

## Matematikai statisztika előadás+gyakorlat informatika BSC „A” szakosoknak

2021/2022 1. félév  
Zempléni András

1

## 1. előadás: Bevezetés

- Irodalom, követelmények
- A félév célja
- Matematikai statisztika tárgya
- Történet
- Alapfogalmak

2

## Irodalom

- Jegyzet  
Bognárné-Göndöcs-...: Matematikai statisztika
- Tankönyv  
Bolla-Krámlí: Statisztikai következtetések elmélete
- Példatár  
Móri-Szeidl-Zempléni: Matematikai statisztika példatár
- Programozáshoz:
  - R nyelv dokumentációi, magyarul például
  - <https://kovacsam.web.elte.hu/R%20bevezetes.R>
  - <http://cran.r-project.org/doc/contrib/Solymosi-Rjegyzet.pdf>
  - Angolul: [http://zempleni.elte.hu/Stat\\_R\\_Prohle\\_Zempleni](http://zempleni.elte.hu/Stat_R_Prohle_Zempleni)

3

## Követelmények

- A tárgy felvételéhez a valószínűségi számítás c. tárgy elvégzése szükséges
- A jelenlét kötelező az előadáson és a gyakorlaton is (3-3 hiányzás lehetséges), a hibrid oktatásnak megfelelően
- Összevont számonkérés részei:
  - 3 alkalommal 60 perces dolgozat, 50 pontért: az 5., 10. héten és a félév vége után, Canvasban, a gyakorlatokon
  - Lesz javítási lehetőség is (a dolgozatoknak legalább 15 pontosoknak kell lenniük)
  - Beadandó önálló feladat (statisztikai elemzés), 50 pontért. Az elemzés több részből áll. Legalább 20 pontot el kell érni!
  - Pontok szerezhetők házi feladatokkal, előadáskvízekkel is
  - Tervezett ponthatárok: 2-es 70 ponttól, ..., 5-ös 160 ponttól

4

## Cél

- Matematikai statisztika alapjainak ismertetése
  - Leíró statisztika (rövid bevezető)
  - Becslélmélet
  - Hipotézisvizsgálat
  - Többdimenziós statisztika elemei
- Alkalmazási készség kialakítása
- R programnyelv használata

5

## A matematikai statisztika tárgya

- Következtetések levonása adatok alapján
  - Ipari termelés
  - Mezőgazdaság
  - Szociológia (közvéleménykutatások)
  - Természettudományok
    - Meteorológia (pl. Klímaváltozás)
    - Genetika (chiptechnológia)
  - Pénzügyi adatok stb.

6

## Történet

- Népszámlálások már az ókorban is voltak
- Táblázatokat a biztosítók már több száz éve használják
- Maga a tudomány fiatal tudomány, alig 100 éves a múltja
  - Angliai mezőgazdasági alkalmazások voltak az elsők
- Fejlődése felgyorsult az utóbbi évtizedekben (számítógépek jóvoltából)

7

## Adatok

- Mintavétel a populációból: eredménye a (statisztikai) minta
- A mintavétel módja is lényeges (legegyszerűbb eset: bármelyik elem ugyanakkora valószínűséggel kerül a mintába)
- A mintavétel eredménye: (statisztikai) minta:  $x_1, x_2, \dots, x_n$  (számsorozat)
- Ugyanakkor egy másik, hasonló mintavételnél más mintát kapnánk, azaz az adott minta véletlen kísérlet eredménye. Ha a minta véletlen jellegét vizsgáljuk:  $X_1, X_2, \dots, X_n$  valószínűségi változó-sorozat. Lényeges különbség az előző félévhez képest: az eloszlása nem (vagy csak részben) ismert.

8

## Matematikai statisztika helye a tudományok között

- Matematikai tudomány, mert a valószínűségszámítás eredményeire épül.
- Ugyanakkor a statisztika mindennapi alkalmazása nem mindig kellően precíz (teljesülnek-e a feltételek?) Ezért lényeges, hogy a valószínűségszámítási eredményeket alkalmazva fogalmazzuk meg következtetéseinket.

9

## Statisztikai elemzés lépései

- Tervezés (mit vizsgálunk, hogyan gyűjtjük az adatokat)
- Adatgyűjtés
- Kódolás (ha szükséges)
- Ellenőrzés: leíró statisztikákkal
- Elemzés: matematikai statisztika módszereivel

10

## Leíró statisztika

- Nem a véletlen hatását vizsgálja, hanem a konkrét minta
  - megjelenítése,
  - jellemzőinek kiszámításaa feladata.
- Adatok elrendezhetőek táblázatban (fontos: forrás feltüntetése), illetve ábrázolhatóak grafikusán.

11

## Adatok típusai (skálák)

- Nominális: csak gyakoriságot tudunk számolni (nem, foglalkozás, nemzetiség)
- Ordinális (rendezett): pl. értékelés szavakkal (rossz-közepes-jó), sorrend egyértelmű, kvantilisok számolhatók
- Intervallum (pl. hőmérséklet: különbség egyértelmű, de hányados nem)
- Arány (itt minden matematikai művelet értelmes), ez szerencsére a leggyakoribb

12

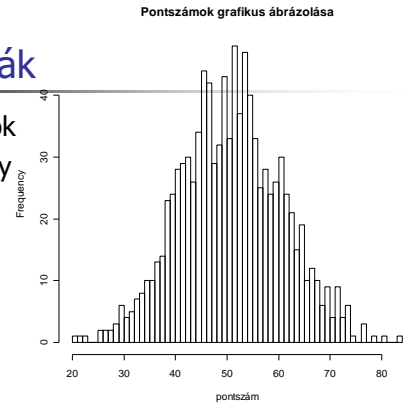
## Hisztogram

- Adatainkat osztályokba soroljuk (mindegyiket pontosan egybe, pl. az  $i$ -edik osztály:  $a_i \leq x < a_{i+1}$ ), a csoportok relatív gyakoriságai ( $r_i$ ) megegyeznek az osztály fölé rajzolt téglalap területével, tehát a téglalap magassága  $m_i = r_i / (a_{i+1} - a_i)$ .
- Összterület: 1 (hasonló a sűrűségfüggvényhez)

13

## Példák

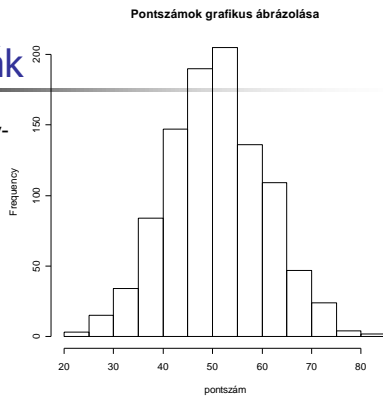
Túl sok osztály



14

## Példák

Jó osztályszám



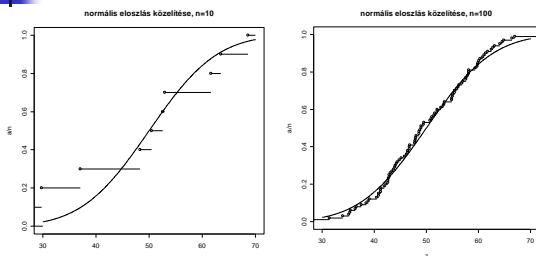
15

## Tapasztalati eloszlás

- Minden megfigyeléshez ( $x_1, x_2, \dots, x_n$ )  $1/n$  súlyt rendel. Valószínűségeloszlás!
- Mintaátlag éppen ennek az eloszlásnak a várható értéke.
- Tapasztalati eloszlás eloszlásfüggvénye: tapasztalati eloszlásfüggvény:  $F_n$  (lépcsősfüggvény).  $F_n(z) = k/n$ , ha  $x_k^{(n)} < z \leq x_{k+1}^{(n)}$   $x_0^{(n)} = -\infty, x_{n+1}^{(n)} = \infty$
- Ha a minta  $X_1, X_2, \dots, X_n$  valószínűségi változó-sorozat,  $F_n(z)$  is valószínűségi változó.

16

## Példa



17

## Középértékek: átlag

- Mintaátlag: 
$$\bar{x} := \frac{x_1 + \dots + x_n}{n}$$
- ha az egyes értékek ( $l_j$ ) gyakoriságai ( $f_j$ ) adottak: 
$$\bar{x} := \frac{f_1 l_1 + \dots + f_k l_k}{n}$$
- Ha csak az osztályközökbe eső értékek gyakoriságát ismerjük, az egyes értékeket becsüljük az osztályközépével és alkalmazzuk az előző képletet.

18

## Medián

- A sorbarendezett minta középső eleme (ha páros sok eleme van: a két középső átlaga).
- Közelítés osztályközös gyakoriságokra (kerékített értékekre):

$$Me = x_i + \frac{\frac{n}{2} - f'_{me-1}}{f_{me}} h$$

- $x_i$ : a mediánt magában foglaló osztály alsó határa
- $f'_{me}$ : kumulált gyakoriság a mediánt megelőző osztályig bezárólag
- $f_{me}$ : a mediánt magában foglaló osztály gyakorisága
- $h$ : a mediánt magában foglaló osztály szélessége.
- $n$ : a minta elemszáma

19

## Módusz

- A leggyakoribb (tipikus) érték.
- Az eloszlás lehet unimodális, bimodális vagy polimodális (egy-, két- vagy többmódusú).
- Meghatározása: A gyakorisági poligon maximumhelye (a modális osztályköz középértéke).

20

## Közelítése osztályközös esetre

$$MO = x_{mo} + \frac{f_0 - f_{0-1}}{2f_0 - f_{0-1} - f_{0+1}} h$$

Ahol

$x_{mo}$  a móduszt tartalmazó osztály alsó határa

$f_0$  a móduszt tartalmazó osztály gyakorisága

$f_{0-1}$  a móduszt tartalmazó osztályt megelőző osztály gyakorisága

$f_{0+1}$  a móduszt tartalmazó osztályt követő osztály gyakorisága

$h$  a móduszt tartalmazó osztály szélessége

Lényeges, hogy ez a három osztály azonos szélességű legyen

21

## Kvantilisek

- Elméleti kvantilis: abszolút folytonos, szigorúan monoton F esetén  $q_z = F^{-1}(z)$
- Általában:  $\inf\{x: F(x) > z\}$
- A tapasztalati eloszlás kvantilisei: tapasztalati kvantilisek. Esetleg lineáris interpolációval lehet pontosítani a becsléseinket.
- $z=1/2$ : medián.
- $z=1/4, 3/4$ : kvantilisek

22

## Kvantilisek kiszámítása

Osztályközös gyakorisági sorból

$$Q_p = x_i + \frac{pn - f'_{i-1}}{f_i} h_i$$

Ahol

$x_i$  a kvantilist tartalmazó osztály alsó határa

$n$  a minta elemszáma

$f'_{i-1}$  kumulált gyakoriság a kvantilist tartalmazó

osztályt megelőző osztállyal bezárólag

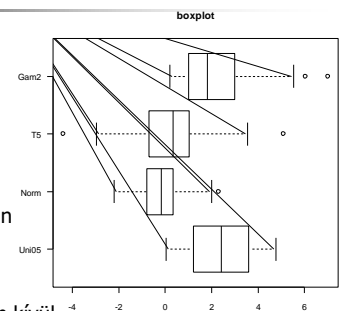
$f_i$  a kvantilist tartalmazó osztály gyakorisága

$h_i$  a kvantilist tartalmazó osztály szélessége

23

## Alapstatisztikák grafikus megjelenítése: boxplot (doboz-diagram)

Az egyes dobozok az alsó kvartilistól a felső kvartilisig tartanak. Középvonal a medián. A vonalak a teljes terjedelmet felölelik, ha ez az egyes irányokban nem nagyobb a kvartilisek közötti különbség 1.5-szeresénél. Ha ezen kívül is vannak pontok, azokat külön-külön jeleníti meg.



24