Szegedi Tudományegyetem
Gazdaságtudományi Kar

Petres Tibor – Tóth László

STATISZTIKA

I. kötet

2001

Szerzők:

Dr. Petres Tibor, PhD
egyetemi docens
Statisztikai és Demográfiai Tanszék

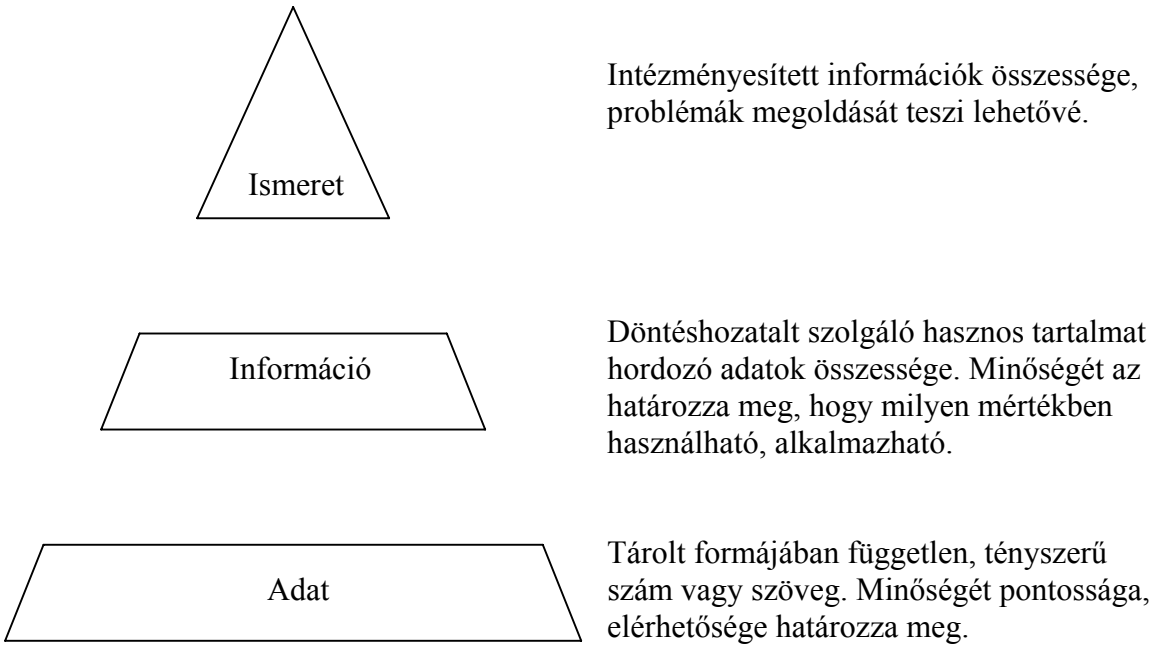
Tóth László
PhD-hallgató
Gazdaságtudományi Kar

Első kötet

Előszó

Könyvünk elsődlegesen közgazdászoknak készült, és általános statisztikával foglalkozik. Ennek részletes taglalása előtt, a kor szellemének megfelelően, néhány (kvantitatív elemzésekkel kapcsolatos) általános összefüggésre hívjuk fel az Olvasó figyelmét.

Az üzleti világ modern, globalizálódó korszakában nagy mértékben növekszik a piacgazdaság szereplőinek információigénye. Az adatok mennyiségének robbanásszerű növekedése nem jár együtt a megfelelő mértékű információ-növekedéssel. A két fogalom közötti jelentős különbséget az alábbi ábra szemlélteti. Igazából az üzleti világ döntéshozóinak nem az adatok hiányával, hanem azok bőségével kell szembenézniük, ugyanis még a legóvatosabb becslések szerint is az elektronikusan tárolt adatok volumene évente legalább megkétszereződik. A rendelkezésre álló adatok nagy mennyisége növeli ezek elemzésének összetettségét és az adatelemzőkkel szemben támasztott elvárásokat. Mivel az adatok információvá alakítása kisebb sebességgel történik, mint azok rendelkezésre bocsátása, a felhasználóknak egyre inkább adatelemzési szakértővé kell válniuk, ismerniük kell azokat a módszereket, amelyekkel az adatok értékelhetőek és hasznosíthatóak.



Egy adathalmazban a rejtett információk feltárásával az adatbányászat foglalkozik. Az

adatbányászati eljárásokat mára az üzleti világ is átvette a tudományos elemzésektől. Számos ilyen eljárás ismert a hagyományos statisztikai elemzésektől a mesterséges intelligencia által használt módszerekig. Ebben a könyvben a szerzők azokkal az alapvető statisztikai módszerekkel kívánják megismertetni az Olvasót, amelyek gazdasági elemzéseknél alkalmazhatóak.

Mind az adatok rendelkezésre bocsátásáról, mind azok elemzéséről szólva nem lehet figyelmen kívül hagyni az informatikai oldalt. Könyvünkben a széles körben hozzáférhető Microsoft Excel szoftver statisztikai funkciói kerülnek felhasználásra. Ez nem egy statisztikai programcsomag, de a bemutatott statisztikai módszerek elvégzésére alkalmas.

Fontossága miatt felhívjuk a figyelmet arra, hogy a kvantitatív elemzéseknél általában az adatok információvá, illetve ismeretté alakítása a cél. Ezért nem elég pusztán a matematikai műveleteket elvégezni, hanem a kapott eredményeket kell megfelelően értelmezni.

Tartalomjegyzék

1. Általában a statisztikáról

1.1. A statisztika fogalma

A **statisztika** kifejezés háromféleképpen is értelmezhető, mint

- gyakorlati számbavételi tevékenység,
- így nyert adatok összessége,
- tömegjelenségek vizsgálatára szolgáló módszerek rendszere.

Mi az utóbbival foglalkozunk részletesen, azt fogjuk megvizsgálni, hogy meghatározott cél érdekében gyűjtött adatokat hogyan lehet feldolgozni, elemezni.

Mivel vizsgálatunk tárgya a gazdasági, társadalmi és természeti jelenségek mennyiségi oldala, nem szakítva el a minőségi oldaltól, így gyakran támaszkodunk alapvető matematikai ismeretekre (például: mértani átlag, normális eloszlás, stb.). Érdemes ezért elhatárolni a matematikát a statisztikától. A matematika tárgya tapasztalattól mentes számabsztrakció, míg a statisztika szintén gyakran dolgozik számokkal, de azok gazdasági, társadalmi vagy természeti aktualitásukban jelennek meg. A statisztika inkább tapasztalati, a posteriori tudomány, míg a matematika tapasztalattól mentes, a priori tételeket alkot. Létezik a matematikai statisztika, mint külön tudomány, mely a valószínűségszámítással együtt fejlődött ki, és a statisztika azon részével foglalkozik, amely matematikai tételekkel alátámasztható.

A mi megközelítésünkben tehát a statisztika a tömegjelenségek jellemzőinek tömör, számszerű megismertetését szolgáló módszertan.

1.2. Alapfogalmak

Sokaság

A statisztikában a vizsgálatunk tárgyát képező egységek, egyedek összességét (statisztikai) **sokaságnak**, vagy **populációnak** nevezzük.

1. példa

Statisztikai sokaság lehet például: Magyarországon bejegyzett kft-k egy meghatározott időpontban, egy üzemben gyártott termékek összessége egy meghatározott időszakban, stb.

A statisztikai sokaságok közötti lényeges különbség, hogy azok időpontra vagy időszakra vonatkoztatva értelmezhetők.

Az olyan statisztikai sokaságot, amely egy adott időpontra vonatkozóan értelmezhető **állósokaságnak** vagy **stock** jellegű sokaságnak nevezzük.

2. példa

A világ népessége 2001. január 1-jén.

Az olyan statisztikai sokaságot, amely egy adott időszakra vonatkozóan értelmezhető **mozgósokaságnak** vagy **flow** jellegű sokaságnak nevezzük.

3. példa

Halálozások, születések alakulása Magyarországon 2001-ben.

Az előző példákból látható, hogy bizonyos álló és mozgó sokaságok összefügghetnek egymással. Ha egy állósokaságra vonatkozó régebbi információnkat úgy tesszük aktuálissá, hogy a kapcsolódó mozgósokaság segítségével növelést vagy csökkentést eszközölünk, akkor **továbbvezetésről** beszélünk.

Ismérv

A sokaság statisztikai egységekből áll, ezek a vizsgált információ hordozói. Meghatározott tulajdonságokkal, jellemzőkkel, vagy más néven ismérvekkel rendelkeznek.

Az **ismérv** a statisztikai egységeknek a statisztikai vizsgálat szempontjából fontos olyan tulajdonsága, amely alapján a sokaság egymást át nem fedő részekre bontható.
Az ismérvek lehetséges értékei az **ismérvváltozatok.** Az ismérveket általában X-szel, míg az ismérvváltozatokat $x_1, x_2,\ldots,x_n$-nel jelöljük.
Azokat a jellemzőket, melyek szerint a sokaság egységei egyformák **közös ismérveknek**, míg azokat melyek szerint különbözőek **megkülönböztető ismérveknek** nevezzük.
Az ismérveknek az alábbi típusai lehetnek:

- területi,
- időbeli,
- minőségi,
- mennyiségi.

A **területi** (**földrajzi**) és **időbeli ismérvek** a statisztikai egységek térbeli, illetve időbeli jellemzését adják.

A **minőségi ismérvek** a sokaság egységeit verbálisan jellemzik. A mindössze két ismérvváltozattal rendelkező ismérveket **alternatív ismérveknek** nevezzük.

A **mennyiségi ismérvek** kvantifikálhatóak, és ismérvváltozatait általában ismérvértékeknek nevezzük. Két fajtájukat különböztetjük meg: a **diszkrét típusú** (csak egész számmal kifejezhető) és a **folytonos típusú** (nem csak egész számmal kifejezhető) ismérvváltozatokat.

4. példa

Statisztikai sokaság: Budapest állandó lakosai 2001. január 1-jén.

A sokaság típusa: állósokaság.

Fontosabb ismérvei és típusai:

lakóhely: területi;
születési idő: időbeli;
életkor: mennyiségi;
foglalkozás: minőségi;
nem: alternatív.

1.3. A statisztikai munka fázisai

A statisztikai munka 4 fázisát különíthetjük el: tervezés, adatgyűjtés, statisztikai adatok feldolgozása, statisztikai adatok elemzése.

Tervezés

Az első feladat annak rögzítése, hogy mi a statisztikai munka célja. Magyarországon az adatvédelmi törvény tartalmazza a **célhoz kötöttség elvét**. Ez azt jelenti, hogy személyes adatot gyűjteni, feldolgozni csak pontosan meghatározott jogszerű célra szabad. A tervezés szakaszában kell eldönteni, hogy milyen típusú adatokat kívánunk begyűjteni, mely megfigyelési egységekről. Meg kell határozni, hogy melyek lesznek a **számbavételi egységek**, amelyekkel kapcsolatot hozunk létre az adatok begyűjtése érdekében. Dönteni kell az adatgyűjtés jellemzőiről: gyakoriságáról, köréről, idejéről, helyéről, módjáról.

Adatgyűjtés

Az adatgyűjtés vagy **adatfelvétel** (röviden: **felvétel**) a statisztikai adatok beszerzését jelenti. Több módja ismeretes.

- Kikérdezés: ez történhet személyes interjúban vagy postai úton kérdőívvel. A piac- és közvéleménykutatásban alkalmazzák leggyakrabban. A kikérdezéseknél gondot okoz a hibás válaszok kezelése.
- Megfigyelés: az adatok rögzítését közvetlen megfigyelés vagy mérőműszer segítségével végezhetjük el. Megfigyelést alkalmaznak pl. forgalomszámlálásnál, testmagasság megállapításánál. A mérési hibának fontos szerepe van.
- Kísérlet: Ennek során valamilyen hipotézis ellenőrzését végezzük. A hipotézis feltételeinek teljesüléséről gyakran külön gondoskodni kell megfelelő beavatkozással, **kezeléssel**. Ismertek az ún. **kontrollált kísérletek**, amelyek esetében valamely változót tervszerűen változtatnak ceteris paribus. A közgazdaságtanban a kísérletezés többnyire nem lehetséges.

Az adatgyűjtés (körét tekintve) lehet teljes vagy részleges. A részleges megfigyelés lehet:

- reprezentatív megfigyelés (mintavétel),
- kontrollált-kísérlet,

– egyéb részleges megfigyelés.

Feldolgozás

Itt kell elvégezni az adatok ellenőrzését és helyesbítését; azok osztályozását, az eredmények táblákba foglalását. Ez történhet kézi vagy gépi eszközökkel.

Elemzés

Matematikai és logikai műveletek végzését jelenti:

– különböző (később ismertetett) módszereket alkalmazunk, mutatószámokat képezünk, összefüggéseket, tendenciákat keresünk;
– elvégezzük a szöveges elemzést, különféle szemléltető eszközöket alkalmazunk.

Az elemzés célját tekintve megkülönböztetünk leíró és induktív (következtető) statisztikákat. A leíró statisztika területe az adatgyűjtésre, adatok ábrázolására, csoportosítására, és egyszerűbb mutatószámok meghatározására terjed ki; míg az induktív statisztikában helyet kap egy általánosítási törekvés. Ez utóbbinak, mivel jóval hasznosabb, nagyobb a gyakorlati alkalmazása.

1.4. Mérési szintek (skálák) és tulajdonságaik

A legegyszerűbb, legkevésbé informatív mérési szint a nominális skála.
A **nominális** (**névleges**) **skálán** az ismérvértékek azonossága vagy különbözősége állapítható csak meg.

5. példa

Vallás, nem, foglalkozás, állampolgárság, stb.

Ha tudjuk két ember állampolgárságát, akkor csak azt tudjuk megállapítani, hogy azok azonos állampolgárságúak-e, vagy sem; egyéb relációnak nyilván nem szabad jelentőséget tulajdonítani. A névleges mérési szintű adatokkal végzett aritmetikai műveletek értelmetlenek.

A következő fokozat az ordinális mérési szint.
Az **ordinális** (**sorrendi**) **skálán** az ismérvértékek közötti sorrendiség is megállapítható.

6. példa

Termékek, szolgáltatások minőségi fokozati, különböző rendfokozatok, stb.
Az ordinális skála ismérvértékei nem tartalmaznak információt azok abszolút nagyságára vonatkozóan, így azok közötti különbség nagysága sem állapítható meg.

Az **intervallum-** vagy **különbségi skálán** már az ismérvértékek közötti mennyiségi különbség is megállapítható, valós információt hordoz.

7. példa

Hőmérséklet, tengerszint feletti magasság, földrajzi szélesség, hosszúság, stb.
Itt a skála kezdőpontja mindig valamilyen önkényesen választott 0-pont, ezért az ismérvértékek aránya nem értelmezhető. Azt mondhatjuk például, hogy 20°C és 10°C között 10°C a hőmérsékletkülönbség; az viszont nem igaz, hogy a 20°C kétszer olyan meleget jelent, mint a 10°C, hiszen ugyanezen hőmérsékletek Kelvin skálán mért értékei között, már más arány adódna. A különbségi skála mindig valamilyen mértékegységgel adott.

A legtöbb információt az **arányskála** nyújtja, itt a kezdőpont is egyértelműen adott.

8. példa

Költségek, jövedelmi adatok, súly, hosszúság mérése, életkor, stb.
Az arányskála adatain minden matematikai és statisztikai művelet értelmes módon elvégezhető.

Skálatranszformáció (Egy lehetséges értelmezése)

A **skálatranszformáció** a skála értékeinek más értékekre történő transzformációja oly módon, hogy a skála tulajdonságai változatlanok maradnak.

Skálatranszformációt hajtunk végre például, amikor valamilyen minőségi ismérvek verbálisan adott értékeit (szám) kódokkal helyettesítjük.

9. példa

A nemek (férfi, nő) új módon való kódolása pl. 0 és 1 számjegyekkel.

A statisztikában az intervallum- és arányskálák összefoglaló neveként gyakran alkalmazzuk a **kardinális skála** vagy **metrikus skála** fogalmakat.

Azokat a skálákat, ahol nominális vagy ordinális skálán mérhető ismérveket valós számértékekkel mérjük **álkardinális skáláknak** nevezzük.

10. példa

Jeles, jó, közepes, elégséges, elégtelen osztályzatok 5,4,3,2,1 valós számértékekre történő transzformációja. Ez nyilvánvalóan ordinális skála, hiszen csak a teljesítmények sorrendje állapítható meg, azt nem lehet tudni, hogy mekkora a különbség két osztályzat között, lehet hogy csak „1 pont”, de lehet, hogy több. Az pedig már végképp nem állítható, hogy például egy 4-es osztályzat eléréséhez kétszer olyan jól kell teljesíteni, mint a ketteshez, hiszen 2-es legtöbbször csak 50% fölötti teljesítményért jár.

Lineáris skálatranszformációról beszélünk, ha a transzformációt pl. $y=ax+b$ alakú lineáris egyenlet szerint hajtjuk végre.

1.5. A statisztikai adatok pontossága

A statisztikai adatok egyik legfontosabb jellemzője a pontosság. Mindig döntenünk kell azonban a pontosság, gyorsaság és gazdaságosság követelményei között, mert egyszerre (általában) nem lehet mindhármat optimalizálni. Gyakran meg kell tehát határoznunk, hogy milyen pontossággal várjuk el a statisztikai adatokat egy adott elemzés esetében. Tökéletesen pontos adatokhoz gyakorlatilag soha sem juthatunk hiszen, ahogy az adatgyűjtés módjainál láttuk, egyfajta felvételi hiba mindig létezik. Ezen kívül, az adatok rögzítése és feldolgozása során is keletkeznek bizonyos hibák. Külön kell beszélnünk a reprezentatív megfigyelésből adódó hibákról. Ezek oka az, hogy nem figyeljük meg a sokaságot teljes körűen. Ez a hiba az eddigiekkel szemben matematikailag kezelhető, számszerűsíthető, ha a megfigyelési egységekből álló minta kiválasztása a követelményeknek megfelelően, véletlenszerűen történik. Ezt a hibát mintavételi hibának nevezzük. (A mintával kapcsolatos törvényszerűségeket, eljárásokat a második kötet tartalmazza.)

Indokolt tehát az adatok és mutatószámok

$$A \mp a \tag{1}$$

alakú megadása, ami úgy értelmezhető, hogy adatunk az $[A-a,A+a]$ intervallumba esik. Az a mennyiséget **abszolút hibakorlátnak** nevezzük. A statisztikai gyakorlatban bevett szokás szerint az adatok pontosságára úgy utalunk, hogy értékét (kerekítve) olyan számjegyekkel közöljük, amelyek még biztosan pontosnak tekinthetőek (az 5-öt és annál nagyobb számjegyeket felfelé, az 5-nél kisebbeket lefelé kerekítjük). Ezek a számjegyek az ún. **szignifikáns számjegyek.** Ha az utolsó szignifikáns számjegy helyértéke 10^{sz}, akkor (a kerekítési konvenció alapján) a hibakorlát becsülhető:

$$\hat{a} = \frac{10^{sz}}{2}. \tag{2}$$

(Megjegyzés: az $\hat{a}$ szimbólum kiejtése „a becsült értéke”.) A ^ jellel mindig arra utalunk, hogy az adatunk becsült értékű.

Gyakran nem az abszolút hanem a **relatív hibakorláttal** dolgozunk:

$$\alpha = \frac{a}{A}. \tag{3}$$

A relatív hibakorlátot, amely az abszolút hibakorlát és a közölt adat hányadosa, általában százalékban kifejezve adjuk meg.
A becsült relatív hibakorlát:

$$\hat{\alpha} = \frac{\hat{a}}{A}. \qquad (4)$$

2. Egyszerű elemzések

2.1. Sokaság nagyságának meghatározása

A statisztikai adat a sokaság valamilyen számszerű jellemzője. Ezek közül a legegyszerűbb a sokaság nagyságát jellemző érték. Azért fontos, mert megadja a vizsgált sokaság súlyát, fontosságát a gazdasági, társadalmi és természeti jelenségek között. Természetesen csak véges sokaságok nagysága adható meg.

- A diszkrét sokaságok nagyságát **megszámlálással** állapítjuk meg.
- A folytonos sokaságok nagyságának meghatározása már csak **méréssel** történhet.

11. példa

A megszámlálás egy klasszikus esete a népszámlálás.
Egy gazdaság adott időszakra vonatkozó tejtermelése csak valamilyen méréssel adható meg.

2.2. Statisztikai sorok, táblák

Statisztikai adatok valamilyen ismérv szerinti felsorolását **statisztikai sornak** nevezzük.

Statisztikai sor keletkezhet:

- azonos fajta adatokból: összehasonlító sor, csoportosító sor;
- különböző fajta adatokból: leíró sor.

12. példa

Egy kft dolgozóinak nemek szerinti megoszlását az 1. táblázat tartalmazza.

Egy kft dolgozóinak nemek szerinti megoszlása

1. táblázat

Nem	Fő
Férfi	15
Nő	5
Összesen	20

Forrás: fiktív példa

Statisztikai sorok összefüggő rendszerét **statisztikai táblának** nevezzük.

13. példa

Egy kft dolgozóinak megoszlását nemek és az adott munkahelyen eltöltött idő szerint a 2. táblázat tartalmazza.

A kft dolgozóinak megoszlása nemek és az adott munkahelyen eltöltött idő szerint

2. táblázat

Munkahelyen eltöltött évek száma	Férfi	Nő	Összesen
–4	2	2	4
5–9	5	2	7
10–14	4	0	4
15–19	3	1	4
20–	1	0	1
Összesen	15	5	20

Forrás: fiktív példa

Minden statisztikai sornak és táblának megkövetelt formai eleme a cím és a forrás megnevezése, illetve kötelező feltüntetése. (Megjegyzés: az egyszerűség kedvéért, a továbbiakban ettől sokszor eltekintünk.)

A sorok és táblák számítógépes tárolása $m \cdot n$-es mátrixokban történik. Egy adatbázisban a mátrix oszlopainak fejlécei foglalják magukba az egyes ismérvek megnevezését, míg a többi sor az ismérvváltozatokat tartalmazza (ezeket nevezzük rekordoknak). Minden egyes rekordban azonos számú mező van. Mi a továbbiakban a Microsoft Excel 7.0 táblázatkezelőt fogjuk használni. Ez egy Windows alapú program, amely alapvetően táblázatkezelő, de hasznos statisztikai műveletek elvégzésére is képes.

Indítsuk el a Microsoft Excelt, és gépeljük be a fenti tábla adatait. A bevitelnél ügyeljünk arra, hogy a hosszabb szövegeket is egyetlen cellába írjuk. A cellák között a kurzormozgató billentyűkkel, illetve egeres kattintással mozoghatunk. Az Excel tulajdonságai közé tartozik az AutoSzámolás funkció. Ezzel folyamatosan visszajelzést kaphatunk az állapotsorban (a képernyő alján) a kijelölt cellák összegéről. Ellenőrizzük ennek segítségével a táblázat összesen sorában levő számok pontosságát!

Összehasonlítás

Az **összehasonlítás** alkalmával több sokaság nagyságának vagy más jellemző adatának egymás mellé rendelését végezzük. Ez történhet egyszerű felsorolással, különbség

képzéssel vagy hányados képzéssel. Összehasonlítás céljából egymás mellé sorolt adatok összességét **összehasonlító sornak** nevezzük.

Csoportosítás (osztályozás)

A statisztikai sokaság egy vagy több ismérv szerinti tagolását csoportosításnak vagy **osztályozásnak** nevezzük. Azt az ismérvet ami alapján a sokaság osztályait elhatároljuk egymástól **csoportképző ismérvnek** nevezzük. Egy osztályozástól azt várjuk el, hogy:

- teljes legyen (a sokaság minden egysége besorolható egy osztályba);
- átfedésmentes legyen (minden sokasági egység csak egy osztályba sorolható be);
- minél homogénebbek legyenek az osztályok (az osztályokon belüli egységek minél jobban hasonlítsanak egymáshoz a vizsgált ismérv szempontjából).

A sokaság egy ismérv szerinti csoportosítását **csoportosító sornak** nevezzük.

A k db osztályból álló csoportosító sor általános alakja a 3. táblázatban látható.

Csoportosító sor

3. táblázat

Ismérvváltozatokat tartalmazó osztályok	Előfordulások száma
C_1	f_1
C_2	f_2
⋮	⋮
C_i	f_i
⋮	⋮
C_k	f_k
Összesen	N

A második oszlopban levő számokat a statisztikában általában **gyakoriságoknak** nevezzük.

A sokaság több ismérv szerinti csoportosításának eredménye a **kontingencia** vagy **kombinációs** tábla.

Az r sorból és c oszlopból álló kétdimenziós kombinációs tábla általános alakját a 4. táblázat tartalmazza.

Kombinációs tábla

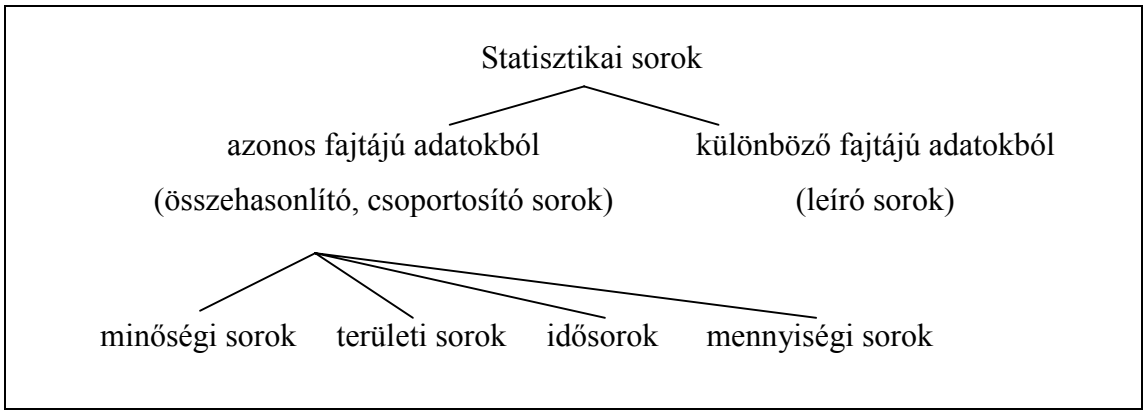
4. táblázat

Az Y ismérv szerinti osztályok / Az X ismérv szerinti osztályok	C_1^Y	C_2^Y	…	C_j^Y	…	C_c^Y	Összesen
C_1^X	f_{11}	f_{12}	…	f_{1j}	…	f_{1c}	$f_{1.}$
C_2^X	f_{21}	f_{22}		f_{2j}		f_{2c}	$f_{2.}$
⋮	⋮						
C_i^X	f_{i1}	f_{i2}		f_{ij}		f_{ic}	$f_{i.}$
⋮	⋮						
C_r^X	f_{r1}	f_{r2}		f_{rj}		f_{rc}	$f_{r.}$
Összesen	$f_{.1}$	$f_{.2}$		$f_{.j}$		$f_{.c}$	$\boldsymbol{N}$

A 4. táblázat utolsó sorában ($f_{.j}$) és oszlopában ($f_{i.}$) szereplő gyakoriságokat a statisztikában **peremgyakoriságoknak** vagy **feltétel nélküli eloszlásoknak** nevezzük, míg a többi gyakoriságot (f_{ij}) **feltételes eloszlásoknak** nevezzük. (Az asszociációról szóló fejezetben ezekkel részletesebben foglalkozunk.)

A statisztikai sorok vázlatos áttekintése az 1. ábrán látható.

A statisztikai sorok vázlatos áttekintése



1. ábra

A továbbiakban az 1. ábra alsó sorában felsoroltakkal foglalkozunk részletesebben.

Minőségi sorok

Minőségi ismérv szerint szerkesztett sort **minőségi sornak** nevezzük.

14. példa

Az 3. táblázatban a vizsgált ismérv legyen minőségi ismérv. (Lásd a 12. példát.)

Területi sorok

Területi sorról akkor beszélünk, ha a sor kialakításakor a rendező elv valamilyen területi hovatartozás.

15. példa

Az 3. táblázatban a vizsgált ismérv legyen területi (földrajzi) ismérv.

Egy kft dolgozóinak lakóhely szerinti megoszlását az 5. táblázat tartalmazza.

Egy kft dolgozóinak lakóhely szerinti megoszlása

5. táblázat

Lakóhely	Fő
Szeged	16
Egyéb	4
Összesen	20

Idősorok

Az **idősoroknak** két fajtája van: állapotidősor (stock típusú) és tartamidősor (flow típusú).

Állapotidősor: egy állósokaság időbeli alakulását jellemzi.

16. példa

Az 3. táblázatban a vizsgált ismérv legyen stock típusú ismérv.

Egy kft foglalkoztatottainak számát (az 1997-1999 közötti időszakban) a 6. táblázat tartalmazza.

Egy kft dolgozóinak száma az év első napján

6. táblázat

Év	Fő
1997	17
1998	19
1999	20

Az ilyen típusú ismérveket tartalmazó táblázatok összesen sorának (az ún. **összegző sornak**) nincs értelme, ezért nem is szerepel.

Tartamidősor: egy mozgósokaság egy-egy időszak alatt bekövetkezett változását jellemzi.

17. példa

Az 3. táblázatban a vizsgált ismérv legyen flow típusú ismérv.

Egy kft forgalmának nagysága 3 év alatt az alábbiak szerint alakult.

Egy kft forgalma (millió Ft)

7. táblázat

Év	Forgalom
1997	16,4
1998	24,0
1999	31,2
Összesen	71,6

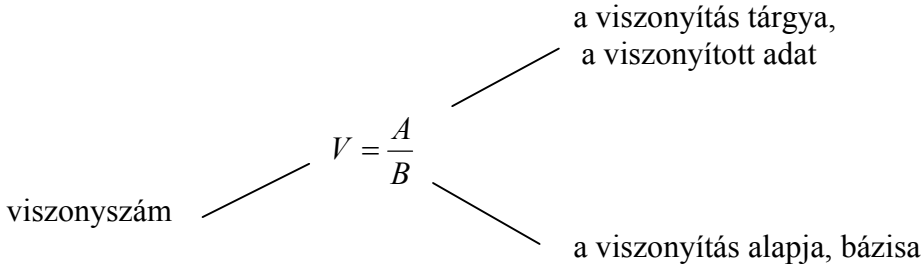
Az ilyen típusú ismérveket tartalmazó táblázatok összesen sorának van értelme. Jelen esetben a teljes vizsgált időszak összforgalmát jelenti.

Mennyiségi sorok

Ezeket a sorokat a harmadik fejezetben majd részletesebben tárgyaljuk.

2.3. Viszonyszámok

Nagytömegű, eredeti formájában áttekinthetetlen adathalmaz kezelésére viszonyszámokat is használhatunk. A viszonyszám nem más, mint adatok vagy mutatószámok hányadosa.



(5)

Három legfontosabb fajtája: dinamikus, megoszlási és intenzitási viszonyszám.

Dinamikus viszonyszám: azonos sokaság, időben különböző adataiból számított hányados, százalékos formában szoktuk megadni. Kettőnél több (i=1,2,…,N) adatból álló idősor esetén kétféle fajtája képezhető.

Bázisviszonyszám:

$$b_i = \frac{x_i}{x_b} \qquad i=1,2,\ldots,N. \tag{6}$$

Láncviszonyszám:

$$l_i = \frac{x_i}{x_{i-1}} \qquad i=2,3,\ldots,N. \tag{7}$$

Gyakran az idősor összes időegységére kiszámítjuk az adott viszonyszámot és a keletkező viszonyszámsort használjuk elemzésre.

A lánc- és bázisviszonyszámokra vonatkozó azonosságokat a (8)-(12) képletek mutatják.

Egymást követő bázisviszonyszámok hányadosa láncviszonyszám.

$$\frac{b_i}{b_{i-1}} = \frac{x_i}{x_b} : \frac{x_{i-1}}{x_b} = \frac{x_i}{x_{i-1}} = l_i \tag{8}$$

Áttérés új bázisra: a bázisviszonyszámokat elosztjuk az új bázishoz tartozó régi viszonyszámmal.

$$\frac{b_i}{b_c} = \frac{x_i}{x_b} : \frac{x_c}{x_b} = \frac{x_i}{x_c} = c_i \tag{9}$$

Bázisidőegységet követő egymás után következő (*m* db) láncviszonyszám szorzata bázisviszonyszámot ad.

$$\prod_{i=b+1}^{m} l_i = \frac{x_{b+1}}{x_b} \cdot \frac{x_{b+2}}{x_{b+1}} \cdot \ldots \cdot \frac{x_{b+m}}{x_{b+m-1}} = \frac{x_{b+m}}{x_b} = b_m \qquad m \le N \tag{10}$$

Láncviszonyszámokból (a vizsgált időszakban) tetszőleges bázisú bázisviszonyszámokat lehet kiszámítani az alábbi összefüggések szerint:

– az időtengelyen jobbra (a jövőbe) haladva

$$b_{i+1} = b_i \cdot l_{i+1}, \tag{11}$$

– az időtengelyen balra (a múltba) haladva

$$b_i = b_{i+1} : l_{i+1}. \tag{12}$$

18. példa

A népesség számát minden év első napjára a 8. táblázat tartalmazza.

Magyarország népessége 1991-1999 között

8. táblázat

Év	Népesség száma, ezer fő
1991	10 355
1992	10 337
1993	10 310
1994	10 277
1995	10 246
1996	10 212
1997	10 174
1998	10 135
1999	10 092

Forrás: Magyar statisztikai zsebkönyv ’98, KSH, Bp., 1999.

Számítsuk ki a fenti idősorból a népesség alakulásának bázisviszonyszámsorát 1991-es és 1999-es bázissal is! Számítsuk ki a láncviszonyszámsort is! Alkalmazzuk az azonosságokat ellenőrzésre! Használjuk a feladat megoldásához az Excelt! Vigyük fel az adatokat! Az eredményt a 2. ábra mutatja.

Az 1991-es bázisévhez tartozó viszonyszámsort úgy tudjuk kiszámítani, hogy az egyes évekhez tartozó adatokat osztjuk az 1991-es év adatával. A cellák feltöltése eredményt szolgáltató képlettel a következőképpen végezhető el: a cella mezőjébe írjuk be egyenlőségjel után annak a műveletnek megfelelő képletet, amelyet a kiindulási cellákkal akarunk elvégezni, úgy, hogy azokra a megfelelő oszlop és sor jelekkel hivatkozunk. (Ez megjelenik a Szerkesztőlécben, a táblázat fölött is.)

Az Excel munkalapjának részlete

	A	B	C	D
1	Év	Népesség	1991=100%	Előző év=100%
2	1991	10 355	=100*B2/B$2	
3	1992	10 337		
4	1993	10 310		
5	1994	10 277		
6	1995	10 246		
7	1996	10 212		
8	1997	10 174		
9	1998	10 135		
10	1999	10 092		

2. ábra

A B2 stílusú jelölés relatív hivatkozás, a B$2 pedig a sorra nézve abszolút hivatkozás.

Magyarország népességének bázis- és láncviszonyszámsorai

9. táblázat

Év	Népesség	1991=100%	Előző év=100%	1999=100%
1991	10 355	100,0	-	102,6
1992	10 337	99,8	99,8	102,4
1993	10 310	99,6	99,7	102,2
1994	10 277	99,2	99,7	101,8
1995	10 246	98,9	99,7	101,5
1996	10 212	98,6	99,7	101,2
1997	10 174	98,3	99,6	100,8
1998	10 135	97,9	99,6	100,4
1999	10 092	97,5	99,6	100,0

Ennek megkülönböztetésére a következők miatt van szükség: ha egy viszonyszámsor első adatát a fenti módon kiszámítottuk, akkor a többi képletet nem kell begépelni, elég az adott cella jobb alsó sarkát az egérrel lefelé a többi cellára húznunk, és a megfelelő képleteket kapjuk a többi cellában is. A megfelelően alkalmazott relatív és abszolút hivatkozások eredményezik azt, hogy a helyes képleteket (értékeket) kapjuk. A többi viszonyszámsor hasonlóan kiszámítható. A kapott eredményeket a 9. táblázat tartalmazza.

Megoszlási viszonyszám: valamely sokaságrésznek az egész sokasághoz viszonyított nagysága, százalékos formában szoktuk megadni.

19. példa

A 12. példa adatai alapján elkészíthető a 10. táblázat.

Egy kft dolgozóinak nemek szerinti megoszlása

10. táblázat

Nem	Megoszlás (%)
Férfi	75
Nő	25
Összesen	100

Intenzitási viszonyszám: két egymással valamilyen kapcsolatban álló sokaság valamilyen adatából képzett hányados. Lehet egyenes vagy fordított, illetve ettől függetlenül, nyers vagy tisztított.

Egyenes intenzitási viszonyszámról beszélünk, ha a társadalmi megítélés szempontjából az lenne a jó, ha a viszonyszám értéke minél nagyobb lenne.

Fordított intenzitási viszonyszámról beszélünk, ha a társadalmi megítélés szempontjából az lenne a jó, ha a viszonyszám értéke minél kisebb lenne.

Ha egy intenzitási viszonyszám esetén a viszonyítás alapjának csak egy része kötődik jobban a viszonyítás tárgyához, akkor gyakran egy új intenzitási viszonyszámot alkotunk, amelyben a viszonyítás alapja az említett részsokaság lesz. Az íly módon létrejövő új viszonyszámot **tisztított intenzitási viszonyszámnak**, míg az elsőt **nyers intenzitási viszonyszámnak** nevezzük.

A nyers intenzitási viszonyszám az (5) képlet szerint legyen $\frac{A}{B}$, a tisztított intenzitási viszonyszám pedig $\frac{A}{b}$. Közöttük felírható a következő összefüggés:

$$\frac{A}{B} = \frac{A}{b} \cdot \frac{b}{B}, \tag{13}$$

ahol $\frac{b}{B}$ a tiszta rész arányát jelenti (ami egy megoszlási viszonyszám).

20. példa

Egy hónap alatt 100 alkalmazott (80 fizikai és 20 szellemi foglalkozású) 120 db terméket állít elő. Vállalti szinten a termelékenységük $\frac{120\ \text{db}}{100\ \text{fő}} = 1{,}2\ \text{db/fő}$. Ez nyers intenzitási viszonyszám. A tisztított pedig: $\frac{120\ \text{db}}{80\ \text{fő}} = 1{,}5\ \text{db/fő}$. A (13) képlet szerint igaz a következő összefüggés:

$$\frac{120}{100} = \frac{120}{80} \cdot \frac{80}{100} = 1{,}2\ \ [\text{db/fő}].$$

2.4. A grafikus ábrázolás eszközei

A grafikus ábrázolás nem kifejezetten elemzési módszer, hanem a statisztikai adatok szemléltető megjelenítésének eszköze, melyben az információsűrítés bizonyos elemei is megjelennek. Minden grafikus ábrázolás lényege az összehasonlítás. Általában pontokat, vonalakat, köröket, oszlopokat használunk.
A grafikus ábrázolás alábbiakban ismertetett fajtáit használjuk a leggyakrabban.

Diagramok

Diagramokon belül megkülönböztetjük a következőket:
- **pontdiagram**: két ismérv szerinti hovatartozást ábrázolunk vele;
- **vonaldiagram**: egyenes szakaszokból álló grafikus ábra;
- **síkdiagram**: gyakoriságokat ábrázolunk vele, területek segítségével (pl. **oszlop-** vagy **kördiagram**).

Kartogramok

Kartogramok: gyakoriságok térképen alapuló ábrázolása.

Sztereogramok

Sztereogramok: három releváns dimenzióban történő ábrázolás, három ismérv szerinti hovatartozást ábrázolunk vele.

Piktogramok

Piktogramok: figurális ábrázolás, gyakoriságok különböző nagyságú vagy számú képszimbólumokkal való ábrázolása.

A grafikus ábrázolásnál figyelnünk kell a következő alapelvekre:

- mindig az alapul vett síkidomok területe kell, hogy arányos legyen az ábrázolni kívánt adat nagyságával;
- mindig legyen címe a grafikus ábrának;
- az adatok forrásának feltüntetése kötelező;
- idősorokat általában vonaldiagrammal, a sokaság szerkezetét általában **osztott oszlop-** vagy **osztott kördiagrammal** szemléltetjük (lásd a 4. és az 5. ábrát!);

– állapotidősornál az *x* tengelyen szereplő időpontokhoz (lásd a 3. ábrát!), tartamidősornál az *x* tengelyen kijelölt időszak közepéhez igazítunk.

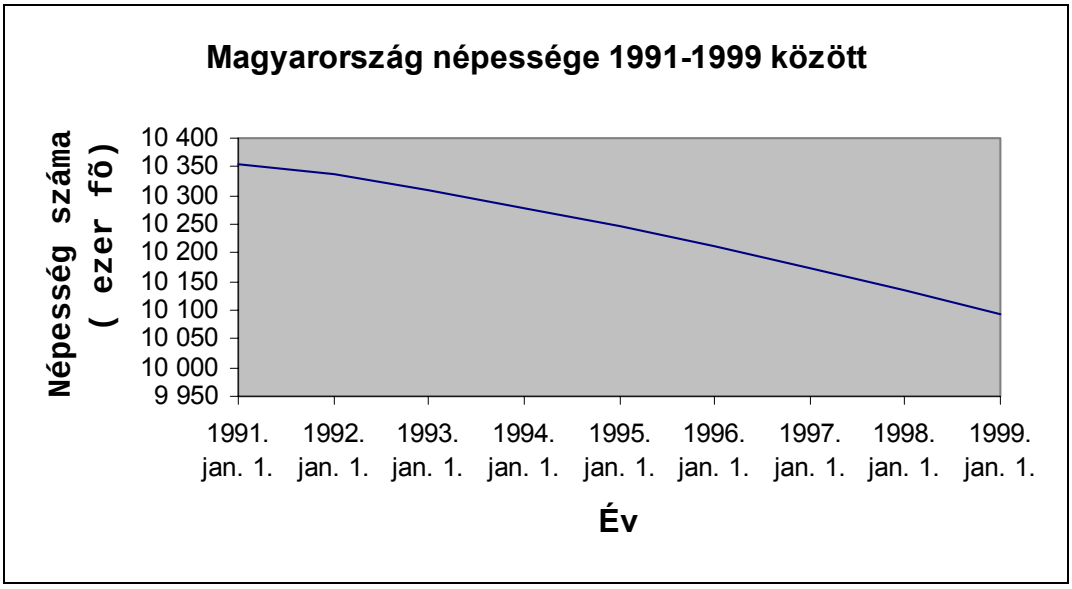
Az ismertetett grafikus ábrázolási módok közül néhány a 3., 4. és 5. ábrán látható.

21. példa

A 18. példa adatai alapján ábrázoljuk vonaldiagram segítségével Magyarország népességének változását 1991-1999 között!

A vonaldiagramot az Excel segítségével készítjük el. A 18. példában már bevittük az adatainkat az **A1-B10** cellatartományba. Jelöljük most ki a **B2-B10** cellákat, és indítsuk el a Diagram Varázslót a **Beszúrás** menü **Diagram...** almenüjének segítségével (ez ikonnal is meghívható)!

Az Excel outputja



3. ábra

Első lépésként válasszuk ki a nekünk megfelelő diagramtípust! A **Tovább>** nyomógomb segítségével léphetünk tovább. Második lépésként az **Adatsorok** menü alatt A kategóriatengely (X) feliratai: mezőbe vigyük be az **A2-A10** cellatartományt a munkalapon történő kijelölésével. Lépjünk tovább. A harmadik lépésben a **Címek** menüben írhatjuk be a diagram címét és a tengelyek megnevezését. A **Rácsvonalak** menüben a vezető és segédrácsokat állíthatjuk be.

Ha ezekre nincs szükségünk, kapcsoljuk ki a jelölőnégyzeteket! **Jelmagyarázat** menüben állíthatjuk be azt, hogy szükségünk van-e jelmagyarázatra, és hogy az hova kerüljön. A negyedik lépésben pedig azt kell eldöntenünk, hogy a diagramunk új lapra (diagramlap) kerüljön, vagy az eredeti munkalapunkra. A kapott diagram a 3. ábrán látható.

A kész diagram beállításait utólag módosíthatjuk a **Formátum** menüjének segítségével, ha a grafikus ábra megfelelő részét aktivizáljuk az arra történő egérkattintással.

22. példa

Az 1999. év első negyedévére vonatkozó kötelező gépjármű-biztosítás díjbevételeinek adatait a 11. táblázat tartalmazza.

1999. első negyedévi díjbevételek

11. táblázat

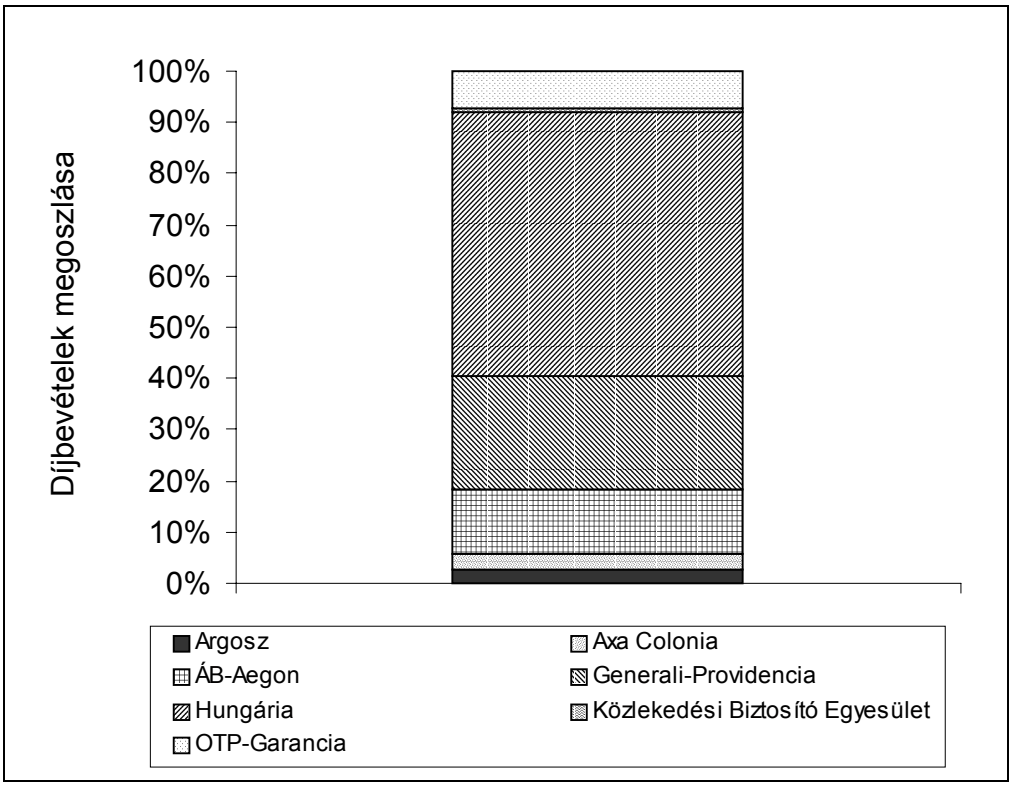
Biztosítók	Díjbevételek (ezer Ft)
Argosz	428 145
Axa Colonia	478 922
ÁB-Aegon	1 986 164
Generali-Providencia	3 455 826
Hungária	8 138 255
Közlekedési Biztosító Egyesület	100 207
OTP-Garancia	1 154 755
Összesen	15 742 274

Forrás: ÁBIF

A sokaság szerkezetének ábrázolására leginkább az osztott oszlop-, illetve az osztott kördiagram alkalmas. Ezeket láthatjuk a 4. és az 5. ábrán.

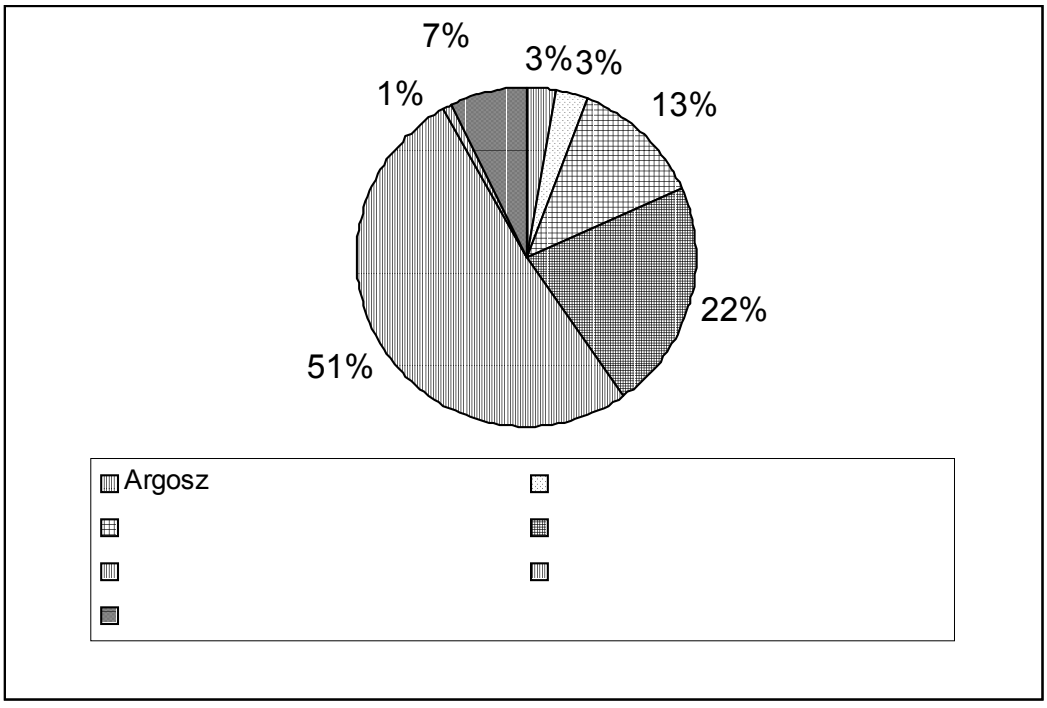
A vonaldiagram rajzolásakor ismertetett menüpontok megfelelő alkalmazásával az Excelben megszerkeszthető a 4. és az 5. ábra is.

A kötelező gépjármű-biztosítások díjbevételeinek megoszlása osztott oszlopdiagramon



4. ábra

A kötelező gépjármű-biztosítások díjbevételeinek megoszlása osztott kördiagramon



5. ábra

3. Sokaság egy ismérv szerinti vizsgálata

3.1. Mennyiségi sorok

Rangsor

A mennyiségi ismérvek lehetséges értékei rendezett halmazt alkotnak, ezért a sokaság egységei sorba rendezhetőek. Ezt monoton nemcsökkenő módon szoktuk megtenni.

A sokaság egységeinek (és a hozzájuk tartozó ismérvértékeknek) mennyiségi ismérv szerinti monoton nemcsökkenő felsorolását **rangsornak** nevezzük.

(Rendezett halmaz elemeinek sorba rendezésére számos rendezési algoritmus létezik: beszúró rendezés, gyorsrendezés, kupac rendezés, stb.)

23. példa

Egy kft dolgozóinak kereseti adatai a következőek (ezer Ft):

70,2; 63,0; 52,5; 77,4; 54,3; 48,1; 42,2; 70,1; 51,0; 63,2; 55,8;

56,7; 36,2; 42,0; 51,0; 53,9; 42,5; 48,0; 53,3; 78,0; 68,6; 47,1.

A sorbarendezést gyorsan elvégezhetjük az Excel segítségével. Vigyük be az adatokat az **A2-A23** cellákba, az **A1** a fejlécet tartalmazza. Jelöljük ki az előbbi 22 cellát! Ezt megtehetjük az egérrel, annak bal gombját lenyomva tartva mozgatva az egérkurzort, vagy billentyűzettel, a SHIFT gomb lenyomása mellett használva a billentyűzet kurzorgombjait. Miután kijelöltük az **A2-A23** cellatartományt, az **Adatok** menü **Sorba rendezés…** almenüvel elvégezhetjük a kívánt rendezést. (A rendezés ikon segítségével is elvégezhető.)

A keresetek rangsora:

36,1; 42,0; 42,2; 42,5; 47,1; 48,0; 48,1; 51,0; 51,0; 52,5; 53,3;

53,9; 54,2; 55,8; 56,7; 63,0; 63,2; 68,6; 70,1; 70,2; 77,4; 78,0.

Gyakorisági sor

Mivel gyakran nagy mennyiségű, és ezért eredeti formájában kezelhetetlen és átláthatatlan adattal kell dolgoznunk, abból elemzést készítenünk, a könnyebb áttekinthetőség érdekében ezeket osztályokba soroljuk. Az osztályok természetesen az X

mennyiségi ismérv lehetséges értékeinek valamilyen alkalmas részhalmazai lesznek. Ezt az osztályozás eredményeként létrejövő csoportosító sort - amely már az alapadatok információjának sűrítését jelenti – **gyakorisági sornak** vagy **gyakorisági eloszlásnak** nevezzük. Az osztályközös gyakorisági sor (a gyakorisági eloszlás legtöbbször előforduló típusa) elkészítése előtt a sokaság egységeit rangsorba rendezzük.

Ahhoz, hogy osztályközös gyakorisági sort készítsünk, először is meg kell határoznunk, hogy hány osztályt alkossunk. Az osztályok optimális számához a (14) képlet ad támpontot.

$$2^k > N, \quad k \to \min \tag{14}$$

Tehát azt a legkisebb *k*-t keressük, amelyre 2 *k*-adik hatványa már nagyobb, mint a sokaság nagysága. Ez nem annyira szigorú szabály, amitől ne lehetne eltérni, inkább csak támpontot ad.

A második eldöntendő kérdés az az, hogy milyen hosszúságú osztályközöket alakítsunk ki. Gyakran egyenlő hosszúságú osztályközöket képezünk, de természetesen ettől is el lehet térni. Azt kell szem előtt tartanunk, hogy a kapott gyakorisági sor

- könnyen áttekinthető legyen, és
- kevés információveszteséggel járjon.

Osztályköz-hosszúságnak a következő értéket nevezzük:

$$h_i = X_{i,1} - X_{i,0}, \tag{15}$$

ahol h_i az *i*-edik osztályköz-hossz, $X_{i,0}$ az *i*-edik osztályköz alsó határát, $X_{i,1}$ az *i*-edik osztályköz felső határát jelöli. A (15) képlet közvetlenül néha nem alkalmazható! Vegyük például a 2. táblázatban szereplő első és utolsó osztályt, ahol ún. **nyílt osztályok** szerepelnek. Ezeknél az osztályköz-hosszúságok kiszámításához előbb meg kell becsülnünk a legkisebb és legnagyobb ismérvértéket. Ez csak valamilyen utólagos, pótlólagos információ alapján lehetséges. Ilyen például az, hogy a munkahelyen eltöltött évek száma nyilvánvalóan nullánál nagyobb. Ha pótlólagos információval nem rendelkezünk, akkor a nyílt osztályok osztályköz-hosszúságát az őket közvetlenül követő, illetve megelőző osztályok osztályköz-hosszával helyettesítjük.

Ha azonos hosszúságú osztályközök kialakítása mellett döntünk, akkor adott k osztályszám esetén a következőképpen határozzuk meg az osztályközök hosszát:

$$h = \frac{x_{\max} - x_{\min}}{k}, \tag{16}$$

ahol $x_{\min}$ és $x_{\max}$ a legkisebb és a legnagyobb előforduló ismérvérték.

24. példa

Készítsünk a 23. példa rangsorából osztályközös gyakorisági sort!

Mivel $2^4 < 22 < 2^5$, így 4 vagy 5 osztályköz kialakítása látszik célszerűnek. Legyen először k=4. Ekkor $h = \frac{78{,}0 - 36{,}1}{4} = 10{,}475$. (Megjegyzés: a rendelkezésünkre álló adatok abszolút hibakorlátjának becslését figyelembe véve; 10,475 helyett; 10,5-del számolunk.) Az egymást követő osztályközök nem érintkezhetnek egymással, tehát egyik osztályköz felső határa sem lehet egyenlő a következő osztályköz alsó határával, mert az osztályozásnak mindig átfedés- és hézagmentesnek kell lennie (lásd a 2.2. pont alattiakat). Mivel az osztályközök határait az alapadatok pontosságával azonos pontossággal határozzuk meg, ezért nem lehetséges, hogy valamelyik egyed ismérvértéke a „hézagba” essen.

A konkrét osztályközök (k=4; h=10,5 esetén) tehát az alábbiak.

36,1-46,5

46,6-57,0

57,1-67,5

67,6-78,0

A táblázatba írt osztályköz-határok értékeit **közölt határoknak** nevezzük, mert ezeket közöljük az olvasó részére. A határok valódi értelmezése adja a **valódi határokat**. Ezek:

$$C_i = \left[X_{i,0} - \frac{1}{2}10^{sz}, X_{i,1} + \frac{1}{2}10^{sz} \right). \tag{17}$$

Az adott példánál a valódi határok az alábbiak.

[36,05;46,55)

[46,55;57,05)

[57,05;67,55)

[67,55;78,05)

A gyakoriságokat a rangsor alapján gyorsan meg tudjuk határozni, de a feladat Excel segítségével is elvégezhető (ehhez nem szükséges rangsor). Nyissuk meg az előző példában használt mappánkat, ahova a rangsort bevittük az **A2-A23** cellákba. Írjuk be az osztályközeink felső határát ($X_{i,1}$) a **C2-C5** cellákba, majd jelöljük ki a **D2-D5** cellákat. Válasszuk ki a gyakoriság függvényt az alábbi módon. A **Beszúrás** menü **Függvény…** almenüjével illeszthetünk be függvényt (ezt ikon segítségével is megtehetjük). Itt válasszuk ki a Statisztikai függvények közül a GYAKORISÁG(adattömb;csoport_tömb) függvényt, majd Adattömbnek adjuk meg az **A2-A23** tömböt, A2:A23 begépelésével. A Csoport_tömb C2:C5 lesz. A Kész ikonra kattintva a szerkesztőlécben a következő jelenik meg: =GYAKORISÁG(A2:A23;C2:C5). Az egérkurzorral a szerkesztőlécre állva a SHIFT, a CTRL és az ENTER billentyűk együttes lenyomása után a **D2-D5** tömb fogja tartalmazni az eloszlást.

A kapott eredményeket a 12. táblázat tartalmazza.

A kft dolgozóinak kereset szerinti eloszlása

12. táblázat

Keresetek (ezer Ft)	Dolgozók száma (f_i)
36,1 – 46,5	4
46,6 – 57,0	11
57,1 – 67,5	2
67,6 – 78,0	5
Összesen	22

Készítsük el a gyakorisági sort k=5 osztályközt alkalmazva! Ekkor $h = \frac{78{,}0 - 36{,}1}{5} = 8{,}38$. (Megjegyzés: a rendelkezésünkre álló adatok abszolút hibakorlátjának becslését figyelembe véve; 8,38 helyett; 8,4-del számolunk).

A kft dolgozóinak kereset szerinti eloszlása

13. táblázat

Keresetek (ezer Ft)	Dolgozók száma (f_i)
36,1 – 44,4	4
44,5 – 52,8	6
52,9 – 61,2	5
61,3 – 69,6	3
69,7 – 78,0	4
Összesen	22

Megjegyzés: a két különböző k érték szemmel láthatóan jelentősen eltérő eloszlást eredményezett.

Értékösszegsor

A mennyiségi sorok egyik altípusa a már ismertetett gyakorisági sor, a másik altípus az **értékösszegsor**. A C_i osztályhoz tartozó értékösszeget S_i-vel jelöljük, és az

$$S_i = \sum_{x_j \in C_i} x_j \qquad i=1,2,...,k \tag{18}$$

képlettel számítható ki. Összegeznünk kell tehát az adott osztályközbe tartozó sokasági egységek ismérvértékeit.

25. példa

A 23. példa értékösszegsorát a 14. táblázat tartalmazza.

Lehetséges azonban, hogy osztályközös gyakorisági sorból kell értékösszegsort készítenünk, mert az eredeti rangsor túl nagy vagy nem is áll rendelkezésre.

A kft dolgozóinak kereset szerinti tényleges értékösszegsora

14. táblázat

Keresetek (ezer Ft)	S_i
36,1 – 44,4	162,80
44,5 – 52,8	297,70
52,9 – 61,2	273,90
61,3 – 69,6	194,80
69,7 – 78,0	295,70
Összesen	1224,90

Ekkor azonban csak becsülni tudjuk az értékösszeget, és valószínűleg az eredeti adatokból számított sortól eltérőt kapunk. Emiatt megkülönböztetjük a **tényleges** és a **becsült értékösszegsort**. Becsült értékösszegsor számításánál az osztályközepeket vesszük figyelembe.

Az

$$X_i = \frac{X_{i,0} + X_{i,1}}{2} \qquad i=1,2,...,k \tag{19}$$

mennyiségeket az *i*-edik osztályhoz tartozó **osztályközepeknek** nevezzük.

Bizonyos számításoknál, így a becsült értékösszegsor számításánál is, azt feltételezzük, hogy az osztályközbe tartozó sokasági egységek ismérvértékei helyettesíthetőek az osztályközéppel. Azt feltételezzük tehát, hogy a sokaság egységeinek az adott ismérv szerinti eloszlása egyenletes az osztályközökben, de legalábbis az egyes osztályközökbe eső ismérvértékek átlaga az osztályközepet adja minden osztályközben.

Az értékösszegsort ennek megfelelően osztályközös gyakorisági sorból a (20) képlettel lehet meghatározni. (Az eredményt a 15. táblázat tartalmazza.)

$$\hat{S}_i = f_i \cdot X_i \qquad i=1,2,...,k \tag{20}$$

A kft dolgozóinak kereset szerinti becsült értékösszegsora

15. táblázat

Keresetek (ezer Ft)	X_i	f_i	$\hat{S}_i$
36,1 – 44,4	40,25	4	161,00
44,5 – 52,8	48,65	6	291,90
52,9 – 61,2	57,05	5	285,25
61,3 – 69,6	65,45	3	196,35
69,7 – 78,0	73,85	4	295,40
Összesen	–	22	1229,90

Hasonlítsuk össze a becsült és a tényleges értékösszegsort! A becsült értékösszegek összege általában igen közel esik a tényleges értékösszegek összegéhez.

A könnyebb áttekinthetőség kedvéért gyakran megoszlási viszonyszámsort számítunk a gyakorisági eloszlásból vagy az értékösszegsorokból. Ezeket, megkülönböztetésül az eddigi abszolút soroktól, **relatív gyakorisági sornak** illetve **relatív értékösszegsornak** nevezzük. Jelölésük és számításuk a (21) - (23) képletek szerint történik.

Relatív gyakoriság:

$$g_i = \frac{f_i}{\sum_{i=1}^{k} f_i} = \frac{f_i}{N}. \tag{21}$$

Relatív értékösszeg:

$$Z_i = \frac{S_i}{\sum_{i=1}^{k} S_i}, \tag{22}$$

illetve ennek becslése:

$$\hat{Z}_i = \frac{f_i X_i}{\sum_{i=1}^{k} f_i X_i}. \tag{23}$$

26. példa

Számítsuk ki az előző példánk relatív sorait az osztályközös gyakorisági sorból!

A képletünk számításait könnyen elvégezhetjük az Excel segítségével is, ha egy régebbi mappánk már tartalmazza az osztályközös gyakorisági sort **C2-C6** cellákban az osztályközök felső határaival, **D2-D6** cellákban a gyakoriságokkal. Készítsünk egy összesen sort a **D7** cellába. A **Beszúrás** menü **Függvény…** almenüjével illeszthetünk be függvényt, itt válasszuk ki a Mat. és trigonom függvények közül a SZUM függvényt, majd Adattömbnek adjuk meg a D2:D6 tömböt (az összeg függvény előhívását közvetlenül a Σ ikon segítségével is megtehetjük). Az **E2** cellában a következő műveletet adjuk ki: =D2/D$7. A cella jobb alsó sarkának lehúzásával a többi cella eredménye könnyen megkapható. A százalékos írásmódot a **Formátum** menü **Cellák…** almenüjében állíthatjuk be (vagy közvetlenül a **%** ikon segítségével). Hasonló módon számíthatjuk ki a becsült relatív értékösszegsort is. Ehhez szükség lesz a osztályközepek értékeire (X_i), amelyekhez a (19) képletet kell az Excel segítségével a cellákban alkalmazni.

A kft dolgozóinak kereset szerinti relatív gyakorisági eloszlása és becsült relatív értékösszegsora

16. táblázat

Keresetek (ezer Ft)	X_i	g_i (%)	$\hat{Z}_i$ (%)
36,1 – 44,4	40,25	18,2	13,1
44,5 – 52,8	48,65	27,3	23,7
52,9 – 61,2	57,05	22,7	23,2
61,3 – 69,6	65,45	13,6	16,0
69,7 – 78,0	73,85	18,2	24,0
Összesen	–	100,0	100,0

Ha csak a relatív gyakorisági sor áll rendelkezésre, akkor is tudunk relatív értékösszegsort számítani a (24) képlettel.

$$\hat{Z}_i = \frac{f_i X_i}{\sum_{i=1}^{k} f_i X_i} = \frac{g_i N X_i}{\sum_{i=1}^{k} g_i N X_i} = \frac{N g_i X_i}{N \sum_{i=1}^{k} g_i X_i} = \frac{g_i X_i}{\sum_{i=1}^{k} g_i X_i} \qquad (24)$$

Kumulálás

Mennyiségi soroknál szoktuk alkalmazni a felfelé illetve lefelé **kumulálás** műveletét.

Ha K_i a C_i osztályhoz tartozó valamilyen adat, akkor a felfelé kumulált adatsor a következő összegek sorozata:

$$K_i' = \sum_{j=1}^{i} K_j \qquad i=1,2,...,k. \tag{25}$$

A lefelé kumulált adatsorozat:

$$K_i'' = \sum_{j=i}^{k} K_j . \tag{26}$$

27. példa

Határozzuk meg a kft kereseteinek felfelé és lefelé kumulált abszolút gyakorisági sorát a 13. táblázatban szereplő adatok alapján! Készítsük el a felfelé kumulált relatív becsült értékösszegsort is (lásd a 15. táblázat utolsó oszlopát)!

Egy kft kereseti adatainak kumulált sorai

17. táblázat

Keresetek (ezer Ft)	f_i	f_i'	f_i''	$\hat{Z}_i'$ (%)
36,1 – 44,4	4	4	22	13,1
44,5 – 52,8	6	10	18	36,8
52,9 – 61,2	5	15	12	60,0
61,3 – 69,6	3	18	7	76,0
69,7 – 78,0	4	22	4	100,0
Összesen	22	–	–	–

A lefelé és felfelé kumulált sorok közötti összefüggés:

$$K_{i-1}' + K_i'' = K_k' , \tag{27}$$

ahol $K_0' = 0$.

Figyeljük meg, hogy a kumulált gyakorisági sorok esetében:

$$f_1' = f_1; \quad f_k' = N; \quad f_1'' = N; \quad f_k'' = f_k .$$

Hasonló azonosságok érvényesek az abszolút értékösszegsorokra és a relatív sorokra is.

A mennyiségi sorok lehetséges fajtái a 6. ábrán vannak feltüntetve.

A mennyiségi sorok grafikus ábrázolása

A mennyiségi sorok közül elsősorban a gyakorisági sorokat és a kumulált gyakorisági sorokat szoktuk ábrázolni, az értékösszegsorokat kevésbé. A gyakorisági sorok szemléltető megjelenítésére háromféle grafikus ábrát használhatunk. Ezek az alábbiak:

- hisztogram,
- gyakorisági poligon,
- gyakorisági görbe.

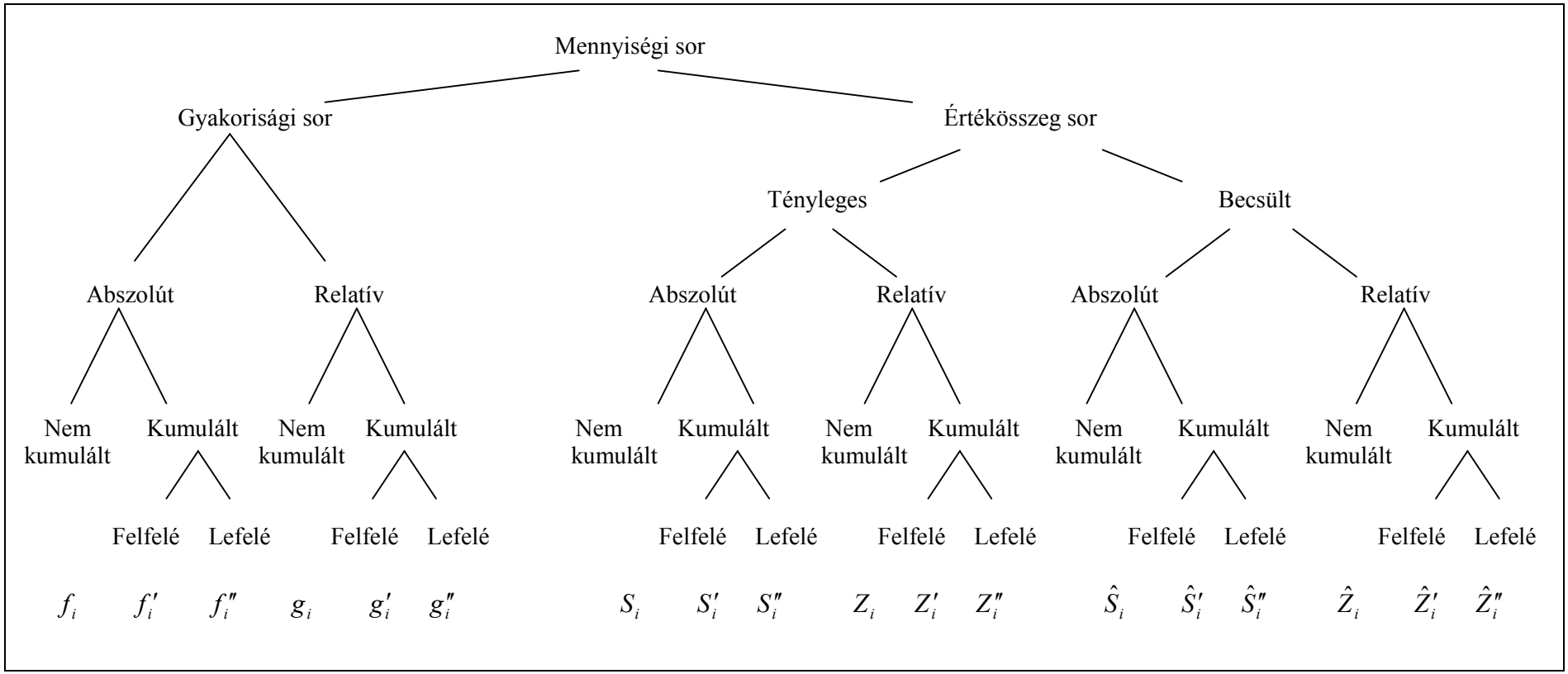
Hisztogramnak nevezzük a gyakorisági sorok (hézag nélküli) oszlopdiagram segítségével történő ábrázolását.

28. példa

Készítsünk a 23. példa adatai alapján hisztogramot az Excel program segítségével!

Válasszuk ki a **Hisztogram** menüpontot az **Eszközök**/**Adatelemzés…** ablakban, és adjuk meg Bemeneti tartománynak a 22 adatból álló tömbünket, amit a 23. példa megoldásakor az **A2-A23** cellákba írtunk. Rekesztartomány a **C2-C6** cellákban megadott felső osztályközhatárok tömbje legyen! (A rekesztartomány megadása nem kötelező, ekkor a program automatikusan hoz létre azonos hosszúságú osztályokat.)

A mennyiségi sorok fajtái

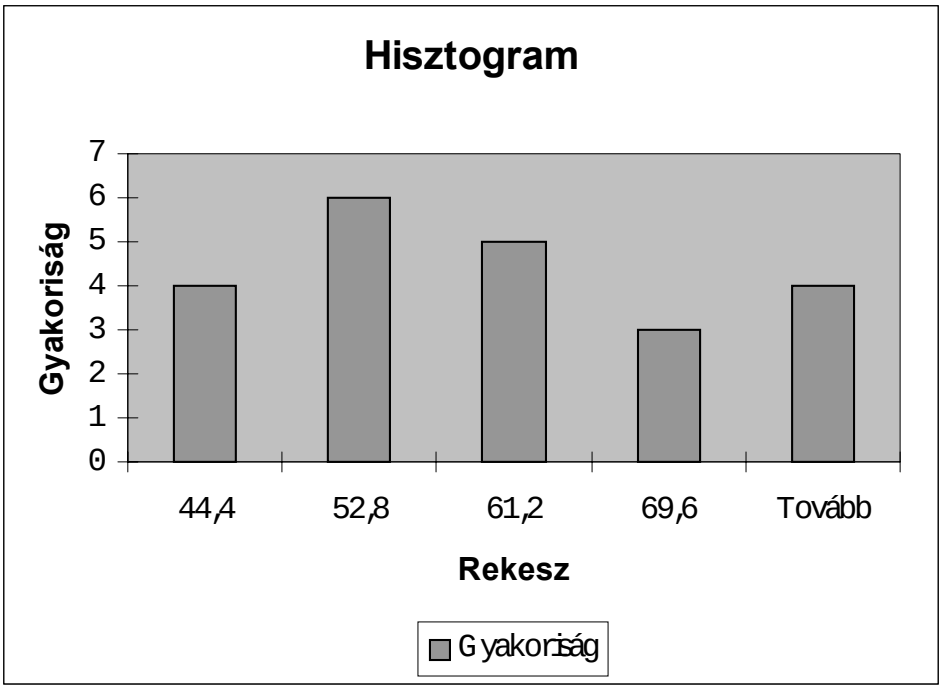


6. ábra

Kimeneti tartományként az adott munkalap egy szabad mezőjét megadva az eredmény az aktuális munkalapunkra kerül, egyébként egy újra. Kapcsoljuk be a Diagramkimenet jelölőnégyzetet! Eredményként osztályközös gyakorisági sort és hisztogramot kapunk. (Lásd a 7. ábrát.)

Az Excel munkalapjának részlete

Rekesz	*Gyakoriság*
44,4	4
52,8	6
61,2	5
69,6	3
Tovább	4



7. ábra

(Megjegyzés: az ismertetett eljárás a hisztogramok ábrázolásának csak egyike, a statisztikai gyakorlatnak megfelelő eljárást a következőkben ismertetjük.)

Minden síkdiagramra, így a hisztogramra is érvényes, hogy az egyes alapul vett síkidomok – itt téglalapok – területének kell arányosnak lennie az ábrázolni kívánt adat nagyságával. Ez azt jelenti, hogy eltérő osztályköz-hosszúságú osztályközös gyakorisági sorok

hisztogramon történő ábrázolásakor a gyakoriságokat azonos osztályköz-hosszúságúra kell átszámítani, és az ennek megfelelő arányos gyakoriságokat kell az *y* tengely mentén felmérni. (Mivel az oszlop alapjának megváltoztatására nincs mód, ezért a magasságot kell átszámítani.) Ha az eredeti gyakoriságokat mérnénk fel az *y* tengelyre, akkor a hosszabb osztályközök nagyobb súlyt kapnának, és az ábra torzítana.

Az átszámításnál vehetjük az $\frac{f_i}{h_i}$, illetve $\frac{g_i}{h_i}$ egységnyi osztályköz-hosszúságra eső gyakoriságokat vagy ezek valamilyen alkalmas többszörösét.

29. példa

Ábrázoljuk hisztogram segítségével Magyarország településeinek (Budapest nélkül) eloszlását népességnagyság-csoportok szerint! (Lásd a 18. táblázatot.)

A települések száma népességnagyság-csoportok szerint, 1998. január 1.

18. táblázat

Népességnagyság-csoport (fő)	A települések száma
– 499	1021
500 – 999	697
1 000 – 1 999	651
2 000 – 4 999	493
5 000 – 9 999	133
10 000 – 19 999	76
20 000 – 49 999	40
50 000 – 99 999	11
100 000 – 199 999	7
200 000 – 300 000	1
Összesen	3130

Forrás: Magyar statisztikai zsebkönyv ’98, KSH, Bp., 1999.

Mivel itt nem azonos hosszúságúak az osztályközök, számítsuk át a gyakoriságokat 500 fő osztályköz-hosszúságra arányítva! A számításokat végezzük el Excelben! Az **A2-A11** cellákba vigyük be az osztályköz-határokat! A **B2-B11** cellákba kerüljenek az osztályköz-

hosszúságok. A **C2:C11** tömb tartalmazza az eredeti, a **D2:D11** tömb pedig az átszámított gyakoriságokat.

A hisztorgram megrajzolásához jelöljük ki a **D2:D11** cellákat, majd a diagram varázslóban válasszuk az oszlopdiagramot! A 2. lépésben A kategóriatengely (X) feliratai: mezőbe írjuk be a következőt: =Munka1!$A$2:$A$11. Az elkészült oszlopdiagramban az **Adatsorok formázása…/Beállítások** ablakban tudjuk az oszlopok közötti távolságot beállítani, megszüntetni.

A települések száma népességnagyság-csoportok szerint

19. táblázat

Népességnagyság-csoport (fő)	h_i	f_i	500 fő osztályköz-hosszúságra eső gyakoriságok
– 499	500	1021	1021,00
500 – 999	500	697	697,00
1 000 – 1 999	1 000	651	325,50
2 000 – 4 999	3 000	493	82,17
5 000 – 9 999	5 000	133	13,30
10 000 – 19 999	10 000	76	3,80
20 000 – 49 999	30 000	40	0,67
50 000 – 99 999	50 000	11	0,11
100 000 – 199 999	100 000	7	0,04
200 000 – 300 000	100 000	1	0,01
Összesen	–	3130	–

Ezek alapján megrajzolható hisztogram a 8. ábrán látható.

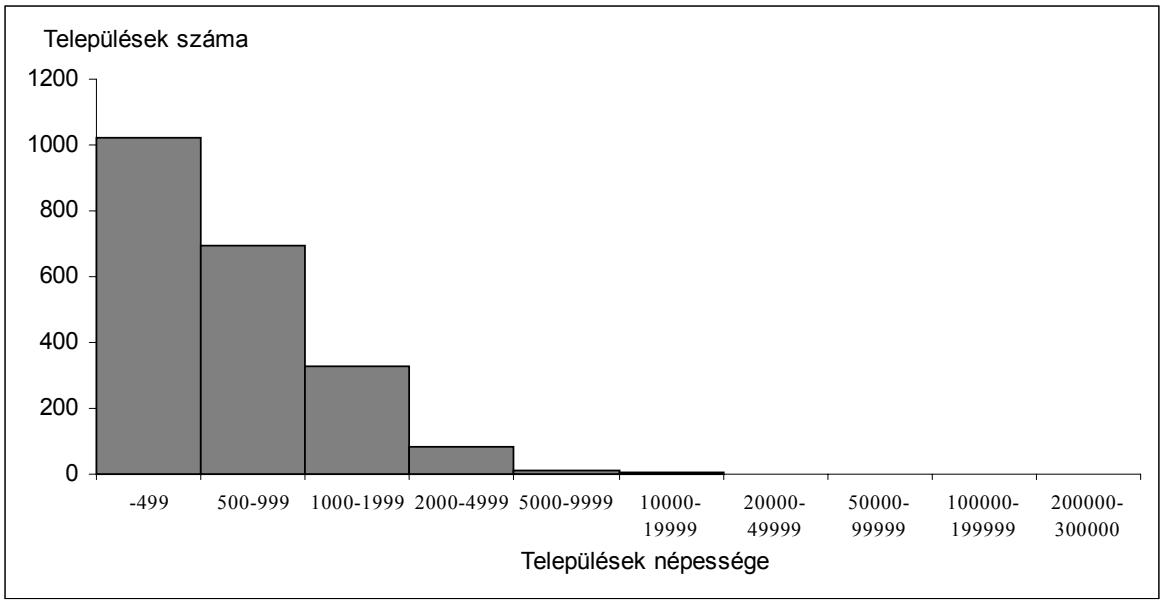
Ha a hisztogramot úgy alakítjuk ki, hogy az alatta levő terület 1 legyen, akkor az X változó **empirikus sűrűségfüggvényéhez** jutunk. A kumulált relatív gyakorisági sor oszlopdiagramja **empirikus eloszlásfüggvényt** ad.

A kumulált gyakorisági sorok vonaldiagramját **ogivának** nevezzük.

A gyakorisági sorok vonaldiagramon történő ábrázolását **gyakorisági poligonnak** nevezzük.

A gyakorisági poligon felrajzolásánál az osztályközepeknél mérjük fel a gyakoriságok pontjait (ez megfelel a hisztogram felső oszlopközepének).

A magyar települések népességszám szerinti eloszlásának grafikus ábrázolása hisztogram segítségével.



8. ábra

Nagy (végtelen) elemszámú sokaság, végtelenül kicsi osztályközökre osztásával a gyakorisági poligon folytonos görbébe megy át. Ezt hívjuk **gyakorisági görbének**. Ha úgy alakítjuk ki a léptéket, hogy a gyakorisági görbe alatti terület 1 legyen, akkor a valószínűségszámításból ismert sűrűségfüggvényhez jutunk. Később többször fogjuk a gyakorisági sorokat (empirikus eloszlásokat) valamilyen nevezetes (elméleti) eloszlással, mint matematikai modellel összevetni.

A mennyiségi sorok ismertetett grafikus ábrázolási lehetőségeit (összefoglalásként) a 20. táblázat tartalmazza.

Mennyiségi sorok nevezetes grafikus ábrázolási lehetőségei

20. táblázat

Ábra típusa	Gyakorisági sorok	Kumulált gyakorisági sorok
Oszlopdiagram	hisztogram	empirikus eloszlásfüggvény (csak relatív esetben)
Vonaldiagram	gyakorisági poligon	ogiva
Görbe	gyakorisági görbe	–

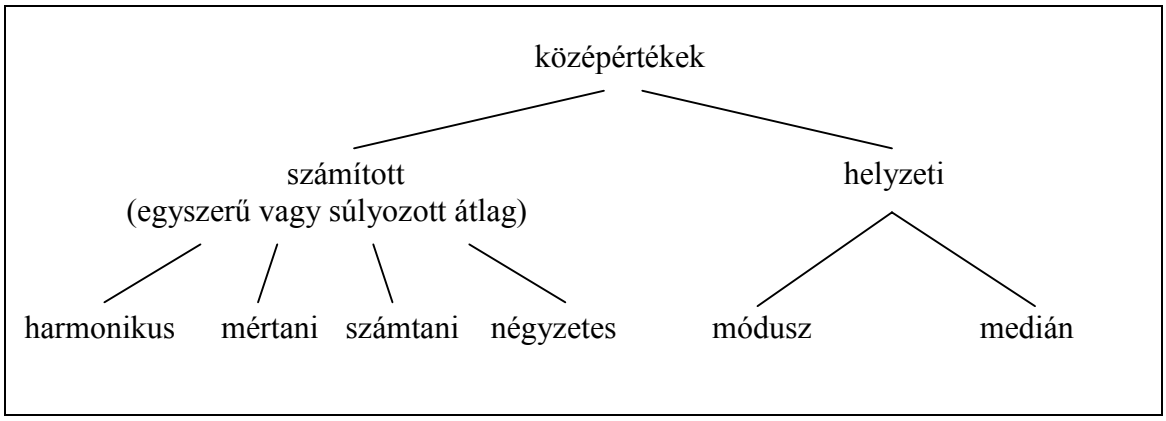
A továbbiakban azzal fogunk foglalkozni, hogy hogyan lehet az empirikus eloszlásokat tömören, egyetlen számba sűrített információt tartalmazó mutatószámokkal jellemezni.

3.2. Helyzet-mutatók, középértékek

A sokaság X ismérv szerinti eloszlásáról az empirikus eloszlásfüggvény és az empirikus sűrűségfüggyvény már sokat elárul. Tovább mélyítené ismereteinket, ha ennek a változónak a jellemzésére egy olyan számadatot keresnénk, amelynek a gyakorlatban jól értelmezhető és szemléletes tartalma van. A **középérték** olyan mutatószám, amely a sokaság valamely tulajdonságát egy számmal fejezi ki. Csak **homogén sokaságnak** (a vizsgált ismérv szempontjából hasonló jellegzetességeket mutató, részekre nem bontható sokaság) lehet jó jellemzője. Mértékegysége az ismérvértékkel azonos.

A középértékek két nagy csoportját szoktuk megkülönböztetni: a **számított** és a helyzeti **középértékeket**. A számított középértékek az **átlagok**. Ezek leggyakrabban használt fajtái a 9. ábrán vannak feltüntetve.

A legismertebb középértékek



9. ábra

Jó volna, ha a középértékek rendelkeznének az alábbi tulajdonságokkal:

- közepes helyzetet foglaljanak el;
- tipikus értékek legyenek;
- könnyen és egyértelműen kiszámíthatóak legyenek;
- jól és könnyen értelmezhetőek legyenek;
- a kiugró szélsőséges értékekre ne legyenek érzékenyek.

A fentiekből következik, hogy minden középértéknek az előforduló legkisebb és legnagyobb ismérvérték közé kell esnie.

Számított középértékek

Ezek az ismérvértékekkel való számszerű összefüggéssel adhatók meg.

Számtani (aritmetikai) átlag

Ez a leggyakrabban használt számított középérték. Az **egyszerű számtani (aritmetikai) átlag** a sokaság ismérvértékei összegének és az elemei számának hányadosa:

$$\bar{x}_a = \frac{\sum_{i=1}^{N} x_i}{N}. \qquad (28)$$

(Megjegyzés: az $\bar{x}$ szimbólum kiejtése „*x* átlag”.)

A továbbiakban $\bar{x}$-gal az átlagforma alkalmazására utalunk, az indexben szereplő jellel pedig annak fajtájára. Itt *a* az aritmetikai rövidítése. Mivel a számtani átlag a leggyakrabban alkalmazott átlagforma, az *a* indexnek a feltüntetését (ha nem okoz zavart) elhagyjuk, és a számtani átlagra egyszerűen $\bar{x}$-gal hivatkozunk.

Ha az egyes ismérvértékek többször is előfordulnak, akkor célszerűbb a **súlyozott átlagformát** használni. Ebben az egyes előforduló ismérvértékek gyakoriságait f_i-vel jelöljük. A súlyozott számtani átlag képlete:

$$\bar{x}_a = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{N} \qquad (29)$$

A fenti képlet jelölésrendszere az osztályközös gyakorisági sorokra emlékeztethet bennünket. Ez nem véletlen. A súlyozott számtani átlag alkalmazásának leggyakoribb esete az osztályközös gyakorisági sorból számított átlag. Ennél azt feltételeztük, hogy az egyes osztályközökbe eső sokasági elemek ismérvértékei az osztályközön belül egyenletesen oszlanak el, ezért azok helyettesíthetőek az osztályközéppel, így többször előforduló értékeket kell átlagolnunk.

30. példa

Számítsuk ki a 23. példában szereplő keresetek átlagát rangsorból és osztályközös gyakorisági sorból is! A számítást Excel segítségével végezzük!

Nyissuk meg azt a mappát amelyikbe előzően már bevittük az **A2-A23** cellákba a kereseteket. Álljunk az **A24** cellára és illesszük be ide a Statisztikai függvények közül az ÁTLAG függvényt. (Argumentumként az A2:A23 tömböt vigyük be.)

$$\bar{x} = 55{,}7 \text{ [ezer Ft]}$$

Az osztályközös gyakorisági sorból számított súlyozott átlag a (29) alapján kiszámítható.

$$\bar{x} = \frac{4 \cdot 40{,}25 + ... + 4 \cdot 73{,}85}{22} = 55{,}9 \text{ [ezer Ft]}$$

Az eltérés a két átlag között abból adódik, hogy csak megközelítően igaz az, hogy az eredeti értékek az osztályközökben egyenletesen oszlanak el.

A számtani átlag néhány jellegzetes tulajdonságának ismertetése következik.

Minden ismérvérték számtani átlaggal való helyettesítésekor elkövetett előjeles hibák kiegyenlítik egymást, vagyis az egyes ismérvértékek számtani átlagtól való eltéréseinek összege 0.

Nem súlyozott esetben

$$\sum_{i=1}^{N} \left(x_i - \bar{x}\right) = 0 .$$

Súlyozott esetben

$$\sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right) = 0 .$$

Minden ismérvérték számtani átlaggal való helyettesítésekor elkövetett hibák négyzetösszege minimális lesz; és fordítva: a számtani átlag az a konstans, amely esetén a négyzetes hiba minimális. Ez az ún. **négyzetes minimum tulajdonság**.

Nem súlyozott esetben

$$\sum_{i=1}^{N}(x_i - a)^2 \to \min \Leftrightarrow a = \bar{x}\,.$$

Súlyozott esetben

$$\sum_{i=1}^{k} f_i (x_i - a)^2 \to \min \Leftrightarrow a = \bar{x}\,.$$

A számtani átlagot a sokasághoz tartozó értékösszeg segítségével is ki tudjuk számítani:

$$\bar{x} = \frac{\sum_{i=1}^{k} S_i}{\sum_{i=1}^{k} f_i} = \frac{S}{N}. \qquad (30)$$

Lehetséges azonban, hogy az átlagolandó adatok összegének nincs statisztikai értelme, és ekkor a számtani átlagnak sincs értelme. Ekkor valamelyik másik átlagformát kell választani.

Bizonyos esetekben célszerű lehet az eredeti ismérvértékek helyett azok lineáris transzformált értékeivel dolgozni. Tekintsük a következő lineáris transzformációt:

$$y_i = \frac{x_i - A}{B} \qquad i=1,2,\ldots,N; \qquad (31)$$

ahol A és B tetszőleges konstansok, $B \neq 0$. Ekkor nyilvánvalóan $x_i = A + B \cdot y_i$.

A transzformált értékek számtani átlaga és az eredeti értékek számtani átlaga között a következő összefüggés áll fenn: $\bar{y} = \frac{\bar{x} - A}{B}$, illetve

$$\bar{x} = A + B \cdot \bar{y}\,. \qquad (32)$$

(Megjegyzés: az $\bar{y}$ az $\bar{x}$-hoz hasonlóan számítható ki.)

A fenti lineáris transzformáció segítségével például egyszerűbbé tehetjük a számtani átlag számítását osztályközös gyakorisági sorból. Ilyenkor B-t az osztályközök hosszával (h_i)

szoktuk egyenlővé tenni, *A*-t pedig úgy választjuk meg, hogy a számtani átlag közelébe essen.

31. példa

Számítsuk ki a 23. példa kereseteinek számtani átlagát osztályközös gyakorisági sorból, az ismérvértékek lineáris transzformációja mellett!

Legyen a (31) transzformációban A=57,05 és B=8,4. Az eredeti és a transzformált értékeket a 21. táblázat tartalmazza

A kft dolgozóinak kereset szerinti eloszlása

21. táblázat

Keresetek (ezer Ft)	x_i	y_i	f_i	$f_i \cdot y_i$
36,1 – 44,4	40,25	-2	4	-8
44,5 – 52,8	48,65	-1	6	-6
52,9 – 61,2	57,05	0	5	0
61,3 – 69,6	65,45	1	3	3
69,7 – 78,0	73,85	2	4	8
Összesen	–	–	22	-3

A (29) szerint

$$\bar{y} = \frac{\sum_{i=1}^{k} f_i y_i}{N} = \frac{-3}{22} = -0{,}136\,.$$

A (32) szerint

$$\bar{x} = A + B \cdot \bar{y} = 57{,}05 + 8{,}4 \cdot \frac{-3}{22} = 55{,}9 \text{ [ezer Ft].}$$

Mértani (geometriai) átlag

Mértani átlagot akkor használunk, ha az átlagolandó értékek szorzata értelmezhető.
Az ún. nem súlyozott vagy egyszerű mértani átlag a (33) képlettel definiált, illetve ennek

logaritmusát véve könnyen kiszámítható.

$$\bar{x}_g = \sqrt[N]{\prod_{i=1}^{N} x_i} \tag{33}$$

A súlyozott mértani átlagot a (34) szerint számíthatjuk ki.

$$\bar{x}_g = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} x_i^{f_i}} \tag{34}$$

(Megjegyzés: empirikus elemzéseknél a (34) képlet fenti alakjában túlcsordulás miatt gyakran nem alkalmazható, ezért kénytelenek vagyunk a logaritmusával számolni.)

32. példa

Számítsuk ki a 18. példa adatai alapján, hogy mekkora volt Magyarország népességének évi átlagos csökkenése 1991-1999 között!

Itt az évről évre bekövetkező népességcsökkenések mértékét kell átlagolnunk. Ezeket a láncviszonyszámok fejezik ki. A láncviszonyszámok összegének nincs statisztikai értelme, szorzatuk azonban a (10) képlet szerint bázisviszonyszámot ad. Ezért mértani átlagformát fogunk használni. A 18. példában kiszámított láncviszonyszámok (lásd a 9. táblázatot) súlyozott mértani átlaga a következő (az egyszerűség kedvéért logaritmusosan számolunk):

$$\ln \bar{x}_g = (\ln 0{,}998 + 4 \cdot \ln 0{,}997 + 3 \cdot \ln 0{,}996)/8 = -0{,}00326\,; \text{ így}$$

$$\bar{x}_g = 0{,}9967\,.$$

Figyeljük meg, hogy összesen 9 eredeti adatunk van, amiből pontosan 8 láncviszonyszám számítható, ezért 8 adat mértani átlagát számítjuk!

A mértani átlagot kiszámíthatjuk az Excel segítségével is. Nyissuk meg azt a mappát, amelyik az eredeti adatokból számított láncviszonyszámokat tartalmazza, majd a MÉRTANI.KÖZÉP(szám1;szám2;…) függvény segítségével számítsuk ki a keresett átlagot. Így pontosabb eredményt fogunk kapni, mert az Excel nagy pontossággal tárolja az adatokat és számol velük. A kapott eredmény:

$$\bar{x}_g = 0{,}9968\,.$$

Azt mondhatjuk tehát, hogy a magyar népesség 1991-1999 között évente átlagosan 0,32%-kal csökkent.

Vegyük észre, hogy a számítást egyszerűbben is el lehet végezni, ha nem csak a láncviszonyszámok adottak. A (10) képlet alapján tudjuk, hogy a láncviszonyszámok szorzata egy megfelelő bázisviszonyszámmal egyenlő:

$$\bar{l} = \sqrt[N-1]{\prod_{i=2}^{N} l_i} = \sqrt[N-1]{b_N} = \sqrt[N-1]{\frac{x_N}{x_1}}. \qquad (35)$$

Ezek alapján az éves átlagos csökkenés mértéke: $\sqrt[8]{0{,}975} = 0{,}9968$.

Harmonikus átlag

Harmonikus átlagot akkor használunk, ha az átlagolandó értékek reciprokaiból kapott összeg értelmezhető. A harmonikus átlag nem súlyozott képlete:

$$\bar{x}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}. \qquad (36)$$

A súlyozott harmonikus átlag képlete:

$$\bar{x}_h = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} \frac{f_i}{x_i}}. \qquad (37)$$

A harmonikus átlag számításának egy tipikus esete az, ha az átlagolandó adatok fordított intenzitási viszonyszámok.

33. példa

Egy kft három titkárnőt foglalkoztat, akik egy adott szöveget (önállóan, egymástól függetlenül) 3,2; 3,3; illetve 3,5 perc alatt gépelnek le. Számítsuk ki, hogy átlagosan mennyi idő alatt gépelnek le egy ilyen szöveget!

Mivel a rendelkezésünkre álló adatok fordított intenzitási viszonyszámok, átlagukat kizárólag a harmonikus átlag segítségével kaphatjuk meg. A (36) képlet szerint a titkárnőknek átlagosan

$$\bar{x}_h = \frac{3}{\frac{1}{3,2} + \frac{1}{3,3} + \frac{1}{3,5}} = 3,3287 \approx 3,33$$

perc szükséges az adott szöveg legépeléséhez.

Négyzetes (kvadratikus) átlag

A **négyzetes átlagot** akkor használjuk, ha nem akarjuk figyelembe venni az átlagolandó értékek előjelét, és azt akarjuk, hogy az átlag a szélsőségesen nagy értékekre érzékenyen reagáljon. A négyzetes átlag tipikus alkalmazása a szóródás mérésénél ismert, ezért ezzel ott foglalkozunk részletesebben. A négyzetes átlag nem súlyozott képlete a következő:

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}. \qquad (38)$$

A súlyozott négyzetes átlag képlete:

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i}}. \qquad (39)$$

Még egyszer hangsúlyozzuk, hogy az eddig említett átlagformák közötti választás nem önkényes. Mindig a statisztikailag, közgazdaságilag értelmezhető formát kell alkalmazni!

Még néhány megjegyzés az átlagokhoz:

– ugyanazon pozitív ismérvértékekből számított négyféle átlag között mindig az alábbi reláció áll fenn:

$$x_{\min} \le \bar{x}_h \le \bar{x}_g \le \bar{x}_a \le \bar{x}_q \le x_{\max}$$

(egyenlőség pontosan akkor áll fenn, ha minden átlagolandó érték egyforma);

– a súlyozott átlag értéke függ:
  a súlyarányoktól, tehát a súlyok relatív nagyságától és
  az átlagolandó értékek abszolút nagyságától;

– a különböző súlyozott átlagformákban az f_i abszolút gyakoriságok felcserélhetők a g_i relatív gyakoriságokkal.

A számtani átlag esetén például:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\frac{1}{N}\sum_{i=1}^{k} f_i x_i}{\frac{1}{N}\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} \frac{f_i}{N} x_i}{\sum_{i=1}^{k} \frac{f_i}{N}} = \frac{\sum_{i=1}^{k} g_i x_i}{\sum_{i=1}^{k} g_i}.$$

Helyzeti középértékek

A **helyzeti középértékek** az ismérvértékek közötti elhelyezkedésükkel adhatók meg. Ezek közül a két legismertebb **a módusz** és a **medián**, amelyeket a későbbiekben részletesen tárgyalunk.

Kvantilisek

Egy sokaságban megkereshetjük azt az ismérvértéket (osztópontot), amelynél az ismérvértékek fele, negyede, stb. kisebb, a többi pedig nagyobb értékű.

Ezek alapján az $x_{i/k}$ i-edik k-ad rendű **kvantilis** az a szám, amelynél az összes előforduló ismérvérték i/k-ad része kisebb, és $(1-i/k)$-ad része nagyobb ($k \geq 2$; $i=1,2,\ldots,k-1$).

(Megjegyzés: itt a k már nem a gyakorisági eloszlás osztályainak számát jelenti!)

A kvantilisek meghatározása egyúttal a sokaság egy osztályozását jelenti. Ezen osztályozás során egyenlő gyakoriságú osztályközöket kapunk.

A k db osztályból álló **kvantilis eloszlás** általános alakját a 22. táblázat tartalmazza.

Kvantilis eloszlás

22. táblázat

Ismérvváltozatokat tartalmazó osztályok	Előfordulások száma
$x_{min} - x_{1/k}$	N/k
$x_{1/k} - x_{2/k}$	N/k
⋮	⋮
$x_{(i-1)/k} - x_{i/k}$	N/k
⋮	⋮
$x_{(k-1)/k} - x_{max}$	N/k
Összesen	N

A fenti definíció alapján a kvantilisek nem mindig határozhatóak meg. Pontosan egyforma gyakoriságú osztályok ugyanis csak akkor képezhetők, ha:

- a sokaság elemeinek száma (N) az osztályok számának (k) egész számú többszöröse, és
- egyik kvantilis érték sem esik egybe valamelyik előforduló ismérvértékkel.

Azért, hogy a kvantilisek (és a kvantilis eloszlás) mindig meghatározhatóak legyenek, a továbbiakban a kvantilis fogalmának a következő módosított definícióját fogjuk használni:

az $x_{i/k}$ i-edik k-ad rendű kvantilis az a szám, amelynél az összes előforduló ismérvérték legalább i/k-ad része nem nagyobb, és legalább $(1-i/k)$-ad része nem kisebb ($k \geq 2$; $i=1,2,\ldots,k-1$).

A legtöbbet használt kvantiliseknek külön elnevezése van, ezeket tartalmazza a 23. táblázat.

A leggyakrabban használt kvantilisek neve és jelölése

23. táblázat

k	Neve	Jelölése
2	Medián	Me
3	**Tercilis**	T_1, T_2
4	**Kvartilis**	Q_1, Q_2, Q_3
5	**Kvintilis**	K_1, K_2, K_3, K_4
10	**Decilis**	D_1, D_2,…,D_9
100	**Percentilis**	P_1, P_2,…,P_{99}

Természetesen bizonyos kvantilis értékek egybeeshetnek. Például: $Me= Q_2=D_5=P_{50}$.

A kvantiliseket legegyszerűbben rangsorból lehet meghatározni. Az $x_{i/k}$ kvantilis a rangsor

$$s_{i/k} = \frac{i}{k}(N+1) \tag{40}$$

sorszámú tagja. Ez nem biztos, hogy egész szám, ezért kvantilisnek a következő értéket tekintjük:

$$x_{i/k} = x_{[s_{i/k}]} + \{s_{i/k}\} \cdot \left(x_{[s_{i/k}]+1} - x_{[s_{i/k}]}\right), \tag{41}$$

ahol $[s_{i/k}]$ az $s_{i/k}$ egész részét, míg $\{s_{i/k}\}$ az $s_{i/k}$ törtrészét jelenti.

A leggyakrabban előforduló kvantilis a medián, ezért most ezzel foglalkozunk részletesebben.

Ha N páratlan, akkor nyilvánvalóan létezik középső elem a rangsorban, amely sorszáma (40) szerint egész számmal egyenlő, így a medián értéke a rangsorban ezen a sorszámon szereplő elem ismérvértéke lesz.

Amennyiben N páros, akkor nem létezik középső elem a rangsorban, és ekkor a két középső elem ismérvértékének számtani átlagát tekintjük mediánnak. Ez megfelel a kvantilisek (rangsorból számított) általános képletének, hiszen:

$$Me = x_{[s_{1/2}]} + \{s_{1/2}\} \cdot \left(x_{[s_{1/2}]+1} - x_{[s_{1/2}]}\right) = x_{N/2} + \frac{1}{2}\left(x_{(N/2)+1} - x_{N/2}\right) = \frac{x_{N/2} + x_{(N/2)+1}}{2}. \tag{42}$$

A medián tehát az az ismérvérték, aminél az összes előforduló ismérvérték legalább fele nem nagyobb, és legalább fele nem kisebb.

A medián egy fontos tulajdonsága: minden ismérvérték mediánnal való helyettesítésekor elkövetett hibák abszolút értékben számított összege minimális lesz; és fordítva: a medián az a konstans, amely esetén az elkövetett előjel nélküli hibák összege minimális, azaz

$$\sum_{i=1}^{N} |x_i - a| \to \min \Leftrightarrow a = Me .$$

Természetesen szükség lehet a medián értékére akkor is, ha csak osztályközös gyakorisági sor áll rendelkezésünkre. Ekkor ezt is becsléssel állapítjuk meg.

Először azt határozzuk meg, hogy melyik osztályközbe esik a medián. Ezt teljes pontossággal meg tudjuk adni. A medián bizonyosan abban az *Me* sorszámú osztályban van, amelyre már igaz, hogy: $f'_{Me} \geq \frac{N}{2}$. Az ezt megelőző osztályközbe ugyanis $f'_{Me-1} < \frac{N}{2}$ elem esik, míg az ezt követőbe $\left(N - f'_{Me}\right) < \frac{N}{2}$ elem. Ezért a mediánt legegyszerűbben a mediánt tartalmazó osztály osztályközepével becsülhetjük. Ezt a durva becslést **nyers mediánnak** nevezzük. Lehetséges azonban egy finomabb becslést adni, amely a mediánt tartalmazó osztályköz hosszának egy arányos osztását jelenti.

A medián osztályközös gyakorisági sorból történő kiszámítására a (43) képletet fogjuk használni.

$$\hat{Me} = X_{Me,0} + \frac{\frac{N}{2} - f'_{Me-1}}{f_{Me}} \cdot h_{Me}, \qquad (43)$$

ahol

Me annak a legelső osztálynak a sorszáma amelyre $f'_{Me} \geq \frac{N}{2}$;

$X_{Me,0}$: a mediánt tartalmazó osztályköz alsó határa;

f'_{Me-1} : a medián osztályközét megelőző osztályközhöz tartozó felfelé kumulált gyakoriság;

f_{Me} : a mediánt tartalmazó osztályköz gyakorisága;

h_{Me} : a mediánt tartalmazó osztályköz hossza.

$X_{Me,0}$ tulajdonképpen valódi határt jelöl, de a gyakorlatban, mivel ez nem okoz nagy tévedést, gyakran a közölt határral számolunk.

A becslés során természetesen abból indultunk ki, hogy az adatok az osztályközökben egyenletes eloszlásúak.

A többi kvantilis (osztályközös gyakorisági sorból történő) becslése a mediánhoz hasonló módon a (44) képlettel történik:

$$\hat{x}_{i/k} = X_{i/k,0} + \frac{\frac{i}{k}N - f'_{(i/k)-1}}{f_{i/k}} \cdot h_{i/k}, \tag{44}$$

ahol *i*/*k* annak a legelső osztálynak a sorszáma, amelyre $f'_{i/k} \geq \frac{i}{k}N$.

(Megjegyzés: a kvantilisek osztályközös gyakorisági sorból történő becslésére ennél finomabb eljárás is ismert.)

A fenti képletek alkalmazhatóak akkor is, ha az abszolút gyakoriságok nem ismertek, csak a relatív gyakoriságok által adott eloszlás.

$$\hat{x}_{i/k} = X_{i/k,0} + \frac{\frac{i}{k}N - f'_{(i/k)-1}}{f_{i/k}} \cdot h_{i/k} = X_{i/k,0} + \frac{\frac{i}{k}N - N \cdot g'_{(i/k)-1}}{N \cdot g_{i/k}} \cdot h_{i/k} =$$

$$= X_{i/k,0} + \frac{\frac{i}{k} - g'_{(i/k)-1}}{g_{i/k}} \cdot h_{i/k},$$

ahol *i*/*k* annak a legelső osztálynak a sorszáma, amelyre $g'_{i/k} \geq \frac{i}{k}$.

A medián képlete osztályközös relatív gyakorisági sorból szintén levezethető.

$$\hat{Me} = X_{Me,0} + \frac{\frac{1}{2} - g'_{Me-1}}{g_{Me}} \cdot h_{Me},$$

ahol *Me* annak a legelső osztálynak a sorszáma, amelyre $g'_{Me} \geq \frac{1}{2}$.

Az előző képletekben a *g* relatív gyakoriságokat természetesen tizedes tört alakjukban kell használni.

34. példa

Számítsuk ki a 23. példában szereplő keresetek mediánját és kvartiliseit rangsorból és az osztályközös gyakorisági sorból is!

A 22 elemből álló sokaság mediánjának sorszáma a (40) képlet szerint: $s_{1/2} = \frac{1}{2} \cdot 23 = 11{,}5$.

A rangsor két középső eleme a 11. és a 12. elem. Ezek átlaga: $\frac{53{,}3 + 53{,}9}{2} = 53{,}6$. Azt mondhatjuk tehát, hogy a kft dolgozóinak fele (11 fő) 53 600 Ft-nál kevesebbet, míg fele (11 fő) 53 600 Ft-nál többet keres.

A kvartilisek közül a második a mediánnal egyenlő, így már csak az ún. **alsó** és **felső kvartilist** kell kiszámítani.

A sorszámok a (40) alapján: $s_{1/4} = \frac{1}{4} \cdot 23 = 5{,}75; \quad s_{3/4} = \frac{3}{4} \cdot 23 = 17{,}25$.

Az alsó és a felső kvartilis a (41) képlet alapján:

$$Q_1 = 47{,}1 + 0{,}75 \cdot (48{,}0 - 47{,}1) = 47{,}775 \approx 47{,}8;$$

$$Q_3 = 63{,}2 + 0{,}25 \cdot (68{,}6 - 63{,}2) = 64{,}550 \approx 64{,}6.$$

Az alsó kvartilis értelmezése: a kft dolgozói közül az első ötnek a keresete 47 800 Ft-nál kisebb, míg a többieknek ettől nagyobb. Értelmezze a felső kvartilis értékét!

Excelben a kvartiliseket a KVARTILIS(tömb;kvart) függvény segítségével („egy adathalmaz negyedszintjét”[1)]) számíthatjuk ki. Tömbként adjuk meg a 22 adatból álló cellatartományunkat, a kvart helyére pedig 1, 2 vagy 3 értéket adjunk attól függően, hogy melyik kvartilist akarjuk kiszámítani.

Az Excelben a többi kvantilist a PERCENTILIS(tömb;k) függvény segítségével („egy tartományban található értékek *k*-adik percentilisét, azaz százalékosztályát”[1)]) tudjuk

[1)] Excel szerinti eredeti értelmezés.

kiszámítani.[2)] Itt k értéke 0 és 1 közé eshet; ezért az alsó kvartilist pl. 0,25-ös érték megadásával kapjuk.

Most az osztályközös gyakorisági sorból számítjuk ki a kvartiliseket. Ehhez a 24. táblázat adataira van szükségünk.

A kft dolgozóinak kereset szerinti megoszlása

24. táblázat

Keresetek (ezer Ft)	f_i	f_i'	g_i' (%)
36,1 – 44,4	4	4	18,2
44,5 – 52,8	6	10	45,5
52,9 – 61,2	5	15	68,2
61,3 – 69,6	3	18	81,8
69,7 – 78,0	4	22	100,0
Összesen	22	–	–

A felfelé kumulált relatív gyakoriságokból közvetlenül megállapítható, hogy Q_1 a második, Q_2=Me a harmadik, míg Q_3 a negyedik osztályban van. A (44) képlet szerint:

$$\hat{Q}_1 = 44{,}5 + \frac{\frac{1}{4}\cdot 22 - 4}{6}\cdot 8{,}4 = 46{,}60 \approx 46{,}6\,;$$

$$\hat{Me} = \hat{Q}_2 = 52{,}9 + \frac{\frac{22}{2} - 10}{5}\cdot 8{,}4 = 54{,}58 \approx 54{,}6\,;$$

$$\hat{Q}_3 = 61{,}3 + \frac{\frac{3}{4}\cdot 22 - 15}{3}\cdot 8{,}4 = 65{,}50 \approx 65{,}5\,.$$

A felső kvartilis becslésének értelmezése: a kft első 17 dolgozójának keresete nem több mint 65 500 Ft, illetve a többiek keresete nem kevesebb mint 65 500 Ft. Értelmezze az első két kvartilis becsült értékét is!

[2)] Megjegyzés: a különböző értelmezések miatt az Excel szerinti eredmények nem azonosak a (41) képlet szerinti eredményekkel!

Az előző három kvartilis becsült érték, így eltérnek a rangsorból számított tényleges értékektől.

Módusz

A módusz szintén a helyzeti középértékek közé tartozó mutató. A tipikus, a divatos, a leginkább jellemző értéket mutatja. E körül sűrűsödnek, tömörülnek az ismérvértékek. (Megjegyzés: nem tévesztendő össze a számtani átlaggal, amely nem minden esetben rendelkezik ezekkel a tulajdonságokkal!)

Diszkrét változó esetén a módusz a leggyakrabban előforduló ismérvérték, míg folytonos változó esetén a gyakorisági görbe maximumhelye.

Míg az eddig ismertetett középértékek mindig egyértelműen meghatározhatóak voltak, addig a módusz nem biztos, hogy mindig létezik (például nem súlyozott diszkrét típusú mennyiségi sor esetén), és ha létezik is, akkor is csak nagy bizonytalansággal határozható meg, hiszen a gyakorisági görbe általában pontosan nem ismert.

A rangsorból történő meghatározásakor a leggyakrabban előforduló értéket tekintjük módusznak.

Osztályközös gyakorisági sor esetén a módusz pontos értékét közvetlenül nem tudjuk kiszámítani, ezért becsülnünk kell. Először meghatározzuk a móduszt tartalmazó osztályt, amit **modális osztálynak** nevezünk. Ez az az osztály amelybe arányosan a legtöbb ismérvérték tartozik. Itt tömörülnek, itt sűrűsödnek az ismérvértékek. Ez azonos hosszúságú osztályközök esetén a legnagyobb gyakoriságú osztályköz. Eltérő hosszúságú osztályközök esetén azonban egységnyi hosszúságúra (vagy ennek konstansszorosára) kell átszámolni a gyakoriságokat, és ezek között kell keresni a maximális értéket.

A nyers mediánhoz hasonóan értelmezzük a **nyers móduszt** is. Nyers módusznak a modális osztály osztályközepét tekintjük. Természetesen a módusz becslésére is ismert finomabb módszer.

A módusz osztályközös gyakorisági sorból történő becslésére a (45) képletet használjuk.

$$\hat{Mo} = X_{Mo,0} + \frac{f_{Mo} - f_{Mo-1}}{\left(f_{Mo} - f_{Mo-1}\right) + \left(f_{Mo} - f_{Mo+1}\right)} \cdot h_{Mo}, \qquad (45)$$

ahol *Mo* a modális osztályt jelenti.

Előfordulhat, hogy a rangsor alapján számított módusz nem esik a rangsorból készített osztályközös gyakorisági sor modális osztályába. Hasonló jelenség a medián és az átlagok között nem lehetséges.

35. példa

Számítsuk ki a 23. példa kereseteinek móduszát!

A rangsorból számítva elméletileg Mo=51 000 Ft adódna, de mivel ez csak kétszer szerepel, és egyébként is csak 22 adatunk van, itt a módusz nem kap értelmet. Ezért alkalmazzuk a (45) képletet. A 13. táblázat adataiból következik, hogy a modális osztály a második, így a módusz értéke:

$$\hat{Mo} = 44{,}5 + \frac{6-4}{(6-4)+(6-5)} \cdot 8{,}4 = 50{,}10 \text{ [ezer Ft].}$$

Ezt úgy értelmezhetjük, hogy a kft dolgozóinak keresetei 50 100 Ft körül sűrűsödnek, tömörülnek, azaz ez tekinthető tipikus (de nem átlagos!) keresetnek.

3.3. Szóródási mutatók

Az előző pontban ismertetett középértékekkel egy gyakorisági eloszlás tömör, számszerű jellemzését adtuk. A középértékekkel meghatároztuk, hogy az ismérvértékek a számegyenesen körülbelül hol helyezkednek el. Az elhelyezkedés azonban csak egyike a mennyiségi sorok jellegzetes tulajdonságainak, mert még számos más tulajdonság is definiálható.

36. példa

Az 1 és 99 ismérvértékeket tartalmazó kételemű sokaság számtani átlaga 50, míg
a 49 és az 51 ismérvértékekkel rendelkező kételemű sokaság számtani átlaga szintén 50.

Ebből a példából is látszik, hogy azonos átlagú sokaságok között nagy különbség lehet abból a szempontból, hogy azok ismérvértékei mennyire térnek el a középértéktől, illetve egymástól.

Szóródásnak nevezzük az ismérvértékek egymáshoz viszonyított különbözőségét, vagy a sokaság egészét jellemző átlagos értéktől való eltérését.

Hasonlóan a helyzet-mutatókhoz, a szóródásnak is vannak mérőszámai. A szóródás abszolút mutatóinak mértékegysége megegyezik a számítás alapjául szolgáló ismérvértékek mértékegységével. A szóródás mértékének jellemzésére relatív mutatókat is használunk, amelyek valamilyen kitüntetett sokasági mutatóhoz viszonyítanak.

Terjedelem-mutatók

A **szóródás terjedelme** (vagy röviden: **terjedelem**) a legnagyobb és a legkisebb ismérvérték közötti különbség.

$$R = x_{\max} - x_{\min} \tag{46}$$

Mivel R csak a legnagyobb és a legkisebb értéktől függ, nagyon érzékeny a kiugróan magas vagy alacsony értékre, ezért helyette a vizsgált jelenség valós szóródásának kvantifikálására jobban alkalmas például az ún. **interkvantilis terjedelem** mutató. Az interkvantilis terjedelem a két szélső kvantilis érték közötti különbség meghatározásán alapul.

Átlagos abszolút különbség

Átlagos abszolút különbségnek (GINI-együttható) nevezzük az összes lehetséges módon párba állított ismérvértékek különbségeinek abszolút értékéből számított számtani átlagát. A GINI-együttható nem súlyozott képlete:

$$G = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right|}{N(N-1)}. \tag{47}$$

A súlyozott képlet:

$$G = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} f_i f_j \left|x_i - x_j\right|}{N(N-1)}. \tag{48}$$

A (47)-(48) képletek nevezőjében azért szerepel $N(N\text{-}1)$, mert az ismérvértékek önmaguktól vett eltéréseit, amik természetesen 0-k, nem vesszük be a számításba. Az átlag és a GINI-együttható között fennáll a következő egyenlőtlenség:

$$0 \le G \le 2\bar{x}.$$

Az átlagos abszolút különbséget leggyakrabban a koncentráció elemzésénél használjuk. A koncentráció fogalmát a későbbiekben részletesebben tárgyaljuk.

Átlagos abszolút eltérés

Átlagos abszolút eltérésnek nevezzük az ismérvértékek számtani átlagtól vett különbségeinek abszolút értékéből számított számtani átlagát.

Az átlagos abszolút eltérés nem súlyozott képlete:

$$\delta = \frac{\sum_{i=1}^{N}\left|x_i - \bar{x}\right|}{N}. \tag{49}$$

A súlyozott átlagos abszolút eltérés képlete:

$$\delta = \frac{\sum_{i=1}^{k} f_i \left| x_i - \bar{x} \right|}{\sum_{i=1}^{k} f_i}. \qquad (50)$$

Az átlagos abszolút eltérés tehát azt fejezi ki, hogy az ismérvértékek átlagosan mennyivel térnek el az átlaguktól.

Az átlagos abszolút eltérést Excelben az ÁTL.ELTÉRÉS(szám1;szám2;…) statisztikai függvény segítségével tudjuk kiszámítani.

Szórás

A **szórás** a szóródás legfontosabb mérőszáma. Szórásnak nevezzük az ismérvértékek számtani átlagtól vett különbségeinek négyzetes átlagát.

A szórás nem súlyozott képlete:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2}. \qquad (51)$$

A súlyozott képlet:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{x} \right)^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i} - \bar{x}^2}. \qquad (52)$$

(Hasonlítsa össze az (51)-(52) képleteket a (38)-(39) képletekkel!)

Megjegyzés: empirikus elemzéseknél az előző két képlet jobb oldalával célszerű elvégezni a számításokat.

A szórás szintén azt fejezi ki, hogy az ismérvértékek átlagosan mennyivel térnek el az átlaguktól, de mivel négyzetes átlagot használ, hangsúlyosabban emeli ki a nagyobb eltéréseket. Mivel a statisztikában a szórás négyzetére is szükségünk van, ezt a négyzetes mutatót **szórásnégyzetnek** vagy **varianciának** nevezzük.

A négyzetes átlag tulajdonságából adódik a következő azonosság, amely néha megkönnyítheti a szórás kiszámítását:

$$\sigma = \sqrt{\bar{x}_q^2 - \bar{x}^2} \ . \tag{53}$$

Az Excelben a szórást a SZÓRÁSP(szám1;szám2;…); míg a varianciát a VARP(szám1;szám2;…) statisztikai függvény segítségével tudjuk kiszámítani.

A szórás két szélső korlátjára (ha $x_i \geq 0$) felírhatjuk a következő összefüggést:

$$0 \leq \sigma \leq \bar{x}\sqrt{N-1} \ .$$

Az alsó korlát $\sigma = 0$ minden esetben fennáll, ha $x_i = x$ (i=1,2,…,N).

A felső korlát $\sigma = \bar{x}\sqrt{N-1}$ csak akkor áll fenn, ha $x_i = 0$ (i=1,2,…,*N-1*) és $x_N = N \cdot \bar{x}$.

Relatív szórás

A **relatív szórást** pozitív ismérvértékekre értelmezzük. Relatív szórásnak nevezzük a szórás és a számtani átlag arányát.

$$\mathrm{v} = \frac{\sigma}{\bar{x}} \tag{54}$$

(Megjegyzés: a fenti képlet értékét legtöbbször százalékban szoktuk kifejezni.)

Ez a mutató az ismérvértékek átlagtól vett átlagos relatív eltérését adja meg.

A relatív szórás a szóródás relatív mutatója, így mértékegység nélküli, értéke százalékos formában is megadható. Ez a dimenzió nélküli mutató alkalmas különböző mértékegységű ismérvek szóródásának összehasonlítására.

Hasonlóan a szóráshoz, a relatív szórásnak is megadhatjuk alsó és felső korlátját:

$$0 \leq \mathrm{v} \leq \sqrt{N-1} \ .$$

Egyenlőséget a szórásnál ismertetett feltételek mellet kapunk.

37. példa

Számítsuk ki a 23. példa kereseteinek szóródási mutatóit a rangsorból és az osztályközös gyakorisági sorból is!

A rangsorból kiszámított szóródási mutatók:

terjedelem: R=78,0-36,1=41,9 [ezer Ft];

átlagos abszolút eltérés: $\delta = \frac{|36{,}1 - 55{,}7| + |42{,}0 - 55{,}7| + ... + |78{,}0 - 55{,}7|}{22} = 9{,}26$ [ezer Ft];

szórás: $\sigma = \sqrt{\frac{(36{,}1 - 55{,}7)^2 + (42{,}0 - 55{,}7)^2 + ... + (78{,}0 - 55{,}7)^2}{22}} = 11{,}38$ [ezer Ft];

relatív szórás: $v = \frac{11{,}4}{55{,}7} = 0{,}205$; illetve 20,5%.

Az osztályközös gyakorisági sorból kiszámított mutatók:

átlagos abszolút eltérés:

$$\delta = \frac{4 \cdot |40{,}25 - 55{,}9| + 6 \cdot |48{,}65 - 55{,}9| + ... + 4 \cdot |73{,}85 - 55{,}9|}{22} = 9{,}65 \text{ [ezer Ft];}$$

szórás:

$$\sigma = \sqrt{\frac{4 \cdot (40{,}25 - 55{,}9)^2 + 6 \cdot (48{,}65 - 55{,}9)^2 + ... + 4 \cdot (73{,}85 - 55{,}9)^2}{22}} = 11{,}41 \text{ [ezer Ft];}$$

relatív szórás:

$$v = \frac{11{,}41}{55{,}9} = 0{,}204 \text{ ; illetve 20,4\%.}$$

Értelmezze az előbbi mutatók értékeit!

A következőkben ismertetjük az átlagos abszolút eltérés és a szórás közötti összefüggést.

Ugyanazon adatokból számított szórás mindig nagyobb (esetleg egyenlő) az átlagos abszolút eltéréstől: $\delta \le \sigma$. A reláció abból következik, hogy a δ az $|x_i - \bar{x}|$ értékek számtani, míg σ ugyanezen értékek négyzetes átlagának tekinthető, hiszen $|x_i - \bar{x}|^2 = (x_i - \bar{x})^2$. (Lásd a számított átlagok közötti nevezetes összefüggést a 3.2. fejezetben!) Az egyenlőség $N = 2$ esetben mindig igaz, valamint $N > 2$ esetén, ha az ismérvértékek egyenlőek.

Megjegyzés: ha minden adat egyforma, akkor az összes eddig ismertetett szóródási mutató értéke 0-val egyenlő!

Standardizált változó

Végezzük el a (31)-es lineáris transzformációt az eredeti adatainkon a következő módon:

$$y_i = \frac{x_i - \bar{x}}{\sigma}. \qquad (55)$$

Az (55) képlet alkalmazásával kapott új változókat **standardizált változóknak** nevezzük.

Ezek fontos tulajdonsága: a standardizált változók átlaga 0, míg szórása 1 egységnyi, azaz

$$\bar{y} = 0 \quad \text{és} \quad \sigma_y = 1.$$

A standardizált változók azt mutatják, hogy az eredeti változók hány szórásnyival térnek el az átlaguktól, ezért szóródási mutatóknak is tekinthetők. (Megjegyzés: ezekkel a változókkal majd jóval részletesebben foglalkozunk a második kötetben.)

38. példa

Standardizáljuk a 23. példa adatait.

A 37. példa adatai alapján a keresetek átlaga 55,90 [ezer Ft]; szórása pedig 11,41 [ezer Ft].

A kft dolgozóinak kereset szerinti eloszlása

25. táblázat

Keresetek (ezer Ft)	x_i	y_i	f_i
36,1 – 44,4	40,25	-1,37	4
44,5 – 52,8	48,65	-0,64	6
52,9 – 61,2	57,05	0,10	5
61,3 – 69,6	65,45	0,84	3
69,7 – 78,0	73,85	1,57	4
Összesen	–	–	22

A 25. táblázat standardizált változójának átlaga és szórása:

$$\bar{y} = \frac{4 \cdot (-1{,}37) + 6 \cdot (-0{,}64) + ... + 4 \cdot 1{,}57}{22} = 0\,;$$

$$\sigma_y = \sqrt{\frac{4 \cdot (-1{,}37 - 0)^2 + 6 \cdot (-0{,}64 - 0)^2 + ... + 4 \cdot (1{,}57 - 0)^2}{22}} = 1\,.$$

Hogyan értelmezhetjük például az $y_5 = 1{,}57$ értéket? A 73 850 Ft-os kereset az 55 900 Ft-os átlagkeresettől 1,57 szórásnyival (tehát, nem forinttal, nem is százalékkal) nagyobb.

Az (55) szerinti standardizálást az Excelben a NORMALIZÁLÁS(x;középérték;szórás) függvény segítségével végezhetjük el.

3.4. A koncentráció vizsgálata

Ha egy sokaságban a teljes értékösszeg jelentős része néhány sokasági egységre összpontosul, akkor **koncentrációról** beszélünk. A koncentráció a szóródás egyfajta megnyilvánulása.

A fenti definíciónak megfelelően a koncentráció foka a kumulált relatív gyakoriság és a kumulált relatív értékösszeg összehasonlításával állapítható meg. Ennek ábrázolására egy speciális grafikus ábrát fogunk használni: a **LORENZ-görbét**. Ez egy egységnyi oldalú négyzetben elhelyezett vonaldiagram, ahol a vízszintes tengelyen a g_i', míg a függőleges tengelyen a Z_i' szerepel. (Lásd a 10. ábrát!)

A LORENZ-görbe a következő pontok által meghatározott görbe:

$$(0,0);\ \left(g_i', Z_i'\right);\ (1,1) \qquad\qquad i=1, 2,\ldots, N\text{-}1.$$

(Megjegyzés: ezek szerint a görbe csak a négyzet alsó háromszögében helyezkedhet el, mint ahogy az a 10. ábrán látható.)

A koncentráció hiányát, azaz az egyenletes eloszlást az jelzi, ha a LORENZ-görbe egybeesik a négyzet bal alsó sarkából a jobb felső sarkába tartó átlójával. A koncentráció nagyságát a LORENZ-görbe és a négyzet átlója közötti terület, a **koncentrációs terület** (t_c) mutatja. A koncentráció mértékének meghatározására ezért a koncenrtrációs terület és a négyzet átlója alatti háromszög területének hányadosát használjuk mutatószámként.

$$L = \frac{t_c}{1/2} = 2 \cdot t_c \tag{56}$$

Ez a mérőszám azonban (57) szerint is felírható.

$$L = \frac{G}{2 \cdot \bar{x}} \tag{57}$$

Megjegyzés: a LORENZ-görbe alapján általában csak szubjektív döntést tudunk hozni, míg az (57) alapján egyértelműen számszerűsíthető a koncentráció mértéke.

A koncentráció vizsgálatára az említetteken kívül még számos eszköz ismert a statisztikai irodalomban.

39. példa

A 22. példa adatait felhasználva rajzoljuk meg a LORENZ-görbét és számítsuk ki az L mutató értékét az (57) képlet alapján!

Első lépésként a biztosító cégeket a díjbevételek alapján rangsorba rendezzük. A szükséges mellékszámításokat a 26. táblázat tartalmazza.

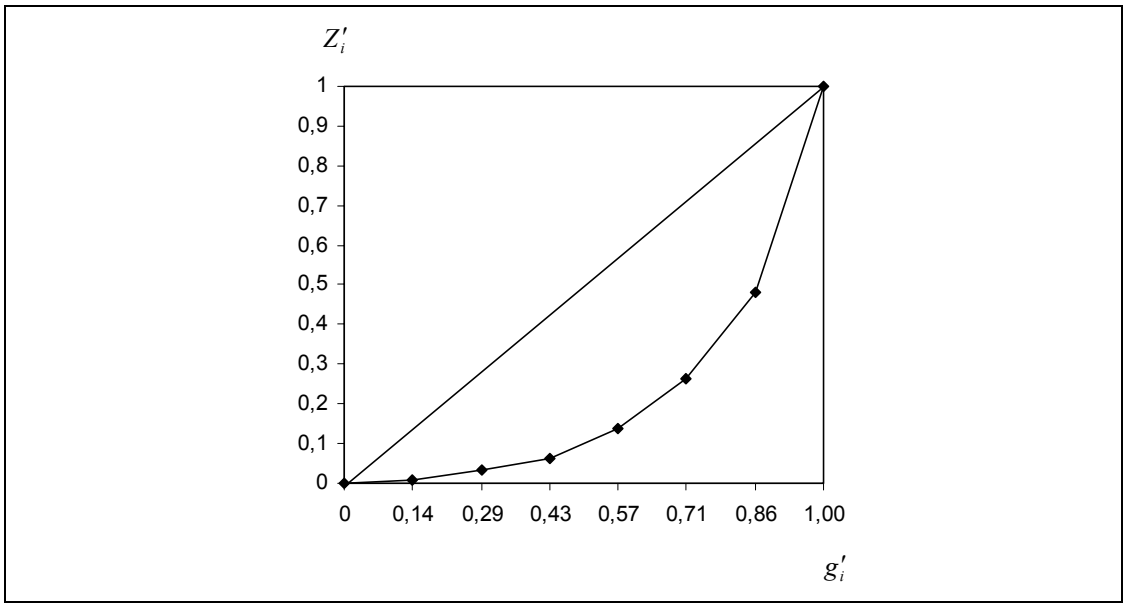
A kötelező gépjármű-biztosítások piacának szereplői

26. táblázat

Biztosítók	g_i' (%)	Z_i' (%)
Közlekedési Biztosító Egyesület	14,29	0,64
Argosz	28,57	3,36
Axa Colonia	42,86	6,40
OTP-Garancia	57,14	13,73
ÁB-Aegon	71,43	26,35
Generali-Providencia	85,71	48,30
Hungária	100,00	100,00

A 26. táblázat adatainak megfelelő LORENZ-görbe a 10. ábrán látható.

A kötelező gépjármű-biztosítások piacának koncentrációját jellemző LORENZ-görbe



10. ábra

A 10. ábrán látható négyzet átlója és a LORENZ-görbe közötti terület az (56) képletben szereplő t_c koncentrációs terület.

A koncentráció mértékét jellemző L mutató meghatározásához szükségünk van a GINI-együtthatóra, amelynek kiszámításához szükséges mellékszámításokat a 27. táblázat tartalmazza.

A GINI-együttható kiszámítását segítő munkatábla

27. táblázat

	100 207	428 145	478 922	1 154 755	1 986 164	3 455 826	8 138 255
100 207	0	327 938	378 715	1 054 548	1 885 957	3 355 619	8 038 048
428 145	327 938	0	50 777	726 610	1 558 019	3 027 681	7 710 110
478 922	378 715	50 777	0	675 833	1 507 242	2 976 904	7 659 333
1 154 755	1 054 548	726 610	675 833	0	831 409	2 301 071	6 983 500
1 986 164	1 885 957	1 558 019	1 507 242	831 409	0	1 469 662	6 152 091
3 455 826	3 355 619	3 027 681	2 976 904	2 301 071	1 469 662	0	4 682 429
8 138 255	8 038 048	7 710 110	7 659 333	6 983 500	6 152 091	4 682 429	0

Így a (47) képlet szerint:

$$G = \frac{327938 + 378715 + ... + 4682429}{7 \cdot 6} = 3016833 .$$

Az átlagos díjbevétel a 11. táblázat adataiból:

$$\bar{x} = \frac{428135 + ... + 1154755}{7} = 2248896 .$$

Az (57) képlet szerint:

$$L = \frac{3016833}{2 \cdot 2248896} = 0{,}67 .$$

Ezek alapján azt mondhatjuk, hogy a kötelező gépjármű-biztosítások piaca meglehetősen koncentrált.

3.5. Momentumok

A **momentumok** a mennyiségi ismérvek vizsgálatának egy egységes átfogó rendszerét alapozzák meg, az átlagok és a szórások bizonyos általánosítását jelentik. A momentumok az ismérvértékek egy A tetszőleges konstanstól vett eltérések hatványait átlagolják (58)-(59) szerint.

A momentumok nem súlyozott képlete:

$$M_r(A) = \frac{\sum_{i=1}^{N} (x_i - A)^r}{N}. \qquad (58)$$

A súlyozott képlet:

$$M_r(A) = \frac{\sum_{i=1}^{k} f_i (x_i - A)^r}{\sum_{i=1}^{k} f_i}. \qquad (59)$$

$M_r(A)$-t az A körüli r-edik momentumnak nevezzük. Ha A=0, akkor egyszerűen ***r*-edik momentumról** beszélünk, jele: M_r; ha pedig $A = \bar{x}$, akkor az ***r*-edik centrális momentumot** kapjuk, jele: $M_r(\bar{x})$.

Néhány eddigi mutatószámunk momentumokkal való kifejezését a 28. táblázat tartalmazza.

Néhány nevezetes momentum

28. táblázat

r	A=0	$A=\bar{x}$
-1	$\frac{1}{\bar{x}_h}$	–
1	$\bar{x}$	0
2	$\left(\bar{x}_q\right)^2$	σ^2

Hasonlítsa össze az (58)-(59) képleteket a 3.2. és a 3.3. fejezetben leírtakkal!

(Megjegyzés: ha az r=0, akkor mindegyik momentum 1-gyel egyenlő.)

Mivel a momentumokat a gyakorisági eloszlások elemzése során gyakran alkalmazzuk, ezért a továbbiakban a momentumok és a lineárisan transzformált ismérvértékek kapcsolatát elemezzük.

Induljunk ki a (31) szerinti lineáris transzformációból.

Ahogy azt már láttuk, a transzformált értékek számtani átlaga és az eredeti értékek számtani átlaga között (32) összefüggés áll fenn.

Ha B=h , akkor a centrális momentumokra az alábbi azonosságok érvényesek.

$$M_1(\bar{x}) = 0$$

$$M_2(\bar{x}) = h^2\left(\frac{\sum_{i=1}^{k} f_i y_i^2}{\sum_{i=1}^{k} f_i} - \bar{y}^2\right) = h^2\left(\bar{y}_q^2 - \bar{y}^2\right) = h^2\sigma_y^2 = \sigma^2 \qquad (60)$$

$$M_3(\bar{x}) = h^3\left(\frac{\sum_{i=1}^{k} f_i y_i^3}{\sum_{i=1}^{k} f_i} - 3\cdot\bar{y}\cdot\frac{\sum_{i=1}^{k} f_i y_i^2}{\sum_{i=1}^{k} f_i} + 2\cdot\bar{y}^3\right) \qquad (61)$$

$$M_4(\bar{x}) = h^4 \left(\frac{\sum_{i=1}^{k} f_i y_i^4}{\sum_{i=1}^{k} f_i} - 4 \cdot \bar{y} \cdot \frac{\sum_{i=1}^{k} f_i y_i^3}{\sum_{i=1}^{k} f_i} + 6 \cdot \bar{y}^2 \cdot \frac{\sum_{i=1}^{k} f_i y_i^2}{\sum_{i=1}^{k} f_i} - 3 \cdot \bar{y}^4 \right) \qquad (62)$$

Megjegyzés: az előző képletekben, az abszolút gyakoriságok helyett, relatív gyakoriságok is szerepelhetnek.

A momentumok ismeretében, az eredeti adatokra vonatkozóan, (53) felírható a (63) módon is.

$$M_2(\bar{x}) = M_2 - M_1^2 \qquad (63)$$

40. példa

A magánnyugdíj-pénztári tagságra vonatkozó adatokat a 29. táblázat tartalmazza. Számítsuk ki a második, a harmadik és a negyedik centrális momentum értékét!

A magánpénztári tagok száma korcsoportonként 1998 végén

29. táblázat

Korcsoport	Magánpénztári tagok száma (f_i)	Magánpénztári tagok relatív gyakorisága (%) (g_i)
15-19	29 546	2,1981
20-24	266 580	19,8324
25-29	294 803	21,9321
30-34	260 958	19,4141
35-39	218 467	16,2530
40-44	182 722	13,5937
45-49	79 702	5,9295
50-54	10 067	0,7489
55-59	1 231	0,0916
60-64	58	0,0043
65-69	24	0,0018
70-74	6	0,0004
Összesen	1 344 164	100,0000

Forrás: Pénztárfelügyelet

Az első oszlopban levő adatok szerint h=5, a tetszőleges konstans pedig legyen A=32. Ezeknek az értékeknek megfelelő transzformált változók értékeit és a mellékszámításokat a 30. táblázat tartalmazza.

A magánpénztári tagok száma korcsoportonként 1998 végén

30. táblázat

y_i	f_i	$f_i \cdot y_i$	$f_i \cdot y_i^2$	$f_i \cdot y_i^3$	$f_i \cdot y_i^4$
-3	29 546	-88 638	265 914	-797 742	2 393 226
-2	266 580	-533 160	1 066 320	-2 132 640	4 265 280
-1	294 803	-294 803	294 803	-294 803	294 803
0	260 958	0	0	0	0
1	218 467	218 467	218 467	218 467	218 467
2	182 722	365 444	730 888	1 461 776	2 923 552
3	79 702	239 106	717 318	2 151 954	6 455 862
4	10 067	40 268	161 072	644 288	2 577 152
5	1 231	6 155	30 775	153 875	769 375
6	58	348	2 088	12 528	75 168
7	24	168	1 176	8 232	57 624
8	6	48	384	3 072	24 576
Σ	1 344 164	-46 597	3 489 205	1 429 007	20 055 085

$$\bar{y} = \frac{\sum_{i=1}^{12} f_i y_i}{\sum_{i=1}^{12} f_i} = \frac{-46597}{1344164} = -0{,}035$$

Így (32) alapján

$$\bar{x} = 32 + 5 \cdot (-0{,}035) = 31{,}83\,.$$

A (60) képlet alapján:

$$M_2(\bar{x}) = h^2\left(\frac{\sum_{i=1}^{12} f_i y_i^2}{\sum_{i=1}^{12} f_i} - \bar{y}^2\right) = 5^2 \cdot \left(\frac{3489205}{1344164} - \left(-0{,}035\right)^2\right) =$$

$$= 25 \cdot \left(2{,}5958 - 0{,}0012\right) = 64{,}87.$$

A (61) képlet alapján:

$$M_3(\bar{x}) = h^3\left(\frac{\sum_{i=1}^{12} f_i y_i^3}{\sum_{i=1}^{12} f_i} - 3 \cdot \bar{y} \cdot \frac{\sum_{i=1}^{12} f_i y_i^2}{\sum_{i=1}^{12} f_i} + 2 \cdot \bar{y}^3\right) =$$

$$= 5^3 \cdot \left(\frac{1429007}{1344164} - 3 \cdot (-0{,}035) \cdot 2{,}5958 + 2 \cdot (-0{,}035)^3\right) =$$

$$= 125 \cdot \left(1{,}0631 - (-0{,}2700) + (-0{,}00009)\right) = 166{,}62.$$

A (62) képlet alapján:

$$M_4(\bar{x}) = h^4\left(\frac{\sum_{i=1}^{12} f_i y_i^4}{\sum_{i=1}^{12} f_i} - 4 \cdot \bar{y} \cdot \frac{\sum_{i=1}^{12} f_i y_i^3}{\sum_{i=1}^{12} f_i} + 6 \cdot \bar{y}^2 \cdot \frac{\sum_{i=1}^{12} f_i y_i^2}{\sum_{i=1}^{12} f_i} - 3 \cdot \bar{y}^4\right) =$$

$$= 5^4\left(\frac{20055085}{1344164} - 4 \cdot (-0{,}035) \cdot 1{,}0631 + 6 \cdot \left(-0{,}035\right)^2 \cdot 2{,}5958 - 3 \cdot \left(0{,}035\right)^4\right) =$$

$$= 9428{,}90.$$

Ellenőrzésképpen, illetve a fenti képletek hasznosságának megítélésére, számítsuk ki a második centrális momentumot (vagyis a szórásnégyzetet) transzformáció nélkül az (52) képlet alapján! A mellékszámításokat a 31. táblázat tartalmazza.

A magánpénztári tagok száma korcsoportonként 1998 végén

31. táblázat

x_i	f_i	$f_i \cdot (x_i - \bar{x})^2$
17	29 546	6 495 100,73
22	266 580	25 741 878,60
27	294 803	6 867 947,58
32	260 958	7 840,11
37	218 467	5 846 909,09
42	182 722	18 911 116,57
47	79 702	18 349 788,84
52	10 067	4 096 899,29
57	1 231	780 080,49
62	58	52 804,93
67	24	29 691,92
72	6	9 683,38
Összesen	1 344 164	87 189 741,53

$$M_2(\bar{x}) = \sigma^2 = \frac{87189741{,}53}{1344164} = 64{,}87$$

A harmadik és negyedik centrális momentumokhoz még ezeknél is nagyobb számokkal kellett volna számolni a transzformáció nélkül.

3.6. Alakmutatók

Az egymóduszú gyakorisági eloszlások alakját gyakran hasonlítjuk a valószínűségszámításból jól ismert normális eloszlás gyakorisági görbéjéhez. Ha egy gyakorisági eloszlásnak több módusza van, akkor ez arra enged következtetni, hogy a jelenség eloszlása a vizsgált ismérv mellett más ismérvektől is jelentősen függ. Ebben az esetben az eddigi mutatószámok nem alkalmasak a jelenség tömör jellemzésére. Ilyenkor ún. **heterogén sokaságról** beszélünk, amelynek vizsgálatát a sokaság részekre bontásával végezzük. (A részekre bontott sokaságok vizsgálatával később részletesen foglalkozunk.)

Az egymóduszú gyakorisági eloszlások alakja kétféleképpen különbözhet az azonos szórású normális gyakorisági görbétől:

- az eloszlás valamelyik irányba hosszan elnyúló, tehát nem szimmetrikus, vagy
- az empirikus eloszlás maximumában nagyobb vagy kisebb, mint a normális eloszlás gyakorisági görbéjének maximum helye, tehát csúcsosabb, vagy lapultabb annál.

Aszimmetria- (ferdeség) mutatók

Az eddigiekben azt vizsgáltuk, és a középértékek segítségével számszerűsítettük, hogy a gyakorisági eloszlások hol helyezkednek el a számegyenesen. Megállapítottuk, hogy az átlag jellemző ereje függ attól, hogy az egyes ismérvértékek mennyire különböznek egymástól. A szóródási mutatóink azonban mindig érzéketlenek voltak az átlagtól való eltérések előjelére, így csak azt mutatták, hogy az ismérvértékek milyen távol helyezkednek el az átlagtól, azt már nem, hogy az átlag körül annak két oldalán egyenlően oszlanak-e meg. Egy empirikus gyakorisági eloszlásról tömör számszerű jellemzők, mutatószámok segítségével nyert információink körét tovább bővíthetjük, ha a fenti tulajdonság jellemzését is megadjuk.

A társadalmi és gazdasági statisztikában igen gyakoriak az aszimmetrikus eloszlások, mert gyakori, hogy valamely átlagtól való eltérést okozó tényező hatása egyirányú és kimagasló a többi hatáshoz képest. Ezek között is gyakoribb a baloldali aszimmetria, mert a 0 érték általában alsó korlátot jelent.

Az aszimmetria szempontjából három lehetséges esetet különböztetünk meg.

Baloldali aszimmetria: az azonos szórású normális eloszlás gyakorisági görbéjéhez képest jobbra hosszan elnyúló eloszlás.

Jobboldali aszimmetria: az azonos szórású normális eloszlás gyakorisági görbéjéhez képest balra hosszan elnyúló eloszlás.

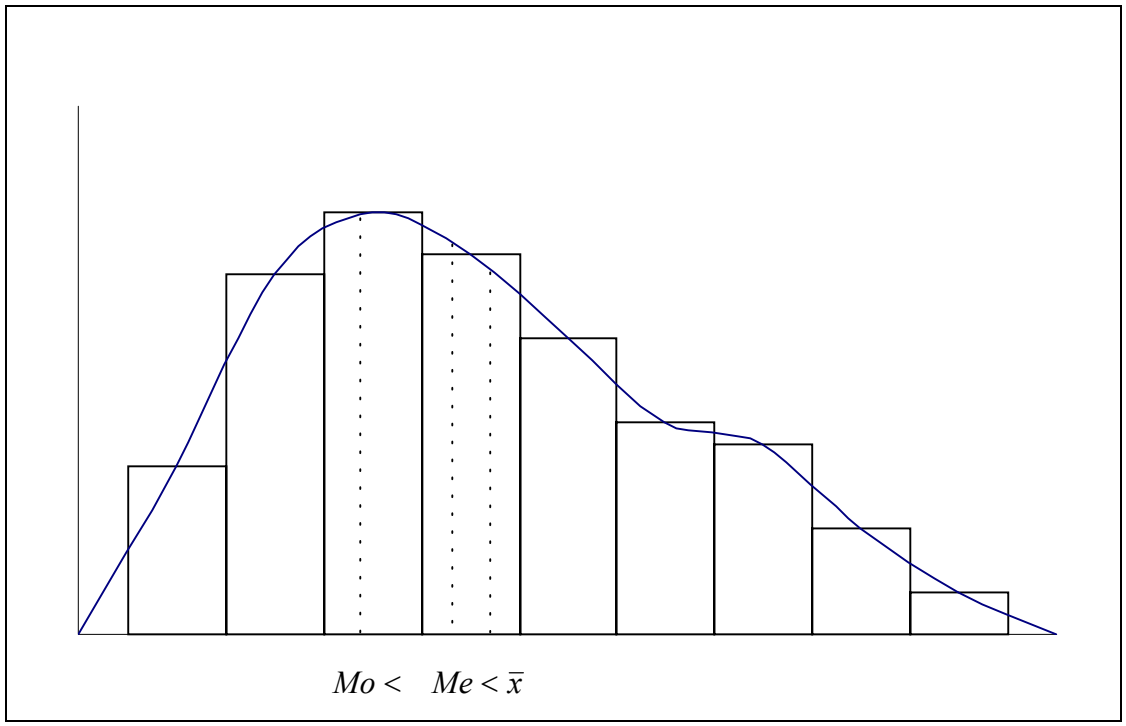
Szimmetrikus eloszlás: nem (baloldali és jobboldali) aszimmetrikus eloszlás.

Az aszimmetria többféleképpen is megragadható, mi a mérésére kétféle típusú mutatót fogunk használni.

Helyzet-mutatókra épülő aszimmetria-mutatók

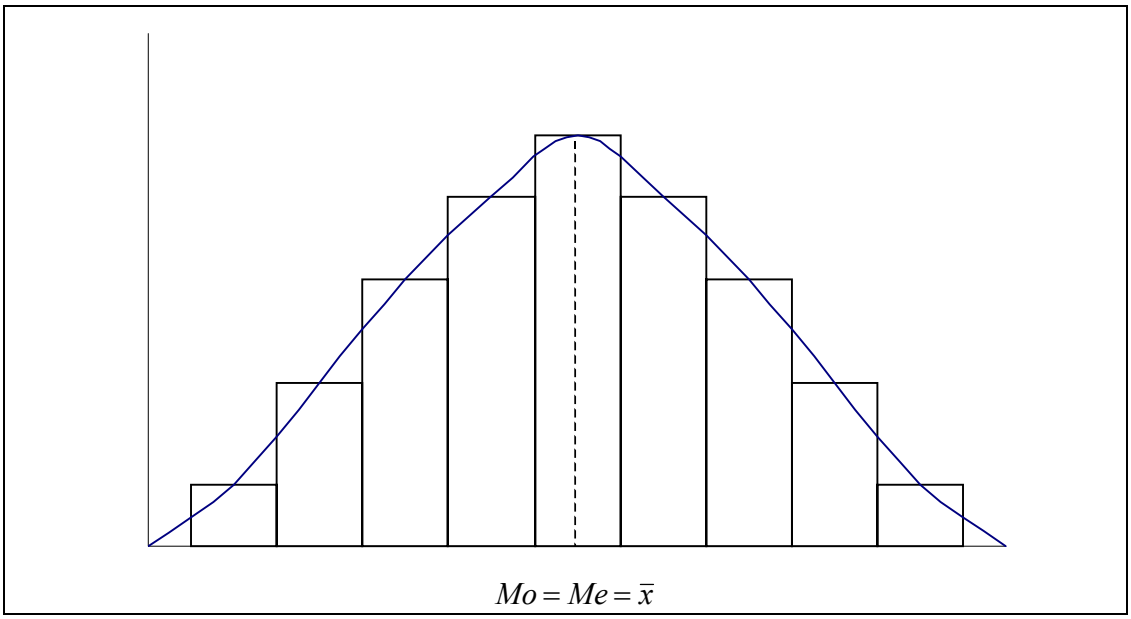
A ferdeség szempontjából vizsgált eloszlások három típusában az átlag, a medián és a módusz elhelyezkedése egyértelmű, így ebből következtethetünk a ferdeségre. Ezeket mutatják a 11.-13. ábrák.

Balra ferdült eloszlás



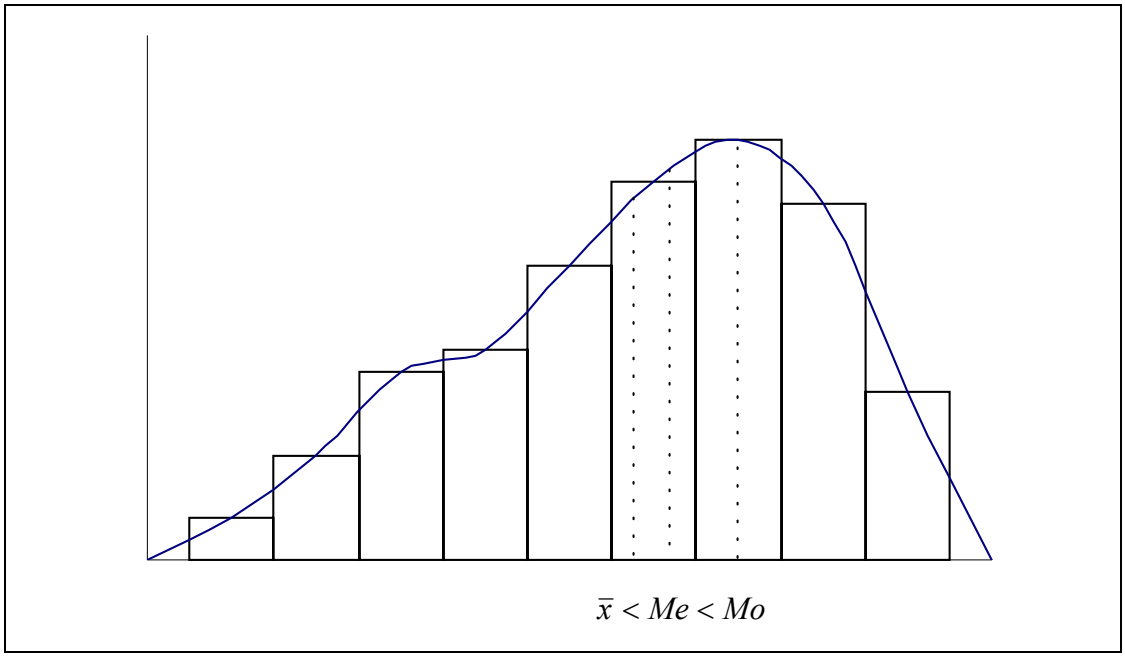
11. ábra

Szimmetrikus eloszlás



12. ábra

Jobbra ferdült eloszlás



13. ábra

(Megjegyzés: a 11.-13. ábráknál az adott középértékeket függőleges szaggatott vonal jelöli.)

Az empirikus eloszlások alakját valójában a gyakorisági görbéjük alapján kell megítélnünk, de a gyakorisági poligon és a hisztogram alapján is lehet következtetéseket levonni.

A **PEARSON-féle aszimmetria-mutató** az átlag és a medián eltérésére alapoz. Szimmetrikus eloszlás esetében ugyanis az átlagnál ugyanannyi kisebb és nagyobb érték van, vagyis a medián és az átlag egybe esik. Ekkor a módusz is nyilvánvalóan azonos lesz velük.

$$P = 3 \cdot \frac{\bar{x} - Me}{\sigma} \tag{64}$$

Normális eloszlás esetén P=0; baloldali aszimmetriát mutató sokaságra értéke pozitív, jobboldali aszimmetriánál negatív. Értékére nem adható alsó, illetve felső korlát, de értéke legtöbbször –3 és 3 közé esik. P -0,5-nél kisebb, illetve 0,5-nél nagyobb értéke erős ferdeségre utal.

A ferdeség F mutatója a kvartilisekre épít :

$$F = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}. \tag{65}$$

Normális eloszlás esetén F=0; baloldali aszimmetriát mutató sokaságra értéke pozitív, jobboldali aszimmetriánál negatív. F abszolút értéke 1-nél nem nagyobb.

Momentumokra épülő aszimmetria-mutató

Az α_3 mutató az eloszlások ferdeségének számszerűsítésére a harmadik centrális momentumot használja.

$$\alpha_3 = \frac{M_3(\bar{x})}{\sigma^3} \tag{66}$$

Ez már az eloszlás összes értékét figyelembe veszi. Értékére nem adható alsó, illetve felső korlát. Normális eloszlás esetén α_3=0; baloldali aszimmetriát mutató sokaságra értéke pozitív, jobboldali aszimmetriánál negatív.

Itt jegyezzük meg, hogy az aszimmetria-mutatók osztályközös gyakorisági sorból történő számításakor szimmetrikus eloszlás esetén is előfordulhat 0 körüli érték, hiszen ekkor a helyzet-mutatók és a momentumok értéke becsült. Az ismertetett mutatókat csak akkor érdemes kiszámítani, ha a gyakorisági poligon jól közelíti a gyakorisági görbét, amihez nagy elemszámú sokaság szükséges.

Az Excel a saját értelmezése szerinti aszimmetria mértékét a FERDESÉG(szám1;szám2;...) statisztikai függvény segítségével számszerűsíti.

41. példa

Vizsgáljuk meg a 40. példa adatai alapján, hogy milyen aszimmetriájú a magánpénztári tagok korcsoport szerinti eloszlása!

A PEARSON-féle aszimmetria-mutató kiszámításához szükségünk lesz a mediánra.

Ez a (43) képlet alapján a következő:

$$\hat{Me} = 30 + \frac{672082 - 590929}{260958} \cdot 5 = 31{,}55\,.$$

A PEARSON-féle aszimmetria-mutatót (64) szerint a 40. példa eredményeinek felhasználásával kapjuk.

$$P = 3 \cdot \frac{31{,}83 - 31{,}55}{8{,}05} = 0{,}104$$

Ez arra utal, hogy a magánpénztári tagok korcsoport szerinti eloszlása az azonos szórású normális eloszlás gyakorisági görbéjéhez viszonyítva baloldali aszimmetriájú.

A 40. példa eredményeinek felhasználásával számítsuk ki (66) szerint az α_3 mutatót is!

$$\alpha_3 = \frac{166{,}62}{522{,}42} = 0{,}32$$

Ez szintén baloldali aszimmetriát jelez az azonos szórású normális eloszlás gyakorisági görbéjéhez képest.

Csúcsossági (kurtózis) mutatók

A szimmetrikus és kevésbé aszimmetrikus empirikus eloszlások jellemzését bővíthetjük azzal, ha a **csúcsosság** szempontjából is összehasonlítjuk az azonos szórású normális eloszlás gyakorisági görbéjével. A csúcsosság is megragadható többféleképpen, mi a mérésére kétféle típusú mutatót fogunk használni.

Helyzet-mutatókra épülő csúcsossági mutató

A csúcsosság *K* mutatója azt használja ki, hogy csúcsosabb eloszlások esetén a két szélső kvartilis különbségének a két szélső decilis különbségéhez viszonyított aránya kisebb, mint lapultabb eloszlásoknál.

$$K = \frac{Q_3 - Q_1}{2(D_9 - D_1)} \tag{67}$$

Normális eloszlás esetén $K \approx 0{,}263$; a normális eloszláshoz képest lapultabb eloszlások *K* értéke 0,263-nél nagyobb; míg a normális eloszlásnál csúcsosabbaké 0,263-nél kisebb.

Momentumokra épülő csúcsossági mutató

Az α_4 mutató az eloszlások csúcsosságának számszerűsítésére a negyedik centrális momentumot használja.

$$\alpha_4 = \frac{M_4(\bar{x})}{\sigma^4} \tag{68}$$

Ez már az eloszlás összes értékét figyelembe veszi. A standard normális eloszlású változó esetén $\alpha_4 = 3$. A normális eloszláshoz képest csúcsosabb eloszlások α_4 értéke 3-nál nagyobb, a normális eloszlásnál lapultabbaké 3-nál kisebb. Ez a mutató mindig 1-nél nagyobb értéket vesz fel.
A csúcsossági mutatókra is érvényes, hogy csak nagy elemszámú sokaságokra érdemes kiszámítani.

Az Excel a saját értelmezése szerinti csúcsosság mértékét a CSÚCSOSSÁG(szám1;szám2;...) statisztikai függvény segítségével számszerűsíti.

42. példa

Vizsgáljuk meg a 40. példa adatai alapján, hogy milyen csúcsosságú a magánpénztári tagok korcsoport szerinti eloszlása!

A *K* mutató értékéhez számítsuk ki az alsó és a felső kvartilist, illetve az első és a kilencedik decilist a (44) képlet segítségével!

A kvartilisek az alábbiak.

$$\hat{Q}_1 = 25 + \frac{336041 - 296126}{294803} \cdot 5 = 25{,}68$$

$$\hat{Q}_3 = 35 + \frac{1008123 - 851887}{218467} \cdot 5 = 38{,}58$$

A decilisek pedig:

$$\hat{D}_1 = 20 + \frac{134416 - 29546}{266580} \cdot 5 = 21{,}97\,;$$

$$\hat{D}_9 = 40 + \frac{1209748 - 1070354}{182722} \cdot 5 = 43{,}81\,.$$

A *K* mutató értéke a (67) képlet alapján:

$$K = \frac{38{,}58 - 25{,}68}{2(43{,}81 - 21{,}97)} = 0{,}295\,.$$

E szerint a magánpénztári tagok korcsoport szerinti eloszlása az azonos szórású normális eloszlás gyakorisági görbéjéhez képest lapultabb.

A 40. példa eredményeinek felhasználásával számítsuk ki az α_4 mutatót is a (68) képlet segítségével!

$$\alpha_4 = \frac{9428{,}90}{4207{,}52} = 2{,}57$$

Ez szintén lapultabb eloszlást jelez az azonos szórású normális eloszlás gyakorisági görbéjéhez képest.

4. Sokaság több ismérv szerinti vizsgálata

4.1. Részekre bontott sokaságok

Az eddigiekben homogén sokaságokkal foglalkoztunk, azt feltételeztük, hogy a vizsgált jelenségünket egy ismérvvel jellemezhetjük. Az eloszlások vizsgálatakor többször hangsúlyoztuk, hogy egymóduszú sokaságokkal foglalkozunk, használt mutatószámaink csak ekkor voltak alkalmasak a sokaság jellegzetességeinek megragadására. (A többmóduszú eloszlások általában heterogén sokaságot jellemeznek, a gyakorisági görbe maximumhelyei a homogén részsokaságok móduszainál jelentkeznek.)

Heterogén sokaságok esetén az eddig ismertetett elemzési technikák összemossák a sokaság jellegzetességeit, ezért ezeket a sokaságokat, az elemzés első lépésében, a heterogenitást előidéző ismérv szerint csoportosítjuk, részekre bontjuk. A megfelelő csoportképző ismérv megtalálása általában igen nehéz; az erre irányuló eljárást **klaszteranalízisnek** nevezzük, amellyel egyelőre nem foglalkozunk. Feltételezzük, hogy a sokaság természetének ismeretében meghatározott ismérv vagy ismérvek alapján a teljes sokaságot (az ún. **fősokaságot)** részekre (ún. **részsokaságokra)** bontottuk.

Részviszonyszámok és összetett viszonyszámok

Ha a fősokaságot M számú részsokaságra bontjuk, akkor a részsokaságokra számított azonos típusú

$$V_j = \frac{A_j}{B_j} \qquad j=1, 2, \ldots, M \tag{69}$$

viszonyszámokat **részviszonyszámoknak,** a fősokaságra vonatkozó

$$\overline{V} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} B_j} \tag{70}$$

viszonyszámot **összetett viszonyszámnak** nevezzük.

A (69) képlet alapján $A_j = B_j \cdot V_j$ és $B_j = \frac{A_j}{V_j}$, így az összetett viszonyszám kiszámítható a részviszonyszámok súlyozott számtani átlagaként:

$$\overline{V} = \frac{\sum_{j=1}^{M} B_j \cdot V_j}{\sum_{j=1}^{M} B_j}, \tag{71}$$

illetve a részviszonyszámok súlyozott harmonikus átlagaként:

$$\overline{V} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} \frac{A_j}{V_j}}. \tag{72}$$

A (71)-(72) képletekben a B_j, illetve az A_j adatok helyettesíthetők a belőlük számított megoszlási viszonyszámokkal is.

43. példa

Egy farmer három egymástól távol eső parcellán termel kukoricát. Az első parcella nagysága 4 ha és a termés mennyisége 16 t; a második parcella nagysága 5 ha és a termés mennyisége 25 t. A harmadik parcellára vonatkozó adatok: 2 ha, 14 t. Számítsuk ki a (69)-(72) szerinti mutatókat!

Az egyes parcellákra (részsokaságokra) vonatkozó átlaghozamok (megoszlási viszonyszámok):

$$V_1 = \frac{16}{4} = 4\ [\text{t/ha}], \qquad V_2 = \frac{25}{5} = 5\ [\text{t/ha}], \qquad V_3 = \frac{14}{2} = 7\ [\text{t/ha}].$$

Együttvéve a három parcellát:

$$\overline{V} = \frac{16 + 25 + 14}{4 + 5 + 2} = 5\ [\text{t/ha}].$$

Az összetett viszonyszámot megkaphattuk volna a részviszonyszámok súlyozott számtani átlagaként:

$$\overline{V} = \frac{4 \cdot 4 + 5 \cdot 5 + 2 \cdot 7}{4 + 5 + 2} = 5\,[\text{t/ha}],$$

vagy a részviszonyszámok súlyozott harmonikus átlagaként:

$$\overline{V} = \frac{16 + 25 + 14}{\frac{16}{4} + \frac{25}{5} + \frac{14}{7}} = 5\,[\text{t/ha}].$$

Részátlagok és főátlagok

Ha a fősokaságot M számú részsokaságra bontjuk, akkor a részsokaságokra számított

$$\overline{x}_j = \frac{\sum_{i=1}^{N_j} x_{ij}}{N_j} = \frac{S_j}{N_j} \qquad j=1, 2, \ldots, M \tag{73}$$

átlagokat **részátlagoknak**, míg a fősokaságra számított

$$\overline{x} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j} x_{ij}}{N} = \frac{\sum_{j=1}^{M} S_j}{N} = \frac{S}{N} \tag{74}$$

átlagot **főátlagnak** nevezzük. A (73)-(74) képletben szereplő S_j a j-edik részsokaság, míg a (74) képletben szereplő S a fősokaság értékösszege. A részsokaságok N_j elemszámaira és a fősokaság N nagyságára érvényes a $\sum_{j=1}^{M} N_j = N$ összefüggés.

Ha az átlagot, mint a sokaság értékösszegének és elemszámának hányadosaként képzett intenzitási viszonyszámot fogjuk fel, akkor a (71)-(72) képletek alapján a főátlag a részátlagokból (75)-(76) szerint is megkapható.

$$\overline{x} = \frac{\sum_{j=1}^{M} N_j \cdot \overline{x}_j}{\sum_{j=1}^{M} N_j} \tag{75}$$

$$\bar{x} = \frac{\sum_{j=1}^{M} S_j}{\sum_{j=1}^{M} \frac{S_j}{\bar{x}_j}} \tag{76}$$

44. példa

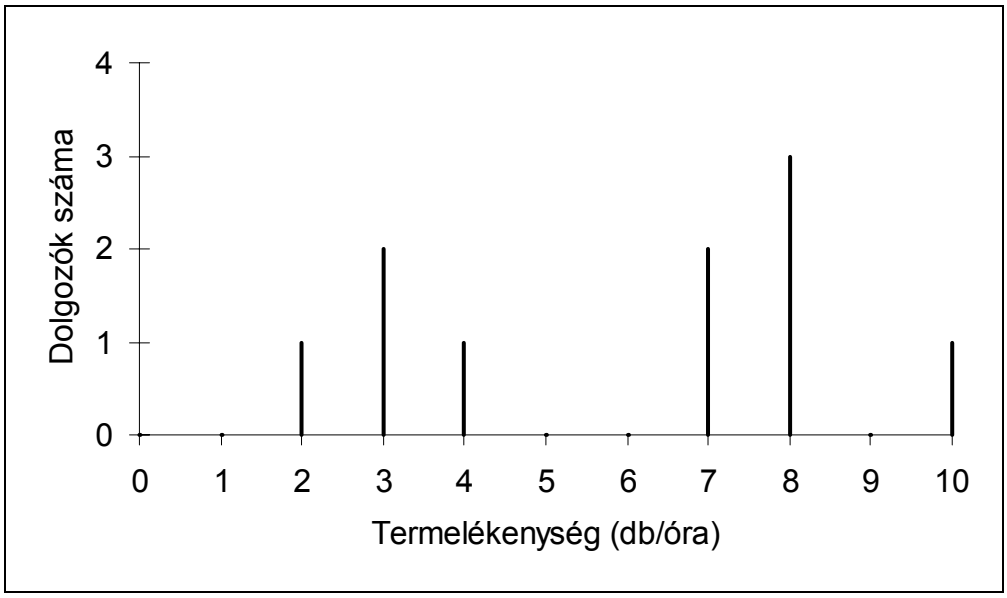
Egy kft 10 fizikai foglalkozású dolgozója azonos fajta terméket állít elő. A termelékenységüket (db/óra) tükröző mutatók - nevük szerinti ábécé sorrendben – a következők: 3, 8, 8, 7, 3, 7, 4, 10, 8, 2.

Termelékenység szempontjából homogénnek tekinthető-e az adott statisztikai sokaság?

A kérdés megválaszolásában sokat segíthet, ha rangsorba állítjuk, vagy grafikusan ábrázoljuk az adatokat. Az adatok rangsora a következő: 2, 3, 3, 4, 7, 7, 8, 8, 8, 10.

Eloszlásuk a 14. ábrán látható.

Termelékenység szerinti eloszlás



14. ábra

A rangsor illetve az ábra alapján könnyen észrevehető, hogy nem homogén sokaságról van szó, mert többmóduszú az eloszlás, azaz a heterogénnek tekinthető fősokaság két homogénebb részsokaságra osztható. Az első elemei: 2, 3, 3, 4; a második elemei: 7, 7, 8,

8, 8, 10. Mi lehet ennek a magyarázata? Egyik kézenfekvő magyarázat lehetne egy új ismérv (a szakképzettség) figyelembevétele: a 10 dolgozó közül 4-en betanított munkások, 6-an pedig szakképzettek.

Ezek szerint M=2, N=10, N_1=4 és N_2=6. Számítsuk ki a két részátlagot és a főátlagot! A (73) képlet szerint:

$$\bar{x}_1 = \frac{2+3+3+4}{4} = 3\,[\text{db/óra}],$$

illetve

$$\bar{x}_2 = \frac{7+7+8+8+8+10}{6} = 8\,[\text{db/óra}].$$

A főátlagot háromféleképpen is kiszámíthatjuk.

A (74) képlet szerint:

$$\bar{x} = \frac{2+3+3+4+7+7+8+8+8+10}{10} = 6\,[\text{db/óra}],$$

a (75) képlet szerint:

$$\bar{x} = \frac{4\cdot 3+6\cdot 8}{4+6} = 6\,[\text{db/óra}],$$

illetve a (76) képlet szerint:

$$\bar{x} = \frac{12+48}{\frac{12}{3}+\frac{48}{8}} = 6\,[\text{db/óra}].$$

Megjegyzés: a (75) képletből következik, hogy a főátlag a részátlagok súlyozott számtani átlaga, ha a súlyozó tényező a részsokaságok nagysága.

A (76) képletből következik, hogy a főátlag a részátlagok súlyozott harmonikus átlaga, ha a súlyozó tényező a részsokaságok értékösszege.

Részsokaságok és fősokaságok szórása

A részekre bontott sokaságok esetén az ismérvértékek különbözőségét kifejező háromféle szórás is számítható. Vizsgálhatjuk

- a fősokaság egységeihez tartozó ismérvértékek főátlagtól való eltéréseit, illetve az $SST = \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}\right)^2$ **teljes eltérés-négyzetösszeget**[3];
- a fősokaság egységeihez tartozó ismérvértékek megfelelő részátlagtól való eltéréseit, illetve az $SSB = \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2$ **belső eltérés-négyzetösszeget**;
- a részátlagok eltérését a főátlagtól, illetve az $SSK = \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(\bar{x}_j - \bar{x}\right)^2$ **külső eltérés-négyzetösszeget**.

Ezek között az eltérés-négyzetösszegek között az alábbi összefüggés áll fenn:

$$SST = SSB + SSK ,$$

illetve

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2 + \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(\bar{x}_j - \bar{x}\right)^2 . \qquad (77)$$

A fenti három eltérés alapján a (78)-(80) képletekkel kifejezett szórásmutatókat használjuk.

A **teljes szórás** képlete:

$$\sigma = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}\right)^2}{N}} . \qquad (78)$$

[3] Az eltérés-négyzetösszegek angol megfelelőjének szokásos rövidítése,
SS: Sum of Squares = négyzetösszeg.

A **belső szórás** képlete:

$$\sigma_B = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2}{N}}. \qquad (79)$$

A **külső szórás** képlete:

$$\sigma_K = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(\bar{x}_j - \bar{x}\right)^2}{N}} = \sqrt{\frac{\sum_{j=1}^{M} N_j\left(\bar{x}_j - \bar{x}\right)^2}{N}}. \qquad (80)$$

A belső szórás azt fejezi ki, hogy a fősokaság egységeihez tartozó ismérvértékek átlagosan mennyivel térnek el a saját részátlaguktól.

A külső szórás azt fejezi ki, hogy a részátlagok átlagosan menyivel térnek el a főátlagtól.

A részsokaságra számított szórást **rész-szórásnak** vagy **csoporton belüli szórásnak** nevezzük.

A **rész-szórások** képlete:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2}{N_j}} \qquad j=1,2,\ldots,M. \qquad (81)$$

A belső szórás megkapható a részsokaságokból számított szórások súlyozott négyzetes átlagaként is:

$$\sigma_B = \sqrt{\frac{\sum_{j=1}^{M} N_j \sigma_j^2}{N}}. \qquad (82)$$

Természetesen az itt említett összes szórástípusnak használjuk a négyzetét is, amelyre most is a megfelelő szórásnégyzet, vagy variancia kifejezéssel utalunk. Használjuk tehát a teljes variancia, a belső szórásnégyzet, stb. fogalmakat.

A (77) képlet mindkét oldalát *N*-nel osztva a teljes, a belső és a külső variancia között az alábbi összefüggés áll fenn:

$$\sigma^2 = \sigma_B^2 + \sigma_K^2 . \qquad (83)$$

A fenti összefüggés úgy értelmezhető, hogy a fősokaság egységeihez tartozó ismérvértékek két hatás miatt ingadoznak a főátlag körül. A belső szórásnégyzet a csoportképző ismérven kívüli egyéb tényezők okozta hatást számszerűsíti, míg a külső szórásnégyzet a csoportképző ismérvnek betudható ingadozást jellemzi. Ez azt jelenti, hogy minél nagyobb részt tesz ki a külső variancia a teljes variancián belül, annál nagyobb részét magyarázza meg a csoportképző ismérv a fősokaság egységei (vizsgált ismérv szempontjából vett) ingadozásainak. Tehát minél nagyobb a $\frac{\sigma_K^2}{\sigma^2} = \frac{SSK}{SST}$ hányados, annál homogénebbek a képzett csoportok. Ez egyben az alkalmazott csoportképző ismérv „jóságát", megfelelőségét is mutatja.

45. példa

Számítsuk ki a 44. példa adataiból a szórás-mutatókat!

A (81) képlet alapján a rész-szórások:

$$\sigma_1 = \sqrt{\frac{(2-3)^2 + ... + (4-3)^2}{4}} = 0{,}707 ; \qquad \sigma_2 = \sqrt{\frac{(7-8)^2 + ... + (10-8)^2}{6}} = 1{,}000 .$$

A belső szórásnégyzet a (82) képlet szerint:

$$\sigma_B^2 = \frac{4 \cdot 0{,}707^2 + 6 \cdot 1^2}{10} = 0{,}8 .$$

A külső szórásnégyzet (80) alapján:

$$\sigma_K^2 = \frac{4 \cdot (3-6)^2 + 6 \cdot (8-6)^2}{10} = 6{,}0 .$$

A teljes szórásnégyzet (78) alapján:

$$\sigma^2 = \frac{(2-6)^2 + \ldots + (10-6)^2}{10} = 6{,}8\,.$$

Könnyen ellenőrizhetjük, hogy fennáll (83). Mivel a teljes szórásnégyzet legnagyobb részét a külső szórásnégyzet teszi ki, azt mondhatjuk, hogy a sokaság részekre bontása a szakképzettség szerint hatékonynak bizonyult. (Megjegyzés: a szakképzettség, mint csoportképző ismérv mellett más vállalkozásoknál a részek homogenitását egyéb ismérvek szerint is valószínűleg elérhetnénk, mint például a nem, a lakóhely, az életkor, stb.)

4.2. Ismérvek közötti kapcsolat

Az előző fejezetben egy sokaság egységeit egyidejűleg két ismérv szerint vizsgáltuk. A csoportképző ismérvet a sokaság részekre bontására, míg a vizsgálat tárgyát képező ismérvet elemzésre használtuk. Feltételeztük, hogy a két ismérv között szoros kapcsolat van, és ezért a keletkező részsokaságok a vizsgálat tárgyát képező ismérv tekintetében többé-kevésbé homogének. Most azzal fogunk foglalkozni, hogy részletesebben megvizsgáljuk az ismérvek közötti kapcsolat jellegét és szorosságát.

Két ismérv között háromféle típusú kapcsolat lehet:

- a két ismérv **független** egymástól, vagyis egy sokasági egység egyik ismérv szerinti hovatartozásának vagy ismérvértékének ismerete semmilyen információt nem szolgáltat a másik ismérv szerinti hovatartozásra vagy ismérvértékre vonatkozóan;

- a két ismérv között **sztochasztikus kapcsolat** van, vagyis egy sokasági egység egyik ismérv szerinti hovatartozásának vagy ismérvértékének ismerete szolgáltat információt a másik ismérv szerinti hovatartozásra vagy ismérvértékre vonatkozóan, de egy egyértelmű következtetés nem lehetséges;

- a két ismérv között **determinisztikus (függvényszerű) kapcsolat** van, vagyis egy sokasági egység egyik ismérv szerinti hovatartozásának vagy ismérvértékének ismerete alapján egyértelműen meghatározható a másik ismérv szerinti hovatartozás vagy ismérvérték.

A háromféle kapcsolat közül a leggyakrabban a sztochasztikus kapcsolattal találkozunk, ezért a továbbiakban arra fektetjük a hangsúlyt, hogy meghatározzuk vajon a két ismérv között a kapcsolat erősebb-e (szorosabb) vagy gyengébb-e (lazább), vagyis az egyik ismérv szerinti hovatartozás vagy ismérvérték ismerete mennyi többletinformációt hordoz a másik ismérv szerinti hovatartozásra vagy ismérvértékre vonatkozóan. Ezek után azt is meg fogjuk vizsgálni, hogy sztochasztikus kapcsolatban álló ismérvek esetén ezt a többletinformációt hogyan tudjuk felhasználni arra, hogy az egyik ismérvértékből következtessünk a másik ismérvértékre.

A továbbiakban két ismérv kapcsolatát vizsgáljuk, de az eddig említett tulajdonságok könnyen általánosíthatók több ismérvre is. (Megjegyzés: a többváltozós modellekkel csak a második kötetben foglalkozunk.)

Itt jegyezzük meg, hogy minden kapcsolatvizsgálat elején, az erre vonatkozó statisztikai eszközök használata előtt, egyéb (az adott témakört érintő szakmai) ismeretek alapján el kell dönteni, hogy van-e valamilyen valóságos alapja a két ismérv közötti kapcsolat létének. Az itt ismertetett eszközök ugyanis csak az ismérvek együtt-mozgásának kimutatására alkalmasak, és ezért fennáll a látszólagos, formális kapcsolatok számszerűsítésének veszélye. Az alábbi formális eljárások tehát csak annyiban értelmezhetők, amennyiben a probléma tartalmilag megalapozott.

A két ismérv típusaitól függően, az ismérvek közötti kapcsolatoknak 4 fajtáját elemezzük.

Asszociáció: két minőségi vagy területi ismérv kapcsolata.

Vegyes kapcsolat: egy mennyiségi és egy minőségi vagy területi ismérv kapcsolata.

Rangkorreláció: két ordinális skálán mért változó kapcsolata.

Korreláció: két mennyiségi ismérv kapcsolata.

Asszociáció

Az asszociációs kapcsolat elemzésénél az alábbi három módszert fogjuk alkalmazni.

Kombinációs tábla

Két minőségi vagy területi ismérv kapcsolatának létezését, illetve a kapcsolat erősségét már az adatok kombinációs táblába rendezésével is feltárhatjuk. Ha ebben a gyakoriságok elhelyezkedése bizonyos szabályosságot mutat, akkor érdemes konkrét mutatószámok segítségével kimutatni a kapcsolat szorosságát.

Az asszociációnál alkalmazott mutatószámokat (a kapcsolat jellege szempontjából) két megközelítés szerint kaphatjuk. Az egyik szempont szerint függetlenséget, a másik szempont szerint, éppen fordítva, függőséget feltételezünk a vizsgált ismérvek között. Az elsőnek említett megközelítés szerinti leggyakrabban alkalmazott mutatók a χ^2 alapú mutatók.

χ^2 alapú mutatók

A χ^2 alapú mutatók vizsgálatához feltételezzük, hogy sokaságunk a két ismérv szerint kombinációs táblába rendezett, összehasonlítjuk a sokaság egységeinek a két minőségi vagy területi ismérv szerinti tényleges eloszlását a függetlenséget feltételező eloszlással. Ehhez a kombinációs tábla minden eleméhez meg kell határoznunk azokat a feltételezett gyakoriságokat, amelyek a két ismérv függetlensége esetén adódnának. A valószínűségszámításból ismert függetlenség alapján ehhez a (84) képletet használjuk.

$$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{N} \tag{84}$$

(Megjegyzés: a képlet számlálójában szereplő gyakoriságok a 4. táblázatnál már ismertetett peremgyakoriságok.)

A tényleges f_{ij} és a függetlenséget feltételező f_{ij}^* gyakoriságok összevetése a χ^2 mutató segítségével történik.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*} \tag{85}$$

A (85) képlet alapján kiszámított értékre érvényes a következő reláció:

$$0 \le \chi^2 \le N \cdot \min\{(r-1),(c-1)\};$$

ahol a szorzandó a kapcsos zárójelben levő számok kisebbike.

χ^2 pontosan akkor lesz 0, amikor a tényleges gyakoriságok megegyeznek a függetlenséget feltételező gyakoriságokkal, vagyis amikor a két ismérv függetlennek tekinthető, és pontosan akkor éri el maximumát, ha a két ismérv között függvényszerű kapcsolat van. Értéke alapján képezhetjük az asszociációs kapcsolat szorosságát jellemző (86) és (87) mutatót.

A **CRAMER-féle asszociációs együtthatót** a (86) képlet szerint definiáljuk.

$$C = \sqrt{\frac{\chi^2}{N \cdot \min\{(r-1),(c-1)\}}} \tag{86}$$

C értéke 1-nél nem nagyobb. A mutató 0 értéke a két ismérv függetlenségére, míg 1-hez közeli értéke pedig nagyon erős kapcsolatra utal.

A **CSUPROV-féle asszociációs együtthatót** a (87) képlet szerint definiáljuk.

$$T = \sqrt{\frac{\chi^2}{N \cdot \sqrt{(r-1)(c-1)}}} \tag{87}$$

A *T* értéke szintén 0 és 1 között mozog.

PRE-mutatók

A **PRE-eljárás**[4] nem a függetlenség felöl közelít, hanem megpróbálja meghatározni azt a többletinformációt, amelyet az egyik ismérv nyújt a másikkal kapcsolatban. Ezek alapján nyilvánvaló, hogy meg kell állapítanunk mindenek előtt, hogy melyik ismérv szerint határozzuk meg a feltételes gyakoriságokat. Legyen ez az *X*-szel jelölt ismérv. Szokás szerint, a kombinációs táblázatban soronként értelmezzük a feltételes eloszlásokat. (Lásd a 4. táblázatot!)

Azt fogjuk számszerűsíteni, hogy mennyivel csökkenti az *X* ismérv szerinti hovatartozás figyelembevétele az *Y* ismérv hovatartozásának meghatározása során elkövetett hibát.

Ha egy sokasági egység *Y* szerinti hovatartozását az *Y* (*X*-től független) feltétel nélküli eloszlására alapozva határozzuk meg, akkor a hibánk

$$E_1 = N - \max_j f_{.j} ,$$

hiszen nyilván a leggyakoribb osztályra „tippelünk”, mert ez a legvalószínűbb.

Ha figyelembe vesszük az adott sokasági egységnek az *X* szerinti hovatartozását is, akkor arra az *Y*-osztályra fogunk „tippelni”, amelyhez a legmagasabb gyakoriság tartozik az adott C_i^X osztályon belül. Az így elkövetett hibánk:

$$E_2 = \sum_{i=1}^{r} \left(f_{i.} - \max_j f_{ij} \right) = N - \sum_{i=1}^{r} \max_j f_{ij} .$$

[4] PRE: Proportional Reduction of Errors = relatív hibacsökkenés.

A PRE-elv a hibacsökkenés mértékének meghatározására a (88) képletet használja.

$$PRE = \frac{E_1 - E_2}{E_1} \tag{88}$$

A már meghatározott E_1 és E_2 alapján, az asszociációs kapcsolatok szorosságának mérésére a (89) szerint definiált PRE-mutató képezhető. Értékét százalékban is kifejezhetjük.

$$\lambda_{Y|X} = \frac{\left(N - \max_j f_{.j}\right) - \left(N - \sum_{i=1}^{r} \max_j f_{ij}\right)}{N - \max_j f_{.j}} = \frac{\sum_{i=1}^{r} \max_j f_{ij} - \max_j f_{.j}}{N - \max_j f_{.j}} \tag{89}$$

λ azt mutatja meg, hogy az X szerinti hovatartozás ismerete $100 \cdot \lambda$ százalékkal csökkenti az Y szerinti hovatartozás becslésekor elkövetett hibánkat. Azt a becslési hibát csökkenti, amit X ismerete nélkül történő „tippeléskor” követünk el.

Megjegyezés: a λ nem szimmetrikus mutató, vagyis általában $\lambda_{X|Y} \neq \lambda_{Y|X}$, és értéke nem nagyobb 1-nél, illetve 100%-nál.

46. példa

Egy város önkormányzata kikérte a lakosság véleményét egy szemétégető felépítéséről. A népszavazás (iskolai végzettség szempontjából rendezett) eredményét a 32. táblázat tartalmazza.

A népszavazás eredménye

32. táblázat

Iskolai végzettség	Válaszok		
	Igen	Nem	Tartózkodott
Kevesebb, mint 8 általános	103	1 612	305
8 általános	2 011	5 320	1 052
Középiskola	4 010	2 013	1 988
Főiskola	1 502	398	101
Egyetem	1 802	95	50
Összesen	9 428	9 438	3 496

Vizsgáljuk meg, hogy milyen szoros összefüggés mutatkozik az iskolai végzettség és az adott kérdésben kialakított vélemény között!

A CRAMER-féle asszociációs együtthatóhoz először számítsuk ki a χ^2 mutatót a (85) képlet szerint.

A függetlenséget feltételező eloszlást a 33. táblázat tartalmazza.

A χ^2 mutató számításához szükséges munkatábla

33. táblázat

Iskolai végzettség	Igen	Nem	Tartózk.	Összesen
Kevesebb, mint 8 ált.	851,6	852,6	315,8	2 020,0
8 általános	3 534,3	3 538,1	1 310,6	8 383,0
Középiskola	3 377,5	3 381,1	1 252,4	8 011,0
Főiskola	843,6	844,5	312,8	2 001,0
Egyetem	820,9	821,7	304,4	1 947,0
Összesen	9 428,0	9 438,0	3 496,0	22 362,0

A táblázat belsejében levő adatok alapján:

$$\chi^2 = \frac{(103-851{,}6)^2}{851{,}6} + \frac{(1612-852{,}6)^2}{852{,}6} + \ldots + \frac{(50-304{,}4)^2}{304{,}4} = 6965{,}4\,.$$

Innen:

$$C = \sqrt{\frac{6965{,}4}{22362 \cdot 2}} = 0{,}39\,.$$

Ez arra utal, hogy az iskolai végzettség és az adott kérdésben kialakított vélemény között a közepesnél gyengébb kapcsolat van.

Számszerűsítsük a kapcsolat erősségét a λ mutató szerint is!

Mivel a λ mutató aszimmetrikusan mér, el kell döntenünk, hogy melyik változó az ok és melyik az okozat. Ebben az esetben egyértelmű, hogy az iskolai végzettség (X) befolyásolhatja a választ (Y); a fordított irányú kapcsolatnak nincs értelme.

A (89) képlet és a 32. táblázat adatai alapján:

$$\lambda_{Y|X} = \frac{(1612 + 5320 + 4010 + 1502 + 1802) - 9438}{22362 - 9438} = \frac{4808}{12924} = 0{,}372\,.$$

A kapott eredményt a következők szerint értelmezhetjük: a válasszal kapcsolatos bizonytalanságunkat 37,2%-kal tudjuk csökkenteni, ha ismerjük a válaszadó iskolai végzettségét.

Vegyes kapcsolat

A vegyes kapcsolat szorosságának mérésére egy PRE-eljárás szerint értelmezhető mutatót fogunk használni, amely levezetésének részletezésével nem foglalkozunk.

Az alkalmazásra kerülő mérőszám a **variancia-hányadosnak** nevezett PRE-mutató. A vegyes kapcsolatok szorosságának mérése a (90) szerint történik.

$$H^2 = \frac{SST - SSB}{SST} = \frac{SSK}{SST} = 1 - \frac{\sigma_B^2}{\sigma^2} = \frac{\sigma_K^2}{\sigma^2} \qquad (90)$$

H^2 azt mutatja meg, hogy a csoportképző (területi vagy minőségi) ismérv a mennyiségi ismérv szórásnégyzetének mekkora részét ($100 \cdot H^2$ százalékát) magyarázza meg.

A statisztikai gyakorlatban ismert a $H = \sqrt{H^2}$ mutató is, amely 0 és 1 közötti értéket vehet fel (ahogy H^2 is), de ez nem értelmezhető megoszlási viszonyszámként, csak a kapcsolat szorosságát jellemző 0 és 1 közötti értékként (százalékban nem fejezhető ki!).

47. példa

A 44. példa adatai alapján számítsuk ki a H^2 mutatót!

Figyelembe véve a 45. példa részeredményeit, (90) szerint:

$$H^2 = \frac{6}{6{,}8} = 0{,}8824\,.$$

A kapott eredmény értelmezése: a 10 munkás termelékenységre vonatkozó adatai nagy mértékben szóródnak. A dolgozók szakképzettségével a teljes szórásnégyzet 88,24%-át

tudjuk értelmezni, míg 11,76%-át a figyelembe nem vett más tényezőkkel (a dolgozók neme, kora, lakóhelye, stb.) és a véletlennel magyarázhatjuk.

A fenti eredmény négyzetgyöke:

$$H = \sqrt{0{,}8824} \;=\; 0{,}9394 \;\approx\; 0{,}94\,.$$

Ennek 1-hez közeli értéke nagyon szoros kapcsolatra utal, azaz a dolgozók szakképzettsége és termelékenységükre vonatkozó adataik között nagyon erős összefüggés van.

Rangkorreláció

A sorrendi mérési szintű ismérvek közötti kapcsolat egy (gyakran alkalmazott) mutatójával foglalkozunk a továbbiakban. Ezen ismérvek sorrendisége, rangsora hordoz információt. A két ordinális skálán mért ismérv 1 és N közötti rangjait (sorszámait) R_{x_i} -vel, illetve R_{y_i} -vel fogjuk jelölni. A kapcsolat szorosságát a (91) képlettel definiált ún. **SPEARMAN-féle rangkorrelációs együtthatóval** mérjük.

$$r_S = 1 - \frac{6\sum_{i=1}^{N}(R_{x_i} - R_{y_i})^2}{N(N^2-1)} \qquad (91)$$

A SPEARMAN-féle rangkorrelációs együttható abszolút értéke 1-nél nem nagyobb. Az $r_S = 0$ arra utal, hogy a rangsorok között nincs kapcsolat. A mutató negatív értéke esetén a két rangsor ellentétesen alakul, míg pozitív értéke esetén a két rangsor azonos irányban mozog. Ha $|r_S| = 1$, akkor a két ismérv rangsorai között determinisztikus kapcsolat van.

Ha egy változónak több egyforma értéke is előfordul, akkor milyen rangszámokat alkalmazzunk? Ilyenkor a megfelelő sorszámok számtani átlagát rendeljük az azonos értékekhez. Ezeket nevezzük **kapcsolt rangoknak**. (Lásd a 35. táblázatot.)

48. példa

Nappali tagozatos közgazdász hallgatók (egy vizsganapján) módszertani szigorlaton elért eredményeit a 34. táblázat tartalmazza.
Számítsuk ki a SPEARMAN-féle rangkorrelációs együtthatót!

A módszertani szigorlat eredményei

34. táblázat

Hallgató sorszáma	Matematika	Statisztika
	Pontszámok	
1.	23	4
2.	32	34
3.	42	32
4.	32	37
5.	45	42
6.	25	21
7.	41	41
8.	26	21
9.	43	27
10.	24	21
11.	43	26
12.	25	31
13.	26	27
14.	40	36
15.	45	43

A (91) képlethez szükséges részeredményeket a 35. táblázat tartalmazza.

A rangszámokhoz szükséges rangsorolást könnyen elvégezhetjük az Excel SORSZÁM(szám;hiv;sorrend) statisztikai függvény segítségével. Mivel ez a függvény nem határozza meg a kapcsolt rangokat, ezeket utólag nekünk kell (most már jóval kevesebb munkával) kiszámítani.

A rangkorrelációs együttható számításához szükséges részeredmények

35. táblázat

R_{x_i}	R_{y_i}	$\left(R_{x_i} - R_{y_1}\right)^2$
15,0	15,0	0,00
8,5	6,0	6,25
5,0	7,0	4,00
8,5	4,0	20,25
1,5	2,0	0,25
12,5	13,0	0,25
6,0	3,0	9,00
10,5	13,0	6,25
3,5	9,5	36,00
14,0	13,0	1,00
3,5	11,0	56,25
12,5	8,0	20,25
10,5	9,5	1,00
7,0	5,0	4,00
1,5	1,0	0,25
–	–	165,00

A 35. táblázat utolsó oszlopának összegét a (91) képletbe behelyettesítve azt kapjuk, hogy:

$$r_S = 1 - \frac{6 \cdot 165}{15^3 - 15} = 0{,}7054 .$$

Ez azt jelenti, hogy a két tantárgyból kapott osztályzatok rangsorai között jelentős pozitív irányú kapcsolat van.

Végezetül néhány megjegyzés a rangkorreláció mérőszámának kiszámításával kapcsolatban:

- a rangsorolásnál a sorbarendezés mindkét változónál azonos (csökkenő vagy növekvő) iránya nem befolyásolja r_S értékét;

- a rangsorolás pontosságát könnyen ellenőrizhetjük a $\sum_{i=1}^{N}\left(R_{x_i} - R_{y_i}\right) = 0$ összefüggés szerint;
- a változók X_i - Y_i, illetve Y_i - X_i jelölése irreleváns.

Korreláció

Két mennyiségi ismérv közötti (nem ok-okozat szerint vizsgált!) kapcsolat jellegére vonatkozó elemzés eszközei közül hármat ismertetünk.

Pontdiagram

Két mennyiségi ismérv közötti kapcsolatról, a mutatószámok számítása előtt, gyakran pontdiagram segítségével igyekszünk többet megtudni. Ekkor az együttesen előforduló (x_i, y_i) ismérvértékeket ábrázoljuk, és a kialakuló „pontfelhőből" következtetünk a kapcsolatra. A pontdiagram segítségével megállapítható a kapcsolat erőssége és iránya is. A korrelációs kapcsolat pozitív irányú, ha a pontdiagramon ábrázolt pontok a bal alsótól a jobb felső sarokig, negatív irányú, ha ezek a bal felsőtől a jobb alsó sarokig húzódnak. Minél keskenyebb a pontfelhő, annál erősebb a kapcsolat, függetlenül az irányától.

A kovariancia

A mennyiségi ismérvek közötti kapcsolat tényét és irányát a (92) alatt definiált ún. **kovariancia** segítségével is kifejezhetjük.

$$C_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N} \qquad (92)$$

Ez az ismérvértékek együtt-mozgását kifejező fontos mérőszám kétváltozós elsőrendű centrális momentumnak tekinthető.

Megjegyzés: a szorzásra érvényes kommutativitás miatt a kovariancia szimmetrikus mérőszám, azaz $C_{xy} = C_{yx}$. Függetlenség esetén 0-val egyenlő.

A továbbiakban az egyes ismérvértékek átlaguktól vett különbségére a következő jelölést vezetjük be:

$$d_{x_i} = x_i - \bar{x}\,, \tag{93}$$

illetve

$$d_{y_i} = y_i - \bar{y}\,. \tag{94}$$

A kovariancia képlete a fenti jelöléssel a következőképpen írható:

$$C_{xy} = \frac{\sum_{i=1}^{N} d_{x_i} d_{y_i}}{N}. \tag{95}$$

A (96) szerint egy ismérv önmagával vett kovarianciája nem más, mint a szórásnégyzete:

$$C_{xx} = \frac{\sum_{i=1}^{N} d_{x_i} d_{x_i}}{N} = \frac{\sum_{i=1}^{N} d_{x_i}^2}{N} = \sigma_x^2\,, \tag{96}$$

illetve

$$C_{yy} = \frac{\sum_{i=1}^{N} d_{y_i} d_{y_i}}{N} = \frac{\sum_{i=1}^{N} d_{y_i}^2}{N} = \sigma_y^2\,.$$

Az empirikus vizsgálatoknál azonban, egyszerűbb számítási módja miatt, gyakran a (97) képletet alkalmazzuk.

$$C_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x} \cdot \bar{y} \tag{97}$$

Az Excelben a kovarianciát a KOVAR(tömb1;tömb2;...) statisztikai függvény segítségével számíthatjuk ki.

A lineáris korrelációs együttható

Amennyiben a két ismérv között lineáris kapcsolat áll fenn, vagyis a pontdiagram pontjai megközelítőleg egy képzeletbeli egyenes körül csoportosulnak, akkor a (98) képlettel definiált ún. **lineáris korrelációs együttható** segítségével számszerűsíthetjük a kapcsolat

erősségét és irányát.

$$r = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{98}$$

A lineáris korrelációs együttható abszolút értéke 1-nél nem nagyobb. A 0-hoz közeli értéke a kapcsolat lazaságára vagy éppen hiányára utal. Az *r* negatív értékéből a két mennyiségi ismérv ellentétes irányú változására, míg pozitív értékéből azonos irányú együttmozgására következtethetünk.

Megjegyzés: a szorzásra érvényes kommutativitás miatt a lineáris korrelációs együttható szimmetrikus mérőszám, azaz $r_{xy} = r_{yx} = r$. Értéke százalékban nem fejezhető ki!

Az Excelben a lineáris korrelációs együtthatót a KORREL(tömb1;tömb2;...) statisztikai függvény segítségével számíthatjuk ki.[5)]

A mennyiségi ismérvek kapcsolatával részletesebben majd a 6. fejezetben foglalkozunk.

49. példa

A népmozgalmi arányszámok közül a csecsemőhalandóságra vonatkozó adatokat (ezrelékben), illetve a Magyarországra belépő külföldiek számát (ezer főben) a 36. táblázat tartalmazza.

Megjegyzés: a csecsemőhalandóság alatt az ezer élve születettre jutó egy éven aluli meghaltak számát értjük.

Számítsuk ki és értelmezzük a két változó közötti lineáris korrelációs együtthatót!

[5)] A KORREL(tömb1;tömb2;…) függvény mellett, az Excel PEARSON(tömb1;tömb2;…) függvénye is a lineáris korrelációs együtthatót számítja ki. Közöttük csak az argumentumok értelmezésében van különbség, az eredményük megegyezik.

A belépő külföldiek és a csecsemőhalandóság adatai

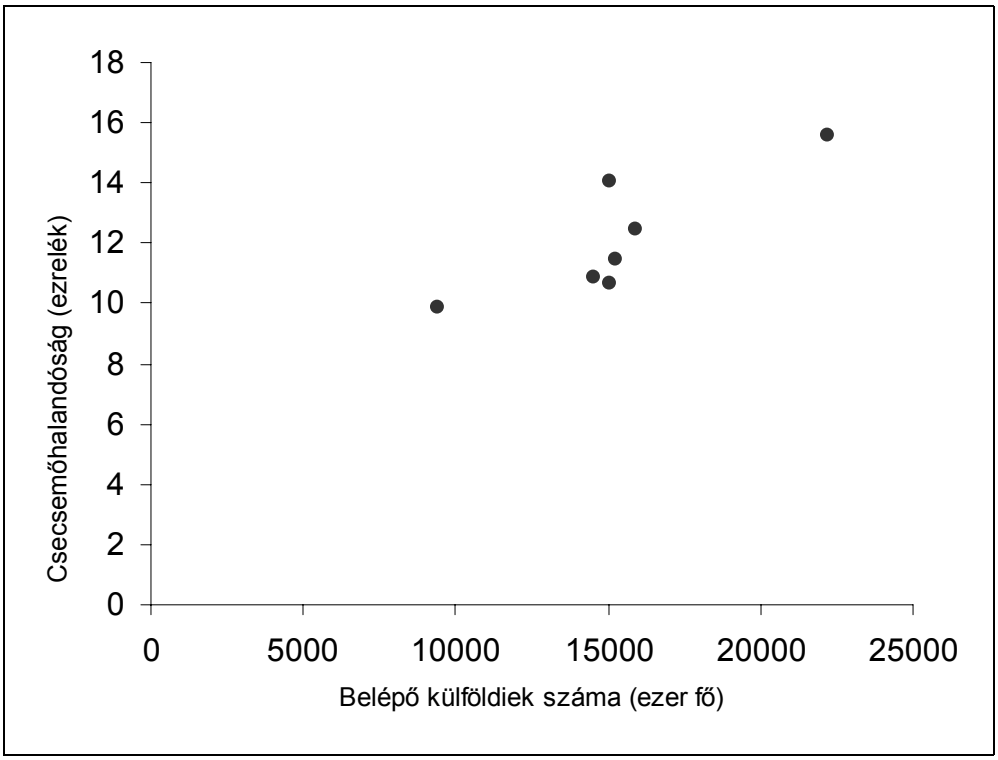
36. táblázat

Év	Belépő külföldiek	Csecsemőhalandóság
1991	22 194	15,6
1992	15 032	14,1
1993	15 901	12,5
1994	15 254	11,5
1995	15 010	10,7
1996	14 503	10,9
1997	9 397	9,9

Forrás: Magyar statisztikai zsebkönyv ’98, KSH, Bp., 1999.

Első lépésként ábrázoljuk az adatokat pontdiagramon.

A belépő külföldiek és a csecsemőhalandóság pontdiagramja



15. ábra

(Megjegyzés: a pontdiagramon csak 7 adatpárt ábrázoltunk, de az empirikus elemzéseknél nem volna szabad kevés számú megfigyelés alapján statisztikai összefüggéseket keresni.)

A lineáris korrelációs együttható kiszámításához szükséges részeredményeket a 37. táblázat tartalmazza.

Munkatábla az *r* kiszámításához

37. táblázat

Belépő külföldiek (x_i)	Csecsemő-halandóság (y_i)	x_i^2	y_i^2	$x_i \cdot y_i$
22 194	15,6	492 573 636	243,36	346 226,4
15 032	14,1	225 961 024	198,81	211 951,2
15 901	12,5	252 841 801	156,25	198 762,5
15 254	11,5	232 684 516	132,25	175 421,0
15 010	10,7	225 300 100	114,49	160 607,0
14 503	10,9	210 337 009	118,81	158 082,7
9 397	9,9	88 303 609	98,01	93 030,3
107 291	85,2	1 728 001 695	1 061,98	1 344 081,1

A (97) képlet szerint (két tizedesre kerekítve a végeredményt):

$$C_{xy} = \frac{1344081,1}{7} - \frac{107291}{7} \cdot \frac{85,2}{7} = 5456,62 .$$

Ha bevittük a 36. táblázat utolsó két oszlopában szereplő adatokat az Excelbe az **A2-B8** cellatartományba (a fejléceket az **A1** és **B1** cellák tartalmazzák), akkor a KOVAR(A2:A8;B2:B8) függvény alkalmazásával ugyanezt az eredményt kapjuk.

(Megjegyzés: a munkalapon a 36. táblázat első oszlopában levő adatok nem szerepelnek, mert a korrelációszámításnál ezekre nincs szükségünk.)

A változók szórásait az (51) képlet felhasználásával kapjuk:

$$\sigma_x = \sqrt{\frac{1728001695}{7} - \left(\frac{107291}{7}\right)^2} = 3454{,}23\,;$$

$$\sigma_y = \sqrt{\frac{1061{,}98}{7} - \left(\frac{85{,}2}{7}\right)^2} = 1{,}89\,.$$

A SZÓRÁSP(A2:A8), illetve a SZÓRÁSP(B2:B8) függvények alkalmazásával ugyanezt az eredményt kapjuk.

A kapott részeredmények felhasználásával, (98) szerint, a lineáris korrelációs együttható:

$$r = \frac{5456{,}62}{3454{,}23 \cdot 1{,}89} = 0{,}8358\,.$$

A KORREL(A2:A8;B2:B8) függvény alkalmazásával eredményként $r = 0{,}8363$ értéket kapunk. Az eltérés a két számítás eredménye között a kerekítésekből adódik.

A lineáris korrelációs együttható értékének értelmezése: *r* pozitív értéke arra utal, hogy a két vizsgált változó között pozitív korrelációs kapcsolat van. Ugyanez a következtetésünk a 15. ábrán látható pontfelhő elhelyezkedése alapján is.
Az *r* 1-hez közeli értéke alapján arra következtethetnénk, hogy a belépő külföldiek száma és a csecsemőhalandóság között nagyon szoros a kapcsolat. Ez azonban csak látszólag igaz, mert a két jelenség között nyilvánvalóan nincs összefüggés. Ez is arra utal, hogy nem szabad összekeverni a korrelációs kapcsolatot (a változók adatainak együttmozgását) az ok-okozati összefüggéssel!

Végezetül néhány megjegyzés a lineáris korreláció mérőszámainak kiszámításával kapcsolatban:

- a változók X_i - Y_i, illetve Y_i - X_i jelölése irreleváns;
- még egyszer hangsúlyozzuk, hogy a változók között lineáris összefüggést feltételezünk (a nem lineáris esetekkel a 6. fejezetben majd részletesebben foglalkozunk);
- a lineáris korrelációs együttható négyzete is értelmezhető, de ezt szintén a 6. fejezetben fogjuk tárgyalni.

5. Standardizálás és indexszámítás

5.1. Standardizálás

Az előző fejezetben a (71)-(72) képleteknél láttuk, hogy az összetett intenzitási viszonyszámok (például: $\overline{V}_0$ és $\overline{V}_1$) a részviszonyszámok súlyozott számtani vagy súlyozott harmonikus átlagaként írhatók fel. Minden súlyozott átlagra érvényes, hogy értékét az átlagolandó értékek abszolút nagysága és a súlyok relatív nagysága, a súlyarány határozza meg. Ezek alapján egy összetett viszonyszám értékét is két tényező befolyásolja:

- a részviszonyszámok nagysága és/vagy
- a részsokaságok súlyaránya, azaz a teljes sokaság összetétele.

Egy jelenség statisztikai elemzésekor gyakran kerül sor heterogén sokaságot jellemző átlagos színvonal időbeli vagy térbeli összehasonlítására. Ebben a fejezetben azzal foglalkozunk, hogy az összetett intenzitási viszonyszámok különbözőségét kialakító tényezők hatását számszerűen külön-külön kimutassuk. Az erre irányuló eljárást nevezzük **standardizálásnak**.

Figyelembe véve az összetett intenzitási viszonyszám értékét befolyásoló két tényezőt, ezek esetleges eltérése is két hatás eredőjével magyarázható:

- a részviszonyszámok különbözőségének hatásával és/vagy
- a súlyarányok (összetétel) különbözőségének hatásával.

A $\overline{V}_0$ és $\overline{V}_1$ összehasonlítása lehet

- térbeli, amikor azt vizsgáljuk, hogy mennyire különbözik két azonos módon részekre bontott sokaság;
- időbeli , amikor azt vizsgáljuk, hogy az összetett intenzitási viszonyszám hogyan változott egy adott időszakról (időpontról) egy másik időszakra (időpontra).

Az első esetben általában a két összetett viszonyszám különbségét ($K = \overline{V}_1 - \overline{V}_0$) bontjuk (a már említett) tényezők összegére, míg a második esetben a két összetett viszonyszám hányadosát ($I = \frac{\overline{V}_1}{\overline{V}_0}$) bontjuk ugyanezen tényezők szorzatára.

Különbség-felbontás

Mivel a két tényező hatása együttesen jelentkezik, ahhoz, hogy hatásuk külön-külön számszerűsíthető legyen, egyiküket mindig változatlannak, standardnak kell tekintenünk.

A részviszonyszámok közötti eltérésből eredő hatást, a fentieknek megfelelően, valamilyen állandó, standard súlyokkal dolgozva számszerűsítjük:

$$K' = \frac{\sum_{i=1}^{M} (B_S)_i (V_1)_i}{\sum_{i=1}^{M} (B_S)_i} - \frac{\sum_{i=1}^{M} (B_S)_i (V_0)_i}{\sum_{i=1}^{M} (B_S)_i}.$$

A továbbiakban, a képletek könnyebb áttekinthetősége érdekében, az összegzésre utaló indexeket elhagyjuk. Ennek megfelelően, a részviszonyszámok illetve az összetétel hatásának mérőszámaiként a (99)-(100) képleteket használjuk.

A **részhatás-különbség**:

$$K' = \frac{\sum B_S V_1}{\sum B_S} - \frac{\sum B_S V_0}{\sum B_S}. \tag{99}$$

Az **összetételhatás-különbség**:

$$K'' = \frac{\sum B_1 V_S}{\sum B_1} - \frac{\sum B_0 V_S}{\sum B_0}. \tag{100}$$

A **teljes különbség**:

$$K = \overline{V}_1 - \overline{V}_0 = K' + K''. \tag{101}$$

A fenti képletek az összetett viszonyszámok (71)-nek megfelelő számtani átlagképletét használják. Természetesen a (72)-nek megfelelő harmonikus átlagképletet használva is elvégezhető a standardizálás, de mi ezzel nem foglalkozunk.

Hányados-felbontás

Két összetett intenzitási viszonyszám hányadosának két index szorzatára bontását a különbségfelbontáshoz hasonló módon végezzük el.

A **részhatás-index**:

$$I' = \frac{\sum B_S V_1}{\sum B_S} : \frac{\sum B_S V_0}{\sum B_S}. \tag{102}$$

Az **összetételhatás-index**:

$$I'' = \frac{\sum B_1 V_S}{\sum B_1} : \frac{\sum B_0 V_S}{\sum B_0}. \tag{103}$$

A **teljes hatás indexe**:

$$I = \frac{\overline{V}_1}{\overline{V}_0} = I' \cdot I''. \tag{104}$$

Kérdés persze, hogy mit használjunk standard súlyoknak a (99) és (102) képletben, és mit használjunk a részviszonyszámok standard sorozatának a (100) és (103) képletben. Ezek megválasztásánál mindenképpen figyelembe kell vennünk, hogy (101) illetve (104) fennálljon. Gyakran használjuk súlyoknak például a következő kombinációkat:

$$B_S = B_0 \quad \text{és} \quad V_S = V_1,$$

illetve

$$B_S = B_1 \quad \text{és} \quad V_S = V_0.$$

Mivel különböző súlyozást használva némileg különböző eredményre juthatunk, a súlyozás módját K' és K'', illetve I' és I'' alsó indexében fogjuk jelölni. Például:

$$K_0' = \frac{\sum B_0 V_1}{\sum B_0} - \frac{\sum B_0 V_0}{\sum B_0}.$$

Megjegyzés: mind a különbség-felbontás, mind a hányados-felbontás során alkalmazott képletekben a súlyként szereplő *B* adatok helyettesíthetőek megoszlási viszonyszámaikkal.

Vigyázzunk arra, hogy a (100) és a (103) képletek az összetételváltozás hatását számszerűsítik, nem pedig magát az összetétel változását. Lehetséges ugyanis, hogy a sokaság szerkezete jelentősen átalakul, és ennek még sincs hatása az összetett viszonyszám változására (például azért, mert minden részviszonyszám azonos értékű).

50. példa

Egy vállalat dolgozóinak számáról és a béralapról a 38. táblázat adatai ismertek.

A vállalat béralapja és dolgozói létszáma

38. táblázat

Állomány-csoport	1998. január		1999. január	
	Béralap (Ft)	Létszám (fő)	Béralap (Ft)	Létszám (fő)
Szak-munkások	4 763 000	110	5 112 900	117
Betanított munkások	1 522 000	40	2 274 300	57
Segéd-munkások	1 652 400	51	4 788 800	146

Hasonlítsuk össze az 1998. januári és 1999. januári átlagbéreket az egyes kategóriákban és a vállalatnál! Mutassuk ki az eltérést okozó tényezők számszerű hatását!

Az átlagbéreket és a dolgozói létszámot a 39. táblázat tartalmazza.

A vállalat béralapja és dolgozói létszáma

39. táblázat

Állomány-csoport	1998. január		1999. január	
	Átlagbér (Ft)	Létszám (fő)	Átlagbér (Ft)	Létszám (fő)
Szak-munkások	43 300	110	43 700	117
Betanított munkások	38 050	40	39 900	57
Segéd-munkások	32 400	51	32 800	146
Összesen	39 490	201	38 050	320

Az összesen sorban szereplő átlagbéreket megkaphatjuk az összes béralap és a teljes dolgozói létszám hányadosaként, vagy az átlagbérek súlyozott átlagaként. Lásd a (70)-(72) képleteket.

$$\overline{V}_{1998} = \frac{4763000 + 1522000 + 1652400}{110 + 40 + 51} = \frac{7937400}{201} = 39490$$

$$\overline{V}_{1999} = \frac{117 \cdot 43700 + 57 \cdot 39900 + 146 \cdot 32800}{320} = 38050$$

A táblázat első látásra meghökkentő eredményt tartalmaz. Minden kategóriában nőttek az átlagbérek, a vállalat egészére nézve azonban csökkent az átlagos bérszínvonal. Az ok nyilvánvalóan az, hogy a gyengébben fizetett segédmunkások aránya nőtt és a legjobban fizetett szakmunkások aránya csökkent a vállaltnál, tehát megváltozott a foglalkoztatottak szerkezete.

Az átlagbér változása:

$$I = \overline{V}_{1999} : \overline{V}_{1998} = 38050 : 39490 = 0{,}9635\,.$$

Azt mondhatjuk, hogy 1998 januárjától 1999 januárjáig 3,65%-kal csökkent az átlagbér a vállalatnál.

Legyen $B_S = B_{1998}$.

A (102) képlet alapján:

$$I'_{1998} = \frac{\sum B_{1998} V_{1999}}{\sum B_{1998}} : \frac{\sum B_{1998} V_{1998}}{\sum B_{1998}} = 40178 : 39490 = 1{,}0174 .$$

Ez azt jelenti, hogy 1998 januárjától 1999 januárjáig az egyes kategóriák átlagbére átlagosan 1,74%-kal nőtt, ha az 1998. januári foglalkoztatási szerkezetet vesszük standardnak (változatlannak).

Legyen $V_S = V_{1999}$.

A (103) képlet alapján:

$$I''_{1999} = \frac{\sum B_{1999} V_{1999}}{\sum B_{1999}} : \frac{\sum B_{1998} V_{1999}}{\sum B_{1998}} = 38050 : 40178 = 0{,}9470 .$$

A foglalkoztatási szerkezetben bekövetkezett változások miatt az átlagbér 5,30%-kal csökkent a vállalatnál, mert nőtt a gyengébben fizetett kategóriában dolgozók aránya.

A (104) szerint a két hatás eredője:

$$I = I' \cdot I'' = 1{,}0174 \cdot 0{,}9470 = 0{,}9635 .$$

Végezetül, fontos szerepük miatt, még egyszer felhívjuk a figyelmet az alábbiakra:
a K'' illetve I'' nem csupán az összetételváltozás tényét fejezi ki, hanem azt, hogy az összetételváltozás hogyan hatott a vizsgált összetett viszonyszám változására.

5.2. Érték-, ár- és volumenindexek

Indexek

A 2.3. fejezetben említett viszonyszámok közül most részletesebben foglalkozunk a dinamikus viszonyszámokkal. Egy vizsgált jelenség (például ár, mennyiség és érték) adott időszakra vonatkozó relatív változása dinamikus viszonyszám. Ebben a fejezetben ezeket **indexeknek** fogjuk nevezni. A viszonyítás tárgyát **tárgyidőszaki**, a viszonyítás alapját **bázisidőszaki** adatnak nevezzük.

Egyfajta termék esetén megkülönböztetünk **egyedi érték-**, **ár-** és **volumenindexeket**. Ha egyidejűleg többfajta terméket vizsgálunk, akkor **együttes érték-**, **ár-** és **volumenindexekről** (vagy röviden **érték-**, **ár-** és **volumenindexekről**) beszélünk.

Egyedi indexek

Egy egyedi index
- egy adott fajta jószág
- bázisidőszakhoz viszonyított,
- tárgyidőszakban bekövetkező
- (rendszerint százalékban kifejezett)
- relatív változását mutatja.

Az egyedi ár-, volumen- és értékindexeket a (105)-(107) képletekkel definiáljuk.

$$i_p = \frac{p_1}{p_0} \tag{105}$$

$$i_q = \frac{q_1}{q_0} \tag{106}$$

$$i_v = \frac{v_1}{v_0} \tag{107}$$

Az egyes szimbólumok jelentése a következő:

q : a vizsgált jószág természetes mértékegységben kifejezett nagysága,

p : egységára,

v : értéke.

Mivel az érték a mennyiség és az ár szorzataként is értelmezhető:

$$\mathrm{v} = q \cdot p\,,$$

az egyedi indexek között fennáll a (108) összefüggés.

$$i_{\mathrm{v}} = i_q \cdot i_p \tag{108}$$

Heterogén sokaság összértékének meghatározása

Statisztikai elemzések során gyakran kell összehasonlítást végeznünk valamilyen heterogén, minőségileg különböző, de valamilyen szempontból mégis összetartozó javak összességei között. Az ilyen sokaságokat **aggregált sokaságoknak** nevezzük. Aggregált sokaság például a nemzeti össztermék, egy ország energiafelhasználása, állatállománya, stb. Ezek összevetése csak úgy lehetséges, ha nagyságukat valamilyen közös mértékegységben határozzuk meg. Kézenfekvő a pénzértékben való számbavétel (például a nemzeti összterméknél), de egyes aggregált sokaságok nagyságát más mértékegységben is kifejezhetjük. Egy ország energiafelhasználását például kőolaj-egyenértékben, vagy az állatállomány nagyságát meghatározott tömegű állatban, ún. számosállatban. A továbbiakban azt feltételezzük, hogy az összesítendő részsokaságok mennyisége és egységára adott, az aggregált sokaság összértéke (az ún. **aggregátum**) pedig:

$$\sum_{i=1}^{N} q_i p_i = \sum_{i=1}^{N} \mathrm{v}_i\,, \tag{109}$$

ahol:

q_i : az i-edik jószágféleség természetes mértékegységben kifejezett nagysága,

p_i : az i-edik jószágféleség egységára,

v_i : az i-edik jószágféleség értéke.

Nyilvánvaló tehát, hogy elemzéseinkben ezen három tényező fog szerepelni. Mivel a q_i mennyiségek általában időszakra vonatkoznak, ezért ekkor a p_i mennyiségekre, mint időszakra vonatkozó átlagárakra tekintünk, nem pedig időponthoz kötődő árra.

(Megjegyzés: a gyakoribb időbeli összehasonlítás mellett, az indexformulák területi összehasonlításra is alkalmasak, ekkor **területi indexekről** beszélünk. Ezekkel részletesebben nem foglalkozunk.)

Az együttes indexek definiálása két módszer szerint történhet. Ezek alapján megkülönböztetünk

- **aggregát-forma** és
- **átlag-forma** szerinti képleteket.

Indexek aggregát-formái

A most következőkben arra keressük a választ, hogy egy adott jószágkosár esetében hogyan változott annak

- értéke,
- mennyisége (volumene),
- árszínvonala.

Heterogén termékek összességére vonatkozó értékváltozást a (110) szerinti értékindex segítségével lehet számszerűsíteni.

$$I_{\mathrm{v}} = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum \mathrm{v}_1}{\sum \mathrm{v}_0} \qquad (110)$$

A (110) képletben az összegzésre utaló indexeket (a képlet könnyebb áttekinthetősége érdekében) elhagytuk. E szerint járunk el a továbbiakban is.

Ahhoz, hogy a volumen relatív változása számszerűsíthető legyen, vagy az árakat, vagy az értékeket változatlannak kell vennünk. Az aggregát-forma az árakat veszi változatlannak. Ha a tárgyidőszakban is a bázisidőszakra vonatkozó árakkal számolunk, akkor **bázisidőszaki súlyozású** vagy más néven **LASPEYRES-féle** (kiejtése: lászpejl) **volumenindexet** kapunk.

$$I_q^0 = \frac{\sum q_1 p_0}{\sum q_0 p_0} \tag{111}$$

Ha a bázisidőszakban is a tárgyidőszakra vonatkozó árakkal számolunk, akkor **tárgyidőszaki súlyozású** vagy más néven **PAASCHE-féle** (kiejtése: páse) **volumenindexet** kapunk.

$$I_q^1 = \frac{\sum q_1 p_1}{\sum q_0 p_1} \tag{112}$$

Az árszínvonal relatív változását hasonló módon, azonos mennyiségeket használó aggregátumok hányadosaiból képzett index segítségével tudjuk kimutatni. Ebben az esetben is használjuk a **bázisidőszaki súlyozású** vagy más néven **LASPEYRES-féle árindexet**:

$$I_p^0 = \frac{\sum p_1 q_0}{\sum p_0 q_0}, \tag{113}$$

valamint a **tárgyidőszaki súlyozású** vagy más néven **PAASCHE-féle árindexet**:

$$I_p^1 = \frac{\sum p_1 q_1}{\sum p_0 q_1}. \tag{114}$$

A (111)-(114) képleteket az indexek aggregát-formáinak nevezzük, mert aggregátumok hányadosai.

Megjegyzés: a (111) és (113) számlálójában, illetve a (112) és (114) képletek nevezőjében szereplő összegek **fiktív aggregátumok**, a többi **valós aggregátum**.

Az említett indexek között fennáll a (115) összefüggés.

$$I_{\mathrm{v}} = I_q^0 \cdot I_p^1 = I_q^1 \cdot I_p^0 \tag{115}$$

Megjegyzés: az empirikus elemzéseknél a (115) alkalmas arra, hogy (a már ismert) bármelyik két index segítségével a harmadikat is kiszámíthassuk.

Az indexek átlag-formái

A (111)-(114) aggregát-formában adott indexek felírhatók az egyedi indexek súlyozott átlagaként a (116)-(119) módon. Ezeket nevezzük az indexek átlag-formáinak.

$$I_q^0 = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0} = \frac{\sum i_q \mathrm{v}_0}{\sum \mathrm{v}_0} \tag{116}$$

$$I_q^1 = \frac{\sum p_1 q_1}{\sum \frac{q_0}{q_1} p_1 q_1} = \frac{\sum \mathrm{v}_1}{\sum \frac{\mathrm{v}_1}{i_q}} \tag{117}$$

$$I_p^0 = \frac{\sum \frac{p_1}{p_0} q_0 p_0}{\sum q_0 p_0} = \frac{\sum i_p \mathrm{v}_0}{\sum \mathrm{v}_0} \tag{118}$$

$$I_p^1 = \frac{\sum q_1 p_1}{\sum \frac{p_0}{p_1} q_1 p_1} = \frac{\sum \mathrm{v}_1}{\sum \frac{\mathrm{v}_1}{i_p}} \tag{119}$$

Megjegyzés: míg az aggregát-forma szerinti képleteknél a súlyozó tényező az ár illetve a volumen, az átlag-forma szerinti képleteknél mindig az érték a súlyozó tényező.

A (116) és (118) képlet a megfelelő egyedi indexek súlyozott számtani átlaga, míg a (117) és (119) a megfelelő egyedi indexek súlyozott harmonikus átlaga.

Az értékindex is kifejezhető átlag-formával:

$$I_{\mathrm{v}} = \frac{\sum \mathrm{v}_0 i_{\mathrm{v}}}{\sum \mathrm{v}_0} = \frac{\sum \mathrm{v}_1}{\sum \frac{\mathrm{v}_1}{i_{\mathrm{v}}}}. \tag{120}$$

Az átlag-formával kapcsolatosan, azaz a (116)-(120) képletek alkalmazását érintően, felhívjuk a figyelmet a következő (gyakorlati szempontból jelentős) tényre: a $p_0 q_0 = \mathrm{v}_0$, illetve $p_1 q_1 = \mathrm{v}_1$ nem csak valamilyen pénznemben kifejezett érték lehet, hanem a bázisidőszaki vagy tárgyidőszaki forgalom (százalékos vagy együtthatós formában adott) szerkezetét is jelentheti.

51. példa

Egy boltban háromféle terméket árusítanak. A forgalomról a 40. táblázat adatai állnak rendelkezésre.

A bolt forgalmának adatai

40. táblázat

Termékek	1998. január		1999. január		2000. január	
	Ár (Ft)	Mennyiség (t)	Ár (Ft)	Mennyiség (t)	Ár (Ft)	Mennyiség (t)
A	53	110	65	96	68	94
B	81	175	96	176	105	162
C	159	23	176	34	180	35

Ezek alapján számítsuk ki a forgalom értékét 1999 januárjára (1998 januárjához viszonyítva) az egyes termékekből külön-külön és a három termékre együttvéve! Számítsunk egyedi érték-, ár- és volumenindexeket! Állapítsuk meg az együttes érték-, ár- és volumenindexet LASPEYRES- és PAASCHE-féle formulával is! Írjuk fel az indexek közötti összefüggéseket!

A forgalom értékét 1998 és 1999 januárjára a 41. táblázat tartalmazza.

A forgalom értéke 1998 januárjában (v_0) és 1999 januárjában (v_1)

41. táblázat

Termékek	$v_0 = q_0 \cdot p_0$	$v_1 = q_1 \cdot p_1$
A	5 830	6 240
B	14 175	16 896
C	3 657	5 984
Összesen	23 662	29 120

A vizsgált időszakban a bolt forgalmának értéke az adott termékcsoportból 23 662 Ft-ról 29 120 Ft-ra növekedett.

Számítsuk ki az egyes termékekre vonatkozó egyedi érték-, ár- és volumenindexeket! Használjuk a (105)-(107) képletek által leírt összefüggéseket! Az eredményeket a 42.

táblázat tartalmazza.

Az egyedi indexek együttható formájában kifejezve

42. táblázat

Termékek	i_{v}	i_p	i_q
A	1,070	1,226	0,873
B	1,192	1,185	1,006
C	1,636	1,107	1,478

A 42. táblázat adatai a következőképpen értelmezhetőek: azt mondhatjuk, hogy az A termék ára 1998 januárjától 1999 januárjáig 22,6%-nőtt, míg az eladott mennyisége 12,7%-kal csökkent. Az A termék forgalmának értéke így 7,0%-kal nőtt. A B és a C termékekre vonatkozó adatok hasonlóan értelmezhetőek.

A három termék együttes forgalmára vonatkozó értékindex a (110) képlet és a 41. táblázat összesen sora alapján:

$$I_{\mathrm{v}} = \frac{29120}{23662} = 1{,}231\,.$$

Az adott boltban a vizsgált termékcsoportra a vásárlók 23,1%-kal költöttek többet 1999 januárjában, mint 1998 hasonló időszakában.

Az együttes árindexeket kiszámíthatjuk a (113)-(114) képletekkel:

$$I_p^0 = \frac{110 \cdot 65 + 175 \cdot 96 + 23 \cdot 176}{23662} = \frac{27998}{23662} = 1{,}183\,;$$

$$I_p^1 = \frac{29120}{96 \cdot 53 + 176 \cdot 81 + 34 \cdot 159} = \frac{29120}{24750} = 1{,}177\,;$$

vagy a (118)-(119) átlag-forma szerinti képletekkel is:

$$I_p^0 = \frac{1{,}226 \cdot 5830 + 1{,}185 \cdot 14175 + 1{,}107 \cdot 3657}{23662} = 1{,}183\,;$$

$$I_p^1 = \frac{29120}{\frac{6240}{1{,}226} + \frac{16896}{1{,}185} + \frac{5984}{1{,}107}} = 1{,}177\,.$$

Ha a forgalom 1999 januárjában ugyanolyan mennyiségű és szerkezetű lett volna, mint 1998 januárjában, akkor csak az árváltozások miatt 18,3%-kal költöttek volna többet a vizsgált termékcsoportra az adott boltban. Ha a fogyasztás már 1998-ban olyan nagyságú és szerkezetű lett volna, mint 1999-ben, az átlagos árszínvonal növekedése 17,7%-os lett volna 1998 januárjához viszonyítva.

Az együttes volumenindexek a (111)-(112) képletek alapján:

$$I_q^0 = \frac{24750}{23662} = 1{,}046;$$

$$I_q^1 = \frac{29120}{27998} = 1{,}040.$$

Ha az 1998. januári árakat tekintjük összehasonlító árnak, akkor 4,6%-kal növekedett a forgalom mennyisége az adott termékcsoportból, az 1999. januári árakkal számolva pedig 4,0%-kal nőtt a vizsgált termékek eladott mennyisége.

Összefüggések:

$$I_{\mathrm{v}} = I_q^0 \cdot I_p^1 = 1{,}046 \cdot 1{,}177 = 1{,}231;$$

$$I_{\mathrm{v}} = I_q^1 \cdot I_p^0 = 1{,}040 \cdot 1{,}183 = 1{,}230.$$

(Megjegyzés: kerekítési hibán belül a két eredmény megegyezik.)

Deflálás

A gazdaságstatisztikában nagy jelentőségű a következő művelet:

$$\frac{\sum q_1 p_1}{I_p}. \qquad (121)$$

A fenti összefüggés számlálójában levő tárgyidőszakra vonatkozó értéket **folyóáras aggregátumnak** nevezzük. Valamely aggregátum árindexszel való osztása a **deflálás**. Egy folyóáras adat deflátor árindexszel való osztásakor bázisidőszaki árszínvonalon kifejezett aggregátumhoz jutunk, melyet a folyóáras aggregátum **reálértékének** nevezünk.

Ha a (121) összefüggés számlálójába 1-et írunk, vagyis képezzük az árindex reciprokát, akkor azt kapjuk meg, hogy egy pénzegység a tárgyidőszakban mennyit ér bázisidőszaki árszínvonalon számítva. Ez az adott pénznem **vásárlóerejének** változását adja meg.

Árollók

A statisztikai elemzésekben előfordul, hogy bizonyos indexek összehasonlítására kerül sor, és ilyenkor ezt szintén hányados-képzéssel tesszük meg. Két árindex hányadosát **árollónak** nevezzük. A két legfontosabb árolló az agrárolló és a külkereskedelmi cserearány-index.

Az **agrárolló** a mezőgazdasági termeléshez felhasznált iparcikkek beszerzési árindexének és a mezőgazdasági termékek értékesítési árindexének hányadosa.

A **külkereskedelmi cserearány-index** az export árváltozásának az import árváltozásához viszonyított arányát mutatja.

Indexpróbák

Mivel az ár- és volumenindexeket többféleképpen is kiszámíthatjuk (attól függően, hogy milyen mennyiséget, vagy milyen árat tekintünk összehasonlítónak), a különböző indexekkel szemben különféle követelményeket fogalmazunk meg. Ezek elősegíthetik egy jelenség tömör, számszerű jellemzésére használható indexek közötti választást.

A fontosabb **indexpróbák** a következőek:

- **összemérhetőségi próba**: az index értéke legyen független a mennyiségi adatok mértékegységétől;
- **időpróba**: az időszakok felcserélésével kapott index és az eredeti index között reciprok összefüggés álljon fenn;
- **tényezőpróba**: az ugyanazon típusú ár- és volumenindex szorzata legyen egyenlő az értékindexszel;
- **átlagpróba**: az index az egyedi indexek valamilyen átlaga legyen;
- **láncpróba**: indexsorok esetében a láncindexek szorzata legyen egyenlő az ugyanazon formulával számított bázisindexszel.

Az eddig megismert LASPEYRES-féle és PAASCHE-féle indexek nem tesznek eleget az időpróbának, a tényezőpróbának és a láncpróbának. A statisztikai irodalomban ismert FISHER-féle index (jele: I^F) eleget tesz a fenti követelményeknek. A (122)-(123) képlettel

definiált index a LASPEYRES-féle és PAASCHE-féle indexek mértani átlagával számol. Lásd a (33) képletet.

FISHER-féle volumenindex

$$I_q^F = \sqrt{I_q^0 I_q^1}\,, \tag{122}$$

FISHER-féle árindex

$$I_p^F = \sqrt{I_p^0 I_p^1}\,. \tag{123}$$

A két index szorzata az értékindexszel egyenlő.

$$I_{\mathrm{v}} = I_q^F \cdot I_p^F$$

Megjegyzés: az említett indexfajták (LASPEYRES-, PAASCHE- és FISHER-féle) mellett a statisztikai irodalomban még sok más indexfajta is ismert, de ezekkel könyvünkben nem foglalkozunk.

Indexsorok

A gyakorlatban gyakran kerül sor kettőnél több aggregátum összevetésére. Indexek kettőnél több időszakra vonatkozó összefüggő sorozatát **indexsornak** nevezzük. Az indexsoroknak az alábbi fajtáit különböztetjük meg.

- Attól függően, hogy milyen jelenség változását mutatják az indexek, beszélünk érték-, ár- és volumen-indexsorokról.
- Az időszakok összehasonlításának módja szerint most is megkülönböztetjük a bázis- és lánc-indexsorokat. Ha mindig egy rögzített aggregátumhoz viszonyítjuk a különböző időszakokhoz tartozó aggregátumokat, akkor bázis-indexsort kapunk. A lánc-indexsorok számításakor a viszonyítás alapja (általában) a megelőző időszakhoz tartozó aggregátum.
- Az indexsorok a súlyozás módja szerint is különbözhetnek egymástól**. Állandó súlyú indexsorról** beszélünk, ha a súlyok (volumenindexek esetén az árak, árindexek esetén a mennyiségek) az egész indexsorban, tehát az összes aggregátumban azonosak. A **változó súlyú indexsor** tagjaiban a súlyok más-más időszakból származnak (az indexeken belül, a számlálóban és a nevezőben természetesen ekkor is azonosak).

Az előbbiek miatt az indexsorok tagjainak megkülönböztetésére a következő összetett

jelölést használjuk:

$$I_b^a(c/d) \qquad a,c,d: 0,1,...,t,...,T \qquad b: p,\ q,\ \text{v}; \tag{124}$$

ahol

a: azt jelöli, hogy a súlyok melyik t időszakból származnak,

b: azt jelöli, hogy az indexek milyen jelenség relatív változását mutatják,

c,d: azt jelöli, hogy mely időszakokhoz tartoznak az egymáshoz viszonyított aggregátumok.

LASPEYRES-, PAASHE- és FISHER-féle indexeknek egyértelműen csak a változó súlyú lánc-indexsorok tagjai nevezhetőek.

Változó súlyú lánc-indexsorok:

	Volumen	Ár
LASPEYRES:	$I_q^{t-1}(t/t-1)=\dfrac{\sum q_t p_{t-1}}{\sum q_{t-1}p_{t-1}}$	$I_p^{t-1}(t/t-1)=\dfrac{\sum q_{t-1} p_t}{\sum q_{t-1}p_{t-1}}$
PAASCHE:	$I_q^{t}(t/t-1)=\dfrac{\sum q_t p_t}{\sum q_{t-1}p_t}$	$I_p^{t}(t/t-1)=\dfrac{\sum q_t p_t}{\sum q_t p_{t-1}}$
FISCHER:	$I_q^F=\sqrt{I_q^{t-1}(t/t-1)\cdot I_q^t(t/t-1)}$	$I_p^F=\sqrt{I_p^{t-1}(t/t-1)\cdot I_p^t(t/t-1)}$

Arra a kérdésre, hogy állandó vagy változó súlyozású indexeket számítsunk-e nem lehet egyértelmű választ adni. Ha a változó súlyú bázisindexekből láncindexeket vagy a lánc-indexsorból bázisindexeket származtatunk, akkor lehetséges, hogy a kapott index nem tesz eleget az átlagpróbának, vagyis nem a megfelelő egyedi indexek átlaga, esetleg nincs is az azok által meghatározott intervallumban. Az állandó súlyú indexsoroknál ilyen probléma nem fordul elő. Ezek alkalmazásakor a gondot a súlyok elavulása okozza. Ez azt jelenti, hogy egy hosszabb időszak esetén a rögzített súlyarányok egyre távolabb kerülnek az összehasonlított időszakokra jellemző tényleges arányoktól, sőt a termékek folyamatos cserélődése miatt szűkülhet az összehasonlítható termékek köre. Az említett problémák miatt a gyakorlatban a szakaszosan állandó súlyú indexsorokat szoktuk alkalmazni. Ekkor a súlyokat 5-10 évenként cseréljük.

52. példa

Számítsuk ki az 51. példa adataiból az 1998. januári bázisú 1998. és 1999. januári állandó súlyú ár-indexsort valamint a FISHER-féle változó súlyú volumen-indexsort.
Az 1998. januári bázisú állandó súlyú ár-indexsor 1998. januári állandó súllyal:

1998: $$I_p^{1998}{}_{(1998/1998)} = \frac{\sum q_{1998} \cdot p_{1998}}{\sum q_{1998} \cdot p_{1998}} = 100{,}0\%$$

1999: $$I_p^{1998}{}_{(1999/1998)} = \frac{\sum q_{1998} \cdot p_{1999}}{\sum q_{1998} \cdot p_{1998}} = 118{,}3\%$$

2000: $$I_p^{1998}{}_{(2000/1998)} = \frac{\sum q_{1998} \cdot p_{2000}}{\sum q_{1998} \cdot p_{1998}} = 126{,}8\%$$

Az 1998. januári bázisú állandó súlyú ár-indexsor 1999. januári állandó súllyal:

1998: $$I_p^{1999}{}_{(1998/1998)} = \frac{\sum q_{1999} \cdot p_{1998}}{\sum q_{1999} \cdot p_{1998}} = 100{,}0\%$$

1999: $$I_p^{1999}{}_{(1999/1998)} = \frac{\sum q_{1999} \cdot p_{1999}}{\sum q_{1999} \cdot p_{1998}} = 117{,}7\%$$

2000: $$I_p^{1999}{}_{(2000/1998)} = \frac{\sum q_{1999} \cdot p_{2000}}{\sum q_{1999} \cdot p_{1998}} = 125{,}8\%$$

A FISHER-féle változó súlyú volumen-indexsor:

1999:
$$I_q^F{}_{(1999/1998)} = \sqrt{I_q^{1998}{}_{(1999/1998)} \cdot I_q^{1999}{}_{(1999/1998)}} =$$
$$= \sqrt{\frac{\sum q_{1999} \cdot p_{1998}}{\sum q_{1998} \cdot p_{1998}} \cdot \frac{\sum q_{1999} \cdot p_{1999}}{\sum q_{1998} \cdot p_{1999}}} = 104{,}3\%$$

2000:
$$I_q^F{}_{(2000/1999)} = \sqrt{I_q^{1999}{}_{(2000/1999)} \cdot I_q^{2000}{}_{(2000/1999)}} =$$
$$= \sqrt{\frac{\sum q_{2000} \cdot p_{1999}}{\sum q_{1999} \cdot p_{1999}} \cdot \frac{\sum q_{2000} \cdot p_{2000}}{\sum q_{1999} \cdot p_{2000}}} = 95{,}5\%$$

5.3. A BORTKIEWICZ-féle összefüggés

Az ugyanazon adatokból különböző típusú formulával kiszámított ár- és volumenindexek általában eltérő eredményt adnak. A továbbiakban a LASPEYRES-féle és PAASCHE-féle indexek közötti összefüggést vizsgáljuk részletesebben. Erre vonatkozik a **BORTKIEWICZ-tétel** néven ismert (125) összefüggés.

$$\frac{I_q^1}{I_q^0} = \frac{I_p^1}{I_p^0} = 1 + \mathrm{v}_{i_q} \cdot \mathrm{v}_{i_p} \cdot r_{i_q i_p}, \tag{125}$$

ahol:

v_{i_q} és v_{i_p} : az egyedi volumen- és árindexek relatív szórása,

$r_{i_q i_p}$: az egyedi volumen- és árindexek közötti lineáris korrelációs együttható.

A (125) összefüggés azonban csak akkor érvényes, ha a jobboldalán szereplő minden mutatószámot a v_0 súlyok segítségével számítjuk ki!

Az egyedi indexek szórása az (52) szórásképlet alapján a (126) illetve a (127) módon írható fel.

$$\sigma_{i_q} = \sqrt{\frac{\sum \mathrm{v}_0 \left(i_q - I_q^0 \right)^2}{\sum \mathrm{v}_0}} \tag{126}$$

$$\sigma_{i_p} = \sqrt{\frac{\sum \mathrm{v}_0 \left(i_p - I_p^0 \right)^2}{\sum \mathrm{v}_0}} \tag{127}$$

Ezeket felhasználva, a relatív szórás (54) képlete alapján, az egyedi indexek relatív szórását a (128) és a (129) képlet tartalmazza.

$$\mathrm{v}_{i_q} = \frac{\sigma_{i_q}}{I_q^0} \tag{128}$$

$$\mathrm{v}_{i_p} = \frac{\sigma_{i_p}}{I_p^0} \tag{129}$$

Az egyedi indexek közötti lineáris korrelációs együttható kiszámításához szükség van a

kovarianciára. A (92) képletnek megfelelően

$$C_{i_q i_p} = \frac{\sum \mathrm{v}_0 \left(i_q - I_q^0\right)\left(i_p - I_p^0\right)}{\sum \mathrm{v}_0}. \qquad (130)$$

A lineáris korrelációs együtthatót (98) alapján a (131) képlet definiálja.

$$r_{i_q i_p} = \frac{C_{i_q i_p}}{\sigma_{i_q}\sigma_{i_p}} \qquad (131)$$

Vizsgáljuk most meg, hogy a BORTKIEWICZ-féle összefüggés baloldalán álló hányadosok mikor lehetnek egynél kisebbek. Ez pontosan akkor áll fenn, ha az egyedi indexek közötti lineáris korrelációs együttható értéke negatív, hiszen a relatív szórások bizonyosan nem negatívak. Az $r_{i_q i_p}$ együttható negatív előjele azt jelzi, hogy az egyedi ár- és volumenindexek (általában) ellentétes irányban változnak. Ez pedig, a közgazdaságtanból ismert helyettesítési hatás következtében, a valóságban majdnem mindig így is van. Emiatt általános jelenség, hogy a bázisidőszaki súlyozású indexek nagyobbak a tárgyidőszaki súlyozású megfelelő indexeknél.

53. példa

Számszerűsítsük az 51. példában szereplő egyedi indexek közötti sztochasztikus kapcsolat erősségét, írjuk fel a BORTKIEWICZ-féle összefüggést!

Először számítsuk ki az egyedi ár- és volumenindexek szórását a (126)-(127) képletek segítségével. (Az egyedi indexeket, a bázisidőszaki forgalom értékét, valamint a LASPEYRES-féle indexeket már az 51. példában kiszámítottuk.)

$$\sigma_{i_p} = \sqrt{\frac{5830\cdot(1{,}226-1{,}183)^2 + 14175\cdot(1{,}185-1{,}183)^2 + 3657\cdot(1{,}107-1{,}183)^2}{23662}} = 0{,}037;$$

$$\sigma_{i_q} = \sqrt{\frac{5830\cdot(0{,}873-1{,}046)^2 + 14175\cdot(1{,}006-1{,}046)^2 + 3657\cdot(1{,}478-1{,}046)^2}{23662}} = 0{,}193.$$

Ezeket az eredményeket felhasználva (128)-(129) szerint az egyedi indexek relatív szórása a következő:

$$v_{i_q} = \frac{0{,}193}{1{,}046} = 0{,}185;$$

$$v_{i_p} = \frac{0{,}037}{1{,}183} = 0{,}031.$$

Az egyedi indexek közötti lineáris korrelációs együttható kiszámításához szükség van az egyedi indexek kovarianciájára. Ezt a (130) képlet segítségével tudjuk kiszámítani.

$$C_{i_q i_p} = \frac{5830 \cdot (0{,}873 - 1{,}046) \cdot (1{,}226 - 1{,}183) + ... + 3657 \cdot (1{,}478 - 1{,}046) \cdot (1{,}107 - 1{,}183)}{23662} =$$

$$= -0{,}007$$

Az egyedi indexek közötti kapcsolat szorosságát kifejező lineáris korrelációs együtthatót a (131) képlet alkalmazásával nyerjük.

$$r_{i_q i_p} = \frac{-0{,}007}{0{,}037 \cdot 0{,}193} = -0{,}980$$

A 42. táblázat alapján már láthattuk, hogy az A termék ára emelkedett legnagyobb mértékben, míg legkevésbé a C termék drágult. A mennyiségi változás ezzel ellentétes irányú volt, az A termékből vásárolt mennyiség 12,7%-kal csökkent, míg a relatív módon leginkább olcsóbbá vált (abszolút mértékben persze drágult) C termékből 47,8%-kal nőtt a kereslet. Ez a már említett helyettesítési hatás. Annak mértékét, hogy milyen erős a kapcsolat az egyedi ár- és volumenindexek között a lineáris korrelációs együtthatóval tudjuk kifejezni. Láthatjuk, hogy a kapcsolat igen erős, és természetesen negatív irányú.

A BORTKIEWICZ-féle összefüggés:

$$\frac{1{,}040}{1{,}046} = \frac{1{,}177}{1{,}183} = 1 + 0{,}185 \cdot 0{,}031 \cdot (-0{,}980) = 0{,}994;$$

azaz a PAASCHE- és a LASPEYRES-féle indexek hányadosa 1-nél kisebb.

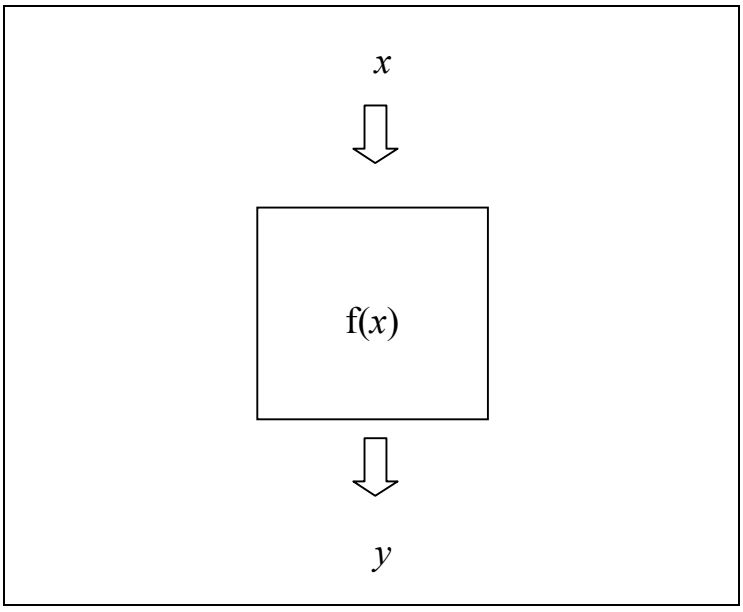
Fontossága miatt kiemeljük a következő törvényszerűséget: az indexek gyakorlati alkalmazásakor általában igaz, hogy a tárgyidőszaki súlyozású indexek (I_p^1 vagy I_q^1) a bázisidőszaki súlyozású indexeknél (I_p^0 vagy I_q^0) kisebbek.

6. Kétváltozós regresszió- és korrelációszámítás

6.1. Lineáris regresszió

A 4. fejezetben már vizsgáltuk a különböző típusú ismérvek közötti kapcsolatokat, így a mennyiségi ismérvek közötti kapcsolatot is. Foglalkoztunk azzal, hogy miként lehet megvizsgálni, hogy két ismérv között létezik-e kapcsolat, és (ha igen, akkor) ennek erősségét (és irányát) hogyan lehet számszerűsíteni. Nem foglalkoztunk viszont részletesen azzal, hogy sztochasztikus kapcsolat esetén az egyik ismérv által hordozott többletinformációt hogyan tudnánk felhasználni a másik ismérv értékeinek meghatározására. A korrelációszámítás során tehát, csak a kapcsolat erősségét vizsgáltuk, és az együttmozgást szimmetrikus mutatókkal számszerűsítettük. Az összefüggéseket ok-okozati kapcsolattal leíró módszert **regressziószámításnak** nevezzük. Ennek megfelelő illusztráció a 16. ábrán látható.

A regressziószámítás grafikus modellje



16. ábra

Amint látható, a bemeneti (ok) és a kimeneti (okozat) adatok összefüggése egyértelmű, azaz szerepük nem cserélhető fel. Az ezeket összekötő f(x) funkcionális operátor egy fekete dobozként is felfogható. A regressziószámítás feladata ennek identifikálása.

A regressziós egyenes

Induljunk ki most is az adatok pontdiagramjából, és tegyük fel, hogy ezekre nagyjából ráilleszthető egy egyenes. Célunk az lesz, hogy a pontokhoz legközelebbi egyenest megtaláljuk. Azt, hogy melyik egyenes tekinthető a legközelebbinek többféleképpen is meghatározhatjuk. Már a pontok és az egyenesek távolságát is többféleképpen mérhetjük. Például, egy pont és egy egyenes távolságát megállapíthatjuk az adott pontból az illesztett egyenesre (ún. **regressziós egyenesre**) bocsátott merőlegesen mérve (ahogy azt geometriailag meghatároznánk), vagy a pontból a regressziós egyenesig húzott vízszintesen (vagyis az X tengellyel párhuzamosan), illetve függőlegesen (vagyis az Y tengellyel párhuzamosan). A legegyszerűbben az utóbbi módon tudjuk meghatározni az egyenes és a pont távolságát, hiszen ez $\left|\overset{*}{y}_i - y_i\right|$, ha $\overset{*}{y}_i$-vel jelöljük az x_i-hez tartozó y_i empirikus érték (regressziós függvény alapján kiszámított) elméleti megfelelőjét.

Még mindig kérdés azonban, hogy ezen távolságokat hogyan összegezzük, és ezt az összeget hogyan minimalizáljuk. A statisztikai gyakorlat erre vonatkozóan legtöbbször az ún. **legkisebb négyzetek módszerét** (LNM) alkalmazza. E szerint a távolságok négyzetösszegét kell minimalizálni, és ez alapján meghatározni a megfelelő egyenest.

Legyen a keresett egyenes $y = \beta_0 + \beta_1 x$ alakú. A β_0 és β_1 együtthatókat **regressziós paramétereknek** nevezzük. A β_1 a **regressziós együttható** vagy **regressziós koefficiens**. Az *y* változót **eredményváltozónak**, az *x* változót **magyarázóváltozónak** hívjuk.

Amennyiben a kapcsolat a két változó között sztochasztikus, akkor a regressziós egyenes által meghatározott értékek (általában) eltérnek a tényleges értékektől. Ezt az eltérést **hibatagnak** nevezzük, és ε-nal jelöljük.

A fentiek szerint felírható a következő összefüggés:

$$y = \beta_0 + \beta_1 x + \varepsilon \, .$$

A gyakorlatban gyakran úgy végezzük el a regressziós illesztést, hogy a sokaságból csak néhány (x_i, y_i) adatpár ismert (i=1,2,…,n < N), ezek alapján (mintabeli információk) határozzuk meg a sokaságra vonatkozó regressziófüggvényt, illetve paramétereit. A

megfigyelések forrásaként rendszerint vagy idősort vagy ún. **keresztmetszeti adatokat** használunk. Ez utóbbi azonos időpontra és különböző helyszínekre vonatkozó információkat jelent.

A továbbiakban (terminológiánkban) feltételezzük, hogy a kapcsolat vizsgálata során csak egy részmegfigyelés eredménye (minta) áll rendelkezésünkre. Ekkor a $\hat{\beta}$ regressziós paraméterek a tényleges β paraméterek becsült értékei. A hibatagok becsült értékeit **reziduumoknak** nevezzük, és *e*-vel jelöljük.

A fentiek szerint

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad i=1,2,\ldots,n \qquad 2<n<N; \tag{132}$$

illetve

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{133}$$

azaz

$$e_i = y_i - \hat{y}_i .$$

A legkisebb négyzetek módszere és a normálegyenletek

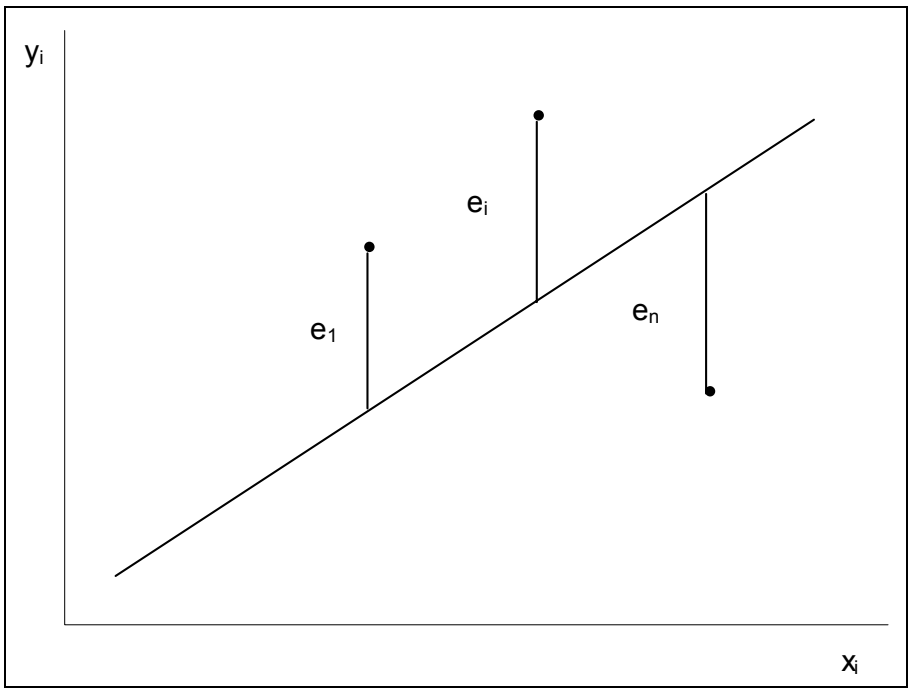
Az LNM szerint minimalizálnunk kell a

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

összeget.

(Lásd a 17. ábrát!)

Regressziós egyenes ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$)



17. ábra

Olyan $\hat{\beta}_0$ és $\hat{\beta}_1$ konstansokat keresünk, amelyre $\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ minimális. Ezt könnyen megkaphatjuk, ha meghatározzuk a fenti összeg β_0 illetve β_1 szerinti parciális deriváltját, és ezeket egyenlővé tesszük 0-val. Az így kapott egyenleteket nevezzük majd **normálegyenleteknek**.

A két normálegyenletből álló egyenletrendszer az alábbi.

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad (134)$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \qquad (135)$$

A normálegyenletek megoldásával a regressziós paraméterek egyértelműen meghatározhatóak.

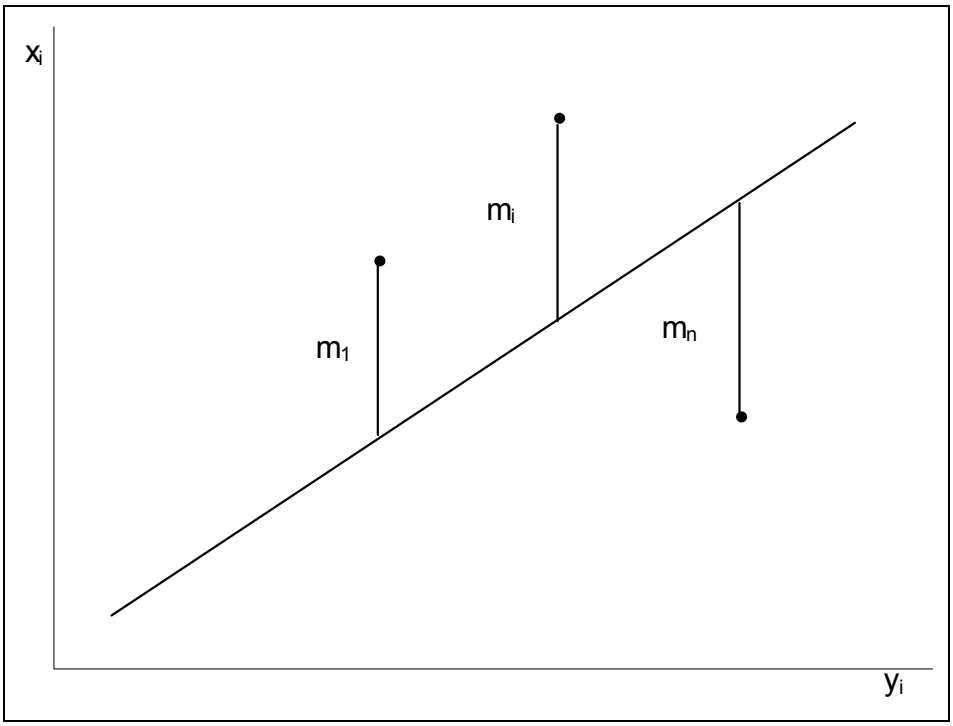
A (134) egyenlet mindkét oldalát *n*-nel osztva:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} . \qquad (136)$$

Tehát az $(\bar{x}, \bar{y})$ pont rajta van a regressziós egyenesen.

Írjuk fel most azt a regressziós egyenest, amelyben *y* a magyarázóváltozó, és *x* az eredményváltozó!

Regressziós egyenes ($\hat{x} = \hat{\gamma}_0 + \hat{\gamma}_1 y$)



18. ábra

Itt most az előzőekhez hasonlóan $\sum_{i=1}^{n} m_i^2$ -et kell minimalizálni. (Lásd a 18. ábrát!) A keresett regressziós egyenes legyen a következő alakú:

$$\hat{x} = \hat{\gamma}_0 + \hat{\gamma}_1 y . \qquad (137)$$

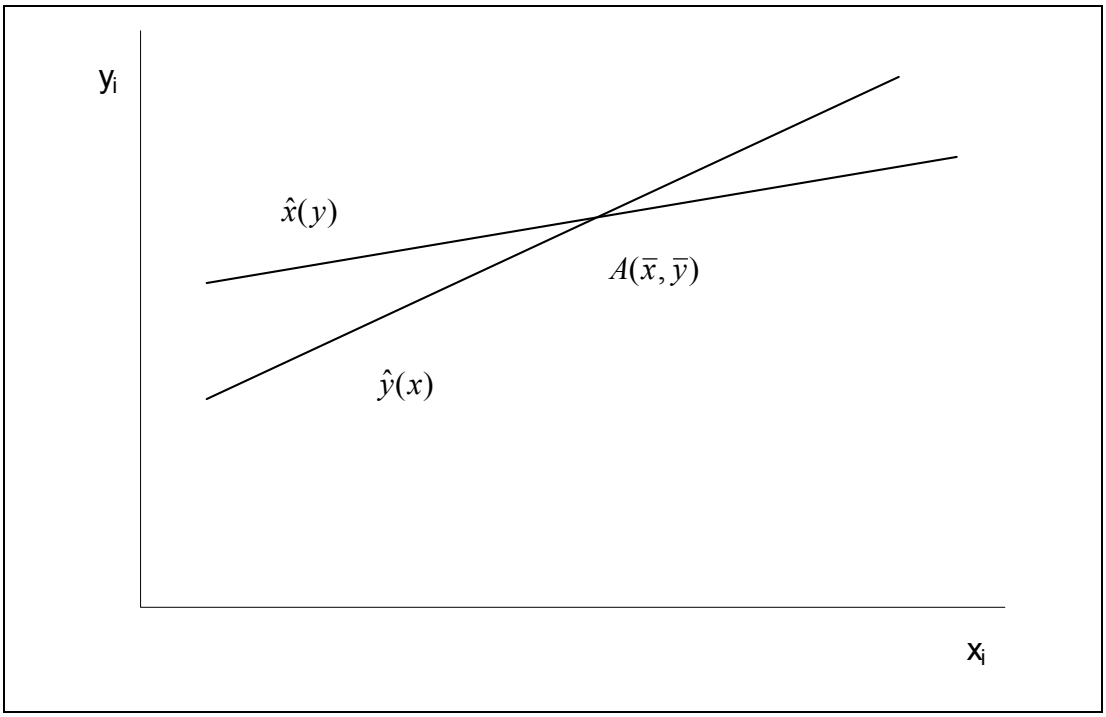
A $\hat{\gamma}$ paraméterek meghatározásához most is a (134)-(135)-höz hasonló

normálegyenletekhez jutunk.

$$\sum_{i=1}^{n} x_i = n\hat{\gamma}_0 + \hat{\gamma}_1 \sum_{i=1}^{n} y_i$$
$$\sum_{i=1}^{n} y_i x_i = \hat{\gamma}_0 \sum_{i=1}^{n} y_i + \hat{\gamma}_1 \sum_{i=1}^{n} y_i^2$$

Megjegyzés: a most meghatározott és a (133) szerinti egyenes általában nem esik egybe, azaz $\hat{y}(x)$ és $\hat{x}(y)$ rendszerint valamilyen szöget zár be. Lásd a 19. ábrát!

Az $\hat{y}(x)$ és az $\hat{x}(y)$ regressziós egyenesek



19. ábra

Az $(\bar{x}, \bar{y})$ pont mindkét regressziós egyenesen rajta van.

A regressziós paraméterek meghatározása a kovariancia módszerével

A továbbiakban, az egyszerűség végett, a futóindexek feltüntetésétől eltekintünk.

Fejezzük most ki $\hat{\beta}_1$-et a (134)-(135) normálegyenletekből!

A megfelelő műveleteket elvégezve

$$\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{\frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \frac{\frac{\sum xy}{n} - \bar{x}\cdot\bar{y}}{\frac{\sum x^2}{n} - \bar{x}^2}.$$

A fenti egyenlet jobboldalán a számlálóban éppen a kovariancia, míg a nevezőben éppen x szórásnégyzete áll. Lásd az (51) és (97) képleteket.

$$\hat{\beta}_1 = \frac{C_{xy}}{\sigma_x^2} \tag{138}$$

A regressziós együttható ismeretében, (136) segítségével, a $\hat{\beta}_0$ is könnyen kiszámítható.

Standardizált változók közötti kapcsolat

Ha az $y(x)$ és $x(y)$ egyenleteit (bizonyos átalakítások után) a megfelelő szórásokkal elosztjuk, akkor a standardizált változók közötti összefüggéshez jutunk.

$$\frac{y-\bar{y}}{\sigma_y} = \frac{C_{xy}}{\sigma_x\sigma_y}\cdot\frac{x-\bar{x}}{\sigma_x}$$

$$\frac{x-\bar{x}}{\sigma_x} = \frac{C_{xy}}{\sigma_x\sigma_y}\cdot\frac{y-\bar{y}}{\sigma_y}$$

A transzformált változókra vezessük be az

$$\frac{y-\bar{y}}{\sigma_y} = Y \qquad \text{és az} \qquad \frac{x-\bar{x}}{\sigma_x} = X$$

jelöléseket. Ekkor a regressziós egyeneseket az alábbi módon írhatjuk fel.

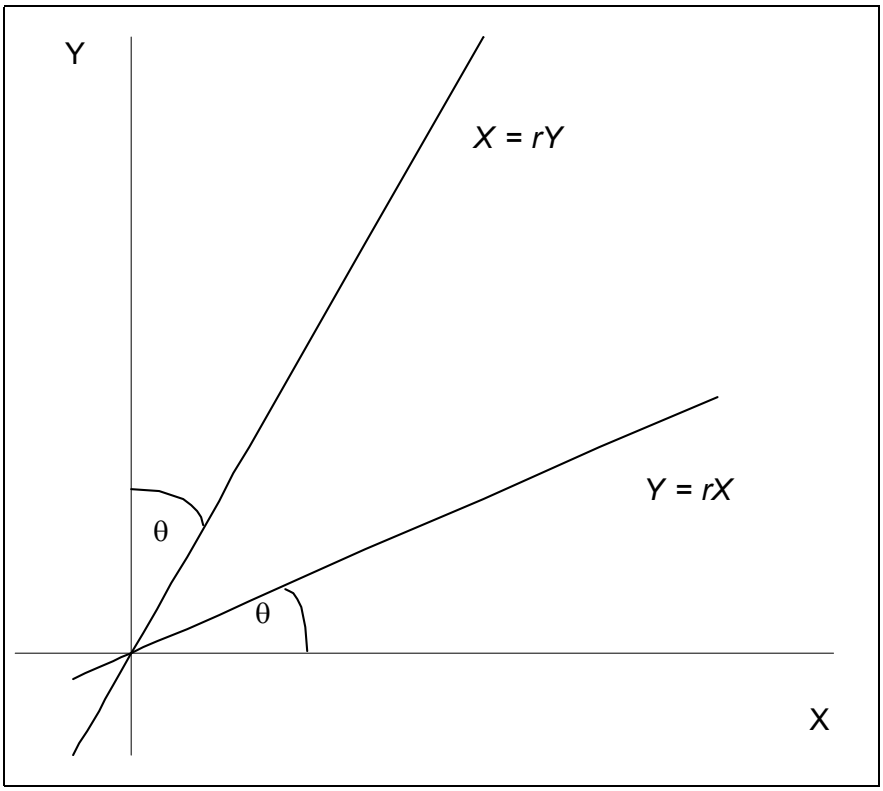
$$Y = rX$$

$$X = rY$$

Megjegyzés: mint tudjuk $r_{xy} = r_{yx} = r$.

Egy diagramon ábrázolva a két egyenest a 20. ábrának megfelelő képet kapjuk.

A standardizált változók közötti regressziós összefüggések



20. ábra

$X = rY$ az Y tengellyel és $Y = rX$ az X tengellyel ugyanakkora θ szöget zár be. Tehát:

$$r = \tan\theta .$$

Nyilvánvaló most már, hogy a két egyenes csak akkor esik egybe, ha az ismérvek közötti kapcsolat determinisztikus, vagyis $\left| r \right| = 1$.

Megjegyzés: könnyen belátható, hogy a standardizált változók és az eredeti változók lineáris korrelációs együtthatója egyenlő.

$$r_{XY} = r_{xy}$$

Összefüggés a regressziós együtthatók és a lineáris korrelációs együttható között

Mivel

$$\hat{\beta}_1 = \frac{C_{xy}}{\sigma_x^2} \qquad \text{és} \qquad \hat{\gamma}_1 = \frac{C_{xy}}{\sigma_y^2},$$

ezek szorzatának négyzetgyöke éppen a lineáris korrelációs együttható abszolút értékével egyenlő.

$$\sqrt{\hat{\beta}_1 \hat{\gamma}_1} = |r| \tag{139}$$

Könnyen belátható az alábbi két összefüggés:

$$\text{ha } \hat{\beta}_1 \text{ és } \hat{\gamma}_1 \text{ pozitív, akkor } r = \sqrt{\hat{\beta}_1 \hat{\gamma}_1};$$

$$\text{ha } \hat{\beta}_1 \text{ és } \hat{\gamma}_1 \text{ negatív, akkor } r = -\sqrt{\hat{\beta}_1 \hat{\gamma}_1}.$$

Paraméterbecslés átlagtól vett eltérések segítségével

Vegyük még egyszer szemügyre a (134)-(135) normálegyenleteket és hajtsuk végre a következő transzformációt: helyettesítsük az x_i és y_i értékeket az átlaguktól vett eltéréseikkel d_{x_i}-vel és d_{y_i}-vel, a (93)-(94) képletek jelöléseinek megfelelően. A transzformált változókra az alábbi normálegyenletek vonatkoznak. (A továbbiakban egy rövid időre eltekintünk a futóindexek feltüntetésétől.)

$$\sum d_y = n\hat{\beta}_0 + \hat{\beta}_1 \sum d_x$$
$$\sum d_x d_y = \hat{\beta}_0 \sum d_x + \hat{\beta}_1 \sum d_x^2$$

Az alkalmazott lineáris transzformációval az $(\bar{x}, \bar{y})$ pont került az origóba, de a regressziós egyenes meredeksége nem változott. Az előző egyenletrendszerből tehát az eredeti, keresett egyenes $\hat{\beta}_1$ paramétere meghatározható. A számtani átlag tulajdonságából adódóan

$$\sum_{i=1}^{n} d_{x_i} = 0 \quad \text{és} \quad \sum_{i=1}^{n} d_{y_i} = 0 .$$

(Lásd a 3.2. fejezetet.) Ennek felhasználásával a második normálegyenletből $\hat{\beta}_1$ könnyen megkapható:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} d_{x_i} d_{y_i}}{\sum_{i=1}^{n} d_{x_i}^2}. \tag{140}$$

A $\hat{\beta}_0$ paramétert (136) segítségével határozhatjuk meg:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} .$$

Paraméterbecslés mátrixegyenletekkel

A most következő módszer egy újabb lehetőséget kínál a regressziós egyenes egyenletének felírására. Írjuk fel a normálegyenleteket mátrixalgebrai jelöléssel. Alkalmazzuk az alábbi vektorokat, illetve mátrixot.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

A regressziófüggvény a fenti jelölésekkel:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} .$$

A mintából kiszámított regressziós egyenes egyenlete a következőképpen írható fel:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} . \tag{141}$$

A paraméterek vektorát a normálegyenleteken alkalmazott mátrixműveletek elvégzése után a (142) alakra hozhatjuk.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\,\mathbf{X}\right)^{-1}\mathbf{X}'\,\mathbf{y} \qquad (142)$$

Megjegyzés: $\mathbf{X}'$ az $\mathbf{X}$ mátrix transzponáltját jelenti.

A most bevezetett jelölésrendszer azért fontos, mert általánosítható többváltozós esetre. A többváltozós regressziószámítással a második kötetben foglalkozunk.

Az említett megoldási módszerek (normálegyenletek, kovariancia, differenciák, mátrixegyenlet) mindegyikére fennállnak az alábbi összefüggések.

$$\sum_{i=1}^{n} y_i = \sum_{\mathrm{i}=1}^{n} \hat{y}_i$$

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right) = 0$$

$$\sum_{i=1}^{n} e_i = 0$$

Megjegyzés: a fenti összefüggések mind az alkalmazott LNM-nek a következménye. A számításaink pontosságának ellenőrzésére bármelyik összefüggés alkalmazható.

A regressziószámítás eredményeinek értelmezése

Először a regressziós egyenes paramétereinek statisztikai, közgazdasági értelmezését kell megadnunk. (A matematikai értelmezés természetesen nem elégséges.)

A $\hat{\beta}_1$ (regressziós együttható) azt mutatja meg, hogy az *x* magyarázóváltozó egységnyi növekedése az eredményváltozó átlagosan mekkora (abszolút) változásával jár együtt. Tehát az *x* változó értékét 1 egységgel növelve az *y* változó értéke átlagosan $\hat{\beta}_1$ értékével növekszik, vagy csökken. A regressziós együttható pozitív vagy negatív előjele a kapcsolat irányát fejezi ki.

A $\hat{\beta}_0$ paraméter az $x = 0$ esetre ad elméleti értéket. Természetesen csak akkor értelmezhető, ha a 0 érték beletartozik *x* ismérvváltozatai közé, vagy még inkább azon *x*-ek

közé, amelyekből a regressziós egyenes egyenletét számítottuk. A $\hat{\beta}_0$ értéknek tehát gyakran nem tulajdonítható statisztikai, közgazdasági tartalom.

Az $\hat{y}$ függvényértékek a megfelelő *x* ismérvértékhez tartozó *y* értékek elméleti megfelelői. A tényleges *y* értékek ezért két tag összegére bonthatóak:

$$y_i = \hat{y}_i + e_i .$$

$\hat{y}_i$ nem más, mint az *y* ismérvérték x_i-vel magyarázható része, míg e_i az a rész, amelyet a magyarázóváltozón kívüli összes többi tényező befolyásol. Ezeket a vizsgálat szempontjából véletlen tényezőknek tekintjük.

Elaszticitás

Az előző pontban azt vizsgáltuk, hogy az x változása átlagosan mekkora abszolút változást idéz elő *y*-ban. A közgazdasági elemzésekben azonban az eredményváltozó relatív változásának van kiemelkedő szerepe. Ennek leggyakrabban alkalmazott mérőszáma a (143) képlettel definiált ún. **elaszticitási** vagy **rugalmassági együttható**.

$$E = \frac{dy}{dx} \cdot \frac{x}{y} \qquad (143)$$

Ez a mutatószám arra ad választ, hogy az *x* magyarázóváltozó adott értékének 1%-os növekedése megközelítőleg és átlagosan milyen relatív változást eredményez az *y* változóban.

A rugalmasság természetesen általában minden *x* értékre más és más. (A gyakorlatban legtöbbször az $\bar{x}$ pontban számítjuk.) A rugalmassági együttható tehát arra ad választ, hogy a vizsgált jelenség *y* ismérvértéke hogyan reagál *x* adott értékről való 1%-os elmozdulására. A (143) alapján kapott eredmény már százalékos formában adott.

Az elaszticitás meghatározása a (133) regressziófüggvény alapján történhet, a (144) képlet szerint.

$$E = \hat{\beta}_1 \frac{x}{\hat{y}} \qquad (144)$$

54. példa

Európa néhány államának egy főre jutó bruttó hazai termékét (folyóáron), valamint az ezer lakosra jutó személyi számítógépek számát a 43. táblázat tartalmazza.

Az egy főre jutó GDP és az ezer lakosra jutó
PC-k száma Európa néhány államában 1995-ben

43. táblázat

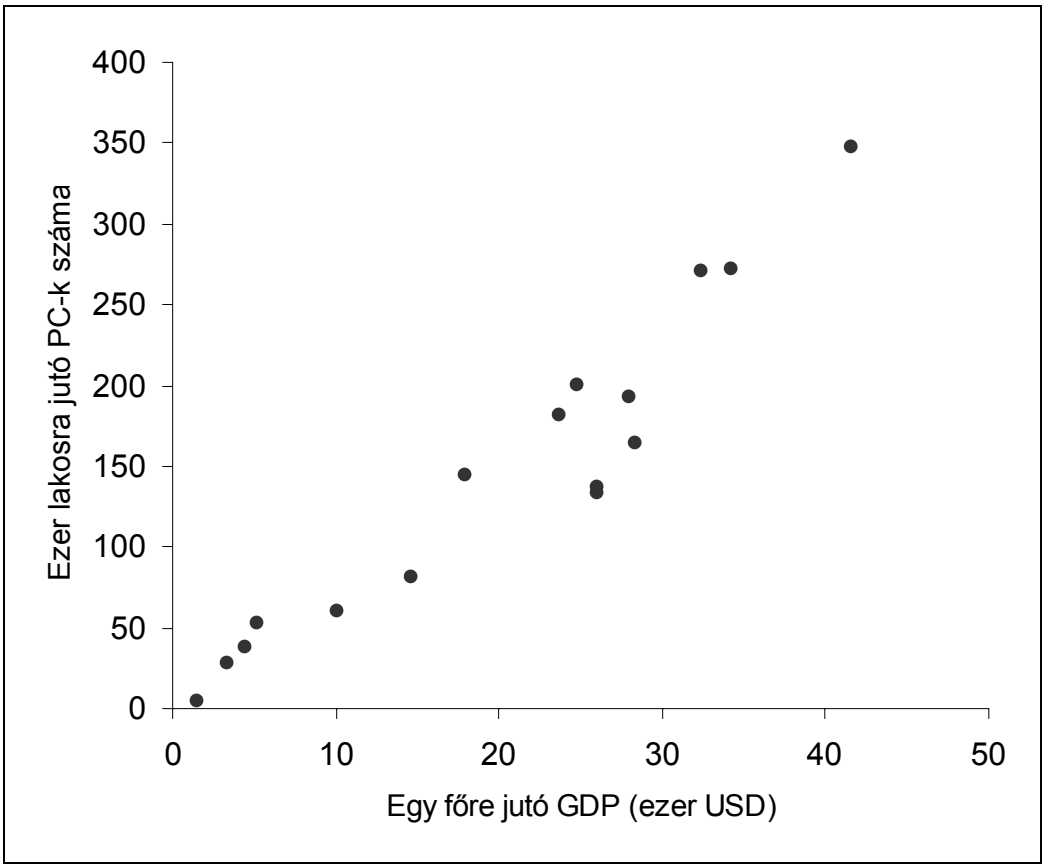
Ország	Egy főre jutó GDP (ezer USD)	Ezer lakosra jutó PC-k száma (db)
Belgium	26,0	138
Csehország	5,2	53
Dánia	32,4	271
Finnország	23,6	182
Franciaország	26,0	134
Hollandia	24,7	201
Írország	17,9	145
Lengyelország	3,3	29
Magyarország	4,4	39
Németország	28,3	165
Norvégia	34,2	273
Portugália	10,1	61
Románia	1,5	5
Spanyolország	14,6	82
Svájc	41,5	348
Svédország	27,9	193

Forrás: Nemzetközi statisztikai zsebkönyv, KSH, Bp., 1999.
Magyar statisztikai évkönyv, KSH, Bp., 1997.

Számítsuk ki és értelmezzük a regressziós függvény paramétereit! Számítsunk rugalmassági együtthatót a magyarázóváltozó átlagértékénél!

Mindenek előtt ábrázoljuk pontdiagramon az adatainkat. Az egy főre jutó GDP-t fogjuk magyarázóváltozónak tekinteni (x), az ezer lakosra jutó PC-k számát pedig eredményváltozónak (y). A 21. ábra alapján a két változó közötti linearitás feltételezhető, azaz alkalmazhatjuk a (133) szerinti lineáris modellt.

Az egy főre jutó GDP és az ezer lakosra jutó
PC-k száma Európa néhány államában 1995-ben



21. ábra

Először írjuk fel a (134)-(135) normálegyenleteket. Az ehhez szükséges mellékszámításokat a 44. táblázat tartalmazza.

A normálegyenletek meghatározásához szükséges mellékszámítások

44. táblázat

x_i	y_i	$x_i \cdot y_i$	x_i^2
26,0	138	3 588,0	676,00
5,2	53	275,6	27,04
32,4	271	8 780,4	1 049,76
23,6	182	4 295,2	556,96
26,0	134	3 484,0	676,00
24,7	201	4 964,7	610,09
17,9	145	2 595,5	320,41
3,3	29	95,7	10,89
4,4	39	171,6	19,36
28,3	165	4 669,5	800,89
34,2	273	9 336,6	1 169,64
10,1	61	616,1	102,01
1,5	5	7,5	2,25
14,6	82	1 197,2	213,16
41,5	348	14 442,0	1 722,25
27,9	193	5384,7	778,41
321,6	2 319	63 904,3	8 735,12

A két normálegyenlet az alábbi.

$$2319 = 16\hat{\beta}_0 + 321{,}6\hat{\beta}_1$$
$$63904{,}3 = 321{,}6\hat{\beta}_0 + 8735{,}12\hat{\beta}_1$$

A fenti kétismeretlenes egyenletrendszerből matematikai átalakításokkal a paraméterek értékeire a következő eredményeket kapjuk:

$$\hat{\beta}_0 = -8{,}1 \quad \text{és} \quad \hat{\beta}_1 = 7{,}6\,.$$

A regressziós egyenes egyenlete:

$$\hat{y} = -8{,}1 + 7{,}6 \cdot x\,.$$

A paramétereket megkaphattuk volna az átlagoktól ($\bar{x} = 20{,}1$ és $\bar{y} = 144{,}9$) vett eltérések segítségével is a (140) és (136) képlet szerint. Az ehhez szükséges mellékszámításokat a 45. táblázat tartalmazza.

A normálegyenletek meghatározásához szükséges mellékszámítások

45. táblázat

d_{x_i}	d_{y_i}	$d_{x_i} \cdot d_{y_i}$	$d_{x_i}^2$
5,9	-6,9	-40,93	34,81
-14,9	-91,9	1369,87	222,01
12,3	126,1	1 550,57	151,29
3,5	37,1	129,72	12,25
5,9	-10,9	-64,53	34,81
4,6	56,1	257,89	21,16
-2,2	0,1	-0,14	4,84
-16,8	-115,9	1 947,75	282,24
-15,7	-105,9	1 663,22	246,49
8,2	20,1	164,51	67,24
14,1	128,1	1 805,68	198,81
-10,0	-83,9	839,38	100,00
-18,6	-139,9	2 602,84	345,96
-5,5	-62,9	346,16	30,25
21,4	203,1	4 345,54	457,96
7,8	48,1	374,89	60,84
0,0	0,0	17 292,40	2 270,96

A regressziós együttható értéke:

$$\hat{\beta}_1 = \frac{17292{,}40}{2270{,}96} = 7{,}6146 \approx 7{,}6 .$$

A (136) összefüggés alapján:

$$\hat{\beta}_0 = 144{,}9375 - 7{,}6146 \cdot 20{,}1 = -8{,}1155 \approx -8{,}1 .$$

A regressziós együtthatót a kovariancia és a magyarázóváltozó szórásnégyzetének segítségével is megkaphattuk volna:

$$C_{xy} = 1080{,}75 \quad \text{és} \quad \sigma_x^2 = 141{,}94 .$$

A (138) képlet alapján:

$$\hat{\beta}_1 = \frac{1080{,}75}{141{,}94} = 7{,}6 .$$

Végül számítsuk ki a paramétereket a mátrixegyenlet megoldásával is.

$$\mathbf{y} = \begin{bmatrix} 138 \\ 53 \\ \vdots \\ 193 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 26{,}0 \\ 1 & 5{,}2 \\ \vdots & \\ 1 & 27{,}9 \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

A paraméterek vektorát a (142) képlet segítségével tudjuk kifejezni.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 26{,}0 & 5{,}2 & & 27{,}9 \end{bmatrix} \cdot \begin{bmatrix} 1 & 26{,}0 \\ 1 & 5{,}2 \\ \vdots & \\ 1 & 27{,}9 \end{bmatrix} = \begin{bmatrix} 16{,}0 & 321{,}6 \\ 321{,}6 & 8735{,}1 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0{,}24040 & -0{,}00885 \\ -0{,}00885 & 0{,}00044 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 26{,}0 & 5{,}2 & & 27{,}9 \end{bmatrix} \cdot \begin{bmatrix} 138 \\ 53 \\ \vdots \\ 193 \end{bmatrix} = \begin{bmatrix} 2319{,}0 \\ 63904{,}3 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0{,}24040 & -0{,}00885 \\ -0{,}00885 & 0{,}00044 \end{bmatrix} \cdot \begin{bmatrix} 2319{,}0 \\ 63904{,}3 \end{bmatrix} = \begin{bmatrix} -8{,}1 \\ 7{,}6 \end{bmatrix}$$

Az **y** vektor és a (141) mátrixegyenlet segítségével könnyen kiszámítható a reziduumok $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ oszlopvektora.

$$\mathbf{e} = \begin{bmatrix} -51{,}9 \\ 21{,}5 \\ \vdots \\ -11{,}3 \end{bmatrix}$$

Megjegyzés: a számításaink pontosságát mutatja, hogy

$$\sum_{i=1}^{16} e_i = 0 .$$

(Lásd a 23. ábra MARADÉK TÁBLA utolsó oszlopát.)

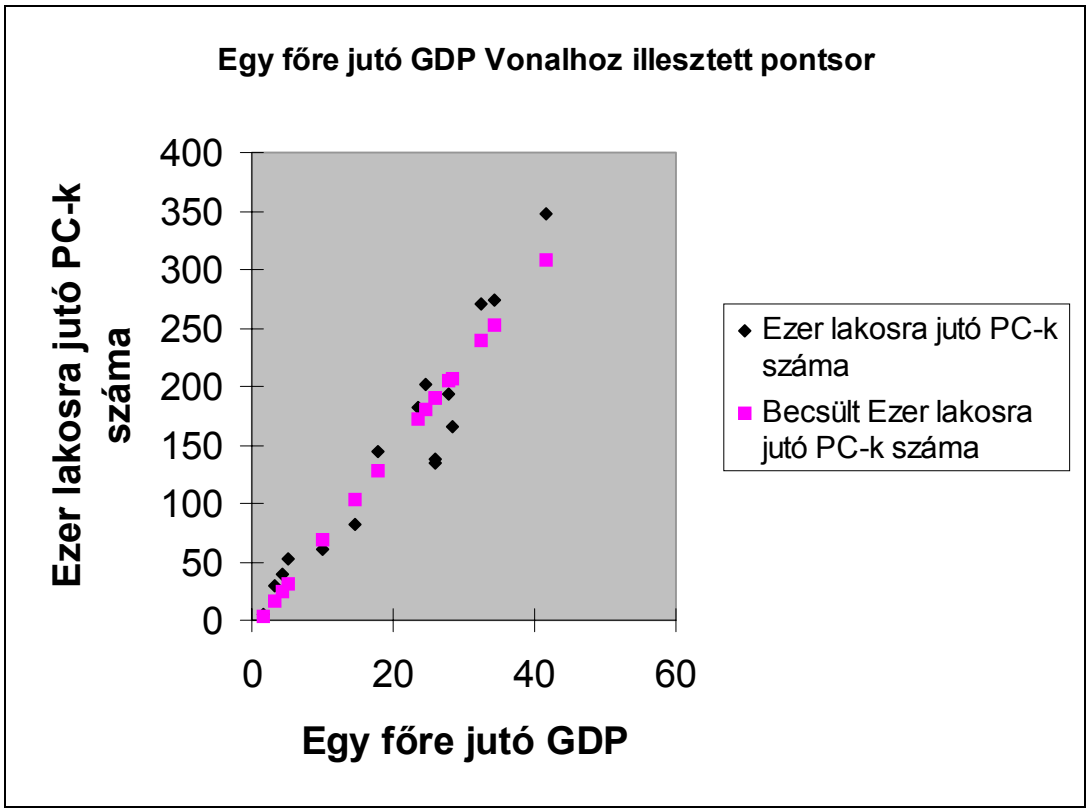
A paraméterek értelmezése:

- az egy főre jutó GDP egy egységnyivel (ezer USD-ral) való növekedése (a 16 vizsgált országban) a PC-k számának átlagosan (megközelítő pontossággal) $\hat{\beta}_1 = 7{,}6$ egységnyi (darab ezer lakosonként) növekedését eredményezi.

- A $\hat{\beta}_0$ paraméter közgazdaságilag ebben az esetben nem értelmezhető, mert ez a 0 dollárnyi GDP-vel rendelkező országok PC állományát mutatja.

A regresszióanalízist az Excelben is elvégezhetjük. Vigyük be az adatainkat egy munkalapra az **A1-B17** cellatartományba (a fejlécekkel együtt). Hívjuk meg az **Eszközök** menü **Adatelemzés…** almenüjét és válasszuk ki a felkínált lehetőségek közül a **Regresszió** menüpontot. Az ekkor megjelenő párbeszédpanellel vigyük be a Bemeneti Y tartományba és a Bemeneti X tartományba az adatainkat tartalmazó megfelelő cellahivatkozásokat. Kapcsoljuk be a Feliratok jelölőnégyzetet, mivel a cellatartományaink első sora fejlécet tartalmaz. A grafikus ábrához a Pontsorok a vonalhoz feliratú jelölőnégyzetet is be kell kapcsolnunk. Az Excel outputja a 22.-23. ábrán látható.

Az Excel outputja



22. ábra

Megjegyzés: a 22. ábra az Excel outputjának első részét, míg a 23. ábra a második részét tartalmazza. A 23. táblázatban levő szöveg nem hibás gépelés eredménye, hanem az Excel szokásos megjelenítési formája. Az itt szereplő fogalmak részletes ismertetésével a második kötetben fogunk foglalkozni.

Az Excel outputja (folytatás)

ÖSSZESÍTŐ TÁBLA

Regressziós statisztika	
r értéke	0,954164
r-négyzet	0,910429
Korrigált r-	0,904031
Standard h	30,41924
Megfigyelé	16

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*szignifikanciája*
Regress	1	131674,3	131674,3	142,2998	1,01E-08
Maradék	14	12954,62	925,3303		
Összese	15	144628,9			

	Koefficiens	*tandard hi*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95*	*Alsó 95,0*
Tengely	-8,1155	14,91482	-0,54412	0,594917	-40,1046	23,87364	-40,1046
Egy főre	7,614577	0,638328	11,92895	1,01E-08	6,245499	8,983655	6,245499

MARADÉK TÁBLA

Megfigyelé	*lakosra ju*	*Maradéko*
1	189,8635	-51,8635
2	31,4803	21,5197
3	238,5968	32,4032
4	171,5885	10,41148
5	189,8635	-55,8635
6	179,9646	21,03545
7	128,1854	16,81457
8	17,0126	11,9874
9	25,38864	13,61136
10	207,377	-42,377
11	252,303	20,69696
12	68,79173	-7,79173
13	3,306366	1,693634
14	103,0573	-21,0573
15	307,8894	40,11055
16	204,3312	-11,3312

23. ábra

Az ÖSSZESÍTŐ TÁBLA adataiból látszik, hogy a lineáris korrelációs együttható értéke igen magas: $r = 0{,}95$. A regressziós paramétereket a VARIANCIAANALÍZIS táblájában a

Koefficiens oszlopban találjuk meg. A MARADÉK TÁBLA második oszlopában az eredményváltozó elméleti értékeit ($\hat{y}_i$) találjuk, a harmadikban pedig a reziduumokat.

Végezetül számítsuk ki a magyarázóváltozó átlagértékéhez tartozó rugalmassági együtthatót a (144) képlet segítségével.

$$E(x = \bar{x} = 20{,}1) \quad = \quad 7{,}6146 \cdot \frac{20{,}1}{-8{,}1155 + 7{,}6146 \cdot 20{,}1} \quad = \quad 1{,}06$$

Az egy főre jutó GDP 1%-os (20 100 dollárról 20 301 dollárra való) növekedése (a megfigyelt országokban) az ezer lakosra jutó PC-k számának átlagosan (megközelítő pontossággal) 1,06%-os növekedésével jár együtt.

6.2. Nemlineáris regresszió

Az előző fejezetben abból indultunk ki, hogy két mennyiségi ismérv kapcsolatát vizsgálva adatpárjainkat egy pontdiagramon ábrázoltuk, és ebben a pontok elhelyezkedése, sűrűsödési helye alapján lineáris modellt feltételeztünk. A pontdiagramon kirajzolódó pontfelhő azonban nem feltétlenül utal egyenesre. A gazdasági jelenségek között gyakran előfordul, hogy az eredményváltozó a magyarázóváltozó 1 egységnyi változására nem állandó változással reagál a különböző x pontokban. A statisztikai gyakorlatban ezért (bizonyos jelenségek vizsgálatánál) gyakran nemlineáris függvényt illesztünk. Ezek közül, valamilyen transzformáció segítségével, néhány visszavezethető a lineáris modellre, míg a többi lineárisra nem transzformálható modell. A nemlineáris (de linearizálható) függvények közül leggyakrabban a hatványkitevős, az exponenciális, a parabolikus és a hiperbolikus függvényeket használjuk. Ezekkel foglalkozunk most részletesebben.

Exponenciális regresszió

Az **exponenciális regressziófüggvény** az alábbi képlettel definiált:

$$\hat{y} = \hat{\beta}_0 \hat{\beta}_1^{\,x}. \qquad (145)$$

E függvénytípus olyan esetekben alkalmazható, ha az y eredményváltozó növekedési üteme függ az x változótól, vagyis egy jelenség változásának üteme függ a jelenség már elért színvonalától. A (145) függvény logaritmikus transzformációval a következő (transzformált változóiban) lineáris összefüggéssé alakítható:

$$\log \hat{y} = \log \hat{\beta}_0 + x \log \hat{\beta}_1 .$$

Megjegyzés: a transzformációhoz tetszőleges alapú logaritmust használhatunk.

Az exponenciális regressziófüggvény paramétereit tehát úgy határozhatjuk meg, hogy alkalmazzuk a lineáris modellnél megismert módszerek valamelyikét a transzformált változókra, majd elvégezzük a kapott eredmények visszatranszformálását.

A $\hat{\beta}_1$ paraméter értelmezése ebben az esetben a következő: a regressziós együttható azt fejezi ki, hogy az x magyarázóváltozó egységnyi növekedése az y eredményváltozó átlagosan hányszoros ($\hat{\beta}_1$-szeres) változásával jár együtt.

Az exponenciális regressziófüggvény elaszticitása a (143) általános képletből adódóan az alábbi:

$$E = x \ln \hat{\beta}_1 .$$

55. példa

Egy árverésen azonos gyártótól származó keleti szőnyegeket értékesítettek, amelyek hasonló anyagból készültek, és az előállításuk idejében sem különböztek jelentősen. A köztük levő legnagyobb eltérés az egy négyzetméterre jutó csomók számában mutatkozott, mivel megközelítőleg azonos nagyságúak voltak. Az árverésen kialakult értékesítési árakat a 46. táblázat tartalmazza.

Az árverésen értékesített szőnyegek adatai

46. táblázat

Minőség (ezer csomó/m^2)	Eladási ár (ezer Ft/m^2)
25	20
60	25
100	30
180	50
220	85
350	110
500	200
750	490

Határozzuk meg az exponenciális regressziófüggvényt (a minőség legyen a magyarázóváltozó, az eladási ár pedig az eredményváltozó), és értelmezzük a regressziós együtthatót!

Az eladási ár (y) logaritmikus transzformációjával kapott részeredményeket a 47. táblázat tartalmazza.

Az exponenciális regressziófüggvény meghatározásához szükséges mellékszámítások

47. táblázat

$\lg y_i$	x_i	$x_i \cdot \lg y_i$	x_i^2
1,30	25	32,53	625
1,40	60	83,88	3 600
1,48	100	147,71	10 000
1,70	180	305,81	32 400
1,93	220	424,47	48 400
2,04	350	714,49	122 500
2,30	500	1 150,51	250 000
2,69	750	2 017,65	562 500
14,84	2185	4 877,05	1 030 025

A részeredményeket például a normálegyenletekbe helyettesítve:

$$\lg \hat{\beta}_0 = 1{,}334755 \quad \text{és} \quad \lg \hat{\beta}_1 = 0{,}001903\,;$$

illetve

$$\hat{\beta}_0 = 21{,}6150 \quad \text{és} \quad \hat{\beta}_1 = 1{,}0043$$

paramétereket kaptuk.

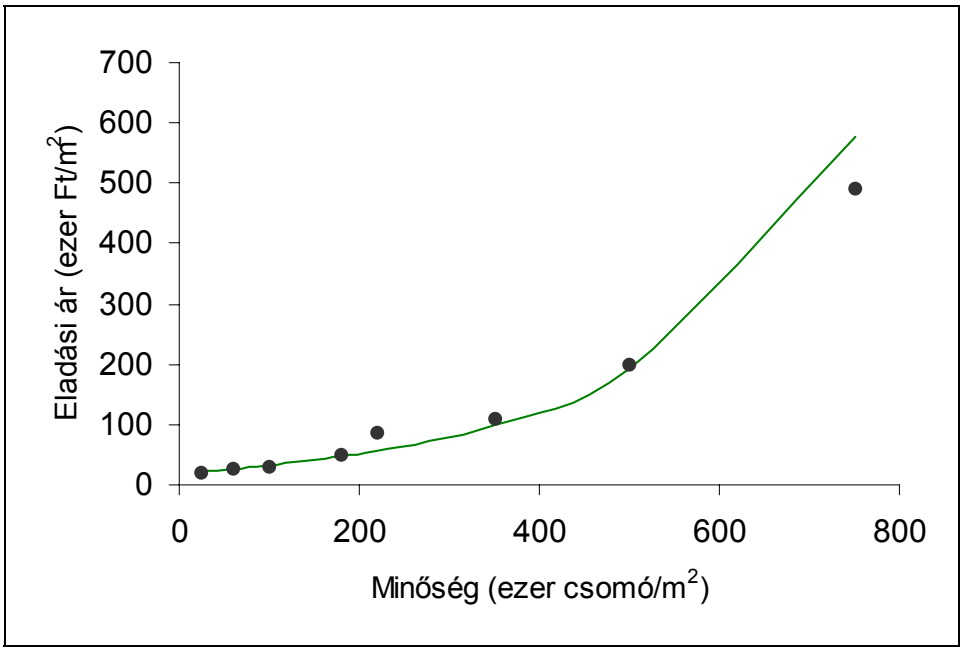
A (145) szerinti exponenciális regressziófüggvény a következő:

$$\hat{y} = 21{,}6 \cdot 1{,}0043^x\,.$$

A $\hat{\beta}_1$ regressziós együttható értelmezése: ha az egy négyzetméterre jutó csomók számát (magyarázóváltozó értékét) 1 egységnyivel növeljük, akkor az értékesített szőnyegek eladási ára (eredményváltozó értéke) átlagosan (megközelítő pontossággal) 1,0043-szorosára növekszik.

Az árverésen kialakult eladási árak empirikus és elméleti adatait a 24. ábra mutatja.

Az exponenciális regressziófüggvény illesztése



24. ábra

Hatványkitevős regresszió

A **hatványkitevős regressziófüggvény** az alábbi képlettel definiált:

$$\hat{y} = \hat{\beta}_0 x^{\hat{\beta}_1}. \qquad (146)$$

Ezt a függvénytípust akkor használjuk, ha az x és y változók logaritmusai között van lineáris összefüggés. A (146) ugyanis logaritmikus transzformációval szintén visszavezethető a lineáris modellre:

$$\log \hat{y} = \log \hat{\beta}_0 + \hat{\beta}_1 \log x.$$

A transzformált modell megoldása után $\hat{\beta}_0$ értékét kell a $\log \hat{\beta}_0$ megfelelő alapú hatványozásával kiszámítani, ugyanis $\hat{\beta}_1$ -et már közvetlenül megkaptuk.

A hatványkitevős regressziófüggvény speciális tulajdonsága, hogy rugalmassági együtthatója nem függ x-től, azaz konstans, és éppen a $\hat{\beta}_1$ paraméterrel egyenlő.

Ebből következik $\hat{\beta}_1$ értelmezése: a regressziós együttható azt fejezi ki, hogy az x magyarázóváltozó (nagyságtól független) 1%-os változása az y eredményváltozó $\hat{\beta}_1$ százaléknyi változásával jár együtt.

56. példa

A budapesti mozik 1999. első félévi látogatóinak és előadásainak számát a 48. táblázat tartalmazza.

Határozzuk meg a hatványkitevős regressziófüggvényt (az előadásszám legyen a magyarázóváltozó, a látogatók száma pedig az eredményváltozó), és értelmezzük a regressziós együtthatót!

A logaritmikus transzformációval kapott részeredményeket a 49. táblázat tartalmazza.

A budapesti mozik mutatói 1999 első félévében

48. táblázat

Mozi	Látogatók száma	Előadás-szám
Hollywood Multiplex (Duna Plaza)	892 558	10 725
Corvin Budapest Filmpalota	670 189	7 285
Hollywood Multiplex (Lurdy Ház)	555 468	9 037
Cineplex Odeon (Pólus Center)	312 015	5 314
Kossuth	181 910	4 334
Cinema City (Csepel Plaza)	176 578	6 087
Puskin	155 170	2 834
Művész	112 811	3 025
Metro	77 038	1 935
Átrium	68 898	814
Horizont	49 786	1 164
Hunyadi	48 700	743
Toldi	41 080	1 132
Duna	40 014	992
Uránia*	39 839	623
Vörösmarty	36 121	736
Szindbád	30 812	1 139
Olimpia	25 418	550
Bem	23 530	743
Alkotás	21 541	997
Európa	17 485	1 549
Örökmozgó	17 424	426
Flórián	17 330	790
Tabán	16 625	534
Hunnia	13 939	659
Ugocsa*	11 263	564
Kőbánya*	10 110	345
Blue Boksz	6 721	502
Sport	512	89
Tátra	423	36
Összesen	3 716 566	65 703

Forrás: Filmforgalmazók Egyesülete
*Időközben bezárt

A hatványkitevős regressziófüggvény meghatározásához szükséges mellékszámítások

49. táblázat

$\lg y_i$	$\lg x_i$	$\lg x_i \cdot \lg y_i$	$\lg^2 x_i$
5,9506364	4,0303973	23,9834291	16,2441024
5,8261973	3,8624296	22,5032766	14,9183621
5,7446590	3,9560243	22,7260107	15,6501281
5,4941755	3,7254216	20,4681197	13,8787657
5,2598566	3,6368889	19,1295140	13,2269609
5,2469366	3,7844033	19,8565242	14,3217084
5,1908078	3,4523998	17,9207439	11,9190647
5,0523514	3,4807254	17,5858479	12,1154492
4,8867050	3,2866810	16,0610403	10,8022718
4,8382066	2,9106244	14,0822023	8,4717344
4,6971072	3,0659530	14,4011099	9,4000677
4,6875290	2,8709888	13,4578432	8,2425768
4,6136304	3,0538464	14,0893188	9,3259780
4,6022120	2,9965117	13,7905819	8,9790822
4,6003084	2,7944880	12,8555069	7,8091634
4,5577598	2,8668778	13,0665404	8,2189884
4,4887199	3,0565237	13,7198788	9,3423373
4,4051414	2,7403627	12,0716851	7,5095877
4,3716219	2,8709888	12,5508777	8,2425768
4,3332659	2,9986952	12,9941434	8,9921727
4,2426656	3,1900514	13,5343215	10,1764280
4,2411479	2,6294096	11,1517149	6,9137948
4,2387986	2,8976271	12,2824575	8,3962428
4,2207617	2,7275413	11,5123015	7,4394813
4,1442316	2,8188854	11,6821141	7,9461150
4,0516541	2,7512791	11,1472312	7,5695367
4,0047512	2,5378191	10,1633340	6,4405258
3,8274339	2,7007037	10,3367649	7,2938006
2,7092700	1,9493900	5,2814238	3,8001214
2,6263404	1,5563025	4,0873801	2,4220775
137,1548829	91,2002408	428,4932383	286,0092024

Megjegyzés: hatványkitevős modellnél fontos minél több tizedes számjeggyel dolgozni, mert különben (a kerekítés következtében) jelentősen torzul a paraméterek értéke.

A részeredményeket például a normálegyenletekbe helyettesítve:

$$\lg \hat{\beta}_0 = 0{,}5665015 \quad \text{és} \quad \hat{\beta}_1 = 1{,}3175386$$

paramétereket kaptuk.

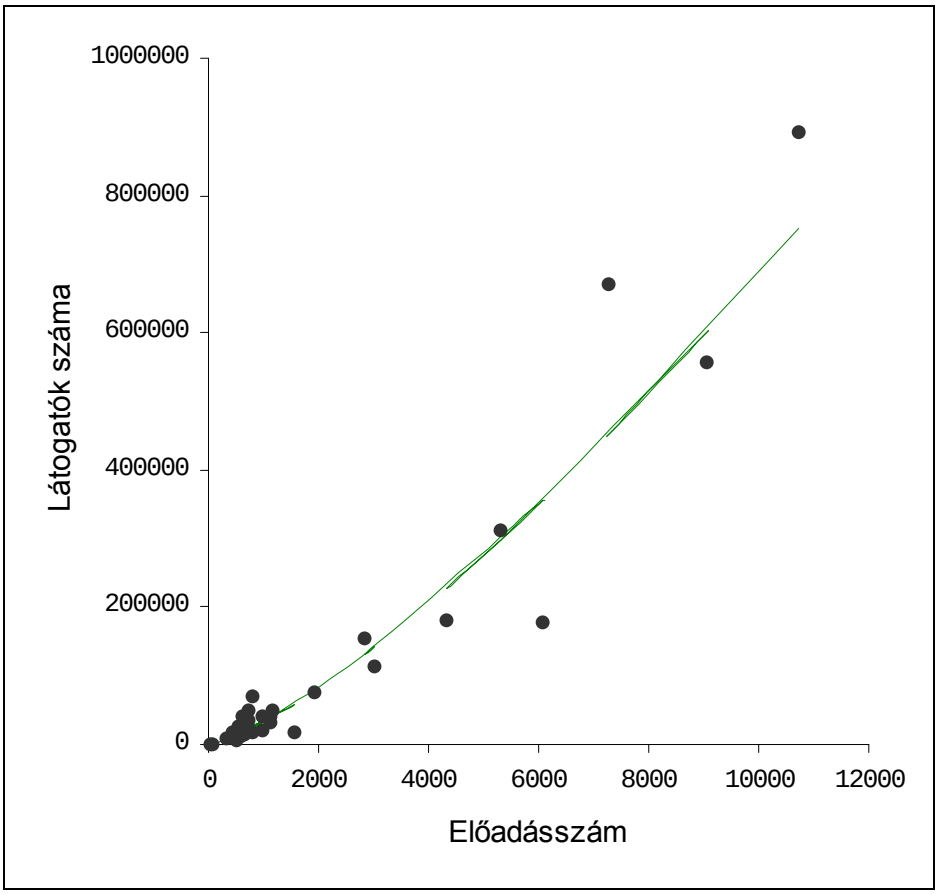
A (146) szerinti hatványkitevős regressziófüggvény a következő:

$$\hat{y} = 3{,}6855 \cdot x^{1{,}3175}.$$

A $\hat{\beta}_1$ regressziós együttható értelmezése: ha az előadásszámot (magyarázóváltozó értékét) bármilyen szintről 1%-kal növeljük, akkor a látogatók száma (eredményváltozó értéke) megközelítő pontossággal átlagosan 1,3%-kal növekszik.

A budapesti mozik látogatóinak empirikus és elméleti adatait a 25. ábra mutatja.

A hatványkitevős regressziófüggvény illesztése



25. ábra

Parabolikus regresszió

A (másodfokú) parabola alakú regressziófüggvény az alábbi képlettel definiált:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2. \tag{147}$$

A LNM szerint ennek paramétereit az alábbi három normálegyenlet alapján tudjuk meghatározni (az egyszerűség végett most is eltekintünk a futóindex feltüntetésétől).

$$\begin{aligned}
\sum y &= n\hat{\beta}_0 + \hat{\beta}_1 \sum x + \hat{\beta}_2 \sum x^2 \\
\sum xy &= \hat{\beta}_0 \sum x + \hat{\beta}_1 \sum x^2 + \hat{\beta}_2 \sum x^3 \\
\sum x^2 y &= \hat{\beta}_0 \sum x^2 + \hat{\beta}_1 \sum x^3 + \hat{\beta}_2 \sum x^4
\end{aligned}$$

A regressziós paraméterek értelmezésének ennél a modellnél nincsen gyakorlati jelentősége.

Hiperbolikus regresszió

A hiperbola alakú regressziófüggvények közül többfélét is alkalmazhatunk. A (148) és a (149) képletekkel definiált regressziófüggvények visszavezethetőek a lineáris modellre, míg például a harmadik függvénytípusnál ez nem lehetséges.

$$\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{x} \tag{148}$$

$$\hat{y} = \frac{\hat{\beta}_0}{\hat{\beta}_1 + x} \tag{149}$$

$$\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{\hat{\beta}_2 + x}$$

A (148) esetében a magyarázóváltozó, míg (149) esetében az eredményváltozó reciprok transzformációja vezet lineáris modellre.

A (148) a következőképpen írható:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{x},$$

amely például $z = \frac{1}{x}$ helyettesítéssel

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 z$$

alakra hozható.

A (149) az alábbi módon írható:

$$\frac{1}{\hat{y}} = \frac{\hat{\beta}_1}{\hat{\beta}_0} + \frac{1}{\hat{\beta}_0} x,$$

amely például $u = \frac{1}{y}$, $\hat{\alpha}_0 = \frac{\hat{\beta}_1}{\hat{\beta}_0}$ és $\hat{\alpha}_1 = \frac{1}{\hat{\beta}_0}$ helyettesítéssel:

$$\hat{u} = \hat{\alpha}_0 + \hat{\alpha}_1 x$$

alakra hozható.

Megjegyzés: a regressziós paraméterek értelmezésének ennél a modellnél sincs gyakorlati jelentősége.

Foglaljuk most össze egy táblázatban a legfontosabb típusú regressziófüggvényeket és elaszticitásaikat.

A lineáris és a legfontosabb nemlineáris (de linearizálható) regressziófüggvények és elaszticitásaik

50. táblázat

Típus	Egyenlete	Elaszticitása
Lineáris	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\frac{\hat{\beta}_1 x}{\hat{\beta}_0 + \hat{\beta}_1 x}$
Exponenciális	$\hat{y} = \hat{\beta}_0 \hat{\beta}_1^x$	$x \ln \hat{\beta}_1$
Hatványkitevős	$\hat{y} = \hat{\beta}_0 x^{\hat{\beta}_1}$	$\hat{\beta}_1$
Parabolikus	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$	$\frac{\hat{\beta}_1 x + 2\hat{\beta}_2 x^2}{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2}$
Hiperbolikus	$\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{x}$	$\frac{-\hat{\beta}_1}{\hat{\beta}_0 x + \hat{\beta}_1}$

A fejezet végén a nem lineáris regressziós elemzéssel kapcsolatosan a következő fontos tényre hívjuk fel a figyelmet: a reziduumok összege jellemzően nullától különböző, azaz

$$\sum_{i=1}^{n} e_i \neq 0 .$$

Ez annak a következménye, hogy a LNM –t nem az eredeti változókra

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$$

alkalmazzuk, hanem a transzformáltakra. Például hatványkitevős regresszió esetén

$$\sum_{i=1}^{n} (\log y_i - \log \hat{y}_i)^2 \to \min ,$$

aminek következményeként a transzformált változók közötti reziduumok összege lesz

nullával egyenlő:

$$\sum_{i=1}^{n}\left(\log y_i - \log \hat{y}_i\right) = 0\,.$$

6.3. Lineáris és nemlineáris korreláció

A 4. fejezetben már tárgyaltuk a mennyiségi ismérvek közötti kapcsolat szorosságának egyik mérőszámát, a lineáris korrelációs együtthatót. Erről azt kell tudni, hogy kizárólag lineáris kapcsolatoknál alkalmazható. (A linearitást pontdiagram alapján szoktuk eldönteni.) Az empirikus elemzéseknél kitüntetett szerepe van a lineáris korrelációs együttható négyzetének, az ún. **lineáris determinációs együtthatónak**.

Lineáris determinációs együttható

Értékét a következő módon számíthatjuk ki:

- a lineáris korrelációs együttható segítségével (*r* négyzetre emelésével),
- a regressziós paraméterek segítségével a (139) alapján ($r^2 = \hat{\beta}_1\hat{\gamma}_1$),
- az eredményváltozó empirikus és elméleti értékeinek segítségével a (150) szerint:

$$r^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \qquad (150)$$

Bármelyik módszer szerint is számítjuk ki, az eredmény a következő zárt intervallumba esik:

$$0 \leq r^2 \leq 1.$$

A lineáris korrelációs együtthatótól eltérően, r^2 értékét százalékban is kifejezhetjük.

A lineáris determinációs együtthatót kétféleképpen értelmezhetjük:

- megoszlási viszonyszámként is és
- a PRE-elv szerint is.

Az előbbi esetben r^2 értelmezése a következő: az eredményváltozó szórásnégyzetének $100 \cdot r^2$ százaléknyi része értelmezhető a magyarázóváltozóval, míg a többi a véletlen (illetve a figyelembe nem vett) tényezők következménye.

A lineáris determinációs együttható PRE-elv szerinti értelmezése a következő: a magyarázóváltozó egy adott értékének ismerete $100 \cdot r^2$ százalékkal csökkenti a hozzá tartozó eredményváltozó elméleti értékének meghatározásánál elkövetett hibát.

Az eddigiekből következik, hogy determinisztikus kapcsolat esetén a lineáris determinációs együttható értéke

$$r^2 = 1, \text{ illetve } r^2 = 100\%.$$

Korrelációs index

Empirikus elemzéseknél gyakran előfordul az az eset, amikor a változók közötti kapcsolat nem lineáris. Az 55. és 56. példánál ilyen esettel találkoztunk. Lásd a 24. és a 25. ábrát. Ha a két vizsgált változó közötti kapcsolat nem lineáris, az *r* helyett az ún. **korrelációs indexet** (I) kell alkalmazni. Ennek definíciója a (150) képlet általánosítása:

$$I = \sqrt{1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}. \qquad (151)$$

A korrelációs index előjelét nem tudjuk értelmezni, csak abszolút nagyságát, amelyre igaz:

$$0 \le I \le 1.$$

Ennek nincs sem megoszlási viszonyszám szerinti, sem PRE-elv szerinti értelmezése.

57. példa

Az 56. példa adatai alapján számítsuk ki a korrelációs indexet és a transzformált változók közötti lineáris korrelációs együtthatót! Értelmezzük a kapott eredményt.

A számításokhoz szükséges részeredményeket az 51. táblázat tartalmazza.

A korrelációs index meghatározásához szükséges mellékszámítások

51. táblázat

y_i	$\hat{y}_i$	e_i^2	$(y_i - \bar{y})^2$
892 558	752 845,0	19 519 735 304,0	593 178 875 451,8
670 189	452 273,9	47 486 978 392,4	300 098 060 385,6
555 468	600 781,0	2 053 271 747,3	187 567 872 026,5
312 015	298 461,4	183 699 278,6	35 962 596 329,1
181 910	228 161,9	2139 234 049,0	3 544 186 026,7
176 578	356 943,1	32 531 557 125,2	2 937 755 627,8
155 170	130 368,2	615 130 676,9	1 075 385 221,4
112 811	142 066,5	855 881 948,8	91 507 080,5
77 038	78 855,2	3 302 360,0	2 055 618 875,8
68 898	25 197,7	1 909 713 132,9	2 859 996 310,5
49 786	40 365,7	88 741 319,8	5 269 443 602,2
48 700	22 342,9	694 695 239,5	5 428 290 505,4
41 080	38 910,1	4 708 579,7	6 609 191 369,4
40 014	32 698,0	53 523 732,9	6 783 652 787,3
39 839	17 715,3	489 460 156,5	6 812 510 438,9
36 121	22 066,0	197 542 920,1	7 440 086 035,2
30 812	39 227,4	70 818 916,4	8 384 137 016,3
25 418	15 032,6	107 855 797,7	9 401 034 753,1
23 530	22 342,9	1 409 139,7	9 770 716 229,4
21 541	32 915,3	129 375 236,3	10 167 885 451,2
17 485	58 819,0	1 708 496 796,2	11 002 317 678,4
17 424	10 736,2	44 726 817,6	11 015 118 215,3
17 330	24 223,5	47 520 460,4	11 034 858 202,7
16 625	14 459,1	4 690 959,7	11 183 471 403,7
13 939	19 076,2	26 390 814,9	11 758 785 385,6
11 263	15 538,8	18 282 601,9	12 346 306 180,8
10 110	8 131,6	3 914 058,4	12 603 864 320,1
6 721	13 328,5	43 659 713,5	13 376 294 915,2
512	1 364,3	726 352,1	14 851 061 976,3
423	414,0	81,2	14 872 761 855,5
3 671 308	3 515 661,4	111 035 043 709,6	1 329 483 641 657,9

A (151) alapján:

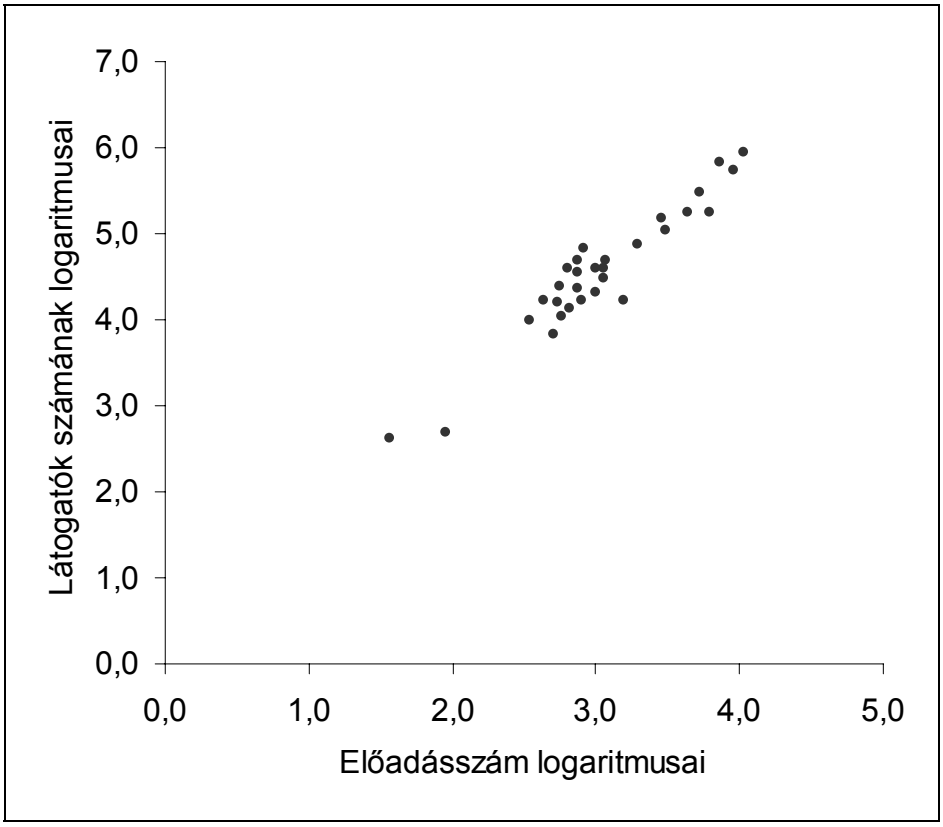
$$I = \sqrt{1 - \frac{111\,035\,043\,709{,}6}{1\,329\,483\,641\,657{,}9}} = 0{,}9573\,.$$

Mivel az *I* normált mutató, az eredmény nagyon erős nemlineáris (hatványkitevős) korrelációs kapcsolatra utal.

Ha az eredeti változók helyett azok logaritmusaival dolgozunk, akkor az eredeti hatványkitevős alakú kapcsolat lineárissá válik.

A korrelációs kapcsolat linearitása a 26. ábra alapján is látható.

A transzformált változók pontdiagramja



26. ábra

A transzformált változók közötti lineáris korrelációs kapcsolatot jellemző mutató a 49. táblázat adatai alapján a (98) képlet felhasználásával kiszámítható:

$$r = 0{,}9567 .$$

Ez azt jelenti, hogy a változók logaritmusai ($\lg x$ és $\lg y$) között nagyon erős, pozitív irányú kapcsolat van.

Tárgymutató

Képletgyűjtemény

1. Általában a statisztikáról

(1) $A \mp a$

(2) $\hat{a} = \frac{10^{sz}}{2}$

(3) $\alpha = \frac{a}{A}$

(4) $\hat{\alpha} = \frac{\hat{a}}{A}$

2. Egyszerű elemzések

(5) $V = \frac{A}{B}$

(6) $b_i = \frac{x_i}{x_b}$ i=1,2,…,N

(7) $l_i = \frac{x_i}{x_{i-1}}$ i=2,3,…,N

(8) $\frac{b_i}{b_{i-1}} = \frac{x_i}{x_b} : \frac{x_{i-1}}{x_b} = \frac{x_i}{x_{i-1}} = l_i$

(9) $\frac{b_i}{b_c} = \frac{x_i}{x_b} : \frac{x_c}{x_b} = \frac{x_i}{x_c} = c_i$

(10) $\prod_{i=b+1}^{m} l_i = \frac{x_{b+1}}{x_b} \cdot \frac{x_{b+2}}{x_{b+1}} \cdot \ldots \cdot \frac{x_{b+m}}{x_{b+m-1}} = \frac{x_{b+m}}{x_b} = b_m$ $m \le N$

(11) $b_{i+1} = b_i \cdot l_{i+1}$

(12) $b_i = b_{i+1} : l_{i+1}$

(13) $\frac{A}{B} = \frac{A}{b} \cdot \frac{b}{B}$

3. Sokaság egy ismérv szerinti vizsgálata

(14) $2^k > N, \quad k \to \min$

(15) $h_i = X_{i,1} - X_{i,0}$ $\quad i=1,2,\ldots,k$

(16) $h = \frac{x_{\max} - x_{\min}}{k}$

(17) $C_i = \left[X_{i,0} - \frac{1}{2}10^{sz}, X_{i,1} + \frac{1}{2}10^{sz} \right)$

(18) $S_i = \sum_{x_j \in C_i} x_j$ $\quad i=1,2,\ldots,k$

(19) $X_i = \frac{X_{i,0} + X_{i,1}}{2}$

(20) $\hat{S}_i = f_i \cdot X_i$

(21) $g_i = \frac{f_i}{\sum_{i=1}^{k} f_i} = \frac{f_i}{N}$

(22) $Z_i = \frac{S_i}{\sum_{i=1}^{k} S_i}$

(23) $$\hat{Z}_i = \frac{f_i X_i}{\sum_{i=1}^{k} f_i X_i}$$

(24) $$\hat{Z}_i = \frac{g_i X_i}{\sum_{i=1}^{k} g_i X_i}$$

(25) $$K_i' = \sum_{j=1}^{i} K_j \qquad i=1,2,...,k$$

(26) $$K_i'' = \sum_{j=i}^{k} K_j$$

(27) $$K_{i-1}' + K_i'' = K_k'$$

(28) $$\bar{x}_a = \frac{\sum_{i=1}^{N} x_i}{N}$$

(29) $$\bar{x}_a = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{N}$$

(30) $$\bar{x} = \frac{\sum_{i=1}^{k} S_i}{\sum_{i=1}^{k} f_i} = \frac{S}{N}$$

(31) $$y_i = \frac{x_i - A}{B} \qquad i=1,2,...,N$$

(32) $$\bar{x} = A + B \cdot \bar{y}$$

(33) $$\bar{x}_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

(34) $$\bar{x}_g = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} x_i^{f_i}}$$

(35) $$\bar{l} = \sqrt[N-1]{\prod_{i=2}^{N} l_i} = \sqrt[N-1]{b_N} = \sqrt[N-1]{\frac{x_N}{x_1}}$$

(36) $$\bar{x}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

(37) $$\bar{x}_h = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

(38) $$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

(39) $$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i}}$$

(40) $$s_{i/k} = \frac{i}{k}(N+1)$$

(41) $$x_{i/k} = x_{[s_{i/k}]} + \{s_{i/k}\} \cdot \left(x_{[s_{i/k}]+1} - x_{[s_{i/k}]}\right)$$

(42) $$Me = \frac{x_{N/2} + x_{(N/2)+1}}{2}$$

(43) $$\hat{Me} = X_{Me,0} + \frac{\frac{N}{2} - f'_{Me-1}}{f_{Me}} \cdot h_{Me}$$

(44) $$\hat{x}_{i/k} = X_{i/k,0} + \frac{\frac{i}{k}N - f'_{(i/k)-1}}{f_{i/k}} \cdot h_{i/k}$$

(45) $$\hat{Mo} = X_{Mo,0} + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \cdot h_{Mo}$$

(46) $$R = x_{\max} - x_{\min}$$

(47) $$G = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|}{N(N-1)}$$

(48) $$G = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} f_i f_j |x_i - x_j|}{N(N-1)}$$

(49) $$\delta = \frac{\sum_{i=1}^{N}|x_i - \bar{x}|}{N}$$

(50) $$\delta = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{x}|}{\sum_{i=1}^{k} f_i}$$

(51) $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2}$$

(52) $$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i} - \bar{x}^2}$$

(53) $$\sigma = \sqrt{\bar{x}_q^2 - \bar{x}^2}$$

(54) $$\mathrm{v} = \frac{\sigma}{\bar{x}}$$

(55) $$y_i = \frac{x_i - \bar{x}}{\sigma} \qquad i=1,2,\ldots,N$$

(56) $$L = \frac{t_c}{1/2} = 2 \cdot t_c$$

(57) $$L = \frac{G}{2 \cdot \bar{x}}$$

(58) $$M_r(A) = \frac{\sum_{i=1}^{N} (x_i - A)^r}{N}$$

(59) $$M_r(A) = \frac{\sum_{i=1}^{k} f_i (x_i - A)^r}{\sum_{i=1}^{k} f_i}$$

(60) $$M_2(\bar{x}) = h^2 \left(\frac{\sum_{i=1}^{k} f_i y_i^2}{\sum_{i=1}^{k} f_i} - \bar{y}^2 \right) = h^2 \left(\bar{y}_q^2 - \bar{y}^2 \right) = h^2 \sigma_y^2 = \sigma^2$$

(61) $$M_3(\bar{x}) = h^3 \left(\frac{\sum_{i=1}^{k} f_i y_i^3}{\sum_{i=1}^{k} f_i} - 3 \cdot \bar{y} \cdot \frac{\sum_{i=1}^{k} f_i y_i^2}{\sum_{i=1}^{k} f_i} + 2 \cdot \bar{y}^3 \right)$$

(62) $$M_4(\bar{x}) = h^4 \left(\frac{\sum_{i=1}^{k} f_i y_i^4}{\sum_{i=1}^{k} f_i} - 4 \cdot \bar{y} \cdot \frac{\sum_{i=1}^{k} f_i y_i^3}{\sum_{i=1}^{k} f_i} + 6 \cdot \bar{y}^2 \cdot \frac{\sum_{i=1}^{k} f_i y_i^2}{\sum_{i=1}^{k} f_i} - 3 \cdot \bar{y}^4 \right)$$

(63) $$M_2(\bar{x}) = M_2 - M_1^2$$

(64) $$P = 3 \cdot \frac{\bar{x} - Me}{\sigma}$$

(65) $$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

(66) $$\alpha_3 = \frac{M_3(\bar{x})}{\sigma^3}$$

(67) $$K = \frac{Q_3 - Q_1}{2(D_9 - D_1)}$$

(68) $$\alpha_4 = \frac{M_4(\bar{x})}{\sigma^4}$$

4. Sokaság több ismérv szerinti vizsgálata

(69) $$V_j = \frac{A_j}{B_j} \qquad j=1,2,\ldots,M$$

(70) $$\overline{V} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} B_j}$$

(71) $$\overline{V} = \frac{\sum_{j=1}^{M} B_j \cdot V_j}{\sum_{j=1}^{M} B_j}$$

(72) $$\overline{V} = \frac{\sum_{j=1}^{M} A_j}{\sum_{j=1}^{M} \frac{A_j}{V_j}}$$

(73) $$\overline{x}_j = \frac{\sum_{i=1}^{N_j} x_{ij}}{N_j} = \frac{S_j}{N_j} \qquad j=1,2,\ldots,M$$

(74) $$\overline{x} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j} x_{ij}}{N} = \frac{\sum_{j=1}^{M} S_j}{N} = \frac{S}{N}$$

(75) $$\overline{x} = \frac{\sum_{j=1}^{M} N_j \cdot \overline{x}_j}{\sum_{j=1}^{M} N_j}$$

(76) $$\bar{x} = \frac{\sum_{j=1}^{M} S_j}{\sum_{j=1}^{M} \frac{S_j}{\bar{x}_j}}$$

(77) $$\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2 + \sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(\bar{x}_j - \bar{x}\right)^2$$

(78) $$\sigma = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}\right)^2}{N}}$$

(79) $$\sigma_B = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2}{N}}$$

(80) $$\sigma_K = \sqrt{\frac{\sum_{j=1}^{M} N_j\left(\bar{x}_j - \bar{x}\right)^2}{N}}$$

(81) $$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N_j}\left(x_{ij} - \bar{x}_j\right)^2}{N_j}} \qquad j=1,2,\ldots,M$$

(82) $$\sigma_B = \sqrt{\frac{\sum_{j=1}^{M} N_j \sigma_j^2}{N}}$$

(83) $$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

(84) $$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{N} \qquad i=1,2,\ldots,r \qquad j=1,2,\ldots,c$$

(85) $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

(86) $$C = \sqrt{\frac{\chi^2}{N \cdot \min\{(r-1),(c-1)\}}}$$

(87) $$T = \sqrt{\frac{\chi^2}{N \cdot \sqrt{(r-1)(c-1)}}}$$

(88) $$PRE = \frac{E_1 - E_2}{E_1}$$

(89) $$\lambda_{Y|X} = \frac{\sum_{i=1}^{r}\max_j f_{ij} - \max_j f_{.j}}{N - \max_j f_{.j}}$$

(90) $$H^2 = \frac{SST - SSB}{SST} = \frac{SSK}{SST} = 1 - \frac{\sigma_B^2}{\sigma^2} = \frac{\sigma_K^2}{\sigma^2}$$

(91) $$r_S = 1 - \frac{6\sum_{i=1}^{N}(R_{x_i} - R_{y_i})^2}{N(N^2-1)}$$

(92) $$C_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

(93) $$d_{x_i} = x_i - \bar{x}$$

(94) $$d_{y_i} = y_i - \bar{y}$$

(95) $$C_{xy} = \frac{\sum_{i=1}^{N} d_{x_i} d_{y_i}}{N}$$

(96) $$C_{xx} = \frac{\sum_{i=1}^{N} d_{x_i} d_{x_i}}{N} = \frac{\sum_{i=1}^{N} d_{x_i}^2}{N} = \sigma_x^2$$

(97) $$C_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

(98) $$r = \frac{C_{xy}}{\sigma_x \sigma_y}$$

5. Standardizálás és indexszámítás

(99) $$K' = \frac{\sum B_S V_1}{\sum B_S} - \frac{\sum B_S V_0}{\sum B_S}$$

(100) $$K'' = \frac{\sum B_1 V_S}{\sum B_1} - \frac{\sum B_0 V_S}{\sum B_0}$$

(101) $$K = \overline{V_1} - \overline{V_0} = K' + K''$$

(102) $$I' = \frac{\sum B_S V_1}{\sum B_S} : \frac{\sum B_S V_0}{\sum B_S}$$

(103) $$I'' = \frac{\sum B_1 V_S}{\sum B_1} : \frac{\sum B_0 V_S}{\sum B_0}$$

(104) $$I = \frac{\overline{V_1}}{\overline{V_0}} = I' \cdot I''$$

(105) $$i_p = \frac{p_1}{p_0}$$

(106) $$i_q = \frac{q_1}{q_0}$$

(107) $$i_{\mathrm{v}} = \frac{\mathrm{v}_1}{\mathrm{v}_0}$$

(108) $$i_{\mathrm{v}} = i_q \cdot i_p$$

(109) $$\sum_{i=1}^{N} q_i p_i = \sum_{i=1}^{N} \mathrm{v}_i$$

(110) $$I_{\mathrm{v}} = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum \mathrm{v}_1}{\sum \mathrm{v}_0}$$

(111) $$I_q^0 = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

(112) $$I_q^1 = \frac{\sum q_1 p_1}{\sum q_0 p_1}$$

(113) $$I_p^0 = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

(114) $$I_p^1 = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

(115) $$I_{\mathrm{v}} = I_q^0 \cdot I_p^1 = I_q^1 \cdot I_p^0$$

(116) $$I_q^0 = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0} = \frac{\sum i_q \mathrm{v}_0}{\sum \mathrm{v}_0}$$

(117) $$I_q^1 = \frac{\sum p_1 q_1}{\sum \frac{q_0}{q_1} p_1 q_1} = \frac{\sum \mathrm{v}_1}{\sum \frac{\mathrm{v}_1}{i_q}}$$

(118) $$I_p^0 = \frac{\sum \frac{p_1}{p_0} q_0 p_0}{\sum q_0 p_0} = \frac{\sum i_p \mathrm{v}_0}{\sum \mathrm{v}_0}$$

(119) $$I_p^1 = \frac{\sum q_1 p_1}{\sum \frac{p_0}{p_1} q_1 p_1} = \frac{\sum \mathrm{v}_1}{\sum \frac{\mathrm{v}_1}{i_p}}$$

(120) $$I_{\mathrm{v}} = \frac{\sum \mathrm{v}_0 i_{\mathrm{v}}}{\sum \mathrm{v}_0} = \frac{\sum \mathrm{v}_1}{\sum \frac{\mathrm{v}_1}{i_{\mathrm{v}}}}$$

(121) $$\frac{\sum q_1 p_1}{I_p}$$

(122) $$I_q^F = \sqrt{I_q^0 I_q^1}$$

(123) $$I_p^F = \sqrt{I_p^0 I_p^1}$$

(124) $$I_b^a(c/d) \qquad a,c,d: 0,1,...,t,...,T \qquad b: p, q, \mathrm{v}$$

(125) $$\frac{I_q^1}{I_q^0} = \frac{I_p^1}{I_p^0} = 1 + \mathrm{v}_{i_q} \cdot \mathrm{v}_{i_p} \cdot r_{i_q i_p}$$

(126) $$\sigma_{i_q} = \sqrt{\frac{\sum \mathrm{v}_0 \left(i_q - I_q^0 \right)^2}{\sum \mathrm{v}_0}}$$

(127) $$\sigma_{i_p} = \sqrt{\frac{\sum \mathrm{v}_0 \left(i_p - I_p^0 \right)^2}{\sum \mathrm{v}_0}}$$

(128) $$\mathrm{v}_{i_q} = \frac{\sigma_{i_q}}{I_q^0}$$

(129) $$\mathrm{v}_{i_p} = \frac{\sigma_{i_p}}{I_p^0}$$

(130) $$C_{i_q i_p} = \frac{\sum \mathrm{v}_0 \left(i_q - I_q^0 \right) \left(i_p - I_p^0 \right)}{\sum \mathrm{v}_0}$$

(131) $$r_{i_q i_p} = \frac{C_{i_q i_p}}{\sigma_{i_q} \sigma_{i_p}}$$

6. Kétváltozós regresszió- és korrelációszámítás

(132) $$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad i=1,2,\ldots,n \qquad 2<n<N$$

(133) $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

(134) $$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

(135) $$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

(136) $$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

(137) $$\hat{x} = \hat{\gamma}_0 + \hat{\gamma}_1 y$$

(138) $$\hat{\beta}_1 = \frac{C_{xy}}{\sigma_x^2}$$

(139) $$\sqrt{\hat{\beta}_1 \hat{\gamma}_1} = \left| r \right|$$

(140) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} d_{x_i} d_{y_i}}{\sum_{i=1}^{n} d_{x_i}^2}$$

(141) $$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

(142) $$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

(143) $$E = \frac{dy}{dx} \cdot \frac{x}{y}$$

(144) $$E = \hat{\beta}_1 \frac{x}{\hat{y}}$$

(145) $$\hat{y} = \hat{\beta}_0 \hat{\beta}_1^x$$

(146) $$\hat{y} = \hat{\beta}_0 x^{\hat{\beta}_1}$$

(147) $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

(148) $$\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{x}$$

(149) $$\hat{y} = \frac{\hat{\beta}_0}{\hat{\beta}_1 + x}$$

(150) $$r^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

(151) $$I = \sqrt{1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Irodalom

Denkinger G.: Valószínűségszámítás, Nemzeti Tankönyvkiadó, Budapest, 1997.

Éltető Ö.-Meszéna Gy.-Ziermann M.: Sztochasztikus módszerek és modellek, Közgazdasági és Jogi Könyvkiadó, Budapest, 1982.

Hunyadi L.-Mundruczó Gy.-Vita L.: Statisztika, Aula Kiadó, Budapest, 1996.

Kerékgyártó Gy.-Mundruczó Gy.: Statisztikai módszerek a gazdasági elemzésben, Aula Kiadó, Budapest, 1994.

Köves P.–Párniczky G.: Általános Statisztika, Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

Lukács O.: Matematikai statisztika, Műszaki Könyvkiadó, Budapest, 1987.

Meszéna Gy.-Ziermann M.: Valószínűségelmélet és matematikai statisztika, Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

Mundruczó Gy.: Alkalmazott regressziószámítás, Akadémiai Kiadó, Budapest, 1981.

Spiegel, M. R.: Statisztika (elmélet és gyakorlat), Panem-McGraw-Hill, Budapest, 1995.