

Extrém-érték modellezés

Zempléni András

Val.modellek

2022. február 18.

Extrém-érték elemzés

- Klasszikus módszerek: évi maximumon alapulnak
- Küszöb feletti értékek elemzése: adott szintet meghaladó minden árvízből használ adatot.
- Többdimenziós módszerek: a közeli mérőállomások összefüggőségét is vizsgálja (extrémumok együttes viselkedése)

Extrém-érték eloszlások

Legyenek X_1, X_2, \dots, X_n független, azonos eloszlású valószínűségi változók. Ha vannak a_n, b_n normáló konstansok, hogy

$$[\max(X_1, X_2, \dots, X_n) - a_n] / b_n$$

nemelfajuló határeloszláshoz közelít, akkor ez a határeloszlás szükségképpen max-stabilis vagy úgynevezett extrém-érték eloszlás.

Extrém-érték eloszlások karakterizációja

- Normalizált maximumok lehetséges határeloszlásai:

Frechet: $F_{\alpha}(x) = \exp(-x^{-\alpha}) \quad (x > 0)$

α pozitív parameter.

Weibull: $F_{\alpha}(x) = \exp(-(-x)^{\alpha}) \quad (x < 0)$

Gumbel: $F(x) = \exp(-\exp(-x))$

Megjegyzések

- Bizonyítás technikai, az eloszlásfüggvény és a maximum művelet kapcsolatán alapul. Az adódó függvényegyenletnek ez a 3 megoldása van.
- Érdekes kérdés: adott F eloszlásfüggvény esetén melyik határeloszláshoz konvergál az F eloszlású minta normalizált maximuma?
- Nem minden esetben lehet normálni: diszkrét eloszlásokra oszcillálhat a maximum eloszlása.

A normálhatóság feltétele

- Folytonos eloszlásokra az eloszlásfüggvény reguláris viselkedése szükséges a felső végpont közelében (teljesül minden fontos eloszlásra):
 - F az α paraméterű Fréchet eloszlás max-vonzási tartományához tartozik ($F \in \text{MDA}(F_\alpha)$), akkor és csak akkor, ha $1-F \sim x^{-\alpha}L(x)$ (L lassú változású függvény: $L(tx)/L(x) \rightarrow 1$ ha $x \rightarrow \infty$)
 - $F \in \text{MDA}(W_\alpha)$, akkor és csak akkor, ha $x_F < \infty$ és $1-F(x_F - 1/x) \sim x^{-\alpha}L(x)$
 - A Gumbel MDA jellemzése bonyolultabb, lényegében az exponenciális lecsengésű eloszlások tartoznak ide (példa: exponenciális, normális).

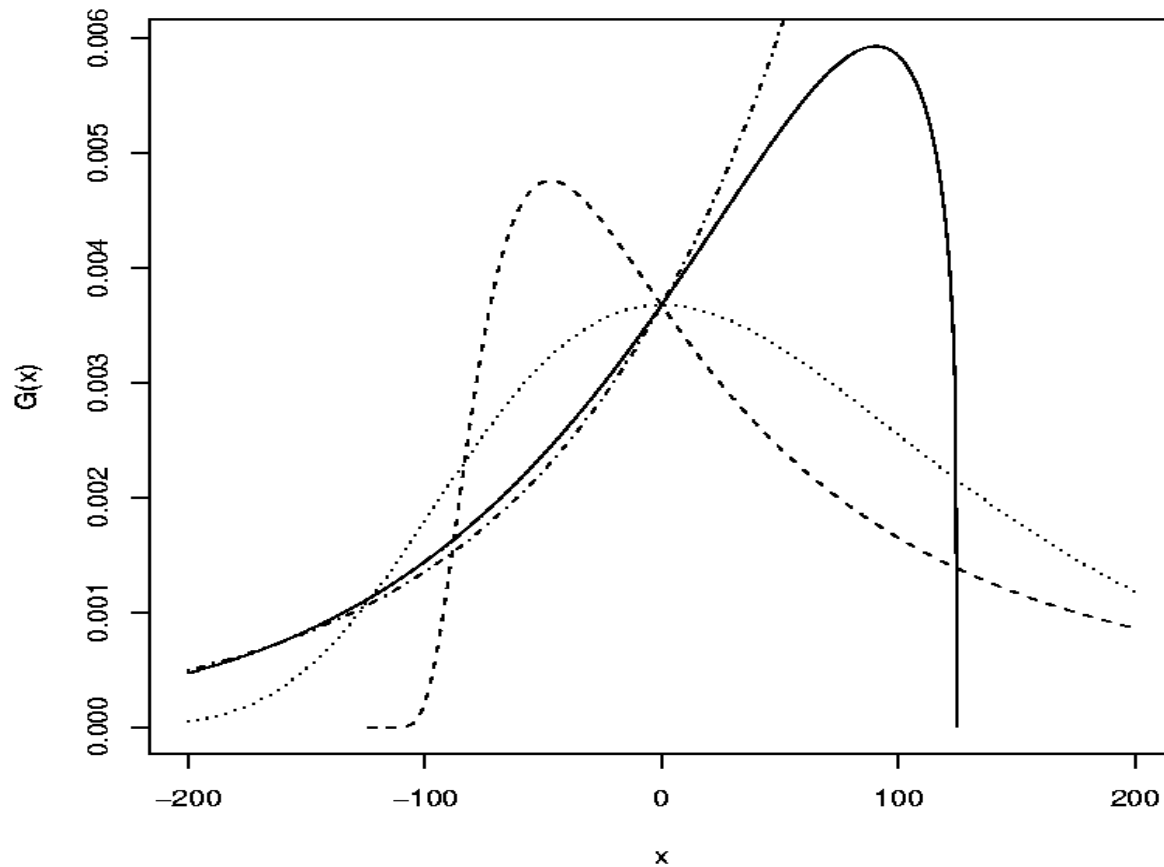
Általánosított extrém-érték (GEV) eloszlás

$$G_{\mu,\sigma,\xi}(z) = \exp\left(-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}\right)$$

ha $1 + \xi \frac{z - \mu}{\sigma} > 0$.

(Lokációs és skála paramétert is tartalmazó
modell.)

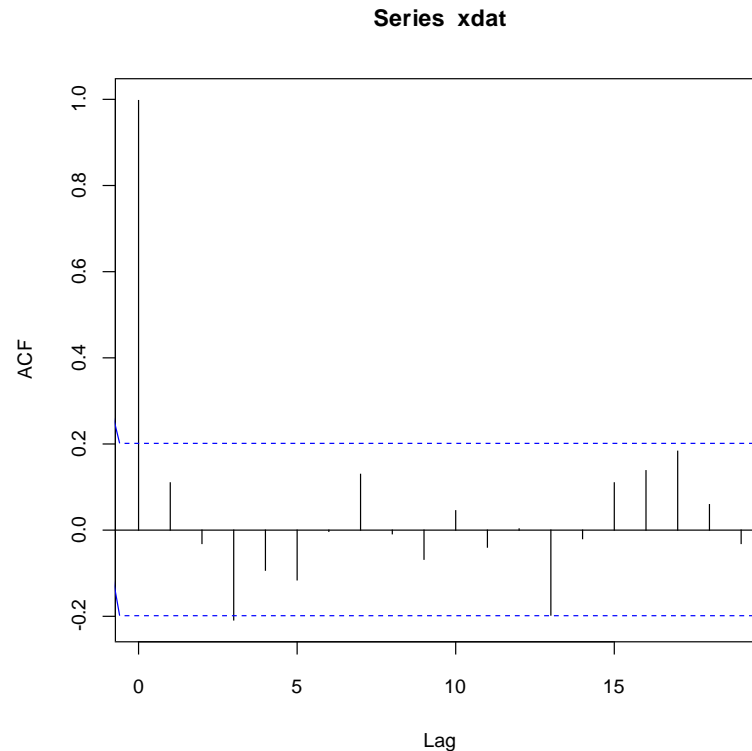
Néhány példa GEV eloszlás sűrűségfüggvényére



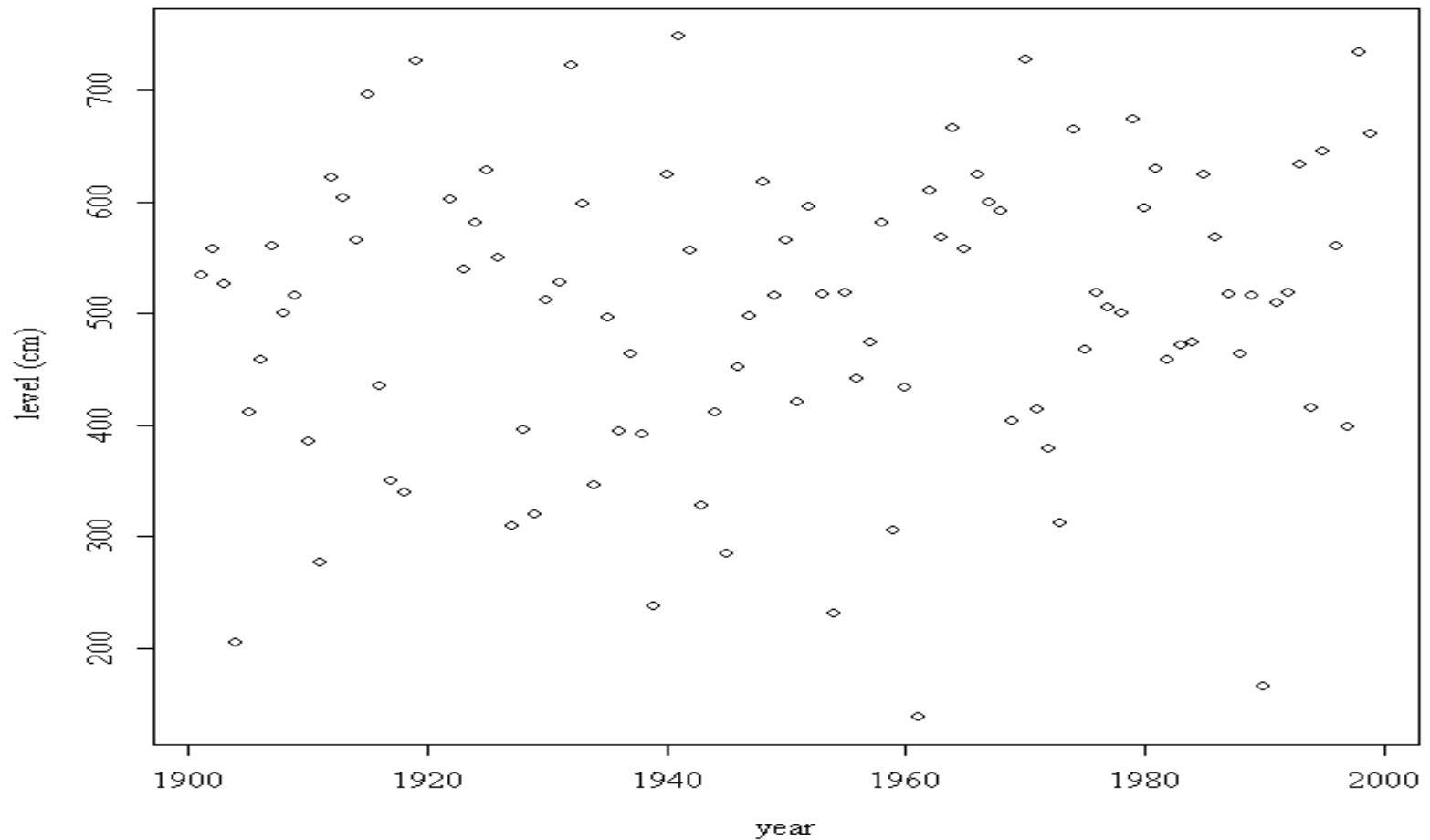
Függetlenség vizsgálata

- Tegyük fel, hogy megtisztítottuk adatainkat.
- A kiindulópont az évi maximumok függetlensége.

Autokorreláció
függvény: $R(X_t, X_{t+k})$
a Záhony vízszint
évi maximum
sorozatára



Évi maximumok, Vásárosnamény



Visszatérési szintek

GEV p -kvantilise: $z_p = \mu - \frac{\sigma}{\xi}(1 - y_p^{-\xi})$, ha $\xi \neq 0$

ahol $y_p = -\log(1 - p)$, $G_{\mu, \sigma, \xi}(z_p) = 1 - p$.
 $\mu - \sigma \log y_p$, ha $\xi = 0$,

az az árvíznagyság, amelyet átlagosan $1/p$ évente egyszer halad meg az éves maximális árvíz.

Annak valószínűsége, hogy $1/p$ évnél előbb megjelenik, nagyobb $1/2$ -nél!

Ha $\hat{\xi} < 0$, akkor az eloszlás becsült felső végpontja

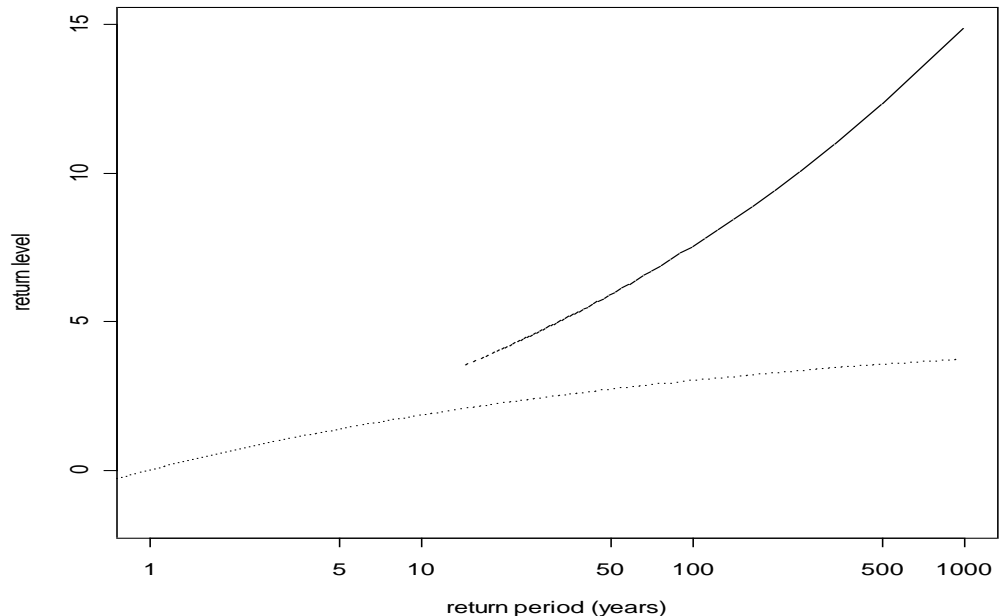
$$\hat{z}_0 = \hat{\mu} - \hat{\sigma} / \hat{\xi}.$$

Visszatérési szint-görbe

- z_p értékét ábrázoljuk $\log(1-p)$ –vel szemben, logaritmikus skálán.
 - Lineáris, ha $\xi = 0$,
 - konkáv, határértéke $\mu - \frac{\sigma}{\xi}$ ha $\xi < 0$
 - konvex, ha $\xi > 0$

Folytonos: $\xi = 0.2$

pontozott: $\xi = -0.2$



Küszöb fölötti csúcsok módszere (POT)

- Azok az események extrémek, amelyek meghaladnak egy rögzített, magas küszöböt
- Előnyei:
 - Több adatot lehet használni
 - A becsléseket nem befolyásolják kicsi “árvizek”
- Hátrányai:
 - Függ a küszöb megválasztásától
 - A declusterezés (annak eldöntése, hogy mely maximumok származnak egy eseményből) nem mindig egyértelmű.

Elméleti alapok

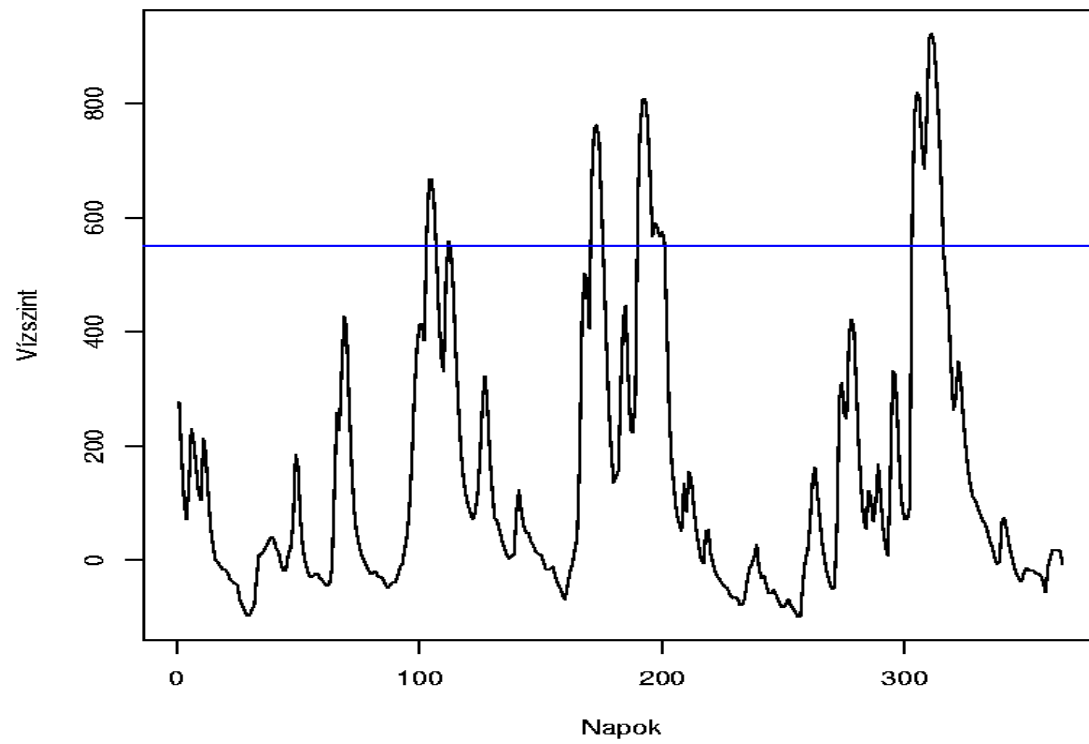
Legyenek X_1, X_2, \dots, X_n független, azonos eloszlású val. változók. Ha ennek a sorozatnak a normalizált maximuma konvergál egy extrém-érték eloszláshoz (μ, σ, ξ paraméterekkel), akkor

$$P(X - u < y \mid X > u) \approx 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}$$

ha $y > 0$ és $1 + \xi y / \tilde{\sigma} > 0$ ahol $\tilde{\sigma} = \sigma + \xi(\mu - u)$

(Általánosított Pareto eloszlás, GPD.)

Az aszimptotika n és u végtelenhez tartása mellett érvényes.

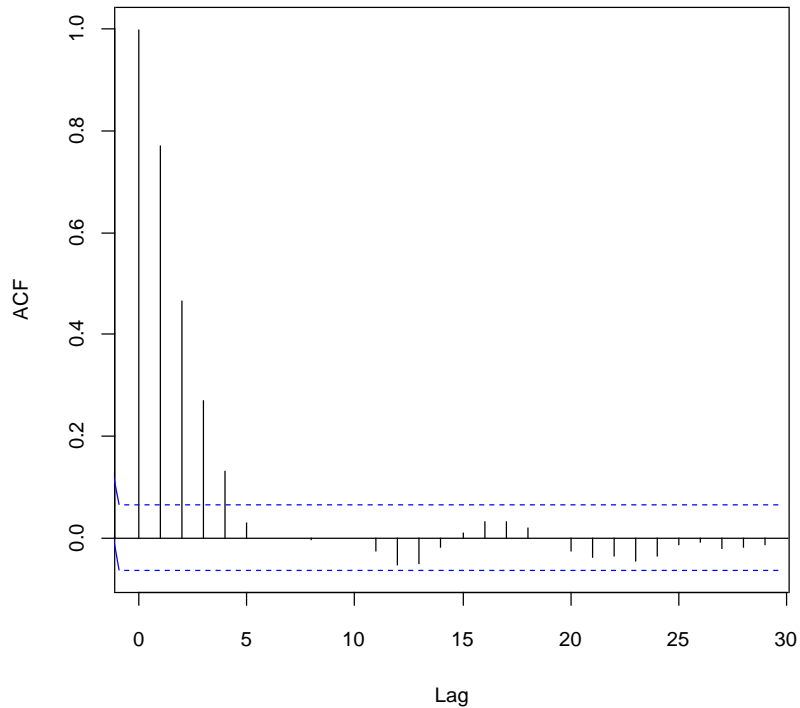


Autokorreláció függvény, Záhony vízszint

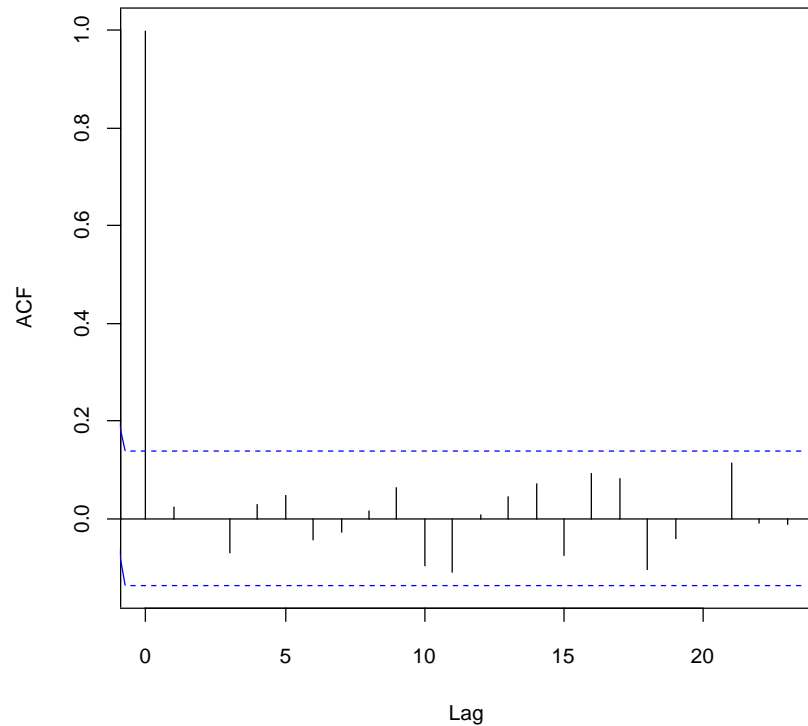
baloldal: minden 400 fölötti értékre

jobb oldal: deklaszterezés után (a csúcsok)

Series dat2[, 2]



Series xdat



Visszatérési szintek

Az általánosított Pareto eloszlás p -kvantilise:

$$x_p = u + \frac{\sigma}{\xi} \left[\left(\frac{1}{p} \zeta_u \right)^\xi - 1 \right], \text{ ha } \xi \neq 0;$$

$$x_p = u + \sigma \log \left(\frac{1}{p} \zeta_u \right), \text{ ha } \xi = 0, \text{ ahol}$$

$$\zeta_u = P(X > u). \quad \hat{\zeta}_u = \frac{n_u}{n}$$

Ha n_y az évente észlelt szint feletti maximumok átlagos száma \Rightarrow T évente visszatérő az $1/T * n_y$ kvantilis.

Ha $\hat{\xi} < 0$, akkor az eloszlás felső végpontja $\hat{x}_1 = u - \hat{\sigma} / \hat{\xi}$.

Küszöb választás

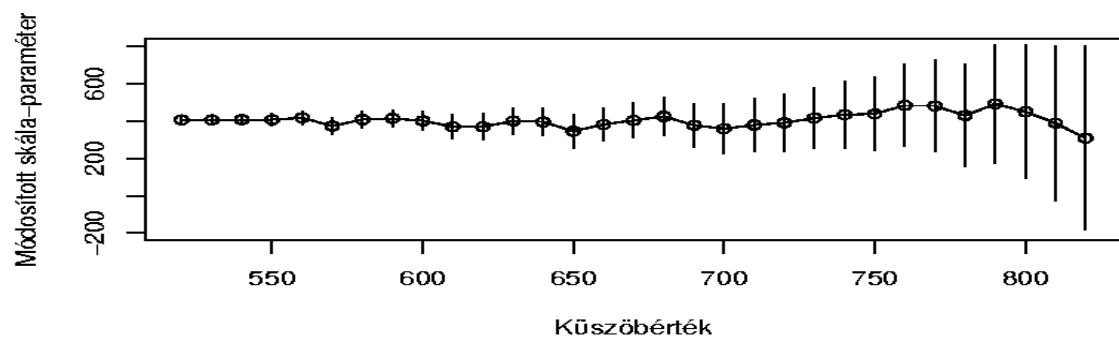
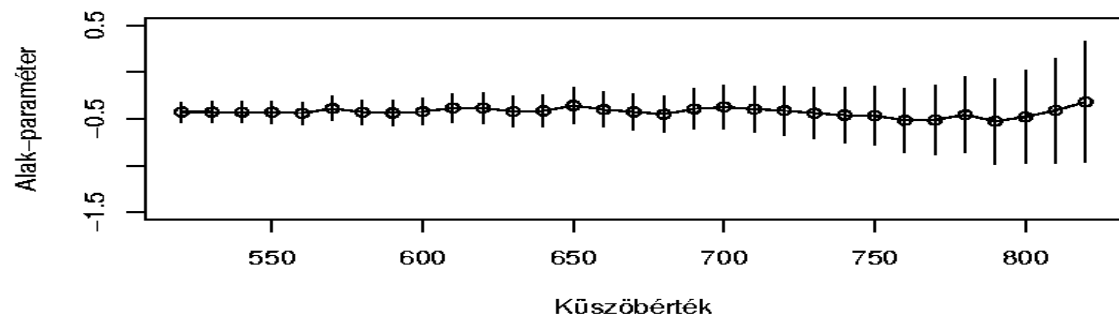
Átlagos meghaladás ábrája:

Tetszőleges u küszöbre ábrázoljuk az $X-u$ átlagát (azokra a megfigyelésekre, amelyekre $X > u$) u függvényében.

Ha a Pareto modell igaz, ez a görbe közel lineáris.

A megmagyarázása nehéz lehet a megfigyelések maximumához közel megfigyelhető nagy ingadozása miatt.

Alternatíva: tekintsük a paraméterbecslések értékeit különböző küszöbök esetén.



Stacionaritás

- Kérdés, hogy az adatok valóban tekinthetők-e stacionáriusnak (időben homogénnek; alternatíva: lehet trend/periodikus komponens).
- Lehet klasszikus tesztekkel vizsgálni (pl. chi-négyzet).
- Az egyik módszert be is mutatjuk.

Nemparaméteres megközelítés

- Mann-Kendall trendteszt

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(x_i - x_j)$$

- Független, azonos eloszlású, egyezések nélküli esetre:

$$ES=0, \quad D^2(S) = \frac{n(n-1)(2n+5)}{18}$$

- Ha n nagy, S közelítőleg normális eloszlású. Alkalmazható már $n > 10$ -re is.

Néhány érték (a standardizált S-statisztika értékei)

- Évi maximumok

	vízállás	vízhozam
Szeged	0.235	-0.199
Záhony	0.203	
V.Namény	0.089	

- A napi maximális vízállásra (Záhony): $S=-22.4$. Miért? Az alacsony vízállásértékek lefelé mutató trendje az ok. Ugyanez az érték a vízhozamra: $S=-2.47$

Stacionárius sorozatok

- Ha nem teljesül a függetlenség (mint pl. az eredeti naponkénti méréseknél), a normalizált maximumok határeloszlása továbbra is GEV eloszlás, ha a függőség a távoli megfigyelések között 0-hoz tart. Az évi maximumokra a GEV modell tehát elméletileg is megalapozott.
- POT modellekre, a klaszter-maximumok használhatóak. (A klasztereket definiálni kell).

A nemstacionaritás esete

- Lineáris regressziós modellek beépíthetők a maximum likelihood megközelítésbe.
- Esetleg szétbontva a megfigyeléseket évszakokra, külön-külön teljesülhet a stacionaritás.

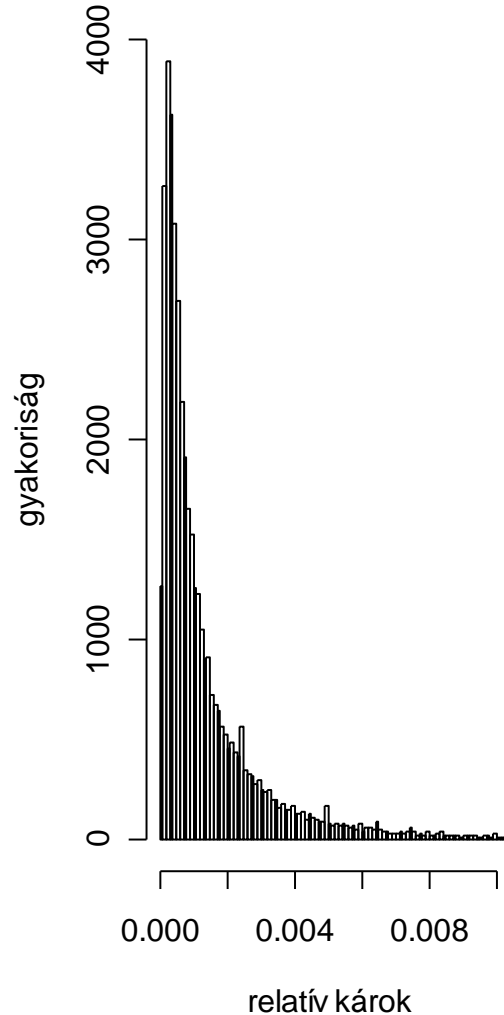
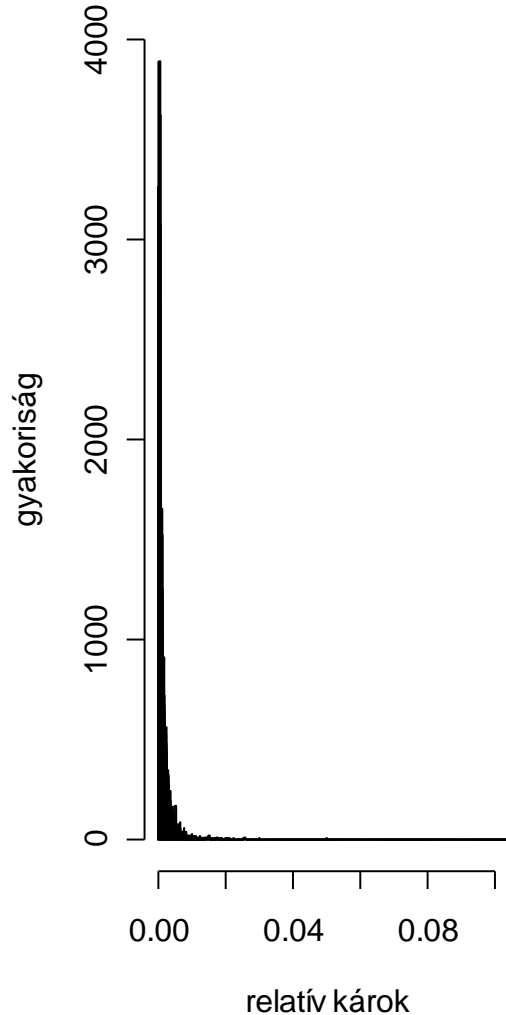
Másik alkalmazás: gépjármű felelősség biztosítás

- Adattisztítás (negatív károk, nem megfelelő időpontok, stb. kiszűrése.)
- Inflációs hatás elemzése (lényeges, mert a mintát azonos eloszlásúnak képzeljük).
 - Negyedéves eltolással 15 db 1 éves részre bontottuk az adatokat.
 - KSH fogyasztói árindex adatok nem megfelelőek (gyorsabb a kárkifizetés növekedése).

Az inflációs hatás becslése

- A kárkifizetési adatok mediánjaira (a kiugró értékek miatt az átlag nem megfelelő!) illesztett nemparaméteres simítás eredményeként adódott az „ágazati kárnövekedési ráta” a vizsgált tartományon.
- Ez tartalmazza
 - az inflációt,
 - a gépjármű-állomány megváltozásának hatását
 - minden más, trend jellegű kárnagyság-módosító hatást.
- Az adatok jelenértékre transzformálásához ezt ki kellett egészíteni az időszak elején és végén.

A jelenértékre transzformált adatok



Károk össz-száma:
kb. 38000.

Néhány alapstatisztika:
Medián: $7.9 \cdot 10^{-4}$
Átlag: 0.002
Felső kvartilis: 0.0018
99%-os kvantilis: 0.0187
99.9%-os kvantilis: 0.1

A biztosítási kockázat és az extrém-érték elemzés

- A legnagyobb kockázatot a nagy károk jelentik. A mi esetünkben:

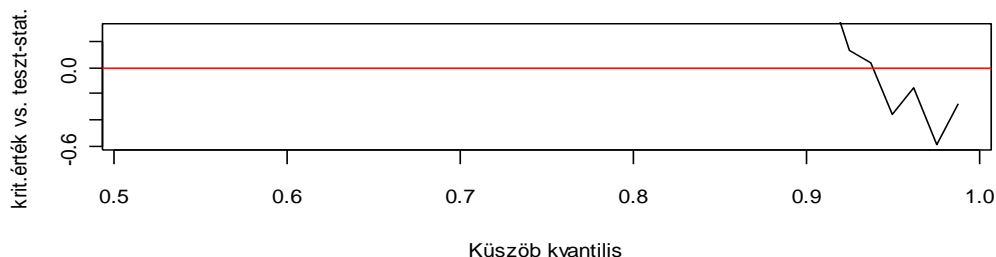
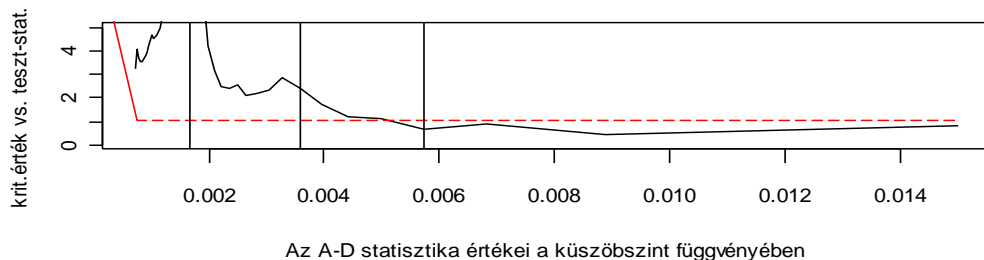
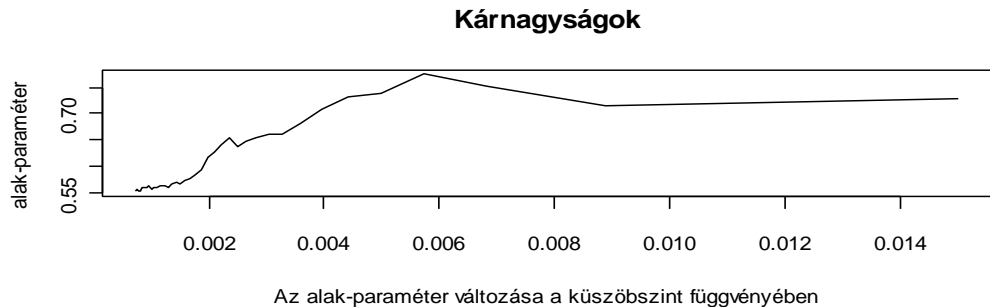
Kvantilis	50%	75%	90%	95%	99%	99.9%
-----------	-----	-----	-----	-----	-----	-------

Részarány	7.7%	22%	41.2%	53.1%	71.8%	87%
-----------	------	-----	-------	-------	-------	-----

Azaz a kárkifizetés közel feléért a legnagyobb 5% a felelős.

- Nincsenek természetesen adódó blokkok (évi maximumok).

Alkalmazás a küszöbválasztáshoz



Olyan küszöböt választunk, melyre a próba elfogadja a GPD modell illeszkedését.

Az ábrából leolvasható, hogy 0.003 feletti szintek jönnek számításba.

A szokásos gond: torzítás (alacsony szintnél) vs. nagy szórás (magas szintnél)

Választásaink: 0.0035, 0.005

Modell diagnosztika

- Valószínűségi ábra (P-P plot), a pontjai:

$$\left\{ \left(\hat{F}(x_i^{(n)}), \frac{i}{n+1} \right) \right\}$$

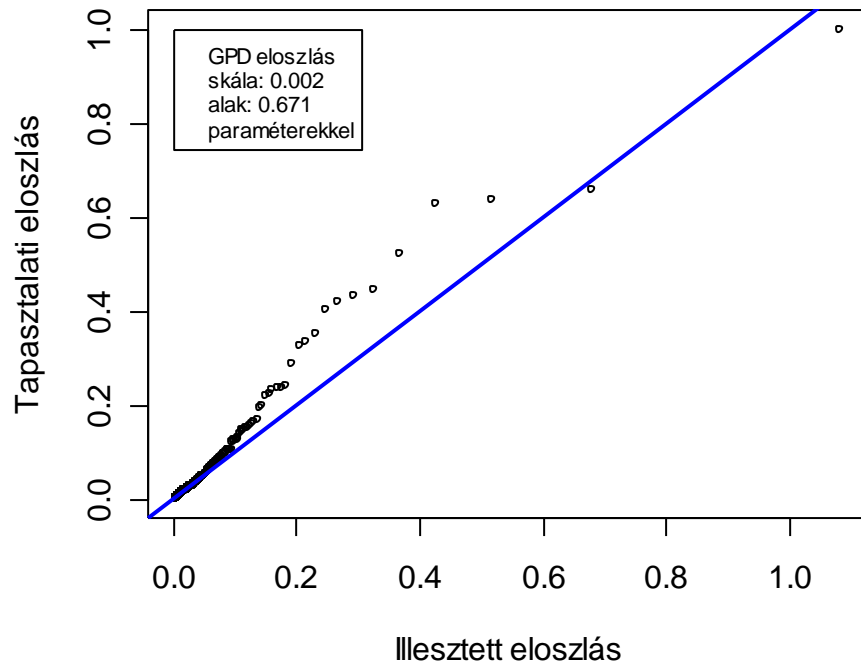
- Kvantilis ábra (Q-Q plot), a pontjai:

$$\left\{ \left(x_i^{(n)}, \hat{F}^{-1}\left(\frac{i}{n+1}\right) \right) \right\}$$

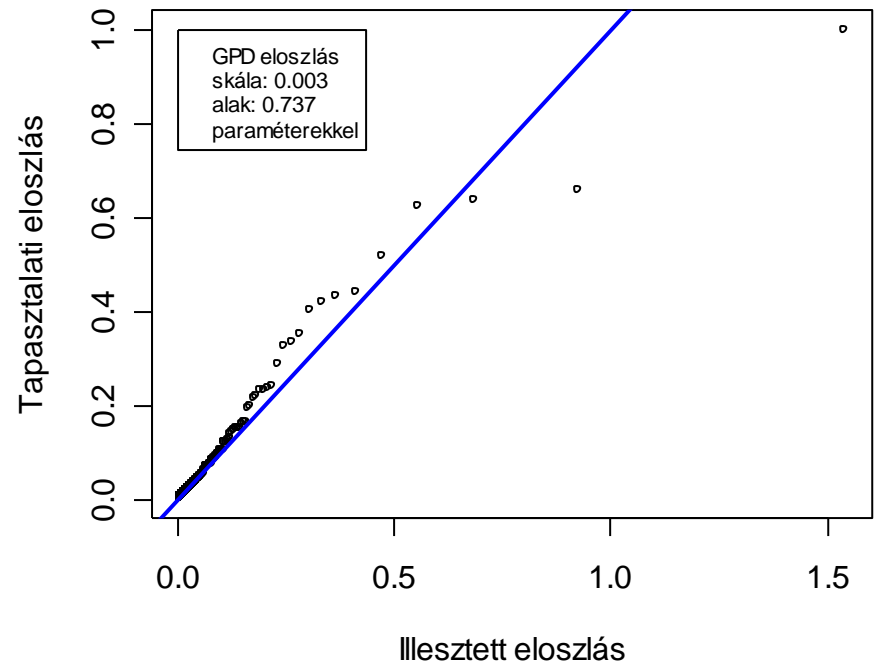
Mindkét esetben a pontok közel kell, hogy legyenek a fődiagonálishoz, ha jó az illeszkedés.

Pareto eloszlás illesztése az adott szintet meghaladó kárkifizetési adatokra

QQ-plot
A küszöbszint: 0.0035



QQ-plot
A küszöbszint: 0.005



Nem jó az illeszkedés

A kapott eloszlások ugyan véges várható értékűek, de a szórás végtelen.

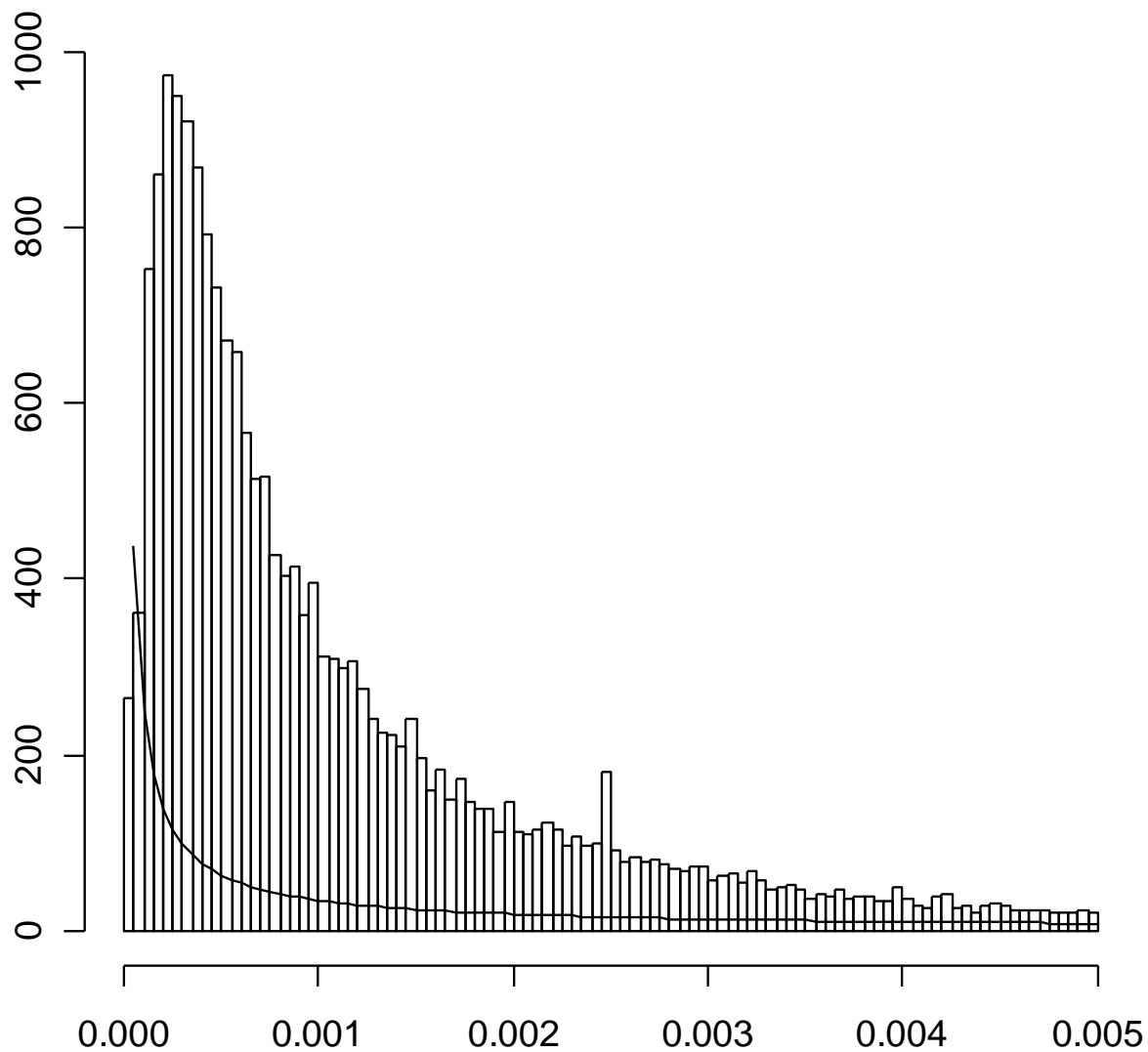
Alternatíva: lognormális eloszlás

- A sűrűségfüggvénye: $f(x) = \frac{1}{\sqrt{2\pi\sigma x}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$
- A paraméterek becslései:

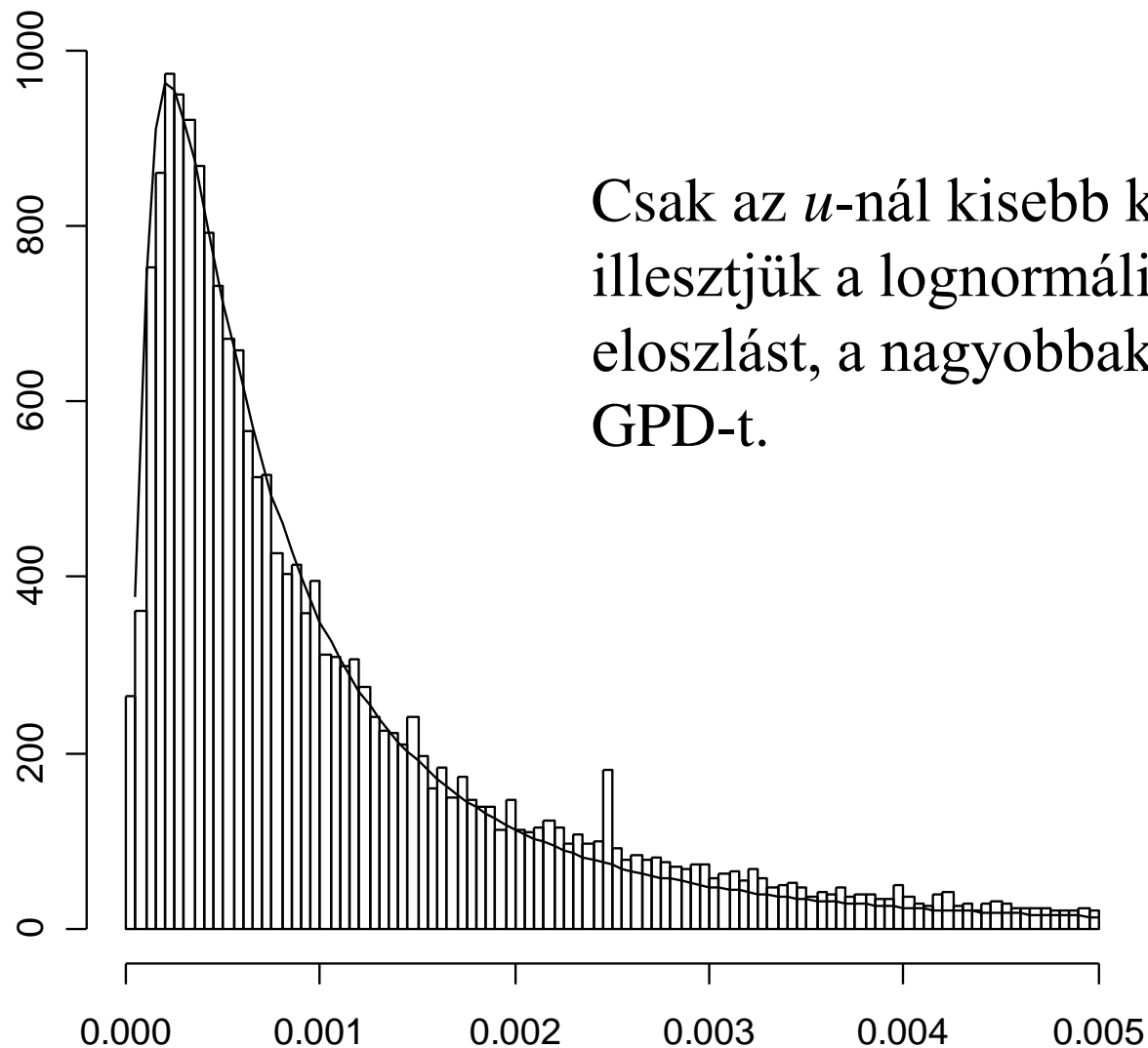
$$\hat{\mu} = \frac{\sum_{i=1}^N \ln(x_i)}{N}, \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N (\ln(x_i) - \mu)^2}{N}}$$

- Az illeszkedés nem jó a teljes adatsorra (túl gyors a lecsengése a ténylegesen fellépő nagy károkhoz képest).

Lognormális eloszlás illesztése a relatív kárnagyságra



Lognormális eloszlás illesztése a relatív kárnagyságra



levágás: $x < 0.005$

Tovább lépés

- A GPD illeszkedés nem volt megfelelő annak ellenére, hogy a statisztika – feltehetően elsősorban a közepesen nagy mintaelemek nagy számának és viszonylag jó illeszkedésének köszönhetően – elfogadta a GPD-modellt.
- Továbbfejlesztett modell: késlekedési időtől való függés figyelembe vétele.

A késlekedési idő

- A néhány, rendelkezésre álló adat egyike.
- Alapstatisztikái:
 - Átlag: 17 nap,
 - Medián: 4 nap,
 - Felső kvartilis: 7 nap
 - Maximum: 1515 nap.

Beépíthető a modellbe, feltételezéseink:

- Skálaparaméter változása: $\sigma(t) = \alpha \cdot t + \beta$
(t a késlekedési idő)
- Alakparaméter nem változik.
- Maximum likelihood becslési módszer most is alkalmazható.

Illeszkedésvizsgálat a háttérváltozót is tartalmazó modellben

- A kvantilis (QQ) plot-ot módosítani kell:

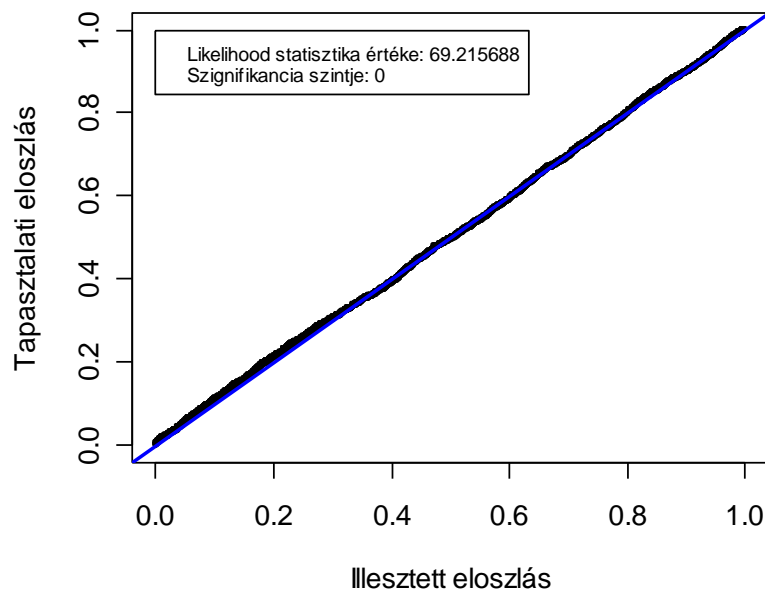
$$\tilde{Y}_{t_i} = \frac{1}{\hat{\xi}} \log \left\{ 1 + \hat{\xi} \left(\frac{Y_{t_i} - u}{\hat{\alpha} \cdot t_i + \hat{\beta}} \right) \right\}$$

standard exponenciális eloszlású, ha teljesül
a modell

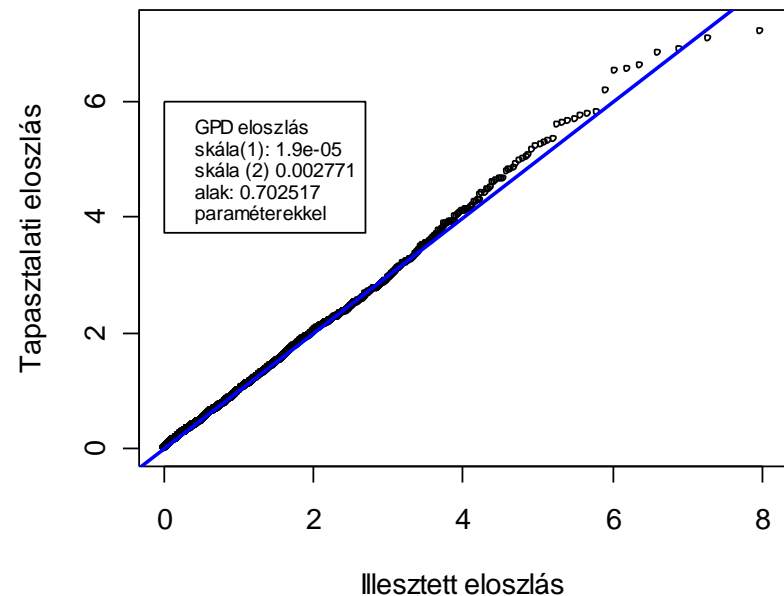
A késlekedési időtől függő modell

- Skálaparaméter változása: $\sigma(t) = \alpha \cdot t + \beta$
(t a késlekedési idő)
- Alakparaméter nem változik
- Maximum likelihood módszer most is alkalmazható

PP-plot, a küszöbszint 0.005
A késlekedési időtől függő



QQ-plot, a küszöbszint 0.005
A késlekedési időtől függő

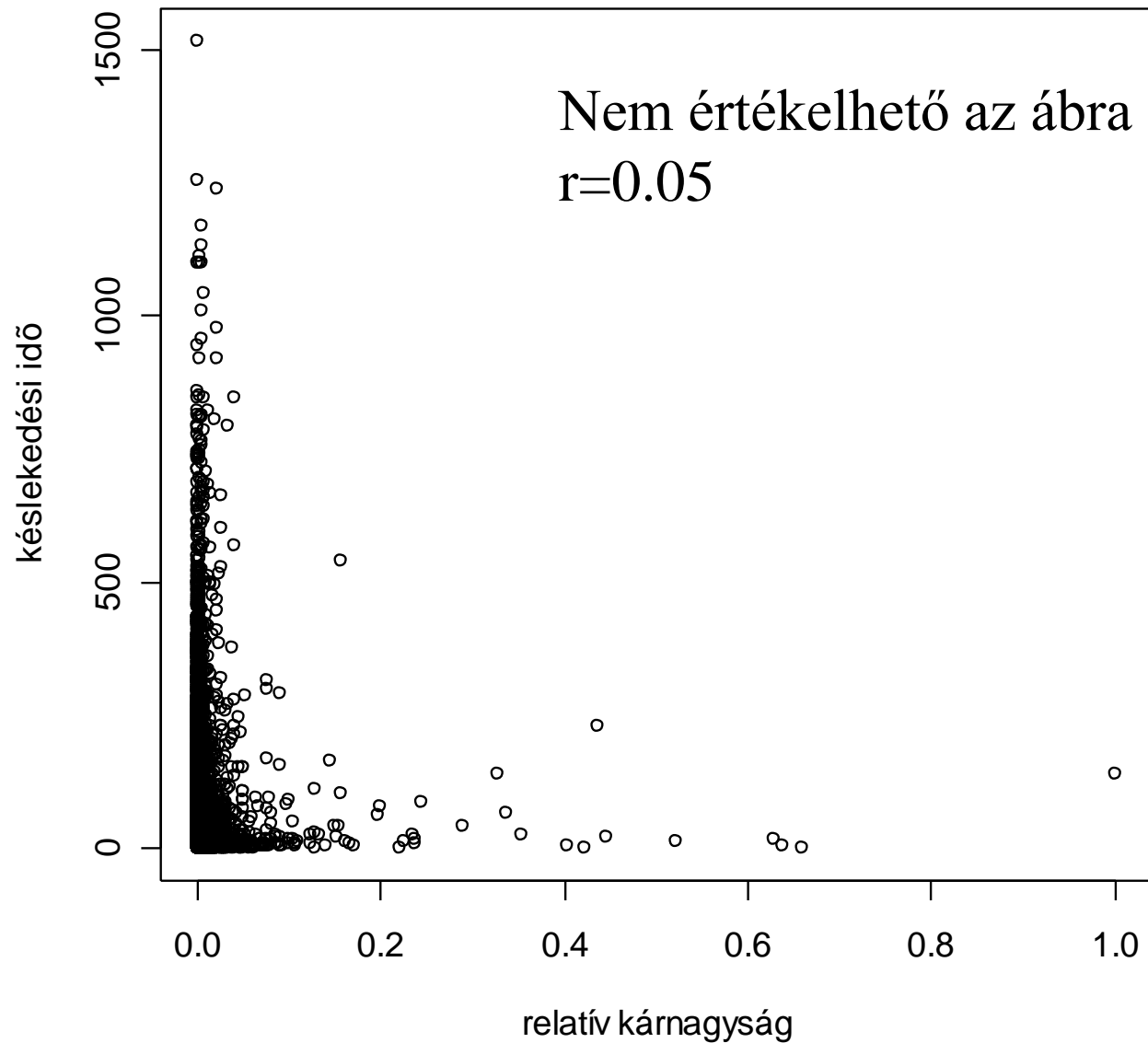


Modell szignifikanciavizsgálata

$$D = 2\{l_1(M_1) - l_0(M_0)\}$$

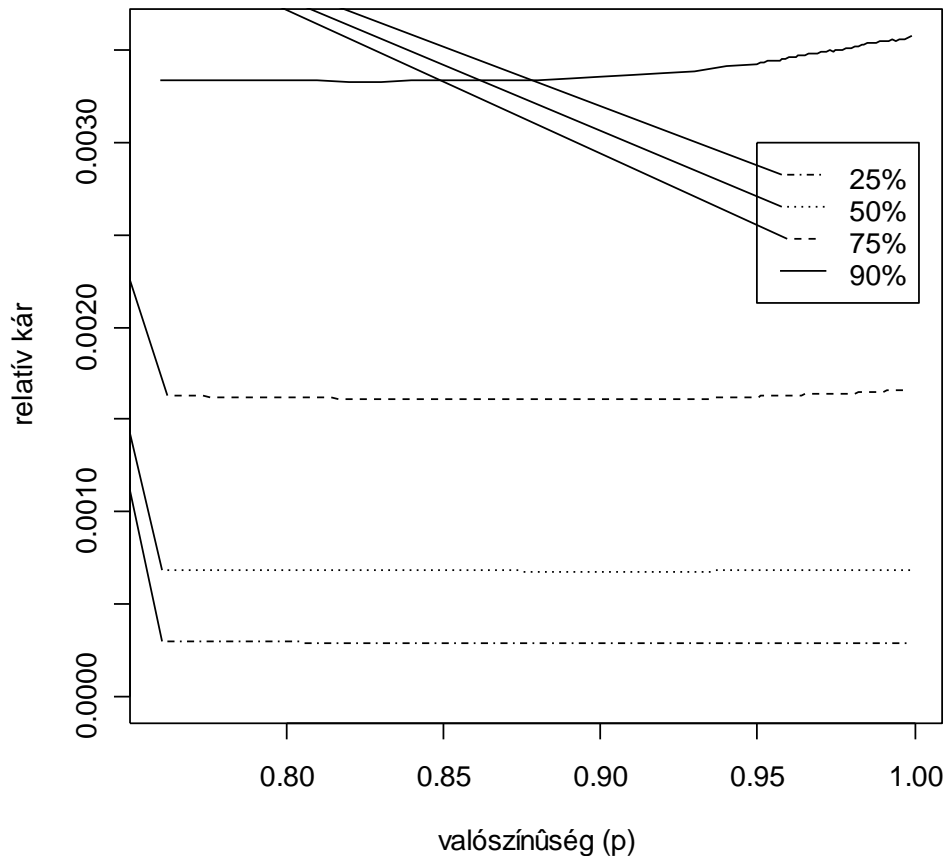
1 szabadságfokú χ^2 eloszlású, ha nincs szignifikáns lineáris trend a skálaparaméterre. Ennek az értéke most $D=102.3$, illetve $D=69.21$, ami gyakorlatilag tetszőlegesen kicsi p mellett szignifikáns hatást mutat.

A késlekedési idő és a kárnagyság

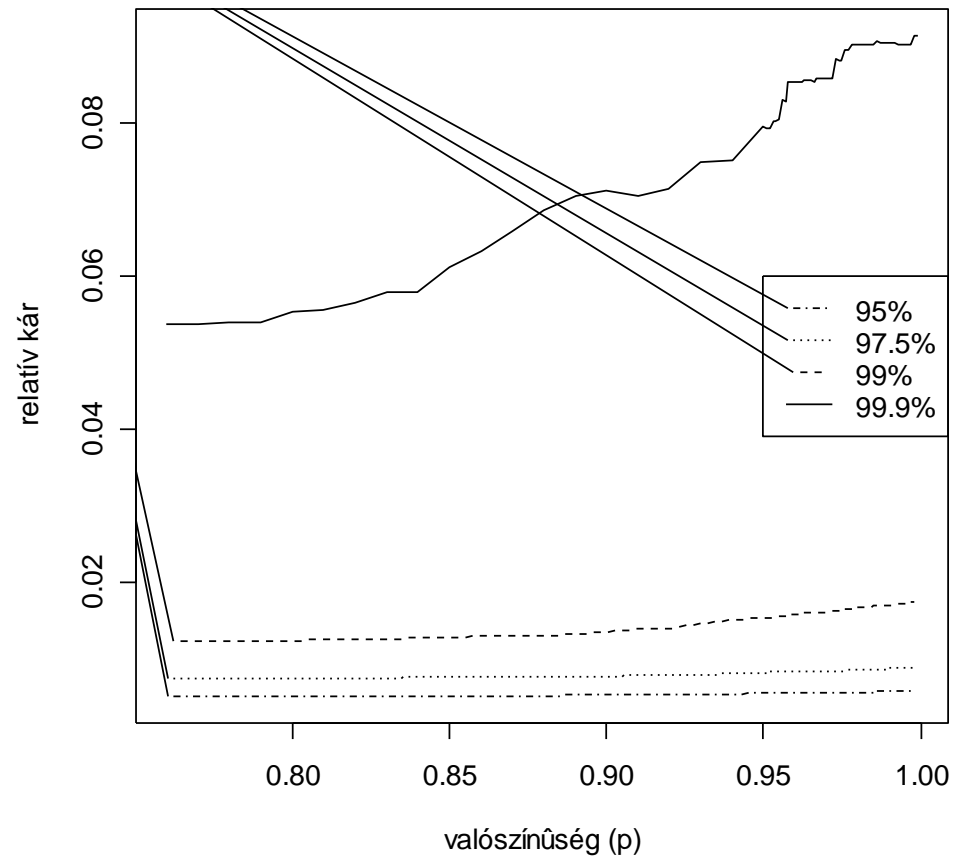


Az ábrák mutatják, hogy csak a magas kvantilisok érzékenyek a késlekedési időre – ezért kaptuk az erős összefüggést a Pareto eloszlás illesztésénél

A kárfizetés feltételes kvantilisei
Feltétel: a késlekedés kisebb a p-kvantilisénél



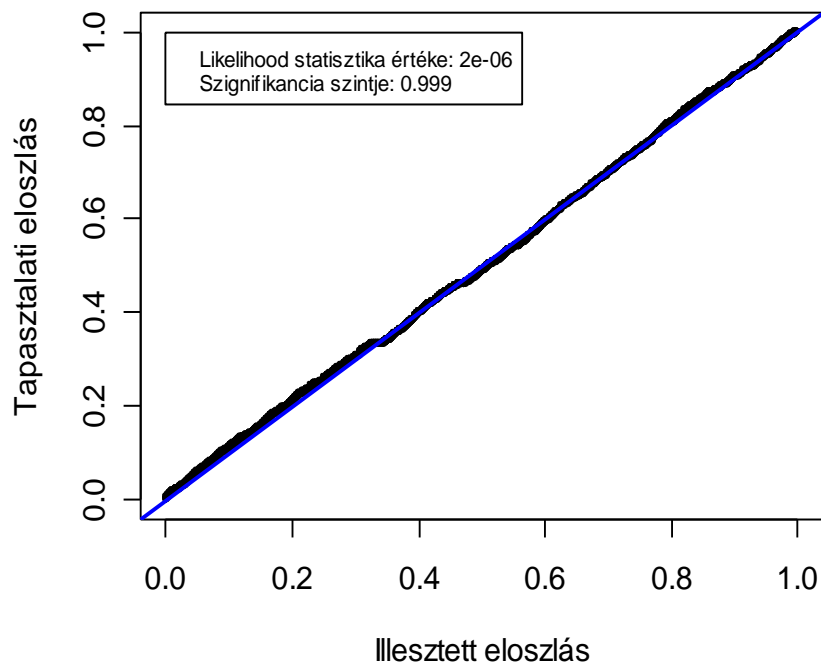
A kárfizetés feltételes kvantilisei
Feltétel: a késlekedés kisebb a p-kvantilisénél



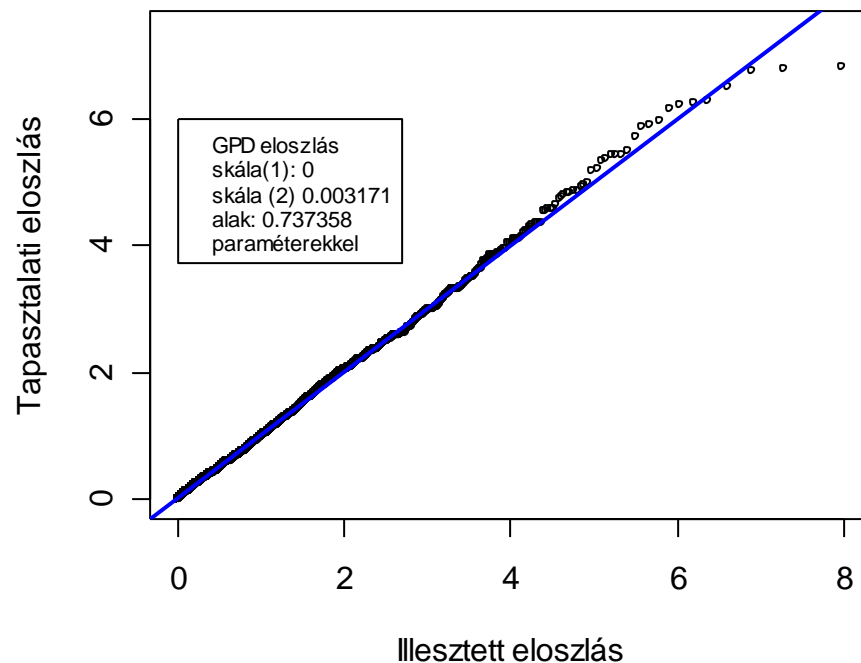
További lehetőségek

- A kárkifizetés ideje már egyáltalán nem jelentkezik tényezőként (azaz elfogadható az infláció kiszűrésére alkalmazott modell)

PP-plot, a küszöbszint 0.005
Bekövetkezés idejétől függő

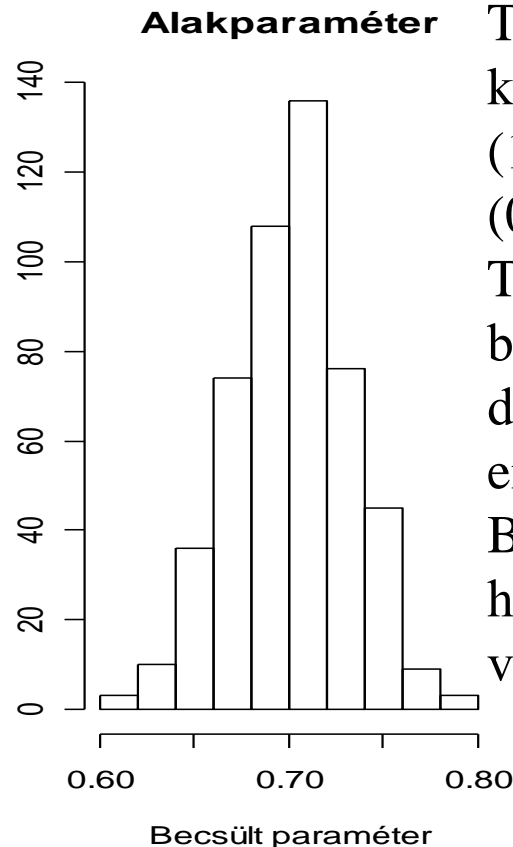
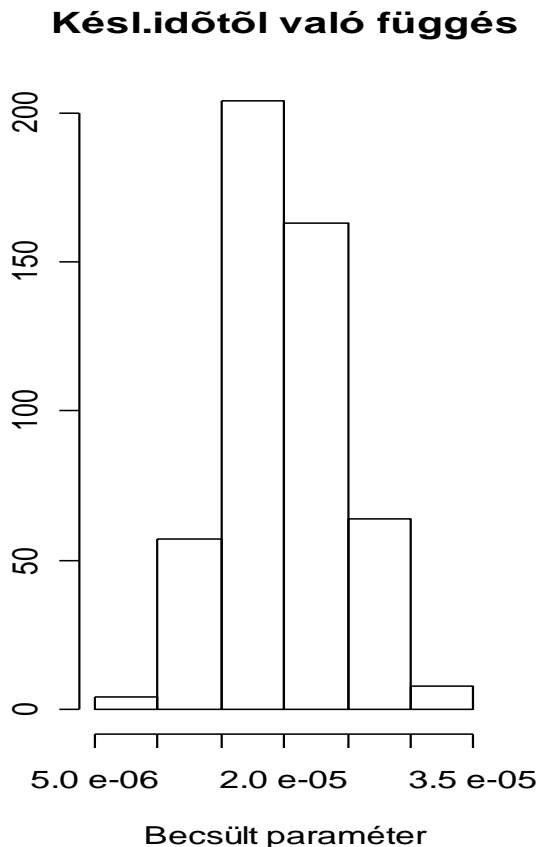


QQ-plot, a küszöbszint 0.005
Bekövetkezés idejétől függő



A becslések bizonytalansága

- Szimulációs vizsgálatok:
 - bootstrap (az eredeti mintából visszatevése mintákat véve) a 0.005 küszöb-értékre



Tapasztalati 95%-os
konfidencia-intervallumok:
 $(1.181; 2.928) \cdot 10^{-5}$

$(0.64; 0.796)$

Tehát az időfüggő paraméter
becslése bizonytalanabb,
de a nagyságrendje
ennek is pontos.

Bizonyosnak tekinthető tehát,
hogy a kárösszeg nem
véges szórású.

Konfidencia intervallumok más esetekre

- Ha nem lenne szerepe a késlekedési időnek, kisebb lenne a szórása a paraméterbecsléseknek, például: $(0.724; 0.750)$ lenne a 95%-os konfidencia intervallum az alakparaméterre.
- A modellt feltételezve, GPD-eloszlásból is generáltunk mintákat. Az alakparaméter ingadozása itt is hasonló volt az általunk kapott értékekhez: $(0.637; 0.76)$ a 95%-os konfidencia intervallum, tehát a modell alkalmazása ebből a szempontból is reális.

További kérdések

- A biztosító számára például az évi összkárkifizetés lényegesebb mennyiség.
- Szimuláció: a
 - nagy károkat a becsült paraméteres modellel közelítve,
 - a késlekedési időt a tapasztalati eloszlásával.
- Eredmények (a megfigyelt kifizetés %-ában):
 - az esetek 0.5 %-ában $> 150\%$
 - az esetek 0.2%-ában $> 200\%$
- Pontosabb szimulációhoz/vizsgálathoz a kisebb károkkal is kell foglalkozni.
- Itt más eloszlás jön szóba.

Tesztkérdés

- Melyik eloszlás tartozik a Gumbel eloszlás max-vonzási tartományába?
- A: binomiális
- B: egyenletes
- C: normális
- D. Poisson

Tesztkérdés/2

- Melyik modell működött a legjobban a felelősségbiztosítási károk modellezésénél?
- A: a lognormális eloszlás
- B: a Pareto eloszlás
- C: A kisebb károkra a lognormális , a nagyokra a Pareto eloszlás
- D: A kisebb károkra a Pareto, a nagyokra a lognormális eloszlás