Eötvös Loránd Tudományegyetem

Természettudományi Kar

Illeszkedésvizsgálati módszerek összehasonlítása

Szakdolgozat

Készítette:

Tóth Alexandra

Matematika BSc.

Matematikai Elemző szakirány

Témavezető:

Zempléni András

egyetemi docens

Valószínűségelméleti és

Statisztika Tanszék

Budapest

2015

Tartalomjegyzék

Bevezetés

Dolgozatom nemparaméteres próbákról szól, azon belül is az illeszkedésvizsgálatot taglalja. A statisztikának ezt az ágát az élet sok területén használják, mint például a pénzügyi szektorban vagy épp árvíz-előrejelzésnél.

Az 1. fejezetben az alapfogalmakat vezetem be. A definíciók logikai sorrendet követnek. A statisztikai próba fogalmától eljutok egészen az illeszkedésvizsgálat definíciójáig. Ezen fogalmak fontossága jelentős, mivel ezekkel dolgozom az azt követő fejezetben. Az 1. fejezet az irodalomjegyzék [3] és [4] könyve felhasználásával készült.

A 2. fejezetben ismertetem azon teszteket, amelyeket illeszkedésvizsgálatra használunk. Sokféle ilyen próba létezik, azonban csak a legfontosabbakat emeltem ki, mégpedig a χ^2 - [1], a Kolmogorov-Szmirnov- [1], a Cramér-von-Mises- [2],[5], az Anderson-Darling- [6] és a Lilliefors-tesztet [5]. Ezeket előbb kifejtem, tisztázom a hozzájuk kapcsolódó definíciókat, tételeket; a tételek közül néhány bizonyítását is bemutatom. Ezek után néhány teszthez írtok egy-egy példát is, ugyanis úgy vélem, ezzel könnyebben lehet szemléltetni az eljárásokat.

Az utolsó fejezetben az `R` nevű programcsomagot használom a 2. fejezetben említett próbák bemutatására és összehasonlítására.

1. fejezet

A *hipotézis*, vagy más néven feltevés jellemzően olyan jelenségek feltételezett magyarázata, illetve lehetséges törvényszerű összefüggések elképzelése, amelyeket még nem bizonyítottuk. A statisztikai vizsgálatok folyamán hipotéziseket konstruálunk, mint például a várható értékre, azonos eloszlásra vagy két valószínűségi változó függetlenségére. Az ilyen feltevéseket *statisztikai hipotéziseknek* nevezzük.

A matematikai statisztikának a *hipotézisvizsgálat* azon része, amely a statisztikai feltevések közötti döntésekhez kapcsolódó módszereket és azok elvi kérdéseit tanulmányozza. A feltevéseket statisztikai módszerek segítségével vizsgálhatjuk meg. Ezen eljárásokat *statisztikai próbáknak* nevezzük. A statisztikai próbák konkrét változókra vett minták, illetve mintaelemek egyes függvényének vizsgálatán alapulnak.

A következőkben bevezetünk néhány definíciót. *Nullhipotézisnek* vagy *alaphipotézisnek* nevezzük azt az esetet, amikor a feltevésünk igaz. Ezt H_0-lal jelöljük. Tegyük fel például, hogy az X valószínűségi változó várható értéke m_0, akkor ezt az esetet nullhipotézisként a következő módon írhatjuk le:

$$H_0 : M(X) = m_0 .$$

Az alaphipotézis ellentettje az *ellenhipotézis* vagy más néven *alternatív hipotézis*, amelyet H_1-gyel jelölünk, amely most ezt jelenti:

$$H_1 : M(X) = m \neq m_0 .$$

Ezen feltevéseket minden esetben aszerint kell létrehozni, hogy ha az egyik hipotézis teljesül, akkor a másik nem teljesülhet. Tehát H_0-nak és H_1-nek egymást kizáró feltevéseknek kell lenniük a vizsgált valószínűségi változóra vonat-

kozólag. Vagyis a hipotézisvizsgálat célja egy már kimondott feltevés helyességének ellenőrzése az aktuális statisztikai minta alapján és döntéshozatal arról, hogy a nullhipotézist elfogadjuk vagy elutasítjuk, vagyis az alternatív hipotézist fogadjuk el.

Hipotézisvizsgálat menete

Az X valószínűségi változó n elemű mintája legyen: $X_1, X_2, \ldots, X_n$, S pedig az n elemű minta lehetséges értékeinek halmaza. Készítsünk egy $h(X_1, X_2, \ldots, X_n)$ függvényt $(h: S \to \mathbb{R})$.

Definiáljunk egy $0 < \alpha < 1$ számot és egy olyan $T \subset \mathbb{R}$ *elfogadási tartományt*, melyre igaz, hogy $h(X_1, X_2, \ldots, X_n) \in T$ nagy valószínűséggel teljesül H_0 fennállásakor, tehát igaz a

$$P(h(X_1, X_2, \ldots, X_n) \in T | H_0) \geq 1 - \alpha,$$

egyenlőtlenség; vagyis az a feltevés, hogy a $h(X_1, X_2, \ldots, X_n)$ függvény értékei a H_0 esetén $1-\alpha$ -nál nagyobb valószínűséggel esnek a T tartományba. A próba *szignifikancia szintjének* az $1-\alpha$ értéket hívjuk.

Legyen $K = \mathbb{R} \setminus T$ *kritikus tartomány*; annak valószínűsége H_0 fennállásakor nagyon kicsi, hogy $h(X_1, X_2, \ldots, X_n) \in K$, azaz a h a kritikus halmazban van, tehát

$$P(h(X_1, X_2, \ldots, X_n) \in K | H_0) \leq \alpha.$$

A hipotézisvizsgálat végkimenetele kétféle lehet:

ha $h(X_1, X_2, \ldots, X_n) \in T$, ebben az esetben elfogadjuk a H_0 hipotézist;

ha $h(X_1, X_2, \ldots, X_n) \in K$, ekkor elutasítjuk a H_0 hipotézist, vagyis elfogadjuk a H_1 feltevést.

A döntésünk jó is, de hibás is lehet, mert az eredményt a kísérletek kimenetele határozza meg.

Elsőfajú hibát akkor ejtünk, ha a H_0 hipotézist visszautasítjuk pedig igaz volt, mert a h a K tartományában van. Ennek a következő a valószínűsége:

$$P(\textit{elsőfajú hiba}) = P(h(X_1, X_2, \ldots, X_n) \in K | H_0) \leq \alpha .$$

Másodfajú hibát ejtünk abban az esetben, ha a hibás H_0 feltevést elfogadjuk, mert a h a T tartományában van. Ennek valószínűsége:

$$P(\textit{másodfajú hiba}) =$$

$$= P(h(X_1, X_2, \ldots, X_n) \in T | H_1) = 1 - P(h(X_1, X_2, \ldots, X_n) \in K | H_1).$$

Ezek után már tudjuk definiálni az erőfüggvényt, amit γ-val jelöltünk. Ezt a következő alakban tudjuk felírni:

$$\gamma = 1 - P(\text{másodfajú hiba}) = 1 - P(h(X_1, X_2, \ldots, X_n) \in T | H_1).$$

A hibás döntések valószínűségét tudjuk csökkenteni mégpedig úgy, hogy az első- és a másodfajú hiba valószínűségét is csökkentjük. Ezt például megtehetjük a minta elemszámának növelésével.

A következőkben bevezetjük a *paraméteres* és a *nemparaméteres* statisztikai próbákat. Az elégséges statisztika definíciójából tudjuk, hogy ez a statisztika az ismeretlen paraméterhez tartozó összes információt tartalmazza, ami az eredeti mintában előfordul. Tehát a paraméteres problémáknál, ahol egy valós paraméterre vonatkozó állítások hozzák létre az egyes feltevéseket, a paraméterre vonatkozó elégséges statisztika segítségével adhatjuk meg a próbát.

Elsődlegesen kiemelendő, amikor a mintaelemek normális eloszlásúak. Az u-próbával tudjuk ellenőrizni azt az esetet, amikor ismert szórású, várható értékre vonatkozó $H_0: m = m_0$ hipotézist vizsgálunk. Az u-próba statisztikája a következő:

$$u = \sqrt{n} \ \frac{\bar{X} - m_0}{\sigma} ,$$

itt a σ a normális eloszlás szórása, m_0 pedig a várható értéke. Ha a H_0 teljesül, akkor a kifejezés standard normális eloszlású és így könnyen meg tudjuk határozni az adott szignifikancia szinthez tartozó kritikus tartományt.

Amikor azonban ismeretlen szórással állunk szemben t-próbát használhatunk. A próbastatisztika ebben az esetben a következő:

$$t_{n-1} = \sqrt{n}\,\frac{\bar{X} - m_0}{s_n^*},$$

ahol $\bar{X} = \sum_{i=1}^{n} X_i / n$ a mintaelemek átlaga, $s_n^{*2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$ a korrigált tapasztalati szórásnégyzet. A próbastatisztika ebben az esetben $n-1$ szabadságfokú t eloszlású a nullhipotézis teljesülése esetén.

Természetesen lehetséges olyan eset is, amikor nem egy mintára vonatkozik a feladat, hanem két minta összehasonlítása a célunk.

Nemparaméteres próbánál a hipotézisek nem az eloszlás valamely paraméterére vonatkoznak, mint a paraméteres próbánál, hanem például olyan esetben alkalmazhatjuk, amikor azt akarjuk eldönteni, hogy két valószínűségi változó azonos eloszlású vagy független-e. Ezen próbák alapja is olyan statisztika, melynek eloszlása meghatározható (legalábbis aszimptotikusan), ha a nullhipotézis teljesül. Tehát kizárólag az elsőfajú hibához alkalmas kritikus tartomány létrehozására törekedünk. Jellemzően nincsenek legerősebb próbák, de olyanok igen, amik elfogadhatóan jó tulajdonságokkal rendelkeznek; ilyen például, hogy konzisztensek, vagyis a másodfajú hiba a mintaelemszám növekedésével 0-hoz tart.

Hasonlóan a paraméteres próbánál itt is többféle vizsgálatot tudunk elkülöníteni. Az egyik ilyen az illeszkedésvizsgálat, amit majd a következő fejezetben láthatunk.

Homogenitásvizsgálat során azt elemezzük, hogy két valószínűségi változó azonos eloszlású-e, tehát az $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_m$ független azonos eloszlású, egymás között is független minták eredhetnek-e ugyanazon sokaságból. A

$$H_0 : P(X < x) = P(Y < x), \quad \forall\, x \in \mathbb{R}$$

nullhipotézist vizsgáljuk a tagadása vagyis a H_1 ellen.

A függetlenségvizsgálat célja, hogy eldöntse a kétdimenziós (X, Y) valószínűségi változó komponenseiről, hogy függetlenek-e. Diszkrét és folytonos esetben is fel tudunk írni olyan $A_1, A_2, \ldots, A_r$ és $B_1, B_2, \ldots, B_s$ teljes eseményrendszereket (ezek X illetve Y értékkészletéből kerülnek ki), amelyekkel a nullhipotézis így írható le:

$$H_0 : P(A_i \cap B_j) = P(A_i) \cdot P(B_j) \quad (i = 1, \ldots, r\,;\, j = 1, \ldots, s).$$

Az alternatív hipotézis esetén ez nem teljesül legalább egy i, j párra.

2. fejezet

2.1. χ^2 - próba

Legyen $A_1, \dots, A_k$ teljes eseményrendszer és

$$H_0: \; P(A_i) = p_i \qquad (i = 1, \dots, k) \tag{1}$$

ahol a $p_i > 0$, $\sum_{i=1}^{k} p_i = 1$ valószínűségek adottak.

Most végezzünk el n darab megfigyelést. Jelölje $\nu_1, \dots, \nu_k$ az $A_1, \dots, A_k$ esemény gyakoriságát $\left(\sum_{i=1}^{k} \nu_i = n \right)$. Ekkor H_0 fennállása esetén a $(\nu_1, \dots, \nu_k)$ valószínűségi változó polinomiális eloszlású:

$$P_{H_0}(\nu_1 = n_1, \dots, \nu_k = n_k) = \begin{cases} \dfrac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k}, & \text{ha } n_1 + \dots + n_k = n, \\ 0, & \text{különben.} \end{cases}$$

2.1. Tétel. [1] *Ha* $(\nu_1, \dots, \nu_k)$ *polinomiális eloszlású* n *és* $p_1, \dots, p_k$ $(p_i > 0)$ *paraméterekkel (tehát az (1)-beli* H_0 *fennállásakor), akkor* $n \to \infty$ *esetén*

$$U = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i} \;\to\; \chi^2(k-1), \tag{2}$$

ahol U $k-1$ *szabadságfokú* χ^2 *eloszlású.*

A tétel bizonyításához ki kell mondanunk egy újabb tételt és egy lemmát is. Ezek segítségével már be tudjuk bizonyítani a tétel állítását.

2.2. Tétel. [10] (a centrális határeloszlás tétel többdimenziós alakja) *Legyenek* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ *független, azonos eloszlású* p*-dimenziós véletlen vektorok, melyek* μ *várható érték vektora és* $\boldsymbol{C}$ *kovarianciamátrixa létezik (ez nem feltétlenül invertálható). Legyen* $\boldsymbol{S}_r = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_r$, $r = 1, 2, \ldots$. *Ekkor a standardizált részletösszegek sorozata, azaz az* $\frac{1}{\sqrt{r}}(\boldsymbol{S}_r - r\mu)$ *véletlen vektor sorozat eloszlása konvergál az* $N_p(0, \boldsymbol{C})$*-eloszláshoz, ha* $r \to \infty$.

Mielőtt bizonyítanánk a többváltozós centrális határeloszlás tételét, ki kell mondanunk egy segédtételt.

2.1. Lemma. [10] (többváltozós véletlen vektorokra az eloszlásuk konvergenciájának a jellemzéséről) *Adott* p*-változós* $\boldsymbol{X}^{(r)} = (X_1^{(r)}, \ldots, X_p^{(r)})$, $r = 1, 2, \ldots$, *véletlen vektorok egy sorozata. Ezen véletlen vektorok akkor és csakis akkor konvergálnak eloszlásban egy* $\boldsymbol{X}^{(0)} = (X_1^{(0)}, \ldots, X_p^{(0)})$ *véletlen vektorhoz, amennyiben koordinátáik bármely* $\sum_{k=1}^{p} a_k X_k^{(r)}$ *lineáris kombinációja eloszlásban konvergálnak a* $\sum_{k=1}^{p} a_k X_k^{(0)}$ *valószínűségi változóhoz.*

A többváltozós centrális határeloszlás tételét most már be tudjuk bizonyítani.

Bizonyítás. A lemma alapján elég annyit bizonyítanunk, hogy bármely $a_1, \dots, a_p$ szám p-esre az $S_r(a_1, \dots, a_p) = \frac{1}{\sqrt{r}} \sum_{k=1}^{p} a_k (S_r^{(k)} - ES_r^{(k)})$ kifejezés eloszlásban konvergál az $Y(a_1, \dots, a_p) = \sum_{k=1}^{p} a_k Y_k$ véletlen vektorhoz, ahol $(Y_1, \dots, Y_p)$ 0 várható értékű és $\boldsymbol{C}$ kovarianciamátrixú normális eloszlású véletlen vektor. Legyenek $Z_i = Z_i(a_1, \dots, a_p) = \sum_{k=1}^{p} a_k X_k^{(i)}$, $1 \le i \le r$, valószínűségi változók. Ebben az esetben $Z_1, \dots, Z_p$ független, egyforma eloszlású valószínűségi változók és $S_r = \sum_{i=1}^{r} (Z_i - EZ_i)$. Ezért a bizonyítandó állításunk következik az egyváltozós centrális határeloszlás tételéből független, egyforma eloszlású valószínűségi változók összegére és abból a megállapításból, hogy $Y(a_1, \dots, a_p)$ 0 várható értékű normális eloszlású valószínűségi változó és $D^2 Y(a_1, \dots, a_p) = D^2 Z_i$. □

2.2. Lemma. [1] *Legyen* $\boldsymbol{X} = (X_1, \dots, X_m)^T \sim N_m(\boldsymbol{0}, \boldsymbol{C})$, *ahol* $\boldsymbol{C}$ *nem feltétlenül invertálható. Akkor* $\sum_{f=1}^{m} X_f^2$ *előáll* $\sum_{f=1}^{m} \lambda_f Y_f^2$ *alakban, ahol* $\boldsymbol{Y} = (Y_1, \dots, Y_m)^T \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)$ m*-dimenziós standard normális eloszlású, továbbá a* $\lambda_1, \dots, \lambda_m$ *nemnegatív valós számok a* $\boldsymbol{C}$ *kovarianciamátrix sajátértékei.*

Bizonyítás. Vegyük az $\boldsymbol{X}$ kovarianciamátrix spektrálfelbontását, vagyis legyen normatartó és $\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U}=\boldsymbol{I}$, $\boldsymbol{C}=\boldsymbol{U}\Lambda\boldsymbol{U}^{T}$, ahol $\boldsymbol{U}$ a normált sajátvektorokat a sajátértékek sorrendjében tartalmazó ortogonális mátrix és $\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_k)$. Ismert, hogy $\boldsymbol{X}$ az $\boldsymbol{X}=\boldsymbol{U}\Lambda^{1/2}\boldsymbol{Y}$ alakban állítható elő; itt $\boldsymbol{Y}$ m-dimenziós standard normális eloszlású. Ebből következően

$$\sum_{f=1}^{m} X_f^2=\|\boldsymbol{X}\|^2=\|\boldsymbol{U}^T\boldsymbol{X}\|^2=\|\boldsymbol{U}^T\boldsymbol{U}\,\Lambda^{1/2}\boldsymbol{Y}\|^2=\|\Lambda^{1/2}\boldsymbol{Y}\|^2=\sum_{f=1}^{m}\lambda_f Y_f^2,$$

mivel az $\boldsymbol{U}^T$ ortogonális transzformáció normatartó. □

Ezen ismeretek tükrében már be tudjuk bizonyítani a **2.1. Tételt**.

Bizonyítás. Először is készítsük el a $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_k)^T$ polinomiális eloszlású véletlen vektor várható érték vektorát és kovarianciamátrixát. Ahhoz, hogy ezt elvégezzük, $\boldsymbol{\nu}$-t határozzuk meg független, azonos eloszlású, k-dimenziós indikátorváltozók összegeként.

Az $A_1,\ldots,A_k$ teljes eseményrendszer k-dimenziós indikátorváltozója az $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_k)^T$ vektor, melynek az a komponense 1, amelyik tagja a teljes eseményrendszernek bekövetkezett, a többi komponense 0. Az f-edik kísérlet $\boldsymbol{\varepsilon}^{(f)}=(\varepsilon_1^{(f)},\ldots,\varepsilon_k^{(f)})^T$ indikátorváltozójára tehát

$$\varepsilon_g^{(f)}=\begin{cases}1, & \text{ha az } f\text{-edik kísérletben } A_g \text{ következett be}\\ 0, & \text{különben}\end{cases}$$

$g=1,\ldots,k\,;\,f=1,\ldots,n$. Könnyen láthatjuk, hogy $\boldsymbol{\nu}=\sum_{f=1}^{n}\boldsymbol{\varepsilon}^{(f)}$.

Egy tetszőleges $\boldsymbol{\varepsilon}$ várható érték vektorának összetevői:

$$E(\varepsilon_g)=P(A_g)=p_g,\qquad g=1,\ldots,k$$

$\boldsymbol{\varepsilon}$ kovarianciamátrixának fődiagonálisában található elemek:

$$c_{gg}=D^2(\varepsilon_g)=E(\varepsilon_g^2)-E^2(\varepsilon_g)=p_g-p_g^2=p_g(1-p_g),\qquad g=1,\dots,k,$$

diagonálison kívüli elemei pedig a következők:

$$c_{hg}=c_{gh}=Cov(\varepsilon_g,\varepsilon_h)=E(\varepsilon_g\varepsilon_h)-E(\varepsilon_g)\cdot E(\varepsilon_h)=0-p_g p_h=-p_g p_h,$$

$1\le g<h\le k.$

Az $\boldsymbol{\varepsilon}$ véletlen vektor kovarianciamátrixa tehát

$$\boldsymbol{C}=\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_m \\ \vdots & & \ddots & \\ -p_1p_m & -p_2p_m & \cdots & p_m(1-p_m) \end{bmatrix}=\boldsymbol{PI_m-pp^T},$$

itt $\boldsymbol{P}=\mathrm{diag}(p_1,\dots,p_k)$ és $\boldsymbol{p}=(p_1,\dots,p_k)^T$. Az $\boldsymbol{\varepsilon}^{(i)}$-k független azonos eloszlásúak, ezért $D^2(\boldsymbol{\nu})=n\boldsymbol{C}$. Egyértelmű, hogy a $\boldsymbol{\nu}$ kovarianciamátrixa, mint a $\boldsymbol{C}$ mátrix szinguláris, vagyis nem invertálható négyzetes mátrix, mivelhogy az elemek között lineáris összefüggés van (például a sorösszegek 0-k).

Most felhasználjuk a **2.2. Tételt**: $\boldsymbol{P}^{-1/2}\boldsymbol{\varepsilon}^{(1)},\dots,\boldsymbol{P}^{-1/2}\boldsymbol{\varepsilon}^{(n)}$ független, azonos eloszlású változókra, melyek összege $\boldsymbol{S}_n=\boldsymbol{P}^{-1/2}\boldsymbol{\nu}$. A **2.2. Tétel** alapján, ha $n\to\infty$, akkor az $\frac{1}{\sqrt{n}}\boldsymbol{P}^{-1/2}(\boldsymbol{\nu}-E(\boldsymbol{\nu}))$ véletlen vektorsorozat eloszlásban konvergál egy $\boldsymbol{X}\sim N_k(\boldsymbol{0},\boldsymbol{P}^{-1/2}\boldsymbol{C}\boldsymbol{P}^{-1/2})$ valószínűségi változóhoz.

Az előzőekben felírt $\boldsymbol{C}$ mátrix konstrukciója miatt

$$\boldsymbol{P}^{-1/2}\boldsymbol{C}\boldsymbol{P}^{-1/2}=\boldsymbol{I_k}-\boldsymbol{Q},$$

itt a $\boldsymbol{Q}$ a $\boldsymbol{q}=(\sqrt{p_1},\dots,\sqrt{p_k})^T$ egységvektorral előállított $\boldsymbol{qq^T}$ diád. Tehát $\boldsymbol{Q}$ a $\boldsymbol{q}$ vektor irányára való 1-rangú vetítés, így tehát $\boldsymbol{I}_k-\boldsymbol{Q}$ a $\boldsymbol{q}$ ortogonális komplementerére való $k-1$-rangú vetítés mátrixa, aminek sajátértékei a következők: $\lambda_1=\dots=\lambda_{k-1}=1$ és $\lambda_k=0$.

Használjuk $\boldsymbol{X}$-re a **2.2. Lemmát**: $\sum_{f=1}^{k} X_f^2$ eloszlása azonos a $\sum_{f=1}^{k} \lambda_f Y_f^2$ eloszlásával, itt az Y_f-k független standard normális eloszlásúak, λ_f-k pedig az $\boldsymbol{I}_k - \boldsymbol{Q}$ mátrix fenti sajátértékei. Ebből következően $\sum_{f=1}^{k} X_f^2 \sim \chi^2(k-1)$.

□

Megjegyzés

- A határeloszlás nem függ a p_i értékektől csak k-tól.
- Csak nagy mintaelemszám esetén tekinthető a (2)-beli statisztika H_0 fennállása esetén χ^2-eloszlásúnak és a normális határeloszláshoz való konvergencia alkalmazhatóságához azt is meg kell követelni, hogy $\nu_i \geqslant 3$.
- Ha a relatív gyakoriságokra az $k_i = \nu_i / n$ jelölést vezetjük be, akkor a (2) képlet a következő alakú lesz:

$$n \sum_{i=1}^{k} \frac{(k_i - p_i)^2}{p_i}$$

Illeszkedésvizsgálat χ^2 - próbával

<u>Tiszta illeszkedésvizsgálat:</u>

- Diszkrét eset:

  Legyen az X diszkrét eloszlású valószínűségi változó értékkészlete véges: $x_1, \ldots, x_k$.

$$H_0 : P(X = x_i) = p_i \qquad (i = 1, \ldots, k),$$

H_1 pedig ennek az ellentéte. A χ^2-vel kifejezett próbastatisztika a (2)-beli, a kritikus tartomány pedig

$$\left\{(\nu_1, \ldots, \nu_k) : \chi^2 \geqslant \chi^2_\alpha(k-1)\right\},$$

ahol $\chi^2_\alpha(k-1)$ az $k-1$ szabadságfokú χ^2-eloszlású $1-\alpha$ kvantilise.

- Abszolút folytonos eset:

  Tegyük fel, hogy X abszolút folytonos eloszlású valószínűségi változó. A

$$H_0 : P(X < x) = F_0(x), \qquad \forall x \in \mathbb{R}$$

  nullhipotézist teszteljük, F_0 adott eloszlásfüggvény. Az ellenhipotézis, hogy nem F_0 az eloszlásfüggvény. Legyen $X_1, \ldots, X_n$ független azonos eloszlású minta és osszuk fel a számegyenest a

$$(-\infty = x_0\ ,\ x_1\ ,\ \ldots\ ,\ x_{k-1}\ ,\ x_k = \infty]$$

  osztópontokkal k részre! A felosztást úgy választjuk meg, hogy minden részbe essen néhány a konkrét mintarealizációkból. Legyen

$$\nu_i = |\{l : X_l \in [\ x_{i-1}, x_i\ )\}|, \quad i = 1, \ldots, k \tag{3}$$

  az i-edik részintervallumba eső mintaelemek száma,

$$p_i = P(X \in [\ x_{i-1}, x_i\ )) = F_0(x_i) - F_0(x_{i-1}), \qquad i = 1, \ldots, k \tag{4}$$

  pedig az i-edik részintervallumba való tartozás elméleti valószínűsége. A H_0 tesztelésére a (3), (4) mennyiségekkel számolt (2)-beli próbastatisztikát alkalmazzuk, jelölje ezt χ^2; a kritikus tartomány legyen

$$\{(\nu_1, \ldots, \nu_k) : \chi^2 \geq \chi^2_\alpha(k-1)\}.$$

  $\chi^2_\alpha(k-1)$ itt is a $k-1$ szabadságfokú χ^2-eloszlás $1-\alpha$ kvantilise.

Becsléses illeszkedésvizsgálat:

Hogyha diszkrét esetben a valószínűségek, abszolút folytonos esetben viszont az eloszlásfüggvény becsültek a próbastatisztika számolása éppúgy megy végbe, mint az előzőekben, csak a kritikus értéket adó χ^2-eloszlás szabadságfoka $f = k - 1 - b$, itt b a becsült paraméterek számát jelöli. Ebben az esetben maximum likelihood becslést használunk.

Példa. A következő példában véletlen számokat veszünk táblázatba rendezve, és a számok „véletlenségét” vizsgáljuk meg, vagyis azt, hogy a számsorozat egyenletes eloszlásból vett mintának tekinthető-e. Az 1. táblázatba 50 kétjegyű, véletlen választott számot írtunk nagyság szerint rendezve.

10	23	31	39	46	58	65	73	84	94
14	25	33	40	47	59	66	75	86	95
18	27	35	41	48	61	67	77	88	96
19	28	36	44	49	62	68	78	90	98
22	29	37	45	56	63	69	79	92	99

1.táblázat

i	Intervallum	n_i	p_i	$n\,p_i$	$\frac{(n_i - n\,p_i)^2}{n\,p_i}$
1	00-29	10	3/10	15	1,666
2	30-49	14	2/10	10	1,6
3	50-69	11	2/10	10	0,1
4	70-99	15	3/10	15	0
		50	1	50	$\chi^2 = 3{,}366$

2. táblázat

A nullhipotézis az, hogy a $\{0,1,\ldots,99\}$ számokon egyenletes az eloszlás. A 2. táblázatban az intervallumot 4 részre osztottuk. Az itt elvégzett számítások alapján a $\chi^2=3{,}366$ értéket kaptuk. Ha az α-t 0,05-nek választjuk, akkor a próba szintje $1-\alpha=0{,}95$. A szabadsági fok, $r=4$ miatt, $r-1=3$ lesz. A χ^2-eloszlás kvantiliseit tartalmazó táblázat alapján a $\chi^2_{3;0{,}95}=7{,}815$ értéket kapjuk. Tehát $\chi^2=3{,}366$ nem esik a kritikus

$$K=\{\chi^2\geq 7{,}815\}$$

tartományba, ezért a nullhipotézisünket elfogadjuk, tehát az adott számok egyenletes eloszlásból vett mintának tekinthetők.

2.2. Kolmogorov-Szmirnov – próba

Mostantól olyan próbákat nézünk, amelyek az itt közölt egyszerű alakjukban folytonos eloszlások tesztelésére alkalmasak.

Illeszkedésvizsgálat

$$H_0 : P(X < x) = F_0(x), \qquad \forall x \in \mathbb{R}$$

- Kétoldali alternatíva:

  $$H_1 : \exists x \in \mathbb{R}, \text{ hogy } P(X < x) \neq F_0(x).$$

  Legyen

  $$D_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F_0(x)|,$$

  ahol F_n^* az n-elemű mintához tartozó empirikus eloszlásfüggvény.

  Ha $x_1^* \leq ... \leq x_n^*$ az $\mathbf{x} = (x_1 \leq ... \leq x_n)$ minta sorba rendezett alakja, akkor

  $$D_n(\boldsymbol{x}) = \max_i \max \{|F_n^*(x_i^*) - F_0(x_i^*)|, |F_n^*(x_i^* + 0) - F_0(x_i^*)|\} =$$

  $$= \max_i \max \left\{ \left| \frac{i-1}{n} - F_0(x_i^*) \right|, \left| \frac{i}{n} - F_0(x_i^*) \right| \right\}.$$

  Kolmogorov tétele alapján tudjuk, hogy H_0 fennállásakor

  $$\lim_{n \to \infty} P(\sqrt{n}\, D_n < z) = K(z), \qquad \forall z \in \mathbb{R},$$

  ahol

  $$K(z) = \begin{cases} 0, & \text{ha} \quad z \leq 0 \\ \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 z^2} = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2}, & \text{ha} \quad z > 0 \end{cases}$$

a Kolmogorov-függvény. Tehát ha n elég nagy, akkor az $1-\alpha$ szinthez tartozó kritikus tartomány:

$$\{\boldsymbol{x} : \sqrt{n}\, D_n(\boldsymbol{x}) \geq k_\alpha\},$$

ahol k_α a Kolmogorov-eloszlás $1-\alpha$ kvantilise.

- Egyoldali alternatíva:

$$H_1 : F(x) \geq F_0(x) \text{ és létezik } x_0, \text{ hogy } F(x_0) \geq F_0(x_0)$$

Legyen

$$D_n^+ = \sup_{x \in \mathbb{R}} \left(F_n^*(x) - F(x)\right) = \max_{1 \leq i \leq n} \left(F_n^*(x_i^* + 0) - F(x_i^*)\right) = \max_{1 \leq i \leq n} \left(\frac{i}{n} - F(x_i^*)\right).$$

A Szmirnov-tétel alapján tudjuk, hogy H_0 fennállásakor

$$\lim_{n \to \infty} P\left(\sqrt{n}\, D_n^+ < z\right) = S(z), \quad \forall\, z \in \mathbb{R},$$

ahol

$$S(z) = \begin{cases} 0, & \text{ha } z \leq 0, \\ 1 - e^{-2z^2}, & \text{ha } z > 0, \end{cases}$$

a Szmirnov-eloszlásfüggvény.

<u>Kétmintás eset</u>

$$H_0 : P(X < x) = P(Y < x), \quad \forall\, x \in \mathbb{R}$$

- Kétoldali alternatíva

$$H_1 : \exists\, x \in \mathbb{R}, \text{ hogy } P(X < x) \neq P(Y < x)$$

Legyen

$$D_{n,m} = \sup_{x \in \mathbb{R}} |F_n^*(x) - G_m^*(x)|,$$

ahol $F_n^*(x)$ az $X_1, \ldots, X_n$, illetve $G_m^*(x)$ az $Y_1, \ldots, Y_m$ olyan empirikus eloszlásfüggvények, amelyek egymástól is független, önmagukon belül független azonos eloszlású mintákból származnak. Elegendő az egyesített

rendezett mintaelemeknél vizsgálni a két empirikus eloszlásfüggvény távolságának maximumát.

2.3. Tétel

$$\lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}|F_n^*(x)-G_m^*(x)|<z\right)=K(z),\quad \forall z\in\mathbb{R}.$$

A **2.3. Tétel** alapján, ha n és m elég nagy, akkor $\sqrt{\frac{nm}{n+m}}D_{n,m}$ eloszlásfüggvénye „közel" a Kolmogorov-függvény, tehát ismét annak segítségével készítjük el a kritikus tartományt.

- Egyoldali alternatíva

$$H_1:\exists x\in\mathbb{R}\text{, hogy } P(X<x)>P(Y<x)$$

Legyen most

$$D_{n,m}^{+}=\sup_{x\in\mathbb{R}}(F_n^*(x)-G_m^*(x)).$$

2.4. Tétel

$$\lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}F_n^*(x)-G_m^*(x)<z\right)=S(z),\quad \forall z\in\mathbb{R}.$$

A **2.4. Tétel** alapján, ha n és m elég nagy, akkor $\sqrt{\frac{nm}{n+m}}D_{n,m}^{+}$ eloszlás függvénye „közel" Szmirnov-féle, tehát ezt az eloszlásfüggvényt használjuk az $1-\alpha$ kvantilis definiálásához.

Tulajdonságok

Az egymintás és a kétmintás próba is alkalmas kis elemszámú mintákra ellentétben a χ^2-val.

Fontos előnye az, hogy eloszlásfüggetlen, és nem csak a normális eloszlás tesztelésére alkalmas. A próbastatisztika az összes folytonos eloszlásra ugyanazt az eloszlást követi, ezért sok helyen alkalmazható. Nagy hátránya azonban, hogy az ereje kicsi. Lehetséges változata a Cramér-von-Mises-teszt, amely egy- és kétmintás esetre is használható, vagy az Anderson-Darling-próba.

Példa. Vegyünk két mintát, melynek elemeit a 3. táblázatba írjuk le véletlen sorrendbe.

X	24,3	10,5	13,4	18,6	11,9
Y	20,3	26,4	23,3	14,5	17,8
X	9,4	15,6	11,1	19,4	28,2
Y	13,1	19,9	18,8	5,9	27,4

3.táblázat

A következő táblázatban a mintákat nagyság szerint rendeztük. A z_i azt jelöli, hogy az adott mennyiség az X illetve az Y mintából származik-e. Ebből az adatból már következik a megfelelő ϑ_i érték, ami vagy $+1$ vagy -1 lehet. Tehát a ϑ_i érték azt jelöli, hogy ha az adott érték az X mintából származik, akkor $+1$ értéket vesz fel, ha pedig Y-ból -1-et. Ezután a részletösszegek (s_i) következnek.

	i	z_i	ϑ_i	s_i		i	z_i	ϑ_i	s_i
5,9	1	Y	-1	-1	18,4	11	X	$+1$	3
9,4	2	X	$+1$	0	18,8	12	Y	-1	2
10,5	3	X	$+1$	1	19,4	13	X	$+1$	3
11,1	4	X	$+1$	2	19,9	14	Y	-1	2
11,9	5	X	$+1$	3	20,3	15	Y	-1	1
13,1	6	Y	-1	2	23,3	16	Y	-1	0
13,4	7	X	$+1$	3	24,3	17	X	$+1$	1
14,5	8	Y	-1	2	26,4	18	Y	-1	0
15,6	9	X	$+1$	3	27,4	19	Y	-1	-1
17,8	10	Y	-1	2	28,2	20	X	$+1$	0

4. táblázat

Ezek után látható, hogy

$$10\mathrm{D}^{+} = \max s_i = 3 \qquad 10\mathrm{D}^{-} = |\min s_i| = 1,$$

így a statisztika értéke

$$10\mathrm{D}_{10,10} = 10 \max\left(D^{+}, D^{-}\right) = 3.$$

Ha az α-t 0,05-nek választottuk, akkor a kritikus érték a kvantilis táblázat alapján $D_{10,10;0,95} = 6$, ezért a két eloszlásfüggvény egyezését 95%-os szinten elfogadjuk.

2.3. Cramér-von-Mises – próba

A Cramér-von-Mises-teszt nagyon hasonló a Kolmogorov-Szmirnov-teszthez, azonban nem a szuprémum eltéréseit használja mint az előző tesztstatisztika, hanem a négyzetes eltéréseket, mint bázist veszi figyelembe. A tesztstatisztika pontos eloszlása, mint a Kolmogorov-Szmirnov-tesztstatisztikánál nem függ az elméleti eloszlás speciális alakjától.

A teszt hipotézisei a következők:

$$H_0 \colon P(X<x) = F_0(x)$$

$$H_1 \colon P(X<x) \neq F_0(x)$$

Tesztstatisztika

$$C^2 = n\int_{-\infty}^{+\infty} (F_n^*(x) - F_0(x))^2 f_0(x)\,dx$$

Kiszámítása

$$C^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(F_0(x_i^*) - \frac{2i-1}{2n}\right)^2$$

Amikor ellenőrizzük a statisztika helyességét a kritikus érték attól függ, hogy az eloszlás paramétereit megadtuk-e vagy a mintából kell azokat becsülni.

Tegyük fel, hogy F_0 normális eloszlás. Ha a paraméterei ismertek a következő közelítő kritikus érték érvényes:

$$C^2_{n,1-\alpha} = C^*_{1-\alpha}\cdot\left[1+\frac{1}{n}\right]^{-1} + 0{,}4/n - 0{,}6/n^2$$

Becsült paraméterekre pedig így számíthatjuk ki:

$$C^2_{n,1-\alpha} = C^{**}_{1-\alpha}\cdot\left[1+\frac{1}{n}\right]^{-1}$$

Mindkét kifejezésben a C^* és a C^{**} n -től független.

Az utolsó lépés a döntési szabály meghatározása. Ha $C^2 \geq C^2_{n,1-\alpha}$ egyenlőtlenség teljesül, akkor a nullhipotézist elvetjük.

Kétmintás teszt

Legyenek X és Y folytonos eloszlású valószínűségi változók F, illetve G eloszlásfüggvénnyel. Itt is a szokásos nullhipotézist vizsgáljuk, amely most

$$H_0: \ P(X<x)=P(Y<x).$$

Az ellenhipotézis pedig a következő:

$$H_1: \ P(X<x)\neq P(Y<x).$$

Tekintsük az $(X_1, X_2, \dots, X_m)$ és $(Y_1, Y_2, \dots, Y_n)$ mintákat. Ekkor a következő lépéseket végezzük el:

Meghatározzuk a két minta empirikus eloszlásfüggvényét. Ezután kiszámítjuk a

$$T=\frac{mn}{(m+n)^2}\left(\sum_{i=1}^{m}(F_m^*(X_i)-G_n^*(X_i))^2+\sum_{j=1}^{n}(F_m^*(Y_j)-G*_n(Y_j))^2\right)$$

statisztikát. α próbaterjedelem választása esetén a H_0 hipotézis kritikus tartománya a H_1 alternatíva mellett $K=\{T\geq w_0\}$, ahol w_0 kielégíti a következő egyenlőtlenséget $P(T<w_0)=1-\alpha$. A statisztika kritikus értékeit táblázatból lehet kiolvasni.

Példa. Vegyünk két változót, legyenek X és Y. Ezekre a változókra vett egy-egy mintát kell összehasonlítanunk. A minták elemszáma $m=6$, illetve $n=11$. A próbaterjedelem legyen $\alpha=0{,}05$. Az 5. táblázat tartalmazza az adatokat és a számításokat is.

Kiszámoljuk a következő értékeket:

$$\sum_{i=1}^{6}(F_6(X_i)-G_{11}(X_i))^2=0{,}479,$$

$$\sum_{j=1}^{11}(F_6(Y_j)-G_{11}(Y_j))^2=0{,}224.$$

Ezeket az eredményeket behelyettesítve T fenti képletébe a következő eredményt kapjuk:

$$T=\frac{11\cdot 6}{17^2}\left(0{,}479+0{,}224\right)=0{,}161.$$

A kvantilis táblázat alapján a kritikus érték $\alpha=0{,}05$ -nél 0,461. Ebből következik, hogy $T=0{,}161<0{,}461=w_{0{,}05}$, vagyis T nem esik a *K* kritikus tartományba, tehát elfogadjuk a nullhipotézist, így a két mintát azonos eloszlásúnak tekinthetjük.

X_i	Y_j	$F_6(X_i)-G_{11}(X_i)$	$F_6(Y_j)-G_{11}(Y_j)$
1,2		$1/6-0/11=1/6$	
1,4		$2/6-0/11=2/6$	
	1,9		$2/6-1/11=8/33$
	2,5		$2/6-2/11=5/33$
	2,7		$2/6-3/11=2/33$
3,0		$3/6-3/11=5/22$	
	3,4		$3/6-4/11=3/22$
	3,8		$3/6-5/11=1/22$
	4,5		$3/6-6/11=-1/22$
	4,9		
5,6		$4/6-7/11=1/33$	
5,7		$5/6-7/11=13/33$	
5,8		$6/6-7/11=4/11$	
	7,3		$6/6-8/11=3/11$
	7,8		$6/6-9/11=2/11$
	8,2		$6/6-10/11=1/11$
	8,8		$6/6-11/11=0$

5. táblázat

2.4. Anderson-Darling – próba

Az Anderson-Darling-teszt ugyanolyan hipotéziseket tesztel, mint az előzőekben használt eljárások, vagyis $H_0: P(X<x)=F_0(x)$. Az ötlet hasonló a Kolmogorov-Szmirnov-tesztéhez: egy távolságmértéket használ F_0 és F_n közötti távolság számára. Elutasítja a nullhipotézist, hogyha az F_0 és az F_n közötti távolság túl nagy. A következő kifejezés, mint távolságmértéket használja a súlyozott integrált távolság négyzetét:

$$A = n\int_{-\infty}^{\infty}(F_0(x)-F_n(x))^2\cdot\psi(x)\cdot f_0(x)\,dx\,.$$

Itt

$$\psi(x)=\frac{1}{F_0(x)(1-F_0(x))}$$

egy súlyfüggvény, f_0 pedig az F_0-hoz tartozó sűrűségfüggvény. A súlyfüggvény azt okozza, hogy az eloszlások „széleihez” a fenti integrálban nagyobb súlyt rendelünk. Az Anderson-Darling-tesztstatisztikát a következő képlettel számolhatjuk ki

$$A=\left(-\frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln z_i+\ln(1-z_{n+1-i})]\right)-n\,,$$

ahol $z_i=F_0(x_{(i)})$.

2.5. Lilliefors-teszt

A Lilliefors-teszt a Kolmogorov-Szmirnov teszt egy kiterjesztett esete, amikor az elméleti eloszlásnak csak a típusát tesszük fel a nullhipotézisben, konkrét paraméter értéket nem tételezünk fel. Ezt a tesztet Kolmogorov-Szmirnov-teszt Lilliefors akadállyal vagy csak egyszerűen Lilliefors-tesztnek is nevezik. A tesztstatisztika olyan, mint a Kolmogorov-Szmirnov-tesztnél, vagyis a mintából becsült paraméterekkel megadott elméleti és tapasztalati eloszlás eltérésének szuprémumán alapul. Csupán a kritikus értékek azok, amelyek mások ennél a tesztnél. Fontos megjegyezni még azt is, hogy a tárgyalt tesztnél minden eloszlástípusra külön táblázatot kell készíteni a kritikus értékek számára.

Nézzük meg ezt a tesztet, ha a feltételezett eloszlás normális eloszlás. A megoldási lépései a következők: először is megbecsüljük az elméleti eloszlás várható értékét és szórását (a mintaátlaggal és a korrigált tapasztalati szórással):

$$\mu = \bar{x} \text{ és } \sigma = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_1-\bar{x})^2}.$$

A teszt további menete a Kolomogorov-Szmirnov-teszthez hasonló.

3. fejezet

Az R egy olyan programozási nyelv, amely statisztikai módszerek alkalmazásával elemzi az adatokat és az eredményeket grafikusan is meg tudja jeleníti.

Minden, már említett próbát lefuttattuk a programban lévő parancsot többféle elemszámra és többféle eloszlásra.

A vizsgált eloszlások a következők voltak:

- normális eloszlás
- lognormális eloszlás
- Cauchy-eloszlás
- Gamma-eloszlás
- Student-eloszlás

Ezeket pedig definiáljuk is, azért, hogy a továbbiakban világosak legyenek a fogalmak.

Normális eloszlás [10]

Az X valószínűségi változó normális eloszlású, ha sűrűségfüggvénye

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

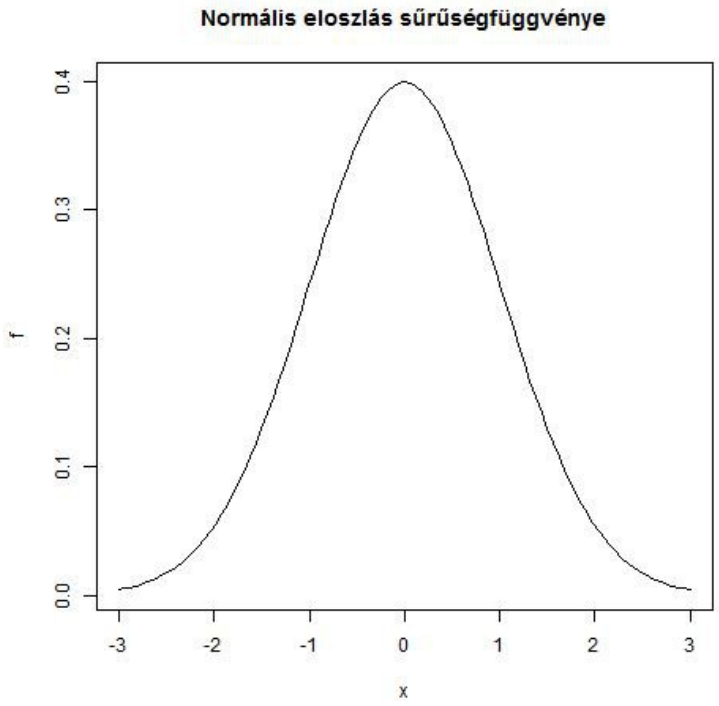
A két paraméter, $m \in \mathbb{R}$ és $\sigma > 0$.

Jelölésben úgy írjuk le, hogy az X valószínűségi változó normális eloszlású, vagyis $X \sim N(m, \sigma^2)$.

Eloszlásfüggvénye

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}} dt = \int_{-\infty}^{x} f(t)\,dt.$$

A standard normális eloszlás a normális eloszlás egy speciális esete akkor, ha $X \sim N(0,1)$.



1. ábra

Lognormális eloszlás [11]

Az Y valószínűségi változó akkor lognormális eloszlású, ha $Y = \exp(X)$, ahol X normális eloszlású. Megfordítva: ha Y lognormális eloszlású, akkor $X = \log(Y)$ normális eloszlású. Sűrűségfüggvénye a következő:

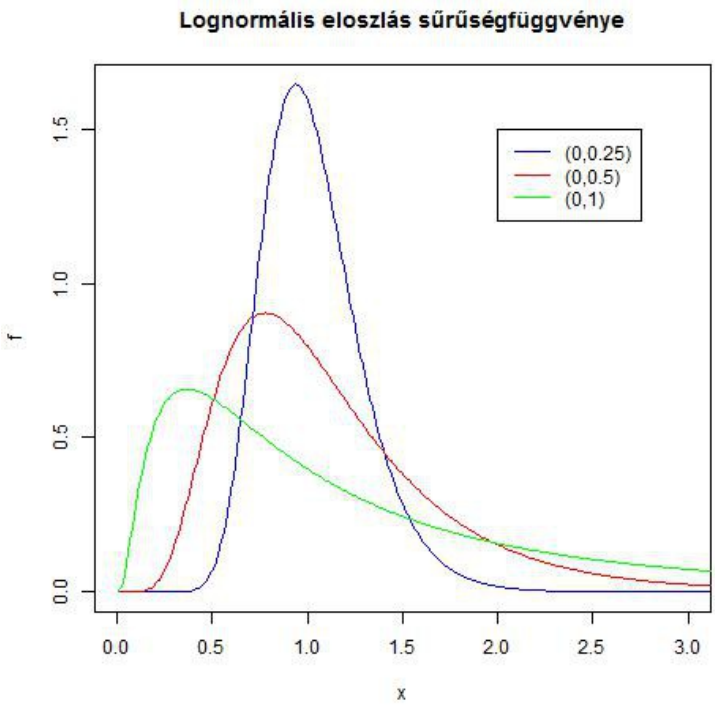
$$f_x(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},$$

$x>0$. μ és σ az X normális eloszlás paraméterei.

Eloszlásfüggvénye

$$F_X(x;\mu,\sigma) = \frac{1}{2}\operatorname{erfc}\left[\frac{-\ln x-\mu}{\sigma\sqrt{2}}\right] = \Phi\left(\frac{\ln x-\mu}{\sigma}\right).$$

Itt az erfc a komplementer hibafüggvény [12], Φ a standard normális eloszlás eloszlásfüggvényét jelöli.



2.ábra

Cauchy-eloszlás [13]

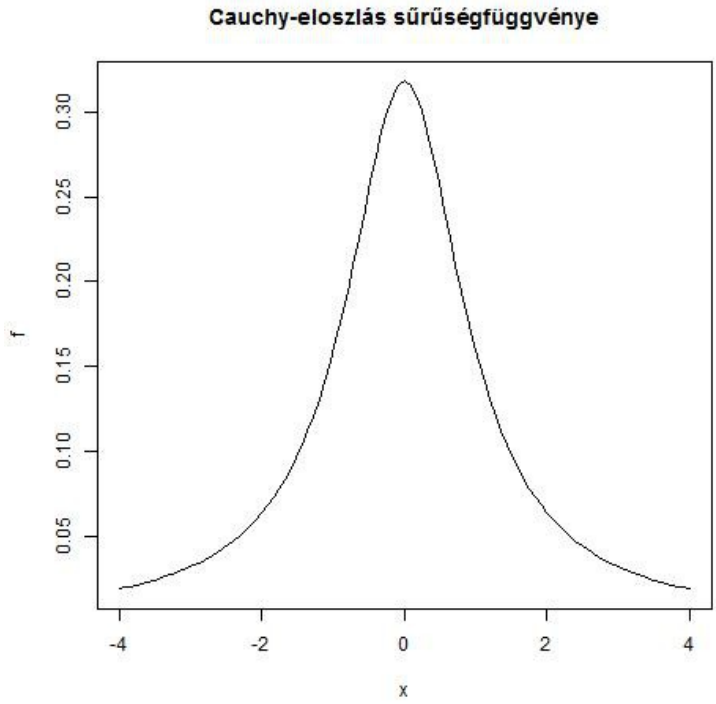
Egy valószínűségi változó Cauchy-eloszlású, ha a sűrűségfüggvénye

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Eloszlásfüggvénye

$$F(x) = \frac{1}{\pi}\operatorname{arctg} x + \frac{1}{2}.$$

Fontos megjegyezni, hogy a Cauchy-eloszlás nem véges várható értékű stabilis eloszlás.



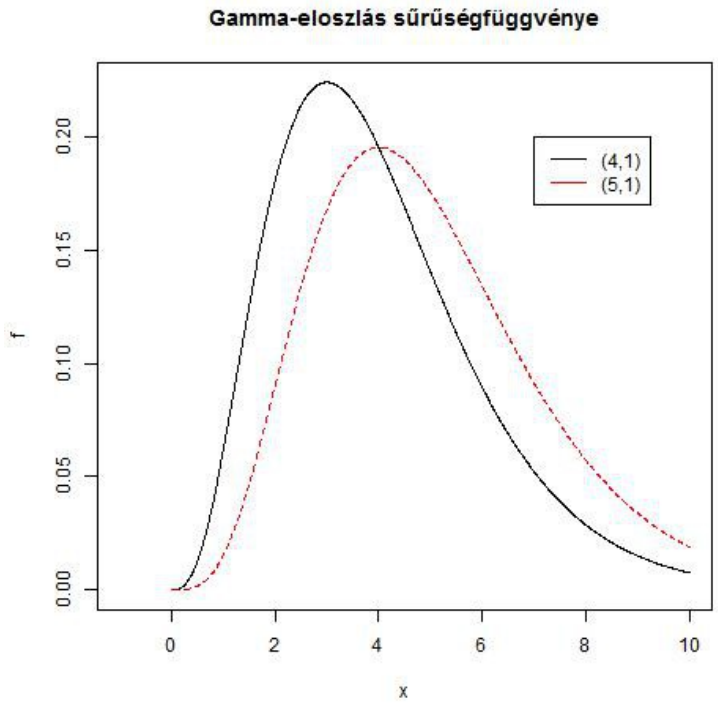
3. ábra

Gamma-eloszlás [14]

Az X valószínűségi változó p-edrendű λ paraméterű gamma-eloszlású pontosan akkor, ha sűrűségfüggvénye

$$f(x)=\begin{cases}\dfrac{\lambda^{p}x^{p-1}e^{-\lambda x}}{\Gamma(p)}, & \text{ha } x>0\\ 0, & \text{ha } x\leqslant 0.\end{cases}$$

ahol $\Gamma(p)$ a gamma-függvény, λ és p pedig pozitívak.

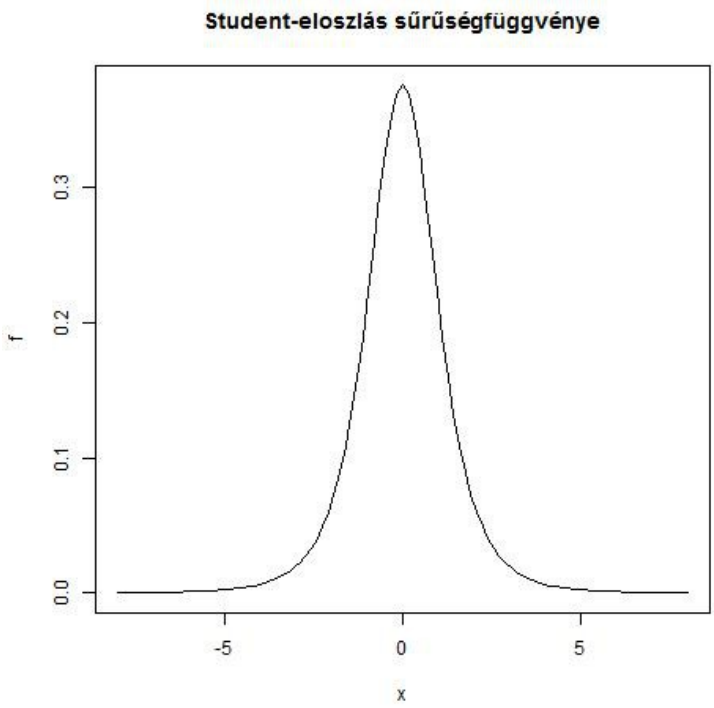


4. ábra

Student-eloszlás [9]

Az X valószínűségi változót Student-eloszlásúnak nevezzük, ha sűrűségfüggvénye:

$$f(x)=\frac{1}{\sqrt{\pi n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\quad x\in\mathbb{R}.$$



5. ábra

3.1. Tétel. *Ha* $\xi\sim(0,1)$, $\eta\sim\chi^2_n$, ξ *és* η *független, akkor*

$$\frac{\sqrt{n}\,\xi}{\sqrt{\eta}}$$

t *-eloszlású* n *szabadságfokú.*

A Student-eloszlás fontosságát az adja, hogy az 1. fejezetben bemutatott t-próbánál ez a statisztika eloszlása a nullhipotézis esetén.

Program bemutatása

A programot úgy futtattuk le, hogy beállítottuk az alfa értékét, a minták számát, a minta nagyságát, és a már említett, definiált eloszlásokat. Ennek során természetesen a 2. fejezetben tárgyalt statisztikai próbákat alkalmaztuk.

A programot azzal kezdtük, hogy behívtuk a `nortest` nevű csomagot. Erre azért volt szükség, mert ebben találhatóak azok a függvények, amelyekkel a továbbiakban dolgozunk. Vagyis az Anderson-Darling-, a Cramér-von Mises- és a χ^2-teszt, továbbá a Lilliefors-teszt. Itt fontos megemlíteni, hogy a nullhipotézisünk a standard normális eloszlás volt.

Ezután a paramétereket adtuk meg. A paraméterek a következők: α, minták száma, minta nagysága. Az eloszlásokra is adtunk meg feltételeket. A gamma-eloszlásra a rendet 4-re, a paramétert pedig 1-re állítottuk be. A Student-eloszlás szabadságfoka 4 lett. A nullhipotézisünk a standard normális eloszlás volt.

A táblázatok azt mutatják, hogy a nullhipotézisünket az esetek mekkora részarányra utasítjuk el. Az α értékét háromféleképpen állítottuk be: az elsőnél $\alpha = 0{,}001$, a másodiknál $\alpha = 0{,}05$; általában ezt az értéket használjuk a feladatok megoldására; az utolsónál pedig $\alpha = 0{,}1$. Ezt fontos volt beállítani a programba, mert azt is láthatjuk a táblázatokból, hogy a különböző α-kra milyen eredményeket kapunk. A minta méretére 4-féle értéket állítottunk be: $n = 20, 50, 200$ és 1000. Az ismétlésszám $J = 10000$ volt.

Az alábbi táblázatok az adott α és mintaelemszám esetén azt mutatják, hogy mekkora volt a nullhipotézis elutasításának a részaránya (10000 ismétlés alapján). Tehát most nézzük a konkrét táblázatokat és értékeket.

$\alpha = 0{,}001$

n=20

	norm	cauchy	lnorm	gamma	t4
KS	0.0008	0.0134	1.0000	1	0.0020
AD	0.0013	0.7091	0.6139	0.0287	0.0522
CvM	0.0012	0.7025	0.5650	0.0219	0.0415
Chiqs	0.0017	0.5701	0.3073	0.0083	0.0122
Lillie	0.0011	0.6286	0.3717	0.0141	0.0267

n=50

	norm	cauchy	lnorm	gamma	t4
KS	0.0017	0.0395	1	1	0.0018
AD	0.0010	0.9836	0.9911	0.1556	0.1458
CvM	0.0010	0.9820	0.9823	0.1153	0.1174
Chiqs	0.0021	0.9384	0.8829	0.0275	0.0347
Lillie	0.0010	0.9594	0.9097	0.0611	0.0672

n=200

	norm	cauchy	lnorm	gamma	t4
KS	0.0007	0.7647	1	1	0.0027
AD	0.0010	1	1	0.9198	0.6157
CvM	0.0009	1	1	0.8380	0.5327
Chiqs	0.0015	1	1	0.3966	0.1474
Lillie	0.0015	1	1	0.5664	0.3296

n=1000

	norm	cauchy	lnorm	gamma	t4
KS	0.0010	1	1	1	0.0653
AD	0.0012	0.8437	1	1	1
CvM	0.0010	1	1	1	0.9997
Chiqs	0.0009	1	1	1	0.9143
Lillie	0.0004	1	1	1	0.9947

$\alpha = 0{,}05$

n=20

	norm	cauchy	lnorm	gamma	t4
KS	0.0530	0.1957	1	1	0.0636
AD	0.0488	0.8827	0.9059	0.2540	0.2202
CvM	0.0483	0.8813	0.8833	0.2248	0.2032
Chiqs	0.0461	0.7781	0.8237	0.1308	0.1172
Lillie	0.0466	0.8451	0.7907	0.1765	0.1661

n=50

	norm	cauchy	lnorm	gamma	t4
KS	0.0468	0.4734	1	1	0.0719
AD	0.0484	0.9966	0.9997	0.5835	0.4196
CvM	0.0502	0.9963	0.9994	0.5194	0.3812
Chiqs	0.0518	0.9832	0.9963	0.2732	0.1940
Lillie	0.0499	0.9939	0.9967	0.4053	0.2965

n=200

	norm	cauchy	lnorm	gamma	t4
KS	0.0448	0.9989	1	1	0.1173
AD	0.0507	1	1	0.9975	0.8909
CvM	0.0509	1	1	0.9901	0.8616
Chiqs	0.0531	1	1	0.8660	0.4833
Lillie	0.0510	1	1	0.9505	0.7613

n=1000

	norm	cauchy	lnorm	gamma	t4
KS	0.0489	1	1	1	0.8103
AD	0.0492	0.8502	1	1	1
CvM	0.0490	1	1	1	1
Chiqs	0.0498	1	1	1	0.9932
Lillie	0.0443	1	1	1	1

$\alpha = 0,1$

n=20

	norm	cauchy	lnorm	gamma	t4
KS	0.1077	0.3303	1	1	0.1208
AD	0.0972	0.9140	0.9449	0.3645	0.3089
CvM	0.0977	0.9121	0.9229	0.3375	0.2891
Chiqs	0.1176	0.8512	0.8877	0.2549	0.2313
Lillie	0.0924	0.8856	0.8630	0.2864	0.2505

n=50

	norm	cauchy	lnorm	gamma	t4
KS	0.0980	0.6543	1	1	0.1314
AD	0.0990	0.9980	1	0.6858	0.5167
CvM	0.0977	0.9981	0.9998	0.6270	0.4789
Chiqs	0.1000	0.9919	0.9987	0.3736	0.2740
Lillie	0.0995	0.9963	0.9984	0.5259	0.4094

n=200

	norm	cauchy	lnorm	gamma	t4
KS	0.0934	0.9999	1	1	0.2233
AD	0.0957	1	1	0.9993	0.9275
CvM	0.0947	1	1	0.9972	0.9058
Chiqs	0.0978	1	1	0.9250	0.5971
Lillie	0.0994	1	1	0.9822	0.8394

n=1000

	norm	cauchy	lnorm	gamma	t4
KS	0.0949	1	1	1	0.9383
AD	0.0996	0.8463	1	1	1
CvM	0.0995	1	1	1	1
Chiqs	0.0958	1	1	1	0.9966
Lillie	0.0987	1	1	1	1

Következtetések

Láthatjuk, hogy általában normális eloszlásnál a program a tőle elvárható értéket adja vissza. Itt mindig az α elsőfajú hiba valószínűséget várjuk. Ez azért lehetséges, mert az ismétlésszámot nagyra állítottuk be. Tehát ebből az következik, hogy minél nagyobb az α, annál nagyobb lesz az elutasítás valószínűsége. Az is megfigyelhető még, hogy a Student-eloszlásnál, úgy nő a valószínűség, ahogyan nő az α. A Cauchy-eloszlásnál is látható ugyanez, de ott minimális mértékben nő.

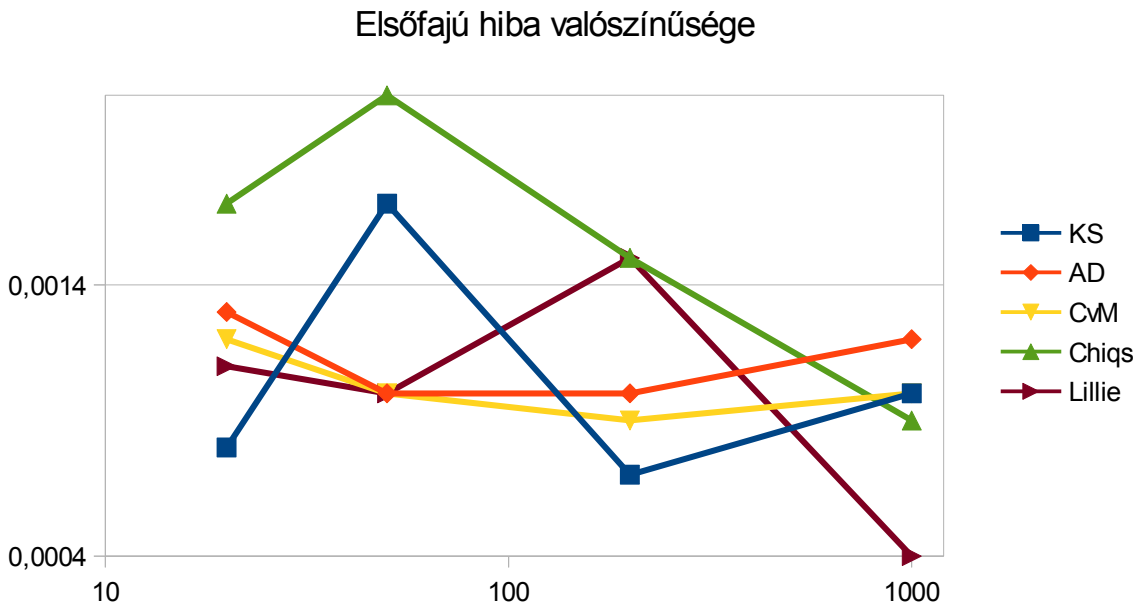
Minden egyes táblázatnál megvizsgálhatjuk, hogy melyik próba lesz az erősebb. Mégpedig az, amelyik a különböző eloszlásokra a nagyobb elutasítási százalékot adja. Lognormális és gamma-eloszlásra a legtöbb az elutasítások száma. A Kolmogorov-Szmirnov-próba ennél a két eloszlásnál egyszer sem fogadta el a nullhipotézist. Megfigyelhetjük még azt is, hogy minden esetben, ahogyan az n értéke nő, úgy nő az elutasítás valószínűsége.

A grafikonok szemléletes képet adnak az egyes próbák különbségét és hasonlóságát egymáshoz képest. A tárgyalt grafikonokról könnyen tudunk következtetéseket levonni az egyes próbákról. Ilyen például, hogy a Kolmogorov-Szmirnov-teszt Cauchy- és Student-eloszlás esetén az összes α értékre a leggyengébb. Viszont lognormális eloszlásnál minden alkalommal a legerősebb próba; gamma-eloszlásnál is majdnem teljesül, csak akkor nem, amikor $\alpha = 0{,}001$ és $n = 200$. Érdekes észrevétel, hogy az elsőfajú hiba a Lilliefors-próbánál úgy viselkedik, mintha a Kolmogorov-Szmirnov-teszt „ellentettfüggvénye” lenne. Abban a tartományban, ahol a Kolomogorov-Szmirnov-teszt hibavalószínűsége nő, ott a Lilliefors-próba csökken és fordítva. Leolvasható még az is, hogy mi is azt kaptuk a szakirodalommal összhangban, hogy a vizsgált próbák közül a szimmetrikus eloszlások esetén többnyire az Anderson-Darling-próba a legerősebb. Például Student-eloszlás esetén minden egyes α-nál ez a legerősebb.

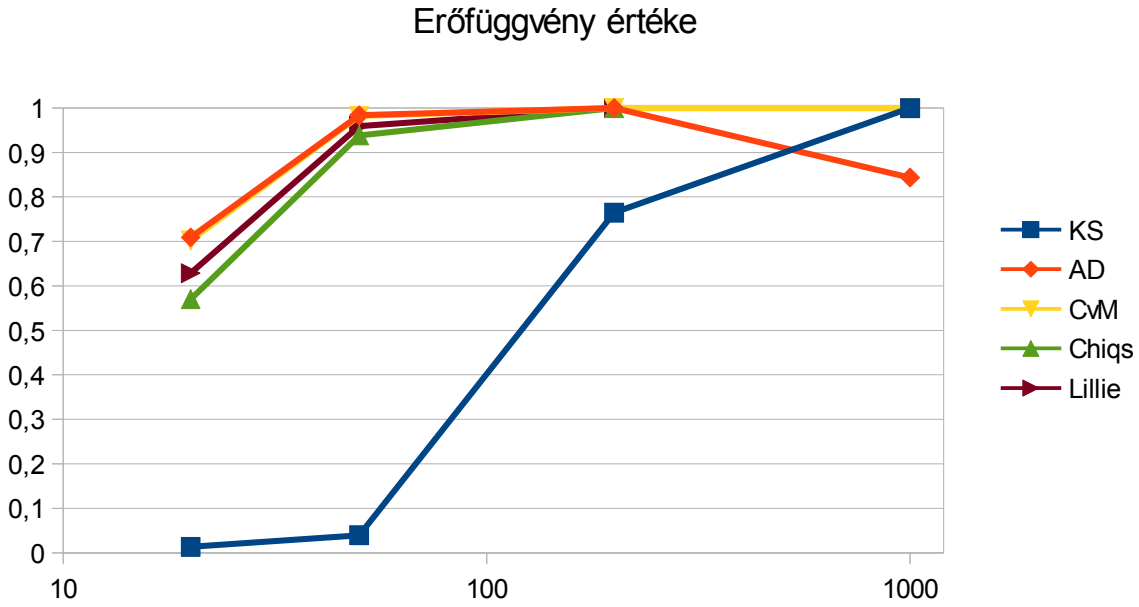
Melléklet

Az itt közölt grafikonok a 3. fejezetben található táblázatok értékeit szemléltetik.

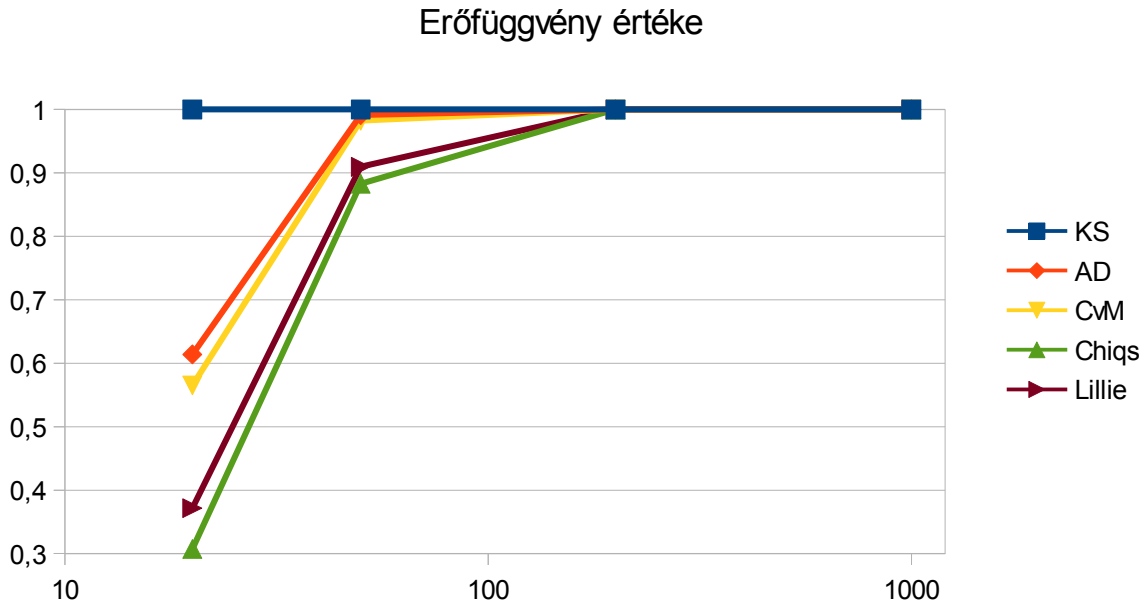
$\alpha = 0{,}001$



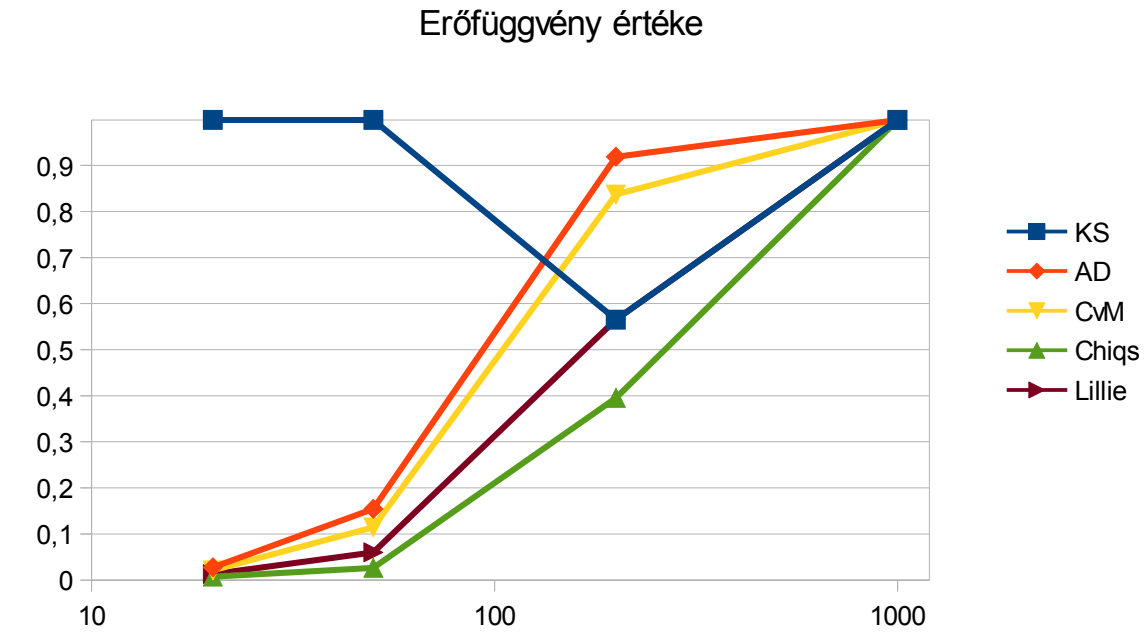
Normális eloszlás



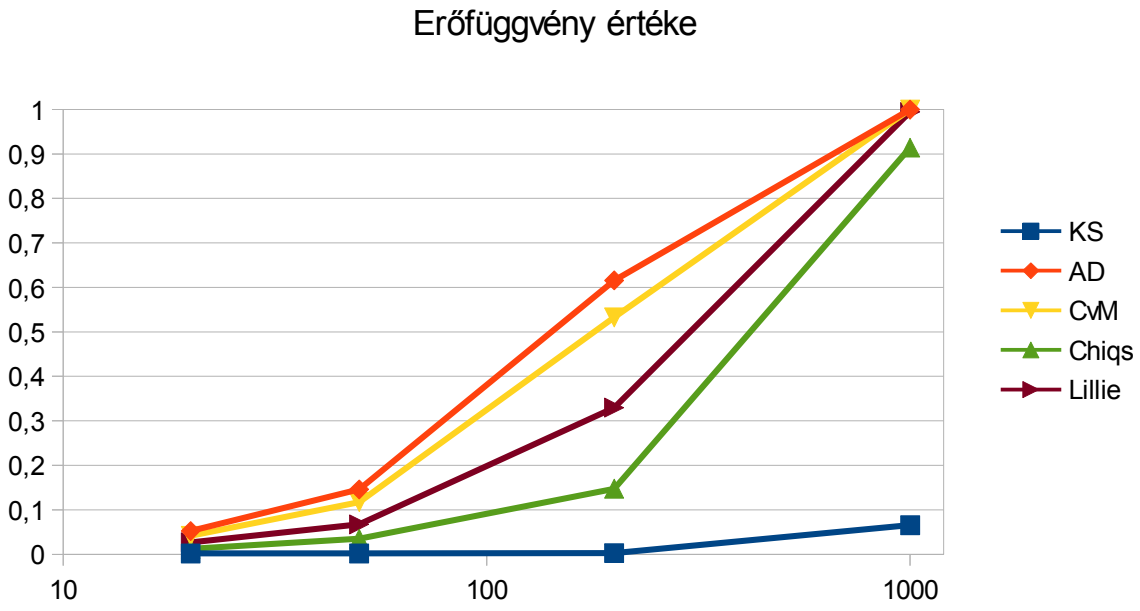
Cauchy-eloszlás



Lognormális eloszlás

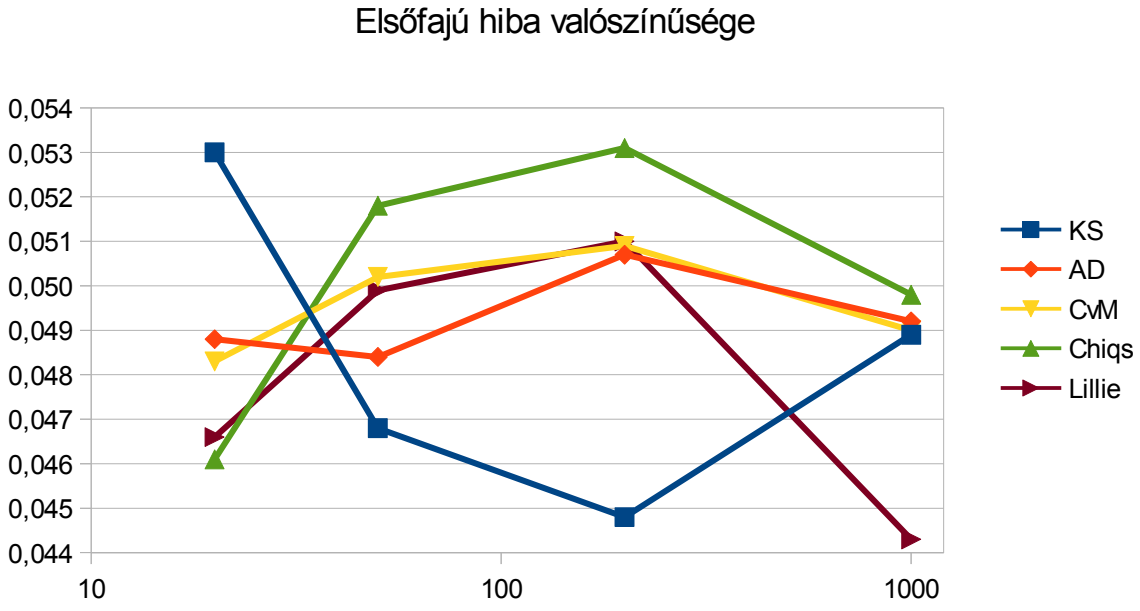


Gamma-eloszlás

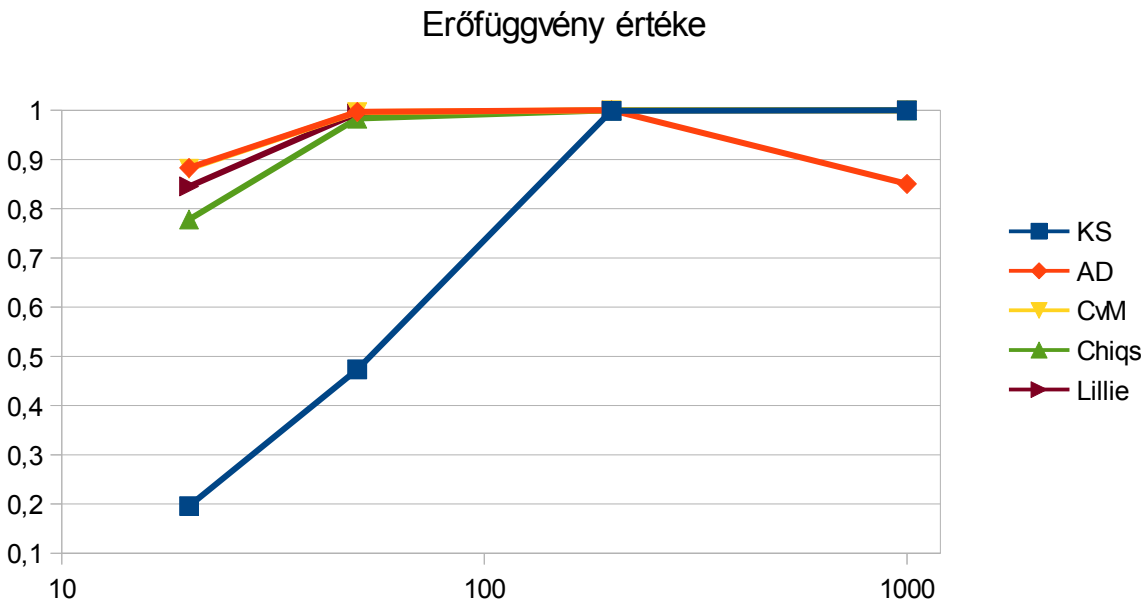


Student-eloszlás

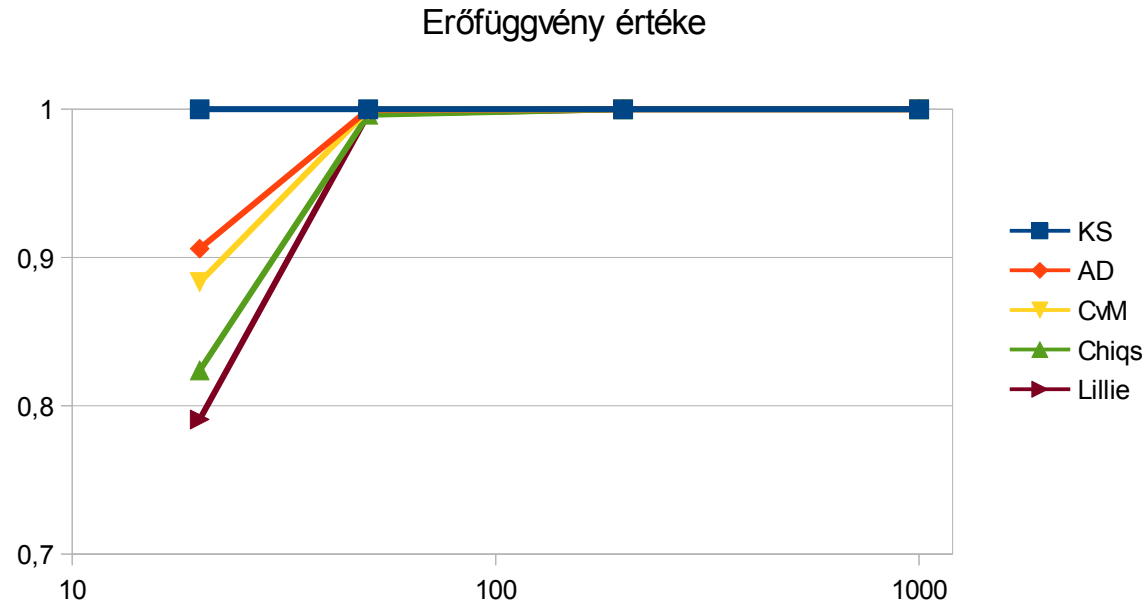
$\alpha = 0{,}05$



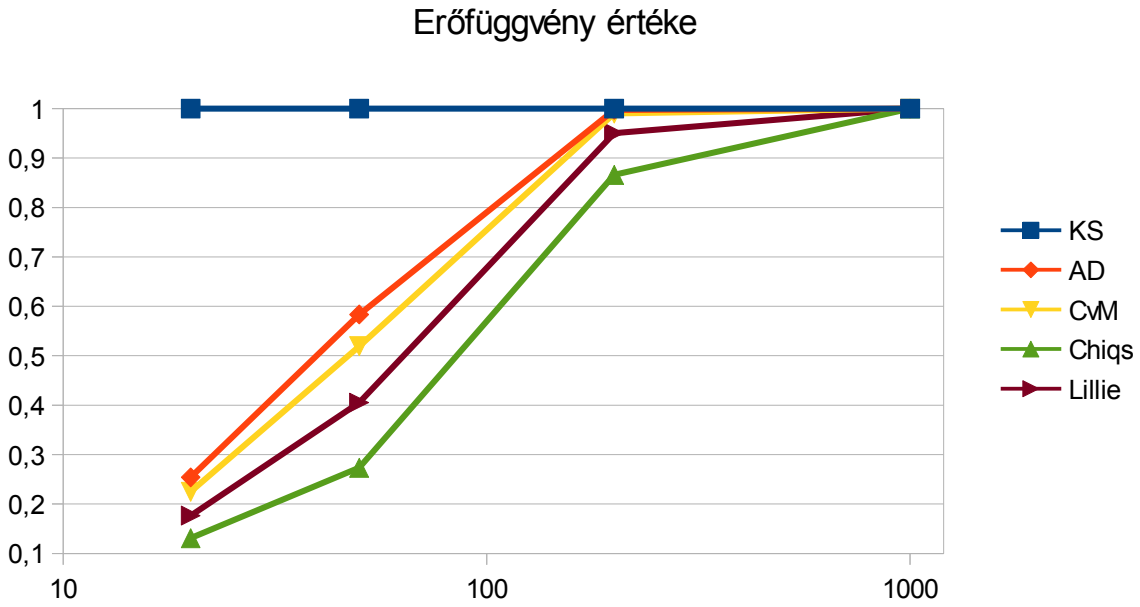
Normális eloszlás



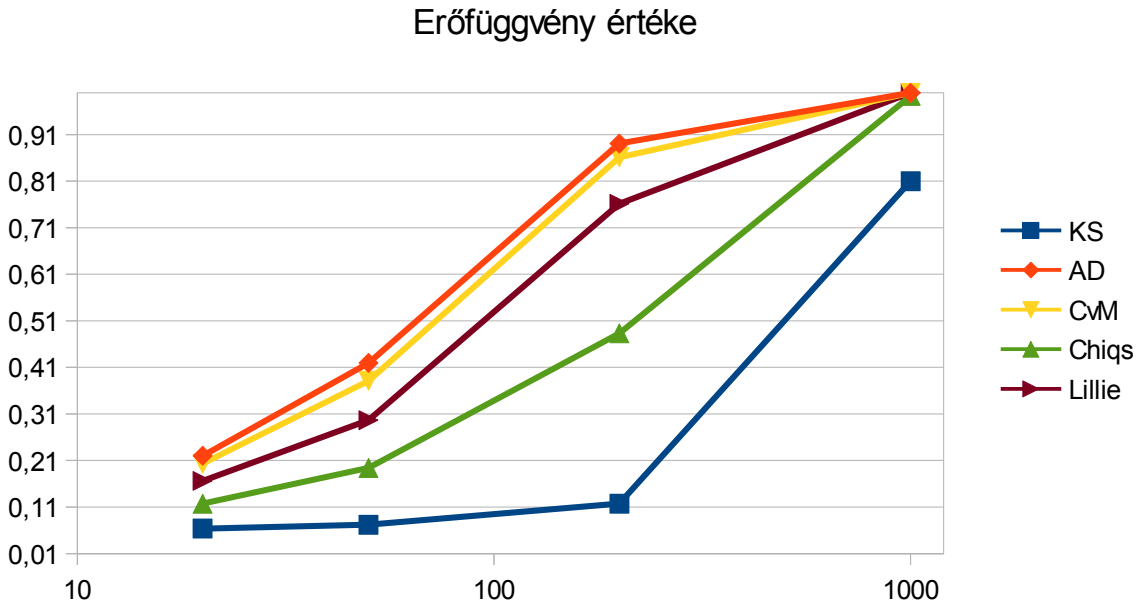
Cauchy-eloszlás



Lognormális eloszlás

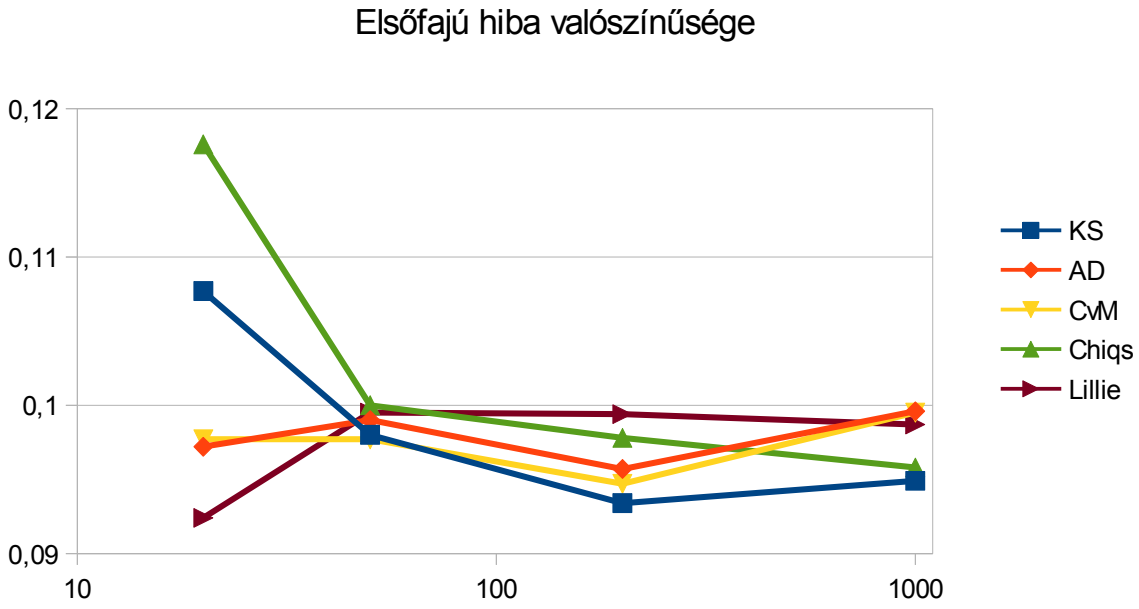


Gamma-eloszlás

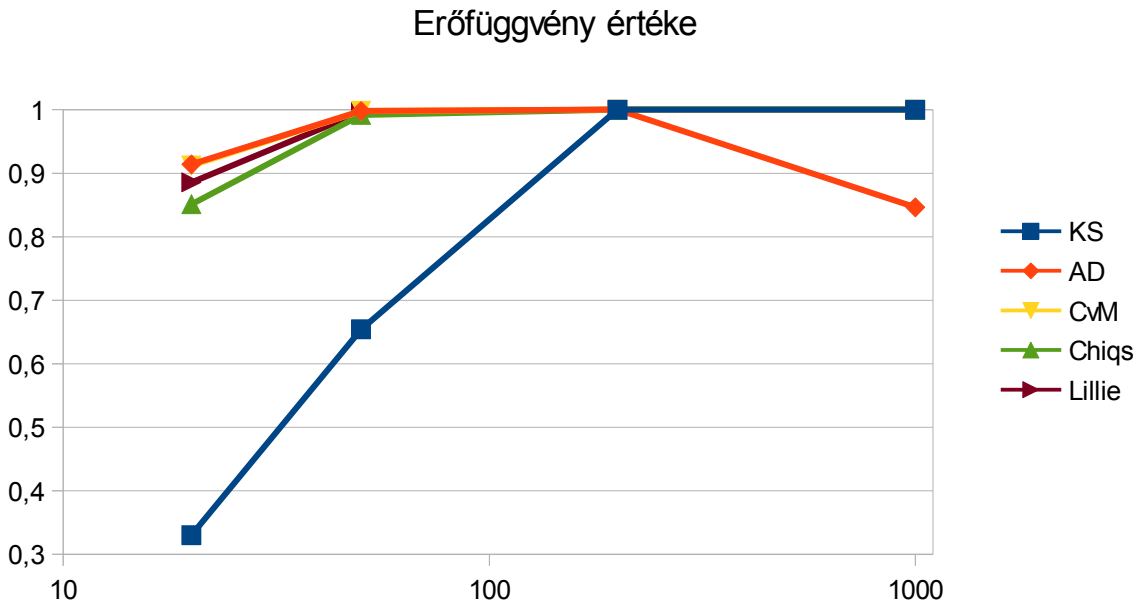


Student-eloszlás

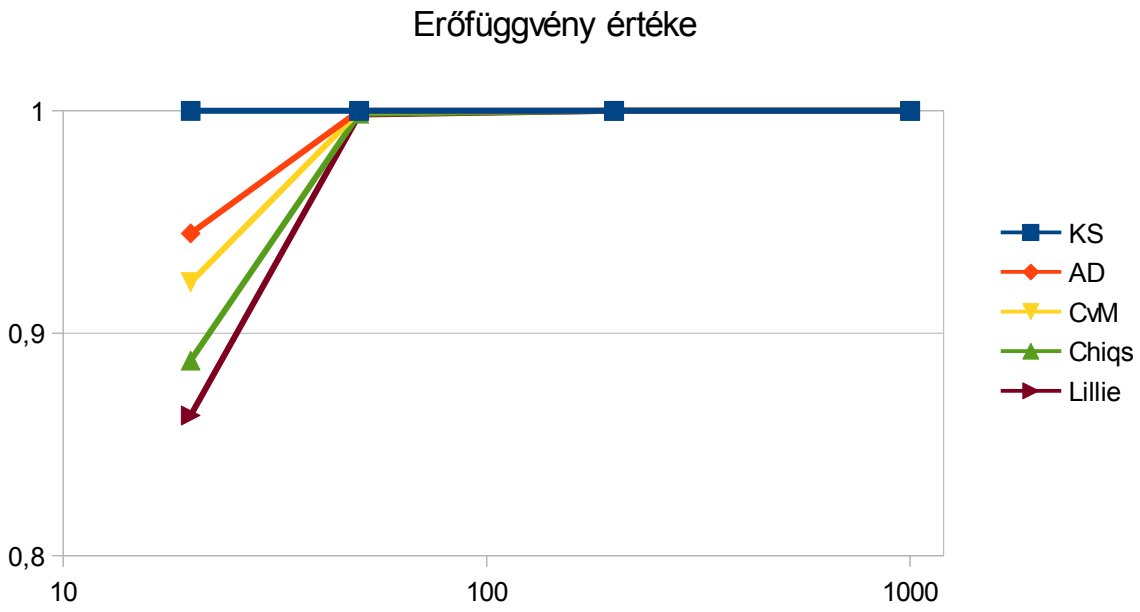
$\alpha = 0,1$



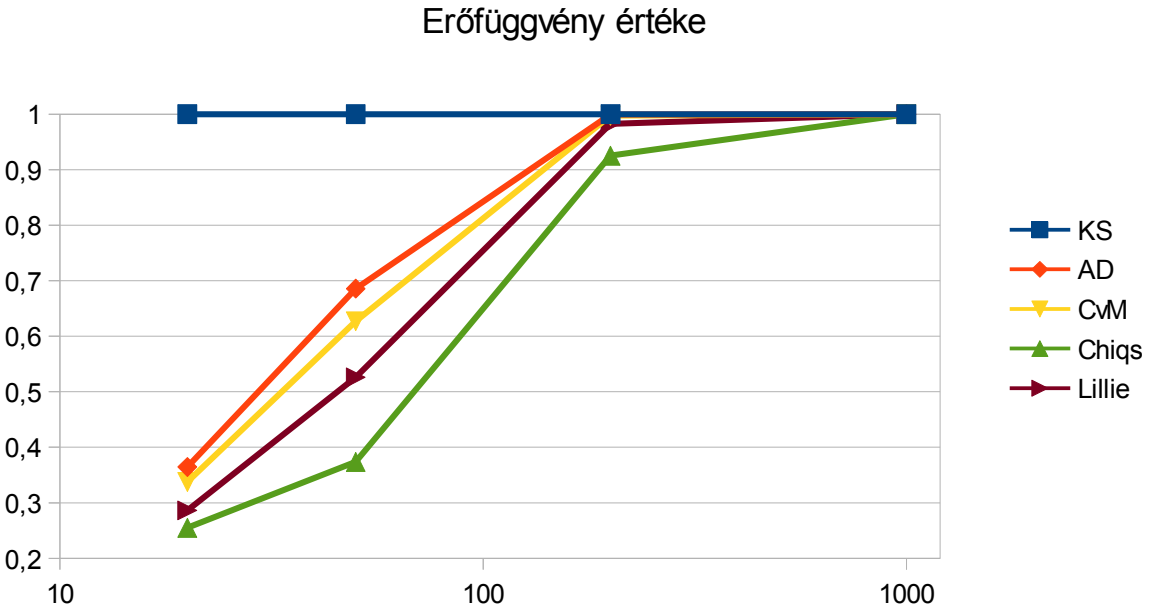
Normális eloszlás



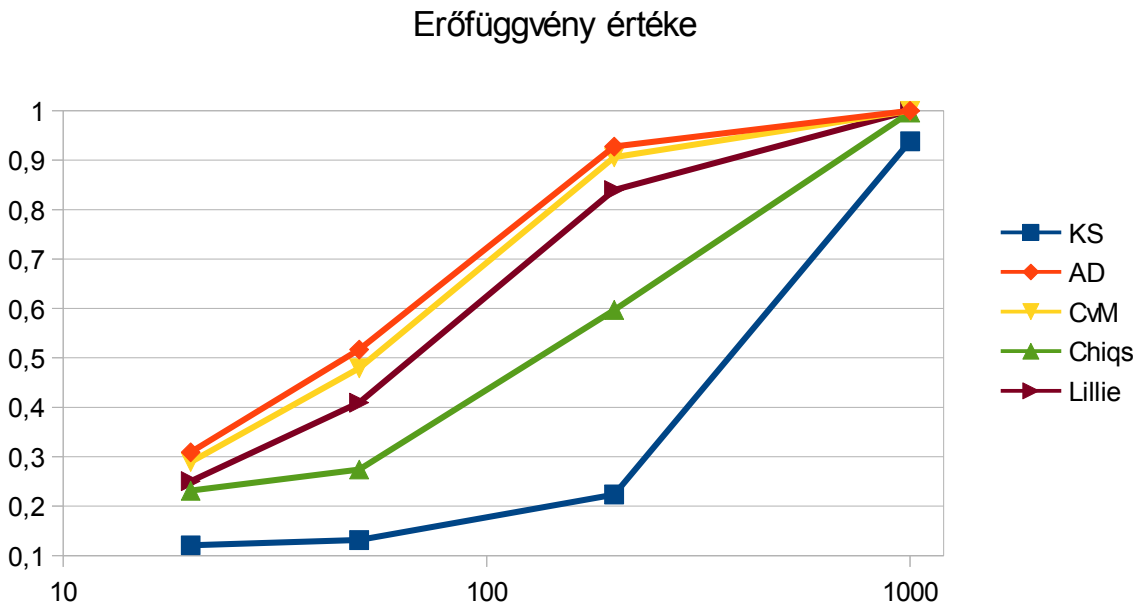
Cauchy-eloszlás



Lognormális eloszlás



Gamma-eloszlás



Student-eloszlás

Köszönetnyilvánítás

Elsősorban szeretném megköszönni témavezetőmnek, Zempléni Andrásnak azt a sok-sok segítséget, melyet dolgozatírásom folyamán nyújtott. Végtelen türelemmel válaszolt minden kérdésemre és az `R` programcsomag elsajátításához is sok támogatást kaptam Tőle.

Nem utolsó sorban köszönöm családomnak a sok támogatást, amit felsőfokú tanulmányaim alatt nyújtottak.

Irodalomjegyzék

[1] Bolla Marianna, Krámli András: Statisztikai következtetések elmélete
Typotex Kiadó, 2005

[2] Vincze István, Varbanova Mária: Nemparaméteres matematikai statisztika; Elmélet és alkalmazások
Akadémia Kiadó, 1993

[3] Móri F. Tamás, Szeidl László, Zempléni András:
Matematikai statisztika példatár
ELTE Eötvös Kiadó, 1997

[4] Obádovics J. Gyula: Valószínűségszámítás és matematikai statisztika
Scolar Kiadó, 2003

[5] Christine Duller: Einführung in die nichtparametrische Statistik mit SAS und R: Ein anwendungsorientiertes Lehr- und Arbeitsbuch
Physica-Verlag, 2008

[6] Claudia Cottin, Sebastian Döhler:
Risikoanalyse: Modellierung, Beurteilung und Management von Risiken mit Praxisbeispielen
Springer Spektrum-Verlag, 2013

[7] http://www.win.tue.nl/~rmcastro/AppStat2013/files/lectures23.pdf

[8] http://www.renyi.hu/~major/debrecen/tobbdim.pdf

[9] http://www.inf.unideb.hu/valseg/JEGYZET/valseg/node176.htm

[10] http://hu.wikipedia.org/wiki/Norm%C3%A1lis_eloszl%C3%A1s

[11] http://hu.wikipedia.org/wiki/Log-norm%C3%A1lis_eloszl%C3%A1s

[12] http://hu.wikipedia.org/wiki/Gauss-f%C3%A9le_hibaf%C3%BCggv%C3%A9ny

[13] Tómács Tibor: Matematikai statisztika (jegyzet, 2011)

[14] http://www.inf.unideb.hu/valseg/JEGYZET/valseg/node159.htm

Nyilatkozat

Név: Tóth Alexandra

Eötvös Loránd Tudományegyetem

Természettudományi Kar

Matematika BSc.

Neptun azonosító: SIKT5T

Szakdolgozat címe:

Illeszkedésvizsgálati módszerek összehasonlítása

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

Budapest, 2015. május 28.

_____

a hallgató aláírása