Normalitásvizsgálati módszerek egy dimenzióban

Szakdolgozat

Írta: Takácová Nikoleta

Matematika BSc

Matematikai elemző szakirány

Témavezető:

Varga László, egyetemi tanársegéd

Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem, Természettudományi Kar

Eötvös Loránd Tudományegyetem

Természettudományi Kar

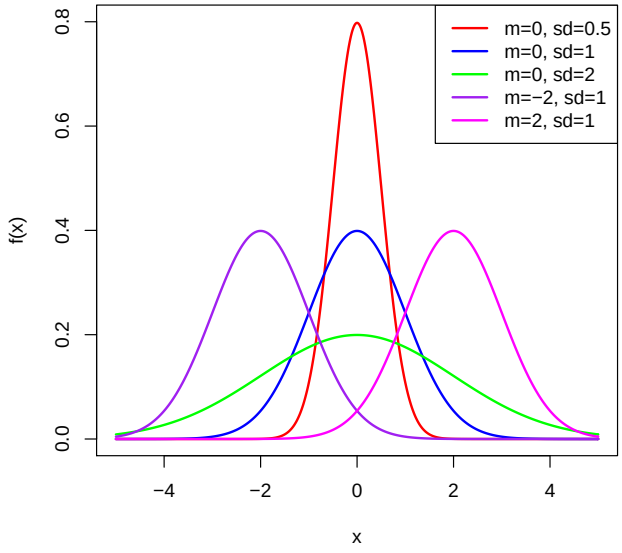
2015

Tartalomjegyzék

1. Bevezetés

A normális eloszlás a valószínűségelmélet és a statisztika egyik legfontosabb abszolút folytonos eloszlása, mely felfedezése óta számos természet- és társadalomtudományi alkalmazásban központi szerepet kap. Karl Pearson alábbi 1920-ban tett kijelentésének köszönhető, hogy normális eloszlás néven vált széles körben ismertté az $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ sűrűségfüggvényű eloszlás: "Many years ago I called the Laplace-Gaussian curve the *normal* curve, which name, while it avoids an international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another abnormal."[1] Stahl [2006]

A normális eloszlás felfedezését több matematikus nevéhez is szokták kötni. Vannak szerzők, akik Fermat és Pascal levelezéséből eredeztetik, vannak, akik de Moivre munkásságából, leggyakrabban pedig Carl Friedrich Gausstól, aki 1809-ben publikált monográfiájában több más fontos statisztikai fogalommal együtt a normális eloszlást is bevezette. Ezért nevezik a normális eloszlást Gauss-eloszlásnak is, sűrűségfüggvényét pedig Gauss-görbének vagy az alakja miatt haranggörbének, ahogy ezt a 1. ábrán láthatjuk.



1. ábra. A normális eloszlás sűrűségfüggvénye különböző m várható érték és sd szórásnégyzet paraméterértékekre.

A normális eloszlás jelentősége egyrészt abban rejlik, hogy sok, a természetben,

[1] "Sok évvel ezelőtt normális görbének neveztem a Laplace-Gauss görbét, amely elnevezés miközben elkerül egy fontos nemzetközi félreértést, az a hátránya, hogy minden másfajta gyakorisági eloszlást valamilyen értelemben szabálytalannak hihetnek."

biológiában, szociológiában és közgazdaságtanban előforduló jelenség legalább megközelítőleg normális eloszlású, másrészt számos statisztikai módszer pontos működésének feltétele a normalitás. Emiatt a normalitás vizsgálatának, amely történhet grafikus módszerekkel vagy statisztikai próbákkal, fontos szerepe van. A leggyakoribb grafikus módszerek leírására a második fejezetben térek ki. Speciálisan a normalitás tesztelésére több alkalmas próba létezik, mint bármely más eloszlás illeszkedésének viszgálatára. A szakirodalomban megtalálható számos teszt csoportokba osztható például az alapján, hogy a normális eloszlás mely jellemzőjének segítségével tesztelik a normalitást. A normalitástesztek így alapvetően négy csoportba sorolhatók: χ^2-típusú próbák, a tapasztalati eloszlásfüggvényen alapuló próbák, regresszión alapuló próbák és momentumtesztek. A szakdolgozat harmadik fejezetében mindegyik csoportból bemutatom a legjelentősebb próbát. Mivel a normalitás sok más statisztikai módszer használatánál szükséges feltétel, ezért az adatok normálissá való transzformálása is egyre nagyobb szerepet játszik a statisztikában. A negyedik fejezetben bemutatom a legismertebb transzformációt, a Box-Cox-transzformációt és ennek néhány lehetséges módozatát. Az ötödik fejezetben Monte Carlo szimulációk segítségével összehasonlítom a próbákat különféle mintaméretekre és eloszlásokra. A hatodik fejezetben pedig a bemutatott normalitásvizsgálati módszereket és transzformációkat valódi adatsorokra alkalmazom, csapadékadatokra, a Happy Planet Index 2012-es adataira és valutaárfolyam adatokra.

A szakdolgozat során a mintát $X_1, \ldots, X_n$ valószínűségi változó sorozat fogja jelölni, a tapasztalati vagy empirikus mintát pedig $x_1, \ldots, x_n$. Az $X \sim N(\mu, \sigma^2)$ azt jelöli, hogy az X valószínűségi változó μ várható értékű és σ^2 szórásnégyzetű normális eloszlást követ, $Y \sim N(0,1)$ pedig a standard normálist.

Az alábbiakban bevezetem a továbbiakban leggyakrabban használt statisztikák jelöléseit:

A mintaátlag: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Az empirikus mintaátlag: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

A tapasztalati szórásnégyzet: $S_n^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$

Az empirikus tapasztalati szórásnégyzet: $s_n^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$

A korrigált tapasztalati szórásnégyzet: $S_n^{2^*} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

Az empirikus korrigált tapasztalati szórásnégyzet: $s_n^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$

2. Grafikus normalitásvizsgálat

A grafikus ábrázolási módszerek, annak ellenére, hogy nem képesek egyetlen mérőszámban megragadni a minta normalitását, hasznos információval szolgálhatnak a mintával támasztott hipotéziseink ellenőrzésére. Célszerű minden statisztikai elemzést azzal kezdeni, hogy valamelyik grafikus módszerrel megvizsgáljuk az adatokat. A számos grafikus módszer közül a szakdolgozatomban a hisztogramra és a Q-Q plotra térnék ki bővebben.

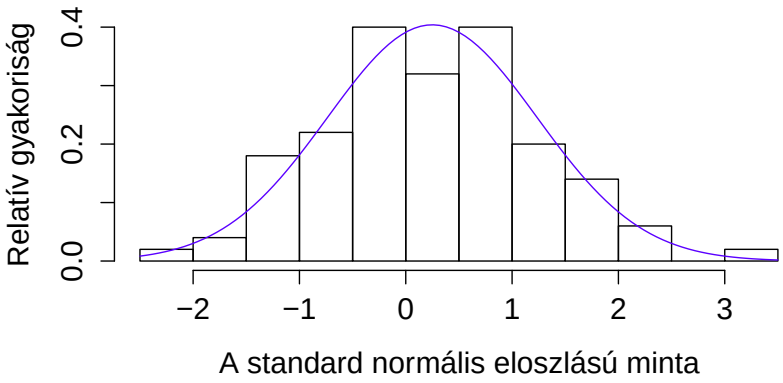
Hisztogram

A hisztogram egy oszlopdiagram, amely a minta előre meghatározott osztályaiba eső elemeinek gyakoriságát szemlélteti. Az oszlopok szélessége az osztály szélességét, magassága a gyakoriságot ábrázolja. Azért hasznos egyenlő nagyságú osztályokat használni, mert így az adatok arányát az egyes oszlopok magassága is érzékelteti. Ha nem egyenlő osztályhosszúságot használunk, a mintában esetlegesen nagyobb számban előforduló szélsőséges értékek miatt félrevezető képet kaphatunk az eloszlásról.

Különböző számítógépes szoftverek alapbeállítása néha nem egyenlő osztályközöket használ a hisztogram ábrázolásakor, ezáltal előfordulhat, hogy hézagos hisztogramot kapunk. Ez szintén félrevezető képet adhat az eloszlásról, ezt is ki tudjuk küszöbölni az egyénileg megszabott osztályhatárokkal.

A hisztogram elkészítéséhez az adatokat először osztályokba soroljuk, mindegyiket pontosan egybe: $(y_j \leq x < y_{j+1})$ a j-edik osztály. Jelölje ν_j a j-edik osztályba eső adatok számát, formálisan $\nu_j = \sum_{i=1}^{n} I(y_j \leq x_i < y_{j+1})$. A relatív gyakoriságok megegyeznek az egyes osztályok fölé rajzolt téglalapok területével, azaz a j-edik osztály fölé rajzolt téglalap magassága $m_j = \frac{\nu_j}{y_{j+1}-y_j}$. A kapott hisztogram a valódi sűrűségfüggvényt közelíti. A téglalapok összterülete a sűrűségfüggvényéhez hasonlóan 1. A hisztogram alakja függ az osztópontok választásától, amelyre általános szabály nincs. Ha az osztópontok túl sűrűn vagy ritkán helyezkednek el a minta elemeihez képest, akkor a hisztogramból nem lehet következtetni a sűrűségfüggvény alakjára.

A hisztogram alapján akkor tekinthető normális eloszlásúnak a vizsgált minta, ha a hisztogram alakja megközelítőleg követi a haranggörbéét. A 2. ábrán egy standard normális eloszlású, 100 elemű véletlen minta hisztogramja látható. A kék görbe a standard normális eloszlás sűrűségfüggvénye.



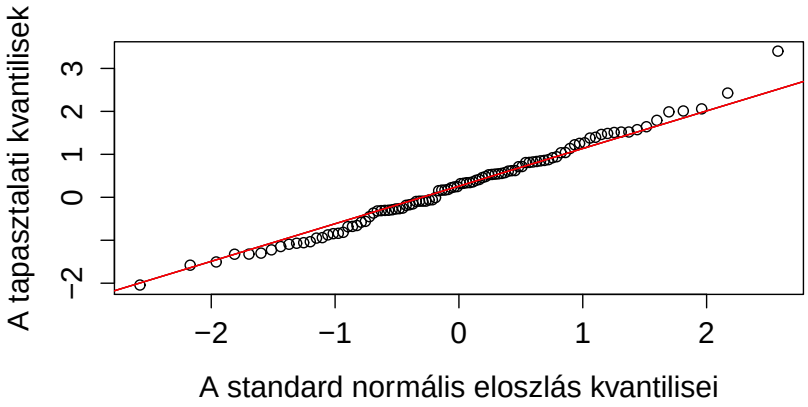
2. ábra. Standard normális eloszlású, 100 elemű véletlen minta hisztogramja.

Q-Q plot

Normalitás vizsgálatakor a leggyakoribb grafikus módszer a Q-Q plot (kvantilis-kvantilis ábra), mely segítségével az $x_1, \ldots, x_n$ minta tapasztalati kvantiliseit vetjük össze az illesztett, azaz a standard normális eloszlás kvantiliseivel. Egy $F_X(x)$ eloszlás z-kvantilise az az érték, amelynél az X valószínűségi változó z valószínűséggel kisebb vagy egyenlő értéket vesz fel, formálisan $q(z) = \sup\{x : F(x) < z\}$. Abszolút folytonos eloszlások esetén ezt az értéket az eloszlásfüggvény inverzéből számíthatjuk.

A Q-Q plot elkészítéséhez először normálást hajtunk végre a mintán: $x'_i = \frac{x_i - \overline{x}}{s^*_n}$, majd sorba rendezzük őket: $x'^*_1 \leq \ldots \leq x'^*_n$, ezek a tapasztalati kvantilis értékek. Az elméleti kvantilis értékek a standard normális eloszlás kvantilisei az $y_i = \frac{i}{n+1}$ $i = 1, \ldots, n$ pontokban. Ezeket a $(\Phi^{-1}(y_i), x'^*_i)$ pontpárokat ábrázoljuk, ahol Φ jelöli a standard normális eloszlás eloszlásfüggvényét.

Amennyiben a minta normális eloszlást követ, a pontok megközelítőleg lineárisan helyezkednek el. Ennek eldöntésére a Q-Q plot ábráján szokás behúzni a 45 fokos egyenest, amelyre minél jobban simulnak a pontok, annál jobbnak tekinthető az illeszkedés. A 3. ábra egy standard normális eloszlású, 100 elemű véletlen minta Q-Q plotját ábrázolja, a 45 fokos egyenest piros színnel jelöljük. Ha számos pont inkább az egyenesen kívül esik, akkor azt mondhatjuk, hogy az adatok szemmel láthatóan nem követnek normális eloszlást.



3. ábra. Standard normális eloszlású, 100 elemű véletlen minta Q-Q plotja.

3. Normalitástesztek

A normalitástesztek a grafikus módszerekkel ellentétben már egy objektív döntést hoznak a vizsgált mintáról. Egy meghatározott α szignifikancia szinten vagy elvetik vagy pedig nem tudják elvetni, hogy a minta normális eloszlást követ. Ebben a fejezetben a normalitás vizsgálatára leggyakrabban használt próbákat fogom bemutatni.

3.1. χ^2-próba

A χ^2-próba az egyik leggyakrabban használt statisztikai próba. Használható illeszkedésvizsgálatra, ahol azt teszteljük, hogy a mintánk származhat-e egy adott eloszlásból, illetve homogenitásvizsgálatra és függetlenségvizsgalatra, melyekkel azt vizsgálhatjuk, hogy két valószínűségi változó tekinthető-e azonos eloszlásúnak illetve függetlennek.

Ez az alfejezet [Bolla and Krámli, 2012] és [Thode, 2002] könyvek alapján íródott.

A χ^2-próba az angol matematikai statisztikus, Karl Pearson nevéhez fűződik, aki a modern statisztika alapjainak megteremtője. 1900-ban publikált cikkében mutatja be először a tesztet, amely a Shapiro-Wilk-próba elterjedéséig a legnépszerűbb próba volt a normalitás tesztelésére. Először általánosan vázolom a χ^2-próbát, majd kitérek arra, hogy miképpen lehet alkalmazni normalitásvizsgálatra.

Legyen $A_1, \ldots, A_r$ teljes eseményrendszer és $p_1, \ldots, p_r$ valószínűségek adottak, amelyekre fennáll, hogy $p_i > 0$ és $\sum_{i=1}^{n} p_i = 1$. Célunk a $H_0: \ P(A_i) = p_i \ \ (i = 1, \ldots, r)$ nullhipotézis tesztelése az ezt tagadó alternatívával szemben. Jelölje $\nu_1, \ldots, \nu_r$ n darab megfigyelésből az $A_1, \ldots, A_r$ események gyakoriságát. A próba azt vizsgálja,

mennyire térnek el a megfigyelt ν_i gyakoriságok az elméleti megfelelőjüktől, az np_i várt gyakoriságoktól.

A próbastatisztika

$$T_n = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i},$$

amely aszimptotikusan $r-1$ szabadságfokú χ^2-eloszlású H_0 teljesülése esetén.

Legyenek $X_1, \ldots, X_n$ független, standard normális eloszlású valószínűségi változók. Ekkor azt mondjuk, hogy az $X = \sum_{i=1}^{n} X_i^2$ valószínűségi változó n szabadságfokú χ^2-eloszlást követ, és χ_n^2-tel jelöljük.

A χ^2-próba végrehajthatóságának feltétele, hogy a tapasztalati gyakoriságok megfelelő nagyságúak legyenek. Kis minta esetén ez a gyakoriság legyen legalább 3, közepes mintánál 5, nagy mintánál pedig 10, és ez legalább a gyakoriságok 80%-ára teljesüljön. Ha ez valamely osztály esetén nem teljesül, akkor azt az osztályt (vagy osztályokat) össze kell vonni más alkalmas osztályokkal.

Diszkrét illeszkedésvizsgálatnál a mintaelemek r különböző értéket vehetnek fel, azaz r darab osztály lesz. Például kockadobásnál az $1, 2, 3, 4, 5, 6$ osztályba sorolhatóak az adatok, a gyakoriságok pedig azt jelölik, hogy hányszor jöttek ki az egyes értékek n darab dobásból. Ekkor χ^2-próbával ellenőrizhető, hogy szabályos-e a kocka, azaz minden p_i $(i = 1, \ldots, 6)$ valószínűség egyenlő-e $\frac{1}{6}$-dal.

Ha tehát az X valószínűségi változó diszkrét eloszlású $x_1, \ldots, x_r$ véges értékkészlettel, akkor a $A_i = \{X = x_i\}$ választással a nullhipotézis a következő alakot ölti: $H_0: \ P(X = x_i) = p_i \ \ (i = 1, \ldots, r)$. Ezt α szignifikanciaszinten elvetjük, ha $\chi^2 \geq \chi^2_{\alpha, r-1}$, ahol $\chi^2_{\alpha, r-1}$ az $r-1$ szabadságfokú χ^2-eloszlás $(1-\alpha)$-kvantilise.

Az előzőleg bevezetett diszkrét illeszkedésvizsgálat a normalitás tesztelésére nem alkalmazható, mivel a normális eloszlás egy folytonos eloszlás.

Ha X abszolút folytonos eloszlású valószínűségi változó, a $H_0 : P(X < x) = F(x) \ \forall x \in \mathbb{R}$ nullhipotézist szeretnénk tesztelni.

A χ^2-próba ennek elvégzésére is alkalmazható, de ekkor úgynevezett diszkretizálást kell végrehajtani – a folytonos eloszlásból származó mintára úgy kell tekinteni, mintha az egy megfelelően megválasztott diszkrét eloszlásból származna. Fontos megjegyezni, hogy a diszkretizálás nem egyértelmű, és vigyázni kell, hogy az egyes mesterségesen képzett osztályokba eső mintaelemek száma kellően magas legyen.

A teljes eseményrendszert a számegyenes r részre való felosztásával kapjuk meg:

$$-\infty = x_0 < x_1 < \ldots < x_{r-1} < x_r = \infty$$

A felosztást két szellemben végezhetjük el: vagy közel azonos hosszúak legyenek az egyes intervallumok, vagy közel azonos valószínűséggel essenek beléjük az egyes értékek (H_0 esetén).

Jelölje $\nu_j = \sum_{l=1}^{n} I(X_l \in [x_{j-1}, x_j)), \quad i = 1, \ldots, r$ az j-edik intervallumba eső megfigyelések számát, és $p_j = P(X \in x_{j-1}, x_j)) = F(x_j) - F(x_{j-1})$ pedig a j-edik intervallumba esés elméleti valószínűségét. Ezután már hagyományos módon elvégezhető az illeszkedésvizsgálat χ^2-próbával.

Amennyiben a nullhipotézisbeli eloszlásunknak egy vagy több paraméterét nem ismerjük, akkor az(oka)t először meg kell becsülnünk maximum likelihood módszerrel, csak utána hajthatjuk végre a χ^2-próbát. Ilyenkor becsléses illeszkedésvizsgálatról beszélhetünk. A próbastatisztika számolásán ez nem változtat, ugyanakkor a szabadságfokból levonódik a becsült paraméterek száma. Tehát a χ^2-eloszlás szabadsági foka $r - 1 - s$ lesz, ahol s a becsült paraméterek száma.

Szakdolgozatomban az egyes minták normális voltát szeretném ellenőrizni, ezért röviden kitérnék a normális minta ismeretlen paramétereinek becslésére. Ha tudjuk, hogy egy $X_1, \ldots, X_n$ független, azonos eloszlású minta normális eloszlásból származik ismeretlen m várható értékkel és ismeretlen σ szórással, akkor ezek maximum likelihood becslése $\hat{m}_{ML} = \overline{X}$ és $\hat{\sigma}^2_{ML} = S_n^2$. Ismert eredmény, hogy a tapasztalati szórásnégyzet nem becsüli torzítatlanul a szórásnégyzetet, ezért σ ML-becslésnek a korrigált tapasztalati szórást szokás választani.

3.2. Tapasztalati eloszláson alapuló próbák

A tapasztalati eloszláson alapuló próbák azt vizsgálják, hogy a tapasztalati eloszlásfüggvény mennyire tér el az elméleti eloszlásfüggvénytől. A tapasztalati eloszlásfüggvény egy lépcsős függvény, amely minden megfigyeléshez $1/n$ súlyt rendel.

Tegyük fel, hogy $X_1, \ldots, X_n$ egy véletlen minta, és legyen $X_1^*, \ldots, X_n^*$ a rendezett minta. A minta tapasztalati eloszlásfüggvénye $F_n(x)$ a következő tiszta ugrófüggvény:

$$F_n(x) = \tfrac{1}{n} \sum_{i=1}^{n} I(X_i < x) = \begin{cases} 0 & \text{ha } x < X_1^* \\ \frac{i}{n} & \text{ha } X_i^* \leq x < X_{i+1}^* \quad \text{ahol } i = 1, \ldots, n-1 \\ 1 & \text{ha } X_n^* \leq x \end{cases}$$

Mint azt a statisztika alaptételének is nevezett Glivenko-Cantelli tételből tudjuk, $n \to \infty$ esetén $\sup_{x \in \mathbb{R}} \mid F_n(x) - F(x) \mid \to 0$, 1 valószínűséggel, azaz a tapasztalati eloszlásfüggvény 1 valószínűséggel egyenletesen tart az elméleti eloszlásfüggvényhez. Az is ismert, hogy a tapasztalati eloszlásfüggvény torzítatlan becslése az eloszlásfüggvénynek. Ezeken a tényeken alapulnak az ebben az alfejezetben tárgyalt próbák.

A tapasztalati eloszláson alapuló tesztek próbastatisztikái a tapasztalati eloszlásfüggvény és az elméleti eloszlásfüggvény közti függőleges távolságot számolják. Ez alapján a tesztek két nagy csoportba sorolhatók:

- maximális távolságon alapuló tesztek (Kolmogorov-Szmirnov, Lilliefors, Kuiper)
- négyzetes távolságon alapuló tesztek (Anderson-Darling, Cramér-von Mises)

Kolmogorov-Szmirnov-próba

Az először 1933-ban publikált tesztet Andrej Nyikolajevics Kolmogorov dolgozta ki. Az egymintás Kolmogorov-Szmirnov próba segítségével ellenőrizhető, hogy egy valószínűségi változó eloszlása valóban az, amit feltételezünk, a kétmintás Kolmogorov-Szmirnov-próba pedig két valószínűségi változó eloszlásának összehasonlítására alkalmas [Razali and Wah, 2011].

Az egymintás Kolmogorov-Szmirnov próbával végzett illeszkedésvizsgálatnál tehát feltételezzük, hogy a minta eloszlása megegyezik egy előre ismert eloszlással. A hipotézisek tehát:
$H_0: \ F_n(x) = F(x)$
$H_1: \ F_n(x) \neq F(x)$

Kolmogorov a következő próbastatisztikát javasolta:

$$D_n = \sup_x |F_n(x) - F(x)| = \max(D^+, D^-),$$

ahol $D^+ = \sup_x (F_n(x) - F(x))$ a legnagyobb pozitív irányú függőleges távolság, $D^- = \sup_x (F(x) - F_n(x))$ pedig a legnagyobb negatív irányú függőleges távolság.

H_0 esetén $\sqrt{n}D_n$ a Kolmogorov-eloszláshoz tart, ha $n \to \infty$. A nullhipotézist elvetjük, ha $\sqrt{n}D_n > K_{1-\alpha}$, ahol $K_{1-\alpha}$-val a Kolmogorov-eloszlás $1-\alpha$-kvantilisét jelöljük. A kvantilisek értékei táblázatba foglalva megtalálhatóak például a [Sachs, 2013] könyvben.

A Kolmogorov-Szmirnov próba eloszlásfüggetlen, nem csak a normalitás tesztelésére alkalmas. További előnye a χ^2-próbával szemben, hogy kis elemszámú minta tesztelésére is alkalmas. Eredetileg folytonos eloszlásokra készült, ugyanakkor használható diszkrét eloszlásokra is, bár akkor ritkábban tudja elvetni a nullhipotézist. A Kolmogorov-Szmirnov-próba felteszi, hogy az eloszlás paraméterei teljesen ismertek, tehát becsléses illeszkedésvizsgálatra nem használható.

Lilliefors-próba

Az 1967-ben megjelent, Hubert Lilliefors, amerikai matematikusról elnevezett teszt a Kolmogorov-Szmirnov-próbát módosítja [Lilliefors, 1967]. Míg az felteszi, hogy ismerjük az eloszlás paramétereit, addig Lilliefors abból indul ki, hogy a para-

méterek ismeretlenek, és a mintából becsüli meg a μ várható értéket és a σ^2 szórásnégyzetet. A módosítás azonban nem változtat a próba eloszlásfüggetlenségén.

A próbastatisztika

$$D_n^* = \sup_x |F_n(x) - F(x)|,$$

ahol $F_n(x)$ a $\hat{\mu}$ várható értékű és $\hat{\sigma}^2$ szórásnégyzetű normális eloszlás eloszlásfüggvénye. A $\hat{\mu} = \overline{X}$ és a $\hat{\sigma}^2 = S_n^2$ a μ várható érték és a σ^2 szórásnégyzet torzítatlan becslései.

Ha a $D_n^* > K_{1-\alpha}^*$, akkor elvetjük a nullhipotézist. A $K_{1-\alpha}^*$ a teszt kritikus értékét jelöli α szignifikanciaszinten.

A Lilliefors-próba nehézsége az eredeti Kolmogorov-Szmirnov-próbáéhoz képest az, hogy a próba kritikus értékei csak szimulációval állapíthatóak meg. Lilliefors $4 \leq n \leq 30$ mintaelemszámra megbecsülte a próba kritikus értékeit, ezek táblázatba foglalva megtalálhatók a [Lilliefors, 1967] cikkben. 30-nál nagyobb mintaelemszámra pedig a kritikus értékek különböző szignifikaciaszinteken különféleképpen becsülhetők, ahogy azt az 1. táblázat mutatja.

	α szignifikanciaszintek				
	0,20	0,15	0,10	0,5	0,1
$n > 30$	$\frac{0{,}736}{\sqrt{n}}$	$\frac{0{,}768}{\sqrt{n}}$	$\frac{0{,}805}{\sqrt{n}}$	$\frac{0{,}886}{\sqrt{n}}$	$\frac{1{,}031}{\sqrt{n}}$

1. táblázat. A Lilliefors-próba kritikus értékeinek becslése $n > 30$ mintaelemszám esetén különböző szignifikanciaszinteken.

Kuiper-teszt

Nicolaas Kuiper holland matematikus 1960-ban szintén a Kolmogorov-Szmirnov tesztet módosította azzal, hogy a legnagyobb negatív és legnagyobb pozitív távolság maximuma helyett a két érték összegét használta [Thadewald and Büning, 2007]. Ezzel a változtatással a teszt az eloszlás széleinél is hasonlóan érzékeny, mint a medián körül. Használható tiszta illeszkedésvizsgálatra, mikor a feltételezett eloszlásfüggvény paraméterei ismertek, illetve becsléses illeszkedésvizsgálatra is, mikor a paraméterek értékét a mintából becsüljük.

A Kuiper-féle próbastatisztika: $V_n = D^+ + D^-$, ahol D^+ a legnagyobb pozitív irányú függőleges távolság, D^- a legnagyobb negatív irányú függőleges távolság az elméleti és a tapasztalati eloszlásfüggvény között, mint a Kolmogorov-Szmirnov-próbánál.

A nullhipotézist elvetjük, ha $V_n \geq k_{1-\alpha}$, ahol $k_{1-\alpha}$ a kritikus értéket jelöli α szignifikaciaszinteken. A kritikus értékek tiszta és becsléses illeszkedésvizsgálat esetére

is megtalálhatóak táblázatba foglalva például a [Stephens, 1965] cikkben.

Cramér - von Mises-teszt

A Cramér-von Mises-teszt Harald Cramér, svéd matematikusról és Richard Edler von Mises, német matematikusról kapta a nevét, akik először foglalkoztak a négyzetes távolságon alapuló tesztekkel. A teszt próbastatisztikája:

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

Cramér 1928-ban és von Mises 1931-ben hasonló próbastatisztikákat javasoltak, mint a W_n^2, de a fentebbi eloszlásfüggetlen változat Nikolai Vasilyevich Szmirnov szovjet matematikustól származik, 1936-ból [Razali and Wah, 2011].

A nullhipotézist elvetjük, ha a $W_n^2 \leq W_\alpha$, ahol W_α a W_n^2 próbastatisztika eloszlásának α kvantilise, amelyek táblázatba foglalva megtalálhatóak a [Thode, 2002] könyvben. A Cramér-von Mises próbastatisztika eloszlása nem függ az elméleti eloszlástól.

Anderson-Darling-próba

A négyzetes távolságon alapuló tesztek próbastatisztikájának általános alakját Theodore Wilbur Anderson és Donald Allan Darling amerikai matematikusok 1952-ben a következőképpen definiálták:

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

ahol $\psi(x)$ súlyfüggvény. A súlyfüggvény azt a célt szolgálja, hogy az eloszlás különböző részeinek fontossága kiemelhető legyen a segítségével. Az Anderson-Darling-próba például az eloszlás széleire fektet nagy hangsúlyt.

A $\psi(x) = 1$ -re pontosan a Cramér-von Mises-teszt próbastatisztikáját kapjuk meg, $\psi(x) = [F(x)(1 - F(x))]^{-1}$ választással pedig az Anderson és Darling által publikált teszt [Anderson and Darling, 1954] próbastatisztikáját:

$A_n^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 [F(x)(1 - F(x))]^{-1} dF(x)$.

$\psi(x) = [F(x)(1-F(x))]^{-1}$ súlyfüggvény választással a teszt nagyobb súlyt helyez az eloszlás széleire, mint a Cramér-von Mises-teszt.

A normalitást abban az esetben vetjük el, ha a próbastatisztika értéke nagy, azaz $A_n^2 \leq A_{1-\alpha}$, ahol $A_{1-\alpha}$ az próbastatisztika eloszlásának kritikus értékeit jelöli.

Az Anderson-Darling-próba kritikus értékei nem függetlenek a vizsgált eloszlástól, a normális eloszláshoz tartozó kritikus értékek táblázata megtalálható a [Thode, 2002] könyvben.

A Cramér-von Mises- és az Anderson-Darling-tesztek próbastatisztikái felírhatók olyan alakban, amely jelentősen megkönnyíti számolásukat. Ezek a következők:

$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_{(i)}) - \frac{2i-1}{2n}\right)$

$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left(2i-1\right)[\log(F(x_{(i)})) + \log(1 - F(x_{(n+1-i)}))]$

Levezetésük az eredeti próbastatisztikákból megtalálható például Móri Tamás jegyzetében (Móri [2011]).

3.3. Regresszión alapuló tesztek

Shapiro-Wilk-próba

Martin Wilk és Samuel Sanford Shapiro Shapiro-Wilk-próbája az egyik legajánlottabb próba a normalitás tesztelésére. A többi próbával összevetve a Shapiro-Wilk-próba általában a legerősebbnek bizonyul, nem túlságosan elnyúló és ferde eloszlások esetén kiemelkedően nagy a próba ereje, hosszan elnyúló eloszlások esetén a teljesítménye pedig még mindig elfogadható erejű [Thode, 2002].

Az 1965-ben publikált cikk [Shapiro and Wilk, 1965] merőben új módszert vezetett be a normalitás tesztelésére. Azt a tényt használta 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.

Legyen $X' = (X_1^*, X_2^*, \ldots, X_n^*)$ n elemű, standard normális eloszlású rendezett minta. Jelöljük $m' = (m_1, m_2, \ldots, m_n)$-vel a rendezett minta várható értékeinek vektorát, $V = (v_{ij})$-vel pedig a rendezett minta kovarianciamátrixát.
Tehát:

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

és

$$\mathrm{Cov}(X_i^*, X_j^*) = v_{ij} \quad (i = 1, \ldots, n)$$

Legyen $Y' = (Y_1, Y_2, \ldots, Y_n)$ n elemű véletlen minta. Annak érdekében, hogy a Y_i-k normalitását tesztelni tudjuk, a mintát növekvő sorrendbe kell rendeznünk. Legyen $Y_1^* < Y_2^* < \ldots < Y_n^*$ a rendezett véletlen minta. Ha az Y_i^* megfigyelésekről feltesszük, hogy normális eloszlásúak ismeretlen ν és σ paraméterrel, akkor az Y_i^*-k kifejezhetőek az $Y_i^* = \nu + \sigma X_i^* \quad (i = 1, 2, \ldots, n)$ alakú regressziós modellel.

A modell ezen formájában a σX_i^* tag tekinthető a hibatagnak, ám az nem 0 várható értékű. Ezért a modellt úgy kell átalakítani, hogy 0 várható értékű hibatagot

kapjunk. Legyen az $\varepsilon_i = \sigma X_i^* - \sigma m_i$ valószínűségi változó. A várható értéke:

$$E(\varepsilon_i) = E(\sigma X_i^* - \sigma m_i) = E(\sigma X_i^*) - E(\sigma m_i) = \sigma m_i - \sigma m_i = 0 \quad (i = 1, 2, \ldots, n)$$

Mivel az ε_i várható értéke 0, a segítségével át tudjuk írni a fenti modellt. A $\sigma X_i^* = \sigma m_i + \varepsilon_i$ behelyettesítésével a $Y_i^* = \nu + \sigma m_i + \varepsilon_i$ $(i = 1, 2, \ldots, n)$ alakú modellt kapjuk. Mátrixos formában felírva: $Y = \nu 1 + \sigma m + \varepsilon$, ahol $Y' = (Y_1, Y_2, \ldots, Y_n)$, $1' = (1, 1, \ldots, 1)$, $m' = (m_1, m_2, \ldots, m_n)$ és $\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ n elemű vektorok.

A ν és a σ legjobb becslései azok, amelyek minimalizálják a $(Y - \nu 1 - \sigma m)' V^{-1} (Y - \nu 1 - \sigma m)$ kvadratikus alakot [Shapiro and Wilk, 1965].

Ezek a becslések a legjobb lineáris torzítatlan becslések:

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

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

Szimmetrikus eloszlásokra fennáll, hogy $1' V^{(} - 1) m = 0$, így a fenti becslések egyszerűsíthetők: $\hat{\nu} = \frac{1}{n} \sum_{i=1}^{n} Y_i^* = \overline{Y}$ és $\hat{\sigma} = \frac{m' V - 1 Y}{m' V - 1 m}$.

A teszt próbastatisztikája

$$W = \frac{\left[\sum_{i=1}^{n} a_i X_i \right]^2}{\sum_{i=1}^{n} (X_i^* - \overline{X})^2},$$

amely a $W = \frac{R^4 \hat{\sigma}^2}{C^2 S^2} = \frac{b^2}{S^2} = \frac{(a' Y)^2}{S^2} = W = \frac{\left[\sum_{i=1}^{n} a_i X_i \right]^2}{\sum_{i=1}^{n} (X_i^* - \overline{X})^2}$ egyenlőségből kapható, ahol

- $S^2 = \sum_{i=1}^{n} (Y_i^* - \overline{Y})^2$ az $(n-1)\sigma^2$ torzítatlan becslése,
- $R^2 = m' V^{-1} m$,
- $C^2 = m' V^{-1} V^{-1} m$ az együtthatókat normalizáló konstans,
- $b = \frac{R^2 \hat{\sigma}}{C}$ a regressziós egyenes meredekségének legjobb torzítatlan becslése,
- $a' = \frac{m' V^{-1}}{[(m' V^{-1})(V^{-1} m)]^{-\frac{1}{2}}}$ az a_i együtthatók vektora.

A próbastatisztika számlálója és nevezője is a szórásnégyzetet becsüli normális eloszlású minta esetén. A számláló az Y_i^* megfigYelések szórásnégyzetének legjobb lineáris torzítatlan becslése, a nevező pedig alkalmas skálázás után a korrigált tapasztalati szórásnégyzet, melyről normális eloszlás esetén tudjuk, hogy szintén torzítatlan becslése a szórásnégyzetnek.

Az a_i együtthatók értéke a [Shapiro and Wilk, 1965] cikk megjelenésekor $n \leq 20$-ig volt ismert, Shapiro és Wilk $n \leq 50$-ig táblázatba foglalt becsléseket közölnek

a cikkben. A táblázatok az a_{n-i+1} értékeket tartalmazzák, mivel teljesül rájuk a szimmetria, azaz $a_i = -a_{n-i+1}$. Royston 1982-ben publikált egy eljárást, amely $n \leq 2000$-ig becsült értékeket ad. Az eljárás javítását 1995-ben közölte, amely már $n \leq 5000$-ig képes megbecsülni az a_i-k értékeit. Ezt az 1995-ös algoritmust használja a legtöbb statisztikai programcsomag az együtthatók becslésére.

Ha a W próbastatisztika értéke kisebb, minta a W_α becsült kvantilis, akkor elvetjük a normalitást. A W_α becsült értékei $n \leq 50$ megtalálhatóak táblázatba foglalva a [Shapiro and Wilk, 1965] cikkben vagy a [Thode, 2002] könyvben.

3.4. Momentumtesztek

A momentumtesztek a normális eloszlástól való különbözőséget a harmadik és negyedik tapasztalati momentumok segítségével vizsgálják.

A harmadik centrális momentum és a szórás harmadik hatványának hányadosaként számítjuk a ferdeséget, ami azt ragadja meg, mennyire tér el a valószínűségi változó eloszlása a szimmetrikustól. Képlete formálisan: $\beta_1 = \frac{E[(X-E(X))^3]}{E[(X-E(X))^2]^{3/2}} = \frac{E[(X-E(X))^3]}{(DX)^3}$. Szimmetrikus eloszlások esetén $\beta_1 = 0$, mivel a centrált páratlanadik momentumok mind 0-val egyenlők, így a ferdeség képletében a számlálóban szereplő várható érték is.

A negyedik centrális momentum és a szórás negyedik hatványának hányadosa, a lapultság (egyes könyvekben csúcsosság), pedig azt mutatja meg, hogy a valószínűségi változó sűrűségfüggvényének csúcsossága hogyan viszonyul a standard normális eloszláséhoz. Számítása: $\beta_2 = \frac{E[(X-E(X))^4]}{E[(X-E(X))^2]^2} - 3 = \frac{E[(X-E(X))^4]}{(DX)^4} - 3$, amiben a 3-at azért szokás kivonni, mert normális eloszlás esetén β_2 éppen 0 értéket vesz fel. Ezt könnyen ellenőrizni is tudjuk: legyen $Z \sim N(0,1)$, ekkor $\beta_2 = \frac{E[(Z-0)^4]}{1}^4 - 3 = \frac{3}{1} - 3 = 0$, mivel ismert, hogy a standard normális eloszlás párosadik momentumai az eggyel kisebb szemifaktoriálisok segítségével számíthatók: $EZ^{2m} = (2m-1)!! = (2m-1)\cdot\ldots\cdot 3\cdot 1$.

A ferdeség és lapultság becsült értékei, a tapasztalati ferdeség ($\sqrt{b_1}$) és a tapasztalati lapultság (b_2) a következőképpen számolhatóak:

$$\sqrt{b_1} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_n}\right)^3$$

és

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_n}\right)^4 - 3$$

A tapasztalati lapultság képletéből, mint az elméleti értékből, szintúgy szokás 3-at levonni.

A momentumtesztek a mintából megbecsült, empirikus momentumokat hasonlítják a standard normális eloszlás elméleti momentumaihoz. Mivel mindkét mennyiség jól jellemzi a normális eloszlást, közkedvelt kiindulási alapok voltak a normalitás tesztelésére, azonban csak az egyik vagy másik adatot vizsgálva félrevezethető eredményt adhatnak. Előfordulhat, hogy míg egy mintánál az egyik elveti a normális eloszlás hipotézisét, a másik nem tudná elvetni. Ezen probléma kiküszöbölésének érdekében kezdték el a kettő különféle függvényeit használni. Az egyik legismertebb és legnépszerűbb momentumteszt a Jarque-Bera-teszt.

Jarque-Bera-teszt

Az 1987-ben Carlos Jarque és Anil K. Bera által publikált teszt [Jarque and Bera, 1987] próbastatisztikája:

$$JB = \frac{n}{6}\left(\left(\sqrt{b_1}\right)^2 + \frac{(b_2-3)^2}{4}\right),$$

ami alapján akkor vetjük el, hogy a minta normális, ha a $JB > \chi^2_{\alpha,2}$, ugyanis a normális eloszlás nullhipotézise mellett a Jarque-Bera próbastatisztika eloszlása 2 szabadságfokú χ^2-eloszlás.

Ez abból következik, hogy a próbastatisztika két aszimptotikusan független standard normális négyzetösszege. Normális eloszlás esetén a tapasztalati ferdeség $\sqrt{n}$-nel skálázott $\sqrt{nb_1}$ értéke aszimptotikusan normális eloszlású $\nu = 0$ és $\sigma^2 = 6$ paraméterekkel, a tapasztalati lapultság $\sqrt{n}$-nel skálázott $\sqrt{n}(b_2 - 3)$ értéke pedig aszimptotikusan normális eloszlású $\nu = 0$ és $\sigma^2 = 24$ paraméterekkel. Ezekből következik, hogy a $\frac{n}{6}(\left(\sqrt{b_1}\right)^2$ és a $\frac{n(b_2-3)^2}{24}$ standard normális eloszlásúak, és négyzetösszegük pedig 2 szabadságfokú χ^2-eloszlású [Frain, 2007].

A próbastatisztika χ^2-eloszlással való közelítése csak nagy méretű minták esetén pontos, 2000-nél kisebb mintákra a kritikus értékeket szimulációkkal becsülik meg.

4. Box-Cox típusú transzformációk

Számos statisztikai módszer, többek között a főkomponens-analízis, a variancia-analízis és a t-próbák egyik feltétele, hogy az adataink normális eloszlásúak legyenek. Gyakorlatban ez általában nem teljesül, ám megfelelő módszer segítségével az adatok megközelítőleg normálissá transzformálhatók.

Adatok transzformálása alatt azt értjük, hogy minden egyes adaton ugyanazt a matematikai műveletet hajtjuk végre, például amikor a hőmérsékleti adatokat Celsius-fokról Fahrenheitre alakítjuk. Ez lineáris transzformációnak tekinthető, mert az eredeti adatokat egyszerűen megszorozzuk egy konstanssal (1, 8-cal), és ehhez hozzáadunk egy másik számot (32-t). A lineáris transzformációk azonban nem változtatják meg az adatok alakját, és ezért nem segítenek azok normálissá tételében.

A normálissá transzformálási módszereknek rengeteg fajtája van, ilyenek például a hagyományos módszerek, mint a négyzetgyök-transzformáció vagy az inverz-transzformáció [Osborne, 2010].

A négyzetgyök-transzformáció minden értéknek a négyzetgyökét veszi. Mivel negatív számokból nem tudunk gyököt vonni, egy megfelelő konstanst hozzáadva előnyösen 0 és 1 közé transzformáljuk az adatainkat. A 0 és 1 közötti számok ugyanis másképp viselkednek, mint a 0, az 1 és az 1-től nagyobb számok. Míg a 0 és az 1 négyzetgyöke marad 0 és 1, az 1 fölötti számok négyzetgyöke mindig kisebb lesz, addig a 0 és 1 közötti számok négyzetgyöke nő. Emiatt ezt a transzformációt nem előnyös folytonos eloszlású adatokra alkalmazni, melyek 0 és 1 közötti, illetve 1-nél nagyobb értéket is felvehetnek, ugyanis máshogy alakítja át a kettőt. A Poisson-eloszlású adatok transzformálása esetén viszont a négyzetgyök-transzformáció egy jónak tartott módszer.

Az inverz-transzformáció az adatokat az $\frac{1}{x}$ függvény segítségével az inverzükre alakítja át. Ezzel a nagyon kicsi adatokat nagyon naggyá, a nagyon nagyokat pedig nagyon kicsivé teszi, megfordítva az adatok sorrendjét.

John Wilder Tukey, amerikai statisztikust szokás az elsőnek tekinteni, akinek 1957-ben definiált transzformációja [Tukey, 1957] hasonló matematikai függvények egy családjának tekinthető. Ha $x > 0$, akkor a transzformáció alakja a következő:

$$x^\lambda = \begin{cases} x^\lambda & \text{ha } \lambda \neq 0 \\ \log x & \text{ha } \lambda = 0 \end{cases}$$

Ebben a függvényben λ – és a későbbiekben is – alkalmasan megválasztott valós paraméternek tekintendő. George Box és Sir David Cox 1964-es publikációjukban [Box and Cox, 1964] módosították a Tukey-féle transzformációcsaládot. Máig az ő

nevükhöz fűződik az egyik legnépszerűbb és leggyakrabban használt transzformáció, az ún. Box-Cox-transzformáció, melynek eredeti képlete az alábbi:

$$y = x^{\lambda} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{ha } \lambda \neq 0 \\ \log x & \text{ha } \lambda = 0 \end{cases} \tag{1}$$

Ugyanezen munkában Box és Cox definiálja a transzformáció egy másik alakját is, ami már képes kezelni a negatív megfigyeléseket is. Ekkor az előbbi λ paraméter egy $\lambda = (\lambda_1, \lambda_2)'$ paramétervektor lesz, a transzformáció pedig

$$y = x^{\lambda} = \begin{cases} \frac{(x+\lambda_2)^{\lambda_1}-1}{\lambda} & \text{ha } \lambda \neq 0 \\ \log(x+\lambda_2) & \text{ha } \lambda = 0 \end{cases}$$

Gyakorlatban a λ_2 megválasztható úgy,

hogy $x + \lambda_2 > 0$ minden x esetén, és ekkor kizárólag a λ_1 tekintendő a modell paraméterének.

A λ értékét úgy kell megválasztanunk, hogy az adatok minél inkább normálisnak tűnjenek. Az ismeretlen paraméter értéke megbecsülhető például maximum likelihood módszerrel [Li, 2005]. Az y megfigyelésekről feltesszük, hogy normális eloszlásúak ν és σ ismeretlen paraméterekkel. Ezek maximum likelihood (ML) becslései: $\hat{\nu}_{ML} = \overline{Y}$ és $\hat{\sigma}^2_{ML} = S^2_n$.

A ν és σ^2 paraméterű normális eloszlású $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. valószínűségi változó likelihood függvénye és log-likelihood függvénye

$L(\nu, \sigma) = f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\nu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\nu)^2}$ és

$\log L(\nu, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \nu)^2$.

A log-likelihood függvény utolsó tagjában lévő szumma az alábbi módon írható tovább:

$$\sum_{i=1}^{n}(x_i-\nu)^2=\sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\nu)^2=\sum_{i=1}^{n}((x_i-\overline{x})+(\overline{x}-\nu))^2=$$
$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+2\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-\nu)+\sum_{i=1}^{n}(\overline{x}-\nu)^2=$$
$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+2\cdot(\overline{x}-\nu)\sum_{i=1}^{n}(x_i-\overline{x})+n(\overline{x}-\nu)^2=$$
$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+2\cdot(\overline{x}-\nu)\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{\sum_{i=1}^{n}x_i-n\overline{x}=n\overline{x}-n\overline{x}=0}+n(\overline{x}-\nu)^2=$$
$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\nu)^2$$

Ennek fényében az ML-becslést behelyettesítve a log-likelihood függvénybe, a következő kifejezést kapjuk:

$$\log L(\overline{x},s_n)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log s_n^2-\frac{1}{2s_n^2}\left[\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})^2}_{ns_n^2}+n(\underbrace{\overline{x}-\overline{x}}_{0})^2\right]=$$
$$=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log s_n^2-\frac{1}{(2s_n^2)}ns_n^2=-\frac{n}{2}\log(2\pi s_n^2)-\frac{n}{2}$$

Ebből pedig megkapjuk a likelihood-függvény egyszerűbb alakját, amit a továbbiakban felhasználunk:

$$L(\overline{x},s_n)=(2\pi s_n^2)^{-\frac{n}{2}}e^{-\frac{n}{2}}$$

A transzformáció λ paraméterének maximum likelihood-becsléséhez szükség van a transzformált minta sűrűségfüggvényére. Egy X valószínűségi változó $Y=g(X)$ függvényének sűrűségfüggvényét az alábbi jól ismert képlet segítségével tudjuk felírni: $f_Y(y)=|(g^{-1}(y))'|f_X(g^{-1}(y))$. Y esetünkben ν várható értékű és σ szórású normális eloszlású valószínűségi változó, és az $X=g^{-1}(Y)$ valószínűségi változó sűrűségfüggvényét szeretnénk felírni, amelyet a következő módon kapunk meg: $f_X(x)=|(g'(x))|f_Y(g(x))$.

Legyen a g függvény az (1)-ben definiált a Box-Cox-transzformáció, ekkor $g'(x)=\partial_x\frac{x^\lambda-1}{\lambda}=x^{\lambda-1}$. Az X sűrűségfüggvénye tehát $f_X(x)=x^{\lambda-1}f_Y(\frac{x^\lambda-1}{\lambda})$.

A ν és σ paraméterek ML-becslései függnek a λ paramétertől, és a transzformált sűrűségfüggvény segítségével már felírható a Box-Cox-transzformáció likelihood

függvénye:

$$L(\overline{x}, s_n, \lambda) = (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} x_i^{\lambda-1} = (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} e^{(\lambda-1)\log x_i}$$

Ennek logaritmusa segítségével pedig már megállapíthatjuk a λ paraméter ML-becslését:

$$\log L(\overline{x}, s_n, \lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log s_n^2(\lambda) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\log x_i \longrightarrow \max_{\lambda}$$

A szakirodalomban a Box-Cox-transzformációnak több módozatát is publikálták. Ezek egyike J.A. John és Norman R. Draper úgynevezett modulus-transzformációja [John and Draper, 1980], mely azon eloszlásokra a legeredményesebb, amelyek megközelítőleg szimmetrikusak:

$$y = x^{\lambda} = \begin{cases} \text{sign}(y)\frac{(|y|+1)^{\lambda}-1}{\lambda} & \text{ha } \lambda \neq 0 \\ \text{sign}(y)\log(|y|+1) & \text{ha } \lambda = 0 \end{cases}$$

A Peter J. Bickel és Kjell A. Docksum által módosított transzformáció [Bickel and Doksum, 1981] minden bemenő értékre alkalmazható, viszont a paraméter értéke szigorúan pozitív lehet:

$$y^{\lambda} = \frac{(|y|^{\lambda}\text{sign}(y)-1)}{\lambda} \quad \text{ha } \lambda > 0$$

In-Kwon Yeo és Richard A. Johnson az ezredfordulón mutatták be cikkükben [Yeo and Johnson, 2000] az általuk létrehozott transzformációt, amely minden x megfigyelésre alkalmazható, nem csak szigorúan pozitív értékűekre, mint az eredeti Box-Cox-transzformáció.

A Yeo-Johnson-transzformáció a következő alakba írható:

$$y = x^{\lambda} = \begin{cases} \frac{(1+x)^{\lambda}-1}{\lambda} & \text{ha } x \geq 0, \lambda \neq 0 \\ \log(1+x) & \text{ha } x \geq 0, \lambda = 0 \\ -\frac{(1-x)^{2-\lambda}-1}{2-\lambda} & \text{ha } x < 0, \lambda \neq 2 \\ -\log(1-x) & \text{ha } x < 0, \lambda = 2 \end{cases}$$

Ha $x > 0$, akkor a Yeo-Johnson-transzformáció x-re ugyanaz, mint a Box-Cox-transzformáció $x + 1$-re. Ha pedig $x < 0$, x Yeo-Johnson-transzformációja at$(1 - x)$ értékekre $(2 - \lambda)$ paraméterű Box-Cox-transzformációval egyezik meg.

A λ paraméter becslését ezen transzformáció esetén is maximum likelihood módszerrel vezetem le, a Box-Cox-transzformációnál eljártakhoz hasonlóan.

Először legyen $x \geq 0$ és $\lambda \neq 0$. Ekkor a transzformált változó sűrűségfüggvénye $f_X(x) = (1 + x)^{\lambda-1} f_Y(\frac{(1+x)^{\lambda}-1}{\lambda})$.

A likelihood-függvény ekkor

$$L(\overline{x}, s_n, \lambda) = (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} (1+x_i)^{\lambda-1} = (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} (1+x_i)^{\lambda-1} =$$
$$= (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} e^{(\lambda-1)\log(1+x_i)},$$

melynek logaritmusából adódik a log-likelihood függvény:

$$\log L(\overline{x}, s_n, \lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log s_n^2(\lambda) - \frac{n}{2} + (\lambda-1)\sum_{i=1}^{n}\log(1+x_i),$$

így a λ ML-becslését a

$$-\frac{n}{2}\log s_n^2(\lambda) + (\lambda-1)\sum_{i=1}^{n}\log(1+x_i)$$

függvény maximalizálásával kapjuk meg.

Végül legyen $x < 0$ és $\lambda \neq 2$. . Ekkor a transzformált változó sűrűségfüggvénye $f_X(x) = (1-x_i)^{1-\lambda} f_Y(-\frac{(1-x)^{2-\lambda}-1}{2-\lambda})$.

Az ezzel kapott likelihood és log-likelihood függvények az alábbiak:

$$L(\overline{x}, s_n, \lambda) = (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} (1-x_i)^{1-\lambda} = (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} (1-x_i)^{1-\lambda} =$$
$$= (2\pi s_n^2(\lambda))^{-\frac{n}{2}} e^{-\frac{n}{2}} \prod_{i=1}^{n} e^{(1-\lambda)\log(1-x_i)}$$

$$\log L(\overline{x}, s_n, \lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log s_n^2(\lambda) - \frac{n}{2} + (1-\lambda)\sum_{i=1}^{n}\log(1-x_i)$$

A λ ML-becslése ekkor a

$$-\frac{n}{2}\log s_n^2(\lambda) + (1-\lambda)\sum_{i=1}^{n}\log(1-x_i)$$

függvény maximuma λ szerint.

A $\log L(\lambda) = -\frac{n}{2}\log s_n^2(\lambda) + (\lambda-1)\sum_{i=1}^{n}\mathrm{sign}(x)\log(1+|x_i|)$ alakban felírt függvény mindkét fenti esetet magában foglalja.

5. A próbák összehasonlítása - szimulációk

Ma már 40-nél is több teszt létezik, amelyik alkalmas normalitás tesztelésére. Ezek különféle esetekben: különböző mintaméretekre, eloszlásokra más-más erővel tudnak működni. Mivel a kutatók általában a kutatásuknak megfelelő legerősebb tesztet szeretnék használni, számos olyan publikáció született, amelyben valamilyen módon összehasonlítják a próbákat. A [Thadewald and Büning, 2007] cikkben a Jarque-Bera-, Kolmogorov-Szmirnov-, súlyozott Kolmogorov-Szmirnov-, Cramér-von Mises-, súlyozott Cramér-von Mises-, Kuiper- és a Shapiro-Wilk-próbák erejét vetik össze a szerzők Monte-Carlo szimulációval.
Az $F = (1-q)U + qV$, ahol $U \sim N(0, 1^2)$, $V \sim N(\nu, \sigma^2)$, $0 < p < 1$ modell segítségével különböző ν, σ és q értékekre, valamint különféle n mintaméretekre 10000 mintát generáltak. A modell használatát azzal indokolták, hogy ezzel az eloszlások széles körét fel tudják ölelni, szimmetrikusakat és asszimetrikusakat egyaránt. A próbastatisztikákat a pontos kritikus értékekkel vetették össze.

Szimmetrikus esetekben, kis q értékekre a Jarque-Bera-teszt teljesített a legjobban, de q növelésével csökken a próba ereje. Nagyobb q értékekre a Kuiper és a súlyozott Cramér-von Mises nyújtotta a legjobb teljesítményt. A Shapiro-Wilk a többi vizsgált teszthez képest nem tekinthető erős tesztnek a szimmetrikus eloszlásokra. Aszimetrikus esetekben, kis q értékekre szintén a Jarque-Bera-próba teljesítménye a legjobb, de a q növelése jelentős csökkenést eredményez a próba erejében, $q = 0,5$-re már a leggyengébben teljesítő teszt a súlyozott Kolmogorov-Szmirnov-próba mellett. Az aszimmetrikus eloszlásokra a Shapiro-Wilk- és a súlyozott Cramér-von Mises-próba teljesítettek a legjobban.

Összegezve, a Jarque-Bera-tesztet szimmetrikus, illetve enyhén ferde, elnyúló eloszlások esetén nagyon jól teljesítőnek találták, nem elnyúló eloszlások esetén viszont nagyon gyenge, sőt sokszor torzított is. Helyette a Cramér-von Mises-, súlyozott Cramér-von Mises- vagy a Shapiro-Wilk-próba használatát javasolják. A Kolmogorov-Szmirnov-, súlyozott Kolmogorov-Szmirnov-, Cramér-von Mises- és a súlyozott Cramér-von Mises-próbákat becsült ν és σ paraméterekkel nagyon konzervatívnak találták az eredeti adatok kritikus értékeit használva, standardizált adatokra szignifikánsan jobban teljesítettek.

A [Razali and Wah, 2011] cikkben a Shapiro-Wilk-, az Anderson-Darling-, a Kolmogorov-Szmirnov- és a Lilliefors-próbák erejét hasonlították össze, ugyancsak Monte-Carlo szimulációk segítségével. Tízezer mintát generáltak különböző mintaméretekre szimmetrikus és aszimmetrikus eloszlásokból. A megfelelő kritikus értékeket minden mintaméretre és teszthez 50.000 standard normális eloszlásból generált mintából állapították meg Monte-Carlo szimulációval. Két szignifikanciaszinten vé-

gezték a próbákat, $\alpha = 0,05$ és $0,1$ azt is tesztelve, hogy ezeknek mekkora hatásuk van a próbák erejére. Hét szimmetrikus eloszlást és hét aszimmetrikus eloszlást vizsgáltak, amelyek változatos ferdeségi és csúcsossági mutatókkal rendelkeznek.

Arra jutottak, hogy szimmetrikus eloszlásnál, ha 3-nál kisebb lapultság értéke, a Shapiro-Wilk-próba teljesítménye jobb, mint a másik 3 teszté, ugyanakkor 30 vagy annál kisebb mintaméret esetén mind a négy teszt ereje 40%-nál kisebb. 3-nál nagyobb csúcsosság esetén szintén a Shapiro-Wilk-próba teljesít a legjobban, kis mintára pedig szintén mindegyik próba gyenge. Szimmetrikus eloszlásokra tehát a legerősebb próba a Shapiro-Wilk, majd az Anderson-Darling, Lilliefors és Kolmogorov-Szmirnov. Aszimmetrikus eloszlásoknál is felülmúlja a Shapiro-Wilk-teszt ereje a másik háromét. Míg az már legalább 50-es mintaméretnél is jól teljesít, addig az Anderson-Darling- és Lilliefors-próbáknak legalább 100-as nagyságú minta kell, hogy jó teljesítményt érjen el. A Kolmogorov-Szmirnov a leggyengébb, és sokkal nagyobb mintaméretre van szükség, hogy a másik hárommal összevethető erőt érjen el.

Általánosságban az a következtetés vonható le, hogy a Shapiro-Wilk-próba a legerősebb, a Kolmogorov-Szmirnov-próba a leggyengébb. Az Anderson-Darling-próba teljesítménye nagyon hasonlít a Shapiro-Wilkéhez. Ugyanakkor az eredményekből az is leszűrhető, hogy mind a négy teszt ereje alacsony kis mintamérereteken esetén.

A [Thode, 2002] könyv 7. fejezetében további eredmények megtalálhatók a különböző normalitástesztek összehasonlításáról.

Szakdolgozatomban a N(0,1), Exp(2), t_7, Cauchy(0,1)[2] és E($-\sqrt{3}$,$\sqrt{3}$) eloszlásokra, $n = 50, 100, 200, 500, 1000$ mintaméretekre a χ^2-, a Kolmogorov-Szmirnov-, az Anderson-Darling-, a Shapiro-Wilk- és a Jarque-Bera-próbák teljesítményét hasonlítottam össze az **R** statisztikai programmal, Monte-Carlo-szimulációval. Mivel a χ^2-próba aszimptotikus próba, csak megfeleően nagy mintaelemszámra használható a χ^2-eloszlás alkalmas kvantilise kritikus értéknek. Ezt a próbát csak 50-nél nagyobb mintákra vizsgáltam. 100000 véletlen mintát generálva megfigyeltem a különféle eloszlásokra különböző mintaméretek esetén az átlagos p-értéket, a medián p-értéket és a próbák elutasítási hányadát. A medián p-értékeket azért figyeltem meg az átlagos értékek mellett, hogy nem befolyásolja-e valamilyen kiugró érték az átlag eredményét. A p-értékek alapján akkor veti el egy teszt a nullhipotézist, ha a p-érték kisebb, mint 0,05. A próbák elutasítási hányadát úgy számoltam, hogy a 0,05-nél kisebb p-értékek számát viszonyítottam az összes elemszámhoz. Ezzel a mennyiséggel próbák erejét becsülöm. Mivel folytonos eloszlásokat vizsgáltam, a χ^2-próba előtt előbb diszkretizálást hajtottam végre.

[2]Standard Cauchy, azaz aminek a sűrűségfüggvénye $f(x) = \frac{1}{\pi}\frac{1}{x^2+1}$ minden $x \in \mathbb{R}$ esetén.

A fentebb összefoglalt cikkekből kiindulva arra számítottam, hogy a Shapiro-Wilk-próba fogja a legjobb teljesítményt nyújtani, az Anderson-Darling- és a Jarque-Bera-teszt hozzá hasonlóan jó eredményeket ér majd, a Kolmogorov-Szmirnov- és a χ^2-próba pedig gyenge erőt mutat majd.

Mivel a legtöbb próbának az **R**-ben többféle implementációja is létezik, a 2. táblázatban összefoglaltam, melyik csomagban található változataikat használtam. Az fBasics csomagban található χ^2-próba a `classes=ceiling(2*(n^{(2/5)}))` képlettel határozza meg az osztályok számát.

Próba	Csomag	Parancs
Shapiro-Wilk	stats	shapiro.test(x)
Jarque-Bera	tseries	jarque.bera.test(x)
Anderson-Darling	nortest	ad.test(x)
Cramér-von Mises	nortest	cvm.test(x)
Lilliefors	nortest	lillie.test(x)
χ^2	fBasics	pchiTest(x)
Kolmogorov-Szmirnov	stats	ks.test(x,y)

2. táblázat. A próbák próbákat tartalmazó csomagok és a próbák **R**-parancsai.

N(0,1) eloszlásból generált véletlen mintáknál azt figyeltem meg, hogy az öt próba elutasítási részarányai a várt 5 százalékhoz képest hogy alakultak a különböző mintaméretek esetén. A legjelentősebben a χ^2-teszt elutasítási részarányai térnek el az elvárttól minden mintaméret esetén.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	4,5	5,0	5,0	3,8
100	7,4	4,6	4,9	4,9	4,2
200	7,4	4,8	4,9	5,0	4,5
500	7,1	4,8	5,0	5,0	4,8
1000	7,1	4,8	4,9	5,2	4,9

3. táblázat. Az elutasítás részaránya százalékban kifejezve különböző méretű N(0,1) minta esetén az egyes próbafajtákra.

Exp(2) eloszlásnál, a 200-nál nagyobb mintaméretek esetén már mindegyik próba 0 átlagos(4. táblázat) és medián(5. táblázat) p-értékkel elveti a minta normalitását, az esetek 100 százalékában. Az Anderson-Darling, Shapiro-Wilk-, és Jarque-Bera-próbáknál már 100-as mintaméretnél is a helyzet áll elő, míg a Kolmogorov-Szmirnov- és a χ^2-próba, bár ugyanúgy elvetik a nullhipotézist 92 és 83 százalékban,

nagyobb (0,017 és 0,043) p-értékekkel. 50-es mintaméretnél a Kolmogorov-Szmirnov-minta már csak 38 százalékban tudja elvetni a nullhipotézist. Mind az átlagos, mind a medián p-értéke (0,113 és 0,074) alapján azt a következtetést lehet levonni, hogy nem tudja elvetni, hogy az exponenciális minta normális eloszlású.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,113	0,001	0,000	0,007
100	0,043	0,017	0,000	0,000	0,000
200	0,004	0,000	0,000	0,000	0,000

4. táblázat. Becsült átlagos p-értékek különböző méretű Exp(2) minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,074	0,000	0,000	0,000
100	0,000	0,008	0,000	0,000	0,000
200	0,000	0,000	0,000	0,000	0,000

5. táblázat. Becsült medián p-értékek különböző méretű Exp(2) minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	37,8	99,7	100	95,4
100	82,8	91,9	100	100	100
200	98,0	100	100	100	100

6. táblázat. Az elutasítás részaránya százalékban kifejezve különböző méretű Exp(2) minta esetén az egyes próbafajtákra.

t_7 eloszlású mintánál a tesztek 500-as mintaelemszám alatt szignifikánsan gyenge teljesítményt nyújtottak. A χ^2-próba medián p-értékei alapján el tudta vetni, hogy a 100-as és 200-as nagyságú minták normális eloszlásúak, de az átlagos p-értékek alapján csak 200-as mintaméretre. A Komogorov-Szmirnov-próba még 1000-es mintanagyság esetén sem tudja elvetni a normalitást se az átlagos, se a medián p-értékei alapján. 200-as mintaméret esetén mindössze 1 százalékban utasítja el a nullhipotézist, 500-as mintára alig az esetek 5 sázalékában, 1000-es mintára pedig 19 százalékban. Az átlagos p-értékek alapján a Shapiro-Wilk, az Anderson-Darling

és Jarque-Bera csak 500 és nagyobb mintaméret esetén, a medián p-értékek szerint a Shapiro-Wilk és a Jarque-Bera már 200-as mintára is elveti, hogy normális a minta.

A [Razali and Wah, 2011] cikkben a szerzők szintén használták a t_7-eloszlást az Anderson-Darling-, Shapiro-Wilk-, Kolmogorov-Szmirnov- és Lilliefors-próbák erejének összevetésére . A publikáció 28. oldalán található táblázatban szereplő értékek és a 9. táblázatban összefoglalt saját eredményeim között legfeljebb néhány tizedesrendű eltérés figyelhető meg.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,673	0,359	0,360	0,397
100	0,125	0,669	0,284	0,269	0,268
200	0,039	0,619	0,179	0,152	0,135
500	0,001	0,476	0,046	0,027	0,018
1000	0,000	0,293	0,004	0,001	0,001

7. táblázat. Becsült átlagos p-értékek különböző méretű t_7 minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,713	0,296	0,288	0,378
100	0,001	0,707	0,184	0,139	0,103
200	0,000	0,642	0,069	0,024	0,003
500	0,000	0,463	0,003	0,000	0,000
1000	0,000	0,256	0,000	0,000	0,000

8. táblázat. Becsült medián p-értékek különböző méretű t_7 minta esetén az egyes próbafajtákra.

A Cauchy(0,1) eloszlás esetén az Anderson-Darling-, Shapiro-Wilk- és Jarque-Bera-próbák már 50-es mintaméret esetén is közel minden esetben elvetették a nullhipotézist, nagyobb mintára pedig 100 százalékban. A Kolmogorov-Szmirnov-próba 50-es mintaméretnél még sem az átlagos, sem a medián p-értéke alapján nem tudta elvetni a normalitást, az elutasítási részaránya mindössze 45 százalék. Nagyobb mintákra már, bár gyengébben, mint a másik három, el tudta vetni. A χ^2-próba teljesít a leggyengébben, 200-as mintaméretnél, ahol a többi teszt már minden esetben elveti a minta normalitását, a χ^2-próba csak 92 százalékban tudja elvetni.

Az $\mathrm{E}(-\sqrt{3},\sqrt{3})$ eloszlású minta esetén 50-es mintanagyságra az átlagos p-értékek alapján csak a Shapiro-Wilk-próba tudja elvetni a normalitást, a medián p-értékek

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,3	18,6	23,2	26,6
100	6,3	0,4	28,0	36,7	44,0
200	9,5	1,0	44,9	57,7	66,1
500	18,2	4,5	80,7	89,9	93,4
1000	100	18,6	98,0	99,4	99,7

9. táblázat. Az elutasítás részaránya százalékban kifejezve különböző méretű t_7 minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,104	0,001	0,001	0,002
100	0,042	0,025	0,000	0,000	0,000
200	0,000	0,001	0,000	0,000	0,000

10. táblázat. Becsült átlagos p-értékek különböző méretű Cauchy(0,1) minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,055	0,000	0,000	0,000
100	0,000	0,009	0,000	0,000	0,000
200	0,000	0,000	0,000	0,000	0,000

11. táblázat. Becsült medián p-értékek különböző méretű Cauchy(0,1) minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	44,1	99,7	99,6	99,4
100	76,7	84,8	100	100	100
200	92,4	99,9	100	100	100

12. táblázat. Az elutasítás részaránya százalékban kifejezve különböző méretű Cauchy(0,1) minta esetén az egyes próbafajtákra.

szerint a Shapiro-Wilk és az Anderson-Darling-próba is. A 200-nál nagyobb méretű mintákra a χ^2-próba medián p-értékétől eltekintve, pedig már minden próba el tudja vetni, hogy a minta normális eloszlású. A χ^2-próba teljesített a leggyen-

gébben, még 200-as nagyságú mintáknál is mindössze az esetek 70 százalékában vetette el a nullhipotézist. Ezzel szemben a legjobb teljesítményt nyújtó Shapiro-Wilk-próba már 50-es nagyságú mintáknál 75 százalékban el tudta vetni, hogy az egyenletes eloszlásból származó minta normális eloszlást követne. Meglepő módon a Jarque-Bera-teszt teljesítménye 50-es mintaméretnél szignifikánsan gyengébb volt, mint a többi próbáé, mindössze az esetek 0,2 százalékában tudta elvetni a nullhipotézist, 0,2-nél nagyobb átlagos és a medián p-értékkel. Nagyobb mintaelemszámokra a Kolmogorov-Szmirnov-próbához hasonlóan teljesített.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,210	0,088	0,044	0,221
100	0,190	0,077	0,011	0,003	0,054
200	0,071	0,012	0,000	0,000	0,003
500	0,002	0,000	0,000	0,000	0,000

13. táblázat. Becsült átlagos p-értékek különböző méretű $\mathrm{E}(-\sqrt{3},\sqrt{3})$ minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	0,141	0,035	0,018	0,202
100	0,094	0,034	0,002	0,001	0,046
200	0,012	0,003	0,000	0,000	0,002
500	0,000	0,000	0,000	0,000	0,000

14. táblázat. Becsült medián p-értékek különböző méretű $\mathrm{E}(-\sqrt{3},\sqrt{3})$ minta esetén az egyes próbafajtákra.

	A végrehajtott próba				
n	χ^2	KS	AD	SW	JB
50	-	25,6	57,5	75,1	0,2
100	39,0	59,1	94,9	99,7	56,4
200	70,2	94,7	100	100	100
500	99,0	100	100	100	100

15. táblázat. Az elutasítás részaránya százalékban kifejezve különböző méretű $\mathrm{E}(-\sqrt{3},\sqrt{3})$ minta esetén az egyes próbafajtákra.

Röviden még egyszer összefoglalom egyes eloszlások esetén a próbák teljesítményéről tett megfigyeléseket:

Standard normális eloszlású minta esetén a tesztek elutasítási részarányai az elvárt 5 százalék körül mozognak, kiemelkedően csak a Jarque-Bera-próbánál kisebbek értékek kis minták esetén.

Exponenciális minta esetén tehát minden mintaméretre a Shapiro-Wilk-próba bizonyult a legerősebbnek, az esetek 100 százalékában elveti, hogy az exponenciális minta normális lenne. Az Anderson-Darling-próba szinte ugyanolyan teljesítményt nyújtott, hasonlóképpen az azt követő Jarque-Bera-teszt is. 100-as mintaméretnél a χ^2-teszt bizonyult a leggyengébbnek. 50-es mintaméretnél pedig a Kolgomorov-Szmirnov-próba a teljesítménye a legrosszabb, az már nem is veti el a normalitást.

t_7 eloszlásnál a Jarque-Bera bizonyult a legerősebbnek, aztán a Shapiro-Wilk és az Anderson-Darling. A Kolmogorov-Szmrinov teljesített a leggyengébben, se kicsi, se nagy mintára nem tudta elvetni a normalitást, a χ-négyzet egy kicsit jobb teljesítményt nyújtott, de a másik három erejéhez képest szignifikánsan gyenge.

Cauchy eloszlás esetén a Shapiro-Wilk, Anderson-Darling és Jarque-Bera egyformán erősen teljesített, a Kolmogorov-Szmirnov már gyengébben, 50-es minta esetén nem vetette el a normalitást, a leggyengébb pedig a χ^2-próba volt.

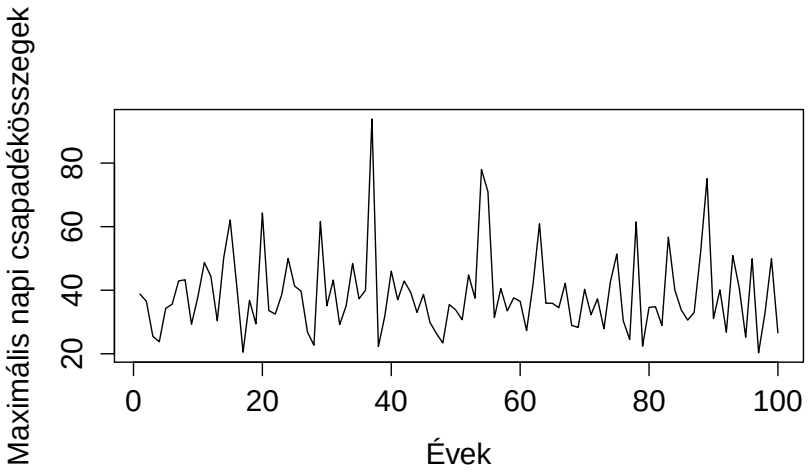
Egyenletes eloszlásnál is a Shapiro-Wilk-próba bizonyult a legjobban teljesítőnek. Minden mintaméret esetén el tudta vetni az egyenletes minta normalitását. Az Anderson-Darling-próba 50-es mintaméretnél még jelentősen gyengébben teljesített, mint a Shapiro-Wilk-próba, de nagyobb mintaméretekre már hasonlóan erős teljesítményt nyújt. A Jarque-Bera teszt, kis mintára, jelentősen gyengébb teljesítményt nyújtott, mint a többi teszt. A leggyengébben a χ^2-próba teljesített, 200-nál kisebb mintaméretekre nem tudta elvetni, hogy normális a minta.

6. Alkalmazások

Ebben a fejezetben az eddig bemutatott módszereket - a grafikus teszteket, a normalitásvizsgálati próbákat és a transzformációkat - valódi adatsorokra alkalmazom az **R** statisztikai program segítségével. A normalitástesztek közül a Shapiro-Wilk(SW)-, Jarque-Bera(JB)-, Anderson-Darling(AD)-, Cramér-von Mises(CVM)-, és χ^2-próbákat használtam, és mivel ismeretlen paraméterű normális eloszláshoz hasonlítom az adatokat, a Kolmogorov-Szmirnov- és Lilliefors-próbák közül a Lilliefors-próbát. A transzformációk paraméterének becslése a car csomag `powerTransform(x)` paranccsal számolható ki az x mintára. Alapbeállításként a Box-Cox-transzformáció paraméterét határozza meg, a Yeo-Johnson-transzformáció paraméterét a `powerTransform(x,family="yjPower")` módosítással kapjuk meg. A kapott lambdák segítségével, a csomag `bcPower(x,lambda)` és `yjPower(x,lambda)` parancsaival hajtottam végre a transzformációkat.

6.1. Csapadékadatok

Az első felhasznált adatsor a maximális napi csapadékösszeget tartalmazza az évben, Budapesten, 1901 és 2000 között.[3] Az 4. ábrán a maximális napi csapadékösszegek évek függvényében való ábrázolása látható.

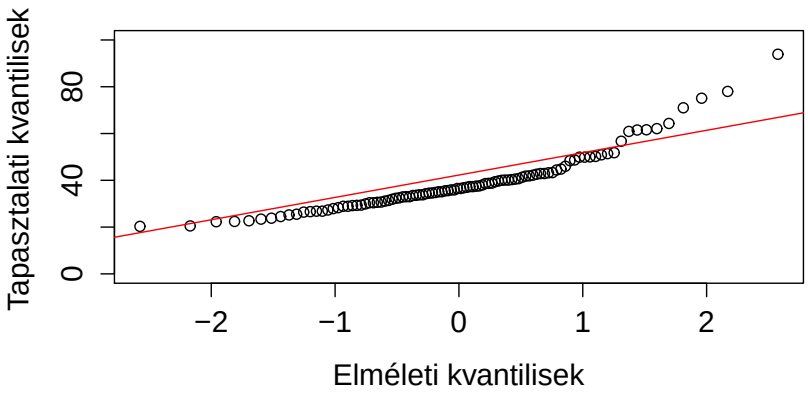


4. ábra. A csapadék adatok ábrája az évek függvényében

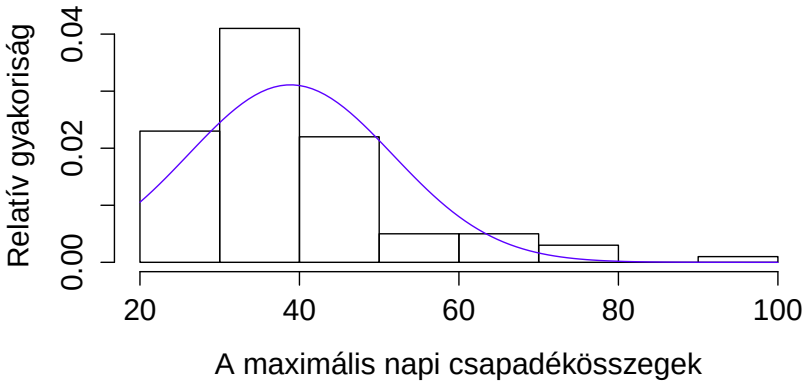
Az adatok Q-Q plotjáról (5. ábra) azt a következtetetés szűrhető le, hogy az adatok nem normális eloszlásúak, hiszen az adatok mintegy fele nem simul rá a 45 fokos egyenesre, leginkább az eloszlás szélein vannak nagyon távol az adatok. A hisztogram

[3]Forrás: `http://owww.met.hu/eghajlat/eghajlati_adatsorok/bp/Navig/Index2.htm` [2015.05.29.]

(6. ábra) alapján is azt gondolhatjuk, hogy az maximális napi csapadékösszegek nem normális eloszlást követnek, ugyanis a hisztogram alakja nem követi a haranggörbe alakját.



5. ábra. A csapadék adatok Q-Q plotja



6. ábra. A csapadék adatok hisztogramja

A grafikus módszerekkel szerzett előzetes benyomást, mely szerint az adatok normalitása elvetendő, a normalitástesztek is megerősítik. A 16. táblázatban összefoglaltam az egyes próbák próbastatisztikáinak értékét, a kapott p-értéket és a döntést, hogy a kapott p-érték alapján elfogadjuk vagy elvetjük a nullhipotézist. A nullhipotézist, azaz a normalitást akkor vetjük el, ha a kapott p-érték kisebb, mint 0,05, ezt a táblázatban H_1-el jelöltem, a nullhipotézis elfogadását pedig H_0-val.
Az eredeti adatokra mindegyik teszt nagyon kicsi p-értékkel veti el a nullhipotézist.

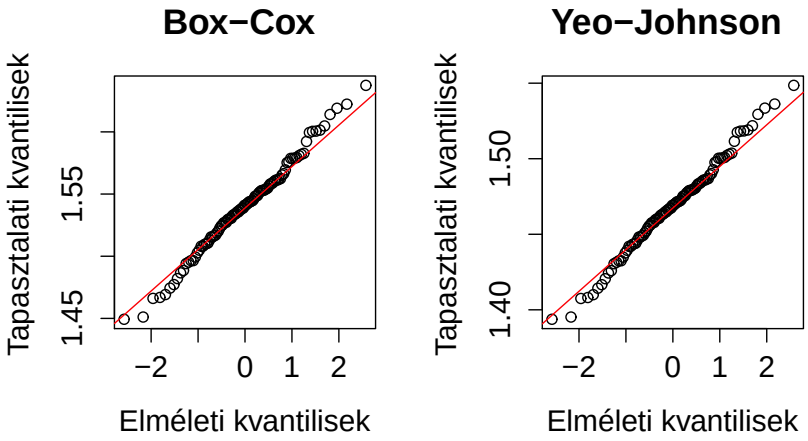
A nem normális eloszlású adatokat a Box-Cox- és a Yeo-Johson-transzformációval megpróbáltam normálissá transzformálni. Először mindkét transzformáció paramé-

próba	Próbastatisztika	p-érték	Döntés
SW	0,8812	2,037e-07	H_1
JB	88,9897	$<$ 2,2e-16	H_1
AD	2,9739	1,599e-07	H_1
CVM	0,5004	2,59e-06	H_1
LF	0,145	2,119e-05	H_1

16. táblázat. A tesztek próbastatisztikái, p-értékei és a döntés a csapadék adatokról.

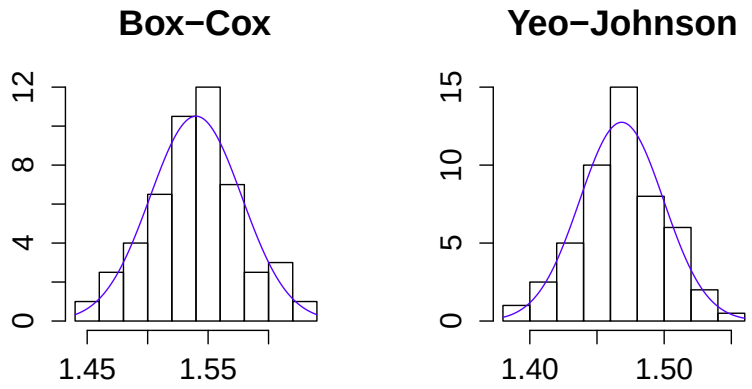
terének megbecsültem az értékét, majd ezekkel transzformáltam az adatokat. A transzformált adatokra aztán elvégeztem minden vizsgálatot, amelyet az eredeti adatokra is. A Box-Cox-transzformáció λ paraméterének becsült értéke -0,563, a Yeo-Johnson-transzformáció λ paraméterének becsült értéke -0,605.

Mindkét transzformált minta Q-Q plotján (7. ábra) az látható, hogy a transzformált adatok az eredetieknél sokkal jobban simulnak az egyenesre, a Q-Q plot alapján mindkettő normálisnak mondható. A hisztogramokon szintén érzékelhető a változás, mind a Box-Cox-, mind a Yeo-Johson-transzformációval transzformált minta hisztogramja (8. ábra) már jól közelíti a harangörbe alakját az eredeti adatokkal ellentétben.



7. ábra. A transzformált csapadék adatok Q-Q plotjai

A transzformált adatokról, mindkét transzformáció esetén, a próbák nem tudják elvetni, hogy normális eloszlásúak (17. táblázat). A p-értékek nagyok, különösen a Jarque-Bera-teszté, amelynek egyik p-értéke közel 1, a másik 1. Mivel a Jarque-Bera-teszt a minta ferdeségének és lapultságának egy függvénye alapján határozza meg a próbastatisztikát, ellenőriztem a transzformált adatok ezen értékeit. A normalitás nullhipotézisének teljesülése mellett a ferdeség 0, a lapultság 3 értéket vesz



8. ábra. A transzformált csapadék adatok hisztogramjai

fel. A Box-Cox-transzformációval transzformált adatok ferdesége 0,0002 és lapultsága 3,0067. A Box-Cox-transzformációval transzformált adatok ferdesége 0,0007 és lapultsága 2,9986. Mind a négy adat csupán néhány ezrednyivel tér el a normális megfelelőitől, így elfogadható, hogy a Jarque-Bera ilyen magas p-értékkel fogadja el a normalitást.

	Box-Cox			Yeo-Johnson		
Próba	Próbastatisztika	p-érték	Döntés	Próbastatisztika	p-érték	Döntés
SW	0,9928	0,8727	H_0	0,9928	0,8745	H_0
JB	2e-04	0,9999	H_0	0	1	H_0
AD	0,2449	0,7551	H_0	0,2419	0,7650	H_0
CVM	0,0396	0,6848	H_0	0,0390	0,6961	H_0
LF	0,0577	0,5680	H_0	0,0575	0,5754	H_0

17. táblázat. A tesztek próbastatisztikái, p-értékei és a döntés a csapadékadatokról a transzformációk után.

A 18. táblázatban a χ^2-próba próbastatisztikája, p-értéke, az osztályok száma és a döntés van összefoglalva a mintákra. Az eredeti adatok normalitását a χ^2-próba is elveti, a Box-Cox-transzformációval transzformált minta (BC) és Yeo-Johnson-transzformációval transzformált minta (YJ) normalitását pedig nem tudja elvetni.

Összegezve, a maximális napi csapadékösszegek eloszlása Budapesten nem normális, de mind a Box-Cox-, mind a Yeo-Johnson-transzformációval normálissá alakíthatóak az adatok.

Adatok	Próbastatisztika	p-érték	Osztályok száma	Döntés
eredeti	27,14	0,0074	13	H_1
BC	3,74	0,9877	13	H_0
YJ	5,30	0,9472	13	H_0

18. táblázat. A χ^2-próba próbastatisztikája, p-értéke, az osztályok száma és a döntés az eredeti és transzformált csapadékadatokról.

6.2. Happy Planet Index

A második vizsgált minta a 2012-es Happy Planet Index adatokat tartalmazza.[4] A New Economics Foundation, melynek célja, egy mind az ember, mind a bolygó érdekeit figyelembe vevő gazdaság létrehozása, 2006-ban bevezette a Happy Planet Indexet(HPI, magyarul: Boldog Bolygó Index), mint az emberi jóllét elérésének ökológiai hatékonyságát mérő mutatószámát. Központi gondolata az volt, hogy a korábbi mutatók, mint a GDP vagy a HDI(Human Development Index) nem tartalmaznak szubjektív adatokat, egy nemzet sikerét az emberek boldogsága és jólléte helyett a produktivitással mérik. A HPI 151 ország adatait veti össze a három különálló mérőszám alapján:

- Experienced well-being (WB, magyarul: Tapasztalati jóllét)
  Az emberek válaszaiból állapították meg. Azt kérték a válaszadóktól, képzeljék el az életüket, mint egy létrát, ahol 0 jelképezi a lehető legrosszabb életminőséget, 10 a lehetséges legjobbat, és jelöljék meg, hogy a jelenlegi életüket a létra melyik fokán érzik.
- Life expectancy (LE, magyarul: Várható élettartam)
  A 2011 UNDP Human Development Report várható élettartam adatait használták fel.
- Ecological Footprint (FP, magyarul: Ökológiai lábnyom)
  A WWF ökológiai lábnyom adatait használták fel, mint az erőforrás-felhasználás mérőszámát.

A HPI-t az alábbi módon számolják:

$$HPI = \frac{WB \cdot LE}{FP}$$

Magyarország a 2012-es felmérés szerint a 151 ország közül a 104. helyen áll. A HPI mutató értéke 37,4, amely a 4,7-es tapasztalati jóllét, 74,4-es várható élettartam és a 3,6-os ökológia lábnyom értékéből tevődik össze.

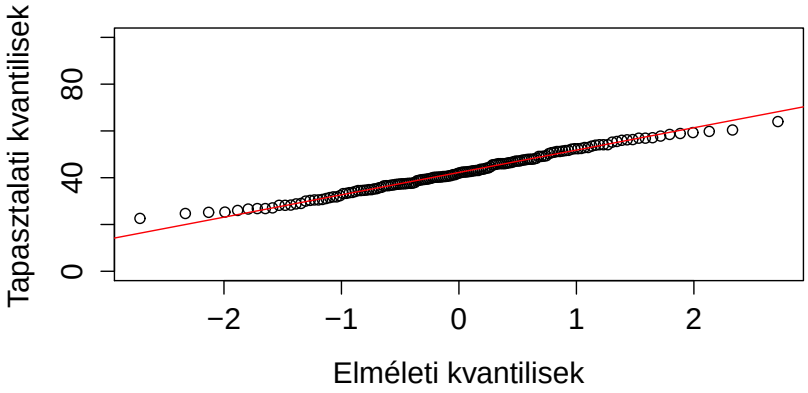
A HPI első tíz helyét meglepő módon nem fejlett, gazdag országok foglalják el, hanem főként közép-amerikai államok, mint Costa Rica, Jamaica, vagy Guatemala.

[4]Forrás:`http://www.happyplanetindex.org/data/` [2015.05.29.]

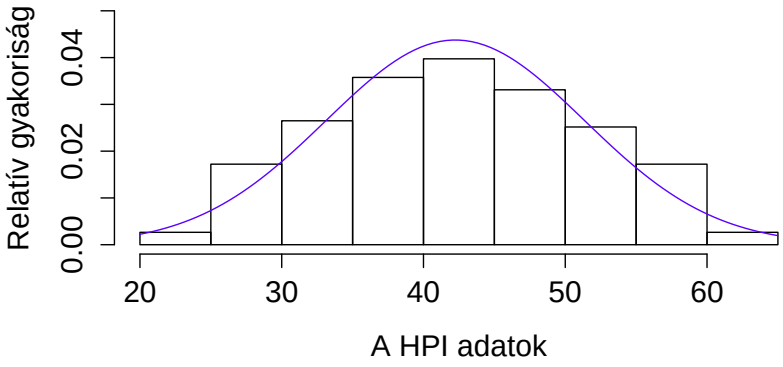
A listavezető Costa Ricának közel kétszer akkora a HPI-értéke mint Magyarországnak.

Először azt vizsgáltam meg, hogy a HPI tekinthető-e normális eloszlásúnak.

A Q-Q ploton (9. ábra) az látható, hogy az adatok túlnyomó részt a 45 fokos egyenesre simulnak. Ebből arra következtetünk, hogy az adatok vélhetően normális eloszlásúak. Hasonló következetésre juthatunk a hisztogram (10. ábra) alakját viszgálva, amelyről elmondható, hogy követi a haranggörbe alakját.



9. ábra. A HPI adatok Q-Q plotja.



10. ábra. A HPI adatok hisztogramja.

Az előzetes benyomást a normalitásról a próbák újra megerősítik, mivel minden próba p-értéke nagyobb, mint 0,05, ahogy az a 19. táblázatban is látható, a HPI adatokat normális eloszlásúnak tekinthetők.

A χ^2-próba 15 osztályba sorolja az adatokat, próbastatisztikája 8,0397, p-értéke 0,8872, tehát ez a próba sem veti el a nullhipotézist.

Próba	Próbastatisztika	p-érték	Döntés
SW	0,9882	0,2320	H_0
JB	3,1006	0,2122	H_0
AD	0,3550	0,4563	H_0
CVM	0,0478	0,5407	H_0
LF	0,0423	0,7300	H_0

19. táblázat. A tesztek próbastatisztikái, p-értékei és a döntés a HPI adatokról.

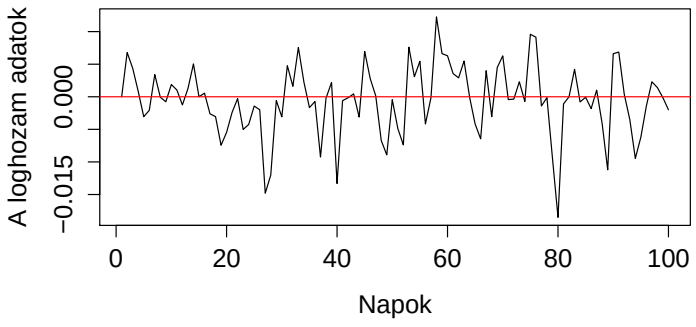
Mivel normális eloszlásúnak tekinthetők az adatok, a transzformációkat nem alkalmaztam rájuk.

6.3. Valutaárfolyam adatok

A harmadik vizsgált mintát a 2011.08.14. és 2014.08.14. közötti euró-dollár valutaárfolyam adatokból[5] számolt loghozam adatok alkotják, Ezeket a következőképpen számoltam ki: $\log(\frac{x_{i+1}}{x_i})$ (x_i az intervallumbeli i. napi valutaárfolyam).

A 11. ábrán a 2011.08.14. – 2011.11.25. időszak közötti napok függvényében vannak a loghozam adatok ábrázolva. Az ábráról azt olvashatjuk le, hogy a adatok eloszlása szabálytalan, nem periodikus.

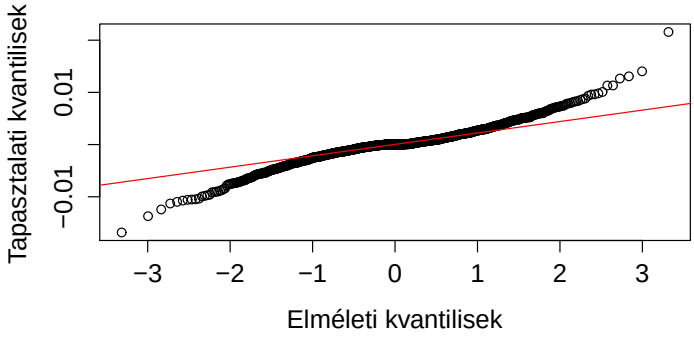
Az adatok hisztogramja az 13. ábrán azt mutatja, hogy a hisztogram nem követi a kék standard normális eloszlású görbe sűrűségfüggvényének alakját a medián adatok környékén. A Q-Q plot a 12. ábrán is arra enged következtetni, hogy az adatok nem normális eloszlásúak, mivel nem simulnak rá a pirossal jelölt 45 fokos egyenesre.



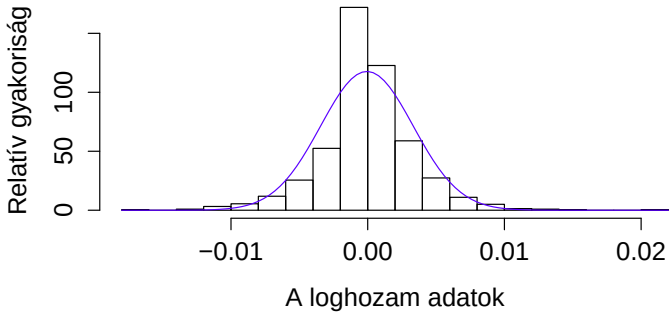
11. ábra. A loghozam adatok ábrája a napok függvényében.

A minta normalitását mindegyik normalitásteszt elveti nagyon kicsi p-értékkel (20. táblázat).

[5]Forrás:`http://www.oanda.com/currency/historical-rates/` [2015.05.29.]



12. ábra. A loghozam adatok Q-Q plotja



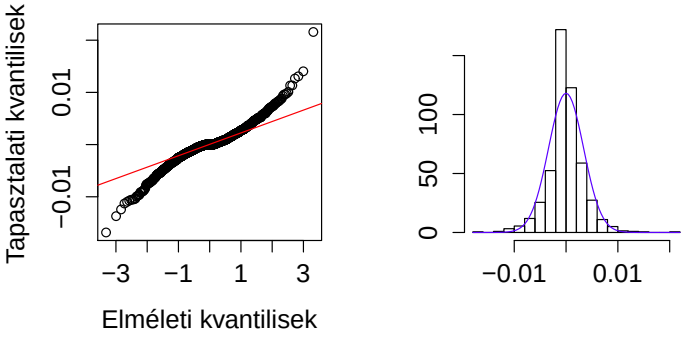
13. ábra. A loghozam adatok hisztogramja

Próba	Próbastatisztika	p-érték	Döntés
SW	0.9436	$<$ 2.2e-16	H_1
JB	574.1004	$<$ 2.2e-16	H_1
AD	20.0245	$<$ 2.2e-16	H_1
CVM	3.9978	$<$ 7.37e-10	H_1
LF	0.0951	$<$ 2.2e-16	H_1

20. táblázat. A tesztek próbastatisztikái, p-értékei és a döntés az loghozam adatokról.

Mivel az adatok nem mind szigorúan pozitívak, az **R** nem tudja végrehajtani a beépített Box-Cox-transzformációt, így csak a Yeo-Johnson-transzformációt tudtam alkalmazni, amely λ paraméterének becsült értéke 11,434. A transzformált adatok Q-Q plotjáról és hisztogramjáról (14. ábra) arra tudunk következtetni, hogy a Yeo-Johnson-transzformáció nem tudta normálissá alakítani az adatokat.

Mint az a 21. táblázatban látható, a p-értékek nem változtak, a transzformáció után továbbra is jelentősen kicsi maradt mindegyik, az alkalmazott próbák a



14. ábra. A Yeo-Johsnson-transzformációval transzformált loghozam adatok Q-Q plotja és hisztogramja.

transzformált adatok normalitását is elvetik.

	Yeo-Johnson		
Próba	Próbastatisztika	p-érték	Döntés
SW	0.9449	$<$ 2.2e-16	H_1
JB	619.8426	$<$ 2.2e-16	H_1
AD	19.4286	$<$ 2.2e-16	H_1
CVM	3.8988	$<$ 7.37e-10	H_1
LF	0.0891	$<$ 2.2e-16	H_1

21. táblázat. A tesztek próbastatisztikái, p-értékei és a döntés a loghozam adatokról a Yeo-Johnson-transzformáció után.

A 22. táblázat azt illusztrálja, hogy a loghozam adatok normalitását a χ^2-próba is egyértelműen elveti mind az eredeti, mind a transzformált esetben.

Adatok	Próbastatisztika	p-érték	Osztályok száma	Döntés
eredeti	495.347	$<$ 2.2e-16	33	H_1
YJ	557.251	$<$ 2.2e-16	33	H_1

22. táblázat. A χ^2-próba próbastatisztikája, p-értéke, az osztályok száma és a döntés az eredeti és transzformált loghozam adatokról.

Tehát sem az eredeti adatok, sem a Yeo-Johnson-transzformációval transzformált minta nem tekinthető normális eloszlásúnak. A loghozam adatok eloszlása leginkább GARCH(1,1) folyamattal modellezhető.

7. Összefoglalás

Összefoglalásként röviden áttekintjük a szakdolgozatban vizsgáltakat. Először bemutattam a két leggyakoribb grafikus módszert, a hisztogramot és a Q-Q plotot. Ezek, bár rendkívül hasznosak, mert képet adnak az adatok eloszlásának alakjáról, ugyanakkor pontos információval nem szolgálnak. Ezután a harmadik fejezetben részletesen bemutattam néhány jelentősebb normalitásvizsgálati próba felépítését és működését a teljesség igénye nélkül. A legtöbb tesztet az 1950-es és 1980-as évek között mutatták be, a Jarque-Bera-teszt 1987-es megjelenése óta csak néhány új teszt jelent meg, de ezek nem tudtak olyan áttörő eredménnyel szolgálni, mint például a szakdolgozatban ismertetettek. Erről a témáról Thode könyve remek áttekintést nyújt. A negyedik fejezetben azt mutattam meg, hogy a nem normális eloszlású adatok hogyan transzformálhatók normálissá. Részletesen a talán legismertebb Box-Cox-transzformációt, és annak egy módosítását, a Yeo-Johnson transzformációt ismertettem. A Box-Cox-transzformáció a transzformációk fejlődésének egy meghatározó pontja, ugyanis az adatok már szélesebb körére alkalmazható, mint a megjelenése előtti transzformációk. A Yeo-Johson-transzformáció pedig kiküszöböli a Box-Cox-transzformáció hiányosságát, és a negatív adatokat is képes kezelni. Az ötödik fejezetben szimulációk segítségével hasonlítottam össze a próbákat. A saját eredményeim is igazolták, hogy a vizsgáltak közül a Shapiro-Wilk-próba tekinthető a legerősebbnek teljesítmény szempontjából. Az Anderson-Darling- és a Jarque-Bera-próba általában a Shapiro-Wilk-próbáéhoz hasonló teljesítményt nyújt, de azt is láthattuk, hogy a Jarque-Bera-teszt kis mintaelemszám esetén nem mindig tud pontos döntést hozni, nem tudja elvetni, hogy az egyenletes eloszlásból származó minta normális eloszlású. Végül a szakdolgozatban bemutatott módszereket három valós adatsorra is alkalmaztam. Az egyik esetben az eredeti adatok normális eloszlásúak voltak. A másik két adatsor eredetei adatai nem voltak normális eloszlásúak, ugyanakkor az egyik adatsort normálissá lehetett transzformálni.

A normalitás vizsgálatának számos különböző módszere létezik a szakdolgozatban bemutatott néhány példán kívül. Az új tesztek felfedezése mellett, fontos kérdés, hogy a már meglévő módszerek javítása lehetséges-e, mint ahogy azt a Kolmogorov-Szmirnov - Lilliefors pár példáján láttuk. Hiszen, ahogy Sir David R. Cox is mondta [Rosenkrantz, 2011]: "There are no routine statistical questions, only questionable statistical routines."[6]

[6] "Nincsenek rutinos statisztikai kérdések, csupán kérdéses statisztikai rutinok."

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Varga Lászlónak, hogy hasznos tanácsaival és iránymutatásával a szakdolgozatom elkészítéséhez nélkülözhetetlen segítséget nyújtott, és családomnak a támogatásukért.

8. Függelék

Ábrák jegyzéke

Táblázatok jegyzéke

Irodalomjegyzék

Theodore W. Anderson and Donald A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

Peter J. Bickel and Kjell A. Doksum. An analysis of transformations revisited. *Journal of the american statistical association*, 76(374):296–311, 1981.

Marianna Bolla and András Krámli. *Statisztikai következtetések elmélete*. Typotex Kft., 2012.

George E.P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

John C. Frain. Small sample power of tests of normality when the alternative is an α-stable distribution. *Trinity Economics Papers*, (207), 2007.

Carlos M. Jarque and Anil K. Bera. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172, 1987.

J.A. John and Norman R. Draper. An alternative family of transformations. *Applied Statistics*, pages 190–197, 1980.

Pengfei Li. Box-cox transformations: An overview. *presentation, http://www. stat. uconn. edu/˜ studentjournal/index_files/pengfi_ s05. pdf*, 2005.

Hubert W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

Tamás F. Móri. *Statisztikai hipotézisvizsgálat*. Typotex Kft., 2011.

Jason W. Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation*, 15(12):1–9, 2010.

Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

Walter A. Rosenkrantz. *Introduction to probability and statistics for science, engineering, and finance*. CRC Press, 2011.

Lothar Sachs. *Angewandte statistik*. Springer-Verlag, 2013.

Samuel Sanford Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, pages 591–611, 1965.

Saul Stahl. The evolution of the normal distribution. *Mathematics magazine*, pages 96–113, 2006.

M.A. Stephens. The goodness-of-fit statistic v_n: Distribution and significance points. *Biometrika*, pages 309–321, 1965.

Thorsten Thadewald and Herbert Büning. Jarque–bera test and its competitors for testing normality–a power comparison. *Journal of Applied Statistics*, 34(1):87–105, 2007.

Henry C. Thode. *Testing for normality*, volume 164. CRC press, 2002.

John W. Tukey. On the comparative anatomy of transformations. *The Annals of Mathematical Statistics*, pages 602–632, 1957.

In-Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.