

Matematikai statisztika

Survey statisztika és adatanalitika mesterszak

Backhausz Ágnes

agnes.backhausz@ttk.elte.hu

Fogadóóra: kedd 8 – 9 és csütörtök 9 – 10, D 3-415

2022/2023. tavaszi félév

Bevezetés

A statisztika céljai

- mérési eredmények, megfigyelések elemzése (leíró statisztika)
- ismeretlen paraméterek becslése (matematikai statisztika, becslélmélet)
- hipotézisek ellenőrzése vagy cáfolata (matematikai statisztika, hipotézisvizsgálat)
- többdimenziós adatok elemzése (többdimenziós statisztika)
- véletlen folyamatok előrejelzése (regresszió, idősorelemzés)

Alkalmazási területek

- társadalomtudományok: szociológia, pszichológia
- élő- és élettelen természettudományok, pl. meteorológia, biológia
- pénzügyi matematika, biztosítás, közgazdaságtan
- mesterséges intelligencia

A kurzus célja és ajánlott irodalom

A kurzus célja a matematikai statisztika főbb módszereinek (például becslésméleti, hipotézisvizsgálati módszerek) és azok matematikai hátterének bemutatása, az alkalmazási készség elsajátítása.

- Bolla–Krámli: Statisztikai következtetések elmélete.
- Johnson–Bhattacharyya: Statistics.
- Móri–Szeidl–Zempléni: Matematikai statisztika példatár.
- Pröhle–Zempléni: Statistical problem solving in R.

Házi feladatok, gyakorló feladatok, tematika, diasor, jegyzet, minta vizsgafeladatsor (később): **moodle.elte.hu**

Követelmények (ld. neptun tárgytematika): 100 pontos írásbeli vizsga, 35, 53, 70, 89 ponthatárokkal. Minden házi feladat 3 pontot ér. Ezekkel legfeljebb egy jegyet lehet javítani, és a vizsga pontszámának legalább 35-nek kell lennie az elégségeshez.

Statisztikai elemzés

- **populáció:** azon egyedek összessége, akiről információt szeretnénk gyűjteni
például: budapesti lakosok, magyar választópolgárok, autótulajdonosok
- ha a teljes populáció adataival dolgozhatunk, big data elemzés végezhető; ha ez nem megvalósítható, véletlenszerűen választott mintákkal dolgozunk

Statisztikai elemzés

- **populáció:** azon egyedek összessége, akiről információt szeretnénk gyűjteni
például: budapesti lakosok, magyar választópolgárok, autótulajdonosok
- ha a teljes populáció adataival dolgozhatunk, big data elemzés végezhető; ha ez nem megvalósítható, véletlenszerűen választott mintákkal dolgozunk
- **minta:** az összegyűjtött adatok összessége
például: ezer megkérdezett budapesti lakos vagy ötven magyar autótulajdonos adatai

A statisztikai elemzés lépései

- tervezés: adatgyűjtés, mérés megtervezése
- adatgyűjtés, mérés
- kódolás: az adatok csoportokba sorolása, ha szükséges
- hibajavítás: olyan kiugró adatok korrekciója vagy elhagyása, amelyek feltehetően mérési hibából keletkeztek
- leíró statisztika: ellenőrzés, főbb jellemzők meghatározása, ábrázolás
- matematikai statisztikai elemzés, következtetések levonása

Statisztikai adatok

Adat: valamely sokaság jellemzőjére vonatkozó mért vagy számított eredmény

- **alapadatok:** méréssel vagy leszámlálással közvetlenül kapott eredmény
például: egy ember testmagassága, jövedelme, egy háztartásban élők száma
- **származtatott adatok:** az alapadatokból műveletek eredményeként kapjuk
például: emberek testmagasságának átlaga, a jövedelmek mediánja, az egy háztartásban élők számának szórása

Statisztikai adatok

Adat: valamely sokaság jellemzőjére vonatkozó mért vagy számított eredmény

- **alapadatok:** méréssel vagy leszámlálással közvetlenül kapott eredmény
például: egy ember testmagassága, jövedelme, egy háztartásban élők száma
- **származtatott adatok:** az alapadatokból műveletek eredményeként kapjuk
például: emberek testmagasságának átlaga, a jövedelmek mediánja, az egy háztartásban élők számának szórása

Az adatok **pontossága** általában korlátozott (mérési hiba, kerekítés, tévedés). Ha ϑ a valós érték, és X a mérés eredménye:

- **abszolút hiba:** a valós érték és a mérés eredményének különbségének abszolút értéke: $|X - \vartheta|$.
- **relatív hiba:** az abszolút hiba és a mért érték hányadosa: $\frac{|X - \vartheta|}{X}$.

Példa: egy mérleg 60 dkg lisztet 57 dkg-nak mér. Az abszolút hiba dkg-ban 3, a relatív hiba $3/57 = 5,3\%$.

Ismérvek, az adatok típusai

Statisztikai ismérv: a populáció egyedeit jellemző tulajdonság. Lehetséges kimenetelei az **ismérvváltozatok**.

Például: családi állapot (házas, özvegy stb.), háztartás létszáma (0, 1, 2, ...), választópolgár pártpreferenciája (pártok).

Ismérvék, az adatok típusai

Statisztikai ismérv: a populáció egyedeit jellemző tulajdonság. Lehetséges kimenetelei az **ismérváltozatok**.

Például: családi állapot (házas, özvegy stb.), háztartás létszáma (0, 1, 2, ...), választópolgár pártpreferenciája (pártok).

Az adatok típusai (skála)

- **nominális:** minőségi ismérv, csak az egyes ismérvváltozatok gyakoriságát tudjuk megszámolni (pl. nem, foglalkozás, nemzetiség)
- **ordinális:** egyértelmű sorrendbe rendezhető változatokkal rendelkező ismérv (pl. jó–közepes–rossz); kvantiliseket lehet számolni
- **intervallum:** az adatok különbsége egyértelmű, de a hányadosuk nem (pl. hőmérséklet – a hányados más, ha Celsius-fok helyett Fahrenheit-fokban számolunk)
- **arány:** az ismérv egy valós számmal jellemezhető, melyek különbsége és hányadosa is egyértelmű (pl. jövedelem, tömeg, csapadékmennyiség)

Matematikai statisztika

Példa matematikai statisztikai kérdésre

- Egy adott helyen húsz éven keresztül feljegyezték, hogy hányszor volt hurrikán. Ezek alapján várhatóan hány hurrikán lesz 2020-ban? Mennyi a becslésünk bizonytalansága? Mennyi a valószínűsége, hogy ötnél több hurrikán lesz?
- Egy közvéleménykutatás során 1000 ember közül 63 választana egy adott pártot. Ez alapján állíthatjuk-e, hogy a párt támogatottsága szignifikánsan magasabb 5%-nál? Mennyi a tévedésünk valószínűsége?
- 10000 ember közül egy véletlenszerűen választott csoport hatóanyagot tartalmazó oltást, a többiek sóoldatot (placebót) kaptak. Az első csoport 4876 tagja közül 45-en betegedtek meg később, a többiek közül 392-en. Állíthatjuk-e, hogy a hatóanyag és a betegség elkerülése között szignifikáns összefüggés van?
- Egy országban húsz éven keresztül figyelik a munkanélküliségi ráta és a bejelentett bűncselekmények számának együttes alakulását. Állíthatjuk-e, hogy szignifikáns összefüggés van a két mennyiség között?

Itt a mérések eredményét mindig **véletlenszerűnek**, arány ismérv esetén **valószínűségi változónak** tekintjük.

Matematikai statisztika

Statisztikai minta: (X_1, X_2, \dots, X_n) valószínűségi változók (azaz: valószínűségi vektorváltozó).

Mintaelemszám: n

A minta **független**, ha az (X_1, X_2, \dots, X_n) valószínűségi változók függetlenek (például ha a megkérdezetteket függetlenül választottuk, vagy ha a mérések nem befolyásolják egymást), azaz

$$\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \dots, X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1) \cdot \mathbb{P}(X_2 \leq t_2) \cdot \dots \cdot \mathbb{P}(X_n \leq t_n)$$

teljesül tetszőleges t_1, t_2, \dots, t_n valós számok esetén.

Az (X_1, X_2, \dots, X_n) valószínűségi változók **eloszlása nem ismert**: nem tudjuk, hogy mennyi $\mathbb{P}(X_1 \leq t)$, vagy mennyi X_1 várható értéke, szórása, vagy hogy két mennyiség között milyen erős a korreláció. A cél az adatok alapján

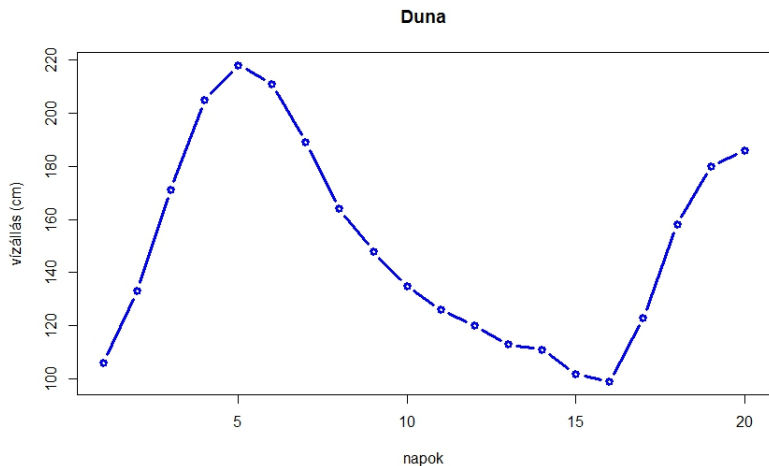
- a valószínűségi változók eloszlásának minél jobb megismerése
- a várható érték, szórás stb. becslése
- az eloszlásra vonatkozó hipotézisek eldöntése
- több valószínűségi változó együttes viselkedésének leírása

Leíró statisztika

Nem a véletlen hatásának megértése és valószínűségszámítási módszereken alapuló következtetések levonása a célja, hanem a megfigyelt adatok **megjelenítése, jellemzőinek kiszámítása**. Ide tartozhat:

- diagramok: kördiagram, oszlopdiaagram, hisztogram (lásd például: <https://www.ksh.hu/heti-monitor/>)
- táblázatok, kontingenciatáblák (például: https://www.ksh.hu/docs/hun/xstadat/xstadat_evkozi/e_odmv002.html)
- középértékek, szórások, egyéb statisztikák kiszámítása
- kvantilisek számítása, boxplot ábra
- indexek számítása

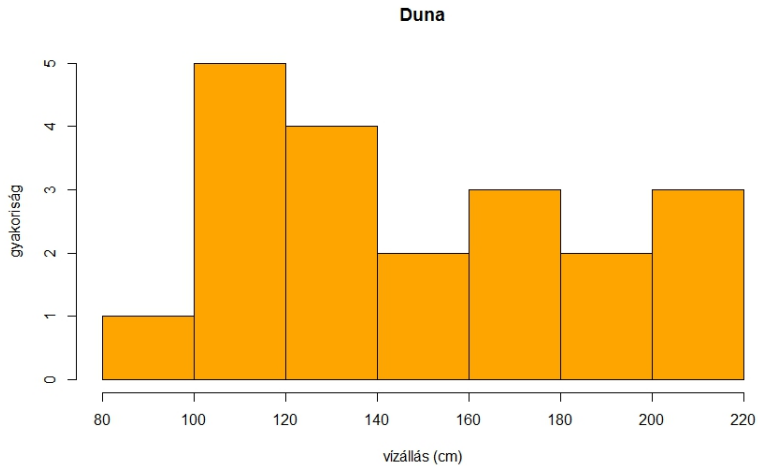
Példa: az adatok ábrázolása



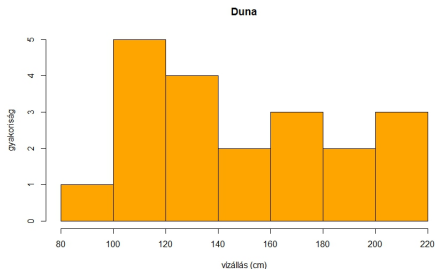
A Duna vízálása 20 napon keresztül (2016. január, adatok forrása: Országos Vízeljáró Szolgálat). Ez **nem független** minta.

Példa: hisztogram

A Duna vízállásának hisztogramja



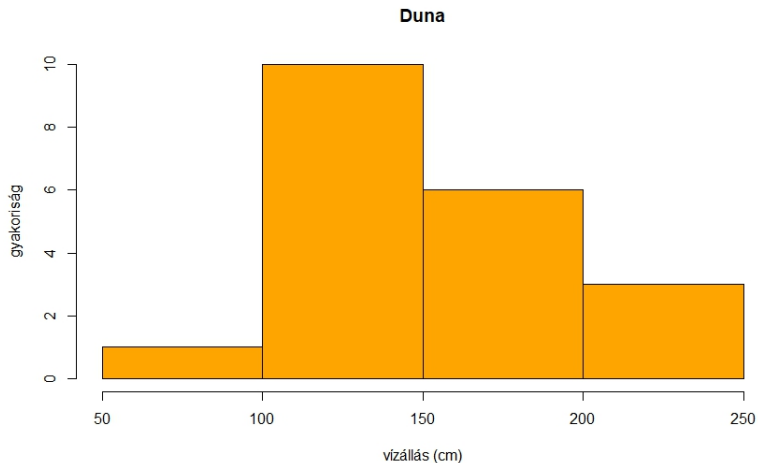
Példa: hisztogram



Választunk egy intervallumot, mely magában foglalja a mérési adatokat. Az intervallumot egyenlő nagyságú részekre osztjuk. Az egyes kis intervallumokba eső mérési adatok számát ábrázoljuk. Sem a túl hosszú, sem a túl rövid intervallumok nem adnak informatív ábrát.

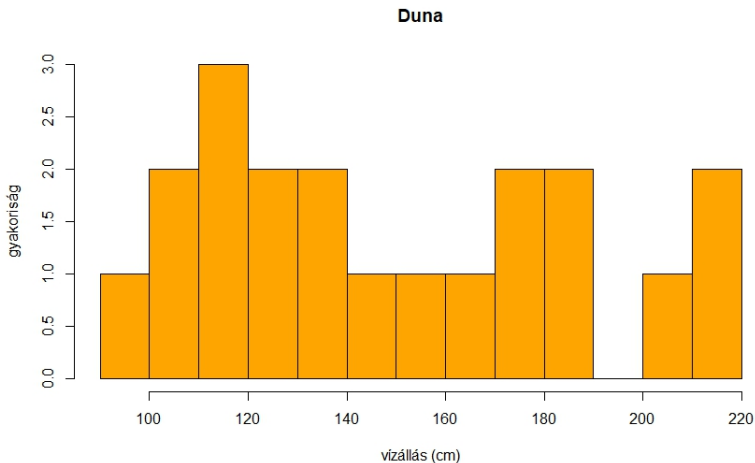
Emlékeztető: megfelelő választás és abszolút folytonos valószínűségi változó esetén a hisztogram a **sűrűségfüggvényhez** közelít.

Példa: túl hosszú intervallumok, túl kevés osztály



```
hist(viz, col="orange", xlab="vízállás (cm)", ylab="gyakoriság",  
main="Duna", breaks=4)
```

Példa: túl rövid intervallumok, túl sok osztály



```
hist(viz, col="orange", xlab="vízállás (cm)", ylab="gyakoriság",  
main="Duna", breaks=15)
```

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

X valószínűségi változó várható értéke: $\mathbb{E}(X)$, szórása: $D(X) = \sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}$.

Alapstatisztikák

X valószínűségi változó várható értéke: $\mathbb{E}(X)$, szórása: $D(X) = \sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}$.

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

- **tapasztalati szórásnégyzet**:

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

- tapasztalati szórás: $s_n = \sqrt{s_n^2}$.

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

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

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

További statisztikák

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

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

- **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

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

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

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)

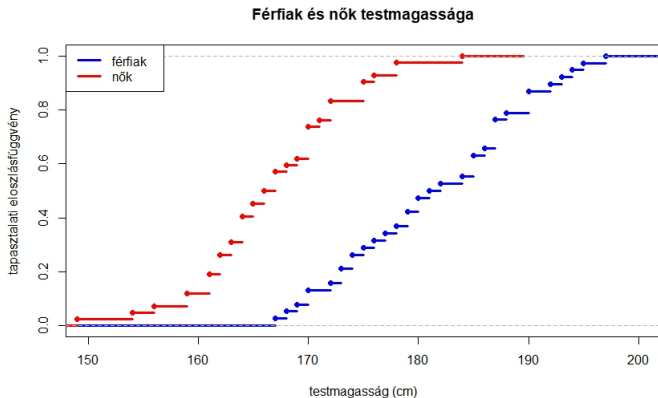
Az X_1, X_2, \dots, X_n minta tapasztalati eloszlásfüggvénye az $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

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

(empirical cumulative distribution function)

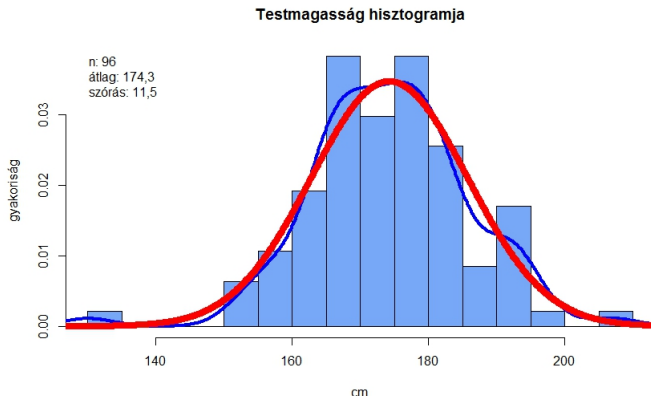
Az R-ben: `plot(ecdf(data))`

Tapasztalati eloszlásfüggvény



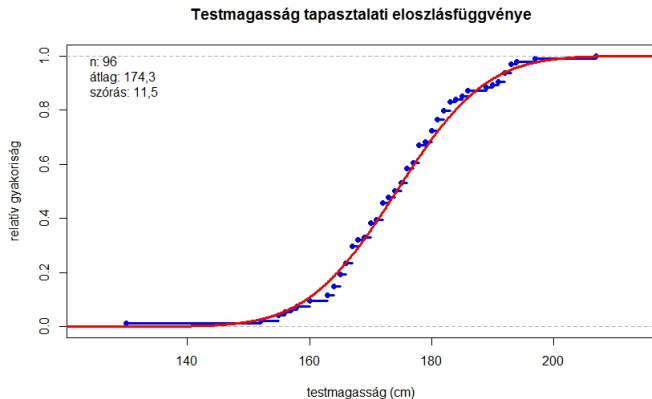
A testmagasság tapasztalati eloszlásfüggvénye $n = 96$ elemű mintából külön a férfiak (kék) és a nők (piros) esetében.

Testmagasság hisztogramja



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

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.

A **nagy számok erős törvénye** szerint független azonos eloszlású véges várható értékű valószínűségi változók átlaga 1 valószínűséggel tart a várható értékhez. Most:

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

ahol $\mathbb{I}_i = 1$, ha $X_i \leq t$, és különben 0. Ezek teljesítik a feltételeket.

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.

A **nagy számok erős törvénye** szerint független azonos eloszlású véges várható értékű valószínűségi változók átlaga 1 valószínűséggel tart a várható értékhez. Most:

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

ahol $\mathbb{I}_i = 1$, ha $X_i \leq t$, és különben 0. Ezek teljesítik a feltételeket.

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

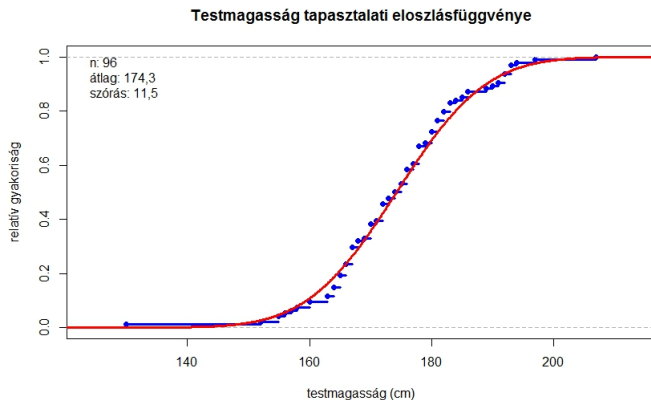
$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0\right) = 1.$$

A **nagy számok erős törvénye** szerint **minden t valós számra**

$$\lim_{n \rightarrow \infty} \hat{F}_n(t) = F(t).$$

Ez a pontonkénti konvergenciát jelenti. A tétel ennél erősebbet állít: ha megnézzük a két függvény közötti „legnagyobb különbséget”, még az is nullához tart.

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.

Házi feladat március 7., kedd, 12:00-ig

Kérdezzünk meg legalább 40 ismerőst az alábbiakról:

- 1 hány sorozatot nézett az elmúlt két hétben (legalább egy részt)
- 2 hetente átlagosan hányszor megy be az iskolásba/egyetemre/munkahelyére
- 3 egy tipikus hétköznapon átlagosan mennyi időt tölt utazással (percben)
- 4 egy tipikus héten mennyi időt tölt sportolással (percben)
- 5 milyen magas

a) Készítsünk hisztogramot az utazással töltött időről. b) Készítsük el a sportolással töltött idő tapasztalati eloszlásfüggvényét azoknál, akik a mediánnál több sorozatot néztek, illetve azoknál, akik legfeljebb annyit, mint a medián érték. Milyen következtetést vonhatunk le az ábrából?

Az adatokra az egész félév során szükség lesz a házi feladatok megoldásához.