A többdimenziós normalitás tesztelése

SZAKDOLGOZAT

Írta: Monori Lilla

Matematika BSc
Matematikai elemző szakirány

Témavezető:
Dr. Zempléni András
Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2024

NYILATKOZAT

Név: Monori Lilla

ELTE Természettudományi Kar, szak: Matematika alapszak

NEPTUN azonosító: F3BSQJ

Szakdolgozat címe:
A többdimenziós normalitás tesztelése

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2024.06.04.

a hallgató aláírása

Köszönetnyilvánítás

Szeretnék köszönetet mondani a témavezetőmnek, Dr. Zempléni Andrásnak a szakmai segítségéért, a türelméért és a támogatásáért, illetve hogy egy ilyen érdekes téma mélyebb megismerését elősegítette számomra.

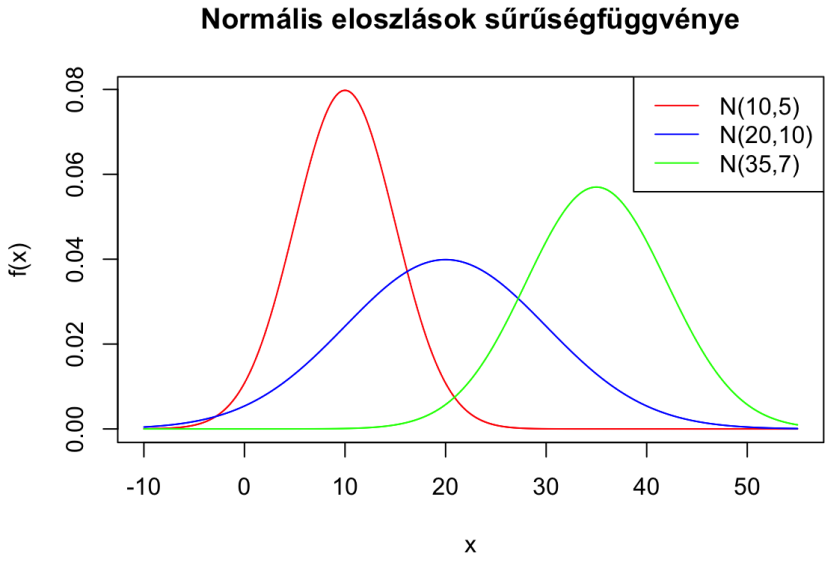
Köszönöm a tanároknak, akiktől volt lehetőségem tanulni az elmúlt 3 évben, mert újra és újra rávilágítottak, hogy miért is szeretem a matematikát minden szépségével és nehézségével együtt.

Továbbá szeretném megköszönni a barátaimnak és a családomnak, akik végig támogattak és mellettem voltak ebben az időszakban.

Tartalomjegyzék

1. Bevezetés

A normális eloszlás a valószínűségszámítás és matematikai statisztika egyik legfontosabb és legtöbbet alkalmazott eloszlása. Ismert még Gauss-eloszlás néven is, mert feltételezhetően Carl Friedrich Gauss volt, aki 1809-ben először használta és nevezte így a normális eloszlást. Az eloszlás sűrűségfüggvényét Gauss-görbének vagy alakjából fakadóan haranggörbének szoktuk hívni. A görbe szimmetriatengelye a várható érték, szélességét pedig a szórás határozza meg. Ennek három példája látható az 1. ábrán.



1. ábra

A normális eloszlással együtt bevezetem majd a standard normális eloszlás fogalmát is, mert többdimenzióban ennek segítségével definiálom a normális eloszlást.

Az biztos, hogy a normális eloszlás az egyik legfontosabb abszolút folytonos eloszlás, mert a természetbeli, közgazdaságtani, szociológiai folyamatok jelentős része közelíthető normális eloszlással, illetve sok statisztikai számítás alapfeltétele a normalitás. Emellett az egyre fontosabb adattudományban és adatbányászatban is nagy jelentősége van annak, hogy az adathalmaz vajon normális eloszlást követ-e.
A többdimenziós normalitás fontos még a faktoranalízis területén is, ahol a maximum likelihood becslés explicit feltétele, hogy független, többdimenziós normális eloszlású legyen a minta. A matematika egy másik területe, ahol alkalmazzuk a többdimenziós normalitást a diszkriminancia analízis.

Ez a klasszifikáció módszer először többdimenziós normális eloszlású adatokra lett kifejlesztve. Aktuálisan pedig a lineáris diszkriminancia analízis azokra az adatokra használható leghatékonyabban, amik minden osztályon belül többdimenziós normális eloszlást követnek.
Ez csak néhány a sokszínű alkalmazási területek közül, így nem véletlen, hogy a mai napig foglalkoznak vele és hogy a tesztelésére sok numerikus és grafikus módszer létezik.

A szakdolgozatom második fejezetében bevezetem a témához szükséges alapfogalmakat, ideértve az egydimenziós normalitást, az illeszkedésvizsgálatot és a bootstrap módszert. A harmadik fejezet a többdimenziós normális eloszlással foglalkozik. Összefoglalom, hogy miért is ilyen fontos, majd bemutatom a d-dimenziós normalitást, illetve egy fontos speciális esetét, a 2-dimenziós változatát. A negyedik fejezet a többdimenziós normalitásvizsgálat két grafikus módszerét tárgyalja, nevezetesen a Q-Q plotot és a Holgersson-féle grafikus tesztet. Az ötödik fejezetben foglalkozom a normalitástesztekkel, konkrétan a Mardia-teszttel, a Henze-Zirkler teszttel, a Székely-Rizzo teszttel és a Doornik-Hansen teszttel. A hatodik fejezetben pedig ezt a négy tesztet hasonlítom össze generált adatok és valós adatok segítségével is. Az utolsó fejezetben pedig összefoglalom a tesztek összehasonlításából levonható konklúziókat.

2. Alapfogalmak

2.1. Az egydimenziós normális eloszlás

Először definiálom a normális és a standard normális eloszlást:

2.1. Definíció (Normális eloszlás)**.** Egy X valószínűségi változó normális eloszlású, ha sűrűségfüggvénye a teljes valós számhalmazon értelmezett alábbi függvény:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ ahol } \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+.$$

Jelölése: $X \sim N(\mu, \sigma)$, ahol μ a várható értéket, σ pedig a szórást jelöli.

2.2. Definíció (Standard normális eloszlás)**.** Egy valószínűségi változó standard normális eloszlású, ha normális eloszlású és $\mu = 0$, $\sigma = 1$, azaz a sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Az egydimenziós normális eloszlás fontos tulajdonsága, hogy szimmetrikus a várható értékre nézve és ekörül helyezkedik el a legtöbb megfigyelt érték, a kiugró értékek pedig kifejezetten ritkák.

A statisztikában kiemelten fontos, mert sok természetesen előforduló jelenséget nagyon jó közelítéssel ír le. Gondolhatunk itt például emberek magasságára, cipőméretére vagy akár vérnyomására is.

Normális eloszlású valószínűségi változókat standardizálhatunk is, és az eredmény is normális eloszlású lesz. A leggyakoribb standardizációs eljárás a Z-érték kiszámolása. Az eljárás során a változókat lineáris transzformációval úgy alakítjuk át, hogy 0 várható értékű, 1 szórású valószínűségi változót kapjunk. Ennek segítségével pedig különböző mértékegységű vagy nagyon eltérő skálázású mintákat is össze tudunk hasonlítani. Emellett rálátást nyerhetünk arra, hogy adott megfigyelés mennyire tér el a várható értéktől vagy összehasonlíthatunk különböző várható értékű és szórású normális eloszlásból származó megfigyeléseket, ami egyébként nehéz feladat lenne.

$X \sim N(\mu, \sigma)$ standardizálása:

$$Z = \frac{X - \mu}{\sigma}$$

Fontos még megemlíteni a centrális határeloszlás-tételt, ami tulajdonképpen az elméleti hátterét mondja ki a normális eloszlás fontosságának. A tétel állítása, hogy sok kis független, véletlen hatás összegződése közelítően normális eloszlású.

2.1. Tétel (Centrális határeloszlás-tétel)**.** *Legyenek $X_1, X_2, \ldots, X_n$ azonos eloszlású, független valószínűségi változók, továbbá tegyük fel, hogy $E(X_i) = \mu < \infty$ és $0 < D(X_i) = \sigma^2 < \infty$ léteznek. Legyen Z_n a következő módon definiálva:*

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma},$$

ekkor

$$\lim_{n \to \infty} P(Z_n \leq x) = \phi(x), \qquad x \in \mathbb{R},$$

ahol $\phi(x)$ a standard normális eloszlásfüggvény.

A tétel megfelelően nagy elemszámú minta esetén alkalmazható. Egzaktul szinte sosem teljesül, hiszen egy nem normális eloszlásból származó minta eléggé nagy elemszám esetén sem lesz pontosan normális eloszlású, csak közelítőleg. De nyilván egy normális eloszlást követő mintára bármekkora elemszám esetén teljesül a tétel. Azonban ha a meghatározott feltételek teljesülnek, bármilyen eloszlású minta esetén meghatározott maximális hibahatárhoz tudunk küszöböt mondani, ehhez használható például a Berry-Esséen tétel. A gyakorlatban talán legtöbbször alkalmazott alsó határ az $n \geq 30$ elemszám, de ez függ a minta eloszlásának alakjától.

A tétel jó tulajdonsága, hogy nem számít, hogy X_i valószínűségi változók milyen eloszlást követnek, a lényeg csupán, hogy független, azonos eloszlásúak és véges szórásúak legyenek. Sőt általánosítható gyengén összefüggő és nem pontosan azonos, csak azonos nagyságrendű összeadandókra is.

A centrális határeloszlás-tétel ebben a formában kimondva alapvetően az X_i valószínűségi változók átlagára, $\overline{X}$-ra állítja, hogy közel normális eloszlású. A gyakorlatban azonban úgy is alkalmazhatjuk, hogy definiáljuk S_n-t, mint X_i-k összegét, és S_n-re mondjuk ki a tételt, tehát a valószínűségi változók összegét közelítjük normális eloszlással.

2.2. Az illeszkedésvizsgálat

A statisztikai illeszkedésvizsálat egy olyan elemzési módszer, melynek segítségével azt vizsgálhatjuk, hogy egy adathalmaz mennyire követ egy adott elméleti eloszlást.

Kétféleképpen is végezhetünk illeszkedésvizsgálatot; grafikusan vagy statisztikai teszttel. A grafikus tesztek vizuálisan segítenek annak kiértékelésében, hogy az adathalmazunk hogyan viselkedik egy elméleti eloszláshoz képest. Azonban ezek segítségével csak egy intuitív választ kapunk, nem pedig egy konkrét számértéket, így a grafikus módszereket inkább kiegészítésként használjuk a statisztikai tesztek mellett.
Az egyik legtöbbet használt grafikus módszer a Q-Q plot, vagyis a Quantile-Quantile plot. Ez egy egyszerű és hatékony eszköz eloszlások vizuális értékeléséhez. Ábrázoljuk a feltételezett elméleti eloszlás kvantiliseit és az adathalmazunk kvantiliseit egy diagramon. Az elméleti és tapasztalati értékek összehasonlításával hasznos információt kaphatunk arról, hogy az adataink eloszlása hogyan viszonyul a feltételezett elméleti eloszláshoz. A legismertebb grafikai tesztek közé sorolható még a boxplot és a hisztogram, de ezeken kívül is sok módszer létezik.

Statisztikai teszteket is használhatunk illeszkedésvizsgálatkor. A hipotézisvizsgálat lényege, hogy felteszünk egy nullhipotézist, ami az alapállapot, és megfogalmazunk egy ellenhipotézist, ami pedig az a feltevés, amit bizonyítani szeretnénk és valamilyen értelemben szignifikáns eltérést fogalmaz meg a nullhipotézisünkkel szemben. Teszteléskor választanunk kell egy szignifikancia szintet, amit általában α-val jelölünk. Ezzel az elsőfajú hiba valószínűségét határozzuk meg, leggyakrabban $\alpha = 0{,}05$ értéknek választjuk. Ennek meghatározása fontos feladat, mert ez befolyásolja a teszt érzékenységét és megbízhatóságát. Ezután kiszámoljuk a próbastatisztikát és a hozzá tartozó p-értéket. Ha ez kisebb, mint az adott szignifikancia szint, akkor elutasítjuk a nullhipotézist, mert szignifikáns eltérést kaptunk a teszt alapján. Ha azonban a p-érték nagyobbb, mint α, akkor nem tudjuk elutasítani H_0-t. Valójában statisztikai bizonyítékot sosem kapunk arra, hogy a nullhipotézisünk igaz, ezekben az esetekben azt tudjuk mondani, hogy nincs szignifikáns eltérés.
Sok illeszkedést vizsgáló tesztet ismerünk, gondoljunk akár a χ^2-próbára, a Kolmogorov-Szmirnov tesztre, az Anderson-Darling tesztre, de talán a legismertebb normalitásvizsgálati módszer Shapiro-Wilk teszt, ezt be is mutatom röviden a következő pontban.

Az illeszkedésvizsgálat fontos szerepet játszik a statisztikai modellezésben és a hipotézisvizsgálatban. A jó illeszkedés azt jelzi, hogy az adatok megfelelően követik az elméleti eloszlást, ami növeli a modell megbízhatóságát és az eredmények érvényességét. Ellenkező esetben, ha az adatok jelentősen eltérnek az elméleti eloszlástól, az azt jelentheti, hogy az alkalmazott modell nem megfelelő adatainkra, és további vizsgálatokra van szükség.

2.2.1. Shapiro-Wilk teszt

Ahhoz, hogy többdimenziós normalitásnak a teszteléséről egyáltalán beszélni tudjunk, elengedhetetlen, hogy az egydimenziós normalitást tesztelni tudjuk.

Martin Wilk és Samuel Sanford Shapiro 1965-ben publikált cikke [1] egy merőben új módszert mutatott be a normalitás tesztelésére. Azt a tényt használták fel, hogy ha egy minta normális eloszlást követ, akkor a minta Q-Q plotján a mintaelemek és a megfelelő standard normális kvantilisek lineárisan helyezkednek el.

A Shapiro-Wilk egy regresszión alapuló eljárás. Kétségtelenül ez az egyik legtöbbet használt normalitásteszt, mert nem túlságosan elnyúló és ferde eloszlások esetén kiemelkedően nagy a próba ereje, de hosszan elnyúló eloszlások esetén a teljesítménye még mindig elfogadható erejű.

Legyen $\mathbf{X}' = (X_1^*, X_2^*, \ldots, X_n^*)$ standard normális, rendezett, n elemű minta $\mathbf{m}' = (m_1, m_2, \ldots, m_n) = 0'$ várható érték vektorral és $\mathrm{V} = (v_{ij})$ kovarianciamátrixszal. Ezek alapján tehát:

$$E(X_i^*) = m_i = 0 \qquad (i = 1, \ldots, n)$$

és

$$cov(X_i^*, X_j^*) = v_{ij} \qquad (i = 1, \ldots, n, j = 1, \ldots, n)$$

Legyen $\mathbf{Y}' = (Y_1^*, Y_2^*, \ldots, Y_n^*)$ rendezett, n elemű véletlen minta. Tudjuk, hogy ha Y_i-k normális eloszlásúak ismeretlen μ és σ paraméterekkel, akkor felírhatóak $Y_i^* = \mu + \sigma X_i^* \quad (i = 1, \ldots, n)$ alakban.

Shapiro és Wilk az 1965-ben kiadott cikkükben levezették, hogy μ és σ legjobb lineáris becslései azok lesznek, amelyek minimálizálják a

$(\mathbf{Y}-\mu\mathbf{1}-\sigma\mathbf{m})'\mathrm{V}^{-1}(\mathbf{Y}-\mu\mathbf{1}-\sigma\mathbf{m})$ kvadratikus alakot, ahol $\mathbf{1}'=(1,1,\ldots,1)$. Ezek a becslések pedig:

$$\hat{\mu}=\frac{\mathbf{m}'\mathrm{V}^{-1}(\mathbf{m}\mathbf{1}'-\mathbf{1}\mathbf{m}')V^{-1}\mathbf{Y}}{\mathbf{1}'\mathrm{V}^{-1}\mathbf{1}\mathbf{m}'\mathrm{V}^{-1}\mathbf{m}-(\mathbf{1}'\mathrm{V}^{-1}\mathbf{m})^2}$$

és

$$\hat{\sigma}=\frac{\mathbf{1}'\mathrm{V}^{-1}(\mathbf{1}\mathbf{m}'-\mathbf{m}\mathbf{1}')\mathrm{V}^{-1}\mathbf{Y}}{\mathbf{1}'\mathrm{V}^{-1}\mathbf{1}\mathbf{m}'\mathrm{V}^{-1}\mathbf{m}-(\mathbf{1}'\mathrm{V}^{-1}\mathbf{m})^2}$$

Szimmetrikus eloszlásoknál igaz, hogy $\mathbf{1}'V^{-1}\mathbf{m}=0$, ekkor a becslések egyszerűsíthetőek.

A Shapiro-Wilk teszt próbastatisztikája:

$$W=\frac{R^4\hat{\sigma}^2}{C^2S^2}=\frac{b^2}{S^2}=\frac{\mathbf{a}'\mathbf{Y}}{S^2}=\frac{\left(\sum_{i=1}^{n}a_iY_i\right)^2}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2},$$

ahol

$\mathbf{a}'=(a_1,\ldots,a_n)=\dfrac{\mathbf{m}'V^{-1}}{\sqrt{\mathbf{m}'V^{-1}V^{-1}\mathbf{m}}}$,

$b=\dfrac{R^2\hat{\sigma}}{C}$,

$R^2=\mathbf{m}'V^{-1}\mathbf{m}$,

$C^2=\mathbf{m}'V^{-1}V^{-1}\mathbf{m}$,

$S^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$ pedig $(n-1)\sigma^2$ torzítatlan becslése.

A teszt statisztika két mutatószám hányadosa, mely normális eloszlás esetén ugyanazt az elméleti paramétert, a variancát becsüli. A számláló egy konstans szorzótól eltekintve a szórás négyzetének a normális eloszlásból származó rendezett statisztikákból számított legjobb lineáris torzítatlan (Best Linear Unbiased Estimation – BLUE) becslése. A nevező pedig az átlagtól való eltérés négyzetösszege, vagyis $(n-1)$-szerese a tapasztalati varianciának.

Ha a W próbastatisztika értéke kisebb, mint a becsült W_α kvantilis, akkor elvetjük a nullhipotézist, miszerint normális eloszlást követ a minta. A W_α becsült értékei a Shapiro-Wilk cikkben is megtalálhatóak $n\leq 50$ esetben.

A Shapiro-Wilk tesztet megvizsgáltam az **R** statisztikai programban is. A

shapiro.test() parancsot használtam. A programban a mintaelemszámnak 3 és 5000 közé kell esnie.

Amennyiben normális eloszlású mintát tesztelek például $\alpha = 0{,}05$ terjedelem mellett, akkor bármekkora mintaszámnál jó közelítéssel 95%-ban fogja a teszt normális eloszlásúnak értékelni a mintát, ahogy ez várható is.

Ezért azt vizsgáltam, ha két normális eloszlású mintát összekeverek és ezt a mintát vizsgálom a Shapiro-Wilk teszttel, akkor $\alpha = 0{,}05$ terjedelem mellett milyen arányban értékeli a teszt normálisnak a mintámat.

Egy "for" ciklusban végeztem a tesztelést, amit 1000-szer futtattam le. A cikluson belül generáltam egy "a" és egy "b" mintát, az első vizsgálat alkalmával mindkettő $\frac{n}{2}$, a második alkalommal $\frac{n}{5}$ és $\frac{4n}{5}$ elemszámúak. Ha a két minta szórása vagy várható értéke között túl nagy az eltérés a teszt nyilván nem fogadja el a nullhipotézist, miszerint a kombinált minta normális eloszlást követ. Ezért úgy generáltam "a" és "b" mintákat, hogy mindkettő várható értéke 0 és "a" szórása 1, míg "b" szórása 0,5. Ezt a két mintát összeraktam egy vektorba, így kaptam a kombinált mintát, aminek elemszáma n lett. Ezen futtattam le a Wilk-Shapiro tesztet.

A "for" ciklusban megszámoltam, hogy az 1000 tesztelésből hány alkalommal fogadta el a nullhipotézist. Ezt kipróbáltam különböző n értékekre, a végeredményt pedig az 1. táblázatban összefoglaltam.

A Wilk-Shapiro teszt az egydimenziós normalitási tesztek közül sok esetben a legerősebb próba, így azt várhatjuk, hogy ilyen minta tesztelésekor is képes elutasítani a nullhipotézist.

	1000-ből hányszor fogadtuk el H_0-t	
mintaelemszám	$50\% - 50\%$ arány	$20\% - 80\%$ arány
$n = 10$	911	947
$n = 50$	818	905
$n = 100$	652	813
$n = 1000$	3	92
$n = 5000$	0	11

1. táblázat

Jól látszik, hogy minél nagyobb elemszámú mintán végezzük el a tesztet, annál több alkalommal képes elutasítani a teszt H_0-t. Azt is látjuk, hogy

amikor a kombinált minta 80%-a ugyanabból a normális eloszlásból származik, akkor a tesztnek nehezebb dolga van és nehezebben utasítja el a nullhipotézist, miszerint az összetett minta is normális eloszlást követ.

2.3. A bootstrap módszer

A bootstrap egy újramintavételezési eljárás, amit általában becslések szórásának a vizsgálatára és modell-illeszkedés ellenőrzésére használunk. A dolgozatomban később látunk majd példát az alkalmazására, azonban ahhoz, hogy ez világos legyen, szeretném most bemutatni ezt a módszert.
Az ötlet a módszer mögött egyszerű: egy populációból sok mintát szeretnénk, azonban csak egy van. A bootstrap során az eredeti adathalmaz mintalemszámának megfelelő adathalmazokat képezünk az eredeti mintából visszatevéses mintavétellel.

Legyenek $X_1, X_2, \ldots, X_n$ független, azonos eloszlású valószínűségi változók, F közös ismeretlen eloszlással. Legyen a vizsgálandó valószínűségi változó $T_n = t_n(X_n, F)$, aminek eloszlását jelöljük G_n-nel. Ekkor a bootstrap módszerrel G_n eloszlást szeretnénk megbecsülni, a következő lépéseket végezzük ehhez el:

1. Veszünk egy adott X-re m (általában $n = m$) elemű mintát visszatevéssel: $\boldsymbol{X_m^*} = (X_1^*, X_2^*, \ldots, X_m^*)$;
2. Az X_i^*-ok közös eloszlása $F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$;
3. Kiszámoljuk $T_{m,n}^* = t_m(X_m, F_n)$-t;
4. Az előző lépéseket B alkalommal megismételjük, majd az összes kapott $T_{m,n}^*$ statisztikából megbecsüljük a G_n eloszlást.

A következő tételre szoktak hivatkozni a bootstrap módszer alaptételeként.

2.2. Tétel (A bootstrap módszer alaptétele)**.** *Ha a fenti esetet nézzük és* $\sigma^2 = Var(X_i) < \infty$ *és a statisztika a standardizált mintaátlag*

$$T_n = \sqrt{n} \frac{\overline{X}_n - \mu}{\sigma},$$

akkor

$$\lim_{n \to \infty} \sup_x |P_*(T_{n,n}^* \leq x) - \Phi(x)| = 0$$

1 valószínűséggel.

A módszer tulajdonságai:

- A bootstrap az eredeti adatok mindegyikét ugyanakkora valószínűséggel választja ki a mintavétel folyamán.
- Visszatevéses mintavételt használ, az eredeti adatok különböző szimulált ismétléses kombinációit hozzá létre.
- A módszer B darab n elemszámú mintát fog létrehozni az eredeti adathalmazból. B-t úgy kell választanunk, hogy elég nagy legyen ahhoz, hogy a mintavételi hiba elhanyagolható legyen.

A módszer előnyei többek között, hogy rugalmas a minta eloszlására vonatkozó feltételek változására, könnyű leprogramozni, illetve jelenleg ez az egyik leggyorsabban fejlődő területe a statisztikának.

3. A többdimenziós normális eloszlás

A többdimenziós normális eloszlás egy alapvető fogalom a statisztikában, ami az egydimenziós normális eloszlás kiterjesztése. Az egydimenziós cenrális határeloszlás tétel kimondható többdimenzióban, ez a 3.1 tétel. Kulcsfontosságú szerepet játszik különböző területeken, többek között:

- a biológiában, ahol például bonyolultabb genetikai interakciókat és öröklődési mintázatokat tanulmányoznak [2],
- tájelemzésben, ahol az olyan befolyásoló tényezők együttes vizsgálata esetén, mint hőmérséklet, szél és csapadék, a többdimenziós normális eloszlás jó közelítés [3],
- a pénzügyi számításokban is, ahol például a csődelőrejelzésben használt diszkriminanciaanalízisben fontos [4].

Minden példához hivatkoztam egy cikkre is, amiben bővebben lehet olvasni az adott alkalmazási területekről. Ezeken kívül pedig sok más olyan esetben is fontos, ahol összetett kapcsolatokat kell vizsgálni.

A többdimenziós normális eloszlás egyik legfontosabb tulajdonsága, hogy modellezni tudja több véletlen változó együttes viselkedését, ahogy ezt majd a többdimeniós centrális határeloszlás tételben látni fogjuk. Az egyváltozós normális eloszlással ellentétben a többdimenziós változat alkalmas arra, hogy vizsgáljuk vele több változó egymástól való függését. Amikor egy bonyolult rendszert szeretnénk modellezni, amit egy változóval nem tudunk megtenni, akkor a többdimenziós eloszlás elengedhetetlen.

Láthatjuk, hogy a többdimenziós normális eloszlás sokoldalú alkalmazhatósága miatt nagyon fontos számos tudományos és analitikai területen, mert lehetővé teszi a komplex többdimenziós rendszerek vizsgálatát. Azonban valódi adatok esetén nem feltételezhetjük biztosan mindig azt, hogy adataink a normális eloszlást követik. Bár megbizonyosodni nem tudunk, számtalan teszt áll rendelkezésünkre, hogy megvizsgáljuk, hogy az illeszkedés elfogadható-e. Ezekből fogok néhányat bemutatni a következő fejezetekben, illetve azt is tárgyalom, hogy melyiket milyen esetben a legcélszerűbb alkalmazni.

3.1. A normális eloszlás d-dimenzióban

Megvizsgáltuk az egydimenziós esetét a normális eloszlásnak, most pedig nézzük meg d-dimenzióban!

Ebben az esetben d darab valószínűségi változónk van, $X_1, X_2, \ldots, X_d$. Ezek alkotnak egy d-dimenziós vektort, amely d-változós normális eloszlást követ. Ezt a várható érték vektorral és a variancia-kovarianca mátrixszal tudjuk jellemezni.

3.1. Definíció (d-dimenziós standard normális eloszlás)**.** Legyenek az $Y_1, Y_2, \ldots, Y_d$ független, standard normális eloszlású valószínűségi változók. Ekkor az $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)$ valószínűségi vektorváltozót d-dimenziós standard normális eloszlásúnak nevezzük.

3.2. Definíció (d-dimenziós normális eloszlás)**.** Ha $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)$ d-dimenziós standard normális eloszlású vektorváltozó, A egy dxd méretű valós mátrix és $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d) \in \mathbb{R}^d$, akkor a

$$\mathbf{X} := \mathbf{Y}\mathrm{A} + \boldsymbol{\mu}$$

valószínűségi vektorváltozót d-dimenziós normális eloszlásúnak nevezzük. Jelölése:

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix} \right], \text{ ahol } \sigma_{ij} = cov(X_i, X_j).$$

Az $\mathbf{X}$ definiálásakor használt feltételeket most ki tudjuk használni, a variancia-kovariancia mátrixot A segítségével ki tudjuk számolni. Mivel $\mathbf{Y}$ d-dimenziós vektor standard normális eloszlású, így a variancia-kovariancia mátrixa a dxd méretű egységmátrix, tehát: $\Sigma = Var(\mathrm{A}\mathbf{Y}) = \mathrm{A}^{\mathrm{T}} Var(\mathbf{Y})\mathrm{A} = \mathrm{A}^{\mathrm{T}}\mathbf{I}\mathrm{A} = \\ = \mathrm{A}^{\mathrm{T}}\mathrm{A}$.

A d-dimenziós normális eloszlás sűrűségfüggvénye:

$$f(x_1, \ldots, x_d) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

A centrális határeloszlás tételt már felírtam egydimenzióban, azonban a tétel d-dimenzióban kimondott változata is legalább ennyire fontos.

3.1. Tétel (Centrális határeloszlás-tétel d-dimenzióban)**.** *Legyenek az* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *azonos eloszlású, független d-dimenziós valószínűségi változók, továbbá tegyük fel, hogy* $E(\mathbf{X}_i) = \boldsymbol{\mu}$ *vektor és* Σ *variancia-kovariancia mátrix véges. Legyen* $\mathbf{Z_n}$ *a következő módon definiálva:*

$$\mathbf{Z_n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}),$$

ekkor

$$\mathbf{Z_n} \to N_d(0, \Sigma), \text{ ahogy } n \to \infty,$$

ahol N_d *a d-dimenziós normális eloszlást jelöli.*

A d-dimenziós normális eloszlás sűrűségfüggvényében az exponenciális tag kitevőjében szereplő $(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$ kifejezés egy kvadratikus formula. Ennek a négyzetgyökét, a $\sqrt{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}$ mennyiséget Mahalanobis-távolságnak hívjuk, ami az $\mathbf{x}$ és μ vektorok távolságát reprezentálja.

Két megemlítendő tulajdonsága az $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ d-dimenziós normális eloszlású vektorváltozónak, amelyek fontosak lehetnek különböző tesztek alkalmazásakor:

- Nem csak az igaz, hogy ha $\mathbf{X}$ koordinátái függetlenek akkor korrelálatlanok, hanem az is, hogy ha $\mathbf{X}$ koorditánái korrelálatlanok akkor függetlenek is.
- $\mathbf{X}$ minden X_i koordinátája egydimenziós normális eloszlású. Ez viszont visszafele nem feltétlenül igaz! Tehát ha van egy $\mathbf{Z}$ vektor, aminek minden Z_i koordinátája normális eloszlást követ, nem biztos, hogy Z_i-k együttes eloszlása is normális.

3.2. A kétdimenziós normális eloszlás

A többdimenziós normális eloszlás egy megemlítendő speciális esete a kétváltozós normális eloszlás. Azért gondolom fontosnak ezt külön tárgyalni, mert a kétdimenziós sűrűségfüggvény jól ábrázolható. Itt értelemszerűen két koordinátánk van:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

A variancia-kovarianca mátrix főátlójában X_1 és X_2 szórásnégyzetét találjuk. A másik átlóban a két valószínűségi változó kovarianciája található, amit a korreláció és a két szórás szorzataként kapunk.

A variancia-kovariancra mátrix determinánsa és inverze ebben az esetben egyszerűen számolható:

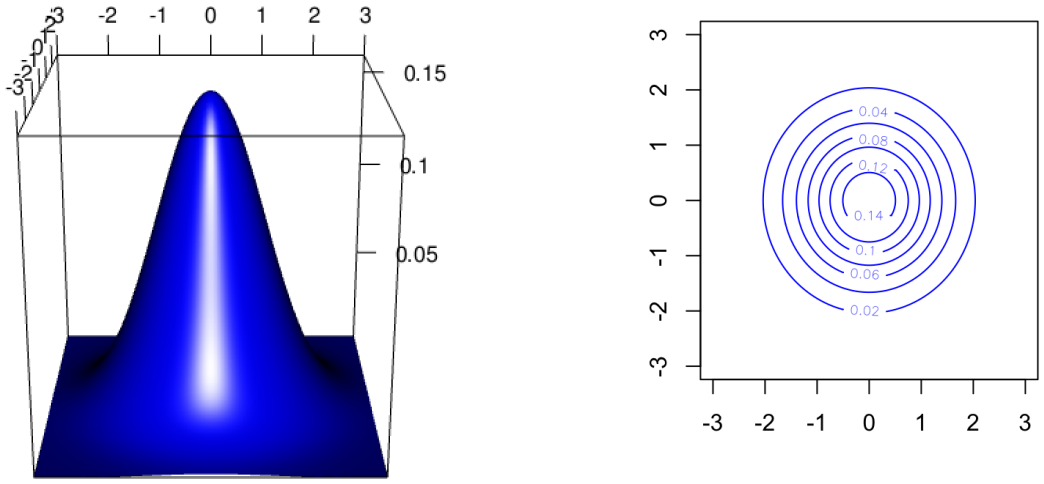
$$|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2), \qquad \Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

Láttuk az előző alfejezetben a többdimenziós normális eloszlás sűrűségfüggvényét. Azonban a Σ determinánsának és inverzének behelyettesítésével kétdimenzióban így is fel tudjuk írni a két valószínűségi változó együttes sűrűségfüggvényét:

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot$$
$$\cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$

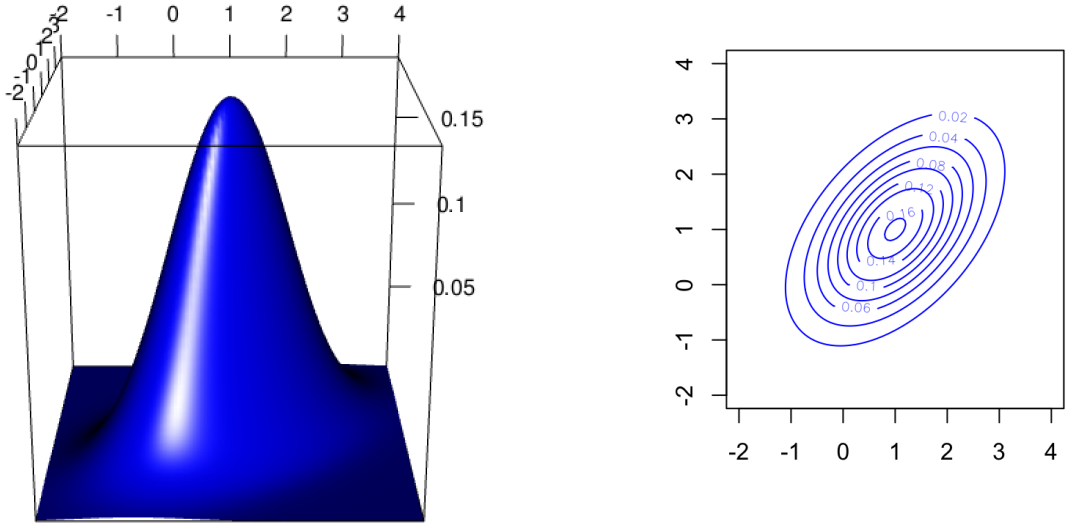
A kétdimenziós normális eloszlás sűrűségfüggvényét megnéztem az **R** statisztikai programban – ezek az (a) jelű ábrák –, majd ennek a szintvonalait is kirajzoltam – ez látható a (b) jelű ábrákon.

A 2. ábra az $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$ esetet ábrázolja, míg a 3. ábra az $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{pmatrix}\right]$ esetet szemlélteti.



(a) (b)

2. ábra. A két valószínűségi változó korrelálatlan



(a) (b)

3. ábra. A két valószínűségi változó között pozitív korreláció van

4. Grafikus többdimenziós normalitásvizsgálat

Azt, hogy az adataink normális eloszlást követnek-e, megvizsgálhatjuk grafikus és statisztikai tesztek segítségével is. A grafikus módszerek lehetővé teszik az adatok vizuális elemzését és értékelését, elősegítve számunkra az összefüggések felismerését és az esetleges kiugró értékek észrevételét.

A grafikus tesztek használata során az a célunk, hogy az adatok illeszkedését az elméleti normális eloszláshoz viszonyítani tudjuk vizuálisan. Ám ezen módszerek használatakor nem kapunk egyértelmű és pontos választ, hiszen egy ábrát vagy grafikont vizsgálunk. Ezért ezeket a teszteket akkor szoktuk inkább alkalmazni, amikor komplex adathalmazzal van dolgunk és nem az a fontos, hogy precíz eredményt kapjunk, hanem inkább az, hogy rálátást nyerjünk az adatainkra és vizualizálni tudjuk ezeket.

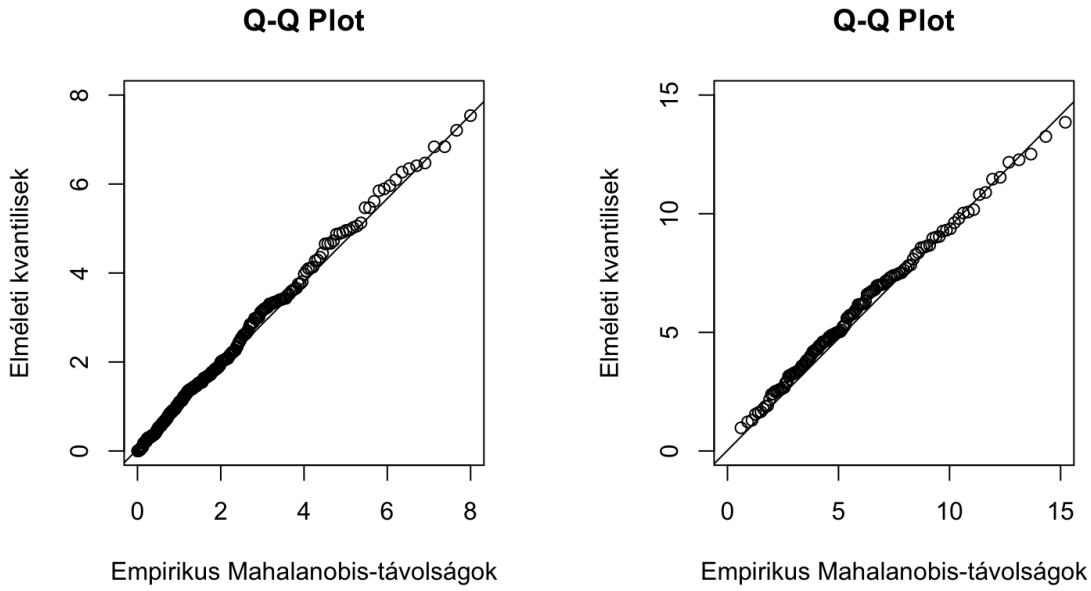
Ebben a fejezetben bemutatok néhány ismert grafikai módszert, amelyek alkalmazhatóak a többdimenziós normális eloszlás tesztelésére.

4.1. Q-Q plot

A Q-Q plot – amit erre az esetre Wilk és Gnanadesikan publikált először 1968-ban – alkalmazható nem csak egydimenziós, hanem többdimenziós eloszlások esetén is, én most ezt a többváltozós normális eloszlás esetére fogom megnézni.

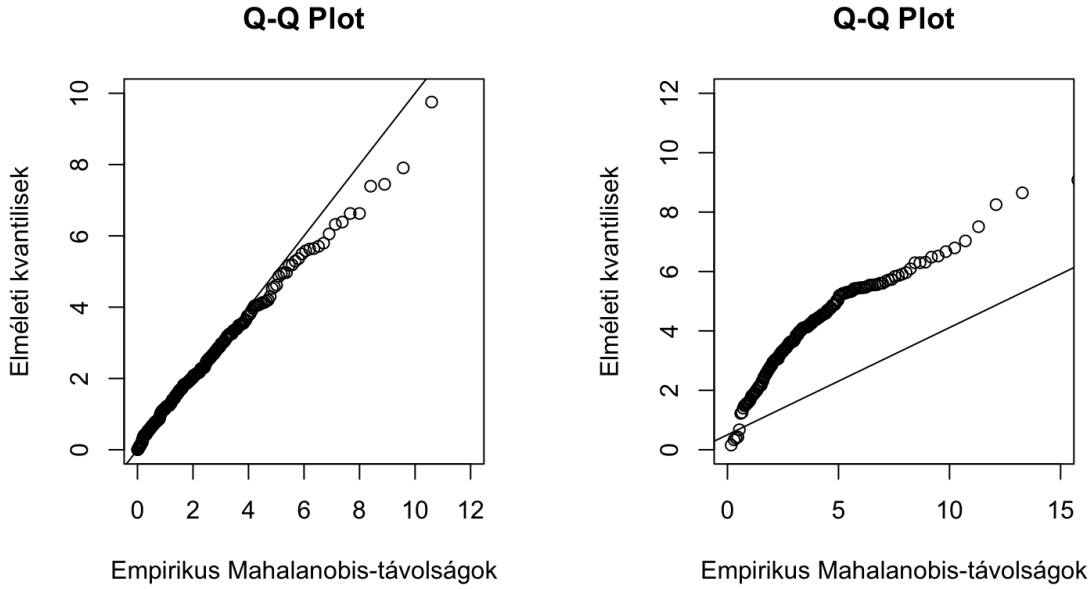
Az előző fejezetben már esett szó a Mahalanobis-távolságról, nézzük meg most azt az esetet, ahol ezt használjuk a Q-Q plotban! Adjunk az érték négyzetének egy jelölést, legyen $\delta^2 = (\mathbf{X}-\mu)^T\Sigma^{-1}(\mathbf{X}-\mu)$. Ez a változó d szabadságfokú χ^2-eloszlást követ d-dimenziós normális eloszlású vektorváltozó esetén. Ezzel tehát vissza tudjuk vezetni a feladatot egydimenziós χ^2 eloszlás tesztelésre.

A Q-Q diagramon a minta Mahalanobis-távolságai fognak szerepelni. Az alapötlet ugyanaz, mint egy normális valószínűségi plot esetén, a rendezett, tapasztalati Mahalanobis-távolság értékeket ábrázoljuk, majd megnézzük, hogy ezek mennyire illeszkednek egy d-szabadságfokú χ^2-eloszlásból származó mintára.



(a) Kétdimenziós normális minta (b) Hatdimenziós normális minta

4. ábra. Többdimenziós normális minták Q-Q plot-ja



(a) Kevert minta (b) Négydimenziós egyenletes minta

5. ábra. Többdimenziós nem normális minták Q-Q plot-ja

A diagramok x-tengelyén az empirikus Mahalanobis-távolságok rendezett értékei szerepelnek, az y-tengelyén pedig a kvantilisek $\frac{i}{n+1}$ felosztásban.

A 4. ábra (a) jelű részén egy kétdimenziós, míg a (b) jelű részén egy hat-

dimenziós generált normális minta Mahalanobis-távolságait hasonlítottam össze egy 2, illetve 6 szabadságfokú χ^2-eloszlással és látható is, hogy valóban egész jól illeszkednek. Ez alapján elfogadhatónak tűnik a feltételezés, hogy valóban normális eloszlásból származik a mintánk.

Az 5.ábra (a) részén ismét elővettem azt a módszert, hogy úgy készítettem egy kétdimenziós mintát, hogy a 20%-a egy kétdimenziós normális eloszlást követ, míg a 80%-a szintén kétváltozós normális eloszlású, azonban más paraméterekkel. Látható, hogy itt azért vannak már szisztematikusan eltérő értékek. Az 5.ábra (b) részén pedig egy négydimenziós egyenletes eloszlású mintát ábrázoltam. Látszik, hogy ez semennyire nem illeszkedik a 4 szabadságfokú χ^2-eloszlásra.

4.2. A Holgersson-féle grafikus teszt

Egy másik grafikus módszer a karakterisztikus függvényt használja fel. Ezt a megközelítést Thomas Holgersson írta le először cikkében [5].

A módszer a normális eloszlás azon tulajdonságán alapszik, hogy a mintaátlag és a minta varianca-kovariancia mátrixának bármilyen lineáris kombinációi függetlenek egymástól akkor és csak akkor, ha a valószínűségi változónk normális eloszlást követ. Holgersson azt mondta, hogy ezen grafikus módszer segítségével azt tudjuk inkább megállapítani, hogy mikor nem normális eloszlású a mintánk. Amikor vizuálisan nincs nagy eltérés az ábrázolt realizált értékek és az elméleti eloszlás között, akkor mondhatjuk, hogy a feltételezett eloszlásból származik a mintánk.

Legyen $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ azonos eloszlású, független minta $\mathbb{R}^p$-ben a bevezetett Σ és $\boldsymbol{\mu}$ jelölésekkel. Továbbá legyen

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X_j} \text{ és } \mathrm{S} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{X_j} - \overline{\mathbf{X}})(\mathbf{X_j} - \overline{\mathbf{X}})^T.$$

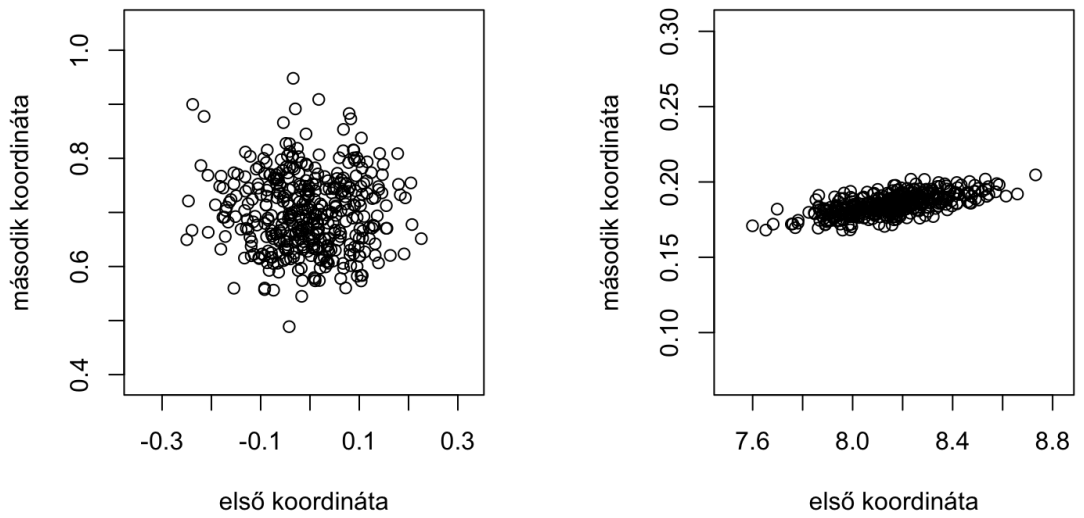
Az együttes eloszlást B bootstrap módszerrel kapott, független mintával fogjuk vizsgálni. Legyen ez az n elemű $\mathbf{X} = (X_1, \ldots, X_n)$ mintából visszatevéses mintavétellel kapott $X_1^*, X_2^*, \ldots, X_B^*$.
B-ből képzett minden $(\mathbf{w^T}\overline{\mathbf{X}}^*{}_b, \mathbf{w^T}\mathbf{S}^*{}_b\mathbf{w})$, $b = 1, \ldots, B$ párt ábrázolunk. Ha a grafikonon korrelációt tudunk felfedezni, akkor az együttes normális eloszlást elvetjük, ellenkező esetben pedig elfogadjuk.

Azt még fontos megjegyezni, hogy $\mathbf{w}$ vektor sokféleképpen választható. Ha az egész mintánk eloszlását szeretnénk vizsgálni, akkor $\mathbf{w}$ egyik komponense sem lehet 0. Ha bizonyos koordinátákat nem szeretnénk a vizsgálat során figyelembe venni, akkor ennek megfelelően kell bizonyos koordinátákat 0-nak választani.

Amikor ábrázoljuk a mintánkat, akkor a B-ből képzett $(\mathbf{w^T}\overline{\mathbf{X}}^*{}_b, \mathbf{w^T}\mathbf{S}^*{}_b\mathbf{w})$ párok $(\mathbf{w^T}\overline{\mathbf{X}}^*{}_b)$ értéke szerepel az x-tengelyen, míg $(\mathbf{w^T}\mathbf{S}^*{}_b\mathbf{w})$ értéke pedig az y-tengelyen.

Ennek a módszernek is – ahogy a grafikus teszteknek általában – hátránya, hogy nem képes egyetlen mérőszámban megragadni a minta normalitását. Ahogy a következő példán látni is fogjuk azonban, szerencsére nemnormális minták esetén a korreláció szemmel látható a grafikonon.

Az egyik minta, amit ábrázoltam Marshall-Olkin féle többdimenziós exponenciális eloszlású. Legyen $\{E_i : \emptyset \neq B \subset i \in \{1,2,\ldots,b\}\}$ λ_b paraméterű exponenciális eloszlású valószínűségi változókból álló sorozat. Legyen továbbá $T_j = \min\{E_B : j \in B\}$, ahol $j = 1,\ldots,b$.
Ekkor $\mathbf{T} = (T_1,\ldots,T_b)$ együttes eloszlása lesz a b-dimenziós Marshall-Olkin exponenciális eloszlás $\{\lambda_B, B \subset \{1,\ldots,b\}\}$ paraméterrel.



(a) Normális eloszlású minta (b) Marshall-Olkin exponenciális minta

6. ábra. Különböző eloszlású minták Holgersson-féle ábrázolása

A 6. ábra (a) jelű részén egy kétdimenziós normális mintát ábrázoltam, ezen látszik, hogy az adataink egymástól viszonylag függetlenül helyez-

kednek el. Az ábra (b) jelű részén egy olyan mintát ábrázoltam, amit a kétdimenziós Marshall-Olkin exponenciális eloszlásból kaptam. Ezen egyértelműen látszik, hogy nem normális minta, mert látszik korreláció az ábrán.

Összességében azt mondanám, hogy nem a leglátványosabb teszt, biztosabb eredményt kapunk, ha inkább kiszámoljuk a korrelációt, bár ez nagy mintaelemszám esetén igencsak költséges lehet. Arra használható jól Holgersson tesztje, hogy vizualizáljuk az adatainkat és az esetleges korrelációt.

5. Normalitástesztek többdimenzióban

A normalitástesztek a grafikus módszerekkel ellentétben, már egy egyértelmű és objektív döntéssel szolgálnak. Meghatározott α szignifikancia szint mellett elutasítják vagy nem tudják elutasítani a nullhipotézist, miszerint a megadott minta normális eloszlásból származik.

5.1. Mardia-teszt

Ezt a módszert Kanti Mardia nevéhez kötjük, ő írta le először a működését [6]. A teszt vizsgálatához először is be kell vezetnem a ferdeség és csúcsosság fogalmát egydimenzióban, majd többdimenzióban is.

5.1. Definíció (Ferdeség)**.** Ha X egy valószínűségi változó, akkor a ferdeségét így definiáljuk:

$$\gamma = \frac{E(X-\mu)^3}{\sigma^3}.$$

5.2. Definíció (Csúcsosság)**.** Ha X egy valószínűségi változó, akkor a csúcsosságát így definiáljuk:

$$\kappa = \frac{E(X-\mu)^4}{\sigma^4} - 3.$$

A csúcsosság képletében azért vonunk le 3-at, mert így lesz normális eloszlás esetén $\kappa = 0$.

5.3. Definíció (Ferdeség d-dimenzióban)**.** Legyen $\mathbf{X}$ és $\mathbf{Y}$ két független, azonos eloszlású d-dimenziós vektor. Ekkor a már ismert μ és Σ jelöléseket használva a ferdeséget így definiáljuk:

$$\gamma = E[\{(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{Y}-\mu)\}^3].$$

5.4. Definíció (Csúcsosság d-dimenzióban)**.** Legyen $\mathbf{X}$ és $\mathbf{Y}$ két független, azonos eloszlású d-dimenziós vektor. Ekkor a csúcsosságot így definiáljuk:

$$\kappa = E[\{(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{Y}-\mu)\}^2].$$

Amikor nem egy valószínűségi változó ferdeségét és csúcsosságát számoljuk – ami egy elméleti érték–, hanem egy megfigyelt mintáét – így tapasztali értéket kapva –, akkor további apró változtatásra van szükségünk.

A kiszámolt ferdeséget $\left(\frac{n}{n-1}\right)^3$-nal, a csúcsosságot pedig $\left(\frac{n}{n-1}\right)^2$-nal szorozzuk be.

A ferdeség azt méri, hogy az eloszlás mennyire nem szimmetrikus, a csúcsosság pedig magától értetődően azt méri, hogy mennyire csúcsos az eloszlás sűrűségfüggvénye.
Fontos tudnunk, hogy a normális eloszlás esetében ezeket a definíciókat alkalmazva $\gamma = 0$ és $\kappa = d(d+2)$. Ez alapján tudjuk a következőkben definiálandó Mardia-teszt segítségével megállapítani, hogy mikor utasíthatjuk el egy adathalmazra a normális eloszlást.

5.1.1. Ferdeség-teszt

Először nézzük meg azt az esetet, amikor ferdeség alapján tesztelünk! Legyen a mintánk $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, ahol $\mathbf{X}_i \in \mathbb{R}^{1xd}$. Ebben az esetben a H_0 az az állítás, hogy a minta többdimenziós normális eloszlásból származik. Ha valóban így van, akkor:

$$\frac{n}{6}\gamma \sim \chi^2(df),$$

ahol

$$df = \frac{d(d+1)(d+2)}{6}.$$

Ha a mintánk kevés megfigyelésből áll (általában ezt $n \leq 20$ esetére értjük), akkor a következő korrigált statisztikát alkalmazzuk:

$$\frac{nc}{6}\gamma \sim \chi^2(df),$$

ahol

$$c = \frac{(n+1)(n+3)(d+1)}{n(n+1)(d+1)-6} \text{ és } df \text{ az előzők alapján definiált.}$$

Ennek kiszámolása után meghatározott szignifikancia szint mellett már látjuk, hogy a nullhopitézist elutasítjuk-e vagy sem.

5.1.2. Csúcsosság-teszt

Most nézzük Mardia azon tesztjét, amikor a $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, ahol $\mathbf{X}_i \in \mathbb{R}^{1xd}$ mintánk csúcsossága alapján tesztelünk. Ha normális eloszlást követ, ak-

kor:

$$(\kappa - d(d+2))\sqrt{\frac{n}{8d(d+2)}} \sim N(0,1)$$

Ez alapján csak úgy, mint a ferdeség tesztelésekor, előre meghatározott α mellett már tudjuk, hogy el tudjuk-e utasítani H_0-t vagy sem. Illetve itt is érvényes az, hogy csak azt tudja a teszt biztosan megállapítani a nullhipotézis elutasítása esetén, hogy a minta adott szignifikancia szinten nem normális eloszlású.

5.2. Empirikus karakterisztikus függvényen alapuló tesztek

Ez egy másik többdimenziós normalitási teszt, amiről Henze és Zinkler írtak először átfogó cikket 1990-ben [7], bár 1988-ban már Baringhaus és Henze is foglalkozott vele.

Legyenek $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ független, azonos eloszlású véletlen vektorok $\mathbb{R}^d$-ben. Legyen emellett $Y_{n,j} = \mathrm{S_n}^{-\frac{1}{2}}(X_j - \overline{X}_n)$ a már használt jelölésekkel.

5.2.1. A BHEP tesztek

T. W. Epps és Lawrence B. Pulley normalitástesztjét, ami az empirikus karakterisztikus függvényen alapul [8], Ludwig Baringhaus és Norbert Karl Henze általánosították többdimenziós esetre először. Innen jön ezeknek a tesztelési módszereknek az összefogó neve és jelölése is, a BHEP betűszó. Ezzel a próbastatisztikával számoló módszer az egyik legtöbbet vizsgált többdimenziós normalitás teszt.

A próbastatisztika így számolandó:

$$\mathrm{BHEP}_{n,\beta} = n \int |\Psi_n(t) - \Psi_0(t)|^2 \omega_\beta(t), dt$$

ahol

$$\Psi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{it^T Y_{n,j}}, \quad t \in \mathbb{R}^d$$

az $Y_{n,1},\ldots,Y_{n,n}$ empirikus karakterisztikus függvényét jelöli, továbbá

$$\Psi_0(t) = e^{-\frac{\|t\|^2}{2}}$$

a $N_d(0,I_d)$ eloszlás karakterisztikus függvényének jelölése. A próbastatisztikában még található egy súlyozó függvény, ami pedig így számolható:

$$\omega_\beta(t) = (2\pi\beta^2)^{\frac{d}{2}} e^{-\frac{\|t\|^2}{2\beta^2}},$$

ahol $\beta > 0$ egy fix konstans.

Bármelyik $\mathrm{BHEP}_{n,\beta}$ próbastatisztikán alapuló teszt jó tulajdonsága, hogy konzisztens bármilyen alternatív hipotézis esetén.

Baringhaus és Henze arról a speciális esetről írtak, amikor $\beta = 1$ [9], az általános esetről már 2 évvel később pedig Henze és Zirkler publikáltak. Én az utóbbit fogom bemutatni.

5.2.2. A Henze-Zirkler módszer

Henze és Zinkler azt vizsgálta, hogy a BHEP próbastatisztika a következőképpen is felírható:

$$\mathrm{BHEP}_{n,\beta} = (2\pi)^{\frac{d}{2}} \beta^{-d} \int_{\mathbb{R}^d} \left(g_{n,\beta}(x) - \frac{1}{(2\pi\tau^2)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2\tau^2}} \right)^2 dx,$$

ahol $\tau^2 = \dfrac{(2\beta^2+1)}{(2\beta^2)}$, és

$$g_{n,\beta}(x) = \frac{1}{nh^d} \sum_{j=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{\|x-Y_{n,j}\|^2}{2h^2}},$$

ahol pedig $h^2 = \dfrac{1}{2\beta^2}$.

A $g_{n,\beta}$ egy nemparaméteres kernel sűrűségbecslés. A magfüggvény a normális eloszlás sűrűségfüggvénye, ezt értékeli ki a függvény az x pontban úgy, hogy kiszámolja minden $Y_{n,j}$ távolságát az adott ponttól. Látszik, hogy azokban a pontokban lesz nagy a kapott érték, ahol az $\|x-Y_{n,j}\|^2$ távolság

kicsi, tehát ahol közel van egymáshoz a két érték. A h választásával pedig azt tudjuk szabályozni, hogy ez milyen széles tartományon hat.

Ha egy zárt alakot szeretnénk a próbastatisztikára, amiben integrálnunk sem kell, akkor a következő gondolatmenetet érdemes követnünk. Legyen a kiinduló próbastatisztikánk

$$T_{n,\beta} = n(4 \cdot \mathbf{1}\{S_n\text{nem invertálható}\} + \text{BHEP}_{n,\beta} \cdot \mathbf{1}\{S_n\text{invertálható}\})$$

Azt már az elején megfigyelhettük, hogy $Y_{n,j}$ csak akkor van értelmezve, ha S_n invertálható, hiszen így definiáltuk. Ebből következik, hogy a $\text{BHEP}_{n,\beta}$ próbastatisztika is csak ebben az esetben számolható. Helyettesítsük tehát $\text{BHEP}_{n,\beta}$-t a legnagyobb értékkel, amit felvehet, amikor S_n nem invertálható, tehát 4-gyel. Ez az ötlet egyébként Csörgő Sándor egyik cikkéből származik [10]. Ekkor ha a β is az eddigiek alapján van definiálva, a próbastatisztika felírható a következő módon:

$$\text{BHEP}_{n,\beta} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{(-\frac{\beta^2}{2}\|Y_{n,i}-Y_{n,j}\|^2)} -$$

$$-2(1+\beta^2)^{-\frac{d}{2}}\frac{1}{n}\sum_{i=1}^{n} e^{(-\frac{\beta^2}{2(1+\beta^2}\|Y_{n,i}\|^2)} + (1+2\beta^2)^{-\frac{d}{2}}$$

Carlos Tenreiro egy 2009-es tanulmányában [11] a teszt erejét vizsgálta $d \in \{2,3,\ldots,10,12,15\}$ dimenziókban $n \in \{20,40,60,80,100\}$ elemszámú mintákra. A konklúziója az lett, hogy $\beta = \frac{1}{2}$ választás adja a legjobb eredményt vastagszélű vagy közepesen csúcsos esetekre, azonban végestartójú eloszlás esetén ez nem egy jó választás. Tenreiro javaslata, hogy $\beta = \sqrt{2}(1{,}376 + 0{,}075d)$ paraméterrel számoljunk, ha nem tudunk semmit a mintánk eloszlásáról.
Ebből a tanulmányból azt is megtudhatjuk, hogy érdekes módon az optimális h választása nem függ a mintaelemszámtól. Inkább azt kell figyelembe vennünk, hogy hány dimenziós adattal dolgozunk és mi az ellenhipotézisünk.

5.3. A Székely-Rizzo teszt

A következő normalitási teszt, amit megvizsgálunk, rendelkezik magyar vonatkozással is. A Székely Gábor és Maria Lizzo által tárgyalt módszert szoktuk energia tesztnek is hívni, amiről két átfogó cikket is írtak, az elsőt 2005-ben [12], a másodikat pedig 2013-ban [13].

5.3.1. Az energia távolság

Többféle távolságot tudunk definiálni statisztikai megfigyelések között. Az egyik legismertebb és leghasználtabb az L_2 távolság, melyet Cramér 1928-as cikkében definiált [14]. Ha van egy valószínűségi változónk F eloszlásfüggvénnyel és F_n tapasztalati eloszlásfüggvénnyel, akkor ezek L_2 távolsága így írható fel:

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dx.$$

Ennek a távolság fogalomnak az a hátránya, hogy nem minden eloszlásra alkalmazható ugyanúgy. Tehát ha például illeszkedésvizsgálat közben szeretnénk ezzel számolni, a kritikus értékek függenek F-től. Ezt a problémát küszöböli ki a Cramér-von – Mises – Smirnov távolság, ami a következőképp definiálható:

$$\int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x).$$

Ám ennek a két távolságnak még mindig fennáll egy hátránya. Amennyiben a mintánk d-dimenziós, ahol $d > 1$, akkor sem az L_2, sem a Cramér-von – Mises – Smirnov távolság nem forgásinvariáns. Ez egy nem elhanyagolható probléma, főleg ha többdimemziós normalitást szeretnénk tesztelni.
Az energia távolság ezt tudja kiküszöbölni.

Legyenek X és Y független valószínűségi változók F és G eloszlásfüggvénnyel. Legyen továbbá $X^{'}$ az X független másolata, ami alatt azt értjük, hogy a két változó ugyanabból az eloszlásból származik, viszont nincs köztük korreláció, sem bármilyen egyéb kapcsolat, tehát teljesen függetlenek egymástól. Ugyanígy képezzük $Y^{'}$ független másolatát is az Y valószínűségi változónak. Ekkor:

$$2\int_{-\infty}^{\infty} (F(x) - G(x))^2 \, dx = 2E|X - Y| - E|X - X^{'}| - E|Y - Y^{'}|.$$

Ahhoz, hogy ezt végig tudjam számolni, először kimondok egy lemmát is, amit használni fogok hozzá.

5.1. Lemma. $E|X - Y| = \int_{-\infty}^{\infty} F(x)(1 - G(x)) \, dx + \int_{-\infty}^{\infty} G(x)(1 - F(x)) \, dx.$

5.1. Bizonyítás. *Tudjuk* $|X-Y| = \int_{-\infty}^{\infty} \mathbb{I}\{X \le u < Y\} + \mathbb{I}\{Y \le u < X\}\, du.$

$$E|X-Y| = \Big(\int_\Omega \int_{-\infty}^{\infty} \mathbb{I}\{X \le u < Y\} + \mathbb{I}\{Y \le u < X\}\Big)\, du\, dP =$$
$$= \int_{-\infty}^{\infty} \int_\Omega (\mathbb{I}\{X \le u < Y\} + \mathbb{I}\{Y \le u < X\})\, dP\, du =$$
$$= \int_{-\infty}^{\infty} (P(X \le u < Y) + P(Y \le u < X))\, du =$$
$$= \int_{-\infty}^{\infty} (P(u < Y)P(u \ge X) + P(u < X)P(u \ge Y))\, du =$$
$$= \int_{-\infty}^{\infty} (G(u)(1-F(u)) + F(u)(1-G(u)))\, du.$$

A bizonyításban Fubini tételét, illetve X és Y függetlenségét használtam ki.

Nézzük tehát:

$$2\int_{-\infty}^{\infty} (F(x)-G(x))^2\, dx = 2\int_{-\infty}^{\infty} (F^2(x) - 2F(x)G(x) + G^2(x))\, dx =$$
$$= 2\int_{-\infty}^{\infty} (F^2(x) - 2F(x)G(x) + G^2(x)) + (F(x)+G(x)) - (F(x)+G(x))\, dx =$$
$$= 2\int_{-\infty}^{\infty} F(x)(1-G(x)) + G(x)(1-G(x)) + F^2(x) + G^2(x) - (F(x)+G(x)) =$$
$$= 2E|X-Y| + 2\int_{-\infty}^{\infty} F(x)(F(x)-1)\, dx + 2\int_{-\infty}^{\infty} G(x)(G(x)-1)\, dx,$$

itt a lemmát felhasználva

$$2\int_{-\infty}^{\infty} F(x)(F(x)-1)\, dx = -\int_{-\infty}^{\infty} F(x)(F(x)-1)\, dx -$$
$$-\int_{-\infty}^{\infty} F'(x)(F'(x)-1)\, dx = -E|X-X'|$$

és

$$2\int_{-\infty}^{\infty} G(x)(G(x)-1)\, dx = -\int_{-\infty}^{\infty} G(x)(G(x)-1)\, dx -$$
$$-\int_{-\infty}^{\infty} G'(x)(G'(x)-1)\, dx = -E|Y-Y'|,$$

így megkapjuk, hogy

$$2\int_{-\infty}^{\infty} (F(x)-G(x))^2\, dx = 2E|X-Y| - E|X-X'| - E|Y-Y'|.$$

Ennek egy természetes kiterjesztése d-dimenzóban az energia távolság fogalma.

5.5. Definíció (Energia távolság)**.** Legyenek $\mathbf{X}$ és $\mathbf{Y}$ független valószínűségi változók, $\mathbf{X}'$ és $\mathbf{Y}'$ független másolatokkal, továbbá $E|\mathbf{X}|_d < \infty$ és $E|\mathbf{Y}|_d < \infty$. Ekkor

$$2E|\mathbf{X}-\mathbf{Y}|_d - E|\mathbf{X}-\mathbf{X}'|_d - E|\mathbf{Y}-\mathbf{Y}'|_d.$$

Nem triviális, de bizonyítható, hogy ez az érték nemnegatív, továbbá akkor és csak akkor 0, ha $\mathbf{X}$ és $\mathbf{Y}$ eloszlása megegyezik. És fontos, hogy erre a mennyiségre már a forgásinvariancia is teljesül.

5.1. Tétel. *Tegyük fel, hogy X és Y független d-dimenziós valószínűségi változók, $E|X|_d + E|Y|_d < \infty$, és ψ és φ jelölik a karakterisztikus függvényeiket. Ekkor az energia távolságuk*

$$2E|X-Y|_d - E|X-X'|_d - E|Y-Y'|_d = \frac{1}{c_d}\int_{\mathbb{R}^d} \frac{|\psi(t)-\varphi(t)|^2}{|t|_d^{d+1}}\,dt,$$

ahol

$$c_d = \frac{\pi^{\frac{d+1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)}.$$

5.3.2. A Székely-Rizzo módszer

Ehhez a módszerhez az energia statisztika fogalmát kell használnunk. Az energia statisztikák lényegében statisztikai megfigyelések közötti távolságok függvényei. Ez a koncepció Newton potenciális energia fogalmán alapul, amely két test távolságának függvénye. Ezért vezettem be részletesen az energia távolság fogalmát, hogy egyszerűen megérthető legyen a Székely és Rizzo által bevezetett módszer.

Legyen $\mathbf{X}$ és $\mathbf{Y}$ két d-dimenziós, független véletlen vektorok $\mathbb{P}^X$ és $\mathbb{P}^Y$ eloszlással, $\mathbf{X}'$ és $\mathbf{Y}'$ pedig $\mathbf{X}$ és $\mathbf{Y}$ független másolatai. Ekkor a négyzetes energia távolságot $\mathbb{P}^X$ és $\mathbb{P}^Y$ között így definiáljuk:

$$D^2(\mathbb{P}^X,\mathbb{P}^Y) = 2E\|\mathbf{X}-\mathbf{Y}\| - E\|\mathbf{X}-\mathbf{X}'\| - E\|\mathbf{Y}-\mathbf{Y}'\|$$

A teszt próbastatisztikája:

$$\varepsilon_n = n\left(\frac{2}{n}\sum_{i=1}^{n} E\|\widetilde{Y}_{n,i} - N_1\| - E\|N_1 - N_2\| - \frac{1}{n^2}\sum_{i,j=1}^{n} E\|\widetilde{Y}_{n,i} - \widetilde{Y}_{n,j}\|\right),$$

ahol $\|\cdot\|$ euklideszi normát jelöl és

$$\widetilde{Y}_{n,i} = \sqrt{\frac{n}{n-1}}Y_{n,i} \quad \text{és} \quad Y_{n,i} = S_n^{-\frac{1}{2}}(x_i - \overline{x}), \quad i = 1,\ldots,n,$$

ahol pedig S a minta varianca-kovarianca mátrixa, $\overline{x}$ pedig a mintaátlag. Továbbá N_1 és N_2 független vektorok $N_d(0, I_d)$ eloszlással.

Az energia távolság mélyebb ismerete segítségével észrevehetjük, hogy $E\|N_1 - N_2\| = 2\Gamma\left(\frac{\frac{d+1}{2}}{\Gamma(\frac{d}{2})}\right)$.

A Székely-Rizzo teszt a nullhipotézist, miszerint a mintánk d-dimenziós normális eloszlást követ, nagy ε_n próbastatisztika érték esetén utasítja el. A két szerző 2005-ös cikkéből azt is megtudhatjuk, hogy a teszt konzisztens minden nemnormális alternatívára.

5.4. A Doornik-Hansen teszt

Az utolsó többdimenziós normalitást tesztelő módszer, amit bemutatok Jurgen A. Doornik és Henrik Hansen cikkében szerepel [15]. A teszt a a vizsgálni kívánt minta ferdeségén és csúcsosságán alapul, ám ezeket az értékeket először sztenderd normális eloszlásúvá konvertáljuk, ezután végezzük el a tesztelést.

A két szerző arra hivatkozva alapozta a tesztet a ferdeségre és csúcsosságra, hogy a leghatékonyabb omnibusz tesztek vagy a rendezett megfigyelések súlyozott összegét használják fel, mint például a Shapiro-Wilk teszt, vagy a minta ferdeségét és csúcsosságát. Viszont véges mintákban a megfigyelések ferdesége és csúcsossága nem független egymástól nagy mintaelemszám esetén sem. Ezért ennek a tesztnek a próbastatisztikáját transzformált értékekből számoljuk.

A Doornik-Hansen teszt használható egyváltozós és többváltozós esetben is. Én a szükséges értékeket definiálom egy- és többdimenziós esetben is,

viszont a működési elvet csak d-dimenziós esetben mutatom be. Az egyváltozós eset hasonló próbastatisztikával számolható.

Az elméleti ferdeséget és csúcsosságot definiáltam már. A tapasztalati ferdeséget és csúcsosságot megkaphatjuk a megfigyelésekből, jelöljük ezeket b_1 és b_2 módon:

$$b_1 = \frac{m_3}{m_2^{\frac{3}{2}}}, \quad b_2 = \frac{m_4}{m_2^2}, \quad \text{ahol} \quad m_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^i.$$

Legyen $\mathrm{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}) \in \mathbb{R}^{nxd}$ mátrix, ahol minden $\mathbf{x_i}$ egy d-dimenziós sorvektor. Legyen $\bar{x}$ mintaátlag és S minta variancia-kovariancia mátrix. Legyen V egy olyan mátrix, aminek főátlójában a szórásnégyzeteket vannak, tehát $\mathrm{V} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$. C pedig legyen az a mátrix, amit így kapunk: $\mathrm{C} = \mathrm{V}^{-\frac{1}{2}}\mathrm{S}\mathrm{V}^{-\frac{1}{2}}$. C sajátértékei legyenek λ_i-k és Λ mátrix tartalmazza ezeket a sajátértékeket a főátlójában, tehát $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$. H mátrix pedig álljon a λ_i-khez tartozó sajátvektorokból.
$Y = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n})$ mátrix pedig álljon azokból a transzformált megfigyelésekből, amiket így kapunk:

$$\mathbf{y_i} = \mathrm{H}\Lambda^{-\frac{1}{2}}\mathrm{H}^{\mathrm{T}}\mathrm{V}^{-\frac{1}{2}}(x_i - \bar{x}).$$

Számítsuk most ki minden transzformált megfigyelésnek a ferdeségét és csúcsosságát, majd rendezzük ezeket két vektorba, $\mathbf{B_1} = (b_{11}, \ldots, b_{1d})$ és $\mathbf{B_2} = (b_{21}, \ldots, b_{2d})$.

Ezután még a b_1 és b_2 értékeket is transzformálnunk kell. A ferdeséget Ralph B. D'Agostino módszere [16] szerint fogjuk számolni. Az átalakítás egyetlen feltétele, hogy $n \geq 8$ legyen:

$$\beta = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$
$$\omega^2 = -1 + (2(\beta-1))^{\frac{1}{2}},$$
$$\delta = \frac{1}{(log(\sqrt{\omega^2}))^{\frac{1}{2}}},$$
$$y = b_1\left(\frac{\omega^2-1}{2}\cdot\frac{(n+1)(n+3)}{6(n-2)}\right)^{\frac{1}{2}},$$
$$z_1 = \delta log(y + (y^2+1)^{\frac{1}{2}}).$$

A csúcsosságot először χ^2 eloszlásúvá, majd továbbalakítva végül sztenderd normális eloszlásúvá transzformáljuk, megkapva így z_2 értéket. Ehhez a Wilson-Hilferty köbgyök módszert használjuk:

$$\begin{gathered}
\delta = (n-3)(n+1)(n^2+15n-4), \\
a = \frac{(n-2)(n+5)(n+7)(n^2+27n-70)}{6\delta}, \\
c = \frac{(n-7)(n+5)(n+7)(n^2+2n-5)}{6\delta}, \\
k = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12\delta}, \\
\alpha = a + b_1 c, \\
\chi = (b_2 - 1 - b_1)2k, \\
z_2 = \left(\left(\frac{\chi}{2\alpha} \right)^{\frac{1}{3}} - 1 + \frac{1}{9\alpha} \right) (9\alpha)^{\frac{1}{2}}.
\end{gathered}$$

Így már tudjuk a próbastatisztikához szükséges vektorokat definiálni, mint $\mathbf{Z_1} = (z_{11}, \ldots, z_{1d})$ és $\mathbf{Z_2} = (z_{21}, \ldots, z_{2d})$.

Ekkor a d-dimenziós próbastatisztika:

$$E_d = \mathbf{Z_1}\mathbf{Z_1^T} + \mathbf{Z_2}\mathbf{Z_2^T} \sim \chi^2(2d).$$

A próbastatisztika előnye, hogy bár hosszadalmas, de nem bonyolult kiszámolni. Emellett a korrelációval számolunk a kovariancia mátrix helyett, így a teszt skála-invariáns, ami szintén egy pozitívum. A módszer jó tulajdonságai közé sorolható még, hogy C mátrix $\mathrm{V}^{\frac{1}{2}}$-ből számolandó, így invariáns a rendezésre.

6. Tesztek összehasonlítása

6.1. Szimulált adatok tesztelése

Mindegyik bemutatott teszt megtalálható az **R** statisztikai program valamelyik beépített csomagjában. A Mardia és a Henze-Zirkler teszt az *mvnnormalTest* csomagból a *mardia()* és *mhz()* parancsaival, a Székely-Rizzo teszt az *energy* csomag *mvnorm.test()* parancsával, a Doornik-Hansen teszt pedig az *mvnTest* csomag *DH.test()* parancsával használható. Mindegyik beépített teszt a feljebb bemutatott próbastatisztikával számol. A Mardia tesztnél van annyi különbség, hogy egy teszten belül vizsgálja az adatok csúcsosságát és ferdeségét, és ha már az egyik érték kritikus, akkor elutasítja a nullhipotézist.

Azzal kezdtem a bemutatott tesztek összehasonlítását, hogy mindegyiket lefuttattam normális eloszlásból származó mintákon, $\alpha = 0{,}05$ szignifikancia szint mellett. Kiindulásképpen ez jól mutatja, hogy melyik teszt mekkora valószínűséggel követ el elsőfajú hibát, mindenhol százalékban megadva.

teszt	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
Mardia	4	4,6	4,4	5,2
Henze-Zirkler	5,4	4,8	3,8	4,8
Székely-Rizzo	8,2	7,6	5,7	5,8
Doornik-Hansen	6	6,8	4,8	5

2. táblázat. Az elsőfajú hiba relatív gyakorisága 5%-os nominális érték és $d = 2$ dimenziós normális minta esetén

teszt	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
Mardia	6,1	6,2	5	4,9
Henze-Zirkler	8,3	7,9	6,1	5,2
Székely-Rizzo	11,2	10,4	7,2	6,8
Doornik-Hansen	7,8	6,7	5,8	5,2

3. táblázat. Az elsőfajú hiba relatív gyakorisága 5%-os nominális érték és $d = 5$ dimenziós normális minta esetén

teszt	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
Mardia	6,8	6,7	5,8	5,1
Henze-Zirkler	8,4	8,1	7,4	5,8
Székely-Rizzo	14,7	12,8	9,4	8,2
Doornik-Hansen	16,2	13,1	9,2	6,9

4. táblázat. Az elsőfajú hiba relatív gyakorisága 5%-os nominális érték és $d = 10$ dimenziós normális minta esetén

Látjuk, hogy magasabb dimenziókban főleg kis mintaelemszámnál akár 16% körül is mozog például a Doornik-Hansen teszt esetén az elsőfajú hiba gyakorisága. Azonban azt is jól megmutatják a táblázatok, hogy magasabb mintaelemszám teszteléskor ez már sokkal kevésbé valószínű, $\alpha = 0{,}05$ esetén mindegyik tesztnél az elsőfajú hiba gyakorisága közelít 5%-hoz.

Ezután generáltam egy olyan mintát, aminek peremeloszlásai normális eloszlásúak, ám az ezekből képzett többdimenziós minta már nem az. Ezt úgy csináltam, hogy d-dimenziós minta esetén így definiáltam X_i-t, ahol $i \in \{1,\dots,d\}: X_1 \sim N(0,1)$ és

$$X_i = \begin{cases} -X_1 & \text{ha } -1 \le X_1 \le 1 \\ X_1 & \text{különben,} \end{cases} \quad \text{ha } i = 2k \text{ alakú, és}$$

$$X_i \sim N(0,1), \text{ ha } i = 2k+1 \text{ alakú, ahol } k \in \mathbb{Z}.$$

Így kapunk d darab egymástól független X_i változót.

Itt is azt tüntettem fel a táblázatban százalékosan, hogy a tesztek milyen arányban tudták elutasítani a nullhipotézist, miszerint d-dimenziós normális eloszlásból származó mintát vizsgálunk.

A lefuttatott tesztek alapján látható, hogy $d = 2$ dimenzióban a Mardia teszt a másik három teszthez képest kisebb arányban tudja elutasítani a nullhipotézist, a Henze-Zirkler, a Székely-Rizzo és a Doornik-Hansen teszt viszont már $n = 500$ elemű minta esetén is 95% körüli arányban képes rá. $d = 5$ és $d = 10$ dimenzió esetén azonban már $n = 500$ és $n = 1000$ elemű minta tesztelésekor a Mardia az a teszt, ami a legnagyobb arányban képes elutasítani H_0-t. A Székely-Rizzo teszt sok esetben észrevehetően nagyobb arányban képes a nullhipotézis elutasítására, ám itt figyelembe kell vennünk, hogy az elsőfajú hiba valószínűsége is ennél a módszernél volt a legnagyobb. Így ennek használatakor mérlegelnünk kell, hogy mekkora szignifikancia szinttel szeretnénk számolni.

teszt	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
Mardia	5,2	34,1	48,7	76,8
Henze-Zirkler	40,7	64,9	94,6	95,2
Székely-Rizzo	49,4	92,6	96,1	95,8
Doornik-Hansen	41,2	84,3	95	94,7

5. táblázat. $d = 2$ dimenziós kevert minta esetén

teszt	$n = 50$	$n = 200$	$n = 500$	$n = 1000$
Mardia	6,4	27,4	64,1	94,6
Henze-Zirkler	5,2	12,8	38,2	69,4
Székely-Rizzo	14,8	32,2	59,7	82,3
Doornik-Hansen	5,2	11,7	41,3	74,6

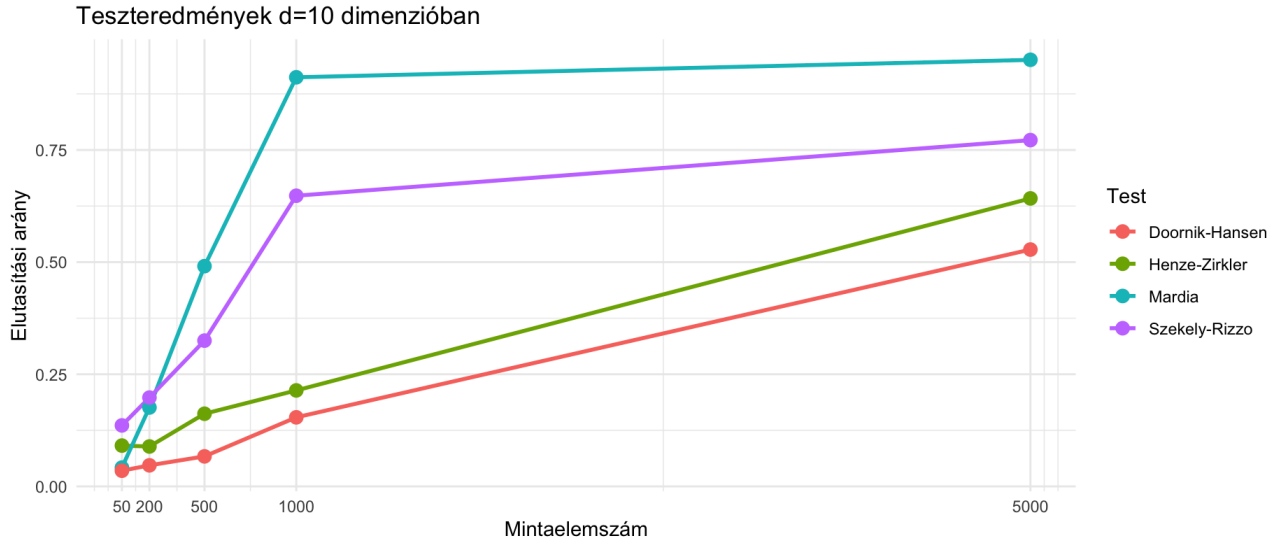
6. táblázat. $d = 5$ dimenziós kevert minta esetén

$d = 10$ dimenzióban lefuttattam a teszteket $n = 5000$ elemszámú mintára is, mert minél magasabb a dimenzió, annál nagyobb elemű minta esetén kapunk megbízható eredményt. Azért néztem $n = 5000$ nagyságú mintára, mert a Mardia teszt az **R** programban maximum ekkora elemszámra fut. Látjuk, hogy ebben az esetben már mindegyik teszt észrevehetően jobb eredményeket ad, mint kisebb minták esetén. Ekkora elemszámú minta esetén már a tesztek futási ideje sem elhanyagolható. 1000-szer futtattam minden tesztet. A Mardia és Henze-Zirkler teszt 5 perc alatt futott le, a Székely-Rizzo teszt pedig majdnem 12 perc alatt. A Doornik-Hansen teszt az, amelyiket a leggyorsabban el tudja végezni a program, ez még ekkora minta esetén is 10 másodperc alatt lefutott. Azt is megállapíthatjuk ezek alapján, hogy minél magasabb dimenziójú adatokat tesztelünk, annál nagyobb elemű mintára van szükségünk a jó teszteredményekhez, akármelyik módszert is választjuk.

teszt	$n = 50$	$n = 200$	$n = 500$	$n = 1000$	$n = 5000$
Mardia	4,2	17,6	49,1	91,2	95
Henze-Zirkler	9,1	8,9	16,2	21,4	64,2
Székely-Rizzo	21,3	27,1	46,5	75,8	84,8
Doornik-Hansen	5,4	6,1	7,2	17,3	63,4

7. táblázat. $d = 10$ dimenziós kevert minta esetén

Láttuk, hogy a Székely-Rizzo és a Doornik-Hansen teszt adott szignifikancia szint mellett jelentősen nagyobb elsőfajú hiba valószínűséggel dolgozik, mint a másik két teszt, főleg magasabb dimenzió és kicsi mintaelemszám esetén. Ezért $d = 10$ dimenziós mintára lefuttattam ezt a két tesztet $\alpha = 0{,}02$ szignifikancia szinttel is, hogy "igazságosabb" legyen az összehasonlításuk a Mardia és a Henze-Zirkler teszttel.



7. ábra

A Székely-Rizzo teszt elég nagy mintaelemszám esetén még így is többször tudta elutasítani a nullhipotézist, mint az $\alpha = 0{,}05$-tel számoló Henze-Zirkler teszt, a Doornik-Hansen pedig maradt a legkevésbé erős teszt a négy közül $d = 10$ dimenzióban. A Mardia teszt az, ami már $n \geq 500$ esetén a többi három tesztnél láthatóan többször tudta elutasítani, hogy a minta együttes eloszlása normális.

6.2. Valódi adatok tesztelése

A teszteket szerettem volna lefuttatni valódi adatokon is. Említettem, hogy fontos szerepet játszik a normális eloszlás a pénzügy területén is, így tőzsdei adatokat választottam. A *Standard&Poor's 500* adatokat töltöttem le. A *Standard & Poor's 500* egy olyan amerikai részvényindex, amely a legnagyobb 500 közép- és nagyvállalatot tartalmazza az Egyesült Államokban. Az *S&P 500* gyakran szolgál referenciaként az amerikai tőzsde teljesítményének mérésére. A befektetők és a pénzügyi szakemberek széles körben használják, hogy megértsék az amerikai részvénypiac alakulását és a gazdasági trendeket.

Az általános feltevés, hogy a részvények árdinamikáját geometriai Brown-mozgás határozza meg, és ebben az esetben az árak logaritmikus hozamai függetlenek és normális eloszlást követnek. Egy fontos példa erre a Black-Scholes modell. Ezt a feltevést teszteltem úgy, hogy 100 olyan vállalatnak letöltöttem az árhozamait az elmúlt 18 évből, amelyek ebben az időszakban konzisztensen részét képezték a *Standard&Poor's* indexnek. Erre a 100 vállalatra nézve kiszámoltam a napi, heti és havi árhozamok logaritmusát. Az így kapott adatokon pedig lefuttattam a már jól ismert négy statisztikai tesztet. A következő táblázatban összefoglaltam, hogy a 100 vállalat hány százalékánál tudták a tesztek elutasítani a nullhipotézist a napi, heti és havi loghozamok alapján. Az előző alfejezetben kapott eredmények miatt a Székely-Rizzo és a Doornik-Hansen tesztet $\alpha = 0{,}02$ szignifikancia szinttel futtattam.

teszt	napi adatok	heti adatok	havi adatok
Mardia	86	42	16
Henze-Zirkler	69	27	9
Székely-Rizzo	72	34	13
Doornik-Hansen	64	27	7

8. táblázat

Ez alapján azt látjuk, hogy a napi adatok alapján nagy arányban tudják elutasítani a tesztek H_0-t, miszerint normális eloszlásúak az adatok. A heti loghozamok vizsgálatakor már kisebb arányban, viszont ha a havi loghozamokat nézzük, azok alapján már az esetek nagy többségében nem tudjuk elutasítani a nullhipotézist.
Ebben tényező lehet a kevesebb adatszám, illetve a korrekció is, miszerint a napi szinten előforduló kiugrások havi szinten kiegyenlítődnek.

7. Konklúzió

A bemutatott tesztek összehasonlítása rámutatott arra, hogy a különböző statisztikai tesztek eltérően viselkednek a többdimenziós normális eloszlás tesztelése során.

Az eredmények alapján elmondható, hogy a Székely-Rizzo és a Doornik-Hansen tesztek használatakor nagyobb az elsőfajú hiba valószínűsége, főleg alacsonyabb dimenzióban, mint amit várunk. Ezért ha ezeket a módszereket használjuk, fontos jól megválasztanunk az α-t. Magasabb dimenzióban a Mardia teszt képes a legnagyobb arányban elutasítani a nullhipotézist, mikor olyan mintákat tesztelünk, amiknek az együttes eloszlása nem normális eloszlást követ. Ez főleg nagy minták esetén látható.

A kevert minták esetén a Székely-Rizzo teszt volt általában a legerősebb, azonban a magasabb elsőfajú hiba valószínűsége miatt érdemes óvatosan kezelni ezeket az eredményeket. A Henze-Zirkler teszt konzisztensen teljesített, bár nem mindig volt olyan erős, mint a Székely-Rizzo, de az alkalmazásakor kisebb az elsőfajú hiba valószínűsége, így megbízhatóbb lehet a gyakorlatban.

A valódi adatok tesztelése során a pénzügyi adatok elemzése is érdekes eredményeket hozott. A kiszámolt napi loghozamok esetén szintén a Mardia teszt volt a legerősebb, 86%-os elutasítási aránnyal, míg a heti és havi értékeknél a tesztek ereje jelentősen csökkent. Ez arra utal, hogy a napi adatokban több olyan tendencia található, amely eltér a normális eloszlástól, míg a heti és havi adatok inkább megfelelnek a normális eloszlás hipotézisének. Ez azt is jelentheti, hogy hosszabb időtávon az adatok szórása csökken, így kevésbé tudjuk elutasítani, hogy normális eloszlásúak az adatok.

Összefoglalva, a különböző módszerek összehasonlítása alapján látható, hogy a megfelelő teszt kiválasztása nagymértékben függ a vizsgált adatok dimenziójától és a minta elemszámától. Az elsőfajú hiba valószínűségére különös figyelmet kell fordítani, különösen magasabb dimenziók és kisebb elemszámú minták esetén. A valódi adatokon végzett tesztek eredményei pedig rámutatnak arra, hogy a többdimenziós normális eloszlás feltételezése nem mindig állja meg a helyét a pénzügyi adatok esetén, különösen rövidebb időszak vizsgálatakor. Ezért a statisztikai tesztek alkalmazása során mindig figyelembe kell venni az adatok jellegzetességeit és a tesztek sajátosságait a legmegbízhatóbb eredmények elérése érdekében.

Hivatkozások

[1] Martin Wilk és Samuel Sanford Shapiro, *An analysis of variance test for normality (complete samples)*, 1965.

[2] John L. Hopper és John D. Mathews, *Extensions to multivariate normal models for pedigree analysis*, 1982.

[3] Thu Pham-Gia, *The multivariate normal distribution, theory and applications,* 2021., 10.7. fejezet

[4] Gordon V. Karels és Arun J. Prakash, *Multivariate normality and forecasting of business bankruptcy,* Journal of Business Finance & Accounting, vol. 14(4) (1987)

[5] H. E. T. Holgerrson, *A graphical technique for assessing multivariate normality,* Computational Statistics, vol. 21 (2006), 141-149. oldal

[6] Kantilal V. Mardia, *Measures of multivariate skewness and kurtosis with applications,* Biometrika, vol. 57(3) (1970), 519-530. oldal

[7] Norbert Henze és Bernd Zirkler, *A class of invariant consistent tests for multivariate normality,* Communications in Statistics-Theory and Methods, vol. 19 (1990), 3595-3617. oldal

[8] T. W. Epps és Lawrence B. Pulley, *A test for normality based on the empirical characteristic function,* Biometrika, vol. 70(3) (1983), 723-726. oldal

[9] Baringhaus és Henze, *A consistent test for multivariate normality based on the empirical characteristic function,* Metrika: International Journal for Theoretical and Applied Statistics, vol. 35(1) (1988), 339-348. oldal

[10] Csörgő Sándor, *Consistency of some tests for multivariate normality,* Metrika: International Journal for Theoretical and Applied Statistics, vol. 36 (1989), 107-116. oldal

[11] Carlos Tenreiro, *On the choice of the smoothing parameter for the BHEP goodness-of-fit test,* 2009.

[12] Székely J. Gábor és Maria Lizzo, *A new test for multivariate normality,* Journal of Multivariate Analysis, vol. 93(1) (2005), 58-80 oldal

[13] Székely Gábor és Maria Lizzo, *Energy statistics: A class of statistics based on distances,* Journal of Statistical Planning and Inference, vol. 143(8) (2013), 1249-1272. oldal

[14] Harald Cramér, *On the composition of elementary errors: II. Statistical applications,* Scandinavian Actuarial Journal, 1928., 141-180. oldal

[15] Jurgen A. Doornik és Henrik Hansen, *An omnibus test for univariate and multivariate normality,* 1994.

[16] Ralph B. D'Agostino, *Transformation to normality of the null distribution of g1,* Biometrika, vol. 57(3) (1970), 679-681. oldal