# Matematikai statisztika

Informatika alapszak, "B" spec.

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: `zempleni.elte.hu`

E-mail: `andras.zempleni@ttk.elte.hu`

Szoba: D 3-310

1. előadás

# A tananyagról

- A statisztika két fő ága:
  - Leíró statisztika
  - Matematikai statisztika
    - becsléselmélet
    - hipotézisvizsgálat
  - Néhol van/lesz átfedés
- **Lényeges, hogy amit kiszámoltunk, értelmezzük szövegesen, értelmes, kerek magyar mondatban - mert laikusoknak is tudnunk kell az eredményeket kommunikálni!**

# A statisztika története

- Kezdetek: népszámlálások az ókorban (Kína, Római Birodalom)
- A statisztika szó eredete (vitatott):
  - *status* [latin]: állapot
  - *Staat* [német], *State* [angol]: állam

  $\rightarrow$ Sokáig a statisztika az állam állapotáról fontos információk begyűjtését jelentette.
- Tudománnyá válásának kezdete: 17. század – demográfia (népesség/társadalomstatisztika)
- A 19. századtól
  - A statisztika mindenféle információ begyűjtésének, feldolgozásának és értelmezésének a tudományává vált
  - Összekapcsolódás a valószínűségelmélettel
- A számítógépek megjelenésével fejlődése felgyorsult és jelentősége megnőtt
- A statisztika megítélése vegyes, az eredményeket mindig kritikusan kell szemlélni $\rightarrow$ Churchill: "*I only believe in statistics that I doctored myself*" (Csak azoknak a statisztikáknak hiszek, amiket én magam hamisítottam.)

# Motiváció

Kérdések, amikre statisztikai eszközökkel – bizonyos mértékig – választ tudunk adni:

- A tavalyi egy nagyon hideg tél volt az USA egyes részein. Igaza van Trumpnak, hogy nincs is globális felmelegedés?
- A dohányzás mennyivel növeli annak az esélyét, hogy valaki 70 éves koráig tüdőrákban betegszik meg?
- A 2016-os USA-beli elnökválasztáson a közvélemény-kutatók Wisconsin államban közvetlenül a választás előtt átlagosan 6,5%-os Clinton-előnyt mértek. Mi az esélye, hogy Wisconsin-ban Trump győz? [$\rightarrow$ 0,7%-kal Trump nyert]
- Vajon állíthatjuk-e, hogy egy év során a bizonyos méretet meghaladó napfoltok száma Poisson-eloszlást követ? Előre tudjuk jelezni a múltbeli adatok alapján, hogy 2022-ben hány napfoltot fognak észlelni?

# A statisztika fogalma és ágai

**Statisztika**: a valóság tömör, számszerű jellemzésére szolgáló tudományos módszertan, illetve gyakorlati tevékenység.
Ágai:

- **Leíró statisztika**: magában foglalja az információk összegyűjtését, összegzését, tömör, számszerű jellemzését szolgáló módszereket. Nem foglalkozik a véletlennel.
- **Matematikai statisztika**: matematikai tudomány, a valószínűségi változókkal jellemezhető jelenségeket leíró adatok feldolgozásáról, értelmezéséről és felhasználásáról szóló tudományos módszertan

Megjegyzés: a *statisztika* szó másik jelentése – matematikai statisztikai értelemben a statisztika egy valószínűségi (vektor)változó, amit a mintából számolunk (később bővebben)

# Leíró statisztikai alapfogalmak I.

- Statisztikai egység: a statisztikai vizsgálat tárgyát képező egyed
- Statisztikai **sokaság**: a megfigyelés tárgyát képező egyedek összessége, halmaza. Röviden: sokaság (populáció). Lehet hipotetikus is (gyár által a jelenlegi körülmények között gyártandó termékek).
- **Statisztikai adat**: valamely sokaság elemeinek száma vagy a sokaságra vonatkozó számszerű jellemző, mérési eredmény.
- Statisztikai **ismérv**: a sokaság egyedeit jellemző tulajdonság. Röviden: ismérv.
- **Ismérvváltozatok**: az ismérvek lehetséges kimenetelei.
- **Minta**: a sokaság véges számosságú részhalmaza.

**Statisztikai következtetés**: a valóságban a teljes sokaságot általában nem tudjuk megfigyelni. A mintára vonatkozó információk alapján szeretnénk a teljes sokaság egészére, egyes jellemzőire, tulajdonságaira érvényes következtetéseket kimondani.

# Leíró statisztikai alapfogalmak II

- Az ismérvek típusai I.
  - minőségi ismérv: az egyedek számszerűen nem mérhető tulajdonsága
  - mennyiségi ismérv: az egyedek számszerűen mérhető tulajdonsága. Két fajtájukat különböztetjük meg:
    - diszkrét: véges vagy megszámlálhatóan sok értéket vehet fel
    - folytonos: egy adott intervallumon belül kontinuum számosságú értéket felvehet
  - időbeli ismérv: az egységek időbeli elhelyezésére szolgáló rendezőelvek
  - területi ismérv: az egységek térbeli elhelyezésére szolgáló rendezőelvek
- Az ismérvek típusai II.
  - közös ismérvek: tulajdonságok, amik szerint a sokaság egyedei egyformák
  - megkülönböztető ismérv: azok a tulajdonságok, amik szerint a sokaság egyedei különböznek egymástól

Legyen a sokaság: a teremben lévő hallgatók. Példák ismérvekre:

| | | | |
|---|---|---|---|
| minőségi: | szemszín, nem | közös: | orrok száma |
| diszkrét mennyiségi: | testvérek száma | megkülönböztető: | testsúly |
| folytonos mennyiségi: | testmagasság | | |
| időbeli: | születési idő | | |
| területi: | születési hely | | |

# Leíró statisztikai alapfogalmak III

**Mérési skálák** (mérési szintek):

- Névleges (nominális): a hozzárendelt számok csak ún. kódszámok, amik a sokaság egyedeinek azonosítására szolgálnak. Ezek között matematikai relációkat és műveleteket nincs értelme végezni. Pl. a hallgatók neme.
- Sorrendi (ordinális): a sokaság egyedeinek valamely tulajdonság alapján sorba való rendezése. Az egyedek tulajdonsága közötti különbséget nem lehet mérni. Pl. a hallgatók jegyei egy tárgyból.
- Intervallumskála: a skálaértékek különbségei is valós információt adnak a sokaság egyedeiről. A skálán a nullpont meghatározása önkényes. Ilyen skálákhoz mértékegység is tartozik. Pl. hőmérséklet (C fokban megadva).
- Arányskála: a skálának van valódi nullpontja is. Minden matematikai művelet elvégezhető ezekkel a számokkal. Pl. a hallgatók magassága.

[Metrikus skála: intervallum- és arányskála közös neve – ritkábban használatos elnevezés]

# Leíró statisztikai alapfogalmak IV

**Statisztikai tábla**: a statisztikai sorok összefüggő rendszere.
A statisztikai táblák fajtái:

- Egyszerű tábla: nem tartalmaz csoportosítást, nincs benne összegző sor
- Csoportosító tábla: egyetlen csoportosító szempontot tartalmaz. A sorok különböző sokaságokat jelentenek, a táblázatban a gyakoriságok találhatóak
- Kombinációs tábla vagy *kontingenciatábla* vagy kereszttábla: legalább két csoportosító szempontot tartalmaz, egy sokaság egyedeit csoportosítjuk, a táblázatban a gyakoriságok találhatóak

# A statisztikai elemzés lépései

1. Tervezés
   a. Mit vizsgálunk, mi a probléma/feladat
   b. Hogyan gyűjtjük az adatokat
   c. Előzetes sejtések, hipotézisek megfogalmazása
2. Terepmunka – adatgyűjtés
3. Adatbevitel, kódolás (ha szükséges)
4. Adatok validálása (biztosan rossz értékek kiszűrése, mint például életkornál a 9999)
5. Adatelemzés, adatellenőrzés: leíró statisztikákkal, grafikonok készítése
6. Hibás adatok kijavítása vagy kihagyása
7. Adatelemzés, statisztikai következtetések levonása – a matematikai statisztika módszereivel
8. Az eredmények értelmezése, visszacsatolás

# Mennyiségi sorok elemzése I

- ha a mennyiségi ismérv diszkrét és az ismérvváltozatok száma "kevés", akkor **gyakoriségi sor**t készítünk:

| Ismérvértékek | Gyakoriságok |
|---|---|
| $x_1$ | $f_1$ |
| $\vdots$ | $\vdots$ |
| $x_k$ | $f_k$ |
| Összesen | $n$ |

- $n$: minta mérete
- $k$: különböző ismérvértékek száma
- $f_i$: hányszor fordul elő az $i$-edik ismérvérték ($i = 1, \ldots, k$)

Ha a mennyiségi ismérv folytonos vagy "sok" ismérvváltozat van, akkor **osztályközös gyakoriségi sor**t készítünk:

| Ismérvértékek | Gyakoriságok |
|---|---|
| $x_{1,a} - x_{1,f}$ | $f_1$ |
| $\vdots \quad \vdots$ | $\vdots$ |
| $x_{k,a} - x_{k,f}$ | $f_k$ |
| Összesen | $n$ |

- $x_{i,a}$: az $i$-edik osztályköz alsó határa
- $x_{i,f}$: az $i$-edik osztályköz felső határa
- Minden megfigyelés pontosan egy osztályba kerüljön!

# Középértékek számítása

Adott az $n$ elemű $\underline{x} = (x_1, x_2, \ldots, x_n)$ tapasztalati minta; osztályközös gyakorisági sor esetén $k$ jelöli az osztályok számát, $x_i$ az osztályközepeket, $f_i$ pedig a gyakoriságokat.

**Mintaátlag**: az adatok átlagos értéke

- Számítása közvetlenül az adatokból: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Számítása osztályközös gyakorisági sorból: $\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$

**Módusz**: a legtöbbször előforduló ismérvérték

**Medián**: azon ismérvérték, amelynél ugyanannyi kisebb vagy egyenlő, mint nagyobb vagy egyenlő ismérvérték fordul elő a mintában (a "középső" elem). Számítása közvetlenül az adatokból:

$$\mathrm{Me} = \begin{cases} x^*_{\frac{n+1}{2}}, & \text{ha } n \text{ páratlan} \\ \frac{x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}}{2}, & \text{ha } n \text{ páros} \end{cases}$$

# Kvantilisek

**Tapasztalati $y$-kvantilis:** azon ismérvérték, amelynél a mintaelemek $y$-ad része kisebb vagy egyenlő, míg $(1-y)$-ad része nagyobb vagy egyenlő, $0 < y < 1$

Számítása nem egyértelmű, többfajta interpolációs módszer lehetséges

Nevezetes kvantilisek (jelölje $q_y$ a tapasztalati $y$-kvantilist):

- tercilisek: $T_1 = q_{1/3}$, $T_2 = q_{2/3}$
- **kvartlisek**:
  - $\mathbf{Q_1} = q_{1/4}$ (alsó kvartilis)
  - $Q_2 = \mathbf{Me} = q_{2/4}$ (középső kvartilis vagy medián)
  - $\mathbf{Q_3} = q_{3/4}$ (felső kvartilis)
- percentilisek: $P_i = q_{i/100}$, $i = 1, 2, \ldots, 99$

# Szóródási mutatók számítása

**Terjedelem**: $R = x_n^* - x_1^*$ ($R$=range)

**Interkvartilis terjedelem**: $IQR = Q_3 - Q_1$

**Tapasztalati szórás**: az átlagtól való átlagos négyzetes eltérés négyzetgyöke

- Számítása közvetlenül az adatokból: $s_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$
- Számítása osztályközös gyakorisági sorból: $s_n = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n}}$

**Korrigált tapasztalati szórás**: az átlagtól való korrigált átlagos négyzetes eltérés négyzetgyöke

- Számítása közvetlenül az adatokból: $s_n^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
- Számítása osztályközös gyakorisági sorból: $s_n^* = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n-1}}$
- ezt "szeretjük" a legjobban, minden szoftver, programcsomag szórás számításánál ezt veszi alapértelmezettnek

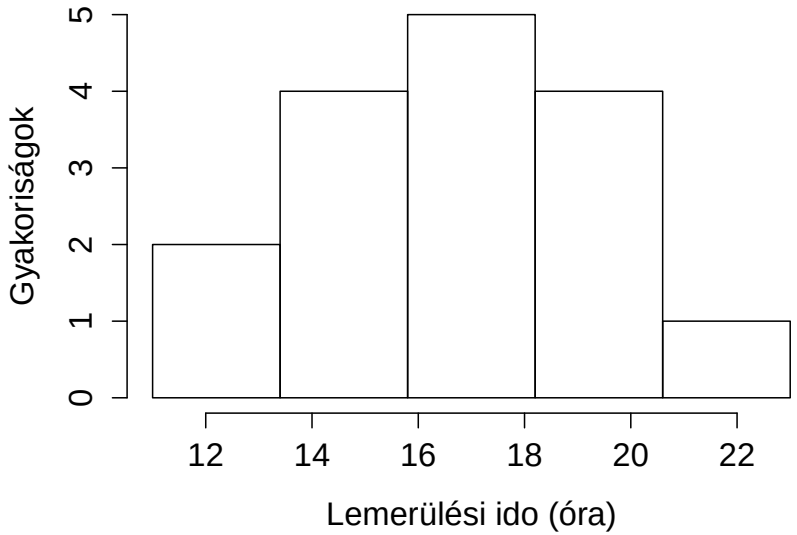
# A grafikus megjelenítés szerepe

- A statisztikus legfőbb kommunikációs eszközei a diagramok.
- Az adatokban rejlő információk gyorsabb kinyerését és feldolgozását segítik az azokból készített különféle ábrák, diagramok:
  - oszlopdiagram: idősorok, megoszlás ábrázolására
  - vonaldiagram: idősorok ábrázolására
  - hisztogram: mennyiségi sorok ábrázolására
  - kördiagram: megoszlás érzékeltetésére (nem ideális)
  - stb.
- Milyen a jó diagram?
  - illeszkedik az ábrázolt adatok fajtájához és a probléma jellegéhez
  - a célközönség meg tudja érteni
  - áttekinthető, olvashatók rajta a feliratok, jelölések
  - kreatív, esztétikus

# Fontos leíró statisztikai ábrák I

**Hisztogram** – Ha a mennyiségi ismérv folytonos vagy sok ismérvérték van, akkor alkalmas módon osztályokat képezünk, majd minden egyes adatot pontosan egy osztályhoz rendeljük. A hisztogram az osztályok gyakoriságait ábrázolja.

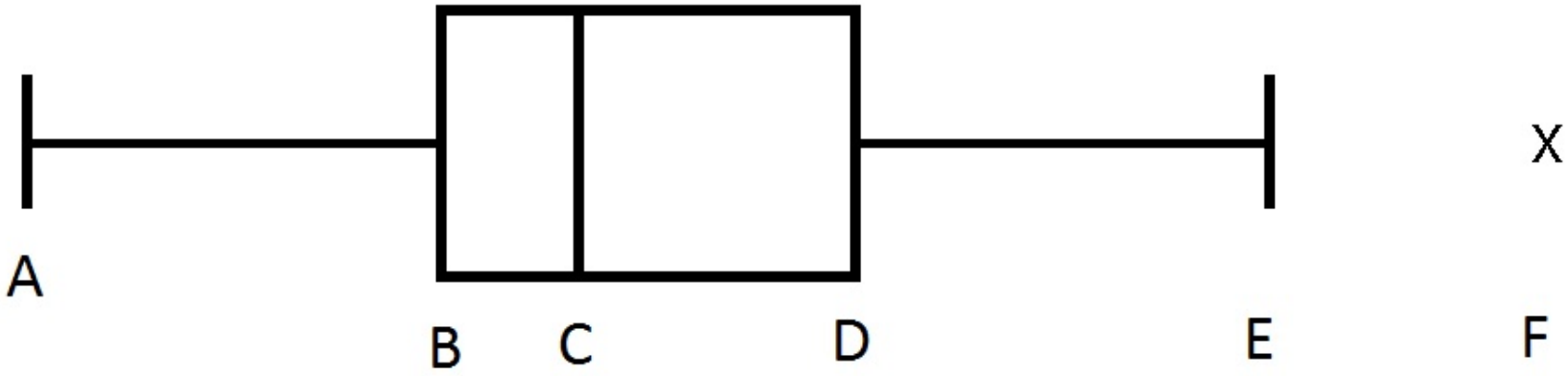
- az osztályok száma (pl.): $k = \lfloor \log_2 n \rfloor$
- ha azonos hosszúságú ($h$) osztályközöket akarunk létrehozni, akkor $h = \frac{x_n^* - x_1^*}{k}$ ($x_n^*$ a max, $x_1^*$ a min)
- az $f_i$ gyakoriságokat ábrázoljuk a függőleges tengelyen
- sűrűséghisztogramnál a $g_i = \frac{f_i}{n h_i}$ relatív gyakoriság/intervallumhossz értéket ábrázoljuk a függőleges tengelyen (területarányos, összterület=1)




- **ha az osztályközök különböző hosszúságúak, akkor a gyakoriságokat egy közös hosszra kell arányosítani**

# Fontos leíró statisztikai ábrák II

**Boxplot ábra** (Box&Whiskers diagram) – ez fekvő, de lehet álló is




A betűk a következő értékeket jelentik:

- $A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$
- $B = Q_1$ (első kvartilis)
- $C = Me$ (medián)
- $D = Q_3$ (harmadik kvartilis)
- $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$
- $F$: kiugró érték (outlier) $\rightarrow$ azokat az adatpontokat tüntetjük fel, amik $A$-n vagy $E$-n kívülre esnek

ahol $IQR = Q_3 - Q_1$ az interkvartilis terjedelem

# Tapasztalati eloszlás

- **Tapasztalati eloszlás**: minden megfigyeléshez azonos, $\frac{1}{n}$ súlyt rendelünk $\Rightarrow$ ez egy diszkrét eloszlás
- A mintaátlag éppen ennek a várható értéke
- A tapasztalati eloszlás eloszlásfüggvényét hívjuk **tapasztalati eloszlásfüggvény**nek, ami egy tiszta ugrófüggvény, értéke minden mintaelem helyén $\frac{1}{n}$ nagyságot ugrik felfelé.
  A tapasztalati eloszlásfüggvény az $x$ helyen:

$$\frac{I(x_1 < x) + I(x_2 < x) + \ldots + I(x_n < x)}{n} = \frac{\sum_{i=1}^{n} I(x_i < x)}{n}$$

Azt mutatja meg, hogy a mintaelemek hányad része kisebb $x$-nél.