

Leíró és matematikai statisztika

Matematika alapszak, matematikai elemző szakirány

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

Tudnivalók a tantárgyról

- Kötelező irodalom: az előadásokon elhangzottak – a bemutatott módszerek, definíciók, tételek, bizonyítások, példák, ellenpéldák, feladatok.
- Ajánlott irodalom:
 - Korpásné: Általános statisztika I. \rightsquigarrow tankönyv leíró statisztikához
 - Molnárné-Tóthné: Általános statisztika példatár I. \rightsquigarrow példatár leíró statisztikához
 - Bolla-Krámlí: Statisztikai következtetések elmélete. \rightsquigarrow tankönyv matematikai statisztikához
 - Móri-Szeidl-Zempléni: Matematikai statisztika példatár.
 - Pröhle-Zempléni: Statistical Problem Solving in R. Elérési helye: http://zempleni.elte.hu/Stat_R_Prohle_Zempleni
 \rightsquigarrow **R** programnyelv bevezető, a benne szereplő statisztikai témák erősen átfednek az előadással
 - Baran: Hipotézisvizsgálat
<https://gyires.inf.unideb.hu/mobiDiak/Baran-Sandor/Hipoteziszvizsgalat/bsstat.pdf>
 - Fazekas: Statisztika
<https://gyires.inf.unideb.hu/AM/statmobi.pdf>

- Gyakjegy szükséges a vizsgához
- Vizsga: írásbeli, 2 és fél órás, 120 pontos
 - Nagy része feladatmegoldás (tesztes és kifejtős feladatok)
 - Definíciók, tételek, bizonyítások, módszerek bemutatása
 - **R** nyelvű számítógépes output-ok, számítások végeredményeinek kiértékelése, szöveges értelmezése
 - Vizsgapontok szereshetők (kb max 25) az előadáshoz kapcsolódóan is (villámkérdések, kvízek megoldásával, az előadáson való aktív részvétellel)

	elégtelen (1)	0	–	44
	elégséges (2)	45	–	64
● Osztályozás (tervezet):	közepes (3)	65	–	84
	jó (4)	85	–	104
	jeles (5)	105	–	

- Tervezett tematika: a honlapon
- A statisztika két fő ága:
 - Leíró statisztika (kb. az első 2-3 hét)
 - Matematikai statisztika (a többi 10 héten keresztül)
 - Néhol van/lesz átfedés
- A levezetések, példamegoldások a táblára kerülnek fel (online verzió: whiteboardfox); a leíró statisztikai anyagrészek nagy része, közérdekű infók, feladatok szövegei, érdekességek, szimulációk, egyéb ábrák lesznek kivetítve
- Az online előadások utólag is visszanezhetőek lesznek
- **Lényeges, hogy amit kiszámoltunk, értelmezzük szövegesen, értelmes, kerek magyar mondatban - mert laikusoknak is tudnunk kell az eredményeket kommunikálni!**

Az elemzésekhez használt szoftver/programnyelv: R

- Statisztikai modellezésre, adatok elemzésére kiválóan alkalmas
- Gyakorlaton mindenki használni fogja
- Nyílt forráskódú, ma már alig van probléma, feladat, aminek a megoldására ne lenne valamilyen csomag – akár több is
- Népszerűsége 2022-ben az összes programozási nyelv mezőnyében:
 - 7. hely – PYPL index
 - 12. hely – TIOBE index
- Jelenleg a legelterjedtebb statisztikai programnyelv
- A gyakorlaton mindenki használni fogja, az előadáson ezzel mutatok be szimulációkat, **a vizsgán kell R-es output-ot elemezni/értelmezni** (a gyakorlatokon is lesznek R-es beadandók)
- Letöltési helye: `https://cran.r-project.org/`
- Szövegszerkesztésre ajánlott szoftver: RStudio
letöltési helye: `https://www.rstudio.com/products/rstudio/download3/`

A statisztika története

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

↪ 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 ↪ Churchill: "*I only believe in statistics that I doctored myself*" (Csak azoknak a statisztikáknak hiszek, amiket én magam hamisítottam.)

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

- Az idei egy nagyon hideg tél 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? [\rightsquigarrow 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 2019-ben hány napfoltot fognak észlelni?

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 (példák)

Példák:

Sokaság:	most a meetingen lévő emberek
Statisztikai egység:	az oktató
Adat:	a legmagasabb hallgató testtömegindexe
Ismérv:	nem
Ismérvváltozatok:	férfi ($\rightarrow 1$), nő ($\rightarrow 2$)
Minta:	5 véletlenül választott hallgató

Sokaság:	az ELTE TTK Matematikai szakgyűjteményében lévő könyvek
Statisztikai egység:	a BF 13873 raktári jelzetű könyv
Adat:	a szakgyűjteményben lévő könyvek száma
Ismérv:	könyv oldalainak száma
Ismérvváltozatok:	631, 321, 153, 463, ...
Minta:	Rényi: Valószínűségszámítás című könyve

A sokaságok csoportosítása:

- 1.) A sokaság egységeinek megkülönböztethetősége szerint:
 - diszkrét: a sokaság egységei elkülönülnek egymástól
 - folytonos: a sokaság egységeit nem tudjuk természetes módon elkülöníteni (pl. áramtermelés)
- 2.) A sokaság időpontra vagy időtartamra értelmezhető-e:
 - álló: csak egy adott *időpontra* értelmezhető
 - mozgó: csak egy adott *időtartamra* értelmezhető
- 3.) A sokaság számossága szerint:
 - véges (a gyakorlatban általában ilyenekkel foglalkozunk)
 - végtelen (hipotetikus)

A statisztikai adatok fajtái:

- Alapadatok: közvetlenül a sokaságból származnak (méréssel, megszámlálással)
- Leszármaztatott adatok: alapadatokból műveletek eredményeként adódnak (pl. átlagolással, osztással)

A statisztikai adatok nem mindig pontosak – a mért és a tényleges adat eltérhet egymástól, például kerekítési okokból.

Egy kivágat a középiskolákban kötelezően alkalmazandó e-Kréta rendszerből:

KRÉTA					
Órarend					
Osztályzatok					
Mulasztások					
Információk					
e-Learning					
OSZTÁLYÁTLAGOK	#	Tantárgy	Tanuló átlaga	Osztályátlag	Különbség
	1	Irodalom	5,00	3,39	1.61
	2	Magyar nyelv	4,75	3,44	1.31
	3	Történelem	4,67	3,76	0.91
	4	Angol I nyelv	5,00	3,94	1.06
	5	Angol II nyelv	0,00	2,67	-2.67
	6	Francia II nyelv	0,00	3,44	-3.44
SZŰRÉS	7	Olasz II nyelv	0,00	4,46	-4.46
	TIPP				

Mi a vélemény?

Minden héten lesz 5 percünk statisztikai furcsaságokra, küldjenek anyagot!

E1.) Döntsük el, hogy az alábbiak egy sokaságot definiálnak, a sokaság egy-egy egyedére vonatkoznak, vagy statisztikai adatok! A sokaságok és az adatok esetében határozzuk meg azok típusát!

- a.) az Allee parkolójában álló autók száma
- b.) az Allee parkolójában álló autók
- c.) az Allee parkolójában álló MSY-766 rendszámú Opel Vectra
- d.) az Allee parkolójában álló Opelek aránya
- e.) az egy hét alatt legyártott selejtes termékek
- f.) a bankszámlánkon jóváírt kamatok
- g.) az őszi Eötvös 5 km-es futáson legjobb időt elérő másodéves hallgató (nem volt holtverseny)

Leíró statisztikai alapfogalmak III

● 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 meetingen lévő hallgatók. Példák ismérvekre:

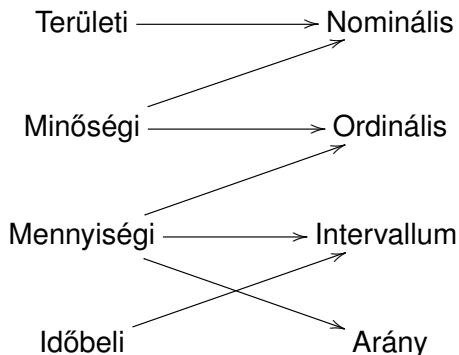
minőségi:	szemszín, nem	közös:	orrok száma
diszkrét mennyiségi:	testvérek száma	megkülönböztető:	testsúly

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]

Az ismérvek és a mérési skálák kapcsolódása:



E2.) Határozzuk meg, hogy a következő ismérvek milyen típusúak és hogy milyen skálán mérhetők! Mennyiségi ismérvek esetén állapítsuk meg, hogy az adott ismerv diszkrét vagy folytonos!

- a.) autó márkája
- b.) testsúly
- c.) cipőméret
- d.) munkahely
- e.) születési év
- f.) egy vállalat bérköltése

Statisztikai sor: a sokaság egyes jellemzőinek felsorolása.

Az ismérvek fajtája szerint beszélhetünk minőségi, mennyiségi, időbeli és területi sorokról.

A statisztikai sorok további csoportosítása:

- Csoportosító sor: a sokaság egy megkülönböztető ismerv szerinti osztályozásának eredménye; az adatok összegezhethők (van 'Összesen' sor)
- Összehasonlító sor: a sokaság *egy részének* hasonlítása a sokaság egészéhez, (példa: mezőgazdaságban dolgozók részaránya az egyes években), vagy változásának adatai (idősorok)
- Leíró sor: különböző fajta, gyakran eltérő mértékegységű statisztikai adatokat tartalmaz

Például ha egy statisztikai sor tartalmazza az előadás hallgatóit nemek szerint, akkor ez a sor minőségi csoportosító sor.

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
- Kombinációs tábla vagy *kontingenciatábla* vagy keresztábla: legalább két csoportosító szempontot tartalmaz, egy sokaság egyedeit csoportosítjuk

E3.) Milyen típusúak az alábbi táblák és milyen típusú sorokat tartalmaznak? Határozzuk meg a táblázatbeli csoportosítás alapját képző ismérvek típusát és azok mérési skáláját!

- a.) Egy vállalatnak 10 telephelye van. Három telephely dolgozóinak megoszlása életkor szerint:

Életkor (év)	2. telephely	8. telephely	9. telephely
18–30	20	20	30
31–40	20	30	20
41–50	20	30	50
51–62	20	20	10
Összesen	80	100	110

- b.) Egy golfklub tagjainak megoszlása nem és testtömegindex szerint:

Testtömegindex	Férfi	Nő	Összesen
–25	30	20	50
25–30	10	5	15
30–	5	2	7
Összesen	45	27	72

Leíró statisztikai alapfogalmak VIII

A statisztikai elemzések egyik legfontosabb eszközei a viszonyszámok (alias: indikátorok). A **viszonyszám** két statisztikai adat hányadosa.

Jelölések:

$$V = \frac{A}{B}$$

ahol V : viszonyszám; A : a viszonyítás tárgya; B : a viszonyítás alapja.

A viszonyszámok fajtái:

- Megoszlási: a sokaság egy részének a sokaság egészéhez való viszonyítása
- Koordinációs: a sokaság egy részének a sokaság egy másik részéhez való viszonyítása
- Dinamikus: két időpont vagy időszak adatának hányadosa
- Intenzitási: különböző fajta adatok viszonyítása egymáshoz; gyakran a mértékegységük is eltérő.

E4.) Az alábbi mondatokban milyen viszonyszámok rejtőznek? Azok milyen típusúak? Adjuk meg kiszámításuk pontos képletét!

- a.) Egy 25 fős csoportban a lányok részaránya 40%.
- b.) Idén 100, a tavalyihoz képest 10%-kal kevesebb hallgató vette fel a Diszkrét matematika tantárgyat.
- c.) Marika összesen 2000 km-es nyaralása alatt autója átlagfogyasztása 7 l/100 km volt.
- d.) Az ELTE-n 1500 oktató van, az egy oktatóra jutó hallgatók száma 20.

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.) Előzetes 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, kommunikáció

A grafikus megjelenítés szerepe

- A statisztikus legfőbb kommunikációs eszközei a diagramok.
- Az emberek többsége utálja a
 - barokkos körmondatokkal teletűzdelt statisztikai jelentéseket.
 - számokkal teli táblázatokat.
- Az adatokban rejlő információk gyorsabb kinyerését és feldolgozását segítik az azokból készített különféle ábrák, diagramok:
 - oszlopdiaagram: 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