Kvantitatív adatelemzési módszerek felsőfokon

Dr. Alpek B. Levente
PTE TTK, Földrajzi Intézet,
Társadalomföldrajzi és Urbanisztikai Tanszék
Kapcsolat: alpeklevente@gmail.com, +36308720003

Tartalom

- Kvantitatív/kvalitatív kutatás és a statisztikai mintavétel
- Statisztikai alapfogalmak
- Leíró statisztika
- Területi statisztika
- Idősorelemzés
- Egy kis kitekintő az indexek világába: a Szenzitivitási-index
- Kapcsolatvizsgálati módszerek
- Adatredukciós módszerek

Kvantitatív/kvalitatív kutatás és a statisztikai mintavétel

Kvalitatív és kvantitatív kutatás

- KVANTITATÍV
  - Mennyiségi, számszerű adatok, hard tényezők
  - Minta jellemzően nagyobb
  - Reprezentativitás kérdése
  - Strukturált
  - Következtetés az alapsokaságra
  - Rendszerességek, szabályszerűségek keresése, trendek
  - **Megbízható**
  - „kívülállás” = émikus jelleg
  - STATISZTIKA szerepe

- KVALITATÍV
  - Minőségi, soft tényezők
  - Minta jellemzően kisebb, „Mélyfúrás” jellegű
  - Reprezentativitás (?)
  - Nem, vagy félig strukturált
  - Specifikumok keresése
  - **Érvényes**
  - „belülállás” = étikus jelleg

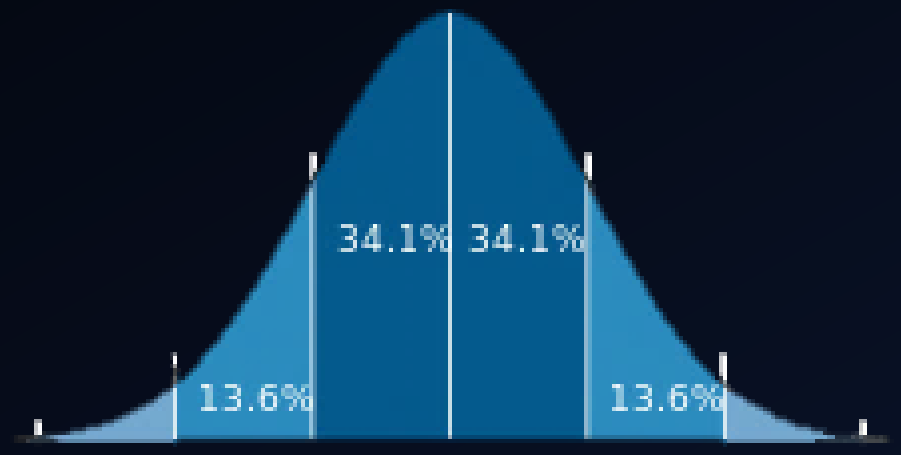
Mintavétel

- Sokaságnak nincsenek elemi térbeli egységei (geológia, meteorológia stb.) → folytonos
- Mintavételi egységek száma ismert → diszkrét (települések, személyek, háztartások, vállalkozások)
- Mintavételi terv (!)
- Primer kutatás – választott információfeldolgozási eszköztár

Statisztikai alapfogalmak

Alapfogalmak

„A statisztika a tömegesen előforduló jelenségekre, folyamatokra vonatkozó információk összegyűjtésének, leírásának, elemzésének, értékelésének és közlésének tudományos módszertana.” (Pintér J. – Ács p. 2007, p. 11)



Csoportosítása

- Leíró statisztika
- Következtetéses statisztika
- Statisztikai döntéselmélet

Alapfogalmak

- Egyed – alapegység, a megfigyelés legkisebb egysége
- Sokaság (populáció) – megfigyelt egyedek összessége
  - álló (stock, eszmei időpont), mozgó (flow)
  - véges, végtelen
  - teljes (alapsokaság), részleges (minta, mintasokaság)
- Ismérv (változó) - egyedekre vonatkozó tulajdonság
  - Közös, megkülönböztető
  - Időbeli, minőségi, területi, mennyiségi
  - Bináris (két ismérvváltozat), diszkrét, folytonos
- Ismérvváltozat - adott változó adott egyedhez tartozó realizációja

A statisztikai adat

- A statisztika számadatokkal dolgozik
- A statisztikai adat
  - Tapasztalati, empirikus
  - Mérés, számlálás útján keletkezik
  - Abszolút („közvetlenül”), származtatott („számítással”)
- Elvárások
  - Megfelelő pontosság
  - Gazdaságosság (alacsony relatív költség)
  - Gyorsaság (mennyire aktuális?)
- Statisztikai mutatószámok, indexek
- Modellek!

Mérési skálák

- Kategóriás változók: jellemzően verbális ismérvváltozatok
  - Nominális: típusokba sorolás (nem):

    $=, \neq$
  - Ordinális: sorbarendezés, rangszám (osztályzatok):

    $=, \neq, <, >$
- Numerikus változók: számszerű ismérvváltozatok
  - Intervallum: azonos távolság az értékek között, választott nulla pont (hőmérséklet, de zéró érték nem azonos a tulajdonság hiányával)

    $=, \neq, <, >, +, -$
  - Arány: értékek aránya azonos, abszolút nulla pont (távolság)

    $=, \neq, <, >, +, -, \times, \div$

Adatállományok

- A változók és/vagy egyedek rendszerezett felsorolása
  - Idősoros adatállomány
    - A változó értékei időrendben felsorolva
    - Jellemzően állandó frekvencia
  - Keresztmetszeti
    - Több egyed jellemzői egy időszakra, időpontra
  - Panel
    - Több egyed jellemzői több időszakra, vagy időpontra

Adatállományok szerkezete

Azonosító	Településnév	Megye	Kistérség	Járás	Jogállás	HH	LHH	Kedvezményezett	Fejlesztendő	Komplex programmal fejlesztendő
1	Ipolyszög	Nógrád	Balassagyarmati	Balassagyarmati	Község	0	0	1	0	0
2	Somoskőújfalu	Nógrád	Salgótarjáni	Salgótarjáni	Község	0	0	1	0	0
3	Ráckeresztúr	Fejér	Martonvásári	Martonvásári	Község	0	0	0	0	0
4	Martonyi	Borsod-Abaúj-Zemplén	Edelényi	Edelényi	Község	0	1	1	1	1
5	Kakasd	Tolna	Bonyhádi	Bonyhádi	Község	0	0	0	0	0
6	Narda	Vas	Szombathelyi	Szombathelyi	Község	0	0	0	0	0
7	Sávoly	Somogy	Marcali	Marcali	Község	1	0	1	0	0
8	Gönyű	Győr-Moson-Sopron	Győri	Győri	Község	0	0	0	0	0
9	Dunaremete	Győr-Moson-Sopron	Mosonmagyaróvári	Mosonmagyaróvári	Község	0	0	0	0	0
10	Mánd	Szabolcs-Szatmár-Bereg	Fehérgyarmati	Fehérgyarmati	Község	0	1	1	1	1
11	Gutorfölde	Zala	Lenti	Zalaegerszegi	Község	0	0	0	0	0
12	Maróc	Zala	Letenyei	Letenyei	Község	1	0	1	1	0
13	Budapest 17. ker.	főváros	Budapesti	Budapest 17. ker.	BP_kerület	0	0	0	0	0
14	Csanádalberti	Csongrád	Makói	Makói	Község	1	0	1	1	0
15	Mihályfa	Zala	Zalaszentgróti	Zalaszentgróti	Község	1	0	1	0	0
16	Foktő	Bács-Kiskun	Kalocsai	Kalocsai	Község	1	0	1	0	0
17	Muhi	Borsod-Abaúj-Zemplén	Miskolci	Tiszaújvárosi	Község	0	0	0	0	0
18	Komádi	Hajdú-Bihar	Berettyóújfalui	Berettyóújfalui	Város	0	1	1	1	1
19	Csopak	Veszprém	Balatonfüredi	Balatonfüredi	Község	0	0	0	0	0
20	Nemesvámos	Veszprém	Veszprémi	Veszprémi	Község	0	0	0	0	0
21	Isztimér	Fejér	Móri	Móri	Község	0	0	0	0	0
22	Balatonfűzfő	Veszprém	Balatonalmádi	Balatonalmádi	Város	0	0	0	0	0
23	Martonfa	Baranya	Pécsváradi	Pécsváradi	Község	1	0	1	0	0
24	Vág	Győr-Moson-Sopron	Csornai	Csornai	Község	0	0	0	0	0
25	Vát	Vas	Szombathelyi	Szombathelyi	Község	0	0	0	0	0
26	Vásárosdombó	Baranya	Sásdi	Hegyháti	Község	0	1	1	1	1
27	Abaújvár	Borsod-Abaúj-Zemplén	Abaúj-Hegyközi	Gönci	Község	0	1	1	1	1
28	Hejőkürt	Borsod-Abaúj-Zemplén	Tiszaújvárosi	Tiszaújvárosi	Község	0	0	0	0	0
29	Tiszabábolna	Borsod-Abaúj-Zemplén	Mezőcsáti	Mezőkövesdi	Község	0	1	1	0	0
30	Kaba	Hajdú-Bihar	Püspökladányi	Püspökladányi	Város	0	0	1	1	1
31	Etyek	Fejér	Bicskei	Bicskei	Község	0	0	0	0	0
32	Baktalórántháza	Szabolcs-Szatmár-Bereg	Baktalórántházai	Baktalórántházai	Város	0	1	1	1	1
33	Kincsesbánya	Fejér	Móri	Móri	Község	0	0	0	0	0
34	Újkígyós	Békés	Békéscsabai	Békéscsabai	Község	0	0	0	0	0
35	Nagybajcs	Győr-Moson-Sopron	Győri	Győri	Község	0	0	0	0	0

Adatforrások

- Primer
  - Pl. megkérdezéssel, megfigyeléssel, kísérlettel „közvetlenül” gyűjtött információk
- Szekunder
  - Fizetős adatállományok, adatbankok
  - Statisztikai hivatalok
  - Weboldalak
  - Minden releváns nyomtatott vagy elektronikus forrás

A statisztikai adatok csoportosítása, osztályozása

- A statisztikai sokaságnak valamely ismérv szerinti rendszerezése.
- Csoportképző ismérv (HALMAZELMÉLETI MEGFONTOLÁSOK)
- Nomenklatúra (TEÁOR)
- Kombinatív csoportosítás

Statisztikai sorok

- Egyedek egy szempont szerinti jellemzése
- Típusai
  - Azonos fajta adatokból áll
    - Csoportosító - a vizsgált sokaságnak valamilyen megkülönböztető ismérv alapján történő csoportosítása (SZUM)
    - Összehasonlító - ugyanazon jelenségre vonatkozó megfigyelt értékek a megkülönböztető ismérv többféle változata esetén (SZUM nem)
  - ÉS
    - Időbeli sor
    - Minőségi sor
    - Mennyiségi sor
    - Területi sor
  - Eltérő fajta (leíró sor)
- Gyakorisági sor (Fi)

Gyakorisági sorok képzése

- Gyakorisági sor fogalma
- Célja: adatrendezés és -tömörítés
- Alapja az egyes ismérvek adott esetben többszöri előfordulása. Egy ismérv gyakorisága:
  - $\sum_{i=1}^{k} f_i = n$, ahol i az ismérvváltozatok száma (k<n)

Tulajdonság	Gyakoriság
x_1	f_1
x_2	f_2
…	…
x_k	f_k
Összesen	N

Relatív gyakoriságok képzése

- Relatív gyakorisági sor fogalma
- Relatív gyakoriság: gyakoriságokból számított megoszlási viszonyszámok:
  - $g_i = \frac{f_i}{n}$
- Osztályközök kialakítása a jelenségek információtartalmának megfelelő tömörítése céljából!

ismérv	kód	gyakoriság (db)	relatív gyakoriság %	érvényes %
a				
b				
c				
…				
$\sum$				
Magyarázat		*gyakorisági megoszlás*	*az egyes válaszok % aránya a hiányt és a nem válaszolt elemeket is figyelembe véve*	*a válaszok %-os aránya az érvényes válaszadók között*

Statisztikai táblák típusai

- Egyszerű tábla
- Csoportosító tábla
- Kombinációs tábla

Statisztikai táblák szerkesztési elvei

- a táblázatnak legyen címe
- %-os táblázatkor a változóba tartozó egyedek db számát is fel kell tűntetni
- a hiányzó esetek számát is meg kell jeleníteni
- adatok forrásának megjelenítése
- adatok időszerűsége, adatgyűjtés éve
- a változó tartalmát is ismertetni kell (szövegben és táblázatban)
- ismertté kell tenni az egyes változók attribútumait

Kontingencia táblák (dimenziószám)

	A1	**A2**	**…**	**Ai**	**Össz.**
B1	F11				F1.
B2					
…					
Bm				Fmi	
Össz.	F.1				N

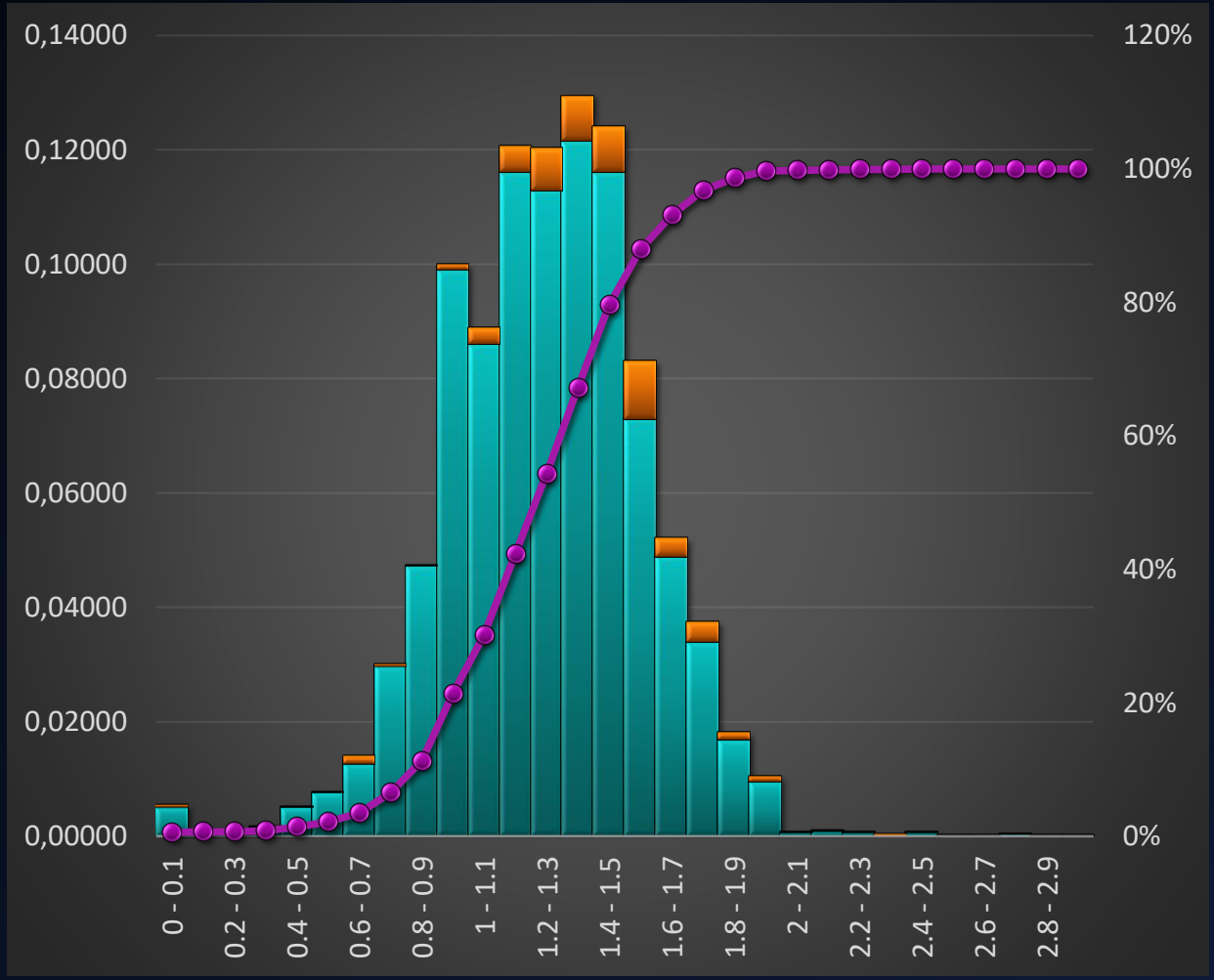
Viszonyszámok

- Gyakorisági sorok viszonyszámai
  - Megoszlási – részt hasonlít egészhez (relatív gyakoriság)
  - Koordinációs – részt hasonlít részhez
  - Leíró sorok viszonyszámai – Intenzitási viszonyszám (két kapcsolódó adat hányadosa)
  - Azonos vagy különböző mértékegységű
  - Egyenes vagy fordított
  - Nyers vagy tisztított
    - A viszonyítandó mennyiséget a teljes sokaságnak egy vele szorosabb kapcsolatban lévő részével osztjuk → tisztított intenzitási viszonyszám
    - A nyers viszonyszám felbontható: A/B=A/b*b/B; b részhalmaza B (egy tisztított és egy megoszlási viszonyszám szorzata)
- Idősorok viszonyszámai – Dinamikus viszonyszámok
  - Bázis (minden értéket egy kitűntetett értékhez)
  - Lánc (minden időszak értékét az előző időszak értékéhez)

Gyakorisági sorok grafikus ábrázolása

- Empirikus sűrűségfüggvény (Hisztogram)
- Empirikus eloszlásfüggvény

+1 relatív gyakorisági sor, kumulált gyakoriságok (alulról, felülről), értékösszegsor



Empirikus sűrűségfüggvény (hisztogram)

A görbe alatti terület az x_1 és x_2 pontok között annak valószínűségét fejezi ki, hogy a kísérlet eredménye az x_1 és x_2 értékek közé esik.

Fontosabb tulajdonságai:

- $f(x) \geq 0$
- A görbe alatti terület egységnyi

- Osztályközök kialakítása!

Osztályközök kialakítása, osztályközös gyakorisági sor

- Miért alkalmazzuk?
  - Adatreprezentációs célt szolgál
  - Adatvédelmi szempontok
  - Kijelölésük elvei:
    - Egyértelmű besorolást tegyenek lehetővé
    - Teljes: nyitott alsó és felső intervallumhatárok
    - Üres osztályok lehetőleg ne legyenek, vagy csak alacsony számban
  - „Szép”, értelmezhető osztályközhatárok
  - Törekedjünk az azonos szélességű osztályközökre, ettől csak nagyon indokolt esetben térjünk el (értelmezhetőség!)
- Hány intervallumot alkalmazzunk?
  - $k = 1 + (3.3 * \lg(n))$, ahol k az osztályközök száma és n a sokaság elemszáma
- Intervallumok hossza?
  - $h = \frac{x_{max} - x_{min}}{k}$, ahol h az osztályköz hossza, x_{max} a legnagyobb ismérvérték, x_{min} a legkisebb ismérvérték

Empirikus sűrűségfüggvény készítése

- Adatok nagyság szerinti sorba rendezése
- Gyakorisági intervallumok meghatározása (képlet!)
- Az intervallumokba eső adatok számának meghatározása
- A sűrűségfüggvény értékeinek meghatározása (a sűrűségfüggvény alatti területet téglalapokkal adjuk meg) EXCEL! → $f(x_i) = \frac{k_i}{n * \Delta_i}$
  - n = adatok száma
  - k_i =intervallumba eső esetek száma
  - Δ_i = i-ik intervallum szélessége
  - $f(x_i)$ = i-ik intervallumhoz tartozó téglalap magassága

Empirikus eloszlásfüggvény

Egy adott x_a értékhez tartozó $F(x_a)$ függvényérték azt fejezi ki, hogy milyen valószínűséggel fordulnak elő az x_a-nál kisebb, vagy azzal egyenlő kísérleti eredmények

Fontosabb tulajdonságai:

- Monoton növekvő $f(x_1) \leq f(x_2)$ ha $x_1 < x_2$
- Minimuma 0, maximuma 1.

Empirikus eloszlásfüggvény készítése

- Adatok nagyság szerinti sorba rendezése
- A keresett valószínűséget a relatív gyakorisággal közelítjük, az eloszlásfüggvény értéke x_a-nál

$$F(x_a) = \frac{k}{n}$$

  - n = adatok száma (összes adat)
  - k_i = azon adatok száma, amelyekre igaz, hogy $x<x_a$

Kumulált gyakoriságok (f_i') és értékösszegsor

Tulajdonság érték	Db (értékösszegek)	Kumulált gyakoriságok (értékösszegek)		Kumulált relatív gyakoriságok (értékösszegek)	
		Alulról ($x \leq x_f$)	Felülről ($x \geq x_f$)	Alulról	Felülről
1	154	154	285	54%	100%
2	73	227	131	80%	46%
3	29	256	58	90%	20%
4	15	271	29	95%	10%
5	9	280	14	98%	5%
6	5	285	5	100%	2%
Összesen	285	-	-	-	-

Leíró statisztika

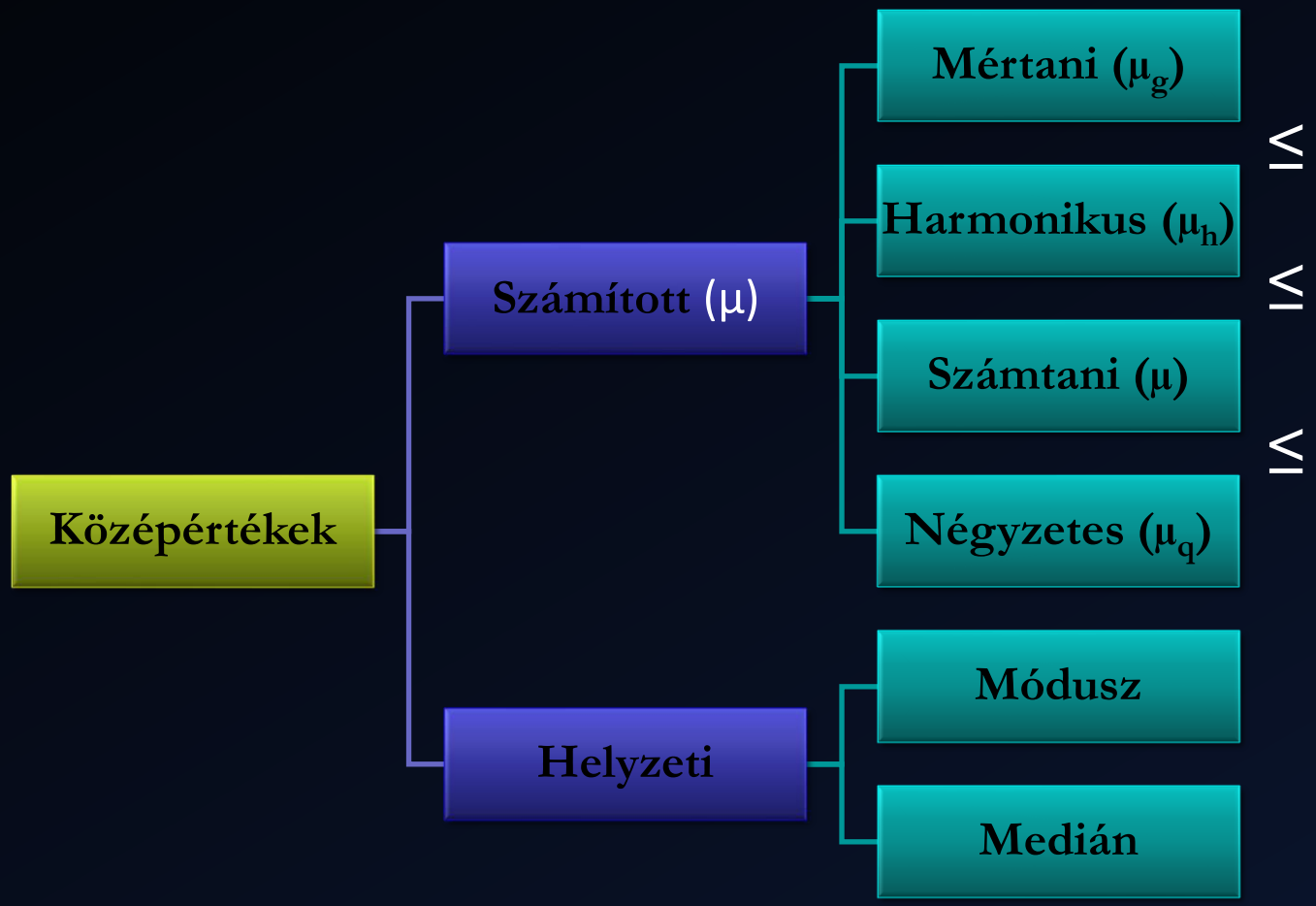
Bevezető

- Nagy mennyiségű adat jellemzése egy-egy érték segítségével
- Várható érték és ingadozás
- Összehasonlítás egy érték mentén (országok, megyék, egyének stb.)
- A sokaság felosztása meghatározott egységekre (mekkora összegnél keres többet a társadalom 50%-a)
- Adatok előkészítése, jellemzése összetett statisztikai vizsgálatok előtt

Középértékek

- Célja: adattömörítés, számszerű jellemzés
- A középértékekkel kapcsolatos főbb elvárások:
  - Közepes helyet foglaljanak el (inkább az átlagok)
  - Tipikus elemek legyenek (inkább a helyzeti középértékek)
  - Jól értelmezhető
  - Könnyen meghatározható (könnyen számítható)
  - Robosztus (nem érzékeny a kiugró értékekre, stabil)
- Dimenziójuk megegyezik az általuk jellemzett ismérv dimenziójával

Középértékek csoportosítása



Számított középértékek – Számtani átlag

- Az átlagolandó értékek helyébe téve azok ÖSSZEGE azonos marad ($n\bar{x} = \sum x$)
- Számítása: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- A számtani átlagtól mért eltérések algebrai összege zéró $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- Számítása gyakorisági sorból (súlyozott átlag formulával): $\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$, ahol f_i az adott súlyszám (a gyakoriságok, vagy relatív gyakoriságok értéke), x_i pl. az osztályközép
- A súlyozott számtani átlag függ:
  - Az átlagolandó értékek (abszolút) nagyságától
  - A súlyok viszonylagos nagyságától (súlyarányok)

Számított középértékek – Harmonikus átlag

- Az átlagolandó értékek helyébe téve azok RECIPROKOK ÖSSZEGE azonos marad ($\sum_{i=1}^{n} \frac{1}{x_i} = n \frac{1}{\bar{x}_h}$)
- Számítása: $\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
- Számítása súlyozott formulával: $\bar{x}_h = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} \frac{1}{x_i}}$, ahol f_i az adott súlyszám (alkalmazása pl. középárfolyam meghat.)
- Alkalmazása (pl. gyártási sebesség)
  - Az értékek reciprokainak összege értelmes
  - Fordított intenzitási viszonyszámok

Harmonikus átlag példa

- Adott 5 gép, amelyen kosárlabdákat gyártanak
- A gépek teljesítménye
  - A – 3 db/perc
  - B – 2.5 db/perc
  - C – 7 db/perc
  - D – 1 db/perc
  - E – 1.5 db/perc
- Mennyi az átlagos megmunkálási idő?
  - A munkafolyamat mekkora részét végzik el az egyes gépek 1 perc alatt? $\frac{1}{3}, \frac{1}{2.5} \dots stb.$
  - Az 5 gép együttesen a munkafolyamat $\sum_{i=1}^{n} \frac{1}{x_i}$, azaz $\frac{89}{35}$, amely szerint az öt gép együtt a munkafolyamat 254%-t végzi el percenként, az egy gépre eső átlagos teljesítmény pedig: $\frac{\frac{89}{35}}{5} = 0.5$ db/perc, az egy termékegységre jutó átlagos munkaidő ennek a reciproka, azaz 1.96 perc/db. Ugyanezt az értéket meghatározhatjuk a harmonikus átlag alkalmazásával: $\bar{x}_h = \frac{5}{\frac{1}{3}+\frac{1}{2.5}+\frac{1}{7}+\frac{1}{1}+\frac{1}{1.5}} = 1.96$ perc/db

Csoportosított adatok átlaga

- Csoportosított adatok alapján kívánjuk meghatározni a csoport egészére vonatkozó átlagot
- A főátlag a csoportátlagok számtani átlaga, azaz:

$$\bar{x} = \frac{\sum_{j=1}^{m} n_j \bar{x}_j}{\sum_{j=1}^{m} n_j}, \text{ ahol}$$

n_j – a megfigyelések száma a j-edik csoportban
$\bar{x}_j$ – a j-edik csoport átlaga
m – 1,2,…m a csoportok száma

vagy az arányok felhasználásával (!)

Csoportosított adatok szórása

- Nem származtatható közvetlenül a részsokaságok szórásaiból
- Szórás felbontása:
  - Belső szórás (belső szórásnégyzet, belső variancia) → egyes megfigyelt értékek átlagos eltérései saját csoportátlaguktól (csoporton belüli szórás) milyen nagyságrendű
  - Külső szórás (külső szórásnégyzet, külső variancia) → a csoportok átlagainak szóródása
  - Teljes szórás (teljes szórásnégyzet) → az egész sokaság szórása

Csoportosított adatok szórása

- **Belső szórásnégyzet**

$$\sigma_B^2 = \frac{\sum_{j=1}^{m} n_j \sigma_j^2}{n}$$

n_j – a megfigyelések száma a j-edik csoportban
n – összes elemszám
m – 1,2,…m a csoportok száma
σ_j^2 – a j-edik csoport szórásnégyzete

- **Külső szórásnégyzet**

$$\sigma_K^2 = \frac{\sum_{j=1}^{m} n_j (\bar{x}_j - \bar{x})^2}{n}$$

n_j – a megfigyelések száma a j-edik csoportban
n – összes elemszám
m – 1,2,…m a csoportok száma
$\bar{x}_j$ – a j-edik csoport átlaga
$\bar{x}$ – a főátlag

- **Teljes szórásnégyzet**

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

Számított középértékek – Mértani(geometriai) és négyzetes átlag (kvadratikus)

- Mértani
  - Az átlagolandó értékek helyébe téve azok SZORZATA azonos marad
  - Számítása: $\bar{x}_q = \sqrt{\frac{\sum x_i^2}{N}}$
  - Számítása súlyozott formulával: $\bar{x}_q = \sqrt{\frac{\sum_j F_j x_j^2}{N}}$, ahol f_i az adott súlyszám és $n = \sum f_i$
  - Alkalmazása, ha az átlagolandó értékek között szorzatszerű viszony van (értek növekedési üteme ismert, pl. %-ban megadva → átlagos növekedés)
- Négyzetes
  - Az átlagolandó értékek helyébe téve azok NÉGYZETÖSSZEGE azonos marad (inkább technikai szerepe van)
  - Számítása: $\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$
  - Számítása súlyozott formulával: $\bar{x}_g = \sqrt[n]{\prod_{i=1}^{k} x^{i^{f_i}}}$, ahol f_i az adott súlyszám

Helyzeti középértékek

- Módusz – A leggyakrabban előforduló (legjellemzőbb) érték
  - Egy vagy több – több móduszú eloszlások
  - Nem minden esetben létezik
  - Nem feltétlenül azonos az átlaggal (!)
- Medián – A rangsorba rendezett (sorba rendezett) adatok középső eleme – számítása páros és páratlan elemszám esetén eltér:
  - Ha N páros - $Me = \frac{X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}}{2}$
  - Ha N páratlan – $Me = X_{\left(\frac{N+1}{2}\right)}$
- Kvantilisek – Osztópontok, a leggyakrabban alkalmazott:
  - Tercilis (harmadoló - T_t), kvartilis (negyedelő - Q_q), kvintilis (ötödölő - K_k), decilis (tizedelő - D_d), perventilis (századoló - P_p)
  - Rangsorba rendezett adatsor felosztása egyenlő egységekre

Helyzeti középértékek gyakorisági sorokból - medián

- Kumulált gyakorisági sor alapján
  - Diszkrét mennyiségi ismérvek esetén a medián értke azonos azzal az értékkel, amelyhez tartozó kumulált gyakoriság tartalmazza a medián sorszámát
    - Pl.: 3, 8, 8, 6, 3, 7, 4, 6, 6, 8, 3, 4, 5, 10, 8, 4, 4, 10, 1, 5, 4, 2, 2, 4, 4, 4, 6 AKKOR Me=4; a sorba rendezett adatok 14. eleme, gyakorisági sorról leolvasva

Sorcímkék	Mennyiség / Értékek	Kumm.gy.
1	1	1
2	2	3
3	3	6
4	**8**	**14**
5	2	16
6	4	20
7	1	21
8	4	25
10	2	27
Végösszeg	27	

  - Osztályközös gyakorisági sor esetében:

$$Me = x_{me,a} + \frac{s - f'_{me-1}}{f_{me}} * \mathrm{h}$$

$X_{me,a}$ – mediánt magában foglaló osztályköz alsó (nem technikai) határa; s – n/2, a medián sorszáma, f'_{me-1} – a mediánt megelőző osztályköz kumm. gy., f_{me} – a mediánt tartalmazó osztályköz gyakorisága, h – a mediánhoz tartozó osztályköz hossza

Helyzeti középértékek gyakorisági sorokból - módusz

- Kumulált gyakorisági sor alapján
  - Diszkrét mennyiségi ismérvek esetén a leggyakoribb érték(ek)
  - Folytonos esetben a gyakorisági görbe maximumhelyei
  - Osztályközös gyakorisági sorból
    - Modális osztályköz – a legnagyobb gyakorisággal rendelkező osztályköz kijelölése
    - Módusz becslése: $x_{mo,a} + \frac{k_1}{k_1+k_2} * h$

- $X_{mo,a}$ – a modális osztályköz alsó (nem technikai) határa; k_1 – a modális osztályköz és a megelőző osztályköz gyakoriságának különbsége, k_1 – a modális osztályköz és az azt követő osztályköz gyakoriságának különbsége , h – a modális osztályköz hossza

Szóródási mérőszámok

- A vizsgált sokaság egyedeinek adott változó szerinti eltéréseinek mérése:
  - Egymástól
  - egy kitüntetett értéktől.
- Szóródás hiánya esetén nulla

Szóródási mérőszámok

- Terjedelem $R = X_{(N)} - X_{(1)}$ (hátránya, hogy csak a szélsőértékekre épít)
- Interkvartilis terjedelem $IQR = Q_3 - Q_1$ (értékek 50%-t fogják közre)
- Átlagos (abszolút) eltérés $\delta = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$
  - Súlyozott formában: $\delta = \frac{1}{n}\sum_{i=1}^{k} f_i|x_i - \mu|$, ahol $n = \sum f_i$
  - Az egyes ismérvek átlagosan mennyivel térnek el az átlagtól
- Szórás
- Variancia
- Relatív szórás

Szóródási mérőszámok – szórás, variancia, relatív szórás

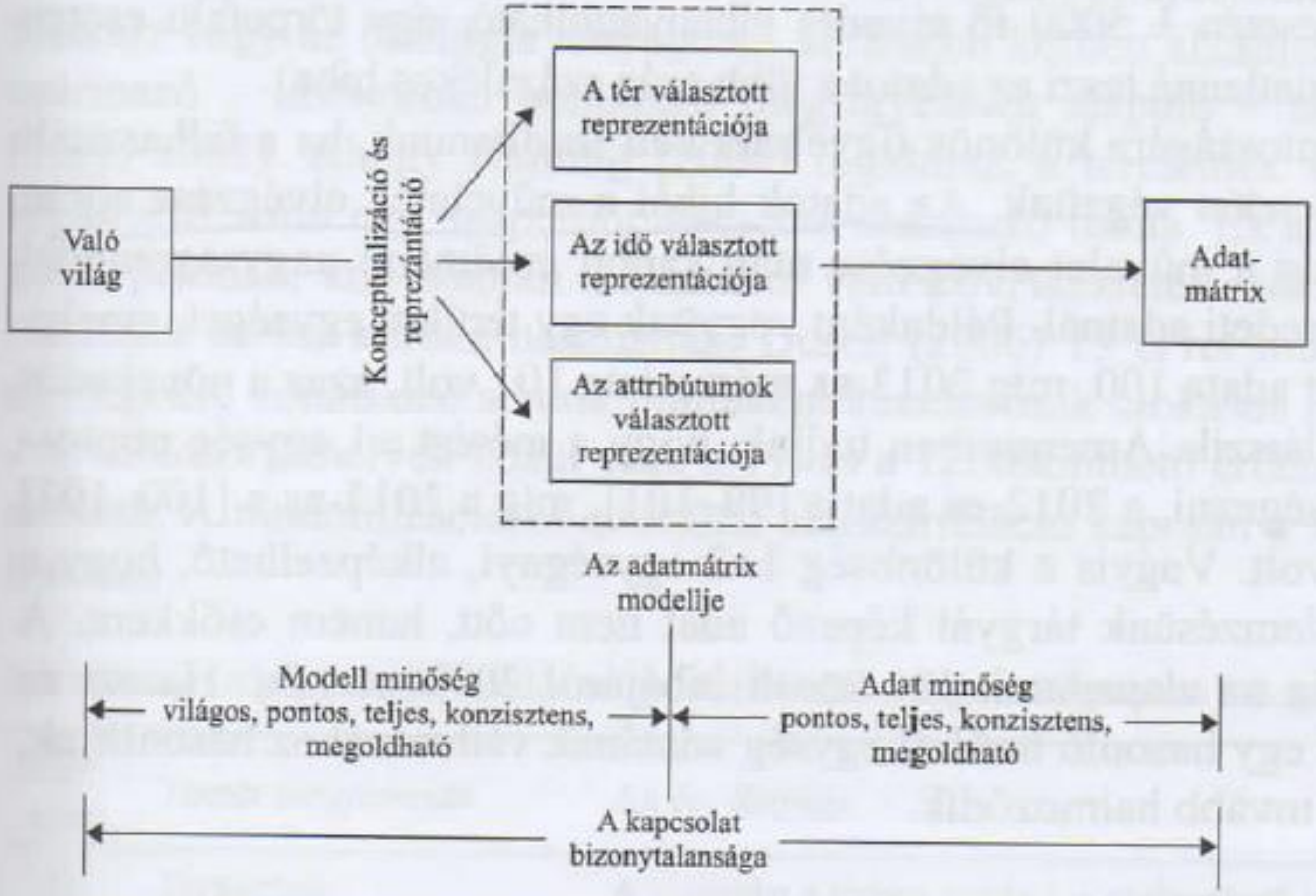
- Előnye, hogy nincs benne szisztematikus torzító hatás
- Szórás → $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Empirikusan korrigált szórás → $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Súlyozott szórás → $\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{k} f}}$
- Variancia → σ^2 (alapvetően technikai jelentőséggel bír)
- Relatív szórás (variációs koefficiens) → $V = \frac{\sigma}{\bar{x}}$ (mértékegység független, lehetővé teszi az összehasonlítást, %)

Alakmutatók

- **Ferdeség**
  - A normális eloszlás görbéjéhez viszonyítva.
- **Csúcsosság**
  - A „normális” eloszláshoz viszonyítva (zérus csúcsosság)
  - Csúcsos eloszlások esetén pozitív csúcsosság
  - Lapos eloszlások esetén negatív csúcsosság

Területi statisztika

Földrajzi valóság – adatmátrix



2.2 ábra. A földrajzi valóságtól az adatmátrixig

Forrás: Haining (2004, 58.)

A tér szerepe

- „A területi elemzés karakterét más társadalomtudományi kutatásokkal szemben az adja, hogy a vizsgálatok középpontjában a társadalom (általában mennyiségi térparaméterekkel is azonosítható) térbelisége áll, a jellegadó megfigyelési egységek pedig a különböző területegységek.”

  - TÁRSADALMI SZFÉRA, VAGY JELENSÉG térbeli elemzése, térszerkezet
  - LEHATÁROLT TÉRRÉSZ, TERÜELTEGYSÉG VIZSGÁLATA (pl. régió)

A TERÜLETI ÖSSZEHASONLÍTÁS DIMENZIÓI:

- *Egy kiválasztott* egység vizsgálata (esettanulmányok)
- *Több, összehasonlított* egység
- *Területegységek rendszerének* vizsgálata (pl. egy ország összes régiója)

(Nemes Nagy 2005)

A (területi) polarizáltság mérőszámai (A) és a szórás-típusú mérőszámok (B)

- A,
- Az adatsor terjedelme (range-arány) - legnagyobb és legkisebb ismérvérték hányadosa
- A szóródás terjedelme (range) - MAX és MIN ismérvérték különbsége
- Relatív range (relatív terjedelem) - MAX és MIN érték különbségét az adatsor átlagához
- Duál-mutató (Éltető – Frigyes index) - a teljes megoszlás átlaga fölötti értékek átlagának és a teljes megoszlás átlaga alatti értékek átlagának a hányadosa
- B,
- Szórás
- Relatív szórás
- Súlyozott szórás
- Súlyozott relatív szórás
- Átlagos (abszolút) eltérés
- Logaritmikus szórás

(Nemes Nagy 2005)

Területi megoszlások eltérését mérő indexek, koncentráció mérése

- Fogalma: jelenségek sűrűsödése, tömörülése, összpontosulása (pl. térbeli koncentráció)
- Vizsgálható az ismérv gyakorisági és értékösszeg-eloszlásának összehasonlításával (Koncentrációs táblázat)
  - Magas relatív gyakoriság – alacsony értékösszeg
  - Alacsony relatív gyakoriság – magas értékösszeg
- Kiszámítjuk a kumulált relatív gyakoriságokat és a kumulált relatív értékösszegeket, majd ezeket egymáshoz viszonyítjuk
- Erős koncentráció: a sokaság nagy hányada a teljes értékösszeg kis hányadát bírja

Koncentráció

- Fogalma: jelenségek sűrűsödése, tömörülése, összpontosulása (pl. térbeli koncentráció)
- Vizsgálható az ismérv gyakorisági és értékösszeg-eloszlásának összehasonlításával (Koncentrációs táblázat)
  - Magas relatív gyakoriság – alacsony értékösszeg
  - Alacsony relatív gyakoriság – magas értékösszeg
- Kiszámítjuk a kumulált relatív gyakoriságokat és a kumulált relatív értékösszegeket, majd ezeket egymáshoz viszonyítjuk
- Erős koncentráció: a sokaság nagy hányada a teljes értékösszeg kis hányadát bírja

Területi megoszlások eltérését mérő indexek, koncentráció mérése

- *Koncentrációs (Hirschman – Herfindahl) – index*

Valamely naturális jellemző területegységek közötti koncentráltságának mértékét számszerűsíti. A megoszlást az index tulajdonképp a teljesen egyenleteshez (amikor minden megfigyelési egység részesedése azonos) viszonyítja. 0,6 feletti értéke már erős koncentráltságra, monopolhelyzetre utal.

$$1/n \leq K \leq 1$$

Koncentráció mérése mutatószámok segítségével – **Hoover-index**, *Robin Hood-index, Disszimilaritási indexnek, Szegregációs indexnek, Krugman index*

- Szerepe: két mennyiségi ismérv területi megoszlásának eltérése, számszerűsítése
- Szimmetrikus, a két összevetett megoszlás sorrendje felcserélhető
- Számítása:

$$h = \frac{\sum_{i=1}^{n} |x_i - f_i|}{2}, \text{ ahol}$$

$\sum x_i = 100$, és $\sum f_i = 100$

  - 0≤h≤100; mértékegysége %
  - Fogalma: megadja, hogy az egyik mennyiségi ismérv hány százalékát kell a területegységek között átcsoportosítani ahhoz, hogy területi megoszlása a másik mennyiségi ismérvével azonos legyen

Koncentráció mérése mutatószámok segítségével – Gini-index

- Szerepe: a Lorenz-görbe és a négyzet átlója által bezárt terület nagyságának mérése
- Számítása:

$$G = \frac{1}{2\bar{x}n^2} \sum \sum \left| x_i - x_j \right|, \text{ ahol}$$

x_i=megoszlási viszonyszámként megadott területi jellemző az i. területegységben,
x_j=megoszlási viszonyszámként megadott területi jellemző a j. területegységben
$\bar{x}$=x_i átlaga

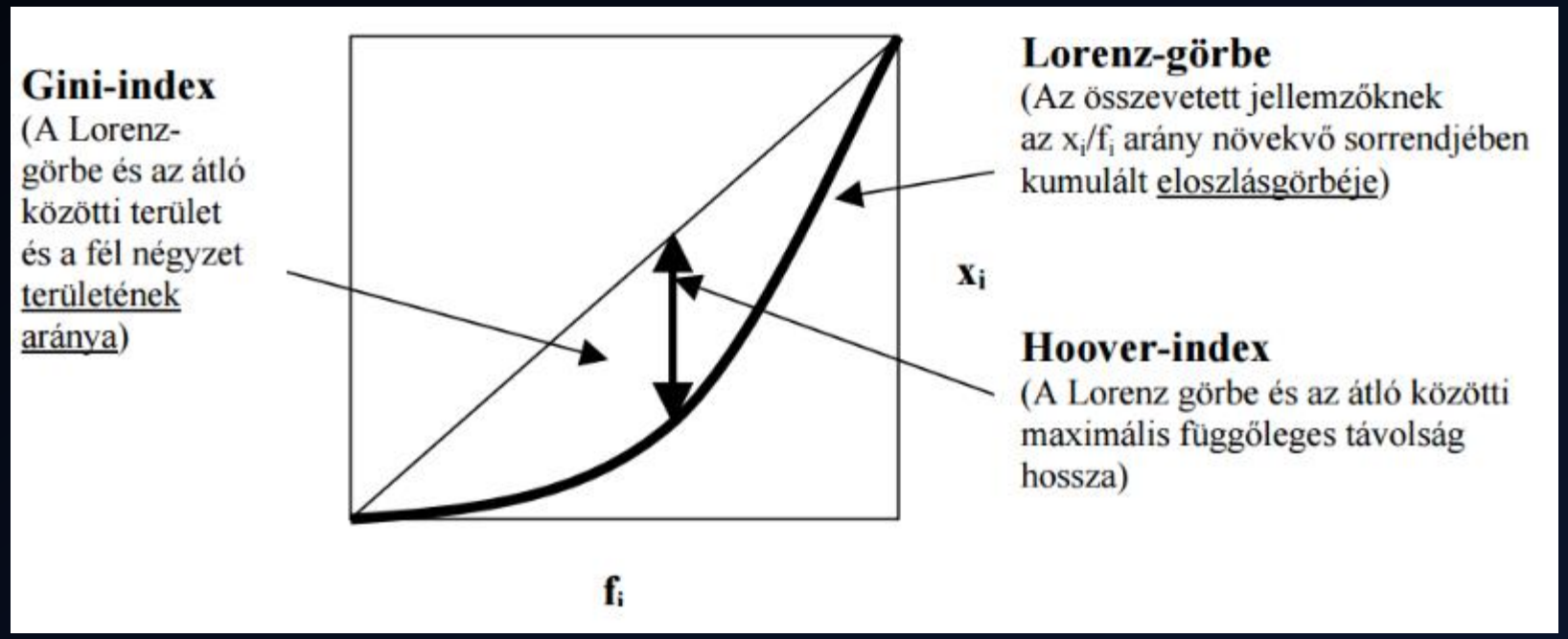
- 0≤G≤1; dimenzió nélküli szám
- Fogalma: minden megfigyelési egység részarányának többi egységtől való átlagos eltérésének viszonyítása az átlaghoz

Koncentráció ábrázolása – **Lorenz-görbe**

- Két tengely (egységoldalú négyzet)
  - Kumulált relatív gyakoriságok (g_i') → x-tengely
  - Kumulált értékösszegek (z_i') → y-tengely
- Amennyiben $g_i' = z_i'$, nincs koncentráció
- Teljes koncentráció esetén a görbe a tengelyekhez „illeszkedik”
- Görbe és átló által bezárt terület nagysága a koncentráció relatív nagyságát érzékelteti
- Több időpontra is elkészíthető, a koncentráció változása nyomon követhető
- Több jelenség koncentrációja összehasonlítható

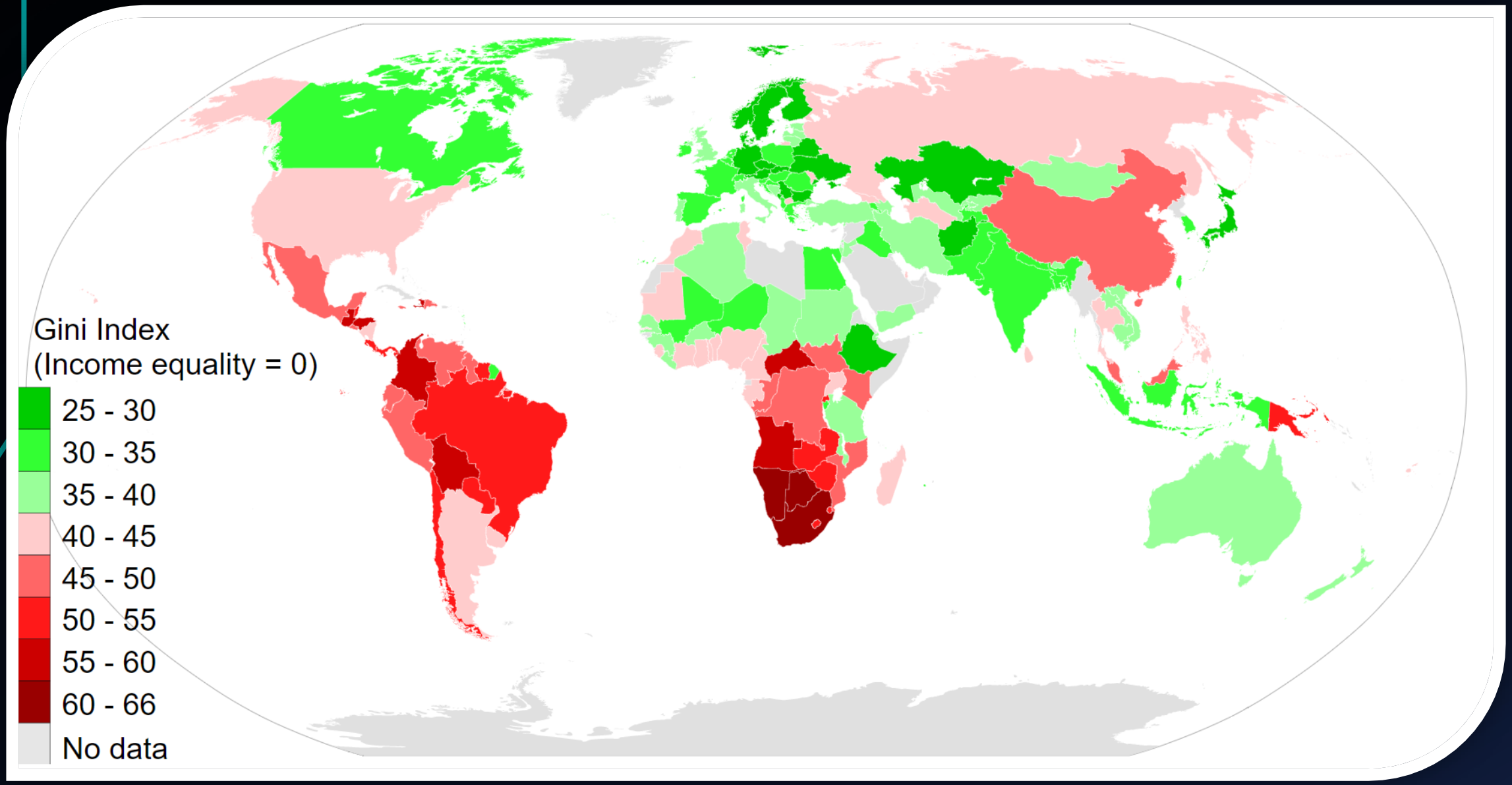
Koncentráció mérése

- A Gini-index, a Lorenz-görbe és a Hoover-index kapcsolata:



http://geogr.elte.hu/REF/REF_Kiadvanyok/REF_RTT_11/RTT-11-03-teregyenlotlenseg.pdf

Gini-index



https://en.wikipedia.org/wiki/Gini_coefficient

A térbeli szegregáció mérőszámai

- **Szegregáció**: a társadalmi csoportok, etnikumok térbeli elkülönülése.

- Disszimilaritási index (lsd. korábban)
- Interakciós index
- Izolációs index

További mutatószámok

Térbeli koncentráció méréséhez

- Entrópia
- Theil-index
- Átlagos logaritmikus eltérés

Térbeli szegregációhoz:

- Diverzitásindex
- Ellison-Glaeser koncentrációs index
- Maurel-Sédillot G_A-mutatója

Idősorelemzés

Bevezetés

- Célja:
  - Különböző időpontok összehasonlítása
  - Változás, változékonyság
  - Az értékek időbeli alakulása mögött húzódó hatások felbontása

Idősorok típusai

- Idősorok típusai:
  - Tartamidősorok – Adott időtartamok alatt bekövetkező változásokat tüntetnek fel (pl. érbevétel 2010-2016)
    - Flow-típusú, mozgó sokaság
    - Az adatokat egy időszakra vonatkoztatjuk
    - Összegzésnek van értelme
  - Állapotidősorok – Az egyes időpontok adatait mutatják (pl. Mo. népessége 2000-2016-ban minden évben, egy cég létszámadatai)
    - Stock-típusú, álló sokaság
    - A megadott adatok az adott időszak végére, vagy elejére vonatkoznak (megadjuk, hogy pontosan melyikre, január 1., vagy dec. 31.)
    - Összegzésnek nincs értelme

Idősorok típusai - példa

- Vegyük egy vállalkozás készletállományát és a készletből történő értékesítés értékének alakulását

Állapot idősor

Tartam idősor

Hónap	Készletállomány nagysága (db)	Értékesítés összege (Ft)
Június	26590	541 926 Ft
Július	24731	429 429 Ft
Augusztus	23204	352 737 Ft
Szeptember	22074	261 030 Ft
Október	19652	559 482 Ft
November	17858	414 414 Ft
December	26590	541 926 Ft
SZUM	Nincs értelme	2 559 018 Ft

A féléves értékesítés összege

Az idősorelemzés eszközei

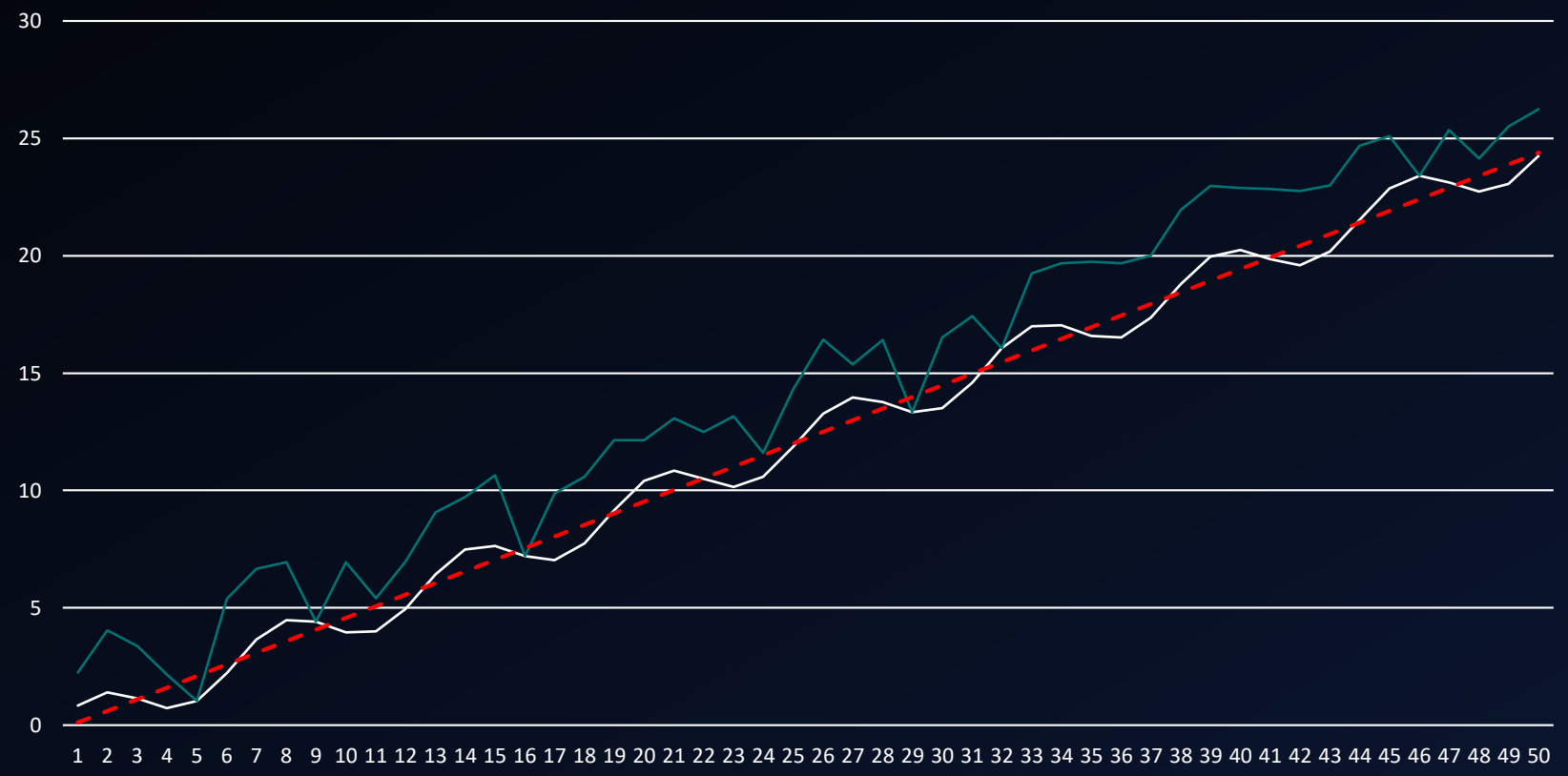
- Grafikus ábrázolás, „látható” jelenségek leolvasása (maximum és minimumhelyek, tendenciák, növekedés, csökkenés, monotonitás, szélsőértékek, tipikus értékek, jellegzetes tartományok)
- Bázis- és láncviszonyszámok
  - Bázis (minden értéket egy kitűntetett értékhez)
  - Lánc (minden időszak értékét az előző időszak értékéhez)
- Átlagok
  - Tartam idősor esetén egyszerű számtani
  - Állapotidősorok esetén kronologikus átlagot kell használni

Az idősor komponensei

Trend – hosszú távú tendencia, az idősor alakulásának elsődleges iránya

Periodikus ingadozások – Hullámzás (szabályos, ezáltal jól modellezhető)

Véletlen ingadozás – NEM MUTAT SEMMIFÉLE SZABÁLYSZERŰSÉGET



Dekompozíciós eljárás

- A determinisztikus modellből indulunk ki
- Felbontjuk a teljes hatást:
  - Trendhatásra
  - Szezonális hatásra
  - Véletlen hatásra
- Felmerül a kérdés: milyen kapcsolat áll fenn az egyes tényezők között (+;*)
  - Amennyiben additív jellegű a kapcsolat additív modellről beszélünk (+); Y a komponensek összegének „eredménye”
  - A multiplikatív modellben a kapcsolat természetesen multiplikatív jellegű (*); Y a komponensek szorzatának „eredménye”

Az additív és a multiplikatív modell

Additív modell

$$y_{ij} = \hat{y}_{ij} + s_j + \varepsilon_{ij}$$

Trendhatás (trendérték)

Szezonális hatás (szezonális eltérés)

Véletlen hatás

Multiplikatív modell

$$y_{ij} = \hat{y}_{ij} \cdot s_j^* \cdot \varepsilon_{ij}$$

Trendhatás (trendérték)

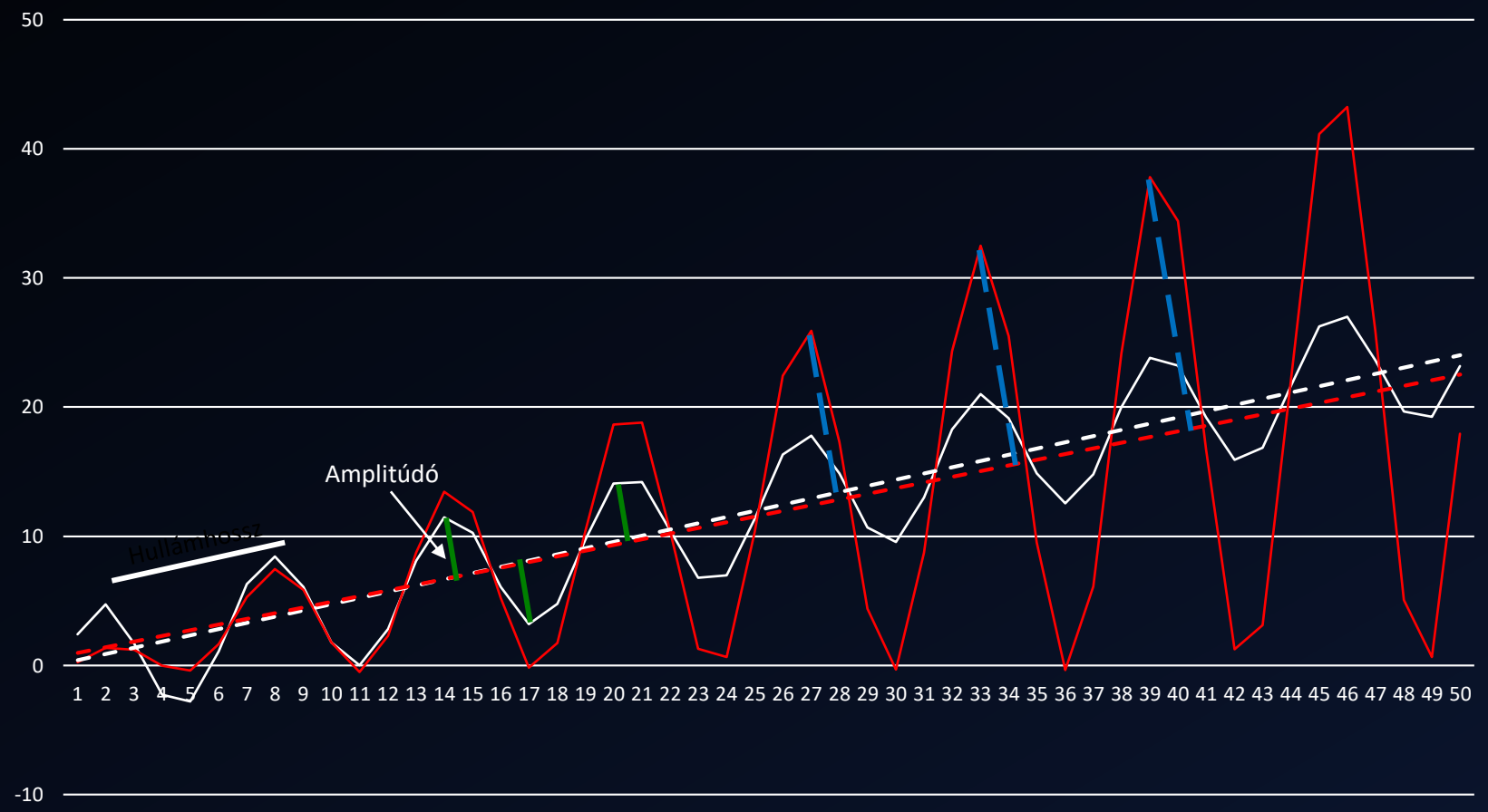
Szezonális hatás (szezonális eltérés)

Véletlen hatás

s_j a j-edik szezonhoz tartozó szezonális komponens (szezonindex)

y_{ij} – a megfigyelés értéke az idősorban, pontosabban az idősor i-edik periódusának j-edik szezonjához tartozó adat; vagy másképp i – a periódusok száma (pl. évek), j – a perióduson belüli időszakok száma (pl. hónapok, negyedévek stb.)

Az additív és a multiplikatív modell



Additív modell: amplitúdók egymással megegyeznek, szezonok eltérítő hatása abszolút állandóságot mutat
Multiplikatív modell: amplitúdók egymással nem egyeznek meg, szezonok eltérítő hatása relatív állandóságot mutat

Trendszámítás MOZGÓÁTLAGOLÁSSAL

- Szezonális és véletlen hatás kiszűrése
- Lényegében az idősor kisimítása
- Egyéb eljárások is ismertek, jelenleg a mozgóátlaggal dolgozunk, az átlagolás segítségével szűrjük ki a „nem kívánatos” hatásokat
- Lépései:
  - Az átlagolandó értékek számának meghatározása (k) → az idősor első előre rögzített ennyi számú eleméből képezünk egy átlagot (k megválasztásának elvei!*)
  - Kiszámítjuk az idősor első k elemének átlagát
  - Az első figyelembe vett adatot elhagyjuk, vesszük az idősor következő elemét
  - Ismétlés az idősor utolsó elemének felhasználásáig
  - Az idősor elejéhez és végéhez nem kapunk átlagot → a trendátlagok száma <, mint az idősor adatainak száma
- * van szezonalitás → k a perióduson belüli szakaszok számával vagy annak egész számú többszörösével
- * minél hosszabb az idősor annál nagyobb k-t érdemes választani, hogy a trend stabilabban mutatkozzon meg (ésszerűség!)

Trendszámítás MOZGÓÁTLAGOLÁSSAL

- „k” értékétől függően két eset:
  - k páratlan (2l+1):

$$\hat{y}_k = \frac{y_{k-l} + \ldots + y_k + \ldots + y_{k+l}}{2l+1}$$

  - k páros (2l):

1.

$$\hat{y}_{k,1} = \frac{y_{k-l} + \ldots + y_k + \ldots + y_{k+l-1}}{2l}$$

$$\hat{y}_{k,2} = \frac{y_{k-l+1} + \ldots + y_k + \ldots + y_{k+l}}{2l}$$

2.

$$\hat{y}_k = \frac{\hat{y}_{k,1} + \hat{y}_{k,2}}{2}$$

Ez a lépés a centírozás

Trendszámítás MOZGÓÁTLAGOLÁSSAL – Példa*

- **Az alábbi táblázat a háztartások számára értékesített gázmennyiség (milló m^3) értékét mutatja negyedéves bontásban Nógrád megyében 1990 és 1994 között.**

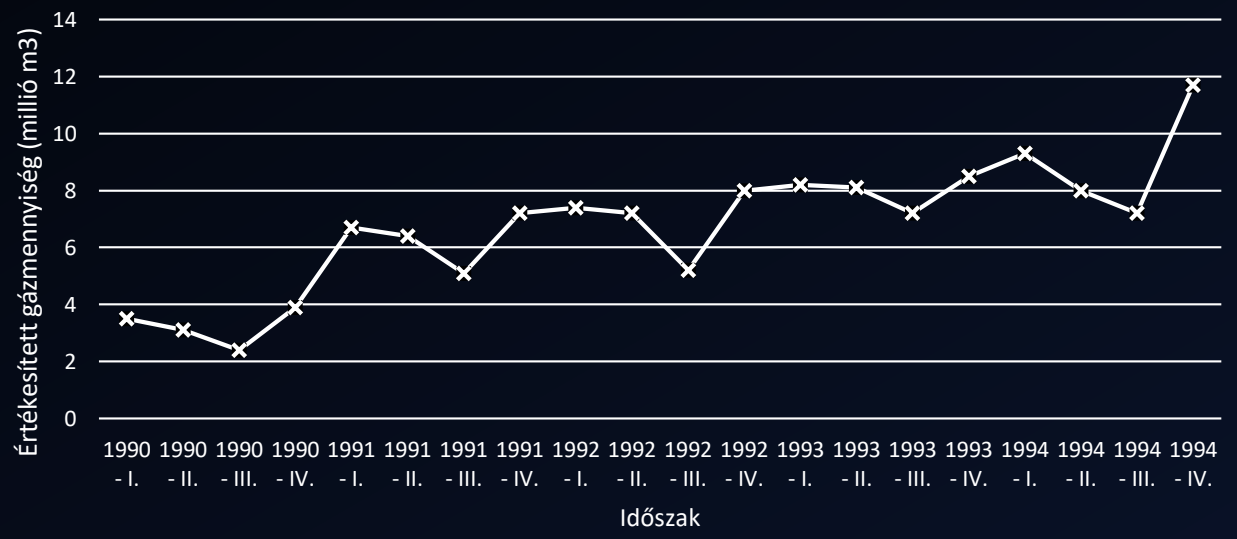
Évek	Negyedévek			
	I.	II.	III.	IV.
1990	3,5	3,1	2,4	3,9
1991	6,7	6,4	5,1	7,2
1992	7,4	7,2	5,2	8,0
1993	8,2	8,1	7,2	8,5
1994	9,3	8,0	7,2	11,7

Határozzuk meg a gázfogyasztás alakulását jellemző trendet mozgóátlagolás alkalmazásával!

* Forrás: Korpás A.-né: Általános statisztika II., Nemzeti Tankönyvkiadó, Budapest, 1997

Trendszámítás MOZGÓÁTLAGOLÁSSAL – Példa

- Célszerű az adatokat először ábrázolni a trendhatás jellegének megállapítása céljából.



- Elvi megfontolások k értékének meghatározására
  - Periódusnak az év tekinthető, amelyen belül negyedéves szezonalitás feltételezhető, így a szezonok száma egy periódusban 4, tehát k=4 célszerű választás

Trendszámítás MOZGÓÁTLAGOLÁSSAL – Példa

Időszak	Értékesített gázmennyiség (millió m3)	4 tagú mozgóátlag	Centrírozás
1990 - I.	3.5		
1990 - II.	3.1	3.23	
1990 - III.	2.4	4.03	3.63
1990 - IV.	3.9	4.85	4.44
1991 - I.	6.7	5.53	5.19
1991 - II.	6.4	6.35	5.94
1991 - III.	5.1	6.53	6.44
1991 - IV.	7.2	6.73	6.63
1992 - I.	7.4	6.75	6.74
1992 - II.	7.2	6.95	6.85
1992 - III.	5.2	7.15	7.05
1992 - IV.	8	7.38	7.26
1993 - I.	8.2	7.88	7.63
1993 - II.	8.1	8.00	7.94
1993 - III.	7.2	8.28	8.14
1993 - IV.	8.5	8.25	8.26
1994 - I.	9.3	8.25	8.25
1994 - II.	8	9.05	8.65
1994 - III.	7.2		
1994 - IV.	11.7		

Az első centírozásba vont értékeket a zöld téglalap jelöli

A centírozott értékek egy lépésben is kiszámíthatók az alábbi módon:

$$\hat{y}_k = \frac{\frac{y_{k-l}}{2} + \sum_{i=k-(l-1)}^{k+(l-1)} y_i + \frac{y_{k+l}}{2}}{2l}$$

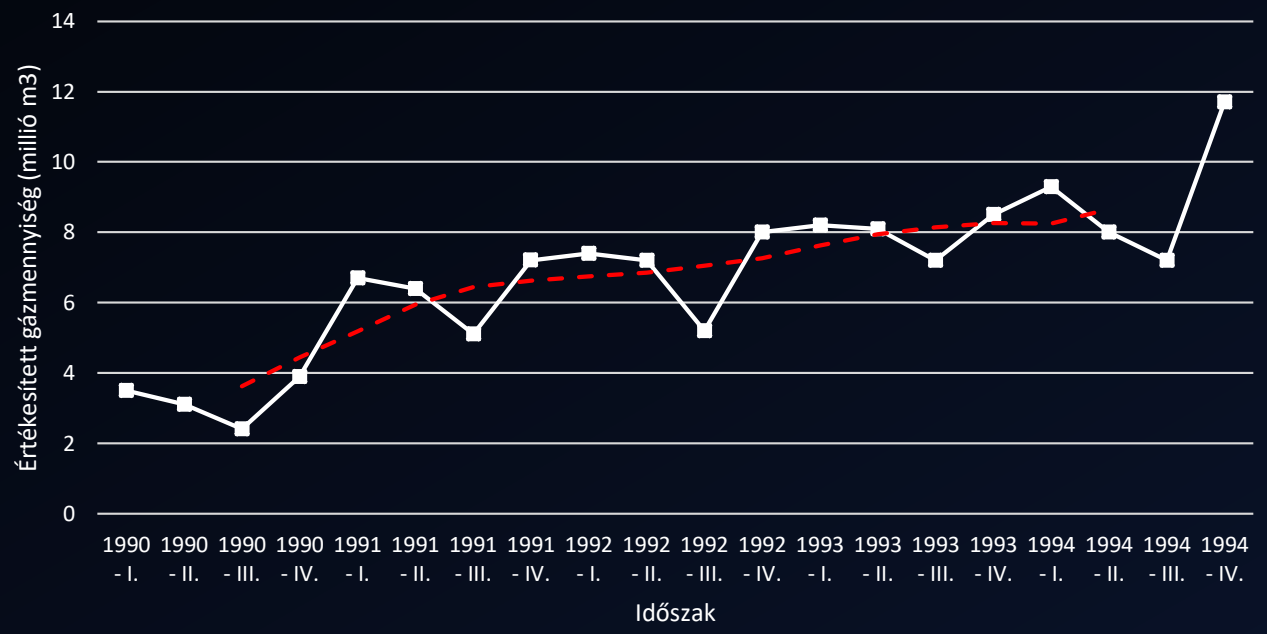
k=4, l=2

$$\hat{y}_3 = \frac{\frac{y_1}{2} + y_2 + y_3 + y_4 + \frac{y_5}{2}}{4} = \frac{\frac{3.5}{2} + 3.1 + 2.4 + 3.9 + \frac{6.7}{2}}{4} = 3.63$$

Számtani átlag (a piros téglalap az első „ablakot” jelöli)

Trendszámítás MOZGÓÁTLAGOLÁSSAL – Példa

- A trendet is érdemes az ábrán feltűntetni



- Az ábrán látható a trendvonal, amely volatilitását tekintve lényegesen kisebb kilengést mutat, mint az eredeti idősor értékei.

ANALITIKUS trendszámítás

- Az idősorban lévő alapirányzat matematikai függvénnyel
- Lépései:
  - A trendhez legjobban illeszkedő függvény meghatározása
  - A függvény paramétereinek becslése
- Jellemző függvénytípusok: lineáris, exponenciális, n-ed fokú polinom, logisztikus görbe

ANALITIKUS trendszámítás

- A megfelelő függvény kiválasztása:
  - idősor képének vizsgálata
  - szomszédos időszakok közötti változás jellegének értelmezése
  - Ha a növekedés/csökkenés üteme állandó, vagy nulla lineáris, illetve konstans függvény
  - Ha a változás relatív nagysága (hányados) állandó → exponenciális függvény
  - Ha a változás iránya, nagysága, arányossága változékony → másodfokú polinom

ANALITIKUS trendszámítás lineáris függvény esetén

- A lineáris trendfüggvény általános alakja:

$$\hat{y}_t = b_1 t + b_0$$

Együttható Konstans

(lényegében az egyenes meredeksége)

- b_1 és b_0 paraméterek meghatározása a cél az alábbi képletek segítségével:

$$b_1 = \frac{\sum(t-\bar{t}) * (y_t - \bar{y})}{\sum(t-\bar{t})^2} = \frac{\sum dtdy}{\sum dt^2}$$

$$b_0 = \bar{y} - b_1\bar{t}$$

Az idősor értékeinek számtani átlaga

Az időtényező átlaga

ANALITIKUS trendszámítás - Példa

Település	Év	Első helyezettek száma (fő)	t	$dt=(t-t_{átlag})$	$dy=(y-y_{átlag})$	dt^2	dtdy
Üllő	2002	8	1	-7.00	-0.80	49	5.6
Vecsés	2003	8	2	-6.00	-0.80	36	4.8
Vasvár	2004	7	3	-5.00	-1.80	25	9
Nyírmada	2005	8	4	-4.00	-0.80	16	3.2
Vásárosnamény	2006	6	5	-3.00	-2.80	9	8.4
Berhida	2007	11	6	-2.00	2.20	4	-4.4
Várpalota	2008	8	7	-1.00	-0.80	1	0.8
Őrbottyán	2009	11	8	0.00	2.20	0	0
Vác	2010	7	9	1.00	-1.80	1	-1.8
Fegyvernek	2011	12	10	2.00	3.20	4	6.4
Törökszentmiklós	2012	6	11	3.00	-2.80	9	-8.4
Tolna	2013	13	12	4.00	4.20	16	16.8
Tokaj	2014	6	13	5.00	-2.80	25	-14
Tiszalök	2015	14	14	6.00	5.20	36	31.2
Tiszaújváros	2016	7	15	7.00	-1.80	49	-12.6
Összesen		**132**	**120**	**0**	**0**	**280**	**45**
Átlag		**8.80**	**8.00**				

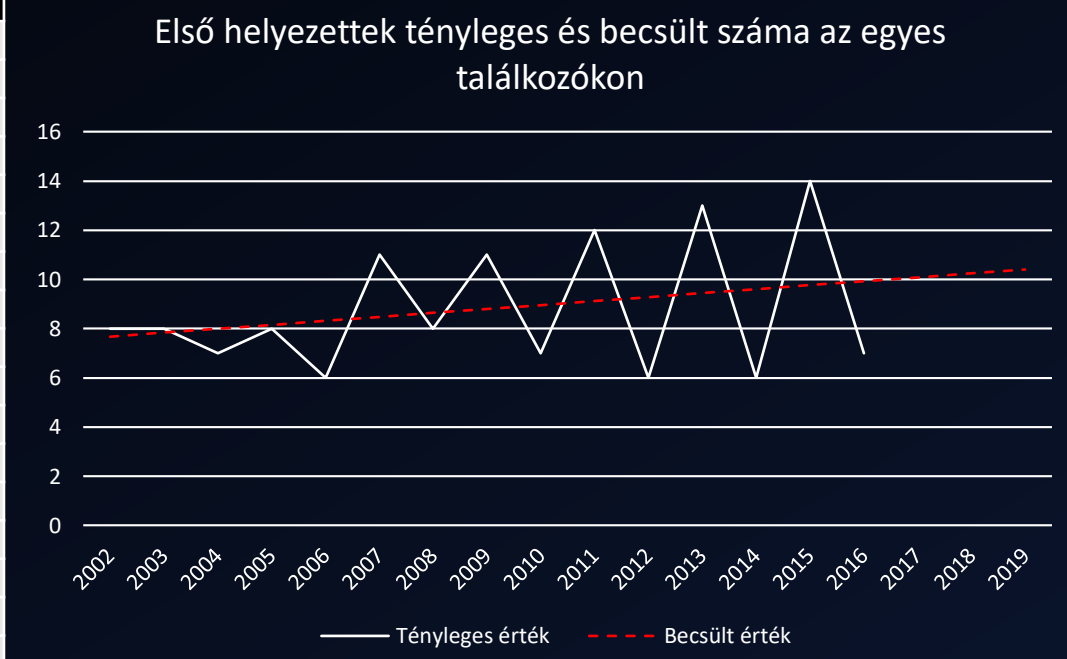
$$\bar{y} = \frac{132}{15} = 8.8 \quad \bar{t} = \frac{120}{15} = 8 \quad b_1 = \frac{45}{280} = 0.16 \quad b_0 = 8.8 - 0.16 * 8 = 7.514$$

$$\boxed{\hat{y}_t = 0.16t + 7.514}$$

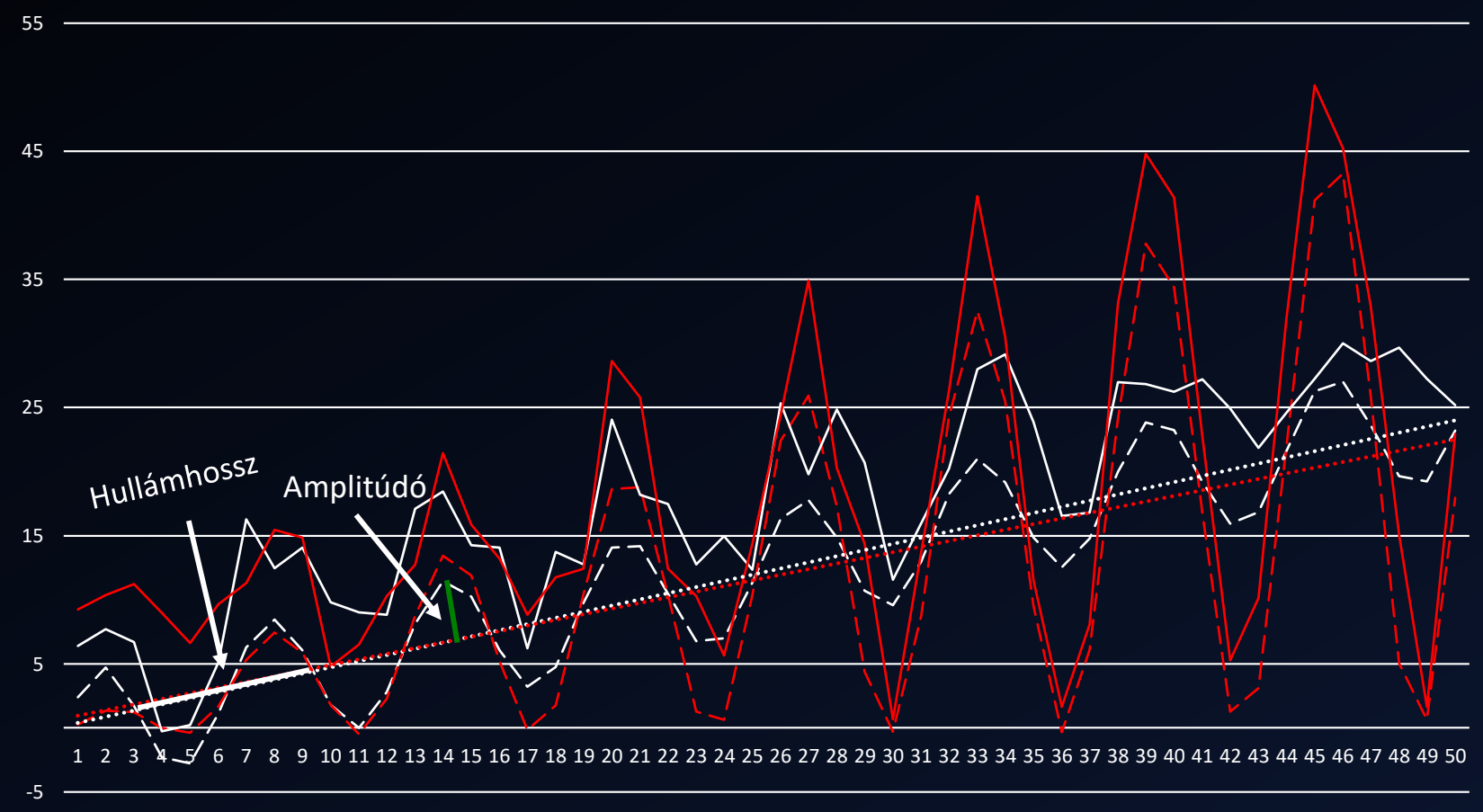
ANALITIKUS trendszámítás - Példa

$$\boxed{\hat{y}_t = 0.16t + 7.514}$$

Év	t	Tényleges érték	Becsült érték
2002	1	8	7.675
2003	2	8	7.835714
2004	3	7	7.996429
2005	4	8	8.157143
2006	5	6	8.317857
2007	6	11	8.478571
2008	7	8	8.639286
2009	8	11	8.8
2010	9	7	8.960714
2011	10	12	9.121429
2012	11	6	9.282143
2013	12	13	9.442857
2014	13	6	9.603571
2015	14	14	9.764286
2016	15	7	9.925
2017	16	-	10.08571
2018	17	-	10.24643
2019	18	-	10.40714



Az additív és a multiplikatív modell, hullámhossz, amplitúdó, véletlen hatás



Bevezető – periodikus ingadozás

- Meghatározott, ismétlődő
- Szezonális ingadozás (időtartama ≤ 1 év)
- Háttérhatások
  - Természeti jelenségek, pl. napszakok, évszakok változása
  - Társadalmi hatások, pl. ünnepek, népszokások, az ízlésvilág változása, idő- és munkarend
- A hatások jellegzetességei és kiváltó okai ismeretében, lehetőségünk van azok mérséklésére, esetlegesen az alkalmazkodásra

Bevezető – periodikus ingadozás

- A szezonalitás eltérítő hatása milyen mértékű (arányú) a trendhez képest?
- Cél: a jelenségről leválasztjuk a véletlen és a trendhatást, hogy csak a periodikus ingadozást vizsgálhassuk.
- A periodikus ingadozás
  - Additív modellben a trendtől való eltérés nagyságával
  - Multiplikatív modellben a trendtől vett relatív eltéréssel jellemehető

Trendhatás kiszűrése az idősor értékeiből

- Attól függ, hogy melyik modellt alkalmazzuk (2 eset)

Additív modell

$$y_{ij} = \hat{y}_{ij} + s_j + \varepsilon_{ij}$$

A trendértékek kivonása
(állandó amplitúdó feltételezése)

Szezonális eltérés

Multiplikatív modell

$$y_{ij} = \hat{y}_{ij} \cdot s_j^* \cdot \varepsilon_{ij}$$

Leosztunk a trendértékekkel
(az amplitúdó értéke változik)

Szezonindex

Véletlen hatás kiszűrése az idősor értékeiből

- A megfelelő szezonokra vonatkoztatva a trendhatástól megtisztított elemek átlagolásával (n=a szezonokban lévő elemek száma), azaz:
  - Szezononként „leválogatjuk” az egyedi szezonális eltéréseket
  - Vesszük az átlagukat

Additív modell

$$s_j = \frac{\sum_{i=1}^{n/p} \left(y_{ij} - \hat{y}_{ij} \right)}{n/p}$$

Nyers szezonális eltérés (j-edik)

Multiplikatív modell

$$s_j^* = \frac{\sum_{i=1}^{n/p} \left(y_{ij} / \hat{y}_{ij} \right)}{n/p}$$

Nyers szezonindex (j-edik)

Véletlen hatás kiszűrése az idősor értékeiből

- A nyers szezonális eltérések és indexek esetében nem feltétlenül teljesül, hogy a szerzonális eltérések összege, illetve átlaga nulla, vagy a multiplikatív modellnél szorzatuk 1 → korrekciót tesz szükségessé

Additív modell

$$s_j' = s_j - \frac{\sum_{j=1}^{p} s_j}{p}$$

Korrigált szezonális eltérés (j-edik)

Multiplikatív modell

$$s_j^{*'} = s_j^{*} \div \frac{\sum_{j=1}^{p} s_j^{*}}{p}$$

Korrigált szezonindex (j-edik)

A korrigált szezonális eltérés és szezonindex értelmezése

- Szezonális eltérések esetén a kapott mutató kifejezi, hogy az idősor egyes értékei az adott periódusban átlagosan mennyivel tér el a trend szerinti értéktől.
- Szezonindexek esetén a kapott mutató kifejezi, hogy az idősor egyes értékei az adott periódusban átlagosan hányszorosa a trend szerinti értéknek.

Példa – Szezonális eltérés (alapadatok és a trend)

Év	Negyedév			
	1	2	3	4
2012	61	71	119	78
2013	63	80	97	83
2014	72	78	109	85
2015	75	72	103	72

Év	Transzponált értékek	4 tagú átlag	Centírozott érték
2012 – I	87	-	-
2012 – II	85	91.5	-
2012 – III	110	88.25	89.875
2012 – IV	84	85	86.625
2013 – I	74	83	84
2013 – II	72	82.5	82.75
2013 – III	102	81.75	82.125
2013 - IV	82	80.5	81.125
2014 – I	71	79.75	80.125
2014 – II	67	78.75	79.25
2014 – III	99	78.5	78.625
2014 – IV	78	78	78.25
2015 – I	70	75.75	76.875
2015 – II	65	74	74.875
2015 – III	90	-	-
2015 – IV	71	-	-

Trendértékek (centírozott értékek)

Év	Negyedév			
	1	2	3	4
2012	-	-	89.875	86.625
2013	84	82.75	82.125	81.125
2014	80.125	79.25	78.625	78.25
2015	76.875	74.875	-	-

Példa – Szezonális eltérés

- Ezen értékek segítségével képezhetők a trendhatástól megtisztított értékek (tényleges érték - trendérték)

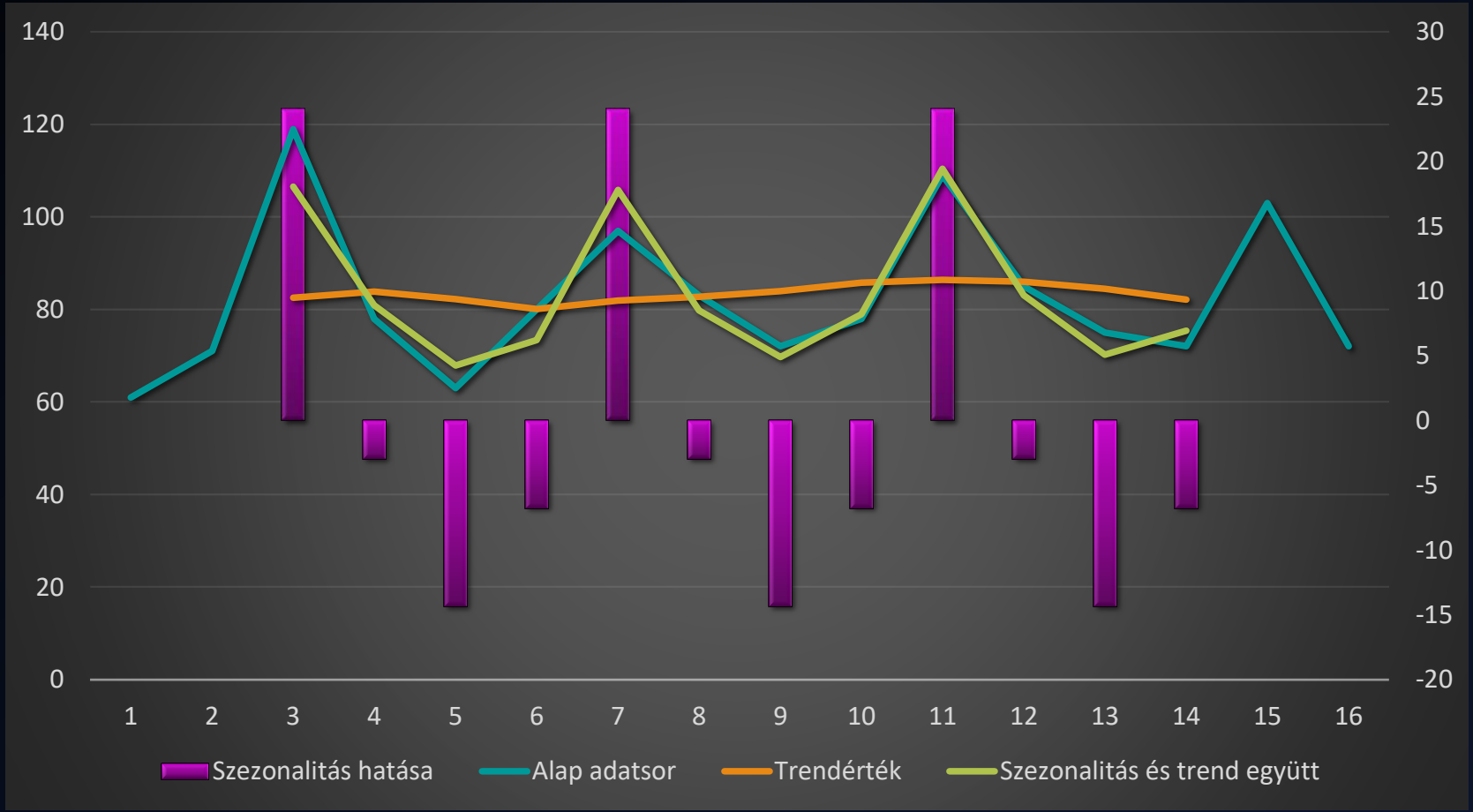
Év	Negyedév			
	1	2	3	4
2012	-	-	29.125	-8.625
2013	-21	-2.75	14.875	1.875
2014	-8.125	-1.25	30.375	6.75
2015	-1.875	-2.875	-	-
A nyers szezonális eltérések átlaga:	-10.333	-2.292	24.792	0.000
A nyers szezonális eltérések átlagának átlaga:		3.042		
Korrigált szezonális eltérések:	-13.375	-5.333	21.750	-3.042

μ

Δ

ÉRTELMEZÉS!

Példa – Szezonális eltérés



Példa – Szezonindex

- Ezen értékek segítségével képezhetők a trendhatástól megtisztított értékek (tényleges érték / trendérték)

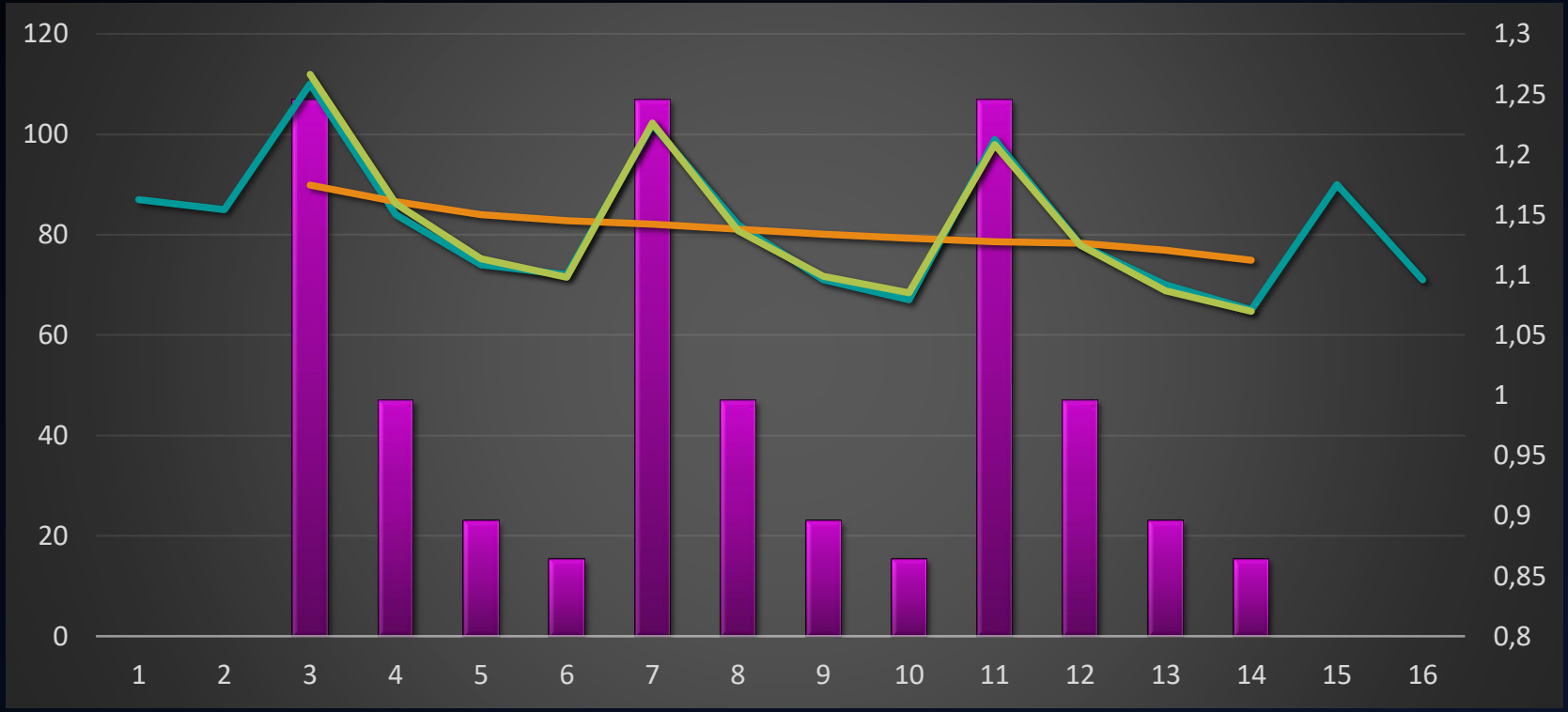
Év	Negyedév			
	1	2	3	4
2012	-	-	1.2239	0.9697
2013	0.8810	0.8701	1.2420	1.0108
2014	0.8861	0.8454	1.2591	0.9968
2015	0.9106	0.8681	-	-
A nyers szezonális eltérések átlaga:	0.8925	0.8612	1.2417	0.9924
A nyers szezonális eltérések átlagának átlaga:		0.997		
Korrigált szezonális eltérések:	0.8953	0.8638	1.2455	0.9954

μ

÷

ÉRTELMEZÉS!

Példa – Szezonindex



Extrapoláció

- Ismeretlen dolgokra történő következtetés a meglévő ismeretek alapján.
- Két fontos fogalom:
  - Interpoláció – belső értékek becslés (pl. hiányzó adatok pótlása)
  - Extrapoláció – külső értékek becslése
- Az idősor alapján felismert összefüggések „kivetítése” a jövőbeli időszakokra
- Becslés arra vonatkozóan, hogy hogyan folytatódnak az értékek?

Extrapoláció a gyarkolatban

- Eszközei
  - A trendegyenlet
  - Periodikus ingadozás
- Lépései:
  - A trendegyenletbe az adott időszakhoz tartozó érték behejettesítése
  - Peirodikus ingadozás kimutatása esetén korrigálunk
    - Additív modellben: trendérték + szezonális eltérés = extrapolációval kapott érték
    - Multiplikatív modellben: trendérték * szezonindex = extrapolációval kapott érték

Kitekintés, az idősorelemzés egyes további kérdései

- Konjunkturális hatásokkal (3-100 év) ciklusok:
  - $y_{ij} = \hat{y}_{ij} + s_j + c_l + v_{ij}$
  - $y_{ij} = \hat{y}_{ij} * s_j^* * c_l^* * v_{ij}^*$
- Egyéb trendtípusok: fél-logaritmikus, másodfokú polinomiális, harmadfokú polinomiális, hatvány alakú, exponenciális, elsőfokú hiperbólikus, telítődési (S-alakú trendek) stb.
- Trend „nélküli” jelenségek

Egy kis kitekintő az indexek világába: a Szenzitivitási-index

Szenzitivitási-index

Szenzitivitás mérési kerete

Preakció fázis
PRA

Akció alappont
AA

Akció fázis
A

Akció végpont
AV

Poszt-akció fázis
POA

Hatásmintázatok és a trendmintázati görbe

Szenzitivitási-index

- Súlytényezők
  - Negatív-súlytényező: $NS_j = NA_j * NM_j * NEA_j * NE_j$

    NA - Negativitási arány
    NM - Negativitási mérték
    NEA - Negatív elmozdulás aránya
    NE - Negatív eltérés
  - Tükör-súlytényező: $TS_j = TA_j * TM_j * TE_j * PE_j$

    TA - Tükör arány
    TM - Tükör mérték
    TE - Tükör elmozdulás
    PE - Pozitív eltérés
- Korrigált szenzitivitási volumenérték
  - $k\eta_j = \eta_j * (NS_j + TS_j)$
  - A domináns reakcióiránytól függően negatív vagy pozitív
- **Szenzitivitási-index**
  - $$\theta_\omega^n = \frac{k\eta_j}{MAX(k\eta_j)} * 100$$

Kapcsolatvizsgálati módszerek

Kapcsolatvizsgálati módszerek, magyarázó-modellek

- Függő és független változó
- A függő változó viselkedését magyarázzuk a független változó(k) segítségével
- A változók mérési szintje → alkalmazandó módszer

		Független változó	
		Nem metrikus	Metrikus
Függő változó	Nem metrikus	Kereszttábla-elemzés	Diszkriminancia-elemzés
	Metrikus	Variancia-elemzés	Korreláció és regresszióelemzés

A kapcsolatok típusai

- A sztochasztikus kapcsolatokat további típusokba soroljuk:
  - Asszociációs kapcsolat: minőségi ismérv – minőségi ismérv
  - **Vegyes kapcsolat: ok – minőségi, okozat – mennyiségi**
  - **Korrelációs kapcsolat: mennyiségi ismérv – mennyiségi ismérv**
- Vegyes kapcsolat → szórásnégyzet felbontása → szóráshányados (H)
- Korrelációs kapcsolat → lineáris korrelációs együttható (r)

Vegyes kapcsolat

Szóráshányados

- A teljes szórásnégyzet meghatározásából következik, hogy:

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

- Amely összefüggést a teljes szórásnégyzettel leosztva

$$1 = \frac{\sigma_B^2}{\sigma^2} + \frac{\sigma_K^2}{\sigma^2}$$

összefüggés adódik.

- A minőségi ismérv hatását a külső szórás értéke közvetíti
  - σ_K^2=0 → a minőségi ismérvnek nincs mérhető hatása a mennyiségi ismérvre
  - σ_B^2=0 → a kapcsolat determinisztikus (σ_K^2=σ^2)

Szóráshányados

- A külső és a teljes szórás aránya segítségével a vegyes kapcsolat erőssége meghatározható, ez a mutatószám a szóráshányados (H):

$$H = \frac{\sigma_K}{\sigma} = \sqrt{1 - \frac{\sigma_B^2}{\sigma^2}} + \sqrt{\frac{\sigma_K^2}{\sigma^2}}$$

- A magyarázóerő kifejezhető H2 értékével (? %-ban magyarázza a minőségi változó a mennyiségi ismérv alakulását)
- A szóráshányados meghatározásának lépései
  - Csoportátlagok meghatározása
  - Főátlag meghatározása
  - Belső szórásnégyzet és belső szórás kiszámítása
  - Külső szórásnégyzet és külső szórás kiszámítása
  - Teljes szórásnégyzet és teljes szórás kiszámítása
  - Szóráshányados meghatározása és értelmezése

Szóráshányados - Példa

Kísérlet sorszáma	Lőfegyver típusa		
	XT12	BZ001	ARC13
1	9	6	6
2	10	6	7
3	14	9	7
4	13	8	13
5	10	7	12
6	10	8	-
7	15	6	-
8	11	9	-
9	12	-	-
Átlag (SZÁMTANI ÁTLAG) $\bar{x}_j$	11.56	7.38	9.00
Szórás σ_j	1.950	1.218	2.898
n_j	9	8	5
Csoportonkénti szórás négyzete	3.8025	1.4844	8.4000
Csoportonkénti átlag – főátlag ($\bar{x}$)	2.1010	-2.0795	-0.4545
(Csoportonkénti átlag-főátlag)^2	4.4142	4.3245	0.2066

Értékek meghatározása

$$\sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

FŐÁTLAG SZÁMÍTÁSA A KÖVETKEZŐ DIÁN

Belső külső és teljes szórásnégyzet meghatározása

- **Főátlag (kétféleképpen kaphatjuk meg)**
  - Az összes mért érték számtani átlagaként (μ=9.455)
  - A csoportonkénti átlagokból súlyozott átlag formulával

- **Belső szórásnégyzet és szórás**

$$\sigma_B^2 = \frac{\sum_{j=1}^{m} n_j \sigma_j^2}{n} = \frac{9 * 3.8 + 8 * 1.48 + 5 * 8.4}{22} = 4$$

$$\sigma_B = 2$$

n_j – a megfigyelések száma a j-edik csoportban
n – összes elemszám (a példában $n_1+n_2+n_3=22$)
m – 1,2,…m a csoportok száma
σ_j^2 – a j-edik csoport szórásnégyzete

- **Külső szórásnégyzet és szórás**

$$\sigma_K^2 = \frac{\sum_{j=1}^{m} n_j (\bar{x}_j - \bar{x})^2}{n} = \frac{9 * 4.41 + 8 * 4.32 + 5 * 0.2}{22} = 3.425$$

$$\sigma_K = 1.85$$

n_j – a megfigyelések száma a j-edik csoportban
n – összes elemszám
m – 1,2,…m a csoportok száma
$\bar{x}_j$ – a j-edik csoport átlaga
$\bar{x}$ – a főátlag

- **Teljes szórásnégyzet (variancia) és szórás**

$$\sigma^2 = \sigma_B^2 + \sigma_K^2 = 4 + 3.43 = 7.43$$

$$\sigma = 2.726$$

Szóráshányados értelmezése

- $H = \frac{\sigma_K}{\sigma} = \frac{1.851}{2.726} = 0.679$
- $H^2 = 0.461$

A szóráshányados (H) értéke alapján megállapítható, hogy a lőfegyver típusa erős hatást gyakorol az elért pontszámokra. A sztochasztikus kapcsolat erőssége közepes, inkább erős. A szórásnégyzet értéke alapján levonható a következtetés, miszerint a lőfegyver típusa 46.1%-ban magyarázza az eredmények szóródását.

Korrelációs kapcsolat

Korrelációs kapcsolat

- „Erőssége” alapján ez esetben is elkülöníthetünk
  - Determinisztikus
  - Sztochasztikus
  - Független kapcsolatokat
- Iránya alapján a változók kapcsolata lehet
  - Pozitív
  - Negatív
- Előzetes értékelés ábra alapján → pontdiagram → kvalitatív következtetéseket fogalmazhatunk meg
- LINEÁRIS KORRELÁCIÓS EGYÜTTHATÓ: alapfeltételezése, hogy a változók közötti kapcsolat lineáris

Lineáris korrelációs együttható meghatározása

- $r = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{\sqrt{\sum_{i=1}^{n}(x_i - M_x)^2 * \sum_{i=1}^{n}(y_i - M_y)^2}}$, vagy $r = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{n * \sigma_x * \sigma_y}$ahol

x_i, y_i - a változók értékpárjai

M_x, M_y – a változók átlagai

- r tulajdonságai
  - $-1 \leq r \leq 1$
  - Minél inkább közelít a „r” abszolút értéke az 1-hez a lineáris kapcsolat annál erősebbnek feltételezhető
- r értékének ismeretében felírhatjuk a kapcsolat jellegét leíró lineáris függvényt: y=ax+b, ahol
  - $a = r * \frac{\sigma_y}{\sigma_x}$ és $b = M_y - a * M_x$, ahol

σ_x és σ_y az adatsorok szórásai

r - Példa

Sorszám	Edzésidő	Teljesítmény	2 $(x_i - \bar{x})$	3 $(y_i - \bar{y})$	4 $(x_i - \bar{x})^2$	5 $(y_i - \bar{y})^2$	6 $(x_i - \bar{x}) * (y_i - \bar{y})$
1	3	16	-19.2	-206.6	368.64	42683.56	3966.72
2	5	22	-17.2	-200.6	295.84	40240.36	3450.32
3	9	62	-13.2	-160.6	174.24	25792.36	2119.92
4	13	154	-9.2	-68.6	84.64	4705.96	631.12
5	15	124	-7.2	-98.6	51.84	9721.96	709.92
6	16	184	-6.2	-38.6	38.44	1489.96	239.32
7	19	204	-3.2	-18.6	10.24	345.96	59.52
8	21	192	-1.2	-30.6	1.44	936.36	36.72
9	23	245	0.8	22.4	0.64	501.76	17.92
10	25	248	2.8	25.4	7.84	645.16	71.12
11	29	313	6.8	90.4	46.24	8172.16	614.72
12	30	296	7.8	73.4	60.84	5387.56	572.52
13	35	378	12.8	155.4	163.84	24149.16	1989.12
14	40	378	17.8	155.4	316.84	24149.16	2766.12
15	50	523	27.8	300.4	772.84	90240.16	8351.12
Összesen	333	3339	0	0	2394.4	279161.6	25596.2
Átlag	22.2	222.6					
Szórás	12.63	136.42					
n	15						

1

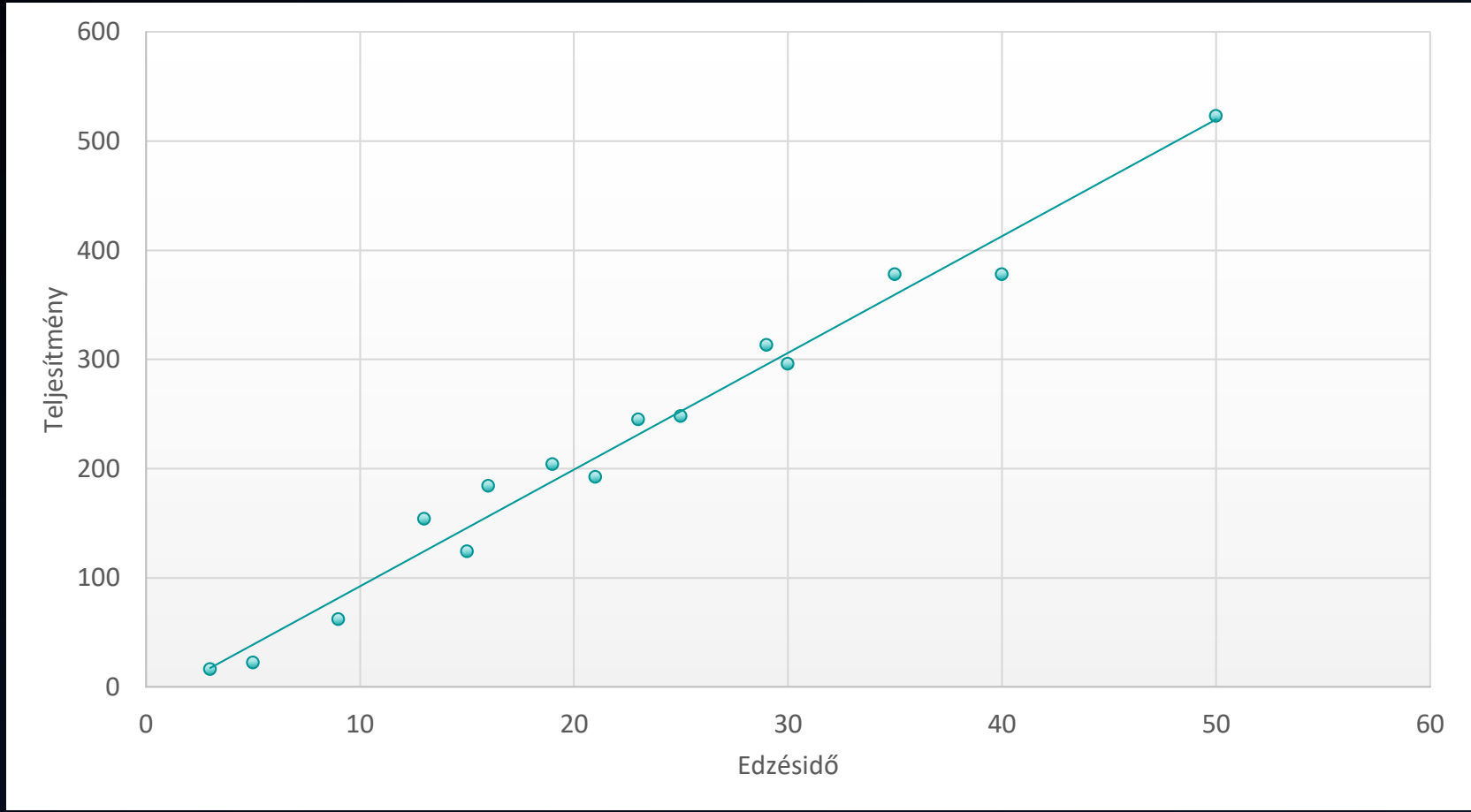
$$\sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

7

$$r = \frac{25596.2}{\sqrt{2394.4 * 279161.6}} \qquad r = \frac{25596.2}{15 * 12.63 * 136.42}$$

ÉRTELMEZÉS! (erősség, irány, jelleg) r=0.99

A mért adatok és r közötti összefüggés



Regresszióanalízis

- Matematikai formula megadása az összefüggés természetére, modellezés
- FONTOS: a változók között jelen esetben lineáris jellegű sztochasztikus kapcsolatot feltételezünk!
- Regresszió függvény meghatározása (legkisebb négyzetek módszere, OLS)

y=ax+b, ahol

$$a = r * \frac{\sigma_y}{\sigma_x}, \text{ vagy } a = \frac{\sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)}{\sum_{i=1}^{n}(x_i - M_x)^2}$$

$$\text{és } b = M_y - a * M_x, \text{ ahol}$$

σ_x és σ_y az adatsorok szórásai

a – regressziós együttható

Lineáris regresszió és extrapoláció

A valódi értékek szórása a becsült érté

A teljes szórás 98%-át magyarázza a regressziós

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,990	,980	,979	1302440,539

The independent variable is Év.

<0,05 esetén létezik kapcsolat

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1,696E+15	1	1,696E+15	999,524	,000
Residual	3,393E+13	20	1,696E+12		
Total	1,729E+15	21			

The independent variable is Év.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Év	1383758,055	43768,697	,990	31,615	,000
(Constant)	-2753609214	87778562,03		-31,370	,000

A bruttó hazai termék (GDP) értéke forintban, euróban, dollárban, vásárlóerő-paritáson (1995–)

Regressziós egyenes képlete:

GDP=-2753609214,071+1383758,055*Év

Elaszticitás

- Rugalmasság
- A változás relatív (%-os) mértéke
- Meghatározása:

$$El = a\frac{\bar{x}}{\bar{y}}$$

- Értelmezése:
  - Amennyiben El>1, rugalmas, amennyiben El<1 rugalmatlan kapcsolatról beszélünk a két változó között

Adatredukciós módszerek

Adatredukciós módszerek

- Összetett jelenségek vizsgálata olykor több száz mutatóval → komplex mutatók, adatredukciós módszerek
- Utóbbi célja:
  - Összevonás, aggregálás, az attribútumok számának csökkentése
  - Eredeti változók összevonása, számuk csökkentése
  - A LEHETŐ LEGNAGYOBB VOLUMENŰ INFORMÁCIÓTARTALOM MEGTARTÁSÁVAL
  - Értelmezhetőség kérdése
  - TÖBBEK KÖZÖTT:
    - Faktoranalízis, főkomponens-elemzés
    - Klaszteranalízis

Faktoranalízis

- Célja a kiinduló változók számának csökkentése → faktorváltozókba történő összevonása (túl sok kérdés a kérdőívből)
  - Könnyebb értelmezhetőség
  - Változók közötti kapcsolatrendszer feltárása
- Korrelációs mátrixból
- Az eredeti változók információtartalmából (variancia) minél többet megőrizni
- Struktúra feltáró módszer (függő és független változó előre nem adott) → mért változók mögött „rejtett” összefüggések
- Index-képzési eljárásként történő alkalmazása

Faktoranalízis

- Továbbá a létrejövő faktorok nem korrelálnak egymással (multikollinearitás kiszűrése)
- Feltételek (!)
  - Többek között minimum 50-100 fős minta
  - Ökölszabály: minimális mintaelemszám=változók száma*5 (10)
- DE!
  - Kapott faktorok tényleg léteznek? Relevánsak?
  - Értelmezhető?

Főkomponens-elemzés

- Tömörítési módszer a faktoranalízisen belül (!)
- A változók számát minimális információveszteség mellett csökkenti
- Az első faktor magyarázza a legnagyobb részt → ezt szokták pl. térképre vinni

Faktoranalízis (1)

Név	Környezet	Bio	Rosszétel	Sport	FAC1_1	FAC2_1
Pataki Tibor	1,00	2,00	5,00	1,00	-1,47343	1,07970
Takács Kri...	2,00	3,00	3,00	1,00	-,42716	-1,44806
Rácz József	3,00	4,00	4,00	5,00	,53544	-,32853
Fábián Fer...	4,00	5,00	5,00	5,00	,89297	,62451
Novák Zoltán	2,00	3,00	3,00	1,00	-,42716	-1,44806
Pintér László	3,00	4,00	4,00	5,00	,53544	-,32853
Kapolcs M...	5,00	5,00	5,00	5,00	1,14245	,56222
Balla Imre	4,00	5,00	5,00	5,00	,89297	,62451
Fazekas D...	4,00	5,00	5,00	5,00	,89297	,62451
Surány Kor...	2,00	3,00	3,00	1,00	-,42716	-1,44806
Takács Bal...	2,00	3,00	3,00	1,00	-,42716	-1,44806
Juhász Gá...	1,00	2,00	5,00	1,00	-1,47343	1,07970
Budai Dávid	2,00	3,00	3,00	1,00	-,42716	-1,44806
Balogh Olivér	5,00	5,00	5,00	5,00	1,14245	,56222
Szőke Ben...	1,00	2,00	5,00	1,00	-1,47343	1,07970
Apród Ger...	3,00	4,00	4,00	5,00	,53544	-,32853
Tamás Mar...	5,00	5,00	5,00	5,00	1,14245	,56222
Balázs Béla	3,00	4,00	4,00	5,00	,53544	-,32853
Hegedüs D...	2,00	3,00	3,00	1,00	-,42716	-1,44806
Dobos Péter	2,00	3,00	3,00	1,00	-,42716	-1,44806
Kozma Do...	1,00	2,00	5,00	1,00	-1,47343	1,07970
Miksa Kat...	5,00	5,00	5,00	5,00	1,14245	,56222
Gál Lilla	1,00	2,00	5,00	1,00	-1,47343	1,07970
Veres Erika	2,00	3,00	3,00	1,00	-,42716	-1,44806
Csatár Liliá...	5,00	5,00	5,00	5,00	1,14245	,56222
Vászoly Iz...	2,00	3,00	3,00	1,00	-,42716	-1,44806

Faktoranalízis (2) – adatok alkalmassága

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,558
Bartlett's Test of Sphericity	Approx. Chi-Square	202,448
	df	6
	Sig.	,000

<0,5 → nem végezhető el a faktoranalízis

<0,05 → a változók alkalmasak lehetnek

Faktorok számának meghatározása:

- A priori kritérium (a kutató dönt)
- Kaiser-kritérium (csak azokat a faktorokat vegyük figyelembe, amelyek sajátértéke legalább 1)
- Varianciahányad-módszer (egy minimális összesített varianciahányad szint elérése a cél)
- Egyéb megoldások is ismertek

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,039	75,981	75,981	3,039	75,981	75,981	2,769	69,229	69,229
2	,832	20,803	96,784	,832	20,803	96,784	1,102	27,555	96,784
3	,114	2,862	99,646						
4	,014	,354	100,000						

Extraction Method: Principal Component Analysis.

A faktorok kapcsolata a megőrzött információ-tartalommal

Faktoranalízis (3) – faktorsúlymátrix

Component Matrix[a]

	Component	
	1	2
Környezet	,963	-,181
Sport	,962	-,023
Bio	,962	-,252
Rosszétel	,512	,858

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

A változók hatása és szerepe a faktorokban

rotáció kérdése

					F1	F2
Apród Ger...	3,00	4,00	4,00	5,00	,53544	-,32853

Klaszteranalízis

- CÉLJA: a megfigyelési egységek viszonylag homogén csoportokba rendezése
- DE!: nincs egyetlen legjobb megoldás, akkor is talál klasztereket ha azok valójában nem léteznek, az eredmény az inputváltozók függvénye
- MEGFELELŐ: klaszterek elemei egymáshoz közel, de a klaszterek egymástól távol esnek
- Tetszőleges számú változó
- FELTÉTELEK:
  - Érzékeny a kiugró adatokra
  - Eltérő skálák problémát okoznak → standardizálás
  - Korreláció optimális esetben kicsi a változók között → faktorok alkalmazásának lehetősége

Klaszteranalízis

- Hierarchikus
  - Összevonó vs. Felosztó
  - Számos eljárás közül választhatunk, a példában Ward-féle eljárás
- K-Közép (nem hierarchikus módszer)
  - n>2000 fölött
  - Kisebb számítási kapacitás igény
  - DE!
    - Klaszterek számát előre meg kell határozni (érdemes hierarchikussal megalapozni)
    - Ismertnek tételezzük fel a klaszterközepeket

Hierarchikus klaszteranalízis (1)

Név	Sport	Környezet	Barátok	Lakás	Demográfia	Jövedelem	CLU4_1
Vass Benjá	5,00	5,00	5,00	5,00	5,00	350000,00	1
Fodor Erik	3,00	1,00	5,00	3,00	4,00	300000,00	2
Biró Valen	1,00	5,00	2,00	4,00	2,00	150000,00	3
Illés Doná	5,00	4,00	1,00	3,00	3,00	200000,00	1
Szücs Báli	1,00	1,00	1,00	1,00	2,00	150000,00	4
Katona Mih	5,00	4,00	4,00	3,00	3,00	175000,00	1
Pap Antal	4,00	1,00	3,00	4,00	3,00	200000,00	2
Fekete Cso	3,00	2,00	5,00	3,00	4,00	350000,00	2
Gulyás Géz	2,00	5,00	4,00	3,00	3,00	400000,00	3
Kis Norber	1,00	5,00	4,00	3,00	2,00	250000,00	3

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
10	100,0	0	,0	10	100,0

a. Squared Euclidean Distance used

b. Ward Linkage

Hierarchikus klaszteranalízis (2)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	8	,500	0	0	5
2	9	10	1,500	0	0	3
3	3	9	5,167	0	2	8
4	4	6	9,667	0	0	6
5	2	7	14,500	1	0	7
6	1	4	24,667	0	4	8
7	2	5	43,333	5	0	9
8	1	3	67,000	6	3	9
9	1	2	97,000	8	7	0

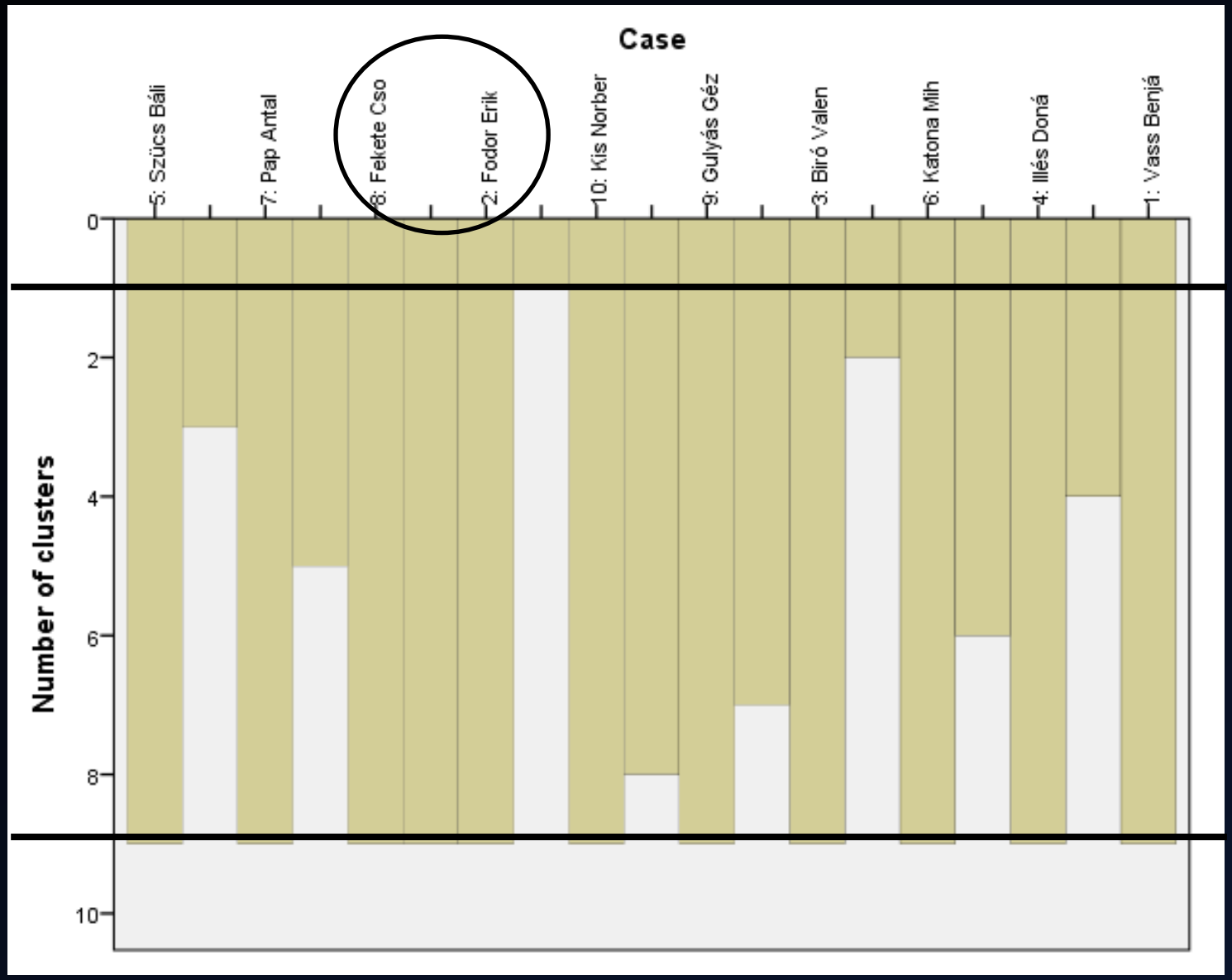
Összevonás lépései

Mit - mivel

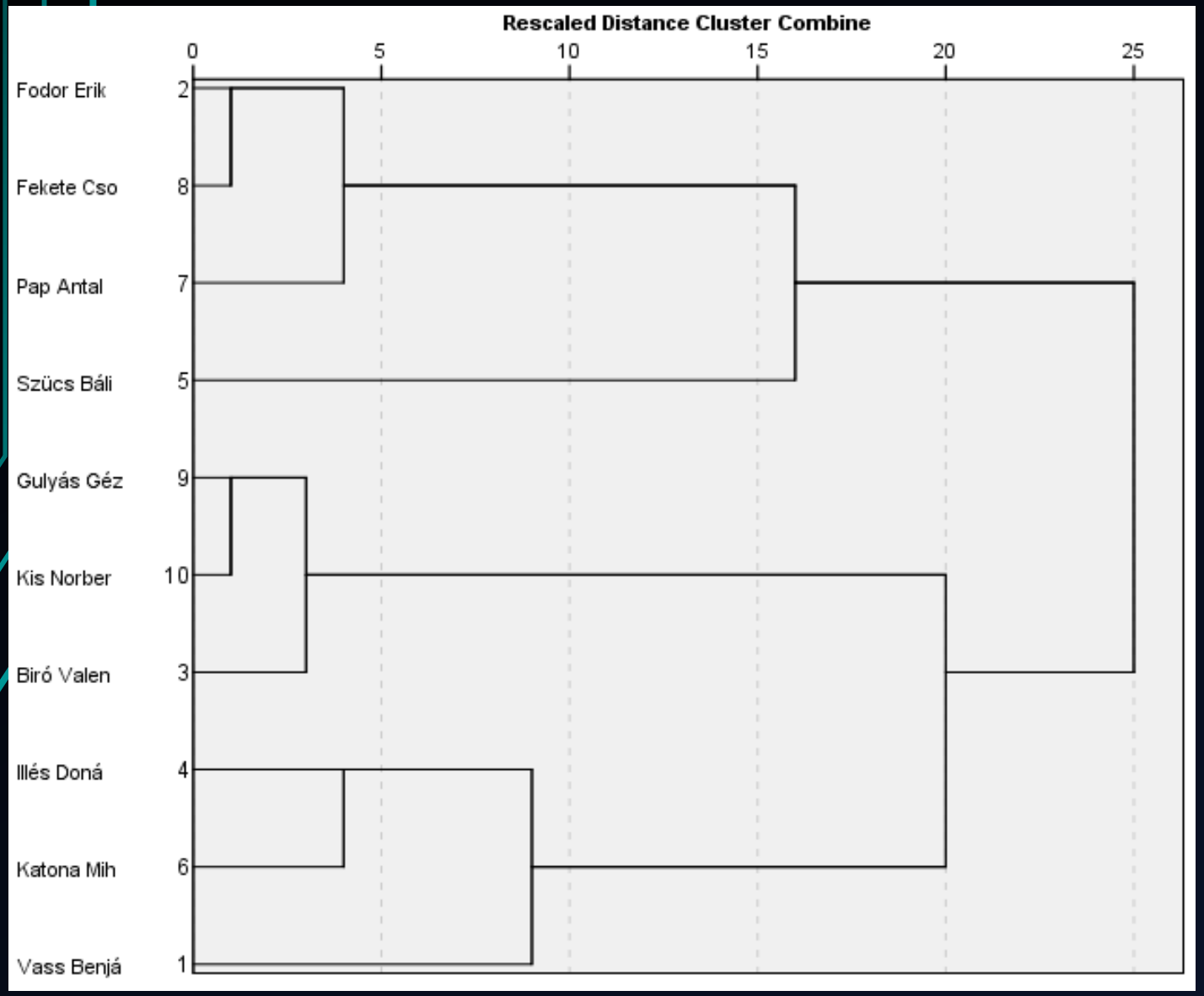
Távolság („ugrás” nagysága a

A két összevonásra kerülő klaszter melyik lépésben jelenik meg először, illetve a közös klaszter hol

Hierarchikus klaszteranalízis (3), jégcsapdiagram



Hierarchikus klaszteranalízis (4), dendogram



Döntés a klaszterek számáról:

- Kutatói döntés, tapasztalat
- Dendogram, vagy jégcsapdiagram alap
- Könyökkritérium a koefficiensek ábráz
- Klaszterek relatív mérete alapján
- Mutatószámok figyelembevételével:
  - Calinski-Harabasz
  - Duda-Hart
  - Stb.

Hierarchikus klaszteranalízis (5), jellemzés

ANOVA (metrikus – nominális (klaszterkód))

Kereszttábla (nem metrikus–nominális (klaszterkód))

Report

Ward Method		Sport	Környezet	Barátok	Lakás	Demográfia
1	Mean	5,0000	4,3333	3,3333	3,6667	3,6667
	N	3	3	3	3	3
	Std. Deviation	,00000	,57735	2,08167	1,15470	1,15470
2	Mean	3,3333	1,3333	4,3333	3,3333	3,6667
	N	3	3	3	3	3
	Std. Deviation	,57735	,57735	1,15470	,57735	,57735
3	Mean	1,3333	5,0000	3,3333	3,3333	2,3333
	N	3	3	3	3	3
	Std. Deviation	,57735	,00000	1,15470	,57735	,57735
4	Mean	1,0000	1,0000	1,0000	1,0000	2,0000
	N	1	1	1	1	1
	Std. Deviation	.	.	.	.	.
Total	Mean	3,0000	3,3000	3,4000	3,2000	3,1000
	N	10	10	10	10	10
	Std. Deviation	1,69967	1,82878	1,57762	1,03280	,99443

Klasztercentroidok

Szórás (mennyire homogén a csoport)

Jövedelem

Ward Method	Mean	N	Std. Deviation
1	480000,0000	3	26457,51311
2	111666,6667	3	37527,76750
3	233333,3333	3	15275,25232
4	1200000,000	1	.
Total	367500,0000	10	331002,2659

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Jövedelem * Ward Method	Between Groups	(Combined)	9,814E+11	3	3,271E+11	419,094	,000
	Within Groups		4683333333	6	780555555,6		
	Total		9,861E+11	9			

<0.05, nullhipotézist elvetjük, kategóriaátlagok

Ajánlott irodalom

- Dusek Tamás – Kotosz Balázs (2015): Területi statisztika. Akadémia Kiadó 285 p.
- Jeney László (2014): Területi fejlettségi különbségek mérése (ppt) – jeney.web.elte.hu/rkem1405.ppt
- Nemes Nagy József (2005): Regionális elemzési módszerek
- Pintér József – Ács Pongrác (2007): Bevezetés a sportstatisztikába. Dialog Campus Kiadó. 161 p.
- Sajtos László – Mitev Ariel (2007): SPSS kutatási és adatelemzési kézikönyv. Alinea Kiadó 402 p.
- Székelyi Mária – Barna Ildikó (2002): Túlélőkészlet az SPSS-hez. Typotex Elektronikus Kiadó Kft. 453 p.