



## 2. előadás



# Áttekintés

- Statisztikai próbák, hipotézis vizsgálat
- Cramér-Rao határ
- Becslés statisztikai hatásfoka



# Statisztikai próba

- Olyan teszt eljárás, amely valamely statisztikai feltevés ellenőrzését teszi lehetővé az adatrendszer (minta) alapján

Paraméteres próbák

Nemparaméteres próbák



# Paraméteres próbák

- *Ismert eloszlástípus* esetén a mintából származó információ alapján döntünk az *eloszlás ismeretlen paramétereire* tett feltevés elfogadásáról
- Egymintás próba:  
egy adatrendszer van
- Kétmintás próba:  
két adatrendszer van
- Többmintás próba:  
varianciaanalízis

# Nemparaméteres próbák

- *Ismeretlen eloszlástípus* esetén alkalmazhatók
- Illeszkedésvizsgálat:

A mérési adatokból előállított tapasztalati sűrűségfüggvény egy adott elméleti eloszlásfüggvénnyel leírható-e?

- Függelenség vizsgálata:

Két külön mérési eljárásból származó adatsor függetlennek tekinthető-e?

- Homogenitás vizsgálata:

Két külön mérési eljárásból származó adatsor azonos eloszlású-e?

# Hipotézis vizsgálat

- Statisztikai hipotézis

A megfigyelt mennyiség eloszlásának a típusára vagy az eloszlásának a paramétereire tett feltevés

- Nullhipotézis ( $H_0$ ):

Az előzetes feltevést igaznak gondoljuk

- Ellenhipotézis ( $H_1$ ):

Bármilyen, a nullhipotézissel szemben álló feltevés

# Hipotézis vizsgálat logikája

- A megfigyelt adat(ok)/minta alapján döntünk a nullhipotézisről ( $H_0$ )

Vigyázat! **NEM** a nullhipotézis vagy az ellenhipotézis valószínűségét mondjuk meg az adatok (minta) alapján!

Helyesen: a megfigyelt **adatok** (minta) valószínűségét határozzuk meg a nullhipotézis vagy az ellenhipotézis fennállása esetén



# Statisztikai függvény, statisztika

- A statisztika olyan számítási utasítás, amely egyetlen adatot számít ki  $n$  db adat alapján
- A statisztikai próba feladata:  

Meg kell találni azt a statisztikai függvényt, amelynek eloszlását  $H_0$  fennállása esetén ismerjük
- Monte Carlo változatban:  

Meg kell határozni  $H_0$  fennállása esetén a statisztikai függvény tapasztalati eloszlását



# Egymintás u-próba

- Feladat: ismert skálaparaméterű ( $\sigma$ -jú) Gauss-eloszlásra annak megvizsgálása, hogy a helyparaméter  $T_0$ -e?

*példa:* egy tartógerenda hossza az előírt érték-e és ismert a mérőeszköz pontossága

- $H_0$ : a helyparaméter értéke  $T_0$

# Egymintás, paraméteres próba (u-próba)

- Ismert az  $x$  mennyiség eloszlása (Gauss típusú) és a mennyiség  $\sigma$  szórása
- A változóra vett mintában az átlag  $E(x)$ .
- Igaz-e, hogy az egész sokaság várható értéke, azaz az eloszlás helyparamétere  $T_0$ ?
- Nullhipotézis  $H_0: E(x) = T_0$
- Ellenhipotézis  $H_1: E(x) \neq T_0$

# Egymintás u-próba statisztikai függvénye

- Legyen az  $u$  változó az  $E(x)$  standardizáltja (egységnyi szórás, zérus átlag), és válasszuk ezt statisztikai függvénynek.
- Az  $u$  is Gauss-eloszlást követ:

$$u = \frac{\frac{1}{n} \sum_{i=1}^n x_i - T_0}{\sigma / \sqrt{n}}$$

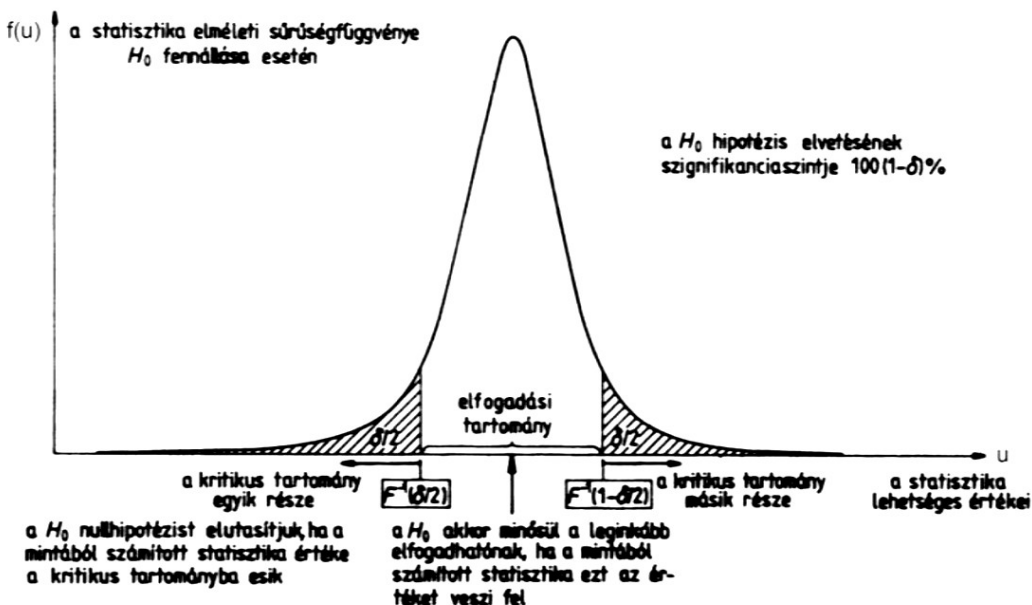
# Megbízhatósági intervallum

- A  $[-u_\delta, u_\delta]$  megbízhatósági intervallum kis  $\delta$  valószínűség esetén nagy valószínűséggel tartalmazza az  $\mu$  értékét
- Az  $1 - \delta$  a szignifikancia szint

$$P\left(-u_\delta \leq \frac{E(x) - T_0}{\sigma / \sqrt{n}} \leq u_\delta\right) = P\left(\left|\frac{E(x) - T_0}{\sigma / \sqrt{n}}\right| \leq u_\delta\right) = 1 - \delta$$

# Egymintás u-próba

- Ha a  $H_0$  nullhipotézis igaz, akkor az  $u$  nagy  $1 - \delta$  valószínűséggel esik a  $[-u_\delta, u_\delta]$  megbízhatósági intervallumba, azaz kis  $\delta$  valószínűséggel a kritikus tartományba



# Döntés

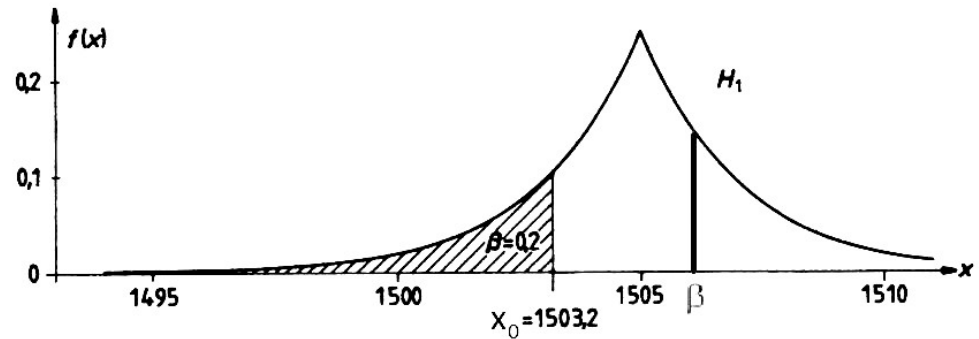
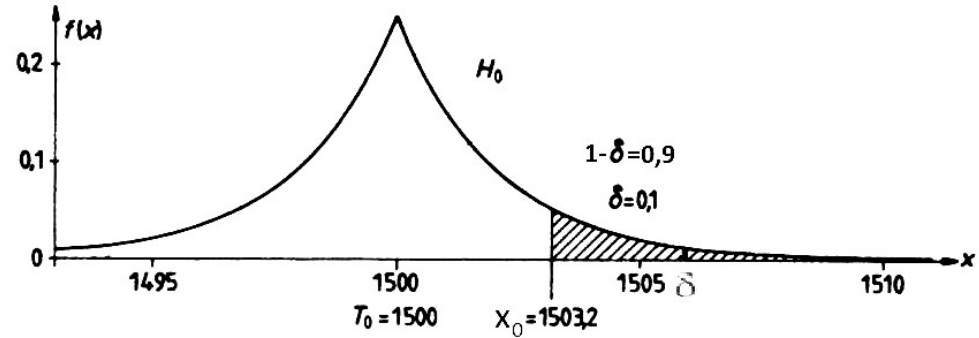
- Ha  $u$  a kritikus tartományban van, akkor a  $H_0$  nullhipotézist *elvetjük*, ha azonban  $u$  a megbízhatósági tartományon belül van, akkor *elfogadjuk*
- $H_0$  hibás elvetése: *elsőfajú hiba*
- $H_0$  hibás elfogadása: *másodfajú hiba*
- **Helytelen** azt mondani, hogy „a próba alapján a  $H_0$  nullhipotézist  $1 - \delta$  szignifikanciaszinten elfogadjuk”

Ez azt sejteti, hogy a nullhipotézis elfogadása nagy biztonsággal történt, holott épp ellenkezőleg! A másodfajú hiba valószínűsége ilyenkor igen nagy lehet.

- **Helyesen**: „ $x$  eltérése  $T_0$ -tól nem annyira lényeges, hogy a nullhipotézist  $1 - \delta$  szignifikanciaszinten elvethessük”

# Hibás döntések

- $H_0$  hibás elvetése a kritikus tartományban: *elsőfajú hiba*
- $H_0$  hibás elfogadása a konfidencia tartományban: *másodfajú hiba*
- $H_0$  elfogadása annál nagyobb kockázattal jár, minél nagyobb az  $1 - \delta$ .





# Cramér-Rao határ

- Az IC-függvény
  - mi a durva hibájú adatok torzító hatása?
  - mi egy mintaelem befolyása a becslésre?
- Becslések aszimptotikus szórása
  - mérések számának növelése mit eredményez?
- Minimális aszimptotikus szórás: a Cramér-Rao határ
  - létezik legkisebb szórású becslés?
- A Cramér-Rao határ különböző eloszlástípusokra

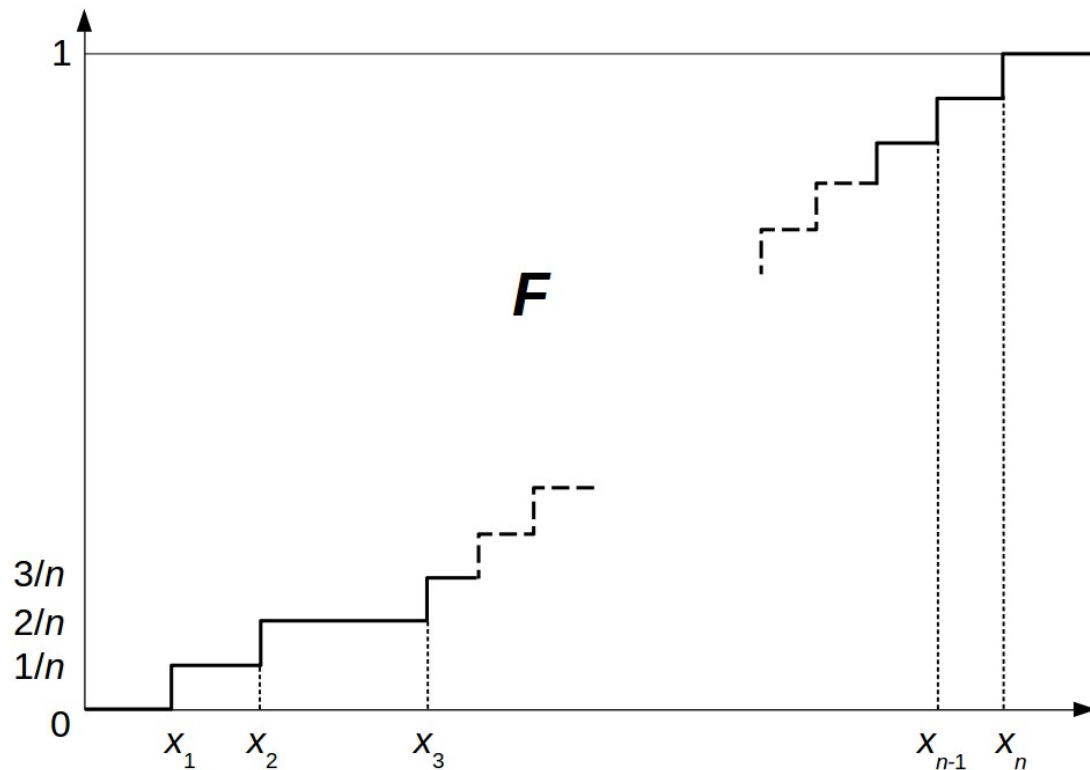


# Az $IC$ -függvény (hatásgörbe)

- Milyen mértékben módosítja egyetlen  $x$  adat a helyparaméter becslés  $T$  eredményét?
- a válasz függ:
  - a becslés algoritmusától (legyen ez is  $T$ )
  - az aktuális  $F$  eloszlástól
  - az adat  $x$  értékétől
- $IC(x, F, T)$  hatásgörbe (influence curve,  $IC$ -görbe,  $IC$ -függvény) adja meg a választ

# Diszkrét eloszlásfüggvény

- $n$  elemű  
sorbarendeztet  
minta:  $x_1, x_2, \dots, x_n$
- a minta értékeknél  
 $1/n$  ugrás

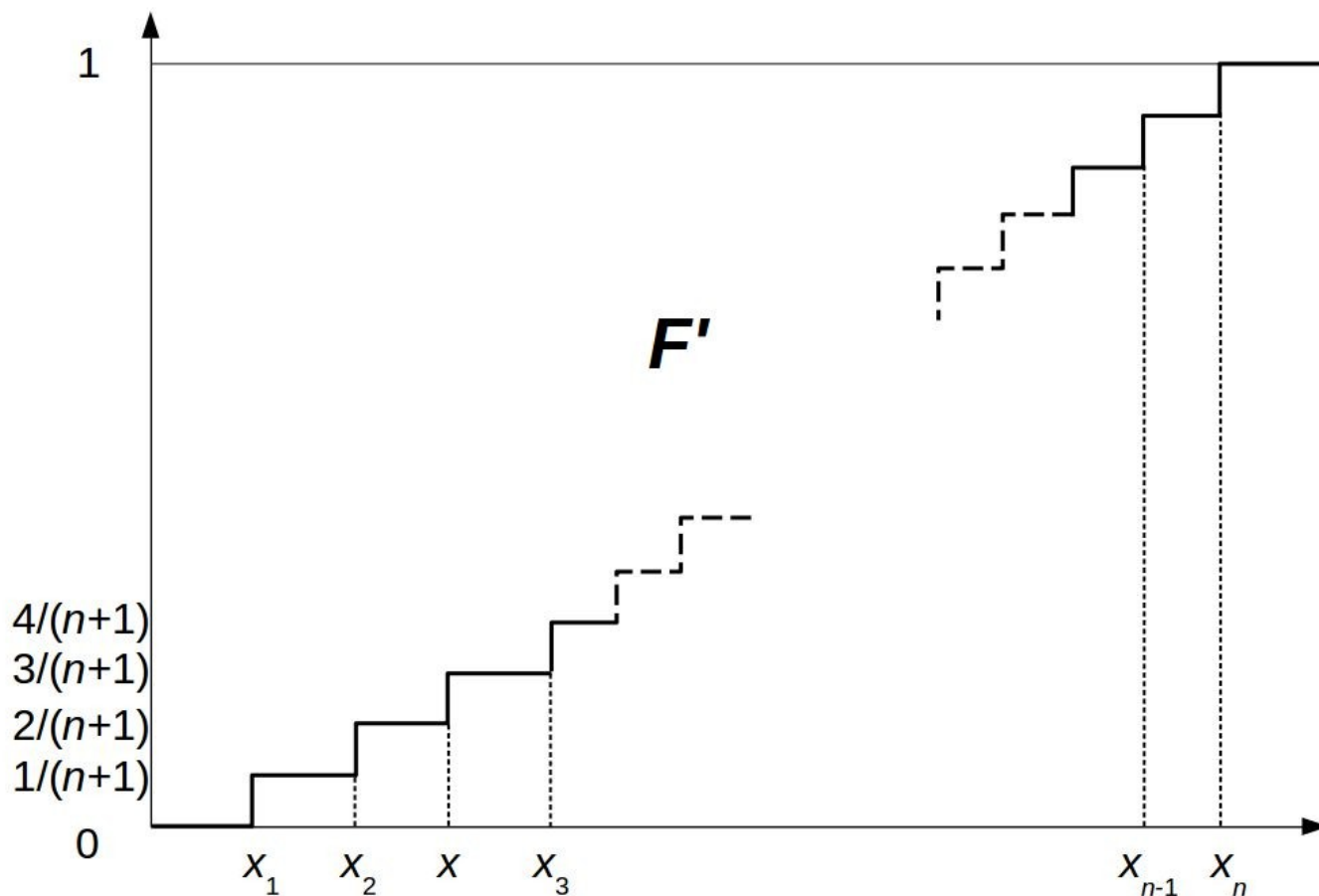


# Diszkrét eloszlásfüggvény

- egy új  $x$  értékkel bővített minta:

$$X_1, X_2, \dots, X, \dots, X_n$$

- a minta értékeknél  $1/(n+1)$  ugrás



# Megváltozott eloszlásfüggvény

- 1) eredeti  $F$  eloszlásfüggvényt  $n/(n+1)$ -el szorozzuk
- 2)  $x$  értékig 0-t, utána  $1/(n+1)$ -et hozzáadunk

$$F' = \frac{n}{n+1} F + \frac{1}{n+1} H(x)$$

- 3) legyen  $n$  nagy érték, ekkor  $t = 1/(n+1)$  kis érték lesz

$$F' = (1-t) F + t H(x)$$

$H(x)$  a Heaviside (egységugrás) függvény  $x$ -ben

# Az $IC$ -függvény definíciója

$$IC(x, F, T) = \lim_{t \rightarrow 0} \frac{T[(1-t) \cdot F + t \cdot H(x)] - T(F)}{t}$$

$H(x)$  az egységugrás (Heaviside) függvény

- Az  $IC$ -függvény megadja egyetlen,  $x$  értékű járulékos észlelésnek a  $T$  értékváltozásában megnyilvánuló hatását
- $T(F)$  az eredeti  $F$  eloszlás alapján becsült  $T$  paraméter
- $[(1-t) \cdot F + t \cdot H(x)]$  a megváltozott eloszlás
- $t$  a járulékos észlelés részaránya a többi adathoz képest

# A becslés értékének változása

- nagy  $n$  esetén közelítőleg  $\Delta T$  értékkel változik meg a  $T$  becslés értéke, ha az  $F$ -eloszlásfüggvényű eloszlásból származó  $n$  elemű mintánkhoz még egyetlen  $x$  értékű észlelést is figyelembe veszünk ( $t = 1/n$ ):

$$IC(x, F, T) = \frac{\Delta T}{1/n} \quad \Delta T = \frac{IC(x, F, T)}{n}$$

# Az $IC$ -görbe alkalmazása

- Közvetlen számszerű információt ad arról, hogy a *durva hibájú adatok milyen mértékben torzítják* az adott algoritmus szerint számított hely (vagy skála) paraméter becslés értékét (*rezisztencia*)
  - a számtani átlag  $IC$ -függvénye  $x$ -szel egyenlő (nem rezisztens)
- *a minta egy  $x$  értékű eleme milyen mértékben vesz részt a helyparaméter becslésében (robosztusság)*

# Helyparaméter becslés

- becsüljük súlyozott átlaggal

$$T = \frac{\sum_{i=1}^n \varphi(x_i) \cdot x_i}{\sum_{i=1}^n \varphi(x_i)}$$

- átrendezéssel a becslést jellemző  $\psi$  függvényt kapjuk

$$\sum_{i=1}^n \varphi(x_i) \cdot (x_i - T) = \sum_{i=1}^n \psi(x_i - T) = 0$$



# Módosult sűrűségfüggvény

- $[(1 - t) \cdot F(x) + t \cdot H(x - x_0)]$  a módosult eloszlásfüggvény
- a módosult  $g(x)$  sűrűségfüggvény az eloszlásfüggvény  $x$  szerinti deriváltja (1-re normált):

$$g(x) = (1 - t) \cdot f(x) + t \cdot \delta(x - x_0)$$

a Dirac-féle  $\delta(x)$  integrálja:

$$\int_{-\infty}^{\infty} f(x) \cdot \delta(x - x_0) dx = f(x_0)$$

# IC-függvény számítása a $\psi$ -függvényből

- becslés integrál megfelelője  $f(x)$  sűrűségfüggvényre

$$\int_{-\infty}^{\infty} \psi(x-T) f(x) dx = 0$$

- $f(x)$  helyére beírva a módosult  $g(x)$  sűrűségfüggvényt,

$$g(x) = (1-t) \cdot f(x) + t \cdot \delta(x-x_0)$$

kapjuk az implicit  $h(T, t) = 0$  függvényt

$$(1-t) \cdot \int_{-\infty}^{\infty} \psi(x-T) \cdot f(x) dx + t \cdot \psi(x_0-T) = 0$$

# IC-függvény számítása a $\psi$ -függvényből

- Az IC-függvény a  $dT/dt$ -nek  $t \rightarrow 0$  helyen felvett értéke, amit  $t$  szerinti differenciálás után átrendezéssel kaphatunk meg

$$\frac{dT}{dt} = \frac{\psi(x_0 - T) - \int_{-\infty}^{\infty} \psi(x - t) \cdot f(x) dx}{(1 - t) \cdot \int_{-\infty}^{\infty} \psi'(x - T) \cdot f(x) dx + t \cdot \psi'(x_0 - T)}$$

- A  $t \rightarrow 0$  határátmenetben megkapjuk az IC-függvényt (hatásgörbét)

# IC-függvény számítása a $\psi$ -függvényből

- Az  $IC$ -függvény a helyparaméter becslésére,  $S = 1$  skálaparaméterrel,  $T = 0$  helyparaméterrel

$$IC(x, F, T) = \frac{\psi(x)}{\int_{-\infty}^{\infty} \psi'(x) \cdot f(x) dx}$$

# A leggyakoribb érték számításának $\psi$ -függvénye

- Az általánosított leggyakoribb érték integrálformulája:

$$M_k = \frac{\int_{-\infty}^{\infty} \frac{x}{(k\varepsilon)^2 + (x - M_k)^2} f(x) dx}{\int_{-\infty}^{\infty} \frac{1}{(k\varepsilon)^2 + (x - M_k)^2} f(x) dx}$$

# A leggyakoribb érték számításának $\psi$ -függvénye

- Az integrálformulából következik az általánosított leggyakoribb érték számításának  $\psi$ -függvénye:

$$\psi(x) = \frac{x}{1+x^2}$$

# A leggyakoribb érték számításának $IC$ -függvénye

- Az általánosított leggyakoribb érték  $IC$ -függvénye:

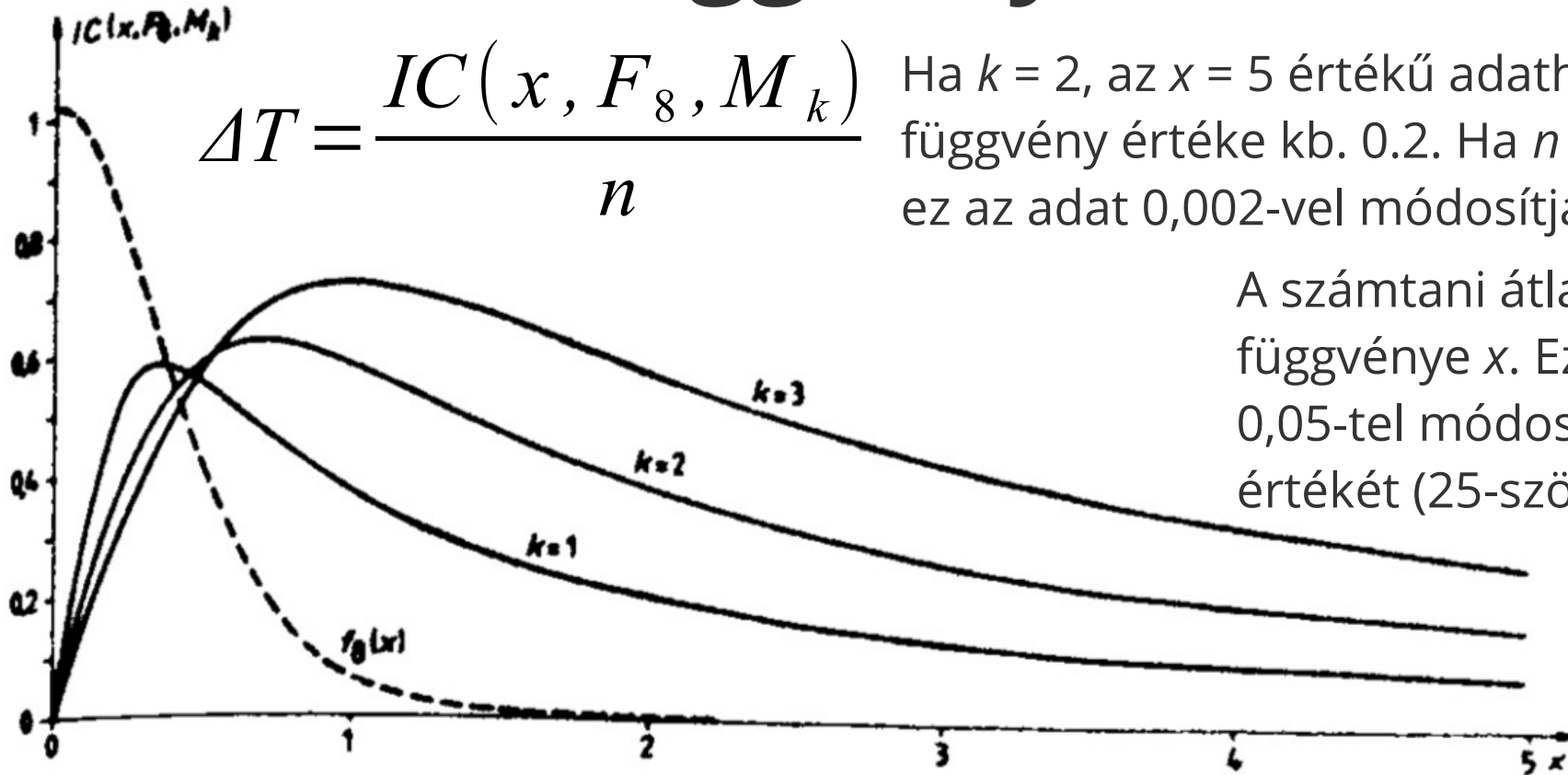
$$IC(x, F, M_k) = \frac{1}{\int_{-\infty}^{\infty} \frac{(k\varepsilon)^2 - y^2}{[(k\varepsilon)^2 + y^2]^2} f(y) dy} \cdot \frac{x}{(k\varepsilon)^2 + x^2}$$

# Az általánosított leggyakoribb érték $IC$ -függvénye

$$\Delta T = \frac{IC(x, F_8, M_k)}{n}$$

Ha  $k = 2$ , az  $x = 5$  értékű adathoz az  $IC$ -függvény értéke kb. 0.2. Ha  $n = 100$ , akkor ez az adat 0,002-vel módosítja a  $T$  értékét

A számtani átlag  $IC$ -függvénye  $x$ . Ezért ez az adat 0,05-tel módosítja az  $E$  átlag értékét (25-szörös hatás)

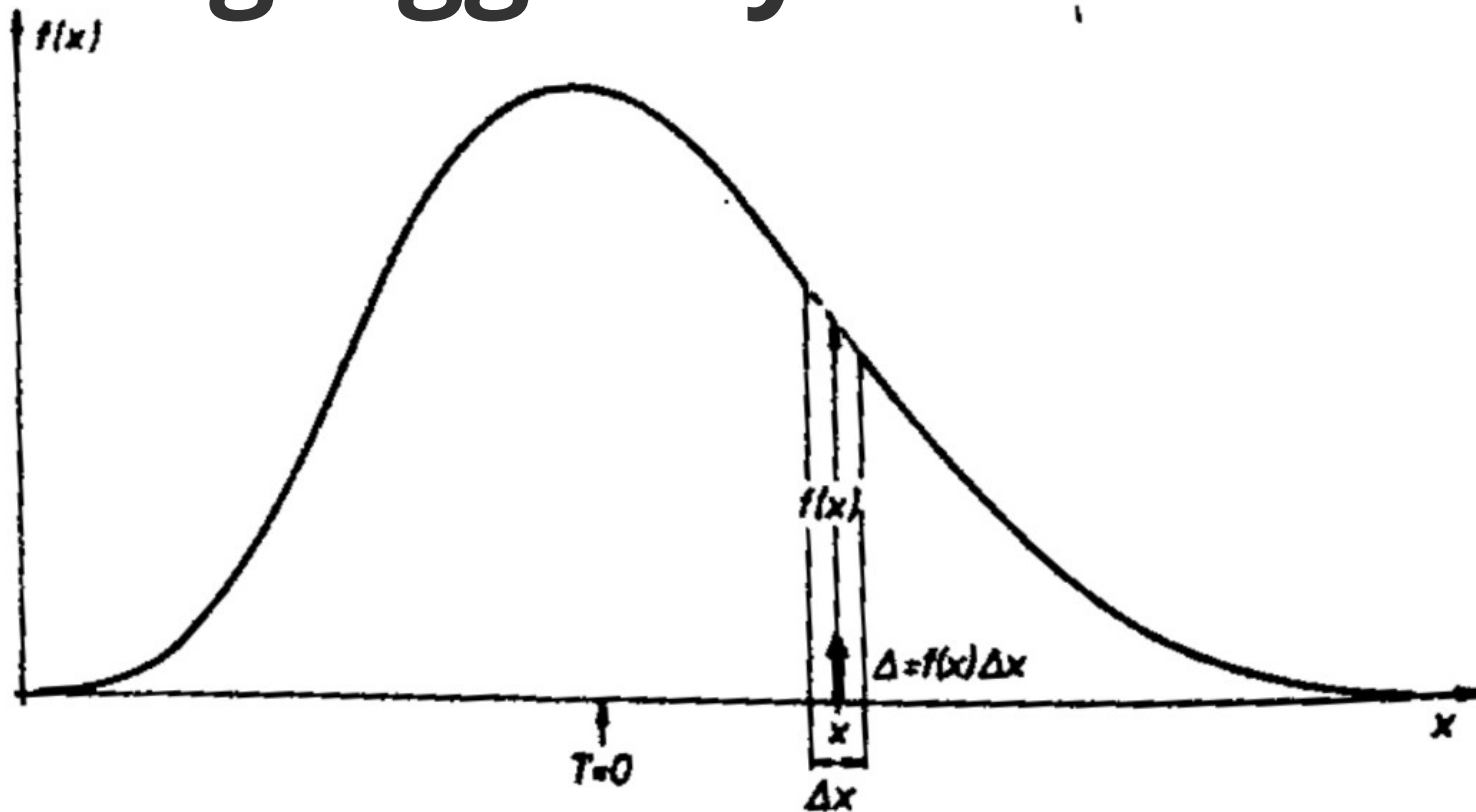




# Becslések aszimptotikus szórása

- a becslés eloszlásának véges a  $\sigma$  szórása (ez *becslésekre* igen gyakran teljesül)
- aszimptotikus szórásnégyzet:  $A^2 = \sigma^2 \cdot n$
- adott egy folytonos  $f(x)$  sűrűségfüggvénnyel jellemzett eloszlás
- valamely  $x$  érték körül nagyon szűk  $\Delta x$  intervallumba a vsz. változó értéke  $f(x) \cdot \Delta x$  valószínűséggel esik

# Sűrűségfüggvény



- az intervallumba esés valószínűsége  $\Delta = f(x) \cdot \Delta x$

# $\Delta x$ intervallum hatása

- Az  $IC$ -görbe definíciójából következik, hogy a  $\Delta$  valószínűséggel fellépő  $x$  jó közelítéssel  $\Delta T$  értékkel járul hozzá a  $T$  értékének a kialakításához:

$$\Delta T = IC(x, F, T) \cdot \Delta$$

# Áttérés $n$ elemű mintára

- A valószínűségi változó egész értéktartományát felosztjuk  $n$  db  $\Delta$  valószínűségű intervallumra
- Mivel  $\Delta = 1/n$  kicsi,  $n \rightarrow \infty$

Tetszőleges mintából szerkesztett empirikus eloszlásfüggvény nagyon közel lesz  $F(x)$ -hez (matematikai statisztika alaptétele)

az intervallumok  $x_i$  helyei  $n$  elemű mintát adnak

# A $j$ -edik mintaelem okozta eltérés szórása

- A  $\Delta T_j$  eltérés empirikus szórásnégyzete, mivel az adott mintavételkor  $j$ -edik elemként az  $x_1, \dots, x_n$  értékeket azonos valószínűséggel kaphatjuk meg:

$$D_{n;j}^2 = \frac{1}{n} \sum_{i=1}^n \left( \frac{IC(x_i)}{n} \right)^2 = \frac{1}{n^3} \sum_{i=1}^n [IC(x_i)]^2$$

# Az összes mintaelem okozta eltérés szórása

- Az  $n$  db mintaelem  $\Delta T_j$  hatásai összegződnek, az azonos eloszlású valószínűségi változók összegének szórására vonatkozó tétel szerint:

$$D_n^2 = n \cdot D_{n;j}^2 = \frac{1}{n} \sum_{i=1}^n [IC(x_i)]^2 = \frac{1}{n} \sum_{i=1}^n [IC(x_i)]^2 \cdot f(x) \cdot \Delta x$$

# Az aszimptotikus szórásnégyzet

- Az összefüggés integrál alakba átírva

$$D_n^2 = \frac{1}{n} \int_{-\infty}^{\infty} [IC(x)]^2 f(x) dx$$

- Az aszimptotikus szórásnégyzet,  $A^2$

$$A^2 = \int_{-\infty}^{\infty} [IC(x)]^2 f(x) dx$$

(mivel általánosan a definíció  $A^2 = \sigma^2 \cdot n$ )

# A helyparaméter-bebecslés aszimptotikus szórása

- A helyparaméter-bebecslés korábban kapott  $IC$ -függvényét beírva


$$A^2 = \frac{\int_{-\infty}^{\infty} \psi^2(x) f(x) dx}{\left[ \int_{-\infty}^{\infty} \psi'(x) f(x) dx \right]^2}$$

(a szórás  $\sigma = A/\sqrt{n}$ )



# A Cramér-Rao határ

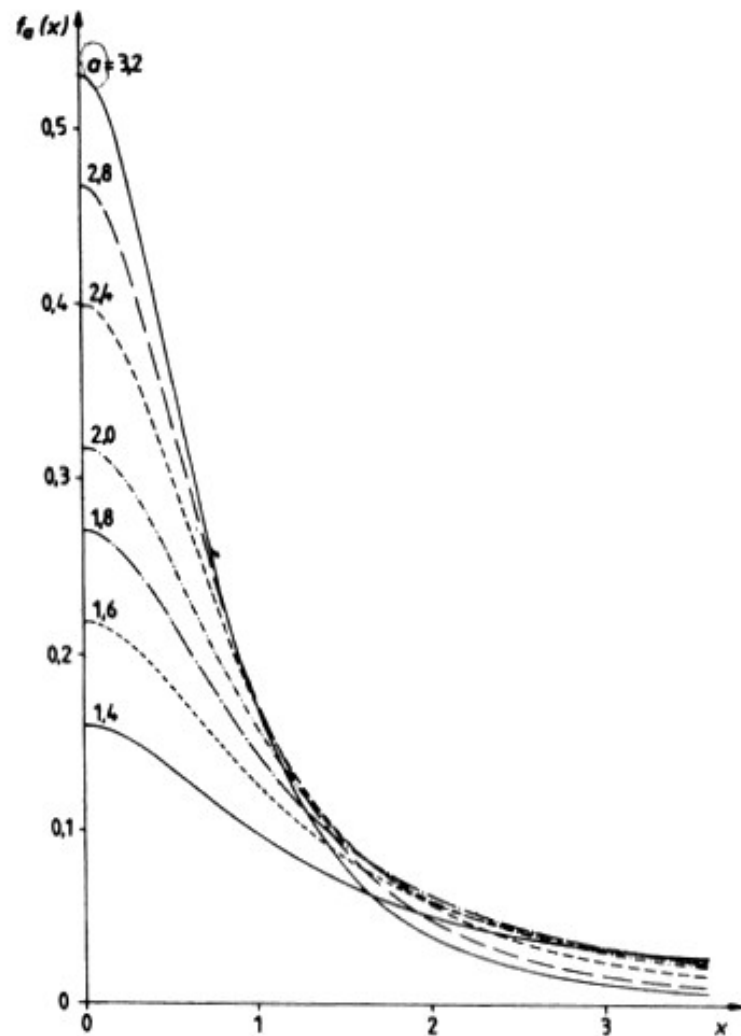
- A  $\psi(x)$  -szel jellemzett becslés aszimptotikus szórásnégyzete mindig nagyobb a Cramér-Rao határnál (minimális aszimptotikus szórásnégyzetnél):

$$\frac{\int_{-\infty}^{\infty} \psi^2(x) f(x) dx}{\left[ \int_{-\infty}^{\infty} \psi'(x) f(x) dx \right]^2} \geq \frac{1}{\int_{-\infty}^{\infty} \left( \frac{f'(x)}{f(x)} \right)^2 \cdot f(x) dx}$$


(a nevezőben az ún. Fisher-információ van;  
az összefüggés levezetését lásd a Steiner könyvben)

# Az $f_a(x)$ szupermodell

- $a$  : típusparaméter
- $a \rightarrow \infty$  : Gauss-eloszlás
- $a = 2$  : Cauchy-eloszlás
- $a = N + 1$  :  $N$  szabadságfokú Student-eloszlás



# Minimális aszimptotikus szórások

Eloszlástípus vagy supermodell	Cramér-Rao határ
általános $f(x)$	$\left[ \int_{-\infty}^{\infty} \left( \frac{f'(x)}{f(x)} \right)^2 \cdot f(x) dx \right]^{-1/2}$
Gauss $f_G(x)$	1
Laplace $f_L(x)$	1
Cauchy $f_C(x)$	$\sqrt{2}$
supermodell $f_a(x)$	$\sqrt{\frac{a+2}{a(a-1)}}$



# A becslés statisztikai hatásfoka

- Az abszolút és relatív hatásfok
- A becslés robusztussága

# A becslés *abszolút* hatásfoka

- Az  $e$  abszolút hatásfok a Cramér-Rao határhoz viszonyított aszimptotikus szórásnégyzet az adott becslési eljárásra (viszonyismként  $e \leq 1$  vagy százalékban kifejezve  $e \leq 100\%$  )

$$e = \frac{A_{\min}^2}{A^2}$$

- Az aktuális  $f(x)$  -hez tartozó optimális eljárással csak  $e$  -szer annyi adat kell ugyanakkora pontossághoz

# A hatások jelentősége

- „A hatások ismerete döntő fontosságú gyakorlati munkánkban, hiszen egy 1-nél lényegesen kisebb  $e$  hatások gyakorlatilag azt jelenti, hogy *adataink*  $100(1 - e)\%$  -át *eltékoztuk*” (Steiner)

# Két becslés *relatív* hatásfoka

- Ha két becslés aszimptotikus szórásai  $A_1$ , illetve  $A_2$ , akkor a második becslésnek az elsőhöz viszonyított relatív hatásfoka

$$e_{\text{rel}} = \frac{A_1^2}{A_2^2}$$

- Ez azt adja meg nagy mintaelemszámokra, hogy a második becsléssel  $n_2$  adatszám esetén elért pontosságot mennyi adattal érnénk el az elsővel:  $n_1 = e_{\text{rel}} \cdot n_2$

# Becslések aszimptotikus szórása $T$ -re szimmetrikus $f(x)$ -eknél

becslési algoritmus	aszimptotikus szórás
medián	$A_{\text{med}} = \frac{1}{2 \cdot f(\text{med})}$
számtani átlag	$A_E = \sqrt{\int_{-\infty}^{\infty} (x - E)^2 f(x) dx}$
leggyakoribb érték $k = 1$	$A_M = \frac{\varepsilon}{\sqrt{n_1(\varepsilon)}}$
leggyakoribb érték $k = 2$	$A_M = 2\varepsilon \frac{\sqrt{n_1(2\varepsilon) - n_2(2\varepsilon)}}{2n_2(2\varepsilon) - n_1(3\varepsilon)}$
megjegyzés: $i = 1$ esetén a súlyok, $i = 2$ esetén a súlynégyzetek átlaga lesz az $n_i(k\varepsilon)$	$n_i(k\varepsilon) = \int_{-\infty}^{\infty} \left[ \frac{(k\varepsilon)^2}{(k\varepsilon)^2 + (x - M)^2} \right]^i f(x) dx$



# Adatszám-többlet

- Cramér-Rao határral kifejezve:

$$\text{adatszám-többlet} = 100 \cdot \left[ \frac{A^2}{A_{min}^2} - 1 \right] \%$$

- Az  $f_a(x)$  szupermodellre az aszimptotikus szórás

$$A_{min} = \frac{a+2}{a(a-1)}$$

# Adatszámítóblet

- Az  $f_a(x)$  szupermodellre, *medián* képzéssel:

$$\text{adatszámítóblet} = 100 \cdot \left[ \frac{\pi a (a-1) \Gamma^2\left(\frac{a-1}{2}\right)}{4 (a+2) \Gamma^2(a/2)} - 1 \right] \%$$

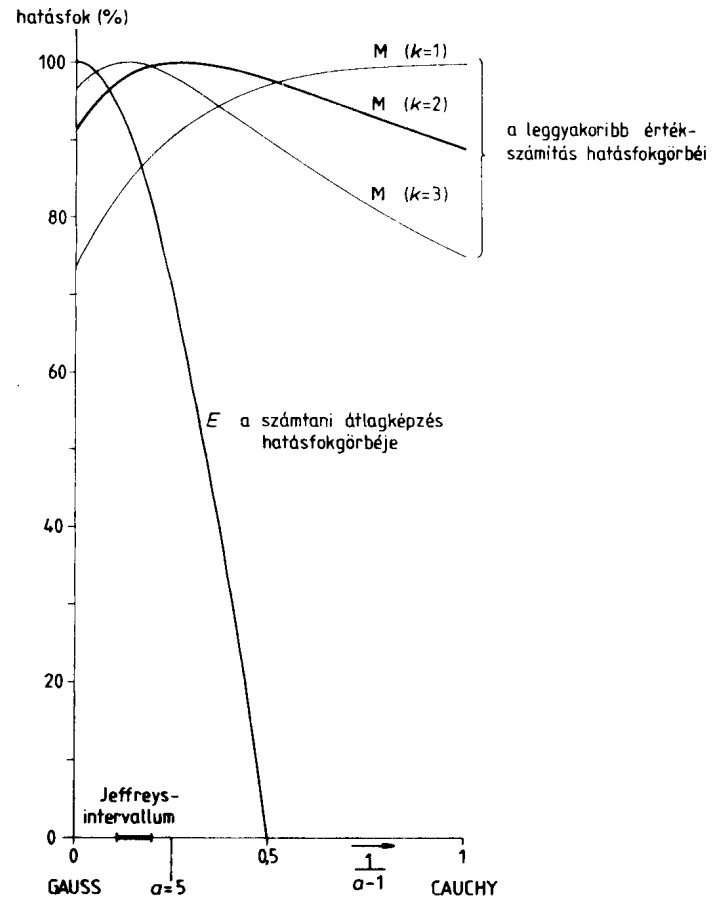
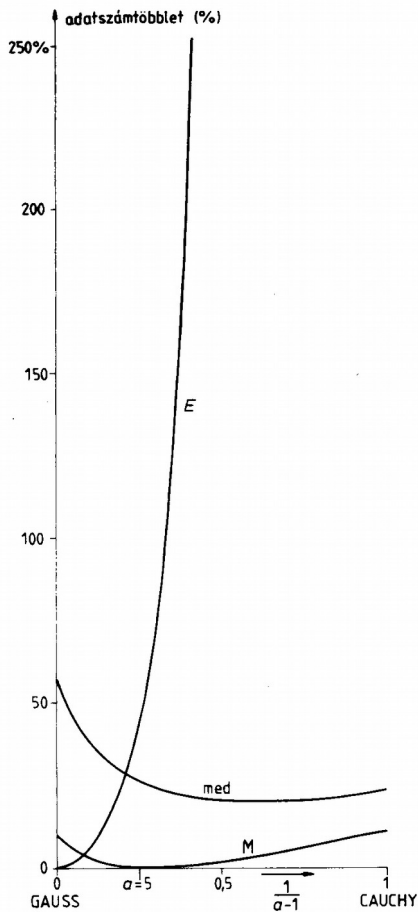
- Az  $f_a(x)$  szupermodellre, *átlag* képzéssel:

$$\text{adatszámítóblet} = 100 \cdot \left[ \frac{a(a-1)}{(a-3)(a+2)} - 1 \right] \%$$

# Leggyakoribb értékek aszimptotikus szórásai az $f_a(x)$ szupermodellre

	a	1/(a-1)	A (k = 2)
<b>Cauchy</b>	2	1	1,5000
	2,5	0,6667	1,1251
	3	0,5	0,9236
	4	0,3333	0,7080
	5	0,25	0,5917
	6	0,2	0,5173
	10	0,1	0,3694
	40	0,0256	0,1699
	100	0,0101	0,1057
<b>Gauss</b>	$\infty$	0	1,0466

# Adatszám-többlet és hatásfok különböző becslésekre az $f_a(x)$ supermodellre

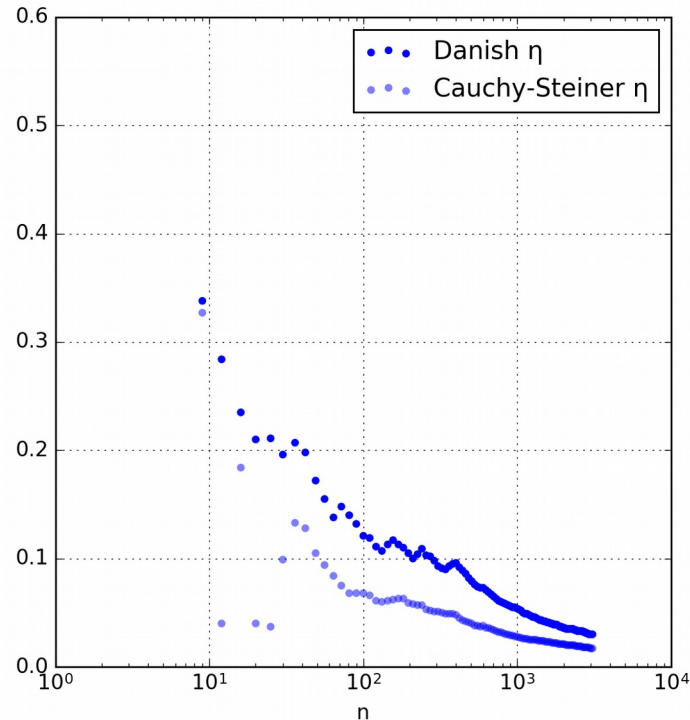
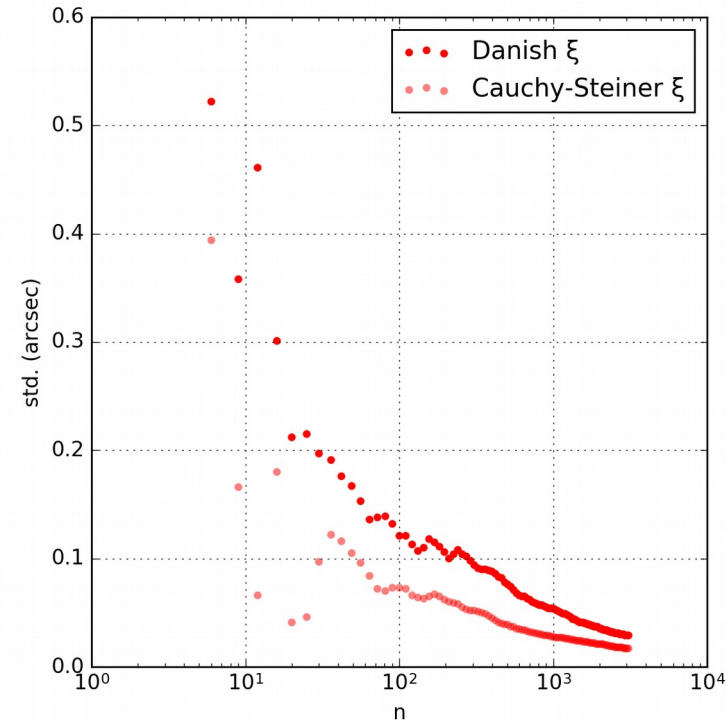


# Robusztusság

- Valamely *becslés* akkor nevezhető robusztusnak, ha az eloszlástípusok széles tartományán a gazdaságossága csak jelentéktelen mértékben csökken
- Egy statisztikai *algoritmus* akkor robusztus, ha az anyaeloszlásban bekövetkező kicsiny eltérés a becslések eloszlásában is csak kis eltérést eredményez

# Gyakorlati példa: QDaedalus feldolgozás eredményei

Pistahegy measurement No.47



A dán módszer és a leggyakoribb érték alapján történő függővonal-elhajlás becslések szórásainak összehasonlítása a mérések száma függvényében. A becslések hatásfoka jelentősen eltér egymástól (Tóth-Völgyesi, 2017)

# A két becslés relatív hatásfoka

- 3080 mérésből a dán módszer aszimptotikus szórása 1.64, a leggyakoribb érték szerintié pedig 0.94, . A relatív hatásfoka a két módszernek 3.0, ami azt jelenti, hogy *ugyanolyan pontosság* eléréséhez a dán módszerrel végzett feldolgozás esetén *háromszor annyi ideig* kellene mérni, mert a módszer háromszor kevésbé hatásos, mint a leggyakoribb érték módszere