# Valószínűségszámítás és Statisztika

9. előadás

2023. április 27.

# Rendezett minta

- A $\xi_1, \dots, \xi_n$ minta elemeit nagyság szerint sorbarendezve kapjuk az $\xi_1^{(n)} \leq \xi_2^{(n)} \leq \cdots \leq \xi_n^{(n)}$ – rendezett mintát.
- Ez $n$-dimenziós statisztika
- Mostantól: a $\xi_1, \dots, \xi_n$ minta elemei független, azonos eloszlásúak.
- Ha feltesszük, hogy a közös eloszlásuk abszolút folytonos, akkor felírható a rendezett minta $k$-adik elemének, $\xi_k^{(n)}$ -nek a sűrűségfüggvénye. (gyakorlat)
- Spec.: minimum, maximum.
- Def.: minta terjedelme: $\xi_n^{(n)} - \xi_1^{(n)}$

# Tapasztalati eloszlásfüggvény

- Tapasztalati eloszlás eloszlásfüggvénye: tapasztalati eloszlásfüggvény:

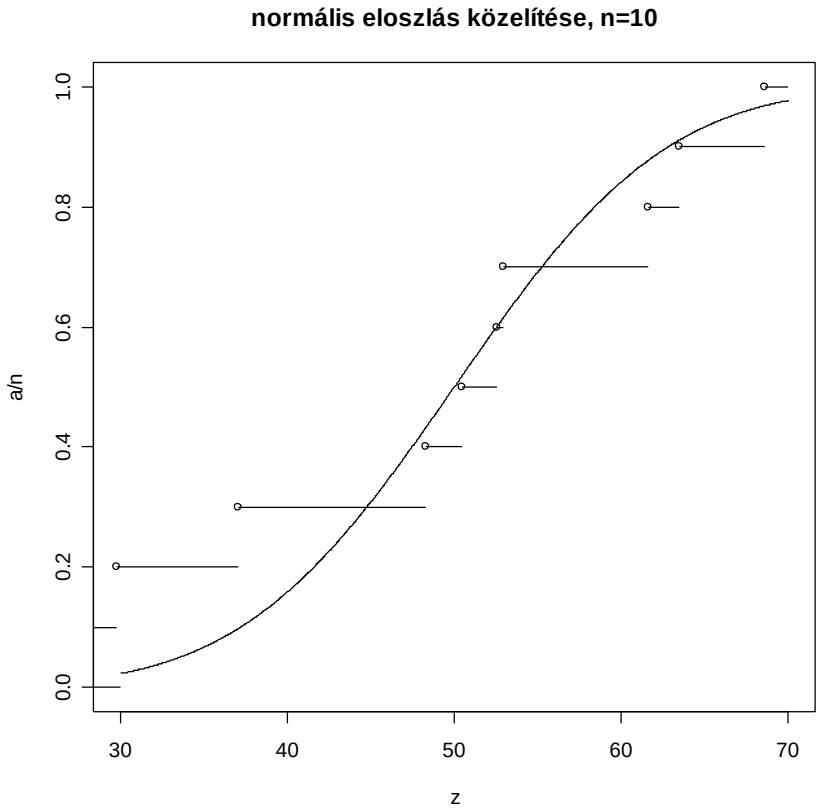
$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i < z\}$$

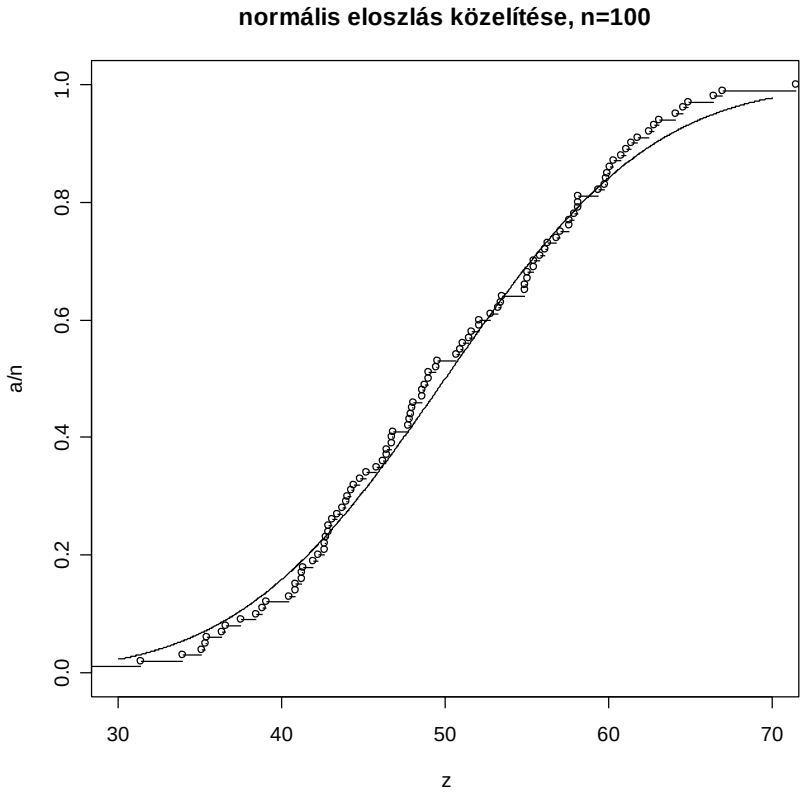
$$F_n(z) = \frac{k}{n}, \text{ha } \xi_k^{(n)} < z \leq \xi_{k+1}^{(n)},$$

$$\xi_0^{(n)} = -\infty, \xi_{n+1}^{(n)} = \infty$$

Mintaátlag éppen ennek az eloszlásnak a várható értéke.

# Példa


normális eloszlás közelítése, n=10
a/n
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
z


normális eloszlás közelítése, n=100
a/n
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
z

# Glivenko-Cantelli tétel (”statisztika alaptétele”)

Tétel: $\xi_1, \ldots, \xi_n$ független, azonos $F$ eloszlásfüggvényűek. Ekkor $\sup_z |F_n(z) - F(z)| \xrightarrow[n\to\infty]{} 0$ majdnem mindenütt (1 vszgel).

Biz.: Csak folytonos $F$ eloszlásfüggvényekre látjuk be. Ebből következik, hogy tetszőleges pozitív egész $N$−hez léteznek olyan valós $z_1, \ldots, z_N$ számok, hogy

$$F(z_0) = 0, F(z_1) = \frac{1}{N}, \ldots, F(z_i) = \frac{i}{N}, \ldots, F(z_{N-1}) = \frac{N-1}{N},$$

$$F(z_N) = 1,$$

$$z_0 = -\infty, z_N = \infty.$$

Ekkor, ha $z \in [z_k, z_{k+1})$, akkor

$$F_n(z) - F(z) \le F_n(z_{k+1}) - F(z_k)$$

$$= F_n(z_{k+1}) - F(z_{k+1}) + \frac{1}{N},$$

$$F_n(z) - F(z) \ge F_n(z_k) - F(z_{k+1}) = F_n(z_k) - F(z_k) - \frac{1}{N}.$$

Ebből következik, hogy

$$\sup_z |F_n(z) - F(z)| \le \max_{0 \le k \le N} |F_n(z_k) - F(z_k)| + \frac{1}{N}.$$

Tudjuk, hogy rögzített $x$ − re

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi\{\xi_i < x\},$$

ahol $\chi\{\xi_i < x\}$ független, azonos eloszlású indikátor valószínűségi változók, melyek várható értéke

$$= E(\chi\{\xi_i < x\}) = P(\xi_i < x) = F(x).$$

Így a nagy számok erős törvénye szerint

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i < x\} \xrightarrow[n\to\infty]{} E\chi\{\xi_i < x\} = F(x) \text{ mm.}$$

Legyen $A_{k,N} = \left\{\omega: \frac{1}{n}\sum_{i=1}^{n} \chi\{\xi_i(\omega) < z_k\} \xrightarrow[n\to\infty]{} F(z_k)\right\}$, ekkor

$P(A_{k,N}) = 1$ és $B_N = \left\{\omega: \max_{0\le k\le N} |F_n(z_k) - F(z_k)| \xrightarrow[n\to\infty]{} 0\right\} = \bigcap_{k=1}^{N-1} A_{k,N}$ .

$B_N$ − en $\limsup_{n\to\infty} |F_n(z) - F(z)| \le \frac{1}{N}$. Ebből következik,

hogy $\bigcap_{N=1}^{\infty} B_N$ −en $\limsup_{n\to\infty} |F_n(z) - F(z)| = 0$ .

1 valószínűségű események metszete is 1 valószínűségű.

Így $\bigcap_{N=1}^{\infty} B_N = \bigcap_{N=1}^{\infty} \bigcap_{k=1}^{N-1} A_{k,N}$ is 1 valószínűségű.

# Becsléselmélet

- A minta eloszlásának ismeretlen paraméterét közelítjük a minta függvényével
- Def.: becslőfüggvény: $\hat{\vartheta}: \mathcal{X} \to \Theta$
- Def.: becslés: $\hat{\vartheta}(\boldsymbol{\xi})$

- A becslések maguk is statisztikák. Szubjektíven: olyan statisztikák, amik jól közelítik az ismeretlen paramétert.

# Példa (Milyen valószínűséggel születik fiúgyermek?)

- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Ekkor $n = 2.644.757$, $\chi = \{0; 1\}^n$.
- Fiúk relatív gyakorisága így 0,5141.
- Mik ennek a becslésnek a tulajdonságai?

$$X_i = \begin{cases} 1, & i.\text{fiú} \\ 0, & i.\text{lány} \end{cases} \Rightarrow$$

$$P_p(X_i = 1) = p,$$

$$n = 2.644.757,$$

$$\hat{p} = \hat{p}(\mathbf{X}) = \frac{\sum_{i=1}^{n} X_i}{n} \Rightarrow$$

$$E_p\hat{p} = p,$$

$$\hat{p} \xrightarrow[n\to\infty]{} p \text{ mm.}$$

# Becslések tulajdonságai

- Def.: *Torzítatlanság:* A paraméter $\hat{\vartheta}(\boldsymbol{\xi})$ becslése torzítatlan, ha

$$E_{\vartheta}\left(\hat{\vartheta}(\boldsymbol{\xi})\right) = \vartheta, \forall \vartheta \in \Theta.$$

- *Konzisztencia*: $\hat{\vartheta}(\boldsymbol{\xi}) \to \vartheta$ sztochasztikusan $(n \to \infty)$ minden paraméterértékre.
- Példák:
  - Valószínűség becslése relatív gyakorisággal.
  - Glivenko tétele: a tapasztalati eloszlásfüggvény egyenletesen is konvergál az elméleti eloszlásfüggvényhez.
  - Várható érték becslése mintaátlaggal

# Konzisztencia

- Elégséges feltétel $E_\vartheta\left(\hat{\vartheta}_n(\boldsymbol{\xi})\right) \to \vartheta$ (aszimptotikus torzítatlanság) és $D_\vartheta{}^2\left(\hat{\vartheta}_n(\boldsymbol{\xi})\right) \to 0$ .

# Példák

- Poisson eloszlás paraméterére: mintaátlag
- Exponenciális eloszlás paraméterére:
  - mintaátlag reciproka: aszimptotikusan torzítatlan, konzisztens
  - $n \cdot \min(X_1, \ldots, X_n)$ torzítatlan, de nem konzisztens
- Szórásnégyzetre

# Becslések összehasonlítása

- Melyik a jobb becslés?

$$X_i = \begin{cases} 1, & i.\,\text{fiú} \\ 0, & i.\,\text{lány} \end{cases}, P_p(X_i = 1) = p,$$

$$\hat{p}_1 = \frac{\sum_{i=1}^{n} X_i}{n},$$
$$\hat{p}_2 = X_1, \text{vagy}$$

$$\hat{p}_3 = \frac{\sum_{i=1}^{[n/2]} X_i}{[n/2]}?$$

# Becslések összehasonlítása (hatásos becslések)

- Torzítatlan becslésekre: $T_1$ hatásosabb becslése $h(\Theta)$-nak a $T_2$-nél, ha

$$D_\theta^2\left(T_1(\underline{X})\right) \leq D_\theta^2\left(T_2(\underline{X})\right)$$

- teljesül minden $\Theta$ paraméterértékre.

Példa: a mintaátlag hatásosabb becslés a várható értékre minden

$$\sum_{i=1}^{n} c_i X_i$$

alakú becslésnél.

# Hatásos becslés

- Def.: A $T$ torzítatlan becslés hatásos, ha minden más torzítatlan becslésnél hatásosabb.
- Miért a torzítatlanokra? Furcsa példa: azonosan 0-val becsüljük az ismeretlen paramétert.
- Ezért érdemes a hatásos becsléseket csak a torzítatlan becslések között keresni.
- Átlagos négyzetes eltérés:

$$E_\theta\left(T(\underline{X}) - \theta\right)^2$$

# Hatásos becslés egyértelműsége

**Áll.:** Amennyiben $T_1$ és $T_2$ hatásos becslései $h(\theta)$-nak, akkor 1 valószínűséggel megegyeznek minden lehetséges paraméter esetén.

$E_\theta T_1 = E_\theta T_2 = h(\theta)$, továbbá $D_\theta T_1 = D_\theta T_2$. Ebből

$$\mathrm{D}_\theta^2(T_1) \le \mathrm{D}_\theta^2\left(\frac{T_1+T_2}{2}\right) = \frac{\mathrm{D}_\theta^2(T_1)+2\,\mathrm{cov}(T_1,T_2)+\mathrm{D}_\theta^2(T_2)}{4} = \frac{\mathrm{D}_\theta^2(T_1)+\mathrm{cov}(T_1,T_2)}{2} \Rightarrow$$

$$\mathrm{D}_\theta^2(T_1) \le \mathrm{cov}(T_1,T_2) = D_\theta T_1 \cdot D_\theta T_2 \cdot R(T_1,T_2) = \mathrm{D}_\theta^2(T_1) \cdot R(T_1,T_2) \le \mathrm{D}_\theta^2(T_1) \Rightarrow$$

$$\mathrm{D}_\theta^2(T_1) = \mathrm{D}_\theta^2(T_2) = \mathrm{cov}(T_1,T_2) \Rightarrow \mathrm{D}_\theta^2(T_1-T_2) = \mathrm{D}_\theta^2(T_1) - 2\,\mathrm{cov}(T_1,T_2) + \mathrm{D}_\theta^2(T_2) = 0.$$

Így $E_\theta(T_1 = T_2) = 1 \;\forall \theta \in \Theta$.

# Mit kell tudni a mintáról?

- Benzinkutas példa. Megfigyelések: 78, 89, 167, 90, 85.
- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.

| Kár-szám | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | > 7 | Össze-sen |
|---|---|---|---|---|---|---|---|---|---|---|
| Veze-tők száma | 129524 | 16267 | 1966 | 211 | 31 | 5 | 1 | 1 | 0 | 148006 |

# Mennyi információt hordoz a statisztika?

Példa: $\xi_1, \ldots, \xi_n$ független $N(m,1)$ minta. Ekkor

$$\overline{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n} \sim N(m, \frac{1}{n})$$ eloszlású (függ $m$-től!),

miközben

$$s^2 = \frac{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}{n}$$ eloszlása nem függ $m$-től!

# Elégséges statisztika

- Minden információt (ugyanannyit mint az eredeti minta) tartalmaz az ismeretlen paraméterre vonatkozóan.
- ”Elég” az ő értékét ismerni.
- Ismeretében már ”nincs bizonytalanság” a mintában (úgy értve, hogy egyértelmű a minta eloszlása, már nem függ az ismeretlen paramétertől).

# Elégséges statisztika diszkrét minta esetén

Def.: A diszkrét $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges $\theta-$ra, ha a $P_\theta(\boldsymbol{\xi} = \mathbf{x}|T(\boldsymbol{\xi}) = t)$ feltételes valószínűség nem függ $\theta-$tól

# Likelihood függvény

Def.: A $\xi_1, ..., \xi_n$ független, azonos eloszlású minta likelihood függvénye

$$L(\mathbf{x}, \theta) = \begin{cases} P_\theta(\boldsymbol{\xi} = \mathbf{x}) = \prod_{i=1}^{n} P_\theta(\xi_i = x_i) & \text{diszkrét minta esetén} \\ f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) & \text{abszolút folytonos minta esetén} \end{cases}$$

ahol $f_\theta$ $\xi_i$ sűrűségfüggvénye.

$l(\mathbf{x}, \theta) = \ln L(\mathbf{x}, \theta)$ a loglikelihood függvény.

# Maximum likelihood becslés

- Definíció heurisztikusan: azt a paraméterértéket keressük, amelyre az adott minta bekövetkezési valószínűsége maximális.

Def.: $\theta$ maximum likelihood becslése $\hat{\theta} = T(\boldsymbol{\xi}) \in \Theta$, ha

$$L(\boldsymbol{\xi}, \hat{\theta}) = \max_{\theta \in \Theta} L(\boldsymbol{\xi}, \theta)$$

# Likelihood egyenlet

Gyakran a loglikelihood függvény maximumhelyét keresik a

$\frac{\partial l(\mathbf{x},\theta)}{\partial \theta} = 0$ egyenletet (vagy egyenletrendszert) megoldva.

Ez diszkrét minta esetén a

$$\sum_{i=1}^{n} \frac{\partial \ln P_\theta(\xi_i = x_i)}{\partial \theta} = 0$$

egyenletet (vagy egyenletrendszert) jelenti.

Abszolút folytonos minta esetén

$$\sum_{i=1}^{n} \frac{\partial \ln f_\theta(x_i)}{\partial \theta} = 0$$

egyenletet (vagy egyenletrendszert) oldjuk meg.

# Koronavírusos halálesetek száma

| nap | Ausztria | Csehország | Magyarország | Szlovákia |
|---|---|---|---|---|
| 01.ápr | 20 | 7 | 1 | 0 |
| 02.ápr | 18 | 8 | 4 | 0 |
| 03.ápr | 12 | 5 | 1 | 0 |
| 04.ápr | 10 | 9 | 11 | 0 |
| 05.ápr | 18 | 6 | 2 | 0 |
| 06.ápr | 18 | 8 | 4 | 0 |
| átlag | 16,0 | 7,2 | 3,8 | 0,0 |

Példa. (Poisson)
Tegyük fel, hogy $\eta_1, \eta_2, \ldots, \eta_n \sim Poisson(\lambda)$. Ekkor

$$\mathrm{L}(\underline{k}, \lambda) = \mathrm{P}_\lambda(\eta_1 = k_1, \ldots, \eta_n = k_n) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \left( \prod_{i=1}^{n} \frac{1}{k_i!} \right) \lambda^{\sum k_i} e^{-n\lambda}$$

$$\ell(\underline{k}, \lambda) = \ln \mathrm{L}(\underline{k}, \lambda) = \left( \sum_{i=1}^{n} \ln \left( \frac{1}{k_i!} \right) \right) + \left( \sum_{i=1}^{n} k_i \right) \ln \lambda - n\lambda$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{\sum k_i}{\lambda} - n = 0 \iff \lambda = \frac{\sum k_i}{n}$$

így az ML becslés

$$\widehat{\lambda} = \frac{\sum \eta_i}{n}$$

Adottak sorszámozott gömbök (lottóhúzás) 1-től *N*-ig.
Visszatevéses húzás esetén becsüljük meg *N*-t!

$$\mathrm{P}_N(\xi_i = k) = \frac{1}{N}\chi\{k \le N\}$$

$$\mathrm{L}(\underline{k}, \lambda) = \mathrm{P}_N(\xi_1 = k_1, \ldots, \xi_n = k_n) = \frac{1}{N^n}\chi\{\max_i k_i \le N\}$$

$$\widehat{N} = \max_i \xi_i$$


Normális eloszlást feltételezve a paraméterek ML becslése:

$$\hat{m} = 405 \text{ és } \widehat{\sigma^2} = 149\ 284$$

# Momentum módszer

- Ha az eloszlást k db paraméter határozza meg, akkor k db egyenletből kaphatunk rájuk becslést. Az egyenletek a tapasztalati és az elméleti momentumok egybevetéséből adódnak:

$$m_i(\underline{\theta}) = E_{\underline{\theta}}(X^i)$$

$$m_i(\underline{\theta}) = \frac{\sum_{j=1}^{n} \left(\xi_j\right)^i}{n}$$

# Konfidenciaintervallum

- Olyan intervallum, mely legalább 1-$\alpha$ valószínűséggel tartalmazza a keresett paramétert:

$$P_\theta\left(T_1(X) < \theta < T_2(X)\right) \geq 1-\alpha$$

# Példa (normális eloszlás)

- A Gyorskenyér Kft automata kenyérsütő készülékei egyszerre 100 kenyeret sütnek ki. Ezek tömegei grammban mérve $N(m,10^2)$ eloszlással közelíthetőek, ahol $m$ a kezelő beállításától függ. Egy ellenőrzésnél megmérték mind a 100 kenyér tömegét. Az átlag 990 g volt. Készítsünk 95%-os megbízhatóságú konfidencia intervallumot $m$-re!

# Konfidencia intervallum normális eloszlás várható értékére (ismert szórás esetén)

$$\xi_1,\ldots,\xi_n \sim N(m,\sigma^2),\ \sigma \text{ ismert},\ \Phi(u_y)=y \Rightarrow$$

$$P\left(\bar{\xi}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}<m<\bar{\xi}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha,$$

$$P\left(m>\bar{\xi}-u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha,$$

$$P\left(m<\bar{\xi}+u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

# Konfidencia intervallum várható értékre (ismert szórás esetén)

$$\xi_1,\ldots,\xi_n,\ E\xi_i = m,\ D^2\xi_i = \sigma^2,\ \sigma \text{ ismert} \Rightarrow$$

$$P\left(\overline{\xi} - \sqrt{\frac{1}{\alpha}}\frac{\sigma}{\sqrt{n}} < m < \overline{\xi} + \sqrt{\frac{1}{\alpha}}\frac{\sigma}{\sqrt{n}}\right) \geq 1-\alpha.$$

| $\alpha$ | $u_{1-\alpha/2}$ | $\sqrt{\frac{1}{\alpha}}$ |
|---|---|---|
| 10% | 1,64 | 3,16 |
| 5% | 1,96 | 4,47 |
| 2,50% | 2,24 | 6,32 |
| 1% | 2,58 | 10,00 |

# Konfidencia intervallum ”sok” megfigyelés esetén

$$\xi_1,\ldots,\xi_n,\ D^2\xi_i=\sigma^2 \text{ ismert} \Rightarrow$$

$$P\left(\bar{\xi}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}<m<\bar{\xi}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\sim 1-\alpha.$$

# Példák (milyen valószínűséggel születik fiúgyermek?)

- Svájcban 1871 és 1900 között a 2.644.757 megszületett gyermekből 1.359.671 fiú és 1.285.086 lány volt.
- Fiúk relatív gyakorisága így 0,5141.

$$p(1-p) \le \frac{1}{4} \Rightarrow$$

$$P\left( \bar{\xi} - \frac{u}{2\sqrt{n}} < p < \bar{\xi} + \frac{u}{2\sqrt{n}} \right) \sim 2\Phi(u) - 1$$

Esetünkben $0{,}9973$ valószínűséggel $0{,}5132 \le p \le 0{,}5150$

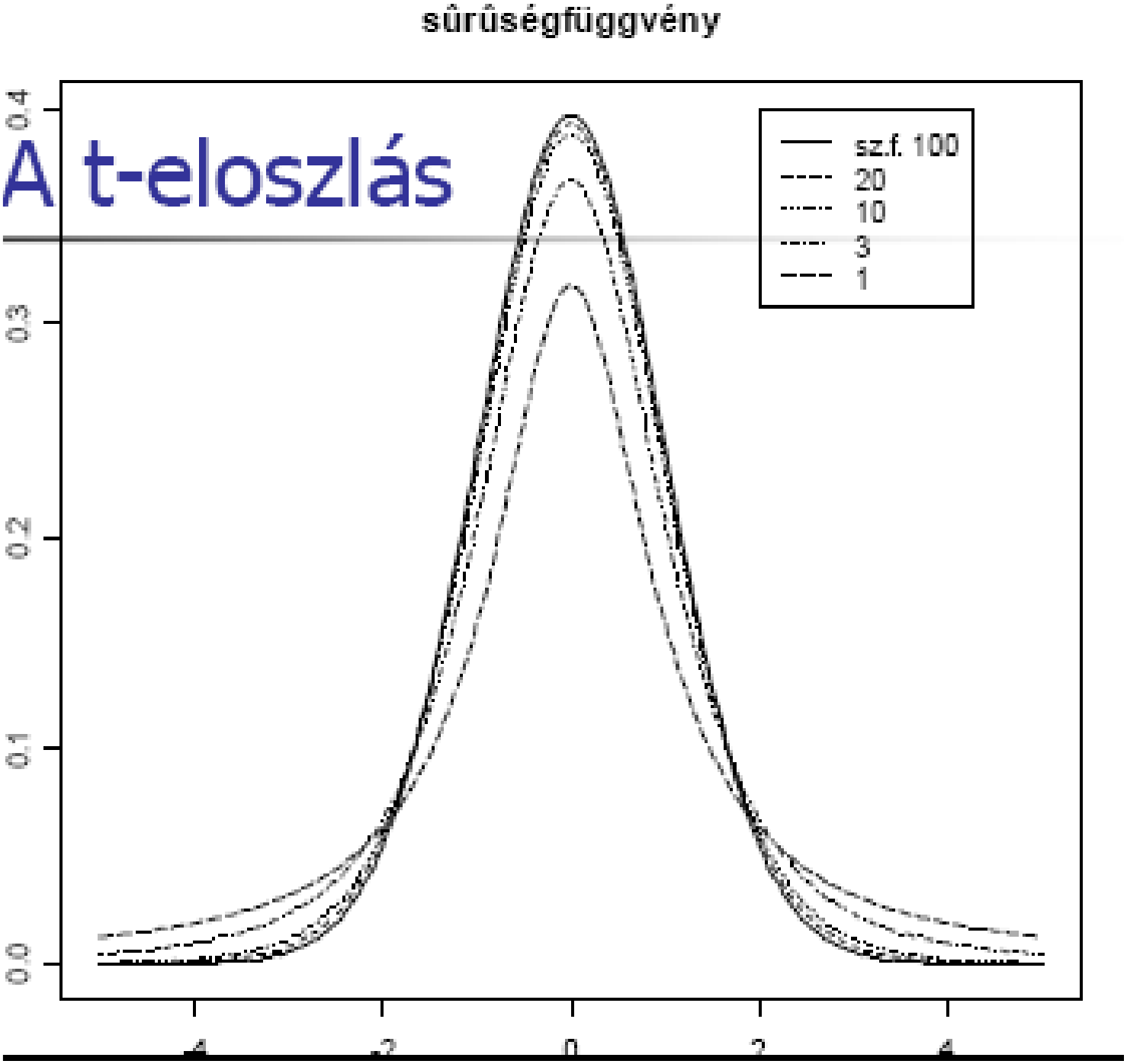
# Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén)

- Ha a szórás nem ismert, becsüljük
- Tétel (biz. nélkül): normális eloszlású minta esetén a mintaátlag és a tapasztalati szórás független
- *n* szabadságfokú t (Student) eloszlás:

$$X_0, X_1, \ldots, X_n \ \textbf{\textit{független}}\ N(0,1)$$

$$\frac{X_0}{\sqrt{\left(X_1^2 + \ldots + X_n^2\right)/n}} \sim t_n$$

# A t-eloszlás



# Konfidencia intervallum normális eloszlás várható értékére (ismeretlen szórás esetén) (folyt.)

$$\xi_1,\ldots,\xi_n \sim N(m,\sigma^2), \tilde{\sigma}^2 = \left(\left(\xi_1-\bar{\xi}\right)^2+\ldots+\left(\xi_n-\bar{\xi}\right)^2\right)/(n-1) \Rightarrow$$

$$\frac{\sqrt{n}\left(\bar{\xi}-m\right)}{\sqrt{\tilde{\sigma}^2}} \sim t_{n-1}$$

$$P(t_{n-1} < t_{n-1,y}) = y$$

$$P\left(\bar{\xi}-t_{n-1,1-\alpha/2}\frac{\tilde{\sigma}}{\sqrt{n}} < m < \bar{\xi}+t_{n-1,1-\alpha/2}\frac{\tilde{\sigma}}{\sqrt{n}}\right) = 1-\alpha,$$

$$P\left(m > \bar{\xi}-t_{n-1,1-\alpha}\frac{\tilde{\sigma}}{\sqrt{n}}\right) = 1-\alpha,$$

$$P\left(m < \bar{\xi}+t_{n-1,1-\alpha}\frac{\tilde{\sigma}}{\sqrt{n}}\right) = 1-\alpha$$

# Példa (kenyér. folyt.)

- Tegyük fel most, hogy nem ismerjük Gyorskenyér Kft kenyereinek szórását. Az átlag 990 g volt.
- Ismert 10 szórásnál 991,6 g volt a 95%-os megbízhatóságú felső konfidencia határ.
- Amennyiben a korrigált tapasztalati szórás is 10, akkor ez a határ csak kis mértékben változik (991,8 g).
- Azonban 50-es korrigált tapasztalati szórásnál ez az érték 999 g-ra változik.

# *u* és *t* együtthatók összehasonlítása

$$u_{1-5\%} = 1,64 \qquad (\Phi(1,64) = 95\%)$$

| $n$ | $t_{n-1,1-5\%}$ |
|---|---|
| 2 | 6,31 |
| 3 | 2,92 |
| 4 | 2,35 |
| 5 | 2,13 |
| 10 | 1,83 |
| 20 | 1,73 |
| 50 | 1,68 |
| 100 | 1,66 |
| 1000 | 1,65 |