4. előadás

Kiegyenlítő számítások MSc 2018/19

Áttekintés

- Extrém érték elmélet
- Monte Carlo eljárások

Extrém érték elmélet

- Bevezetés
- Alapvető módszerek (GEV és POT)
- Extrém érték eloszlások
- Alkalmazás: GNSS integritás vizsgálat

Bevezetés

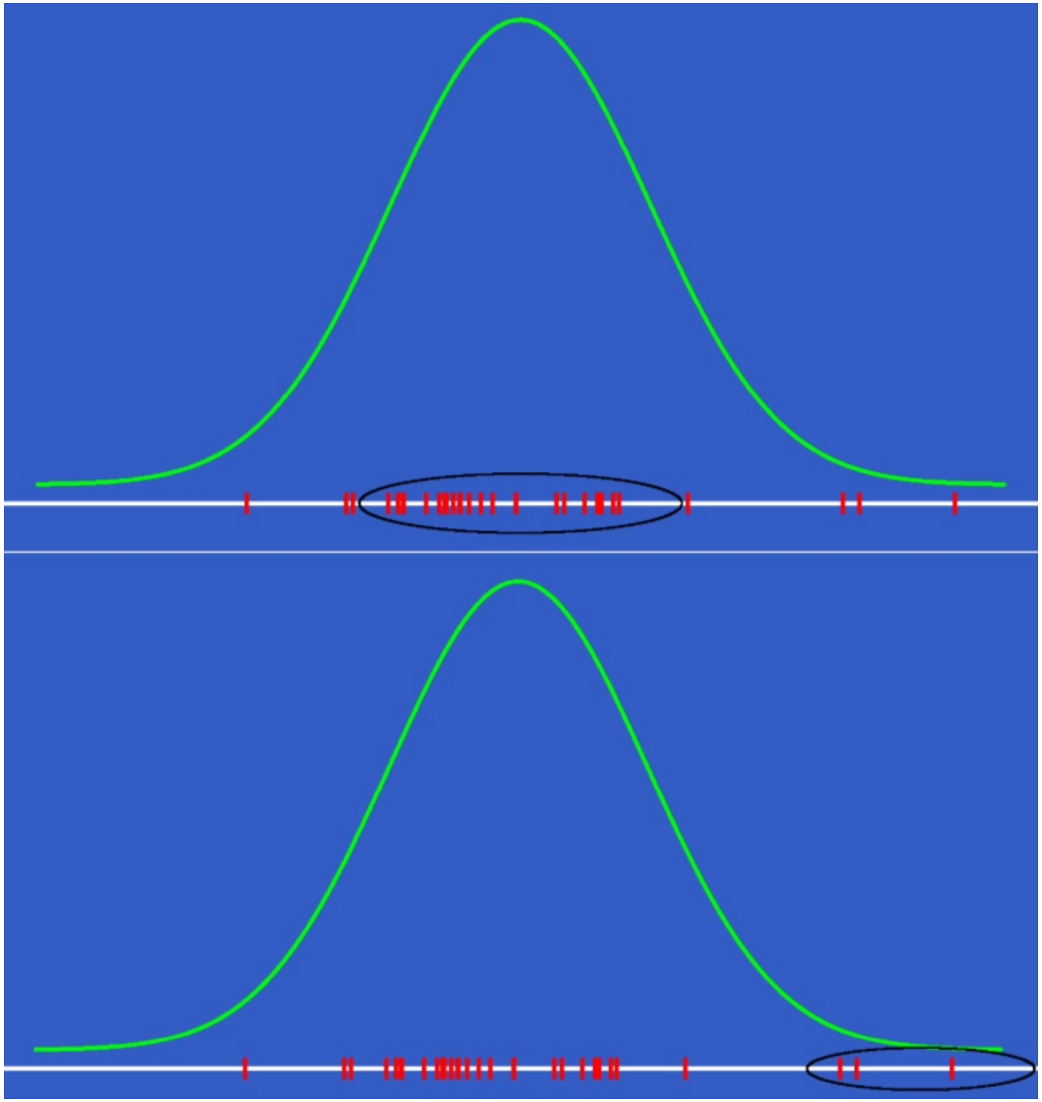
- Az extrém érték elmélet (extreme value theory, EVT) célja olyan *események valószínűségének a becslése*, amelyek kívül esnek a megfigyelt adatok körén
- ilyen események lehetnek
  - szélsőséges időjárási események
  - földrengések, erdőtüzek
  - GNSS életvédelmi célú szolgáltatás meghibásodása
  - nagy kárértékű biztosítási események
  - gépalkatrészek tönkremenetele (kifáradás, korrózió)
  - tőzsdekrach

Bevezetés

- Az extrém érték elmélet alapvetően extrapolációt jelent a mérési adatok körén kívülre
  - könnyen kritizálható, természeténél fogva nem megbízható
  - modell feltevéseken alapszik, a valóság ennél mindig összetettebb
- Szilárd matematikai alapokon nyugszik
  - nem javasoltak még más hihető alternatívát helyette
  - a meglevő adatok legjobb felhasználására törekszik az extrém események valószínűségének kiszámítása céljából
- Óvatosságra van szükség az alkalmazása során

Hagyományos és extrém érték statisztika

- hagyományos statisztika: az átlagos viselkedésre összpontosít
  - központi határeloszlás tétel
- extrém érték elmélet: a rendkívüli és ritka értékekre összpontosít
  - Fisher-Tippett tétel



Alapvető módszerek

- Klasszikus (blokk) módszerek
  (GEV = Generalized Extreme Value)
  - adott időszakon (blokk) belüli maximum értékekkel foglalkozik – mi legyen az adott időszak?
- Küszöb feletti értékek elemzési módszere
  (POT = Peak Over Threshold)
  - adott küszöbérték feletti értékekkel foglalkozik – mi legyen az adott küszöbérték?

Közös jellemzők

- az extrém érték elmélet általánosan alkalmazható függetlenül a mért adatok eloszlásának (pl. Gauss eloszlás) feltételezésétől
- az eloszlások szélei az adatok és néhány általánosságban igaz feltételezés alapján meghatározhatók akkor is, ha nincs adat ebből a tartományból

Eloszlásfüggvények

- **Blokk maximum**

$$M_n = \max\{X_{1,} \ldots , X_n\}$$

ha $n \longrightarrow \infty$, akkor

M_n általánosított extrém érték (GEV) eloszlást követ

- **Küszöb feletti értékek**

$$\{X_i - u | X_i > u\}$$

nagyon nagy u küszöb esetén

általánosított Pareto-féle eloszlást követ

GEV – Fisher-Tippett tétel

$M_n = \max\{X_1, \ldots, X_n\}$ eloszlása $(n \to \infty)$

$$\xi \neq 0 \qquad G(y) = \exp\left(-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}\right)$$

$$\xi = 0 \qquad G(y) = \exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)$$

ami az általánosított extrém érték eloszlás (GEV)

paraméterek: μ hely, σ skála, ξ alak

GEV eloszlás típusok

ξ alak paramétertől függően 3 típusa van

Gumbel $\xi = 0$

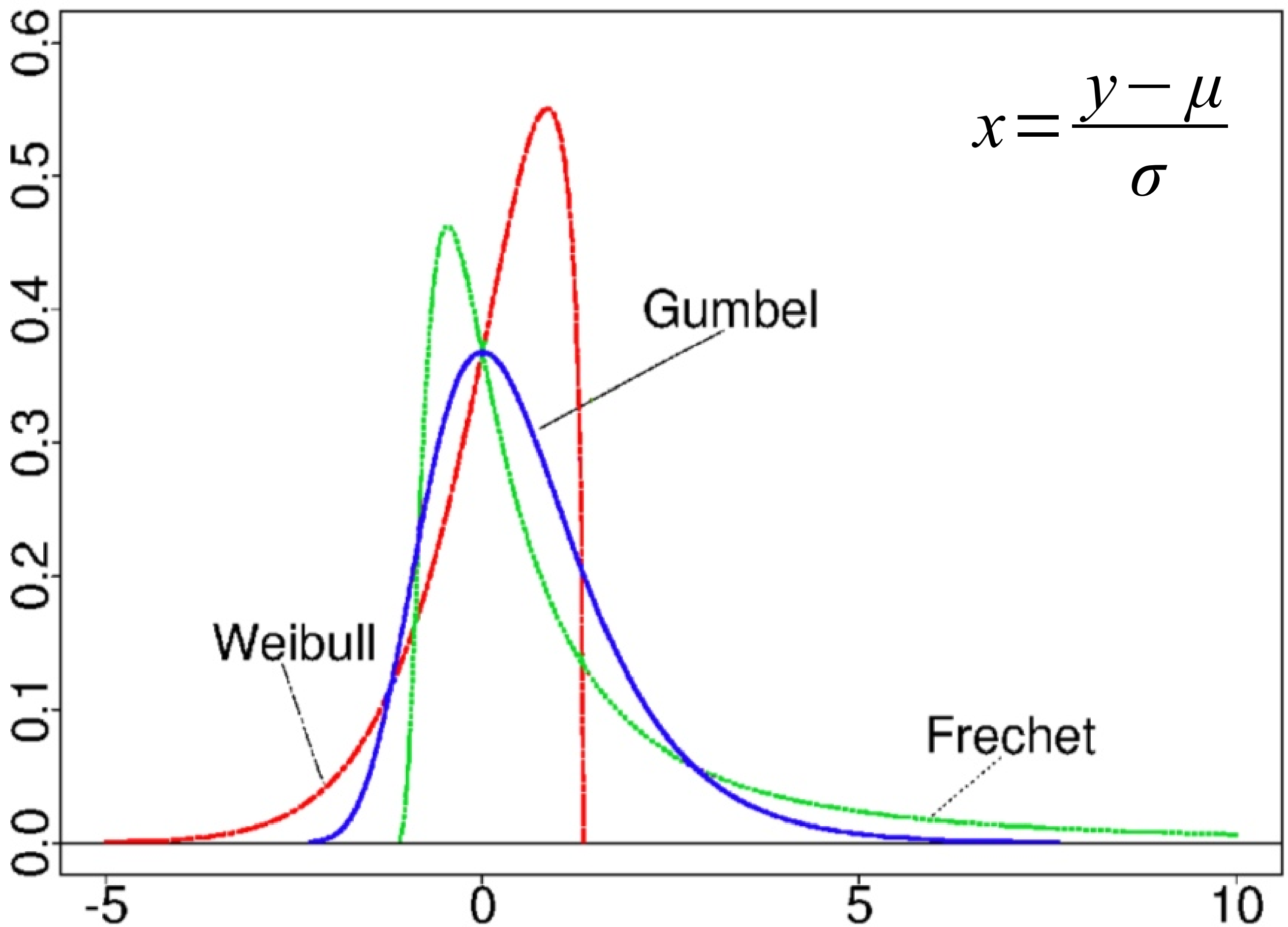
$$G(x) = \exp\left(-\exp[x]\right)$$

Fréchet $\xi = 1/\alpha > 0$

$$G(x) = \exp\left(-\left[1 + \frac{x}{\alpha}\right]^{-\alpha}\right)$$

Weibull $\xi = -1/\alpha < 0$

$$G(x) = \exp\left(-\left[1 + \frac{x}{\alpha}\right]^{\alpha}\right)$$



$$x = \frac{y - \mu}{\sigma}$$

GEV eloszlás típusok

ξ alak paramétertől függően 3 típusa van

Gumbel $\xi=0$

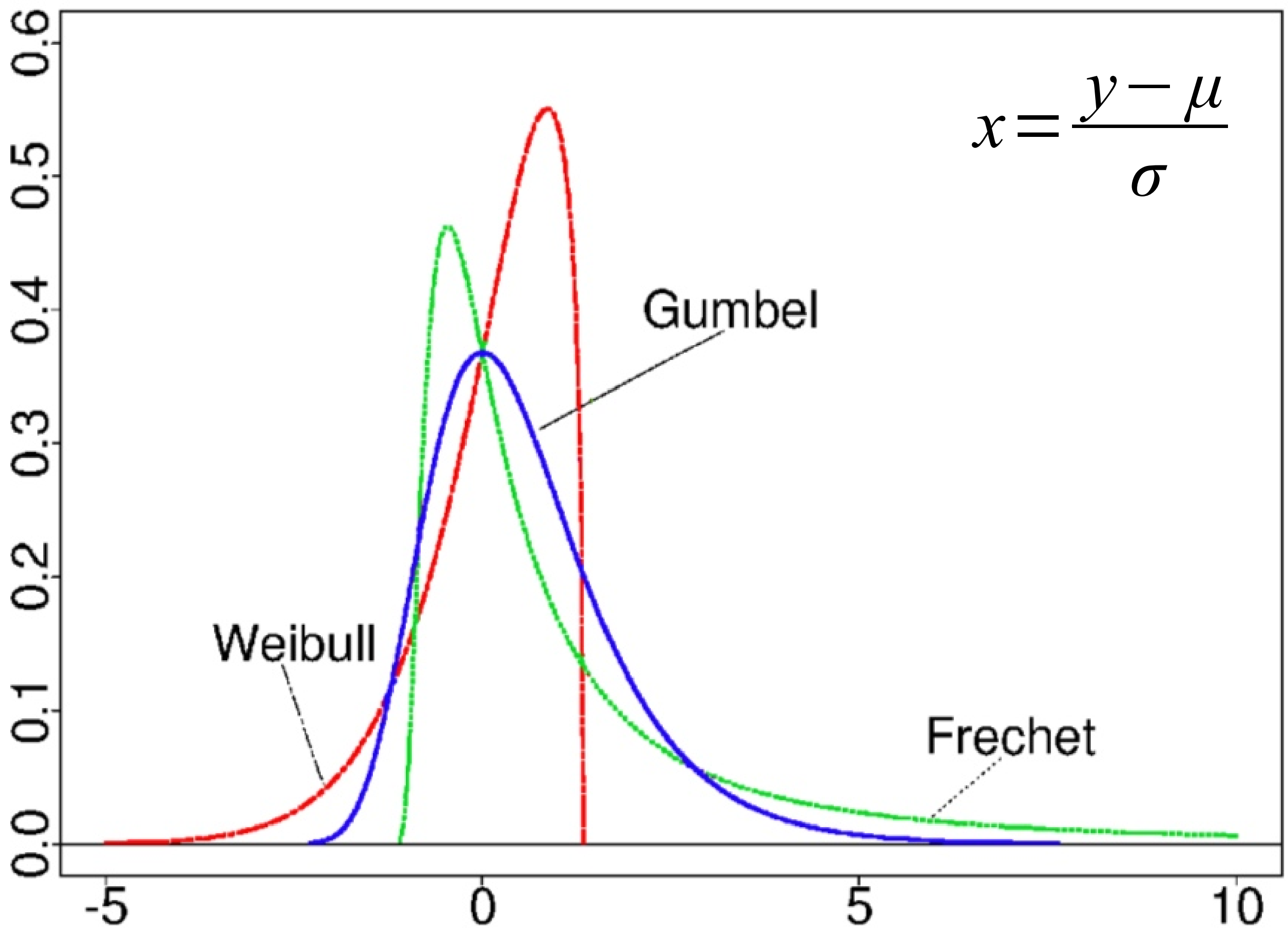
exponenciális szárny

Fréchet $\xi=1/\alpha>0$

súlyos szárny

Weibull $\xi=-1/\alpha<0$

felső végpont



GEV eloszlások

- A $\{X_1, \ldots, X_n\}$ mintára, melyből a maximumok származnak:
- X_i -k azonos eloszlású, egymástól független valószínűségi változók legyenek

Visszatérési szintek

- Adott idő (pl. 25, 50, 100 év) alatt várhatóan egyszer kapunk ilyen vagy magasabb értéket
- Ha *E* év alatt egyszer tér vissza: $p = 1/E$
- GEV p-kvantilise: $G(z_p) = 1 - p$

$$\xi \neq 0 \qquad z_p = \mu - \frac{\sigma}{\xi}\exp\left(1 - y_p^{-\xi}\right)$$

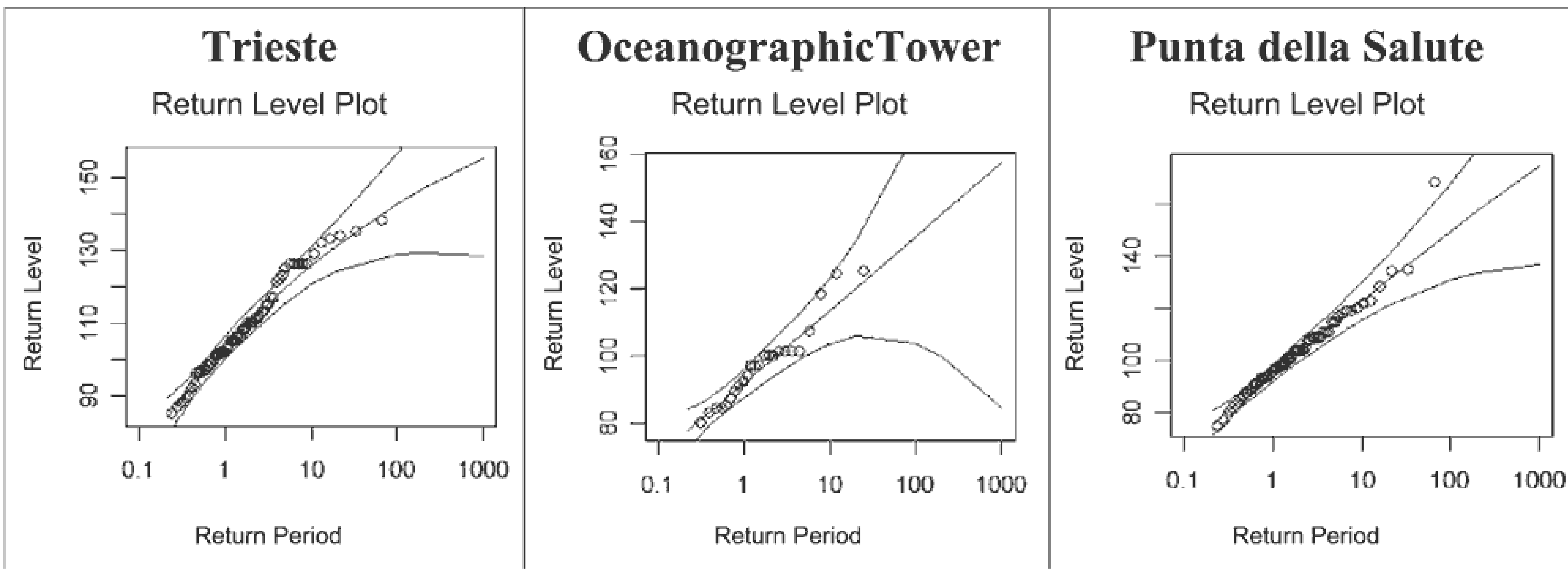
$$\xi = 0 \qquad z_p = \mu - \sigma \log y_p$$

ahol $\quad y_p = \log(1 - p)$

Visszatérési szint görbe

- ábrázoljuk z_p -t $\log(1-p)$ -vel szemben logaritmikus skálán
  - lineáris (egyenes), ha $\xi = 0$
  - konvex görbe, határértéke $\mu - \dfrac{\sigma}{\xi}$, ha $\xi < 0$
  - konkáv görbe, ha $\xi > 0$

Pirazzoli et al. 2007



Küszöb feletti értékek (POT)

- Azok az események extrémek, amelyek meghaladnak egy küszöböt
- Ha a küszöb elég magas, az azt meghaladó adatok hasonlóan viselkednek
- Előnye, hogy általában több adatunk van
- A küszöbérték helyes megválasztása kritikus pont

POT eloszlása

- Az $Y_i = \{X_i - u \mid X_i > u\}$, vagyis egy u küszöbértéket meghaladó X_i értékek eloszlása nagy u értékekre $(u \to \infty)$

általánosított Pareto-féle eloszlást követ (GPD)

paraméterek: σ skála, ξ alak

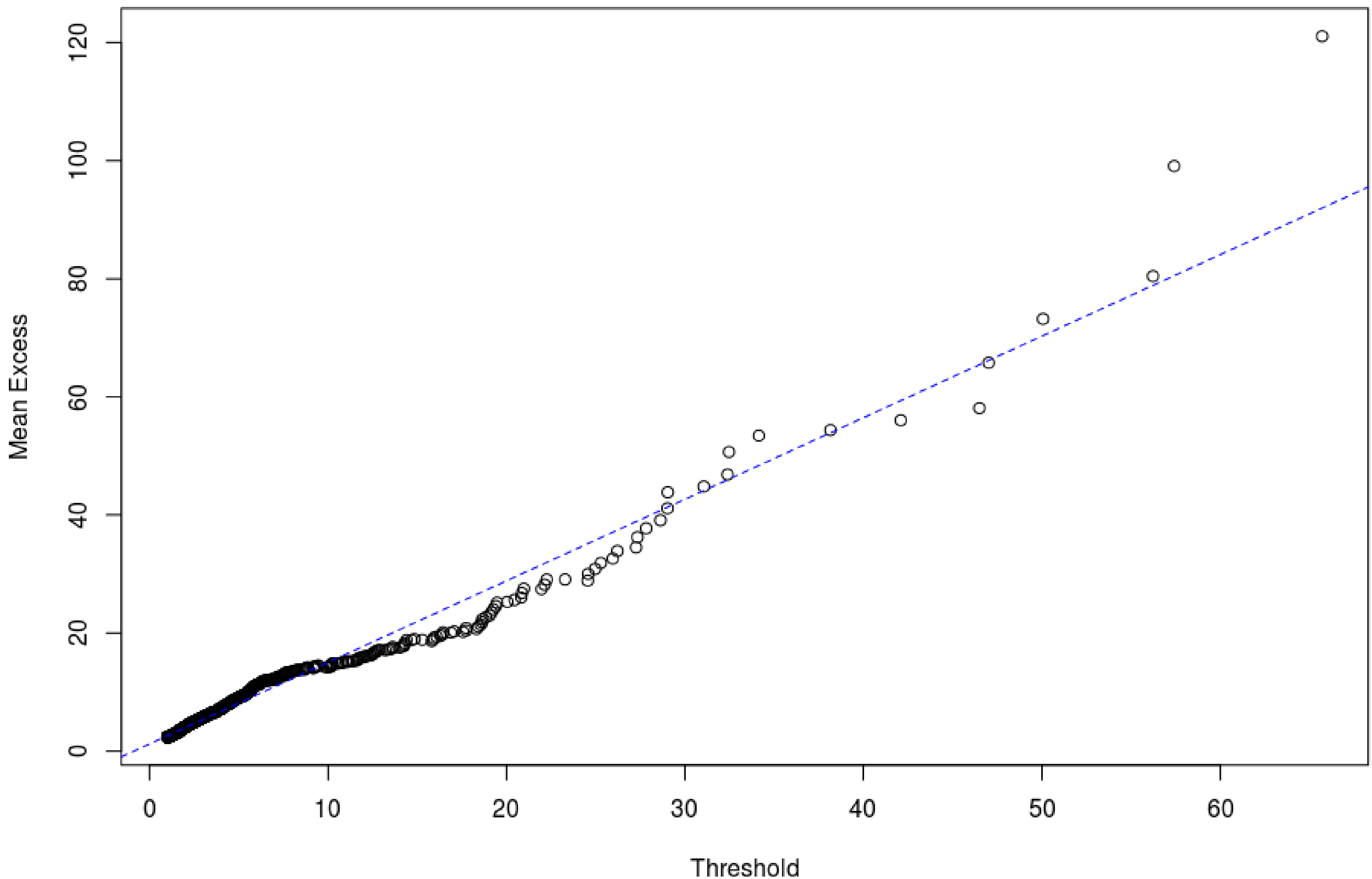
$$H\left(y \mid X_i > u\right) = 1 - \left(1 + \xi \frac{y}{\sigma_u}\right)^{-\frac{1}{\xi}}$$

Küszöb választása: átlagos meghaladás ábra

- Tetszőleges u küszöbre ábrázoljuk az $X_i - u$ várható értékét (átlagát) azokra a megfigyelésekre, melyekre $X_i > u$ az u függvényében
- Ha a Pareto modell igaz, akkor ez a görbe közelítőleg *lineáris* lesz
- Tehát úgy kell megválasztanunk a küszöb értékét, hogy a görbe már lineáris legyen

Átlagos meghaladás ábra

- R csomag: 'evir', meplot(danish)



POT eloszlás típusok

ξ alak paramétertől függően 3 típusa van

Gumbel $\xi = 0$ exponenciális szárny

$$H(y) = 1 - \exp\left(-\frac{y}{\sigma_u}\right)$$

Pareto(Fréchet) $\xi > 0$ súlyos szárny

$$1 - H(y) \sim c y^{-\frac{1}{\xi}}$$

Weibull $\xi < 0$ felső végpont

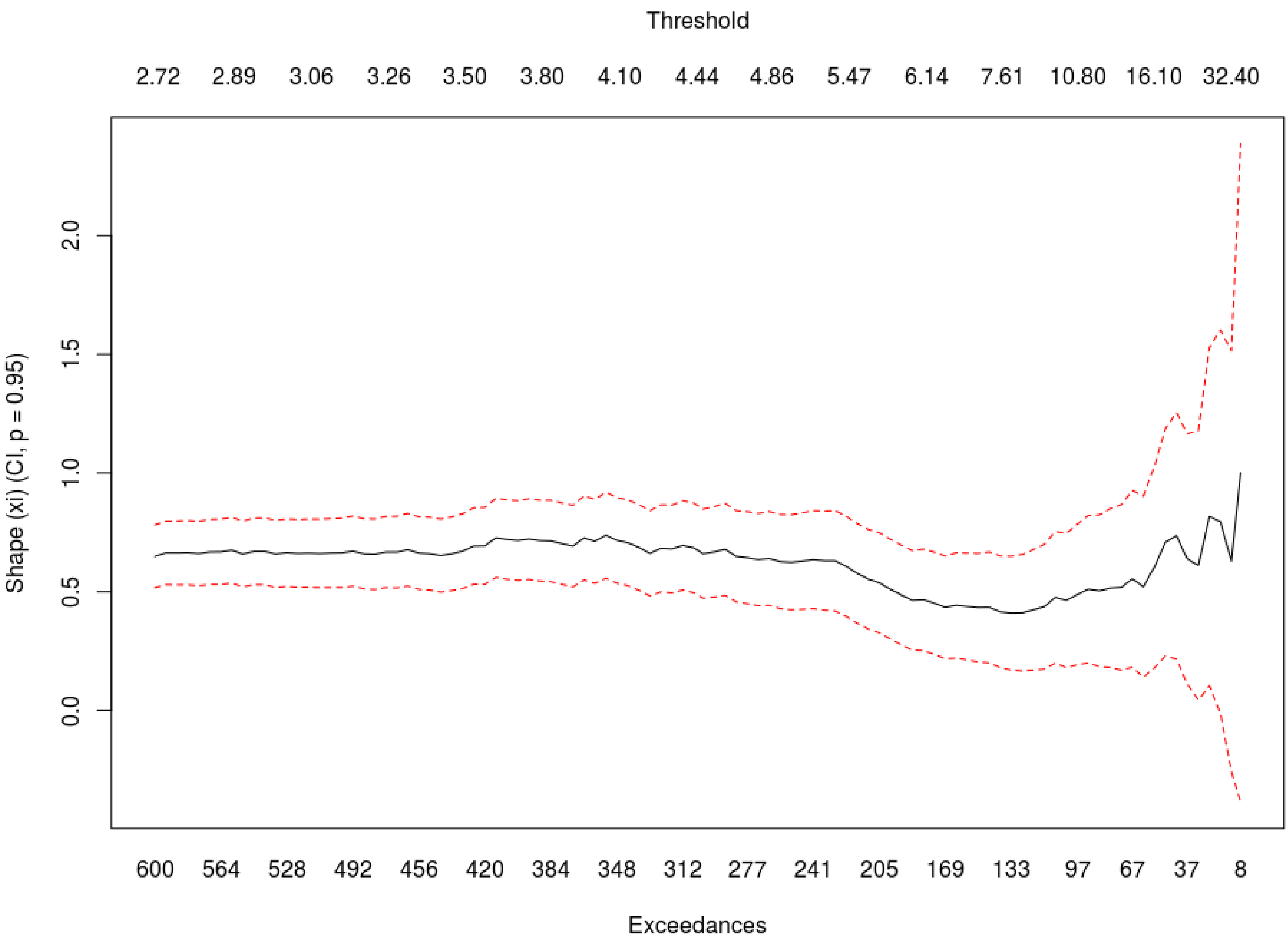
$$\omega_F = \frac{\sigma_u}{|\xi|}$$

Küszöb választás: paraméter - u ábra

- Tetszőleges u küszöbre ábrázoljuk a ξ vagy a σ paraméter becsült értékét azokra a megfigyelésekre, melyekre $X_i > u$ az u függvényében
- Ha a Pareto modell igaz, akkor ez a görbe közelítőleg *konstans* lesz
- Tehát úgy kell megválasztanunk a küszöb értékét, hogy a görbe konstans legyen

ξ(u) stabilitás ábra (parameter stability plot)

- R csomag: 'evir', shape(danish)

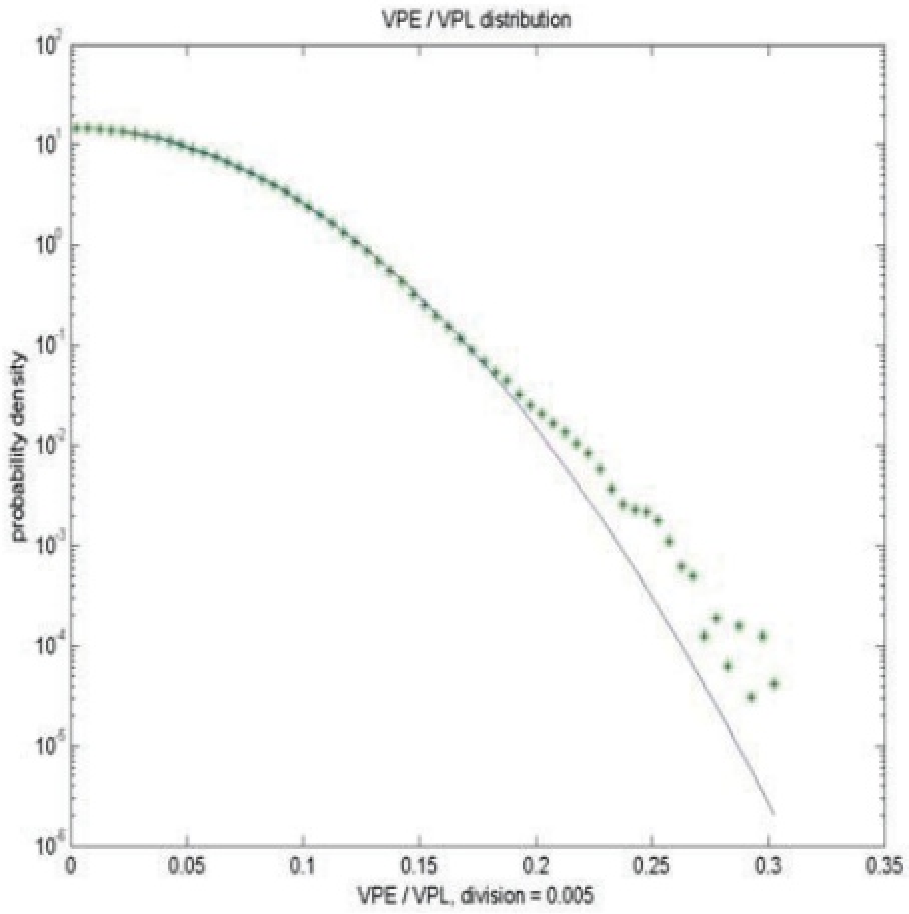


Alkalmazás: GNSS integritás vizsgálat

- légi irányítás szempontjából (ICAO LPV) a hiba (MI: Misleading Information) elfogadható maximális valószínűsége $2 \cdot 10^{-7}$ 150 s alatt (24 évente 1)
  - klasszikus statisztikai módszerekkel nem kezelhető
- extrém érték elmélet jól használható az MI valószínűség becsléséhez
  - GIMAT: GNSS Integrity Monitoring and Analysis Tool

VPE/VPL

- VPE: magassági pozíció hiba
- VPL: magassági biztonsági szint
- VPE/VPL > 1 := ML (félrevezető információ)



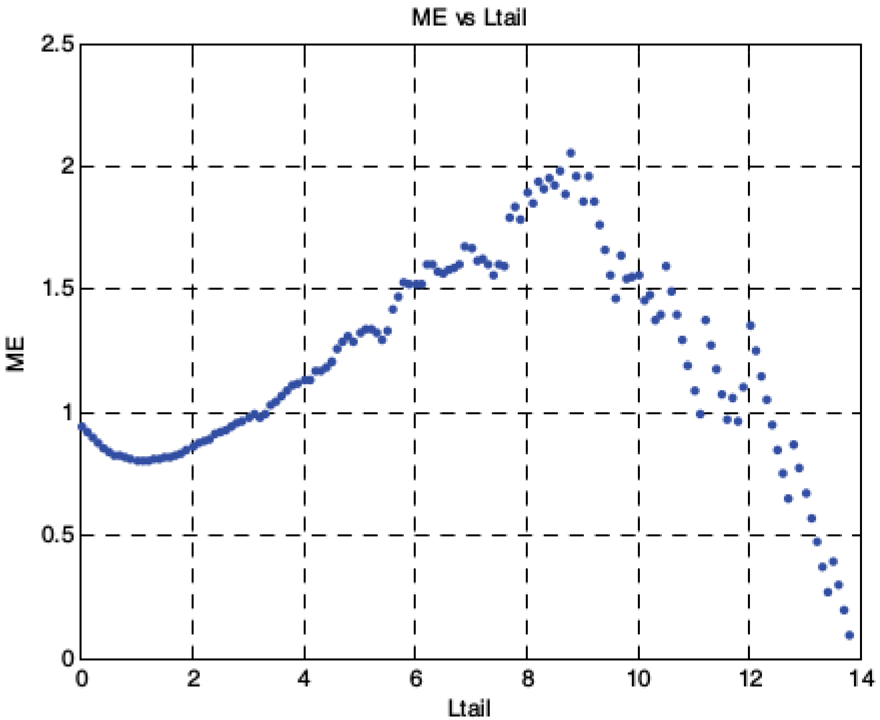
3 havi GPS/EGNOS adatok
(> 6 millió epocha)

vonal: Gauss-eloszlás

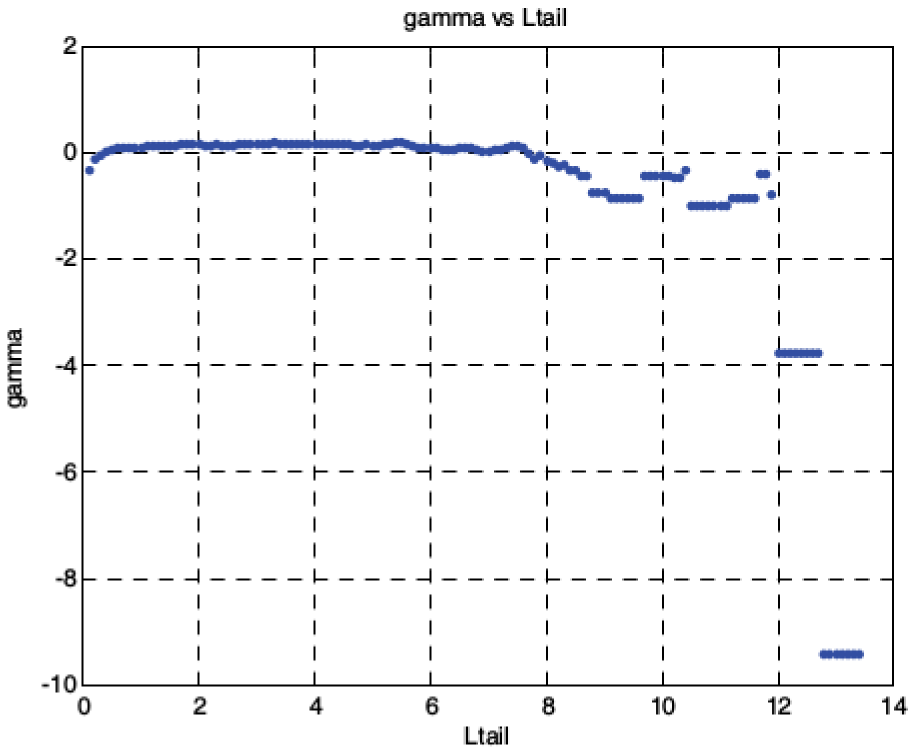
Veerman et al. 2009

Extrém érték elmélet alkalmazása (GIMAT)

átlagos meghaladás ábra



paraméter stabilitás ábra



Monte Carlo eljárások

- Számítási algoritmusok széles köre, melyek ismételt véletlen mintavételezéssel állítanak elő számszerű eredményeket
- Hagyományos eszközökkel nehezen vagy egyáltalán nem vizsgálható problémák
- Fő alkalmazási területek:
  - optimalizáció
  - numerikus integrálás
  - adott eloszlás szerinti minta generálása

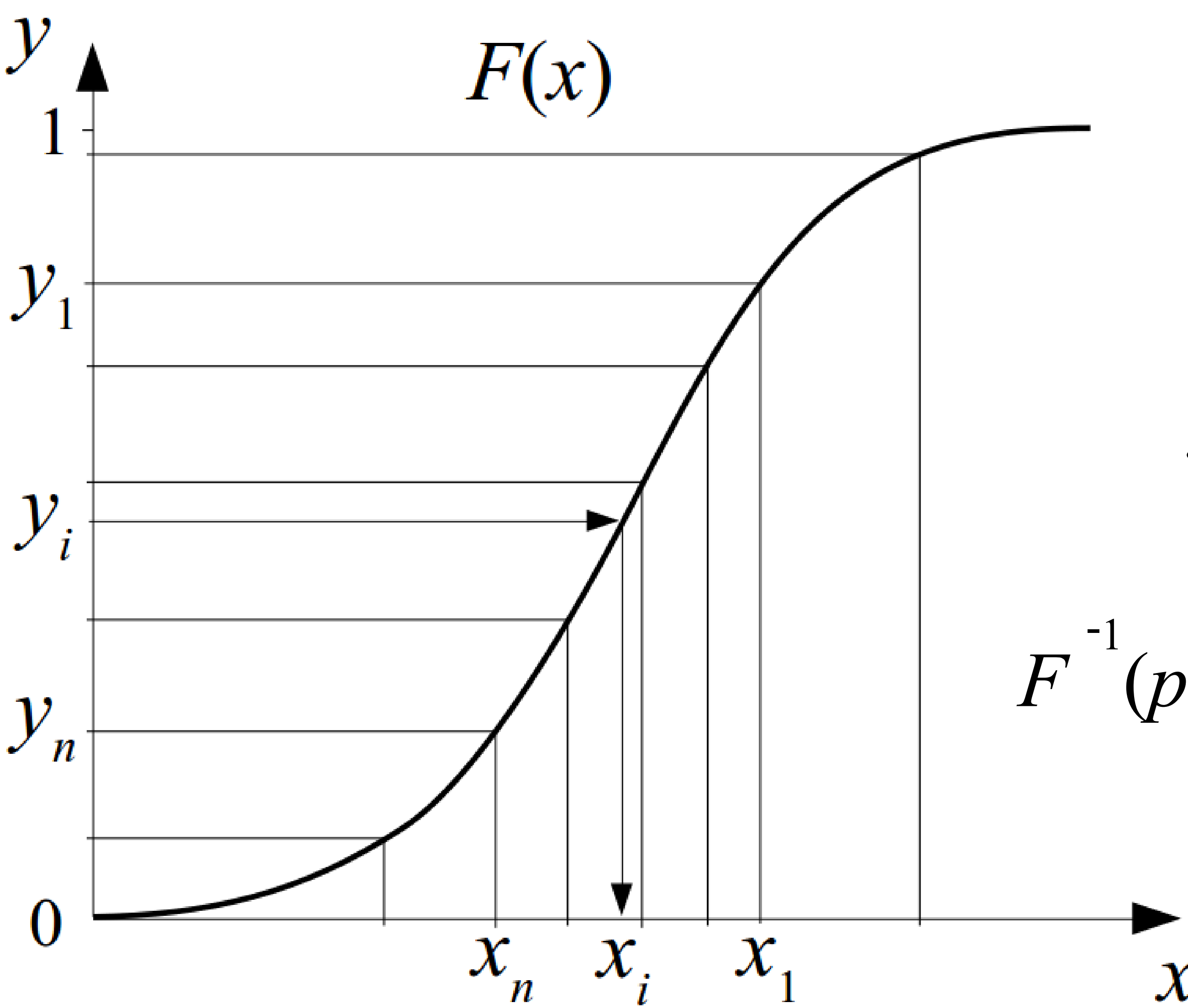
Lépések

1) Meghatározzuk a lehetséges értékek tartományát
2) A tartományon megadott eloszlás szerint véletlen mintákat állítunk elő
3) A mintákon valamilyen előírt számítást végzünk
4) Összesítjük a kapott eredményeket

Mintavétel tetszőleges valószínűségeloszlásból

- Rendelkezünk a (0, 1) intervallumban egyenletes eloszlású véletlenszámokat szolgáltató eljárással
- Az $F(x)$ eloszlásfüggvényű valószínűségeloszlásból szeretnénk n elemű mintát venni
- Ismerjük az $F(x)$ eloszlásfüggvény inverzét, az $F^{-1}(y) = F^{-1}(p)$ kvantilis függvényt
- Előállítunk n db y_i egyenletes eloszlású véletlenszámot $(i = 1,\ldots,n)$
- Az $x_i = F^{-1}(y_i)$ a keresett n elemű minta

Mintavétel tetszőleges valószínűségeloszlásból



$$y_i = F\left(x_i\right)$$

$$x_i = F^{-1}\left(y_i\right)$$

$F^{-1}(p)$ kvantilis függvény

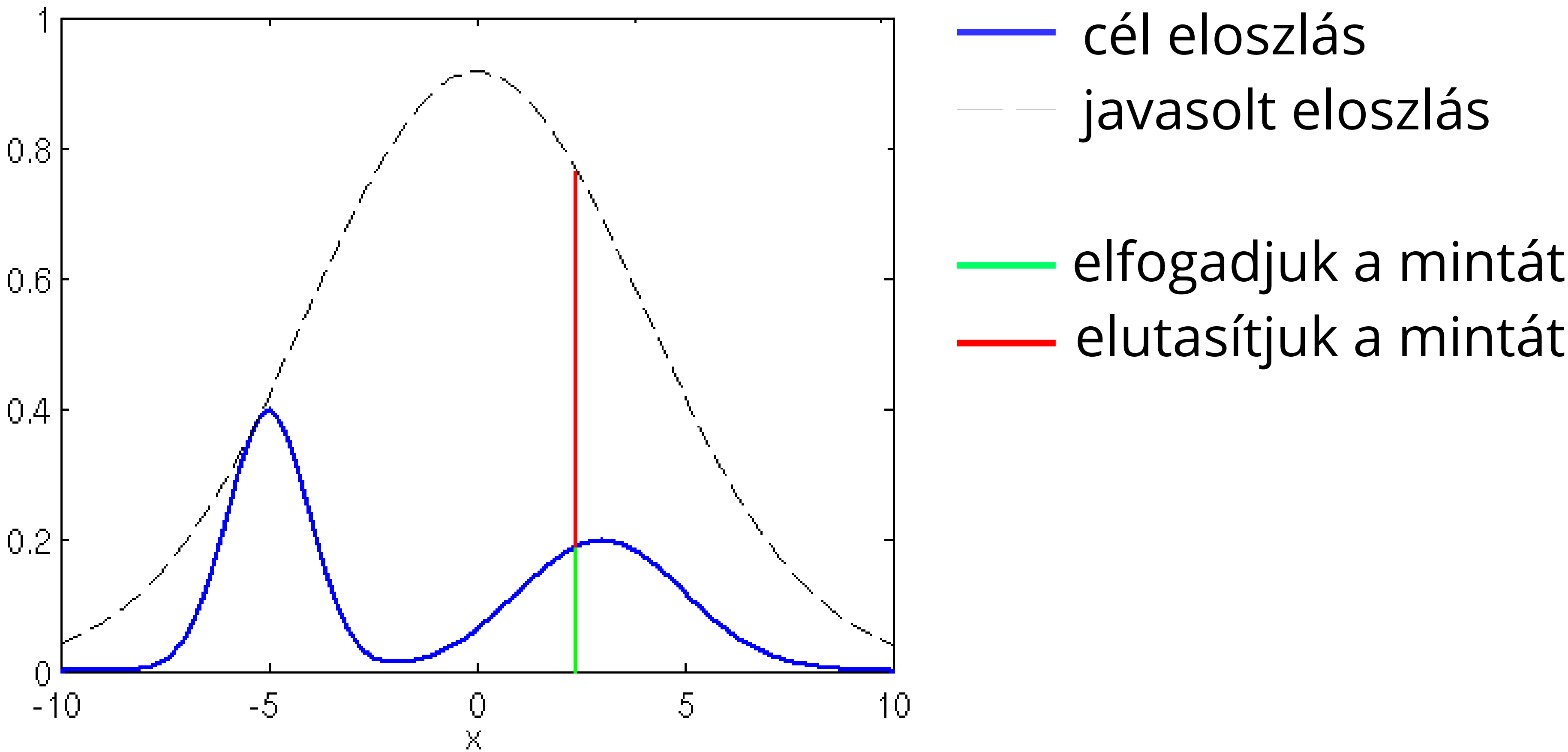
Analitikusan megadható kvantilis függvények

Density $f(x)$	$F(x)$	$X=F^{-1}(U)$	Simplified form
Exponential(λ) $\lambda e^{-\lambda x}$, $x \geq 0$	$1-e^{-\lambda x}$	$-\frac{1}{\lambda}\log(1-U)$	$-\frac{1}{\lambda}\log(U)$
Cauchy(σ) $\frac{\sigma}{\pi(x^2+\sigma^2)}$	$\frac{1}{2}+\frac{1}{\pi}\arctan(\frac{x}{\sigma})$	$\sigma\tan(\pi(U-\frac{1}{2}))$	$\sigma\tan(\pi U)$
Rayleigh(σ) $\frac{x}{\sigma}e^{-\frac{x^2}{2\sigma^2}}$, $x \geq 0$	$1-e^{-\frac{x^2}{2\sigma^2}}$	$\sigma\sqrt{-\log(1-U)}$	$\sigma\sqrt{-\log(U)}$
Triangular on$(0,a)$ $\frac{2}{a}(1-\frac{x}{a})$, $0 \leq x \leq a$	$\frac{2}{a}(x-\frac{x^2}{2a})$	$a(1-\sqrt{1-U})$	$a(1-\sqrt{U})$
Tail of Rayleigh $xe^{\frac{a^2-x^2}{2}}$, $x \geq a > 0$	$1-e^{\frac{a^2-x^2}{2}}$	$\sqrt{a^2-2\log(1-U)}$	$\sqrt{a^2-2\log U}$
Pareto(a,b) $\frac{ab^a}{x^{a+1}}$, $x \geq b > 0$	$1-(\frac{b}{x})^a$	$\frac{b}{(1-U)^{1/a}}$	$\frac{b}{U^{1/a}}$

Devroye (1986)

Elutasító mintavétel

- javasolt eloszlából vett minta elfogadása/elutasítása



Véletlenszám generátorok

- fontos szempontok:
  - statisztikai minőség
  - előrejelezhetőség
  - reprodukálható eredmények
  - ismétlési periódus
  - futásidő
  - tárigény
  - kódméret, bonyolultság
  - tetszőleges dimenzióban egyenletes eloszlás

	Statistical Quality	Prediction Difficulty	Reproducible Results	Multiple Streams	Period	Useful Features	Time Perfomance	Space Usage	Code Size & Complexity	k-Dimensional Equidistribution
PCG Family	Excellent	Challenging	Yes	Yes (e.g. 2^{63})	Arbitrary	Jump ahead, Distance	Very fast	Very compact	Very small	Arbitrary*
Mersenne Twister	Some Failures	Easy	Yes	No	Huge 2^{19937}	Jump ahead	Acceptable	Huge (2 KB)	Complex	623
Arc4Random	Some Issues	Secure	Not Always	No	Huge 2^{1699}	No	Slow	Large (0.5 KB)	Complex	No
ChaCha20†	Good	Secure	Yes	Yes (2^{128})	2^{128}	Jump ahead, Distance	Fairly Slow	Plump (0.1 KB)	Complex	No
Minstd (LCG)	Many Issues	Trivial	Yes	No	Tiny $< 2^{32}$	Jump ahead, Distance	Acceptable	Very compact	Very small	No
LCG 64/32	Many Issues	Published Algorithms	Yes	Yes 2^{63}	Okay 2^{64}	Jump ahead, Distance	Very fast	Very compact	Very small	No
XorShift 32	Many Issues	Trivial	Yes	No	Small 2^{32}	Jump ahead	Fast	Very compact	Very small	No
XorShift 64	Many Issues	Trivial	Yes	No	Okay 2^{64}	Jump ahead	Fast	Very compact	Very small	No
RanQ	Some Issues	Trivial	Yes	No	Okay 2^{64}	Jump ahead	Fast	Very compact	Very small	No
XorShift* 64/32	Excellent	Unknown?	Yes	No	Okay 2^{64}	Jump ahead	Fast	Very compact	Very small	No

http://www.pcg-random.org/

Eloszlásfüggvények terjedése nem lineáris rendszeren keresztül

- Ismert a vizsgált rendszer modellje: $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$
- Bemeneti x_i ($i = 1 ,\ldots, m$) mennyiségek (együttes) vsz. eloszlás függvényei is ismertek: $F_i(x_i)$ ($F(\boldsymbol{x})$)
- Előállítunk az ismert $F_i(x_i)$ eloszlások ($F(\boldsymbol{x})$) alapján véletlenszerűen egy lehetséges $\boldsymbol{x}$ bemenetet és meghatározzuk a $\boldsymbol{g}$ függvény alapján az y_k ($k = 1 \ \ldots, n$) kimeneteket
- Monte Carlo: Az eljárást sokszor (N-szer) megismételve az y_k kimenetek mindegyikéből N elemű mintát kapunk
- A kimeneti mennyiségek tapasztalati eloszlásait meghatározzuk
- Az eloszlások segítségével az y_k kimenetek legjellemzőbb értékeit és bizonytalanságait kiszámítjuk

Monte Carlo módszerek alkalmazása a mérési bizonytalanság meghatározására

- A GUM Supplement 1: „Propagation of distributions using a Monte Carlo method” 5.1.1

a) modell felállítása

1. a mérendő mennyiség, Y definiálása
2. az $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)^T$ bemeneti mennyiségek meghatározása, amiktől Y függ
3. az $\boldsymbol{X}$-et és Y-t összekapcsoló modell megalkotása
4. a meglevő ismeretek alapján X_i-k eloszlásfüggvényeinek (Gauss, egyenletes, háromszög, stb.) megadása, illetve nem független X_i-k együttes eloszlásfüggvényinek a megadása

b) terjedés

c) összefoglaló statisztikák készítése Y eloszlásfüggvényéről

Monte Carlo módszerek alkalmazása a mérési bizonytalanság meghatározására

- A GUM Supplement 1: „Propagation of distributions using a Monte Carlo method” 5.1.1

a) modell felállítása

b) terjedés

   1. az X_i-k eloszlásfüggvényeinek terjedése a modellen keresztül, hogy megkapjuk Y eloszlásfüggvényét

c) összefoglaló statisztikák készítése Y eloszlásfüggvényéről

   1. az Y várható értéke lesz a mennyiség y becslése (nem minden alkalmazás számára megfelelő)
   2. az Y standard bizonytalansága lesz az y mennyiség $u(y)$ standard bizonytalansága (Cauchy eloszlás esetén nem létezik)
   3. megbízhatósági tartomány készítése, mely Y-t előírt megbízhatósági valószínűséggel tartalmazza

Eloszlásfüggvények terjedése

- a mérendő mennyiség, Y eloszlásfüggvénye:

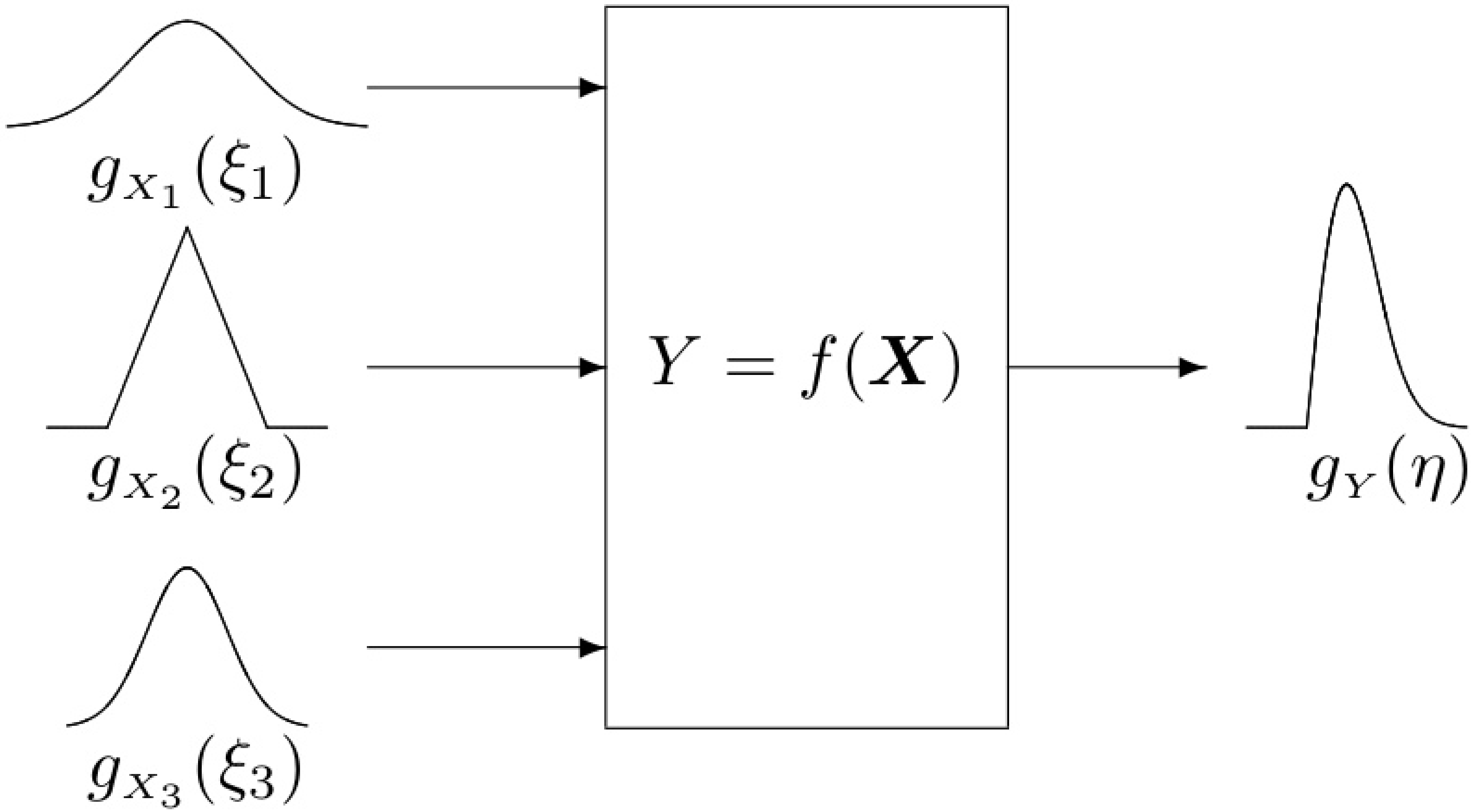
$$G_Y(\eta)=\int_{-\infty}^{\eta} g_Y(z)\,dz$$

- a sűrűségfüggvény formális definíciója (δ a Dirac delta):

$$g_Y(\eta)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_{\mathbf{X}}(\boldsymbol{\xi})\,\delta(\eta-f(\boldsymbol{\xi}))\,d\xi_N\cdots d\xi_1$$

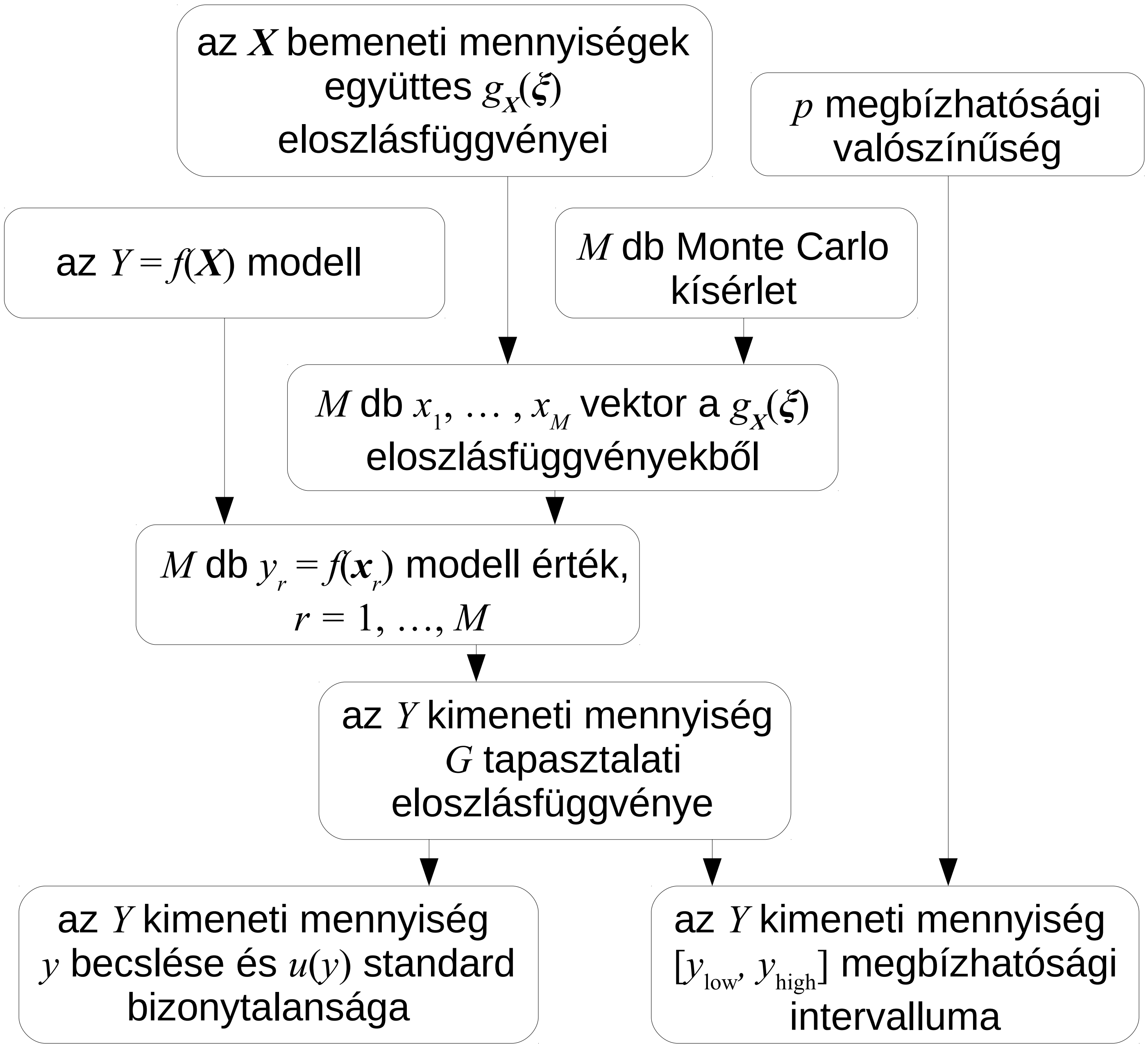
- analitikusan az integrálás nem végrehajtható, helyette Monte Carlo számításra (MCM) van szükség

Eloszlásfüggvények terjedése



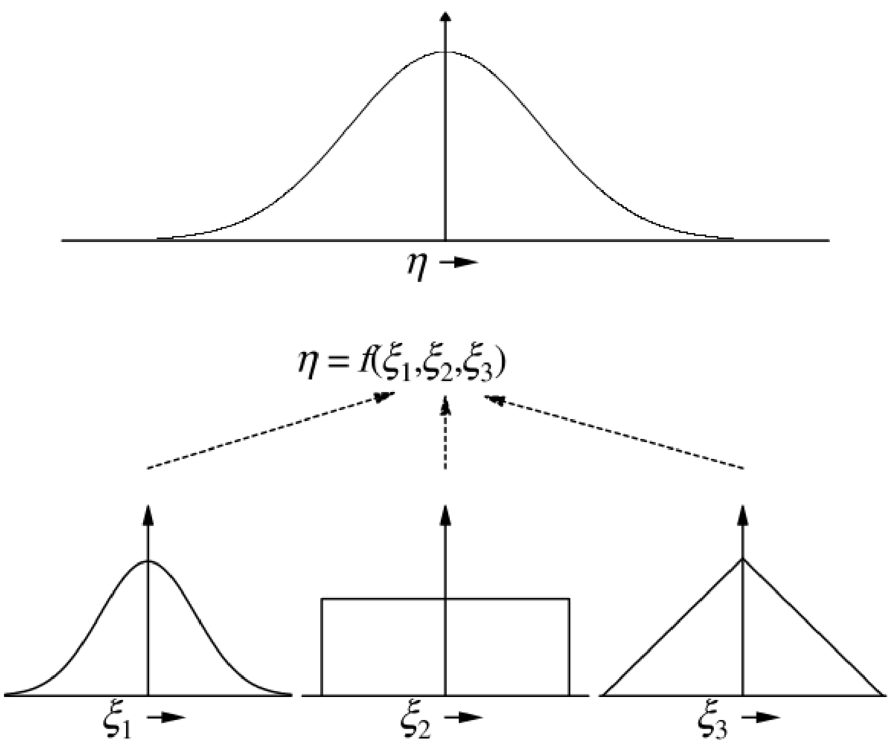
Az eloszlások terjedésének illusztrációja $N = 3$ független bemeneti mennyiségre

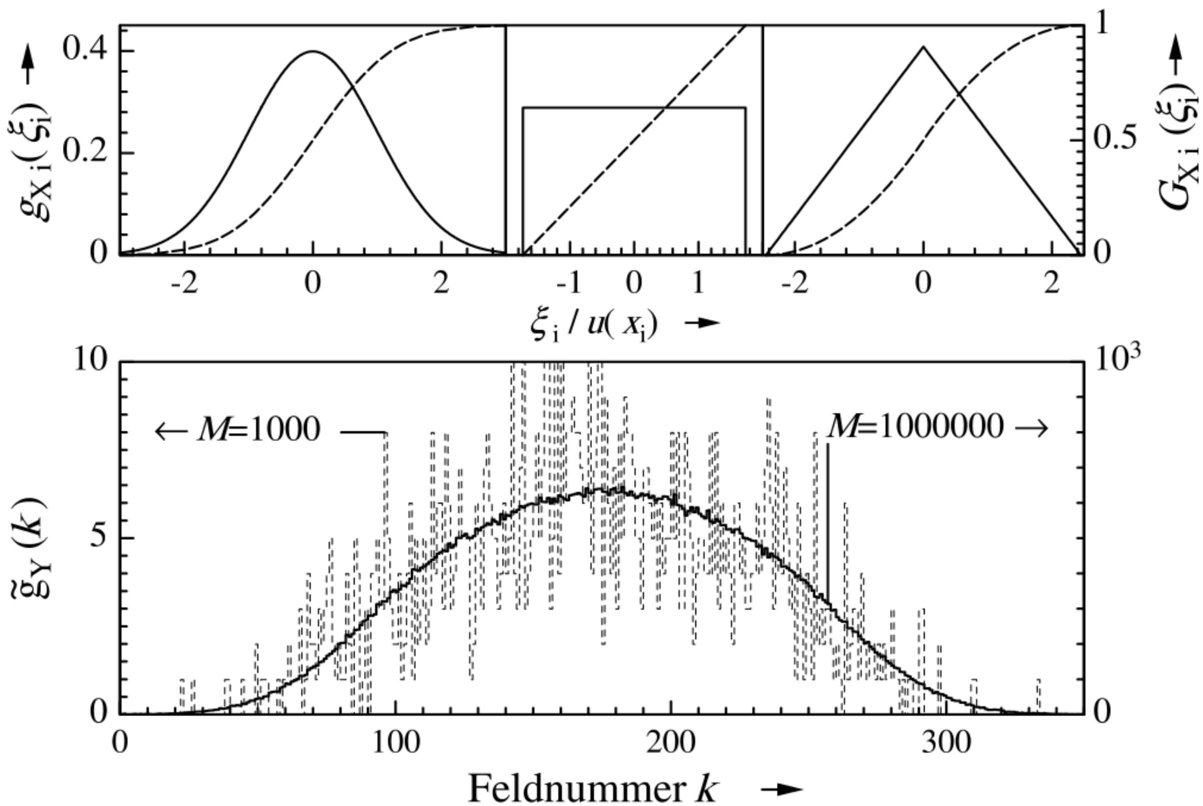
Monte Carlo számítás



MCM példa

$$\eta = \xi_1 + \xi_2 + \xi_3$$





(Siebert – Sommer, 2004)

Irodalom

- Steiner F. (1990): 5.1.3 pont
- Barabás B. (2010): Extrémérték-elmélet BME TTK
- Brodin E. (2006): A Non-mathematical Introduction to Statistics of Extremes NFT 3/2006
- Friederichs P. (2007): An Introduction to Extreme Value Theory University of Bonn, COPS Summer School