Eloszláscsaládokhoz való illeszkedés vizsgálata

Ph. D. értekezés tézisfüzete

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

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

1. 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 2. fejezetben a disszertáció szempontjából fontos történeti előzményeket gyűjtöttük össze. Felidézzük az első módszereket, amelyekkel rögzített eloszláshoz való illeszkedést lehet tesztelni, valamint azt is, hogy hogyan találták meg ezen tesztstatisztikák határeloszlásait. Majd a számunkra érdekes első összetett illeszkedésvizsgálati módszereket és határeloszlásukat elevenítjük fel. 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.

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 [6] 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. Aszimptotikus χ^2-tesztet kapunk egyszerű, illetve összetett nullhipotézis ellenőrzésére. Elvégeztük az új tesztek szimulációs vizsgálatát.

A 4. fejezetben az L^2-Wasserstein távolságot használó del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [10] által bevezetett normalitás teszt szimulációs vizsgálatát mutatjuk be. Egy eltolás- és skálamentes tesztstatisztikát kaptak a $\mathcal{H}_0 : F \in \mathbf{N}$ nullhipotézis ellenőrzésére, ahol $\mathbf{N}$ a normális eloszláscsaládot jelöli. Ennek a normalitástesztnek számos alternatívával szembeni erővizsgálatát végeztük el szimuláció segítségével, valamint összehasonlítottuk más normalitástesztek viselkedésével.

Az 5. fejezetben Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [10], valamint del Barrio, Cuesta-Albertos és Matrán [9] által bevezetett kvantilis korreláció teszt súlyozott változatát vezetjük be logisztikus eloszláscsalád esetében. A súlyfüggvény használatát a tesztstatisztikában egymástól függetlenül de Wet [7, 8] és Csörgő S. [4, 5] különböző motivációból javasolta. Mi a Csörgő-féle [5] eredményt a de Wet által, eltolás eloszláscsalád esetére javasolt konkrét súlyfüggvénnyel bizonyítjuk logisztikus eltolás-skála

eloszláscsalád esetében. Del Barrio, Cuesta-Albertos és Matrán [9] a tesztstatisztika határeloszlását megadták súlyozott Brown-hidak Karhunen–Loève-sorfejtéseként. Ugyanezen technikával meghatározzuk az általunk kapott határeloszlás soros alakját. Majd bemutatjuk az új teszttel kapcsolatos szimulációs vizsgálat eredményét.

A disszertáció három cikk eredményeit tartalmazza. Az Osztényiné Krauczi [16] tartalmazza az illeszkedésvizsgálat eredményeit egyenletes eloszlás esetében. A szimulációs vizsgálat eredményei a normális eloszláscsalád esetében a Krauczi [14] cikkben találhatók. A Balogh, Krauczi [2] tartalmazza a logisztikus eloszláscsalád esetében kapott súlyozott kvantilis korreláció teszt bevezetését, aszimptotikus és szimulációs vizsgálatát.

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

2. Történeti előzmények

Az első alfejezetben felidézzük az első teszteket, amelyekkel rögzített eloszláshoz való illeszkedést lehet ellenőrizni valamint, hogy hogyan találták meg ezen tesztstatisztikák határeloszlását. Az első illeszkedésvizsgálatra használt eljárás a Pearson-féle χ^2-teszt [17], amely aszimptotikusan χ^2 eloszlású megfelelő szabadsági fokkal 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 eloszláscsaládhoz 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 Wilk–Shapiro [19] normalistásztesztjét, ennek további változatait, valamint, hogy hogyan sikerült meghatározni a határeloszlását.

3. Illeszkedésvizsgálat egyenletes eloszlás esetében

Bevezetés és előzmények

Ebben a fejezetben egy eljárást vezetünk be egyenletesség tesztelésére klaszterszámok segítségével. 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.

Csörgő S. és Wu [6] három különböző aszimptotikus viselkedésű távolságszint sorozat mellett bizonyították a klaszterek számának aszimptotikus normalitását, és még rátát is adtak az eloszlásfüggvények konvergenciájának sebességére.

1. Tétel (Csörgő és Wu [6]). *(i) Ha* $nd_n \to 0$ *és* $n^2 d_n \to \infty$, *akkor*

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

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*

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

Ennek a tételnek bizonyítjuk a többdimenziós változatait különböző intervallumon egyenletes eloszlások esetében, majd használjuk egyenletesség tesztelésére ismert és ismeretlen intervallumon.

Elméleti eredmények

Megvizsgáltuk 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. A $K_{nj}(d_{nj})$ jelölje a d_{nj} távolságszinthez tartozó klaszterek számát minden n és j esetén. Tekintsünk a

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

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

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

centralizáló és normalizáló sorozattal. A következő határeloszlástételt kapjuk az (1) vektor viselkedésére.

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

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

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

$$\mathbf{K}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma), \tag{3}$$

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

Egy következménye ennek a tételnek, ha a távolságszintek típusai szerint sorbarendezük a klaszterszámokat, valamint ha a különböző típusokhoz tartozó távolságszintek jól viselkednek, akkor blokkdiagonális kovarianciamátrixú határeloszlását kapjuk a normált klaszterszám vektornak.

2.1. Következmény. *Speciálisan tegyük fel, hogy $J \geq 2$ és $0 \leq J_1 \leq J_2 \leq J$ olyanok, hogy minden $j \leq J_1$ 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 $i < j \leq J_1$ esetén $s_{ij} := \lim_{n\to\infty} \sqrt{d_{ni}/d_{nj}} \in \mathbb{R}$ létezik.

(ii) Minden $J_1 < j \leq J_2$ esetén $c_j := \lim_{n\to\infty} nd_{nj} \in \mathbb{R}$ szintén létezik. Ekkor $J_1 < i < j \leq J_2$ 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 $s_{ij} := \lim_{n\to\infty} e^{-n(d_{nj} - d_{ni})/2} \in \mathbb{R}$ is létezik.

Legyen továbbá $s_{ji} := s_{ij}$ és $s_{jj} := 1$. Ekkor a (3) 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.*

Csörgő S. és Wu [6] mutat jól viselkedő távolságszint sorozatokat mindhárom típushoz. 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. Hasonlóan egy tipikus $(d_n)_{n=1,2,\ldots}$ 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 $\beta_{J_2+1} < \beta_{J_2+2} < \cdots < \beta_J$ paraméterválasztással szintén a $s_{ij} = 0$ értékeket eredményezik minden $J_2 < i < j < J$ esetén. Végül, legyen $0 \leq J_2 - J_1 \leq 2$. 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. A $J_2 - J_1 = 2$ esetben pedig a $c_{J_2} = (e^{c_{J_1+1}} - 1)/c_{J_1+1}$ összefüggés teljesül. Azáltal, hogy a sorozatokban lévő paramétereket a fenti módon választjuk, diagonális kovarianciamátrixot kapunk. Így ezekkel a sorozatokkal a 2.1. Következmény a következő alakot ölti.

2.2. 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.*

Vizsgáltuk továbbá ismert $[a,b]$ intervallumon egyenletes eloszlású véletlen változók esetén adott távolságszintekhez tartozó klaszterszámok együttes viselkedését. Ebben az esetben is meg tudunk adni a 2. Tétel megfelelőjét, mivel $[a,b]$ intervallumon egyenletes eloszlású minta könnyen [0,1] intervallumon egyenletesé transzformálható.

Legyenek $V_1, V_2, \ldots, V_n$ 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 $V_1, V_2, \ldots, V_n$ 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. Legyen $J \geq 1$ természetes szám, és legyenek $d_{n1} \leq d_{n2} \leq \ldots \leq d_{nJ}$ távolságszint sorozatok. A $K_{nj}^{a,b}(d_{nj})$ jelöli a megfelelő d_{nj} távolságszinthez tartozó klaszterszámot, $j = 1, \ldots, J$. Legyenek

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

valamint

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

Ekkor a következő eredményt bizonyítottuk.

3. 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*

$$e^{-\frac{nd_{ni}}{b-a}-\frac{nd_{nj}}{b-a}}\left(e^{\frac{nd_{ni}}{b-a}}-1-\frac{nd_{ni}}{b-a}\frac{nd_{nj}}{b-a}\right)/\sigma_{ni}^{a,b}\sigma_{nj}^{a,b} \to s_{ij}, \qquad 1 \le i < j \le J, \tag{4}$$

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

$$\mathbf{K}_n^{a,b} \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma) \tag{5}$$

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

Végül pedig legyenek $V_1, \dots, V_n$ független, ismeretlen $[a, b]$ intervallumon egyenletes eloszlású véletlen változók, $a, b \in \mathbb{R}$, $a < b$, valamint legyen $V_{1,n}, \dots, V_{n,n}$ a hozzá tartozó rendezett minta. Ebben az esetben is megkaptuk a 2. és a 3. Tételek megfelelőit azáltal, hogy az intervallum végpontjait becsüljük az $\hat{a}_n$ legkisebb és $\hat{b}_n$ legnagyobb mintaelemmel.

Hasonlóan az eddigi jelölésekhez, adott $J \ge 1$ természetes szám és adott $d_{n1} < \dots < d_{nJ}$ távolságszintek esetén $\hat{K}_{nj}(d_{nj})$ jelöli 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

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

4. Tétel. *Tegyük fel, hogy teljesülnek a 3. Tétel feltételei, és tekintsük az ott definiált* Σ *kovarianciamátrixot. Ekkor*

$$\widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \mathcal{N}_J(0, \Sigma). \tag{6}$$

Statisztikai eredmények

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.

Tetszőleges $J \ge 1$ esetén legyenek a $d_{n1} \le \dots \le d_{nJ}$, $n \in \mathbb{N}$, távolságszint sorozatok olyanok, hogy mindegyik sorozat kielégíti a *(T1), (T2)* vagy *(T3)* feltételek valamelyikét. Továbbá tegyük fel, hogy a (2) feltétel teljesül, és a 2. Tételbeli Σ kovarianciamátrix nem szinguláris. Legyen $\mathbf{K}_n$ az (1)-ben definiált vektor. Ekkor a (3) konvergenciából a nullhipotézis mellett következik, hogy a tesztstatisztika

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

ahol χ^2_J a J szabadsági fokú khi-négyzet eloszlás. Így a C_n próbastatisztikával tesztelhetjük a $\mathcal{H}_0$ nullhipotézist. Ezt a tesztet nevezzünk *klaszterteszt*nek.

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. Legyenek $d_{n1} \leq \ldots \leq d_{nJ}$, $n \in \mathbb{N}$, távolságszint sorozatok olyanok, melyek kielégítik a 4. Tétel feltételeit. Ekkor teljesül

$$\widehat{C}_n := \widehat{\mathbf{K}}_n^\top \Sigma^{-1} \widehat{\mathbf{K}}_n \xrightarrow{\mathcal{D}} \chi^2_J. \tag{7}$$

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 a tetszőleges intervallumból származó $V_1, \ldots, V_n$ mintát a [0,1] intervallumba transzformáljuk a következőképpen:

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

(A fenti formulában $V_{1,n}, \ldots, V_{n,n}$ a $V_1, \ldots, V_n$ mintaelemekhez tartozó rendezett mintát jelöli.) 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. Ezt nevezzünk *módosított klaszterteszt*nek.

Meghatároztuk ezen tesztek erejét különböző [0,1] intervalumon folytonos alternatívákkal szemben szimulációval, valamint összehasonlítottuk az új tesztek erejét Inglot és Ledwina [12] által bevezetett data driven smooth teszttel. Az erővizsgálat konkluziója, hogy a klaszter tesztek rosszabbul viselkednek, mint más egyenletesség tesztek, kivéve a nagyon oszcilláló alternatívák esetében. A pontos eredmények táblázatok és ábrák formájában találhatók a disszertációban.

4. Illeszkedésvizsgálat normális eloszláscsalád esetén

Bevezetés és előzmények

Az L^2-Wasserstein távolságot használó del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [10] által bevezetett normalitás teszt szimulációs vizsgálatát mutatjuk be.

Egy eltolás- és skálamentes tesztstatisztikát kaptak a $\mathcal{H}_0 : F \in \mathbf{N}$ nullhipotézis ellenőrzésére, ahol $\mathbf{N}$ a normális eloszláscsaládot jelöli. Ez az eljárás egyrészt úgy tesztel normális eloszláscsaládhoz való tartozást, hogy a teststatisztika a minta empirikus kvantilisfüggvényének egy funkcionálja; másrészt aszimptotikusan ekvivalens egy korrelációteszttel. A két különböző megközelítésből származik az elnevezése: kvantilis korrelációteszt.

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 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 $X_1, \ldots, X_n$ 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:

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

ahol S_n^2 az empirikus szórásnégyzet.

Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [10] megvizsgálták a tesztstatisztika nullhipotézis melletti aszimptotikus viselkedését. Két alakban sikerült előállítaniuk a határeloszlást. Az első Brown-híd funkcionáljaként, a második véletlen változók végtelen soraként. Jelölje φ a standard normális eloszlás sűrűségfüggvényét, és legyen

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

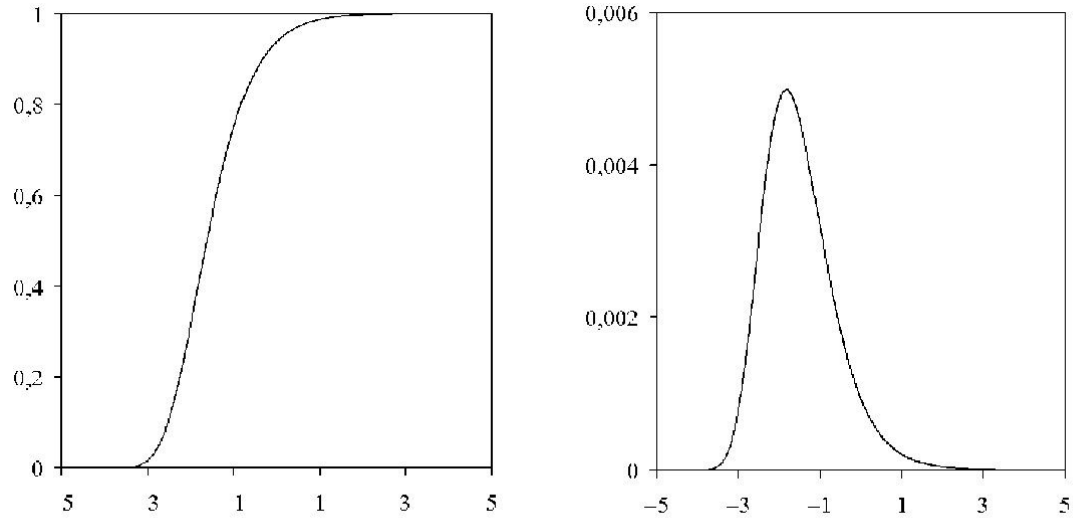
5. Tétel (del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez[10]). *Ha $F \in \mathbf{N}$, akkor*

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

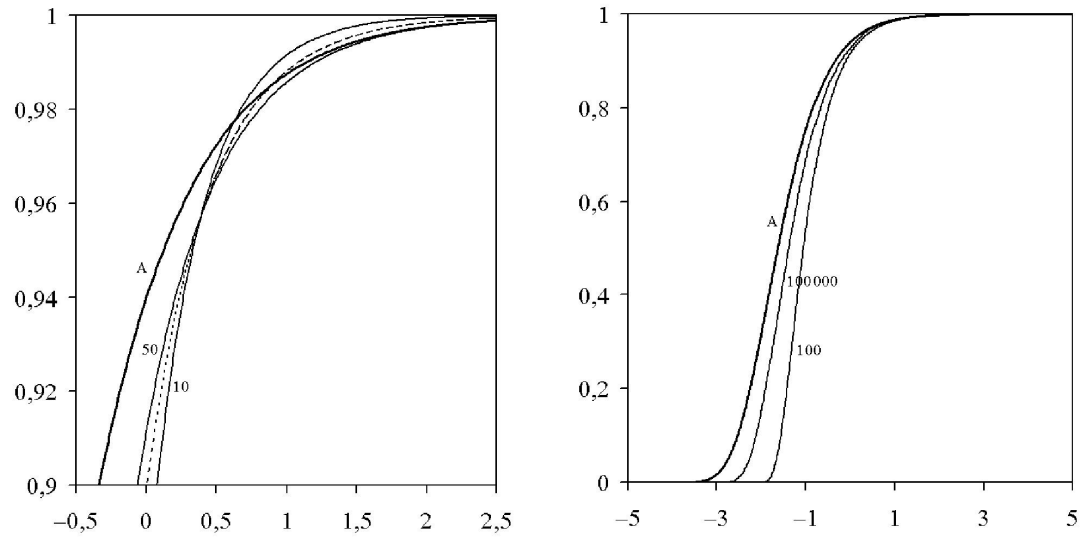
ahol $(Z_j)_{j=3}^\infty$ független, standard normális eloszlású véletlen változók sorozata.

Szimulációs eredmények

A szimulációs vizsgálatban először az 5. Tételben definiált határ véletlen változó eloszlását határoztuk meg numerikusan az ott megadott végtelen soros alak segítségével. Eztán $n = 10$-től $n = 100\,000$-ig többféle mintaméret mellett Monte Carlo szimuláció alkalmazásával meghatároztuk az $n(T_n - a_n)$ tesztstatisztika eloszlásfüggvényét, és megfigyeltük a konvergencia sebességét. A vizsgálat eredményeit a 1. és a 2. ábra tartalmazza.

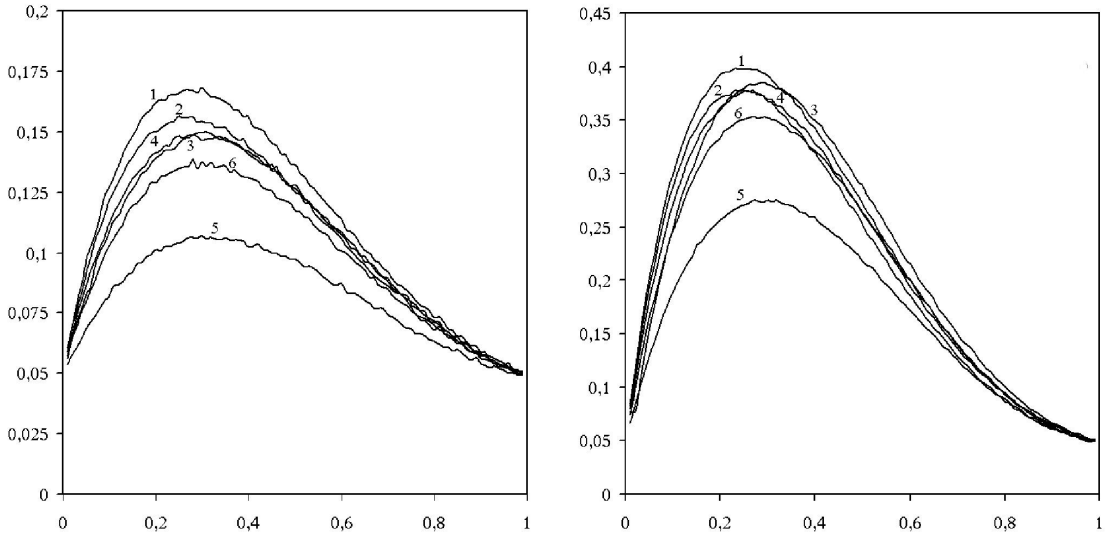


1. ábra. Az aszimptotikus eloszlásfüggvény (balra) és a sűrűségfüggvény (jobbra)



2. ábra. Az $n(T_n - a_n)$ 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).

Továbbá a szimulációs vizsgálatban kiértékeltük a BCMR-teszt (a szerzők kezdőbetűiből) számos alternatívával szembeni erejét és öt másik normalitás teszttel összehason-



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

lítottunk. Ezen tesztek közül az első Shapiro–Wilk W-tesztje [19], amit $n = 20$ és $n = 50$ esetén használtuk az összehasonlításban. Mivel a W-teszt együtthatói az $n = 100$ mintaméret esetén nagyon nehezen számolhatók, ezért ebben az esetben a Shapiro–Francia [18] W'-tesztet használtuk. Az EDF-tesztek közül a Kolmogorov–Smirnov [13] D-teszt Stephens [20] által javasolt módosított változatát, és az Anderson–Darling [1] A^2-tesztet választottuk. A negyedik teszt, amit bevettünk az összehasonlításba, egy sűrűségbecslésre alapozott teszt, Bowman és Foster [3] integrált négyzetes hiba ISE-tesztje. Az ötödik teszt Epps és Pulley [11] BHEP-tesztje. A jobb összehasonlíthatóság céljából a 3. ábrán felvettük a hat tesztnek a kontaminált normális alternatívákkal szembeni erejét a λ paraméter függvényében. Jelölje $\mathrm{CN}(\lambda, \sigma^2)$, $0 < \lambda < 1$ és $\sigma > 0$ paraméterekkel a kontaminált normális eloszlást, amely a következő eloszlásfüggvénnyel van definiálva

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

A szignifikanciaszint 0,05; a mintaméret $n = 20$ mindkét esetben.

Általános konkluziója ennek a vizsgálatnak, 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 nagyon hasonlítanak egymáshoz.

5. Illeszkedésvizsgálat logisztikus eloszláscsalád esetén

Del Barrio, Cuesta-Albertos, Matrán és Rodríguez-Rodríguez [10], valamint del Barrio, Cuesta-Albertos és Matrán [9] által bevezetett kvantilis korreláció teszt súlyozott

változatát vezetjük be logisztikus eloszláscsalád esetében. A súlyfüggvény használatát a tesztstatisztikában egymástól függetlenül de Wet [7], [8] és Csörgő S. [4], [5] különböző motivációból javasolta. Mi a Csörgő-féle [5] 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

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

mennyiséggel. A $\mathcal{W}_w(F,\mathcal{G}_{l,s})=\inf\{\mathcal{W}_w(F,G):G\in\mathcal{G}_{l,s}\}$ az F eloszlás és a $\mathcal{G}_{l,s}$ eltolás-skála család közötti súlyozott L_2-Wasserstein távolságot és a súlyozott variancia hányadosát

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

Csörgő S. [5] cikkéből származtatjuk.

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. Csörgőtől [5] származik a következő eredmény a V_n statisztika aszimptotikus viselkedéséről.

6. Tétel (Csörgő [5]). *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*

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

és az

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

feltételek teljesülnek, akkor a következő állítás érvényes:

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

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

Ennek a tételnek a segítségével találtuk meg a tesztstatisztika határeloszlását logisztikus eloszláscsalád esetében.

Eredmények

Tekintsünk a $G(x)=1/(1+e^{-x}),\ x\in\mathbb{R}$, logisztikus eloszlásfüggvényt, és jelölje $\mathcal{G}_{l,s}$ a kapcsolatos eltolás-skála családot. Direkt számolással megmutatható, hogy a de Wet [8] által eltolás család esetére javasol $w(t)=6t(1-t),\ 0<t<1$ súlyfüggvénnyel a súlyozott első és második momentum $\mu_1(G,w)=0$ és $\mu_2(G,w)=\pi^2/3-2$. Ekkor az eltolás-skála mentes tesztstatisztika logisztikus eltolás-skála család esetében

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

ahol az együtthatók pontosan számolhatóak az alábbi alakban:

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

Csörgő aszimptotikus eredményének [5] a következményeként kapjuk a V_n tesztstatisztika határeloszlását.

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

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

ahol a határérték 1 valószínűséggel létezik.

Del Barrio, Cuesta-Albertos, Matrán [9] a súly nélküli tesztstatisztika határeloszlását megadták súlyozott Brown-hidak Karhunen–Loève-sorfejtéseként. Ugyanezen technikával meghatároztuk a határeloszlás végtelen soros alakját.

8. Tétel. *A V határeloszlás felírható*

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

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

Szimulációs eredmények

Hasonlóan az előbbi fejezetekhez egy szimulációs erővizsgálatot hajtottunk végre, majd összehasonlítottuk az új tesztet az empirikus karakterisztikus függvényre és empirikus momentum generáló függvényre alapozott Meintanis-tesztekkel [15]. A kapott eredményeket az 1. táblázat foglalja össze. Általános konkluziója ennek a szimulációs vizsgálatnak, hogy könnyen számolható tesztstatisztikájú, akár az aszimptotikus kritikus értékeket is használható, közepes erősségű tesztet kaptunk.

Hivatkozások

[1] 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.

[2] F. Balogh and É. Krauczi. Weighted quantile corelation test for the logistic family. *Acta Scientiarum Mathematicarum.(Szeged)*, 80(1-2):307–326, 2014.

[3] 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.

[4] S. Csörgő. Weighted correlation tests for scale families. *Test*, 11(1):219–248, 2002.

[5] 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.

[6] S. Csörgő and W. B. Wu. On the clustering of independent uniform random variables. *Random Structures Algorithms*, 25(4):396–420, 2004.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] T. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70:723–726, 1983.

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

Alternatívák	20	50	100	20	50	100
$N(0,1)$	5	6	8	2	2	4
Egyenletes	13	47	93	5	29	82
Cauchy	88	99	*	84	99	*
Laplace	26	39	55	17	29	43
Exp(1)	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)	40	81	99	27	69	98
Lognormal	86	*	*	79	*	*
Student(5)	16	19	21	10	12	13
χ^2_1	94	*	*	88	*	*
Negatív Exp	69	99	*	56	97	*
α		0,10			0,05	

[12] 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.

[13] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale del Istituto Italiano degli Attuari*, 4:83–91, 1933.

[14] É. Krauczi. A study of the quantile correlation test of normality. *Test*, 18(1):156–165, 2009.

[15] 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.

[16] K. É. Oszténiyiné. Joint cluster counts from uniform distribution. *Probability and Mathematical Statistics*, 33(1):93–106, 2013.

[17] E. S. Pearson. A further development of tests for normality. *Biometrika*, 22:239–249, 1930.

[18] 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.

[19] S. Shapiro and M. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[20] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.