

## Rendezett minta (2. előadás)

**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 (2. előadás)

**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:

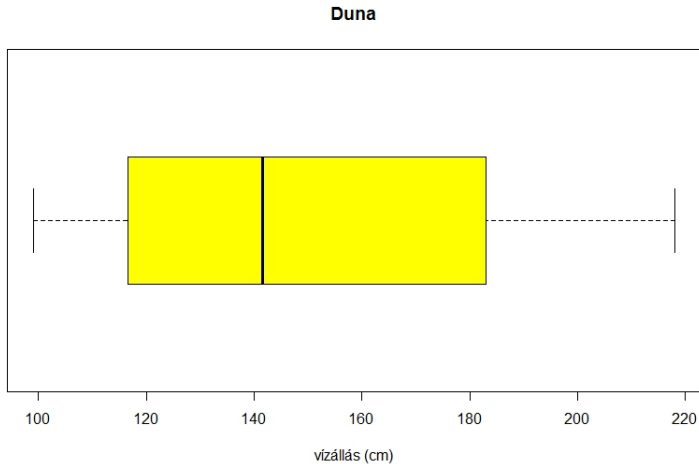
99	102	106	111	113	120	123	126	133	135
148	158	164	171	180	186	189	205	211	218

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

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

## Példa: boxplot

A Duna vízállásáról szóló minta boxplotja a húsznapos adatsorból



# 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).

# Tapasztalati eloszlásfüggvény

Az  $X$  valószínűségi változó eloszlásfüggvénye az  $F : \mathbb{R} \rightarrow [0, 1]$  függvény, melyre

$$F(t) = \mathbb{P}(X \leq t)$$

minden  $t \in \mathbb{R}$ -re.

# Tapasztalati eloszlásfüggvény

Az  $X$  valószínűségi változó eloszlásfüggvénye az  $F : \mathbb{R} \rightarrow [0, 1]$  függvény, melyre

$$F(t) = \mathbb{P}(X \leq t)$$

minden  $t \in \mathbb{R}$ -re.

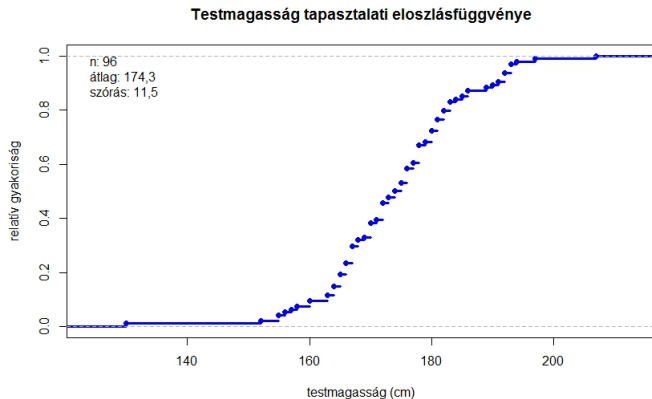
## Definíció (Tapasztalati eloszlásfüggvény)

Az  $X_1, X_2, \dots, X_n$  minta tapasztalati eloszlásfüggvénye az  $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$  függvény, melyre

$$\hat{F}_n(t) = \frac{t\text{-nél nem nagyobb mintaelemek száma}}{n}.$$

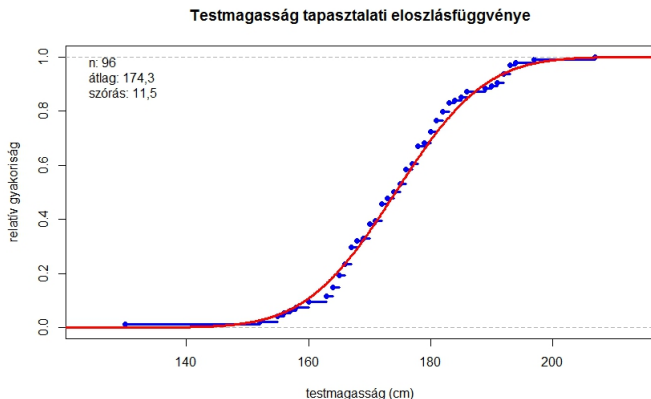
(empirical cumulative distribution function)

# Tapasztalati eloszlásfüggvény



A testmagasság tapasztalati eloszlásfüggvénye  $n = 96$  elemű mintából.

# Tapasztalati eloszlásfüggvény



A testmagasság tapasztalati eloszlásfüggvénye  $n = 96$  elemű mintából, és az  $\bar{X} = 174,3$  várható értékű és  $s_n^* = 11,5$  szórású normális eloszlás eloszlásfüggvénye.

## A statisztika alaptétele

Az  $X$  és  $Y$  valószínűségi változók azonos eloszlásúak, ha eloszlásfüggvényük megegyezik, azaz  $\mathbb{P}(X \leq t) = \mathbb{P}(Y \leq t)$  minden  $t \in \mathbb{R}$ -re.

### Tétel (Glivenko–Cantelli, 1933)

Legyenek  $X_1, X_2, \dots, X_n$  **független azonos eloszlású** valószínűségi változók, melyek közös eloszlásfüggvénye  $F$ . Ekkor az  $\hat{F}_n$  tapasztalati eloszlásfüggvényekből álló sorozat 1 valószínűséggel egyenletesen tart  $F$ -hez, azaz

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0\right) = 1.$$

## A statisztika alaptétele

Az  $X$  és  $Y$  valószínűségi változók azonos eloszlásúak, ha eloszlásfüggvényük megegyezik, azaz  $\mathbb{P}(X \leq t) = \mathbb{P}(Y \leq t)$  minden  $t \in \mathbb{R}$ -re.

### Tétel (Glivenko–Cantelli, 1933)

Legyenek  $X_1, X_2, \dots, X_n$  **független azonos eloszlású** valószínűségi változók, melyek közös eloszlásfüggvénye  $F$ . Ekkor az  $\hat{F}_n$  tapasztalati eloszlásfüggvényekből álló sorozat 1 valószínűséggel egyenletesen tart  $F$ -hez, azaz

$$\mathbb{P}\left(\limsup_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0\right) = 1.$$

Ha  $t$  rögzített: legyen  $\mathbb{I}_i = 1$ , ha  $X_i \leq t$ , és 0 különben. Ezek összege épp a  $t$ -nél nem nagyobb mintaelemek száma, ezért

$$\hat{F}_n(t) = \frac{\sum_{i=1}^n \mathbb{I}_i}{n} \rightarrow \mathbb{E}(\mathbb{I}_1) = \mathbb{P}(X_1 \leq t) = F(t)$$

1 valószínűséggel, hiszen  $\mathbb{I}_1, \mathbb{I}_2, \dots$  független, azonos eloszlású, véges várható értékű valószínűségi változók, a nagy számok erős törvénye alkalmazható. A tétel ennél erősebbet állít, a hibát  $t$ -től függetlenül felülről korlátozza.

# Statisztikai mező

## Definíció

Az  $(\Omega, \mathcal{A}, \mathcal{P})$  hármast **statisztikai mezőnek** nevezzük, ha minden  $\mathbb{P} \in \mathcal{P}$ -re  $(\Omega, \mathcal{A}, \mathbb{P})$  Kolmogorov-féle valószínűségi mező.

Paraméteres statisztika mező:  $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ . Ekkor  $\vartheta$  az ismeretlen paraméter, mely egy  $\Theta \subseteq \mathbb{R}^q$  ismert halmaz eleme.

Például:  $\mathcal{P}$  lehet például

- a  $\lambda$  paraméterű Poisson-eloszlások halmaza;
- a normális eloszlások halmaza (ekkor  $\vartheta = (m, \sigma)$  az ismeretlen paraméter);
- az  $[a, b]$  intervallumon egyenletes eloszlások halmaza.

# Minta és statisztika

## Definíció (Minta)

Legyen  $(\Omega, \mathcal{A}, \mathcal{P})$  statisztikai mező. Egy

$$\underline{X} = (X_1, X_2, \dots, X_n) : \Omega \rightarrow B \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót ( $n$  elemű) **mintának** nevezünk. Itt  $B$  a mintatér,  $n$  a minta elemszáma vagy nagysága. A minta független, ha az  $X_1, X_2, \dots, X_n$  valószínűségi változók függetlenek.

# Minta és statisztika

## Definíció (Minta)

Legyen  $(\Omega, \mathcal{A}, \mathcal{P})$  statisztikai mező. Egy

$$\underline{X} = (X_1, X_2, \dots, X_n) : \Omega \rightarrow B \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót ( $n$  elemű) **mintának** nevezünk. Itt  $B$  a mintatér,  $n$  a minta elemszáma vagy nagysága. A minta független, ha az  $X_1, X_2, \dots, X_n$  valószínűségi változók függetlenek.

## Definíció (Statisztika)

Legyen  $T : B \rightarrow \mathbb{R}^k$  függvény. Ekkor a  $T(X_1, X_2, \dots, X_n)$  valószínűségi változót statisztikának nevezzük.

Például:  $T(X_1, \dots, X_n) = \bar{X}$  a mintaátlag, vagy  $T(X_1, \dots, X_n) = s_n^*$  a korrigált tapasztalati szórásnégyzet.

# Becslések és tulajdonságaik

- $(\Omega, \mathcal{A}, \mathcal{P})$  statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$  valamely  $\Theta$  halmazzal ( $\Theta$  a paramétertér);
- $\psi : \Theta \rightarrow \mathbb{R}$  függvény.
- Cél: olyan  $T$  statisztika keresése, amire a  $T(X)$  valószínűségi változó és a  $\psi(\vartheta)$  érték valamilyen értelemben közel esnek egymáshoz.

# Becslések és tulajdonságaik

- $(\Omega, \mathcal{A}, \mathcal{P})$  statisztikai mező;
- $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$  valamely  $\Theta$  halmazzal ( $\Theta$  a paraméterter);
- $\psi : \Theta \rightarrow \mathbb{R}$  függvény.
- Cél: olyan  $T$  statisztika keresése, amire a  $T(X)$  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.*

# 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).

## Házi feladat február 27, 9:00-ig

Generáljunk  $n = 100000$  elemű exponenciális eloszlású mintákat  $\lambda = 1/2$ ,  $\lambda = 1$  és  $\lambda = 5$  paraméterekkel.

- Mindhárom különböző paraméterre ábrázoljuk  $k$  függvényében (de csak a  $k = 1, 2, \dots, 1000$  értékekre) az első  $k$  mintaelem átlagának reciprokát, és ezt hasonlítsuk össze a mintához tartozó paraméterrel.
- Csoportosítsuk a mintaelemeket ezresével, és legyen  $Y_i$  az  $i$ . csoport (vagyis az  $(i-1) \cdot 1000 + 1, (i-1) \cdot 1000 + 2, \dots, i \cdot 1000$  indexű mintaelemek) átlagának reciproka. Számítsuk ki az  $Y_1, Y_2, \dots, Y_{100}$  átlagát, és ezt hasonlítsuk össze a  $\lambda$  paraméterrel.
- Értelmezzük az eredményeket: mennyire látható, hogy exponenciális eloszlásnál az átlag reciproka konzisztens, de nem torzítatlan becslése a paraméternek?

# Házi feladat február 20-ig: megoldás

Kérdezzünk meg legalább húsz felnőtt ismerőst az

- testmagasságukról;
- nemükről;
- cipőjük méretéről.

Ábrázoljuk az adatokat boxplot ábrán.

## Házi feladat február 20-ig: megoldás

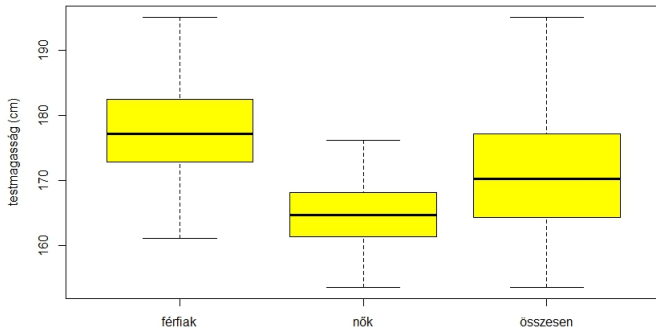
Kérdezzünk meg legalább húsz felnőtt ismerőst az

- testmagasságukról;
- nemükről;
- cipőjük méretéről.

Ábrázoljuk az adatokat boxplot ábrán.

```
boxplot(ferfi, no, együtt, col="yellow", names=c("férfiak", "nők",  
"összesen"), ylab="testmagasság (cm)")
```

## Házi feladat február 20-ig: megoldás



A testmagasság boxplotja 20 férfi és 20 nő adatai alapján (nem valós adatokból).