Leíró és matematikai statisztika

Matematika alapszak, matematikai elemző szakirány

Varga László

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: `vargal4.elte.hu`

E-mail: `vargal4@cs.elte.hu`

Szoba: D 3-309

2017. július 3.

Tudnivalók a tantárgyról

- Kötelező irodalom: az előadásokon elhangzottak – a bemutatott módszerek, definíciók, tételek, bizonyítások, példák, ellenpéldák, feladatok, feladatok, feladatok, feladatok. Ja, és a feladatok.
- Ajánlott irodalom:
  - Korpásné: Általános statisztika I. $\rightsquigarrow$ tankönyv leíró statisztikához
  - Molnárné-Tóthné: Általános statisztika példatár I. $\rightsquigarrow$ példatár leíró statisztikához
  - Bolla-Krámli: Statisztikai következtetések elmélete. $\rightsquigarrow$ tankönyv matematikai statisztikához
  - Fazekas (szerk.): Bevezetés a matematikai statisztikába. $\rightsquigarrow$ 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

Tudnivalók a tantárgyról

- Gyakorlat
  - Gyakjegy szükséges ahhoz, hogy vizsgázhass.
  - A gyakjegy feltételeiről majd a gyakorlaton...
  - **A folyamatos gyakorlás, feladat-, problémamegoldás nagyon fontos**, anélkül nem lehet elsajátítani a tananyagot!
- Vizsga: írásbeli, 2-3 órás, 100 pontos
  - Nagy része feladatmegoldás (tesztes és kifejtős feladatok)
  - Definíciók, tételek, bizonyítások, módszerek bemutatása
  - **R** nyelvű számítógépes output-ok, számítások végeredményeinek kiértékelése, szöveges értelmezése
  - Két részes:
    - I. rész: 40 pontos, minimum 60%-ot (24 pontot) el kell érni – tesztfeladatok, definíciók, tételek, rövid keresztkérdések
    - II. rész: 60 pontos: főleg kifejtős kérdések
  - **Számológép** ($\neq$ mobiltelefon) használható
  - Osztályozás:

karó (1)	0	–	34,99
elégséges (2)	35	–	49,99
közepes (3)	50	–	64,99
jó (4)	65	–	79,99
jeles (5)	80	–	100

A tananyagról

- Tervezett tematika: a honlapomon – ld. syllabus
- A Statisztika két fő ága:
  - Leíró statisztika (3-4 héten keresztül)
  - Matematikai statisztika (9-10 héten keresztül)
  - Néhol van/lesz átfedés
- A matematika a táblán fog megszületni; a leíró statisztikai anyagrészek nagy része, közérdekű infók, feladatok szövegei, érdekességek, szimulációk, egyéb ábrák lesznek kivetítve
- **Mindig gondold végig, hogy amit kiszámoltál, hogyan tudod szövegesen, értelmes és szép, kerek magyar mondatban értelmezni!**

Szimulációkhoz használt szoftver/programnyelv: **R**

- Statisztikai modellezésre, adatok elemzésére kiválóan alkalmas programnyelv
- Gyakorlaton mindenki használni fogja
- Nyílt forráskódú, ma már alig van probléma, feladat, aminek a megoldására ne lenne valamilyen package – akár több is
- Népszerűsége 2017 februárjában az összes programozási nyelv mezőnyében:
  - 9. hely – PYPL index
  - 16. hely – TIOBE index
- Jelenleg a legelterjedtebb matematikai célú programnyelv
- A gyakorlaton mindenki használni fogja, az előadáson ezzel mutatok be szimulációkat, **a vizsgán kell R-es output-ot elemezni/értelmezni** (a gyakomon 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]: állam

  $\rightsquigarrow$ 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 $\rightsquigarrow$ 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:

- Az idei egy nagyon hideg január volt a Kárpát-medencében. Állíthatjuk-e, hogy nincs is globális felmelegedés?
- A dohányzás mennyivel növeli annak az esélyét, hogy valaki 70 éves koráig tüdőrákban betegszik meg?
- A legutóbbi 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 fog győzni? [$\rightsquigarrow$ 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 2017-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
- **Matematikai statisztika**: matematikai tudomány, a valószínűségi változókkal jellemezhető jelenségek leíró adatainak 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.
- **Statisztikai adat**: valamely sokaság elemeinek száma vagy a sokaság valamilyen másféle számszerű jellemzője, mérési eredmény.
- Statisztikai **ismérv**: a sokaság egyedeit jellemző tulajdonság. Röviden: ismérv.
- **Ismérvváltozatok**: az ismérvek lehetséges kimenetelei.
- **Minta**: a sokaság véges számosságú részhalmaza. [A minta más értelmezéseiről később...]

Statisztikai következtetés: a valóságban a teljes sokaságot nem tudjuk vagy akarjuk megfigyelni, ezért csak az egyedek egy szűkebb csoportját figyeljük meg. A viszonylag kisszámú egyedre vonatkozó információk alapján szeretnénk a teljes sokaság egészére, egyes jellemzőire, tulajdonságaira érvényes következtetéseket kimondani.

Leíró statisztikai alapfogalmak

Példák:

Sokaság:	most a teremben lévő homo sapiensek
Statisztikai egység:	a teremben lévő oktató
Adat:	a legmagasabb hallgató testtömegindexe
Ismérv:	nem
Ismérvváltozatok:	férfi ($\to 1$), nő ($\to 0$)
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:	oldalak száma
Ismérvváltozatok:	631, 321, 153, 463, . . .
Minta:	a Rényi: Valószínűségszámítás című könyvek

Leíró statisztikai alapfogalmak

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. bauxittermelés)

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

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.

Feladatok

E**1.)** Döntsd el, hogy az alábbiak egy sokaságot definiálnak, a sokaság egy-egy egyedére vonatkoznak, vagy statisztikai adatok! A sokaságok és az adatok esetében határozd meg azok típusát!

a.) az épület melletti parkolóban álló autók száma
b.) az épület melletti parkolóban álló autók
c.) az épület melletti parkolóban álló ASY-766 rendszámú Opel Vectra
d.) az épület melletti parkolóban álló Opelek aránya
e.) az egy hét alatt legyártott selejtes termékek
f.) bankszámlámon jóváírt kamatok
g.) az őszi ELTE 5 km-en legjobb időt elérő másodéves hallgató (nem volt holtverseny)

Leíró statisztikai alapfogalmak

- 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 sok. 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		

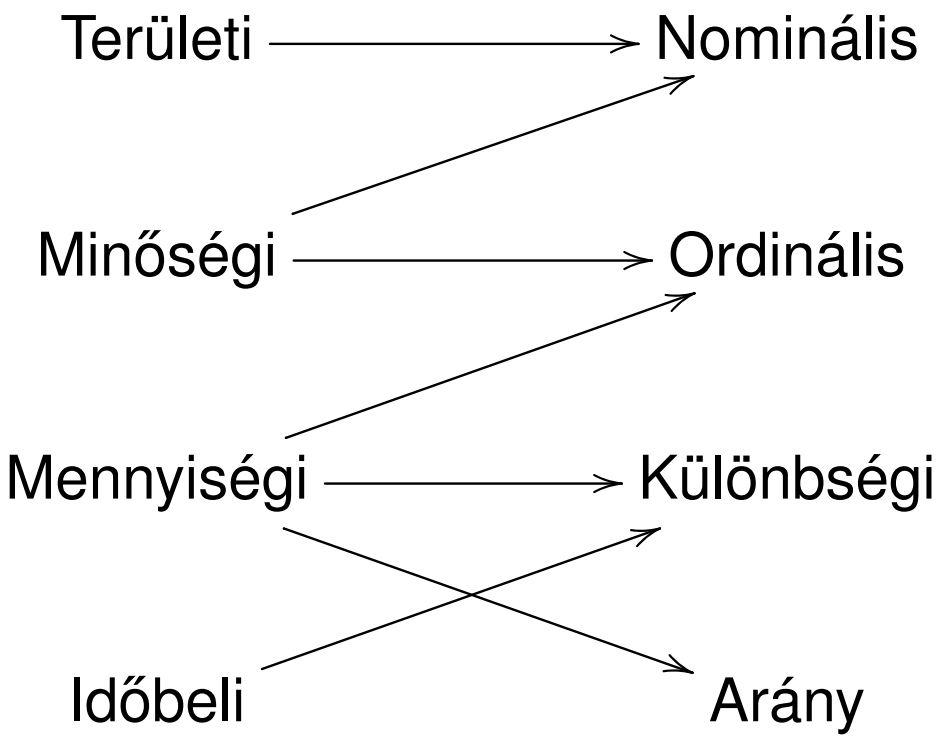
Leíró statisztikai alapfogalmak

Mérési skálák (mérési szintek):

- Névleges (nominális): a 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.
- 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:



Feladatok

E**2.)** Határozd meg, hogy a következő ismérvek milyen típusúak és hogy milyen skálán mérhetők! Mennyiségi ismérvek esetén állapítsd meg, hogy az adott ismérv diszkrét vagy folytonos!

a.) szemszín
b.) testmagasság
c.) hőmérséklet
d.) munkahely
e.) születési idő
f.) egy vállalat bérköltsége

Leíró statisztikai alapfogalmak

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 ismérv szerinti osztályozásának eredménye; az adatok összegezhetők (van 'Összesen' sor)
- Összehasonlító sor: a sokaság *egy részének* a sokaságot egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok nem összegezhetők
- Leíró sor: különböző fajta, gyakran eltérő mértékegységű statisztikai adatokat tartalmaz

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

Leíró statisztikai alapfogalmak

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

- Egyszerű tábla: nem tartalmaz csoportosítást, nincs benne összegző sor
- Csoportosító tábla: egyetlen csoportosító sort tartalmaz
- Kombinációs tábla vagy *kontingenciatábla* vagy kereszttábla: legalább két csoportosító sort tartalmaz

Feladatok

E**3.)** Milyen típusúak az alábbi táblák és milyen típusú sorokat tartalmaznak? Határozd 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
50–62	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

Feladatok

c.) A szobámban lévő sok növény közül két növény néhány jellemzője:

A növény neve	Súly (kg)	Magasság (cm)	Tüskék száma (db)
Miss Jukka	2,4	52	0
Mr. Kaktusz	1,1	43	1111

E**4.)** Az alábbi mondatokban milyen viszonyszámok rejtőznek? Azok milyen típusúak? Add meg kiszámításuk pontos képletét!

a.) Egy 25 fős csoportban a lányok részaránya 40%.

b.) Idén 180, a tavalyihoz képest 10%-kal kevesebb hallgató vette fel a Diszkrét matematika tantárgyat.

c.) Marika összesen 2000 km-es nyaralása alatt autója átlagfogyasztása 8 l/100 km volt.

d.) Az ELTE-n 4000 oktató van, az egy oktatóra jutó hallgatók száma 20.

Leíró statisztikai alapfogalmak

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.

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

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:
  - kördiagram: megoszlás érzékeltetésére
  - oszlopdiagram: idősorok ábrázolására
  - vonaldiagram: idősorok ábrázolására
  - hisztogram: mennyiségi sorok ábrázolására
  - 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

Összetett viszonyszámok

- Adott egy sokaság és annak m része.
- Jelölje $V_i = \frac{A_i}{B_i}$ az i. részsokaságra vonatkozó viszonyszámot $(i = 1, \ldots, m)$.
  Nevük: **részviszonyszámok**.
- Jelölje a teljes sokaságra számolt viszonyszámot $\overline{V}$.
  Neve: **összetett viszonyszám**.
- Kiszámítási lehetőségek: $\overline{V} = \frac{\sum_{i=1}^{m} A_i}{\sum_{i=1}^{m} B_i} = \underbrace{\frac{\sum_{i=1}^{m} B_i V_i}{\sum_{i=1}^{m} B_i}}_{\text{súlyozott számtani átlag}} = \underbrace{\frac{\sum_{i=1}^{m} A_i}{\sum_{i=1}^{m} \frac{A_i}{V_i}}}_{\text{súlyozott harmonikus átlag}}$
- A leíró statisztikai szakirodalomban az i indexeket – pongyola módon – le szokták hagyni: $\overline{V} = \frac{\sum A}{\sum B} = \frac{\sum BV}{\sum B} = \frac{\sum A}{\sum \frac{A}{V}}$

Feladatok

E**5.)** Egy szálloda 2016-os vendégforgalmáról az alábbiakat ismerjük:

Származási ország szerint a vendég	Vendég-éjszakák száma (éj)	Egy vendég-éjszakára jutó szállás díja (Ft/éj)	Egy vendégre jutó vendég-éjszakák száma (éj/fő)
Belföldi	5000	16000	4
Külföldi	4000	12000	2
Összesen	9000	...	...

Határozd meg a teljes hotelre vonatkozóan az egy vendégéjszakára jutó szállás díjat és az egy vendégre jutó vendégéjszakák számát!

Idősorok elemzése [leíró statisztikai eszközök]

Véges idősor: $y_1, y_2, \ldots, y_n$

Az idősorok fajtái:

- állapotidősor: a benne lévő adatok egy-egy adott időpontra vonatkoznak (pl. egy cég raktárkészlete adott napokon);
- tartamidősor: a benne lévő adatok időszakra vonatkoznak (pl. egy cég havi nyereségei).

Az idősor értékeiből számítható dinamikus viszonyszámok:

- Bázisviszonyszámok: $b_t = \frac{y_t}{y_B}$, ahol $t = 1, \ldots, n$; B fix, neve: bázisidőszak (tipikusan $B = 1$);
- Láncviszonyszámok: $l_t = \frac{y_t}{y_{t-1}}$, ahol $t = 2, \ldots, n$.
  l_1-et nem értelmezzük, táblázatban kihúzandó!

A bázisviszonyszámokból ki lehet számítani a láncviszonyszámokat és fordítva:

- láncból bázis: $b_t = l_{B+1} \cdot l_{B+2} \cdot \ldots \cdot l_t \quad (t = 1, \ldots, n)$;
- bázisból lánc: $l_t = \frac{b_t}{b_{t-1}} \quad (t = 2, \ldots, n)$.

Idősorok elemzése [leíró statisztikai eszközök]

Az idősor átlagos értékének kiszámítása:

- tartamidősor esetén sima számtani átlaggal: $\overline{y} = \frac{\sum_{t=1}^{n} y_t}{n}$
- állapotidősor esetén kronologikus átlaggal: $\overline{y}_K = \frac{\frac{1}{2}y_1 + \sum_{t=2}^{n-1} y_t + \frac{1}{2}y_n}{n-1}$

Az idősor átlagos változásának vizsgálata (állapotidősor esetén értelmes):

- a fejlődés átlagos mértéke: $\overline{d} = \frac{y_n - y_1}{n-1}$
- a fejlődés átlagos üteme: $\overline{I} = \sqrt[n-1]{\frac{y_n}{y_1}}$

Feladatok

E6.) Egy vállalat bankszámláján lévő pénz az egyes években január 1-jén (millió forint):

Év	Pénzösszeg (M Ft)	Év	Pénzösszeg (M Ft)
2009	190	2013	260
2010	200	2014	280
2011	210	2015	300
2012	225	2016	310

a.) Határozd meg a fent látható statisztikai sor típusát!
b.) Határozd meg a bázisviszonyszámokat 2010-es bázissal, valamint a láncviszonyszámokat!
c.) 2010-ben átlagosan hány forint volt a cég bankszámláján?
d.) A 2010. január elsejét követő 5 évben (2010 és 2015 között) átlagosan hány forint volt a cég bankszámláján?
e.) 2010. január elsejéről 2016. január elsejére évente átlagosan mennyivel változott a cég pénzvagyona? Értelmezd szövegesen az eredményt!

Mennyiségi sorok elemzése

Adatok mennyiségi sorba helyezése:

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

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

  - $x_{i,a}$: az i-edik osztályköz alsó határa
  - $x_{i,f}$: az i-edik osztályköz felső határa

Mennyiségi sorok elemzése

Osztályközös gyakorisági sor létrehozása nem egyértelmű

- Hány osztályköz legyen?
- Mik legyenek az osztályközök?

Hüvelykujjszabály:

- Osztályközök száma: $k = \lfloor \log_2 n \rfloor$
- Azonos hosszúságú osztályközök, hosszuk: $h = \frac{x_{max} - x_{min}}{k}$

Jelölések (osztályközös) gyakorisági soroknál:

- $x_i = \frac{x_{i,a} + x_{i,f}}{2} \rightsquigarrow$ az i. osztályközép/ismérvérték
- $f_i \rightsquigarrow$ gyakoriság
- $f_i' = \sum_{k=1}^{i} f_k \rightsquigarrow$ kumulált gyak.
- $g_i = \frac{f_i}{\sum_i f_i} \rightsquigarrow$ relatív gyak.
- $g_i' = \sum_{k=1}^{i} g_k \rightsquigarrow$ kumulált rel. gyak.
- $s_i = x_i \cdot f_i \rightsquigarrow$ értékösszeg
- $s_i' = \sum_{k=1}^{i} s_k \rightsquigarrow$ kum. értékösszeg
- $z_i = \frac{s_i}{\sum_i s_i} \rightsquigarrow$ relatív értékösszeg
- $z_i' = \sum_{k=1}^{i} z_k \rightsquigarrow$ kum. rel. értékö.

Példa - magyarországi jövedelmek 2015-ben

Decilis jövede-lemosztály	Bruttó éves átlag-jövedelem (e Ft)	Személyek száma (e fő)	g_i (%)	g_i' (%)	s_i (Mrd Ft)	z_i (%)	z_i' (%)
1	366	965	10	10	353	3	3
2	652	967	10	20	630	5	7
3	886	970	10	30	859	6	13
4	1025	964	10	40	989	7	20
5	1210	971	10	50	1175	8	29
6	1368	964	10	60	1318	9	38
7	1489	969	10	70	1443	10	49
8	1771	965	10	80	1710	12	61
9	2164	967	10	90	2092	15	76
10	3470	967	10	100	3355	24	100
Összesen	1440	9669	100	–	13924	100	–

Megjegyzések:

- A KSH (Központi Statisztikai Hivatal) háztartásokra összegezte a jövedelmeket, majd számolt átlagjövedelmet, így az átlagkeresetek gyerekekre is vonatkoznak, pedig ők nyilván nem dolgoznak.
- Ezek csak a legális jövedelmek, nincs bennük becslés az illegális jövedelmekre.
- A táblázatban lévő számok kerekített értékek.

Forrás: `https://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_zhc014a.html?down=1634`

Mennyiségi sorok elemzése

Koncentráció: a sokasághoz tartozó teljes értékösszeg jelentős része a sokaság kevés egységére összpontosul.
Legyen a sokaság n elemű, a különböző ismérvértékek $x_1, \ldots, x_k$, ezek gyakoriságai $f_1, \ldots, f_k$.

Gini-együttható: $G = \frac{1}{n(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} f_i f_j |x_i - x_j|$.

A koncentráció mutatószámai:

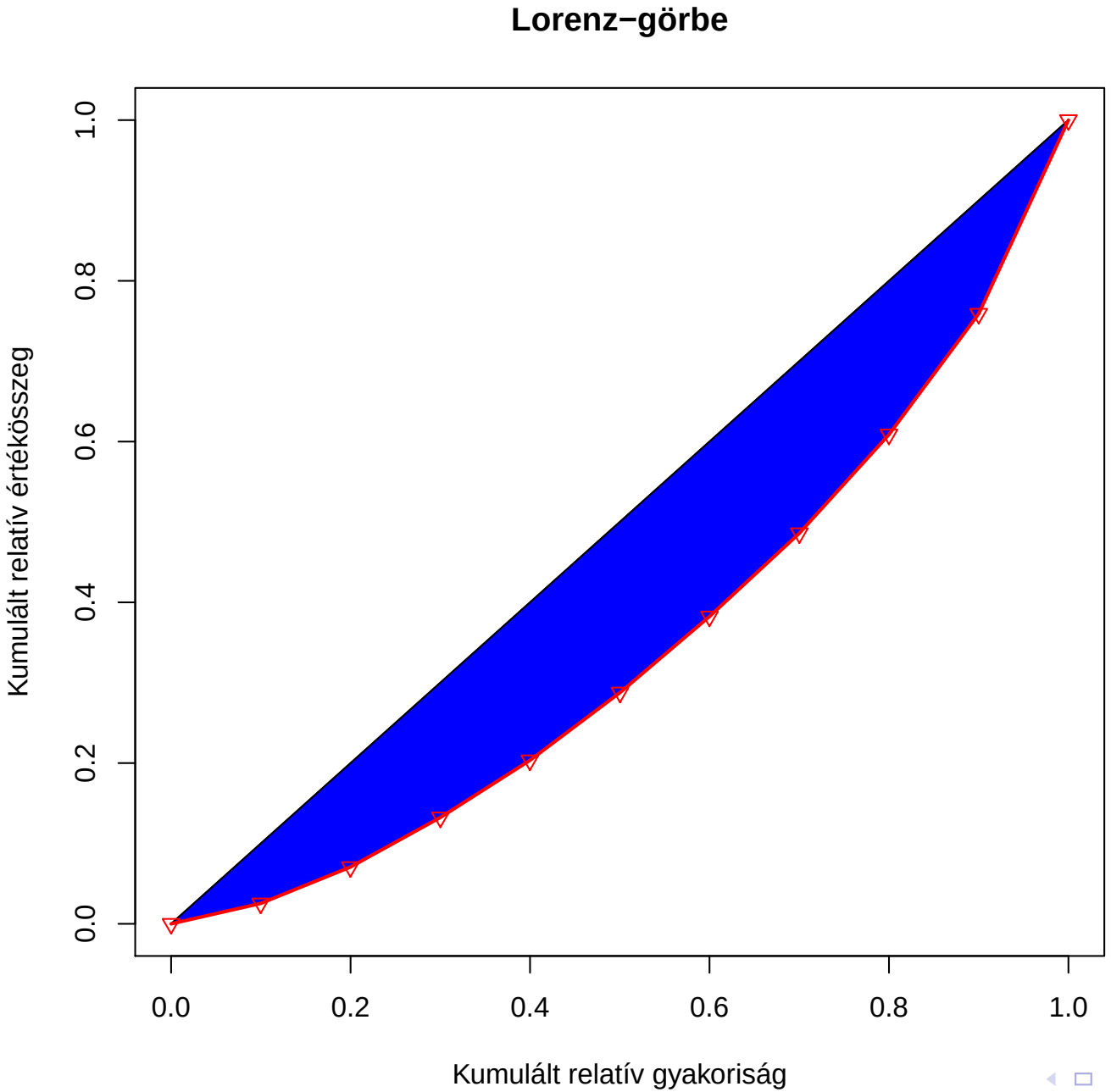
- *Koncentrációs együttható*: $L = \frac{G}{2\bar{x}}$
  Ez nem más, mint a koncentrációs terület (ld. következő fólia) 2-szerese.
  Értéke 0 és 1 között van; minél nagyobb, annál erősebb a koncentráció.
- *Herfindahl-index*: $HI = \sum_{i=1}^{k} z_i^2$
  Értéke $\frac{1}{k}$ és 1 közötti; minél nagyobb, annál erősebb a koncentráció.

Mennyiségi sorok elemzése

Lorenz-görbe – a koncentráció mértékét szemléltető ábra

- Vízszintes tengely: g'_i kumulált relatív gyakoriságok
- Függőleges tengely: z'_i kumulált relatív értékösszegek
- A 45 fokos egyenest (átló) berajzolása
- *Koncentrációs görbe* berajzolása: $(0;0), (g'_1; z'_1), (g'_2; z'_2), \ldots, (g'_{k-1}; z'_{k-1}), (g'_k; z'_k) = (1;1)$ pontok összekötésével kapott töröttvonal
- *Koncentrációs terület*: a koncentrációs görbe és az átló által közbezárt terület
- Erős a koncentráció, ha a koncentrációs görbe közel van a négyzet oldalaihoz. Gyenge a koncentráció, ha a koncentrációs görbe közel van az átlóhoz.

Példa – magyarországi jövedelmek 2015-ben



Kék: koncentrációs terület

Piros: koncentrációs görbe (törröttvonal)

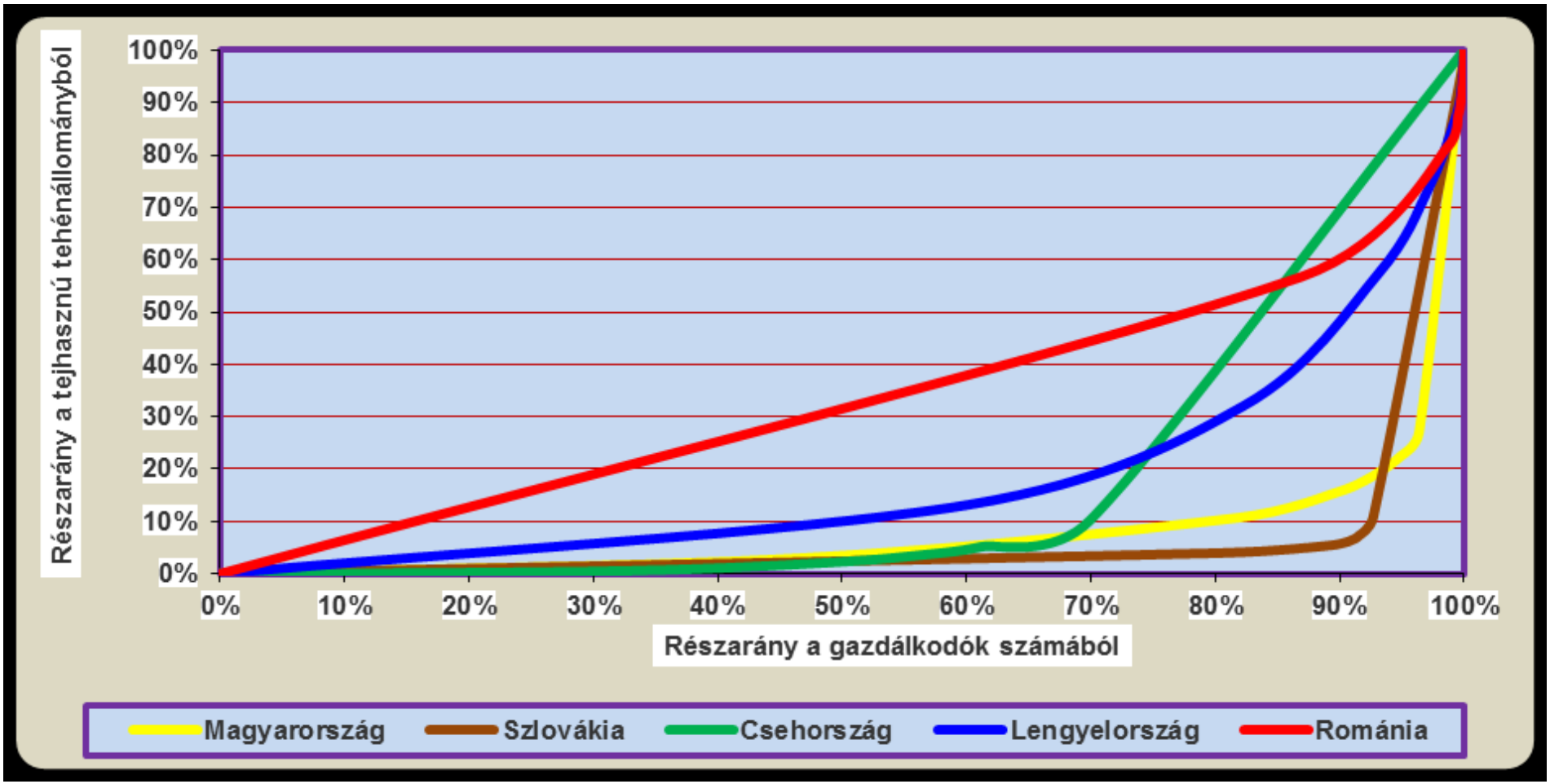
$L = 0,3089$

$HI = 0,134$
$0,1 \leq HI \leq 1$

Mihez viszonyítsuk a koncentráció mértékét?

Példa – Lorenz-görbe

A tejágazat koncentrációja különböző országokban



Forrás:
http://docplayer.hu/189090-A-magyar-tejagazat-helyzete-es-fejlodesenek-lehetseges-iranya.html

Feladatok

E**7.)** Egy piacon 4 azonos méretű vállalat működik (a piaci forgalomból azonos mértékben részesednek). Számszerűsítsük a Herfindahl-indexszel a piaci koncentráció változását, ha az egyik cég felvásárolja a másikat!

E**8.)** Legyen az X valószínűségi változó
a.) eloszlása $P(X = 0) = P(X = 2) = P(X = 3) = \frac{1}{3}$;
b.) sűrűségfüggvénye $f(x) = (2x - 2)I(1 < x < 2)$.
Határozd meg X kvantilisfüggvényét!

E**9.)** Határozzuk meg a standard normális eloszlás móduszát, mediánját, ferdeségét és lapultságát!

E**10.)** Határozzuk meg a standard Cauchy-eloszlás (Cauchy(0;1)) és a Pareto-eloszlás várható értékét!

E**11.)** Legyenek $X \sim \Gamma(\alpha, \lambda)$, $Y \sim \Gamma(\beta, \lambda)$ függetlenek! Mutassuk meg, hogy $U = X + Y$ és $V = \frac{X}{X+Y}$ függetlenek, és határozzuk meg az eloszlásukat!

Eddig megismert eloszlások

Jelölése	Eloszlása	EX	D^2X
Ind(p)	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$
Hipgeo(N,M,n)	$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ $k=0,1,\ldots,\min(n,M)$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$
Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,\ldots,n$	np	$np(1-p)$
Geo(p)	$P(X=k)=p(1-p)^{k-1}$ $k=1,2,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
NegBin(n,p)	$P(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}$ $k=n,n+1,\ldots$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poi(λ)	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ $k=0,1,\ldots$	λ	λ

Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
$E(a,b)$	$\begin{cases}0 & \text{ha } x\le a\\ \frac{x-a}{b-a} & \text{ha } a<x\le b\\ 1 & \text{ha } b<x\end{cases}$	$\begin{cases}\frac{1}{b-a} & \text{ha } a<x\le b\\ 0 & \text{különben}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$N(m,\sigma^2)$	$\ldots$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ $x\in\mathbb{R}$	m	σ^2
Exp(λ)	$\begin{cases}1-e^{-\lambda x} & \text{ha } x\ge 0\\ 0 & \text{különben}\end{cases}$	$\begin{cases}\lambda e^{-\lambda x} & \text{ha } x\ge 0\\ 0 & \text{különben}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

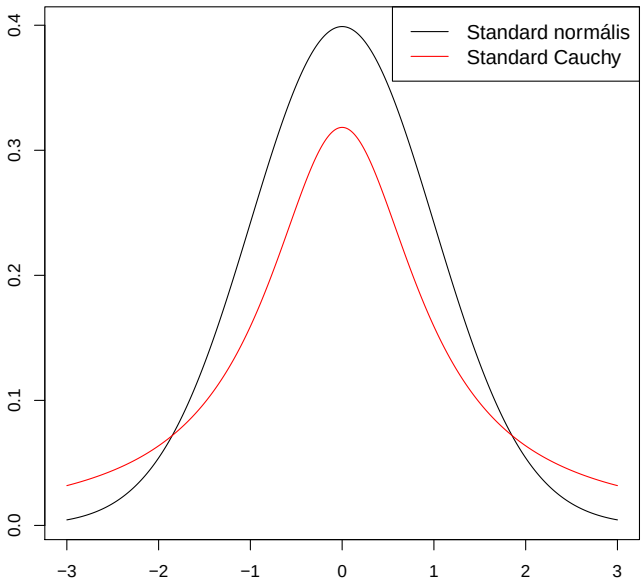
További abszolút folytonos eloszlások

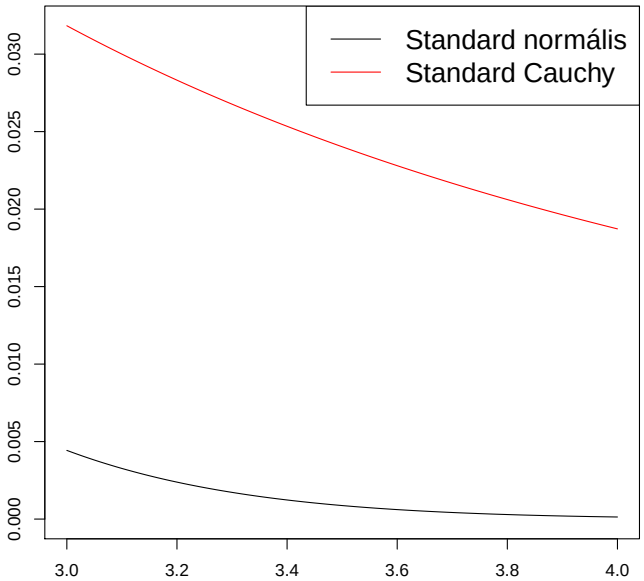
Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	D^2X
Cauchy	*Cauchy*(a, b) $a \in \mathbb{R}$, $b > 0$	$\frac{1}{\pi}\operatorname{arctg}\left(\frac{x-a}{b}\right) + \frac{1}{2}$	$\frac{1}{\pi b\left[1+\left(\frac{x-a}{b}\right)^2\right]}$ $x \in \mathbb{R}$	$\nexists$	$\nexists$
Pareto*	*Pareto*(α, β) $\alpha, \beta > 0$	$\begin{cases} 1 - \left(\frac{\beta}{x}\right)^{\alpha} & \text{ha } x \geq \beta \\ 0 & \text{ha } x < \beta \end{cases}$	$\begin{cases} \frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1} & \text{ha } x \geq \beta \\ 0 & \text{ha } x < \beta \end{cases}$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$

* A Pareto-eloszlásnak akkor van véges várható értéke a képletnek megfelelően, ha $\alpha > 1$, szórásnégyzete pedig akkor, ha $\alpha > 2$.

Eloszlás neve	Jelölése	Sűrűségfüggvény	EX	D^2X
Lognormális	LN(m, σ^2) $m \in \mathbb{R}$, $\sigma > 0$	$\begin{cases} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log x - m)^2}{2\sigma^2}} & \text{ha } x \leq 0 \\ 0 & \text{ha} x < 0 \end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2m+\sigma^2}$
Gamma	$\Gamma(\alpha, \lambda)$ $\alpha, \lambda > 0$	$\begin{cases} \frac{1}{\Gamma(\alpha)}\lambda^{\alpha} e^{-\lambda x} x^{\alpha-1} & \text{ha } x \geq 0 \\ 0 & \text{ha } x < 0 \end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Béta	*Beta*(α, β) $\alpha, \beta > 0$	$\begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in [0,1] \\ 0 & \text{különben} \end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Khí-négyzet	χ^2_k $k \in \mathbb{N}$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$ $x \in \mathbb{R}$	k	$2k$
Student	t_ν $\nu > 0$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	0 (ha $\nu > 1$)	$\frac{\nu}{\nu-2}$ (ha $\nu > 2$)
Fisher	F_{d_1,d_2} $d_1, d_2 > 0$	$\frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1}\left(1+\frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}$	$\frac{d_2}{d_2-2}$ (ha $d_2 > 2$)	$\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ (ha $d_2 > 2$)

Vastag szélű (fat tailed) eloszlások





- végtelen vagy nem létezik a szórás
- fontos vastag szélű eloszlások:
  - Cauchy-eloszlás
  - Pareto-eloszlás
  - Student-féle t-eloszlás alacsony szabadságfok esetén
- "extrém" események, például
  - nagy természeti katasztrófák,
  - atomerőmű-katasztrófák,
  - globális pénzügyi válságok,
  - az Internet összeomlása,
  - sejtekben ritka mutációk stb.

  valószínűségének becslésére jóval alkalmasabbak a normális eloszlásnál

Standard normális versus vastag szélű eloszlások

Legyenek $X \sim \mathrm{N}(0;1)$, $Y \sim \mathrm{Cauchy}(0;1)$, $U \sim \mathrm{Pareto}\left(2;\frac{1}{2}\right)$ és $V \sim t_2$

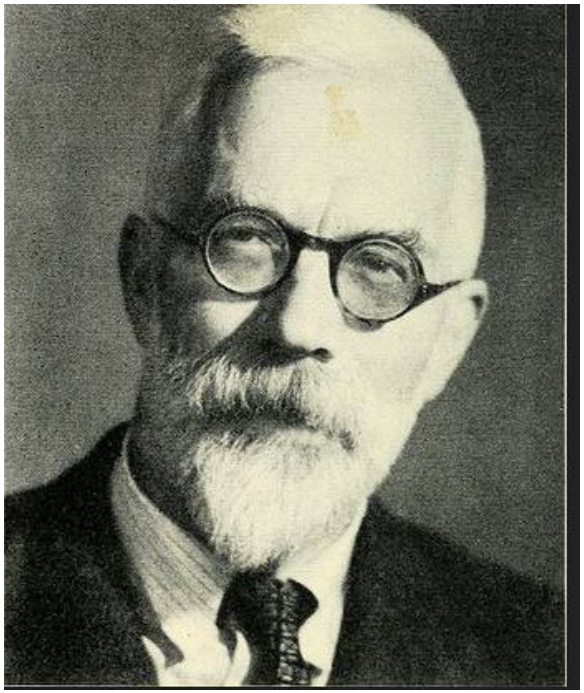
Ekkor a $h_i^Y = \frac{P(Y>i)}{P(X>i)}$, $h_i^U = \frac{P(U>i)}{P(X>i)}$ és $h_i^V = \frac{P(V>i)}{P(X>i)}$ hányadosok alakulása:

i	h_i^Y	h_i^U	h_i^V
2	$6,5$	$2,7$	$4,0$
3	$7,6 \cdot 10^1$	$2,1 \cdot 10^1$	$3,5 \cdot 10^1$
4	$2,5 \cdot 10^3$	$4,9 \cdot 10^2$	$9,0 \cdot 10^2$
5	$2,2 \cdot 10^5$	$3,4 \cdot 10^4$	$6,6 \cdot 10^4$
6	$5,3 \cdot 10^7$	$7,0 \cdot 10^6$	$1,4 \cdot 10^7$
7	$3,5 \cdot 10^{10}$	$4,0 \cdot 10^9$	$7,7 \cdot 10^9$
8	$5,9 \cdot 10^{13}$	$5,9 \cdot 10^{12}$	$1,1 \cdot 10^{13}$

Karl Pearson (1857 – 1936)

- angol matematikus, statisztikus
- a matematika statisztika atyja
- hisztogram
- Pearson-féle korreláció és kapcsolata a lineáris regresszióval
- momentum módszer
- hipotézisvizsgálat elméletének lefektetése, *p*-érték
- χ^2-próba
- főkomponens analízis (principal component analysis, PCA)
- "Statistics is the grammar of science."

Ronald Fisher (1890 – 1962)



- angol statisztikus és biológus
- F-eloszlás, Student-féle t-eloszlás
- elégséges statisztika
- Fisher-információ
- a statisztika bayes-i megközelítése

- diszkriminancia analízis
- extrémérték-elmélet (extreme value theory)
- újramintavételezés – Fisher-féle permutációs teszt

Mintavétel a gyakorlatban

Az adatokkal szemben támasztott követelmények:

- pontosság – ne legyenek hibásak és a szükséges pontosságban álljanak rendelkezésre
- gyorsaság – hamar be lehessen őket szerezni
- gazdaságosság – az adatgyűjtés legyen "olcsó"

Az adatgyűjtés fajtái:

- teljes körű – például a népszámlálás
- részleges – a gyakorlatban ez a jellemző

A részleges adatgyűjtés fajtái:

- reprezentatív (mintavételes): a teljes sokaság jellemzőit megfelelően tükröző részsokaságból, ún. mintasokaságból szerezzük be az adatokat
- monográfia: egy vagy néhány kiemelt egyed részletes vizsgálata
- egyéb – például önkéntes kitöltésen alapuló internetes teszt

Az adatelemzés elemei

1.) Adathibák keresése, irreális adatok, értékek törlése. Ha lehet, akkor a hibák korrigálása.
2.) Ha sok a különböző adat, akkor alkalmas osztályközös gyakorisági sor készítése
3.) Középértékek kiszámítása:
   - átlag (számtani vagy mértani – amelyiknek értelme van)
   - helyzeti középértékek: módusz (az osztályközös gyakorisági sorból) és medián
4.) Szóródási mutatók kiszámítása:
   - szórás és relatív szórás
   - terjedelem és interkvartilis terjedelem
5.) Alakmutatók kiszámítása:
   - ferdeség
   - csúcsosság
6.) Ábrák készítése:
   - hisztogram/sűrűséghisztogram
   - boxplot ábra
   - Lorenz-görbe (értékösszeg sor esetén)
7.) Visszacsatolás $\rightsquigarrow$ a felfedezett adathibák javítása

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.

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

- Számítása osztályközös gyakorisági sorból:

$$\text{Mo} = x_{mo,a} + \frac{d_a}{d_a + d_f} \cdot h_{mo}, \text{ ahol}$$

  - a móduszt tartalmazó osztályköz: amelyikben egységnyi osztályköz hosszra a legnagyobb gyakoriság jut ($\rightsquigarrow$ *korrigált gyakoriságok!*)
  - $x_{mo,a}$: a móduszt tartalmazó osztályköz alsó értéke
  - h_{mo}: a móduszt tartalmazó osztályköz hossza
  - d_a: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül megelőző osztályköz korrigált gyakorisága
  - d_f: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül követő osztályköz korrigált gyakorisága

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

- Számítása osztályközös gyakorisági sorból:

$$\text{Mo} = x_{mo,a} + \frac{d_a}{d_a + d_f} \cdot h_{mo}, \text{ ahol}$$

  - a móduszt tartalmazó osztályköz: amelyikben egységnyi osztályköz hosszra a legnagyobb gyakoriság jut ($\rightsquigarrow$ *korrigált gyakoriságok!*)
  - $x_{mo,a}$: a móduszt tartalmazó osztályköz alsó értéke
  - h_{mo}: a móduszt tartalmazó osztályköz hossza
  - d_a: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül megelőző osztályköz korrigált gyakorisága
  - d_f: a móduszt tartalmazó osztályköz korrigált gyakorisága mínusz a móduszt közvetlenül követő osztályköz korrigált gyakorisága

Középértékek számítása

Jelölje $x_1^* \leq x_2^* \leq \ldots \leq x_n^*$ a rendezett tapasztalati mintát.

Medián: azon ismérvérték, amelynél ugyanannyi kisebb vagy egyenlő, mint nagyobb vagy egyenlő ismérvérték fordul elő a mintában (a "középső" elem)

- Számítása közvetlenül az adatokból:

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

- Számítása osztályközös gyakoriságí sorból – két lépésben lineáris interpolációval:
  1. Melyik osztályközben van a medián: azon i, amire $f'_{i-1} \leq \frac{n}{2}$ és $f'_i \geq \frac{n}{2}$
  2. $\text{Me} = x_{i,a} + \frac{\frac{n}{2} - f'_{i-1}}{f_i} \cdot h_i$, ahol
     - $x_{i,a}$: a mediánt tartalmazó osztályköz alsó értéke
     - h_i: a mediánt tartalmazó osztályköz hossza
     - f'_{i-1}: a mediánt közvetlenül megelőző osztályköz kumulált gyakorisága
     - f_i: a mediánt tartalmazó osztályköz gyakorisága

Tapasztalati kvantilisek számítása

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, mi mindig az egyik interpolációs módszert alkalmazzuk két lépésben:

1. hányadik mintaelem a keresett kvantilis $\rightsquigarrow$ sorszám: $s := (n+1)y$
2. lineáris interpolációval a kvantilis kiszámítása

- Számítása közvetlenül az adatokból
  1. Sorszám: $s = e + t$ (e: egészrész, t: törtrész)
  2. $q_y = x^*_e + t(x^*_{e+1} - x^*_e)$
- Számítása osztályközös gyakorisági sorból – két lépésben lineáris interpolációval:
  1. Melyik osztályközben van az s-edik elem: jelölje ezt i, azaz $f'_{i-1} \leq s$ és $f'_i \geq s$
  2. $q_y = x_{i,a} + \frac{s - f'_{i-1}}{f_i} h_i$, ahol
     $x_{i,a}$, h_i, f'_{i-1} és f_i ugyanazokat jelöli, mint az előző fólia alján, csak az adott y-kvantilisre vonatkozóan

Tapasztalati kvantilisek számítása

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, mi mindig az egyik interpolációs módszert alkalmazzuk két lépésben:

1. hányadik mintaelem a keresett kvantilis $\rightsquigarrow$ sorszám: $s := (n+1)y$
2. lineáris interpolációval a kvantilis kiszámítása

- Számítása közvetlenül az adatokból
  1. Sorszám: $s = e + t$ (e: egészrész, t: törtrész)
  2. $q_y = x^*_e + t(x^*_{e+1} - x^*_e)$
- Számítása osztályközös gyakoriságú sorból – két lépésben lineáris interpolációval:
  1. Melyik osztályközben van az s-edik elem: jelölje ezt i, azaz $f'_{i-1} \leq s$ és $f'_i \geq s$
  2. $q_y = x_{i,a} + \frac{s - f'_{i-1}}{f_i} h_i$, ahol
     $x_{i,a}$, h_i, f'_{i-1} és f_i ugyanazokat jelöli, mint az előző fólia alján, csak az adott y-kvantilisre vonatkozóan

Nevezetes kvantilisek

A szakirodalomban a tapasztalati és az elméleti értékek között nem tesznek különbséget, mindegyiket nagy betűvel írják (ami néha meglehetősen zavaró...). 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)
- kvintilisek: $K_1 = q_{1/5}$, $K_2 = q_{2/5}$, $K_3 = q_{3/5}$, $K_4 = q_{4/5}$
- decilisek: $D_i = q_{i/10}$, $i = 1, 2, \ldots, 9$
- 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 eltérés abszolút mértékegységben

- 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 átlagos eltérés abszolút mértékegységben

- 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

Szóródási mutatók számítása

Relatív szórás vagy **szórási együttható**: az átlagtól való átlagos eltérés százalékban; lehet a korrigált és a korrigálatlan tapasztalati szórásnégyzetből is számítani:

$$V = \frac{s_n^*}{\overline{x}} \text{ vagy } V = \frac{s_n}{\overline{x}}$$

Kevésbé gyakran használt, szóródást mérő mutatók:

- átlagos abszolút eltérés: $\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$
- Gini-együttható: $G = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$.

Alakmutatók számítása

A szórást ezeknél is választhatjuk a tapasztalati vagy a korrigált tapasztalati szórásnak egyaránt.

Tapasztalati ferdeség

- Számítása közvetlenül az adatokból: $\frac{\sum_{i=1}^{n}(x_i-\overline{x})^3}{(s_n)^3}$
- Számítása osztályközös gyakorisági sorból: $\frac{\sum_{i=1}^{n} f_i(x_i-\overline{x})^3}{(s_n)^3}$

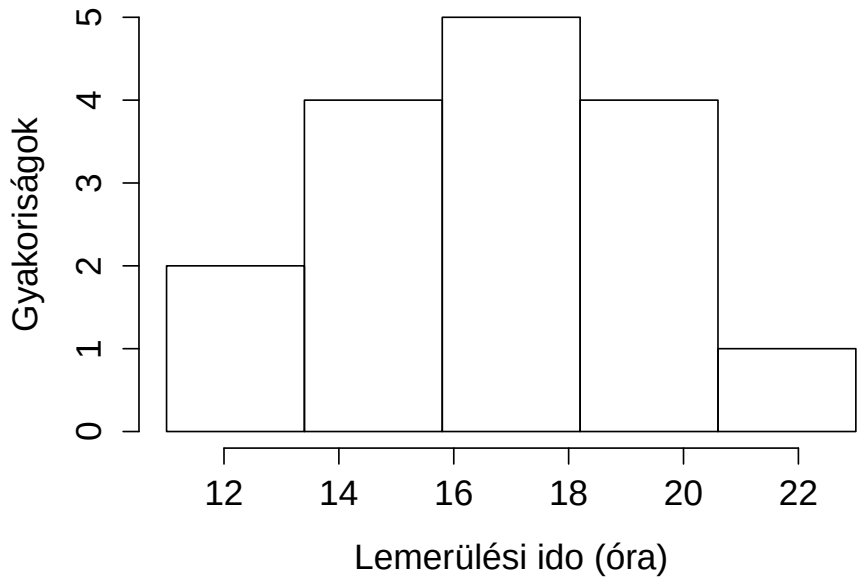
Tapasztalati csúcsosság

- Számítása közvetlenül az adatokból: $\frac{\sum_{i=1}^{n}(x_i-\overline{x})^4}{(s_n)^4} - 3$
- Számítása osztályközös gyakorisági sorból: $\frac{\sum_{i=1}^{n} f_i(x_i-\overline{x})^4}{(s_n)^4} - 3$

Fontos leíró statisztikai ábrák

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.

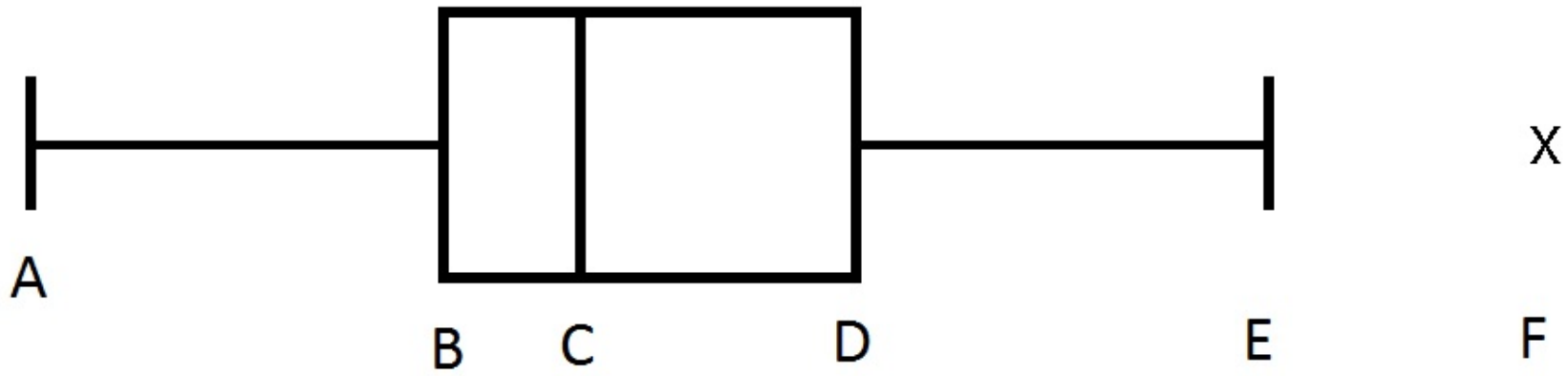
- az osztályok száma: $k = \lfloor \log_2 n \rfloor$
- ha azonos hosszúságú (h) osztályközöket akarunk létrehozni, akkor $h = \frac{x_n^* - x_1^*}{k}$
- az f_i gyakoriságokat ábrázoljuk a függőleges tengelyen
- sűrűséghisztogramnál a $g_i = \frac{f_i}{n}$ relatív gyakoriságokat ábrázoljuk a függőleges tengelyen



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

Fontos leíró statisztikai ábrák

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



A betűk a következő értékeket jelentik:

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

ahol $IQR = Q_3 - Q_1$ az interkvartilis terjedelem

Feladatok

E**12.)** Azonos felhasználási körülmények között megmérték 15 azonos típusú mobiltelefon akkumulátorának lemerülési idejét teljes feltöltöttségről: (óra)

18	16	15	20	12	16	-15	23
14	11	17	15	200	19	18	20

a.) Nézd át nagy vonalakban az adatokat, reálisak-e! Próbáld meg kijavítani az esetleges adathibákat!

b.) Ábrázold a tapasztalati eloszlásfüggvényt! Számítsd ki és értelmezd a 16 helyen!

c.) Készíts alkalmas sávszélességű hisztogramot!

d.) Elemezd a lemerülési időt az alapstatisztikák: az átlag, a korrigált tapasztalati szórás, szórási együttható és boxplot ábra (kvartilisek) segítségével! Számítsd ki a tapasztalati ferdeséget és csúcsosságot! Értelmezd is az eredményeket!

Megoldás (értelmezések)

a.) Adatjavítás: -15 és 200, a helyes értékek vélhetően 15 és 20

b.) Az akkumulátorok 3/8-ad része 16 óránál hamarabb merült le.

d.) Az akkumulátorok átlagosan 16,8 óra alatt merültek le. Az egyes akkumulátorok lemerülési ideje az átlagos lemerülési időtől átlagosan 3,19 órával, azaz 18,96%-kal tért el.
Az akkumulátorok egynegyede legfeljebb 15 óra alatt lemerült, míg háromnegyede legalább 15 órán keresztül ébren volt. Az akkumulátorok egyik fele legfeljebb 16,5 óra alatt lemerült, míg másik fele legalább 16 és fél órán keresztül tudta árammal ellátni a telefont. Az akkumulátorok 75%-a legfeljebb 19,75 óra alatt lemerült.
Az akkumulátorok lemerülési idejének eloszlása nagyjából szimmetrikus, csúcsossága a normális eloszláséhoz viszonyítva laposabb.

Tipikus rossz értelmezések: (válogatások a tavalyi beadandókból, a mostani feladat szövegkörnyezetébe ültetve)

- Az akkumulátorok átlaga 16,8 óra. A lemerülési idő szórása 3,19 óra. Az alsó kvartilis 15 óra.
- Az adatok átlaga 16,8 óra.
- A minta relatív szórása 4,8%, tehát az átlagtól való eltérés 4,8%.
- A relatív szórás azt fejezi ki, hogy az egyes ismérvértékek átlagosan hány százalékkal térnek el az átlagtól.
- A medián 16,5, ami a középső mintaelem a rendezett mintában (az elemek nemcsökkenő sorban helyezkednek el). Az alsó kvartilis 15, ami a legkisebb és a medián között középen elhelyezkedő adat számértéke a rendezett mintában.
- A standard normális elosztáshoz képest a nikkel elosztása 0,002%-kal jobbra ferdül és 0,836%-kal laposabb.
- A hármas kvartilis értéke 19,75 óra, azaz ...

Feladatok

E**13.)** Egy megyében a kistermelő gazdaságok termőterület szerinti megoszlása:

Termőterület (hektár)	Gazdaságok száma
– 4	200
4 – 10	90
10 – 20	80
20 – 30	60
30 – 50	10
50 –	10
Összesen	450

a.) Készíts hisztogramot! Milyen az eloszlás ferdesége?

b.) Jellemezd (szövegesen is) a kistermelők termőterület szerinti eloszlását alapstatisztikák (mintaátlag, korrigált tapasztalati szórás, tapasztalati módusz és kvartilisek) segítségével!

c.) Mennyire koncentrálódik a termőterület? Készíts Lorenz-görbét!

Motiváció – becsléselmélet

Az Asus kicseréli táblagépeit, amennyiben a vevők 8-nál több pixelhibát jelentenek be vásárlástól számítva 3 napon belül. A Samsung már egyetlen, 3 napon belül bejelentett pixelhiba esetén is új készüléket biztosít. A Sony-nál legalább 2 pixelhiba esetén jár új táblagép.
Hogyan tudnánk megbecsülni, hogy a gyártónak éves szinten milyen mértékű vesztesége származik ezekből a cserékből?

- Kulcskérdés: mi az esélye, hogy egy, a gyártósorról véletlenszerűen leemelt készüléket pixelhiba miatt ki kell cserélni?
- Ha X a pixelhibák száma, akkor a kérdéses valószínűség például a Sony-nál: $P(X \geq 2)$
- Milyen eloszlású lehet X (Poisson?) $\rightsquigarrow$ *illeszkedésvizsgálat*
- Ha tudom, hogy Poisson-eloszlású, akkor hogyan becsüljem meg a paramétert? $\rightsquigarrow$ *pontbecslés*
- Milyen intervallumban lesz "nagy" valószínűséggel a becsült paraméter? $\rightsquigarrow$ *intervallumbecslés*
- Ezután készíthető a kérdéses valószínűségre intervallumbecslés, abból pedig egy intervallumbecslés a várható veszteségre.

Forrás: http://www.origo.hu/techbazis/20130329-ra-lehet-fazni-a-pixelhibas-okostelefonnal.html

Motiváció – becsléselmélet

Az Asus kicseréli táblagépeit, amennyiben a vevők 8-nál több pixelhibát jelentenek be vásárlástól számítva 3 napon belül. A Samsung már egyetlen, 3 napon belül bejelentett pixelhiba esetén is új készüléket biztosít. A Sony-nál legalább 2 pixelhiba esetén jár új táblagép.
Hogyan tudnánk megbecsülni, hogy a gyártónak éves szinten milyen mértékű vesztesége származik ezekből a cserékből?

- Kulcskérdés: mi az esélye, hogy egy, a gyártósorról véletlenszerűen leemelt készüléket pixelhiba miatt ki kell cserélni?
- Ha X a pixelhibák száma, akkor a kérdéses valószínűség például a Sony-nál: $P(X \geq 2)$
- Milyen eloszlású lehet X (Poisson?) $\rightsquigarrow$ *illeszkedésvizsgálat*
- Ha tudom, hogy Poisson-eloszlású, akkor hogyan becsüljem meg a paramétert? $\rightsquigarrow$ *pontbecslés*
- Milyen intervallumban lesz "nagy" valószínűséggel a becsült paraméter? $\rightsquigarrow$ *intervallumbecslés*
- Ezután készíthető a kérdéses valószínűségre intervallumbecslés, abból pedig egy intervallumbecslés a várható veszteségre.

Forrás: `http://www.origo.hu/techbazis/20130329-ra-lehet-fazni-a-pixelhibas-okosmobillal.html`

Fontos becslések tulajdonságai

Tétel. Legyen $X_1, \ldots, X_n$ i.i.d. minta egy ϑ paraméterű eloszláscsaládból, $h : \mathbb{R} \to \mathbb{R}$ (mérhető) függvény. Tegyük fel, hogy a táblázatban szereplő összes várható érték/szórás létezik minden ϑ esetén.

Mit becsülünk? $g(\vartheta)$	Ha mivel becsüljük? $T_n(\mathbf{X})$	Torzítatlan?	Aszimptotikusan torzítatlan?	Gyengén/erősen konzisztens?
$E_\vartheta X_1$	$\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$	igen	igen	igen
$D^2_\vartheta X_1$	$S_n^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$	nem	igen	igen
$D^2_\vartheta X_1$	$(S_n^*)^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$	igen	igen	igen
$F_\vartheta(x)$	$F_n(x) = \frac{\sum_{i=1}^n I(X_i < x)}{n}$	igen	igen	igen
$E_\vartheta h(X_1)$	$\frac{\sum_{i=1}^n h(X_i)}{n}$	igen	igen	igen

Feladatok

E14.) El szeretnénk dönteni egy érméről, hogy az szabályos-e, avagy cinkelt. Írjuk fel a problémát leíró statisztikai mezőt!

E15.) Legyen $X_1, \ldots, X_n$ i.i.d. Bin$(4; p)$ eloszlású valószínűségi változó, ahol $p \in (0; 1)$ ismeretlen valós paraméter.

a.) Adjuk meg a mintateret és a paraméterteret!
b.) Határozzuk meg az ismeretlen paraméter ML-becslését!
c.) Határozzuk meg az ismeretlen paraméter momentum becslését!
d.) Torzítatlan, illetve konzisztens az ML-becslés? Amennyiben nem torzítatlan, tegyük azzá!
e.) Adj torzítatlan becslést $g(p) = p^2(1-p)^2$-re!

E16.) Legyen X_1 Bin$(2; p)$ eloszlású (egyelemű) minta, ahol $p \in (0; 1)$ ismeretlen valós paraméter. Adj X_1 segítségével torzítatlan becslést $g(p) = \frac{1}{p}$-re!

Feladatok

E**17.)** Minden nap a Mester utca megállónál szállok fel a 4-es/6-os villamosok valamelyikére. E hét munkanapjain az alábbi várakozási időket mértem (perc): 1,2 2 1,5 3 2,1
A várakozási időről tegyük fel, hogy exponenciális eloszlású.

a.) Adjuk meg a mintateret és a paraméterteret!
b.) Határozzuk meg az ismeretlen paraméter ML-becslését!
c.) Határozzuk meg az ismeretlen paraméter momentum-becslését!
d.) Szimulációval vizsgáljuk meg, hogy 10, 20, 50 és 100 elemű exponenciális mintából számolt ML-becslés torzítatlanul becsüli-e az ismeretlen paramétert!
e.) Torzítatlan, illetve konzisztens az ML-becslés? Amennyiben nem torzítatlan, tegyük azzá!
f.) Mutassuk meg, hogy az $S(\mathbf{X}) = n \cdot X_1^*$ statisztika torzítatlan, de nem konzisztens becslése $g(\vartheta) = \frac{1}{\lambda}$-nak!

E18.) Legyen $X_1, \ldots, X_n$ i.i.d. abszolút folytonos valószínűségi változók sorozata.

a.) Adjuk meg az i-edik rendezett mintaelem, azaz X_i^* eloszlás- és sűrűségfüggvényét ($1 \leq i \leq n$)!

b.) Milyen eloszlású X_i^*, amennyiben a mintaelemek $(0; 1)$-en egyenletesek? Határozzuk meg az EX_i^* mennyiséget!

E19.) Egy véletlen szám generátorral 20 véletlen számot állítunk elő egy ismeretlen (a, b) intervallumból. A kapott véletlen számok sorrendbe téve és (egyszerűség kedvéért) egészre kerekítve:

5	11	12	13	13	14	17	19	21	22
23	24	25	27	31	31	32	35	36	38

a.) Adjuk meg a mintateret és a paraméterteret!

b.) Határozzuk meg az ismeretlen paraméterek maximum likelihood becslését!

c.) Határozzuk meg a paraméterek momentum-becslését! Értékeljük a kapottakat!

Sűrűségfüggvény becslése – magfüggvényes módszer (Parzen-Rosenblatt becslés)

$$f_n(x) = \frac{1}{n \cdot h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \text{ ahol}$$

- $K : \mathbb{R} \to \mathbb{R}$ páros függvény, neve: magfüggvény
- h_n sávszélesség, rendszerint $h_n = n^c$, ahol $-1 < c < 0$ valós szám

A leggyakoribb magfüggvények sűrűségfüggvény becslésére:

Magfüggvény neve	$K(x)$
Gauss	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
Téglalap (rectangular)	$\frac{1}{2} I(\lvert x \rvert \leq 1)$
Háromszög (triangular)	$(1 - \lvert x \rvert) \cdot I(\lvert x \rvert \leq 1)$
Bartlett–Epanechnikov	$\frac{3}{4}(1 - x^2) \cdot I(\lvert x \rvert \leq 1)$
Cosinus	$\frac{\pi}{4} \cos\left(\frac{\pi}{2} x\right) \cdot I(\lvert x \rvert \leq 1)$

Mi az "optimális" sávszélesség? Mi az, hogy "optimális"? $\rightsquigarrow$ amivel $f_n(x)$ "legjobban" közelíti a valódi sűrűségfüggvényt.

A Fisher-információ alkalmazásai

- hatásos becslés keresése az információs határ segítségével

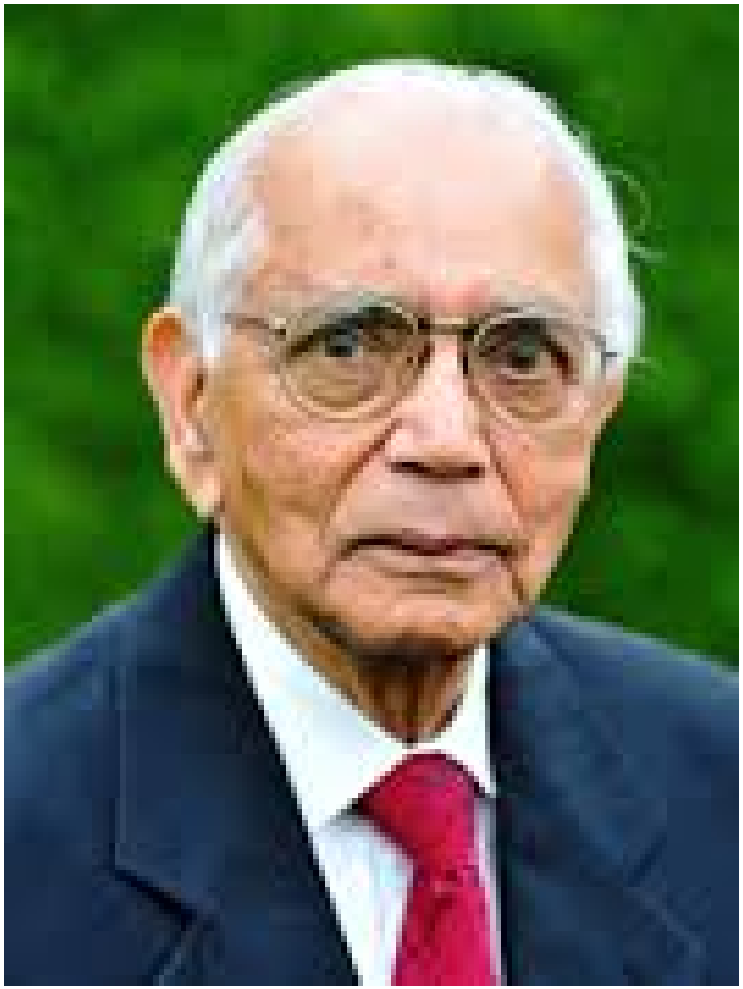
  Cramér-Rao egyenlőtlenség: $D^2_\vartheta(T(\mathbf{X})) \geq \underbrace{\frac{(g'(\vartheta))^2}{I_n(\vartheta)}}_{\text{információs határ}}$

  Ha egy $g(\vartheta)$-ra nézve torzítatlan T statisztika esetén egyenlőség teljesül, akkor az a statisztika hatásos becslése $g(\vartheta)$-nak.
- az ML-becslés eloszlásban egy olyan normális eloszláshoz tart, aminek a szórásnégyzete a Fisher-információ inverze
- intervallumbecslés az ML-becslésre
- kísérlettervezés
- bayes-i statisztika – Jeffrey-féle apriori eloszlás számításához
- neurális hálók, machine learning
- számítógépes agykutatás

Az információs határ felfedezői/névadói

Harald Cramér (1893 – 1985)

C. R. Rao (1920 –)



Hatásos becslés keresése az információs határral

E**20.)** Legyen $X_1, \ldots, X_n$ i.i.d. Ind(p) eloszlású.

a.) Határozzuk meg a mintában lévő Fisher-információ értékét!

b.) Mutassuk meg, hogy a relatív gyakoriság hatásos becslése a valószínűségnek!

E**21.)** Legyen $X_1, \ldots, X_n$ i.i.d. Exp(λ) eloszlású.

a.) Határozzuk meg a mintában lévő Fisher-információ értékét!

b.) Mutassuk meg, hogy a mintaátlag hatásos becslése $\frac{1}{\lambda}$-nak!

Intervallumbecslések normális eloszlású minta esetén

Legyen $X_1, \ldots, X_n \sim N(m, \sigma^2)$ i.i.d. minta, $\alpha > 0$ "kicsi" valós szám.

Kétoldali $(1-\alpha)$-konfidenciaintervallumok:

- m-re
  - ha σ ismert, akkor $\boxed{\overline{X} \pm u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$ [68.1.]
  - ha σ ismeretlen, akkor $\overline{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{S_n^*}{\sqrt{n}}$ [68.2.]
- σ^2-re: $\left[\frac{(n-1)\cdot(S_n^*)^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \; , \; \frac{(n-1)\cdot(S_n^*)^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \right]$ [68.3.]

Egyoldali (alsó) $(1-\alpha)$-konfidenciaintervallumok:

- m-re
  - ha σ ismert, akkor $\left[-\infty \; , \; \overline{X} + u_\alpha \frac{\sigma}{\sqrt{n}} \right]$ [68.4.]
  - ha σ ismeretlen, akkor $\left[-\infty \; , \; \overline{X} + t_{n-1,\alpha} \frac{S_n^*}{\sqrt{n}} \right]$ [68.5.]
- σ^2-re: $\left[-\infty \; , \; \frac{(n-1)\cdot(S_n^*)^2}{\chi^2_{n-1,\alpha}} \right]$ [68.6.]

Intervallumbecslések lefedési valószínűsége

E**22.)** Generáljunk $n = 5, 10, 20, 50, 100$ elemű mintát
a.) $N(1, 2^2)$;
b.) Exp(2);
c.) $E(1; 5)$
eloszlásból 10^4 alkalommal, majd becsüljük meg a várható értékre adott $\overline{x} \pm \frac{\sigma}{\sqrt{n}} u_{\alpha/2}$ intervallum lefedési valószínűségét $\alpha = 0,01,\ 0,05$ és $0,1$ esetén, ahol $u_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$!

E**23.)** Generáljunk $n = 5, 10, 20, 50, 100, 200, 500, 1000$ elemű mintát Exp(λ) eloszlásból 10^4 alkalommal, majd adjunk az ismeretlen λ paraméter ML-becslésére intervallumbecslést annak aszimptotikus eloszlása segítségével! Becsüljük meg az intervallumok lefedési valószínűségét $\alpha = 0,01,\ 0,05$ és $0,1$ esetén!
Kifejtés után: $\hat{\lambda}_n \pm \frac{\lambda_0 \cdot u_{\alpha/2}}{\sqrt{n}} \approx \hat{\lambda}_n \pm \frac{\hat{\lambda}_n \cdot u_{\alpha/2}}{\sqrt{n}} = \hat{\lambda}_n \cdot \left(1 \pm \frac{u_{\alpha/2}}{\sqrt{n}}\right)$

Hipotézisvizsgálati alapfogalmak

- **Hipotézis**: egy állítás, aminek igazságát vizsgálni szeretnénk. Egy hipotézist vagy elfogadunk, vagy elutasítunk/elvetünk.
- A paraméterteret diszjunkt részekre bontjuk: $\Theta = \Theta_0 \cup^* \Theta_1$
- A hipotézisvizsgálati alapfeladat (absztraktul, a gyakorlatban konkretizálni szoktuk)

  $H_0 : \vartheta \in \Theta_0$ $\rightsquigarrow$ nullhipotézis

  $H_1 : \vartheta \in \Theta_1$ $\rightsquigarrow$ ellenhipotézis vagy alternatív hipotézis
- A nullhipotézis esetén az elfogadás helyett helyesebb azt mondani, hogy nem tudjuk elvetni. Az okokról később.
- A H_0 hipotézisnek azon állítást szokás választani,
  - ami sok éves tapasztalatnak felel meg
  - amit "remélünk", hogy teljesül
  - amit elutasítva, gyakran negatív következményekkel jár (büntetés, bírság, jobb modell keresésének kényszere stb.)

Hipotézisvizsgálati alapfogalmak

Hogyan döntsünk? Vajon H_0 igaz, vagy H_1? $\rightsquigarrow$ jó lenne valamilyen matematikai eljárás

- Statisztikai próba vagy röviden **próba**: az a módszer/eljárás, amely során a minta segítségével döntést hozunk a hipotézis(ek)ről.
- **Paraméteres próba**: Olyan próba, amely során a feladatban lévő ismeretlen eloszlás jellege ismert, és a nullhipotézis az eloszlás valamely paraméterére (vagy annak egy minket érdeklő függvényére) vonatkozik.

- Mintatér felbontása két diszjunkt részre: $\mathcal{X} = \mathcal{X}_e \cup^* \mathcal{X}_k$
- $\mathcal{X}_k$: **kritikus tartomány** – azon **x** megfigyelések halmaza, amikre *elutasítjuk* a nullhipotézist
- $\mathcal{X}_e$: **elfogadási tartomány** – azon **x** megfigyelések halmaza, amikre *elfogadjuk* a nullhipotézist

Ez oké, de mi alapján rakjunk egy **x** megfigyelt mintát $\mathcal{X}_k$-ba vagy $\mathcal{X}_e$-be? $\rightsquigarrow$ 2 fóliával később

Hipotézisvizsgálati alapfogalmak

- Döntési mátrix hipotézisvizsgálat esetén:

Döntés / "Valóság"	H_0-t elfogadjuk ($\mathcal{X}_e$)	H_0-t elutasítjuk ($\mathcal{X}_k$)
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

- **Elsőfajú hiba** (type I. error): a nullhipotézist elvetettük, de nem szabadott volna, mert a H_0-beli állítás igaz
  Valószínűsége: $\alpha(\vartheta) := P_\vartheta(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_0$
  További szokásos jelölések: $\alpha(\vartheta) = P_{\vartheta\in\Theta_0}(\mathcal{X}_k) = P_{H_0}(\mathcal{X}_k) = P_0(\mathcal{X}_k)$
- **Másodfajú hiba** (type II. error): a nullhipotézist elfogadtuk, de nem szabadott volna, mert a H_0-beli állítás hamis
  Valószínűsége: $\beta(\vartheta) := P_\vartheta(\mathcal{X}_e)$, ahol $\vartheta \in \Theta_1$
  További szokásos jelölések: $\beta(\vartheta) = P_{\vartheta\in\Theta_1}(\mathcal{X}_e) = P_{H_1}(\mathcal{X}_e) = P_1(\mathcal{X}_e)$
- **Erőfüggvény**: $\psi(\vartheta) := P_\vartheta(\mathcal{X}_k)$, ahol $\vartheta \in \Theta_1$

Hipotézisvizsgálati alapfogalmak

- **Terjedelem**: $\alpha := \sup_{\vartheta \in \Theta_0} \alpha(\vartheta)$

  Hosszabban: a próba pontos terjedelmének is hívják
- A hipotézisvizsgálati feladat elején rögzíteni szokás a terjedelmet, tipikusan 5%-on (esetleg más szám 1% és 10% között). Ezáltal döntésünket
  - 5%-os elsőfajú hiba valószínűsége mellett, vagy másképp:
  - 95%-os megbízhatósággal

  fogjuk meghozni.

Hipotézisvizsgálati alapfogalmak

- Legyen $H_0 : \vartheta \in \Theta$
  - H_0 **egyszerű**, ha $|\Theta_0| = 1$ (egyelemű)
  - H_0 **összetett**, ha $|\Theta_0| > 1$ (legalább kételemű)
- **Kétoldali próba**: $H_0 : \vartheta = \vartheta_0 \quad H_1 : \vartheta \neq \vartheta_0$
- **Egyoldali próba**: $H_0 : \vartheta = \vartheta_0 \quad H_1 : \vartheta > \vartheta_0$ (vagy $H_1 : \vartheta < \vartheta_0$)
- **Próbastatisztika**: Olyan alkalmas statisztika, amely segítségével a kritikus tartományt meghatározzuk.
  - Ez jellemzően úgy szokott menni, hogy valós értékű $T : \mathcal{X} \to \mathbb{R}$ próbastatisztikát választunk, majd az alábbi alakú kritikus tartományok közül keressük valamelyiket:

    $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \ T(\mathbf{x}) > c\}$ (egyoldali próbánál)
    $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \ T(\mathbf{x}) < c\}$ (egyoldali próbánál)
    $\mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \ |T(\mathbf{x})| > c\}$ (kétoldali próbánál)
  - c neve: **kritikus érték**, ami jellemzően függ a próba terjedelmétől, ezért c_α-val jelöljük. Ez általában arra utal, hogy c_α a $T(\mathbf{X})$ valószínűségi változó α-kvantilise.
  - A próba meghatározása: előre rögzített α terjedelemhez azt a c_α értéket keressük, amire a próba terjedelme éppen α: $\sup_{\vartheta \in \Theta_0} P_\vartheta(T(\mathbf{X}) > c_\alpha) = \alpha$.

Feladatok

E**24.)** Minőségellenőrként az a feladatunk, hogy a gyártósorokat szükség esetén leállítsuk, amennyiben túl sok selejtes termék kerül le róluk. A megengedett selejtarány legfeljebb 5%. Összesen 25 terméket vizsgálunk meg. Ha a selejtesek száma legalább k, akkor leállítjuk a gyártósort. Amennyiben kevesebb selejtesünk van k-nál, akkor a termelés mehet tovább.

a.) Írjuk fel a hipotéziseket és a próbát!

b.) Határozzuk meg az elsőfajú hibát!

c.) Mely k érték esetén lesz az elsőfajú hiba valószínűsége legközelebb 5%-hoz? Adjuk meg azt a véletlenített próbát, amivel a terjedelem pontosan 5%!

A továbbiakban tegyük fel, hogy a vizsgált gyártósoron a termékek meghibásodásának valószínűsége 0,1.

d.) Határozzuk meg a másodfajú hibát!

e.) Ábrázoljuk az erőfüggvényt különböző k értékek esetén!

f.) Mely k érték esetén lesz az elsőfajú és a másodfajú hiba valószínűségének összege minimális?

Feladatok

E25.) Az (A) gép által termelt termékek egy bizonyos jellemzője $N(11, 1^2)$, míg a (B) gépen termelt termékeké $N(13, 4^2)$ eloszlású.
Legyenek
H_0 : a mintánk az (A) gépen készült
H_1 : a mintánk a (B) gépen készült
Ha egy 16 elemű minta átlaga legfeljebb 12, akkor elfogadjuk H_0-t, különben elvetjük.

a.) Mekkora az elsőfajú és a másodfajú hiba valószínűsége?
b.) Milyen c értéket adjunk meg a 12 helyett ahhoz, hogy $0,05$ legyen a próba terjedelme? Ekkor mennyi a másodfajú hiba valószínűsége?
c.) Milyen c értéket adjunk meg a 12 helyett ahhoz, hogy az elsőfajú és a másodfajú hiba valószínűségének összegét minimalizáljuk?

A hipotézisvizsgálat menete I.

1.) A terjedelem (α) lefixálása, ami jellemzően 1% és 10% közötti, tipikusan 5%
Megbízhatóság$= 1 - \alpha$, általában %-osan írjuk
2.) Nullhipotézis (H_0) felírása – sokévi, megszokott, elvárt értékeknek megfelelő paramétertartomány
3.) Alternatív hipotézis (H_1) felírása – a minta alapján bennünket érdeklő kérdésnek megfelelő paramétertartomány
4.) A probléma megoldására alkalmas próba vagy próbák kiválasztása – feltételek ellenőrzése
5.) Próbastatisztika kiszámítása
6.) Kritikus érték kiszámítása, kritikus tartomány ($\mathcal{X}_k$) megállapítása
7.) Döntés:

- $\mathbf{x} \in \mathcal{X}_k \rightsquigarrow$ **erős döntés**, H_1-et elfogadjuk, H_0-t elvetjük/elutasítjuk
- $\mathbf{x} \in \mathcal{X}_e \rightsquigarrow$ **gyenge döntés**, H_1-et elutasítjuk, H_0-t nem tudjuk elutasítani

A hipotézisvizsgálat menete II.

1.) A terjedelem (α) lefixálása
2.) Nullhipotézis (H_0) felírása
3.) Alternatív hipotézis (H_1) felírása
4.) A probléma megoldására alkalmas próba vagy próbák kiválasztása
5.) Számítógéppel dolgozva, az előző fólián lévő 5.)-6.)-7.) helyett dönthetünk az ún. p-érték alapján is:

$$\boxed{p\text{-érték} < \alpha \quad \Leftrightarrow \quad \mathbf{x} \in \mathcal{X}_k \quad \Leftrightarrow \quad H_1\text{-et elfogadjuk}}$$

p-érték: az a terjedelem, amire a kritikus érték megegyezik a próbastatisztikával

Ha például p-érték$= 0.06$, akkor 5%-os elsőfajú hiba valószínűsége mellett nem tudjuk elvetni H_0-t, de 10%-os elsőfajú hiba valószínűsége esetén már elvetjük H_0-t.
Ha például p-érték$= 0.16$, akkor a hagyományos, értelmes – 90% és 99% közötti – megbízhatósági szinteken nem tudjuk elvetni H_0-t.

Nevezetes paraméteres próbák – áttekintés

Próbák a normális eloszlás várható értékére vonatkozóan:

- Egymintás *u*-próba, egymintás *t*-próba
- Kétmintás próbák:

	a két minta független		a két minta nem független
σ_1 és σ_2 ismert	kétmintás *u*-próba		egymintás *u*-próba a különbségekre
σ_1 és σ_2 ismeretlen	előzetes *F*-próba		egymintás *t*-próba a különbségekre
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	
	kétmintás t-próba	Welch-próba	

Próbák normális eloszlás szórásnégyzetére vonatkozóan:

- Egymintás próba: χ^2-próba
- Kétmintás próba: *F*-próba

Összefüggő (páros) minták: X_i és Y_i ugyanahhoz, az i-edik személyhez, tárgyhoz, objektumhoz tartozó véletlen mennyiség, $i = 1, 2, \ldots$

Nevezetes paraméteres próbák

Egymintás u-próba

$X_1, \ldots, X_n \sim N(m, \sigma^2)$, ahol σ ismert, m ismeretlen paraméter

Kétoldali: $H_0 : m = m_0$
$H_1 : m \neq m_0$

Próbastatisztika: $T(\mathbf{X}) = u := \sqrt{n}\frac{\overline{X} - m_0}{\sigma} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$
Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : |u| > u_{\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak:	$H_1 : m > m_0$	$H_1 : m < m_0$
Kritikus tartomány:	$\mathcal{X}_k = \{\mathbf{x} : u > u_\alpha\}$	$\mathcal{X}_k = \{\mathbf{x} : u < -u_\alpha\}$

Áttekintés

Nevezetes paraméteres próbák

Egymintás t-próba

$X_1, \ldots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen paraméterek

Kétoldali: $H_0 : m = m_0$
$H_1 : m \neq m_0$

Próbastatisztika: $T(\mathbf{X}) = t := \sqrt{n}\frac{\overline{X} - m_0}{S_n^*} \overset{H_0 \text{ esetén}}{\sim} t_{n-1}$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : |t| > t_{n-1,\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak:	$H_1 : m > m_0$	$H_1 : m < m_0$
Kritikus tartomány:	$\mathcal{X}_k = \{\mathbf{x} : t > t_{n-1,\alpha}\}$	$\mathcal{X}_k = \{\mathbf{x} : t < -t_{n-1,\alpha}\}$

Áttekintés

Nevezetes paraméteres próbák

Kétmintás u-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták
ahol m_1, m_2 ismeretlen paraméterek, σ_1, σ_2 ismert

Kétoldali: $H_0 : m_1 = m_2$
$H_1 : m_1 \neq m_2$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = u := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |u| > u_{\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Kritikus tartomány:	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : u > u_\alpha\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : u < -u_\alpha\}$

Áttekintés

Nevezetes paraméteres próbák

Kétmintás t-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták
ahol m_1, m_2, $\sigma_1 = \sigma_2$ ismeretlen paraméterek

Kétoldali: $H_0 : m_1 = m_2$
$H_1 : m_1 \neq m_2$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = t := \sqrt{\frac{nm}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)(S_1^*)^2 + (m-1)(S_2^*)^2}{n+m-2}}} \overset{H_0 \text{ esetén}}{\sim} t_{n+m-2}$

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |t| > t_{n+m-2,\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak:	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$
Krit. tart.:	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t > t_{n+m-2,\alpha}\}$	$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t < -t_{n+m-2,\alpha}\}$

Áttekintés

Nevezetes paraméteres próbák

Welch-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$ és $Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták ahol m_1, m_2, $\sigma_1 \neq \sigma_2$ ismeretlen paraméterek

Kétoldali: $H_0 : m_1 = m_2$

$H_1 : m_1 \neq m_2$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = t' := \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(S_1^*)^2}{n} + \frac{(S_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} t_f$, ahol

$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$, $\quad c = \frac{\frac{(s_1^*)^2}{n}}{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}$, ha $s_1^* > s_2^*$ (így csináljuk)

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |t| > t_{f,\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak: $H_1 : m_1 > m_2$ $\qquad$ $H_1 : m_1 < m_2$

Krit. tartomány: $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t > t_{f,\alpha}\}$ $\qquad$ $\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : t < -t_{f,\alpha}\}$

Áttekintés

Nevezetes paraméteres próbák

χ^2**-próba** (normális eloszlás szórására)

$X_1, \ldots, X_n \sim N(m, \sigma^2)$, ahol m és σ ismeretlen paraméterek

Kétoldali: $H_0 : \sigma = \sigma_0$
$H_1 : \sigma \neq \sigma_0$

Próbastatisztika: $T(\mathbf{X}) = h := \frac{(n-1)(S_n^*)^2}{\sigma_0^2} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{n-1}$

Kritikus tartomány: $\mathcal{X}_k = \left\{ \mathbf{x} : h < \chi^2_{n-1,\alpha/2} \text{ vagy } h > \chi^2_{n-1,1-\alpha/2} \right\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak: $H_1 : \sigma > \sigma_0$ $\quad$ $H_1 : \sigma < \sigma_0$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : h > \chi^2_{n-1,1-\alpha}\}$ $\quad$ $\mathcal{X}_k = \{\mathbf{x} : h < \chi^2_{n-1,\alpha}\}$

Áttekintés

Nevezetes paraméteres próbák

F-próba

$X_1, \ldots, X_n \sim N(m_1, \sigma_1^2)$
$Y_1, \ldots, Y_m \sim N(m_2, \sigma_2^2)$ független minták
ahol m_1, m_2, σ_1, σ_2 ismeretlen paraméterek

Kétoldali: $H_0 : \sigma_1 = \sigma_2$
$\underline{H_1 : \sigma_1 \neq \sigma_2}$

Próbastatisztika: $T(\mathbf{X}, \mathbf{Y}) = F = \frac{(S_1^*)^2}{(S_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1}$

Kritikus tartomány:
$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : F < F_{n-1,m-1,\alpha/2} \text{ vagy } F > F_{n-1,m-1,1-\alpha/2}\}$

Egyoldali próbák esetén H_0 és a próbastatisztika ugyanaz marad, csak H_1, és ezáltal a kritikus tartomány változik.

Egyoldaliak:	$H_1 : \sigma_1 > \sigma_2$	$H_1 : \sigma_1 < \sigma_2$
Krit. tart.:	$\{(\mathbf{x}, \mathbf{y}) : F > F_{n-1,m-1,1-\alpha}\}$	$\{(\mathbf{x}, \mathbf{y}) : F < F_{n-1,m-1,\alpha}\}$

Áttekintés

Feladatok

E**26.)** A fogyasztóvédelmi hatóság többszöri lakossági bejelentést kapott, hogy a Portokall nevű, fél literes kiszerelésű narancsitalokban a flakonra írt 500 ml-nél jóval kevesebb üdítő van. Ez alapján vizsgálatot kezdtek, a fogyasztóvédelem munkatársa vásárolt a boltban 10 darabot, majd megnézte a benne lévő édes nedű térfogatát (ml):
483, 502, 498, 496, 502, 483, 494, 491,505, 486.
Tegyük fel, hogy egy fél literes üdítős üvegbe töltött narancslé mennyisége normális eloszlást követ.
Állíthatjuk-e 95%-os megbízhatóság esetén, hogy a Portokall gyártója át akarja verni a vevőket?

E**27.)** Használjuk az első előadáson kitöltött kérdőív eredményeit, és próbáljunk meg az alapján a TTK hallgatóira vonatkozólag következtetéseket levonni! Állíthatjuk-e, hogy a TTK-n
a.) a fiúk legalább 10 cm-rel magasabbak a lányoknál;
b.) a lányok magasságának szórása k cm, ahol $k = 1, 2, \ldots, 15$?

Feladatok

E**28.)** Bálint gazdának 66 tehene van, teheneit reggel kitereli nagy birtokára, és egész nap ott legelésznek. Este összefut a helyi kocsmában a szomszéd gazdálkodóval, Máté gazdával, aki elmeséli, a tehenei tejének tejzsírszázaléka jelentősen megnőtt, mióta szilázzsal is eteti őket minden nap. Ezen felbuzdulva, Bálint gazda úgy dönt, hogy 6 kedvenc tehenén kipróbálja ezt a "diétát" – egy hónapon keresztül szilázzsal is etette őket, majd megnézte a tejük tejzsírszázalékát:

Mit ettek	Julcsa	Bogár	Riska	Csendes	Bimbó	Mula
Csak füvet	3,84	3,79	3,78	4,00	3,83	3,84
Szilázst is	3,90	4,05	3,8	4,01	3,81	3,9

Vizsgáljuk meg alkalmas statisztikai próbával, hogy a szilázs növeli-e a tej tejzsírszázalékát!

Feladatok

Az előző feladat végeredménye (számítógéppel megoldva):
p-érték$= 0,0867$ (egymintás t-próba a különbségekre)

Helyes, a szövegkörnyezetbe ágyazott szöveges értelmezések:

- 95%-os megbízhatósággal elutasíthatjuk, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát. ($\rightsquigarrow$ H_1-et elvetjük)
- 5%-os elsőfajú hiba valószínűsége mellett nem tudjuk elutasítani, hogy a szilázzsal való etetés nem változtat a tehenek tejének tejzsírszázalékán. ($\rightsquigarrow$ H_0-t nem tudjuk elvetni)
- 90%-os megbízhatóság esetén azt mondhatjuk, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát.

Nem teljesen korrekt értelmezés:

- 5%-os elsőfajú hiba valószínűsége mellett azt mondhatjuk, hogy a szilázzsal való etetés nem változtat a tehenek tejének tejzsírszázalékán. ($\rightsquigarrow$ H_0-t elfogadjuk)

Feladatok

Helytelen, hibás szöveges értelmezések:

- 95%-os valószínűséggel elvethetjük, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát.
- 5%-os valószínűséggel nem tudjuk elutasítani, hogy a szilázzsal való etetés nem változtat a tehenek tejének tejzsírszázalékán.
- $8,67\%$ az esélye, hogy a szilázzsal való etetés növeli a tehenek tejének tejzsírszázalékát.

A próbafüggvény

Próbafüggvény: $\varphi\colon \mathcal{X} \to [0;1] \rightsquigarrow$ ennyi valószínűséggel vetem el a minta alapján a nullhipotézist

- $\varphi(\mathbf{x}) := I(\mathbf{x} \in \mathcal{X}_k) = \begin{cases} 1 & \text{ha } \mathbf{x} \in \mathcal{X}_k \\ 0 & \text{ha } \mathbf{x} \in \mathcal{X}_e \end{cases}$
- a próbafüggvény is egy statisztika
- a próbafüggvény egyértelműen meghatározza a próbát, ezért gyakran a próbát magával a φ függvénnyel azonosítják
- tipikusan $\varphi(\mathbf{x}) = \begin{cases} 1 & \text{ha } T(\mathbf{x}) \geq c_\alpha \\ 0 & \text{ha } T(\mathbf{x}) < c_\alpha \end{cases}$ alakú,
  ahol T egy alkalmas statisztika,
  c_α pedig a kritikus érték, amit úgy határozunk meg, hogy
  $P_{\vartheta\in\Theta_0}(\mathcal{X}_k) = E_{\vartheta\in\Theta_0}(\varphi(\mathbf{X})) = P_{\vartheta\in\Theta_0}(T(\mathbf{X}) \geq c_\alpha) = \alpha$ teljesüljön
- diszkrét eloszlású minták esetén rendszerint nem lehet úgy meghatározni c_α-t, hogy a terjedelem pontosan α legyen, ezért a próbafüggvény fogalmának általánosítására, úgynevezett **véletlenítés**re (randomizálásra) van szükség ilyen esetek miatt.

A próbafüggvény

Véletlenített próbafüggvény: $\varphi\colon \mathcal{X} \to [0;1]$

- $\varphi(\mathbf{x}) := I(\mathbf{x} \in \mathcal{X}_k) = \begin{cases} 1 & \text{ha } \mathbf{x} \in \mathcal{X}_k \\ p & \text{ha } \mathbf{x} \in \mathcal{X}_r \\ 0 & \text{ha } \mathbf{x} \in \mathcal{X}_e \end{cases}$, ahol

  $\mathcal{X}_r$ neve: véletlenítési vagy "randomizálási" tartomány; $p \in [0;1]$

- Tipikusan $\varphi(\mathbf{x}) = \begin{cases} 1 & \text{ha } T(\mathbf{x}) > c_\alpha \\ p_\alpha & \text{ha } T(\mathbf{x}) = c_\alpha \\ 0 & \text{ha } T(\mathbf{x}) < c_\alpha \end{cases}$ alakú,

  ahol T egy alkalmas statisztika,
  c_α a kritikus érték és $p_\alpha \in [0;1]$, amiket úgy határozunk meg, hogy
  $E_{\vartheta \in \Theta_0}(\varphi(\mathbf{X})) = P_{\vartheta \in \Theta_0}(T(\mathbf{X}) > c_\alpha) + p_\alpha \cdot P_{\vartheta \in \Theta_0}(T(\mathbf{X}) = c_\alpha) = \alpha$
  teljesüljön
  Az ilyen próbafüggvénnyel végrehajtott próbát **véletlenített próbá**nak hívjuk.

Próbák tulajdonságai

- **Torzítatlan próba (legfeljebb α terjedelemmel)**:
  $P_\vartheta(\mathbf{X} \in \mathcal{X}_k) \leq \alpha \quad \forall \vartheta \in \Theta_0$-ra és
  $P_\vartheta(\mathbf{X} \in \mathcal{X}_k) \geq \alpha \quad \forall \vartheta \in \Theta_1$-re
  Megfontolás a definíció mögött: ha nem teljesül H_0, akkor a minta alapján az elvetés valószínűsége legalább annyi legyen, mintha igaz lenne H_0.
- **Konzisztens próba (α terjedelemmel)**: olyan próba, aminek α a terjedelme és a mintaméret növelésével az erőfüggvény 1-hez konvergál. Formálisan felírva:
  $P_{\vartheta \in \Theta_0}(\mathbf{X} \in \mathcal{X}_k) = \alpha$ és
  $\psi_n(\vartheta) \overset{n\to\infty}{\longrightarrow} 1 \quad \forall \vartheta \in \Theta_1$-re, ahol
  ψ_n az n elemű mintához tartozó erőfüggvény
  Megj.: ha az erőfüggvény 1-hez konvergál, akkor ebből következik, hogy a másodfajú hiba valószínűsége 0-hoz tart.

Próbák tulajdonságai

- Legyenek φ_1 és φ_2 torzítatlan próbák. A φ_1 **próba erősebb** φ_2 **próbánál**, ha φ_1 próba erőfüggvénye $\forall \vartheta \in \Theta_1$ esetén nagyobb vagy egyenlő, mint φ_2 próba erőfüggvénye
  Megj.: nem biztos, hogy két próba közül az egyik erősebb a másiknál
- **Egyenletesen legerősebb próba**: az adott hipotézisvizsgálati feladat esetén minden más torzítatlan próbánál erősebb
  Megj.: nem biztos, hogy létezik egyenletesen legerősebb próba az adott feladatra
- Mikor létezik egyenletesen legerősebb próba? Ha létezik, akkor hogyan találjuk meg?

Legerősebb próba keresése

Ha mind H_0, mind H_1 egyszerű, akkor adott α terjedelemhez lehet legerősebb próbát találni, ezt pedig úgy hívják, hogy **valószínűség-hányados próba**.

A hipotézisek folytonos esetre (diszkrétre a sűrűségfüggvény helyett a konkrét eloszlást kell írni):

$H_0 : \ f = f_0$

$H_1 : \ f = f_1$

A próba kritikus tartománya: $\mathcal{X}_k = \left\{ \mathbf{x} : \overbrace{\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}}^{T(\mathbf{x})} > c_\alpha \right\}$

Tehát azokat az $\mathbf{x}$-eket, amikre a $T(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ statisztika nagy, bepakoljuk a kritikus tartományba egészen addig, míg az adott α terjedelmet el nem érjük. Diszkrét esetben ehhez általában véletlenítésre van szükség, azaz bizonyos $\mathbf{x}$-ek esetén nem 1 vagy 0, hanem egy, e két szám közé eső (jelöljük p_α-val) valószínűséggel vetjük el a nullhipotézist.

A valószínűség-hányados próba elméleti hátterét a *Neyman-Pearson* (alap)*lemma* biztosítja.

Feladatok

E29.) Legyen $X_1, \ldots, X_n$ minta az $f(x) = \begin{cases} \frac{2x}{a^2} & \text{ha } 0 \leq x \leq a \\ 0 & \text{különben} \end{cases}$ sűrűségfüggvényű eloszlásból, ahol a ismeretlen paraméter. Adjunk meg α terjedelemhez egyenletesen legerősebb próbát az alábbi hipotézisek esetén:

$H_0 : a = 1$

$H_1 : a = \frac{1}{2}$

E30.) Legyen két megfigyelésünk a $(3; p)$ paraméterű binomiális eloszlásból. Adjuk meg a legjobb (egyenletesen legerősebb) $0,04$ terjedelmű próbát az alábbi hipotézisekre:

$H_0 : \ p \geq \frac{1}{2}$

$H_1 : \ p < \frac{1}{2}$

E31.) Mutassuk meg, hogy a kétoldali, egymintás u-próba torzítatlan és konzisztens!

Feladatok

E32.) Keressünk n elemű $N(m, \sigma^2)$ független minta esetén egyenletesen legerősebb α terjedelmű próbát a
$H_0 : m = m_0$
$H_1 : m > m_0$
hipotézisvizsgálati feladatra, ha σ ismert!
Határozzuk meg a kritikus értéket, ha a próbastatisztika
$T(\mathbf{X}) = \sqrt{n}\frac{\overline{X}-m_0}{\sigma}$! Ez alapján kimondható a következő

Tétel: az egyoldali u-próba egyenletesen legerősebb a $H_0 : m = m_0$, $H_1 : m > m_0$ hipotézisvizsgálati feladatra. ($H_1 : m < m_0$-re is)

Mi történik, ha az u-próbák, illetve t-próbák feltételei közül nem teljesül az, hogy a minta normális eloszlású, mi mégis hagyományos módon végrehajtjuk a próbát?

E33.) Vizsgáljuk meg szimulációval a kétoldali, egymintás t-próba terjedelmét, amennyiben a minta $n = 10, 20, 50, 100, 200$ elemű, független normális, illetve exponenciális eloszlású!

Motiváció – χ^2-próbával vizsgálható problémák

- Szabályos-e egy érme/kocka?
- Normális eloszlást követ-e a magyar nők testmagassága?
- Lehet-e exponenciális eloszlású az az idő, amit a Blaha Lujza téri megállóban a következő villamosra várakozással töltünk?
- Állíthatjuk-e, hogy a nők és a férfiak vérnyomása ugyanolyan eloszlást követ?
- Független-e a diákok matematika és irodalom érdemjegye? Aki jó matekból, általában jó magyarból is?
- Független-e egymástól az emberek szemszíne és hajszíne? Igaz-e, hogy a szőke hajúak főleg kék szeműek?

A χ^2-próba

Legyen $A_1, \ldots, A_r$ teljes eseményrendszer.
Végezzünk n darab független megfigyelést, jelölje az i-edik esemény bekövetkezési gyakoriságát N_i $(i = 1, \ldots, r)$. A megfigyelések egyes eredményei segítségével definiálható az X_i valószínűségi változó, ami vegyen fel olyan értéket, amelyik számú esemény a teljes eseményrendszerből bekövetkezett. Ezáltal formálisan

$$N_i = \sum_{j=1}^{n} I(X_j = i) \text{ és } \sum_{i=1}^{r} N_i = n$$

H_0: $P(A_i) = p_i$, $i = 1, \ldots, r$ $\rightsquigarrow$ tfh. $p_i > 0 \ \forall i, \ p_1 + \ldots + p_r = 1$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_n(\mathbf{X}) := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \left\{\mathbf{x} : T_n(\mathbf{X}) > \chi^2_{r-1;1-\alpha}\right\}$

A χ^2-próba

Alkalmazásai:

- illeszkedésvizsgálat: egy minta adott eloszlást követ-e
- homogenitásvizsgálat: két minta eloszlása megegyezik-e
- függetlenségvizsgálat: két szempont, ismérv, tulajdonság független-e egymástól

Megjegyzések:

- a χ^2-próba **aszimptotikus** próba, ami azt jelenti, hogy "nagy" mintaelemszámra lehet végrehajtani. "Kicsi" minták esetén a kritikus érték nem használható, azt szimulálni kell a konkrét minta alapján.
- Mikor elég "nagy" már egy minta – hüvelykujjszabály: ha legalább 100 elemű. Egyébként eloszlásfüggő, legalább mekkora n-re van szükség, hogy kritikus értéknek a χ^2-eloszlás kvantiliseit lehessen használni.
- Végrehajtásának további feltétele, hogy minden osztályban "elegendő" mennyiségű gyakoriság legyen.
- A próbastatisztikában lévő összeg tagjai $\frac{(O-E)^2}{E}$ alakúak, ahol E: elméleti gyakoriságok, O: tapasztalati gyakoriságok

Illeszkedésvizsgálat

H_0: a minta egy adott eloszlásból származik
H_1: a minta nem ilyen eloszlású

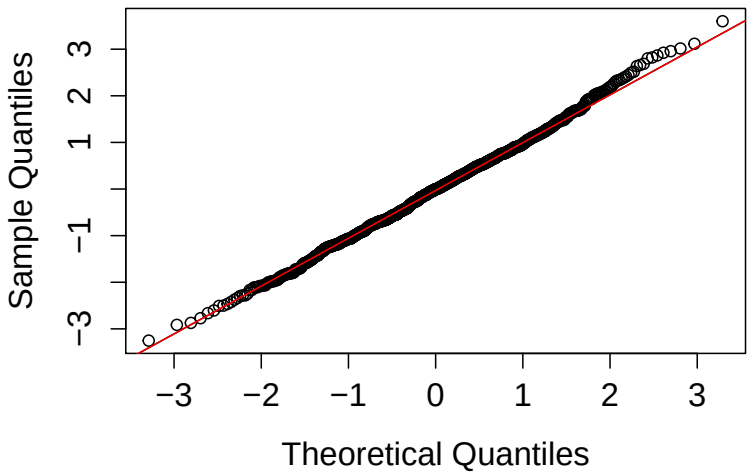
Végrehajtása:

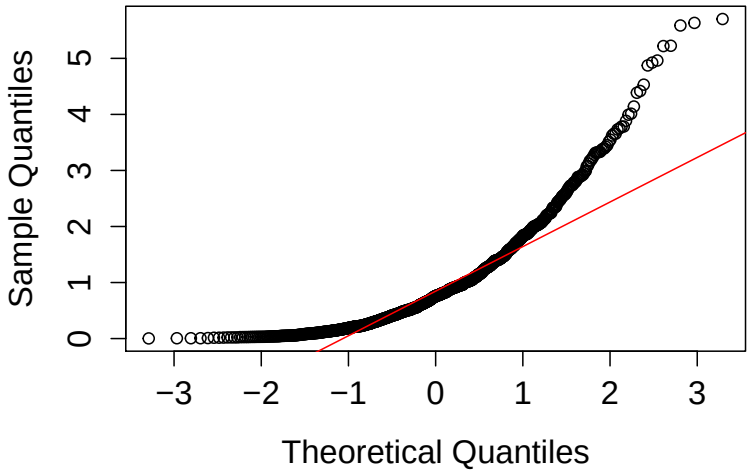
- grafikusan módszerek ("szemmel" jónak tűnik-e az illeszkedés):
  - Q-Q plot
  - P-P plot
  - hisztogram/magfüggvényes sűrűségfüggvény-becslés, valamint az illesztett sűrűségfüggvény egy ábrában
- statisztikai próbák:
  - diszkrét eloszlás esetén χ^2-próba
  - folytonos eloszlás esetén több statisztikai próba közül lehet választani
    - diszkretizálás (mesterséges osztályok létrehozása) révén χ^2-próba
    - Kolmogorov-Szmirnov próba
    - Cramér-von Mises próba
    - Anderson-Darling próba
    - Shapiro-Wilk próba: kizárólag normalitásvizsgálatra, amire ez a legjobb

Illeszkedésvizsgálat grafikusan

Q-Q plot (kvantilis-kvantilis ábra)

- Az illesztett eloszlás kvantiliseit vetjük össze a tapasztalati kvantilisekkel, azaz a következő pontokat ábrázoljuk:
  $\left(F^{-1}\left(\frac{k}{n+1}\right), x_k^*\right) \quad k = 1, \ldots, n$
  ahol
  - F: az illesztett eloszlás eloszlásfüggvénye
  - x_k^* a k. rendezett mintaelem
- Be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.
- Felnagyítja az eloszlás szélein az eltéréseket, ezért szinte mindig előnyben részesítik a P-P plot-tal szemben.





Illeszkedésvizsgálat grafikusan

P-P plot (percentilis-percentilis ábra)

- Az illesztett eloszlás egyes valószínűségeit vetjük össze a tapasztalati valószínűségekkel, azaz a következő pontokat ábrázoljuk:
  $\left(\frac{k}{n+1}, F\left(x_k^*\right)\right) \quad k = 1, \ldots, n \qquad$ ahol
  - F: az illesztett eloszlás eloszlásfüggvénye
  - x_k^* a k. rendezett mintaelem
- Be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.
- Felnagyítja az eloszlás közepén az eltéréseket

A Q-Q plot és P-P plot nem helyettesíti a formális tesztelést, inkább kiegészíti azt!

Illeszkedésvizsgálat χ^2-próbával

Osztályok	1	2	...	r	Összesen
Valószínűségek	p_1	p_2	...	p_r	1
Gyakoriságok	N_1	N_2	...	N_r	n

H_0 : a valószínűségek: $\mathbf{p}=(p_1,\ldots,p_r)$

H_1: nem ezek a valószínűségek

Próbastatisztika: $T_n(\mathbf{X}) = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ elo.-ban, ha $n \to \infty$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n(\mathbf{X}) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

A χ^2-próba végrehajtásának feltételei (hüvelykujjszabály): $N_i \geq 4$ és $np_i \geq 4$ minden i-re. Ha ezek nem teljesülnek, akkor osztályokat kell összevonni.

Illeszkedésvizsgálat Kolmogorov-Szmirnov próbával

$H_0 : F_{X_1}(x) = F(x) \ \ \forall x \in \mathbb{R}$ ahol F egy adott eloszlás elofv.-e

H_1: a nullhipotézis tagadása

Próbastatisztika: $D_n(\mathbf{X}) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$

A próbastatisztika $\sqrt{n}$-szeresének eloszlása H_0 esetén az ún. Kolmogorov-eloszláshoz tart ($n \to \infty$). Jelöljük K_α-val a Kolmogorov-eloszlás α-kvantilisét.

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : \sqrt{n} D_n(\mathbf{x}) > K_{1-\alpha}\}$

Megjegyzések:

- D_n kiszámításához elég csak a mintapontokban tekinteni az eltérést.
- Nem lehet használni a határeloszlást, ha paramétereket kell becsülnünk! Ilyen esetben a kritikus értéket szimulációval kaphatjuk meg.
- A Kolmogorov-eloszlás eloszlásfüggvénye: $1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$

E34.) Egy gyártó megfigyelte, hogy 100, általa előállított SSD merevlemezen 5 év használat után hány hibás szektort talál az ezek felkutatására készített szoftver:

Hibás szektorok száma	0	1	2	3	4	5	7	Összesen
Gyakoriságok	45	35	12	5	1	1	1	100

Vizsgáljuk meg, hogy a szektorhibák száma Poisson-eloszlást követ-e!

E35.) Nézzük meg P-P plot-tal és Q-Q plot-tal, majd diszkretizálás után χ^2-próbával, valamint Kolmogorov-Szmirnov próbával, hogy a következő minta:

4,3 2.0 5,6 8,1 3,2 0,6 5,4 8,9 7,5 9,3
9,6 6,7 4,4 2,9 1,0 6,5 4,0 6,6 4,2 1,9

származhat-e az alábbi eloszlásokból:

a.) $E(0; 10)$;

b.) $N\left(5; \left(\frac{5}{\sqrt{3}}\right)^2\right)$.

Homogenitásvizsgálat

Adott két független minta, mindkettő egy közös szempont szerint r osztály egyikébe sorolva.

	Osztályok	1	2	...	r	Összesen
1. minta	Valószínűségek	p_1	p_2	...	p_r	1
	Gyakoriságok	N_1	N_2	...	N_r	n
2. minta	Valószínűségek	q_1	q_2	...	q_r	1
	Gyakoriságok	M_1	M_2	...	M_r	m

H_0: a két minta azonos eloszlású, azaz $(p_1, \ldots, p_r) = (q_1, \ldots, q_r)$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_{n,m}(\mathbf{X}, \mathbf{Y}) = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : T_{n,m}(\mathbf{X}, \mathbf{Y}) > \chi^2_{r-1,1-\alpha}\}$

Függetlenségvizsgálat

Feladat: van egy minta, két ismérv szerint csoportosítva. Azt kell eldönteni, hogy a két szempont független-e egymástól.

$p_{i,j} = P$(egy megfigyelés az (i,j) osztályba kerül)
$N_{i,j} =$ennyi megfigyelés kerül az (i,j) osztályba

		2. szempont					
		1	...	j	...	s	Összesen
1. szempont	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
	i	N_{i1}	...	N_{ij}	...	N_{is}	$N_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	N_{r1}	...	N_{rj}	...	N_{rs}	$N_{r\bullet}$
Összesen		$N_{\bullet 1}$	...	$N_{\bullet j}$	...	$N_{\bullet s}$	n

ahol $N_{i\bullet} = \sum_{j=1}^{s} N_{ij}$ és $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$

Függetlenségvizsgálat

Itt formálisan a mintánk két dimenziós: a megfigyelések az $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ párok, ahol az X-ek r, az Y-ok pedig s különböző értéket vehetnek fel nemnulla valószínűséggel:
$p_{i,j} = P(X_1 = x_i, Y_1 = y_j)$, ahol $i = 1, \ldots, r$ és $j = 1, \ldots, s$.
Továbbá $N_{i,j} = \sum_{k=1}^{r} \sum_{l=1}^{s} I(X_k = x_i, Y_l = y_j)$.

H_0 : az ismérvek függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j} \quad \forall i, j$-re
H_1: az ismérvek nem függetlenek

Próbast.: $T_n(\mathbf{X}, \mathbf{Y}) = n \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{N_{i,j}^2}{N_{i\bullet} N_{\bullet j}} - 1 \right) \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{(r-1)(s-1)}$ elo.-ban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : T_n(\mathbf{X}, \mathbf{Y}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

Ha $r = s = 2$, akkor a próbastatisztika $T_n = n \cdot \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}$-re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú χ^2.

E**36.)** Egy webtervező azt gyanítja, hogy az általa létrehozott internetes vásárlás honlapján a vásárlások mértéke összefügg azzal, hogy milyen nap van a héten. Ennek a sejtésnek az ellenőrzésére egy héten keresztül adatokat gyűjt – összesen 3758 látogatót számlált meg:

Vásárlás	H	K	Sz	Cs	P	Sz	V	Össz.
Nem vásárolt	399	261	284	263	393	531	502	2633
1 vásárlás	119	72	97	51	143	145	150	777
Több vásárlás	39	50	20	15	41	97	86	348
Összesen	557	383	401	329	577	773	738	3758

Alkalmas statisztika próbával döntsünk arról, hogy helyes-e a webtervező sejtése!

Regresszióelemzés, lineáris modell – motiváció

- Gyakorlati szempontból a félév egyik legfontosabb témája!
- Az 1. órai kérdőíves felmérés alapján mennyire magyarázható jól
  - a hallgatók testmagassága a súlyuk segítségével?
  - a hallgatók testmagassága a súlyuk és a cipőméretük segítségével?
  - a hallgatók statisztika érdemjegye a testmagasságuk segítségével?
- Egy részvény holnapi árfolyamát hogyan jelezzük előre a tegnapi, tegnapelőtti, stb. árfolyamadatok segítségével?
- Egy gazda földvásárlási dilemmája – egy bizonyos földterületen a várható termésátlag mennyire jelezhető előre a földterület fontosabb jellemzői alapján (a talaj kémhatása, a $CaCO_3$ megjelenési mélysége, a humusztartalom, topográfiai helyzet)?
- Meg lehet-e becsülni annak az esélyét, hogy valaki élete során megbetegszik tüdőrákban? Hogyan modellezzük ezt? Például: megbetegedés esélye $\longleftarrow$ dohányzik-e, hány éven át dohányzott élete során, van-e tüdőrákos a közeli rokonságban, van-e egyéb tüdőbetegsége, poros/füstös helyen dolgozik-e

Regresszióelemzés – bevezetés

Legyenek $Y, X_1, \ldots, X_p$ véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői.

A regresszióelemzés célja: a bennünket különösen érdeklő Y valószínűségi változó "minél jobb" közelítése az $X_1, \ldots, X_p$ valószínűségi változók segítségével.

Y elnevezései: eredményváltozó, függő változó, endogén változó
X_i-k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az $(Y, X_1, \ldots, X_p)^T$ valószínűségi vektorváltozó realizációinak tekinthetők:
$(y_i, x_{i,1}, \ldots, x_{i,p})^T \quad i = 1, 2, \ldots, n \qquad$ általában $n >> p$
Feltehetjük, hogy az y_i megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát ε_i-vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és egy véges σ szórású valószínűségi változó.

Regresszióelemzés – bevezetés

Legyenek $Y, X_1, \ldots, X_p$ véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői.

A regresszióelemzés célja: a bennünket különösen érdeklő Y valószínűségi változó "minél jobb" közelítése az $X_1, \ldots, X_p$ valószínűségi változók segítségével.

Y elnevezései: eredményváltozó, függő változó, endogén változó
X_i-k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az $(Y, X_1, \ldots, X_p)^T$ valószínűségi vektorváltozó realizációinak tekinthetők:
$(y_i, x_{i,1}, \ldots, x_{i,p})^T \quad i = 1, 2, \ldots, n \qquad$ általában $n >> p$
Feltehetjük, hogy az y_i megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát ε_i-vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és egy véges σ szórású valószínűségi változó.

Regresszióelemzés

Legyenek Y, X, $X_1, \ldots, X_p$ véges szórású valószínűségi változók, $c, a, b_1, \ldots, b_p$ valós számok.
Jelölje $\mathbf{X} = (X_1, \ldots, X_p)^T$, $\mathbf{b} = (b_1, \ldots, b_p)^T$ vektorokat.

	Feladat	Megoldás
a.)	$\min\limits_{c} E(Y-c)^2$	$\hat{c} = EY$ [Steiner-tétel]
b.)	$\min\limits_{f:\mathbb{R}\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X))^2$	$\widehat{f}(X) = E(Y\|X)$
c.)	$\min\limits_{a,b} E(Y-(a+bX))^2$	$\hat{b} = \frac{cov(X,Y)}{D^2X}$, $\hat{a} = EY - \hat{b}EX$
d.)	$\min\limits_{f:\mathbb{R}^p\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X_1,\ldots,X_p))^2$	$\widehat{f}(X_1,\ldots,X_p) = E(Y\|X_1,\ldots,X_p)$
e.)	$\min\limits_{a,b_1,\ldots,b_p} E\left(Y-(a+\sum\limits_{i=1}^{p} b_iX_i)\right)^2$ [Többváltozós lineáris regresszió]	$\hat{\mathbf{b}} = (cov(\mathbf{X},\mathbf{X}))^{-1} cov(\mathbf{X}, Y)$ $\hat{a} = EY - \sum\limits_{i=1}^{p} \hat{b}_i EX_i$

$E(Y|X)$: feltételes várható érték – jövő félévben

Lineáris modell

- A modell: $\mathbf{y} = X\mathbf{b} + \varepsilon$
- $F := \mathrm{Im}X \quad \rightsquigarrow X$ képtere
- $r :=$rang(X), általánosan $r \leq p$, teljes rangú esetben $r = p$
- Paraméterbecslés: $\widehat{\mathbf{b}} = (X^T X)^{-1} X^T \mathbf{y}$
- Projekció az F altérre: $P_F = X(X^T X)^{-1} X^T$
- Becsült értékek: $\widehat{\mathbf{y}} := X\widehat{\mathbf{b}}$
- Reziduálisok: $\widehat{\varepsilon} = \mathbf{y} - \widehat{\mathbf{y}}$
- Reziduális négyzetösszeg: $\mathrm{RNÖ} := \|\widehat{\varepsilon}\|^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$
- Teljes négyzetösszeg: $\mathrm{NÖ} = \sum_{i=1}^{n} (y_i - \overline{y})^2$
- Determinációs együttható: $R^2 = 1 - \frac{\mathrm{RNÖ}}{\mathrm{NÖ}} = \frac{\mathrm{NÖ} - \mathrm{RNÖ}}{\mathrm{NÖ}} \quad \rightsquigarrow$ az eredményváltozó változékonyságának hány %-át magyarázza a regressziós modell
  Értéke 0 és 1 között lehet. Minél nagyobb, annál jobb.

- Korrigált determinációs együttható: $R^2_{\text{adj}} = 1 - \frac{n-1}{n-r-1}\frac{SSR}{SS}$ $\rightsquigarrow$ egy lehetséges modellválasztási kritérium, minél nagyobb, annál jobb
- Akaike-féle információs kritérium: $AIC = 2(p+1) - 2\log\widehat{L}$, ahol $\widehat{L}$ a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel)
  Ez is egy lehetséges modellválasztási kritérium, minél kisebb, annál jobb.

Regresszióelemzés

A regresszióelemzés lépései

- az eredményváltozó(k) és a lehetséges magyarázóváltozók kiválasztása
- adatgyűjtés
- adattisztítás, adathibák korrekciója
- pontdiagrammal a potenciális modellek kiválasztása (lineáris, négyzetes, logisztikus stb.)
- paraméterbecslés
- modelldiagnosztika – az együtthatók szignifikanciája, a modell együttes jósága
- legjobb modell kiválasztása, "modellépítés" – több módszer/mutató közül választhatunk: korrigált R^2, cross-validation, AIC/BIC információs kritériumok stb.
- előrejelzés

A regressziós modell "felépítése"

Ha p magyarázóváltozónk van, akkor 2^p modell közül kell a legjobbat kiválasztani. Nagy számú magyarázóváltozó esetén ez nem lehetséges, több módszer közül lehet választani:

- Nagyról kicsire (hátulról előre): először az összes magyarázóváltozót bevesszük, majd egyenként a legkevésbé szignifikánsat kivesszük egészen addig, míg mindegyik szignifikáns lesz
- Kicsiről nagyra (elölről hátulra): egyesével azt vesszük hozzá, amelyikkel a legjobban illeszkedő modellt kapjuk a következő lépésben.
  Vége: ha bármelyik, még a modellen kívüli magyarázóváltozót bevéve, már nem javul a modell illeszkedése.

Feladatok

E37.) Tekintsünk az alábbi regressziós modellekre lineáris modellként, és becsüljük meg a paramétereket! Jelölések: (y_i, x_i) a megfigyelések, ε_i a mérési hiba $(i = 1, \ldots, n)$, a becsülendő paraméterek pedig a, b, c.

a.) $y_i = a + bx_i + \varepsilon_i \quad \rightsquigarrow$ (egyszerű) kétváltozós regresszió
b.) $y_i = a + bx_i + cx_i^2 + \varepsilon_i \quad \rightsquigarrow$ négyzetes regresszió
c.) $y_i = a + b \sin x_i + c \cos x_i + \varepsilon_i \quad \rightsquigarrow$ harmonikus regresszió

Határozzuk meg a becsült paramétereket **R** segítségével és ábrázoljuk a megfigyeléseket az illesztett görbével együtt, ha a megfigyelések a következők:

y_i	-0,82	1,72	2,72	1,14	0,96	0,93	-0,08	0,29	3,38	3,36
x_i	3,92	2,63	1,68	2,57	2,61	2,78	3,81	2,89	0,28	0,94

Értékeljük az egyes modelleket önmagukban, és egymáshoz képest is!

Feladatok

E38.) Az év eleji kérdőíves felmérés alapján célunk a hallgatók testmagasságának minél jobb magyarázása.

a.) Tekintsük az alábbi modelleket:

I.) Testmagasság = Testsúly + Hiba , ami a Testmagasság $= a_0 + a_1 \cdot$ Testsúly + Hiba kiírás rövidített változata (a többinél hasonlóan)

II.) Testsúly = Testmagasság + Hiba

III.) Testmagasság = Testsúly + Lábméret + Hiba

IV.) Testmagasság = Nem + Hiba

b.) Vizsgáljuk meg a korrelációs mátrixot! Keressük meg a legjobban illeszkedő modellt!

c.) Adjunk előrejelzést a legjobbnak tűnő modell(ek) alapján egy olyan fiú hallgató testmagasságára, aki 70 kg-os, 45-ös a cipőmérete, 5-öse volt valszámból, 25 percet utazik az egyetemre és heti 12 órát tanul!

Ismérvek közti kapcsolat vizsgálata (leíró statisztika)

Jelölje a D ismérv lehetséges (elméleti) értékeit X, az E ismérv lehetséges értékeit pedig Y.

Az ismérvek közti kapcsolat fajtái az összefüggés iránya szerint:

- Ok-okozati összefüggés – az egyik ismérv magyarázó változónak, a másik függő változónak tekinthető (pl. tanulási idő $\rightarrow$ ZH-n szerzett pontszám) $\rightsquigarrow$ jóval gyakoribb
- Kölcsönös kapcsolat – mindegyik ismérv hatással van a másikra (pl. ár és kereslet)

A kapcsolat a kapcsolat erőssége (E-t magyarázzuk D-vel) szerint:

- függetlenség: attól, hogy ismerjük D értékeit, semmit se tudunk mondani E értékeiről
- függvényszerű kapcsolat: ha ismerjük D értékeit, abból egyértelműen meg lehet mondani E értékeit; lehet lineáris, négyzetes, exponenciális stb.
- sztochasztikus kapcsolat: a fenti kettő közötti, D értékeiből tudunk következtetni arra, hogy E milyen értéket vesz fel, de ez a következtetés bizonytalansággal jár

Ismérvek közti kapcsolat vizsgálata

A kapcsolat fajtája az ismérvek típusa (E-t magyarázzuk D-vel) szerint:

- asszociáció: E és D egyaránt nominális/sorrendi skálán mért ismérvek ($\approx$ Y és X diszkrét v.v.-k)
- vegyes kapcsolat: E intervallum vagy arányskálán mért ismérv ($\approx$ Y abszolút folytonos valószínűségi változó), D nominális/sorrendi skálán mért ismérv ($\approx$ X diszkrét v.v.)
- korreláció: E és D egyaránt intervallum vagy arányskálán mért ismérvek (Y és X is abszolút folytonos v.v.)

Az alkalmazandó megközelítések/módszerek összefoglalása:

		Az eredményváltozó	
		diszkrét	abszolút folytonos
A magyarázó-változó	diszkrét	asszociáció χ^2-próba	vegyes kapcsolat t-próba, ANOVA
	absz. folyt.	osztályozási eljárások, diszkriminancia analízis	korreláció regresszió

Asszociáció

		E ismérv					
		1	...	j	...	s	Összesen
D ismérv	1	f_{11}	...	f_{1j}	...	f_{1s}	$f_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
	i	f_{i1}	...	f_{ij}	...	f_{is}	$f_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	f_{r1}	...	f_{rj}	...	f_{rs}	$f_{r\bullet}$
Összesen		$f_{\bullet 1}$	...	$f_{\bullet j}$	...	$f_{\bullet s}$	n

ahol $f_{i,j}$: gyakoriságok száma az (i,j) osztályban

$$f_{i\bullet} = \sum_{j=1}^{s} f_{ij} \quad \text{és } f_{\bullet j} = \sum_{i=1}^{r} f_{ij}$$

- függetlenség esetén $f_{ij} = \frac{f_{i\bullet} \cdot f_{\bullet j}}{n}$ minden i-re és j-re
- függvényszerű kapcsolat esetén minden sorban egyetlen gyakoriság nem nulla, a többi nulla

Az asszociáció mérőszámai

Cramér-féle asszociációs együttható:

$$C = \sqrt{\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij}-f_{ij}^*)^2}{f_{ij}^*}}{n\cdot(\min(r,s)-1)}}, \quad \text{ahol } f_{ij}^* = \frac{f_{i\bullet}\cdot f_{\bullet j}}{n}$$

Yule-féle asszociációs együttható: $r = s = 2$ esetén számolható

$$Y = \frac{f_{11}\cdot f_{22} - f_{12}\cdot f_{21}}{f_{11}\cdot f_{22} + f_{12}\cdot f_{21}}$$

Tulajdonságaik:

- $C = 0$ vagy $Y = 0$ esetén a két ismérv független egymástól
- $C = 1$ vagy $Y = \pm 1$ esetén a két ismérv között függvényszerű kapcsolat van
- $0 < C < 1$ vagy $0 < |Y| < 1$ esetén a két ismérv között sztochasztikus kapcsolat van
- erős a kapcsolat, ha C vagy $|Y|$ 1-hez van közel és gyenge a kapcsolat, ha 0-hoz

A vegyes kapcsolat elemzése – szóráselemzés

- Elnevezései: szóráselemzés = variancia-analízis = ANOVA (analysis of variance)
- A lineáris modell egyik legfontosabb alkalmazása
- Gyakorlatban az egyik leggyakoribb probléma

- Motivációs példák:
  - Hatással van-e egy vállalatnál a (bruttó) fizetésekre az, hogy valaki nő-e, avagy férfi?
  - Egymástól eltérő összetételű trágyák esetén megnézték a termésátlagot egy nagyobb földterület különböző részein. Vajon hatással van-e a trágya fajtája a termésátlagra?
  - Hatással van-e a valszám gyakorlati összpontszámra, hogy a hallgatónak ki a gyakorlatvezetője?

A vegyes kapcsolat mérőszáma

Szórásnégyzet-hányados

$$H^2 = 1 - \frac{SSB}{SST} = \frac{SSK}{SST}$$

Megjegyzés: ez nem más, mint a regressziónál az R^2

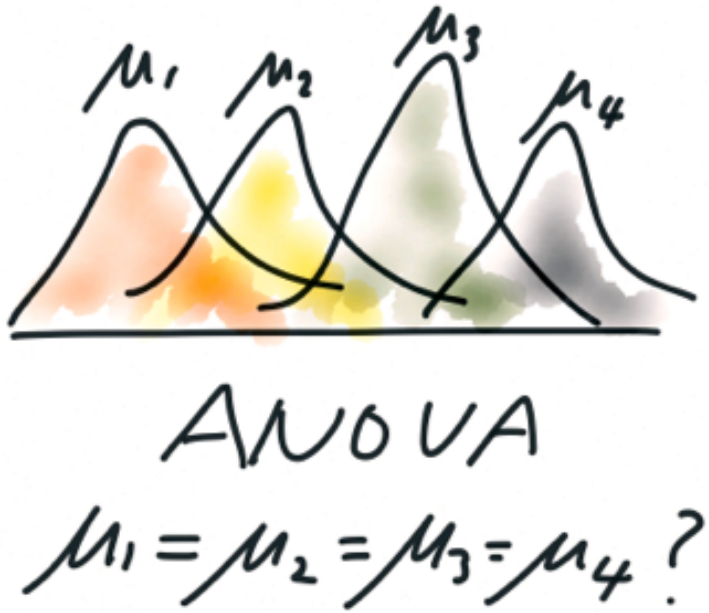
Tulajdonságai:

- $H^2 = 0$ esetén a két ismérv között nincs kapcsolat, DE (!!) ekkor nem feltétlen függetlenek egymástól (analógia: korrelálatlanságból nem következik a függetlenség)
- $H^2 = 1$ esetén a két ismérv között függvényszerű kapcsolat van
- $0 < H^2 < 1$ esetén a két ismérv között sztochasztikus kapcsolat van
- erős a kapcsolat, ha H^2 közel van 1-hez és gyenge a kapcsolat, ha 0-hoz

Szóráselemzés (ANOVA)

A szóráselemzési feladat fő kérdése: hatással van-e az eredményváltozó értékére, hogy a faktor melyik szintjén vagyunk?

$H_0 : b_1 = b_2 = \ldots = b_p$
$H_1 :$ nem igaz H_0



ANOVA táblázat:

Szóródás forrása	Szabadság-fok	Négyzet-összegek	Tapasztalati szórásnégyzetek	
Külső	$p-1$	SSK	$MSK = \frac{SSK}{p-1}$	$F = \frac{\frac{SSK}{p-1}}{\frac{SSB}{n-p}}$
Belső	$n-p$	SSB	$MSB = \frac{SSB}{n-p}$	
Teljes	$n-1$	SST	–	

Szóráselemzés (ANOVA)

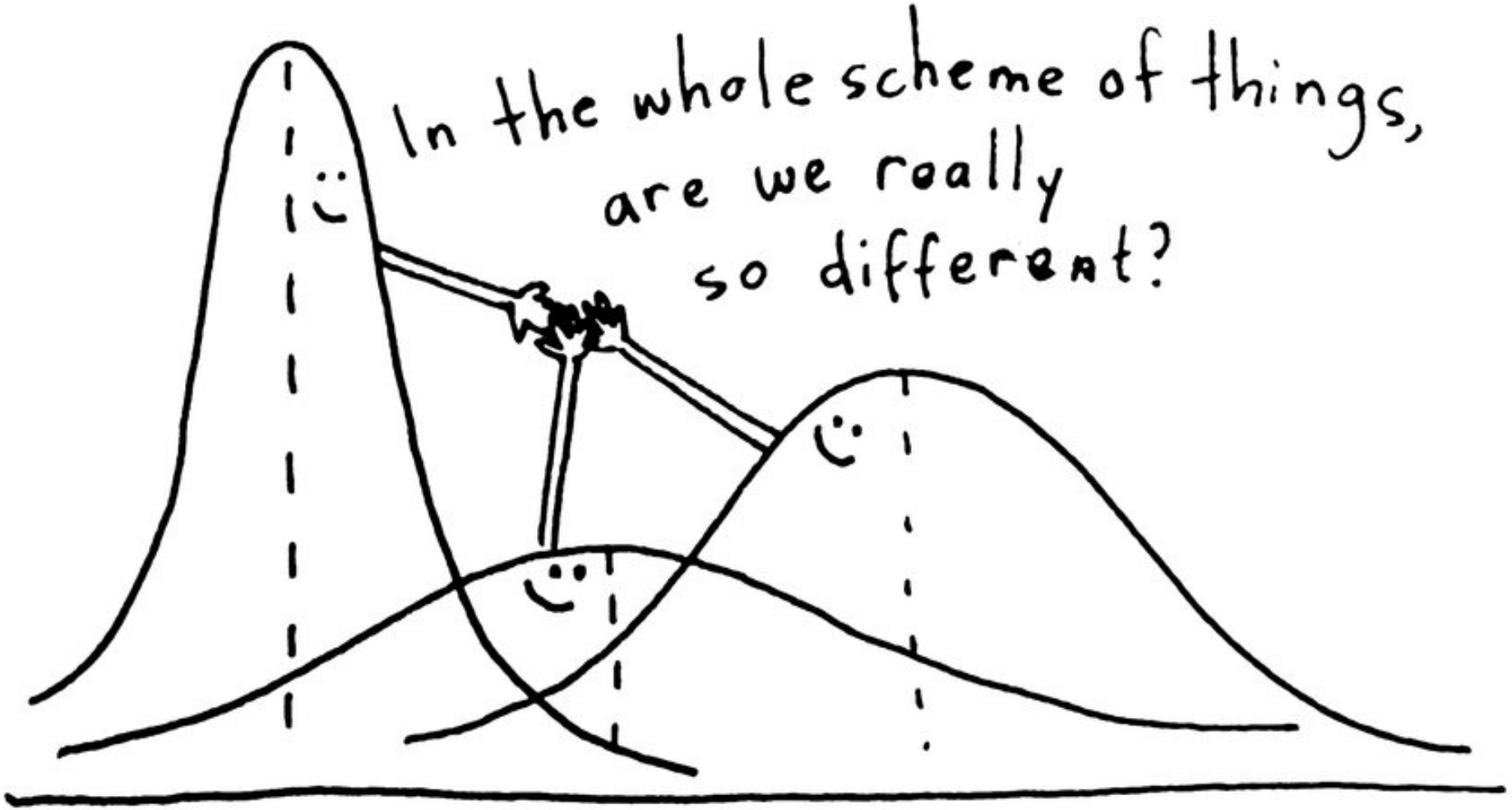
Állítás: $\frac{\overline{y}_{i\bullet}-b_i}{\sqrt{MSB}}\sqrt{n_i} \sim t_{n-p}$ és $\frac{\overline{y}_{i\bullet}-\overline{y}_{j\bullet}-(b_i-b_j)}{\sqrt{MSB}}\sqrt{\frac{n_i n_j}{n_i+n_j}} \sim t_{n-p}$

Ezek alapján konfidenciaintervallumokat lehet készíteni b_i együtthatókra és a $b_i - b_j$ különbségekre.

Konfidenciaintervallumok:

- b_i-re: $\overline{y}_{i\bullet} \pm t_{n-p;\alpha/2}\sqrt{\frac{MSB}{n_i}}$
- $b_i - b_j$-re: $\overline{y}_{i\bullet} - \overline{y}_{j\bullet} \pm t_{n-p;\alpha/2}\sqrt{MSB}\sqrt{\frac{n_i+n_j}{n_i n_j}}$

A szóráselemzés egy lehetséges kiterjesztése – a szórás az egyes csoportokban eltérhet



Feladatok

E39.) A következő táblázatok a 2016/2017-es őszi félév Valószínűségszámítás nevű tárgy hallgatóinak megoszlását mutatják aszerint, hogy a hallgató milyen szakos és a vizsgán hányast szerzett (csak azok szerepelnek, akik legalább 1-szer próbálkoztak).

a.)

Szakirány	Karó	2-es	3-as	4-es	5-ös	Összesen
Elemző	16	13	10	0	2	41
Infó A	4	3	7	1	7	22
Összesen	20	16	17	1	9	63

b.)

Szakirány	Karó	Legalább 2-es	Összesen
Elemző	16	25	41
Infó A	4	18	22
Összesen	20	43	63

Vizsgáljuk meg alkalmas mutatószámmal, a megszerzett érdemjegyre hatással volt-e az, hogy a hallgató milyen szakra jár!

Feladatok

E**40.)** A következő táblázat az előző félévben tanár szakos BSc-s hallgatóknak tartott 4 bevezető valszám gyakorlat év végi, 100-ra skálázott végső pontszámait tartalmazza:

Gyakvezér	Pontszámok											
Cs. V.	98	87	102	92	52	46	95	60	81	55	60	94
	81	58	80	93	70	66	49	94	50	88	74	
W. G.	77	46	54	57	50	45	39	63	26	107	75	
	66	52	109	91	35	65						
B. Á.	86	94	54	61	42	59	88	81	81	80	102	72
	88	96	58	90	110	58	80	90	84	80	94	
V. L.	66	60	72	49	52	54	80	56	36	91	68	
	60	51	40	38	54	62						

a.) Vizsgáljuk meg, az év végi pontszám függ-e attól, hogy a hallgató melyik csoportba jár! Hány %-ban magyarázza a pontszámok változékonyságát az, hogy a hallgatók melyik csoportba járnak?

b.) Adjunk intervallumbecslést az egyes csoportok várható pontszámára!

c.) Állíthatjuk-e, hogy Cs. V. és B. Á. csoportjának átlagpontszámai (statisztikailag) egyenlők?

Érték-, ár- és volumenindexek

Index vagy **indexszám**: közvetlenül nem összesíthető, de gazdaságilag összetartozó adatok átlagos változását mutató összetett viszonyszám.

Tegyük fel, hogy m különböző terméket értékesítünk két különböző időszakban, és az értékesítés árbevételét szeretnénk elemezni.

Jelölések:

- $q_{0,j}$: a j. termékből eladott mennyiség a bázisidőszakban
- $q_{1,j}$: a j. termékből eladott mennyiség a tárgyidőszakban
- $p_{0,j}$ ($p_{1,j}$): az j. termék egységára a bázis- (tárgy)időszakban
- $v_{0,j}$: a j. termék értékesítéséből származó árbevétel (tágabb értelemben *termelési érték*) a bázisidőszakban, számítása: $v_{0,j} = q_{0,j} \cdot p_{0,j}$
- $v_{1,j}$: a j. termék értékesítéséből származó árbevétel a tárgyidőszakban, számítása: $v_{1,j} = q_{1,j} \cdot p_{1,j}$

Érték-, ár- és volumenindexek

Egyedi indexek – mostantól a j indexeket lehagyjuk

- Egyedi volumenindexek: $i_{q,j} = \frac{q_{1,j}}{q_{0,j}} \rightsquigarrow i_q = \frac{q_1}{q_0}$
- Egyedi árindexek: $i_{p,j} = \frac{p_{1,j}}{p_{0,j}} \rightsquigarrow i_p = \frac{p_1}{p_0}$
- Egyedi értékindexek: $i_{v,j} = \frac{v_{1,j}}{v_{0,j}} = \frac{q_{1,j} \cdot p_{1,j}}{p_{1,j} \cdot p_{0,j}} \rightsquigarrow i_v = \frac{v_1}{v_0} = \frac{q_1 p_1}{q_0 p_0} = i_p \cdot i_q$

Összetett indexek:

Index fajtája	Bázisidőszaki súlyozású vagy Laspeyres-féle	Tárgyidőszaki súlyozású vagy Paasche-féle	Fisher-féle
Árindexek	$I_p^0 = \frac{\sum q_0 p_1}{\sum q_0 p_0}$	$I_p^1 = \frac{\sum q_1 p_1}{\sum q_1 p_0}$	$I_p^F = \sqrt{I_p^0 \cdot I_p^1}$
Volumenindexek	$I_q^0 = \frac{\sum q_1 p_0}{\sum q_0 p_0}$	$I_q^1 = \frac{\sum q_1 p_1}{\sum q_0 p_1}$	$I_q^F = \sqrt{I_q^0 \cdot I_q^1}$
Értékindex	$I_v = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum_{j=1}^{m} q_{1,j} p_{1,j}}{\sum_{j=1}^{m} q_{0,j} p_{0,j}}$		

Érték-, ár- és volumenindexek

Állítás: az előző fóliákon bevezetett jelölésekkel igazak a következő összefüggések (indexek különböző átlagformulái):

- $I_v = I_q^0 \cdot I_p^1 = I_q^1 \cdot I_p^0 = \frac{\sum q_0 p_0 \cdot i_v}{\sum q_0 p_0} = \frac{q_1 p_1}{\sum \frac{q_1 p_1}{i_v}}$
- $I_p^0 = \frac{\sum q_0 p_0 \cdot i_p}{\sum q_0 p_0} = \frac{\sum q_0 p_1}{\sum \frac{q_0 p_1}{i_p}}$
- $I_q^1 = \frac{\sum q_0 p_1 \cdot i_q}{\sum q_0 p_1} = \frac{\sum q_1 p_1}{\sum \frac{q_1 p_1}{i_q}}$

Megjegyzés: az egyes összetett indexek a megfelelő saját egyedi indexeik átlagai

Az indexek képleteiben lévő osztások helyett különbségeket is lehet képezni, ekkor az I és i helyett K-t és k-t írunk. Például

$k_q = q_1 - q_0$

$K_p^0 = \sum q_0 p_1 - \sum q_0 p_0$.

Érték-, ár- és volumenindexek

Gyakorlati alkalmazásaik:

- Értékindexek:
  - vállalatok árbevételének, forgalmának alakulása
  - export és import értékének változása
  - energiafelhasználás értékének változása
- Árindexek:
  - fogyasztói árindex – az infláció mérőszáma, a lakosság által vásárolt fogyasztási cikkek és szolgáltatások árainak átlagos változását fejezi ki
  - cserearányindex: egy ország által exportált termékek árindexe osztva az általa importált termékek árindexével (itt tehát árindexekből számítunk további indexeket)

Feladatok

E41.) Ágnes asszony háromféle, saját termelésű termékeket árul a Lehel téri piac őstermelői részlegén: tojásokat, krumplikat és hagymákat. Az idei (2017-es) és a tavalyi (2016-os) áprilisi forgalmát szeretné összevetni egymással. Eladásairól a következőket jegyezte fel:

Termékfajta	A forgalom értéke 2017-ben (e Ft)	Az árak alakulása, 2017/2016 (%)	Az eladott mennyiség alakulása, 2017/2016 (%)
Tojás	280	95	110
Krumpli	80	110	160
Hagyma	20	120	120

a.) Számíts érték-, ár- és volumenindexet a kofa forgalmára vonatkozóan! Értelmezd szövegesen az egyes indexeket!

b.) A forgalom értékének növekedéséből hány forint volt az ár- és a volumenváltozás hatása?

Feladatok

Szöveges értelmezések I_v-re, I_p^1-re és I_q^0-ra:
Ágnes asszony 2017 áprilisában 16,1%-kal több bevételre tett szert, mint 2016 áprilisában. Ez két részből tevődött össze: egyrészt az eladott termékek árai – 2017-es eladott mennyiségekkel számolva – átlagosan 1,1%-kal csökkentek, másrészt az értékesítés volumene – bázisévi árakkal számolva – 17,4%-kal bővült.

Szöveges értelmezések K_p^1-re és K_q^0-ra:
Az árak átlagos csökkenése – 2017-es eladott mennyiségekkel számolva – a forgalom értékének csökkenéséhez 4.131 Ft-tal járult hozzá. Az eladott mennyiségek átlagos növekedése – bázisévi árakkal számolva – a forgalom értékének növekedéséhez 56.845 Ft-tal járult hozzá.

A vizsga

- Gyakorlati jegy megléte szükséges.
- Nem elvárás az ünnepi öltözet. Nyugodtan gyertek kényelmes, hétköznapi utcai ruhában.
- A 2-eshez a teljes anyagból kell minimális ismeretekkel rendelkezni. Ha megtanulod 5-ösre az első 7 előadás anyagát, a többiből pedig nem tudsz semmit, akkor $P(\text{megbuksz a vizsga I. részén}) = 1$.
- Mindenki kap majd egy vizsgasegédletet, amin a legfontosabb képletek szerepelnek. A képletek jelentését, használatát magadtól kell tudni, az azokban lévő "betűk", jelek értelmezését a vizsgasegédlet nem tartalmazza!
- **Számológép** ($\neq$ mobiltelefon) használható
- 100 pontos, osztályozás:

karó (1)	0	–	34,99
elégséges (2)	35	–	49,99
közepes (3)	50	–	64,99
jó (4)	65	–	79,99
jeles (5)	80	–	100

A vizsga

- Bizonyos anyagrészeket nem kérek számon. Az előadásnaplóban pirossal kiemeltem, melyek ezek az anyagrészek, lásd `http://varga14.elte.hu/EaNaploLeiroMatStat17.pdf`
- írásbeli, legfeljebb 180 percig írhatjátok
- Feladattípusok:
  - tesztes vagy kifejtős számolós példák – a nagy többség
  - Definíciók, tételek, bizonyítások, módszerek bemutatása
  - A tananyag átlátását, megértését lemérő keresztkérdések
  - **R** nyelvű számítógépes output-ok, ábrák kiértékelése, szöveges értelmezése
- Két részes:
  - I. rész: 40 pontos, minimum 60%-ot (24 pontot) el kell érni – tesztfeladatok, definíciók, tételek, rövid keresztkérdések az alapvető ismeretekről
  - II. rész: 60 pontos: nagyobb lélegzetvételű feladatok, bizonyítások, számítógépes output-ok kiértékelése
- Egy próbavizsga: a honlapomon lesz, hamarosan

Konzultációs és vizsgaidőpontok

	Nap	**Időpont**	**Hely**
konzultáció	máj. 29.,hétfő	10:00	D 3-309
1. vizsga	máj. 30., kedd	10:00	D 0-822
konzultáció	jún. 12., hétfő	10:00	D 3-309
2. vizsga	jún. 13., kedd	10:00	D 0-822
konzultáció	jún. 26.,hétfő	10:00	D 3-309
3. vizsga	jún. 27., kedd	10:00	D 0-822

A vizsga megtekintése: a vizsga napján 14:00-kor a D 3-309-es szobában (kicsit később, ha addigra nem végeztem a javításokkal) vagy a konzultáció időpontjában. Ha egyik se jó, akkor egy E-mail-ben egyeztetett, mindkettőnknek megfelelő időpontban.