Matematikai statisztika előadás, 1. hét, február 10.

Bevezetés

1. A kurzus célja és ajánlott irodalom

A statisztika céljai:

- mérési eredmények, megfigyelések elemzése (leíró statisztika)
- ismeretlen paraméterek becslése (matematikai statisztika, becsléselmélet)
- 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 a matematikai statisztika főbb módszereinek (például becsléselméleti, 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.

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: $\frac{|X-\vartheta|}{X}$.

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 $(0, 1, 2, \ldots)$, választópolgár pártpreferenciája (pártok).

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

- **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)

Például t-próba minőségi vagy ordinális ismérv 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).

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

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

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

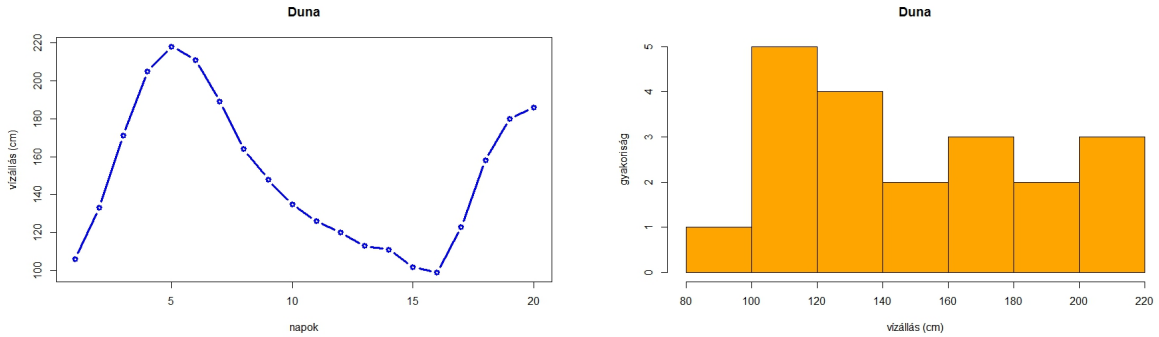
- a valószínűségi változók eloszlásának minél jobb megismerése
- a várható érték, szórás stb. becslése
- az eloszlásra vonatkozó hipotézisek eldöntése
- több valószínűségi változó együttes viselkedésének

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`)



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

- 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 $X_1, X_2, \ldots, X_{20}$ az egyes napokon mért értékek, akkor $(X_1, X_2, \ldots, X_{20})$ egy valószínűségi vektorváltozó;
- az $X_1, X_2, \ldots, X_{20}$ valószínűségi változók, vagyis a mért értékek nem függetlenek egymástól, ez fontos lehet, ha az adatok feldolgozására matematikai statiszikai 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.

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.

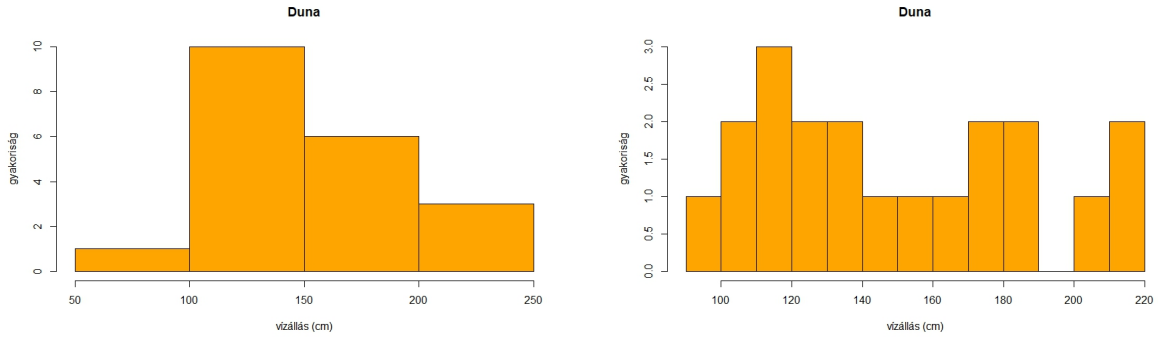
```
hist(viz, col="orange", xlab="vízállás (cm)", ylab="gyakoriság", main="Duna", breaks=4)
hist(viz, col="orange", xlab="vízállás (cm)", ylab="gyakoriság", main="Duna", breaks=15)
```

3.2. Alapstatisztikák

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

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

- **minimum**: a legkisebb mintaelem, azaz $\min(X_1, X_2, \ldots, X_n)$.
- **maximum**: a legnagyobb mintaelem, azaz $\max(X_1, X_2, \ldots, X_n)$.



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

- **terjedelem** (range): a legnagyobb és legkisebb mintaelem különbsége, azaz

$$\max(X_1, X_2, \ldots, X_n) - \min(X_1, X_2, \ldots, X_n).$$

- **medián**: a **nagyság szerinti középső** 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): $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{X_1+\ldots+X_n}{n}$.
- **tapasztalati szórásnégyzet**:

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

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

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

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

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

- **korrigált tapasztalati szórás** (standard deviation, sd): $s_n^* = \sqrt{s_n^{*2}}$.
- **relatív szórás** (relative standard deviation, rsd): $\frac{s_n^*}{\overline{X}}$.
- **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: $X_1 = 106, X_2 = 133, \ldots, X_{20} = 186.$

átlag: $\overline{X} = 149,9$

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:

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

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

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

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

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

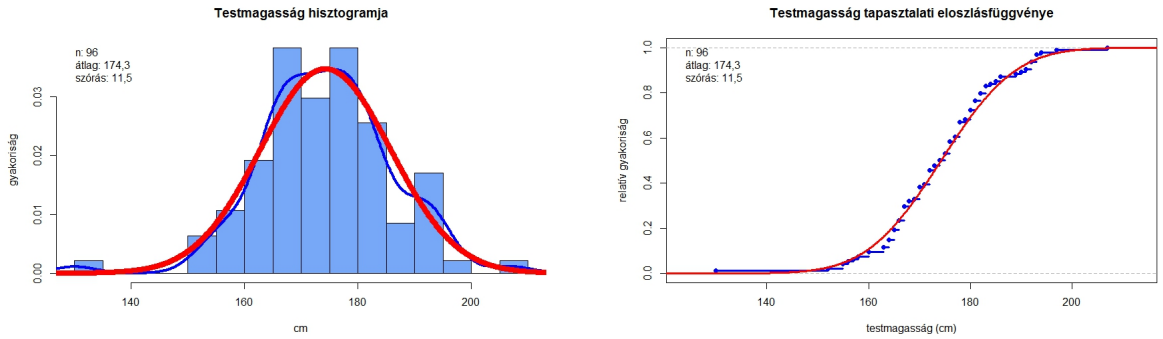
Az X valószínűségi változó *eloszlásfüggvénye* az $F : \mathbb{R} \to [0,1]$ függvény, melyre

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

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

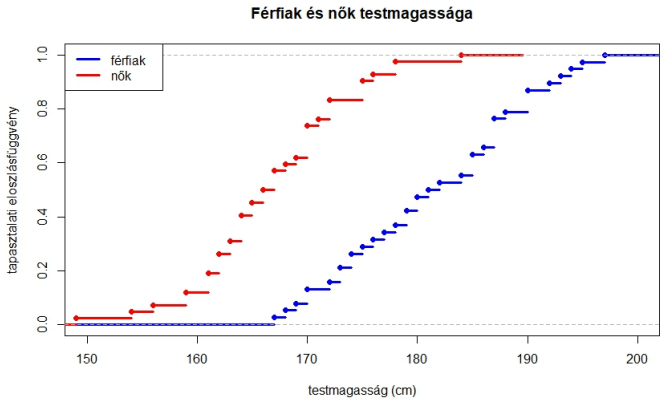
4.1. Definíció (Tapasztalati eloszlásfüggvény (empirical cumulative distribution function)). *Az $X_1, X_2, \ldots, X_n$ minta tapasztalati eloszlásfüggvénye az $\hat{F}_n : \mathbb{R} \to [0,1]$ függvény, melyre*

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

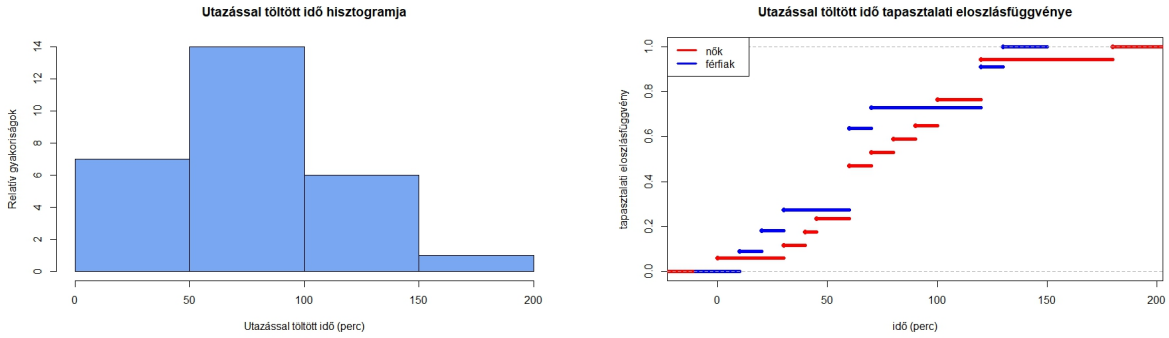


3. ábra. Egy $n = 96$ elemű, testmagasság adatsor hisztogramja és tapasztalati eloszlásfüggvénye

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



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



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

az átlaggal: $\overline{X} = 174,3$, a szórást pedig a korrigált tapasztalati szórással, és ábrázoltuk annak a normális eloszlásnak az eloszlásfüggvényét (balra), illetve sűrűségfüggvényét (jobbra, pirossal) is, melynek éppen ezek a becsült értékek a paraméterei. Az ábra alapján azt láthatjuk, hogy a becsült 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 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).

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 változók átlaga 1 valószínűséggel tart a várható értékhez. Most:

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

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 $X_1, X_2, \ldots, X_n$* **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*

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

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.