NORMALITÁS-VIZSGÁLAT

Írta: MAJOR PÉTER és TUSNÁDY GÁBOR

BEVEZETÉS

Ez a tanulmány az MTA Számítástechnikai és Automatizálási Kutatóintézetének a felkérésére az MTA Matematikai Kutató Intézetének Matematikai statisztikai osztályán készült. Célja a normalitás-vizsgálat irodalmának az ismertetése, és az egyes módszerek összehasonlítása.

A normalitás-vizsgálat alapja a következő

MODELL. A $\xi_1, \xi_2, \ldots, \xi_n$ valószínűségi változók függetlenek, egyforma eloszlásúak, és közös eloszlásfüggvényük, $F(x)$ folytonos.

Ebben a modellben azt kívánjuk eldönteni, hogy igaz-e a következő

HIPOTÉZIS. A ξ_i változók normális eloszlásúak, azaz

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

alkalmas (de ismeretlen) μ és σ mellett, ahol $\Phi(u)$ a standard normális eloszlásfüggvény:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}}\, dt.$$

Lényeges, az egész vizsgálatot meghatározó körülmény, hogy az eloszlás várható értéke, μ, és szórása, σ ismeretlen. Ha ez nem volna így, feladatunk speciális esete volna az ún. tiszta illeszkedés-vizsgálati feladatnak, amely szerint azt kell eldöntenünk, hogy a fenti modellben igaz-e az a hipotézis, hogy

$$F(x) = F_0(x),$$

ahol $F_0(x)$ tetszőleges, adott (folytonos) függvény.

A normalitás-vizsgálat természetes általánosítása a következő. Döntsük el, hogy a fenti modellben igaz-e a következő

HIPOTÉZIS. A ξ_i változók közös eloszlásfüggvénye

$$F(x) = F(x; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)$$

alakú, ahol $(\vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ ismeretlen paraméterek. Ezt a feladatot becsléses illeszkedés-vizsgálatnak nevezzük. Már ez az elnevezés is sejteti, hogy a megoldást legtermészetesebb a következő irányban keresni. A tiszta illeszkedés-vizsgálat megoldása általában valamilyen

$$\Delta_n = \delta(F_n(x), F_0(x))$$

alakú statisztikán alapszik, ahol $F_n(x)$ az empirikus eloszlásfüggvény:

$$F_n(x) = \frac{1}{n} \sum_{i:\xi_i < x} 1 = \frac{\nu_n(x)}{n},$$

$(\nu_n(x)$ az x-nél kisebbek száma a $\xi_1, \xi_2, \ldots, \xi_n$ mintaelemek között), és $\delta(G(x), H(x))$ a G és H függvények eltérését mérő alkalmas funkcionál. Ezt figyelembe véve a becsléses illeszkedés-vizsgálati feladat megoldását kereshetjük a következő alakban. Adjuk meg először a ϑ_i paraméterek alkalmas

$$\hat{\vartheta}_i = \hat{\vartheta}_i(\xi_1, \xi_2, \ldots, \xi_n) \quad (i = 1, 2, \ldots, k)$$

becsléseit, majd mérjük a vizsgált hipotézis teljesülésének a mértékét a

$$\hat{\Delta}_n = \delta(F_n(x), F(x; \hat{\vartheta}_1, \hat{\vartheta}_2, \ldots, \hat{\vartheta}_k))$$

statisztikával. A normalitás-vizsgálat esetében a paraméterek becsléséül nyilván a

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

mintaátlagot, és az

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}$$

tapasztalati szórást választjuk. Mint ismeretes, ezek a μ, σ paraméterpár torzítatlan becslései között a legkisebb szórásúak (σ-nál egy konstans szorzótól eltekintve).

Bármennyire természetes is ez a megoldás, sok szempontból nem látszik megnyugtatónak. Az első szempont esztétikai. Akik a tiszta illeszkedés-vizsgálati módszereket ismerik, tudják, milyen nehéz e módszerek közül a legjobbat kiválasztani. A normális eloszlásnak az egész valószínűségszámításban elfoglalt központi szerepe arra vezette a statisztikusokat, hogy önálló, a tiszta illeszkedés-vizsgálattól független módszereket dolgozzanak ki, amelyektől — ha egzaktul ezt róluk bizonyítani nem is lehet — legalább heurisztikusan azt várhatjuk, hogy közvetlenül az eloszlás normalitását, normális voltát ellenőrzik. Hogy ez a törekvés milyen régi a statisztikusok között, azt talán legjobban PEARSON 1930-as(!) cikke [16] bizonyítja (címe: A further development of tests for normality, amiből a „further" jelzőt külön érdemes kiemelni). Először a megoldást a momentumok módszerével kívánták megadni. Ennek az irányzatnak a legjobban kidolgozott eredményeit GEARY [7] cikke tartalmazza. Hamarosan kiderült azonban, hogy azzal, hogy az eloszlás ferdeségét, csúcsosságát teszteljük, tulajdonképpen nem az eloszlás normalitását ellenőrizzük. Hiszen számtalan eloszlás van, amelynek ferdesége, vagy csúcsossága 0, és az eloszlás mégis igen messze van a normálistól.

A becsléses illeszkedés-vizsgálati módszerek lényegesebb hiányossága a fenti, esztétikai kifogáson túl az, hogy nem ad megoldást a következő feladatra. Ennek alapja az alábbi

TÖBBMINTÁS-MODELL. A $\xi_{i1}, \xi_{i2}, \ldots, \xi_{in_i}$ $(i=1, 2, \ldots, k)$ valószínűségi változók függetlenek, az azonos kezdő indexű $\xi_{i1}, \xi_{i2}, \ldots, \xi_{in_i}$ változók egyforma eloszlásúak, és közös eloszlásfüggvényük, $F_i(x)$ folytonos $(i=1, 2, \ldots, k)$.

Ebben a modellben azt vizsgáljuk, hogy igaz-e a következő

HIPOTÉZIS. A ξ_{ij} változók normális eloszlásúak, azaz

$$F_i(x) = \Phi\left(\frac{x-\mu_i}{\sigma_i}\right) \quad (i = 1, 2, \ldots, k)$$

alkalmas (de ismeretlen) μ_i, σ_i paraméterekkel.

A statisztikai gyakorlatban ugyanis az esetek többségében nem egyetlen minta normalitását kell megvizsgálnunk, hanem sok kis mintáról együtt kell eldöntenünk, hogy normális-e az eloszlásuk, vagy sem. Ezek a kis minták sokszor csak a lokációs és a skála paraméterekben különböznek egymástól, vagyis az eredeti modell helyettesíthető egy speciálisabb változatával:

SPECIÁLIS TÖBBMINTÁS MODELL. A többmintás modell feltevésein túl feltesszük, hogy a ξ_{ij} változók eloszlása lineáris transzformációval kapható meg az $F_0(x)$ eloszlásból, azaz

$$F_i(x) = F_0\left(\frac{x-\mu_i}{\sigma_i}\right) \quad (i = 1, 2, \ldots, k)$$

alkalmas (de ismeretlen) μ_i, σ_i paraméterekkel (ahol $F_0(x)$ folytonos eloszlásfüggvény).

Ebben a speciális modellben elsősorban az $F_0(x)$ eloszlásfüggvény érdekel minket, a μ_i, σ_i paraméterek értéke közömbös, emiatt e paramétereket szokás zavaró paramétereknek is nevezni. Természetes gondolat ebben az esetben az, hogy valamilyen módon megszabaduljunk a zavaró paraméterektől, szemléletes kifejezéssel élve alkalmas transzformációval „kidobjuk" a paramétereket. Ilyen eljárást a normális eloszlás esetében DURBIN [6], SARKADI [21], és STÖRMER [27] adott meg, és ezzel megalapozták a normalitás-vizsgálat új, és szerintünk leghatásosabb módszerét.

Dolgozatunk első és második része a tiszta és a becsléses illeszkedés-vizsgálattal foglalkozik, a harmadik részben pedig ezeket a transzformációs módszereket ismertetjük.

1. TISZTA ILLESZKEDÉS-VIZSGÁLAT

MODELL. A $\xi_1, \xi_2, \ldots, \xi_n$ valószínűségi változók függetlenek, egyforma eloszlásúak, és közös eloszlásfüggvényük, $F(x)$ folytonos.

HIPOTÉZIS. A ξ_i változók közös eloszlásfüggvénye az ismert $F_0(x)$, azaz

$$F(x) = F_0(x),$$

(ahol F_0 természetesen folytonos).

Ennek a hipotézisnek a fenti modellben történő ellenőrzését tiszta illeszkedés-vizsgálatnak nevezzük. A tiszta illeszkedés-vizsgálat módszereit SAHLER [20] cikke alapján ismertetjük, ebben a cikkben található meg e témakör részletes irodalomjegyzéke is.

Hipotézisünket statisztikai próbával ellenőrizhetjük. A statisztikai próba alapja — a fenti modell esetében — egy n-változós függvény, jelöljük $\pi(x_1, x_2, \ldots, x_n)$-nel, amelynek lehetséges értékei 0 és 1 közöttiek:

$$0 \leqq \pi(x_1, x_2, \ldots, x_n) \leqq 1.$$

Ennek alapján eljárásunk a következő: $\pi(\xi_1, \xi_2, \ldots, \xi_n)$ valószínűséggel elutasítjuk a vizsgált hipotézist, $1-\pi(\xi_1, \xi_2, \ldots, \xi_n)$ valószínűséggel pedig elfogadjuk azt. Az esetek többségében π csak 0-val, vagy 1-gyel egyenlő, ekkor az n dimenziós térnek azt a részét, ahol $\pi=0$, elfogadási tartománynak, a $\pi=1$ feltétellel definiált részt pedig elutasítási, vagy kritikus tartománynak nevezzük. A $\pi(\xi_1, \xi_2, \ldots, \xi_n)$ függvény maga is valószínűségi változó, ennek a várható értéke a ξ_i-k eloszlásfüggvényétől, F-től függ. Jelöljük ezt a várható értéket $\beta_\pi(F)$-fel:

$$\beta_\pi(F) = E_F \pi(\xi_1, \xi_2, \ldots, \xi_n).$$

Ezt az értéket a π próba erejének nevezzük, $\beta_\pi(F_0)$ adja meg annak a valószínűségét, hogy a vizsgált hipotézis igaz, mégis elutasítjuk azt. Ezt a hibát elsőfajú hibának, az elsőfajú hiba $\beta_\pi(F_0)$ valószínűségét pedig a próba terjedelmének nevezzük, és ε-nal jelöljük:

$$\varepsilon = \beta_\pi(F_0).$$

Tetszőleges szintű próbát kapunk, ha valamilyen (folytonos) $T(x_1, x_2, \ldots, x_n)$ függvényt veszünk alapul, és T alapján π-t a

$$\pi(x_1, x_2, \ldots, x_n) = \begin{cases} 0, & \text{ha} \quad T(x_1, x_2, \ldots, x_n) < c \\ 1, & \text{ha} \quad T(x_1, x_2, \ldots, x_n) \geqq c \end{cases}$$

összefüggéssel definiáljuk, ahol c értékét úgy határozzuk meg, hogy a próba szintje ε legyen, azaz

$$P_{F_0}\big(T(\xi_1, \xi_2, \ldots, \xi_n) < c\big) = 1 - \varepsilon$$

teljesüljön.

A statisztikai próbák elméletében fontos szerepet játszik az invariancia: bizonyos transzformációkkal ugyanis a legtöbb hipotézis-vizsgálati feladatot átfogalmazhatjuk; előnyben részesítjük azokat az eljárásokat, amelyek e transzformációk hatására érzéketlenek, illetve amelyekre a transzformációk természetes módon hatnak. Esetünkben a transzformáció alapja tetszőleges monoton növő $g(x)$ függvény lehet, ennek segítségével az eredeti mintához az

$$\eta_i = g(\xi_i) \quad (i = 1, 2, \ldots, n)$$

mintát rendelhetjük. Hipotézisünk teljesülése esetén az η_i-k eloszlásfüggvénye

$$G_0(x) = P(\eta_i < x) = P\big(g(\xi_i) < x\big) = P\big(\xi_i < g^{-1}(x)\big) = F_0\big(g^{-1}(x)\big).$$

Jelöljük az F_0 eloszlás ellenőrzésére szolgáló próbafüggvényt $\pi(x_1, x_2, \ldots, x_n; F_0)$-lal. Természetes követelmény a π próbával szemben, hogy ha a mintát a g függvénnyel transzformáljuk, akkor a transzformált G_0 eloszlást a transzformált $\eta_1, \eta_2, \ldots, \eta_n$

mintán ugyanígy ellenőrizze, mint az eredeti $\xi_1, \xi_2, \dots, \xi_n$ mintán F_0-t, azaz

$$\pi\big(g(x_1), g(x_2), \dots, g(x_n); G_0\big) = \pi(x_1, x_2, \dots, x_n; F_0) \tag{1}$$

legyen, ahol $G_0(x) = F_0\big(g^{-1}(x)\big)$.

Definíció. A $\pi(x_1, x_2, \dots, x_n; F_0)$ próbafüggvényt — és a megfelelő próbát — eloszlásmentesnek nevezzük, ha π-re teljesül (1) minden monoton növekvő g függvény mellett.

Eloszlásmentes eljárást kapunk, ha a próbához az

$$F_n(x) = \frac{1}{n} \sum_{i:\xi_i<x} 1$$

empirikus eloszlásfüggvény alapján a

$$T(\xi_1, \xi_2, \dots, \xi_n) = \delta\big(F_n(x), F_0(x)\big)$$

statisztikát használjuk, ahol $\delta(G, H)$ a G és H eloszlásfüggvények távolságát mérő tetszőleges funkcionál.

Az eloszlásmentes próbák vizsgálatánál mindig feltehetjük, hogy a ξ_i-k a vizsgált hipotézis szerint a $(0, 1)$ intervallumban egyenletes eloszlásúak, hiszen a $g=F_0$ transzformáció tetszőleges F_0 mellett erre az esetre vezet. Emiatt a továbbiakban feltehetjük, hogy

$$F_0(x) = \begin{cases} 0, & \text{ha} \quad x < 0; \\ x, & \text{ha} \quad 0 \leqq x \leqq 1; \\ 1, & \text{ha} \quad x > 1. \end{cases}$$

Annak érdekében, hogy az ε szinthez tartozó c szignifikancia-határt megkapjuk, meg kell határoznunk a $\delta\big(F_n(x), F_0(x)\big)$ statisztika eloszlását abban az esetben, ha a ξ_i-k a $(0, 1)$ intervallumban egyenletes eloszlásúak. Vannak δ-k, amelyekre ez zárt formában megadható, vannak, amikre csak a határeloszlás adható meg, és vannak, amelyekre csak Monte-Carlo-módszerrel készíthetjük el (vagy készítették el) a szükséges táblázatokat. Az alábbiakban áttekintjük a legfontosabb funkcionálokat — ezzel együtt a legfontosabb tiszta illeszkedés-vizsgálati módszereket.

1.1. A Kolmogorov—Szmirnov próba

A próba statisztikája egyoldali esetben

$$D_n^+ = \sup_{-\infty<x<\infty} [F_n(x) - F_0(x)],$$

kétoldali esetben

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|.$$

A próba általánosítása a tetszőleges pozitív ψ-hez tartozó

$$D_n^+(\psi) = \sup_{a<F_0(x)<b} \psi\big(F_0(x)\big)[F_n(x) - F_0(x)],$$

illetve

$$D_n(\psi) = \sup_{a<F_0(x)<b} \psi(F_0(x))\,|F_n(x)-F_0(x)|$$

statisztikákon alapszik. Ennek speciális esete a *Rényi*-próba, amelynél

$$\psi(x) = \frac{1}{x}, \quad a = \alpha, \quad b = 1;$$

vagy

$$\psi(x) = \frac{1}{1-x}, \quad a = 0, \quad b = 1-\alpha;$$

ahol $0<\alpha<1$.

E statisztikák közül sokáig csak D_n^+ eloszlása volt ismeretes, D_n eloszlására csak igen bonyolult összegezéseket tartalmazó formulák voltak. Könnyen látható, hogy az összes *Kolmogorov—Szmirnov*-típusú próba-statisztika eloszlása meghatározható STECK [25] alábbi eredménye alapján.

TÉTEL. *Legyenek a ξ_i-k a* (0, 1)*-ben egyenletes eloszlású, független valószínűségi változók* ($i=1, 2, \ldots, n$), *és ξ_i^* legyen közöttük a nagyság szerint i-edik. Legyen továbbá*

$$0 \leqq a_1 \leqq a_2 \leqq \ldots \leqq a_n \leqq 1,$$

$$0 \leqq b_1 \leqq b_2 \leqq \ldots \leqq b_n \leqq 1,$$

$$a_i \leqq b_i \quad (i = 1, 2, \ldots, n).$$

Akkor

$$P(a_i \leqq \xi_i^* \leqq b_i, \quad i = 1, 2, \ldots, n) =$$

$$= n!\,\det\left[(b_i - a_j)_+^{j-i+1}/(j-i+1)!\right] =$$

$$= n! \begin{vmatrix} (b_1-a_1)_+ & \dfrac{(b_1-a_2)_+^2}{2!} & \ldots & \dfrac{(b_1-a_n)_+^n}{n!} \\ 1 & (b_2-a_2)_+ & \ldots & \dfrac{(b_2-a_n)_+^{n-1}}{(n-1)!} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & (b_n-a_n)_+ \end{vmatrix}$$

(*ahol* $(a)_+ = a$, *ha* $a>0$, *és* $(a)_+ = 0$, *ha* $a \leqq 0$, és $0^0 = 1$).

Erre a tételre igen egyszerű bizonyítást adott SARKADI [22]. A tétel alapján D_n kis n-ek melletti eloszlása számítógépen könnyen meghatározható.

Az eredeti D_n^+ és D_n, valamint a *Rényi*-féle $D_n^+(\psi)$ és $D_n(\psi)$ statisztikák határeloszlására zárt formulák ismeretesek, ezekre az eloszlásokra táblázatok is találhatók, például a [2], [14], [17] táblázat-gyűjteményekben. A konvergencia gyorsasága kiolvasható az alábbi lemmából.

LEMMA.

$$P(\sqrt{n}\, D_n \leqq z) = K(z) + \frac{1}{6\sqrt{n}} K'(z) + o\left(\frac{1}{n}\right),$$

ahol

$$K(z) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2z^2}.$$

Itt említjük meg a

$$V_n = \sup_{-\infty<x<\infty} [F_n(x) - F_0(x)] + \sup_{-\infty<x<\infty} [F_0(x) - F_n(x)]$$

statisztikát, amelynek határeloszlása

$$\lim_{n\to\infty} P\left(\sqrt{n}\, V_n \leqq z\right) = 1 + \sum_{j=1}^{\infty} (2 - 8j^2z^2)e^{-2j^2z^2},$$

(természetesen itt is, és az előbbi lemmában is z pozitív). Úgy gondoljuk, puszta véletlen, hogy a D_n statisztika terjedt el a gyakorlatban, a magunk részéről V_n és D_n között semmi különbséget nem látunk.

1.2. A Cramér—Mises-próba

A próba statisztikája

$$\Omega_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x) = \frac{1}{12n} + \sum_{i=1}^{n} \left[F_0(\xi_i^*) - \frac{2i-1}{2n}\right]^2,$$

ahol ξ_i^* a rendezett minta i-edik eleme. A próba általánosítása a tetszőleges pozitív $\psi(x)$ súlyfüggvényhez tartozó

$$\Omega_n^2(\psi) = n \int_{-\infty}^{\infty} \psi(F_0(x)) [F_n(x) - F_0(x)]^2 dF_0(x)$$

statisztikán alapszik. Ennek speciális esete az *Anderson—Darling*-próba, amelynél

$$\psi(x) = \frac{1}{x(1-x)}.$$

Ezeknek a statisztikáknak az egzakt eloszlása nem ismeretes, de a határeloszlásuk igen. E határeloszlások ugyanis, akárcsak az előző pontban szereplő határeloszlások, meghatározhatók a következő tétel segítségével.

TÉTEL (DONSKER). *Legyen $\psi(f)$ a* [0, 1]*-beli, szakaszonként folytonos függvényeken értelmezett, folytonos funkcionál.* (Ez az utóbbi feltétel azt jelenti, hogy tetszőleges, a [0, 1]-ben szakaszonként folytonos f_0 függvényhez, és tetszőleges pozitív ε-hoz van olyan pozitív δ, hogy

$$|\psi(f) - \psi(f_0)| < \varepsilon,$$

teljesül minden olyan, a [0, 1]-ben szakaszonként folytonos f függvényre, amelyre

$$|f(x) - f_0(x)| < \delta$$

minden $0 \leqq x \leqq 1$ mellett.) *Legyenek továbbá a* $\xi_1, \xi_2, \ldots, \xi_n$ *valószínűségi változók függetlenek, és a* $(0, 1)$*-ben egyenletes eloszlásúak. Akkor a*

$$\Delta_n = \psi(\sqrt{n}\,[F_n(x) - x])$$

statisztika határeloszlása egyenlő a $\Delta = \psi(W(t))$ *változó határeloszlásával, azaz*

$$\lim_{n=\infty} P(\Delta_n < x) = P(\Delta < x);$$

ahol $F_n(x)$ *az empirikus eloszlásfüggvény, és* $W(t)$ *az ún. Brown-bridge, azaz* $W(t)$ *olyan Gauss-folyamat, amelyre* $EW(t) = 0$, *és* $EW(t)W(s) = t(1-s)$, *ha* $0 \leqq t \leqq s \leqq 1$.

E tétel alapján tetszőleges ψ funkcionállal generálhatunk próbát, hiszen a tétel alapján ahhoz, hogy a próba-statisztika határeloszlását meghatározzuk, „csak" a $\Delta = \psi(W(t))$ valószínűségi változó eloszlását kell meghatároznunk. Amíg azonban a *Kolmogorov—Szmirnov*-típusú statisztikák esetében ezeknek az eloszlásoknak a meghatározására nem ismerünk általános módszert, a *Cramer—Mises*-típusú statisztikák határeloszlása előállítható a következő, KAC—SIEGERTtől [9] származó lemma alapján.

LEMMA. *Legyen* $\xi(t)$ *a* $0 \leqq t \leqq 1$ *szakaszon értelmezett Gauss folyamat, a várható értéke legyen* 0, *és jelöljük a kovariancia függvényét* $B(s, t)$*-vel:*

$$B(s, t) = E\xi(s)\xi(t).$$

Akkor az

$$\eta = \int_0^1 \xi^2(t)\,dt$$

valószínűségi változó eloszlása megegyezik az

$$\tilde{\eta} = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \eta_i^2$$

összeg eloszlásával, ahol az η_i*-k függetlenek, standard eloszlásúak, és a* λ_i *számok az*

$$\int_0^1 \psi(s) B(s, t)\,ds = \lambda \psi(t)$$

integrálegyenlet sajátértékei.

1.3. Elemi statisztikák

Mint láttuk, feltehetjük, hogy az a hipotézis, hogy a ξ_i-k a $(0, 1)$-ben egyenletes eloszlásúak, és a ξ_i-k értékkészlete a modell szerint is része a $(0, 1)$ szakasznak. Azt kell tehát ellenőriznünk, hogy a $(0, 1)$-beli

$$0 = \xi_0^* \leqq \xi_1^* \leqq \xi_2^* \leqq \ldots \leqq \xi_n^* \leqq \xi_{n+1}^* = 1$$

n elemű rendezett minta (ξ_0^*, ξ_{n+1}^* szerepeltetése csak a jelöléseket egyszerűsíti) mennyire egyenletesen osztja fel a $(0, 1)$ intervallumot, azaz az

$$F_n(x) = \frac{i}{n}, \quad \text{ha} \quad \xi_i^* < x \leqq \xi_{i+1}^* \quad (i = 0, 1, \ldots, n)$$

empirikus eloszlásfüggvény milyen mértékben közelíti meg az $F_0(x)=x$ függvényt, vagyis a

$$W_n(x) = \sqrt{n}\,[F_n(x)-x]$$

függvény mennyire tér el 0-tól. Az eddigi statisztikák W_n nagyságát vizsgálták: a *Kolmogorov—Szmirnov*-típusú statisztikák a legnagyobb értéket, a *Cramér—Mises*-típusú statisztikák a globális eltérést ellenőrizték. Elvileg azonban akármilyen funkcionált használhatunk, a további lehetőségek közül mutatunk be itt néhányat.

a) *Az eloszlásfüggvény árnyéka*

Képzeljük azt, hogy az origóban egy fényforrás van, és az $(x, F_n(x))$ görbe árnyékát az $y=1$ egyenesen látjuk: jelöljük az árnyék kezdőpontját α_n^+-szal:

$$\alpha_n^+ = \min_{0<x<1} \frac{x}{F_n(x)},$$

és legyen e statisztika párja, α_n^- az (1, 1) pontbeli fényforrásból származó árnyék az x tengelyen:

$$\alpha_n^- = \min_{0<x<1} \frac{1-x}{1-F_n(x)}.$$

Mindkettő egyenletes eloszlású (0, 1)-ben, és mindkettőnek a nagy értékei szólnak a hipotézis teljesülése ellen. A D_n, W_n statisztikák mintájára ezekből is előállíthatjuk a

$$\alpha_n = \max(\alpha_n^+, \alpha_n^-), \quad \tilde{\alpha}_n = \alpha_n^+ + \alpha_n^-$$

statisztikákat, ezek eloszlása azonban már bonyolultabb, és függ a mintaelemszámtól.

b) *Pozitív szakaszok*

Ez a statisztika azt méri, mennyire szimmetrikus a $W_n(t)$ folyamat értékkészlete a 0-ra: a statisztika értéke annak a halmaznak a mértéke, ahol $W_n(t)$ pozitív:

$$\pi_n = \int_0^1 P(W_n(t))\,dt,$$

ahol

$$P(u) = \begin{cases} 1, & \text{ha} \quad u > 0; \\ 0, & \text{ha} \quad u \leqq 0. \end{cases}$$

A π_n statisztika értéke tehát azoknak a szakaszoknak az összhossza, amelyek felett $W_n(t)$ pozitív, vagyis $F_n(x)>x$, az empirikus eloszlásfüggvény az $F_0(x)=x$ eloszlás felett van. A π_n statisztika eloszlása egyenletes, túl kicsi és túl nagy értékei a hipotézis ellen szólnak.

c) *Átmetszések száma*

A statisztika értéke azoknak a (0, 1)-beli x-eknek a száma, amelyekre $F_n(x)=x$, azaz $W_n(x)=0$ teljesül. A statisztika eloszlása, általánosításai, és megfelelő határeloszlások megtalálhatók a [4] dolgozatban. A statisztika alacsony értékei szólnak a hipotézis teljesülése ellen.

d) *Kvázi—Cramér—Mises-statisztikák*

A *Cramér—Mises*-statisztika kissé módosított változata a

$$\beta_n(\lambda) = \sum_{i=1}^{n} \left| \xi_i^* - \frac{i}{n+1} \right|^\lambda$$

statisztika, ahol $\lambda \geqq 1$. Ez a statisztika azt méri, mennyire maradnak a ξ_i^* rendezett mintaelemek a várható értékük, $i/n+1$ közelében. Előnye, hogy az értéke könnyen kiszámítható, és a határeloszlása egyszerű: ugyanis a tetszőleges $\lambda \geqq 1$ mellett a határeloszlás normális. Hibája viszont, hogy nem veszi figyelembe azt a tényt, hogy a különböző rendezett minta elemek szórása igen eltérő lehet. Ezt a hibát küszöböli ki a

$$\gamma_n = \sum_{i=1}^{n} \frac{\left(\xi_i^* - \frac{i}{n+1} \right)^2}{i(n+1-i)}$$

statisztika, amely már a szórásokat is figyelembe veszi. Ebből a szempontból még egzaktabb a

$$\delta_n = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \left(\xi_i^* - \frac{i}{n+1} \right) \left(\xi_j^* - \frac{j}{n+1} \right)$$

statisztika felépítése, ahol az $A=(a_{ij})$ mátrix a rendezett minta elemek C kovarianciamátrixának az inverze: $A=C^{-1}$. A $C=(c_{ij})$ mátrix elemei tehát

$$c_{ij} = \frac{i(n+1-j)}{(n+1)^2}, \quad \text{ha} \quad 1 \leqq i \leqq j \leqq n;$$

(és természetesen C szimmetrikus).

e) *A Moran-statisztikák*

Ismeretes, hogy ha pozitív számok összege állandó, négyzetösszegük akkor minimális, ha egyenlőek. Ennek alapján mérhetjük a minta egyenletes eloszlását a

$$\mu_n(\lambda) = \sum_{i=0}^{n} |\xi_{i+1}^* - \xi_i^*|^\lambda$$

statisztikával, ahol $\lambda \geqq 0$. E statisztikák sok szempontból hasonlítanak a kvázi-*Cramér—Mises*-statisztikákra, határeloszlásuk normális, és éppúgy lehetne őket továbbfejleszteni, mint a β statisztikákat.

Ezzel korántsem ért véget a lehetséges statisztiák felsorolása, ez nem is volt célunk, csak a lehetőségeket akartuk érzékeltetni. E statisztikák inkább színező jellegűek, valamely „fő" próba mellett kiegészítésül használhatjuk őket. Utalunk ezzel kapcsolatban VINCZE [28] cikkére, aki azt vizsgálta, milyen mértékben lehet a *Kolmogorov—Szmirnov*-próba erejét növelni az első maximum hely (index) figyelembevételével.

1.4. Sűrűségfüggvények

Eddigi próbáink mind azt ellenőrizték valamilyen formában, milyen közel van az F_n empirikus eloszlásfüggvény a hipotetikus F_0 eloszlásfüggvényhez. Most olyan próbákról lesz szó, amelyek a sűrűségfüggvényt ellenőrzik. A sűrűségfüggvényt lényegében háromféleképpen becsülhetjük:

— hisztogrammal,

— ortogonális sorfejtéssel,

— Parzen-féle magfüggvénnyel.

Ebben a sorrendben tárgyaljuk tehát a lehetséges próbákat.

a) *Gyakorisági hisztogram: χ^2-próba*

Legyen

$$-\infty = a_0 < a_1 < a_2 < \ldots < a_k < a_{k+1} = \infty$$

egy beosztása a számegyenesnek, és jelöljük a j-edik intervallumot I_j-vel: $I_j=(a_j, a_{j+1})$ ($j=0, 1, 2, \ldots, k$), a $\xi_1, \xi_2, \ldots, \xi_n$ mintának az I_j-be eső elemeinek a számát pedig ν_j-vel:

$$\nu_j = \sum_{i:\xi_i \in I_j} 1 = \nu(a_{j+1}) - \nu(a_j),$$

ahol $\nu(x)$ az x-nél kisebb mintaelemek számát jelöli. A beosztáshoz tartozó gyakorisági hisztogram

$$f_n(x) = \frac{\nu_j}{n(a_{j+1}-a_j)}, \quad \text{ha} \quad x \in I_j \quad (j = 0, 1, 2, \ldots, k).$$

(Eszerint I_0-ban és I_k-ban f_n értéke definíció szerint 0.) Azt, hogy ez a sűrűségfüggvény milyen közel van a teoretikus

$$f_0(x) = F_0'(x)$$

sűrűségfüggvényhez (amelynek a létezését ebben a pontban feltételezzük) az előzőek alapján a

$$d_n^+ = \sup_{-\infty<x<\infty} [f_n(x) - f_0(x)],$$

$$d_n = \sup_{-\infty<x<\infty} |f_n(x) - f_0(x)|,$$

$$\omega_n^2 = \int_{-\infty}^{\infty} [f_n(x) - f_0(x)]^2 f_0(x)\, dx,$$

statisztikákkal mérhetnénk. Ezek azonban általában nem eloszlásmentesek: vagy a beosztást kell F_0 figyelembevételével alkalmasan megválasztani, vagy a

$$d_n^+(\psi) = \sup_{-\infty<x<\infty} \psi(f_0(x))\,[f_n(x)-f_0(x)],$$

$$d_n(\psi) = \sup_{-\infty<x<\infty} \psi(f_0(x))\,|f_n(x)-f_0(x)|,$$

$$\omega_n^2(\psi) = \int_{-\infty}^{\infty} \psi(f_0)\,[f_n(x)-f_0(x)]^2 f_0(x)\,dx$$

általánosítsukban szereplő ψ súlyfüggvényt kell alkalmasan megválasztani. A d-típusú statisztikákkal először RÉVÉSZ foglalkozott, eredményeivel kapcsolatban [18] alatti cikkére hivatkozunk. Az ω^2-típusú statisztikák között központi szerepet játszik az, amelyik lényegében a $\psi(t)=1/t^2$ súlyfüggvénynek felel meg, és amelyiket χ^2-próbaként ismerünk:

$$\chi_k^2 = \sum_{j=0}^{k} \frac{(\nu_j-np_j)^2}{np_j},$$

ahol

$$p_j = F_0(a_{j+1})-F_0(a_j) = \int_{a_j}^{a_{j+1}} f_0(x)\,dx.$$

Ez a próba ugyan egzaktul nem eloszlásmentes, de a határeloszlása már az: ha n tart a végtelenbe, a statisztika határeloszlása a k szabadságfokú χ^2-eloszlás. A próba legrészletesebb elemzését COCHRAN [3] dolgozatában találjuk meg. MANN és WALD [12] a χ^2-próba és a *Kolmogorov—Szmirnov*-próba erejét hasonlította össze.

b) *Ortogonális sorfejtés: Neyman—Barton-próba*

Módosítsuk most úgy a modellt, hogy még azt is a feltevések közé számítjuk, hogy a közös sűrűségfüggvény:

$$f(x) = f_0(x) \sum_{j=0}^{k} a_j\psi_j(x)$$

alakú, ahol az a_j-k alkalmas együtthatók, a ψ_j-k pedig az f_0 súlyfüggvényre nézve ortonormált rendszert alkotnak:

$$\int_{-\infty}^{\infty} \psi_i(x)\psi_j(x)f_0(x)\,dx = \delta_{ij},$$

ahol δ_{ij} értéke 1, vagy 0 aszerint, hogy i egyenlő-e j-vel, vagy sem. Feltesszük továbbá, hogy $\psi_0(x)=1$. Ebben a modellben az a_j együtthatók torzítatlan becslése

$$\alpha_j = \frac{1}{n}\sum_{i=1}^{n}\psi_j(\xi_i) \qquad (j = 1, 2, \dots, k),$$

hiszen a ψ rendszer ortonormált volta miatt

$$E\psi_j(\xi_i) = \int_{-\infty}^{\infty} \psi_j(x) f(x)\,dx = \int_{-\infty}^{\infty} \psi_j(x) \sum_{\nu=0}^{k} a_\nu \psi_\nu(x) f_0(x)\,dx =$$

$$= \sum_{\nu=0}^{k} a_\nu \delta_{\nu j} = a_j.$$

(Mivel $\psi_0 \equiv 1$, $a_0 = \alpha_0 = 1$, így csak $j \geqq 1$ mellett kell az a_j-ket becsülnünk.)

A vizsgált hipotézis most ekvivalens azzal, hogy

$$\sum_{i=1}^{k} a_i^2 = 0,$$

azaz a_0 kivételével az összes együttható 0 (mivel f is sűrűségfüggvény, ebből már következik, hogy $a_0 = 1$). Ha ez a hipotézis teljesül, az α_j becslések korrelálatlanok, és a szórásnégyzetük $1/n$:

$$\operatorname{cov} \alpha_i, \alpha_j = \frac{1}{n^2} \sum_{\nu=1}^{n} \operatorname{cov} \psi_i(\xi_\nu), \psi_j(\xi_\nu) =$$

$$= \frac{1}{n} \int_{-\infty}^{\infty} \psi_i(x) \psi_j(x) f_0(x)\,dx = \delta_{ij}.$$

Mindegyik α_j független valószínűségi változók számtani közepe, tehát a

$$B_k^2 = n \sum_{j=1}^{k} \alpha_j^2 = \frac{1}{n} \sum_{j=1}^{k} \left\{ \sum_{i=1}^{n} \psi_j(\xi_i) \right\}^2$$

statisztika határeloszlása k szabadságfokú χ^2 eloszlás.

Ez a statisztika lényegében NEYMANtól származik (vö.: [13]), aki azt az esetet vizsgálta, amikor f_0 a (0, 1)-ben azonosan 1, és a ψ függvényrendszer a *Hermite*-polinomrendszer. Később BARTON foglalkozott sokat e statisztikákkal (vö. pl. [1]), jelenleg a sűrűségfüggvények becslésének az aszimptotikus viselkedésének a vizsgálatánál ismét az érdeklődés középpontjába került. Ezzel kapcsolatban hivatkozunk még a KENDALL—STUART [10] könyvre, mint az egyetlen nem hazai statisztika-kézikönyvre, ahol a normalitásvizsgálatról szó esik.

c) *A sűrűségfüggvény Parzen-féle becslése*

A módszer inkább PARZEN nevéhez fűződik (vö. [15]), bár PARZEN előtt már ROSENBLATT is foglalkozott vele. Legyen g tetszőleges sűrűségfüggvény, b_n alkalmasan választott 0-hoz tartó sorozat, akkor a sűrűségfüggvény *Parzen*-féle becslése az

$$f_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} g\left(\frac{x-\xi_i}{b_n}\right)$$

függvény. Sokan ezt tartják a sűrűségfüggvény egyetlen becslésének, amit vitathatatlan egyszerűsége, áttekinthetősége, kezelhetősége indokol. Rögzített n mellett b_n természetesen egyetlen szám, értékét tanácsos $n^{-1/5}$ közelében választani. Szemléletesen szólva b_n annak a perturbáló tényezőnek, mesterséges hibának a szórása, amelyet a mintához rendelünk, hogy a diszkrét elemekből álló minta helyett folytonos, összemosott görbét kapjunk. Ezért b_n se túl kicsi nem lehet (ekkor a mintaelemek izoláltak maradnának), se túl nagy (ekkor a valódi f helyett a becsléshez használt g hatása dominálna a becslésben).

A becslés illeszkedésvizsgálati célokra való felhasználásának elvi alapjait Rosenblatt adta meg a [19] cikkben.

2. BECSLÉSES NORMALITÁS-VIZSGÁLAT

Modell. A $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók függetlenek, egyforma eloszlásúak, és közös eloszlásfüggvényük, $F(x)$ folytonos.

Hipotézis. A ξ_i változók normális eloszlásúak, azaz

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

alkalmas (de ismeretlen) μ és σ mellett.

Ennek a hipotézisnek a fenti modellben történő ellenőrzését normalitás-vizsgálatnak nevezzük. Mint a bevezetőben már láttuk, ez a feladat két különböző módon vezethető vissza a tiszta illeszkedés-vizsgálatra: becsléssel és transzformációval. Ebben a fejezetben a becsléses illeszkedés-vizsgálati módszerrel foglalkozunk. Becslésül a

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

mintaátlagot, és az

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}$$

tapasztalati szórást használjuk. Azt fogjuk ellenőrizni, milyen közel van az

$$F_n(x) = \frac{1}{n}\sum_{i:\xi_i<x} 1$$

tapasztalati eloszlásfüggvény a $\bar{\xi}$ várható értékű, s szórású normális eloszláshoz, vagyis azt, hogy mennyire marad a 0 közelében a

$$w_n(x) = \sqrt{n}\left[F_n(x) - \Phi\left(\frac{x-\bar{\xi}}{s}\right)\right]$$

folyamat.

2.1. A Donsker-tétel megfelelője

Ha most is az előző pontban definiált ψ funkcionálokat kívánjuk használni, felmerül a kérdés, milyen mértékben befolyásolja a határeloszlásukat az a körülmény, hogy az eloszlás paramétereit becsültük. Ezt a kérdést először a *Cramér—Mises*-próbával kapcsolatban Darling [5] vizsgálta, nem sokkal utána Kac, Kiefer, WOLFOWITZ [8] a becsléses *Kolmogorov—Szmirnov*-próba határeloszlását határozta meg. Dolgozataikban néhány — *Monte-Carlo*-módszerrel készített — táblázatot is megadnak, ezeket csak itt, az eredeti cikkekben lehet megtalálni. Könnyen látható, hogy a

$$W_n(x) = w_n\big(\Phi^{-1}(sx+\bar{\xi})\big)$$

folyamat eloszlása nem függ a μ, σ paraméterektől, tehát a $\psi\big(W_n(x)\big)$ statisztika eloszlása sem függ μ, σ értékétől egyetlen ψ funkcionál mellett sem, így kellő számú szimulálással — standard normális eloszlásból kiindulva — a

$$P\big(\psi(W_n(x)) < \lambda\big)$$

valószínűségek *Monte-Carlo*-módszerrel meghatározhatók. (Minket ez elsősorban az általunk választott ψ, valamint a mintabeli n és λ mellett érdekel, úgy látjuk, kis mintaelemszám mellett ez a *Monte-Carlo*-módszer a legjobb.)

Visszatérve a határeloszlás meghatározására, mint láttuk, a kérdés az, hogy az ismert határeloszlások változtatás nélkül használhatók-e, vagy sem. Várható ugyanis, hogy a becsült paraméterekhez tartozó normális eloszlás jobban meg tudja közelíteni az empirikus eloszlást, mint maga a teoretikus eloszlás, hiszen a paraméterek becslése épp olyan irányban tér el a valódi értékektől, ahogy azt a minta diktálja. A kérdés csak az, hogy ez a hatás lényeges-e a határeloszlás szempontjából, vagy sem. Az említett szerzők azt találták, hogy ez a hatás lényeges, eredményük a következő tételben foglalható össze.

TÉTEL. *Legyen $\Psi(f)$ a* [0, 1]*-beli, szakaszonként folytonos függvényeken értelmezett folytonos funkcionál. Legyenek továbbá a $\xi_1, \xi_2, \ldots, \xi_n$ valószínűségi változók függetlenek, normális eloszlásúak μ, σ paraméterekkel. Legyen $\bar{\xi}$, és s a minta átlaga, és szórása, az $\eta_1, \eta_2, \ldots, \eta_n$ minta pedig legyen az eredeti minta standardizáltja:*

$$\eta_i = \frac{\xi_i - \bar{\xi}}{s} \qquad (i = 1, 2, \ldots, n).$$

Jelöljük $G_n(x)$-szel az η_i mintához tartozó empirikus eloszlásfüggvényt:

$$G_n(x) = \frac{1}{n} \sum_{i:\eta_i < x} 1 = F_n(sx+\bar{\xi}),$$

ahol F_n az erdeti empirikus eloszlásfüggvény. Akkor a

$$\Delta_n = \Psi\big(\sqrt{n}\,[G_n(x) - \Phi(x)]\big)$$

statisztika határeloszlása egyenlő a $\Delta = \psi\big(W(t)\big)$ változó eloszlásával, azaz

$$\lim_{n\to\infty} P(\Delta_n < x) = P(\Delta < x),$$

ahol $W(t)$ olyan Gauss-folyamat, amelyre $EW(t)=0$, és

$$EW(t)W(s) = \Phi(t)\big(1-\Phi(s)\big) - \frac{1}{2\pi}\left(1+\frac{st}{2}\right)e^{-\frac{s^2+t^2}{2}},$$

ha $-\infty < t \leqq s < \infty$.

Ez persze csak a kérdés egyik fele, a konkrét funkcionálok esetében még meg kell határozni a Δ változó eloszlását. A *Kolmogorov—Szmirnov*-statisztikára ez nem ismeretes, a *Cramér—Misesre* viszont DARLING a már ismert *Kac—Siegert*-lemma alapján megadta a teoretikus eloszlást.

Összehasonlítás céljából megadjuk a következő táblázatokat:

Kolmogorov—Szmirnov-próba $P(\sqrt{n}D_n \leqq z) = \alpha$

Elfogadási szint	Kritikus érték (tiszta illeszkedésvizsgálat)	Kritikus érték (becsléses illeszkedésvizsgálat)
$\alpha=0{,}9$	$z=1{,}23$	$z=0{,}82$
$\alpha=0{,}95$	$z=1{,}36$	$z=0{,}90$
$\alpha=0{,}99$	$z=1{,}63$	$z=1{,}04$

Cramér—Mises-próba $P(\Omega_n^2 \leqq z) = \alpha$

Elfogadási szint	Kritikus érték (tiszta illeszkedésvizsgálat)	Kritikus érték (becsléses illeszkedésvizsgálat)
$\alpha=0{,}9$	$z=0{,}35$	$z=0{,}10$
$\alpha=0{,}95$	$z=0{,}46$	$z=0{,}13$
$\alpha=0{,}99$	$z=0{,}75$	$z=0{,}18$

2.2. A Wilk—Shapiro-próba

Külön figyelmet érdemel az így kapott eljárások közül az, amelyet lényegében az 1.3.d. pontban δ_n-nel jelölt statisztikából kapunk. Ez a módszer WILKtől és SHAPIROtól (vö.: [23]) származik, akik később a [24] dolgozatukban *Monte-Carlo*-módszerrel összehasonlították az eljárásukat más normalitásvizsgálati próbákkal és azt kapták, hogy az egyértelműen jobb minden más eljárásnál. Igaz ugyan, hogy csak a próbák erejét vizsgálták, a szükséges gépidőt nem, és az is elképzelhető, hogy ez az eredmény csak az általuk figyelembe vett alternatívák speciális választásától függ, mégis elfogadható volna a módszerük, ha kiterjeszthető volna a többmintás feladatra. A szerzők ezt meg is kísérelték a [29] dolgozatukban, úgy gondoljuk azonban, ez a kérdés a módszer esetleges alkalmazása előtt további vizsgálatot igényelne.

A *Wilk—Shapiro*-statisztika formálisan a következő:

$$W_n = \frac{1}{s^2}\left(\sum_{i=1}^{n} a_{ni}\xi_i^*\right)^2,$$

ahol az a_{ni} együtthatókat a szerzők $2 \leqq n \leqq 50$ mellett táblázatosan megadják. Ugyanezekre a mintaelemszámokra adják meg a megfelelő szignifikancia-határokat is.

2.3. Sűrűségfüggvények: χ^2 próba

Ha valaki nagyon keveset tud a normalitás-vizsgálattal kapcsolatban felmerülő nehézségekről, a „Hogyan teszteljük a normalitást" kérdésre valószínűleg a következő választ adja: „A *Kolmogorov—Szmirnov*-próba ugyanúgy használható, csak a paramétereket a becsült értékükkel kell helyettesíteni, a χ^2 próbában pedig 2-vel kell csökkenteni a szabadságfokot." Láttuk már, hogy ennek a válasznak az első fele mennyire nem igaz, most arról lesz szó, hogy a második fele sem igaz. Pedig ennek látszatra még teoretikus alapja is van: a maximum-likelihood becsléseknél megadható a χ^2-próba megfelelő általánosítása, és itt azt a körülményt, hogy néhány paramétert a mintából becsülünk, valóban a szabadságfok csökkentésével vehetjük figyelembe. A baj csak az, hogy ha valamely beosztáshoz tartozó f_n empirikus sűrűségfüggvényt veszünk alapul, a paraméterek maximum-likelihood becslése már nem $\bar{\xi}$ és s lesz, és a helyes becsléseket igen nehéz numerikusan meghatározni. KOLLER [11] megad ugyan könnyebben kiszámolható, a maximum likelihood becslésekkel ekvivalens becsléseket, ezek azonban jelen formájukban nem használhatók, hiszen a szerző fix beosztást vesz alapul, pedig a normálitás-vizsgálatnál a beosztást is célszerű volna a minta függvényében megválasztani.

Több szempontból előnyösebbnek látszik a másik két sűrűségfüggvény-becslési eljárásból normalitás-vizsgálati módszert kifejleszteni. A *Neyman—Barton*-próbának $k=4$ mellett a legősibb normalitás-vizsgálati módszer felel meg: a ferdeség és a csúcsosság ellenőrzése. A mintából kiszámolható

$$\beta_n = \frac{1}{s^3}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^3$$

empirikus ferdeség, és

$$\gamma_n = \frac{1}{s^4}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^4-3$$

empirikus csúcsosság értéke ugyanis normális eloszlás esetén várható közel van 0-hoz. Ennek az ellenőrzésére szolgáló szignifikancia-határok megtalálhatók például a [17] táblázatban. Itt említjük még meg az ún. *Geary*-indexet:

$$\alpha_n = \frac{1}{s}\sum_{i=1}^{n}|\xi_i-\bar{\xi}|,$$

amelyre szintén található [17]-ben táblázat.

Ezekkel a táblázatokkal persze csak külön-külön ellenőrizhetjük az egyes momentumokat, az együttes ellenőrzés csak egy megfelelő ortogonális függvényrendszer alapján volna lehetséges. Ilyet azonban az irodalomban nem találtunk, előállítása nem látszik könnyű feladatnak.

Lényegében a *Wilk—Shapiro*-próbához hasonló kvadratikus kifejezést kapunk, ha a *Parzen*-féle sűrűségfüggvényre becsléses *Cramér—Mises*-próbát kívánunk alkalmazni. Ez az eljárás nincs kidolgozva, bár a *Wilk—Shapiro*-módszerrel szemben kétségtelen előnye volna, hogy alkalmazása esetén nem kell a rendezett mintát előállítani.

2.4. Összefoglalás

A becsléses normalitás-vizsgálati módszerekről összefoglalva elmondhatjuk, hogy a lehetőségek száma igen nagy, de a megvalósításuk, különösen a minket érdeklő többmintás esetben nem látszik könnyűnek. A feladattal foglalkozó statisztikusok túlnyomó többsége ebben az irányban kereste a megoldást. Ez sok szempontból érthető is. Olyan mérőszámot kívántak előállítani, amely valamilyen értelemben optimálisan méri a minta normalitását, és így a többmintás esetben jó képet adna arról, milyen mértékben távolodik el a minta egyik vagy másik része a normális eloszlástól. A nehézségek, amikkel szembetalálták magukat, a következők voltak:

— nincs egyértelműen megadható optimalitási kritérium;

— a javasolt statisztikák értékének a kokrét kiszámolása sokszor igen bonyolult;

— a javasolt statisztikák eloszlása és határeloszlása nehezen meghatározható.

Nem tartjuk kizártnak, hogy a megoldást ebben az irányban meg lehet találni. Sok szempont szól amellett, hogy az ismertetett törekvések, vizsgálatok helyes irányban haladnak, és a szükséges gépi tapsztalatok megszerzése után sikerül majd a számítógépen gyorsan realizálható, statisztikai szempontból egzakt módszert megtalálni. A jelen körülmények között azonban elsősorban a következő fejezetben ismertetésre kerülő transzformációs módszer realizálását javasoljuk.

A teljesség kedvéért mégis megjegyezzük, hogy elvileg semmi akadálya nincs annak, hogy a fenti módszerek bármelyikét a többmintás feladat megoldására használjuk. Legyen ugyanis egy tetszőleges módszer alapja az S_n statisztika, amelynek a

$$H_n(x) = P(S_n < x)$$

eloszlása nem függ a μ, σ paraméterektől, ha a minta normális eloszlású, és amelynek nagy értékei szólnak a normalitás ellen. Definiáljuk a ϱ változót a

$$\Phi(\varrho) = H_n(S_n)$$

összefüggéssel, ahol Φ a standard normális eloszlás. Ez a ϱ — mint az könnyen látható — normális eloszlású, és a nagy értékei szólnak a hipotézis teljesülése ellen. Ha tehát a többmintás feladat esetében mindegyik minta-egységből meghatározzuk ennek a változónak az értékét, és a kapott számokat összegezzük, egyszerű u-próbával ellenőrizhetjük a minta-egységek együttes normalitását.

3. NORMALITÁSVIZSGÁLAT TRANSZFORMÁCIÓVAL

Ugyanazt a hipotézist vizsgáljuk ugyanabban a modellben, mint a 2. fejezetben, csak a módszer lesz más: alkalmas transzformációval fogunk megszabadulni a zavaró paraméterektől.

DEFINÍCIÓ. Az n-változós $\psi_i(x_1, x_2, \ldots, x_n)$ $(i=1, 2, \ldots, k)$ függvényeket megengedhető transzformációnak (MT) nevezzük, ha az

$$\eta_i = \psi_i(\xi_1, \xi_2, \ldots, \xi_n)$$

valószínűségi változók a vizsgált modellben függetlenek, és standard normális eloszlásúak. Az $(n+t)$-változós $\tilde{\psi}_i(x_1, x_2, \ldots, x_n; Z_1, \ldots, Z_t)$ $(i=1, 2, k)$ függvényeket megengedhető randomizált transzformációknak nevezzük (MRT), ha az

$$\tilde{\eta}_i = \tilde{\psi}_i(\xi_1, \xi_2, \ldots, \xi_n; \zeta_1, \zeta_2, \ldots, \zeta_k) \qquad (i = 1, 2, \ldots, k)$$

valószínűségi változók a vizsgált modellben alkalmasan választott, a ξ_i-ktől független ζ_i változók mellett függetlenek, és standard normális eloszlásúak.

Definíciónkkal kapcsolatban megjegyezzük, hogy nincs különösebb jelentősége benne annak a feltételnek, hogy az η_i-k eloszlása éppen standard normális legyen: lehetne az η_i-k közös eloszlása bármilyen ismert eloszlás, hiszen az MT-k és az MRT-k szerepe az, hogy a normalitás-vizsgálatot visszavezesse a tiszta illeszkedés-vizsgálatra, ott viszont — mint láttuk — nincs semmilyen jelentősége, hogy a vizsgált, ellenőrizni kívánt F_0 eloszlás milyen alakú, hiszen az ellenőrzést mindig eloszlásmentes módszerrel végezzük.

Az alábbiakban felsorolunk néhány MT-t és MRT-t. Ezek helyes voltának a bizonyításában szükségünk lesz a normális eloszlásnak a következő tulajdonságára.

TÉTEL. *Ha a $\xi_1, \xi_2, \ldots, \xi_n$ változók függetlenek, normális eloszlásúak μ, σ paraméterekkel, akkor az*

$$\alpha = \frac{\bar{\xi} - \mu}{\sigma}\sqrt{n}, \quad \beta = \frac{s^2}{\sigma^2}(n-1)$$

változók függetlenek egymástól, és a

$$\gamma_i = \frac{\xi_i - \bar{\xi}}{s} \qquad (i = 1, 2, \ldots, n)$$

koordinátájú $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ vektorváltozótól; és α, β, γ eloszlása nem függ a μ, σ paraméterektől. (Nevezetesen α standard normális eloszlású β pedig χ^2 eloszlású $(n-1)$ szabadságfokkal.)

Célszerűbb a transzformációkat két lépésben megadni: MTA-nak vagy MRTA-nak nevezzük a transzformációt, ha független μ, σ paraméterű normálisakhoz független, 0, σ paraméterű normálisakat rendel, és MTB-nek, vagy MRTB-nek nevezzük, ha független, 0, σ paraméterű normálisakhoz független, 0, 1 paraméterű normálisakat rendel. (Ha egy MRTA után MRTB-t hajtunk végre — e két transzformáció együtt nyilván MRT.)

Az MTA-kat a lineáris függvények között keressük: legyen

$$\psi_i(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} a_{ij}x_j \qquad (i = 1, 2, \ldots, k).$$

Ez akkor és csakis akkor MTA, ha

$$\sum_{j=1}^{n} a_{ij} = 0 \qquad (i = 1, 2, \ldots, k)$$

$$\sum_{j=1}^{n} a_{ij}a_{\nu j} = \delta_{i\nu} \qquad (i, \nu = 1, 2, \ldots, k).$$

Emiatt, ha $\psi = (\psi_1, \psi_2, \ldots, \psi_k)$ lineáris MTA, akkor $k \leqq n-1$, és ha $k = n-1$, az $A = a_{ij}$ mátrix az $a_{nj} = 1/\sqrt{n}$ kiegészítő sorral ortonormált négyzetes mátrix: $AA' = I$.

Ezzel természetesen A nincs egyértelműen meghatározva. Ilyen például az ún. *Helmert*-mátrix:

$$A = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 & \dots & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & 0 & \dots & 0 \\ 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & -3/\sqrt{12} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & \dots & -(n-1)/\sqrt{n(n-1)} \\ 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & \dots & 1/\sqrt{n} \end{pmatrix}.$$

Megfelelő A mátrixot kapunk, ha az a, b, c paramétereket úgy határozzuk meg, hogy

$$\psi_i = ax_i + b\bar{x} + cx_n \qquad (i = 1, 2, \dots, n-1)$$

MTA legyen $\left(\text{ahol } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$. A fentiek szerint ekkor

$$a+b+c = 0,$$

$$a^2 + \frac{b^2}{n} + c^2 + \frac{2ab}{n} + \frac{2bc}{n} = 1,$$

$$\frac{2ab}{n} + \frac{b^2}{n} + \frac{2bc}{n} + c^2 = 0,$$

amiből $a=1$, $b = -\frac{\sqrt{n}}{\sqrt{n}-1}$, $c = \frac{1}{\sqrt{n}-1}$ következik:

$$\psi_i = x_i - \tilde{x}, \quad \text{ahol} \quad \tilde{x} = \frac{\sqrt{n}\,\bar{x} - x_n}{\sqrt{n}-1}.$$

Ez a transzformáció szemléletesen úgy mondható el, hogy az x_i mintaelemekből levonjuk — lényegében — a mintaátlagot, hogy a μ paramétertől megszabaduljunk. Hogy mégsem pontosan az $\bar{x}$ mintaátlagot vonjuk le, annak az az oka, hogy úgy a transzformált elemek nem volnának függetlenek: a legutolsó mintaelemet tehát arra használjuk fel, hogy a vele módosított átlagot már levonhassuk, vagyis hogy az $x_i - \bar{x}$ különbségek függetlenek legyenek. Ezzel elvesztünk egy mintaelemet — de hát úgyis csak $(n-1)$ transzformált mintaelemet kaphatunk. Ez a transzformáció SARKADItól [21] származik.

Nem vesztünk el mintaelemet, és a transzformáció is szimmetrikussá válik, ha felhasználunk egy normális eloszlású, 0, σ paraméterű, a mintától független ζ valószínűségi változót. Az előbb kimondott tétel állításai alapján ugyanis könnyen belátható, hogy az

$$\eta_i = \xi_i - \bar{\xi} + \frac{1}{\sqrt{n}}\zeta \qquad (i = 1, 2, \dots, n)$$

változók függetlenek, és normális eloszlásúak $0, \sigma$ paraméterekkel. Ez a transzformáció DURBINtől származik (vö. [6]). Érdemes megjegyezni, hogy ebből a transzformációból visszakaphatjuk az előző transzformációt: ha ugyanis csak az első $(n-1)$ mintaelemet használjuk, a

$$\hat{\xi} = \frac{1}{n-1} \sum_{i=1}^{n-1} \xi_i$$

mintaátlag független a $\{\xi_i - \hat{\xi}\}_{i=1}^{n-1}$ vektortól, tehát

$$\zeta = \frac{\sqrt{n-1}}{\sqrt{n}} (\xi_n - \hat{\xi})$$

is független a $\{\xi_i - \hat{\xi}\}_{i=1}^{n-1}$ vektortól, és normális eloszlású $0, \sigma$ paraméterekkel, és ezek szerint

$$\eta_i = \xi_i - \hat{\xi} + \frac{1}{\sqrt{n-1}} \zeta = \xi_i - \left(1 + \frac{1}{\sqrt{n}}\right) \hat{\xi} + \frac{1}{\sqrt{n}} \xi_n = \xi_i - \frac{\sqrt{n}\,\hat{\xi} - \xi_n}{\sqrt{n} - 1},$$

ami valóban azonos az előző transzformációval.

SARKADI megmutatta, hogy az általa használt transzformáció esetében a transzformált, és az eredeti mintaelemek közti korrelációs együttható maximális az alábbi értelemben.

LEMMA. *Ha* $\bar{\Psi}_i$ $(i = 1, 2, \dots, n-1)$ *tetszőleges MTA, és* ψ_i *a Sarkadi-féle transzformáció, akkor*

$$\min_{1 \leq i \leq n-1} R(\xi_i, \bar{\eta}_i) \leqq \min_{1 \leq i \leq n-1} R(\xi_i, \eta_i),$$

ahol $R(\xi, \eta)$ *a* ξ, η *változók korrelációs együtthatóját jelöli, és* $\eta_i = \Psi_i(\xi_1, \xi_2, \dots, \xi_n)$, $\bar{\eta}_i = \bar{\Psi}_i(\xi_1, \xi_2, \dots, \xi_n)$.

Rátérünk az MTB-k és MRTB-k konstrukciójára. A fenti tétel szerint, ha ξ_1 és ξ_2 normális eloszlásúak $(0, \sigma)$ paraméterekkel, akkor az

$$\eta = \frac{1}{\pi} \operatorname{arc\,tg} \frac{\xi_1}{\xi_2}, \quad \zeta = \frac{1}{\sigma^2} [\xi_1^2 + \xi_2^2]$$

változók függetlenek, η egyenletes eloszlású a $(0, 1)$ szakaszon, ζ pedig exponenciális eloszlású $\lambda = 1$ paraméterrel. Ha tehát n páros, a mintaelemekből képezett párokból $n/2$ ilyen η-t és $n/2$ ilyen ζ-t kapunk. Az η-kból alkalmas transzformációval normális eloszlású változókat állíthatunk elő, a ζ-kra pedig alkalmazhatjuk STÖRMER [26] módszerét.

DURBIN [6] a következő módszert javasolta: legyen most (mivel $\mu = 0$)

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \xi_i^2,$$

akkor a

$$\left(\frac{\xi_1}{s}, \frac{\xi_2}{s}, \dots, \frac{\xi_n}{s}\right)$$

hányadosok függetlenek s-től, és eloszlásuk nem függ σ-tól. Ha tehát ζ a ξ_i-któl független χ-eloszlású valószínűségi változó n szabadságfokkal, akkor az

$$\eta_i = \frac{\xi_i}{s}\zeta \qquad (i = 1, 2, \ldots, n)$$

változók függetlenek, és standard normális eloszlásúak.

Ebből az MRTB-ből ugyanúgy kaphatjuk meg a *Sarkadi*-féle MTB-t, mint ahogy az előbb előállítottuk a *Sarkadi*-féle MTA-t a *Durbin*-féle MRTA-ból (megjegyezzük, hogy a visszavezetés is SARKADItól származik). Legyen

$$z^2 = \frac{1}{n-1}\sum_{i=1}^{n-1} \xi_i^2,$$

akkor a

$$\left(\frac{\xi_1}{z}, \frac{\xi_2}{z}, \ldots, \frac{\xi_{n-1}}{z}\right)$$

hányadosok függetlenek Z-től, és eloszlásuk nem függ σ-tól. Emiatt ezek a hányadosok függetlenek a ξ_n/Z hányadostól is, és ξ_n/Z eloszlása $(n-1)$-szabadságfokú t-eloszlás. Ha tehát G_{n-1}-gyel jelöljük a $|\xi_n|/Z$ hányados eloszlásfüggvényét:

$$G_{n-1}(x) = P\left(\frac{1}{z}|\xi_n| < z\right) = 2F_{n-1}(x) - 1,$$

(ahol $F_t(x)$ az $(n-1)$ szabadságfokú t-eloszlást jelöli), és H_{n-1}-gyel az $(n-1)$ szabadságfokú χ-eloszlást, akkor a

$$G_{n-1}\left(\frac{1}{z}|\xi_n|\right) = H_{n-1}(\zeta)$$

összefüggéssel definiált változó χ eloszlású $(n-1)$ szabadságfokkal. Így ez a ζ alkalmazható a *Durbin-féle* transzformációhoz: az

$$\eta_i = \frac{\xi}{z}\zeta \qquad (i = 1, 2, \ldots, n-1)$$

változók független standard normálisak lesznek.

A fenti *Sarkadi*-féle MTA és MTB egymás után való alkalmazásából kapjuk a *Sarkadi-féle* MT-t. Legyen tehát $(\xi_1, \xi_2, \ldots, \xi_n)$ az eredeti minta, ebből először

$$\bar{\xi}_1 = \frac{1}{n-2}\sum_{i=1}^{n-2} \xi_i, \quad \bar{\xi}_2 = \frac{1}{2}\sum_{i=n-1}^{n} \xi_i, \quad \bar{\bar{\xi}} = (\bar{\xi}_1 - \bar{\xi}_2)\sqrt{\frac{2}{n}}$$

átlagokat és az

$$\eta_i = \xi_i - \bar{\xi}_1 + \bar{\bar{\xi}} \qquad (i = 1, 2, \ldots, n-2)$$

transzformált mintaelemeket határozzuk meg. Ezektől eltérő módon határozzuk meg a transzformált minta utolsó elemét, az

$$\eta_{n-1} = \frac{\xi_n - \xi_{n-1}}{\sqrt{2}}$$

számot. Az $\eta_1, \eta_2, \ldots, \eta_{n-1}$ mintára ezután a *Sarkadi*-féle MTB-t alkalmazzuk: legyen

$$z^2 = \frac{1}{n-2} \sum_{i=1}^{n-2} \eta_i^2,$$

és legyen u a

$$H_{n-2}(u) = 1 - G_{n-2}\left(\frac{1}{z}|\eta_{n-1}|\right)$$

egyenlettel definiált szám, ahol H_f az f szabadságfokú χ-eloszlás, G_f az f szabadságfokú t-eloszlás abszolút értékének az eloszlásfüggvénye. Akkor a

$$\zeta_i = \frac{\eta_i}{z} u \qquad (i = 1, 2, \ldots, n-2)$$

változók függetlenek és standard normálisak.

A *Durbin*-féle MRT nem kapható meg közvetlenül a fenti MRTA és MRTB egymás utáni alkalmazásából, hiszen a *Durbin*-féle MRTA-hoz σ szórású normális eloszlású véletlen szám kellene, viszont σ ismeretlen. Visszatérünk az eredeti tételhez, annak alapján könnyen látható, hogy ha ζ_1 a mintától független, standard normális eloszlású véletlen szám, és ζ_2 a mintától, és ζ_1-től független, $(n-1)$ szabadságfokú χ^2-eloszlású véletlen szám, akkor az

$$\eta_i = \frac{\xi_i - \bar{\xi}}{s} \sqrt{\frac{\zeta_2}{n-1}} + \frac{1}{\sqrt{n}} \zeta_1 \qquad (i = 1, \ldots, n)$$

transzformált elemek független standard normálisak.

Foglalkozzunk az eljárás hatékonyságával. Arra vagyunk kíváncsiak mi történik akkor, amikor a hipotézis nem teljesül. E célból fogalmazzuk meg az eljárást általánosabban.

Az $x_1, x_2, \ldots, x_n$ mintáról el akarjuk dönteni, hogy az elemei $F(x, \vartheta)$ eloszlásúak-e, ϑ ismeretlen paraméter. E célból alkalmazzunk egy $y_1 = T_1(x_1, \ldots, x_n)$ $y_2 = T_2(x_1, \ldots, x_n), \ldots, y_m = T_m(x_1, \ldots, x_n)$ transzformációt, mely az $x_1, x_2, \ldots, x_n$ $F(x, \vartheta)$ eloszlású független mintát valamely $y_1, y_2, \ldots, y_m$ ismert $G(x)$ eloszlású független vagy esetleg rendezett mintába viszi át. Ezután például *Kolmogorov—Szmirnov*-próbával ellenőrizzük, hogy $y_1, \ldots, y_m$ valóban $G(x)$ eloszlású minta-e Itt azonban a következő kérdés merül fel:

Amikor az $x_1, x_2, \ldots, x_n$ minta nem $F(x, \vartheta)$ eloszlású, tehát a hipotézist el kellene utasítanunk, meg fogjuk-e ezt tenni nagy valószínűséggel? A következő példa azt mutatja, hogy ily módon a rossz eloszlású mintát nagy minta esetén is nagy valószínűséggel jónak fogadhatjuk el.

Legyen $x_1, x_2, \ldots, x_n$ exponenciális eloszlású minta (eloszlásfüggvénye $1 - e^{-\lambda x}$, ha $x < 0$) ismeretlen λ paraméterrel. Ismeretes, hogy ekkor az $y_i = (x_1 + \ldots + x_i)/ /(x_1 + \ldots + x_n)$ valószínűségi változók a $[0, 1]$ intervallumban egyenletes eloszlású rendezett mintát alkotnak. Ha azonban az $x_1, x_2, \ldots, x_n$ elfajult eloszlású, azaz $P(x_i = a) = 1$, $i = 1, \ldots, n$, ahol a rögzített konstans, akkor $y_i = \frac{i}{n}$ $(i = 1, \ldots, n)$ 1 valószínűséggel, és a *Kolmogorov—Szmirnov*-próbát alkalmazva a $x_1, x_2, \ldots, x_n$ mintát még kis n esetén is elfogadjuk exponenciális eloszlásúnak, noha nem az. E hibát elkerülendő bevezetjük a konzisztens transzformáció fogalmát, és a továbbiakban konzisztens transzformációkkal foglalkozunk.

DEFINÍCIÓ. Azt mondjuk, hogy egy MT, vagy MRT konzisztens, ha az $\eta_i(i=1, 2, \dots, k)$ változók G_n empirikus eloszlásfüggvénye $n\to\infty$ mellett tart a vizsgált modellben (ha tehát a ξ_i mintaelemek függetlenek, és egyforma eloszlásúak) a

$$G(x) = F\left(\frac{x-\mu}{\sigma}\right)$$

eloszlásfüggvényhez, ahol F a ξ_i-k közös eloszlásfüggvénye, és $\mu=E\xi_i$, $\sigma=D\xi_i$.

Megjegyezzük, hogy a fenti definíció pontatlan, nem tartalmazza, milyen értelemben kell G_n-nek G-hez tartania. Erős vagy gyenge konzisztenciáról beszélünk aszerint, hogy a

$$\sup_{-\infty<x<\infty} |G_n(x)-G(x)|$$

valószínűségi változó 1 valószínűséggel, vagy sztochasztikusan tart-e 0-hoz.

Ez a tulajdonság, amely a *Sarkadi-* és a *Durbin*-féle transzformációt kiemeli a többi közül, hiszen amint azt STÖRMER [27] lényegében megmutatta, mind a két transzformáció konzisztens.

A kérdés tehát az, hogy e két transzformáció közül melyik jobb. SARKADI a fenti lemma állításához hasonló optimális tulajdonságot bizonyított az általa javasolt MTB, és MT transzformációkról is. Mi viszont az általa használt optimalitási kritériumot nem tartjuk elég meggyőzőnek, és a *Sarkadi*-módszer lényeges hiányosságának tartjuk, hogy nem szimmetrikus. Azt mondhatjuk tehát, hogy pusztán elméleti szempontból nem tudunk különbséget tenni a két módszer között, így a számítástechnikai szempontot kell alapul vennünk. Ez viszont egyértelműen DURBIN módszere mellett szól, ezért mi ezt a módszert javasoljuk.

IRODALOM

[1] BARTON, D. E.: Neyman's ψ_k^2 test of goodness of fit when the null hypothesis is composite. *Skand. Aktuarietidskr.* **38** (1956) 216—245.

[2] BOLSEV, L. N.—SZMIRNOV, N. W.: *Matematikai statisztikai táblázatok* (oroszul), Nauka, Moszkva 1965.

[3] COCHRAN, W. G.: The χ^2 test of goodness of fit. *Ann. Math. Stat.* **23** (1952) 315—345.

[4] CSÁKI, E.—TUSNÁDY, G.: On the number of intersections and the ballot theorem. *Periodica Math. Hung.* **2** (1972) 5—13.

[5] DARLING, D. A.: The Cramér—Smirnov test in the parametric case. *Ann. Math. Stat.* **26** (1955). 1—20.

[6] DURBIN, S.: Some methods of constructing exact tests. *Biometrika* **48** (1961) 41—55.

[7] GEARY, R. C.: Testing for normality. *Biometrika* **34** (1947) 209—242.

[8] KAC, M.—KIEFER, J.—WOLFOWITZ, J.: On tests of normality and other tests of goodness of fit based on distance methods. *Ann. Math. Stat.* **26** (1955) 189—211.

[9] KAC, M.—SIEGERT, A. J. F.: An explicit representation of a stationary Gaussian proces. *Ann. Math. Stat.* **18** (1947) 438—442.

[10] KENDALL, M. G.—STUART, A.: *The advanced theory of statistics.* Vol. 2. Griffin, London 1967.

[11] KOLLER, D.: Prüfung der Normalität einer Verteilung. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **2** (1964) 147—166.

[12] MANN, H. B.—WALD, A.: On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math. Stat.* **13** (1942) 306—317.

[13] NEYMAN, J.: „Smooth test" for goodness of fit. *Skand. Aktuarietidskr.* **20** (1937) 149—199.

[14] Owen, D. B.: *Handbook of statistical tables.* Addison—Wesley, 1962.
[15] Parzen, E.: On estimation of a probability density function and mode. *Ann. Math. Stat.* **33** (1962) 1065—1076.
[16] Pearson, E. S.: A further development of tests for normality. *Biometrika* **22** (1930) 239—249.
[17] Pearson, E. S.—Hartley, H. O.: *Biometrika tables for statisticians.* Vol. I. Cambridge University Press 1954.
[18] Révész, P.: On empirical density functions. *Periodica Math. Hung.* **2** (1972) 85—110.
[19] Rosenblatt, M.: Curve estimates. *Ann. Math. Stat.* **42** (1971) 1815—1842.
[20] Sahler, W.: A survey on distribution-free statistics based on distances between distribution functions. *Metrika* **13** (1968) 149—169.
[21] Sarkadi, K.: On testing for normality. *Proc. of the Fifth Berkeley Symp. on Math. Stat. and Prob. I.* 1967. 373—387.
[22] Sarkadi, K.: On the exact distribution of statistics of Kolmogorov—Smirnov type. *Periodica Math. Hung.* **3.** (1973) 9—12.
[23] Shapiro, S. S.—Wilk, M. B.: An analysis of variance test for normality (complete samples) *Biometrika* **52** (1965) 591—611.
[24] Shapiro, S. S.—Wilk, M. B.—Chen, H. J.: A comparative study of various tests for normality. *Jour. Amer. Stat. Assoc.* **63** (1968) 1343—1372.
[25] Steck, G. P.: Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. *Ann. Math. Stat.* **42** (1971) 1—11.
[26] Störmer, H.: Ein Test zum Erkennen von Exponentialverteilungen. *Metrika* **5** (1962) 128—137.
[27] Störmer, H.: Ein Test zum Erkennen von Normalverteilungen. *Zeitschrift für Wahrscheinlichkeitstehorie und verwandte Gebiete* **2** (1964) 420—428.
[28] Vincze, I.: On some questions connected with two-sample tests of Smirnov-type. *Proc. of the Fifth Berkeley Symp. on Math. Stat. and Prob. I.* (1967) 657—666.
[29] Wilk, M. B.—Shapiro, S. S.: The joint assessment of normality of several independent samples. *Technometrics* **10** (1968) 825—840.
[30] Pearson, E. S. Hartley, H. O.: *Biometrika Tables for statisticians*, Vol. II. Cambridge, The University Press 1972.

TESTING FOR NORMALITY

by

P. Major and G. Tusnády

Summary

The first part of the paper investigates the goodness of fit testing a simple hypothesis. Then different methods for testing normality with unknown parameters are considered. One of these is to estimate the parameters, another method is to eliminate the unknown parameters by transformation. The goodness of these methods is investigated.

(Beérkezett: 1973. I. 25.)