

Matematikai statisztika

Survey statisztika és adatanalitika mesterszak

Backhausz Ágnes

agnes.backhausz@ttk.elte.hu

Fogadóóra: hétfő 10 – 11 és szerda 13 – 14, D 3-415

2019/2020. tavaszi félév

Bevezetés

A statisztika céljai

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

Alkalmazási területek

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

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, 30, 49, 68, 86 ponthatárokkal. Minden házi feladat 3 pont. Ezekkel legalább elégséges jegy esetén egy jegyet lehet javítani legfeljebb (a vizsga és a házi feladat pontok összeadódnak, ha a vizsga legalább elégséges).

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?
- Megmérték 100 férfi és 60 nő testmagasságát. Állíthatjuk-e az adatok alapján, hogy a férfiak szignifikánsan magasabbak a nőknél? Mennyi a tévedésünk valószínűsége?
- 100 ember közül 27 télen, 22 tavasszal, 34 nyáron, a többiek őszele születtek. Állíthatjuk-e az adatok alapján, hogy a születések eloszlása szignifikánsan eltér az egyenletes eloszlástól (amikor minden évszaknak $1/4$ a valószínűsége)?

Leíró statisztika

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

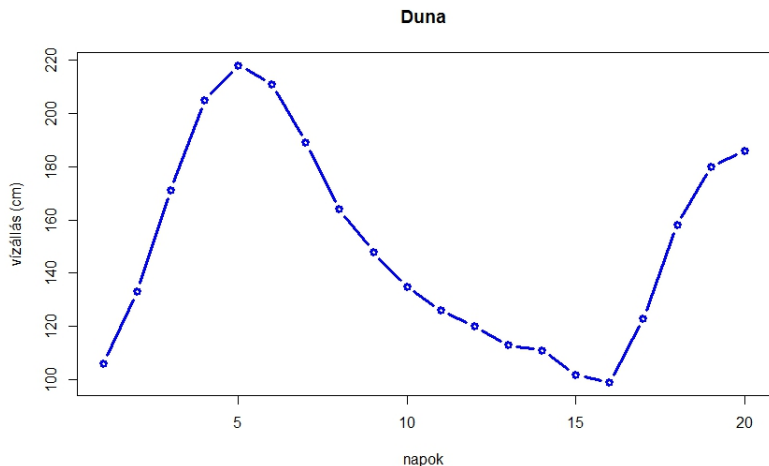
- diagramok: kördiagram, oszlopdiaagram, hisztogram
- táblázatok, kontingenciatáblák
- középértékek, szórások kiszámítása
- kvantilisok számítása, boxplot ábra
- indexek számítása

Indexek számítása

Cél: egyetlen értéket rendelni egy olyan jellemzőhöz, ami több komponensből áll össze, például: intelligencia, gazdasági állapot, tőzsde

- a komponensek kiválasztása
- a komponensek együttes viselkedésének megértése (például a korreláció segítségével)
- a súlyok és szorzók meghatározása: a komponensek lehetséges értékei alapján annak meghatározása, hogy melyiket milyen szorzóval vegyük figyelembe, hogy végül is megfelelő súllyal szerepelhessenek
- tesztelés: egy meglévő adathalmazon annak kipróbálása, hogy az index alapján előrejelezhető-e olyan jellemzők, amik az indexbe nincsenek beépítve

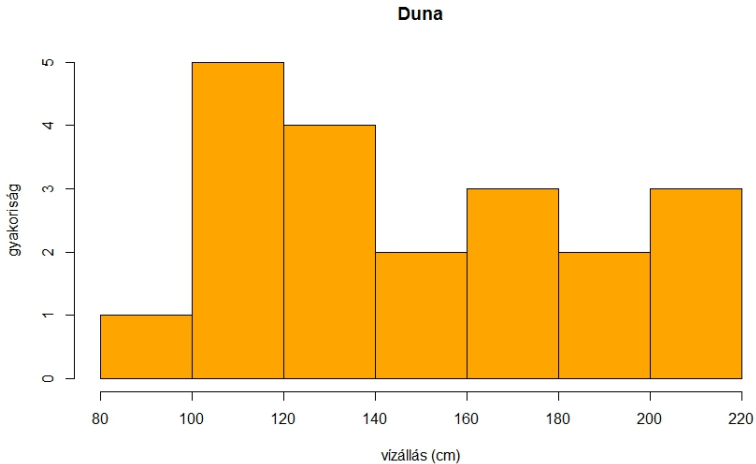
Példa: az adatok ábrázolása



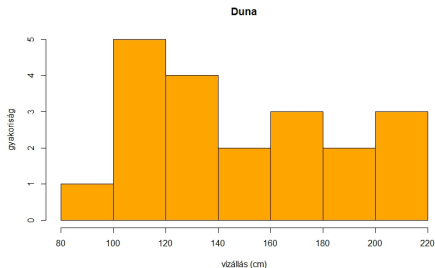
A Duna vízállása 20 napon keresztül (2016. január, adatok forrása: Országos Vízelvezető Szolgálat). Ez **nem független** minta.

Példa: hisztogram

A Duna vízállásának hisztogramja



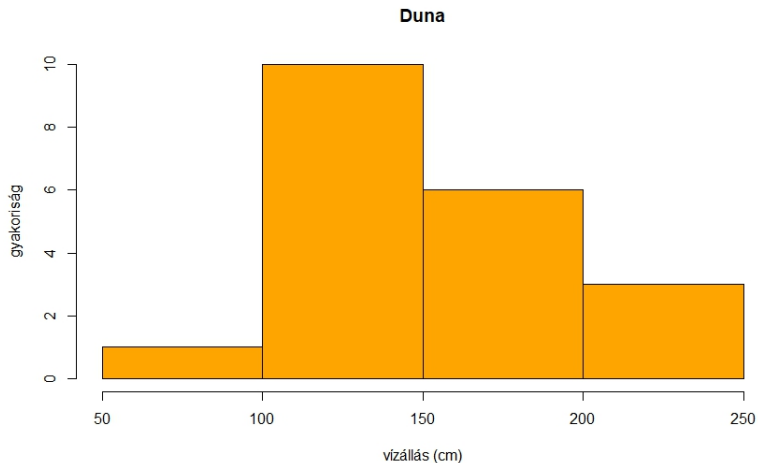
Példa: hisztogram



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.

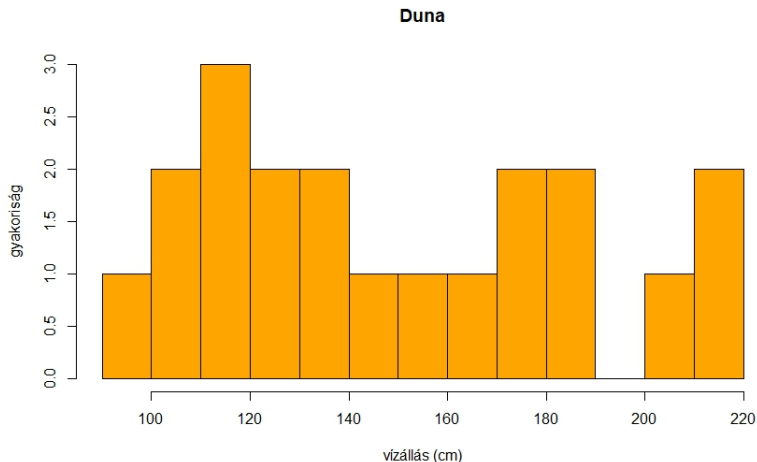
Emlékeztető: megfelelő választás és abszolút folytonos valószínűségi változó esetén a hisztogram a **sűrűségfüggvényhez** közelít.

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)
```

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

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

Mintaelemszám: n

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

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.

Matematikai statisztika

A mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk.

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. A cél a valószínűségi változók eloszlásának a becslése, rá vonatkozó hipotézisek eldöntése a megfigyelések, vagyis az adatok alapján.

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.

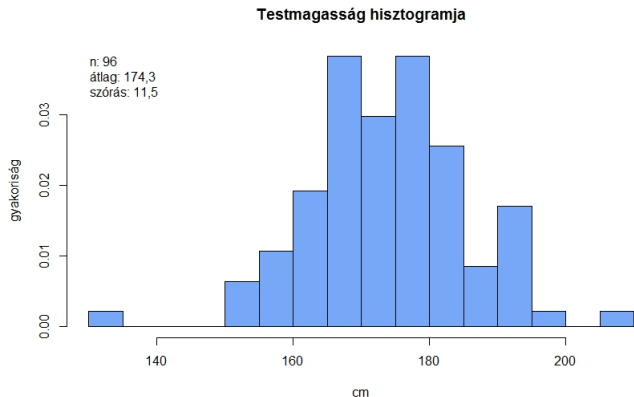
Definíció (Tapasztalati eloszlásfüggvény)

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

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

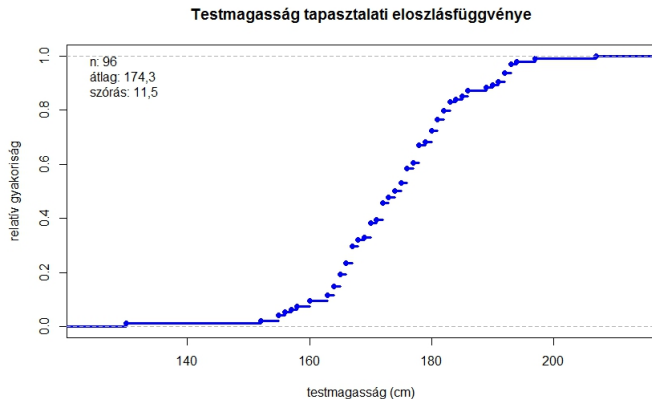
(empirical cumulative distribution function)

Testmagasság hisztogramja



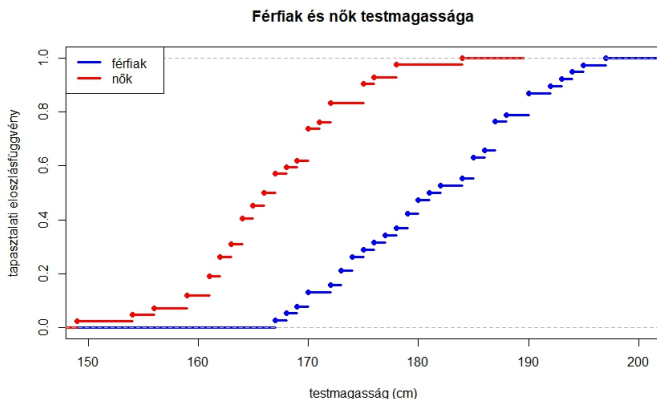
A testmagasság hisztogramja $n = 96$ elemű mintából.

Tapasztalati eloszlásfüggvény



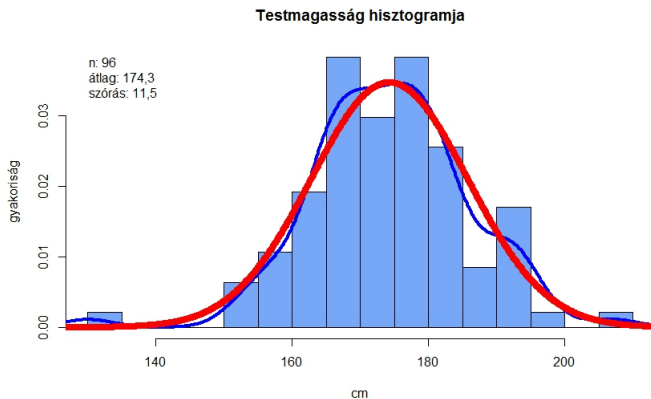
A testmagasság tapasztalati eloszlásfüggvénye $n = 96$ elemű mintából.

Tapasztalati eloszlásfüggvény



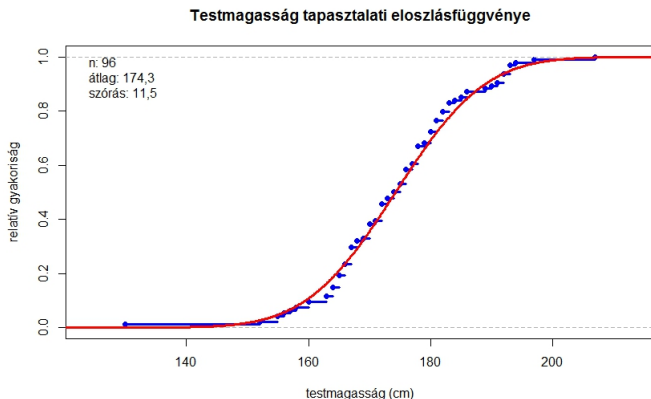
A testmagasság tapasztalati eloszlásfüggvénye $n = 96$ elemű mintából külön a férfiak (kék) és a nők (piros) esetében.

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

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

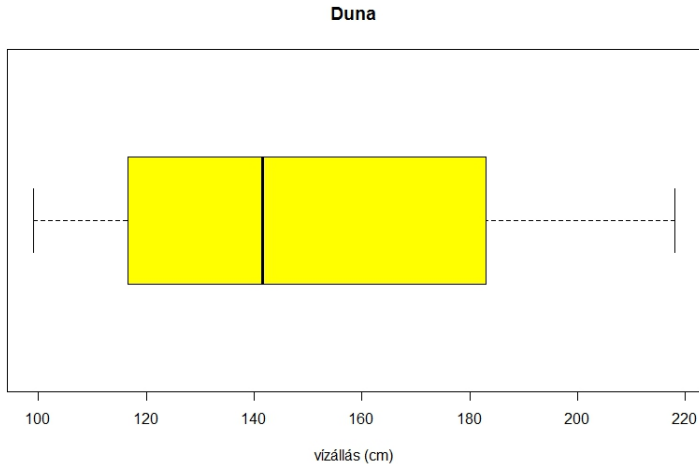
A **nagy számok erős törvénye** szerint **minden t valós számra**

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

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

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 $\mathbb{P}(X \leq q) = F(q) \geq z$.

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:

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

Kvantilisek

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

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.

Ehhez nagyjából azt kell megnézni, hogy a tapasztalati eloszlásfüggvény hol éri el z -t.

Boxplot

Definíció (Tapasztalati kvantilis)

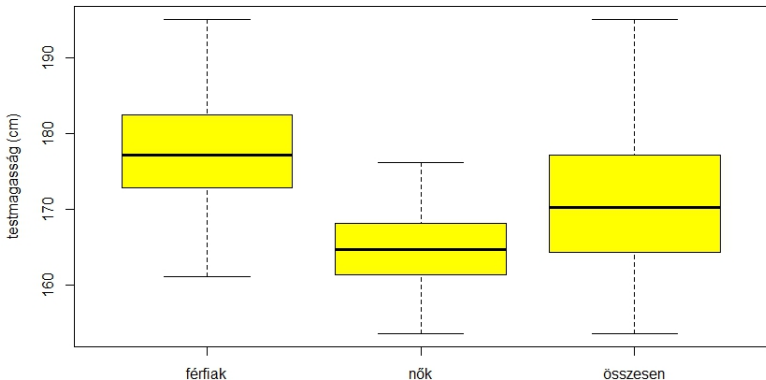
Legyen X_1, X_2, \dots, X_n minta, és $z \in [0, 1]$ adott szám. Ekkor a minta tapasztalati z -kvantilise:

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

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).

Példa: boxplot



Testmagasság adatok boxplotja $n = 96$ elemű mintából külön a férfiak és nők esetében, illetve összesítve

Középértékek: medián

Minta: (X_1, X_2, \dots, X_n) , mintaelemszám: n .

Definíció (medián)

Ha n páratlan: a rendezett minta középső, $(n+1)/2$. elemét, azaz $X_{(n+1)/2}^*$ -t a minta **mediánjának** nevezzük.

Ha n páros: a rendezett minta $n/2$. és $n/2 + 1$. elemének átlagát, azaz a

$$\frac{X_{n/2}^* + X_{n/2+1}^*}{2}$$

mennyiséget a minta mediánjának nevezzük.

Megjegyzés: páros n esetén a teljes $[X_{n/2}^*, X_{n/2+1}^*]$ intervallumot (vagy annak bármely elemét) is a minta mediánjának lehet hívni.

Példa: a Duna vízállásáról kapott húszelemű minta mediánja:

$$\frac{1}{2}(X_{10}^* + X_{11}^*) = \frac{1}{2}(135 + 148) = 141,5.$$

Középértékek: az átlag és a medián összehasonlítása

Normális eloszlás

500 elemű független minta: X_1, X_2, \dots, X_{500} függetlenek, eloszlásuk normális eloszlás $m = 1$ várható értékkel és $\sigma = 1$ szórással

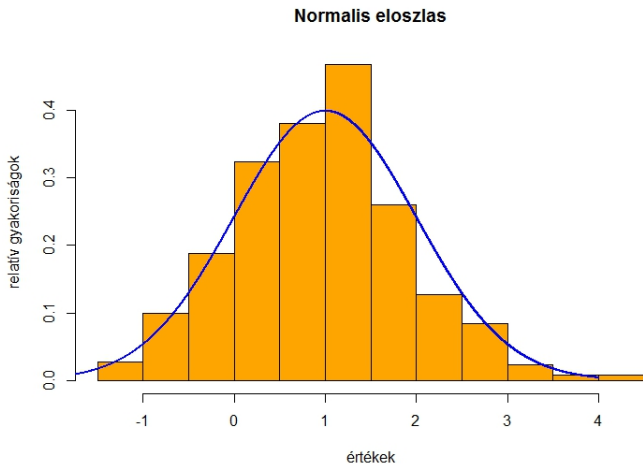
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.9840	0.2847	0.9842	0.9863	1.6930	3.6110

Exponenciális eloszlás

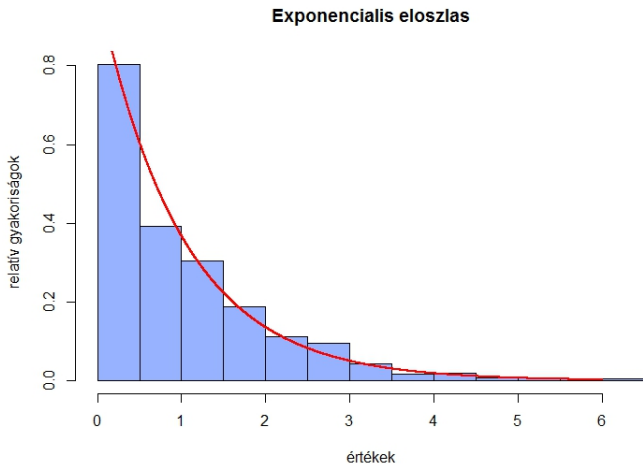
500 elemű független minta: Y_1, Y_2, \dots, Y_{500} függetlenek, eloszlásuk exponenciális eloszlás $b = 1$ paraméterrel. $\mathbb{E}(Y_k) = 1$ és $D(Y_k) = 1$ minden $k = 1, 2, \dots, 500$ -ra.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001326	0.282700	0.637300	0.984900	1.349000	5.895000

A normális eloszlású minta hisztogramja



Az exponenciális eloszlású minta hisztogramja



Az átlag és a medián összehasonlítása

Az átlag

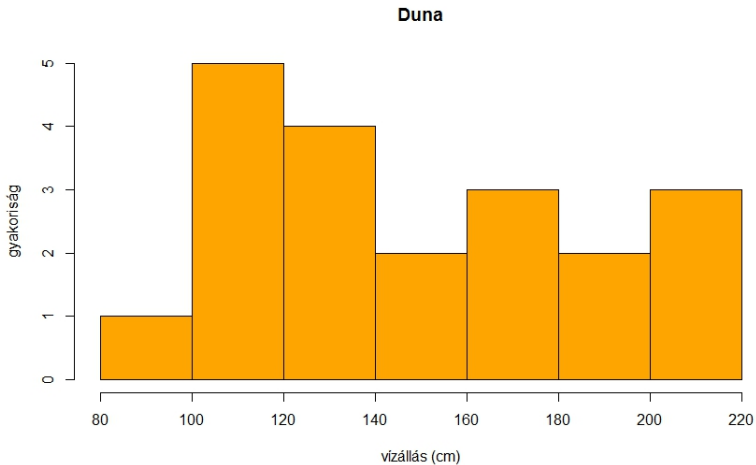
- "több információt használ"
- érzékenyebb a kiugró adatokra, azaz egy hibás mérés is könnyen megváltoztathatja
- nem szimmetrikus esetben eltérhet a leggyakrabban megfigyelt értékektől

A mediánt is érdemes használni, ha

- vannak kiugró (esetleg hibás) adatok;
- ha az eloszlás nem szimmetrikus, és az átlag és a medián jelentősen különbözik (mint a fenti példában az exponenciális eloszlás esetén).

Példa: hisztogram

A Duna vízállásának hisztogramja (medián: 141,5)



Középértékek közelítése osztályközös gyakoriságokkal

Tegyük fel, hogy az adatokat nem ismerjük pontosan, csak a hisztogramot, vagyis hogy az egyes osztályokba, intervallumokba hány megfigyelés esik. Legyen x_j a j . osztályközép (az alsó és felső határ átlaga), és f_j a j . osztályba eső megfigyelések száma, továbbá $n = f_1 + f_2 + \dots + f_k$ az összes megfigyelés száma. Ekkor

- az átlag közelítése:

$$\frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{n};$$

Középértékek közelítése osztályközös gyakoriságokkal

Tegyük fel, hogy az adatokat nem ismerjük pontosan, csak a hisztogramot, vagyis hogy az egyes osztályokba, intervallumokba hány megfigyelés esik. Legyen x_j a j . osztályközép (az alsó és felső határ átlaga), és f_j a j . osztályba eső megfigyelések száma, továbbá $n = f_1 + f_2 + \dots + f_k$ az összes megfigyelés száma. Ekkor

- az átlag közelítése:

$$\frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{n};$$

- a medián közelítése:

$$t_{\text{me}} + \frac{n/2 - F_{\text{me}-1}}{f_{\text{me}}} \cdot h_{\text{me}},$$

ahol t_{me} a mediánt tartalmazó osztály alsó határa, $F_{\text{me}-1}$ a mediánt tartalmazó osztályt megelőző osztályok gyakoriságainak összege, f_{me} a mediánt tartalmazó osztály gyakorisága, h_{me} a mediánt tartalmazó osztály szélessége.

Házi feladat február 19., 8:30-ig

Kérdezzünk meg legalább huszonöt felnőtt ismerőst a

- a nemükről;
- arról, hogy hány filmsorozatot (vagy tévésorozatot) néztek rendszeresen az elmúlt egy hónapban (az nem kell, hogy melyik sorozatokat)
- arról, hogy egy hétköznapon átlagosan hány percet töltenek közlekedéssel.

Az adatokra az egész félév során szükség lesz a házi feladatoknál, azzal együtt, hogy melyik válaszok tartoznak össze.

Tekintsük a közlekedéssel töltött időket. Készítsünk (a) hisztogramot; (b) boxplotot; (c) ábrázoljuk a tapasztalati eloszlásfüggvényt külön a férfiak és a nők esetében, illetve összesítve. Milyen következtetéseket vonhatunk le a kapott ábrákról?