

Matematikai statisztika előadás

Bevezetés, leíró statisztika

1. A kurzus célja és ajánlott irodalom

- mérési eredmények, megfigyelések elemzése (leíró statisztika)
- ismeretlen paraméterek becslése (matematikai statisztika, becslésmélett)
- 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 a matematikai statisztika főbb módszereinek (például becslésméleti, hipotézisvizsgálati módszerek) és azok matematikai hátterének bemutatása, az alkalmazási készség elsajátítása.

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

2. Statisztikai elemzés

A statisztikai elemzés szempontjából fontos megkülönböztetni a vizsgálni kívánt csoportokat, egyedeket, illetve a mérésekből származó információt.

- **populáció:** azon egyedek összessége, akikről 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

2.1. Statisztikai adatok

Miután a mérések, mintavételezés során összegyűjtöttük az adatokat, ezekből további számokat, mennyiségeket határozhatunk meg.

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: $\frac{|X - \vartheta|}{X}$.
Másik változat: a valódi értékkel osztunk: $\frac{|X - \vartheta|}{\vartheta}$

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

Hasonlóképpen, ha egy statisztikai eljárással egy ismeretlen ϑ mennyiséget az általunk a mintából kiszámított X mennyiséggel becsülünk, ugyanígy értelmezhetjük az abszolút, illetve relatív hiba fogalmát.

2.2. 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).

Az adatok alábbi típusait (skáláját) különböztethetjük meg. Ettől függ, hogy milyen statisztikai módszereket alkalmazhatunk egy adott feladatban.

- **nominális:** minőségi ismerv, csak az egyes ismervváltozatok gyakoriságát tudjuk megszámlálni (pl. nem, foglalkozás, nemzetiség)
- **ordinális:** egyértelmű sorrendbe rendezhető változatokkal rendelkező ismerv (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 ismerv egy valós számmal jellemezhető, melyek különbsége és hányadosa is egyértelmű (pl. jövedelem, tömeg, csapadékmennyiség)

Például t -próba minőségi vagy ordinális ismerv esetén nem végezhető, ott fontos, hogy az adatok számszerűsíthetőek legyenek. Viszont például többféle χ^2 -próba végezhető bizonyos fajta nominális adatok esetén (például annak ellenőrzésére, hogy a nemzetiség és a foglalkozás független-e egymástól).

2.3. Matematikai statisztika

Az adatok feldolgozása során többféle matematikai módszert is alkalmazhatunk. Mindezek során azt feltételezzük, hogy az adataink mérések véletlen eredményeként álltak elő, és ezeknek a véletlen mennyiségeknek, valószínűségi változóknak az eloszlását szeretnénk minél jobban megismerni, majd ezekből következtetéseket levonni.

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 ősszel születtek. Állíthatjuk-e az adatok alapján, hogy a születések eloszlása szignifikánsan eltér az egyenletes eloszlástól (amikor minden évszaknak $1/4$ a 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?

A matematikai statisztika alapfeltevése, hogy a mintavétel eredményeként kapott adatok véletlenek: véletlenszerűen választjuk a megkérdezetteket, mérési hibát követünk el stb. A kísérlet megismétlésénél más eredményeket kapnánk. Ezt a valószínűségszámítás fogalmaival a következőképpen tudjuk leírni, elsősorban arány típusú adatok esetén.

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

Minta elemszám: 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

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

a megfigyelések, vagyis az adatok alapján.

3. Leíró statisztika

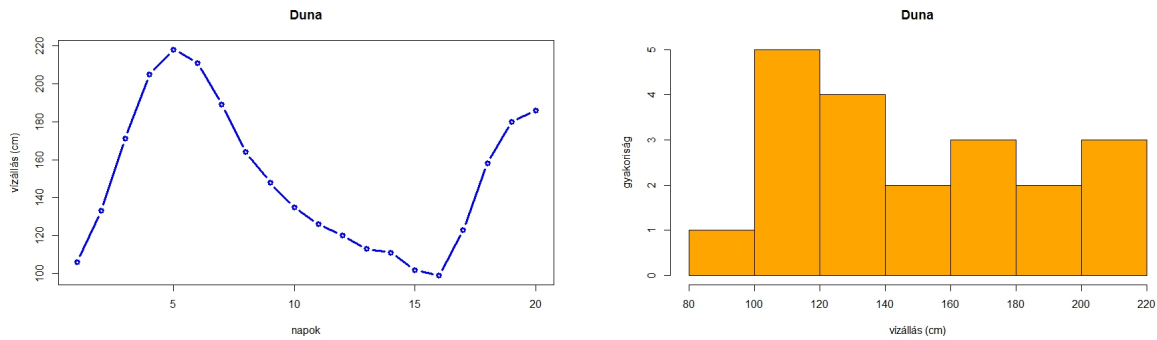
A leíró statisztika módszereinek (a matematikai statisztikával ellentétben) nem a véletlen hatásának megértése, az eloszlások megismerése a célja, hanem a megfigyelt adatok **megjelenítése, jellemzőinek kiszámítása**. Ide tartozhat:

- diagramok: kördiagram, oszlopdiagram, 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

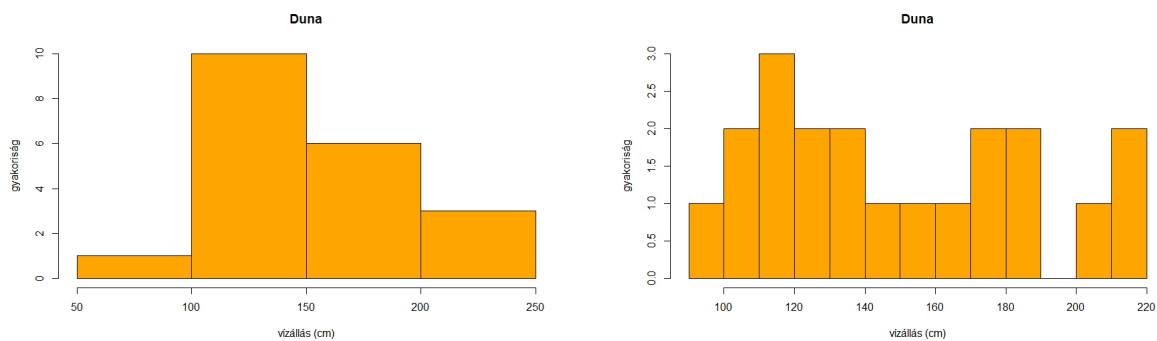
3.1. Példák: az adatok ábrázolása

A következő néhány ábrán ugyanannak az adatsornak többféle ábrázolási módját figyelhetjük meg. Vegyük észre, hogy

- ezek arány típusú adatok, valós számokkal jellemezhetők;
- a méréseket megfeleltethetnénk valószínűségi változóknak, sőt ha X_1, X_2, \dots, X_{20} az egyes napokon mért értékek, akkor $(X_1, X_2, \dots, X_{20})$ egy valószínűségi vektorváltozó;
- az X_1, X_2, \dots, X_{20} valószínűségi változók, vagyis a mért értékek nem függetlenek egymástól, ez fontos lehet, ha az adatok feldolgozására matematikai statisztikai módszereket választunk (például egy egyszerű t -próba nem lenne jó annak eldöntésére, hogy a vízállás várható értéke több-e 250 cm-nél). Az adatok ábrázolása, mint leíró statisztika módszer ekkor is rendelkezésre áll.



1. ábra. A Duna vízállása 20 napon keresztül: adatok és hisztogram (2016. január, adatok forrása: Országos Vízelző Szolgálat)



2. ábra. Hisztogram a Duna vízállásának adataiból, különböző intervallumhosszakkal

Hisztogram készítése: választunk egy intervallumot, mely magában foglalja a mérési adatokat. Az intervallumot egyenlő nagyságú részekre osztjuk. Az egyes kis intervallumokba eső mérési adatok számát ábrázoljuk (ezt gyakran inkább oszlopdiagramnak nevezik), vagy úgy készítjük el az oszlopokat, hogy a magasságuk arányos legyen a gyakoriságokkal, az összterület azonban 1 legyen.

Emlékeztető: megfelelő választás és abszolút folytonos valószínűségi változó esetén a hisztogram a [sűrűségfüggvényhez](#) közelít. Később látni fogjuk, hogy a hisztogramhoz hasonló objektumok használhatók a sűrűségfüggvény becslésére is.

Sem a túl hosszú, sem a túl rövid intervallumok nem adnak informatív ábrát.

```
hist(viz, col="orange", xlab="vízállás (cm)", ylab="gyakoriság", main="Duna", breaks=4)
hist(viz, col="orange", xlab="vízállás (cm)", ylab="gyakoriság", main="Duna", breaks=15)
```

3.2. Alapstatisztikák

Az alábbi mennyiségeket mind leíró statisztikában, mind a matematikai statisztikában gyakran használjuk.

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 közép**ső mintaelem, vagy a középső kettő átlaga (ha n páros).
- **módusz** (mode): a leggyakrabban előforduló mintaelem.

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

Ehhez kapcsolódó statisztikák, melyek a várható érték és a szórás becslésére használhatók, illetve néhány további gyakori statisztika:

- **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): $s_n^* = \sqrt{s_n^{*2}}$.
- **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): $\frac{s_n^*}{\sqrt{n}}$

Nézzük meg, hogyan alakulnak ezek a korábban látott adatsor esetében.

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: $8,62$

4. Tapasztalati eloszlásfüggvény, a statisztika alaptétele

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

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

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

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

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

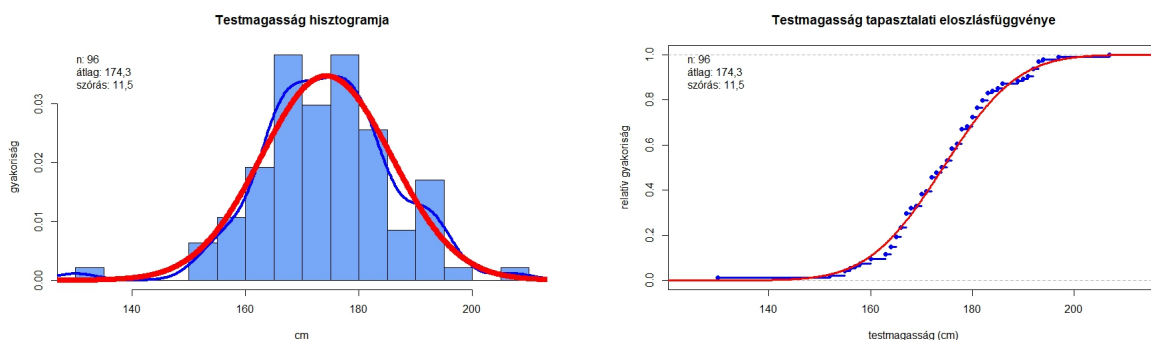
minden $t \in \mathbb{R}$ -re.

Az eloszlásfüggvény becslésénél a valószínűségeket a relatív gyakoriságokkal becsljük, így kapjuk az alábbi definíciót.

4.1. Definíció (Tapasztalati eloszlásfüggvény (empirical cumulative distribution function)).

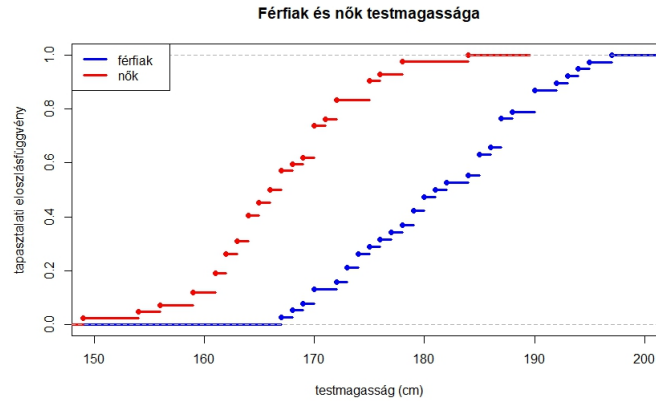
Az X_1, X_2, \dots, X_n minta *tapasztalati eloszlásfüggvénye* az $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

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



3. ábra. Egy $n = 96$ elemű, testmagasság adatsor hisztogramja és tapasztalati eloszlásfüggvénye

Tekintsünk egy példát (3. ábra). Itt egy $n = 96$ elemű, emberek testmagasságából származó adatsorból készítettünk hisztogramot, és tapasztalati eloszlásfüggvényt. Ezen kívül megbecsültük a várható értéket az átlaggal: $\bar{X} = 174,3$, a szórást pedig a korrigált tapasztalati szórással, és ábráztuk annak a normális eloszlásnak az eloszlásfüggvényét (balra), illetve sűrűségfüggvényét (jobbra, pirossal) is, melynek éppen ezek a becslt értékek a paraméterei. Az ábra alapján azt láthatjuk, hogy a becslt normális eloszlás jól illeszkedik a megfigyelésekre (ennek pontosítására a Kolmogorov–Szmirnov-próba használható). A 4. ábrán pedig azt látjuk, hogyan lehet összehasonlítani ugyanannak a mennyiségnek a viselkedését két különböző csoportban a tapasztalati



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

eloszlásfüggvény segítségével (emlékeztetőül: $F(t) = \mathbb{P}(X \leq t)$, így minél nagyobb F , annál nagyobb valószínűséggel vesz fel X "kicsi" értékeket, nevezetesen t -nél kisebbet).

```

nosorozat<-c(1, 0, 3, 2, 0, 10, 0, 2, 0, 0, 1, 2, 1, 2, 0, 2, 1)
noutazas<-c(0, 45, 120, 120, 60, 60, 30, 60, 100, 70, 180, 40, 60, 100, 80, 90, 120)
ferfiutazas<-c(60, 30, 70, 60, 20, 60, 10, 120, 60, 120, 130)
ferfisorozat<-c(1, 0, 1, 5, 2, 2, 2, 0, 0, 0, 1)
utazas<-c(ferfiutazas, noutazas)

hist(utazas, col="#79a7f2", xlab="Utazással töltött ido (perc)",
ylab="Relatív gyakoriságok", main="Utazással töltött ido a férfiak esetében")

nosorozat<-c(1, 0, 3, 2, 0, 10, 0, 2, 0, 0, 1, 2, 1, 2, 0, 2, 1)
noutazas<-c(0, 45, 120, 120, 60, 60, 30, 60, 100, 70, 180, 40, 60, 100, 80, 90, 120)
ferfiutazas<-c(60, 30, 70, 60, 20, 60, 10, 120, 60, 120, 130)
ferfisorozat<-c(1, 0, 1, 5, 2, 2, 2, 0, 0, 0, 1)
utazas<-c(ferfiutazas, noutazas)

plot(ecdf(noutazas), lwd='5', col='red', main='Utazással töltött ido tapasztalati
eloszlásfüggvénye', xlab="ido (perc)", ylab="tapasztalati eloszlásfüggvény")
lines(ecdf(ferfiutazas), lwd="5", col="blue")

legend("topleft", c("nok", "férfiak"), col=c("red", "blue"), lwd="3")

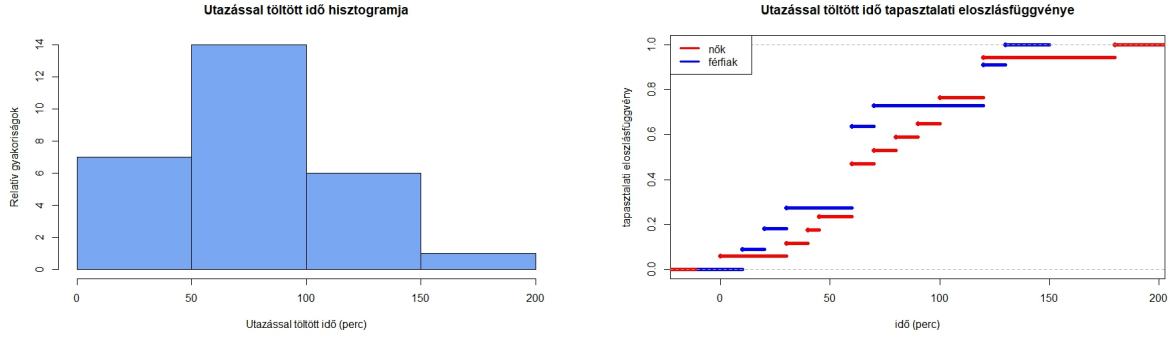
```

4.1. A statisztika alaptétele (Glivenko–Cantelli-tétel)

A statisztika alaptétele célja annak megfogalmazása, hogy ha független azonos eloszlású minta esetén a minta méretével (a mérések számával) végtelenhez tartunk, akkor a minta eloszlását végül "tökéletesen" megismerhetjük: a tapasztalati eloszlásfüggvény határértéke az "igazi" eloszlásfüggvény, vagyis a megfigyelt valószínűségi változók közös eloszlásfüggvénye lesz.

Emlékeztetőül: 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



5. ábra. Hétköznap utazással töltött idő histogramja a teljes mintára ($n = 28$, illetve tapasztalati eloszlásfüggvény külön ($n_1 = 17$ nő, $n_2 = 11$ férfi), 2020. februári adatok

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 nagy számok erős törvénye szerint tehát minden rögzített t esetén az $\hat{F}_n(t)$ tapasztalati eloszlásfüggvény, vagyis az eloszlásfüggvény t -ben felvett értékének becslése 1 valószínűséggel tart $F(t)$ -hez, amit becsülni szeretnénk. Az alábbi tétel ennél abban az értelemben erősebb, hogy nem minden t -re külön-külön állítja a konvergenciát, hanem azt mondja, hogy a tapasztalati és "igazi" eloszlásfüggvény legnagyobb különbsége is nullához tart 1 valószínűséggel. Tehát 1 valószínűséggel igaz az, hogy ha $\varepsilon > 0$ adott és n elég nagy, akkor bármilyen t -re legfeljebb ε -t tévedünk annak valószínűségének becslésekor, hogy a valószínűségi változó értéke legfeljebb t .

4.1. 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 tételre úgy is gondolhatunk, hogy ha tudnánk a testmagasság valódi eloszlásfüggvényét, és azt ábrázonánk együtt a tapasztalati eloszlásfüggvénnyel, akkor a 3. ábra jobb oldali részéhez hasonló ábrát kapnánk.

5. Kvantilisek, boxplot

Ahogy láttuk, a tapasztalati eloszlásfüggvény a mintaelemszám növekedésével a valódi eloszlásfüggvényhez tart (Glivenko–Cantelli-tétel). Gyakran azonban nem közvetlenül az eloszlásfüggvényre, vagyis a $\mathbb{P}(X \leq t)$ valószínűsége vagyunk kíváncsiak (ahol most az X megfigyelt mennyiség eloszlása ismeretlen), hanem a kvantilisekre vagyunk kíváncsiak, azaz arra, hogy adott z esetén mi lehet az a q , amire $\mathbb{P}(X \leq q) = z$ teljesül. Például ha X egy folyó legnagyobb vízállása egy évben, és $z = 0,95\%$, akkor q mondja meg, hogy mi az a legnagyobb vízállás, aminél a folyó csak 5%-kal megy magasabbra – ha ilyen magas gátat építünk, az 95% valószínűséggel megfelelő lesz. Vagy, ha X a szükséges kórházi ágyak számának maximuma egy városban egy év alatt, és q az eloszlás z -kvantilise $z = 95\%$ -kal, akkor q kórházi ágy 95% valószínűséggel lesz elég (a példában nem számolva azzal, hogy nem minden beteget tudnak bármelyik egységben megfelelően ellátni).

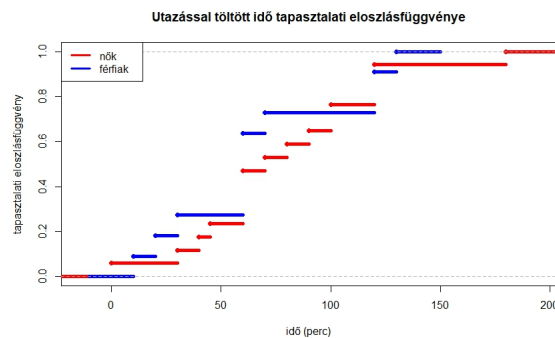
Emlékeztetőül: az X valószínűségi változó z -kvantilise a legkisebb olyan q szám, melyre teljesül, hogy $\mathbb{P}(X \leq q) = F(q) \geq z$.

Kérdés, hogy a kvantiliseket hogyan tudjuk a megfigyelt X_1, X_2, \dots, X_n mintából becsülni. Itt tehát az a feltételezésünk, hogy az X_1, X_2, \dots, X_n valószínűségi változók függetlenek, azonos eloszlásúak, azonban ezt az eloszlást nem ismerjük.

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

5.1. 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:

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



6. ábra. Hétköznap utazással töltött idő histogramja a teljes mintára ($n = 28$, illetve tapasztalati eloszlásfüggvény külön ($n_1 = 17$ nő, $n_2 = 11$ férfi), 2020. februári adatok

Ehhez nagyjából azt kell megnézni, hogy a tapasztalati eloszlásfüggvény hol éri el z -t, mivel pedig tipikusan z két felvett érték közé esik, azoknak az értékeknek a lineáris kombinációját vesszük, amiknél még éppen kisebb, illetve már nagyobb z -nél a tapasztalati eloszlásfüggvény. Például az 6. ábra alapján keressük meg a férfiak utazási idejének 40%-os kvantilisét, vagyis azt az értéket, ami azt a q időt becsüli, aminél a férfiak 40%-a tölt kevesebbet utazással. Látjuk, hogy a 60 gyakran szerepel, és 59 percnél kevesebb időt a mintában a férfiak kevesebb, mint 30%-a szán utazásra, 61 percnél kevesebb időt pedig több, mint a 60 százalékuk. Ez alapján mondhatnánk a 60-at is becslésnek, hiszen a tapasztalati eloszlás függvény itt lépi át a 40%-ot, és az R ezt is adja vissza:

```
> ferfi<-c(60, 30, 70, 60, 20, 60, 10, 120, 60, 120, 130)
> quantile(ferfi, probs=c(0.4, 0.8))
40% 80%
60 120
> boxplot(ferfi)
```

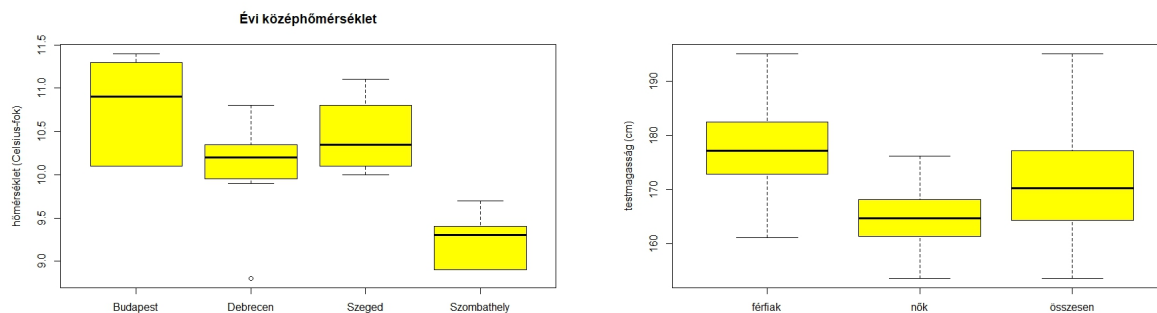
Ha pedig a fenti képletet alkalmazzuk (bár az R nem pontosan ezt használja): $z = 0,4$ és $n = 11$, így $z(n+1) = 4,4$ alsó egészrésze: $\lfloor z(n+1) \rfloor = 4$. Ez alapján

$$\hat{q}_z = X_4^* + 0,4 \cdot (X_5^* - X_4^*) = 60 + 0,4 \cdot (60 - 60) = 60,$$

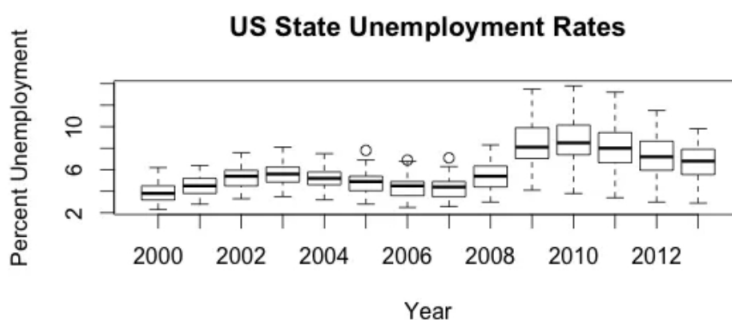
hiszen a rendezett minta 4. és 5. eleme is 60-nal egyenlő.

A 80%-os kvantilist is hasonlóképpen számolhatnánk ki.

A kvantilisek közül az alábbiak gyakran előfordulnak, többek között a boxplot ábrán :



7. ábra. Boxplot ábra négy város éves középhőmérséklet adataiból (forrás: Országos Meteorológiai Szolgálat), illetve testmagasság adatok boxplotja $n = 96$ elemű mintából külön a férfiak és nők esetében és összesítve



8. ábra. Az USA államaiban mért munkanélküliségi ráta boxplotja (forrás: <https://www.r-bloggers.com/2015/11/free-webinar-learn-to-map-unemployment-data-in-r/>)

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

Ahogy a 7. ábrán láthatjuk, a boxplot ábra segítségével több adatsort hasonlíthatunk össze.

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).

Az egyes dobozok határait az első és a harmadik kvartilis adja meg, a középső vonal a medián, a vonalak pedig a legkisebb, illetve legnagyobb mintaelemig tartanak.

Megállapíthatjuk például, hogy Szombathelyen még a maximum is kisebb volt, mint a másik három városban bármelyik megfigyelés, vagy hogy Szegeden a megfigyelések negyede kb. 10,1 és 10,3 fok közé esett (ez az első kvartilis és a medián közötti tartomány). A jobb oldali ábráról pedig azt vehetjük például észre, hogy a mintában szereplő legalacsonyabb férfnál a nők nagyjából negyedrésze alacsonyabb.

6. Középértékek

A mintát, különösen, ha más adatsorokkal akarjuk összehasonlítani vagy az időbeli változást figyeljük, gyakran csak egy, rá jellemző számmal, középértékkel jelenítjük meg. Erre is több lehetőség van, a következőkben azt nézzük meg, hogy a két leggyakoribb középérték egymáshoz viszonyítva hogyan viselkedik.

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

6.1. 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

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

menyiséget 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.

6.1. Az átlag és a medián összehasonlítása

Normális eloszlás

Tekintsünk egy 500 elemű független mintát: 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

Vegyünk egy másik, 500 elemű független mintát is: 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.

```
> exp=rexp(n=500, rate=1)
```

```
> hist(exp, col="blue", main="Exponencialis eloszlas", xlab="ertekek", ylab="gyakorisagok")
```

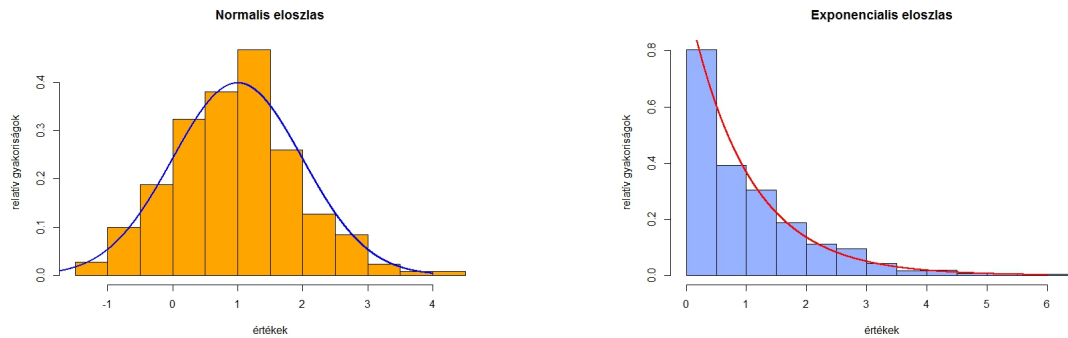
```
> summary(exp)
```

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

A két mintát összehasonlítva azt vehetjük észre, hogy a normális eloszlású, szimmetrikusabb esetben az átlag és a medián értéke majdnem megegyezik, lényegében mindegy, hogy melyiket használjuk. Az exponenciális eloszlás sűrűségfüggvénye viszont nem szimmetrikus, ilyenkor az átlag és a medián eltér, ilyenkor érdemes lehet mindkettőt feltüntetni, ha pedig az aszimmetriát többek között kiugró, részben hibásnak vélt mérések okozzák, akkor az átlag helyett a mediánt használni.

Az átlag

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



9. ábra. A normális, illetve exponenciális eloszlású minták hisztogramja

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).

6.2. Közéértékek 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:

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

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

7. Indexek számítása

Indexeket különböző mennyiségek összehasonlítására, hatások összehasonlítására, vagy közvetlenül nem összemérhető mennyiségek változásának leírására szoktak használni (például: a GDP közvetlenül nem összehasonlítható mennyiségek keveréke).

A leíró statisztikának ebből a témaköréből egy példát nézünk meg alaposabban.

Tegyük fel, hogy egy időben változó mennyiség egy időszakban (tárgyidőszakban) mért értékeit szeretnénk egy korábbi, hasonló időszakban (bázisidőszakban) mért értékekkel összehasonlítani, hogy az átlagos változást leírhatjuk. Például tekinthetjük a fogyasztói árindexet (például: https://www.ksh.hu/stadat_files/ara/hu/ara0040.html vagy http://www.ksh.hu/interaktiv/fogyar_radar/index.html), ami az infláció mérőszáma, a lakosság által vásárolt termékek és szolgáltatások árainak átlagos változását fejezi ki. A bázisidőszak lehet az „előző” év, a tárgyidőszak a vizsgált év.

Tegyük fel, hogy az árindexbe az $1, 2, \dots, n$ termékek forgalmát építik be. Legyen

- $q_{0,j}$ a j . termékből eladott mennyiség a bázisidőszakban;
- $q_{1,j}$ a j . termékből eladott mennyiség a tárgyidőszakban;
- $p_{0,j}$ a j . termék egységára a bázisidőszakban;
- $p_{1,j}$ a j . termék egységára a tárgyidőszakban.

Ekkor

- bázisidőszaki súlyozású vagy Laspeyres-féle árindex (annak hányadosa, hogy az új árakkal, de a bázisidőszak fogyasztásával mennyivel nőtt az összes kiadás a régebbi időszakhoz képest):

$$\frac{\sum_{j=1}^n q_{0,j} p_{1,j}}{\sum_{j=1}^n q_{0,j} p_{0,j}}$$

- tárgyidőszaki súlyozású vagy Paasche-féle árindex (annak hányadosa, hogy az új árakkal és az új fogyasztással mennyivel nőtt az összes kiadás):

$$\frac{\sum_{j=1}^n q_{1,j} p_{1,j}}{\sum_{j=1}^n q_{1,j} p_{0,j}}$$

- bázisidőszaki súlyozású vagy Laspeyres-féle volumenindex (a régi árakkal számolva hányszorosára nőtt az összes kiadás, vagyis a régi árakkal számolva mennyivel nőtt a fogyasztás):

$$\frac{\sum_{j=1}^n q_{1,j} p_{0,j}}{\sum_{j=1}^n q_{0,j} p_{0,j}}$$

- tárgyidőszaki súlyozású vagy Paasche-féle volumenindex (az új árakkal számolva hányszorosára nőtt az összes kiadás):

$$\frac{\sum_{j=1}^n q_{1,j} p_{1,j}}{\sum_{j=1}^n q_{0,j} p_{1,j}}$$

A témáról részletesebben: https://regi.tankonyvtar.hu/hu/tartalom/tamop425/2011_0001_519_42491/ch05s02.html

8. Sűrűségfüggvény becslése

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 $\mathbb{P}(X_1 \leq t)$, vagyis nem ismerjük az eloszlásfüggvényt, vagy mennyi X_1 várható értéke, szórása.

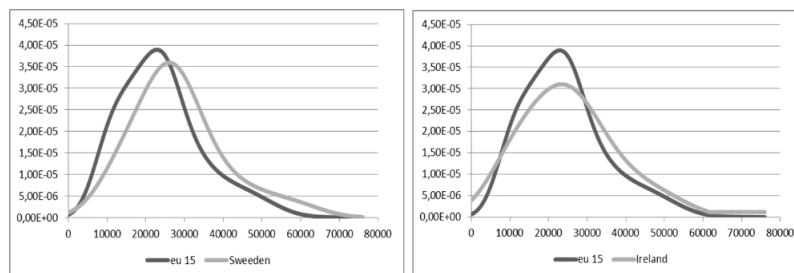
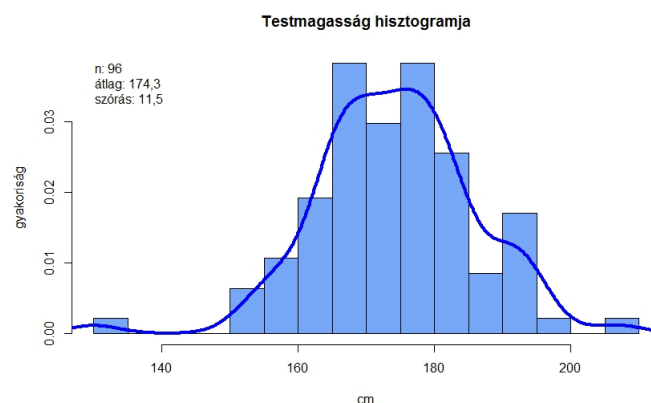


Figure 10: Density Function Estimations of EU Countries and Sweden (left) and Ireland (right) in 2001.

10. ábra. A svédországi és írországi jövedelmek sűrűségfüggvényének becslése egy összetettebb módszerrel, a megfelelő ponton Gauss-magfüggvénnyel (forrás: [1])



11. ábra. A testmagasság histogramja $n = 96$ elemű mintából (valós adatokból), a sűrűségfüggvény becslése Gauss-magfüggvénnyel.

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

Abban az esetben, amikor az ismeretlen eloszlásról feltételezhetjük, hogy abszolút folytonos eloszlású, vagy ilyen módon modellezzük (például nincsenek olyan kitüntetett értékek, amik a valószínűségi változó pozitív valószínűséggel venne fel, amire onnan következtethetünk, hogy a megfigyelések mind vagy majdnem mind különbözőek), az eloszlás sűrűségfüggvényét sem ismerjük, viszont az adatok alapján megpróbálhatjuk megbecsülni.

Ahogy több példán láttuk (akár a 9. ábrán), elég sok megfigyelés esetén a histogram és a sűrűségfüggvény közel esik egymáshoz. Ezt használhatjuk ki a sűrűségfüggvény pontosabb becslésekor, ugyanakkor itt is kérdés, hogy milyen intervallumhosszat használjunk.

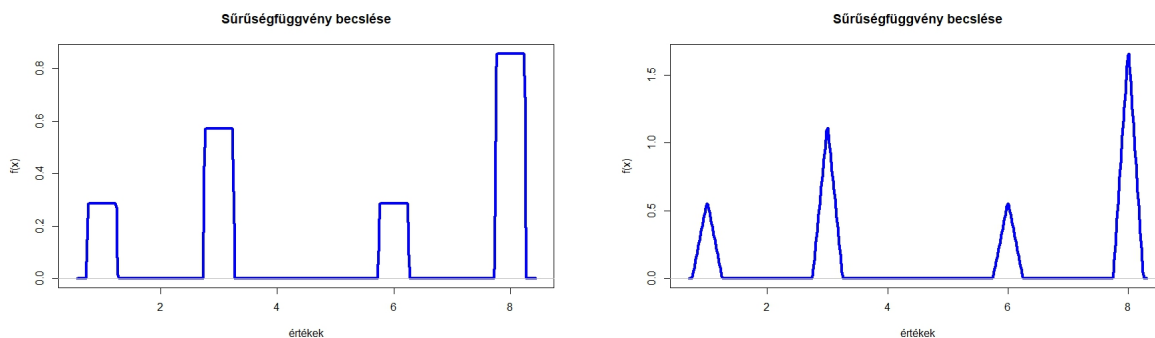
Ahhoz, hogy 1 legyen az alatta lévő terület:

```
> hist(magassag, freq=F)
```

Az alábbiakban a sűrűségfüggvény becslésének Parzen–Rosenblatt-féle módszerét ismertetjük.

X_1, X_2, \dots, X_n független azonos eloszlású abszolút folytonos minta. A sűrűségfüggvény f , azaz

$$\mathbb{P}(a \leq X_1 \leq b) = \int_a^b f(t)dt \quad \text{minden } a < b\text{-re.}$$



12. ábra. A sűrűségfüggvény becslése téglalapos és háromszöges magfüggvénnyel az 1, 3, 3, 6, 8, 8, 8 mintából

Az f függvény ismeretlen. Hogyan tudjuk $f(t)$ értékét becsülni az X_1, \dots, X_n megfigyelések segítségével?

Ehhez az alábbiakból indulhatunk ki:

$$\mathbb{P}(a \leq X_1 \leq b) = \int_a^b f(t) dt \approx \frac{1}{n} \sum_{j=1}^n \mathbb{I}(a < X_j \leq b),$$

azaz a becslés a és b közé eső mintaelemek aránya, azaz annak relatív gyakorisága, hogy a megfigyelés az $(a, b]$ -be esik.

Az indikátorfüggvény értéke 1, ha $a < X_j \leq b$, és 0 különben. Annyi darab egyest adunk össze, ahány X_j esik bele az $(a, b]$ intervallumba. Így kapjuk a relatív gyakoriságot.

A jobb oldalon éppen az $(a, b]$ intervallumba eső mintaelemek relatív gyakorisága szerepel, ez hasonló, mint ami a hisztogramban is szerepel. Ez az alábbi definícióhoz vezet.

8.1. A sűrűségfüggvény becslése különböző magfüggvényekkel

A 12. ábra bal oldalán az alábbi mintából készített becslést láthatjuk (X_1, \dots, X_7): 1, 3, 3, 6, 8, 8, 8. Minden oszlop közepe egy megfigyelés, a magasság a gyakoriságtól függ, az oszlop szélessége pedig $2h$, ahol h az ablakszélesség, ez választható a becslés során.

Ezt általánosabban az alábbi módon írhatjuk fel.

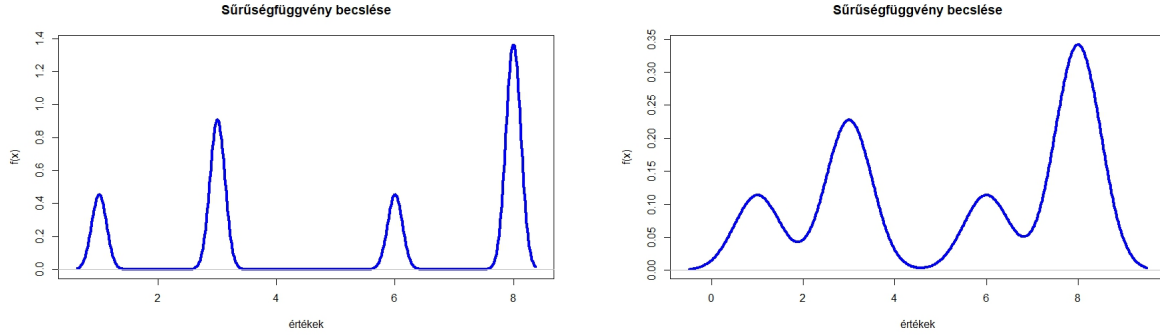
Téglalap magfüggvény: $k(y) = 1/2$, ha $-1 \leq y \leq 1$, nulla különben, azaz $k(y) = \frac{1}{2} \mathbb{I}(|y| \leq 1)$ és h az **ablakszélesség**.

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{j=1}^n \frac{1}{2} \mathbb{I}(|t - X_j| < h) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right).$$

A téglalapon kívül más alakú függvényeket is szoktak használni a becsléshez, ezt például a 12. ábra jobb oldalán, illetve a 13. ábrákon láthatjuk. Ilyenkor a fenti leírásban k szerepe ugyanaz, csak k -t választjuk másképpen.

Háromszöges magfüggvény: $k(y) = \max(1 - |y|, 0)$ és $h = 1/2$ az ablakszélesség.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right).$$



13. ábra. A sűrűségfüggvény becslése Gauss-féle magfüggvénnyel az 1, 3, 3, 6, 8, 8, 8 mintából $h = 0,5$ és $h = 2$ -es ablakszélességgel

Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ és $h = 1/2$ az ablakszélesség.

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right) = \frac{1}{n \cdot h \cdot \sqrt{2\pi}} \sum_{j=1}^n \exp\left(-\frac{(y - X_j)^2}{2h^2}\right).$$

Minta (X): 1, 3, 3, 6, 8, 8, 8. Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$ és $h = 2$ **ablakszélesség** (sávszélesség, bandwidth).

$$\hat{f}_n(t) = \frac{1}{n \cdot h} \sum_{j=1}^n k\left(\frac{t - X_j}{h}\right) = \frac{1}{n \cdot 2 \cdot \sqrt{2\pi}} \sum_{j=1}^n \exp\left(-\frac{(y - X_j)^2}{2 \cdot 2^2}\right).$$

8.2. A sűrűségfüggvény Parzen–Rosenblatt-féle becslése

A fenti módszer általánosított változata a Parzen–Rosenblatt-féle becslés. Ezt a következőképpen definiálhatjuk:

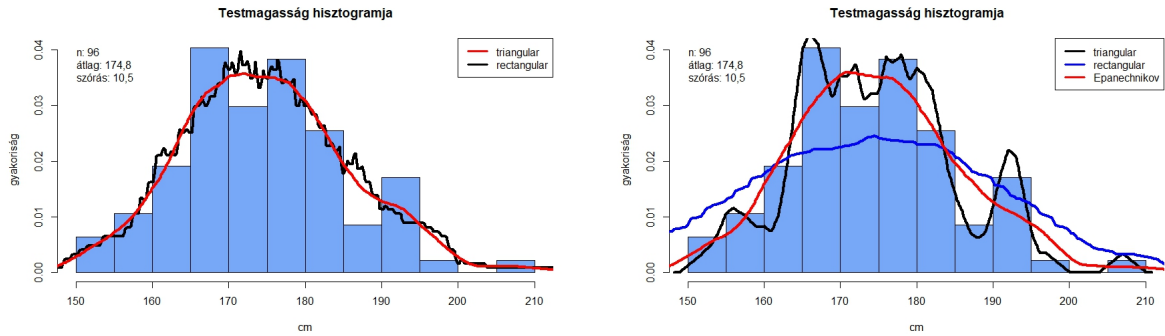
Legyen $k : \mathbb{R} \rightarrow \mathbb{R}_+$ olyan függvény, mely korlátos, $\lim_{y \rightarrow \infty} yk(y) = 0$, továbbá h_n olyan számsorozat, melyre $\lim_{n \rightarrow \infty} h_n = 0$ és $\lim_{n \rightarrow \infty} nh_n = \infty$. A sűrűségfüggvény becslése a t pontban a Parzen–Rosenblatt-módszerrel a k magfüggvénnyel és h_n sávszélességgel az X_1, \dots, X_n független minta alapján:

$$\hat{f}_n(t) = \frac{1}{n \cdot h_n} \sum_{j=1}^n k\left(\frac{t - X_j}{h_n}\right).$$

Itt a k egy standardizált eloszlás sűrűségfüggvénye. Ezt a $t - X_j$ eltolja úgy, hogy X_j legyen a várható értéke (másképpen, az X_j -re legyen szimmetrikus). A skálázás (h_n -nel osztás) pedig úgy állítja be, hogy h_n legyen a szórása.

A mintaelemszám növelésével a fenti módszer határértékben pontos eredményt ad, mindegyik fent bemutatott magfüggvényre. Megfelelő feltételek mellett $\hat{f}_n(t) \rightarrow f(t)$ minden t -re, ha $n \rightarrow \infty$ (szükséges például, hogy f folytonos legyen). Szokásos magfüggvények például (ebből hármát már láttunk):

- Gauss-magfüggvény: $k(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$.
- Háromszög magfüggvény: $k(y) = (1 - |y|)$, ha ez nemnegatív, nulla különben.
- Epanechnikov-magfüggvény: $k(y) = \frac{3}{4}(1 - y^2)$, ha ez nemnegatív, nulla különben.



14. ábra. A testmagasság sűrűségfüggvényének becslése $n = 96$ megfigyelésből; a bal oldalon: háromszöges (piros) és téglalapos (fekete) magfüggvénnyel (ez utóbbihoz túl kicsi a sáv szélesség); a jobb oldalon: háromszöges magfüggvény $1/3$ -szoros sáv szélességgel (fekete), téglalapos magfüggvény 3 -szoros sáv szélességgel (kék), Epanechnikov-magfüggvény alapértelmezett sáv szélességgel (piros)

- Téglalap magfüggvény: $k(y) = 1/2$, ha $-1 \leq y \leq 1$, nulla különben.

A magfüggvények mind olyanok, hogy az alattuk lévő terület 1. Változócserevel (helyettesítéses integrálással) belátható, hogy emiatt

$$\int_{-\infty}^{\infty} \frac{1}{h} k(t/h) dt \stackrel{s=t/h}{=} \int_{-\infty}^{\infty} \frac{1}{h} k(s) \cdot h ds = \int_{-\infty}^{\infty} k(s) ds = 1.$$

A h tényező a helyettesítéses integrálból adódik.

Az eltolás nem változtat. Utána n ilyen összeadunk, de n -nel osztunk is (átlagolunk), így \hat{f}_n integrálja a teljes számegegyenesen is 1.

Szokásos sáv szélesség-választások (normális eloszlás és Gauss-magfüggvény esetén az első optimális), ezekre $h_n \rightarrow 0$, de $nh_n \rightarrow \infty$:

$$h_n = 0,7 \cdot \frac{s_n^*}{n^{1/5}}; \quad h_n = 0,7 \cdot \frac{\min(s_n^*, q)}{n^{1/5}},$$

ahol s_n^* a korrigált tapasztalati szórás, q a harmadik és első kvartilis távolsága.

Ugyanúgy, mint a hisztogramnál, a túl nagy sáv szélesség túl kevés részletes ábrához, a túl kicsi sáv szélesség túl részletes ábrához vezet. Ezt figyelhetjük meg a 14. ábrán: a túl kicsi sáv szélesség esetén a minta esetlegességei is benne maradnak a becslésben, túl nagy sáv szélesség esetén azoknak a tartományoknak túl nagy súlya lesz, ahová csak néhány megfigyelés esik.

Kapcsolódó irodalom: [1, 2]

Házi feladat február 21., hétfő, 10:15-ig Tekintsük az összegyűjtött adatokat, és készítük el a lakóhely és munkahely/egyetem közti távolság sűrűségfüggvény becslését Epanechnikov-magfüggvénnyel

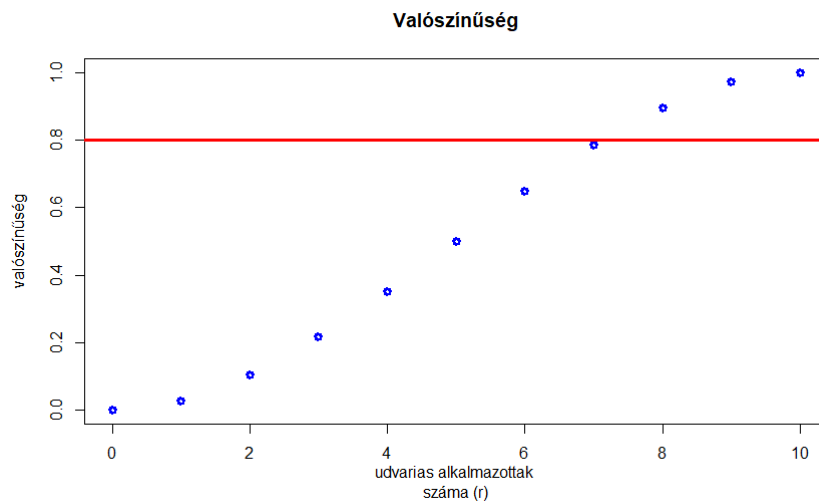
- az összes adatra egyben;
- azokra, akik a mediánnál több vagy ugyanannyi sorozatot néztek (a medián itt a nézett sorozatok számára vonatkozik) az elmúlt két hétben;
- azokra, akik a mediánnál kevesebb vagy ugyanannyi sorozatot néztek az elmúlt két hétben;
- azokra, akik áttestek koronavírus-fertőzésen az elmúlt fél évben.

Hasonlítsuk össze az így kapott görbéket. Milyen következtetéseket vonhatunk le ez alapján? Hasonlít-e a becsült sűrűségfüggvény valamilyen nevezetes eloszlás sűrűségfüggvényére?

R-ben: "density"

Excelből kijelölve az adatokat:

```
> adatok=read.table(file="clipboard", sep="\t", header=TRUE)
> plot(density(adatok$tavolsag, kernel="epanechnikov"), lwd="3", col="blue",
main="Utazási ido surusegfuggvenye", xlab="ido (perc)",
ylab="becsült surusegfuggveny")
```

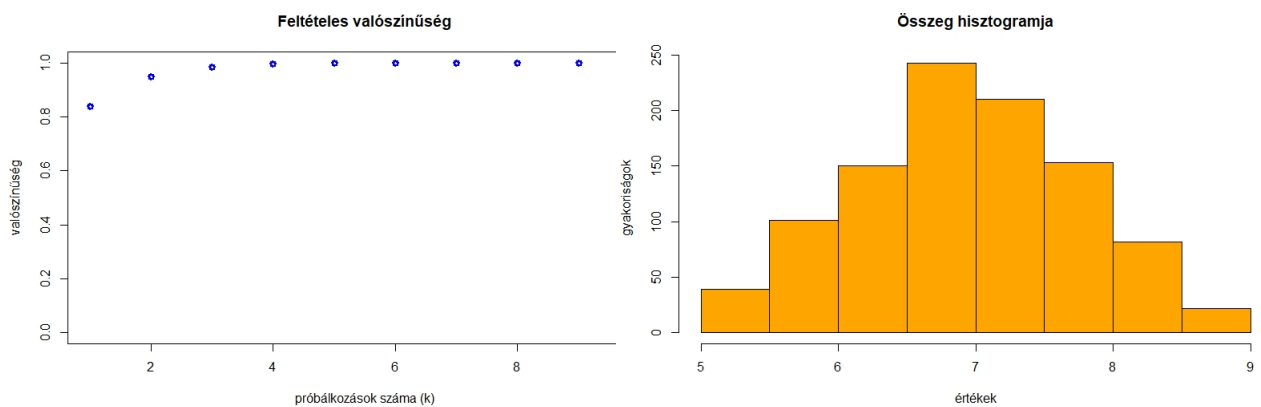


15. ábra. Az utazási távolság becsült sűrűségfüggvénye a teljes adatsorból

```
> summary(adatok$sorozat) Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 1.000 1.000
1.925 3.000 10.000
```

```
> plot(density(adatok$tavolsag[adatok$sorozat>=1], kernel="epanechnikov"), lwd="3",
col="blue", main="Távolság surusegfuggvenye", xlab="tavolsag (km)", ylab="becsult
surusegfuggveny")
```

```
> lines(density(adatok$tavolsag[adatok$sorozat<=1], kernel="epanechnikov"), lwd="3",
col="red") > legend("topright", c("legalább 1 sorozat", "legfeljebb 1 sorozat"),
col=c("blue", "red"), lwd="3")
```



16. ábra. Az utazási távolság becsült sűrűségfüggvénye a sorozatok száma, illetve a covid-fertőzöttség szerint

```
> plot(density(adatok$tavolsag, kernel="epanechnikov"), lwd="3", col="blue", main="Távolság
```

```

suruségfüggvénye", xlab="idő (perc)", ylab="becsült suruségfüggvény")
> lines(density(adatok$stavolsag[adatok$covid==1], kernel="epanechnikov"), lwd="3",
col="red") > legend("topright", c("összes", "covid"), col=c("blue", "red"), lwd="3")

```

9. Statisztikai mező

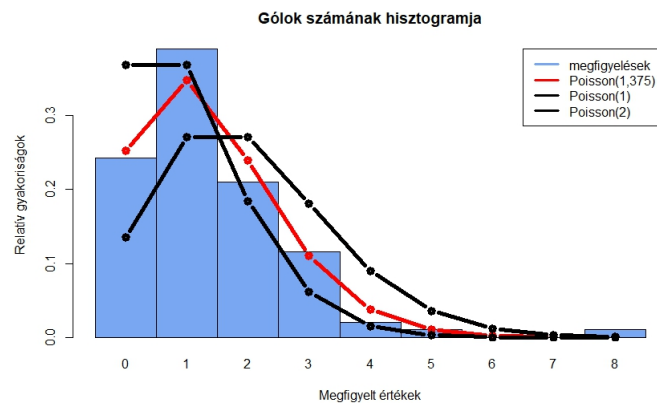
A matematikai statisztika egyik fő célja, hogy ha az X_1, X_2, \dots, X_n ismeretlen eloszlású minta, akkor erről az ismeretlen eloszlásról minél több információt nyerjen. Azt, hogy az eloszlás ismeretlen, úgy fogalmazhatjuk meg, hogy olyan valószínűségi mezőket tekintünk, ahol az eseménytér és az események halmaza ugyanaz, de a valószínűség, ami megmondja, hogy melyik esemény mennyire valószínű, eltérő.

Például annak valószínűsége, hogy egy véletlenszerűen választott ember jövedelme több 500000 forintnál, egész más lehet, ha a jövedelem (mondjuk) 300000 várható értékű és 100000 szórású normális eloszlású, vagy ha a jövedelem ezzel azonos várható értékű, de például $\alpha = 3$ rendű, vagyis végtelen szórású Pareto-eloszlással írható le.

Így juthatunk el az alábbi definícióhoz.

9.1. 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.

Ennek fontos speciális esete, amikor feltételezzük, hogy az eloszlás egy néhány paraméterrel leírható eloszláscsaládból származik, azon belül azonban nem tudjuk, hogy melyik eloszlásról van szó. Például feltételezzük, hogy a jövedelem eloszlása Pareto-eloszlás, de nem ismerjük a paramétereit, vagy feltételezzük, hogy egy betegség lappangási ideje normális eloszlású, de nem ismerjük a várható értéket és a szórást. Az ismeretlen paramétert vagy paramétereket általában ϑ -val jelöljük.



17. ábra. A gólok számának hisztogramja $n = 95$ mérkőzésen, és különböző paraméterű Poisson-eloszlások

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

Például: \mathcal{P} lehet például

- a Poisson-eloszlások halmaza, $\vartheta = \lambda$ az ismeretlen paraméter, $\Theta = (0, \infty)$ a paraméter lehetséges értékeinek halmaza;
- a normális eloszlások halmaza, ekkor $\vartheta = (m, \sigma)$ az ismeretlen paraméter (ilyenkor $\Theta \subset \mathbb{R}^2$);
- az $[a, b]$ intervallumon egyenletes eloszlások halmaza, ekkor $\vartheta = (a, b)$ az ismeretlen paraméter.

Az 17. ábra azt mutatja, hogy ha például a gólok számát Poisson-eloszlásúnak feltételezzük, de a paramétert ismeretlennek tekintjük, akkor néhány különböző λ érték mellett mennyire jól illeszkedik a λ -hoz tartozó eloszlás a megfigyelésekhez.

9.3. Definíció. *Statisztikai minta:* X_1, X_2, \dots, X_n valószínűségi változók az $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mezőn.

A minta független, ha ezek a valószínűségi változók függetlenek.

Statisztika: ha $T : \mathbb{R}^n \rightarrow \mathbb{R}^k$ egy n változós függvény, akkor a $T(X_1, \dots, X_n)$ valószínűségi változót statisztikának nevezzük.

A statisztika tehát olyan mennyiség, amit a megfigyelésekből, a mintából egy megfelelő, előre rögzített függvény alkalmazásával ki tudunk számolni. Vegyük észre, hogy a valószínűségi változók, és így a statisztika értelmezéséhez Ω és \mathcal{A} elég, hiszen $X_j : \Omega \rightarrow \mathbb{R}$ függvény volt, amire olyan feltétel volt, amit \mathcal{A} -val lehetett megfogalmazni. Viszont $\mathbb{P}(X_j \leq t)$ már függ a \mathbb{P} -től, vagyis attól, hogy a statisztikai mező melyik elemét tekintjük.

Például $k = 1$ -re példa: $T(X_1, \dots, X_n) = \frac{X_1 + \dots + X_n}{n}$ esetén a statisztika az átlag.

Vagy $k = 2$ -re példa: $T(X_1, \dots, X_n) = (\bar{X}, s_n^*)$ az a statisztika, ami a mintából az átlagot és a korrigált tapasztalati szórást számítja ki.

10. Torzítatlanság és hatásosság

Tegyük fel, hogy a $[0, \vartheta]$ intervallumon egyenletes eloszlás ismeretlen ϑ paraméterét szeretnénk becsülni (ez analóg a német tankok problémájával: https://en.wikipedia.org/wiki/German_tank_problem, lényegében annak a folytonos változata, ahol az ismeretlen paraméter nem csak egész értékeket vehet fel).

Minta: X_1, X_2, \dots, X_5 , például 4, 3; 5, 6; 3, 2; 1, 8; 2, 2

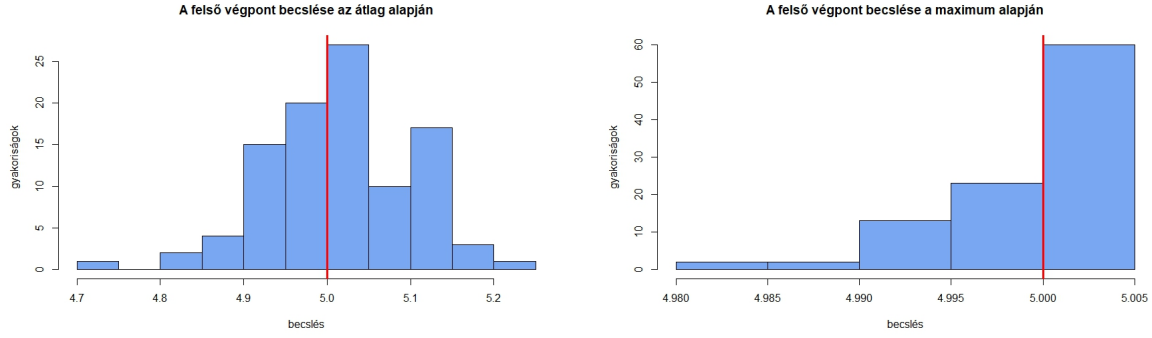
Átlag: $\bar{X} = 3,42$

Vegyük észre, hogy egy megfigyelés várható értéke $\vartheta/2$, így az átlag várható értéke is $\vartheta/2$, az átlag kétszeresének várható értéke éppen ϑ .

Ebből kiindulva tekintsük az átlag kétszeresét, mint becslést ϑ -ra:

$$\begin{aligned} T_1(X_1, \dots, X_n) &= 2\bar{X}; \\ \mathbb{E}_\vartheta(T_1) &= 2\mathbb{E}_\vartheta(\bar{X}) = 2 \cdot \mathbb{E}_\vartheta(X_1) = 2 \cdot \frac{\vartheta}{2} = \vartheta; \\ D_\vartheta(T_1) &= 2D_\vartheta(\bar{X}) = \frac{2}{\sqrt{n}}D_\vartheta(X_1) = \frac{\vartheta}{\sqrt{3n}}, \end{aligned}$$

felhasználva az egyenletes eloszlásról, illetve az átlag várható értékéről és szórásáról valószínűség-számításból tanultakat.



18. ábra. A $[0, \vartheta]$ intervallumon egyenletes eloszlás paraméterének becslése $2\bar{X}$, illetve $(n+1)X_n^*/n$ alapján

Itt az \mathbb{E}_ϑ és D_ϑ jelölés arra utal, hogy a várható érték és a szórás függ attól, hogy milyen eloszlásúnak tételezzük fel az X_1, X_2, \dots, X_n -t, amit pedig ϑ mond meg.

Másrészt, felhasználva, hogy a legnagyobb megfigyelésnek, $X_n^* = \max(X_1, \dots, X_n)$ -nek sűrűségfüggvénye: $f_\vartheta(t) = (nt^{n-1}/\vartheta^n)\mathbb{I}(0 \leq t \leq \vartheta)$, egy másik becslést is találhatunk (a számolás részleteit mellőzve), felhasználva, hogy a legnagyobb megfigyelés várható értéke $\vartheta \cdot n/(n+1)$:

$$T_2(X_1, \dots, X_n) = \frac{n+1}{n} \cdot X_n^*; \quad \mathbb{E}_\vartheta(T_2) = \frac{n+1}{n} \cdot \frac{n\vartheta}{n+1} = \vartheta.$$

$$D_\vartheta(T_2) = \sqrt{\frac{n \cdot \vartheta^2}{(n+2)(n+1)^2}} \leq \frac{\vartheta}{n+1}.$$

A két becslés összehasonlítása látható a 18. ábrán. Itt $n = 1000$ elemű mintát használtunk, száz alkalommal kisorsolva, és elvégezve a becslést. Az ábrák a száz becslés hisztogramját mutatják, a bal oldalon a $2\bar{X}$, a jobb oldalon az $(n+1)X_n^*/n$ becsléssel. Az igazi paraméter $\vartheta = 5$. Az első esetben a száz becslés átlaga: **5,015**, korrigált tapasztalati szórása: **0,086**. A második esetben a száz becslés átlaga: **4,9999**, korrigált tapasztalati szórása: **0,0049** < **0,086**.

Tehát azt látjuk, hogy bár várható érték szempontjából a két becslés hasonló, a második esetben a szórás lényegesen kisebb. Ezt az elméleti számítások is alátámasztják, a korábbiak alapján:

$$\mathbb{E}_\vartheta(T_1) = \mathbb{E}_\vartheta(T_2) = \vartheta$$

teljesül minden lehetséges $\vartheta \in \Theta$ -ra, azaz mindkét becslés **torzítatlan becslés** ϑ -ra. Ugyanakkor a második, a legnagyobb mintaelemet használó becslés szórása kisebb, ez **hatásosabb** a másiknál:

$$D_\vartheta(T_1) = \frac{\vartheta}{\sqrt{3n}} > \frac{\vartheta}{n+1} > D_\vartheta(T_2)$$

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

10.1. Torzítatlanság

A fenti példa alapján egy becslésre az alábbi tulajdonságokat vizsgálhatjuk. A g függvényre azért lehet szükség, mert nem mindig magát a paramétert közvetlenül szeretnénk becsülni. Például lehet, hogy a Poisson-eloszlás paramétere λ , mi azonban $\sqrt{\lambda}$ -t, vagyis az eloszlás szórását szeretnénk torzítatlanul megbecsülni, ekkor g lehet a gyökvonás. Vagy normális eloszlásnál lehet, hogy paraméternek a σ szórást tekintjük ismeretlen paraméternek, de a σ^2 szórásnégyzetet szeretnénk torzítatlanul becsülni, ekkor g a négyzetre emelés.

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

10.1. Definíció (Torzítatlanság). A T statisztika torzítatlan becslés g -re, ha minden $\vartheta \in \Theta$ -ra

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

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

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

10.2. A várható érték és a szórásnégyzet torzítatlan becslése

10.1. 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 becslése a várható értéknek.

Ez az alábbi, valószínűségiszámításból ismert állításnak a következménye.

10.2. Állítás. Legyen X_1, \dots, X_n azonos eloszlású minta, és $m = \mathbb{E}(X_i) < \infty$. Ekkor

$$\mathbb{E}(\bar{X}) = m.$$

Bizonyítás.

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n}\mathbb{E}(X_1 + \dots + X_n) = \frac{1}{n} \cdot nm = m.$$

Felhasználtuk a várható érték linearitását, és hogy csak eloszlástól függ:

- $\mathbb{E}(cX) = c\mathbb{E}(X)$, ha $c \in \mathbb{R}$;
- $\mathbb{E}(Y + Z) = \mathbb{E}(Y) + \mathbb{E}(Z)$;
- ha Y és Z eloszlása megegyezik, akkor $\mathbb{E}(Y) = \mathbb{E}(Z)$

□

Ebből következik, hogy a **mintaátlag** torzítatlan becslés a várható értékre.

Speciálisan: a **relatív gyakoriság** torzítatlan becslés egy esemény valószínűségére.

A szórásra teljes általánosságban nem találhatunk torzítatlan becslést, a szórásnégyzetre azonban igen (ehhez emlékeztetőül: általában $\mathbb{E}(T)^2 \neq \mathbb{E}(T^2)$, ezért nem igaz, hogy ha T^2 torzítatlan a szórásnégyzetre, akkor T torzítatlan a szórásra, nem elég gyököt vonni).

10.3. 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 szórásnégyzetre.

Ennek bizonyításához idézzük fel az alábbi állítást.

10.4. Állítás. Legyen X_1, \dots, X_n független azonos eloszlású minta, és $D^2(X_i) < \infty$ létezik. Ekkor

$$D(\bar{X}) = \frac{D(X_1)}{\sqrt{n}}.$$

Bizonyítás.

$$D(\bar{X}) = D\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{D(X_1 + \dots + X_n)}{n} = \frac{\sqrt{nD^2(X_1)}}{n} = \frac{D(X_1)}{\sqrt{n}}.$$

Felhasználtuk a szórás alábbi tulajdonságait:

- $D(cX) = |c|D(X)$, ha $c \in \mathbb{R}$ valós szám;
- $D(Y + Z) = \sqrt{D^2(Y) + D^2(Z)}$, ha Y és Z függetlenek;
- ha Y és Z eloszlása megegyezik, akkor $D(Y) = D(Z)$.

□

Szintén segítségünkre lesz, ha a tapasztalati szórásnégyzetet nem definíció alapján, hanem egy másik alakban írjuk fel.

10.5. Állítás (A tapasztalati szórásnégyzet másik alakja).

$$s_n^2 = \frac{1}{n} \left[\sum_{k=1}^n X_k^2 \right] - \bar{X}^2.$$

Bizonyítás. Átrendezéssel kapjuk, hogy

$$\sum_{k=1}^n (X_k - \bar{X})^2 = \sum_{k=1}^n [X_k^2 - 2X_k \cdot \bar{X} + \bar{X}^2] = \sum_{k=1}^n X_k^2 - 2n\bar{X} \cdot \bar{X} + n \cdot \bar{X}^2 = \sum_{k=1}^n X_k^2 - n \cdot \bar{X}^2.$$

Ebből adódik, hogy

$$s_n^2 = \frac{1}{n} \left[\sum_{k=1}^n (X_k - \bar{X})^2 \right] = \frac{1}{n} \left[\sum_{k=1}^n X_k^2 \right] - \bar{X}^2,$$

a tapasztalati szórásnégyzet definíciója alapján.

Most már kiszámíthatjuk a korrigált tapasztalati szórásnégyzet várható értékét.

$$s_n^{*2} = \frac{n}{n-1} s_n^2 = \frac{n}{n-1} \left[\frac{1}{n} \left[\sum_{k=1}^n X_k^2 \right] - \bar{X}^2 \right] = \frac{1}{n-1} \left[\sum_{k=1}^n X_k^2 \right] - \frac{n}{n-1} \bar{X}^2.$$

Ennek várható értékére vagyunk kíváncsiak.

Az első tag várható értéke a szórásnégyzet definíciója alapján:

$$\mathbb{E}_\vartheta \left(\sum_{k=1}^n X_k^2 \right) = \sum_{k=1}^n \mathbb{E}_\vartheta(X_k^2) = n \cdot \mathbb{E}_\vartheta(X_1^2) = n \cdot [D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2].$$

A második tag várható értéke szintén a szórásnégyzet definíciója, valamint az átlag várható értéke (10.2. állítás) és szórása (10.4. állítás) alapján:

$$\mathbb{E}_\vartheta(\bar{X}^2) = D_\vartheta^2(\bar{X}) + \mathbb{E}_\vartheta(\bar{X})^2 = \frac{1}{n} D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2.$$

Vagyis valóban s_n^{*2} torzítatlan becslés a szórásnégyzetre:

$$\mathbb{E}_\vartheta(s_n^{*2}) = \frac{n}{n-1} [D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2] - \frac{n}{n-1} \left[\frac{1}{n} D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2 \right] = \left(\frac{n}{n-1} - \frac{1}{n-1} \right) D_\vartheta^2(X_1) = D_\vartheta^2(X_1).$$

11. Hatásosság

Ahogy a bevezető példában is láttuk, két, a várható érték szempontjából egyformán jó becslés szórása (bizonytalansága) között lényeges különbség is lehet. Sőt, ha csak a várható értéket vennénk figyelembe, egy mintaelem, X_1 , ugyanolyan jó becslés lenne, mint 1000 mintaelem átlaga, pedig ez utóbbi sokkal informatívabb. A várható érték szempontjából egyformán jó becsléseket a szórás alapján hasonlíthatjuk össze.

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

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

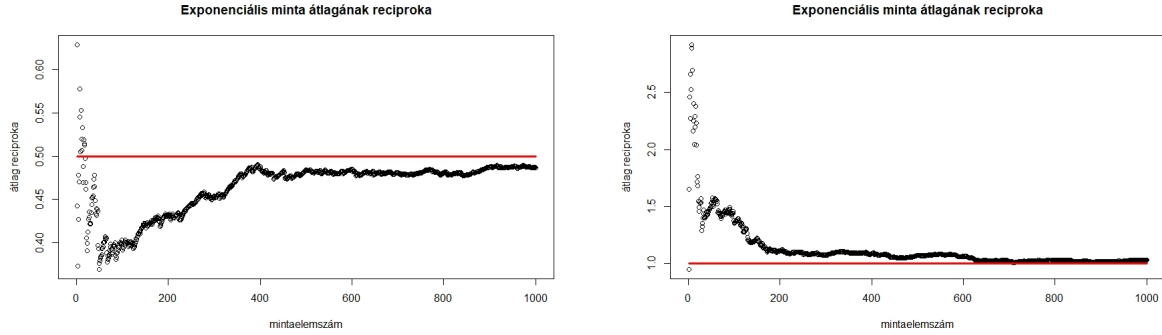
teljesül minden $\vartheta \in \Theta$ -ra.

A T_1 becslés **hatásos** $g(\vartheta)$ -ra, ha $g(\vartheta)$ 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 $\sum_{j=1}^n c_j = 1$).
- **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.

12. Konzisztencia

Az eddigiekben csak azt vizsgáltuk, hogy rögzített mintaelemszám esetén milyen tulajdonságai lehetnek egy becslésnek. Ahogy azonban például a Glivenko–Cantelli-tételnél, a statisztika alaptételénél is láttuk, az is fontos kérdés, hogy hogyan viselkedik egy becslésekből álló sorozat, ha a mintaelemszám végtelenhez tart. Ehhez a 19. ábrán látunk egy példát: ha X_1, X_2, \dots független, exponenciális eloszlású valószínűségi változókból álló minta, akkor $1/\bar{X}$, azaz az átlag reciprokának sorozata paraméterhez tart, legalábbis a két vizsgált paraméter esetén. Ha ez



19. ábra. $\lambda = 0,5$ paraméterű exponenciális eloszlást generálva a mintaátlag reciproka $0,5$ -höz tart (bal oldali ábra), $\lambda = 1$ paraméter esetén ugyanez a mennyiség 1 -hez tart (jobb oldali ábra)

minden paraméterértékre fennáll, vagyis a becslések sorozata tart a valódi, becsülni kívánt paraméterhez, azt mondjuk, hogy a becslés konzisztens.

A példában ez teljesül, hiszen a nagy számok erős törvénye szerint \bar{X} a várható értékhez tart 1 valószínűséggel, ami exponenciális eloszlás esetén $1/\lambda$. Ebből következik, hogy $1/\bar{X} \rightarrow \lambda$ teljesül 1 valószínűséggel $n \rightarrow \infty$ esetén, és így sztochasztikusan is. Vagyis a paraméter reciproka konzisztens becslése λ -nak.

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

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

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

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

Elégséges feltétel:

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

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

12.1. Példák torzítatlan, konzisztens becslésekre

X_1, X_2, \dots független azonos eloszlású véges szórású minta. Ekkor

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

teljesül $n \rightarrow \infty$ esetén sztochasztikusan a nagy számok gyenge törvénye szerint, vagyis az **átlag** konzisztens becslés a **várható értékre**, és torzítatlan is. A nagy számok erős törvénye is használható ugyanerre, abból (véges várható érték esetén) következik az 1 valószínűségű konvergencia, abból pedig a sztochasztikus.

Speciális eset: a **relatív gyakoriság** konzisztens becslés a **valószínűségre**, és torzítatlan is.

Az s_n és s_n^* közül mindkettő konzisztens becslés a szórásra, a négyzeteik pedig konzisztens becslései a szórásnégyzetnek. Azonban torzítatlanság szempontjából csak azt állíthatjuk általánosan, hogy s_n^{*2} torzítatlan becslése a szórásnégyzetnek.

Nevezetes eloszlások:

- Poisson-eloszlás λ paraméterére az átlag torzítatlan, konzisztens

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

Házi feladat 2022. február 28., hétfő, 10:30-ig Tekintsük a $\lambda = 0.1, 0.2, \dots, 1$ értékeket. Minden ilyen λ -ra sorsoljunk egy ezer elemű független mintát a megadott λ paraméterű exponenciális eloszlásból, és számítsuk ki a mintából az alábbi mennyiségeket: (a) az átlag reciproka; (b) a legkisebb mintaelem 1001-szerese; (c) a korrigált tapasztalati szórás.

Ábrázoljuk λ függvényében az így kapott értékeket (lehet egy ábrán, de három külön ábrán is). Hasonlítsuk össze a becsléseket az ábra alapján.

Hivatkozások

- [1] Ignacio Moral-Arce, Antonio de las Heras Perez, Stefan Sperlich. Recovering income distributions from aggregated data via micro-simulations. Spanish Journal of Statistics, Vol.1, No.1 (2019) 13–29, doi:<https://doi.org/10.37830/SJS.2019.1.0>
- [2] Adriano Z. Zambom and Ronaldo Dias. A Review of Kernel Density Estimation with Applications to Econometrics. <https://arxiv.org/pdf/1212.2812.pdf>