

# Becslések és tulajdonságaik (11. előadás)

- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$  valamely eloszlások halmaza valamilyen  $\Theta$  halmazzal ( $\Theta$  a paramétertér, a paraméter összes lehetséges értékeiből álló halmaz);
- $\psi : \Theta \rightarrow \mathbb{R}$  függvény.
- Cél:  $\psi(\vartheta)$  becslése, azaz olyan  $T$  statisztika keresése, amire a  $T(X_1, \dots, X_n)$  valószínűségi változó és a  $\psi(\vartheta)$  érték valamilyen értelemben közel esnek egymáshoz.

# Becslések és tulajdonságaik (11. előadás)

- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$  valamely eloszlások halmaza valamilyen  $\Theta$  halmazzal ( $\Theta$  a paramétertér, a paraméter összes lehetséges értékeiből álló halmaz);
- $\psi : \Theta \rightarrow \mathbb{R}$  függvény.
- Cél:  $\psi(\vartheta)$  becslése, azaz olyan  $T$  statisztika keresése, amire a  $T(X_1, \dots, X_n)$  valószínűségi változó és a  $\psi(\vartheta)$  érték valamilyen értelemben közel esnek egymáshoz.

## Definíció (Torzítatlanság)

A  $T$  statisztika torzítatlan becslés  $\psi$ -re, ha minden  $\vartheta \in \Theta$ -ra

$$\mathbb{E}_\vartheta(T(X_1, \dots, X_n)) = \psi(\vartheta).$$

A  $T$  statisztika torzítása a  $b_T(\vartheta) = \mathbb{E}_\vartheta(T(X_1, \dots, X_n)) - \psi(\vartheta)$  függvény.

**Példa.**  $X_1, X_2, \dots, X_n$  független minta a  $[0, \vartheta]$  intervallumon egyenletes eloszlásból. Ekkor  $2\bar{X}$  torzítatlan becslés  $\psi(\vartheta) = \vartheta$ -ra.

# Torzítatlan becslések

## Állítás (A várható érték torzítatlan becslése)

Legyen  $X_1, \dots, X_n$  független azonos eloszlású véges várható értékű minta. Ekkor

$$\mathbb{E}_\vartheta(\bar{X}) = \mathbb{E}_\vartheta(X_1) \quad \text{minden } \vartheta \in \Theta\text{-ra,}$$

vagyis a **mintaátlag** torzítatlan becslés  $\psi$ -re.

## Állítás (A szórásnégyzet torzítatlan becslése)

$X_1, \dots, X_n$  független azonos eloszlású véges szórású minta. Ekkor Ekkor

$$\mathbb{E}_\vartheta(s_n^{*2}) = D_\vartheta^2(X_1) \quad \text{minden } \vartheta \in \Theta\text{-ra,}$$

vagyis a **korrigált tapasztalati szórásnégyzet** torzítatlan becslés a szórásnégyzet-re.

# Becslések összehasonlítása

## Definíció (Hatásosság)

Legyenek  $T_1, T_2$  **torzítatlan** becslései a paraméter  $\psi(\vartheta)$  függvényének.  $T_1$  **hatásosabb**  $T_2$ -nél, ha

$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

teljesül minden  $\vartheta \in \Theta$ -ra.

A  $T_1$  becslés **hatásos**  $\psi(\vartheta)$ -ra, ha  $\psi(\vartheta)$  minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).

# Becslések összehasonlítása

## Definíció (Hatásosság)

Legyenek  $T_1, T_2$  **torzítatlan** becslései a paraméter  $\psi(\vartheta)$  függvényének.  $T_1$  **hatásosabb**  $T_2$ -nél, ha

$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

teljesül minden  $\vartheta \in \Theta$ -ra.

A  $T_1$  becslés **hatásos**  $\psi(\vartheta)$ -ra, ha  $\psi(\vartheta)$  minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).

- Nem mindig létezik hatásos becslés, és lehetséges, hogy  $T_1$  és  $T_2$  közül egyik sem hatásosabb a másiknál.
- A várható értékre nézve a mintaátlag hatásosabb minden  $\sum_{j=1}^n c_j X_j$  alakú becslésnél (ahol  $\sum_{j=1}^n c_j = 1$ ).

# Becslések összehasonlítása

## Definíció (Hatásosság)

Legyenek  $T_1, T_2$  **torzítatlan** becslései a paraméter  $\psi(\vartheta)$  függvényének.  $T_1$  **hatásosabb**  $T_2$ -nél, ha

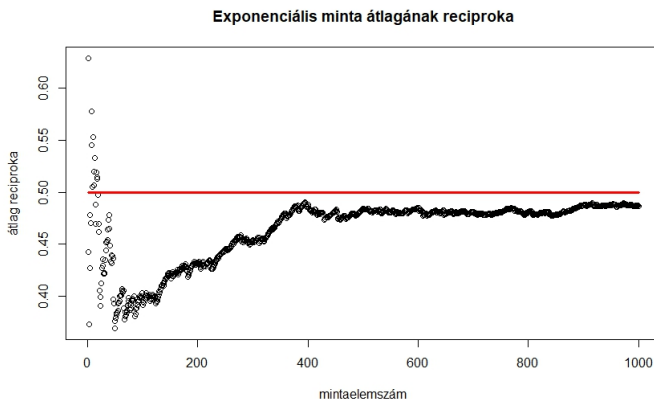
$$D_{\vartheta}^2(T_1) \leq D_{\vartheta}^2(T_2)$$

teljesül minden  $\vartheta \in \Theta$ -ra.

A  $T_1$  becslés **hatásos**  $\psi(\vartheta)$ -ra, ha  $\psi(\vartheta)$  minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).

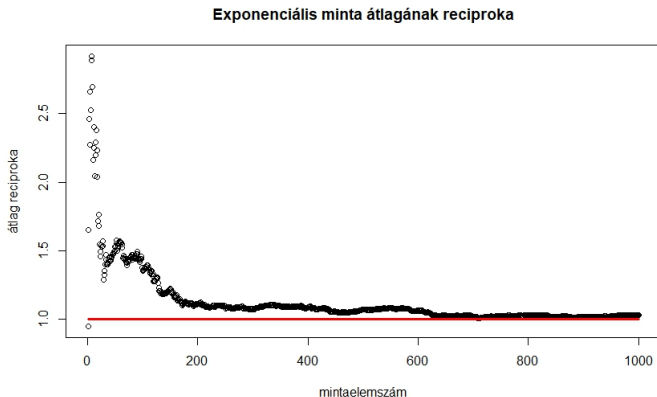
- Nem mindig létezik hatásos becslés, és lehetséges, hogy  $T_1$  és  $T_2$  közül egyik sem hatásosabb a másikonál.
- A várható értékre nézve a mintaátlag hatásosabb minden  $\sum_{j=1}^n c_j X_j$  alakú becslésnél (ahol  $\sum_{j=1}^n c_j = 1$ ).
- **Bizonyos feladatokban lehet a mintaátlagnál hatásosabb becslés a várható értékre:** A  $[0, b]$  intervallumon egyenletes eloszlás esetén  $b$ -re  $\frac{n+1}{n} \max(X_1, \dots, X_n)$  hatásosabb a mintaátlag kétszeresénél.

# Konzisztens becslés



$\lambda = 0,5$  paraméterű exponenciális eloszlást generálva a mintaátlag reciproka 0,5-höz tart, azaz **konzisztens** becslés, hiszen ez minden  $\lambda$ -ra teljesül.

# Konzisztens becslés



$\lambda = 1$  paraméterű exponenciális eloszlást generálva a mintaátlag reciproka 1-hez tart, azaz **konzisztens** becslés, hiszen ez minden  $\lambda$ -ra teljesül.

# Konzisztencia

## Definíció

A  $T_n = T_n(X_1, \dots, X_n)$  **konzisztens** becsléssorozat  $\psi(\vartheta)$ -ra, ha minden  $\vartheta \in \Theta$ -ra

$$(T_n(X_1, \dots, X_n)) \rightarrow \psi(\vartheta)$$

$n \rightarrow \infty$  esetén sztochasztikusan, azaz minden  $\vartheta \in \Theta$  és  $\varepsilon > 0$ -ra teljesül, hogy

$$\mathbb{P}_\vartheta(|T_n - \psi(\vartheta)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty).$$

# Konzisztencia

## Definíció

A  $T_n = T_n(X_1, \dots, X_n)$  **konzisztens** becsléssorozat  $\psi(\vartheta)$ -ra, ha minden  $\vartheta \in \Theta$ -ra

$$(T_n(X_1, \dots, X_n)) \rightarrow \psi(\vartheta)$$

$n \rightarrow \infty$  esetén sztochasztikusan, azaz minden  $\vartheta \in \Theta$  és  $\varepsilon > 0$ -ra teljesül, hogy

$$\mathbb{P}_\vartheta(|T_n - \psi(\vartheta)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty).$$

Elégséges feltétel:

$$\mathbb{E}_\vartheta(T(X)) \rightarrow \vartheta \quad \text{és} \quad D_\vartheta(T(X)) \rightarrow 0$$

minden  $\vartheta \in \Theta$ -ra.

## Példák torzítatlan, konzisztens becslésekre

$X_1, X_2, \dots$  független azonos eloszlású minta. Ekkor

$$T_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}_\theta(X_1)$$

teljesül  $n \rightarrow \infty$  esetén sztochasztikusan a nagy számok gyenge törvénye szerint, vagyis az **átlag** konzisztens becslés a **várható értékre**.

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűsége**re.

## Példák torzítatlan, konzisztens becslésekre

$X_1, X_2, \dots$  független azonos eloszlású minta. Ekkor

$$T_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}_\theta(X_1)$$

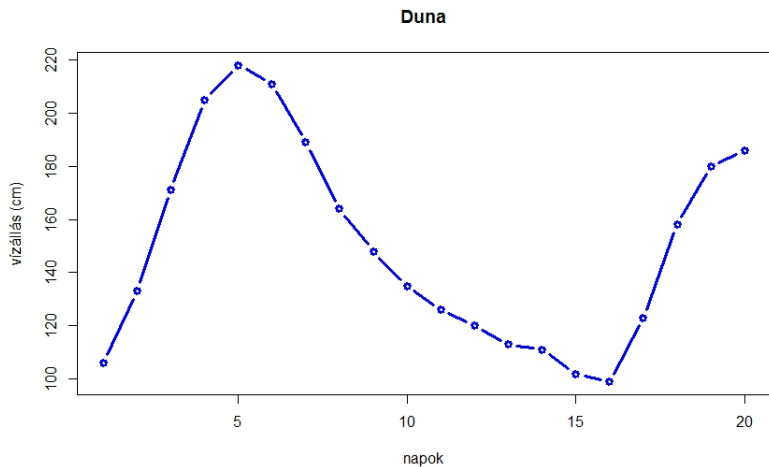
teljesül  $n \rightarrow \infty$  esetén sztochasztikusan a nagy számok gyenge törvénye szerint, vagyis az **átlag** konzisztens becslés a **várható értékre**.

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűségre**.

Nevezetes eloszlások:

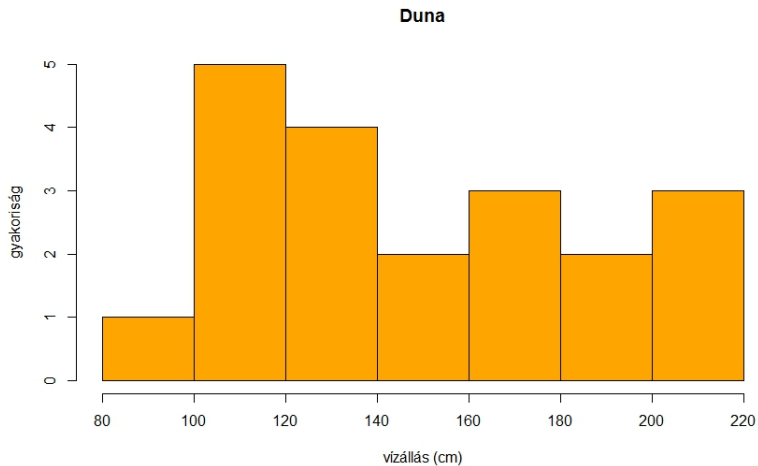
- Poisson-eloszlás  $\lambda$  paraméterére az átlag torzítatlan, konzisztens
- a normális eloszlás  $m$  paraméterére az átlag torzítatlan és konzisztens; a  $\sigma$  paraméterre a tapasztalati szórás és a korrigált tapasztalati szórás konzisztensek, de nem torzítatlanok;  $\sigma^2$ -re  $s_n^{*2}$  torzítatlan
- exponenciális eloszlás:  $1/\bar{X}$  konzisztens  $\lambda$ -ra, de nem torzítatlan a paraméterre
- exponenciális eloszlás:  $n \cdot \min(X_1, \dots, X_n)$  torzítatlan, de nem konzisztens a várható értékre (vagyis  $1/\lambda$ -ra).

## Példa: az adatok ábrázolása



# Példa: hisztogram

## A Duna vízállásának hisztogramja



# Alapstatisztikák

Minta:  $X_1, \dots, X_n$  (a példában  $X_1 = 106, X_2 = 133, \dots, X_{20} = 186$ )

- **minimum**: a legkisebb mintaelem, azaz  $\min(X_1, X_2, \dots, X_n)$ .
- **maximum**: a legnagyobb mintaelem, azaz  $\max(X_1, X_2, \dots, X_n)$ .
- **terjedelem** (range): a legnagyobb és legkisebb mintaelem különbsége, azaz
$$\max(X_1, X_2, \dots, X_n) - \min(X_1, X_2, \dots, X_n).$$
- **medián**: a **nagyság szerinti középső** mintaelem, vagy a középső kettő átlaga (ha  $n$  páros).
- **módusz** (mode): a leggyakrabban előforduló mintaelem.

# Alapstatisztikák

Minta:  $X_1, X_2, \dots, X_n$ .

- **mintaátlag** (mean):  $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j = \frac{X_1 + \dots + X_n}{n}$ .

- **tapasztalati szórásnégyzet**:

$$s_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2 = \left( \frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 = \frac{X_1^2 + X_2^2 + \dots + X_n^2}{n} - \bar{X}^2.$$

- tapasztalati szórás:  $s_n = \sqrt{s_n^2}$ .
- **korrigált tapasztalati szórásnégyzet** (variance):

$$s_n^{*2} = \frac{n}{n-1} \cdot s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{n}{n-1} \left( \left( \frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 \right).$$

- **korrigált tapasztalati szórás** (standard deviation, sd):  $s_n^* = \sqrt{s_n^{*2}}$ .

## További statisztikák

- **korrigált tapasztalati szórásnégyzet** (variance):

$$s_n^{*2} = \frac{n}{n-1} \cdot s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 = \frac{n}{n-1} \left( \left( \frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 \right).$$

- **korrigált tapasztalati szórás** (standard deviation, sd):  $s_n^* = \sqrt{s_n^{*2}}$ .
- **relatív szórás** (relative standard deviation, rsd):  $\frac{s_n^*}{\bar{X}}$ .
- **standard hiba (standard error)**:  $\frac{s_n^*}{\sqrt{n}}$ .

## Példa: alapstatisztikák

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

mintaelemszám:  $n = 20$

minta:  $X_1 = 106, X_2 = 133, \dots, X_{20} = 186$ .

átlag:  $\bar{X} = 149,9$

tapasztalati szórásnégyzet:  $s_n^2 = 1412,09$

tapasztalati szórás:  $s_n = 37,58$

korrigált tapasztalati szórásnégyzet:  $s_n^{*2} = 1486,411$

korrigált tapasztalati szórás:  $s_n^* = 38,55$

relatív szórás:  $0,257$

standard hiba:  $8,62$

## Rendezett minta

**Rendezett minta:** a mintaelemeket nagyság szerint növekvő sorrendbe állítjuk.

Jelölés:

$$(X_1^*, X_2^*, \dots, X_n^*).$$

Vagyis  $\{X_1^*, X_2^*, \dots, X_n^*\} = \{X_1, X_2, \dots, X_n\}$  és  $X_1^* \leq X_2^* \leq \dots \leq X_n^*$ .

A minimum  $X_1^*$ , a maximum  $X_n^*$ . A  $k$ . legkisebb mintaelem  $X_k^*$ .

## Rendezett minta

**Rendezett minta:** a mintaelemeket nagyság szerint növekvő sorrendbe állítjuk.  
Jelölés:

$$(X_1^*, X_2^*, \dots, X_n^*).$$

Vagyis  $\{X_1^*, X_2^*, \dots, X_n^*\} = \{X_1, X_2, \dots, X_n\}$  és  $X_1^* \leq X_2^* \leq \dots \leq X_n^*$ .

A minimum  $X_1^*$ , a maximum  $X_n^*$ . A  $k$ . legkisebb mintaelem  $X_k^*$ .

---

**Példa:** a Duna vízállásáról kapott húszelemű adatsor rendezett mintája:

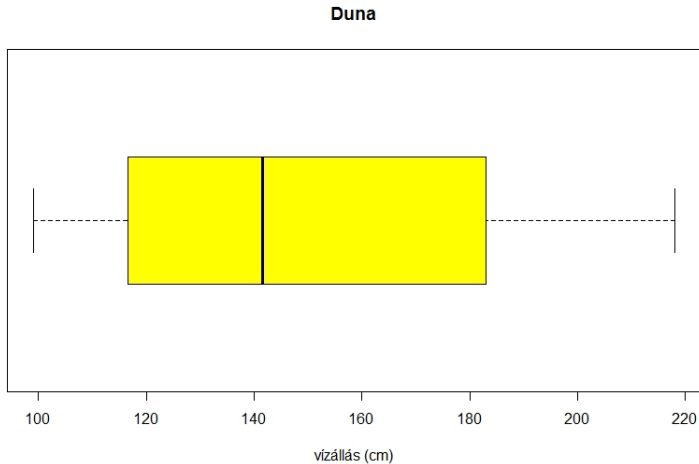
99	102	106	111	113	120	123	126	133	135
148	158	164	171	180	186	189	205	211	218

$X_1^* = 99, X_2^* = 102, X_3^* = 106, \dots, X_6^* = 120, \dots, X_{10}^* = 135$

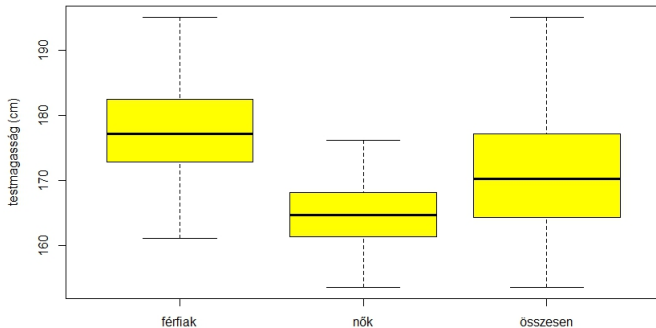
$X_{11}^* = 148, \dots, X_{14}^* = 171, \dots, X_{20}^* = 218.$

## Példa: boxplot

A Duna vízállásáról szóló minta boxplotja a húsznapos adatsorból



## Példa: boxplot



A testmagasság boxplotja  $n = 96$  elemű mintából, balról jobbra: férfiak, nők, összesen

# Kvantilisek

Az  $X$  valószínűségi változó  $z$ -kvantilise a legkisebb olyan  $q$  szám, melyre teljesül, hogy  $\mathbb{P}(X \leq q) \geq z$ .

A tapasztalati  $z$ -kvantilise több definíciót is szoktak használni, egy lehetőség:

## Definíció (Tapasztalati kvantilis)

Legyen  $X_1^* \leq X_2^* \leq \dots \leq X_n^*$  rendezett minta, és  $z \in [0, 1]$  adott szám. Ekkor a minta tapasztalati  $z$ -kvantilise:

$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

# Kvantilisek

Az  $X$  valószínűségi változó  $z$ -kvantilise a legkisebb olyan  $q$  szám, melyre teljesül, hogy  $\mathbb{P}(X \leq q) \geq z$ .

A tapasztalati  $z$ -kvantilisre több definíciót is szoktak használni, egy lehetőség:

## Definíció (Tapasztalati kvantilis)

Legyen  $X_1^* \leq X_2^* \leq \dots \leq X_n^*$  rendezett minta, és  $z \in [0, 1]$  adott szám. Ekkor a minta tapasztalati  $z$ -kvantilise:

$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

Első kvartilis:  $z = 1/4$ -kvantilis, harmadik kvartilis:  $z = 3/4$ -kvantilis, a medián pedig a  $z = 1/2$ -hez tartozó tapasztalati kvantilis.

# Boxplot

## Definíció (Tapasztalati kvantilis)

Legyen  $X_1, X_2, \dots, X_n$  minta, és  $z \in [0, 1]$  adott szám. Ekkor a minta tapasztalati  $z$ -kvantilise:

$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

A boxplot készítéséhez szükséges adatok:

- **minimum**: a legkisebb mintaelem (99);
- **első kvartilis**: a  $z = 1/4$ -hez tartozó kvantilis ( $118,2 = X_5^* + 0,25 \cdot (X_6^* - X_5^*)$ );
- **medián** (141,5);
- **harmadik kvartilis**: a  $z = 3/4$ -hez tartozó kvantilis (181,5);
- **maximum**: a legnagyobb mintaelem (218).

# Konfidenciaintervallumok

Legyen  $(\Omega, \mathcal{A}, \mathcal{P})$  statisztikai mező,  $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$  és  $\underline{X} = (X_1, \dots, X_n)$  független azonos eloszlású minta. Tegyük fel, hogy  $\vartheta$  valós paraméter, vagyis  $\Theta \subseteq \mathbb{R}$ .

## Definíció

Azt mondjuk, hogy a  $(T_1(\underline{X}), T_2(\underline{X}))$  intervallum legalább  $1 - \alpha$  megbízhatósági szintű konfidenciaintervallum  $\vartheta$ -ra, ha minden  $\vartheta \in \mathbb{R}$  esetén teljesül, hogy

$$\mathbb{P}_\vartheta(T_1(\underline{X}) < \vartheta < T_2(\underline{X})) \geq 1 - \alpha.$$

A konfidenciaintervallum megbízhatósági szintje:  $\inf_{\vartheta \in \Theta} \{\mathbb{P}_\vartheta(\vartheta \in (T_1, T_2))\}$ .

# Konfidenciaintervallum

Példa: hatvan különböző mintából megmértük a talajvíz pH-értékét egy adott helyen.

A minta egy részlete:

5,98    6,1    5,99    6,21    5,97    6,23    ...    5,85

Alapstatisztikák:  $n = 60$  (méret),  $\bar{X} = 5,99$  (átlag),  $s_n^* = 0,18$  (korigált tapasztalati szórás)

# Konfidenciaintervallum

Példa: hatvan különböző mintából megmértük a talajvíz pH-értékét egy adott helyen.

A minta egy részlete:

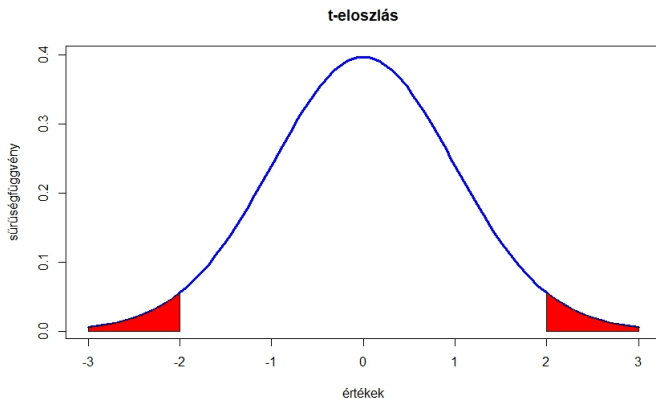
5,98    6,1    5,99    6,21    5,97    6,23    ...    5,85

Alapstatisztikák:  $n = 60$  (méret),  $\bar{X} = 5,99$  (átlag),  $s_n^* = 0,18$  (korigált tapasztalati szórás)

Cél: a várható érték becslése az adatok alapján

pontosabban: adjunk meg egy olyan intervallumot, ami legalább 95% valószínűséggel tartalmazza az "igazi" várható értéket – ezt fogjuk **95 % megbízhatósági szintű konfidenciaintervallumnak** hívni

## $t$ -eloszlás kritikus értékei



Az  $f = 59$  szabadsági fokú  $\alpha = 0,05$  terjedelmű kétoldali  $t$ -próba kritikus értéke:  
 $t_{59,0,05} = 2,001$ .

## Konfidenciaintervallum a várható értékre

Legyenek  $Z_0, Z_1, \dots, Z_n$  független  $N(0, 1)$  eloszlásúak, és  $t_{f, \alpha}$  az  $f$  szabadsági fokú  $\alpha$  terjedelmű kétoldali  $t$ -próba kritikus értéke, azaz az  $f$  szabadsági fokú  $t$ -eloszlás  $1 - \alpha/2$ -kvantilise:

$$1 - \alpha/2 = \mathbb{P}(Y \leq t_{f, \alpha}) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \dots + Z_f^2}} \leq t_{f, \alpha}\right).$$

Az  $Y = \frac{Z_0}{\sqrt{Z_1^2 + \dots + Z_f^2}}$  valószínűségi változó eloszlása  $f$  szabadsági fokú  **$t$ -eloszlás**.

## Konfidenciaintervallum a várható értékre

Legyenek  $Z_0, Z_1, \dots, Z_n$  független  $N(0, 1)$  eloszlásúak, és  $t_{f, \alpha}$  az  $f$  szabadsági fokú  $\alpha$  terjedelmű kétoldali  $t$ -próba kritikus értéke, azaz az  $f$  szabadsági fokú  $t$ -eloszlás  $1 - \alpha/2$ -kvantilise:

$$1 - \alpha/2 = \mathbb{P}(Y \leq t_{f, \alpha}) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \dots + Z_f^2}} \leq t_{f, \alpha}\right).$$

Az  $Y = \frac{Z_0}{\sqrt{Z_1^2 + \dots + Z_f^2}}$  valószínűségi változó eloszlása  $f$  szabadsági fokú  **$t$ -eloszlás**.

### Állítás (Konfidenciaintervallum a várható értékre, ismeretlen szórás)

*Tegyük fel, hogy  $X_1, \dots, X_n$  független  $N(m, \sigma^2)$  normális eloszlású valószínűségi változók ( $m, \sigma$  ismeretlenek). Ekkor a*

$$(T_1, T_2) = \left( \bar{X} - t_{n-1, \alpha} \cdot \frac{s_n^*}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha} \cdot \frac{s_n^*}{\sqrt{n}} \right)$$

*intervallum  $1 - \alpha$  megbízhatósági szintű kétoldali konfidenciaintervallum az eloszlás várható értékére.*

## Példa konfidenciaintervallumra

Minta: megmértük a talajvíz pH-értékét, tegyük fel, hogy ez **normális eloszlású**, szórása nem ismert.

$n = 60$  (méret),  $\bar{X} = 5,99$  (átlag),  $s_n^* = 0,18$  (korrigált tapasztalati szórás)

Adjunk meg olyan intervallumot, ami legalább 95% valószínűséggel tartalmazza a pH-érték valódi várható értékét – bármi is a valódi várható érték.

megbízhatósági szint:  $1 - \alpha = 95\%$ , azaz  $\alpha = 0,05$ . Az  $f = 59$  szabadsági fokú  $\alpha = 0,05$  terjedelmű kétoldali  $t$ -próba kritikus értéke:  $t_{59,0,975} = 2$ .

**95 %-os megbízhatósági szintű konfidenciaintervallum a várható értékre:**

$$\left( 5,99 - 2 \cdot \frac{0,18}{\sqrt{60}}, 5,99 + 2 \cdot \frac{0,18}{\sqrt{60}} \right) = (5,94; 6,04).$$