, ahol Σ_1, Σ_2 *és* Σ_3 *blokkok rendre* $J_1\times J_1$, $(J_2-J_1)\times(J_2-J_1)$ *és* $(J-J_2)\times(J-J_2)$ *dimenziósak. A* Σ *mátrix blokkjaiban található komponensek a fent definiált* s_{ij} *értékek.*

Bizonyítás. A 3.5. Következmény bizonyítása a 3.2. Tételből következik azáltal, hogy ellenőrizük az

$$s_{nij}:=\frac{e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}}-1-n^2d_{ni}d_{nj})}{\sigma_{ni}\sigma_{nj}}=\frac{e^{nd_{ni}}-1-n^2d_{ni}d_{nj}}{\sqrt{(e^{nd_{ni}}-1-n^2d_{ni}^2)(e^{nd_{nj}}-1-n^2d_{nj}^2)}},$$

$1\le i<j\le J$, sorozat konvergenciáját. Ehhez elég azt megmutatni, hogy $s_{nij}\sim s_{ij}$. Először legyen mindkét távolságszint sorozat (T1) típusú. Használva az $e^x-1\sim x$, $x\to 0$, aszimptotikus ekvivalenciát és az (i) feltételt, a következő aszimptotikus relációt kapjuk:

$$s_{nij}\sim\frac{nd_{ni}-n^2d_{ni}d_{nj}}{\sqrt{(nd_{ni}-n^2d_{ni}^2)(nd_{nj}-n^2d_{nj}^2)}}=\frac{\sqrt{nd_{ni}(1-nd_{nj})}}{\sqrt{(1-nd_{ni})nd_{nj}}}\sim\sqrt{\frac{d_{ni}}{d_{nj}}}\sim s_{ij}.$$

Amennyiben a távolságszint sorozatok kielégítik a (T2) feltételt, akkor az

$$s_{nij}\sim\frac{(e^{c_i}-1-c_ic_j)}{\sqrt{(e^{c_i}-1-c_i^2)(e^{c_j}-1-c_j^2)}}=s_{ij},$$

egyenlőség nyilvánvaló a (ii) feltétel miatt. Amennyiben a távolságszint sorozatok a (T3) feltételt teljesítik, akkor használva az $e^{-x}x^2\to 0$, $x\to\infty$, konvergenciát, továbbá a (iii) feltételt, azt kapjuk, hogy

$$\begin{aligned}s_{nij}&=\frac{\sqrt{e^{-nd_{ni}}e^{-nd_{nj}}}e^{nd_{ni}}(1-e^{-nd_{ni}}-e^{-nd_{ni}}n^2d_{ni}d_{nj})}{\sqrt{\big(1-e^{-nd_{ni}}-e^{-nd_{ni}}(nd_{ni})^2\big)\big(1-e^{-nd_{nj}}-e^{-nd_{nj}}(nd_{nj})^2\big)}}\\&=\frac{\sqrt{e^{n(d_{ni}-d_{nj})}}\big(1-e^{-nd_{ni}}-e^{-nd_{ni}}nd_{ni}n(d_{nj}-d_{ni})-e^{-nd_{ni}}(nd_{ni})^2\big)}{\sqrt{\big(1-e^{-nd_{ni}}-e^{-nd_{ni}}(nd_{ni})^2\big)\big(1-e^{-nd_{nj}}-e^{-nd_{nj}}(nd_{nj})^2\big)}}\\&\sim\sqrt{e^{n(d_{ni}-d_{nj})}}\sim s_{ij}.\end{aligned}$$

A különböző típusú távolságszint sorozatok esetén az s_{nij} sorozatoknak nullához kell tartaniuk. Az alábbi három esetben jelölje mindig i az első, j pedig a második feltételt kielégítő távolságszint sorozatot. A (T1) és (T2) feltételt teljesítő távolságszint sorozatok esetén ismét használjuk az $e^x - 1 \sim x$, $x \to 0$, aszimptotikát. Ekkor

$$s_{nij} \sim \frac{nd_{ni} - nd_{ni}c_j}{\sqrt{(nd_{ni} - n^2d_{ni}^2)(e^{c_j} - 1 - c_j^2)}} = \frac{\sqrt{nd_{ni}}(1-c_j)}{\sqrt{(1-d_{ni})(e^{c_j} - 1 - c_j^2)}} \to 0.$$

Ha a távolságszint sorozatok (T1) és (T3) típusúak, akkor

$$\begin{aligned} s_{nij} &\sim \frac{nd_{ni} - n^2 d_{ni} d_{nj}}{\sqrt{(nd_{ni} - n^2d_{ni}^2)e^{nd_{nj}}(1 - e^{-nd_{nj}} - e^{-nd_{nj}}(nd_{nj})^2)}} \\ &= \sqrt{\frac{nd_{ni}}{(1-nd_{ni})}} \frac{(1-nd_{nj})\sqrt{e^{-nd_{nj}}}}{\sqrt{1 - e^{-nd_{nj}} - e^{-nd_{nj}}(nd_{nj})^2}} \sim \sqrt{\frac{0}{1}\frac{0}{1}} = 0. \end{aligned}$$

Ha pedig (T2) és (T3) típusúak, akkor

$$\begin{aligned} s_{nij} &\sim \frac{(e^{c_i} - 1 - c_i nd_{nj})}{\sqrt{(e^{c_i} - 1 - c_i^2)\left(e^{nd_{nj}} - 1 - n^2 d_{nj}^2\right)}} \\ &= \frac{(e^{c_i} - 1 - c_i nd_{nj})\sqrt{e^{-nd_{nj}}}}{\sqrt{(e^{c_i} - 1 - c_i^2)\left(1 - e^{-nd_{nj}} - e^{-nd_{nj}}(nd_{nj})^2\right)}} \to 0. \end{aligned}$$

Ezzel bebizonyítottuk a 3.5. Következményt. □

Csörgő és Wu [23] mutat jól viselkedő távolságszint sorozatokat mindhárom típushoz, nevezzük ezeket tipikus sorozatoknak. A 3.5. Következményt fogjuk alkalmazni ezekre a tipikus sororozatokra. Azáltal, hogy a sorozatokban lévő paramétereket jól választjuk, diagonális kovarianciamátrixot kapunk. Egy tipikus $(d_n)_{n=1,2,\ldots}$ távolságszint sorozat (T1) esetben a $d_n = n^{-\alpha}$ sorozat tetszőleges $\alpha \in (1,2)$ paraméterrel. J_1 darab ilyen $d_{nj} = n^{-\alpha_j}$, $j \leq J_1$, sorozatot véve, $\alpha_1 > \alpha_2 > \cdots > \alpha_{J_1}$ paraméterrel a kovarianciamátrixban $s_{ij} = 0$ adódik minden $i < j \leq J_1$ esetén. Ennek az az oka, hogy

$$\begin{aligned} s_{nij} &= \frac{e^{nn^{-\alpha_i}} - 1 - n^2 n^{-\alpha_i} n^{-\alpha_j}}{\sqrt{\left(e^{nn^{-\alpha_i}} - 1 - n^2(n^{-\alpha_i})^2\right)\left(e^{nn^{-\alpha_j}} - 1 - n^2(n^{-\alpha_j})^2\right)}} \\ &= \frac{e^{n^{1-\alpha_i}} - 1 - n^{1-\alpha_i} n^{1-\alpha_j}}{\sqrt{\left(e^{n^{1-\alpha_i}} - 1 - (n^{1-\alpha_i})^2\right)\left(e^{n^{1-\alpha_j}} - 1 - (n^{1-\alpha_j})^2\right)}} \\ &\sim \frac{e^{n^{1-\alpha_i}} - 1}{\sqrt{\left(e^{n^{1-\alpha_i}} - 1\right)\left(e^{n^{1-\alpha_j}} - 1\right)}} \sim \sqrt{\frac{e^{n^{1-\alpha_i}} - 1}{e^{n^{1-\alpha_j}} - 1}} \\ &\sim \sqrt{\frac{e^{n^{1-\alpha_i}}(1-\alpha_i)n^{-\alpha_i}}{e^{n^{1-\alpha_j}}(1-\alpha_j)n^{-\alpha_j}}} = \sqrt{\frac{e^{n^{1-\alpha_i}}(1-\alpha_i)}{e^{n^{1-\alpha_j}}(1-\alpha_j)}} n^{\alpha_j - \alpha_i} \to 0, \end{aligned}$$

ahol a gyök alatti sorozat viselkedését a L'Hospital-szabály segítségével vizsgáltuk. Hasonlóan egy tipikus $(d_n)_{n=1,2,\dots}$ távolságszint sorozat a (T3) esetben a $d_n = \beta(\log n)/n$ sorozat tetszőleges $\beta \in (0,1)$ paraméterrel. Így a $d_{nj} = \beta_j(\log n)/n$, $j > J_2$, sorozatok, a $\beta_{J_2+1} < \beta_{J_2+2} < \dots < \beta_J$ paraméterválasztással szintén a $s_{ij} = 0$ értékeket eredményezik minden $J_2 < i < j < J$ esetén, mivel

$$\begin{aligned} s_{nij} &= \frac{e^{\beta_i(\log n)} - 1 - \beta_i(\log n)\beta_j(\log n)}{\sqrt{\left(e^{\beta_i(\log n)} - 1 - (\beta_i(\log n))^2\right)\left(e^{\beta_j(\log n)} - 1 - (\beta_j(\log n))^2\right)}} \\ &= \frac{n^{\beta_i} - 1 - \beta_i(\log n)\beta_j(\log n)}{\sqrt{\left(n^{\beta_i} - 1 - (\beta_i(\log n))^2\right)\left(n^{\beta_j} - 1 - (\beta_j(\log n))^2\right)}} \\ &= \frac{n^{\beta_i}\left(1 - n^{-\beta_i} - n^{-\beta_i}\beta_i(\log n)\beta_j(\log n)\right)}{\sqrt{n^{\beta_i}\left(1 - n^{-\beta_i} - n^{-\beta_i}(\beta_i(\log n))^2\right)n^{\beta_j}\left(1 - n^{-\beta_j} - n^{-\beta_j}(\beta_j(\log n))^2\right)}} \\ &\sim \frac{n^{\beta_i}}{\sqrt{n^{\beta_i}n^{\beta_j}}} = \sqrt{\frac{n^{\beta_i}}{n^{\beta_j}}} \to 0. \end{aligned}$$

Végül, legyen $0 \le J_2 - J_1 \le 2$, ami azt jelenti, hogy a Σ_2 legfeljebb 2×2-es mátrix. A $J_2 - J_1 = 0$ esetben nincs (T2) típusú távolságszint sorozat, míg a $J_2 - J_1 = 1$ esetén egy ilyen típusú sorozat van. Ezekben az esetekben 3.5. Következmény (ii) feltétele automatikusan teljesül. A $J_2 - J_1 = 2$ esetben pedig ha a $c_{J_2} = (e^{c_{J_1+1}} - 1)/c_{J_1+1}$ összefüggés teljesül, akkor algebrailag ellenőrizhető, hogy $s_{J_1+1,J_2} = 0$, így a 3.5. Következmény (ii) feltétele teljesül. Ezekkel a tipikus sorozatokkal a 3.5. Következmény a következő alakot ölti.

3.6. Következmény. *Az előző bekezdésben szereplő távolságszint sorozatok esetén*

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, E_J),$$

ahol E_J a J dimenziós egységmátrix.

Jegyezzük meg, hogy diagonális kovarianciamátrixot távolságszintek más sorozatára is kaphatuk.

3.2.2. Adott intervallumon egyenletes eloszlásból származó klaszterszámok együttes aszimptotikus viselkedése

Legyenek $V_1, V_2, \dots, V_n$ független, egy ismert $[a,b]$ intervallumon egyenletes eloszlású véletlen változók, ahol $a, b \in \mathbb{R}$, $a < b$. Jelölje $K_n^{a,b} := K_n^{a,b}(d_n)$ az $[a,b]$ intervallumból származó $V_1, V_2, \dots, V_n$ mintához és a d_n távolságszinthez tartozó klaszterszámot, amely mennyiséget ugyanúgy definiáljuk, mint a $[0,1]$ intervallumon a $K_n^{0,1}(d_n) = K_n(d_n)$ klaszterszámot. Továbbra is a három típusból származó távolságszintekhez tartozó $K_n^{a,b}(d_n)$ klaszterszámok együttes viselkedésével foglalkozunk. Ebben az esetben is belátható egy, a 3.2. Tételhez hasonló állítás.

Legyen $J \ge 1$ természetes szám, és legyenek $d_{n1} \le d_{n2} \le \dots \le d_{nJ}$ távolságszint sorozatok. A $K_{nj}^{a,b}(d_{nj})$ jelöli a megfelelő d_{nj} távolságszinthez tartozó klaszterszámot, $j = 1, \dots, J$.

Legyenek

$$m_{nj}^{a,b} = ne^{-\frac{nd_{nj}}{b-a}}, \qquad \sigma_{nj}^{a,b} = \sqrt{e^{-2\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{nj}}{b-a}} - 1 - \left(\frac{nd_{nj}}{b-a}\right)^2\right)}, \tag{3.11}$$

valamint

$$\mathbf{K}_n^{a,b} = \frac{1}{\sqrt{n}}\left(\frac{K_{n1}^{a,b}(d_{n1}) - m_{n1}^{a,b}}{\sigma_{n1}^{a,b}}, \ldots, \frac{K_{nJ}^{a,b}(d_{nJ}) - m_{nJ}^{a,b}}{\sigma_{nJ}^{a,b}}\right)^\top. \tag{3.12}$$

Ekkor igaz a következő állítás:

3.7. Tétel. *Tegyük fel, hogy a d_{nj} sorozatok mindegyike kielégíti a (T1), a (T2) vagy a (T3') feltétel valamelyikét, ahol*

(T3') $nd_{nj} \to \infty$, $ne^{-\frac{nd_{nj}}{b-a}} \to \infty$.

Tegyük fel továbbá, hogy létezik s_{ij} valós szám, amire

$$e^{-\frac{nd_{ni}}{b-a} - \frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{ni}}{b-a}} - 1 - \frac{nd_{ni}}{b-a}\frac{nd_{nj}}{b-a}\right)/\sigma_{ni}^{a,b}\sigma_{nj}^{a,b} \to s_{ij}, \qquad 1 \le i < j \le J, \tag{3.13}$$

és legyen $s_{ii} := 1$ és $s_{ji} := s_{ij}$. Ekkor érvényes a

$$\mathbf{K}_n^{a,b} \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma) \tag{3.14}$$

konvergencia a $\Sigma = (s_{ij})_{i,j=1,\ldots,J}$ kovarianciamátrixszal.

Bizonyítás. A 3.7. Tétel közvetlen következménye a 3.2. Tételnek, köszönhetően az $[a, b]$ és a [0,1] intervallumok közötti lineáris transzformálciónak. Természetesen mind a mintát, mind a távolságszinteket transzformálni kell,

$$U_i = \frac{V_i - a}{b-a}, \qquad d_{ni}^{0,1} = \frac{d_{ni}}{b-a}$$

aminek következtében az új változóra az új távolságszint sorozatokkal teljesülnek a 3.2. Tétel feltételei, ekkor

$$K_n^{a,b}(d_n) = K_n^{0,1}\left(\frac{d_n}{b-a}\right) \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma),$$

amivel a tételt bebizonyítottuk. □

A 3.5. Következmény megfelelőjét ebben az esetben is be lehet bizonyítani.

3.8. Következmény. *Tegyük fel, hogy $J \ge 2$ és $0 \le J_1 \le J_2 \le J$ olyanok, hogy minden $j \le J_1$ esetén a d_{nj} távolságszintek (T1) típusúak, és minden $j > J_2$ esetén pedig (T3') típusúak. Továbbá tegyük fel, hogy teljesülnek az alábbi feltételek:*

(i) Minden $i < j \le J_1$ esetén $s_{ij} := \lim_{n\to\infty} \sqrt{d_{ni}/d_{nj}} \in \mathbb{R}$ létezik.

(ii') Minden $J_1 < j \le J_2$ esetén $c_j := \lim_{n\to\infty} \frac{nd_{nj}}{b-a} \in \mathbb{R}$ szintén létezik. Ekkor $J_1 < i < j \le J_2$ esetén

$$s_{ij} := \frac{(e^{c_i} - 1 - c_i c_j)}{\sqrt{(e^{c_i} - 1 - c_i^2)(e^{c_j} - 1 - c_j^2)}}.$$

(iii') Minden $J_2 < i < j$ *esetén pedig* $s_{ij} := \lim_{n\to\infty} e^{-n(d_{nj}-d_{ni})/2(b-a)} \in \mathbb{R}$ *is létezik. Legyen továbbá* $s_{ji} := s_{ij}$ *és* $s_{jj} := 1$. *Ekkor a (3.14) konvergencia érvényes, a*

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix} \tag{3.15}$$

blokkdiagonális kovarianciamátrixszal, ahol Σ_1, Σ_2 *és* Σ_3 *blokkok rendre* $J_1 \times J_1$, $(J_2 - J_1) \times (J_2 - J_1)$ *és* $(J - J_2) \times (J - J_2)$ *dimenziósak. A* Σ *mátrix blokkjaiban található komponensek a fent definiált* s_{ji} *értékek.*

3.2.3. Ismeretlen intervallumon egyenletes eloszlásból származó klaszterszámok együttes aszimptotikus viselkedése

Legyenek $V_1, V_2, \ldots, V_n$ független, egy ismeretlen $[a, b]$ intervallumon egyenletes eloszlású véletlen változók, ahol $a, b \in \mathbb{R}$, $a < b$, valamint legyen $V_{1,n}, \ldots, V_{n,n}$ a hozzá tartozó rendezett minta. A 3.2. és 3.7. Tételek megfelelőit keressük úgy, hogy az intervallum végpontjait becsüljük az $\hat{a}_n = V_{1,n}$ legkisebb, és a $\hat{b}_n = V_{n,n}$ legnagyobb mintaelemmel.

Hasonlóan az eddigi jelölésekhez, adott $J \geq 1$ természetes szám és adott $d_{n1} < \cdots < d_{nJ}$ távolságszintek esetén $\hat{K}_{nj}(d_{nj})$ jelöli a megfelelő d_{nj} távolságszinthez tartozó klaszterszámot, $j = 1, \ldots, J$. Legyenek

$$\hat{m}_{nj} = ne^{-\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}}, \qquad \hat{\sigma}_{nj} = \sqrt{e^{-2\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}} \left(e^{\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}} - 1 - \left(\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n} \right)^2 \right)}$$

valamint

$$\widehat{\mathbf{K}}_n = \frac{1}{\sqrt{n}} \left(\frac{\hat{K}_{n1}(d_{n1}) - \hat{m}_{n1}}{\hat{\sigma}_{n1}}, \ldots, \frac{\hat{K}_{nJ}(d_{nJ}) - \hat{m}_{nJ}}{\hat{\sigma}_{nJ}} \right)^\top. \tag{3.16}$$

3.9. Tétel. *Tegyük fel, hogy teljesülnek a 3.7. Tétel feltételei, és tekintsük az ott definiált* Σ *kovarianciamátrixot. Ekkor*

$$\widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma). \tag{3.17}$$

A 3.9. Tétel bizonyítása előtt kimondunk két lemmát, amit használni fogunk. Az első a Szluckij-lemma egy általánosítása:

3.10. Lemma. *Legyenek* $X_n = (X_{n1}, X_{n2}, \ldots, X_{nJ})^\top$, $L_n = (L_{n1}, L_{n2}, \ldots, L_{nJ})^\top$ *és* $S_n = (S_{n1}, S_{n2}, \ldots, S_{nJ})^\top$, n=1,2 $\ldots$, $\mathbb{R}^J$*-értékű véletlen vektorokból álló sorozatok, és legyenek* $l_n = (l_{n1}, l_{n2}, \ldots, l_{nJ})^\top$ *és* $s_n = (s_{n1}, s_{n2}, \ldots, s_{nJ})^\top \in \mathbb{R}^J$ *determinisztikus sorozatok. Tegyük fel, hogy létezik* Y $\mathbb{R}^J$*-értékű véletlen vektor úgy, hogy*

$$\left(\frac{X_{n1} - l_{n1}}{s_{n1}}, \ldots, \frac{X_{nJ} - l_{nJ}}{s_{nJ}} \right)^\top \xrightarrow{\mathcal{D}} Y,$$

és tetszőleges $0 \leq j \leq J$ *esetén* $(L_{nj} - l_{nj})/s_{nj} \to_P 0$ *és* $S_{nj}/s_{nj} \to_P 1$. *Ekkor*

$$\left(\frac{X_{n1} - L_{n1}}{S_{n1}}, \ldots, \frac{X_{nJ} - L_{nJ}}{S_{nJ}} \right)^\top \xrightarrow{\mathcal{D}} Y.$$

Bizonyítás. A Cramér–Wold-lemma szerint elegendő azt bebizonyítani, hogy tetszőleges $c = (c_1, \dots, c_J)^\top \in \mathbb{R}^J$ vektor esetén

$$\sum_{j=1}^{J} c_j \frac{X_{nj} - L_{nj}}{S_{nj}} \xrightarrow{\mathcal{D}} c^\top Y.$$

A feltevésből következik, hogy

$$\sum_{j=1}^{J} c_j \frac{X_{nj} - l_{nj}}{s_{nj}} \xrightarrow{\mathcal{D}} c^\top Y,$$

ekkor a Szluckij-lemma szerint elegendő azt belátni, hogy

$$\left| \sum_{j=1}^{J} c_j \frac{X_{nj} - l_{nj}}{s_{nj}} - \sum_{j=1}^{J} c_j \frac{X_{nj} - L_{nj}}{S_{nj}} \right| \xrightarrow{\mathbf{P}} 0,$$

A háromszög-egyenlőtlenség alkalmazásával a lemma feltételeiből következik, hogy

$$\begin{aligned}
&\left| \sum_{j=1}^{J} c_j \frac{X_{nj} - l_{nj}}{s_{nj}} - \sum_{j=1}^{J} c_j \frac{X_{nj} - L_{nj}}{S_{nj}} \right| \leq \sum_{j=1}^{J} c_j \left| \frac{X_{nj} - l_{nj}}{s_{nj}} - \frac{X_{nj} - L_{nj}}{S_{nj}} \right| \\
&= \sum_{j=1}^{J} c_j \left| \frac{L_{nj}}{s_{nj}} - \frac{l_{nj}}{s_{nj}} + \frac{L_{nj}}{S_{nj}} - \frac{L_{nj}}{s_{nj}} + \frac{X_{nj}}{s_{nj}} - \frac{X_{nj}}{S_{nj}} \right| \\
&= \sum_{j=1}^{J} c_j \left| \frac{L_{nj} - l_{nj}}{s_{nj}} + \frac{L_{nj} - l_{nj}}{s_{nj}} \left(\frac{s_{nj}}{S_{nj}} - 1 \right) + \frac{X_{nj} - l_{nj}}{s_{nj}} \left(1 - \frac{s_{nj}}{S_{nj}} \right) \right| \\
&\leq \sum_{j=1}^{J} c_j \left(\left| \frac{L_{nj} - l_{nj}}{s_{nj}} \right| + \left| \frac{L_{nj} - l_{nj}}{s_{nj}} \right| \left| \frac{s_{nj}}{S_{nj}} - 1 \right| + \left| \frac{X_{nj} - l_{nj}}{s_{nj}} \right| \left| 1 - \frac{s_{nj}}{S_{nj}} \right| \right) \xrightarrow{\mathbf{P}} 0,
\end{aligned}$$

amit bizonyítani akartunk. □

Ismert, hogy az $n(b - \hat{b}_n)$ és az $n(\hat{a}_n - a)$ változóknak van nemdegenerált határeloszlása. Ennek bizonyítása például megtalálható [54]-ben. Ebből következik az alábbi lemma.

3.11. Lemma. *Minden $\alpha < 1$ esetén*

$$n^\alpha (b - \hat{b}_n) \xrightarrow{\mathbf{P}} 0 \quad \text{és} \quad n^\alpha (\hat{a}_n - a) \xrightarrow{\mathbf{P}} 0.$$

Most már foglalkozhatunk a 3.9. Tétel bizonyításával.

A 3.9. Tétel bizonyítása. Vegyük észre, hogy tetszőleges j esetén $\hat{K}_{nj}(d_{nj}) = K_{nj}^{a,b}(d_{nj})$. Ekkor a 3.7. Tétel és a 3.10. Lemma szerint elegendő azt megmutatni, hogy

$$\frac{\hat{m}_{nj} - m_{nj}^{a,b}}{\sqrt{n}\sigma_{nj}^{a,b}} \xrightarrow{\mathbf{P}} 0 \qquad \text{és} \qquad \frac{\hat{\sigma}_{nj}^2}{(\sigma_{nj}^{a,b})^2} \xrightarrow{\mathbf{P}} 1, \quad j = 1, \dots, J. \tag{3.18}$$

Mivel $\hat{b}_n-\hat{a}_n<b-a$ majdnem biztosan teljesül, és a Lagrange-tétel miatt minden $x\le y$ esetén $|e^{-y}-e^{-x}|\le|x-y|e^{-x}$ érvényes, így a következő becslést kapjuk:

$$|\hat{m}_{nj}-m_{nj}^{a,b}| = n\left|e^{\frac{-nd_{nj}}{\hat{b}_n-\hat{a}_n}}-e^{\frac{-nd_{nj}}{b-a}}\right| \le n\left|\frac{nd_{nj}}{b-a}-\frac{nd_{nj}}{\hat{b}_n-\hat{a}_n}\right| e^{-\frac{nd_{nj}}{b-a}}$$
$$\le n^2 d_{nj}\frac{|\hat{b}_n-\hat{a}_n-(b-a)|}{(b-a)(\hat{b}_n-\hat{a}_n)}e^{-\frac{nd_{nj}}{b-a}} \le n^2 d_{nj}\frac{|\hat{b}_n-b|+|a-\hat{a}_n|}{(b-a)(\hat{b}_n-\hat{a}_n)}e^{-\frac{nd_{nj}}{b-a}}.$$

Ennélfogva

$$\left|\frac{\hat{m}_{nj}-m_{nj}^{a,b}}{\sqrt{n}\sigma_{nj}^{a,b}}\right| \le \frac{n^2 d_{nj}\frac{|\hat{b}_n-b|+|a-\hat{a}_n|}{(b-a)(\hat{b}_n-\hat{a}_n)}e^{-\frac{nd_{nj}}{b-a}}}{\sqrt{n}\sqrt{e^{-2\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{nj}}{b-a}}-1-\left(\frac{nd_{nj}}{b-a}\right)^2\right)}}$$
$$= \sqrt{n}\frac{|\hat{b}_n-b|+|a-\hat{a}_n|}{\hat{b}_n-\hat{a}_n}\frac{nd_{nj}}{b-a}e^{-\frac{nd_{nj}}{b-a}}\Bigg/\sqrt{e^{-2\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{nj}}{b-a}}-1-\left(\frac{nd_{nj}}{b-a}\right)^2\right)}$$
$$= \frac{\sqrt{n}|\hat{b}_n-b|+\sqrt{n}|a-\hat{a}_n|}{(\hat{b}_n-\hat{a}_n)}\varphi\left(\frac{nd_{nj}}{b-a}\right),$$

ahol $\varphi(x)=x/\sqrt{e^x-1-x^2}$, $x>0$. Vizsgáljuk meg a $\varphi(x)$ folytonos függvényt a $(0,\infty)$ intervallumon. A L'Hospital-szabály kétszeri alkalmazásával

$$\lim_{x\to0}(\varphi(x))^2=\lim_{x\to0}\frac{x^2}{e^x-1-x^2}=\lim_{x\to0}\frac{2x}{e^x-2x}=\lim_{x\to0}\frac{2}{e^x-2}=0,$$

továbbá elemi határérték számolási módszerekkel

$$\lim_{x\to\infty}\varphi(x)=\lim_{x\to\infty}\frac{x}{\sqrt{e^x-1-x^2}}=\lim_{x\to\infty}\frac{xe^{-\frac{x}{2}}}{\sqrt{1-e^{-x}-x^2e^{-x}}}=0.$$

Így $\varphi(x)$ egy korlátos függvény. Ekkor a (3.18) formula első konvergenciája következik a 3.11. Lemmából $\alpha=1/2$ paraméterrel.

A második konvergencia bizonyításához legyen $\psi(x)=x/(1-e^{-x}-x^2e^{-x})$, $x>0$. Szintén vizsgáljuk meg a $\psi(x)$ folytonos függvényt a $(0,\infty)$ intervallumon. Először is, mivel $1-e^{-x}<x$ és $0<xe^{-x}<1$, így

$$\psi(x)=\frac{x}{1-e^{-x}-x^2e^{-x}}>\frac{x}{x-x^2e^{-x}}=\frac{1}{1-xe^{-x}}>1, \qquad x>0.$$

Másrészt, mivel a ψ függvény deriváltja folytonos függyvény a pozitív félegyenesen véges határértékekkel a 0 és ∞ helyeken, ezért a derivált korlátos függvény a $(0,\infty)$ intervallumon. Legyen K a $|\psi'|$ függyvény egy korlátja. Ekkor a Lagrange-tételből következik, hogy

$$|\psi(y)-\psi(x)|\le K|y-x|, \qquad x,y>0,$$

tehát a ψ függyvény Lipschitz-folytonos a $(0,\infty)$ intervallumon.

Most már foglalkozhatunk a (3.18) formula második konvergenciájával. Azt kapjuk, hogy

$$\begin{aligned}\frac{\hat\sigma^2_{nj}}{(\sigma^{a,b}_{nj})^2}-1 &= \frac{e^{-2\frac{nd_{nj}}{\hat b_n-\hat a_n}}\left(e^{\frac{nd_{nj}}{\hat b_n-\hat a_n}}-1-\left(\frac{nd_{nj}}{\hat b_n-\hat a_n}\right)^2\right)}{e^{-2\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{nj}}{b-a}}-1-\left(\frac{nd_{nj}}{b-a}\right)^2\right)}-1\\ &= e^{-nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)}\frac{1-e^{-\frac{nd_{nj}}{\hat b_n-\hat a_n}}-\left(\frac{nd_{nj}}{\hat b_n-\hat a_n}\right)^2 e^{-\frac{nd_{nj}}{\hat b_n-\hat a_n}}}{1-e^{-\frac{nd_{nj}}{b-a}}-\left(\frac{nd_{nj}}{b-a}\right)^2 e^{-\frac{nd_{nj}}{b-a}}}-1\\ &= \frac{b-a}{\hat b_n-\hat a_n}e^{-nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)}\frac{\psi(\frac{nd_{nj}}{b-a})}{\psi(\frac{nd_{nj}}{\hat b_n-\hat a_n})}-1\\ &= \frac{b-a}{\hat b_n-\hat a_n}e^{-nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)}\frac{\psi(\frac{nd_{nj}}{b-a})-\psi(\frac{nd_{nj}}{\hat b_n-\hat a_n})}{\psi(\frac{nd_{nj}}{\hat b_n-\hat a_n})}+\frac{b-a}{\hat b_n-\hat a_n}e^{-nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)}-1.\end{aligned}$$

Mivel mindhárom esetben $nd_{nj}<\log n$ elég nagy n esetén, ezért a 3.11. Lemmából következik, hogy

$$0\le nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)=nd_{nj}\frac{(b-a)-(\hat b_n-\hat a_n)}{(\hat b_n-\hat a_n)(b-a)}\le\frac{\log n(b-\hat b_n)+\log n(\hat a_n-a)}{(\hat b_n-\hat a_n)(b-a)}\xrightarrow{\mathbf{P}}0.$$

Ekkor a ψ függvény tulajdonságai miatt a (3.18) második konvergenciája is teljesül:

$$\begin{aligned}\left|\frac{\hat\sigma^2_{nj}}{(\sigma^{a,b}_{nj})^2}-1\right| \le{}& \frac{b-a}{\hat b_n-\hat a_n}e^{-nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)}K\cdot nd_{nj}\left(\frac{1}{b-a}-\frac{1}{\hat b_n-\hat a_n}\right)\\ &+\left|\frac{b-a}{\hat b_n-\hat a_n}e^{-nd_{nj}\left(\frac{1}{\hat b_n-\hat a_n}-\frac{1}{b-a}\right)}-1\right|\xrightarrow{\mathbf{P}}0.\end{aligned}$$

Ezzel bebizonyítottuk a 3.9. Tételt. □

3.3. Statisztikai eredmények és szimuláció

3.3.1. Tesztstatisztikák

Adott $X_1,\dots,X_n$ minta egy ismeretlen $F(x),\ x\in\mathbb{R}$, eloszlásfüggvényű véletlen változóból. Tesztelni szeretnénk azt az egyszerű nullhipotézist, hogy

$$\mathcal{H}_0: F=F_0,$$

ahol most F_0 a [0,1] intervallumon egyenletes eloszlás eloszlásfüggvényét jelöli.

Tetszőleges $J\ge 1$ esetén legyenek a $d_{n1}\le\ldots\le d_{nJ},\ n\in\mathbb{N}$, távolságszint sorozatok olyanok, hogy mindegyik sorozat kielégíti a (T1), (T2) vagy (T3) feltételek valamelyikét.

Továbbá tegyük fel, hogy a (3.3) feltétel teljesül, és a 3.2. Tételbeli Σ kovarianciamátrix nem szinguláris. Legyen $\mathbf{K}_n$ a (3.1)-ben definiált vektor. Ekkor a (3.4) konvergenciából a nullhipotézis mellett következik, hogy a tesztstatisztika

$$C_n := \mathbf{K}_n^\top \Sigma^{-1} \mathbf{K}_n \xrightarrow{\mathcal{D}} \chi_J^2, \tag{3.19}$$

ahol χ_J^2 a J szabadsági fokú khi-négyzet eloszlás. Így a C_n próbastatisztikával tesztelhetjük a $\mathcal{H}_0$ nullhipotézist. Ezt a tesztet nevezzünk *klaszterteszt*nek. A (3.19) formulából következik, hogy ennek a tesztnek az aszimptotikus kritikus értékei a J szabadsági fokú khi-négyzet eloszlás kvantilisei. Mivel ez a konvergencia nagyon lassú, célszerű inkább a tesztstatisztika empirikus kvantiliseit használni. Erről részletesen a 3.3.2. fejezetben írunk.

Jelölje $\mathcal{F}$ a véges zárt intervallumon vett egyenletes eloszlások családját. Tekintsük azt az összetett nullhipotézist, hogy a minta valamelyik egyenletes eloszlásból származik, tehát

$$\mathcal{H}_0 : F \in \mathcal{F}.$$

Legyenek a $d_{n1} \leq \ldots \leq d_{nJ},\ n \in \mathbb{N}$, távolságszint sorozatok olyanok, melyek kielégítik a 3.9. Tétel feltételeit. Ekkor teljesül

$$\widehat{C}_n := \widehat{\mathbf{K}}_n^\top \Sigma^{-1} \widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \chi_J^2. \tag{3.20}$$

Ez alapján úgy tűnhet, hogy az összetett nullhipotézist lehet tesztelni az előző bekezdéshez hasonlóan. A probléma az, hogy mivel nem ismertjük az a és b pontos értékét, ezért a Σ kovarianciamátrix komponenseit se tudjuk meghatározni, emiatt a $\widehat{C}_n$ statisztika egy adott minta alapján nem számolható ki. Éppen emiatt az összetett nullhipotézist egy másik módszerrel fogjuk tesztelni. Egy lehetséges megoldás, hogy az adatokat a [0,1] intervallumba transzformáljuk, ami a következő lemma alapján lehetséges.

3.12. Lemma. *Legyenek $V_1, \ldots, V_n$ független, az $[a, b]$ intervallumon egyenletes eloszlású véletlen változók, és legyen $V_{1,n} \leq \ldots \leq V_{n,n}$ a rendezett minta. Ekkor minden rögzített n esetén*

$$\left(\frac{V_{2,n} - V_{1,n}}{V_{n,n} - V_{1,n}}, \ldots, \frac{V_{n-1,n} - V_{1,n}}{V_{n,n} - V_{1,n}} \right) \overset{\mathcal{D}}{=} (U_{1,n-2}, \ldots, U_{n-2,n-2}), \tag{3.21}$$

amely eloszlásbeli egyenlőség jobb oldalán a [0,1] intervallumon egyenletes eloszlású $U_1, \ldots, U_{n-2}$ változókhoz tartozó rendezett minta áll.

Bizonyítás. Ezen eloszlásbeli egyenlőséget a már használt (3.8) egyenlőséggel bizonyítjuk. Ehhez legyenek $Y_1, Y_2, \ldots$ független, λ=1 paraméterű exponenciális véletlen változók, (mint a 3.2. Tétel bizonyításában,) és jelölje $S_k := Y_1 + \cdots + Y_k$ a részletösszegeket. Továbbá $U_{1,n} \leq \ldots \leq U_{n,n}$ az $U_1, \ldots, U_n$ a [0,1] intervallumon egyenletes eloszlásból származó mintához tartozó rendezett minta. Jegyezzük meg ismét, hogy ekkor

$$(U_{1,n}, \ldots, U_{n,n}) \overset{\mathcal{D}}{=} \left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}} \right),$$

és ebből következik, hogy

$$\left(\frac{V_{2,n}-V_{1,n}}{V_{n,n}-V_{1,n}},\dots,\frac{V_{n-1,n}-V_{1,n}}{V_{n,n}-V_{1,n}}\right)=\left(\frac{\frac{V_{2,n}-a}{b-a}-\frac{V_{1,n}-a}{b-a}}{\frac{V_{n,n}-a}{b-a}-\frac{V_{1,n}-a}{b-a}},\dots,\frac{\frac{V_{n-1,n}-a}{b-a}-\frac{V_{1,n}-a}{b-a}}{\frac{V_{n,n}-a}{b-a}-\frac{V_{1,n}-a}{b-a}}\right)$$
$$\stackrel{\mathcal{D}}{=}\left(\frac{\frac{S_2}{S_{n+1}}-\frac{S_1}{S_{n+1}}}{\frac{S_n}{S_{n+1}}-\frac{S_1}{S_{n+1}}},\dots,\frac{\frac{S_{n-1}}{S_{n+1}}-\frac{S_1}{S_{n+1}}}{\frac{S_n}{S_{n+1}}-\frac{S_1}{S_{n+1}}}\right)=\left(\frac{Y_2}{Y_2+\dots+Y_n},\dots,\frac{Y_2+\dots+Y_{n-1}}{Y_2+\dots+Y_n}\right)$$
$$\stackrel{\mathcal{D}}{=}\left(\frac{S_1}{S_{n-1}},\dots,\frac{S_{n-2}}{S_{n-1}}\right)=(U_{1,n-2},\dots,U_{n-2,n-2}),$$

ami pontosan az, amit bizonyítani akartunk. □

Tetszőleges $J\geq 1$ esetén tekintsünk a $d_{n1}\leq\ldots\leq d_{nJ}$ távolságszint sorozatokat úgy, hogy minden sorozat teljesíti (T1), (T2) vagy (T3) feltételek valamelyikét. Tegyük fel, hogy a (3.3) feltétel érvényes, és alkalmazzuk rá a (3.1) formulában bevezetett statisztikát az átskálázott $\big((V_{2,n}-V_{1,n})/(V_{n,n}-V_{1,n}),\dots,(V_{n-1,n}-V_{1,n})/(V_{n,n}-V_{1,n})\big)$ mintára. Tehát jelölje $\tilde{K}_{n-2,j}(d_{nj})$ a d_{nj} távolságszinthez tartozó klaszterszámot az átskálázott minta esetén, $j=1,\dots,J$, és legyen

$$\tilde{\mathbf{K}}_{n-2}:=\frac{1}{\sqrt{n}}\left(\frac{\tilde{K}_{n-2,1}(d_{n1})-m_{n-2,1}}{\sigma_{n-2,1}},\dots,\frac{\tilde{K}_{n-2,J}(d_{nJ})-m_{n-2,J}}{\sigma_{n-2,J}}\right)^{\top} \tag{3.22}$$

az átskálázott mintához tartozó normalizált klaszterszám vektor. Legyen továbbá $\tilde{\Sigma}$ a 3.2. Tételben definiált kovarianciamátrix az átskálázott minta esetén. Ekkor (3.21)-ből és a 3.2. Tételből a nullhipotézis mellett következik, hogy

$$C_n^{\mathrm{mod}}:=\tilde{\mathbf{K}}_{n-2}^{\top}\tilde{\Sigma}^{-1}\tilde{\mathbf{K}}_{n-2}\xrightarrow{\mathcal{D}}\chi^2_J. \tag{3.23}$$

Ezek szerint a C_n^{mod} próbastatisztikával tesztelhető a $\mathcal{H}_0$ összetett nullhipotézis. Ezt nevezzünk *módosított klaszterteszt*nek. Mivel a (3.23)-beli konvergencia lassú, hasonlóan C_n-hez, ezért az empirikus kritikus értékek használatát javasoljuk.

3.3.2. A távolságszint sorozatok optimális választása és a kritikus értékek

A C_n és C_n^{mod} tesztstatisztikák nullhipotézis melletti pontos eloszlása túl bonyolult ahhoz, hogy meghatározzuk. Továbbá Csörgő és Wu [23] megmutatták, hogy a 3.2.1. fejezet végén bevezetett távolságszint sorozatok esetén a $K_{nj}(d_{nj})$ statisztika konvergencia sebessége $O(n^{-1/4}\log n)$ vagy rosszabb. Így a konvergencia minden j-re és emiatt együttesen is nagyon lassú. Ezt támasztja alá a szimulációs eredményeket tartalmazó 3.2. táblázat is. Emiatt az aszimptotikus kritikus értékek nem alkalmazhatók a tesztelésre, és ezért szimulációval határoztuk meg a C_n és C_n^{mod} tesztstatisztikák empirikus kritikus értékeit.

Első körben azt vizsgáltuk meg, mely távolságszint sorozatok mellett legnagyobb a C_n ereje. Ehhez számos esetet összehasonlítottunk, a J értéke 2 és 6 között mozgott, a korábbiaknak megfelelően $J_2-J_1\leq 2$, valamint a választott távolságszint sorozatok 3.6. Következménybeliek voltak. Az eredményeket a 3.1. táblázat tartalmazza. A J, az α, a c

3.1. táblázat. A kritikus értékek ($u_{0,05}$) és a C_n klaszter teszt ereje (%-ban megadva) a g_1 és g_2 alternatívákkal szemben különböző J és különböző paraméterű távolságszint sorozatok esetén 0,05 szignifikanciaszint, $n = 100$ mintaméret és 200 000 ismétlés mellett.

J	α	c	β	$u_{0,05}$	$g_1,\ \varrho = 3/2$	$g_2,\ \varrho = 0{,}9,\ j = 5$
2	1,5	-	0,5	6,52	6	14
2	-	1	0,5	6,75	9	56
2	1,5	1	-	6,20	13	70
2	1,3	1	-	6,06	14	77
2	1,1	1	-	6,41	16	84
3	1,5	1	0,5	8,38	10	61
3	1,1	1	0,9	10,68	14	85
4	1,1	1	0,9			
	-	1,7	-	11,96	14	88
4	1,1	0,5	0,9			
	-	1,3	-	12,34	16	88
4	1,9	1	0,1			
	-	-	0,9	12,01	13	85
6	1,9	0,5	0,1			
	1,1	1,3	0,9	16,65	16	87
6	1,9	1	0,1			
	1,1	1,7	0,9	15,74	16	89

és a β oszlop a választott paramétereket jelöli, míg az $u_{0,05}$ oszlopban az adott paraméterekhez és 0,05 szignifikanciaszinthez tartozó empirikus kritikus értékek találhatók.

Ezek után azt vizsgáltuk meg, hogy a kiválasztott paraméter beállítások mellett a C_n tesztnek mekkora az ereje az alábbi két alternatívával szemben, melyeket sűrűségfüggvényükkel definiálunk.

1. $g_1(t) = \begin{cases} 2^{\varrho-1}\varrho t^{\varrho}, \text{ ha } 0 \le t < 1/2 \\ 2^{\varrho-1}\varrho(1-t)^{\varrho}, \text{ ha } 1/2 \le t \le 1, \end{cases}$ ahol $\varrho > 0$,

2. $g_2(t) = 1 + \varrho\cos(\pi j t), t \in [0,1]$, ahol $\varrho \in [-1,1]$.

Az első alternatívát a ϱ=3/2, a második alternatívát a ϱ=0,9, j=5 paraméterrel vizsgáltuk. A klasztertesztet %-ban kifejezett erejét a táblázat utolsó két oszlopa tartalmazza. Látható, hogy a C_n klaszter teszt a legnagyobb erőt a $J = 6$ és $\alpha_1 = 1{,}9, \alpha_2 = 1{,}1, c_1 = 1, c_2 = e - 1, \beta_1 = 0{,}1, \beta_2 = 0{,}9$ paraméter választás esetén éri el. A C_n^{mod} módosított klaszter teszt hasonlóan viselkedik.

Előzetes szimuláció után meghatároztuk a C_n és C_n^{mod} statisztikák α_1=1,9, α_2=1,1, c_1= $= 1, c_2 = e - 1, \beta_1 = 0{,}1, \beta_2 = 0{,}9$ paraméterekhez tartozó kritikus értékeit különböző mintaméretek és szignifikanciaszintek mellett. Az eredményeket a 3.2. táblázat tartalmazza. A táblázat utolsó sora, az $n = \infty$ eset tartalmazza a χ^2_J eloszlásból származó kritikus értékeket, ami a két teszt esetén megegyezik.

3.2. táblázat. A C_n klaszter teszt és a C_n^{mod} módosított klaszter teszt kritikus értékei különböző mintaméret (n) és szignifikanciaszintek (0,10; 0,05 és 0,01) esetén, az ismétlések száma 200 000.

	C_n				C_n^{mod}		
n	0,90	0,95	0,99	n	0,90	0,95	0,99
20	13,05	15,69	21,86	20	16,29	19,32	25,90
50	12,36	15,26	22,32	50	13,33	16,07	22,65
100	12,71	15,74	23,34	100	13,03	15,96	23,22
200	12,77	15,98	23,68	200	12,91	16,00	23,63
500	12,59	15,68	23,24	500	12,65	15,71	23,19
1 000	12,34	15,38	22,64	1 000	12,29	15,34	22,34
				∞	10,65	12,59	16,81

3.3.3. A tesztek ereje

A 3.1. táblázat tartalmazza a C_n és C_n^{mod} statisztikák erejét a g_1 és g_2 alternatívákkal szemben. A következő lépésben azt vizsgáltuk meg, hogy a teszteknek mekkora az ereje az alábbi további alternatívákkal szemben. Minden alternatív eloszlást vagy a sűrűségfüggvényével (g_3, g_4) vagy a kvantilisfüggvényével (G_5^{-1}) adunk meg. A következő alternatívákat használjuk:

3. $g_3(t) = c(\theta^{(j)}) e^{\sum_{k=1}^{j} \theta_k b_k(t)}, t \in [0,1]$, ahol b_k függvények Legendre-polinomok a [0,1] intervallumon (lásd Abramowitz és Stegun [1], 22.7.10), és $\theta^{(j)} = (\theta_1, \ldots, \theta_j) \in \mathbb{R}^n$, $c(\theta^{(j)})$ normalizáló konstans.

4. Béta eloszlással kontaminált egyenletes eloszlás:

$$g_4(t) = 1 - \varrho + \varrho\Gamma(p+q)/(\Gamma(p)+\Gamma(q)) t^{p-1}(1-t)^{q-1}, \qquad t \in [0,1], \qquad \varrho \in [0,1],$$

5. $G_5^{-1}(t) = 1/2 + (t - (1-t)^{\varrho})/2), t \in [0,1]$, ahol $\varrho > 0$.

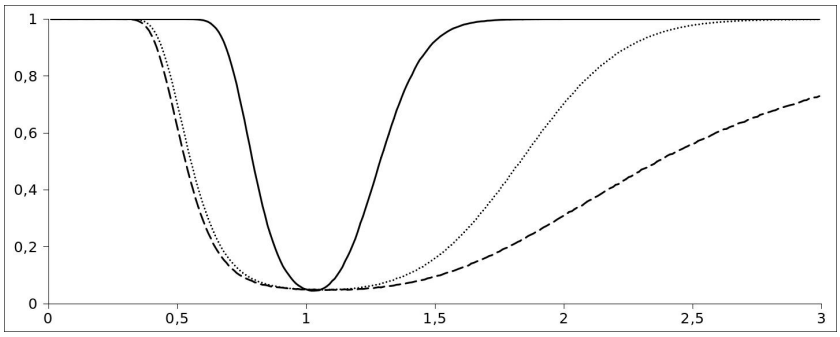
Azért ezeket az alternatívákat választottuk, mert össze akartuk hasonlítani az új C_n és C_n^{mod} teszteket az Inglot és Ledwina [48] által bevezetett „data driven smooth" N_{T1} teszttel, amiről ismert, hogy számos alternatívával szemben nagy erőt képvisel. A hipotézisek vizsgálatát mi magunk nem hajtottuk végre, hanem a [48] cikkből a 2., 3. és 4. táblázatokból vettük. Általában a „data driven smooth" N_{T1} teszt tűnik a legerősebbnek a szimulációs eredmények alapján. A klaszter és a módosított klaszter teszt egyenletesen gyengébben teljesít, kivéve a nagyon oszcilláló sűrűségfüggvénnyel rendelkező alternatívák esetében, ahol a klaszter tesztek majdnem olyan jól vagy jobban viselkednek mint N_{T1} teszt. Például $\varrho = 1{,}00$ és $j = 10$ paraméterű g_2 alternatívával szembeni ereje a két klaszter tesztnek 100% és 99% (lásd 3.3. táblázat).

Ezek után megrajzoltuk a három összehasonlított teszt 1. (3.1. Ábra) és 5. (3.2. Ábra) alternatívával szembeni erőfüggvényét a jobb összehasonlíthatóság céljából. Mindkét esetben az erőt az alternatíva paraméterének függvényében ábrázoltuk a [0,3] intervallum

3.3. táblázat. Az N_{T1}, C_n és C_n^{mod} tesztek ereje (%-ban megadva) az g_2, g_3 és g_4 alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 100$ mintaméret és 200 000 ismétlés mellett.

Alt	ϱ	j	p	q	θ	N_{T1}	C_n	C_n^{mod}
g_2	0,45	1				78	15	12
g_2	0,60	4				71	34	29
g_2	0,75	7				81	62	54
g_2	1,00	10				75	100	99
g_3		2			(-0,2,-0,3)	73	12	9
g_3		5			(0 ;0 ;0 ;0 ;0,4)	76	22	18
g_3		8			(0 ;0 ;0 ;0 ;0 ;0 ;0 ;-0,5)	90	42	36
g_4	0,25		2,0	10,0		73	16	15
g_4	0,50		0,8	1,5		61	10	09
g_4	0,10		0,1	0,1		68	36	26

300 helyen véve a paraméter értékét, 0,05 elsőfajú hibavalószínűség és $n = 100$ mintaméret mellett. Az 5. alternatíva a $\varrho = 0$ esetben a [0,1/2] intervallumon egyenletes eloszlást, míg a $\varrho = 1$ esetben pedig a [0,1] intervallumon egyenletes eloszlást ad, ezért adja a módosított klaszter teszt mindkét paraméter érték mellett az elsőfajú hibavalószínűséget (lásd 3.2. Ábra).

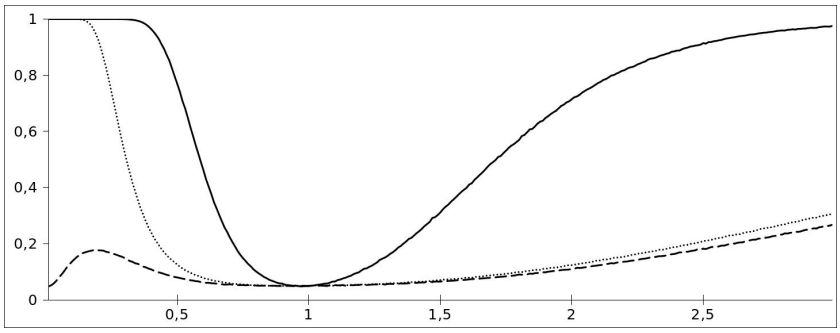


3.1. ábra. Az N_{T1} (vastag vonal), a C_n (pontozott vonal) és a C_n^{mod} (szaggatott vonal) tesztek ereje a g_1 alternatíva ϱ paraméterének függvényében. Az elsőfajú hibavalószínűség 0,05, a mintaméret $n = 100$, az ismétlések száma 200 000.

Mindkét klaszter teszt konzisztencája nehéz kérdés, mivel egy Csörgő és Wu típusú tételt kellene bizonyítani nem egyenletes minta esetén. A szimulációból úgy tűnik, hogy a két új teszt konzisztens, mivel a növekvő mintaméret nagyobb erőt eredményez. Például a C_n teszt ereje a $\varrho = 0{,}80$ és $j = 8$ paraméterű g_2 alternatívával szemben $n = 20, 50$ és 100 mintaméret mellett 26%, 48% és 75%; valamint ugyanezen alternatívával szemben, $\varrho = 1{,}00$ és $j = 12$ paraméterekkel, $n = 20, 50$ és 100 mintaméret mellett 31%, 85% és 100%.

Az erővizsgálat konkluziója, hogy a klaszter tesztek rosszabbul viselkednek, mint más egyenletesség tesztek, kivéve a nagyon oszcilláló alternatívák esetében. Azok a minták

természetüknél fogva jól klaszteresednek, amelyek periodikus sűrűségfüggvényű alternatívából valók.



3.2. ábra. Az N_{T1} (vastag vonal), a C_n (pontozott vonal) és a C_n^{mod} (szaggatott vonal) tesztek ereje a 5. alternatíva ϱ paraméterének függvényében. Az elsőfajú hibavalószínűség 0,05, a mintaméret $n = 100$, az ismétlések száma 200 000.

4. fejezet

Illeszkedésvizsgálat normális eloszláscsaládra

4.1. A kvantilis korrelációteszt normális eloszláscsalád esetében

Ebben a fejezetben az a célunk, hogy a normális eloszláscsaládhoz való illeszkedést teszteljük. Erre a célra a del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] által bevezetett normalitástesztet fogjuk használni, mely az eloszlások L^2-Wasserstein-távolságán alapul. Jegyezzük meg, hogy ezt a tesztet általánosította del Barrio, Cuesta-Albertos és Matrán [33] nem normális eloszláscsaládokra is.

Legyen $\mathcal{P}_2(\mathbb{R})$ azon valószínűségi mértékek halmaza $\mathbb{R}$-en, melyeknek létezik a második momentumuk. A P_1 és $P_2 \in \mathcal{P}_2(\mathbb{R})$ valószínűségi mértékek L^2-Wasserstein távolsága

$$\mathcal{W}(P_1, P_2) := \inf \left\{ [E(X_1 - X_2)^2]^{1/2} : \mathcal{L}(X_1) = P_1, \mathcal{L}(X_2) = P_2 \right\},$$

ahol $\mathcal{L}(X)$ az X véletlen változó eloszlását jelöli. Kvantilisfüggvények segítségével pontosan számolható ez a távolság (lásd például Bickel és Freedman [7]):

$$\mathcal{W}(P_1, P_2) = \left[\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 \mathrm{d}t \right]^{1/2},$$

ahol F_1^{-1} illetve F_2^{-1} a P_1 illetve a P_2 eloszlásokhoz tartozó kvantilisfüggvények.

Egy eloszláscsalád és egy adott eloszlás távolságát úgy definiáljuk, mint az adott eloszlásnak az eloszláscsalád tagjától vett távolságainak infimumát. Legyen $P \in \mathcal{P}_2(\mathbb{R})$ tetszőleges valószínűségi mérték, és legyen az eloszlásfüggvénye F, várható értéke μ_0 és a szórása σ_0. Jegyezzük meg, hogy ekkor

$$\int_0^1 F^{-1}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x \mathrm{d}F(x) = \mu_0 \qquad \text{és} \qquad \int_0^1 \left(F^{-1}(t)\right)^2 \mathrm{d}t = \int_{-\infty}^{\infty} x^2 \mathrm{d}F(x) = \sigma_0^2 + \mu_0^2.$$

Ekkor a P eloszlás távolságnégyzete az $\mathbf{N}$ normális eloszláscsaládtól

$$\mathcal{W}^2(P,\mathbf{N}) := \inf\{\mathcal{W}^2(P,N_\sigma^\mu) : N_\sigma^\mu \in \mathbf{N}\} = \inf_{\substack{\mu\in\mathbb{R}\\ \sigma>0}} \int_0^1 \Big(F^{-1}(t) - \big(\mu+\sigma\Phi^{-1}(t)\big)\Big)^2 \,\mathrm{d}t$$

$$= \inf_{\substack{\mu\in\mathbb{R}\\ \sigma>0}} \left\{ \int_0^1 \big(F^{-1}(t)\big)^2 \,\mathrm{d}t - 2\int_0^1 F^{-1}(t)\big(\mu+\sigma\Phi^{-1}(t)\big)\,\mathrm{d}t + \int_0^1 \big(\mu+\sigma\Phi^{-1}(t)\big)^2\,\mathrm{d}t \right\}$$

$$= \inf_{\substack{\mu\in\mathbb{R}\\ \sigma>0}} \left\{ (\sigma_0^2+\mu_0^2) - 2\mu_0\mu - 2\sigma\int_0^1 F^{-1}(t)\Phi^{-1}(t)\,\mathrm{d}t + (\sigma^2+\mu^2) \right\}$$

$$= \inf_{\substack{\mu\in\mathbb{R}\\ \sigma>0}} \left\{ (\mu_0-\mu)^2 + \sigma_0^2 + \left(\sigma - \int_0^1 F^{-1}(t)\Phi^{-1}(t)\,\mathrm{d}t\right)^2 - \left(\int_0^1 F^{-1}(t)\Phi^{-1}(t)\,\mathrm{d}t\right)^2 \right\}$$

$$= \sigma_0^2 - \left(\int_0^1 F^{-1}(t)\Phi^{-1}(t)\mathrm{d}t\right)^2.$$

A számolásból látható, hogy P ahhoz a normális eloszláshoz van a legközelebb, amelyiknek $\mu=\mu_0$ a várható értéke és $\sigma=\int_0^1 F^{-1}(t)\Phi^{-1}(t)\mathrm{d}t$ a szórása. Megjegyezzük, hogy a $\mathcal{W}^2(P,\mathbf{N})/\sigma_0^2$ hányadosra nincs hatással P eltolás illetve skála változása. Ennélfogva jó mértéke lehet a nem-normalitásnak, ugyanis minél nagyobb ennek a törtnek az értéke, a P annál távolabb van a normális eloszláscsaládtól.

Ha adott egy F eloszlásfüggvényű $X_1,\dots,X_n$ véletlen minta, akkor a $\mathcal{H}_0 : F\in\mathbf{N}$ összetett nullhipotézis ellenőrzésére megadható a $\mathcal{W}(P,\mathbf{N})/\sigma_0$ hányados empirikus változata. Az empirikus változatot úgy definiáljuk, hogy a P eloszlását az empirikus eloszlással helyettesítjük. Ekkor egy eltolás- és skálamentes statisztikát kapunk:

$$T_n := \frac{\mathcal{W}^2(F_n,\mathbf{N})}{S_n^2} = 1 - \frac{\left[\int_0^1 Q_n(t)\Phi^{-1}(t)dt\right]^2}{S_n^2} = 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi^{-1}(t)\,\mathrm{d}t\right]^2}{S_n^2}. \quad (4.1)$$

Érdekessége a tesztnek, hogy az illeszkedésvizsgálat két nagy osztályához is tartozik. Egyrészt úgy tesztel eloszláscsaládhoz való tartozást, hogy a mimimum távolság módszerét használja. Másrészt legyen $\boldsymbol{\nu}_n = (\nu_{1n},\dots,\nu_{nn})^\top$ a $\nu_{kn} = \int_{(k-1)/n}^{k/n}\Phi^{-1}(t)\,\mathrm{d}t, k=1,\dots,n$, komponensekből álló vektor. Ekkor a 2.2.2. fejezetben használt jelöléssekkel tekintsük a

$$\rho^2(\boldsymbol{\nu}_n,\mathbf{X}_n) = \frac{\big(n\cdot\boldsymbol{\nu}_n{}^\top\mathbf{X}_n - 1^\top\boldsymbol{\nu}_n\cdot 1^\top\mathbf{X}_n\big)^2}{\big(n\cdot\boldsymbol{\nu}_n{}^\top\boldsymbol{\nu}_n - (1^\top\boldsymbol{\nu}_n)^2\big)\big(n\cdot\mathbf{X}_n^\top\mathbf{X}_n - (1^\top\mathbf{X}_n)^2\big)}$$

statisztikát. Mivel a standard eloszlás várható értéke 0, teljesül az $1^\top\boldsymbol{\nu}_n = \int_{-\infty}^{\infty}\Phi^{-1}(t)\,\mathrm{d}t = 0$ egyenlőség, amiből

$$\rho^2(\boldsymbol{\nu}_n,\mathbf{X}_n) = \frac{\big(\sum_{k=1}^n \nu_{kn}X_{k,n}\big)^2}{n\boldsymbol{\nu}_n^\top\boldsymbol{\nu}_n S_n^2}.$$

Mivel a $\rho^2(\boldsymbol{\nu}_n,\mathbf{X}_n)$ egy korrelációteszt, és mivel $\boldsymbol{\nu}_n^\top\boldsymbol{\nu}_n\to 1$, ezért a „spanyolok” T_n tesztje aszimptotikusan ekvivalens ezzel a korrelációteszttel.

Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] megvizsgálták a tesztstatisztika nullhipotézis melletti aszimptotikus viselkedését. Két alakban sikerült előállítaniuk a határeloszlást. Az első Brown-híd funkcionáljaként, a második véletlen változók soraként. Jelölje φ a standard normális sűrűségfüggvényét, ekkor az eredményüket a következő tételben foglaljuk össze.

4.1. Tétel (del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34]). *Legyen*

$$a_n = \frac{1}{n}\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{\left[\varphi(\Phi^{-1}(t))\right]^2}dt.$$

Ha $F \in \mathbf{N}$, *akkor*

$$n(T_n - a_n) \xrightarrow{\mathcal{D}} \int_0^1 \frac{B^2(t)-E(B^2(t))}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t - \left[\int_0^1 \frac{B(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 - \left[\int_0^1 \frac{B(t)\Phi^{-1}(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2$$
$$\stackrel{\mathcal{D}}{=} -\frac{3}{2} + \sum_{j=3}^{\infty}\frac{Z_j^2-1}{j},$$

ahol $(Z_j)_{j=3}^{\infty}$ *független, standard normális eloszlású véletlen változók sorozata.*

A következő fejezetben ennek a tesztnek az erővizsgálata található.

4.2. Szimuláció

4.2.1. A határeloszlás és a szimulált kritikus értékek

A 4.1. Tételben szereplő határ véletlen változó eloszlásfüggvényét kétféleképpen számítottuk ki, mindkét esetben numerikusan a határ véletlen változó soros alakjából kiindulva. Az első alkalommal a határ változó karakterisztikus függvényéből indultunk ki, és de Wet és Venter [31] technikáját használtuk. Meghatároztuk a határ véletlen változó karakterisztikus függvényét, majd numerikus inverzióval kaptuk a határ eloszlásfüggvényt. Jelölje ϕ az aszimptotikus karakterisztikus függvényt. Ekkor a függetlenség, a majoráns konvergenciatétel és a χ_1^2 eloszlás karakterisztikus függvényének alkalmazásával

$$\phi(t) = E\left(e^{it\left(-\frac{3}{2}+\sum_{j=3}^{\infty}\frac{Z_j^2-1}{j}\right)}\right) = e^{-\frac{3}{2}it}\prod_{j=3}^{\infty}e^{-i\frac{t}{j}}E\left(e^{it\frac{Z_j^2}{j}}\right) = e^{-\frac{3}{2}it}\prod_{j=3}^{\infty}e^{-i\frac{t}{j}}\frac{1}{\sqrt{1-2i\frac{t}{j}}},$$

minden $t \in \mathbb{R}$ esetén. Szeretnénk olyan alakban felírni ezt a karakterisztikus függvényt, amely számítógépes numerikus számolással könnyebben megkapható. Ehhez keressük $\phi(t) = r(t)e^{i\vartheta(t)}$ alakban, ahol $r(t) = |\phi(t)|$ az origótól való távolsága és $\vartheta(t)$ a $\phi(t)$ komplex szám irányszöge. Elemi számolásból ellenőrizhető, hogy $\left|(1-2it/j)^{1/2}\right| = (1+4t^2/j^2)^{1/4}$, továbbá $\sinh(x) = x\prod_{k=1}^{\infty}(1+x^2/(k^2\pi^2))$, $x \in \mathbb{R}$, (lásd Abramowitz és Stegun [1], 4.5.68). Ezekből azt kapjuk, hogy

$$r(t) = \left|e^{-\frac{3}{2}it}\prod_{j=3}^{\infty}e^{-i\frac{t}{j}}\frac{1}{\sqrt{1-2i\frac{t}{j}}}\right| = \prod_{j=3}^{\infty}\frac{1}{\left|\sqrt{1-2i\frac{t}{j}}\right|} = \prod_{j=3}^{\infty}\left(1+4\frac{t^2}{j^2}\right)^{-\frac{1}{4}}$$
$$= (\sinh(2\pi t))^{-\frac{1}{4}}\left(2\pi t(1+4t^2)(1+t^2)\right)^{\frac{1}{4}}, \qquad t \in \mathbb{R}.$$

Továbbá, mivel az $e^{-i\frac{3}{2}t}, e^{-i\frac{t}{j}}, (1-2it/j)^{-1/2}$ komplex számokhoz rendre a $-3t/2, -t/j, 1/2\arctan(2t/j)$ irányszögek tartoznak, ezért

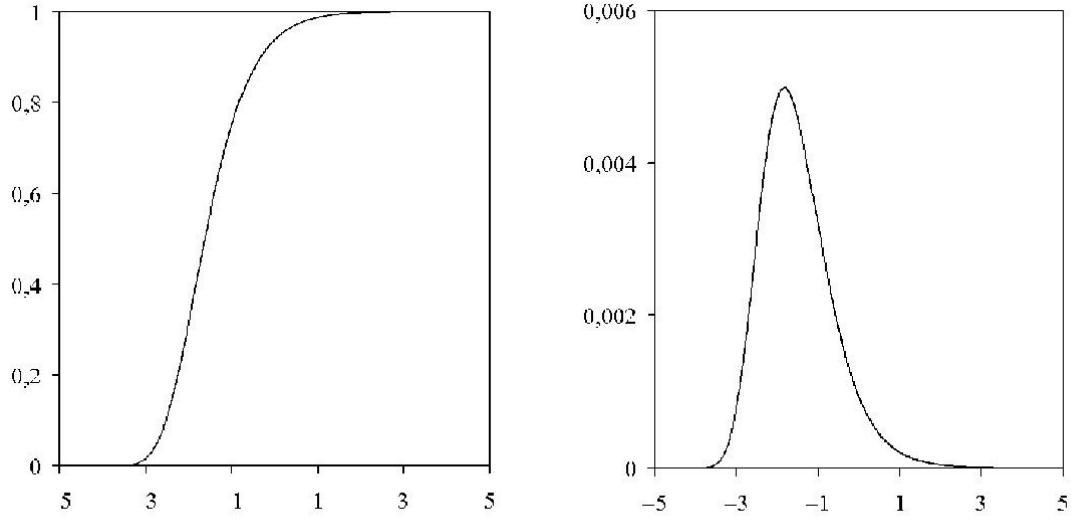
$$\vartheta(t) = -\frac{3}{2}t + \frac{1}{2}\sum_{j=3}^{\infty}\left(\arctan\frac{2t}{j} - \frac{2t}{j}\right), \qquad t \in \mathbb{R}. \tag{4.2}$$

A r sugárfüggvény alakjából könnyen látható, hogy $\int_{\mathbb{R}} |t|^k r(t)dt < \infty$ minden $k = 1,2,\ldots$ esetén, így a H határ eloszlásfüggvény végtelen sokszor differenciálható, és az inverziós formula szerint előáll

$$H(y) - H(0) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{\phi(t)}{it}\left(1 - e^{-ity}\right)dt = \frac{1}{\pi}\int_0^{\infty} \frac{r(t)}{t}\left[\sin\vartheta(t) + \sin\left(ty - \vartheta(t)\right)\right]dt,$$

$y \in \mathbb{R}$, alakban. Ezt az integrált számítógép segítségével számoltuk ki következő módon. A ϑ függvényt a (4.2) sor első 10 000 tagjával közelítettük, és a fenti improprius integrált 0 és 100 között numerikusan integráltuk. Első körben y értékét kellően nagynak választva, közelítőleg megkaptuk $1 - H(0)$ értékét, amiből $H(0)$ már könnyen számolható volt. Ezek után a H függvényt a $[-5,5]$ intervallumon meg tudtuk határozni. Az eredményekből kitűnik, nem érdemes bővebb intervallumon dolgozni, ugyanis $H(-5) \approx 0$ és $H(5) \approx 1$.

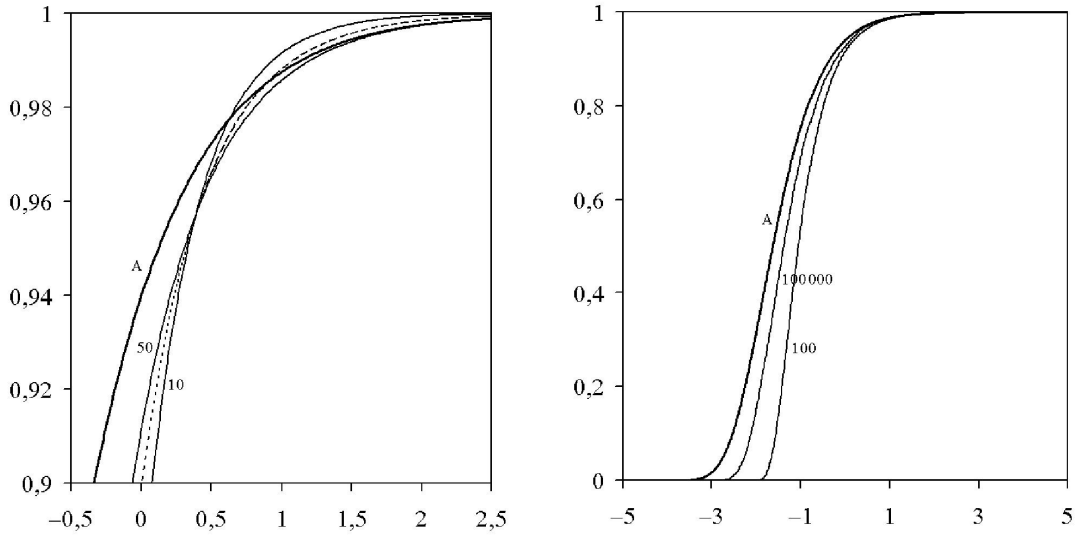
A másik út, ahogyan a H eloszlásfüggvényt meghatároztuk, magának a 4.1. Tételbeli határ véletlen változónak a szimulációja volt, természetesen a soros alakot használva. Számítógépen legeneráltuk a változót 1 000 000 példányban úgy, hogy a változót definiáló sor első 5 000 tagját vettük, majd felírtuk a kapcsolatos empirikus eloszlásfüggvényt. A két eljárásból származó empirikus eloszlásfüggvények 3 tizedes pontossággal megegyeztek, ami arra utal, hogy sikerült nagy pontossággal meghatározni a H elméleti eloszlásfüggvényt. A vizsgálat eredményeit a 4.1. ábra tartalmazza.



4.1. ábra. Az aszimptotikus eloszlásfüggvény (balra) és a sűrűségfüggvény (jobbra)

Ezek után a del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] által bevezetett $n(T_n - a_n)$ tesztstatisztikát vizsgáltuk meg, melyet a továbbiakban a szerzők

neve után csak BCMR-próbának fogunk nevezni. Először a tesztstatisztika empirikus eloszlásfüggvényét határoztuk meg, abból a célból, hogy megkapjuk a statisztika kritikus értékeit. Az $n \leq 500$ mintaméretekre 1 000 000 generálást végeztünk, és ezek alapján írtuk fel az eloszlásfüggvényt. Az 500-nál nagyobb mintaméretek esetén, a futási idő kordában tartása céljából, az ismétlések számát fokozatosan csökkentettük, de minden mintaméret esetén legalább 5 000-szer elvégeztük. Az N oszlopa tartalmazza az ismétlések számát. A kapott empirikus eloszlásfüggvényt a 4.2. ábra, a kapcsolatos kritikus értékeket a 4.1. táblázat tartalmazza. Az ábrán és a táblázatban az előző bekezdésben kapott aszimptotikus eloszlásfüggvényt és kritikus értéket is feltüntettük. Látható, hogy a konvergencia sebessége mindenhol lassú. Ez különösen igaz, a kicsi kvantilisekre, de a tesztelés szempontjából fontos kvantilisek magas tartományában is.



4.2. ábra. Az $n(T_n - a_n)$ BCMR tesztstatisztika eloszlásfüggvénye $n = 10$, 20 (pontozott vonal), 50 mintaméret esetén és az A-val jelölt vastagabb vonal bal oldalon az aszimptotikus eloszlásfüggvény (balra). Ugyanez $n = 100$ és 100 000 mintaméret esetén (jobbra).

Amint a 4.1. táblázatból kitűnik a 0,15, 0,10 és 0,05 szignifikanciaszintek esetén az aszimptotikus kritikus érték alacsonyabb, mint az adott mintaméretthez tartozó kritikus érték. Ez azt jelenti, hogy a tesztelés során az aszimptotikus kritikus értékeket használva az elvetések aránya magasabb, mintha az adott mintaméretthez tartozó egzakt kritikus értékeket használnánk, vagyis a teszt elsőfajú hibája nagyobb a tervezettnél. A kritikus értékek hasonlóan viselkednek a 0,01 szignifikanciaszint esetén, ha a mintaméret nagyobb, mint 35. Mindez arra világít rá, hogy helyesebb az adott mintaméretthez tartozó kritikus értéket használni.

4.2.2. A teszt erejének vizsgálata

Egy szimulációs vizsgálatban kiértékeltük a BCMR-teszt erejét, és hét másik normalitás teszttel hasonlítottunk össze. A hétből öt tesztnek az erejét szimulációs tanulmány

4.1. táblázat. Az $n(T_n - a_n)$ BCMR tesztstatisztika kritikus értékei 0,15; 0,10; 0,05 és 0,01 szignifikanciaszintek esetén.

n	N	0,85	0,90	0,95	0,99
10	1 000 000	-0,08	0,07	0,32	0,93
15	1 000 000	-0,15	0,02	0,31	1,02
20	1 000 000	-0,19	0,00	0,31	1,07
35	1 000 000	-0,25	-0,05	0,30	1,15
50	1 000 000	-0,28	-0,07	0,30	1,19
100	1 000 000	-0,33	-0,10	0,29	1,24
200	1 000 000	-0,37	-0,13	0,29	1,27
500	1 000 000	-0,40	-0,15	0,28	1,29
1 000	200 000	-0,42	-0,16	0,27	1,29
2 000	100 000	-0,44	-0,18	0,26	1,27
5 000	100 000	-0,45	-0,20	0,25	1,27
10 000	100 000	-0,46	-0,20	0,26	1,32
20 000	100 000	-0,46	-0,21	0,24	1,23
50 000	5 000	-0,49	-0,22	0,21	1,17
100 000	5 000	-0,49	-0,21	0,22	1,18
∞	1 000 000	-0,63	-0,35	0,11	1,13

keretei között mi magunk vizsgáltuk meg, az utolsó két tesztre vonatkozó eredményeket más forrásból gyűjtöttük össze. Ezen tesztek közül az első Shapiro–Wilk W-tesztje [65], amit $n = 20$ és $n = 50$ esetén használtuk az összehasonlításban. Ez a teszt azért is különösen érdekes, mert a BCMR-teszt és W-teszt aszimptotikusan ekvivalens. Mivel a W-teszt együtthatói az $n = 100$ mintaméret esetén nagyon nehezen számolhatók, ezért ebben az esetben a Shapiro–Francia [63] W'-tesztet használtuk. Az EDF-tesztek közül a Kolmogorov–Smirnov [51] D-teszt Stephens [71] által javasolt módosított változatát, és az Anderson–Darling [4] A^2-tesztet választottuk. A negyedik teszt, amit bevettünk az összehasonlításba, egy sűrűségbecslésre alapozott teszt, Bowman és Foster [9] integrált négyzetes hiba ISE-tesztje fix maggal. Az ötödik teszt Epps és Pulley [38] BHEP-tesztje α=1 paraméterrel, ami az empirikus karakterisztikus függvényt használja. Végül bevettük az összehasonlításba Kallenberg és Ledwina [50] „data driven smooth” tesztjét és Cabaña és Cabaña [11] „focused” tesztjét. Mivel az utolsó két publikációban az alternatíváknak meglehetősen széles halmazára számítanak erőket, ezért n=20 és n=50 mintaméret esetén a megfelelő Table V és Table 4 táblázatokból vettük az erőket.

A szimulációs vizsgálatba azon alternatív eloszlásokat vettünk be, amelyeket Shapiro, Wilk és Chen [63] valamint Gan és Koehler [42] használtak az ő szimulációs vizsgálatukban. Jelölje U illetve Z a [0,1] intervallumon egyenletes illetve a standard normális eloszlású véletlen változót.

Az alternatív eloszlások:

1. Beta(p,q), $p,q>0$, jelölje a Béta eloszlást, melynek sűrűségfüggvénye

$$f(t)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}t^{p-1}(1-t)^{q-1}, \qquad t\in(0,1),$$

ahol $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x$, $\alpha\in(0,\infty)$.

2. CN(λ,σ^2), $0<\lambda<1$ és $\sigma>0$ paraméterekkel a kontaminált normális eloszlás, amely a következő eloszlásfüggvénnyel van definiálva

$$F(x)=(1-\lambda)\Phi(x)+\lambda\Phi(x/\sigma), \qquad x\in\mathbb{R}.$$

3. A Gumbel-eloszlás, melynek eloszlásfüggvénye $F(x)=1-e^{-e^x}$, $x\in\mathbb{R}$.

4. A HalfN(0,1) eloszlás, amely a $|Z|$ véletlen változó eloszlása.

5. A Laplace-eloszlás, melynek sűrűségfüggvénye $f(t)=e^{-|t|/2}$, $t\in\mathbb{R}$.

6. A Lognormal (magyarul lognormál) eloszlás a e^Z véletlen változó eloszlása.

7. A Logistic (magyarul logisztikus) eloszlás, melynek sűrűségfüggvénye

$$f(t)=e^t(1+e^t)^{-2},\ t\in\mathbb{R}.$$

8. SB(γ,δ), $\gamma\in\mathbb{R},\delta>0$, egy korlátos Johnson-eloszlás, a $e^{(Z-\gamma)/\delta}/(1+e^{(Z-\gamma)/\delta})$ véletlen változó eloszlása, valamint SU(γ,δ) egy nemkorlátos Johnson-eloszlás, melynek eloszlása a $\sinh((Z-\gamma)/\delta)$ véletlen változó eloszlása.

9. Két háromszög eloszlás, Triangle(I) és Triangle(II), melyek rendre az alábbi sűrűségfüggvényekkel vannak definiálva:

$$f(t)=1-|t|,\ t\in[-1,1], \qquad \text{és} \qquad f(t)=2-2t,\ t\in[0,1].$$

10. TruncN(a,b), $a,b\in\mathbb{R}$, $a<b$, legyen az a és b helyeken levágott standard normális eloszlás, a $Z\,\mathbf{I}_{\{a\le Z\le b\}}$ eloszlása.

11. T(λ), $\lambda>0$, a Tukey-eloszlás, az $U^\lambda-(1-U)^\lambda$ véletlen változó eloszlása.

12. A Weibull(k), $k>0$, eloszlás, melynek sűrűségfüggvénye

$$f(t)=kt^{k-1}e^{-t^k}, \qquad t>0.$$

A bemutatott teszteket a fenti alternatívákkal szemben egy olyan szimulációs vizsgálat során tanulmányoztuk, ahol a szignifikanciaszint 0,10 és 0,05, a mintaméret $n=20,50$ és 100 volt, és 200 000 mintát generáltunk le. A vizsgált alternatívák a $\sqrt{\beta_1}$ ferdeség és β_2 lapultság értékek alapján lettek besorolva a táblázatokba.

Az alábbi táblázatokban találhatók ezen szimulációs vizsgálat eredményei.

4.2. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje szimmetrikus vékony szélű, közel normális és vastag szélű alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 20$ mintaméret és 200 000 ismétlés mellett.

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(0,5;0,5)	0	1,50	67	73	50	47	32	59
Beta(1;1)	0	1,80	16	20	14	13	10	16
Beta(1,1;1,1)	0	1,85	13	16	12	11	8	13
Beta(1,3;1,3)	0	1,93	9	11	9	8	7	9
Beta(1,5;1,5)	0	2,00	7	8	7	6	6	8
Beta(2;2)	0	2,14	4	5	5	5	5	5
T(0,7)	0	1,92	9	11	9	8	7	10
T(1,5)	0	1,75	21	25	17	16	12	20
T(3)	0	2,06	5	6	5	5	5	6
SB(0; 0,5)	0	1,63	38	44	30	28	19	35
SB(0; 0,707)	0	1,87	12	14	11	10	9	13
Triangle(I)	0	2,40	3	3	3	3	4	4
TruncN(-1;1)	0	1,94	8	10	8	8	7	9
TruncN(-2;2)	0	2,36	4	3	4	4	4	4
TruncN(-3;3)	0	2,84	4	4	4	4	5	4
T(0,1)	0	3,21	6	6	6	6	6	6
SU(0; 3)	0	3,53	8	8	7	7	6	7
SU(0; 2)	0	4,51	13	12	12	12	9	11
Logistic	0	4,20	13	12	11	11	9	10
Student(10)	0	4,00	10	10	9	9	7	9
T(10)	0	5,38	82	81	79	72	90	90
Laplace	0	6,00	28	26	28	27	22	26
SU(0; 1)	0	36,2	44	43	47	42	35	41
SU(0; 0,9)	0	82,1	52	50	52	50	43	49
Cauchy	0	∞	88	87	88	87	84	87
Student(2)	0	∞	55	53	54	53	45	51
Student(3)	0	∞	35	34	34	34	26	32
Student(4)	0	∞	25	24	23	23	18	22
Student(5)	0	9,00	20	19	18	18	13	16

4.3. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje aszimmetrikus vékony- és vastag szélű alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 20$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(2;1)	-0,57	2,40	28	30	25	26	18	18
Beta(3;2)	-0,29	2,36	7	7	7	7	7	5
TruncN(-2;1)	-0,32	2,27	8	10	9	9	7	6
TruncN(-3;1)	-0,55	2,78	14	14	12	12	9	7
TruncN(-3;2)	-0,18	2,65	4	4	4	4	5	3
Weibull(4)	-0,09	2,75	4	4	4	5	5	4
Weibull(3,6)	0,00	2,72	4	4	4	4	5	4
Weibull(2,0)	0,63	3,25	15	15	14	15	10	18
SB(0,533; 0,5)	0,65	2,13	69	72	59	59	44	70
SB(1; 1)	0,73	2,91	29	30	27	29	19	34
SB(1; 2)	0,28	2,77	6	6	6	6	6	8
Half N(0;1)	0,97	3,78	43	44	37	39	24	44
SU(1; 1)	-5,37	93,4	73	73	72	73	61	62
SU(1; 2)	-0,87	5,59	21	20	20	21	15	12
Triangle(II)	0,57	16,4	28	30	25	26	18	32
Gumbel	1,14	5,40	31	32	31	29	20	17
χ^2_n	2,83	15,0	98	98	95	96	88	98
Exp(1/2)	2,00	9,00	83	84	76	78	58	84
χ^2_4	1,41	6,00	52	53	48	50	33	56
Lognormal	6,18	113,9	93	93	90	91	79	93
Weibull(0,5)	6,62	87,7	*	*	99	99	98	*

4.4. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje szimmetrikus vékony szélű, közel normális és vastag szélű alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 50$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(0.5,0.5)	0	1.50	*	*	98	98	80	99
Beta(1,1)	0	1.80	68	88	58	54	26	56
Beta(1.1,1.1)	0	1.85	58	81	50	46	22	47
Beta(1.3,1.3)	0	1.93	39	65	37	33	16	33
Beta(1.5,1.5)	0	2.00	27	50	28	24	12	24
Beta(2,2)	0	2.14	12	27	16	13	8	13
$T(0.7)$	0	1.92	40	67	38	34	17	34
$T(1.5)$	0	1.75	80	94	67	63	32	66
$T(3)$	0	2.06	21	45	18	15	8	16
SB(0,0.5)	0	1.63	96	99	89	87	55	90
SB(0,0.707)	0	1.87	52	75	48	43	21	44
Triangle(I)	0	2.40	4	9	6	5	4	5
TruncN(-1,1)	0	1.94	39	64	33	29	14	30
TruncN(-2,2)	0	2.36	4	10	6	5	5	5
TruncN(-3,3)	0	2.84	3	5	5	4	5	4
$T(0.1)$	0	3.21	7	6	6	6	6	6
SU(0,3)	0	3.53	11	8	8	9	7	8
SU(0,2)	0	4.51	23	16	16	18	12	17
Logistic	0	4.20	22	14	15	16	12	16
Student(10)	0	4.00	16	12	11	13	9	12
$T(10)$	0	5.38	*	99	*	98	*	*
Laplace	0	6.00	55	42	54	52	44	54
SU(0,1)	0	36.2	78	67	75	76	65	75
SU(0,0.9)	0	82.1	86	79	85	85	76	85
Cauchy	0	∞	*	99	*	*	99	*
Student(2)	0	∞	87	81	86	86	78	85
Student(3)	0	∞	66	56	59	61	49	60
Student(4)	0	∞	49	38	40	43	31	41
Student(5)	0	9.00	37	28	29	31	21	30

4.5. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje aszimmetrikus vékony- és vastag szélű alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 50$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(2,1)	-0.57	2.40	81	90	69	72	46	64
Beta(3,2)	-0.29	2.36	18	30	19	20	12	14
TruncN(-2,1)	-0.32	2.27	29	47	26	27	15	20
TruncN(-3,1)	-0.55	2.78	43	53	30	35	20	24
TruncN(-3,2)	-0.18	2.65	5	8	6	6	5	4
Weibull(4)	-0.09	2.75	4	6	6	5	5	4
Weibull(3.6)	0.00	2.72	3	5	5	4	5	4
Weibull(2.0)	0.63	3.25	40	44	30	36	21	38
SB(0.533,0.5)	0.65	2.13	*	99	98	98	90	99
SB(1,1)	0.73	2.91	77	84	67	72	47	75
SB(1,2)	0.28	2.77	9	12	10	11	8	12
Half N(0,1)	0.97	3.78	93	95	79	83	57	89
SU(1,1)	-5.37	93.4	98	97	97	98	94	96
SU(1,2)	-0.87	5.59	44	39	36	41	28	30
Triangle(II)	0.57	16.4	81	90	69	71	45	77
Gumbel(0,1)	1.14	5.40	68	68	57	65	44	50
χ^2_1	2.83	15.0	*	*	*	*	*	*
Exp(1/2)	2.00	9.00	*	*	99	*	96	*
χ^2_4	1.41	6.00	94	96	87	91	70	93
Lognormal	6.18	113.9	*	*	*	*	*	*
Weibull(0.5)	6.62	87.7	*	*	*	*	*	*

4.6. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje szimmetrikus vékony szélű, közel normális és vastag szélű alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 100$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(0.5,0.5)	0	1.50	*	*	*	*	99	*
Beta(1,1)	0	1.80	*	96	95	95	59	96
Beta(1.1,1.1)	0	1.85	98	89	89	89	50	90
Beta(1.3,1.3)	0	1.93	90	70	76	76	36	75
Beta(1.5,1.5)	0	2.00	76	49	61	62	26	59
Beta(2,2)	0	2.14	40	18	35	34	15	32
T(0.7)	0	1.92	92	72	77	78	37	77
T(1.5)	0	1.75	*	99	98	98	69	99
T(3)	0	2.06	69	39	42	42	15	43
SB(0,0.5)	0	1.63	*	*	*	*	92	*
SB(0,0.707)	0	1.87	96	84	87	87	48	87
Triangle(I)	0	2.40	8	3	9	8	5	8
TruncN(-1,1)	0	1.94	90	69	73	72	31	72
TruncN(-2,2)	0	2.36	10	3	10	9	6	9
TruncN(-3,3)	0	2.84	3	2	5	5	5	5
T(0.1)	0	3.21	7	9	6	7	6	7
SU(0,3)	0	3.53	15	18	9	10	8	10
SU(0,2)	0	4.51	36	41	23	27	17	26
Logistic	0	4.20	33	37	22	25	16	24
Student(10)	0	4.00	25	28	14	17	11	16
T(10)	0	5.38	*	*	*	*	*	*
Laplace	0	6.00	81	84	82	80	70	82
SU(0,1)	0	36.2	96	97	95	95	89	95
SU(0,0.9)	0	82.1	98	99	98	98	96	98
Cauchy	0	∞	*	*	*	*	*	*
Student(2)	0	∞	99	99	98	99	96	98
Student(3)	0	∞	89	90	84	86	73	85
Student(4)	0	∞	73	76	62	67	49	65
Student(5)	0	9.00	58	63	43	50	33	48

4.7. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje aszimmetrikus vékony- és vastag szélű alternatívákkal szemben 0,05 szignifikanciaszint esetén, $n = 100$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(2,1)	-0.57	2.40	*	*	97	98	82	97
Beta(3,2)	-0.29	2.36	49	30	40	45	23	35
TruncN(-2,1)	-0.32	2.27	77	55	55	61	30	53
TruncN(-3,1)	-0.55	2.78	87	76	59	69	28	59
TruncN(-3,2)	-0.18	2.65	8	4	8	8	7	6
Weibull(4)	-0.09	2.75	5	3	7	7	6	5
Weibull(3.6)	0.00	2.72	4	2	6	5	5	5
Weibull(2.0)	0.63	3.25	77	67	53	66	39	67
SB(0.533,0.5)	0.65	2.13	*	*	*	*	*	*
SB(1,1)	0.73	2.91	99	98	95	97	81	98
SB(1,2)	0.28	2.77	18	13	15	19	13	20
Half N(0,1)	0.97	3.78	*	*	98	99	91	*
SU(1,1)	-5.37	93.4	*	*	*	*	*	*
SU(1,2)	-0.87	5.59	70	71	56	66	48	56
Triangle(II)	0.57	16.4	*	99	97	98	82	99
Gumbel(0,1)	1.14	5.40	94	92	84	92	73	84
χ^2_1	2.83	15.0	*	*	*	*	*	*
Exp(1/2)	2.00	9.00	*	*	*	*	*	*
χ^2_4	1.41	6.00	*	*	99	*	95	*
Lognormal	6.18	113.9	*	*	*	*	*	*
Weibull(0.5)	6.62	87.7	*	*	*	*	*	*

4.8. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje szimmetrikus vékony szélű, közel normális és vastag szélű alternatívákkal szemben 0,10 szignifikanciaszint esetén, $n = 20$ mintaméret és 200 000 ismétlés mellett.

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(0.5,0.5)	0	1.50	82	85	68	68	48	74
Beta(1,1)	0	1.80	31	36	27	27	19	28
Beta(1.1,1.1)	0	1.85	26	30	23	23	17	24
Beta(1.3,1.3)	0	1.93	19	21	18	18	14	19
Beta(1.5,1.5)	0	2.00	15	18	15	15	12	15
Beta(2,2)	0	2.14	11	12	12	11	10	12
T(0.7)	0	1.92	20	23	18	19	14	19
T(1.5)	0	1.75	37	42	32	32	21	33
T(3)	0	2.06	12	15	12	12	10	12
SB(0,0.5)	0	1.63	57	63	48	48	32	51
SB(0,0.707)	0	1.87	24	28	23	23	17	23
Triangle(I)	0	2.40	7	8	8	7	8	8
TruncN(-1,1)	0	1.94	18	21	17	17	13	17
TruncN(-2,2)	0	2.36	8	8	8	8	9	9
TruncN(-3,3)	0	2.84	9	9	9	9	9	9
T(0.1)	0	3.21	12	11	11	11	11	11
SU(0,3)	0	3.53	14	13	13	13	12	13
SU(0,2)	0	4.51	20	19	19	19	16	18
Logistic	0	4.20	19	18	18	18	15	17
Student(10)	0	4.00	16	16	15	16	13	15
T(10)	0	5.38	89	88	86	80	95	94
Laplace	0	6.00	37	35	38	36	32	36
SU(0,1)	0	36.2	52	50	52	51	44	50
SU(0,0.9)	0	82.1	60	58	60	58	53	58
Cauchy	0	∞	91	90	91	90	88	91
Student(2)	0	∞	61	60	61	60	54	59
Student(3)	0	∞	43	41	42	41	35	40
Student(4)	0	∞	33	31	32	31	26	30
Student(5)	0	9.00	27	26	25	25	21	24

4.9. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje aszimmetrikus vékony- és vastag szélű alternatívákkal szemben 0,10 szignifikanciaszint esetén, $n = 20$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(2,1)	-0.57	2.40	43	46	38	40	28	29
Beta(3,2)	-0.29	2.36	14	15	14	15	13	11
TruncN(-2,1)	-0.32	2.27	17	19	17	17	14	12
TruncN(-3,1)	-0.55	2.78	24	25	22	23	17	14
TruncN(-3,2)	-0.18	2.65	9	9	9	9	9	8
Weibull(4)	-0.09	2.75	9	9	9	9	10	8
Weibull(3.6)	0.00	2.72	8	7	9	9	9	9
Weibull(2.0)	0.63	3.25	24	25	23	24	18	29
SB(0.533,0.5)	0.65	2.13	82	84	73	74	58	81
SB(1,1)	0.73	2.91	43	45	40	43	30	48
SB(1,2)	0.28	2.77	13	12	12	12	11	15
Half N(0,1)	0.97	3.78	57	59	50	52	36	60
SU(1,1)	-5.37	93.4	79	79	79	80	70	70
SU(1,2)	-0.87	5.59	29	29	28	29	23	19
Triangle(II)	0.57	16.4	43	46	38	40	28	47
Gumbel	1.14	5.40	41	42	39	41	30	26
χ^2_1	2.83	15.0	99	99	98	98	94	99
Exp(1/2)	2.00	9.00	90	90	85	86	70	90
χ^2_4	1.41	6.00	64	65	60	62	45	68
Lognormal	6.18	113.9	96	96	94	94	86	96
Weibull(0.5)	6.62	87.7	*	*	*	*	99	*

4.10. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje szimmetrikus vékony szélű, közel normális és vastag szélű alternatívákkal szemben 0,10 szignifikanciaszint esetén, $n = 50$ mintaméret és 200 000 ismétlés mellett.

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(0.5,0.5)	0	1.50	*	*	99	99	91	*
Beta(1,1)	0	1.80	84	95	73	73	42	72
Beta(1.1,1.1)	0	1.85	76	91	66	65	36	63
Beta(1.3,1.3)	0	1.93	59	80	53	52	28	49
Beta(1.5,1.5)	0	2.00	45	68	42	41	23	38
Beta(2,2)	0	2.14	25	43	27	25	16	23
T(0.7)	0	1.92	60	82	54	53	29	50
T(1.5)	0	1.75	91	98	80	80	49	80
T(3)	0	2.06	38	64	30	29	16	28
SB(0,0.5)	0	1.63	99	*	95	95	72	94
SB(0,0.707)	0	1.87	70	88	64	63	35	60
Triangle(I)	0	2.40	9	19	11	11	9	10
TruncN(-1,1)	0	1.94	58	80	49	48	25	46
TruncN(-2,2)	0	2.36	10	20	13	12	10	11
TruncN(-3,3)	0	2.84	7	10	9	9	10	9
T(0.1)	0	3.21	13	11	12	12	13	12
SU(0,3)	0	3.53	18	13	14	15	13	15
SU(0,2)	0	4.51	31	22	25	26	20	26
Logistic	0	4.20	30	20	24	25	20	25
Student(10)	0	4.00	24	17	19	20	16	19
T(10)	0	5.38	*	*	*	99	*	*
Laplace	0	6.00	64	50	65	63	56	64
SU(0,1)	0	36.2	83	74	82	82	74	81
SU(0,0.9)	0	82.1	90	83	89	89	83	89
Cauchy	0	∞	*	*	*	*	*	*
Student(2)	0	∞	90	85	90	90	84	89
Student(3)	0	∞	72	62	68	68	59	68
Student(4)	0	∞	56	45	50	52	41	51
Student(5)	0	9.00	46	35	38	40	31	39

4.11. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje aszimmetrikus vékony- és vastag szélű alternatívákkal szemben 0,10 szignifikanciaszint esetén, $n = 50$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(2,1)	-0.57	2.40	91	96	81	84	61	77
Beta(3,2)	-0.29	2.36	32	46	31	33	21	24
TruncN(-2,1)	-0.32	2.27	47	65	39	42	25	32
TruncN(-3,1)	-0.55	2.78	60	69	44	50	31	37
TruncN(-3,2)	-0.18	2.65	11	16	12	12	11	9
Weibull(4)	-0.09	2.75	9	12	11	11	10	9
Weibull(3.6)	0.00	2.72	8	11	10	9	10	9
Weibull(2.0)	0.63	3.25	54	59	42	49	32	52
SB(0.533,0.5)	0.65	2.13	*	*	99	99	95	*
SB(1,1)	0.73	2.91	88	92	79	83	61	86
SB(1,2)	0.28	2.77	17	21	17	19	15	21
Half N(0,1)	0.97	3.78	97	98	88	91	71	94
SU(1,1)	-5.37	93.4	99	98	98	99	96	98
SU(1,2)	-0.87	5.59	53	47	46	51	39	40
Triangle(II)	0.57	16.4	91	96	81	83	60	87
Gumbel(0,1)	1.14	5.40	77	76	68	75	57	61
χ^2_1	2.83	15.0	*	*	*	*	*	*
Exp(1/2)	2.00	9.00	*	*	*	*	98	*
χ^2_4	1.41	6.00	97	98	92	95	81	96
Lognormal	6.18	113.9	*	*	*	*	*	*
Weibull(0.5)	6.62	87.7	*	*	*	*	*	*

4.12. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje szimmetrikus vékony szélű, közel normális és vastag szélű alternatívákkal szemben 0,10 szignifikanciaszint esetén, $n = 100$ mintaméret és 200 000 ismétlés mellett.

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(0.5,0.5)	0	1.50	*	*	*	*	*	*
Beta(1,1)	0	1.80	*	99	98	98	76	99
Beta(1.1,1.1)	0	1.85	99	96	95	96	67	95
Beta(1.3,1.3)	0	1.93	96	86	86	88	55	86
Beta(1.5,1.5)	0	2.00	88	70	75	78	42	74
Beta(2,2)	0	2.14	59	35	50	53	27	47
T(0.7)	0	1.92	98	88	88	90	55	88
T(1.5)	0	1.75	*	*	99	99	84	*
T(3)	0	2.06	85	61	58	61	27	60
SB(0,0.5)	0	1.63	*	*	*	*	97	*
SB(0,0.707)	0	1.87	99	94	94	95	66	94
Triangle(I)	0	2.40	19	8	16	16	10	15
TruncN(-1,1)	0	1.94	96	85	83	86	48	84
TruncN(-2,2)	0	2.36	21	9	19	19	12	17
TruncN(-3,3)	0	2.84	7	5	10	9	10	9
T(0.1)	0	3.21	14	16	12	13	12	13
SU(0,3)	0	3.53	22	26	16	18	14	17
SU(0,2)	0	4.51	45	51	33	38	27	36
Logistic	0	4.20	42	47	32	35	25	34
Student(10)	0	4.00	33	38	23	26	19	25
T(10)	0	5.38	*	*	*	*	*	*
Laplace	0	6.00	87	90	89	87	80	88
SU(0,1)	0	36.2	97	98	97	97	93	97
SU(0,0.9)	0	82.1	99	99	99	99	98	99
Cauchy	0	∞	*	*	*	*	*	*
Student(2)	0	∞	99	99	99	99	98	99
Student(3)	0	∞	92	93	89	90	81	89
Student(4)	0	∞	79	82	70	75	60	73
Student(5)	0	9.00	66	71	55	60	45	59

4.13. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek százalékban megadott empirikus ereje aszimmetrikus vékony- és vastag szélű alternatívákkal szemben 0,10 szignifikanciaszint esetén, $n = 100$ mintaméret és 200 000 ismétlés mellett (a $*$ jelöli a 100% empirikus erőt).

Alternatívák	$\sqrt{\beta_1}$	β_2	BCMR	W	ISE	BHEP	D	A^2
Beta(2,1)	-0.57	2.40	*	*	99	99	91	99
Beta(3,2)	-0.29	2.36	67	48	55	61	37	50
TruncN(-2,1)	-0.32	2.27	89	74	70	76	45	68
TruncN(-3,1)	-0.55	2.78	95	88	73	81	53	73
TruncN(-3,2)	-0.18	2.65	17	10	15	16	13	12
Weibull(4)	-0.09	2.75	10	7	13	13	12	11
Weibull(3.6)	0.00	2.72	8	5	11	11	10	10
Weibull(2.0)	0.63	3.25	87	81	67	78	53	79
SB(0.533,0.5)	0.65	2.13	*	*	*	*	*	*
SB(1,1)	0.73	2.91	*	99	98	99	89	99
SB(1,2)	0.28	2.77	29	23	25	30	21	31
Half N(0,1)	0.97	3.78	*	*	99	*	96	*
SU(1,1)	-5.37	93.4	*	*	*	*	*	*
SU(1,2)	-0.87	5.59	77	78	66	75	59	65
Triangle(II)	0.57	16.4	*	*	99	99	91	*
Gumbel(0,1)	1.14	5.40	97	96	90	95	82	90
χ^2_1	2.83	15.0	*	*	*	*	*	*
Exp(1/2)	2.00	9.00	*	*	*	*	*	*
χ^2_4	1.41	6.00	*	*	*	*	98	*
Lognormal	6.18	113.9	*	*	*	*	*	*
Weibull(0.5)	6.62	87.7	*	*	*	*	*	*

A táblázatokból látható, hogy a teszteknek ereje 0,10 szignifikanciaszint mellett nagyobb, mint 0,05 szignifikanciaszint esetében, de a viselkedésük nagyon hasonló. Ezért a tesztek összehasonlítását csak 0,05 szignifikanciaszint mellett fogjuk elvégezni. Ebből a célból a 4.14. táblázatban rendeztük a teszteket az átlagos erejük szerint az alternatívák fenti öt csoportjára. A W és W' tesztek kombinációja és a BCMR-teszt tűnik a legjobb, a D-teszt a legrosszabb teljesítményűnek. Érdekes kivétel a T(10) alternatíva, amellyel szemben viszont a D-tesztnek van a legnagyobb ereje. Az $n = 20$ esetben a W-teszt valamivel nagyobb erővel bír a szimmetrikus vékony szélű alternatívákkal szemben mint a BCMR-teszt, a BCMR-teszt pedig jobb, mint a többi. A szimmetrikus közel normális és vastag szélű alternatívákkal esetében a BCMR-teszt teljesít kicsivel jobban, mint a többi teszt. Aszimmetrikus alternatívákra a legjobb teszt a W-teszt, amihez a BCMR-teszt is nagyon közel van. Az $n = 50$ mintaméret esetén a tesztek viselkedése nagyon hasonló az n=20 esethez; valójában a vezető tesztek elsőbbsége még inkább erősödik. Érdekes kivétel az n=50 esetben, hogy a W-teszt hátraesik és a BHEP-teszt erősebbé válik a szimmetrikus közel normális és vastag szélű alternatívákkal szemben. Az $n = 100$ mintaméret mellett a BCMR-tesztnek van a legnagyobb ereje a szimmetrikus vékony szélű alternatívákkal szemben és a W'-teszt teljesít a legjobban minden szimmetrikus közel normális és vastag szélű alternatívával szemben. A többi teszt teljesítménye közel hasonló egymáshoz, kivéve a D-tesztet szimmetrikus alternatívák esetén, amely teszt kisebb erővel bír. Aszimmetrikus alternatívák ellen a legjobb teszt, $n = 100$ mintaméret mellett, a BCMR-teszt.

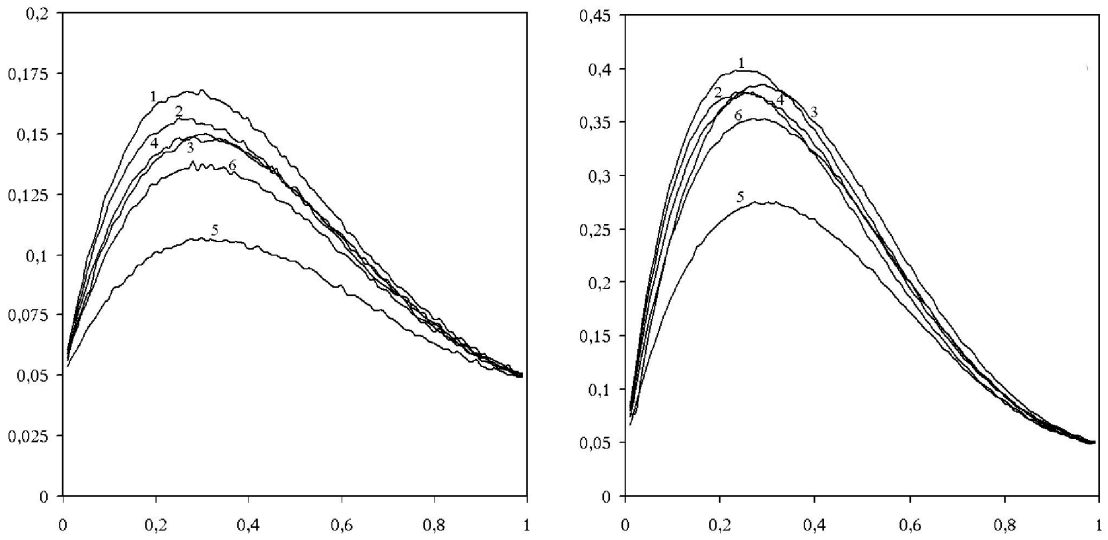
A két nem szimulált teszttel összehasonlítva a BCMR-tesztet, a következő eredményeket kaptuk. Az $n = 20$ mintaméret mellett Cabaña és Cabaña [11] megfelelő „focused" tesztjének jobb az ereje a szimmetrikus vékony szélű alternatívákkal szemben, mint a BCMR-tesztnek. Ellenben a BCMR-teszt teljesít jobban a szimmetrikus vastag szélű és aszimmetrikus alternatívákkal szemben. Az $n = 50$ mintaméret esetében a BCMR-teszt átveszi a vezetést még a szimmetrikus vékony szélű alternatívákkal szemben is. Kallenberg és Ledwina [50] megfelelő „data driven smooth" tesztjének a teljesítménye gyengébb mint a BCMR-teszté szimmetrikus vékony szélű alternatívák esetében, de épp fordított a helyzet szimmetrikus vastag szélű alternatívákra. Aszimmetrikus alternatívákkal szemben nagyon hasonló a két teszt viselkedése.

A jobb összehasonlíthatóság céljából a 4.3. ábrán felvettük a hat tesztnek a kontaminált normális alternatívákkal szembeni erejét a λ paraméter függvényében. A szignifikanciaszint 0,05; a mintaméret $n = 20$ mindkét esetben. A BCMR-teszt egyenletesen a legjobb teszt a $CN(\lambda, 4)$ alternatíva esetén, de a $CN(\lambda, 9)$ esetében az ISE-teszt legyőzi $\lambda > 0{,}3$ paraméterértékekre.

A szimulációs vizsgálat általános konklúziója hogy a BCMR-teszt általában jobban teljesít, mint más tesztek, kivéve a Wilk–Shapiro- és Shapiro–Francia-teszteket. Valamint a legtöbb esetben a W–W' kombinált teszt tulajdonságai és a BCMR kvantilis korrelációteszt tulajdonságai, amikor a pontos kritikus értékeket használuk, nagyon hasonlítanak egymáshoz. Nem meglepő módon, hiszen a két teszt aszimptotikusan ekvivalens.

4.14. táblázat. A BCMR, W, ISE, BHEP, D és A^2 tesztek sorrendje átlagos erejük alapján a különböző alternatíva csoportokra.

	1	2	3	4	5	6
n=20						
szimmetrikus vékony szélű	W	BCMR	A^2	ISE	BHEP	D
szimmetrikus közel normális	BCMR	W	ISE	BHEP	A^2	D
szimmetrikus vastag szélű	BCMR	ISE	W	A^2	BHEP	D
n=50						
szimmetrikus vékony szélű	W	BCMR	ISE	A^2	BHEP	D
szimmetrikus közel normális	BCMR	BHEP	A^2	ISE	W	D
szimmetrikus vastag szélű	BCMR	BHEP	A^2	ISE	W	D
n=100						
szimmetrikus vékony szélű	BCMR	ISE	BHEP	A^2	W'	D
szimmetrikus közel normális	W'	BCMR	BHEP	A^2	ISE	D
szimmetrikus vastag szélű	W'	BCMR	BHEP	A^2	ISE	D
n=20						
aszimmetrikus vékony szélű	W	BCMR	A^2	BHEP	ISE	D
aszimmetrikus vastag szélű	W	BCMR	BHEP	ISE	A^2	D
n=50						
aszimmetrikus vékony szélű	W	BCMR	BHEP	A^2	ISE	D
aszimmetrikus vastag szélű	W	BCMR	BHEP	A^2	ISE	D
n=100						
aszimmetrikus vékony szélű	BCMR	BHEP	W'	A^2	ISE	D
aszimmetrikus vastag szélű	BCMR	W'	BHEP	A^2	ISE	D



4.3. ábra. A BCMR, W, ISE, BHEP, D és A^2 tesztek ereje a $CN(\lambda, 4)$ alternatíva λ paraméterének függvényében (balra) és ugyanez a $CN(\lambda, 9)$ alternatívára (jobbra), jelölések: 1=BCMR-teszt; 2=W-teszt; 3=ISE-teszt; 4=BHEP-teszt; 5=D-teszt; 6=A^2-teszt

5. fejezet

Illeszkedésvizsgálat logisztikus eloszláscsaládra

5.1. Súlyozott kvantilis korreláció tesztek

Ebben a fejezetben a logisztikus eloszláscsaládhoz való illeszkedést szeretnénk tesztelni. A 4.1. fejezetben bemutattuk a del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] valamint del Barrio, Cuesta-Albertos és Matrán [33] által bevezetett kvantilis korreláció tesztet, mellyel normalitást teszteltünk. A továbbiakban ennek a tesztnek a súlyozott változatát ismertetjük, majd alkalmazzuk logisztikus eloszláscsalád esetére. A súlyfüggvény használatát a tesztstatisztikában egymástól függetlenül de Wet [28, 29] valamint Csörgő S. [19, 20] javasolta. Csörgő és Szabó [21, 22] számos eloszláscsaládra bevezette az új tesztet.

Két típusú eloszláscsalád, eltolás-skála valamint eltolás esetére vezetjük be ezeket a súlyozott teszteket. Létezik a skála eloszláscsaládra is súlyozott kvantilis korreláció teszt, de ezt mi nem használjuk a későbbiekben. Adott $G(x),\ x\in\mathbb{R}$, eloszlásfüggvényre valamint $\theta\in\mathbb{R}$ és $\sigma>0$ eltolás és skála paraméterekre legyen $G_\sigma^\theta(x)=G((x-\theta)/\sigma)$, $x\in\mathbb{R}$, továbbá tekintsük a következő eltolás-skála és eltolás családokat:

$$\mathcal{G}_{l,s}=\{G_\sigma^\theta:\theta\in\mathbb{R},\sigma>0\},\qquad \mathcal{G}_l=\{G_1^\theta:\theta\in\mathbb{R}\}.$$

Jelölje

$$Q_G(t)=G^{-1}(t)=\inf\{x\in\mathbb{R}:G(x)\geq t\},\qquad 0<t<1,$$

a G kvantilisfüggvényét. Legyen a $w\ :\ (0,1)\to[0,\infty)$ súlyfüggvény olyan, amely a $\int_0^1 w(t)\,\mathrm{d}t=1$ feltételt kielégíti, és definiáljuk az r-edik súlyozott momentumot:

$$\mu_r(G,w):=\int_0^1 (Q_G(t))^r\,w(t)\,\mathrm{d}t=\int_{-\infty}^{\infty}x^r w\left(G(x)\right)\mathrm{d}G(x).$$

A továbbiakban feltesszük, hogy $\mu_1(G,w)$ és $\mu_2(G,w)$ véges, és definiáljuk a súlyozott szórásnégyzetet is:

$$\nu(G,w):=\mu_2(G,w)-\mu_1^2(G,w)\geq 0\ .$$

Két eloszlásfüggvény, F és G, súlyozott L^2-Wasserstein-távolságát definiáljuk a

$$\mathcal{W}_w(F,G):=\left[\int_0^1\left(Q_F(t)-Q_G(t)\right)^2 w(t)\,\mathrm{d}t\right]^{\frac{1}{2}}$$

mennyiséggel. Továbbá jelölje

$$\mathcal{W}_w(F,\mathcal{G}_l) := \inf\{\mathcal{W}_w(F,G) : G \in \mathcal{G}_l\} \quad \text{és} \quad \mathcal{W}_w(F,\mathcal{G}_{l,s}) := \inf\{\mathcal{W}_w(F,G) : G \in \mathcal{G}_{l,s}\}$$

az F eloszlásnak a $\mathcal{G}_l$ illetve $\mathcal{G}_{l,s}$ családtól vett a súlyozott L^2-Wasserstein-távolságát. Csörgő S. [20] megmutatta, hogy

$$\begin{aligned}\mathcal{W}_w^2(F,\mathcal{G}_l) &= \int_0^1 (Q_F(t)-Q_G(t))^2\, w(t)\mathrm{d}t - \left[\int_0^1 (Q_F(t)-Q_G(t))\, w(t)\mathrm{d}t\right]^2 \\ &= \nu(F,w)+\nu(G,w)-2\int_0^1 Q_F(t)Q_G(t)w(t)\mathrm{d}t+2\mu_1(F,w)\mu_1(G,w),\end{aligned}$$

illetve

$$\frac{\mathcal{W}_w^2(F,\mathcal{G}_{l,s})}{\nu(F,w)} = 1 - \frac{\left[\int_0^1 Q_F(t)Q_G(t)w(t)\mathrm{d}t-\mu_1(F,w)\mu_1(G,w)\right]^2}{\nu(F,w)\nu(G,w)}.$$

Tekintsünk egy $X_1,\ldots,X_n$ véletlen mintát egy ismeretlen F eloszlásfüggvénnyel, és legyen G egy rögzített eloszlásfüggvény. Szeretnénk tesztelni a $\mathcal{H}_0 : F \in \mathcal{G}_{l,s}$ nullhipotézist. Ebből a célból definiálni fogjuk a $\mathcal{W}_w^2(F,\mathcal{G}_{l,s})/\nu(F,w)$ hányadosnak az empirikus változatát a következő módon:

$$\begin{aligned}V_n &:= 1 - \frac{\left[\int_0^1 Q_n(t)Q_G(t)w(t)\mathrm{d}t-\mu_1(G,w)\int_0^1 Q_n(t)w(t)\mathrm{d}t\right]^2}{\nu(G,w)\left[\int_0^1 Q_n^2(t)w(t)\mathrm{d}t-\left(\int_0^1 Q_n(t)w(t)\mathrm{d}t\right)^2\right]} \\ &= 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\left\{\int_{\frac{k-1}{n}}^{\frac{k}{n}} Q_G(t)w(t)\mathrm{d}t-\mu_1(G,w)\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right\}\right]^2}{\nu(G,w)\left[\sum_{k=1}^n X_{k,n}^2\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t-\left(\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right)^2\right]}. \qquad (5.1)\end{aligned}$$

Hasonló módon a $\mathcal{H}_0 : F \in \mathcal{G}_l$ nullhipotézis tesztelésére a $\mathcal{W}_w^2(F,\mathcal{G}_l)$ empirikus változatát definiáljuk:

$$\begin{aligned}W_n &:= \int_0^1 \{Q_n(t)-Q_G(t)\}^2\, w(t)\mathrm{d}t - \left[\int_0^1 \{Q_n(t)-Q_G(t)\}\, w(t)\mathrm{d}t\right]^2 \\ &= \nu(G,w)+\sum_{k=1}^n X_{k,n}^2\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t-\left[\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right]^2 \\ &\quad -2\sum_{k=1}^n X_{k,n}\left\{\int_{\frac{k-1}{n}}^{\frac{k}{n}} Q_G(t)w(t)\mathrm{d}t-\mu_1(G,w)\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right\}.\end{aligned}$$

Jegyezzük meg, hogy a V_n eltolás- és skálamentes, a W_n pedig eltolásmentes. A G eloszlásfüggvény segítségével legyen

$$-\infty \le a_G := \sup\{x\in\mathbb{R} : G(x)=0\} \le \inf\{x\in\mathbb{R} : G(x)=1\} =: b_G \le \infty,$$

vagyis $a_G = \inf(\mathrm{supp}(G))$, $b_G = \sup(\mathrm{supp}(G))$, ahol $\mathrm{supp}(G)$ a G tartója, azaz az a legszűkebb $\mathrm{supp}(G)\subset\mathbb{R}$ halmaz, melynek mértéke G szerint 1. Legyen $Y_1,\ldots,Y_n$ a G eloszlásfüggvényből származó minta, és jelölje $Y_{1,n}\le\ldots\le Y_{n,n}$ a kapcsolatos rendezett mintát. Csörgőtől [20] származik a következő eredmény a V_n és W_n statisztikák aszimptotikus viselkedéséről. A [20] 2. és 3. Tételének azt a részét idézzük, amelyet használni fogunk.

5.1. Tétel (Csörgő [20]). *Legyen w egy nemnegatív, a $(0,1)$ intervallumon integrálható függvény, amelyre $\int_0^1 w(t)\,\mathrm{d}t=1$. Tegyük fel, hogy G olyan eloszlásfüggvény, amelynek van véges súlyozott második momentuma, és kétszer folytonosan differenciálható az (a_G,b_G) nyitott intervallumon, továbbá $g(x)=G'(x)>0$ minden $x\in(a_G,b_G)$ esetén, legyen továbbá B a Brown-híd. Ha a*

$$\sup_{0<t<1}\frac{t(1-t)|g'(Q_G(t))|}{g^2(Q_G(t))}<\infty,\qquad \int_0^1\frac{t(1-t)}{g^2(Q_G(t))}w(t)\mathrm{d}t<\infty, \tag{5.2}$$

és az

$$n\int_0^{\frac{1}{n+1}}\left[Y_{1,n}-Q_G(t)\right]^2 w(t)\mathrm{d}t\xrightarrow{\mathbf{P}}0,\qquad n\int_{\frac{n}{n+1}}^1\left[Y_{n,n}-Q_G(t)\right]^2 w(t)\mathrm{d}t\xrightarrow{\mathbf{P}}0, \tag{5.3}$$

feltételek teljesülnek, akkor a következő állítások érvényesek:

(i) Ha F a G által generált $\mathcal{G}_l$ eltoláscsaládhoz tartozik, akkor

$$nW_n\xrightarrow{\mathcal{D}}W_g:=\int_0^1\frac{B^2(t)}{g^2(Q_G(t))}w(t)\mathrm{d}t-\left[\int_0^1\frac{B(t)}{g(Q_G(t))}w(t)\mathrm{d}t\right]^2.$$

(ii) Ha F a G által generált $\mathcal{G}_{l,s}$ eltolás-skála családhoz tartozik, akkor

$$\begin{aligned}nV_n\xrightarrow{\mathcal{D}}V_g:=&\frac{1}{\nu(G,w)}\left\{\int_0^1\frac{B^2(t)}{g^2(Q_G(t))}w(t)\mathrm{d}t-\left[\int_0^1\frac{B(t)}{g(Q_G(t))}w(t)\mathrm{d}t\right]^2\right\}\\&-\left[\frac{1}{\nu(G,w)}\int_0^1\frac{B(t)Q_G(t)}{g(Q_G(t))}w(t)\mathrm{d}t-\frac{\mu_1(G,w)}{\nu(G,w)}\int_0^1\frac{B(t)}{g(Q_G(t))}w(t)\mathrm{d}t\right]^2.\end{aligned} \tag{5.4}$$

A következőkben ennek a tételnek a segítségével fogjuk a logisztikus eloszláscsaládhoz tartozó kvantilis korreláció teszt aszimptotikus viselkedését bizonyítani.

5.2. Elméleti eredmények

5.2.1. Súlyozott kvantilis korreláció tesztek logisztikus eloszláscsaládok esetén

A logisztikus eloszlást a

$$G(x)=\frac{1}{1+e^{-x}},\qquad x\in\mathbb{R}, \tag{5.5}$$

eloszlásfüggvénnyel definiáljuk. A G logisztikus növekedési görbét a 19. század közepén populációdinamikai munkájában Verhulst [73] vezette be. A logisztikus eloszlás első tisztán statisztikai értelmezése Gumbel [46] nevéhez fűződik, aki 1944-ben megmutatta, hogy szimmetrikus folytonos eloszlásból származó $X_{1,n}\le X_{2,n}\le\ldots\le X_{n,n}$ rendezett minta esetén az $X_{1,n}+X_{n,n}$ „mid-range” aszimptotikus eloszlása logisztikus. Balakrishnan [5] egy könyvet szentelt a logisztikus eloszlásnak, amely könyv tartalmaz a logisztikus eloszlásra

vonatkozó teszteket is a 13. fejezetben. Ezek között a szokásos technikák is megtalálhatók: χ^2-teszt, EDF-tesztek valamint a regresszió- és korrelációtesztek. További teszteket javasolt Aguirre és Nikulin [2] és Meintanis [58] a logisztikus eloszláshoz való illeszkedés vizsgálatára. Mi a súlyozott kvantilis korreláció tesztet szeretnénk bevezetni logisztikus eloszláscsaládhoz való illeszkedés vizsgálatára.

A logisztikus eloszlás sűrűségfüggvénye és kvantilisfüggvénye

$$g(x) = \frac{e^{-x}}{(1+e^{-x})^2}, \quad x \in \mathbb{R}, \qquad \text{és} \qquad Q_G(t) = \ln \frac{t}{1-t}, \quad 0 < t < 1. \tag{5.6}$$

A $\mathcal{G}_{l,s}$ jelölje a logisztikus eltolás-skála családot, és a $\mathcal{G}_l$ jelölje a logisztikus eltoláscsaládot az előző fejezetbeli definíciókkal. De Wet [29] eltoláscsaládok esetében javasolt egy $w(t) = = L_1'(Q_G(t))/I_1$ alakú általános súlyfüggvényt, ahol

$$L_1(x) := \frac{-g'(x)}{g(x)}, \ x \in \mathbb{R}, \qquad \text{és} \qquad I_1 := \int_{\mathbb{R}} L_1'(x) g(x) \, \mathrm{d}x.$$

A logisztikus esetben azt kapjuk, hogy

$$L_1(x) = -\frac{\big(e^{-x}(-1)(1+e^{-x})^2 - e^{-x}2(1+e^{-x})e^{-x}(-1)\big)/(1+e^{-x})^4}{e^{-x}/(1+e^{-x})^2} = \frac{1-e^{-x}}{1+e^{-x}},$$

a deriváltja pedig $L_1'(x) = 2g(x)$. Ekkor egy parciális integrálás után

$$\begin{aligned} I_1 &= \int_{\mathbb{R}} 2\frac{e^{-x}}{(1+e^{-x})^2}\frac{e^{-x}}{(1+e^{-x})^2}\,\mathrm{d}x = \lim_{y\to\infty}\frac{2}{3}\int_{-y}^{y} e^{-x}(-3)(1+e^{-x})^{-4}e^{-x}(-1)\,\mathrm{d}x \\ &= \lim_{y\to\infty}\left[\frac{2}{3}e^{-x}(1+e^{-x})^{-3}\right]_{-y}^{y} + \lim_{y\to\infty}\frac{1}{3}\int_{-y}^{y}(-2)(1+e^{-x})^{-3}e^{-x}(-1)\,\mathrm{d}x \\ &= \lim_{y\to\infty}\left[\frac{1}{3}(1+e^{-x})^{-2}\right]_{-y}^{y} = \frac{1}{3}, \end{aligned}$$

ami a

$$w(t) = 6\frac{(1-t)/t}{(1+(1-t)/t)^2} = 6t(1-t), \qquad 0 < t < 1, \tag{5.7}$$

súlyfüggvényt eredményezi.

Megjegyezzük, hogy de Wet különböző súlyfüggvényeket javasolt eltolás-, illetve skálacsaládok esetén. Motivációja az volt, hogy a tesztstatisztika határeloszlásának soros előállításában „szabadságifok vesztést" idézzen elő, amit az eloszláscsalád paraméterének Cramér-Rao értelemben aszimptotikusan hatékony becslésével ért el. Mi most az általa javasolt eltoláscsaládhoz gyártott (5.7) súlyfüggvényt szeretnénk az eltolás-skála család esetében is használni.

Ahhoz, hogy a tesztstatisztikákat bevezethessük, először meghatározuk a súlyozott első és második momentumot. Az első momentum értékének meghatározásához vegyük észre, hogy a $t \mapsto 1-t$ helyettesítés alkalmazásával az

$$\int_0^1 \ln(t) 6t(1-t)\,\mathrm{d}t = \int_0^1 \ln(1-t) 6t(1-t)\,\mathrm{d}t$$

egyenlőség teljesül. Továbbá egy parciális integrálás és a L'Hospital-szabály segítségével kapjuk, hogy

$$\int_0^1 |\ln(t)t(1-t)|\,\mathrm{d}t = \int_0^1 (-1)\ln(t)t(1-t)\,\mathrm{d}t \le (-1)\int_0^1 \ln(t)t\,\mathrm{d}t$$
$$= \lim_{\varepsilon\to 0}\left[(-1)\ln t\frac{t^2}{2}\right]_\varepsilon^1 + \int_0^1 \frac{1}{t}\frac{t^2}{2}\,\mathrm{d}t = \frac{1}{4},$$

így az első momentum értéke az (5.5) és (5.7) formulák behelyettesítésével, a logaritmus tulajdonságainak segítségével a következőképpen adódik:

$$\mu_1(G,w) = \int_0^1 \ln\left(\frac{t}{1-t}\right)6t(1-t)\,\mathrm{d}t = \int_0^1 \ln(t)6t(1-t)\,\mathrm{d}t - \int_0^1 \ln(1-t)\,6t(1-t)\,\mathrm{d}t = 0.$$

A második momentumot először két parciális integrálás segítségével alakítjuk át. Ehhez vegyük észre, hogy a következő összefüggések érvényesek. Az első parciális integráláshoz

$$(3t^2-2t^3)' = 6t(1-t) \qquad \text{és} \qquad \left(\left(\ln\left(\frac{t}{1-t}\right)\right)^2\right)' = 2\ln\left(\frac{t}{1-t}\right)\frac{1}{t(1-t)},$$

továbbá a második parciális integráláshoz

$$(t(t-1)-\ln(1-t))' = \frac{3t-2t^2}{1-t} \qquad \text{és} \qquad \left(\ln\left(\frac{t}{1-t}\right)\right)' = \frac{1}{t(1-t)}.$$

Ezután szükségünk lesz egy $t\mapsto 1-t$ helyettesítésre, és az

$$\frac{\ln t}{t(t-1)} = \frac{\ln t}{t-1} - \frac{\ln t}{t}$$

felbontásra. Végül a második momentum egy fontos részét az

$$\int_0^1 \frac{\ln t}{t-1}\,\mathrm{d}t = \frac{\pi^2}{6}$$

azonosság adja, ami megtalálható Abramowitz és Stegun [1] 4.1.55 alatt, és szükség lesz még a $((\ln t)^2)' = 2\ln t/t$ deriváltra. A parciálás integrálás során kapott kifejezések határértéke nullává válik, mivel $\lim_{\varepsilon\to 0}\varepsilon\ln\varepsilon = 0$ és $\lim_{\varepsilon\to 0}\ln(1-\varepsilon)\ln\varepsilon = 0$. Ahhoz, hogy ez látható legyen, először alakítsuk át ezeket a tagokat a logaritmus tulajdonságait használva, majd helyettesítsünk be és vonjunk össze, ekkor

$$\lim_{\varepsilon\to 0}\left[\left(\ln\left(\frac{t}{1-t}\right)\right)^2(3t^2-2t^3) - 2\ln\left(\frac{t}{1-t}\right)\big(t(t-1)-\ln(1-t)\big) - (\ln t)^2\right]_\varepsilon^{1-\varepsilon}$$
$$= \lim_{\varepsilon\to 0}\Big[\big((\ln t)^2 - 2\ln t\ln(1-t) + (\ln(1-t))^2\big)(3t^2-2t^3)$$
$$-2\big(\ln t-\ln(1-t)\big)\big(t(t-1)-\ln(1-t)\big) - (\ln t)^2\Big]_\varepsilon^{1-\varepsilon}$$
$$= \lim_{\varepsilon\to 0}\Big[(\ln t)^2(3t^2-2t^3-1) - 2\ln t\ln(1-t)(3t^2-2t^3-1) + (\ln(1-t))^2(3t^2-2t^3-2)$$
$$-2(\ln t)t(1-t) + 2\ln(1-t)t(1-t)\Big]_\varepsilon^{1-\varepsilon}$$

$$
\begin{aligned}
&= \lim_{\varepsilon\to 0}\Big\{ (\ln(1-\varepsilon))^2\left(3(1-\varepsilon)^2-2(1-\varepsilon)^3-1\right)-2\ln(1-\varepsilon)\ln\varepsilon\big(3(1-\varepsilon)^2-2(1-\varepsilon)^3-1\big)\\
&\quad +(\ln\varepsilon)^2\left(3(1-\varepsilon)^2-2(1-\varepsilon)^3-2\right)-2\ln(1-\varepsilon)\ (1-\varepsilon)\varepsilon+2\ln\varepsilon(1-\varepsilon)\varepsilon\\
&\quad -(\ln\varepsilon)^2\left(3\varepsilon^2-2\varepsilon^3-1\right)+2\ln\varepsilon\ln(1-\varepsilon)(3\varepsilon^2-2\varepsilon^3-1)-(\ln(1-\varepsilon))^2\left(3\varepsilon^2-2\varepsilon^3-2\right)\\
&\quad +2(\ln\varepsilon)\varepsilon(1-\varepsilon)-2\ln(1-\varepsilon)\varepsilon(1-\varepsilon)\Big\}\\
&= \lim_{\varepsilon\to 0}\Big\{ (\ln(1-\varepsilon))^2\left(-3\varepsilon^2+2\varepsilon^3\right)-2\ln(1-\varepsilon)\ln\varepsilon(-3\varepsilon^2+2\varepsilon^3)+(\ln\varepsilon)^2\left(-3\varepsilon^2+2\varepsilon^3-1\right)\\
&\quad -(\ln\varepsilon)^2\left(3\varepsilon^2-2\varepsilon^3-1\right)+2\ln\varepsilon\ln(1-\varepsilon)(3\varepsilon^2-2\varepsilon^3-1)-(\ln(1-\varepsilon))^2\left(3\varepsilon^2-2\varepsilon^3-2\right)\\
&\quad -4\ln(1-\varepsilon)\ (1-\varepsilon)\varepsilon+4\ln\varepsilon(1-\varepsilon)\varepsilon\Big\}\\
&= \lim_{\varepsilon\to 0}\Big\{ (\ln(1-\varepsilon))^2\left(-6\varepsilon^2+4\varepsilon^3-2\right)-2\ln(1-\varepsilon)\ln\varepsilon(-6\varepsilon^2+4\varepsilon^3+1)+(\ln\varepsilon)^2\left(-6\varepsilon^2+4\varepsilon^3\right)\\
&\quad -4\ln(1-\varepsilon)\ (1-\varepsilon)\varepsilon+4\ln\varepsilon(1-\varepsilon)\varepsilon\Big\}=0.
\end{aligned}
$$

Ekkor

$$
\begin{aligned}
\mu_2(G,w) &= \int_0^1\left(\ln\left(\frac{t}{1-t}\right)\right)^2 6t(1-t)\,\mathrm{d}t = \lim_{\varepsilon\to 0}\int_\varepsilon^{1-\varepsilon}\left(\ln\left(\frac{t}{1-t}\right)\right)^2 6t(1-t)\,\mathrm{d}t\\
&= \lim_{\varepsilon\to 0}\left\{\left[\left(\ln\left(\frac{t}{1-t}\right)\right)^2(3t^2-2t^3)\right]_\varepsilon^{1-\varepsilon}-\int_\varepsilon^{1-\varepsilon}2\ln\left(\frac{t}{1-t}\right)\frac{1}{t(1-t)}(3t^2-2t^3)\,\mathrm{d}t\right\}\\
&= \lim_{\varepsilon\to 0}\left\{\left[\left(\ln\left(\frac{t}{1-t}\right)\right)^2(3t^2-2t^3)-2\ln\left(\frac{t}{1-t}\right)\big(t(t-1)-\ln(1-t)\big)\right]_\varepsilon^{1-\varepsilon}\right.\\
&\qquad \left.+2\int_\varepsilon^{1-\varepsilon}\frac{1}{t(1-t)}\big(t(t-1)-\ln(1-t)\big)\,\mathrm{d}t\right\}\\
&= \lim_{\varepsilon\to 0}\left\{\left[\left(\ln\left(\frac{t}{1-t}\right)\right)^2(3t^2-2t^3)-2\ln\left(\frac{t}{1-t}\right)\big(t(t-1)-\ln(1-t)\big)\right]_\varepsilon^{1-\varepsilon}\right.\\
&\qquad \left.-2+2\int_\varepsilon^{1-\varepsilon}\left(\frac{\ln t}{t-1}-\frac{\ln t}{t}\right)\,\mathrm{d}t\right\}\\
&= \lim_{\varepsilon\to 0}\left\{\left[\left(\ln\left(\frac{t}{1-t}\right)\right)^2(3t^2-2t^3)-2\ln\left(\frac{t}{1-t}\right)\big(t(t-1)-\ln(1-t)\big)-(\ln t)^2\right]_\varepsilon^{1-\varepsilon}\right\}\\
&\qquad -2+\frac{\pi^2}{3}=\frac{\pi^2}{3}-2.
\end{aligned}
$$

Ebből következik, hogy $\nu(G,w)=\pi^2/3-2$.

Az (5.1) eltolás-skála mentes V_n tesztstatisztika logisztikus eltolás-skála családra

$$V_n = 1 - \frac{\left[\sum_{k=1}^{n} a_{k,n} X_{k,n}\right]^2}{\left(\frac{\pi^2}{3} - 2\right)\left[\sum_{k=1}^{n} b_{k,n} X_{k,n}^2 - \left(\sum_{k=1}^{n} b_{k,n} X_{k,n}\right)^2\right]},$$

ahol az együtthatók az alábbi módon kaphatók meg:

$$\begin{aligned}
a_{k,n} &:= \int_{\frac{k-1}{n}}^{\frac{k}{n}} 6t(1-t)\ln\left(\frac{t}{1-t}\right)\mathrm{d}t \\
&= \left[(3t^2-2t^3)\ln\left(\frac{t}{1-t}\right)\right]_{\frac{k-1}{n}}^{\frac{k}{n}} - \int_{\frac{k-1}{n}}^{\frac{k}{n}} (3t^2-2t^3)\frac{1}{t(1-t)}\,\mathrm{d}t \\
&= \left[(3t^2-2t^3)\ln\left(\frac{t}{1-t}\right) - \ln(1-t) - t^2 + t\right]_{\frac{k-1}{n}}^{\frac{k}{n}} \\
&= \frac{k^2(3n-2k)}{n^3}\ln\frac{k}{n-k} - \frac{(k-1)^2(3n-2k+2)}{n^3}\ln\frac{k-1}{n-k+1} \\
&\quad + \ln\frac{n-k}{n-k+1} + \frac{1-2k}{n^2} + \frac{1}{n}, \\
b_{k,n} &:= \int_{\frac{k-1}{n}}^{\frac{k}{n}} 6t(1-t)\mathrm{d}t = \frac{3(2k-1)}{n^2} + \frac{2(-3k^2+3k-1)}{n^3}.
\end{aligned}$$

5.2. Megjegyzés (de Wet [29]). *Megjegyezzük, hogy az eltolásmentes tesztstatisztika logisztikus eltoláscsalád esetében*

$$W_n = \left(\frac{\pi^2}{3} - 2\right) + \sum_{k=1}^{n} b_{k,n} X_{k,n}^2 - \left[\sum_{k=1}^{n} b_{k,n} X_{k,n}\right]^2 - 2\sum_{k=1}^{n} a_{k,n} X_{k,n},$$

ahol $a_{k,n}$ és $b_{k,n}$ a fent definiált együtthatók. Ekkor de Wet a következőt állítja: Ha $F \in \mathcal{G}_l$, akkor

$$nW_n \xrightarrow{\mathcal{D}} W := \int_0^1 \frac{6B^2(t)}{t(1-t)}\,\mathrm{d}t - \left[\int_0^1 6B(t)\,\mathrm{d}t\right]^2 \tag{5.8}$$

Csörgő [20] aszimptotikus eredményének a következményeként kapjuk a V_n tesztstatisztika határeloszlását. Ez az új eredmény:

5.3. Tétel. *Ha a minta F eloszlásfüggvénye a $\mathcal{G}_{l,s}$ logisztikus eltolás-skála családhoz tartozik, akkor*

$$\begin{aligned}
nV_n \xrightarrow{\mathcal{D}} V := &\frac{1}{\pi^2/3-2}\left\{\int_0^1 \frac{6B^2(t)}{t(1-t)}\,\mathrm{d}t - \left[\int_0^1 6B(t)\,\mathrm{d}t\right]^2\right\} \\
&- \left[\frac{1}{\pi^2/3-2}\int_0^1 6B(t)\ln\left(\frac{t}{1-t}\right)\,\mathrm{d}t\right]^2,
\end{aligned} \tag{5.9}$$

ahol határérték 1 valószínűséggel létezik.

Vegyük észre, hogy a $\{\}$ zárójelben pontosan az (5.8) formulában definiált W változó jelenik meg. Ahhoz, hogy az 5.3. Tétel bizonyításával folytathassuk, szükségünk van a következő lemmára.

5.4. Lemma. *Tetszőleges $k \geq 0$ esetén*

$$n \int_0^{\frac{1}{n+1}} \left(\ln \left(n \frac{t}{1-t} \right) \right)^k t(1-t) \mathrm{d}t = (-1)^k \frac{k!}{2^{k+1}} \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^2} \right),$$

és

$$n \int_{\frac{n}{n+1}}^1 \left(\ln \left(n \frac{1-t}{t} \right) \right)^k t(1-t) \mathrm{d}t = (-1)^k \frac{k!}{2^{k+1}} \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^2} \right).$$

Bizonyítás. Először az $x = nt/(1-t)$ helyettesítést alkalmazzuk, amely leképezés szigorúan növekvő módon képezi bele az $(0,1/(n+1))$ intervallumot a $(0,1)$ intervallumba, ezzel azt kapjuk, hogy

$$n \int_0^{\frac{1}{n+1}} \left(\ln \left(n \frac{t}{1-t} \right) \right)^k t(1-t)\, \mathrm{d}t = \frac{1}{n} \int_0^1 (\ln x)^k \, \frac{x}{(1+\frac{x}{n})^4} \, \mathrm{d}x.$$

Használva, hogy $0 < x < 1$, a következő becslést kapjuk

$$\begin{aligned} \left| \frac{1-(1+\frac{x}{n})^4}{(1+\frac{x}{n})^4} \right| &= \frac{\left|1-\left(1+\frac{x}{n}\right)\right| \left|1+\left(1+\frac{x}{n}\right)\right| \left|1+(1+\frac{x}{n})^2\right|}{\left|1+\frac{x}{n}\right|^4} \\ &\leq \left|-\frac{x}{n}\right| \left|2+\frac{x}{n}\right| \left|2+2\frac{x}{n}+\frac{x^2}{n^2}\right| \leq \frac{1}{n} \cdot 3 \cdot 5 = \frac{15}{n}, \end{aligned}$$

amely becsléssel

$$\begin{aligned} \frac{1}{n} \int_0^1 (\ln x)^k \, \frac{x}{(1+\frac{x}{n})^4} \, \mathrm{d}x &= \frac{1}{n} \int_0^1 x \, (\ln x)^k \, \mathrm{d}x + \frac{1}{n} \int_0^1 x \, (\ln x)^k \, \frac{1-(1+\frac{x}{n})^4}{(1+\frac{x}{n})^4} \, \mathrm{d}x \\ &= \frac{1}{n} \int_0^1 x \, (\ln x)^k \, \mathrm{d}x \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right). \end{aligned}$$

Parciálisan integrálva az

$$\int_0^1 x \, (\ln x)^k \, \mathrm{d}x = \lim_{\varepsilon \to 0} \left[\frac{x^2}{2} \, (\ln x)^k \right]_\varepsilon^1 - \int_0^1 \frac{x^2}{2} k \, (\ln x)^{k-1} \, \frac{1}{x} \, \mathrm{d}x = -\frac{k}{2} \int_0^1 x \, (\ln x)^{k-1} \, \mathrm{d}x$$

rekurzió érvényes, ami azt jelenti, hogy

$$\int_0^1 x \, (\ln x)^k \, \mathrm{d}x = (-1)^k \frac{k!}{2^k} \int_0^1 x \, \mathrm{d}x = (-1)^k \frac{k!}{2^{k+1}}.$$

Ezzel bebizonyítottuk az állítás első egyenlőségét. A második egyenlőség a $t \mapsto 1-t$ helyettesítéssel jön az első egyenlőségből. □

Az 5.3. Tétel bizonyítása. A konvergencia eredmény bizonyításához az (5.2) – (5.3) feltételeket kell ellenőriznünk. Az (5.6) formulák alkalmazásával azonnal kapjuk, hogy

$$\sup_{0<t<1} \frac{t(1-t)|g'(Q_G(t))|}{g^2(Q_G(t))} = \sup_{0<t<1} \frac{t(1-t)|t(1-t)(1-2t)|}{t^2(1-t)^2} \leq 1,$$

és

$$\int_0^1 \frac{t(1-t)}{g^2(Q_G(t))} w(t)\,\mathrm{d}t = \int_0^1 \frac{t(1-t)}{t^2(1-t)^2} 6t(1-t)\,\mathrm{d}t = 6,$$

tehát az (5.2) feltétel teljesül. Már csak az (5.3) feltétel ellenőrzése van hátra. Tetszőleges F eloszlásfüggvényű $X_1, \ldots, X_n$ minta és minden $x \in \mathbb{R}$ esetén

$$P(\min\{X_1, \ldots, X_n\} > x) = P(\cap_{i=1}^n \{X_i > x\}) = (P(X_i > x))^n = (1-F(x))^n.$$

és

$$P(\max\{X_1, \ldots, X_n\} \leq x) = P(\cap_{i=1}^n \{X_i \leq x\}) = (P(X_i \leq x))^n = (F(x))^n.$$

Ekkor az

$$A_n := Y_{1,n} + \ln n \qquad \text{és} \qquad B_n := Y_{n,n} - \ln n \qquad n = 1,2,\ldots$$

véletlen változók sorozatára

$$\begin{aligned} P(A_n \leq x) =& P(Y_{1,n} + \ln n \leq x) = P(\min\{Y_1, \ldots, Y_n\} \leq x - \ln n) = 1 - \left(1 - \frac{1}{1+e^{-x+\ln n}}\right)^n \\ =& 1 - \left(\frac{e^{-x}n}{1+e^{-x}n}\right)^n = 1 - \frac{1}{\left(\frac{e^x}{n}+1\right)^n} \to 1 - e^{-e^x}, \end{aligned}$$

és

$$\begin{aligned} P(B_n \leq x) =& P(Y_{n,n} - \ln n \leq x) = P(\max\{Y_1, \ldots, Y_n\} \leq x + \ln n) = \left(\frac{1}{1+e^{-x-\ln n}}\right)^n \\ =& \frac{1}{\left(1+\frac{e^{-x}}{n}\right)^n} \to e^{-e^{-x}}. \end{aligned}$$

Ennélfogva $(A_n)_{n=1,2\ldots}$ és $(B_n)_{n=1,2\ldots}$ sorozatok sztochasztikusan korlátosak, amiből következik az 5.4. Lemma miatt

$$\begin{aligned} & n \int_0^{\frac{1}{n+1}} \left[Y_{1,n} - \ln\left(\frac{t}{1-t}\right)\right]^2 6t(1-t)\mathrm{d}t \\ &= n \int_0^{\frac{1}{n+1}} \left[Y_{1,n} + \ln n - \left(\ln n + \ln\frac{t}{1-t}\right)\right]^2 6t(1-t)\mathrm{d}t \\ &= 6A_n^2 n \int_0^{\frac{1}{n+1}} t(1-t)dt - 12A_n n \int_0^{\frac{1}{n+1}} \ln\left(n\frac{t}{1-t}\right) t(1-t)\mathrm{d}t \\ &+ 6n \int_0^{\frac{1}{n+1}} \left(\ln\left(n\frac{t}{1-t}\right)\right)^2 t(1-t)\mathrm{d}t \\ &= A_n^2\left(\frac{3}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)\right) + A_n\left(\frac{3}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)\right) + \frac{3}{2n} + \mathcal{O}\left(\frac{1}{n^2}\right) \xrightarrow{\mathbf{P}} 0, \end{aligned}$$

hasonlóan

$$
\begin{aligned}
& n\int_{\frac{n}{n+1}}^{1}\left[Y_{n,n}-\ln\left(\frac{t}{1-t}\right)\right]^2 6t(1-t)\mathrm{d}t \\
& = n\int_{\frac{n}{n+1}}^{1}\left[Y_{n,n}-\ln n+\left(\ln n-\ln\frac{t}{1-t}\right)\right]^2 6t(1-t)\mathrm{d}t \\
& = B_n^2\left(\frac{3}{n}+\mathcal{O}\left(\frac{1}{n^2}\right)\right)-B_n\left(\frac{3}{n}+\mathcal{O}\left(\frac{1}{n^2}\right)\right)+\frac{3}{2n}+\mathcal{O}\left(\frac{1}{n^2}\right)\xrightarrow{\mathbf{P}}0.
\end{aligned}
$$

Ezzel kész a bizonyítás, mivel a feltételeket ellenőriztük. □

5.2.2. A határeloszlás végtelen soros alakja

A 4. fejezetben bemutattuk del Barrio, Cuesta-Albertos és Matrán [33] BCMR normalitástesztjét, és a 4.1. Tételben megadtuk a tesztstatisztika határeloszlását egyrészt egy Brown-híd funkcionáljaként, másrészt független χ_1^2 eloszlású változók végtelen lineáris kombinációjaként. Del Barrio, Cuesta-Albertos és Matrán [33] a végtelen soros alakot főkomponens analízis segítségével határozták meg, mely módszer ebben az esetben a sztochasztikus folyamatok Karhunen–Loève-sorfejtésére épül. Erről a sorfejtésről általánosan, részletesen Shorack és Wellner [66] 5. fejezetében olvashatunk. Hasonló módon de Wet [29] megmutatta, hogy az (5.8) formulában definiált W változóra

$$
W\overset{\mathcal{D}}{=}\sum_{k=2}^{\infty}\frac{6}{k(k+1)}Z_k^2, \tag{5.10}
$$

ahol $(Z_k)_{k=1}^{\infty}$ független, standard normális eloszlású véletlen változók végtelen sorozata, és a sor 1 valószínűséggel konvergál. A továbbiakban szeretnénk meghatározni az (5.9) formulában definiált V változó végtelen soros alakját, és bebizonyítjuk a következő tételt:

5.5. Tétel. *A V határeloszlás felírható*

$$
V\ \overset{\mathcal{D}}{=}\ \frac{1}{\pi^2/3-2}\sum_{k=2}^{\infty}\frac{6}{k(k+1)}Z_k^2\ -\left[\frac{1}{\pi^2/3-2}\sum_{l=1}^{\infty}\frac{3\sqrt{4l+1}}{l(l+1)(2l-1)(2l+1)}Z_{2l}\right]^2 \tag{5.11}
$$

alakban, ahol $(Z_m)_{m=1}^{\infty}$ független, standard normális eloszlású véletlen változók végtelen sorozata, és a sor 1 valószínűséggel konvergál.

Vegyük észre, hogy az (5.11) formulában megjelenik a de Wet által megadott sorfejtés, ami nem meglepő annak tükrében, hogy a V változó (5.9) formulában szereplő definíciójában a W változó teljesen megjelenik. Az (5.9) formulában, a V változó definíciójában az első tag sorfejtése a (5.10) formula alapján azonnal jön. A teljesség kedvéért mi most mégis levezetjük (5.10) formulát is, ugyanis erre az eredményre szükségünk van az (5.9) formula második tagjában szereplő kifejezés sorfejtéséhez. Ezért a tétel bizonyításához először a

$$
Z(t):=\frac{1}{\sqrt{t(1-t)}}B(t),\qquad 0<t<1\ , \tag{5.12}
$$

Gauss-folyamat Karhunen–Loève-sorfejtését fogjuk meghatározni, aminek a kovarianciafüggvénye

$$K(s,t) := \mathrm{Cov}(Z(s),Z(t)) = \frac{\min(s,t)-st}{\sqrt{t(1-t)s(1-s)}}. \tag{5.13}$$

Ez a K kovarianciafüggvény eleme az $L^2\left((0,1)^2\right)$ térnek, de nem terjeszthető ki folytonosan a zárt $[0,1]^2$ egységnégyzetre. Ez azért baj, mert a standard Karhunen–Loève-sorfejtés (Shorack és Wellner [66] Section 5.2.) csak olyan kovarianciafüggvényű folyamatokra alkalmazható, melyek kovarianciafüggvénye a $[0,1]^2$ egységnégyzeten négyzetesen integrálható, tehát elelme az $L^2\left([0,1]^2\right)$ térnek. Anderson és Darling [4] bizonyos regularitási feltételek mellett kiterjesztette az elméletet olyan mértékben, ami már a Z folyamatot is lefedi. Cikkükben a Z folyamat meg is jelenik egy példaként. Az általános eset, vagyis súlyozott Brown-hidak Karhunen–Loève-sorfejtése Deheuvels és Martynov [32] cikkében található. A következő tételt bizonyították:

5.6. Tétel (Deheuvels és Martynov [32]). *Legyen $\psi(t), t\in(0,1)$, pozitív és folytonos függvény, melyre*

$$\lim_{t\downarrow 0} t\psi(t) = \lim_{t\uparrow 1}(1-t)\psi(t) = 0 \qquad \text{és} \qquad \int_0^1 t(1-t)\psi^2(t)\,\mathrm{d}t < \infty. \tag{5.14}$$

Tekintsük a $Z(t) := \psi(t)B(t), t\in(0,1)$, folyamatot.

(i) Léteznek $\lambda_k > 0$ konstansok és $f_k(t), t\in(0,1)$, valós függvények, $k=1,2,\ldots$, továbbá $Z_1, Z_2, \ldots$ független, standard normális eloszlású véletlen változók, hogy

$$Z(t) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} Z_k f_k(t), \qquad t\in(0,1), \tag{5.15}$$

majdnem biztosan.

(ii) Tetszőleges $k=1,2,\ldots$, esetén az f_k függvény megkapható $f_k(t)=y_k(t)\psi(t)$, $t\in(0,1)$, alakban, ahol az y_k a $(0,1)$ intervallumon kétszer folytonosan differenciálható, és megoldása az

$$y_k''(t) + \frac{1}{\lambda_k}\psi(t)y_k(t) = 0, \qquad t\in(0,1), \tag{5.16}$$

differenciálegyenletnek az $y(0)=y(1)=0$ kezdetiérték feltételek mellett.

Itt valójában λ_k és $f_k(t), t\in(0,1)$, $k=1,2,\ldots$, a $Z(t)$ folyamathoz tartozó kovariancia operátor sajátértékei és sajátfüggvényei. Legyen $T : L^2(0,1) \to L^2(0,1)$ operátor, amely minden $f\in L^2(0,1)$ esetén a

$$Tf(t) = \int_0^1 K(s,t)f(s)\,\mathrm{d}s, \qquad t\in(0,1),$$

hozzárendeléssel van megadva, ahol a K függvény a Z folyamat kovarianciafüggvénye. Ekkor a T operátornak megszámlálható sok, egymástól különböző $\lambda_k > 0$, $\lambda_k \downarrow 0$, sajátértéke és a hozzá tartozó $f_k(t), t\in(0,1)$, $L^2(0,1)$-beli sajátfüggvénye van, melyek teljes, ortonormált megoldásrendszere a $Tf_k(t) = \lambda_k f_k(t), t\in(0,1)$ egyenletnek.

A mi esetünkben

$$\psi(t) := \frac{1}{\sqrt{t(1-t)}}, \qquad t \in (0,1), \tag{5.17}$$

a súlyfüggvény. Erre az esetre alkalmazva az 5.6. Tételt a következő állítás érvényes.

5.7. Állítás. *Tekintsük a*

$$Z(t) = \frac{1}{\sqrt{t(1-t)}} B(t), \qquad t \in (0,1)$$

folyamatot. Ekkor létezik $(Z_k)_{k=1}^{\infty}$ független, standard normális eloszlású véletlen változók sorozata úgy, hogy

$$Z(t) = \sum_{k=1}^{\infty} \sqrt{\frac{1}{k(k+1)}} Z_k \sqrt{\frac{(2k+1)(k+1)}{k}} P_{k-1}^{(1,1)}(2t-1)\sqrt{t(1-t)}, \tag{5.18}$$

ahol $P_k^{(1,1)}(x), x \in (-1,1), k \in \mathbb{N}$, a k-adik, (1,1) paraméterű Jacobi ortogonális polinomokat jelöli.

Bizonyítás. Az állítás bizonyításához először az (5.14) feltételt kell ellenőriznünk. Az (5.17) definíciójából kapjuk, hogy a

$$\lim_{t \downarrow 0} t \frac{1}{\sqrt{t(1-t)}} = \lim_{t \downarrow 0} \sqrt{\frac{t}{1-t}} = 0 \qquad \lim_{t \uparrow 1} (1-t) \frac{1}{\sqrt{t(1-t)}} = \lim_{t \uparrow 1} \sqrt{\frac{1-t}{t}} = 0$$

és

$$\int_0^1 t(1-t) \left(\frac{1}{\sqrt{t(1-t)}} \right)^2 \mathrm{d}t = 1 < \infty$$

feltételek teljesülnek. Ekkor az 5.6. Tétel szerint léteznek olyan $\lambda_k > 0$ sajátértékek $f_k(t)$, $t \in (0,1), k = 1,2,\ldots,$ sajátfüggvényekkel, melyek megkaphatók az

$$f(t) = y(t) \frac{1}{\sqrt{t(1-t)}}$$

formulából, továbbá az y függvényt definiáló (5.16) differenciálegyenlet az

$$y''(t) + \frac{1}{\lambda} \frac{1}{t(1-t)} y(t) = 0 \tag{5.19}$$

alakot ölti a

$$y(0) = 0 \quad \text{és} \quad y(1) = 0 \tag{5.20}$$

kezdetiérték feltételekkel. A $t = \frac{x+1}{2}$ és $u(x) = y(\frac{x+1}{2})$ helyettesítéssel a (5.19) differenciálegyenlet az

$$u''(x) + \frac{1}{\lambda} \frac{1}{1-x^2} u(x) = 0, \qquad -1 < x < 1, \tag{5.21}$$

alakra hozható. Ez az egyenlet a Jacobi egyenlet az $\alpha = 1$ és $\beta = 1$ paraméterekkel, és $u(-1) = u(1) = 0$ kezdetiérték feltételekkel. Abramowitz és Stegun [1], 22.6.2 szerint ennek az egyenletnek pontosan akkor van megoldása, ha a λ

$$\lambda_k = \frac{1}{k(k+1)}, \qquad k = 1,2,\dots, \tag{5.22}$$

alakú. Továbbá ennek a kezdeti érték problémának a teljes megoldáshalmaza felírható a $P_k^{(1,1)}(x), x \in (-1,1),\ k = 0,1\dots,$ Jacobi ortogonális polinomok kifejezéseként

$$u(x) = (1-x)(1+x)P_k^{(1,1)}(x), \qquad x \in (-1,1),$$

formában (lásd Abramowitz és Stegun [1], 22.6.2). Ennélfogva az (5.19)-(5.20) eredeti kezdetiérték probléma megoldása

$$y(t) = 2(1-t)2tP_k^{(1,1)}(2t-1), \qquad t \in (0,1).$$

Abramowitz és Stegun [1], 22.2.1 alapján

$$\int_{-1}^{1} \left(P_k^{(1,1)}(x)\right)^2 (1-x)(1+x)\,\mathrm{d}x = \frac{2^3}{2k+3}\frac{k+1}{k+2},$$

amiből azonal származtathatók az

$$f_k(t) = \sqrt{\frac{(2k+1)(k+1)}{k}} P_{k-1}^{(1,1)}(2t-1)\sqrt{t(1-t)}, \quad k = 1,2,\dots, \tag{5.23}$$

ortonormált sajátfüggvények az (5.22) sajátértékekkel.

Továbbá, az 5.6. Tétel szerint létezik $(Z_k)_{k=1}^{\infty}$ független, standard normális eloszlású véletlen változók sorozata úgy, hogy a Z sztochasztikus folyamatnak van Karhunen–Loève kiterjesztése. Ezzel beláttuk az állítást. □

Szükségünk van még a következő lemmára ahhoz, hogy meghatározzuk a V eloszlásának végtelen soros reprezentációjában szereplő együtthatókat.

5.8. Lemma. *A következő formula érvényes:*

$$\int_{-1}^{1} P_n^{(1,1)}(x)(1-x^2)\ln\left(\frac{1+x}{1-x}\right)\,\mathrm{d}x = \begin{cases} \frac{8}{(2k+1)(2k+3)(k+2)}, & \text{ha} \quad n = 2k+1, \\ 0, & \text{ha} \quad n = 2k. \end{cases}$$

Bizonyítás. A lemma bizonyításához szükségünk van az (1,1) paraméterű Jacobi-polinomok generátorfüggvényére, ami Abramowitz és Stegun [1], 22.9 alapján írható fel:

$$\sum_{n=0}^{\infty} P_n^{(1,1)}(x)z^n = \frac{4}{R(1-z+R)(1+z+R)}, \qquad x, z \in (-1,1), \tag{5.24}$$

ahol

$$R = R(x,z) = \sqrt{1-2zx+z^2} = \sqrt{(x-z)^2+(1-x^2)} \in \mathbb{R}.$$

Tekintsünk az

$$f(x) = (1-x^2)\ln\left(\frac{1+x}{1-x}\right), \qquad -1 < x < 1,$$

függvényt. Az Abramowitz és Stegun [1], 4.1.28 alapján kapjuk a következő sorfejtést

$$\ln\left(\frac{1+x}{1-x}\right) = \sum_{l=0}^{\infty} \frac{2}{2l+1} x^{2l+1}, \qquad |x| < 1. \tag{5.25}$$

Ekkor minden $-1 < x < 1$ esetén

$$|f(x)| \le (1-x^2)|x| \sum_{l=0}^{\infty} \frac{2}{2l+1}|x|^{2l} \le (1-x^2)\left(\sum_{l=0}^{\infty}(x^2)^l + 1\right) = (1-x^2)\left(\frac{1}{1-x^2}+1\right) \le 2.$$

Legyen

$$a_n := \int_{-1}^{1} P_n^{(1,1)}(x) f(x) \mathrm{d}x, \qquad n = 0,1,\ldots,$$

a keresett integrál, és legyen az $a_0, a_1, \ldots$ valós sorozat generátorfüggvénye

$$g(z) = \sum_{n=0}^{\infty} a_n z^n, \qquad -1 < z < 1. \tag{5.26}$$

Mivel Abramowitz és Stegun [1], 22.14.1 alapján $|P_n^{(1,1)}(x)| \le n+1,\ -1 \le x \le 1$, ezért $|a_n| \le 4(n+1)$, tehát ez a hatványsor abszolút és egyenletesen konvergens a $[-1/2,1/2]$ kompakt halmazon. Rögzítsünk egy $z \in (0,1/2]$ számot. Ekkor az (5.24) azonosság és a majoráns konvergenciatétel alkalmazásával kapjuk, hogy

$$\begin{aligned} g(z) &= \sum_{n=0}^{\infty} \int_{-1}^{1} P_n^{(1,1)}(x) z^n f(x)\,\mathrm{d}x = \int_{-1}^{1} \sum_{n=0}^{\infty} P_n^{(1,1)}(x) z^n f(x)\,\mathrm{d}x \\ &= \int_{-1}^{1} \frac{4}{R(1-z+R)(1+z+R)} f(x)\,\mathrm{d}x = \int_{-1}^{1} \frac{4(1-x)(1+x)}{R(1-z+R)(1+z+R)} \ln\left(\frac{1+x}{1-x}\right)\,\mathrm{d}x. \end{aligned}$$

Ahhoz, hogy megtaláljuk ennek az integrálnak az értékét, alkalmazzuk az $u = \sqrt{1-2zx+z^2}$ helyettesítést. Ekkor

$$x = \frac{z^2+1-u^2}{2z}, \quad 1-x = \frac{(u-z+1)(u+z-1)}{2z}, \quad 1+x = \frac{(u+z+1)(1+z-u)}{2z},$$

valamint az $x = x(u)$ leképezés szigorúan csökkenő módon képezi bele az $[1-z,1+z]$ intervallumot a $[-1,1]$ intervallumba. Azt kapjuk, hogy

$$g(z) = \int_{1-z}^{1+z} \frac{(u+z-1)(z+1-u)}{z^3} \ln\left(\frac{(z+1+u)(z+1-u)}{(u-z+1)(u+z-1)}\right)\mathrm{d}u.$$

A következő lépésben egy parciális integrálást végzünk el. Ehhez vegyük észre, hogy

$$\left(\frac{z^2(u-1) - \frac{1}{3}(u-1)^3}{z^3}\right)' = \frac{(u+z-1)(z+1-u)}{z^3},$$

$$\left(\ln\left(\frac{(z+1+u)(z+1-u)}{(u-z+1)(u+z-1)}\right)\right)' = \frac{8uz}{(u^2-(z+1)^2)(u^2-(1-z)^2)}.$$

A parciális integrálás után azt kapjuk, hogy

$$g(z) = \lim_{\varepsilon\to 0}\left[\frac{z^2(u-1)-\frac{1}{3}(u-1)^3}{z^3}\ln\left(\frac{(z+1+u)(z+1-u)}{(u-z+1)(u+z-1)}\right)\right]_{1-z+\varepsilon}^{1+z-\varepsilon}$$
$$-\int_{1-z}^{1+z}\frac{z^2(u-1)-\frac{1}{3}(u-1)^3}{z^3}\frac{8uz}{(u^2-(z+1)^2)(u^2-(1-z)^2)}\mathrm{d}u = g_1(z)-g_2(z)$$

Az első tagra azt kapjuk, hogy

$$g_1(z) = \lim_{\varepsilon\to 0}\frac{z^2(z-\varepsilon)-\frac{1}{3}(z-\varepsilon)^3}{z^3}\left(\ln\left(\frac{(2+2z-\varepsilon)\varepsilon}{(2-\varepsilon)(2z-\varepsilon)}\right)+\ln\left(\frac{(2+\varepsilon)(2z-\varepsilon)}{(2-2z+\varepsilon)\varepsilon}\right)\right)$$
$$= \lim_{\varepsilon\to 0}\frac{z^2(z-\varepsilon)-\frac{1}{3}(z-\varepsilon)^3}{z^3}\ln\left(\frac{(2+\varepsilon)(2+2z-\varepsilon)}{(2-\varepsilon)(2-2z+\varepsilon)}\right) = \frac{2}{3}\ln\left(\frac{1+z}{1-z}\right),$$

illetve a második tag parciális törtekre bontható, és

$$g_2(z) = \int_{1-z}^{1+z}\left\{-\frac{8}{3z^2}-\frac{2z^3-12z}{3z^3}\left(\frac{1}{u+z+1}+\frac{1}{u-z+1}\right)\right.$$
$$\left.+\frac{8}{3z^3}\left(\frac{1}{u+z+1}-\frac{1}{u-z+1}\right)+\frac{2}{3}\left(\frac{1}{u+z-1}+\frac{1}{u-z-1}\right)\right\}\mathrm{d}u$$
$$= \lim_{\varepsilon\to 0}\left[-\frac{8}{3z^2}u-\frac{2z^3-12z}{3z^3}\ln((u+z+1)(u-z+1))\right.$$
$$\left.+\frac{8}{3z^3}\ln\left(\frac{u+z+1}{u-z+1}\right)+\frac{2}{3}\ln((u+z-1)(u-z-1))\right]_{1-z+\varepsilon}^{1+z-\varepsilon}$$
$$= -\frac{16}{3z}-\frac{2z^3-12z}{3z^3}\ln\left(\frac{1+z}{1-z}\right)+\frac{8}{3z^3}\ln\left((1+z)(1-z)\right).$$

A fentiek összegzése azt adja, hogy

$$g(z) = \frac{4}{3z^3}\left[4z^2+(z^3-3z)\ln\left(\frac{1+z}{1-z}\right)-2\ln\left((1+z)(1-z)\right)\right], \qquad z\in(0,1/2].$$

Ezután az (5.25) sorfejtést ismételten alkalmazva

$$g(z) = \frac{4}{3z^3}\left[4z^2+(z^3-3z)\sum_{l=0}^{\infty}\frac{2}{2l+1}z^{2l+1}-2\sum_{l=0}^{\infty}-\frac{2}{2l+2}z^{2l+2}\right]$$
$$= \frac{16}{3}\frac{1}{z}+\sum_{l=0}^{\infty}\frac{8}{3}\frac{1}{2l+1}z^{2l+1}-\sum_{l=0}^{\infty}\frac{8}{2l+1}z^{2l-1}+\sum_{l=0}^{\infty}\frac{16}{3}\frac{1}{2l+2}z^{2l-1}$$
$$= \frac{16}{3}\frac{1}{z}+\sum_{k=0}^{\infty}\left(\frac{8}{3}\frac{1}{2k+1}-\frac{8}{2k+3}+\frac{8}{3}\frac{1}{k+2}\right)z^{2k+1}-8\frac{1}{z}+\frac{8}{3}\frac{1}{z}$$
$$= 8\sum_{k=0}^{\infty}\frac{z^{2k+1}}{(2k+1)(2k+3)(k+2)}.$$

Ebből és az (5.26) formulából azonnal következik, hogy

$$a_{2k}=0 \qquad \text{és} \qquad a_{2k+1}=\frac{8}{(2k+1)(2k+3)(k+2)}, \qquad k=0,1,\dots.$$

Ezzel beláttuk a lemmát. □

A 5.5. Tétel bizonyítása. Az (5.15) Karhunen–Loève-sorfejtés a Z folyamat $L^2(0,1)$ Hilbert-térben megadott sorfejtése az $(f_k)_{k=1}^{\infty}$ ortonormált bázisra nézve. A Parseval-azonosságot alkalmazva

$$\int_0^1 \frac{B^2(t)}{t(1-t)}\mathrm{d}t = \|Z\|_{L^2(0,1)} = \sum_{k=1}^{\infty}(\sqrt{\lambda_k}Z_k)^2 = \sum_{k=1}^{\infty}\frac{1}{k(k+1)}Z_k^2.$$

Vegyük észre, hogy $f_1(t)=\sqrt{6}\sqrt{t(1-t)}, 0<t<1$, így

$$\int_0^1 B(t)\mathrm{d}t = \frac{1}{\sqrt{6}}\int_0^1 Z(t)\sqrt{6t(1-t)}\,\mathrm{d}t = \frac{1}{\sqrt{6}}\langle Z,f_1\rangle_{L^2(0,1)} = \frac{1}{\sqrt{6}}\sqrt{\lambda_k}Z_1,$$

amiből

$$\left[\int_0^1 B(t)\mathrm{d}t\right]^2 = \frac{1}{6}\lambda_1 Z_1^2.$$

Tekintsük a

$$h(t) := \sqrt{t(1-t)}\ln\left(\frac{t}{1-t}\right), \qquad 0<t<1,$$

függvényt, ami az $L^2(0,1)$ tér eleme. Először meghatározzuk a h függvény sorfejtését. Az együtthatókat az 5.8. Lemma segítségével kapjuk meg:

$$\begin{aligned}
\langle h,f_k\rangle_{L^2(0,1)} &= \int_0^1 h(t)f_k(t)\,\mathrm{d}t \\
&= \int_0^1 \sqrt{t(1-t)}\ln\left(\frac{t}{1-t}\right)\sqrt{\frac{(2k+1)(k+1)}{k}}P_{k-1}^{(1,1)}(2t-1)\sqrt{t(1-t)}\,\mathrm{d}t \\
&= \sqrt{\frac{(2k+1)(k+1)}{k}}\frac{1}{8}\int_{-1}^1 P_{k-1}^{(1,1)}(x)(1-x^2)\ln\left(\frac{1+x}{1-x}\right)\,\mathrm{d}x \\
&= \sqrt{\frac{(2k+1)(k+1)}{k}}\begin{cases}\frac{1}{(2l+1)(2l+3)(l+2)}, & \text{ha} \quad k-1=2l+1,\\ 0, & \text{ha} \quad k-1=2l.\end{cases} \\
&= \begin{cases}\sqrt{\frac{4l+5}{(2l+1)^2(2l+2)(2l+3)(l+2)^2}}, & \text{ha} \quad k=2l+2,\\ 0, & \text{ha} \quad k=2l+1,\end{cases}
\end{aligned}$$

tehát

$$h(t)=\sum_{l=0}^{\infty}\sqrt{\frac{4l+5}{(2l+1)^2(2l+2)(2l+3)(l+2)^2}}f_{2l+2}(t), \qquad 0<t<1.$$

És végül az 5.7. Állításból és a Parseval-azonosságból azt kapjuk, hogy

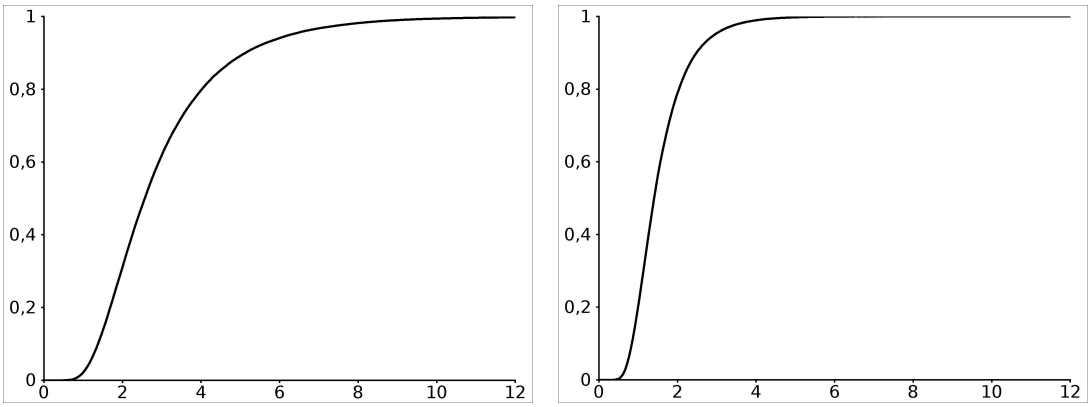
$$\int_0^1 B(t)\ln\left(\frac{t}{1-t}\right)\,\mathrm{d}t = \int_0^1 Z(t)\sqrt{t(1-t)}\ln\left(\frac{t}{1-t}\right)\,\mathrm{d}t = \int_0^1 Z(t)h(t)\,\mathrm{d}t$$
$$= \langle Z,h\rangle_{L^2(0,1)} = \sum_{l=0}^{\infty}\sqrt{\frac{1}{(2l+2)(2l+3)}}Z_{2l+2}\sqrt{\frac{4l+5}{(2l+1)^2(2l+2)(2l+3)(l+2)^2}}$$
$$= \sum_{l=1}^{\infty}\frac{\sqrt{4l+1}}{2l(l+1)(2l-1)(2l+1)}Z_{2l}.$$

Összekapcsolva a fenti eredményeket a megfelelő konstans együtthatókkal, az (5.11) végtelen soros reprezentációt kapjuk. □

5.3. Szimuláció

5.3.1. Az nV_n és nW_n tesztstatisztikák eloszlásai és aszimptotikus eloszlásai

A határ véletlen változók eloszlásfüggvényét numerikusan számítottuk ki az (5.10) és (5.11) végtelen soros reprezentációkat használva. A W és V határ véletlen változókat 200 000 példányban generáltuk le, a változókat definiáló sorok első 10 000 tagját vettük, és numerikusan számítottuk ki a H_l és $H_{l,s}$ határeloszlásfüggvények empirikus változatát. Ezeket a mennyiségeket (ismétlések számát, levágás helyét) úgy választottuk meg, hogy ezen paraméterek mellett a H_l és $H_{l,s}$ határeloszlásfüggvények empirikus változatai két tizedesjegy pontossággal stabilizálódtak. A 5.1. ábrán láthatók a határeloszlások.



5.1. ábra. A W határ véletlen változó eloszlásfüggvénye (balra) és ugyanez a V véletlen változóra (jobbra).

Különböző mintaméretek mellett, $n=20$-tól $n=500$-ig, az nW_n és nV_n tesztstatisztikák empirikus eloszlásfüggvényét szimuláltuk ugyancsak 200 000 ismétléssel. Amint a 5.2. ábrán látható, a konvergencia mindenhol nagyon gyors. A 5.1. táblázat részletesen mutatja az nW_n és nV_n tesztstatisztikák empirikus kritikus értékeit 0,15, 0,10, 0,05 és 0,01

szignifikanciaszintek mellett. Az utolsó sor, az $n{=}\infty$, mindkét teszt aszimptotikus kritikus értékeit tartalmazza, melyeket a W és V változók eloszlásából határoztunk meg.

5.1. táblázat. Az nW_n és nV_n tesztstatisztikák empirikus kritikus értékei különböző mintaméretek és különböző szignifikanciaszintek mellett.

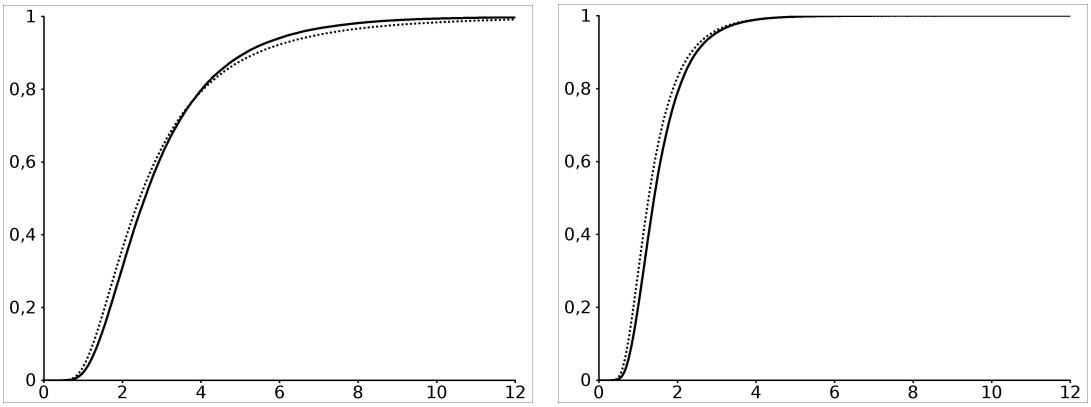
	nW_n					nV_n			
n	0,15	0,10	0,05	0,01	n	0,15	0,10	0,05	0,01
20	4,60	5,43	7,00	11,40	20	2,07	2,34	2,83	4,02
50	4,52	5,25	6,66	10,76	50	2,21	2,49	2,99	4,17
100	4,49	5,20	6,50	10,40	100	2,24	2,52	2,99	4,13
200	4,48	5,15	6,39	9,87	200	2,24	2,52	2,99	4,14
500	4,47	5,13	6,31	9,39	500	2,23	2,51	2,97	4,06
∞	4,47	5,12	6,26	8,98	∞	2,22	2,49	2,95	4,02

A konvergencia gyorsasága miatt kis mintaméret esetén is használhatóak az aszimptotikus kritikus értékek. A következő fejezetben bemutatunk egy további szimulációs tanulmányt, melyben az nW_n és nV_n tesztstatisztikák néhány alternatívával szembeni erejét vizsgáljuk. Ezen tanulmányban a véges kritikus értékeket használtuk.

5.3.2. Az nV_n és nW_n tesztek ereje

Elvégeztünk egy szimulációs vizsgálatot, hogy meghatározzuk a tesztek erejét néhány folytonos alternatívával szemben. Az eloszlások pontos definiciója található a következő listában, ahol $Z \sim N(0,1)$ a standard normális véletlen változót jelöli.

Az alternatív eloszlások:



5.2. ábra. (balra) Az nW_n tesztstatisztika eloszlásfüggvénye $n = 20$ mintaméretnél (pontozott vonal) és a W határeloszlásfüggvénye (vastagabb vonal), valamint (jobbra) a nV_n tesztstatisztika eloszlásfüggvénye $n = 20$ mintaméretnél (pontozott vonal) és a V határeloszlásfüggvénye (vastagabb vonal).

1. Beta(p,q), $p,q>0$, jelölje a Béta eloszlást, melynek sűrűségfüggvénye

$$f(t)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}t^{p-1}(1-t)^{q-1}, \qquad 0<t<1,$$

ahol $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x$, $\alpha\in(0,\infty)$.

2. A Cauchy-eloszlás, melynek sűrűségfüggvénye

$$f(t)=\frac{1}{\pi}\frac{1}{1+t^2},\ t>0.$$

3. Az Egyenletes eloszlás, melynek sűrűségfüggvénye $f(t)=1,\ 0\le t\le 1$.

4. Az Exponenciális(λ) eloszlás, melynek sűrűségfüggvénye $f(t)=\lambda e^{-\lambda t},\ t>0$.

5. A Gamma(α,λ), $\alpha,\lambda>0$, eloszlás, melynek sűrűségfüggvénye

$$f(t)=\frac{\lambda^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t}, \qquad t>0.$$

6. A χ^2_n-eloszlás az

$$X_1^2+X_2^2+\cdots+X_n^2$$

eloszlása, ahol $n>1$, és $X_1,X_2,\dots X_n$ független standard normális véletlen változók.

7. A Laplace-eloszlás, melynek sűrűségfüggvénye $f(t)=e^{-|t|/2},\ t\in\mathbb{R}$.

8. A Lognormal eloszlás a e^Z véletlen változó eloszlása.

9. A Negatív Exponenciális eloszlás, melynek sűrűségfüggvénye $f(t)=\lambda e^{\lambda t},\ t<0$.

10. Az Student(n)-eloszlás az

$$\frac{Y}{\sqrt{\frac{X_1^2+X_2^2+\cdots+X_n^2}{n}}}$$

eloszlása, ahol $n>1$, és $Y,X_1,X_2,\dots X_n$ független standard normális véletlen változók.

11. Két háromszög eloszlás, Triangle(I) és Triangle(II), melyek rendre az alábbi sűrűségfüggvényekkel vannak definiálva:

$$f(t)=1-|t|,\ -1\le t\le 1, \qquad \text{és} \qquad f(t)=2-2t,\ 0\le t\le 1.$$

12. A Weibull(k), $k>0$, eloszlás, melynek sűrűségfüggvénye

$$f(t)=kt^{k-1}e^{-t^k}, \qquad t>0.$$

Minkét teszt és minden mintaméret esetében az adott mintaérethez tartozó empirikus kritikus értékeket használtuk. Az empirikus erők 200 000 ismétlésből származnak minden mintaméret ($n = 20, 50$ és 100) és mindkét teszt esetében. A részletek a 5.2. táblázatban találhatók. Összehasonlítottuk az új nV_n tesztet eltolás-skála család esetében az empirikus karakterisztikus függvényre és empirikus momentum-generáló függvényre alapozott Meintanis-tesztekkel [58]. Az összehasonlításhoz a Table 3 értékeit használtuk a [58] cikkből. Ez a táblázat a Meintanis-tesztek ereje mellett tartalmazza a klasszikus EDF-tesztek (Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling, Watson) erejét $n = 20$ és $n = 50$ mintaméret, valamint 0,10 szignifikanciaszint mellett. A [58] cikkben minden teszt esetén az eltolás és skála paramétereket momentum vagy maximum likelihood módszerrel becsülik, ezáltal válnak alkalmassá összetett nullhipotézis tesztelésére. A Cauchy és Laplace alternatívákkal szemben az új nV_n tesztnek a legnagyobb az ereje. Ugyanezen alternatívák esetében az EDF-tesztek jobban teljesítenek, mint a Meintanis-tesztek. Az összes többi alternatíva esetében a Meintanis-tesztek a legerősebbek és az új tesztnek van a legkisebb ereje.

Ha a logisztikus eltolás családot teszteljük az nW_n teszt segítségével, akkor jobb erőket kapunk, mint a logisztikus eltolás-skála család esetében, kivéve a Gamma, Lognormal és χ^2_1 alternatívákkal szemben.

Általános konkluziója ennek a szimulációs vizsgálatnak, hogy minkét esetben könnyen számolható tesztstatisztikával és akár az aszimptotikus kritikus értékekkel is dolgozhatunk. A logisztikus eltolás család esetében erősebb, mialatt a logisztikus eltolás-skála család esetében kevésbé erős tesztet kapunk.

5.2. táblázat. Az nW_n és nV_n tesztek %-ban megadott empirikus ereje néhány alternatívával szemben, $n = 20, 50$ és 100 mintaméret és α szignifikanciaszint mellett ($*$ a 100% empirikus erőt jelöli).

	nW_n			nW_n			nV_n			nV_n		
Mintaméret	20	50	100	20	50	100	20	50	100	20	50	100
$N(0,1)$	47	99	*	22	96	*	5	6	8	2	2	4
Egyenletes	*	*	*	*	*	*	13	47	93	5	29	82
Cauchy	88	99	*	84	99	*	88	99	*	84	99	*
Laplace	27	76	97	12	61	93	26	39	55	17	29	43
Exp(1)	88	*	*	69	*	*	70	99	*	56	97	*
Triangle(I)	*	*	*	*	*	*	4	7	13	2	3	6
Triangle(II)	*	*	*	*	*	*	21	61	97	11	43	91
Beta(2;2)	*	*	*	*	*	*	6	15	40	2	7	24
Weibull(2)	*	*	*	*	*	*	12	25	54	5	15	38
Gamma(2,1)	25	83	*	10	62	99	40	81	99	27	69	98
Lognormal	80	*	*	61	*	*	86	*	*	79	*	*
Student(5)	27	82	99	11	67	98	16	19	21	10	12	13
χ^2_1	88	*	*	71	*	*	94	*	*	88	*	*
Negatív Exp	88	*	*	69	*	*	69	99	*	56	97	*
α	0,10			0,05			0,10			0,05		

Összefoglalás

Bevezetés

A disszertációban illeszkedésvizsgálattal kapcsolatos eredményeket taglalunk. Legyen $X_1, \ldots, X_n$ minta (független, azonos eloszlású véletlen változók) egy ismeretlen $F(x),\ x \in \in \mathbb{R}$, eloszlásfüggvényű véletlen változóból. Több különböző módszerrel, több eloszlás esetén tesztelni szeretnénk azt az egyszerű nullhipotézist, hogy

$$\mathcal{H}_0 : F = F_0,$$

ahol $F_0(x),\ x \in \mathbb{R}$, egy rögzített eloszlásfüggvény, valamint azt az összetett nullhipotézist, hogy

$$\mathcal{H}_0 : F \in \mathcal{F},$$

ahol $\mathcal{F}$ egy eloszláscsaládot jelöl.

A disszertáció a következőképpen épül fel. A 2. fejezetben a disszertáció szempontjából fontos történeti előzményeket gyűjtöttük össze. A 3. fejezetben egy eljárást javasoltunk egyenletes eloszlás esetén egyszerű, illetve összetett illeszkedésvizsgálatra, valamint az új teszteket megvizsgáljuk egy szimulációs tanulmányban. A 4. fejezetben az L^2-Wasserstein távolságot használó del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] által bevezetett normalitás teszt szimulációs vizsgálatát mutatjuk be. Az 5. fejezetben Csörgő S. [19], [20] által kidolgozott súlyozott kvantilis korreláció tesztet vezetjük be logisztikus eloszláscsalád esetében, és bemutatjuk az új teszttel kapcsolatos szimulációs vizsgálat eredményét.

Történeti előzmények

Az első alfejezetben felidézzük az első módszereket, amelyekkel rögzített eloszláshoz való illeszkedést lehet tesztelni valamint azt is, hogy hogyan találták meg ezen tesztstatisztikák határeloszlását. Az első illeszkedésvizsgálatra használt eljárás a Pearson-féle χ^2-teszt, amely aszimptotikusan χ^2 eloszlású a nullhipotézis teljesülése mellett. Majd az empirikus és a hipotetikus eloszlásfüggvény különböző távolságait használó tesztek, az EDF-tesztek bemutatása következik határeloszlásaik izgalmas megtalálásával. A második alfejezetben a számunkra érdekes első összetett illeszkedésvizsgálati módszereket és határeloszlásukat elevenítjük fel. Az első vizsgálatok normális eloszláscsalád esetében történtek. Majd bemutatjuk, hogy az első alfejezetbeli rögzített eloszláshoz való illeszkedésvizsgálatra használt

módszerek alkalmasak parametrikus eloszládcsaládhoz való illeszkedés ellenőrzésére. A paraméterek becslése után egy a becsült paraméterű eloszláshoz való illeszkedést kell vizsgálni, illetve a becsléses tesztstatisztikák aszimptotikus viselkedését. Végül a regresszió-, illetve korrelációteszteket idézzük fel. Bemutatjuk a Wilk–Shapiro normalistástesztet [65], ennek további változatait, valamint, hogy hogyan sikerült meghatározni a határeloszlását.

Illeszkedésvizsgálat egyenletes eloszlás esetében

A 3. fejezet tartalmazza a Krauczi [59] cikk eredményeit. Egy eljárást vezetünk be egyenletesség tesztelésére klaszterszámok segítségével. Legyenek $U_1, \ldots, U_n$ független, a [0,1] intervallumon egyenletes eloszlású véletlen változók, egy minta. Emellett adott egy determinisztikus $d_n \in (0,1)$ távolságszint minden mintaméréthez. A [0,1] intervallumon húzzuk végig ezt a távolságszintet, és figyeljük meg, hogy a rendezett minta elemei hány osztályba esnek. Egy klaszterbe azok az elemei tartoznak a rendezett mintának, amelyekre teljesül az, hogy az egymást követő elemek távolsága nem nagyobb, mint d_n. Egy adott mintához és távolságszinthez tartozó osztályok számát nevezzük klaszterszámnak.

Az első alfejezetben felelevenítjük, hogy Csörgő S. és Wu [23] három különböző asszimptotikus viselkedésű távolságszint sorozat mellett bizonyították a klaszterek számának aszimptotikus normalitását, és még rátát is adtak az eloszlásfüggvények konvergenciájának sebességére. Ennek a tételnek bizonyítjuk a többdimenziós változatait különböző intervallumon egyenletes eloszlások esetében, majd használjuk egyenletesség tesztelésére ismert és ismeretlen intervallumon.

A második alfejezet az elméleti eredményeket tartalmazza. Ebben bebizonyítjuk a Csörgő–Wu-féle, különböző távolságszintekhez tartozó klaszterszámok együttes aszimptotikus normalitását három esetben: ha a minta a [0,1], ha az ismert $[a,b]$, illetve ha egy ismeretlen intervallumon egyenletes eloszlásból származik.

Tekintsünk $J \geq 1$ darab $d_{n1} \leq d_{n2} \leq \ldots \leq d_{nJ}$, $n \in \mathbb{N}$, távolságszint sorozatot. Ha a minta a [0,1] intervallumon egyenletes eloszlásból származik, akkor $K_{nj}(d_{nj})$ jelölje a d_{nj} távolságszinthez tartozó klaszterek számát minden n és j esetén. Tekintsük a

$$\mathbf{K}_n = \frac{1}{\sqrt{n}} \left(\frac{K_{n1}(d_{n1}) - m_{n1}}{\sigma_{n1}}, \ldots, \frac{K_{nJ}(d_{nJ}) - m_{nJ}}{\sigma_{nJ}} \right)^\top, \qquad n \in \mathbb{N},$$

a véletlen vektorváltozók sorozatát az $m_{nj} = ne^{-nd_{nj}}$ és

$$\sigma_{nj} = \sqrt{e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2)}, \qquad j = 1, \ldots, J,$$

centralizáló és normalizáló sorozattal. Tegyük fel, hogy a távolságszint sorozatok mindegyike kielégíti az alábbi feltételek valamelyikét:

(T1) $nd_{nj} \to 0$, $n^2 d_{nj} \to \infty$;

(T2) $0 < \liminf_n nd_{nj} \leq \limsup_n nd_{nj} < \infty$;

(T3) $nd_{nj} \to \infty$, $ne^{-nd_{nj}} \to \infty$.

Továbbá, tegyük fel, hogy léteznek

$$s_{ij} := \lim_{n\to\infty} \frac{e^{-nd_{ni}-nd_{nj}}(e^{nd_{ni}} - 1 - n^2 d_{ni} d_{nj})}{\sigma_{ni}\sigma_{nj}} \in \mathbb{R}, \qquad 1 \leq i < j \leq J,$$

határértékek, és legyen $s_{jj} := 1$ és $s_{ji} := s_{ij}$. Vezessük be $\Sigma := (s_{ij})_{i,j=1,\dots,J}$ mátrixot. Ezen feltételek mellett a

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma)$$

konvergenciát bizonyítjuk. Továbbá bebizonyítjuk egy következményben, hogy diagonális kovarianciamátrixú normális határeloszlás is kapható megfelelő távolságszint sorozatok esetén.

Ha a minta ismert $[a, b]$ intervallumon egyenletes eloszlásból származik, akkor bebizonyítjuk az $[a, b]$ és [0,1] intervallumok közötti lineáris transzformáció segítségével, hogy a transzformált klaszterszám vektor ugyancsak normális eloszlású lesz megfelelően transzformált feltételek mellett.

A harmadik esetben a minta egy ismeretlen intervallumon egyenletes eloszlásból származik. Legyenek $V_1, V_2, \dots, V_n$ független, egy ismeretlen $[a, b]$ intervallumon egyenletes eloszlású véletlen változók, ahol $a, b \in \mathbb{R},\ a < b$, valamint legyen $V_{1,n}, \dots, V_{n,n}$ a hozzá tartozó rendezett minta. Az intervallum végpontjait becsüljük az $\hat{a}_n = V_{1,n}$ legkisebb, és a $\hat{b}_n = V_{n,n}$ legnagyobb mintaelemmel. Jelölje $\hat{K}_{nj}(d_{nj})$ a megfelelő d_{nj} távolságszinthez tartozó klaszterszámot, $j = 1, \dots, J$. Legyenek

$$\hat{m}_{nj} = ne^{-\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}}, \qquad \hat{\sigma}_{nj} = \sqrt{e^{-2\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}} \left(e^{\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}} - 1 - \left(\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n} \right)^2 \right)}$$

valamint

$$\widehat{\mathbf{K}}_n = \frac{1}{\sqrt{n}} \left(\frac{\hat{K}_{n1}(d_{n1}) - \hat{m}_{n1}}{\hat{\sigma}_{n1}}, \dots, \frac{\hat{K}_{nJ}(d_{nJ}) - \hat{m}_{nJ}}{\hat{\sigma}_{nJ}} \right)^\top.$$

Ekkor ugyanazon feltételek mellett, mint a [0,1] intervallumon egyenletes eloszlásból származó minta esetében bebizonyítjuk, hogy

$$\widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma).$$

A harmadik alfejezet a statisztikai eredményeket és a szimulációt tartalmazza. Az elméleti eredményekből adódóan aszimptotikus χ^2-tesztet kapunk egyszerű, illetve összetett nullhipotézis ellenőrzésére.

Adott $X_1, \dots, X_n$ minta egy ismeretlen $F(x),\ x \in \mathbb{R}$, eloszlásfüggvényű véletlen változóból. Tesztelni szeretnénk azt az egyszerű nullhipotézist, hogy

$$\mathcal{H}_0 : F = F_{0,1},$$

ahol most $F_{0,1}$ a [0,1] intervallumon egyenletes eloszlás eloszlásfüggvényét jelöli. Ezen nullhipotézis és a megfelelő feltételek mellett azt kapjuk, hogy a tesztstatisztika

$$C_n := \mathbf{K}_n^\top \Sigma^{-1} \mathbf{K}_n \xrightarrow{\mathcal{D}} \chi^2_J.$$

Jelölje $\mathcal{F}$ a véges zárt intervallumon vett egyenletes eloszlások családját. Tekintsük azt az összetett nullhipotézist, hogy a minta valamelyik egyenletes eloszlásból származik, tehát

$$\mathcal{H}_0 : F \in \mathcal{F} = \{F_{a,b} : a, b \in \mathbb{R}, a < b\},$$

ahol $F_{a,b}$ az $[a,b]$ intervallumon vett egyenletes eloszlás eloszlásfüggvényét jelöli. Ekkor ezen nullhipotézis és a megfelelő feltételek mellett

$$\widehat{C}_n := \hat{\mathbf{K}}_n^\top \Sigma^{-1} \hat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \chi^2_J.$$

Ez alapján úgy tűnhet, hogy az összetett nullhipotézist lehet tesztelni az előző bekezdéshez hasonlóan. A probléma az, hogy mivel nem ismerjük az a és b pontos értékét, ezért a Σ kovarianciamátrix komponenseit se tudjuk meghatározni, emiatt a $\hat{C}_n$ statisztika egy adott minta alapján nem számolható ki. Éppen emiatt az összetett nullhipotézist egy másik módszerrel fogjuk tesztelni. Egy lehetséges megoldás, hogy a tetszőleges intervallumból származó $V_1, \ldots, V_n$ mintát a [0,1] intervallumba transzformáljuk a következőképpen:

$$\left(\frac{V_{2,n} - V_{1,n}}{V_{n,n} - V_{1,n}}, \ldots, \frac{V_{n-1,n} - V_{1,n}}{V_{n,n} - V_{1,n}} \right).$$

Jelölje $\tilde{K}_{n-2,j}(d_{nj})$ a d_{nj} távolságszinthez tartozó klaszterszámot az átskálázott minta esetén, $j = 1, \ldots, J$, és legyen

$$\tilde{\mathbf{K}}_{n-2} := \frac{1}{\sqrt{n}} \left(\frac{\tilde{K}_{n-2,1}(d_{n1}) - m_{n-2,1}}{\sigma_{n-2,1}}, \ldots, \frac{\tilde{K}_{n-2,J}(d_{nJ}) - m_{n-2,J}}{\sigma_{n-2,J}} \right)^\top$$

az átskálázott mintához tartozó normalizált klaszterszám vektor. Továbbá jelölje $\tilde{\Sigma}$ a kovarianciamátrixot az átskálázott minta esetén. Ekkor

$$C_n^{\text{mod}} := \tilde{\mathbf{K}}_{n-2}^\top \tilde{\Sigma}^{-1} \tilde{\mathbf{K}}_{n-2} \xrightarrow{\mathcal{D}} \chi^2_J.$$

Az így kapott tesztstatisztika már számolható, és ezáltal összetett nullhipotézis ellenőrzésére alkalmas.

Meghatároztuk a tesztek erejét különböző [0,1] intervallumon folytonos alternatívákkal szemben szimulációval, valamint összehasonlítotjuk az új tesztek erejét az Inglot és Ledwina [48] által bevezetett data driven smooth teszttel.

Illeszkedésvizsgálat normális eloszláscsaládra

A 4. fejezet tartalmazza a Krauczi [52] cikk eredményeit. Az L^2-Wasserstein távolságot használó del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] által bevezetett normalitás teszt szimulációs vizsgálatát mutatjuk be. Egy eltolás- és skálamentes tesztstatisztikát kaptak, amely egyrészt úgy tesztel normális eloszláscsaládhoz való tartozást, hogy minimális távolságot keres kvantilisfüggvények távolságának segítségével; másrészt aszimptotikusan ekvivalens egy korrelációteszttel.

Legyen $\mathcal{P}_2(\mathbb{R})$ azon valószínűségi mértékek halmaza $\mathbb{R}$-en, melyeknek létezik a második momentumuk. A P_1 és $P_2 \in \mathcal{P}_2(\mathbb{R})$ valószínűségi mértékek L^2-Wasserstein távolsága

$$\mathcal{W}(P_1, P_2) := \inf \left\{ [E(X_1 - X_2)^2]^{1/2}, \mathcal{L}(X_1) = P_1, \mathcal{L}(X_2) = P_2 \right\},$$

ahol $\mathcal{L}(X)$ az X véletlen változó eloszlását jelöli. Kvantilisfüggvények segítségével pontosan számolható ez a távolság:

$$\mathcal{W}(P_1, P_2) = \left[\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 \mathrm{d}t \right]^{1/2},$$

ahol F_1^{-1} illetve F_2^{-1} a P_1 illetve a P_2 eloszlásokhoz tartozó kvantilisfüggvények. Egy eloszláscsalád és egy adott eloszlás távolságát úgy definiáljuk, mint az adott eloszlásnak az eloszláscsalád elemeitől vett távolságainak infimumát. Legyen $P \in \mathcal{P}_2(\mathbb{R})$ tetszőleges valószínűségi mérték, és legyen F az eloszlásfüggvénye, μ_0 a várható értéke és σ_0 a szórása. Ekkor a P eloszlás távolságnégyzete az $\mathbf{N}$ normális eloszláscsaládtól

$$\mathcal{W}^2(P,\mathbf{N}) := \inf\{\mathcal{W}^2(P,N_\sigma^\mu), N_\sigma^\mu \in \mathbf{N}\} = \sigma_0^2 - \left(\int_0^1 F^{-1}(t)\Phi^{-1}(t)\mathrm{d}t\right)^2,$$

ahol Φ^{-1} a standard normális kvantilisfüggvényt jelöli. Ha adott egy F eloszlásfüggvényű $X_1,\dots,X_n$ véletlen minta, akkor a $\mathcal{H}_0 : F \in \mathbf{N}$ összetett nullhipotézis ellenőrzésére megadható a $\mathcal{W}(P,\mathbf{N})/\sigma_0$ hányados empirikus változata. Ekkor egy eltolás- és skálamentes statisztikát kapunk:

$$T_n := \frac{\mathcal{W}^2(F_n,\mathbf{N})}{S_n^2} = 1 - \frac{\left[\int_0^1 Q_n(t)\Phi^{-1}(t)dt\right]^2}{S_n^2} = 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi^{-1}(t)\,\mathrm{d}t\right]^2}{S_n^2},$$

ahol S_n^2 az empirikus szórásnégyzet.

Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34] megvizsgálták a tesztstatisztika nullhipotézis melletti aszimptotikus viselkedését. Két alakban sikerült előállítaniuk a határeloszlást. Az első Brown-híd funkcionáljaként, a második véletlen változók soraként. Jelölje φ a standard normális eloszlás sűrűségfüggvényét, és legyen

$$a_n = \frac{1}{n}\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{\left[\varphi(\Phi^{-1}(t))\right]^2}dt.$$

Ha $F \in \mathbf{N}$, akkor

$$\begin{aligned} n(T_n - a_n) &\xrightarrow{\mathcal{D}} \int_0^1 \frac{B^2(t) - E(B^2(t))}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t - \left[\int_0^1 \frac{B(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 - \left[\int_0^1 \frac{B(t)\Phi^{-1}(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 \\ &\overset{\mathcal{D}}{=} -\frac{3}{2} + \sum_{j=3}^{\infty}\frac{Z_j^2 - 1}{j} \end{aligned}$$

ahol $(Z_j)_{j=3}^\infty$ független, standard normális eloszlású véletlen változók sorozata.

Ennek a normalitástesztnek számos alternatívával szembeni erővizsgálatát végeztük el szimuláció segítségével, valamint összehasonlítotjuk más normalitástesztek viselkedésével. Mivel a Wilk–Shapiro-teszttel aszimptotikusan ekvivalens a „spanyolok” [34] tesztje, nem meglepő az erővizsgálat eredménye.

Illeszkedésvizsgálat logisztikus eloszláscsaládra

Az 5. fejezet tartalmazza Balogh és Krauczi [6] cikk eredményeit. Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [34], valamint del Barrio, Cuesta-Albertos és Matrán [33] által bevezetett kvantilis korreláció teszt súlyozott változatát vezetjük be

logisztikus eloszláscsalád esetében. A súlyfüggvény használatát a tesztstatisztikában egymástól függetlenül de Wet [28], [29] és Csörgő S. [19], [20] különböző motivációból javasolta. Mi a Csörgő-féle [20] eredményt a de Wet által eltolás eloszláscsalád esetére javasolt, konkrét súlyfüggvénnyel bizonyítjuk logisztikus eltolás-skála eloszláscsalád esetében.

Adott $G(x),\ x \in \mathbb{R}$, eloszlásfüggvényre valamint $\theta \in \mathbb{R}$ és $\sigma > 0$ eltolás és skála paraméterekre legyen $G_\sigma^\theta(x) = G((x-\theta)/\sigma)$, $x \in \mathbb{R}$, valamint tekintsük a

$$\mathcal{G}_{l,s} = \{G_\sigma^\theta : \theta \in \mathbb{R}, \sigma > 0\}$$

eltolás-skála családot. Jelölje $Q_G(t) = G^{-1}(t), 0 < t < 1$, a G kvantilisfüggvényét. Legyen a $w : (0,1) \to [0,\infty)$ súlyfüggvény olyan, amely a $\int_0^1 w(t)\,\mathrm{d}t = 1$ feltételt kielégíti, és definiáljuk az r-edik súlyozott momentumot

$$\mu_r(G,w) := \int_0^1 (Q_G(t))^r w(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x^r w(G(x))\mathrm{d}G(x).$$

A továbbiakban feltesszük, hogy $\mu_1(G,w)$ és $\mu_2(G,w)$ véges, és definiáljuk a súlyozott szórásnégyzetet is:

$$\nu(G,w) := \mu_2(G,w) - \mu_1^2(G,w) \geq 0\ .$$

Két eloszlásfüggvény, F és G, súlyozott L^2-Wasserstein-távolságát definiáljuk a

$$\mathcal{W}_w(F,G) := \left[\int_0^1 (Q_F(t) - Q_G(t))^2\, w(t)\,\mathrm{d}t\right]^{\frac{1}{2}}$$

mennyiséggel.

Tekintsünk egy $X_1, \ldots, X_n$ véletlen mintát egy ismeretlen F eloszlásfüggvénnyel, és legyen G egy rögzített eloszlásfüggvény. Szeretnénk tesztelni a $\mathcal{H}_0 : F \in \mathcal{G}_{l,s}$ nullhipotézist. Ebből a célból definiáljuk a minta empirikus eloszlása és a $\mathcal{G}_{l,s}$ eltolás-skála család súlyozott L^2-Wasserstein-távolságából származtatott

$$\begin{aligned} V_n &:= 1 - \frac{\left[\int_0^1 Q_n(t)Q_G(t)w(t)\mathrm{d}t - \mu_1(G,w)\int_0^1 Q_n(t)w(t)\mathrm{d}t\right]^2}{\nu(G,w)\left[\int_0^1 Q_n^2(t)w(t)\mathrm{d}t - \left(\int_0^1 Q_n(t)w(t)\mathrm{d}t\right)^2\right]} \\ &= 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\left\{\int_{\frac{k-1}{n}}^{\frac{k}{n}} Q_G(t)w(t)\mathrm{d}t - \mu_1(G,w)\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right\}\right]^2}{\nu(G,w)\left[\sum_{k=1}^n X_{k,n}^2\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t - \left(\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right)^2\right]} \end{aligned}$$

tesztstatisztikát, ahol Q_n az empirikus kvantilisfüggvényt jelöli.

A logisztikus eloszlás esetében $\mathcal{G}_{l,s}$ jelölje a logisztikus eltolás-skála családot. De Wet [29] eltoláscsaládok esetében javasolt

$$w(t) = 6t(1-t), \qquad 0 < t < 1,$$

súlyfüggvényét fogjuk használni. Ekkor a tesztstatisztika a logisztikus eltolás-skála családra

$$V_n = 1 - \frac{\left[\sum_{k=1}^{n} a_{k,n} X_{k,n}\right]^2}{\left(\frac{\pi^2}{3} - 2\right)\left[\sum_{k=1}^{n} b_{k,n} X_{k,n}^2 - \left(\sum_{k=1}^{n} b_{k,n} X_{k,n}\right)^2\right]},$$

ahol az $a_{k,n}$ és $b_{k,n}$ együtthatók explicit módon számolhatók. Csörgő S. [20] aszimptotikus eredményének a következményeként kapjuk a V_n tesztstatisztika határeloszlását. Bebizonyítjuk, hogy ha a minta F eloszlásfüggvénye a logisztikus eltolás-skála családhoz tartozik, akkor

$$\begin{aligned} nV_n \xrightarrow{\mathcal{D}} V := & \frac{1}{\pi^2/3-2}\left\{\int_0^1 \frac{6B^2(t)}{t(1-t)}\,\mathrm{d}t - \left[\int_0^1 6B(t)\,\mathrm{d}t\right]^2\right\} \\ & - \left[\frac{1}{\pi^2/3-2}\int_0^1 6B(t)\ln\left(\frac{t}{1-t}\right)\,\mathrm{d}t\right]^2, \end{aligned}$$

ahol határérték 1 valószínűséggel létezik.

Del Barrio, Cuesta-Albertos és Matrán [33]-ben a tesztstatisztika határeloszlását megadták súlyozott Brown-hidak Karhunen–Loève-sorfejtéseként. Ugyanezen technikával meghatározzuk az általunk kapott határeloszlás soros alakját. Bebizonyítjuk, hogy a V határeloszlás felírható

$$V \overset{\mathcal{D}}{=} \frac{1}{\pi^2/3-2}\sum_{k=2}^{\infty}\frac{6}{k(k+1)}Z_k^2 - \left[\frac{1}{\pi^2/3-2}\sum_{l=1}^{\infty}\frac{3\sqrt{4l+1}}{l(l+1)(2l-1)(2l+1)}Z_{2l}\right]^2$$

alakban, ahol $(Z_m)_{m=1}^{\infty}$ független, standard normális eloszlású véletlen változók végtelen sorozata, és a sorok 1 valószínűséggel konvergálnak.

Majd ugyancsak egy szimulációs erővizsgálatot hajtottunk végre, valamint összehasonlítottuk az új teszt erejét az empirikus karakterisztikus függvényre és az empirikus momentum-generáló függvényre alapozott Meintanis-tesztekkel [58].

Summary

In the thesis the results connected with goodness of fit are covered. Let $X_1, \ldots, X_n$ be a sample (independent identically distriduted random variables) from an unknown distribution with distribution function F. The simple hypothesis is

$$\mathcal{H}_0 : F = F_0 \,,$$

where F_0 is a given distribution function, and the composite hypothesis is

$$\mathcal{H}_0 : F \in \mathcal{F} \,,$$

where $\mathcal{F}$ denotes the family of probability distributions.

The thesis is organized as follows. In Chapter 2 we collect the historical preliminaries. In Chapter 3 we suggest a goodness of fit procedure to the uniform distribution on [0,1] and to the uniform family, and we investigate the new tests in a simulation study. In Chapter 4 we demonstrate a simulation study of the goodness of fit test to the normal family, based on the L^2-Wasserstein distance, proposed by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [34]. In Chapter 5 we introduce the weighted version of the quantile correlation test proposed by S. Csörgő [19], [20] for the logistic family, and we present the results of the simulation study connedted with the new test.

Historical preliminaries

For the overview in Section 2.1. for the overview we recall the first tests which are suitable for goodness of fit to a fixed distribution paying special attention to the development of the asymptotic theory of goodness of fit tests. The first goodness of fit procedure is the χ^2-test proposed by Pearson [61]. Under the null-hypothesis, this test has asymptotic distribution χ^2. The EDF-tests and the recovery of their asymptotic distribution have received special attention. These tests use different functional distances to measure the discrepancy between the hypothesized distribution function and the empirical distribution function. Section 2.2. is devoted to the problem of the goodness of fit to the family of distributions and their asymptotic theories. The first studies are occurred in the most interesting case, for the Gaussian family. Then we adapt all the procedures considered in the first subsection for the case of the parametric family. The simple idea is choosing some adequate estimator of the parameter and replacing the fixed distribution by the distribution with the estimated parameter. Finally we recall the regression and correlation tests, the very popular Wilk–Shapiro-test of normality [65], it's further modifications and asymptotic results.

Goodness of fit to the uniform family

The results of Chapter 3 are from Krauczi [59]. We suggest a goodness of fit procedure to the uniform distribution on [0,1] and to the uniform family. The idea is the following: let $U_1, \dots, U_n$ be a random uniform sample (independent uniformly distributed on [0,1] random variables). Moreover, there is a given deterministic distance level $d_n \in (0,1)$ for all n. We push through this distance level on [0,1] and we observe how many nonempty disjoint classes breaks up the elements of the order statistics into. The elements of the order statistics belong to the same class, where the distance between any two neighbouring elements is not greater than d_n. The classes belong to a given sample at a given distance level is called the number of clusters.

In Section 3.1. we recall that Csörgő and Wu showed that the number of clusters is asymptotically normal for three different distance level sequences. We extend the results of Csörgő and Wu [23] to multivariate limit theorems for uniform distributions on different intervals. These theorems are applied for testing uniformity on a known and an unknown interval.

Section 3.2. consists of the theoretical results. We prove that the joint cluster count vector is asymptotically normal in three different cases: the sample comes from the uniform distribution on [0,1], on a known $[a,b]$ and an unknown interval.

Set $J \geq 1$ and let $d_{n1} \leq d_{n2} \leq \ldots \leq d_{nJ},\ n \in \mathbb{N}$, be distance levels. If the sample comes from the uniform distribution on the unit interval [0,1], then $K_{nj}(d_{nj})$ denote the numbers of clusters corresponding to the distance levels d_{nj} for all n and j. Consider the random vector

$$\mathbf{K}_n = \frac{1}{\sqrt{n}} \left(\frac{K_{n1}(d_{n1}) - m_{n1}}{\sigma_{n1}}, ..., \frac{K_{nJ}(d_{nJ}) - m_{nJ}}{\sigma_{nJ}} \right)^\top, \qquad n \in \mathbb{N},$$

with the sequences $m_{nj} = ne^{-nd_{nj}}$ and

$$\sigma_{nj}^2 = e^{-2nd_{nj}}(e^{nd_{nj}} - 1 - n^2 d_{nj}^2), \qquad j = 1, \dots, J.$$

Suppose the distance levels satisfying one of the following conditions:

(T1) $nd_{nj} \to 0$, $n^2 d_{nj} \to \infty$;

(T2) $0 < \liminf_n nd_{nj} \leq \limsup_n nd_{nj} < \infty$;

(T3) $nd_{nj} \to \infty$, $ne^{-nd_{nj}} \to \infty$.

In addition the limits

$$s_{ij} := \lim_{n \to \infty} \frac{e^{-nd_{ni} - nd_{nj}}(e^{nd_{ni}} - 1 - n^2 d_{ni} d_{nj})}{\sigma_{ni}\sigma_{nj}} \in \mathbb{R}, \qquad 1 \leq i < j \leq J,$$

exist, and let be $s_{jj} := 1$ and $s_{ji} := s_{ij}$. Introduce the matrix $\Sigma := (s_{ij})_{i,j=1,\dots,J}$.

Under the above notations and assumptions the convergence

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma)$$

is proved.

One of the corollary of this theorem is that we can obtain the limiting distribution with the diagonal covariance matrix Σ for special distance level sequences. Csörgő and

Wu give well-behaving examples called typical sequences. A typical sequence $(d_n)_{n=1,2...}$ for the case (T1) is $d_n = n^{-\alpha}$ for some $\alpha \in (1,2)$. In the case (T2) the existence of the limit $c := \lim_{n\to\infty} nd_n \in \mathbb{R}$ gives the typical sequence $(d_n)_{n=1,2...}$. A typical sequence $(d_n)_{n=1,2...}$ for the case (T3) is $d_n = \beta(\log n)/n$ for some $\beta \in (0,1)$.

If the sample comes from the uniform distribution on the known interval $[a,b]$ with $a,b \in \mathbb{R},\ a<b$, then we prove with applying a linear transformation of the interval $[a,b]$ onto the interval [0,1], that the transformed cluster count vector is also asymptotically normal distributed under the correctly transformed assumptions.

Finally, the sample comes from the uniform distribution on the unknown interval. Let $V_1,\dots,V_n$ be independent, uniformly distributed random variables on the interval $[a,b]$ with $a<b$ being unknown and let $V_{1,n},\dots,V_{n,n}$ be the ordered sample. The endpoints of the interval are estimated by $\hat{a}_n = V_{1,n}$ and $\hat{b}_n = V_{n,n}$. In an analogue to the previous notations, for given $J \geq 1$ and distance levels $d_{n1} < \dots < d_{nJ}$ set

$$\hat{m}_{nj} = ne^{-\frac{nd_{nj}}{\hat{b}_n - \hat{a}_n}}, \qquad \hat{\sigma}_{nj} = \sqrt{e^{-2\frac{nd_{nj}}{\hat{b}_n-\hat{a}_n}}\left(e^{\frac{nd_{nj}}{\hat{b}_n-\hat{a}_n}} - 1 - \left(\frac{nd_{nj}}{\hat{b}_n-\hat{a}_n}\right)^2\right)}$$

and

$$\widehat{\mathbf{K}}_n = \frac{1}{\sqrt{n}}\left(\frac{\hat{K}_{n1}(d_{n1}) - \hat{m}_{n1}}{\hat{\sigma}_{n1}}, \dots, \frac{\hat{K}_{nJ}(d_{nJ}) - \hat{m}_{nJ}}{\hat{\sigma}_{nJ}}\right)^\top.$$

Under the assumptions as on the interval [0,1] we prove that

$$\widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0,\Sigma).$$

Section 3.3. consists of the statistical results and simulations. It follows from theoretical results that we obtain asymptotically χ^2 test for goodness of fit under the simple and the composite null hypotheses.

First consider the simple null hypothesis asserting that a sample $X_1,\dots,X_n$ has the uniform distribution on [0,1]. Under the simple null hypothesis and the convenient assumptions we get

$$C_n := \mathbf{K}_n^\top \Sigma^{-1} \mathbf{K}_n \xrightarrow{\mathcal{D}} \chi^2_J.$$

Now, consider the composite null hypothesis asserting that a sample comes from the family of all uniform distributions on $\mathbb{R}$. Then under the simple null hypothesis and the convenient assumptions we get

$$\widehat{C}_n := \widehat{\mathbf{K}}_n^\top \Sigma^{-1} \widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \chi^2_J.$$

Accordingly it may seemed, that the composite hypothesis may be tested like the previous paragraph. The problem is that as we don't know the explicit value a and b, so the component of the covariance matrix Σ can't be determined, hence the test statistics $\widehat{C}_n$ can't be counted based on a given sample. Therefore we test the composite null hypothesis with another procedure. Here, we propose a possible solution based on the random transformation of the sample $V_1,\dots,V_n$ coming from an unknown interval into the unit interval as follows:

$$\left(\frac{V_{2,n} - V_{1,n}}{V_{n,n} - V_{1,n}}, \dots, \frac{V_{n-1,n} - V_{1,n}}{V_{n,n} - V_{1,n}}\right).$$

Here $\tilde{K}_{n-2,j}(d_{nj})$ denote the numbers of clusters corresponding to the distance levels d_{nj} for the randomly transformed sample, $j = 1, \ldots, J$, and let

$$\tilde{\mathbf{K}}_{n-2} := \frac{1}{\sqrt{n}} \left(\frac{\tilde{K}_{n-2,1}(d_{n1}) - m_{n-2,1}}{\sigma_{n-2,1}}, \ldots, \frac{\tilde{K}_{n-2,J}(d_{nJ}) - m_{n-2,J}}{\sigma_{n-2,J}} \right)^{\top}$$

be a vector of normalized numbers of clusters of the randomly transformed sample. In addition let $\tilde{\Sigma}$ be the covariance matrix computed using the randomly transformed sample. Then

$$C_n^{\mathrm{mod}} := \tilde{\mathbf{K}}_{n-2}^{\top} \tilde{\Sigma}^{-1} \tilde{\mathbf{K}}_{n-2} \xrightarrow{\mathcal{D}} \chi_J^2.$$

Thus, these tests define asymptotically χ^2 tests for a uniform distribution or for the uniform family.

We simulated powers of the new tests against some continuous alternative distributions on [0,1] and we compared these tests with the data driven smooth test introduced in Inglot and Ledwina [48].

Goodness of fit to the normal family

Chapter 4 is devoted to the paper of Krauczi [52]. In this chapter we perform a simulation study of the goodness of fit test to the normal family based on the L^2-Wasserstein distance, proposed by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [34]. They obtained the location- and scale-free test statistic for the null hypothesis $\mathcal{H}_0 : F \in \in \mathbf{N}$, where $\mathbf{N}$ denotes the normal family. This testing procedure belongs to the class of minimum distance tests (using the distance of quantile functions); on the other hand it is asymptotically equivalent with a correlation test. The name of this test derives from these two different approaches: the quantile correlation test.

To describe their proposal, let $\mathcal{P}_2(\mathbb{R})$ be the set of probabilities on $\mathbb{R}$ with a finite second moment. For probabilities P_1 and P_2 in $\mathcal{P}_2(\mathbb{R})$ the L_2-Wasserstein distance between P_1 and P_2 is

$$\mathcal{W}(P_1, P_2) = \inf \left\{ [E(X_1 - X_2)^2]^{1/2}, \mathcal{L}(X_1) = P_1, \mathcal{L}(X_2) = P_2 \right\},$$

where $\mathcal{L}(X)$ denotes the probability distribution of the random variable X. It can be explicitly obtained in terms of quantile functions:

$$\mathcal{W}(P_1, P_2) = \left[\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 \mathrm{d}t \right]^{1/2},$$

where F_1^{-1} and F_2^{-1} are quantile function associated with the probabilities P_1 and P_2.

If P is a probability distribution in $\mathcal{P}_2(\mathbb{R})$ with distribution function F, mean μ_0 and standard deviation σ_0, then L_2-Wasserstein distance-square between F and the class of all normal laws $\mathbf{N}$ is

$$\mathcal{W}^2(P, \mathbf{N}) := \inf \{ \mathcal{W}^2(P, N_\sigma^\mu), N_\sigma^\mu \in \mathbf{N} \} = \sigma_0^2 - \left(\int_0^1 F^{-1}(t) \Phi^{-1}(t) \mathrm{d}t \right)^2,$$

where Φ^{-1} is the standard normal quantile function. Thus, the law in $\mathbf{N}$ closest to F is given by $\mu = \mu_0$ and $\sigma = \int_0^1 F^{-1}(t)\Phi^{-1}(t)\mathrm{d}t$. The ratio $\mathcal{W}^2(P, \mathbf{N})/\sigma_0^2$ is not affected by

location or scale changes of F. Hence, it can be considered as a measure of dissimilarity between F and $\mathbf{N}$.

Given a random sample $X_1, \ldots, X_n$ from F, now the empirical version of the ratio $\mathcal{W}(P, \mathbf{N})/\sigma_0$ may be obtained. Then the location- and scale-free BCMR-test statistic for the null hypothesis $H_0 : F \in \mathbf{N}$ is

$$T_n := \frac{\mathcal{W}^2(F_n, \mathbf{N})}{S_n^2} = 1 - \frac{\left[\int_0^1 Q_n(t)\Phi^{-1}(t)dt\right]^2}{S_n^2} = 1 - \frac{\left[\sum_{k=1}^n X_{k,n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \Phi^{-1}(t)\,\mathrm{d}t\right]^2}{S_n^2}.$$

Del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [34] investigated the asymptotic distribution of the test statistic under the null-hypothesis. They managed to produce the limit distribution in two different forms. The first form is functionals of the Brownian bridge, the second is a series of random variables. Let φ denote the standard normal density function, and let

$$a_n = \frac{1}{n} \int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{\left[\varphi(\Phi^{-1}(t))\right]^2} dt.$$

If $F \in \mathbf{N}$, then

$$\begin{aligned} n(T_n - a_n) \xrightarrow{\mathcal{D}} & \int_0^1 \frac{B^2(t) - E(B^2(t))}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t - \left[\int_0^1 \frac{B(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 - \left[\int_0^1 \frac{B(t)\Phi^{-1}(t)}{\varphi^2(\Phi^{-1}(t))}\mathrm{d}t\right]^2 \\ & \stackrel{\mathcal{D}}{=} -\frac{3}{2} + \sum_{j=3}^{\infty} \frac{Z_j^2 - 1}{j}, \end{aligned}$$

where $(Z_j)_{j=3}^{\infty}$ is a sequence of independent standard normal random variables.

A simulation study was performed to evaluate the power of the BCMR-test and to make comparisons with other tests of normality. Since under the null hypothesis the asymptotic distribution for Wilk–Shapiro-test is the same as for the BCMR-test, thus the result of the power study isn't surprising.

Goodness of fit to the logistic family

The results of Chapter 5 are from Balogh and Krauczi [6]. In this chapter we present the weighted version of the quantile correlation test statistics for goodness of fit to the logistic family, introduced by del Barrio, Cuesta-Albertos, Matrán and Rodríguez-Rodríguez [34], and del Barrio, Cuesta-Albertos and Matrán [33]. The use of weight functions in the test statistics were suggested independently from each other by de Wet in [28] and [29] and by S. Csörgő in [19] and [20]. It is an interesting fact that there the authors' motivations were considerably different. S. Csörgő showed that the suitably weighted versions of the correlation tests have limiting distribution for more family of probability distributions; de Wet expected „the loss of degrees of freedom" in the limiting null distribution (in the case of the normal family this means that the first two terms are missing in the infinite series representation of the asymptotic distribution). We prove the results of S. Csörgő [20] for

location and scale logistic family with the weight function for location family suggested by de Wet.

For a given distribution function $G(x)$, $x \in \mathbb{R}$, and for $\theta \in \mathbb{R}$ and $\sigma > 0$, let $G_\sigma^\theta(x) = G((x-\theta)/\sigma)$, $x \in \mathbb{R}$, and consider the location-scale family

$$\mathcal{G}_{l,s} = \{G_\sigma^\theta : \theta \in \mathbb{R}, \sigma > 0\}.$$

Denote by $Q_G(t) = G^{-1}(t), 0 < t < 1$, the quantile function of G. Consider a weight function $w : (0,1) \to [0,\infty)$ satisfying $\int_0^1 w(t)\,\mathrm{d}t = 1$, and define the weighted r-th moment

$$\mu_r(G,w) := \int_0^1 (Q_G(t))^r w(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x^r w(G(x))\mathrm{d}G(x).$$

Assume that $\mu_1(G,w)$ and $\mu_2(G,w)$ are finite, and define also the weighted variance:

$$\nu(G,w) := \mu_2(G,w) - \mu_1^2(G,w) \geq 0.$$

The weighted L_2-Wasserstein distance with weight function w of two distributions F and G can be defined as

$$\mathcal{W}_w(F,G) := \left[\int_0^1 \left(Q_F(t) - Q_G(t)\right)^2 w(t)\,\mathrm{d}t\right]^{\frac{1}{2}}.$$

Therefore the weighted L_2-Wasserstein distance $\mathcal{W}_w(F,\mathcal{G}_{l,s}) = \inf\{\mathcal{W}_w(F,G) : G \in \mathcal{G}_{l,s}\}$ between F and location-scale family $\mathcal{G}_{l,s}$, scaled to F is

$$\frac{\mathcal{W}_w^2(F,\mathcal{G}_{l,s})}{\nu(F,w)} = 1 - \frac{\left[\int_0^1 Q_F(t)Q_G(t)w(t)\mathrm{d}t - \mu_1(F,w)\mu_1(G,w)\right]^2}{\nu(F,w)\nu(G,w)},$$

as derived in [20].

Consider a random sample $X_1, \ldots, X_n$ with common distribution function F, and let a fixed distribution function G. We would like to test the null hypothesis $\mathcal{H}_0 : F \in \mathcal{G}_{l,s}$. Letting Q_n be the sample quantile function, in order to define the following test statistics

$$\begin{aligned} V_n &:= 1 - \frac{\left[\int_0^1 Q_n(t)Q_G(t)w(t)\mathrm{d}t - \mu_1(G,w)\int_0^1 Q_n(t)w(t)\mathrm{d}t\right]^2}{\nu(G,w)\left[\int_0^1 Q_n^2(t)w(t)\mathrm{d}t - \left(\int_0^1 Q_n(t)w(t)\mathrm{d}t\right)^2\right]} \\ &= 1 - \frac{\left[\sum_{k=1}^n X_{k,n}\left\{\int_{\frac{k-1}{n}}^{\frac{k}{n}} Q_G(t)w(t)\mathrm{d}t - \mu_1(G,w)\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right\}\right]^2}{\nu(G,w)\left[\sum_{k=1}^n X_{k,n}^2 \int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t - \left(\sum_{k=1}^n X_{k,n}\int_{\frac{k-1}{n}}^{\frac{k}{n}} w(t)\mathrm{d}t\right)^2\right]}, \end{aligned}$$

derived from the weighted L^2-Wasserstein distance between the empirical distribution of the sample and the location-scale family $\mathcal{G}_{l,s}$.

For the logistic location family $\mathcal{G}_l$ de Wet suggested in [29] the use of the weight function

$$w(t) = 6t(1-t), \qquad 0 < t < 1.$$

The above introduced location-scale-free test statistic specializes to

$$V_n = 1 - \frac{\left[\sum_{k=1}^{n} a_{k,n} X_{k,n}\right]^2}{\left(\frac{\pi^2}{3} - 2\right)\left[\sum_{k=1}^{n} b_{k,n} X_{k,n}^2 - \left(\sum_{k=1}^{n} b_{k,n} X_{k,n}\right)^2\right]},$$

where the coefficients $a_{k,n}$ and $b_{k,n}$ are given explicitly. We obtain the following limit distribution of the test statistics V_n as a consequence to the asymptotic result by Csörgő S. [20]. We prove that if the sample comes from the logistic location-scale family, then

$$nV_n \xrightarrow{\mathcal{D}} V := \frac{1}{\pi^2/3-2}\left\{\int_0^1 \frac{6B^2(t)}{t(1-t)}\,\mathrm{d}t - \left[\int_0^1 6B(t)\,\mathrm{d}t\right]^2\right\} - \left[\frac{1}{\pi^2/3-2}\int_0^1 6B(t)\ln\left(\frac{t}{1-t}\right)\,\mathrm{d}t\right]^2,$$

where the integrals exists with probability 1.

Del Barrio, Cuesta-Albertos and Matrán [33] obtained the asymptotic distribution as the Karhunen–Loève expansion of the weighted Brownian-bridge. With the same technique we determine the infinite series representation of our limiting distribution. The limiting distribution V can be represented alternatively as

$$V \overset{\mathcal{D}}{=} \frac{1}{\pi^2/3-2}\sum_{k=2}^{\infty}\frac{6}{k(k+1)}Z_k^2 - \left[\frac{1}{\pi^2/3-2}\sum_{l=1}^{\infty}\frac{3\sqrt{4l+1}}{l(l+1)(2l-1)(2l+1)}Z_{2l}\right]^2,$$

where $(Z_m)_{m=1}^{\infty}$is an infinite sequence of independent identically distributed standard normal random variables, the series converge with probability 1.

Similarly to previous results a simulation study was performed to evaluate the power of the tests. We compare the new test with the Meintanis-tests based on the empirical characteristic function and the empirical momentum generating function from [58].

Köszönetnyilvánítás

Szeretnék köszönetet mondani témavazetőmnek, Csörgő Sándornak, hogy kiváló előadásaival megszeretette velem a valószínűségszámítást. Hálás vagyok Neki azért a rengeteg emberségért, amit kaptam Tőle.

Köszönettel tartozok mostani témavazetőmnek, Pap Gyulának, akitől bátorítást, támogatást kaptam ahhoz, hogy Csörgő Tanár Úr halála után befejezzem az elkezdett munkát.

Köszönöm Szűcs Gábornak a disszertáció megírásához nyújtott hatalmas segítségét.

Irodalomjegyzék

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] N. Aguirre and M. Nikulin. Goodness-of-fit test for the family of logistic distributions. *Qüestiió. Quaderns d'Estadística i Investigació Operativa. Segona Època*, 18(3):317–335, 1994.

[3] M. M. Ali. Stochastic ordering and kurtosis measure. *Journal of the American Statistical Association*, 69:543–545, 1974.

[4] T. W. Anderson and D. A. Darling. Asymptotic theory of certain „goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

[5] N. Balakrishnan. *Handbook of the logistic distribution.* Dekker, New York, 1992. Statist. Textbooks Monogr. 123.

[6] F. Balogh and É. Krauczi. Weighted quantile corelation test for the logistic family. *Acta Scientiarum Mathematicarum.(Szeged)*, 80(1-2):307–326, 2014.

[7] P. J. Bickel and D. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9:1196–1217, 1981.

[8] P. Billingsley. Convergence of probability measures. New York-London-Sydney-Toronto: John Wiley and Sons, Inc. XII, 1968.

[9] A. Bowman and P. Foster. Adaptive smoothing and density-based tests of multivariate normality. *JASA. Journal of the American Statistical Association*, 88:529–537, 1993.

[10] M. Burke, M. Csörgő, S. Csörgő, and P. Révész. Approximations of the empirical process when parameters are estimated. *The Annals of Probability*, 7(5):790–810, 1979.

[11] A. Cabaña and E. M. Cabaña. Tests of normality based on transformed empirical processes. *Methodology and Computing in Applied Probability*, 5(3):309–335, 2003.

[12] H. Chernoff and E. Lehmann. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, 25:579–586, 1954.

[13] W. Cochran. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23:315–345, 1952.

[14] H. Cramér. On the composition of elementary errors. I. Mathematical deductions. II. Statistical applications. *Skandinavisk Aktuarietidskrift*, 11:13–74, 141–180, 1928.

[15] S. Csörgő. Limit behaviour of the empirical characteristic function. *The Annals of Probability*, 9:130–144, 1981.

[16] S. Csörgő. The empirical characteristic process when parameters are estimated. Contributions to probability, Collect. pap. dedic. E. Lukacs, 215-230, 1981.

[17] S. Csörgő. Testing for normality in arbitrary dimension. *The Annals of Statistics*, 14:708–723, 1986.

[18] S. Csörgő. Consistency of some tests for multivariate normality. *Metrika*, 36:107–116, 1989.

[19] S. Csörgő. Weighted correlation tests for scale families. *Test*, 11(1):219–248, 2002.

[20] S. Csörgő. Weighted correlation tests for location-scale families. *Mathematical and Computer Modelling*, 38(7-9):753–762, 2003. Hungarian applied mathematics and computer applications.

[21] S. Csörgő and T. Szabó. Weighted correlation tests for gamma and lognormal families. *Tatra Mountains Mathematical Publications*, 26(part II):337–356, 2003. Probastat '02. Part II.

[22] S. Csörgő and T. Szabó. Weighted quantile correlation tests for Gumbel, Weibull and Pareto families. *Probability and Mathematical Statistics*, 29(2):227–250, 2009.

[23] S. Csörgő and W. B. Wu. On the clustering of independent uniform random variables. *Random Structures Algorithms*, 25(4):396–420, 2004.

[24] R. B. D'Agostino. An omnibus test of normality for moderate and large sample sizes. *Biometrika*, 58:341–348, 1971.

[25] D. Darling. The Cramér–Smirnov test in the parametric case. *Annals of Mathematical Statistics*, 26:1–20, 1955.

[26] F. David and N. Johnson. The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35:182–190, 1948.

[27] H. David, H. Hartley, and E. Pearson. The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika*, 41:482–493, 1954.

[28] T. de Wet. Discussion of "Contributions of empirical and quantile processes to the asymptotictheory of goodness-of-fit tests". *Test*, 9(1):74–79, 2000.

[29] T. de Wet. Goodness-of-fit tests for location and scale families based on a weighted L_2-Wasserstein distance measure. *Test*, 11(1):89–107, 2002.

[30] T. de Wet and J. Venter. Asymptotic distributions of certain test criteria of normality. *South African Statistical Journal*, 6:135–149, 1972.

[31] T. de Wet and J. Venter. A goodness of fit test for a scale parameter family of distributions. *South African Statistical Journal*, 7:35–46, 1973.

[32] P. Deheuvels and G. Martynov. Karhunen–Loève expansions for weighted Wiener processes and Brownian bridges via Bessel functions. In *High dimensional probability, III (Sandjberg, 2002)*, volume 55 of *Progr. Probab.*, pages 57–93. Birkhäuser, Basel, 2003.

[33] E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, 9(1):1–96, 2000. With discussion.

[34] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[35] M. D. Donsker. An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*, 6:12, 1951.

[36] J. L. Doob. Heuristic approach to the Kolmogorov–Smirnov theorems. *Annals of Mathematical Statistics*, 20:393–403, 1949.

[37] J. Durbin. Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, 1:279–290, 1973.

[38] T. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70:723–726, 1983.

[39] W. Feller. On the Kolmogorov–Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics*, 19:177–189, 1948.

[40] A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *The Annals of Statistics*, 5:88–97, 1977.

[41] R. A. Fisher. The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London. Series A*, 130:16–28, 1930.

[42] F. Gan and K. Koehler. Goodness of fit tests based on P-P probability plots. *Technometics*, 32:289–303, 1990.

[43] R. Geary. Testing for normality. *Biometrika*, 34:209–242, 1947.

[44] E. Godehardt and J. Jaworski. On the connectivity of a random interval graph. *Random Structures Algorithms*, 9:137–161, 1996.

[45] E. J. Gumbel. On the reliability of the classical chi-square test. *Annals of Mathematical Statistics*, 14:253–263, 1943.

[46] E. J. Gumbel. Ranges and midranges. *Annals of Mathematical Statistics*, 15:414–422, 1944.

[47] P. Hall and A. H. Welsh. A test for normality based on the empirical characteristic function. *Biometrika*, 70:485–489, 1983.

[48] T. Inglot and T. Ledwina. Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Applications*, 417(1):124–133, 2006.

[49] M. Kac, J. Kiefer, and J. Wolfowitz. On tests of normality and other tests of goodness of fit based on distance methods. *Annals of Mathematical Statistics*, 26:189–211, 1955.

[50] W. Kallenberg and T. Ledwina. Data driven smooth tests for composite hypotheses: comparison of powers. *Journal of Statistical Computation and Simulation*, 59:101–121, 1997.

[51] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale del Istituto Italiano degli Attuari*, 4:83–91, 1933.

[52] É. Krauczi. A study of the quantile correlation test of normality. *Test*, 18(1):156–165, 2009.

[53] V. LaRiccia and D. M. Mason. Cramér–von Mises statistics based on the sample quantile function and estimated parameters. *Journal of Multivariate Analysis*, 18:93–106, 1986.

[54] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, 1998.

[55] J. Leslie, M. Stephens, and S. Fotopoulos. Asymptotic distribution of the Shapiro–Wilk W for testing for normality. *The Annals of Statistics*, 14:1497–1506, 1986.

[56] H. W. Lilliefors. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

[57] H. Mann and A. Wald. On the choice of the number of class intervals in the application of the chi square test. *Annals of Mathematical Statistics*, 13:306–317, 1942.

[58] S. G. Meintanis. Goodness-of-fit tests for the logistic distribution based on empirical transforms. *Sankhyā. The Indian Journal of Statistics*, 66(2):306–326, 2004.

[59] K. É. Oszténiyiné. Joint cluster counts from uniform distribution. *Probability and Mathematical Statistics*, 33(1):93–106, 2013.

[60] E. Pearson, R. D'Agostino, and K. Bowman. Tests for departure from normality: Comparison of powers. *Biometrika*, 64:231–246, 1977.

[61] E. S. Pearson. A further development of tests for normality. *Biometrika*, 22:239–249, 1930.

[62] D. Pollard. The minimum distance method of testing. *Metrika*, 27:43–70, 1980.

[63] M. W. Shapiro, S.S. and H. Chen. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 63:1343–72, 1968.

[64] S. Shapiro and R. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67:215–216, 1972.

[65] S. Shapiro and M. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[66] G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics.* Wiley, New York, 1986.

[67] N. Smirnov. Sur la distribution de ω^2 (Critérium de M.R. von Mises). *Comptes Rendus de l'Académie des Sciences Paris*, 202:449–452, 1936.

[68] N. Smirnov. Sur la distribution de ω^2 (Critérium de M.R. von Mises). *Matematicheskij Sbornik*, 2:973–993, 1937.

[69] N. Smirnov. Sur les écarts de la courbe de distribution empirique. *Matematicheskij Sbornik*, 6:3–26, 1939.

[70] N. Smirnov. Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, 10:179–206, 1941.

[71] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.

[72] S. Sukhatme. Fredholm determinant of a positive definite kernel of a special type and its application. *Annals of Mathematical Statistics*, 43:1914–1926, 1972.

[73] P.-F. Verhulst. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathèmatique et physique*, 10:113–121, 1838.

[74] S. Verrill and R. Johnson. The asymptotic equivalence of some modified Shapiro–Wilk statistics — complete and censored sample cases. *Annals of Statstics*, 15:413–419, 1987.

[75] R. von Mises. *Wahrscheinlichkeitsrechnung.* Wein, Leipzig, 1931.

[76] G. Watson. The χ^2 goodness-of-fit test for normal distributions. *Biometrika*, 44:336–348, 1957.

[77] G. Watson. On chi-square goodness-of-fit tests for continuous distributions. *Journal of the Royal Statistical Society. Series B*, 20:44–72, 1958.

[78] S. Weisberg and C. Bingham. An approximate analysis of variance test fornon-normality suitable for machine calculation. *Technometrics*, 17:133–134, 1975.

[79] P. Williams. Note on the sampling distribution of $\sqrt{\beta_1}$ where the population is normal. *Biometrika*, 27:269–271, 1935.