Bevezetés a sportstatisztikába

Ács Pongrác – Pintér József

Bevezetés a sportstatisztikába

Ács Pongrác – Pintér József

Publication date 2011

Tartalom

Az ábrák listája

A táblázatok listája

Bevezetés a sportstatisztikába

Ács Pongrác – Pintér József

Pécsi Tudományegyetem • Pécs, 2011

Kézirat lezárva: 2011. november 30.

ISBN: 978-963-642-412-1

Pécsi Tudományegyetem

A kiadásért felel: Dr. Bódis József

Felelős szerkesztő: Schenk Borbála

Műszaki szerkesztő: Dialóg Campus Kiadó – Nordex Kft.

1. fejezet - A statisztika tárgya, a statisztika és sportstatisztika szerepe

A bennünket körülvevő világban számszerű adatok, információk, mértékek nélkül nehezen tudnánk eligazodni. Mindennapi életünkbe ezek a kvantitatív információk már oly szervesen beépültek, hogy alkalmazásuk automatikussá vált. Elég, ha ezen a helyen a tömegkommunikáció információáradatát említjük meg. A különféle információk felhasználása során nem mindig gondolunk arra, hogy ezek nagy része olyan gyakorlati és elméleti tevékenység eredménye, amit közismerten statisztikának hívunk.

A társadalmi-gazdasági fejlődés már a történelem korai szakaszában igényelte a számszerű információkat, és a statisztika[1] szó etimológiai értelemben is szorosan kapcsolódik az államhoz, mivel kezdetben elsődleges feladata annak felépítéséről, berendezkedéséről és állapotáról való ismeretek gyűjtése és közlése volt. Ez a tevékenység azonban fokozatosan önálló módszertudománnyá vált, amely már nem kötődik csupán a társadalmi–gazdasági jelenségekhez. A modern statisztika mint módszertudomány már közvetlenül kapcsolódik más tudományágakhoz; nemcsak a társadalomtudományok, hanem a természettudományok is igénybe veszik sajátos logikáját, felhasználják a módszerek által kapott eredményeket. A számszerű adatok megfigyelése, feldolgozása, elemzése, közzététele és felhasználása napjainkban önálló tudományos módszereket igényel.

Egy új diszciplína tanulmányozásának kezdetén a tudományág pontos definiálásán túl nehézséget jelenthet a terminológia megismerése. Az előzőekben tulajdonképpen egyfajta definíciót is adtunk, amelynek pontosítása, megismétlése mindenképpen indokolt.

A **statisztika** *a tömegesen előforduló jelenségekre, folyamatokra vonatkozó információk összegyűjtésének, leírásának, elemzésének, értékelésének és közlésének tudományos módszertana.*

A definícióban igen fontos hangsúlyt kap a **tömeges** jelző. A statisztika ugyanis csak nagy számban, tömegesen előforduló jelenségek, folyamatok vizsgálatából tud levonni következtetéseket. A statisztikusi munka magában foglalja a numerikus információk tömegéből a legfontosabb jellemzők kiragadását és szintézisét is.

A nemzetközi irodalomban elterjedt csoportosítás szerint megkülönböztethetjük a **leíró statisztikát,** a **következtetéses statisztikát** és a **statisztikai döntéselméletet**.

A **leíró statisztika** alapvetően a numerikus információk összegyűjtését, az információk összegezését, és tömör jellemzését szolgáló módszereket foglalja magában. A leíró statisztika legfontosabb területei:

- az adatgyűjtés,
- az adatok ábrázolása,
- az adatok csoportosítása, osztályozása,
- az adatokkal végzett egyszerűbb aritmetikai műveletek,
- az eredmények megjelenítése.

[1] A statisztika szóban felismerhető a görög eredet, a **στατζειμυ,** és a latin status szó, ami az állam, illetve az állapot gyökérre vall.

A leíró statisztika eredményeinek szemléltetésére álljon itt két egyszerű példa:

A PTE-PEAC asztalitenisz-szakosztály játékosainak létszáma 2005. január 1-jén 107 fő. A férfiak a szakosztály 88%-át alkotják.

A **következtetéses statisztika** segítségével a jelenségekre, folyamatokra vonatkozóan olyan megállapításokat tehetünk, amelyek nem csak a közvetlen megfigyelésen alapulnak. Igen leegyszerűsítve, alkalmazásával közvetlenül nem mérhető, csak összetett statisztikai, matematikai eljárásokkal megszerezhető számszerű információkat nyerhetünk. A következtetéses statisztika szorosan épít a matematikai statisztikára és a valószínűség-elméletre.

- 20 európai ország sportbajnokságainak adatai alapján végzett vizsgálatból megállapítható, hogy általában az 5%-kal magasabb gólátlag átlagosan 0,11%-os nézőszám-növekedéssel jár együtt.

- Egy közvélemény-kutatás előrejelzése alapján a válaszadók 60%-a szerint a következő labdarúgó-bajnokságot a Ferencváros csapata nyeri.

- Egy sportcsarnokban lévő világítótestek várható élettartama 3000 óra.

A fenti példák mindegyikében érződik a véletlenszerűség szerepe, amely a jelenségek kapcsolatában, az események kimenetelében egyaránt érvényesülhet. Vagyis a fenti események egy meglévő – ismeretlen vagy megközelítően ismert – valószínűség mellett következnek be.

A **statisztikai döntéselmélet** a véletlen környezet által bekövetkező események figyelembevétele mellett, több cselekvési lehetőség közül az optimálisnak vélt kiválasztásához ad számszerű információkat. Az empirikus statisztikai megfigyeléseken, és a következtetéseken túl szubjektív, szakértői értékítéleteknek is teret enged. A valószínűség-elmélet és a játékelmélet elemeit kombinálja a statisztikai megfigyelések eredményeivel.

A sportgazdasági életből számtalan példát hozhatunk a statisztikai döntéselmélet alkalmazására. Pl.: beruházási döntések (új sportcsarnok építésére vonatkozó döntés), új termékek bevezetésére (újfajta síléc bevezetése), profilváltásra vonatkozó döntések, pénzügyi döntések (igazolások, átigazolások) stb.

Természetesen találkozhatunk a statisztikát felölelő módszerek és eszközök más irányú, más rendszerű csoportosításával is. Különösen az elmúlt évtizedekben volt kiemelkedő szerepe a nemzetgazdaság különféle ágazatainak kérdéseit tárgyaló ún. szakstatisztikáknak (pl.: ipari, mezőgazdasági statisztika stb.), de ma is megkülönböztetünk népességstatisztikát, egészségügyi statisztikát, igazságügyi statisztikát, sportstatisztikát stb. Különös figyelmet kap minden országban a nemzetgazdaság legfőbb elszámolásait, összefüggéseit feltáró és közlő gazdaságstatisztika.

A statisztika szerepének egyre erőteljesebb növekedését figyelhetjük meg a gazdasági problémák megoldásán túl az orvosi, a mezőgazdasági és a biológiai kérdések megválaszolása során is. Különösen „előkelő” szerepet tölt be a statisztika az üzleti problémák elemzése terén. A számszerű információk iránti igény napjainkban egyre inkább növekszik, ennek kielégítését a számítógépek elterjedése jelentősen megkönnyíti. A számítógépes háttér az adatok mennyiségének és komplexitásának növelését teszi lehetővé.

A döntési problémák köre – amely szintén épít a statisztikára – igen szerteágazó, ezért csak néhányat említünk itt meg az üzleti, közgazdasági életből.

A **marketing** hatékony művelése elképzelhetetlen statisztikai módszerek és eljárások nélkül. A piac megismerése az állami statisztikai adatszolgáltatáson túl konkrét, célra orientált statisztikai – többnyire reprezentatív – adatfelvételt, és feldolgozást, valamint

elemzést is igényel.

A vállalati **management** az emberi erőforrások hatékony működtetése és az irányítás területén nem nélkülözheti sem a statisztikai adatbázisokat, sem a statisztika módszereit, amelyekkel az erőforrások működésének mélyebb összefüggéseit lehet megismerni.

A **pénzügyi** problémák kezelése során mindennapos a statisztikai adatok és módszerek felhasználása. A beruházási, finanszírozási döntések mellett igen fontos helyet foglal el a különféle pénzügyi előrejelzések statisztikai módszerekkel történő megalapozása.

Természetesen a fenti felsorolást tovább folytatva a gazdasági élet szinte valamennyi területe igényli a statisztikát. Nélküle sem hatékony gyakorlati, sem eredményes elméleti munkát nem végezhetünk.

2. fejezet - A sportstatisztika szerepe

A sportban, mint az élet más területén is találkozunk mérhető, összehasonlítható adatokkal.

A sport egyik fő mozgatóeleme az elért eredmények, rekordok megdöntése. Ezekhez mindenképpen szükséges, hogy a már meglévő adatokat ismerjük, összegyűjtsük, rendszerezzük és értékeljük. A mai modern sporttudománynak egyik elengedhetetlen és minden területen megtalálható része: a sportstatisztika. A sportban fellelhető számszerű információkat már évszázadok óta gyűjtik és rendszerezik. Ezeknek a feldolgozásához és megértéséhez gyakran használjuk a statisztika módszereit. Gondoljunk csak bele, hogy egy edző miként készítené el a korszerű edzéstervét, ha nem alkalmazna statisztikai módszereket (pl. a felkészülés második hetében a maximális erő 60%-ával végezzünk fekvenyomást).

De nem kell ilyen messze mennünk, elég, ha a sportorvosnál lévő adatokra gondolunk, ha ki akarjuk számolni, hogy a sportoló az elmúlt öt évben átlagosan mennyit hízott vagy nőtt.

Tényként kell leszögeznünk, hogy a sportteljesítmény-adatok sokassága ellenére, a sportgazdasági kutatásokhoz felhasználható alapstatisztikai adatbázisok terén óriási lemaradásunk van (Ács, 2009). A Központi Statisztikai Hivatal sportot is érintő adatállománya roppant szűkös, a mai napig nem tudjuk pontosan, hogy mennyien vannak az igazolással rendelkező sportolók (élsportolók) hazánkban. Az évek óta emlegetett Sportinformációs rendszer nem tudja még (mindig) a célként kitűzött feladatát ellátni.

A sportközgazdaságtan is gyakran használja a statisztikát, mivel itt a kvantitatív adatoknak kiemelkedő szerepe van. Gondoljunk akár a játékosok béreinek, juttatásainak megállapítására. Napjainkban ezeket az információkat már többnyire számítógépeken tároljuk és a szükséges számításokat is ezeknek a segítségével végezzük.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Fogalmazza meg a statisztika és sportstatisztika definícióját, csoportosítsa és jellemezze a nemzetközi irodalmaknak megfelelően!
- Mondjon három konkrét sportpéldát a következtetéses statisztikára!
- Milyen területeken találkozhatunk a hétköznapokban a statisztikával?
- Mivel indokolja a sportstatisztika létjogosultságát?
- Mondjon három sportágat, ahol a sportstatisztikát alkalmazhatjuk!

3. fejezet - Néhány statisztikai alapfogalom

Mint minden tudományág, úgy a statisztika is sajátos nyelvezettel bír, amelynek elsajátítása nélkülözhetetlen a tárgy megismerése során. Itt csak a legszükségesebb fogalmakat vezetjük be, amelyek mindenütt általános érvénnyel alkalmazhatók. A további speciális definíciókkal a konkrét kérdéskörök megismerésekor találkozunk majd.

Statisztikai sokaságnak nevezzük a statisztikai megfigyelés tárgyát képező egyedek összességét. A sokaság fogalmilag meghatározott. A sokaság legkisebb részeit, egyedeit megfigyelési **egységek**nek nevezzük. A sokasági egység az az egyed, amelyre a sokaság fogalmi meghatározottsága még ráillik. A sokaság egységei lehetnek élőlények, szervezetek, tárgyak, események, képzett egységek stb.

Statisztikai sokaságot alkothatnak emberek, sportolók csoportjai, pl.: a magyar iparban foglalkoztatottak; egy adott bajnokságban szereplő játékosok; az ország stadionjai; Magyarország rally versenyautóinak állománya, a felsőoktatásban használt sportszerek.

A sokaságnak alapvetően két típusát szokás megkülönböztetni. Eszerint beszélhetünk **álló sokaság**ról (stock) és **mozgó sokaság**ról (flow). Az álló sokaság állapotot fejez ki, élőlényekből, tárgyakból és szervezetekből állhat, és az adatok időpontra vonatkoztatva értelmezhetőek (ezt az időpontot szokás ún. eszmei időpontnak is nevezni). A mozgó sokaság folyamatot fejez ki, és időtartamra értelmezhető.

Álló sokaság a 2005/2006-os vízilabda-bajnokságban szereplő csapatok száma, például 2006. január 12-én.

Mozgó sokaság például a 2005-ben külföldről érkezett sportolók száma.

Az álló és a mozgó sokaság természetesen nem független egymástól. Ha folyamatok eredményeit egy adott időpontban mérjük, már álló sokaságról beszélhetünk. Például az évente igazolt játékosok számának (flow) összege alkotja egy adott időpont igazolt játékosállományát (stock).

A sokaság tartalmazhat **véges** vagy **végtelen** számosságú egyedeket. A társadalmi-gazdasági vizsgálatok általában véges számú egyeddel operálnak, mivel területileg és időben pontosan körülhatárolhatók a sokaságok. A különféle kísérleti statisztikákban, a mintavételes eljárásokban és a folyamatok modellezése során találkozhatunk végtelen egyedeket tartalmazó sokasággal.

A sokaság egyedeit a statisztika módszereivel számba vehetjük, megfigyelhetjük. Amennyiben a sokaság egészének megfigyelését valamilyen előre elhatározott szempont szerint végezzük, **teljes körű megfigyelést** hajtunk végre (klasszikusan teljes körű megfigyelést, a népszámlálás nyomán szokás cenzusnak is nevezni). Ha a megfigyelés a sokaságnak csak meghatározott egyedeire terjed ki, **részleges megfigyelést** végzünk. A részleges megfigyeléseken belül kiemelkedően fontos szerepe van a **reprezentatív megfigyelés**nek. A részleges, de különösen a reprezentatív megfigyelés során célunk az alapsokaság egy részének, a **mintasokaság**nak a segítségével a teljes sokaságra vonatkozó következtetések megfogalmazása.

Teljes körű adatfelvétel például a népszámlálás. De nem képzelhető el minden egyedre vonatkozó felvétel egy roncsolásos minőség-ellenőrzési vizsgálat esetén, ezért ebben az esetben reprezentatív megfigyelést végzünk. Mégis mindkét esetben a sokaság egészére vonatkoztatjuk megállapításainkat.

A reprezentatív megfigyelésnek több fajtája van. A **véletlen (valószínűségi)** minta garantálja azt, hogy az alapsokaság bármely eleme azonos eséllyel kerüljön a mintába,

ezért lehetőség nyílik a statisztikai eszközök széles körű felhasználására.

Monográfiának hívjuk azt a részleges adatfelvételt, amely a sokaság néhány jellemzőjének előzetes, a priori ismeretében egy (vagy kevés) kiemelt egyed részletes statisztikai vizsgálatát adja. Az ilyen jellegű vizsgálatok eredményei matematikai–statisztikai módszerekkel nem, vagy csak nehezen elemezhetőek. Különösen a szociológiában népszerű a monografikus adatfelvétel.

A reprezentatív megfigyelés létjogosultsága a gazdasági jelenségek vizsgálatánál korábban sem volt elhanyagolható. Szerepe, jelentősége napjainkban mindenképpen növekszik. Ebbe az irányba hat a gazdasági jelenségek és folyamatok gyors változása, és ennek nyomon követési igénye, ami a gyors és megbízható reprezentatív adatgyűjtés nélkül megoldhatatlan.

Már itt le kell szögezni, hogy a statisztika módszertana alapvetően támaszkodik a reprezentatív adatfelvételre. A mintavétel és az ezen alapuló következtetés egyfajta statisztikai látásmódot jelent, amely segíti a nagy számosságú alapsokaságról levonható főbb, tendenciózus következtetések megfogalmazását. Gondoljunk a manapság egyre többször hangoztatott, olimpiai rendezéssel kapcsolatos megkérdezésekre. Egy közvéleménykutató intézet estleges felmérése alapján (amely reprezentálja az ország lakosságát) lehet következtetni arra, hogy a lakosság szeretné-e vagy nem az olimpia megrendezését.

A statisztikai sokasággal szervesen összefüggő fogalom az ismérv. **Statisztikai ismérv**nek nevezzük a statisztikai sokaság egyedeire vonatkozó tulajdonságokat, jellemzőket. A statisztikai sokaság az általa hordozott fogalom tekintetében homogén, de sok tulajdonsága tekintetében heterogén. Ezeknek a különbözőségeknek a kifejezői az ismérvek. A sokaság egységei az ismérvek hordozói. Az ismérv lehetséges kimenetelei az **ismérvváltozatok**.

Ismérv lehet például a sportolók kora, neme, területi elhelyezkedése, sportága, vállalatoknál a termelés értéke, az elektromos energiafogyasztás időbeli alakulása. Ismérvváltozatok lehetnek a területi elhelyezkedés vonatkozásában például Budapest, illetve vidék; vagy városok, községek. A sportszektorban termelési érték esetében ismérvváltozat lehet egy sportesemény bevétele (pl. a belépők eladásából bejövő konkrét számadat).

Amennyiben egy ismérv csupán két ismérvváltozattal rendelkezik, **alternatív** ismérvről beszélhetünk.

Alternatív ismérvre jó példa a nemhez való tartozás: férfi, nő; egy csapat szerkezetének megoszlása kapus, mezőnyjátékos

Általánosságban az ismérvek lehetnek: **időbeli**, **területi**, **minőségi** és **mennyiségi ismérvek**.

Időbeli ismérv: egy csapat tagjainak születési éve. Területi ismérv a sokaság egységeinek valamilyen földrajzi (pl. megyei, városi) jellemzője, tulajdonsága. Minőségi ismérv: csapatok minősítése. Mennyiségi ismérv a sokaság egységeire vonatkozó számszerű megjelölés: gólpasszok száma, jövedelem, életkor stb.

Azokat az ismérveket, amelyek a statisztikai sokaság valamennyi egyedére jellemzőek, definiálják a sokaságot, **közös ismérvek**nek hívjuk; míg azokat az ismérveket, amelyek szerint az egyedek különböznek egymástól, **megkülönböztető ismérvek**nek nevezzük. Természetesen ebből a definícióból is látható, hogy a két ismérv fogalma relatív viszonyban van egymással.

Az ismérvek megjelölésére gyakran a változó kifejezést is használják. A **mennyiségi változó** (méréses jellemző) az ismérvváltozatokat számokkal, míg a **minőségi változó** (minősítéses jellemző) minőségi jegyekkel, fogalmakkal jellemzi. A mennyiségi változók mindig számértékek; amelyek lehetnek **folytonos változók**, amelyeket végtelen

pontossággal mérhetünk és a sokféle méréssorozat közül csak egy fogadható el (ilyen, pl. a magasság, súly stb.), de beszélhetünk **diszkrét változókról** is, amelyek véges ismérvváltozattal rendelkeznek. Ezek általában egész számok (pl: a lőtt gólok száma csapatonként).

A társadalmi-gazdasági jelenségek kölcsönhatásban vannak egymással, függnek egymástól, feltételezik egymást. A jelenségek és folyamatok az ismérvek segítségével jellemezhetőek, ezért célszerűnek tűnhet, ha a köztük levő kapcsolatok típusait – amelyek az ismérvek közötti kapcsolatként jelennek meg – már itt definiáljuk.

Ha az egyik statisztikai ismérv egyértelműen meghatározza a másik ismérv egy konkrét változatához való tartozást, **függvényszerű kapcsolat**ról beszélünk.

Egy sportszaküzletben a munkaviszony kezdete (időbeli ismérv) egyértelműen meghatározza a munkában eltöltött időt (mennyiségi ismérv), tehát közöttük függvényszerű kapcsolat van.

Amennyiben az egyik ismérv szerinti hovatartozásból nem következtethetünk a másik ismérv konkrét változatára, **függetlenség**ről beszélünk. Ha az ismérvek között tendenciaszerűen érvényesülő kapcsolatot észlelünk, tehát az egyik ismérvváltozathoz való tartozásból csak **valószínűségi** jelleggel következtethetünk a másik ismérvváltozatra, annak átlagos bekövetkezésére, **sztochasztikus kapcsolat**ról van szó.

A bajnokság minőségének javulása (gólok számának növekedése) a nézők számának emelkedésével jár együtt, ami általános szabálynak tekinthető. Ez azonban nem jelenti azt, hogy a sokaság (lakosság, szurkolók) valamennyi egyede törvényszerűen így is viselkedik. Elképzelhető, hogy egy csapat eredményeinek javulása ellenére is csökken a nézőszám. A többség a fenti törvényszerűség szerint viselkedik, és az esetleges „kivétel” erősíti a szabályt.

Hasonló módon interpretálhatjuk egy olimpiára való felkészülés ideje és az elért eredmény közötti összefüggést. Nagy valószínűséggel számíthat jó teljesítményre az a sportoló, aki több időt töltött edzéssel és felkészüléssel.

A teljes, determinisztikus meghatározottságra jó példa egy versenyen megtett út és a sebesség viszonya. Nagyobb sebességgel – adott idő alatt – több utat lehet megtenni. (Természetesen itt ki kell zárni a véletlen baleset és egyéb tényezők szerepét.)

A sztochasztikus összefüggés a függetlenség és a teljes meghatározottság között foglal helyet. A sztochasztikus kapcsolatnak alapvetően háromféle típusát különböztethetjük meg:

- **asszociációs** kapcsolatnak hívjuk a minőségi ismérvek kapcsolatát;
- **vegyes típusú** a sztochasztikus kapcsolat, ha az egyik oldalon minőségi ismérv – mint ok – a másik oldalon mennyiségi ismérv – mint okozat – szerepel;
- **korrelációs** kapcsolatról beszélhetünk, ha a kapcsolatot mennyiségi ismérvek közvetítik.

Asszociációs kapcsolatra példa lehet a csapatok szimpátia vizsgálata: lakóhely és a kedvenc csapatok kapcsolata közötti viszony; vegyes kapcsolatként értékelhetjük a bajnokság szintje és a jövedelem közötti összefüggést; korrelációs kapcsolatra példa, hogy hosszabb időtávon a gólok száma függ a csatárok számától.

A statisztikai megfigyelés során az adatfelvétel tárgyát, a sokasági egyedeket **megfigyelési egység**nek nevezzük. A megfigyelési egységek adatait általában kérdőívek segítségével gyűjthetjük össze. (Nem szabad megfeledkezni a modern technika segítségével megvalósítható korszerűbb eljárásokról sem, pl. telekommunikáció). A kérdőíveknek két fő típusa van: **egyéni kérdőív** és **lajstrom**. Az egyéni kérdőívre egy megfigyelési egység, a lajstromra több megfigyelési egység adatai

kerülnek. A kérdőívek tartalmazzák a kérdéseket, lehetséges válaszokat, azonosítási adatokat és a feldolgozást segítő kódokat. Az adatfelvétel során az adatszolgáltató vagy maga tölti ki a kérdőívet, amit **önszámlálás**nak hívunk, vagy számlálóbiztost bízunk meg a szakszerű adatfelvétellel, ez utóbbit **kikérdezéses eljárás**nak nevezzük. Ez az adatfelvételi mód akkor indokolt, ha nem várhatjuk el a megkérdezettektől a szakszerű válaszokat. Természetesen mindkét eljárásnál sok függ a jól szerkesztett, logikus kérdőívtől. A kérdőív szerkesztésének általános elvei ugyan megadhatók, azonban az adatgyűjtés tárgyát képező jelenség alapos megismerése elengedhetetlen. A kérdőív „jóságának” ellenőrzése érdekében gyakorta végeznek ún. próbafelvételeket, ami a tervezett statisztikai adatfelvétel minőségét is jelentősen javíthatja.

A megfigyelési egységek adatainak összegyűjtése során törekedni kell a bizonylati elv érvényesülésére, azaz az adatok valóságtartalma – lehetőség szerint – bizonylatokkal legyen alátámasztható.

Az adatok keletkezése, gyűjtése során igen fontos szerepet tölt be a közvetlen észlelés, pl. halálokok, közlekedési balesetek stb. Az észlelés eredményeit – természetesen – bizonylatokon rögzítik.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Példákon keresztül mutassa be a sztochasztikus összefüggéseket!
- Az Újpest labdarúgócsapata az első osztályban szerepel a hazai labdarúgó-bajnokságban. Milyen ismérvről beszélhetünk itt?
- Mutassa be az ismérvek fajtáit egy-egy sportpéldán keresztül!

4. fejezet - Statisztikai adatok

A sportstatisztika számadatokkal dolgozik, melyekhez számlálás, mérés útján jutunk. Természetesen az adatok keletkezésének körülményei behatárolják az adatok jellegét, felhasználhatóságának körét.

A Statisztikai adat tehát olyan tapasztalati, empirikus szám, amely **mérés** vagy **számlálás** útján keletkezik. A statisztikai adat a sokaság valamilyen számszerű jellemzője. A statisztikai adat sajátossága – ellentétben a matematikai ún. tiszta számmal –, hogy a numerikus értékhez tartozik a sokaság, a hely és idő megjelölése, valamint a mértékegység.

Az Alba Volán jégkorongcsapatának a játékoskerete a 2005/2006-os bajnoki idényben 27 fő volt. A 2005/2006-os első osztályú labdarúgó-bajnokságban a külföldi játékosok aránya 11,37%-ra csökkent, 2005-ben hazánkban 111 951 db kerékpárt gyártottak.

A statisztikai adatok között megkülönböztetünk **abszolút** és **származtatott adatok**at. Az abszolút adatok (a kosárlabdázók száma és a kerékpártermelés adata) számlálás, mérés útján jönnek létre. A származtatott adatokhoz az abszolút adatokkal végzett műveletek segítségével juthatunk. Ezek, mint később látni fogjuk, viszonyszámok, átlagok stb.

Például ha a 2009-es első osztályú bajnokságban szereplő labdarúgócsapatok játékoskeretét (387 főt) a 2010-es játékoskerethez (381 főhöz) viszonyítjuk, akkor a játékoskeret az elmúlt évhez képest 1,0157 szeresére, vagyis 101,57%-ra változott, tehát a játékoskeret 1,57%-kal bővült.

A rendszeresen ismétlődő társadalmi, gazdasági jelenségek elemzésére, jellemzésére szolgáló statisztikai mérőszámokat **statisztikai mutatószámok**nak, **statisztikai mutatók**nak nevezünk.

A sportban a termelékenység mérőszámaként használhatjuk a rúgott gólok számát (pl.: egy csapatnál a rúgott gólok száma egy évben 237 db). Gyakran használt mutató az átlagosan pályán töltött idő (pl. egy adott évben egy játékos átlagosan 82 percet töltött a pályán). A sport területi sűrűségének a mutatószáma lehet az ezer főre jutó első osztályú csapatok száma (2004-ben a dél-dunántúli régióban 1,22). Természetesen folytathatnánk a sort mind a társadalmi, mind a gazdasági jelenségek vizsgálata terén.

A gazdasági és társadalmi jelenségek összefüggő, komplex rendszerben történő vizsgálata igen hasznos információkat szolgáltat mind az elemzés, mind a döntés-előkészítés számára. A különféle tudományágakban használt **modell** fogalom a statisztikai jellegű vizsgálódásokban is megjelenik. A társadalmi–gazdasági jelenségek és folyamatok vizsgálata során alkalmazott **modell** a vizsgált valóságnak – a vizsgálat szempontjából – legfontosabb vonásait, összefüggéseit kifejező **logikai–matematikai konstrukció**. A valóság formalizált, szimbolikus ábrázolását a statisztikai modellekben egyenletek, egyenletrendszerek segítségével adjuk meg. Jellemző sajátosságuk, hogy a belőlük levont következtetések sztochasztikus, valószínűségi jellegűek, ezért a modellekben a véletlen hatást is szerepeltetjük.

A statisztikai adatoknak minőségi követelményeknek is meg kell felelniük. A statisztikai adattokkal szemben támasztott követelmények általában az elfogadható **pontosság**, a **gazdaságosság** és a **gyorsaság**. El kell mondani, hogy minden feltételnek egy időben tökéletesen megfelelni nem lehet (például a gyorsaság általában a pontosság ellen hat). A fentieknek megfelelően csupán egy **optimum**ot lehet és kell találni a feltételek között.

Tudatában kell lenni annak, hogy a statisztikai adatok legtöbbször csak **korlátozottan pontosak**.

A pontatlanság egy része a sokaság méretéből adódik, illetve a számlálás során

keletkezik. Például a Nemzeti Sport című napilapból ismerjük, hogy a 2009/2010-os magyar férfi asztalitenisz-bajnokságban szereplő csapatok száma tíz; itt a sokaság terjedelme és definíciója sem engedi meg a pontatlanságot. Más esetekben – és ezek vannak többségben – számolni kell azzal, hogy az adatok csak korlátozottan pontosak. Pl. a 2001-es Bajnokok Ligája elődöntő mérkőzést 130 millió ember látta a világon. Ez egy összeírt nézőszámnak felel meg.

Mivel az adatok ezen hibái a mérés természetével, szubjektív elemekkel, emberi hibákkal magyarázhatók, ezért – elvileg – ki is javíthatóak. Gondoljunk csak arra, hogy a népszámlálás során a kimaradt egyedek feltárásával és utólagos adatfelvétellel az adatok pontossága jelentősen javulhat. Ugyancsak növeli az ilyen típusú hibák elkerülésének esélyét az ellenőrzések gyakoriságának növelése. Mindezek azonban olyan faktorok, amelyek költségkihatással járnak, ezért az adatfelvétel során mérlegelni kell, hogy milyen megbízhatóságot kívánunk elérni.

Az adatok pontatlanságának másik része a gazdasági, társadalmi jelenségek összetett jellegéből ered. Például 2005-ben a fogyasztói árindex az előző évhez képest 103,6% volt. Az árak együttes és átlagos 3,6%-os növekedése nem pontos adat, csupán az ármozgás hozzávetőleges nagyságát mutatja. Módszertani finomítással, a megfigyelési egységek körének és számának még pontosabb megválasztásával ez a számadat is pontosítható, azonban így is csak a tendenciát képes közvetíteni.

Mint láthattuk a statisztikai adatokat mérés, számlálás útján kapjuk. A mérés meghatározott szabályok szerinti hozzárendelést jelent dolgokhoz, tulajdonságokhoz. Ezek a hozzárendelési szabályok a **mérési skálákat** határozzák meg. A mérési skáláknak négy fontos **típusa** különböztethető meg:

- nominális skála;
- ordinális skála;
- intervallumskála;
- arányskála.

Nominális (névleges) **skálán** a szimbólumok, számok csak az azonosítást szolgálják, amelyek segítségével elvégezhető a jelenségek, folyamatok osztályozása. Nominális skálát alkalmazunk tipikusan a minőségi ismérv szerinti megfigyeléseknél. Az egyedeket aszerint osztályozzuk ezen a skálán, hogy milyen osztályba, kategóriába tartoznak.

Néhány gyakran használt nominális skála: nemek, hajszín, rendszám, állampolgárság stb.

Bizonyos mennyiségi ismérvek és a segítségükkel képzett intervallumok is visszavezethetők nominális skálára, pl. az egyes ismérvértékek helyett magas értékekről, illetve alacsony értékekről beszélhetünk. Ezen a skálán csak az egyenlőség értelmezhető, amely szerint két megfigyelési egység vagy egyenlő, vagy különböző.

Az **ordinális skálán** sorrendiségre vonatkozó relációk alapján rangsorba rendezzük a megfigyelt objektumokat, egyedeket. A sorrendi skálán az egyes egyedek egymástól nem biztos, hogy egyenlő távolságban helyezkednek el. Ez még nem tekinthető tiszta kvantitatív skálának, habár használja a numerikus értékeket, és gyakran további műveletek is végezhetők a rangszámokkal. Itt már az egyenlőség mellett a sorrendiség relációja is érvényes.

Ilyen típusú skálák az osztályzatok, minősítések, egyéni ranglisták, sporttermékek minőségi osztályai stb.

Az **intervallumskála** már tiszta kvantitatív mértékeket használ, tartalmazza azokat a számértékeket, amelyek jellemzik az egyes egyedek értékeit. A sorrend mellett itt a skála bármely két pontja közötti távolság is értelmezhető. Az intervallumskálának egy nagyon jellegzetes tulajdonsága, hogy nem rendelkezik igazi zéró ponttal. Ez azt jelenti,

hogy a zéró pont meghatározása önkényes, nem tükrözi a zéró érték egyben a tulajdonság hiányát.

Az intervallumskálán mért adatokra vonatkozó példák között elsőként szokták emlegetni a tengerszint feletti magasságot és a hőmérsékletet. Az utóbbi jól jellemzi a skálát. A 20 °C nem fele a 40 °C hőmérsékletnek. Ugyanakkor a skála segítségével konzisztens intervallum-kapcsolat mérhető, mivel 45 °C–40 °C = 25 °C–20 °C.

Intervallumskála a sportban a gólkülönbség, mivel létezhet negatív gólkülönbség is, és a zéró érték is tulajdonság.

Az **arányskála**, vagy hányadosskála igazi kvantitatív skála, a numerikus értékek úgy jellemzik az objektumok, egyedek elrendeződését, hogy azok egyértelműen behatárolhatóak. A skálának igazi zéró pontja van. Mindez azt is jelzi egyben, hogy a nulla érték az ismérv, a tulajdonság hiányát egyértelműen jellemzi. A skála értékei multiplikatív módon transzformálhatóak, bármely két pont aránya független a mértékegységtől, valamennyi matematikai, statisztikai művelet az arányadatokkal elvégezhető.

Az arányskálára sok példát lehetne hozni, csupán jelzésszerűen néhány: hosszúság, súly, költség, magasugrásban elért eredmények stb.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Mit nevezünk a statisztikában adatnak? Mondjon néhány sportadatot!
- Példákon keresztül mutassa be a mérési skálák fajtáit!

5. fejezet - A statisztikai adatok csoportosítása

A statisztikai adatok feldolgozásának alapvetően fontos egyszerű eszköze a csoportosítás vagy osztályozás.

A **csoportosítás** vagy **osztályozás** a statisztikai sokaságnak valamely ismérv szerinti tagolása, rendszerezése.

A csoportosítás lényegében azt is jelenti, hogy a statisztikai sokaságot minőségileg különböző részekre, csoportokra bontjuk, és így tanulmányozzuk szerkezetét, felépítését.

A csoportosítás a gyakorlatban úgy történik, hogy a **csoportképző ismérv** alapján az ismérv változatainak megfelelően a sokaság egyes tagjait a konkrét ismérvváltozatokhoz rendeljük. A csoportképző ismérvek a sokaság lényeges tulajdonságait tükrözik, ezek alapján lehetőség nyílik a sokaságon belül az alapvető különbségek, eltérések feltárására, elemzésére. A csoportosításhoz felhasznált csoportképző ismérv változatai sok esetben adottak (pl. nemek, sportági minősítés), más esetben gyakorlati és elméleti megfontolást igényel a sokféle egyedi tulajdonság tömörítése, a lényeges jegyek megragadása (pl. pályán töltött szerepkör (posztok), keresetek szerinti csoportosítás). Fontos szempont a csoportosítás során, hogy az adatok **egyértelműen besorolhatók** legyenek valamelyik csoportba. Ez annyit jelent, hogy valamennyi egyed **egy és csak egy csoportba** kerülhessen.

A gyakorta ismétlődő csoportosítások – különösen minőségi ismérvek – esetén a rendszeresen használt ismérvváltozatok felsorolását **nomenklatúrá**nak nevezzük. Ismert nomenklatúra pl. a foglalkozások egységes rendszere, a sporttermékjegyzékek, a sportszolgáltatások, sportágak jegyzéke.

Egy statisztikai sokaság egyidejűleg több ismérv szerinti csoportosítását **kombinatív csoportosítás**nak hívjuk. A kombinatív csoportosítás – igen gazdag információtartalma miatt – fontos helyet foglal el a statisztika módszertanában.

A kombinatív csoportosításra az alábbi példát adjuk, amely az ökölvívókat osztályozza:

5.1. ábra - Kombinatív csoportosítás

Forrás: saját szerkesztés

Természetesen további kombinatív csoportosítás is elképzelhető, amely más elemzéseket is lehetővé tesz (pl. az ökölvívók is tovább csoportosíthatók nemzetiség, lakóhely, bajnokság szerint). A kombinatív csoportosításnál fontos, hogy az eredmény áttekinthető legyen.

A csoportosítás során figyelemmel kell lenni arra is, hogy a széttagolás mellett szükségünk lehet az adatok összevonására is. Mindez erősen behatárolja a kombinatív csoportosítás mélységét.

A csoportosítás időbeli, mennyiségi, minőségi és területi ismérvek alapján egyaránt történhet, tipikus csoportosító ismérvek a minőségi és mennyiségi ismérvek.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy női NB II-es kosárlabda-bajnokságban szereplő csapat idei játékoskerete 24 fő. A csapat jelenleg három helyezést javított a tavalyi szerepléshez képest. A szakértők ezt a javulást a játékoskeret 20%-os emelkedésével magyarázzák. Mekkora volt a csapat tavalyi játékoskerete?

- Mondjon példát az abszolút és származtatott adatra!

- Mi a nomenklatúra?

- A következő ábra a Cornexi-Alcoa HSB Holding női kézilabdacsapatának játékosaira vonatkozó adatokat tartalmazza.

Név	**Poszt**	**Magasság/Súly**
Trimmel Brigitta	kapus	176/62
Csenus Krisztina	beálló	180/68
Kiss Olívia	beálló	165/61
Tápai Szabina	irányító	174/62
Löw Andrea	szélső	175/67
Kornyuk Jelena (RUS)	átlövő	174/70
Vaszari Virág	átlövő	176/70
Uljanics Viktória (RUS)	irányító	181/71
Kenyeres Fanni	szélső	174/73
Szabó Valéria	beálló	181/70
Sugár Tímea	kapus	171/63
Siti Beáta	irányító	175/68
Őri Edina	átlövő	171/69
Tápai Andrea	átlövő	173/62
Pádár Margit	átlövő	184/69
Vijuanite Sonata (LTU)	beálló	178/62
Fekete Csilla	kapus	185/81
Balogh Beatrix	átlövő	165/64
Tilinger Tamara	átlövő	170/63
Brigovácz Nikolett	átlövő	180/71

Forrás: http://www.nemzetisport.hu

- Az ábra (táblázat) alapján csoportosítsa (egyszerűsítse) az adatokat.

- Készítsen táblákat, amelyekben az adatok csoportosításával átlátható összefüggéseket figyelhetünk meg.

6. fejezet - Statisztikai sorok, statisztikai táblák

A statisztikai adatok valamilyen ismérv szerinti felsorolását **statisztikai sor**nak nevezzük. A statisztikai sorokat csoportosítás eredményeként vagy összehasonlítás céljából állíthatjuk elő. Az **azonos fajta adatok**ból álló statisztikai sorokat – amelyek általában csoportosító vagy összehasonlító sorok – az ismérvek típusai szerint is osztályozhatjuk. Így beszélhetünk **időbeli, minőségi, mennyiségi** és **területi** statisztikai sorokról.

A **különböző fajta,** de egymással összefüggő adatokat tartalmazó sort **leíró sor**nak nevezzük. Egy társadalmi–gazdasági egység adott vizsgálat szempontjából fontos, rendezett adatait fogja keretbe a leíró sor. Az alábbiakban a különböző típusú sorokra mutatunk be egy-egy példát.

Idősor

6.1. táblázat - Magyar sportolók száma az egyes olimpiákon

Az olipmia sorszáma	Év, helyszín	Magyar sportolók száma
I.	1896, Athén	7
II.	1900, Párizs	17
III.	1904, St. Louis	4
IV.	1908, London	63
V.	1912, Stockholm	119
VIII.	1924, Párizs	89
IX.	1928, Amszterdam	110
X.	1932, Los Angeles	54
XI.	1936, Berlin	216
XIV.	1948, London	128
XV.	1952, Helsinki	189
XVI.	1956, Melbourne	111
XVII.	1960, Róma	180
XVIII.	1964, Tokió	182
XIX.	1968, Mexikóváros	167
XX.	1972, München	232
XXI.	1976, Montreal	178
XXII.	1980, Moszkva	263
XXIV.	1988, Szöül	188
XXV.	1992, Barcelona	217
XXVI.	1996, Atlanta	214
XXVII.	2000, Sydney	178
XXVIII.	2004, Athén	219

Forrás: http://www.mob.hu

Minőségi sor

6.2. táblázat - A Dunaferr S.E. kézilabdacsapatának játékoskerete posztok szerint a 2005–2006-os bajnoki évben

Játékban betöltött szerep	Játékosok száma (fő)
kapus	3
átlövő	4
szélső	5
beálló	6
irányító	3
összesen	21

Forrás: http://www.nemzetisport.hu

Mennyiségi sor

6.3. táblázat - A góllövőlista állása (labdarúgás) 2005. december. 11-én

Gólok száma (darab)	Góllövők száma (fő)
1	68
2	28
3	11
4	11
5	7
6	3
6 és több	11
összesen	139

Forrás: http://www.nemzetisport.hu

6.4. táblázat - Szabadidejükben sportoló 15-64 éves féfiak korcsoportok szerinti aránya, Egészségi Állapot Felvétel, 2002.

Korcsoport (év)	Férfiak (%)
15–19	52,9
20–24	30,7
25–29	26,8
30–34	17,2
35–39	20,7
40–44	10,8
45–50	8,4
50–54	9,6
55–59	9,7
60–64	9,0
együttesen	21,4

Forrás: Életminőség és egészség, KSH, 2002.

Területi sor

6.5. táblázat - A sportegészségügyi ellátás fontosabb adatai terület szerint (megyénként) 2004-ben

Megye	Sportegészségügyi rendelők száma (db)
Budapest	16
Baranya	6
Bács-Kiskun	10
Békés	5
Borsod-Abaúj-Zemplén	6
Csongrád	6
Fejér	0
Győr-Moson-Sopron	5
Hajdú-Bihar	8
Heves	4
Jász-Nagykun-Szolnok	11
Komárom-Esztergom	5
Nógrád	4
Pest	21
Somogy	9
Szabolcs-Szatmár-Bereg	7
Tolna	6
Vas	6
Veszprém	8
Zala	6
Összesen	149

Forrás: Egészségügyi Statisztikai Évkönyv, 2004

Leíró sor

6.6. táblázat - A PTE-PEAC asztalitenisz-szakosztály főbb adatai, 2009

Megnevezés	A mutató értéke
Országos csapatbajnokságban szereplő csapatok száma (év végén, db)	5
Szakosztályi létesítmények száma (év végén, db)	1
Igénybevett szállítóeszközök	11

Megnevezés	A mutató értéke
(db)	
Támogatói kör nagysága (db)	9
Havi tagdíjbefizetések átlagos nagysága (Ft/fő)	4000
Sportolók összlétszáma (fő)	75
Sportolók havi átlagos keresete (Ft/ fő)	35000

Forrás: saját gyűjtés

Az idősoroknál kell megemlíteni, hogy a vizsgált sokaság természete szerint – amely lehet álló és mozgó sokaság – megkülönböztetünk **állapot-** és **tartamidősort**. Az előbbihez tartozó ismérvváltozatok ún. eszmei időpontokat jelölnek, míg a tartamidősor ismérvváltozatai az időtartamok.

A közölt minőségi, mennyiségi és területi sorok csoportosító sorok, így lehetővé teszik a sokaság belső struktúrájának vizsgálatát. A statisztikai sorokban szereplő adatok többségükben abszolút adatok voltak, amelyek mérés, számlálás útján keletkeztek, de találkoztunk származtatott adatokkal is, pl.: a sportolók havi átlagos keresete, vagy a szabadidejükben sportoló férfiak aránya.

Statisztikai táblának nevezzük a statisztikai sorok összefüggő rendszerét. A gyakorlatban a jelenségek vizsgálata gyakran megköveteli, hogy a sorokat ne egyedileg, hanem összefüggésükben vizsgáljuk. A táblának fontos formai elemei a **cím**, a **forrás**, és a **magyarázó szöveg**ek. A statisztikai tábla legalább két statisztikai sorból áll. Annak függvényében, hogy hány statisztikai sor szerepel a táblában, beszélhetünk **kétdimenziós, háromdimenziós** stb. **tábláról**.

Az alábbiakban különféle kétdimenziós statisztikai táblákat mutatunk be példaszerűen.

6.7. táblázat - A gyermek- és ifjúsági pszichiátriai gondozók adatai

Megnevezés	1994	1999
A gondozók száma (fő)	42,0	41,0
Teljesített évi szakorvosi munkaórák száma (1000 óra)	89,4	86,2
Betegforgalom (1000 fő)	156,5	158,0
Egy gondozottra jutó évi megjelenések átlaga	4,6	4,6

Forrás: Életminőség és egészség, 2002.

A 7. táblázat statisztikai tábla, amely idősorokat és leíró sorokat tartalmaz. A táblákat, melyekben csak összehasonlító, és/vagy leíró sorok szerepelnek, **egyszerű táblá**nak nevezzük.

6.8. táblázat - A 2005/2006-os női kosárlabda bajnokság csapatai

A csapatok székhely szerint	A lehetséges maximális nézőszám, (fő)	Játékosok száma, (fő)
Szekszárd	1200	15

A csapatok székhely szerint	A lehetséges maximális nézőszám, (fő)	Játékosok száma, (fő)
Diósgyőr	2500	14
Nagykanizsa	1200	11
Sopron	2600	14
Győr	2500	12
Szeged	3200	14
Szolnok	2500	11
Zalaegerszeg	4000	12
BEAC-Újbuda	500	16
BSE	600	12
Pécs	4000	12
Összesen	24800	143

Forrás: saját gyűjtés

A 8. tábla egy irányban már tartalmaz csoportosítást. Azokat a táblákat, melyekben csak az egyik ismérv szempontjából végeztünk csoportosítást, **csoportosító tábl**ának nevezzük.

6.9. táblázat - Az országos jégkorongbajnokságban szereplő sportolók megoszlása 2005/2006-os idényben (fő)

A csapat neve	Magyar állampolgárságú sportoló	Külföldi állampolgárságú sportoló	Összesen (játékoskeret)
Alba Volán Székesfehérvár	21	6	27
Dunaújváros	20	7	27
Ferencváros	22	1	23
Győr	18	2	20
Újpest	17	7	24
Miskolc	19	5	24
Összesen	117	28	145

Forrás: saját gyűjtés

A 9. táblában a minőségi statisztikai sorok csoportosító jellegűek, a tábla kombinatív csoportosítást tartalmaz. A táblákat, melyekben az adatokat legalább két ismérv alapján csoportosítjuk, **kombinációs** vagy **kontingenciatábl**ának hívjuk. A kombinációs táblák szerkesztése során körültekintően kell eljárni. A sok ismérv szempontjából kombinált csoportosítás (4-5 dimenzió) már az áttekinthetőség rovására is mehet.

1. Ellenőrző feladatok, gyakorló példák a

fejezethez

Néhány sportág területi megoszlása 2005/2006:

Régió	Kézilabda			Kosárlabda			Röplabda			Összesen		
férfi	nő	össz.	férfi	nő	össz.	férfi	nő	össz.	férfi	nő	össz.	
Közép-Magyarország	2	4	6	1	2	3	0	4	4	3	10	13
Közép-Dunántúl	3	2	5	1	0	1	2	0	2	6	2	8
Nyugat-Dunántúl	1	1	2	4	4	8	0	0	0	5	5	10
Dél-Dunántúl	1	0	1	4	2	6	2	0	2	7	2	9
Észak-Magyarország	1	0	1	0	1	1	1	1	2	2	2	4
Észak-Alföld	2	1	3	3	1	4	2	2	4	7	4	12
Dél-Alföld	2	3	5	1	1	2	1	1	2	4	5	9
Összesen	12	11	23	14	11	25	8	8	16	34	30	64

- Mi jellemzi ezt a táblatípust?
- Szerkesszen egy sporttal kapcsolatos egyszerű táblát!
- Az asztaliteniszezőket a játékstílus alapján három csoportba sorolhatjuk. Létezik a kimondottan támadó, ún. pörgető játékos, létezik a másik hibáját váró védőjátékos, és a kettő közti átmenetet képviselik a „droppoló” játékosok. Az asztaliteniszezők első osztályát vizsgálva a 2003-as létszámuk 200 fő volt. Ennek 58%-a férfi, 50%-a pörgető stílusú játékos. A férfi védőjátékosok 18%-át, míg a droppoló játékosok 10%-át adják az összlétszámnak. A női pörgető és droppoló játékosok száma megegyezik. 2004-re a pörgető játékosok száma 10%-kal nőtt, de változatlan maradt a nők és a férfiak aránya. A droppoló játékosok létszáma 20%-kal csökkent, ebből 7 fővel a férfiak száma. Az összes játékos száma nem változott, továbbra is 58%-a férfi.
- A következő adatok figyelembevételével szerkesszen statisztikai táblát! Nevezze meg a feltüntetett években – játékstílus alapján – a játékosok számát.

7. fejezet - A statisztikai adatok összehasonlítása

Az összehasonlítás mint vizsgálati módszer nemcsak a közgazdaságtan, menedzsment, business tudományok sajátja, eszközrendszerének egyik jellemző eleme, hanem fontos szerepet kap mind a műszaki, mind a biológiai, mind az orvostudományok, de a sport területén is.

A statisztikai módszerek minden eleménél, a legegyszerűbbektől a legbonyolultabb eljárásokig, találkozunk az összehasonlíthatóság biztosításának alapvető követelményével. A statisztikai módszerekkel nyert információk csak környezetükben, időbeli változásukban vagy keresztmetszeti „elhelyezkedésükben” értékelhetőek, és ezen megállapítások nem nélkülözhetik a „zavaró” hatásoktól történő elvonatkoztatást. Mindez egyszerűbben fogalmazva azt is jelenti, hogy minden összehasonlítás csak **viszonylagosan értékelhető,** jelentőségüket az támasztja alá, hogy mihez mérjük, mihez viszonyítjuk a kapott számszerű információkat. Más megfogalmazásban csak azokat az adatokat szabad egymás mellé állítani, amelyek eleget tesznek az összehasonlíthatóság követelményeinek.

Az összehasonlíthatóság statisztikai fogalmán azt értjük, hogy az egymással összemért jelenségek vagy folyamatok csak azokból az okokból következően térhetnek el egymástól, amelyeket a vizsgálat során elemezni kívánunk. Mindebből közvetlenül adódik, hogy az összehasonlíthatóság biztosítása érdekében minden olyan egyéb tényező hatását ki kell szűrni a vizsgálatból, amely zavarja az összehasonlíthatóságot.

Az, hogy a jelenségpár vagy folyamatpár összehasonlíthatóságát mely tényezők befolyásolják alapvetően, és melyek azok, amelyeknek hatását az ún. mellékhatások (zavaró hatások) közé soroljuk, nem természettől fogva adott, vagy eleve elrendelt, hanem sok esetben a vizsgálatot végző, azt előkészítő szubjektív elhatározásától függ.

A fentiekből is világosan következik, hogy az **összehasonlíthatóság** nem abszolút, hanem **relatív fogalom**. Mindez egyben azt is jelenti, hogy az összehasonlíthatóság biztosítására nem lehet általános, mindenre vonatkozóan egyértelműen érvényesíthető receptet adni. Mindaz, amire vállalkozhatunk, csupán azok a többé-kevésbé általánosítható elvek, és a hozzájuk kapcsolható módszerek, amelyeknek megfelelő alkalmazása világosabbá, érthetőbbé és reálisabbá teheti a vizsgálat egészét.

Az összehasonlítások során alapvető szempont a jelenségeket közvetlenül vagy közvetve befolyásoló tényezők felkutatása, azonosítása. Ki kell választani azokat a legfontosabb elemeket, faktorokat, amelyek mellékhatásoknak, zavaró tényezőknek tekinthetők, és ezek eliminálása után alkothatunk csak „tiszta” képet az összehasonlítandó jelenségekről, folyamatokról.

Az összehasonlíthatóság általános követelményrendszeréhez szorosan kapcsolódik a **pontosság** kérdése. Mivel egy-egy számszerű statisztikai információ csak bizonyos intervallumon belül tekinthető pontosnak, ez a kérdés az összehasonlíthatóságnál fokozottan jelentkezik. Általánosságban elmondhatjuk, hogy az aggregáltabb mutatószámok összemérése során „nagyvonalúbbak” lehetünk, mint az elemi mutatószámok esetén.

A statisztikai adatok feldolgozásának gyakorta alkalmazott elemi módszere az **összehasonlítás**, ez tulajdonképpen a statisztikai adatok egymás mellé rendelését jelenti elemzési célból. Az összehasonlítással a mindennapi életünkben gyakran találkozunk, és szinte semmilyen megállapítást nem teszünk nélküle.

Gondoljunk arra, hogy szívesen és gyakran hasonlítjuk össze más országok polgárainak életkörülményeit a miénkkel, ilyenkor pl. az összehasonlítást az egy főre jutó jövedelem

alapján végezhetjük el.

Ha tudjuk, hogy valamely országban az átlagos havi jövedelem 3000 $, ez még nem mond számunkra sokat, ha nem ismerjük a napi kiadások mértékét és az árak színvonalát stb. Amennyiben pontosabb összehasonlításra van szükségünk, minden olyan zavaró tényezőt ki kell küszöbölnünk, melyek torzíthatják az összemérést.

Gyakorta szinte „megoldhatatlan” helyzet elé állítja az elemzőt az összehasonlítás igénye. Gondoljunk például két ország labdarúgó-bajnokságának összehasonlítására. Nehéz minden „zavaró” tényező kiszűrése (pl. finanszírozás módja, átlagos nézőszám, kultúra stb.), amely lehetővé tenné egy viszonylag reális kép kialakítását. Természetesen törekedni kell a befolyásoló faktorok azonosítására, a hatások standardizálására, azonban mindezek csak korlátokkal oldhatóak meg.

Az időbeli összehasonlítások során gyakorta kell ellenőrizni azt, hogy nem történt-e módszertani változás a vizsgált mutatószámok képzésénél. Sokszor okoz nehézséget a nem egyenlő hosszúságú időtartam, valamint a külső tényezők által előidézett strukturális változás.

Az összehasonlítás kérdéskörének jelentős irodalma van, amely részletesen foglalkozik azokkal a teendőkkel, melyeket el kell végezni az összehasonlíthatóság érdekében. Általános korrekciós eljárások, elvek nem adhatók meg; a vizsgálat jellege és szempontja dönti el a konkrét lépések szükségességét. Ezért ügyelni kell az adott szakterület különbözőségeire, specifikumaira.

Az összehasonlítás műveletével valamennyi ismérv esetében találkozunk, de a statisztikai elemzésekben általában az időbeli és térbeni összehasonlítások dominálnak. Az összehasonlítás műveletének két alapvető módja az összehasonlítandó adatokból történő **hányados-,** illetve **különbségképzés**.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- A tavalyi bajnokság befejeztével a rendelkezésükre bocsájtották a Kométa-Kaposvár és a Dunaferr férfi röplabdacsapatának játékoskeretére vonatkozó főbb adatait.

Kométa-Kaposvár férfi röplabda csapatának játékoskerete (2004–2005)

Mezszám	Név	Magasság/Súly	Poszt	Válogatott mérkőzések száma (db)
1	Laszczik István	186/76	feladó	0
4	Bánhegyi István	200/94	center	15
5	Gubik János	187/81	feladó	23
7	Pásztor Attila	202/97	center	92
9	Kovács Balázs	200/93	univerzális	0
10	Németh Szabolcs	196/75	ütő	0
11	Gelencsér Balázs	204/93	center	72
14	Mészáros Péter	183/73	feladó	96
15	Dávid Zoltán	191/88	ütő	8

Mezszám	Név	Magasság/Súly	Poszt	Válogatott mérkőzések száma (db)
16	Pampuch Csaba	196/89	ütő	16
17	Mózer Zoltán	186/78	ütő	11
18	Molnár Dávid	194/79	liberó	0

Dunaferr férfi röplabda csapatának játékoskerete (2004–2005)

Mezszám	Név	Magasság/Súly	Poszt	Válogatott mérkőzések száma (db)
1	Horváth Attila	188/75	ütő	0
2	Bodor Elek	195/85	ütő	0
3	Kecskeméti Péter	198/83	feladó	0
5	Németh Gergely	190/85	ütő	0
6	Kovács Zoltán	204/93	center	0
7	Nagy József	199/75	ütő	0
8	Németh Zoltán	190/73	ütő	0
9	Tóth Gábor	186/73	feladó	0
10	Pajor Arnold	206/100	center	2
11	Forgó Nándor	196/87	univerzális	6
12	Hackenmüller Richárd	201/78	ütő	0
13	Tóth Tibor	193/85	ütő	0
14	Czintula Mihály	194/73	ütő	0
16	Lukács Attila	198/83	ütő	0
0	Szabó Gábor	192/95	liberó	28

Forrás: http://www.nemzetisport.hu

- A táblázat adatait felhasználva hasonlítsa össze a két röplabdacsapatot a játékoskeret alapján (pl. válogatott játékosok száma, legmagasabb játékos, legkönnyebb játékos, a játékosok átlagos magassága stb.)!
- Állapítsa meg, melyik csapat volt eredményesebb a bajnokságban. Indokolja meg röviden válaszát!

8. fejezet - A viszonyszámok

A korábbi példáink többségében a statisztikai adatok abszolút számértékek voltak. A statisztikai elemzőmunkában gyakorta kell élnünk a származtatott számok használatával, amelyek képesek az arányok szemléltetésére, a változások, különbözőségek relatív nagyságának kimutatására. A származtatott számok körében a viszonyszámok igen fontos helyet foglalnak el.

Viszonyszámnak nevezzük **két,** egymással kapcsolatban álló **statisztikai adat hányadosát**.

A viszonyszám általános definíciója: V = A/B

ahol:

- V – viszonyszám,
- A – viszonyított adat,
- B – viszonyítási alap.

A viszonyszám kifejezési formája többféle lehet:

- együtthatós forma,
- százalékos, ezrelékes forma,
- képzett egység.

Többféle viszonyszámot ismer a szakirodalom, mi azonban csupán a legfontosabbakat ismertetjük, nagyon vázlatosan. A viszonyszámok képzési elvének ismeretében a konkrét problémának megfelelő viszonyszám meghatározása nem jelenthet gondot.

A **dinamikus viszonyszámok** – amelyek két időszak vagy időpont adatainak hányadosai –, fontos szerepet töltenek be a statisztikai elemző munkában. A viszonyítás alapját képező időpontot, időszakot bázisidőszaknak, míg a viszonyítás tárgyát tárgyidőszaknak szokták nevezni. Amennyiben kettőnél több időszak vagy időpont adataival rendelkezünk, a viszonyítás alapja lehet állandó vagy változó; ezen utóbbi esetben mindig a megelőző időszak (időpont) adatát tekintjük viszonyítási alapnak. Az első esetben **bázisviszonyszám**ot, a második esetben **láncviszonyszám**ot számítunk.

A bázisviszonyszám[1] képlete:

$$b_i = \frac{y_i}{y_0}$$

ahol: i = 0, 1, 2, 3... n.

A láncviszonyszám képlete:

$$l_i = \frac{y_i}{y_{i-1}}$$

Az alábbi képletek segítségével a bázisviszonyszámokból osztással láncviszonyszámot számíthatunk, míg az m-edik időszak bázisviszonyszáma m darab láncviszonyszám szorzatává alakítható

$$\frac{b_m}{b_{m-1}} = l_m$$

$l_1 \times l_2 \times l_3 \times \ldots \times l_m = b_m$

A dinamikus viszonyszámokat szemléltetjük az alábbi példával.

[1] Megjegyezzük, hogy az időszak első tagját sokszor jelöljük 0 indexszel.

8.1. táblázat - Sportolók vizsgálata 2000–2004.

Év	Minősítés céljából megvizsgált sportolók száma (db)	Viszonyszámok	
Bázis, 2000 = 100%	Lánc, előző év = 100%		
2000	250 422	100,00	–
2001	252 721	100,92	100,92
2002	260 969	104,21	103,26
2003	266 926	106,59	102,28
2004	270 098	107,86	101,18

Forrás: Egészségügyi Statisztikai Évkönyv 2004, 2005.

A példa adataiból világosan látható, hogy a bázisviszonyszám a változás relatív mérésére, míg a láncviszonyszám a változás ütemének nyomonkövetésére alkalmas.

Az egyes csoportok elemszámának a teljes sokaság nagyságához viszonyított arányát **megoszlási viszonyszám**nak nevezzük[2]

$$p_j = \frac{n_j}{\sum_{j=1}^{m} n_j}$$

ahol: n_j – a j-edik csoport elemszáma, j – 1, 2, …, m a csoportok száma.

A fentiekből következik, hogy

$$0 \le p_j \le 1 \text{ és } \sum_{j=1}^{m} p_j = 1$$

A megoszlási viszonyszám a sokaság belső szerkezetének, struktúrájának kimutatására, elemzésére alkalmas származtatott számérték. Az ismérvváltozatok alapján a sokaságot részsokaságokra bontjuk, így magától értetődően a csoportok száma megegyezik az ismérvváltozatok számával.

8.2. táblázat - Egy kézilabdacsapat játékosainak szerepkör szerinti megoszlása

A játékban betöltött szerep	A jétékosok szerepkör szerinti megoszlása (%)
Kapus	14,3
Átlövő	19,0
Szélső	23,8
Beálló	28,6
Irányító	14,3
Összesen	100,0

Forrás: http://www.nemzetisport.hu

8.3. táblázat - Olimpiai sportágak doppinglistája

[2]A nevezőben lévő kifejezés a Σ (nagy szigma) az összegzés (szumma) jele. Így olvassuk: szumma j tart 1-től m-ig, n_j.

Header 1	Összes vizsgálat száma (db)	Pozitív vizsgálati eredmények (db)	A pozitív eredmények aránya (%)
Atlétika	11 266	108	0,96
Biatlon	206	1	0,49
Birkózás	845	16	1,89
Bob	214	3	1,40
Cselgáncs	1 277	6	0,47
Evezés	1 338	7	0,52
Íjászat	245	2	0,82
Kajak-kenu	1 221	3	0,25
Labdarúgás	9 936	39	0,39
Lövészet	1 457	12	0,82
Műugrás	103	2	*1,94*
Ökölvívás	1 018	8	0,79
Súlyemelés	4 164	86	*2,07*
Úszás	2 262	9	0,40

Forrás: Népszabadság, 1993. július 28.

A fenti tábla alapján egy újságíró levonta a következtetést: „A lebukottak listáját – sokak számára nem meglepő módon – a súlyemelők vezetik, … Meglepően jelentős a műugrásban vétkezők aránya…”. Amennyiben csak a származatott adatokat (megoszlási viszonyszámokat) tekintjük, igaz a megállapítás. Figyelembe kell azonban venni ilyen esetben az összehasonlítandó sokaságok (sportágak) eltérő nagyságát is; tehát abszolút értékek nélkül nem értékelhetők korrekt módon a viszonyszámok. A műugrás kis elemszáma a többi sportághoz képest tehát önmagában torzítja az összehasonlítást.

Amennyiben az egyes csoportok elemszámát nem a teljes sokasághoz, hanem valamelyik csoport nagyságához viszonyítjuk, ún. **koordinációs viszonyszám**ot számítunk. Erre általában alternatív ismérvek esetén kerül sor. Például az 1000 férfire jutó nők vagy 1000 nőre jutó férfiak száma (az 1000 férfire jutó nők száma 1998-ban 1093 fő volt) fontos mutatószámok a népességstatisztikában. Képezhetünk továbbá koordinációs viszonyszámot, ha egy sporttermékeket gyártó cégnél azt vizsgáljuk, hogy 100 fő fizikai alkalmazottra hány fő szellemi alkalmazott jut.

Az **intenzitási viszonyszám** két különböző, de egymással kapcsolatban álló statisztikai adat hányadosa. Megkülönböztetünk **nyers** és **tisztított intenzitási viszonyszám**ot. Más szempontból beszélhetünk **egyenes** és **fordított intenzitási viszonyszám**ról. Tisztított intenzitási viszonyszám esetén a viszonyítási alappal szorosabb kapcsolatban van a viszonyítandó számérték, mint a nyers viszonyszámmal. Ez jelzi természetesen a jelenség viszonylagos jellegét.

Az egyenes intenzitási viszonyszám esetén a mutatószám értékének növekedése pozitív irányú változást jelez, míg csökkenése negatív hatású!

- 2004-ben az 1000 lakosra jutó élveszületések száma 9,4 fő/1000 lakos volt. Ugyanebben az évben az 1000 15–49 éves nőre jutó élveszületések száma 38,4. Az első viszonyszám a másodikhoz képest nyers intenzitási viszonyszám.
- 2005-ben a PTE-PEAC asztaltenisz-szakosztálynál 66 játékosra 3 edző jutott, míg az egy edzőre jutó játékosok száma 22 volt. Az első egyenes, míg a második fordított intenzitási viszonyszám.

A különféle intenzitási viszonyszámok eredményesen alkalmazhatóak a sport világában. Jól kifejezi egy adott ország, térség „sportfejlettségét” például az ún. **sportsűrűségi arányszám**, amely a lakosság valamely egységére jutó igazolt, vagy aktívan sportolók számát jelzi. Ugyancsak hasznos információt nyújthat az edzők számára az **edzéshatékonyság arányszáma** (az elért eredmények és az edzésre fordított idő hányadosa). Az intenzítási viszonyszámok logikájának felhasználásával természetesen a mutatószámok köre tovább specifikálható, bővíthető.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy csapat az tavalyi évben 110 gólt lőtt és az idei évben a teljesítménye 120%-os volt az elmúlt évhez képest. Mennyivel lőtt több gólt a csapat idén?

Év	**Nézőszám (fő)**	**Viszonyszámok (%)**	
bázis	**lánc**		
2000	20 542	100	
2001		108,9	108,9
2002			113,8
2003		133,3	

- Töltse ki a hiányzó adatokat!
- Mit nevezünk megoszlási viszonyszámnak?

9. fejezet - Gyakorisági sorok

A mennyiségi ismérvek igen nagy szerepet játszanak a statisztika gyakorlati tevékenységében és módszertanában egyaránt. A mennyiségi ismérvek, változók – mint korábban szóltunk róla – folytonos és diszkrét változók lehetnek. A folytonos ismérvek bizonyos határok között bármilyen értéket felvehetnek, míg a diszkrét ismérv változatai csak meghatározott (véges számú) számértékek, elkülönített számok lehetnek. A mérési skáláknak megfelelve, alapvetően a mennyiségi ismérv intervallum- és arányskálán keletkező statisztikai adatot jelent.

A mennyiségi ismérvnek formalizált kifejezését is adhatjuk. Egy n elemű sokaság egyedei x_1, x_2, x_3, …, x_n mennyiségi ismérvértékeket vehetnek fel. A gyakorlatban a feljegyzett ismérvértékek a megfigyelések, feljegyzések sorrendjében követik egymást.

Szemléltető példánk legyen egy kondicionáló terem napi jegyeladásainak száma 30 egymás után következő munkanapon:

9.1. táblázat - Napi jegyeladások száma (db)

38	35	76	58	48	59
67	63	33	69	53	51
28	25	36	32	61	53
49	78	48	42	72	52
47	66	58	44	44	53

Forrás: saját szerkesztés

A jelenség megismerésében az első lépést megtehetjük, ha a csoportosítatlan adatokat valamilyen szempont szerint rendezzük. A legegyszerűbb rendezés egy **rangsor** készítése lehet.

Rendezzük növekvő sorrendbe a napi jegyeladásokra vonatkozó adatokat!

9.2. táblázat - A jegyvásárlások emelkedő sorrendben (fő)

25	28	32	33	35	36
38	42	44	44	47	48
48	49	51	52	53	53
53	58	58	59	61	63
66	67	69	72	76	78

Forrás: saját szerkesztés

Mind a két (9-1. és 9-2.) tábla mennyiségi sort tartalmaz. Az utóbbi közelebb visz a jelenség természetének vizsgálatához, de könnyen belátható, hogy kis elemszám esetén (példánkban az elemszám csupán 30) még elképzelhető a sorbarendezés; azonban nagyobb sokaság – és általában a statisztikában ez a gyakoribb – nem tesz lehetővé ilyen módon gyors értékelést, áttekinthetetlen. Mindez aláhúzza az információk tömörítésének szükségességét.

A mennyiségi ismérvek alapján végzett adatrendezés, adattömörítés legelterjedtebb módja a **gyakorisági sorok** képzése. Az egyes ismérvértékek többször is előfordulhatnak. Az előfordulások **gyakoriságát** a továbbiakban jelölje f_i, ahol fennáll az alábbi összefüggés:

$$\sum_{i=1}^{k} f_i = n$$

ahol: i – 1, 2, ..., k – az ismérvváltozatok száma.

Itt kell megjegyezni, hogy gyakorisági sor esetén egy-egy ismérvváltozat többször fordul elő, de k < n.

A **gyakorisági sor** a mennyiségi ismérv változatainak a gyakoriságok segítségével történő felírása. Általános sémája:

9.3. táblázat - A gyakorisági sor általános sémája

Ismérvváltozat	Gyakoriság
x_1	f_1
x_1	f_2
x_1	f_3
.	.
.	.
.	.
x_k	F_k
Összesen	N

Forrás: saját szerkesztés

Természetesen a gyakoriságokból számíthatunk megoszlási viszonyszámokat, amiket ún. **relatív gyakoriságok**nak (jele: g_i) nevezünk

$$g_i = \frac{f_i}{n}$$

A segítségükkel felírt mennyiségi sort **relatív gyakorisági sor**nak hívjuk.

Az előzőekben bemutatott jegyeladási adatokból készítsünk gyakorisági sort!

9.4. táblázat - A napi jegyeladások megoszlása

Eladott jegyek száma (db)	Napok száma
25	1
28	1
32	1
33	1
35	1
36	1
38	1
42	1
44	2
47	1
48	2
49	1
51	1
52	1
53	3

Eladott jegyek száma (db)	Napok száma
58	2
59	1
61	1
63	1
66	1
67	1
69	1
72	1
76	1
78	1
Összesen	30

Forrás: saját szerkesztés

A 16. táblázatban lévő gyakorisági sor már tömörebb formában reprezentálja az információkat, mint az ismérvértékek egyszerű, vagy rangsorba rendezett felsorolása; azonban még így sem lehetünk elégedettek, mivel a jelenség fő összefüggéseinek felismerése a túlzottan részletes sor segítségével nem könnyű feladat. Amennyiben nagyszámú ismérvértékkel rendelkezünk, célszerű olyan **osztályköz**ök kialakítása, amelyek jól tömörítik a jelenség információtartalmát, de ugyanakkor még nem eredményeznek számottevő információveszteséget. Az utóbbi feltétel az osztályközök számának szaporítását, míg az előbbi csökkentését indokolja. A két ellentmondó feltételnek kell megfelelni, amiben alapvetően a jelenséggel összefüggő szakmai információkra kell támaszkodni, de ugyanakkor orientálhatnak megelőző számítások. Az alábbi meghatározási mód csupán egy lehetséges megoldás az osztályközök számának meghatározására:[1]

$k = 1 + (3{,}3 \times \lg n)$

ahol: k – az osztályközök száma, és n – a sokaság elemszáma.

Az osztályközök számának ismeretében az egyes osztályközök hosszát az alábbi módon közelíthetjük:

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

ahol: h – az osztályköz hossza, x_{max} – a legnagyobb ismérvérték, x_{min} – a legkisebb ismérvérték.

Az osztályközök azonos nagysága olyan kívánalom, amelynek sokszor célszerűségből indokolt megfelelni. A gyakoriságok változásának értelmezése, az ábrázolás és a gyakorisági sorokkal való számolás sokkal könnyebben végrehajtható, ha azonos hosszúságúak az osztályközök. Ettől az elvtől azonban néha – praktikus okokból – el kell térni, amit a racionális tömörítés igénye indokol. (Ilyen indok lehet például, ha a sokaság elemei egy bizonyos érték körül jelentősen tömörülnek, ugyanakkor más kiugró értékek is jelentőséggel bírnak.)

Az osztályközös gyakorisági sorok készítésénél fokozott figyelmet igényel az osztályközhatárok megállapítása is. Általános alapelvként fogalmazhatjuk meg, hogy a határok mindenkor tegyék lehetővé az egyértelmű besorolást. Kifejezésre kell juttatni, hogy egy adott határérték mely osztályközbe tartozik. Ez különösen a folytonos ismérvértékek használata esetén okozhat gondot, itt fokozott óvatossággal kell eljárni (pl. csoportosíthatjuk az embereket a naponta sportolással eltöltött idő alapján, így a 15–20 perc, illetve 20–25 perc két minőségileg eltérő osztályközt jelölhet, azonban egy

[1] Meg kell jegyezni, hogy itt a 10-es alapú logaritmussal számolunk.

személy besorolásánál, aki pontosan 20 percet tölt a sportolással, már gondban lehetünk. Ilyen esetben követhető eljárás, ha az osztályközhatárokat a megfigyeltnél nagyobb pontossággal adjuk meg: 15,0–20,0 és 20,1–25,0.[2]) Sokat segíthet a folytonos ismérvértékek kerekítése is, amely végül egyértelművé teheti a besorolást.

Folytassuk a korábbi példánkat és készítsünk egy 20 eladott jegy hosszúságú osztályközöket tartalmazó gyakorisági sort!

9.5. táblázat - A napi jegyeladások megoszlása

Eladott jegyek száma (db)	Napok száma
–40	7
41–60	15
61–	8
Összesen:	30

Forrás: saját szerkesztés

A 17. táblázatban a gyakorisági sor túlzottan tömör, kissé összemossa a napi jegyeladások eloszlását, hiszen az eredeti adatokból láthatjuk, hogy az egyes számértékek nem egyenletesen helyezkednek el egy-egy intervallumon belül. Használjuk fel az osztályközök meghatározására megismert eljárást.

$k = 1 + (3{,}3 \times \lg(30)) = 5{,}87 \approx 6$

$h = \frac{78-25}{6} = 8{,}8 \approx 10$

Ezek alapján a gyakorisági sor:

9.6. táblázat - A napi jegyeladások megoszlása

Napi jegyeleladások (db)	Napok száma
–30	2
31–40	5
41–50	7
51–60	8
61–70	5
71–	3
Összesen:	30

Forrás: saját szerkesztés

A 17. és 18. táblázatban leírt gyakorisági sorban az alsó és a felső intervallum ún. **nyitott intervallum**. Ezt a megoldást általában akkor szokták használni, ha az adatok között extrém, kiugró szélsőértékek szerepelnek. Példánkban ez a tény nem áll fenn, azonban egy megismételt vizsgálat esetén – ahol nem kizárt szélsőségesebb értékek keletkezése – a fenti csoportosítás jó lehetőséget ad az összehasonlításra. A további számítások, elemzések során ezeket a nyitott intervallumokat úgy kezeljük, mintha zártak lennének, az első intervallumot ugyanolyan hosszúságúnak tételezzük fel, mint az azt követőt; az utolsót pedig olyan hosszúnak, mint az azt megelőzőt.

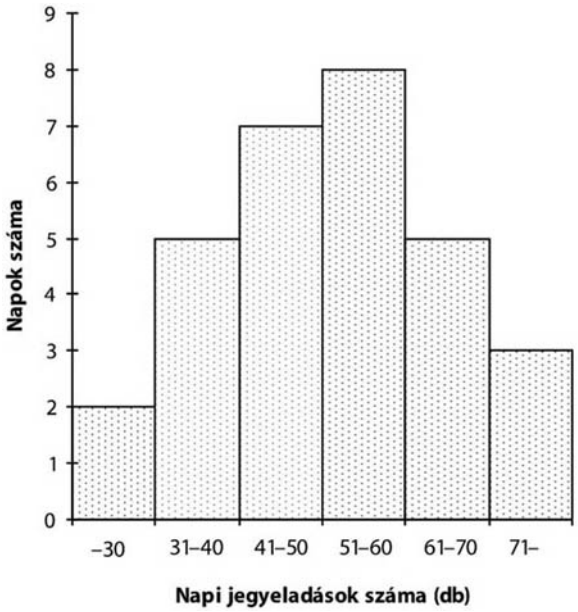
A gyakorisági sorok szemléltetéséhez háromféle **grafikus ábrát** használhatunk.

Hisztogramnak hívjuk azt a grafikus ábrát, amely a derékszögű koordináta rendszerben hézag nélküli oszlopdiagram segítségével szemlélteti a gyakorisági

[2]Találkozhatunk olyan megoldással is, amikor az osztályköz felső határa a „technikai” szám, pl. 15,0–19,9.

sorokat. A napi jegyeladás hisztogramja az alábbi:

9.1. ábra - A napi jegyeladások hisztogramja



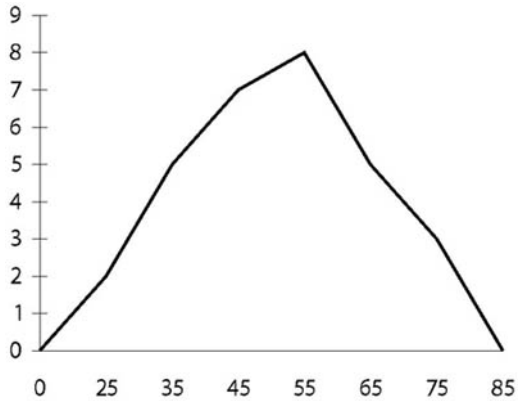
Forrás: saját szerkesztés

A hisztogram oszlopainak területe arányos a gyakoriságokkal. Egyenlő hosszúságú osztályközök esetén az ábrázolás nem okoz gondot, mert csupán az oszlopok magasságára kell figyelni (példánkban ez reprezentálja az adott osztályközbe tartozó jegyeladások számát).

Eltérő hosszúságú osztályközök esetén azonban a gyakoriságok függvényében történő automatikus ábrázolás torzítana – a hosszabb osztályköz nagyobb súlyt kapna –, ezért módosítani kell az ábrázolás adatait. Mivel az oszlopok alapjának megváltoztatására nincs mód, ezért a magasságot kell átszámítani. A megoldás a gyakoriságok (magasságok) korrigálása. Megállapítjuk, hogy a legrövidebb osztályköz hányszorosai a hosszabb osztályközök, és az így kapott számértékekkel elosztva a gyakoriságokat, nyerjük azokat a korrigált értékeket, amelyek alkalmasak az oszlopok magasságának ábrázolására.

A folytonos ismérvértékek alapján készült gyakorisági sort vonaldiagrammal is lehet ábrázolni, amit **gyakorisági poligon**nak nevezünk. Természetesen a gyakorisági poligon minden olyan esetben elkészíthető, amikor osztályközös gyakorisági sorral dolgozunk. A gyakorisági poligon felrajzolása során az osztályközepeknél felmért gyakoriságok pontjait (ezek a hisztogramok oszlopközepének felelnek meg) összekötjük. A szemléltető példánk adataiból készült gyakorisági poligon az alábbi:

9.2. ábra - A napi jegyeladások hisztogramja



Forrás: saját szerkesztés

Igen nagy elemszámú sokaság nagyszámú osztályközét tekintve a hisztogram és a poligon tovább „finomítható”, az így készült ábrát **gyakorisági görbének** nevezzük. A gyakorisági görbe a – matematikai statisztikából ismert –, folytonos valószínűségi változó sűrűségfüggvényének empirikus megfelelője.

Korábban már szóltunk arról, hogy nemcsak a gyakorisági sor, hanem a megoszlási viszonyszámok segítségével **relatív gyakorisági sor** is készíthető. Bővíti a választékot, ha arra gondolunk, hogy mind a gyakorisági sor, mind a relatív gyakorisági sor értékei halmozottan összegezhetők. Az így képzett ún. **kumulált gyakoriságok** (ezeket f_i'-vel jelöljük) azt mutatják meg, hogy az adott osztályköz felső határának megfelelő vagy annál kisebb ismérvérték hányszor fordul elő, vagyis hány esetben teljesül az $x \leq x_f$ egyenlőtlenség. Az így készült sort **alulról kumulált sor**nak nevezzük. Természetesen a fordított esetnek, a **felülről történő kumulálás**nak is van létjogosultsága.

Előző példánknál maradva néhány gyakorisági sort tartalmaz a következő tábla.

9.7. táblázat - A napi jegyeladásra vonatkozó adatok

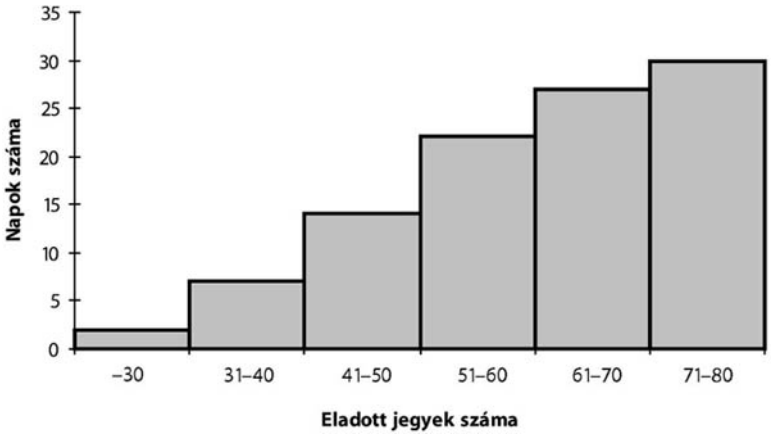
Eladott jegyek száma (db)	Napok száma		Relatív gyakoriság (%)	
–30				
31–40				
41–50				
51–60				
61–70				
71–				
Összesen:				

Forrás: saját szerkesztés

Az adatok világosan mutatják, hogy pl. 50, vagy annál kevesebb jegyértékesítés a vizsgált időszakban 14 napon volt, ez az összes megfigyelt nap 46,7%-a.

A kumulált gyakorisági, illetve relatív gyakorisági értékeket is ábrázolhatjuk hisztogram segítségével. Az így képzett lépcsőzetes ábra a folytonos valószínűségi változó eloszlásfüggvényének empirikus megfelelője.

9.3. ábra - Kumulált gyakorisági sor grafikus megfelelője



Forrás: saját szerkesztés

A mennyiségi ismérvek segítségével előállítható további mennyiségisor-típus az ún. **értékösszegsor**.

Az értékösszegsor tartalmazza a mennyiségi ismérv változatait, számértékek vagy osztályközök formájában (x_i), valamint a hozzájuk rendelhető értékek összegeit (s_i). Az osztályközös gyakorisági sorok esetén – különösen folytonos ismérvértékeknél – a fenti értékösszegeket legtöbbször nem ismerjük. Ilyen esetekben az értékösszegeket becsüljük a gyakoriságok és az osztályközepek szorzataként (f_ix_i). Itt jegyezzük meg, hogy az osztályközepek – amelyek fontos szerepet töltenek be a további számításokban is –, az intervallumok alsó és felső határainak átlagolásával képezhetők (kiszámításuk során már nem vesszük figyelembe a csak az egyértelmű besorolás érdekében megkülönböztetett, technikai alsó- és felső osztályközhatárokat, pl. 11–20 intervallum esetén 10–20 értékekkel számolunk).

Példánkat folytatva elkészítettük az alábbi statisztikai táblát, amely az értékösszegsorokat tartalmazza.

9.8. táblázat - Értékösszegsor

Eladott jegyek száma (db)	**Osztályközép**	**Napok száma**	**Összes jegyeladás (db)**	
x_i	**f_i**	**Tényleges s_i**	**Becsült F_ix_i**	
–30	25	2	53	50
31–40	35	5	174	175
41–50	45	7	322	315
51–60	55	8	437	440
61–70	65	5	326	325
71–	75	3	226	225
Összesen:	–	30	1 538	1 530

Forrás: saját szerkesztés

Láthatjuk, hogy a becsült és a tényleges értékösszegek eltérnek egymástól, mert a becslés során valamennyi esetben feltételeztük, hogy minden osztályközben csak azonos értékek fordulnak elő, az osztályközepek. A gyakorlatban a legtöbb esetben nem rendelkezünk a tényleges értékösszeg adataival, csupán a becslések állnak rendelkezésünkre. A becslés az osztályozás minőségétől erőteljesen függ. Példánkban elfogadhatónak tekinthetjük a becsléseket, mivel itt tudjuk, hogy ténylegesen 1538 darab jegyet értékesítettek a vizsgált időszakban és a közelítő számítással ettől

szignifikánsan nem eltérő, 1530 darab eladott jegyet állapítottunk meg.

Természetesen az értékösszegsorok esetében is van létjogosultsága a relatív értékösszegek (megoszlási viszonyszámok) kiszámításának, valamint a kumulált tényleges és relatív érték-összegsorok meghatározásának.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- A következő tábla egy sportszergyár alkalmazottainak havi átlagos bérek alapján történő megoszlását szemlélteti egy adott évben.

Átlagos havi bér (Ft)	**Alkalmazottak**	
Létszám (fő)	**Megoszlása (%)**	
–50 000	12	8,0
50 001–60 000	27	18,0
60 001–70 000	33	22,0
70 001–80 000	42	28,0
80 001–90 000	21	14,0
90 001–100 000	9	6,0
100 001–	6	4,0
Összesen:	150	100,0

- Készítsen kumulált sorokat!
- Állapítsa meg, hogy hány fő, illetve az alkalmazottak hány százaléka keres 60 000 Ft-nál többet?
- Mondja meg, hogy ezeknek a foglalkoztatottaknak mennyi a részesedésük a cég béralapjából!

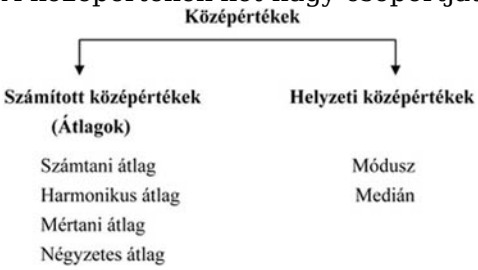
10. fejezet - Középértékek

A mennyiségi ismérvértékek, számadatok a sokaság lényeges tulajdonságainak hordozói, önmagukban is igen hasznos tájékoztatást adnak a vizsgált jelenségről, folyamatról. Az adatok rendezése, csoportosítása nagymértékben megkönnyíti a megértést. A számszerű információk azt is lehetővé teszik, hogy segítségükkel általános, tömör jellemzést adjunk, egyetlen számadatba sűrítsük fontos jellemzőjüket. Ezt a számadatot **középérték**nek hívjuk. A középérték **az azonos fajta adatok tömegének számszerű jellemzője.**

A kosárlabdában egy játékos leigazolásakor a teljesítményéről információt szolgáltathat az előző évi bajnokságban szerzett pontjainak az átlaga. Lehetséges, hogy a játékos sohasem szerzett annyi pontot, mint az éves átlaga, tehát meccsenkénti teljesítménye nem azonos az átlagos értékkel. Tudjuk, hogy a játékos a mérkőzéseinek egy részén az átlagnál több, míg a mérkőzések másik részén az átlagnál kevesebb pontot dobott. Ugyancsak hasonlóan hasznos információ a labdarúgó-bajnokságokban a meccsenként elért gólok átlagos száma is.

Az átlagos értékek vonzereje tömörségükből, általános voltukból fakad, dimenziójuk megegyezik az általuk jellemzett ismérv dimenziójával.

A középértékek két nagy csoportját szokás megkülönböztetni:



A középértékekkel szemben, általában az alábbi követelményeket lehet támasztani:

1. Közepes helyet foglaljanak el.

2. A számszerű értékek halmazának legyenek tipikus értékei.

3. Jól kezelhető matematikai formulával legyenek meghatározhatók.

4. Legyenek jól értelmezhetők.

5. Ne legyenek érzékenyek a kiugró értékekre.

A fentiekben megfogalmazott feltételeknek a középértékek nem egyformán felelnek meg. Különösen áll ez az első két feltételre. A közepes értéket elsősorban az átlagok veszik fel, ezek algebrai kapcsolatban vannak az összes előforduló számértékkel. A helyzeti középértékek viszont tipikus számértékek, melyek nem függnek az összes számértéktől, nagyságukat az értékek elhelyezkedésének rendje határozza meg. A vizsgált jelenségek természete alapvetően megszabja az alkalmazható középértékek típusát, a gyakorlatban előforduló leggyakrabban használt középértékfajták: a számtani átlag, a módusz és a medián.

1. Számtani átlag

A **számtani átlag** az a szám, amelyet az átlagolandó értékek helyébe téve azok összege azonos marad.

A definícióból közvetlenül adódik, hogy a számtani átlag (x) a *megfigyelt értékek* (x_1, x_2,

x_3, …, x_n) *összegének, és az elemek számának (n) hányadosa:*[1]

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A számtani átlag képlete igazolja a definíciót: az átlagolandó értékek helyébe téve a számtani átlagot az értékek összege nem változik, mivel

$$n\bar{x} = \sum x$$

. A számtani átlag közepes értéket vesz fel, szemléletes és jól értelmezhető. A számtani átlag definíciójából belátható az a fontos matematikai tulajdonsága is, amely szerint a számtani átlagtól mért eltérések algebrai összege zéró:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

A gyakorlatban sokszor kell **gyakorisági sorból számtani átlag**ot számítani. Ilyenkor az ismérvértékek többször fordulnak elő, amit a gyakoriságok jeleznek. A számtani átlag számításához itt fel kell használni ezeket a gyakoriságokat, amit a **súlyozott számtani átlag** formulájával érünk el.

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

Könnyen belátható, hogy a súlyozott számtani átlag nagyságát nem befolyásolja, ha a súlyadatokat egy konstans számmal osztjuk vagy szorozzuk. Ezért a relatív gyakoriságok *(g_i)* is használhatók a számtani átlag kiszámítása során.

$$\bar{x} = \sum_{i=1}^{k} g_i x_i \text{, mivel} \sum_{i=1}^{k} g_i = 1$$

A fenti képlet segít megérteni a súlyszámoknak az átlagszámításban betöltött szerepét. A súlyozott **számtani átlag nagyságát két tényező határozza meg**:

1. az átlagolandó értékek (abszolút) nagysága,

2. a súlyok viszonylagos nagysága, más szóval a súlyarányok.

A korábban már bemutatott példánk segítségével illusztráljuk a számtani átlag kiszámításának menetét!

10.1. táblázat - Munkatábla a számtani átlag kiszámításához

A naponta eladott jegyek száma (db)	Osztályközép	Napok száma	Értékösszegek, összes értékesített jegy (db)	
x_i	f_i	becsült ($f_i x_i$)	tényleges (s_i)	
–30	25	2	50	53
31–40	35	5	175	174
41–50	45	7	315	322
51–60	55	8	440	437
61–70	65	5	325	326
71–	75	3	225	226
Összesen:	–	30	1 530	1 538

Forrás: saját szerkesztés

[1] Megjegyezzük, hogy a továbbiakban, ahol a megértést ez nem nehezíti, a Σ jelhez tartozó határokat elhagyjuk.

$$\bar{x} = \frac{\sum_{i=1}^{6} f_i x_i}{n} = \frac{1.530}{30} = 51 \text{ jegy/nap,}$$
$$\text{illetve } \bar{x} = \frac{1.538}{30} = 51{,}27 \text{ jegy/nap}$$

A kétféle számítás között az eltérés abból ered, hogy az első esetben a becsült osztályközepeket használtuk fel, a különbség azonban nem számottevő.

A számtani átlag felhasználását elemzéseinkben általában az indokolja, hogy az átlagolandó értékek összegének van tárgyi értelme. Például a fenti példánkban az értékesített összes jegynek, továbbá egy adott szakosztálynál kifizetett összes bérnek, vagy a lőtt gólok átlagának kiszámításánál az összes gólnak van tárgyi tartalma. Sok példát hozhatunk arra is, amikor az értékösszegek nem rendelkeznek tartalommal. Elég, ha például a vizsgaeredmények összegére, vagy az életkorok összegére gondolunk. Ilyen esetben is számtani átlagot használunk a számadatok tömör jellemzésére, amit az magyaráz, hogy számítása egyszerű, ugyanakkor más átlag használata sok esetben félreérthető lehet.

2. Harmonikus átlag

A **harmonikus átlag** *az a szám, amelyet az átlagolandó értékek helyébe téve, azok reciprokainak összege változatlan marad*.

Ez a felismerés tömören:
$$\sum_{i=1}^{n} \frac{1}{x_i} = n \frac{1}{\bar{x}_h}$$

Amiből *a harmonikus átlag* kiszámításának *formulája*:
$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

A harmonikus átlagnak is létezik súlyozott formája:
$$\bar{x}_h = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} f_i \frac{1}{x_i}}$$

A harmonikus átlagot elsősorban olyan esetekben alkalmazzuk, ha értelmezhető az átlagolandó értékek reciprokainak összege. Használata főleg akkor indokolt, ha az átlagolandó értékek fordított intenzitási viszonyszámok.

Tegyük fel, hogy egy teniszlabdát gyártó vállalatnál ugyanazt a labdát három gépen állítják elő. A teniszlabda elkészítésének időtartama „A” gépen 2,5 perc/darab, „B” gépen 4 perc/darab, „C” gépen 10 perc/darab. A munkaszervezés során kíváncsiak lehetünk az átlagos megmunkálási időre. Amennyiben az egyszerű számtani átlagszámítás szabályai szerint járunk el, akkor a keresett érték 5,5 perc/darab. Ennél a számításnál azonban azonos súlyt adtunk valamennyi átlagolandó értéknek, vagyis feltételeztük, hogy valamennyi gép egy bizonyos idő alatt ugyanannyi darab labdát készít el. Ez ellentmond a vizsgált jelenség természetének. Indokolatlanul „kedvezőtlenebb” elbírálásban részesül a gyorsabban dolgozó, azaz fajlagosan kevesebb időt felhasználó gép, ugyanakkor az átlagolandó értékek összegének sincs tárgyi értelme. Az azonos súlyszámok használatának gondolatát vonatkoztathatjuk az átlagolandó értékek reciprokaira. Az első gép egy labda gyártását 2,5 perc alatt végzi el, tehát a munkafolyamat 1/2,5 azaz négy tized részével végez 1 perc alatt; ugyanígy a másik két gépre is megállapíthatjuk a reciprokértékeket. Az együttes teljesítmény így:
$$\frac{1}{2{,}5} + \frac{1}{4} + \frac{1}{10} = \frac{7{,}5}{10}$$

Tehát a három gép együttesen egy labda gyártásának 3/4 részét végzi el percenként.

Ebből egy gép átlagos teljesítménye:

$$\frac{\frac{7,5}{10}}{3} = \frac{7,5}{30} \text{ azaz } \frac{1}{4} \text{ darab/perc.}$$

Eredeti kérdésünk azonban az egy termékegységre jutó munkaidő-felhasználás átlagos értékére vonatkozott, ami a fenti érték reciproka: 4 perc/darab. A számítás menetét végiggondolva láthatjuk, hogy ezt a két lépésben végrehajtott számítást egy lépésben is elvégezhettük volna a harmonikus átlag számítási algoritmusának felhasználásával:

$$\overline{x}_h = \frac{3}{\frac{1}{2,5}+\frac{1}{4}+\frac{1}{10}} = 4 \text{ perc / darab}$$

A statisztikai gyakorlat az egyszerű harmonikus átlagot ritkán használja. A súlyozott harmonikus átlag alkalmazására alapvetően gyakorlati, praktikus okokból kerül sor. Például gyakorisági sor adataiból harmonikus átlagot általában akkor számítunk, ha az átlagolandó értékeket az értékösszegekkel súlyozzuk, mivel közvetlenül ezek állnak rendelkezésünkre. Ugyancsak ennek az átlagnak a segítségével kapunk gyors eredményt, ha intenzitási viszonyszámokat azok mértékegységének, dimenziójának számlálójával súlyozzuk.

Az utóbbi illusztrálására nézzük az alábbi példát! Egy valutaváltó egy adott héten az alábbi forgalmat érte el euró (EUR) váltása kapcsán:

10.2. táblázat - Az euró (€) -forgalom adatai egy adott héten

Ügylet	Árfolyam (Ft/€)	Forgalom (Ft)
Vétel	243	400 950
Eladás	257	308 400

Forrás. Saját szerkesztés

Határozzuk meg a középárfolyamot!

Az árfolyam – mint viszonyszám – dimenziójának számlálójára vonatkoznak közvetlenül a forgalmi adatok mint súlyszámok. Az átlagolandó értékek és a súlyszámok közötti viszony nem teszi lehetővé közvetlenül a számtani átlag használatát, viszont a harmonikus átlag jól alkalmazható.

$$\frac{400.950+308.400}{\frac{400.950}{243}+\frac{308.400}{257}} = \frac{709.350}{1650+1200} = 248,89 \text{ Ft/€}$$

Természetesen az így kiszámított átlagot számtani átlagként kell értelmezni, hiszen – a súlyszámok jellege miatt – a számítás módszerében tértünk csak el a megszokott számtani átlagtól.

3. Mértani átlag

A **mértani** (geometriai) **átlag** *az a szám, amelyet az átlagolandó számértékek helyébe téve azok szorzata változatlan marad.*[2]

$$\overline{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

A fenti definícióból és képletből következik, hogy a mértani átlagot elsősorban akkor használjuk, ha az átlagolandó értékek között szorzatszerű viszony van.

Természetesen a mértani átlagnak is felírható a súlyozott átlag formája:

[2] A Π (nagy pi) a produktum, az adatok szorzását jelöli.

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

ahol: $n = \sum f_i$

A Komló NB I-es kézilabdacsapatának nézőszáma 2002-ről 2003-ra 2%-kal, 2003-ról 2004-re 3,9%-kal, míg 2004-ről 2005-re 8,5%-kal nőtt. Határozzuk meg az éves átlagos nézőszám-növekedés mértékét! Itt a mértani átlagot használhatjuk, mivel a láncviszonyszámok közötti szorzatszerű viszony értelmezhető.

$$\sqrt[3]{1,02 \times 1,039 \times 1,085} = \sqrt[3]{1,15} = 1,048$$

Tehát 2002 és 2005 között a nézőszám évről évre átlagosan 4,8%-kal nőtt.

4. Négyzetes átlag

A **négyzetes** (kvadratikus) **átlag** *az a szám, amelyet az átlagolandó számértékek helyébe téve azok négyzetösszege változatlan marad*. A négyzetes átlagot meghatározhatjuk, ha az átlagolandó értékek négyzeteinek számtani átlagából négyzetgyököt vonunk. Ennek az átlag-típusnak általában csak technikai értelmet tulajdonítunk. A négyzetes átlag közvetlen alkalmazására csak korlátozottan nyílik lehetőség, azonban a későbbi fejezetben megismerendő szórás mutatószámának kiszámítása során (amikor eltérő előjelű értékeket átlagolunk) kiemelt szerephez jut.

5. Medián

A **medián** a szó legszorosabb értelmében közepes érték, *a mennyiségi ismérvnek azon értéke, amelynél ugyanannyi kisebb, mint amennyi nagyobb érték fordul elő.*

A medián meghatározásához az összes megfigyelt értéket figyelembe vesszük, de nagyságát a szélső értékek nem befolyásolják. Ha az összes xi értéket ismerjük, akkor a medián meghatározásának első lépéseként a **számértékeket rangsorba rendezzük** és

- ha n páratlan, akkor az (n + 1)/2. sorszámú egyed ismérvváltozatának értéke lesz a medián;
- ha n páros, akkor az n/2. és (n/2) + 1. egyed ismérvváltozatainak egyszerű számtani átlaga lesz a medián.

Kissé összetettebb a medián meghatározása, ha a számértékek gyakorisági sorba vannak rendezve. Ilyen esetben a kumulált gyakorisági sor segíti a felhasználót.

Diszkrét mennyiségi ismérv esetén a medián értéke megegyezik azzal az értékkel, amelyhez tartozó kumulált gyakoriság tartalmazza a medián sorszámát.

A mediánt osztályközös gyakorisági sorból mindössze becsülni tudjuk. A medián osztályközös gyakorisági sorból történő meghatározásának széleskörűen elterjedt számítási módját írja le az alábbi képlet:

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

ahol: $x_{me,\,a}$ – mediánt magába foglaló osztályközalsó (nem technikai) határa, s – n/2 – a medián sorszáma, f'_{me-1} – a mediánt megelőző osztályköz kumulált gyakorisága, f_{me} – a mediánt tartalmazó osztályköz gyakorisága, h – a mediánt tartalmazó osztályköz hossza.

A fenti közelítő számítás tulajdonképpen a mediánt tartalmazó osztályköz arányos osztását jelenti. A becslés során feltételezzük, hogy az osztályközben az értékek egyenletesen helyezkednek el.

A medián közelítő meghatározását az alábbi példa segítségével illusztráljuk. Egy sportágban 50 igazolt versenyző életkorát jellemzik az alábbi adatok:

10.3. táblázat - A versenyzők életkorának megoszlása

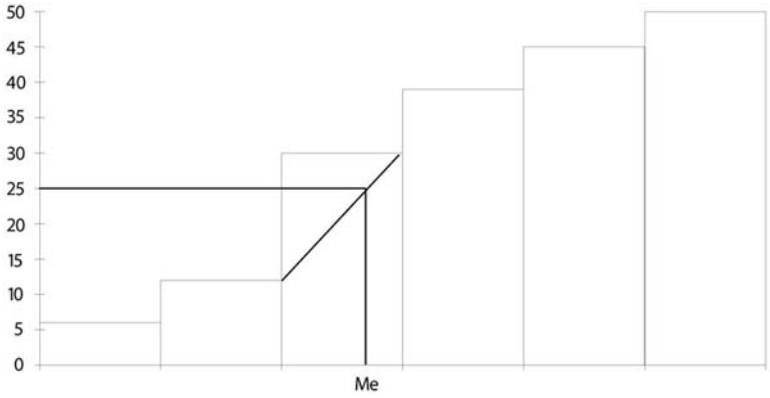
Életkor (év)	Versenyzők száma, fő (f_i)	Kumulált gyakoriság, fő (f_i')
–20	6	6
21–25	7	13
26–30	18	31
31–35	11	42
36–	8	50
Összesen:	50	–

Forrás: Saját szerkesztés

$$Me = 25 + \frac{25-13}{18} \times 5 = 28{,}33 \text{ év.}$$

A medián becslését jól szemléltethetjük az 5. ábra segítségével:

10.1. ábra - A medián meghatározását szemléltető grafikus ábrázolás



Forrás: Saját szerkesztés

A korábban bemutatott példánkban – ahol az eladott jegyek megoszlását vizsgáltuk – szintén meghatározható a medián. Itt azzal az előnnyel is rendelkezünk, hogy az egyes megfigyelések külön-külön is ismertek. Ezért mind a becsült, mind a tényleges medián értéke kiszámítható.[3] A tényleges érték 51,5 db, azaz a megfigyelt napok felében 51,5 darabnál kevesebb, illetve felében 51,5 darabnál több jegyet értékesítettek. Az osztályközös gyakorisági sorból becsült medián 51,25 db, ami nem túl jelentős eltérést mutat a tényleges értékhez képest.

6. Módusz

A **módusz** *az ismérvértékek tipikus, leginkább jellemző értékét jelöli.* Diszkrét értékekkel rendelkező mennyiségi ismérv módusza a sokaságban leggyakrabban előforduló ismérvérték.

Gyakorisági sorban – diszkrét ismérvértékek esetén – a módusz meghatározása nem jár különös nehézséggel, csak le kell olvasni a leggyakrabban előforduló ismérvértéket. Ez a számérték a jegyeladásokat taglaló példánkban 53 darab, mivel ez az érték fordult elő legtöbbször, háromszor.

Folytonos mennyiségi ismérv módusza az az érték, amely körül az előforduló **értékek**

[3] Mivel a sokaság elemszáma 30, ezért példánkban a medián sorszáma (30 + 1)/2, azaz 15,5. A 15. megfigyelés 51 db, a 16. elem 52 db, így a belőlük képzett egyszerű számtani átlag értéke: 51,5 db.

legjobban **sűrűsödnek**, ahol a **gyakorisági görbe maximuma** van. Ez a definíció egyben azt is jelenti, hogy a módusz egzakt meghatározására osztályközös gyakorisági sor esetén nincs mód, értékét itt is csak közelítő számítással becsülni tudjuk.

Osztályközös gyakorisági sor esetében a módusz közelítő meghatározása az ún. **modális osztályköz** meghatározásával kezdődik. A modális osztályköz – amely a móduszt magában foglalja – a legnagyobb gyakorisággal rendelkező osztályköz. Természetesen csak egyenlő hosszúságú osztályközök esetén tudjuk a modális osztályközt rögtön meghatározni. Ha az osztályközök nem egyenlőek, akkor korrekciót kell végezni, át kell számolni a gyakoriságokat azonos hosszúságú osztályközre. A modális osztályköz kijelöléséhez, valamint a további számításokhoz ezeket a korrigált gyakoriságokat használjuk.

A modális osztályköz kijelölése után azt az értéket kell meghatározni az intervallumon belül, amely az értékek sűrűsödési helyének tekinthető. Ehhez segítséget adnak a modális osztályközzel szomszédos osztályközök gyakoriságai. Feltehetjük ugyanis, hogy a sűrűsödési hely a modális osztályköznek ahhoz a határához esik közelebb, amelyik irányban a szomszédos gyakoriságok nagyobbak. Mindezek figyelembevételével a móduszt megbecsülhetjük, ha a modális osztályköz hosszát a gyakoriságok különbségei alapján, arányos osztással felosztjuk, és az így kapott értéket a modális osztályköz alsó határához hozzáadjuk. A becslés az alábbi képlet segítségével végezhető el:

$$Mo = x_{mo,a} + \frac{k_1}{k_1 + k_2} \times h$$

Ahol: $x_{mo,a}$ – a modális osztályköz alsó határa, k_1 – a modális osztályköz és a megelőző osztályköz gyakoriságának különbsége, k_2 – a modális osztályköz és az azt követő osztályköz gyakoriságának különbsége, h – a modális osztályköz hossza.

A sportolók életkorát elemző példánkban, ahol azonosak az osztályközök, a legfontosabb adatok:

modális osztályköz: 25– 30 év,

osztályköz hossza: 5 év,

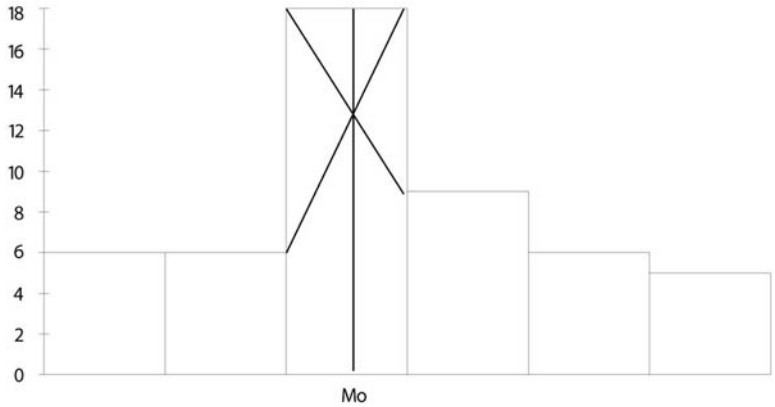
k_1 = 18 – 7, és k_2 = 18 – 11,

$$Mo = 25 + \frac{11}{11+7} \times 5 = 28{,}08\ év.$$

A tipikus életkor tehát 28,08 év.

A módusz becslésének gondolatmenetét szemlélteti a 6. ábra.

10.2. ábra - A módusz meghatározása



Forrás: Saját szerkesztés

Természetesen a módusz osztályközös gyakorisági sor esetén más módszerrel is

becsülhető. Ilyen például a parabolaillesztés módszere.

Szólni kell végezetül arról az esetről is, amikor a gyakorisági görbének több helyi maximuma van. Ilyen esetben a legnagyobb gyakorisággal rendelkező értéket **főmódusz**nak, és az ilyen sokaságot többmóduszú sokaságnak nevezzük.

7. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy magasugróversenyen 20 érvényes ugrást regisztráltak. Az eredmények a következőek voltak (cm).
  - Számítsa ki az átlagos magasságot és a mediánt a versenyen elért eredmények alapján!

165	171	168	183
168	160	171	171
171	165	168	176
166	176	176	165
178	181	165	181

- Egy szakosztály alkalmazottainak keresetük szerinti megoszlását szemlélteti a következő tábla.
  - Számítsa ki a szakosztály alkalmazottainak átlagkeresetét!
  - Jellemezze keresetüket a helyzeti középértékek segítségével (medián, módusz)!

Kereset (Ft/fő)	**Alkalmazottak létszáma (fő)**
–50 000	2
50 001–60 000	6
60 001–70 000	7
70 001–80 000	21
80 001–90 000	19
90 001–100 000	14
100 001–	6
Összesen:	75

11. fejezet - Kvantilisek

A kvantilis értékek a mennyiségi ismérv értékeinek rendezésére szolgálnak, néhány számszerű információ alapján segítik az eligazodást. A kvantilisek nem tartoznak szorosan a középértékekhez, azonban az **egyik nevezetes kvantilis érték, a medián** magyarázata lehet annak, hogy miért itt tárgyaljuk röviden ezeket a mutatószámokat.

Ha a **rangsorba** rendezett sokaságot 2, 3, 4, ..., k **egyenlő részre osztjuk**, az osztópontoknak megfelelő ismérvértékeket **kvantiliseknek** hívjuk. Másképpen a kvantilis értékek azok az ismérvértékek, amelyeknél az összes előforduló érték $\frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k}$, röviden $\frac{j}{k}(j = 1,2,\ldots,k-1)$-ad

része kisebb, illetve 1 – (j/k)-ad része nagyobb. Néhány fontosabb kvantilis értéknek sajátos elnevezése is van:

11.1. táblázat - Néhány nevezetes kvantilis

K	Elnevezés	Jele
2	Medián	M_e
3	Tercilis	T_j
4	Kvartilis	Q_j
5	Kvintilis	K_j
10	Decilis	D_j
100	Percentilis	P_j

Forrás: Saját szerkesztés

Például a 40. percentilis (P_{40}) a mennyiségi ismérvnek az az értéke, amelyiknél az összes érték 40%-a kisebb, és 60%-a nagyobb.

A kvantilis értékek közül gyakran találkozunk a kvartilis értékekkel (Q_j). A kvartilisek használata során általában a **felső kvartilis**re (Q_3), illetve **alsó kvartilis**re (Q_1) gondolunk, mivel Q_2 = Me. Ezek szerint a kvartiliseket a mediánnál kisebb, illetve nagyobb értékek mediánjaiként is felfoghatjuk. Az alsó kvartilisnél az előforduló ismérvértékek egynegyede kisebb, háromnegyede nagyobb értéket; a felső kvartilisnél az előforduló ismérvértékek háromnegyede kisebb, egynegyede nagyobb értéket vesz fel.

Értékes következtetéseket lehet levonni a népesség, illetve a háztartások jövedelmi tizedek, decilisek szerinti csoportosításából.

A sport világában – különösen az edzéstervek készítése, a versenyre való felkészülés során – hasznos információkkal szolgálhatnak a kvantilisek. Gondoljunk a sporteredményekből nyerhető percentilis, decilis értékekre, amelyek ismerete orientálhatja a sportolókat felkészülésük során.

A kvantilisek meghatározásának módszere azonos a mediánnál megismert eljárással. Általában a j-edik kvantilis a rangsor j(n + 1)/k-adik tagja.[1]Amennyiben a sorszám nem egész szám, a két szomszédos adat egyszerű számtani átlaga a kvantilis érték.

Osztályközös gyakorisági sor esetén a kvantiliseket a medián kapcsán megismert közelítő eljárással lehet becsülni.

A jegyeladásokat taglaló példánkban az alsó kvartilis:

[1]Ez megfelel a medián esetében a sorszámot meghatározó képletnek.

$$Q_1 = 40 + \frac{7{,}5 - 7}{7} \times 10 \approx 41 \text{ darab/nap}$$

A felső kvartilis $Q_3 = 61$ darab/nap.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Nevezze meg az alábbi kvartiliseket!

  K = 2.....

  K = 4.....

  K = 10.....

  K = 5.....

  K = 3.....

- Állítsa emelkedő sorrendbe a következő kvantiliseket!

  Nyolcadik decilis

  Első kvartilis

  Medián

  Tizedik percentilis

12. fejezet - Szóródási mérőszámok

Az előzőekben megismerkedtünk a középértékekkel és tudjuk azt, hogy segítségükkel lehetőség nyílik a sokaság egészének tömör jellemzésére. Mivel a valóságot tükröző számadatok különböznek általában egymástól, és eltérnek a jellemzésükre használt középértékektől is, ezt a különbözőséget is vizsgálnunk kell.

Szóródásnak *nevezzük a statisztikában az adatok (általában a mennyiségi ismérvértékek) eltérését egymástól, vagy egy meghatározott, a sokaság egészét jellemző értéktől*. A szóródás vizsgálata igen fontos helyet tölt be a statisztika módszertanában, szinte valamennyi módszer kapcsolódik hozzá.

A szóródásról már képet alkothattunk a gyakorisági sor és a hisztogram értékelése során, azonban óhatatlanul felmerül annak igénye, hogy a szóródás jelenségét egyetlen számértékbe tömörítsük és értékeljük.

A szóródás jelenségének gyors mérésére természetesen jól alkalmazhatók azok a mutatószámok, amelyek csak néhány, a helyzetüknél fogva jelentős számadatot hasznosítanak. Az egymástól eltérő, tehát szóródó adatok egy-egy középértéktől, így a számtani átlagtól mint közepes értéktől is eltérnek. Mindez azt is jelenti, hogy a szóródás mérésére szerkesztett mutatószámok egy része ezt a tulajdonságot használja fel.

Valamennyi szóródást mérő **mutatószám**mal szemben megfogalmazódik az a követelmény, hogy **értékük a szóródás hiánya esetén nulla**, a szóródás megléte esetén nullától különböző számérték legyen. Mi csak néhány gyakrabban használt szóródási mérőszámot mutatunk be vázlatosan:

1. Szóródás terjedelme (T)

2. Interkvartilis terjedelem (TQ)

3. Átlagos eltérés[1] (δ)

4. Szórás[2] (σ) és a variancia [szórásnégyzet] (σ^2)

5. Relatív szórás (V).

A szóródás különböző mutatószámai ugyanazt a jelenséget különbözőképpen közelítik meg, eltérően mérik. A mutatószámok felhasználása során különösen ügyelni kell arra, hogy csak azonos tartalmú értékeket hasonlítsunk össze. Törekedni kell arra, hogy egy-egy konkrét vizsgálat során a vizsgálati célnak legjobban megfelelő mérőszámot alkalmazzuk.

1. A szóródás terjedelme

A **szóródás terjedelme** az előforduló legnagyobb és legkisebb érték különbsége:

$T = x_{max} - x_{min}$

A jegyeladásokkal foglalkozó példánkban a szélső értékek rendre 25 db/nap, illetve 78 db/nap voltak, így T = 78 – 25 = 53 db/nap. Amennyiben osztályközös gyakorisági sorral dolgozunk, és az alsó, illetve felső intervallumok nyitottak, a legnagyobb és a legkisebb osztályközepet használhatjuk fel. Példánkban tehát T = 75 – 25 = 50 db/nap. Az eltérést

[1] δ (kis delta) – az átlagos eltérést jelöli.
[2] σ (kis szigma, egyszerűen: szigma) – a szórás jelölésére szolgál.

a tényleges illetve a becsült adatok közötti különbség okozza.

A szóródás terjedelme könnyen számítható, jól értelmezhető mérőszám, azonban hátránya, hogy csak a szélső értékekre épít, így egy-egy kiugró érték nagyságát számottevően befolyásolhatja.

Némely esetben fontos lehet számunkra egy kiugró érték. A sportolók felkészülése során nem elhanyagolható információ a **legjobb eredménytől (világcsúcstól) való eltérés** nagysága. Az így definiált terjedelmi mutatószám csökkenése jelzi a sportoló formájának javulását.

2. Interkvartilis terjedelem

A terjedelem mutatójának fenti hátrányát kísérli meg kiküszöbölni az **interkvartilis terjedelem** mutatója. Az interkvartilis terjedelem azt az intervallumot jelöli, ahol az összes érték középső 50%-a helyezkedik el. Az interkvartilis terjedelem képlete:

$TQ = Q_3 - Q_1$

A jegyeladásokat taglaló példánkban tudjuk, hogy a megfigyelt napok 75%-ában 61 darabnál kevesebb, míg a napok 25%-ában 41 darabnál kevesebb jegyet értékesítettek. Az interkvartilis terjedelem:

TQ = 61 – 41 = 20 jegy/nap.

3. Átlagos eltérés

Az **átlagos** (abszolút) **eltérés** épít arra a gondolatmenetre, hogy a számértékeknek egy középértéktől való eltéréseiből következtetni tudunk a szóródás nagyságára. Ezeket az eltéréseket „sűríthetjük” egy középérték segítségével. Az értékeknek a számtani átlagtól mért eltérése közvetlenül nem használható, mivel azok összege nulla,
$\sum(x_i - \bar{x}) = 0$
. Ezért csak az **eltérések abszolút értékeiből számított átlag**nak van értelme:

$$\delta = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$

Az átlagos eltérés megmutatja, hogy az egyes ismérvértékek átlagosan mennyivel térnek el az átlaguktól.

Átlagos eltérést természetesen gyakorisági sor adataiból és osztályközös gyakorisági sorból is számíthatunk. Ez utóbbi esetben az osztályközepek eltérését mérjük az átlagtól. A mutatószám **súlyozott** formája:

$$\delta = \frac{1}{n}\sum_{i=1}^{k} f_i|x_i - \bar{x}|$$

ahol: $n = \sum f_i$

A jegyeladások átlagos eltérését az alábbi módon számíthatjuk ki:

$$\delta = \frac{2|25-51| + 5|35-51| + 7|45-51| + 8|55-51| + 5|65-51| + 3|75-51|}{30} = 11{,}6\ \text{db/nap}$$

A szóródás jelenségének mérése szempontjából kedvező, ha a mérőszám nagysága alapvetően csak a vizsgált jelenségtől függ, nem zavarja más külső tényező. Az átlagos eltérés mutatószáma azonban rendelkezik egy „torzító” tényezővel, ugyanis bebizonyítható, hogy a különbségek abszolút értékeinek összege akkor minimális, ha az eltéréseket a mediántól mérjük. Amennyiben a mediánt az adott vizsgálat során kiszámítjuk és értelmezzük, a számtani átlagot a mediánnal jól helyettesíthetjük. A gyakorlatban a mediánnál azonban többször alkalmazzuk a jelenségek vizsgálatára a számtani átlagot.

4. Szórás

Az egyes értékek számtani átlagtól való eltéréseinek négyzetes átlagát **szórás**nak nevezzük.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ha gyakorisági sorból számítjuk a szórást, a mutatószám **súlyozott** formáját kell alkalmazni:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f}} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}$$

A számtani átlag rendelkezik egy olyan „kedvező” tulajdonsággal, amely alapján nem kell számítani a szórás mérőszámában szisztematikus torzító hatásra.

A szórás négyzetét **variancia**nak (σ^2) hívjuk. Önálló tartalommal nem bír, bizonyos statisztikai eljárásokban azonban nagyon fontos szerepet tölt be.

A statisztikai módszerek további megismerése során többször fogunk találkozni a szórás fogalmával, mutatószámával. Szinte valamennyi összetett statisztikai módszer épít erre a mérőszámra. A szórás mutatószáma – alapvetően kedvező matematikai tulajdonságai miatt – a szóródás mérésének elsőrendű fontossággal bíró eszköze, sokszor a szóródás jelenségét a szórás fogalmával helyettesítik (helytelenül!) a napi gyakorlatban.

A szórás kiszámításának menetét a jegyértékesítések példájának segítségével mutatjuk be. A számításhoz szükséges adatokat a 12-1. tábla tartalmazza.

12.1. táblázat - A szórás kiszámításának munkatáblája

A naponta eladott jegyek száma (db/nap)	Osztályközép x_i	Napok száma f_i	$(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$
21–30	25	2	–26	1 352
31–40	35	5	–16	1 280
41–50	45	7	–6	252
51–60	55	8	4	128
61–70	65	5	14	980
71–	75	3	24	1 728
Összesen:	–	30	–	5 720

Forrás: saját számítás

$$\sigma^2 = \frac{5.720}{30} = 190{,}67 \qquad \sigma = \sqrt{190{,}67} = 13{,}8 \text{ db/nap}$$

Az egyes napokon értékesített jegyek száma átlagosan 13,8 darabbal tér el az átlagtól. Természetesen az alapadatokból is kiszámítható a variancia és a szórás. A tényleges számtani átlaghoz (51,3 db/nap) mérve, a fenti mutatószámok sorrendben 192,45 illetve 13,9 db/nap. Láthatjuk, hogy a kétféle számítási mód eltérése nem jelentős.

A szórás és – amint azt említettük – a variancia mutatószáma különösen kedvelt a statisztika módszertanán belül, és a statisztika alkalmazói körében. Alapvetően jó elméleti bázison közelítve méri a szóródás jelenségét, továbbá a valószínűségelméletben

definiált elméleti szórás empirikus megfelelője.

5. Relatív szórás

A szóródás eddig megismert mérőszámai a mennyiségi ismérv mértékegységében fejezik ki a szóródás nagyságát. Sok esetben szükség lehet arra, hogy elvonatkoztassunk a mértékegységektől (és/vagy nagyságrendektől), és ezáltal összehasonlíthatóvá tegyük a különböző jelenségek, különböző mértékegységben kifejezett szóródását (Például a sportteljesítmények szóródását összevethessük a bérek szóródásával egy adott sportágban.). A megoldást az adja, ha a szóródási mérőszámot egy középértékhez, – értelemszerűen – a számtani átlaghoz viszonyítjuk. A leggyakrabban használt ilyen jellegű mérőszám a **relatív szórás**, vagy más néven variációs koefficiens, amelynek képlete:

$$V = \frac{\sigma}{\bar{x}}$$

Ezek alapján a relatív szórás kifejezi azt, hogy az egyes értékek átlagosan hány%-kal térnek el az átlagtól.

A jegyértékesítéssel foglalkozó példánkban a relatív szórás:

$$V = \frac{13,8}{51} = 0,27$$

azaz, a szórás az átlagos értéknek mintegy 27%-a.

6. A szórás felhasználásának néhány további lehetősége

A szórás mérőszámának használata számos alkalmazási lehetőséget kínál. A teljesség igénye nélkül itt csupán egy-két olyan területre szeretnénk felhívni a szakemberek figyelmét, amely – megítélésünk szerint – hatékony segítséget adhat a mindennapi munkában.

Az edzők számára fontos információt jelenthet a sportolók **teljesítményingadozás**a. A folyamatosan vezetett eredménylistából számított szórás közvetlen választ ad erre a kérdésre. Eltérő edzéstervet és technikát kell kidolgozni annak az úszónak, aki a 100 méteres távot átlagosan 49,12 másodperc alatt, de 5,15 másodperces szórással (mint teljesítményingadozással) teljesíti, mint annak a versenyzőnek, aki ugyanezen átlagos értéket 3,15 másodperces szórással éri el.

A versenyzők teljesítményének számszerű vizsgálata segíthet az atlétika területén is. Például a távolugrás eredményeiből számított szórás hatékonyan alkalmazható egy **minimumküszöb** meghatározásánál. Feltéve, hogy az adott versenyző(k) ugrásai 30 cm-es szórással rendelkeznek 800 – 30 = 770 cm lehet egy elfogadható célként megjelölt küszöbérték. (Itt a az átlagot jelöli.)

Általánosságban a szórás mérőszáma a sportolók eredményeinek „kockázati” megjelenítőjeként is felfogható.

7. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy asztaliteniszütő-ragasztót gyártó cégnek a napi forgalmi adatait tartalmazza a következő statisztikai tábla.
  - Számítsa ki, hogy a vásárlások mennyivel térnek el az átlagostól (szórás)!
  - Számítsa ki a relatív szórás mértékét!

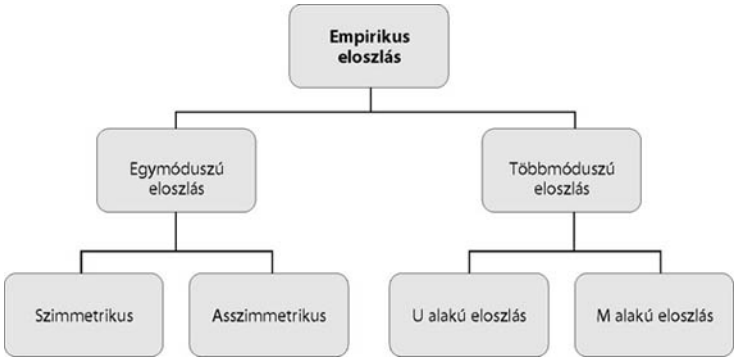
- Ön mint edző tevékenykedik egy női kosárlabdacsapatnál. A félév végén összehív egy szakosztály-értekezletet, ahol az elmúlt félév játékos teljesítményét értékeli. A hét mérkőzésre vonatkozó teljesítményadatokat a következő táblázatban olvashatjuk.
  - Értékelje a játékosokat a tejesítményeik alapján!
  - Mondja el, ki volt a legponterősebb, kinek a játékában volt a legkisebb teljesítményingadozás, és az milyen mértékű volt!
- Egy általános iskolában a 8. évfolyamon (A. és B. osztály) egy falhoz pattintási tesztet végeztek. A 60 főnek egy méterről kellett a kézilabdát minél többször egy perc alatt a falhoz pattintania. Az eredményeket a következő táblázat közli.
  - Határozza meg a falhoz pattintások átlagát!
  - Számítsa ki a leggyakrabban előfordulő értéket!
  - Az adatok alapján számolja ki a mediánt!
  - Mennyi a teljesítmények szórása, ha ismert a relatív szórás mértéke (V = 1,56)?

13. fejezet - Empirikus eloszlástípusok. Aszimmetria mérése

A gyakorisági sorok poligonjának alakja, formája igen fontos információt szolgáltat a vizsgált jelenségről. A gyakorisági sorokat ábrázolva megállapítható, hogy a görbék igen változatosak lehetnek, de nagy többségük bizonyos szabályszerűséget mutat. A különböző görbék közül igen nagy jelentősége van azoknak, amelyek valamilyen ismert elméleti eloszlás empirikus megfelelői.

Az empirikus eloszlástípusok megnyugtatóan a gyakorisági görbe birtokában azonosíthatóak (ami egyben feltételezi a nagyszámú megfigyeléseket és a véletlen hatásának nem meghatározó voltát), azonban a gyakorlatban sokszor meg kell elégedni a gyakorisági poligon, a hisztogram, illetve a gyakorisági sor alapján történő besorolással. Az eligazodást nagymértékben segíti az eloszlástípusok ismerete. Az osztályozás történhet az alábbi séma alapján:

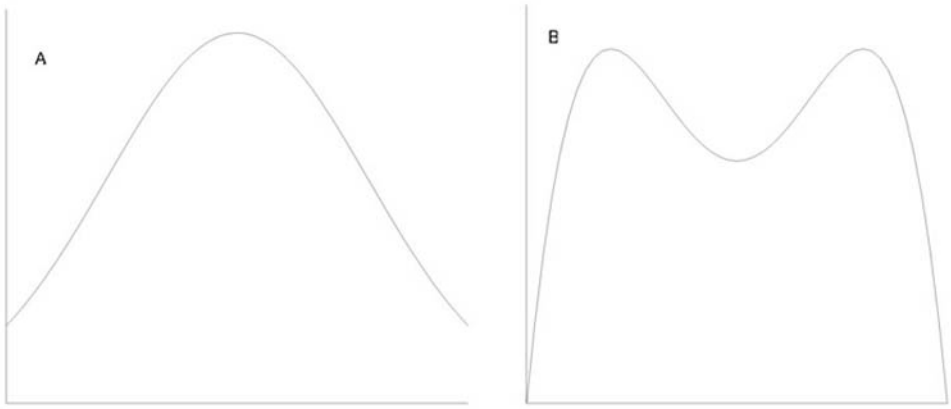
13.1. ábra - Az empirikus eloszlás felosztása



Forrás: Saját szerkesztés

Az egymóduszú, illetve a többmóduszú eloszlások görbéjének egy-egy típusát találjuk a 7. ábrán.

13.2. ábra - A móduszok grafikus ábrája



Forrás: saját szerkesztés

A fenti ábra A. részében egy egymóduszú, **szimmetrikus** eloszlás görbéjét vázoltuk fel.

A többmóduszú eloszlások általában heterogén[1] sokaságot jellemeznek. A gyakorisági görbe helyi maximumai a homogénebb részsokaságok móduszainál jönnek létre. Jellemző példa lehet – az ábra B. részében szereplő – M alakú eloszlásra az alkalmazottak keresetének eloszlása, ami a képzettség, nemek stb. szerint különböző helyi maximumokkal bírhat. Viszonylag ritkán fordul elő a gyakorlatban egy olyan kétmóduszú, U alakú eloszlás, amelynek jellemzője, hogy a két módusz egyben a két szélső érték.

Az **egymóduszú gyakorisági sorok** lehetnek **szimmetrikus** vagy **aszimmetrikus** eloszlásúak. A szimmetrikus gyakorisági sorok jellemzője, hogy grafikus ábrájuk a módusz értékénél felvehető tengelyre szimmetrikus. Az ilyen eloszlásoknál a **módusz**, a **medián** és a **számtani átlag** egyenlő egymással. Szimmetrikus eloszlású pl. a felnőtt férfiak és nők testmagasság szerinti megoszlása; de szimmetrikus eloszlású lehet egy adott munkahelyen a teljesítménybérben dolgozók egyéni teljesítményének eloszlása is. A szimmetrikus empirikus eloszlás legtöbbször a matematikai statisztikából ismert normális eloszlás gyakorlati megfelelője.

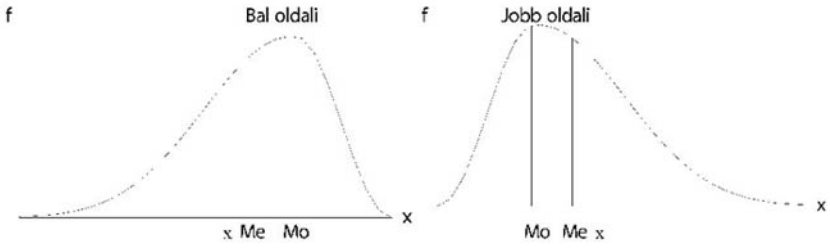
Aszimmetrikus eloszlások esetén a módusz a két szélső érték közül az egyikhez esik közelebb. A három alapvető középérték nagyságrendje jellemzi a ferdeséget. **Bal oldali**[2] **aszimmetria** esetén:

$Mo > Me > \bar{x}$

Jobb oldali aszimmetria esetén a nagyságrendi reláció:

$Mo < Me < \bar{x}$

13.3. ábra - A kétféle jellemző aszimmetrikus eloszlás



Forrás: Saját szerkesztés

A ferde, aszimmetrikus eloszlások közül a gazdasági életben, de a sportban is a jobb oldali aszimmetriát mutató eloszlás a gyakoribb. Tipikus példa a keresetek, jövedelmek eloszlása, de a sportolók teljesítményének eloszlása is jobb oldali aszimmetriájú, ugyanis kevesen érnek el kimagasló eredményeket, a többség csak szerényebb teljesítményre képes. Általában elmondhatjuk, hogy a jobb oldali aszimmetriájú eloszlások esetén a mennyiségi ismérvértékek alsó határát szigorúbb törvények szabják meg mint a felsőt.

Bal oldali aszimmetria esetén a viszonylag magasabb értékek nagyobb gyakorisággal fordulnak elő. Jól szemlélteti ezt az eloszlást, pl. a halálozások életkor szerinti megoszlása, de a sport területéről is vehetünk példát, ha arra gondolunk, hogy a kitűzött edzési szintek már elavultak, a sportolók többsége túlteljesíti azokat.

Az aszimmetrikus eloszlások szélsőséges típusa az ún. **J alakú eloszlás**, amelynek jellemzője, hogy a módusz valamelyik szélső értékkel esik egybe.

A sokaság mennyiségi ismérv szerinti tömör jellemzését tudjuk adni mind a

[1]Heterogénnek tekintjük a sokaságot, ha valamely ismérv szerint homogénebb, egyneműbb részekre bontható.
[2]A bal- és jobboldali elhatárolás szubjektív megítélésen alapszik. Több szakkönyv a fentiekkel ellentétesen jelöli az aszimmetriát. Mi az angolszász terminológiát vettük alapul, amelyet a statisztikai szoftverek többsége használ.

középértékek, mind a szóródási mérőszámok segítségével. A gyakorisági görbe alakjának tanulmányozása a vizsgált jelenség további jellemzőit segít feltárni. (Előfordulhat, hogy két jelenség összehasonlítása során azonos számtani átlag és azonos szórás mögött eltérő eloszlás rejlik.) Az eloszlások empirikus vizsgálata a középértékek, illetve a szóródási mérőszámok kiszámításával együtt más-más megközelítésű vizsgálatát adják az adott jelenségeknek. Így az elemzés komplexebbé válik.

Az egymóduszú gyakorisági sorok esetén felmerül az aszimmetria (ferdeség) egzakt mérésének igénye. Többféle mutatószámot használnak az aszimmetria mérésére, amelyeknek közös tulajdonságai:

1. értékük nulla legyen, ha az eloszlás szimmetrikus,

2. jobb oldali aszimmetria estén pozitív, míg ellenkező esetben negatív értéket vegyenek fel,

3. dimenzió nélküliek legyenek.

Széleskörűen ismert, és igen gyakran alkalmazott az aszimmetria alábbi mérőszáma (jele: A):

$$A = \frac{\bar{x} - Mo}{\sigma}$$

A mutatószám azon a tényen alapul, hogy szimmetrikus eloszlásoknál a számtani átlag és a módusz értéke biztosan megegyezik. Szimmetrikus eloszlás esetén a mutató értéke nulla, jobb oldali aszimmetriánál pozitív, míg bal oldali aszimmetria esetén negatív az előjele. A mérőszámnak abszolút értékben nincs felső korlátja, azonban 1-nél nagyobb abszolút érték már erőteljes aszimmetriát jelez.

A jegyeladással foglalkozó példában az osztályközös gyakorisági sor alapján becsült nevezetes mérőszámok:

$\bar{x} = 51$ db/nap, $\sigma = 13{,}8$ db/nap, $Mo = 52{,}5$ db/nap.

Az aszimmetria foka:

$$A = \frac{51 - 52{,}5}{13{,}8} = -0{,}11$$

A jegyeladások eloszlása enyhén bal oldali aszimmetriájú.

Az aszimmetria egyik további mérőszámának (jele: F) logikája feltételezi, hogy szimmetrikus eloszlású gyakorisági sorok esetén a medián az alsó és a felső kvartilistől egyenlő távolságra helyezkedik el. Ha a medián közelebb esik az alsó kvartilishez jobb oldali, ellenkező esetben bal oldali az aszimmetria. A mutatószám képlete:

$$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

Jobb oldali aszimmetriánál pozitív, míg bal oldali aszimmetriánál negatív lesz a mutató értéke, amely szimmetrikus eloszlású gyakorisági sor esetén zéró értékkel bír. A mutató határozott alsó és felső határokkal bír:

$-1 \leq F \leq 1$

Az F mérőszám általában 0,3 tizednél nagyobb abszolút érték esetén már jelentős ferdeséget jelez.

A jegyeladások esetén a kvartilisek az alábbi értékeket vették fel: $Q_3 = 61$ $Me = 51{,}5$ $Q_1 = 41$

Az F mérőszám:

$$F = \frac{(61 - 51{,}5) - (51{,}5 - 41)}{(61 - 51{,}5) + (51{,}5 - 41)} = -0{,}05$$

Az eloszlás nagyon enyhe bal oldali aszimmetriájú.

Az aszimmetria vizsgálatának gyakorlati hasznosítására jó lehetőség nyílik a teljesítmény-sportok területén. Például egy úszóedző célként tűzheti ki a 100 méteres távnak 55 másodperc alatti teljesítését. Amennyiben „tanítványainak” eredményei alapján bal oldali aszimmetriát tapasztal, tehát több az 55 másodpercnél hosszabb idő, növelni kell különféle eszközökkel a sportolók fizikai felkészültségét. Ellenkező esetben (jobb oldali aszimmetria esetén) csökkenteni kell a kitűzött szintidőt, ezzel inspirálva a sportolókat a jobb teljesítmény elérésére.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy asztaliteniszütő-ragasztót gyártó cégnek a napi forgalmi adatait tartalmazza a következő statisztikai tábla.
  - Határozza meg a mediánt!
  - Határozza meg az aszimmetriát!
- Egy 88 fős főiskolai évfolyamon az 50 méteres gyorsúszást mérték fel, melynek eredményeit a következő tábla szemlélteti.
  - Határozza meg a középértékéket, és az aszimmetriát minimum egy mutatószám segítségével! Értékelje a kapott eredményt!

14. fejezet - A koncentráció mérése

Koncentráción általában a jelenségek tömörülését, összpontosulását értjük. A koncentráció fogalma mind a gazdasági, társadalmi folyamatokat, mind az azok eredményeként létrejött állapotokat jellemzi. Így beszélhetünk pl. a termelés, a forgalom, a beruházások koncentrációjáról, de az elemzés tárgya lehet a jövedelmek, a vagyon, a munkavállalói létszám, a sportolók lakóhely szerinti koncentrációja.

A koncentrációt a gyakorisági, mennyiségi sorok alapján mérhetjük, ugyanis jó megközelítést adhatunk, ha egy adott X ismérv gyakorisági és értékösszeg-eloszlását hasonlítjuk össze. Amennyiben a relatív gyakoriságok nagy értékeihez alacsony relatív értékösszegek tartoznak (illetve fordítva), a koncentráció meglétéről beszélhetünk.

A koncentráció jelenlétéről gyorsan tájékozódhatunk, ha a sokaságot egy mennyiségi ismérv szerint csoportosítjuk, és egy statisztikai táblában helyezzük el a kumulált relatív gyakoriságokat és a kumulált relatív értékösszegeket. A kumulált relatív gyakoriságok és értékösszegek viszonya ugyanis szemléletesen fejezi ki a koncentráció létezését. Amennyiben azt tapasztaljuk, hogy az egyes intervallumokhoz rendelhető kumulált relatív gyakoriságokhoz rendre kisebb kumulált relatív értékösszegek tartoznak, koncentrációról beszélhetünk.

14.1. táblázat - A városok (Budapest nélkül) népességmegoszlása Magyarországon, 1997. év végi népességszámuk szerint

Népesség (fő)	Városok száma
2 000–4 999	20
5 000–9 999	61
10 000–49 999	105
50 000–99 999	11
100 000–	8
Összesen:	205

Forrás: Magyar Statisztikai Évkönyv 1997.

Számítsuk ki a relatív gyakorisági- és a relatív értékösszegsor adatait!

A relatív gyakoriságok:

14.2. táblázat - Munkatábla

Népesség (fő)	Városok száma (f_i)	Relatív gyakoriság, % (g_i)
2 000–4 999	20	9,8
5 000–9 999	61	29,8
10 000–49 999	105	51,1
50 000–99 999	11	5,4
100 000–	8	3,9
Összesen:	205	100,0

Forrás: Saját számítás

14.3. táblázat - A relatív érték-összegek munkatáblája

Népesség (fő)	Városok száma (f_i)	Osztályközép (x_i)	Értékösszeg, Lakosság, fő (f_ix_i)	Relatív értékösszeg, % (z_i)
2 000–4 999	20	3 500	70 000	1,3
5 000–9 999	61	7 500	457 500	8,3
10 000–49 999	105	30 000	3 150 000	57,2
50 000–99 999	11	75 000	825 000	15,0
100 000–	8	125 000	1 000 000	18,2
Összesen:	205	–	5 502 500	100,0

Forrás: Saját számítás

A 29. tábla a kumulált relatív gyakoriságokat és a kumulált relatív értékösszegeket tartalmazza:

14.4. táblázat - A kumulált relatív gyakoriságok és értékösszegek munkatáblája

Népesség (fő)	Kumulált relatív	
Gyakoriság, % (g'_i)	Értékösszeg, % (z'_i)	
2 000–4 999	9,8	1,3
5 000–9 999	39,6	9,6
10 000–49 999	90,7	66,8
50 000–99 999	96,1	81,8
100 000–	100,0	100,0

Forrás: Saját számítás

A kumulált relatív gyakorisági értékek jelentősen meghaladják a kumulált relatív értékösszeg adatait. A városi népesség koncentrációját érzékelhetjük, ha tetszés szerint egy adatot kiragadunk. Például az összes vidéki város 90,7%-a 50 ezer főnél kisebb lélekszámú volt, de az összes városi népességnek csupán 66,8%-a lakott ezekben a városokban.

Erősnek tekintjük a koncentrációt, ha a sokaság nagy hányadához a teljes értékösszeg kis hányada tartozik, ugyanakkor a sokaság kis hányada az értékösszeg jelentős hányadát mondhatja magáénak.

A koncentráció ábrázolására és elemzésére szolgáló speciális grafikus ábrát, megalkotójáról, **Lorenz-görbé**nek nevezték el (Kerékgyártó–Mundruczó–Sugár, 2001). A Lorenz-görbe egységoldalú négyzetben elhelyezett ábra, amely a kumulált relatív gyakoriságok (g'_i) függvényében ábrázolja a kumulált relatív értékösszegeket (z'_i).

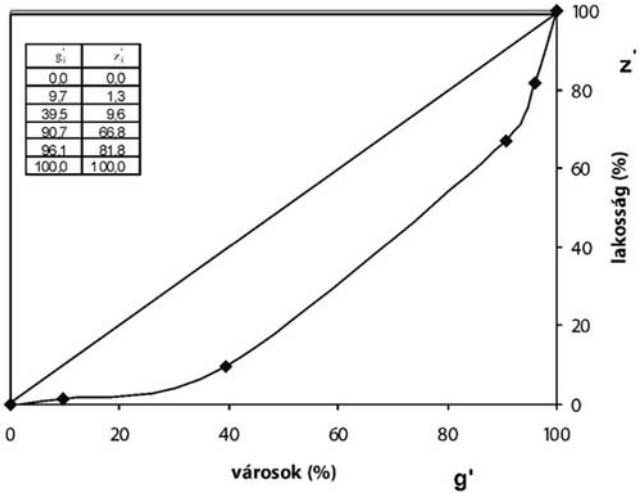
Amennyiben az egységeknek az értékösszegből való részesedése egyforma, a kumulált relatív gyakoriságok és a kumulált relatív értékösszegek rendre megegyeznek ($g'_i = z'_i$). Mindez a koncentráció hiányára utal. Ilyen esetben a görbe a négyzet átlójával egybeesik.

Ha a sokaságban létezik olyan egység, amely az értékösszeg igen nagy hányadát leköti, a relatív gyakoriságok és relatív értékösszegek igen jelentősen eltérnek egymástól, a görbe a koordinátatengelyekhez igen közel kerülhet. A teljes koncentráció esetén a görbe egybeesik a koordinátatengelyekkel.

A görbe és az átló által bezárt terület a koncentráció relatív nagyságát jellemzi. Léteznek olyan mutatószámok, amelyek az átló és a koordinátatengelyek által meghatározott háromszöghöz mérik a görbe és az átló által bezárt terület arányát, azonban ezek tárgyalásától itt eltekintünk.

A vidéki városi népességre vonatkozó kumulált relatív gyakoriságok és értékösszegek alapján elkészíthető a Lorenz-görbe, amely mint a 9. ábráról leolvasható, közepesnél gyengébb mértékű koncentrációt jelez.

14.1. ábra - A városok koncentrációjának Lorenz-görbéje



Forrás: saját szerkesztés

A következőkben egy gyakorlati példával illusztráljuk a sport területi koncentrációját. Érdeklődésre tarthat számot az, hogy mutat-e koncentrációt a hazai olimpiai keret sportolók területi elhelyezkedése vagy egyenletes a területi eloszlásuk. A számításokat regionális bontásban végeztük el.

14.5. táblázat - A tehetséges sportolók területi koncentrációjának számítása

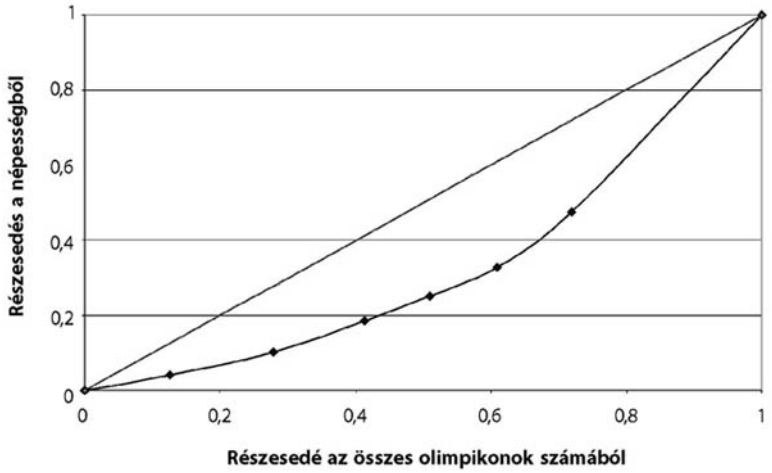
Régió	Olimpiai keretsportolók száma	Népesség (eFő)	1 millió főre jutó olimpikonok száma	Kumulált népesség (eFő)	Részesed és a népességből	Részesed és az olimpikonok számából
Észak-Magyarország	36	1 271	28,32	1 271	0,13	0,04
Észak-Alföld	53	1 542	34,37	2 813	0,28	0,10
Dél-Alföld	73	1 355	53,87	4 168	0,41	0,19
Dél-Dunántúl	56	977	57,32	5 145	0,51	0,25
Nyugat-Dunántúl	68	1 000	68,00	6 145	0,61	0,33
Közép-Dunántúl	128	1 111	115,21	7 256	0,72	0,48

Régió	Olimpiai keretsportolók száma	Népesség (eFő)	1 millió főre jutó olimpikonok száma	Kumulált népesség (eFő)	Részesed és a népességből	Részesed és az olimpikonok számából
Közép-Magyarország	457	2 841	160,86	10 097	1,00	1,00

Forrás: Saját számítás

A kapott számítási eredményeket szemléltethetjük a *Lorenz-görbe* segítségével, grafikus módon is.

14.2. ábra - Az olimpiai keret sportolók Lorenz-görbéje



Forrás: Saját szerkesztés

Megállapíthatjuk, hogy a tehetséges sportolók területi eloszlása nem egyenletes, mivel a görbe nem esik egybe az átlóval. Tehát elmondható, hogy a sportolók területi elhelyezkedésében koncentráció tapasztalható, mivel a görbe láthatóan eltér az átlótól és közelít a koordinátatengelyhez.

A Lorenz-görbéről és a területi egyenlőtlenségek mérésekor használt mutatókról bővebben Ács, 2007 írásában olvashatnak.[1]

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy vizsgálat során a hazai sportcsarnokok befogadóképességét vizsgálták. Az eredményeket a következő táblázat közli.
  - Számolja ki, hogy az itt feltüntetett sportcsarnokok hány százaléka képes legalább 4000 nézőt egy sporteseményen befogadni?
  - Hány százaléka ez az összes sportcsarnoknak?
- A következő táblázatban az első osztályú sportklubok számát láthatjuk, regionális bontásban.

[1]Ács (2007): A területi egyenlőtlenségek feltérképezése során leggyakrabban alkalmazott mérőszámok bemutatása, a sporttehetségek területi elhelyezkedésének példáján, Egy életpálya három dimenziója-Tanulmánykötet Pintér József emlékére (ISBN 978-963-642-195-3), Pécsi Tudományegyetem Közgazdaságtudományi Kar, Pécs, 10–22. o.

- Az adatok felhasználásával készítsen grafikus ábrát az első osztályú csapatok területi eloszlásáról.

15. fejezet - Csoportosított adatok átlaga, szórása

A korábbiakban megismerkedtünk a viszonyszámok, középértékek és szóródási mérőszámok számításának alapvető kérdéseivel. A vizsgált jelenségek, folyamatoknak a fenti módszerekkel történő vizsgálata során kimondatlanul is fel kellett tételezni, hogy a sokaság **homogén**. Természetesen a gyakorlatban többször találkozunk olyan problémával, amikor **heterogén** (összetett) a vizsgálandó sokaság. Általánosságban *heterogénnek nevezzük a sokaságot, ha valamilyen ismérv alapján viszonylag homogén részekre (csoportokra) bontható.*

A vizsgálandó sokaság természetének ismeretében található meg az az ismérv, amely alapján egy heterogén sokaság homogén, de minőségileg egymástól különböző csoportokra bontható.

A sportolók vizsgálata során gyakorta észlelhetjük, hogy a sportolók keresetük szerint nem homogén sokaságot alkotnak. Például nemek szerint csoportosítva a keresetek szempontjából homogénebb, egyneműbb csoportok képezhetők. Lényeges csoportképző ismérv a sportoló napi munkahelyre (sportlétesítménybe) történő utazásának időtartama szempontjából a lakóhely (a sportoló helyben sportol vagy ingázó). Viszonylag kézenfekvő a sportolók teljesítményének nemek szerinti bontása, aminek segítségével homogénebb csoportok képezhetőek.

A csoportosított sokaságban is érdeklődésre tarthat számot a korábban megismert valamennyi mutatószám. Ebben a fejezetben csupán két módszer – a számtani átlag és szórás – számítási sajátosságát mutatjuk be, csoportosított sokaság esetén. Természetesen az összetett sokaság elemzése során is fontos információkat szolgáltatnak a csoportok egyéb mérőszámai, amelyeket a már megismert módszerekkel határozhatunk meg. A sokaság egészére meghatározható mutatószámok tovább színesítik az elemzési eszköztárat.

A csoportosított sokaságból számított átlag és szórás számítását vázlatosan, egy példa segítségével mutatjuk be.

Az edzés időtartamát vizsgáltuk 200 élsportolót megkérdezve. A napi sportolás időtartamára vonatkozóan a 15-1. táblában közölt fontosabb értékeket kapták:

31. táblázat:

15.1. táblázat - A sportra fordított napi időmennyiség

Megnevezés	Megkérdezettek száma (fő)	Megoszlás (%)	Napi átlagos idő (óra)	A sportolás időtartamának szórása (óra)
Férfi	120	60	3	1,5
Nő	80	40	2	0,5

Forrás: Saját szerkesztés

Számítsuk ki a csoport egészére vonatkozóan a napi sportolás átlagos időtartamát és annak szórását!

Meg kell jegyezni, hogy a homogénebb csoportokban (részsokaságokban; itt férfi és nő) mind a számtani átlagokat, mind a szórásokat a korábban megismert módon számítottuk ki.

A fenti példában a nemek csoportképző ismérvek. Segítségükkel létrejött homogén **csoportok átlagos értékei** ugyanúgy kezelhetők, mintha egy mennyiségi ismérv értékei lennének, *a számtani átlag számításának szabályai szerint átlagolhatók. A főátlag a csoportátlagok számtani átlaga.*

A főátlag (x) képlete:

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

ahol: n_j – a megfigyelések száma a j-edik csoportban, x_j – a j-edik csoport átlaga, m – 1,2…m – a csoportok száma.

A főátlag:

$$\bar{x} = \frac{120 \times 3 + 80 \times 2}{200} = 2{,}6 \quad \text{óra.}$$

A számításhoz az arányokat felhasználva:

$$\bar{x} = 0{,}6 \times 3 + 0{,}4 \times 2 = 2{,}6 \text{ óra.}$$

A vizsgált sokaság naponta átlagosan 2,6 órát tölt sportolással.

A csoportosított sokaságban a teljes sokaságra vonatkozóan kiszámítható szórás azonban nem közvetlenül származtatható a részsokaságok (csoportok) szórásaiból. Ennek a megállapításnak a megértését segíti, ha végiggondoljuk a teljes sokaság adatainak szerkezetét. A heterogén (teljes) sokaság *egy-egy megfigyelt számadata* (pl. adott egyén sportolásának időtartama) *eltérhet a saját csoportjának átlagától és egyben a főátlagtól* is. Ugyanakkor – mivel valóságos csoportokról van szó – *a csoportok átlagai is eltérnek a főátlagtól*. Általánosságban az eltéréseket az alábbi módon írhatjuk fel:

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

ahol: x_{ij} – az i-edik megfigyelt egyedi érték a j-edik csoportban, x_j – a j-edik csoport átlaga, x – a főátlag.

Az adatbázis összetett jellege a szóródást kifejező mérőszámokat is jellemzi.

Választ kaphatunk arra a kérdésre, hogy a csoportokon belüli szórások (egyes megfigyelt értékek átlagos eltérései saját csoportátlaguktól) együttesen milyen nagyságrendűek. Ezt az ún. **belső szórás,** illetve a **belső szórásnégyzet (variancia)** mutatójával számszerűsíthetjük. A belső szórásnégyzet meghatározható a csoportok szórásnégyzetének átlagaként.

Képlete:

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

ahol: n_j – a j-edik csoport elemeinek száma, n – az összes elemszám, m – a csoportok száma,

σ_B^2
– a j-edik csoport szórásnégyzete.

Belső szórásnégyzet:

$$\sigma_B^2 = \frac{120 \times 1{,}5^2 + 80 \times 0{,}5^2}{200} = 1{,}45$$

illetve, $\sigma^2_B = 0{,}6 \text{ x } 1{,}5^2 + 0{,}4 \text{ x } 0{,}5^2 = 1{,}45$.

Belső szórás:

$$\sigma_B = \sqrt{1{,}45} = 1{,}204 \text{ óra.}$$

A nők/férfiak sportolással töltött ideje saját csoportátlaguktól átlagosan 1,204 óra/nap értékkel tér el.

Természetesen a belső szórás – mivel csak a csoportokon belüli eltéréseket fejezi ki – nem egyezik meg a teljes szórással. Az adatok szóródásában a heterogén sokaság esetén ugyanis számolni kell a csoportok átlagainak szóródásával is, amit az ún. **külső szórás** illetve **külső szórásnégyzet (variancia)** fejez ki. A külső szórásnégyzet () meghatározásához a csoportátlagokat úgy tekintjük, mintha azok nem átlagok, hanem mért értékek lennének.

Képlete:

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

ahol: n_j – a j-edik csoport elemeinek száma, n – az összes elemszám, m – a csoportok száma, x_j – a j-edik csoport átlaga, x – a főátlag.

Külső szórásnégyzet:

$$\sigma_K^2 = \frac{120(3-2{,}6)^2 + 80(2-2{,}6)^2}{200} = 0{,}24$$

illetve, $\sigma^2_K = 0{,}6 \text{ x } (3 - 2{,}6)^2 + 0{,}4 \text{ x } (2 - 2{,}6)^2 = 0{,}24$.

Külső szórás:

$$\sigma_K = \sqrt{0{,}24} = 0{,}49 \text{ óra}$$

A csoportátlagok a főátlagtól (és egymástól) a sokaság egészében 0,49 óra/nap nagyságrenddel térnek el átlagosan.

A kétféle megközelítéssel mért szórás lehetőséget ad arra, hogy számszerű értékeik birtokában az egész sokaság szórását, a teljes szórást (σ) is meghatározzuk.

Bebizonyítható ugyanis, hogy a *teljes szórásnégyzet egyenlő a belső szórásnégyzet és a külső szórásnégyzet összegével:*

$\sigma^2 = \sigma^2_B + \sigma^2_K$

Példánkban:

$\sigma^2 = 1{,}45 + 0{,}24 = 1{,}69$

amiből:

$$\sigma = \sqrt{1{,}69} = 1{,}3 \text{ óra.}$$

Tehát a megfigyelt sokaság egyedei átlagosan 1,3 óra/nap értékkel szóródnak a főátlag körül. Ugyanezt az értéket kaptuk volna, ha a megfigyelt adatokkal a sokaság minden értékére vonatkozóan rendelkezünk. (Amennyiben a 200 megfigyelés értékéből „hagyományos” módon – csoportosítás nélkül – számítottunk volna szórást.)

A csoportosított adatokból számított szórás segítséget ad a sokaság jobb megismeréséhez. Amennyiben a *külső szórás értéke nulla,* azaz a részátlagok nem térnek el egymástól, a csoportosításnak nincs értelme, a sokaságot az adott csoportképző ismérv szempontjából homogénnek tekinthetnénk. Mindez *a csoportképző és a mennyiségi ismérv függetlenségét* is jelentené egyben.

Abban az esetben, ha azt tapasztaljuk, hogy a csoportokon belül az adatok nem szóródnak, tehát a **belső szórás értéke** is **nulla,** de a csoportátlagok különböznek, a csoportképző ismérv egyértelműen meghatározza a mennyiségi ismérv értékét. Ilyen esetben az **ismérvek között determinisztikus kapcsolat**ot állapíthatnánk meg.

A fentiekben elmondottak jelzik, hogy a heterogenitás megállapításában, valamint ahogy később látni fogjuk a kapcsolatok mérésében, a szórásnégyzet összetevőkre bontásának

kiemelkedő szerepe van.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy népességcsoportban 4000 főt megkérdezve vizsgálták a sportolási szokásokat. A napi sportolás időtartamára vonatkozóan a következő adatok keletkeztek. A megkérdezettek 70%-a férfi volt. Ők átlagosan 3 órát töltöttek sportolással 1,5 órás szórással. A nők átlagosan 2 órát sportolnak, melynek szórása 0,5 óra.
  - Készítse el a fenti adatok alapján a statisztikai táblát!
  - Számítsa ki a népességcsoport egészére vonatkozóan a napi átlagos sportolási időt és annak szórását!

16. fejezet - Kapcsolatvizsgálatok

A társadalmi és a gazdasági jelenségek és folyamatok egymással összefüggő rendszert alkotnak. Tudjuk, hogy a jelenségeket és folyamatokat ismérvek segítségével azonosíthatjuk, jellemezhetjük és ezek a tulajdonságok nagyon fontos információk hordozói, de egyben lehetőséget teremtenek arra is, hogy a különböző összefüggéseket, hatásokat számszerűsítsük.

A jelenségek, folyamatok közötti kapcsolatok számos szempont alapján csoportosíthatók, azonban a gyakorlati megközelítés szempontjából elterjedt a következőkben leírt tipizálási rendszer:

Az ismérvek (és az általuk jellemzett jelenségek, folyamatok) lehetnek egymástól **függetlenek**, állhatnak egymással **sztochasztikus kapcsolat**ban, illetve lehet a kapcsolatuk **determinisztikus**. Statisztikai módszerekkel elsősorban a sztochasztikus kapcsolatokat vizsgáljuk, így a szélsőséges esetek – a függetlenség és a determinisztikus eset – is óhatatlanul vizsgálatunk tárgyává válhatnak. **Sztochasztikus kapcsolaton** a statisztikai ismérvek között tendenciaszerűen, *valószínűségi jelleggel* érvényesülő kapcsolatot értünk.

A bennünket környező világ megismerése során a kapcsolatok megismerése igen fontos, és az így nyert információk fontos szerephez jutnak egy-egy döntés előkészítése, illetve a döntések hatásának mérése kapcsán. Nem közömbös ugyanis, hogy a döntés hatása egy adott jelenségre jelentős-e vagy jelentéktelen.

A sztochasztikus kapcsolatok csoportosításának legelfogadottabb fajtája az ismérvek típusa szerint épül fel. Beszélhetünk **asszociációs** (mindkét[1] ismérv minőségi), **vegyes** (az ok szerepét minőségi, de az okozat szerepét mennyiségi ismérv tölti be) és **korrelációs** (mind az ok, mind az okozat vagy okozatok mennyiségi ismérvek) kapcsolatról.

Mindhárom típusú kapcsolatot valamilyen mutatószám segítségével célszerű számszerűsíteni. Fontos szerepet töltenek be az ún. kapcsolat intenzitását kifejező mérőszámok, amelynek általános sémáját az alábbiakkal írhatjuk le (az intenzitást *általánosságban mérő mutatószám* jele legyen T):

$0 \leq T \leq 1$.

Általánosságban a T mutató abszolút értékére fogalmazzuk meg a fenti intervallumot, de bizonyos esetben – főleg a korrelációs kapcsolatokban – az előjel is fontos információ hordozója, ugyanis a kapcsolat pozitív vagy negatív irányát mutatja. Természetesen ilyen esetben a mutatószám intervalluma: [–1; 1].

A mutatószámok értelmezése mindig függ az adott problémától; ismerni kell a vizsgált összefüggés jellegét, természetét. Az általános elemzéshez ad segítséget az alábbi séma:

$T = 0$ – nincs kapcsolat,

$0 < T < 0{,}3$ – gyenge kapcsolat,

$0{,}3 < T < 0{,}7$ – közepes szorosságú kapcsolat,

$0{,}7 < T < 1$ – erős a kapcsolat,

$T = 1$ – függvényszerű vagy determinisztikus a kapcsolat,

[1] Természetesen a több ismérv (változó) közötti kapcsolatot is lehet elemezni statisztikai módszerekkel, azonban könyvünkben csak a legegyszerűbb eseteket említjük meg.

1. Asszociációs kapcsolat

Az asszociációs kapcsolat mérését egy példa segítségével mutatjuk be.

Egy labdarúgócsapat hazai és idegenbeli bajnoki mérkőzését kísértük figyelemmel. 80 mérkőzés eredményéből vizsgáltuk a csapat hazai és idegenbeli teljesítményét. 48 mérkőzés hazai pályán, 32 idegenben volt. A győztes, illetve vesztes mérkőzések számát az alábbi tábla tartalmazza:

16.1. táblázat - A felmérés eredményei

Játékhely	**A vizsgált csapat eredménye**		**Összesen**
Győzelem	**Vereség**		
Otthon	39	9	48
Idegenben	11	21	32
Összesen:	50	30	80

Forrás: saját szerkesztés

Határozzuk meg az eredményesség és a játékhely közti kapcsolatot!

Az asszociációs kapcsolat esetében az adatokat egy kombinációs táblába rendezzük, amely a minőségi ismérvek szerinti gyakoriságokat tartalmazza. Az ilyen típusú táblákat – mint már szóltunk róla – ún. kontingenciatábláknak nevezi a statisztikai irodalom. Felírhatjuk a tábla általános formáját (33. tábla):

16.2. táblázat - A kontingenciatábla

A ismérv változatai	**B ismérv változatai**		**Összesen**
B_1	B_2		
A_1	f_{11}	f_{12}	S_1
A_2	f_{21}	f_{22}	S_2
Összesen:	O_1	O_2	n

Forrás: saját szerkesztés

n – az összes elemszám,

f_{11} – az A ismérv első és a B ismérv első változatához rendelt gyakoriság (hasonlóan értelmezhetők a többi cella gyakoriságai!),

S_1 – az első sor (az A ismérv első változatához tartozó) gyakoriságok összege,

O_1 – az első oszlop (a B ismérv első változatához tartozó) gyakoriságok összege.

Belátható az alábbi összefüggés:

$S_1 + S_2 = O_1 + O_2 = n$.

A sorok, illetve az oszlopok összegeit **peremgyakoriságoknak** nevezzük.

Alternatív ismérvek esetén a kapcsolat mérésére alkalmazhatjuk az ún. **Yule-féle mutató**t, ami a táblában szereplő gyakoriságok „keresztszorzataiból” állítható elő:

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

A mutatószám – mivel két adat különbségének és ugyanazon adatok összegének hányadosa – minden esetben –1 és +1 közötti értéket vesz fel.

Példánkban a Yule-mutató:

$$Y = \frac{39 \times 21 - 9 \times 11}{39 \times 21 + 9 \times 11} = 0{,}784$$

A mutató ismeretében megállapíthatjuk, hogy erős sztochasztikus kapcsolat van a játékhely és az adott labdarúgócsapat teljesítménye között. Az előjelnek nem tulajdonítunk jelentőséget, mivel a táblában a sorok vagy oszlopok kicserélése – amire semmilyen ellenérv nem hozható fel – megváltoztatná, negatívvá tenné az előjelet.

Természetesen a fenti kapcsolat szorosságára vonatkozó megállapítás statisztikai jellegű, csak tendenciaszerűen, valószínűségi jelleggel értelmezhető.

A mutatószám alkalmazása során azonban fokozottan figyelni kell arra, hogy valamennyi átlóban lévő elem különbözzön nullától. Ha csak egy esetben nulla a gyakoriság, a mutatószám akkor is determinisztikus kapcsolatot jelez, ha az egyébként nem áll fenn.

Tételezzük fel, hogy egy megfigyelés során az alábbi eredményt kaptuk:

16.3. táblázat - A megfigyelés alapadatai

Játékhely	**A vizsgált csapat eredménye**		**Összesen:**
Győzelem	**Vereség**		
Otthon	39	0	39
Idegenben	11	30	31
Összesen:	50	30	80

Forrás: Saját szerkesztés

A Yule-féle mérőszám:

$$Y = \frac{39 \times 30 - 0 \times 11}{39 \times 30 + 0 \times 11} = 1$$

A fenti esetben nem áll fenn a determinisztikus kapcsolat, ugyanis a csapat otthon nem veszít (csak idegenben), azonban a győzelmek száma erősen megoszlik a játék helye szerint.

Kettőnél több ismérvváltozat esetén más mérőszámot kell alkalmazni. A **Cramer-együttható** feloldja az alternatív ismérvek dilemmáját és ugyanakkor érzéketlen a kirívó (egyik cellában nulla értékkel bíró) esetekkel szemben, alapgondolata az alábbi:

Amennyiben a független viszonyt feltételező gyakoriságok és a tényleges gyakoriságok között eltéréseket találunk, akkor a sztochasztikus kapcsolat meglétére gondolhatunk. A kétféle gyakoriság eltérése közötti különbségeket egy mérőszámban kell kifejezni.

Az ún. peremgyakoriságok segítségével kiszámíthatjuk a függetlenség esetén feltételezett gyakoriságokat, amelyeket *-gal különböztetünk meg:

$$\frac{S_1 \times O_1}{n} = f_{11}^* \qquad \frac{S_1 \times O_2}{n} = f_{12}^*$$

$$\frac{S_2 \times O_1}{n} = f_{21}^* \qquad \frac{S_2 \times O_2}{n} = f_{22}^*$$

Az előző példa adatai alapján készítsük el a további számításokat!

A peremgyakoriságok segítségével a függetlenség esetén feltételezett gyakoriságok:

$\frac{48\times50}{80}=30 \quad \frac{48\times30}{80}=18$ stb.

A feltételezett, független gyakoriságokat az eredeti táblához hasonlóan foglalhatjuk össze:

16.4. táblázat - A függetlenség esetén feltételezett gyakoriságok

Játékhely	A vizsgált csapat eredménye		Összesen:
Győzelem	Vereség		
Otthon	30	18	48
Idegenben	20	12	32
Összesen:	50	30	80

Forrás: Saját számítás

Ha a kiinduló és a fenti tábla belső adatait összehasonlítjuk, látjuk, hogy a gyakoriságok eltérnek egymástól, ezért feltételezhetjük a sztochasztikus kapcsolatot.

A tényleges és feltételezett gyakoriságok közötti eltéréseket egyetlen mutatószámba kell „sűríteni”, amihez az alábbi számítás segítségével jutunk el.

Elsőként minden cellában kiszámítjuk az alábbi relatív különbséget:

$$\frac{(f_{ij}-f_{ij}^{*})^2}{f_{ij}^{*}}$$

ahol: az f_{ij} az i-edik sorának és j-edik oszlopának gyakorisága.

Az eltérésekből képzett összeg (valamennyi cellát figyelembe véve, amit a dupla szummázás jelöl!) a (Khi-négyzet) néven ismert matematikai–statisztikai eloszlás értéke.

$$\chi^2=\sum\sum\frac{(f_{ij}-f_{ij}^{*})^2}{f_{ij}^{*}}$$

A önmagában még nem felel meg a sztochasztikus kapcsolatok mérőszámaival szemben megfogalmazott feltételnek. Alsó határa ugyan 0, de felső határa jelentősen meghaladhatja az 1-et. Ezt a dilemmát oldja fel a Cramer-féle mutatószám, amelynek képlete:

$$C=\sqrt{\frac{\chi^2}{n\times(s-1)}}$$

Ahol a tört nevezőjében az s a két változó ismérvváltozatainak minimumát (a kevesebb ismérvváltozat számát) jelöli. (Ez alternatív ismérvek esetén nem tér el, mindkét ismérv esetében kettő.)

Természetesen a feltételezett gyakoriságok kiszámítása kettőnél több ismérvváltozatra is kiterjeszthető, így a Cramer-mutató kiszámításának lehetősége általánosan adott.

A Cramer-féle mutató eleget tesz a sztochasztikus kapcsolati mérőszámokkal szemben támasztott követelménynek is, mivel:

$0 \le C \le 1$

Folytassuk a számításokat bemutató példánk adataival!

$$\chi^2=\frac{(39-30)^2}{30}+\frac{(9-18)^2}{18}+\frac{(11-20)^2}{20}+\frac{(21-12)^2}{12}=18$$

A Cramer-féle mutatószám példánkban:

$$C=\sqrt{\frac{18}{80\times(2-1)}}=0{,}47$$

A mérőszám szerint a játékhely jellege és a csapatok teljesítménye közötti sztochasztikus kapcsolat közepesnek mondható. A C2 mérőszám is értelmezhető, amely azt mutatja meg, hogy – esetünkben – a játékhely mintegy 23%-ban ($0{,}47^2 = 0{,}23$) determinálja a labdarúgócsapat teljesítményét.

A korábbi Yule-féle mérőszámnál most alacsonyabb intenzitású kapcsolatot számszerűsítettünk. A kétféle mérőszám eredményét egymással nem lehet összemérni, a *Cramer-együttható* „szigorúbban” mér. Előnye azonban az utóbbinak, hogy *nemcsak alternatív ismérvek esetén* használható.

Az asszociációs kapcsolat mérését szemlélteti az alábbi példa, felhasználva a Cramer-féle mutató előnyeit. Itt ugyanis egyik minőségi ismérv nem alternatív.

Tételezzük fel, hogy a teljesítmény szerint vizsgált kapcsolatot kiterjesztették a döntetlenre is. A vizsgálat eredményét a 35. tábla tartalmazza:

16.5. táblázat - A csapat eredményei

Játékhely	**A vizsgált csapat eredménye**			**Összesen**
Győzelem	**Vereség**	**Döntetlen**		
Város	30	9	9	48
Község	6	15	11	32
Összesen:	36	24	20	80

Forrás: Saját számítás

A függetlenséget feltételező gyakoriságok:

16.6. táblázat - Gyakoriságok függetlenség esetén

Játékhely	**A vizsgált csapat eredménye**			**Összesen**
Győzelem	**Vereség**	**Döntetlen**		
Otthon	21,6	14,4	12	48
Idegenben	14,4	9,6	8	32
Összesen:	36	24	20	80

Forrás: Saját számítás

Ahol pl. $f_{11}^{*} = \frac{48 \times 36}{80} = 21{,}6$

A χ^2-eloszlás értékét a 37. tábla segítségével határozhatjuk meg:

16.7. táblázat - Munkatábla

Játékhely	**A vizsgált csapat eredménye**			**Összesen**
Győzelem	**Vereség**	**Döntetlen**		
Otthon	3,266	2,025	0,75	6,041
Idegenben	4,9	3,037	1,125	9,062
Összesen:	8,166	5,062	1,875	15,103

Forrás: Saját számítás

Pl. $\frac{(9-12)^2}{12} = 0{,}75$

$\chi^2 = 15{,}103$

$C^2 = \frac{15{,}103}{80 \times (2-1)} = 0{,}0944$

$C = \sqrt{0{,}0944} = 0{,}307$

A játékhely meghatározó képessége mintegy 9,4%-os, az ismérvek közötti kapcsolat gyengének mondható.

2. Vegyes kapcsolat

Mint korábban már utaltunk rá, a vegyes kapcsolatokban az ok szerepét mindig a minőségi ismérv tölti be, míg az okozat(ok)ét a mennyiségi ismérv(ek). A mennyiségi ismérv lehetőséget teremt arra, hogy a számítási eljárásokat kibővítsük, így például többirányú szórásszámítás segítségével számszerűsítsük a kapcsolatot. Itt elsősorban a szórás felbontásának összefüggésére tudunk építeni. Korábbi tanulmányainkból tudjuk, hogy – heterogén sokaság esetén – a sokaság szórásnégyzete egyenlő a belső és külső szórásnégyzet összegével:

$\sigma^2 = \sigma^2_B + \sigma^2_K$

Ismerjük továbbá a belső és külső szórásnégyzet kiszámításának módját.

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

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

A szórásnégyzetek összefüggését kifejező képletet átrendezhetjük, ha elosztjuk az egyenlőség mindkét oldalát a teljes szórásnégyzettel:

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

A csoportosító (minőségi) ismérv – ami egyben a sztochasztikus kapcsolat ok szerepét is betölti – hatását a külső szórás közvetíti. Könnyen belátható, hogy amennyiben a külső szórás nulla, a minőségi ismérvnek semmilyen mérhető hatása nincs; a két ismérv (a minőségi és mennyiségi) független. Az ellenkező szélsőséges esetben – amennyiben a belső szórás nulla – a külső szórás megegyezik a teljes szórással, tehát a kapcsolat determinisztikus. Ezek alapján a külső és a teljes szórás segítségével számszerűsíthető a vegyes kapcsolatot mérő **szóráshányados** mutatója:

$$H = \frac{\sigma_K}{\sigma}$$

A fenti összefüggéseket világítsuk meg egy példával!

Tegyük fel, hogy egy evezős szakosztály versenyzői részére új kétpárevezős hajót kíván beszerezni. A döntés előtt három típust vizsgáltak meg. Az egyes típusok (minőségi ismérv) és a hajókkal elért eredmény (mennyiségi ismérv) között sztochasztikus összefüggést lehet számszerűsíteni. Hajótípusonként elért eredményeket tartalmazza a 16-8. tábla.

16.8. táblázat - A különböző hajótípusokkal elért eredmények (perc)

Sorszám	Típusok		
A	B	C	
1.	10,1	8,0	12,1
2.	9,5	8,2	9,1
3.	8,0	7,1	10,0
4.	9,2	6,9	–
5.	9,0	–	–
Átlag	9,16	7,55	10,4
Szórás	0,689	0,559	1,257

Forrás: Saját számítás

A táblából jól látható, hogy típusonként az átlagok szembetűnően eltérnek, ami a sztochasztikus kapcsolat meglétére utal.

A csoportonkénti átlagok segítségével (is) kiszámíthatjuk a 12 mérés alapján a főátlagot:
$$\bar{x} = \frac{5 \times 9{,}16 + 4 \times 7{,}55 + 3 \times 10{,}4}{12} = 8{,}93 \text{ perc}$$

A belső szórásnégyzet és a belső szórás a csoportszórásokat felhasználva az alábbi módon határozható meg:
$$\sigma_B^2 = \frac{5 \times 0{,}689^2 + 4 \times 0{,}559^2 + 3 \times 1{,}257^2}{12} = 0{,}6969$$
$$\sigma_B = \sqrt{0{,}6969} = 0{,}835$$

A külső szórásnégyzet és szórás:
$$\sigma_K^2 = \frac{5 \times (9{,}16 - 8{,}93)^2 + 4 \times (7{,}55 + 8{,}93)^2 + 3 \times (10{,}4 - 8{,}93)^2}{12} = 1{,}1971$$
$$\sigma_K = \sqrt{1{,}1971} = 1{,}094$$

A teljes sokaságra vonatkozó szórásnégyzet és szórás (az additív összefüggést felhasználva):
$$\sigma^2 = 0{,}6969 + 1{,}1971 = 1{,}894$$
$$\sigma = \sqrt{1{,}894} = 1{,}376$$

A futási eredmények együttes (teljes) szórása: 1,376 perc.

A fenti adatok lehetővé teszik a kapcsolat szorosságának mérését a szóráshányados segítségével.
$$H^2 = \frac{1{,}1971}{1{,}894} = 0{,}632$$
$$H = \frac{1{,}094}{1{,}376} = 0{,}795$$

A számítások alapján azt mondhatjuk, hogy a sporthajók típusa számottevően meghatározza a velük elérhető eredményeket, mivel a kapcsolat szoros (0,795) és a típus mintegy 63,2%-ban befolyásolja az eredmények szóródását.

A 15. fejezetben a csoportosított átlagok és szórások számítása során megismerkedtünk egy példával. Itt most lehetőségünk van ennek a példának a folytatására.

A férfiak és nők napi edzési időtartamát vizsgálva megállapítottuk, hogy az együttes szórás 1,3 óra/nap, míg a csoportátlagok a főátlagtól 0,49 óra/nap nagyságrenddel térnek el átlagosan. Kiszámíthatjuk a nemek és az edzés időtartama közötti kapcsolat szorosságát.
$$H = \frac{0{,}49}{1{,}3} = 0{,}377$$

Tehát a nemek csak gyenge mértékben befolyásolják a sportolás napi időtartamát.

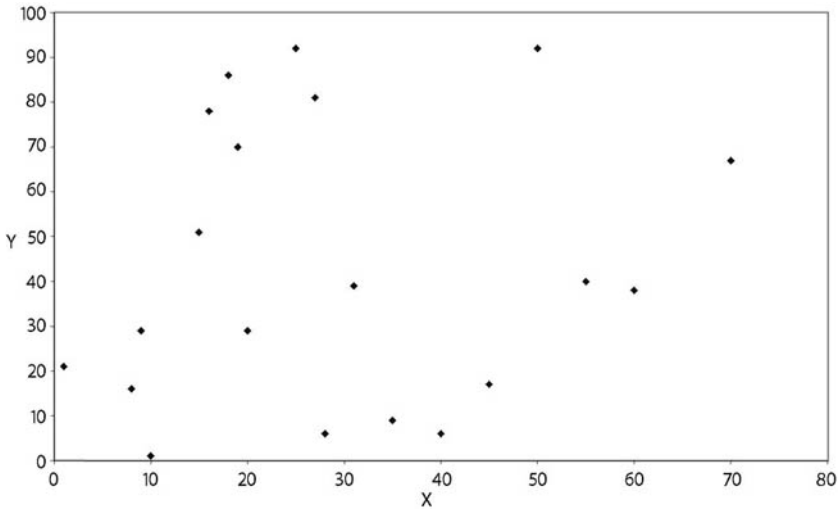
3. Korrelációs kapcsolat

Amennyiben mind az ok(ok), mind az okozat szerepét is mennyiségi ismérvek közvetítik, **korrelációs kapcsolat**ról beszélünk. Ebben a fejezetben csak nagyon röviden, jelzésszerűen érintjük a témakört. Itt is utalni szeretnék arra, hogy a valóságban általában nem egy, hanem több tényező együttes igen összetett hatására alakul ki egy-egy jelenség, folyamat. A korrelációs kapcsolat mérése során azonban több ok együttes hatásának vizsgálatát is viszonylag könnyen meg lehet oldani.

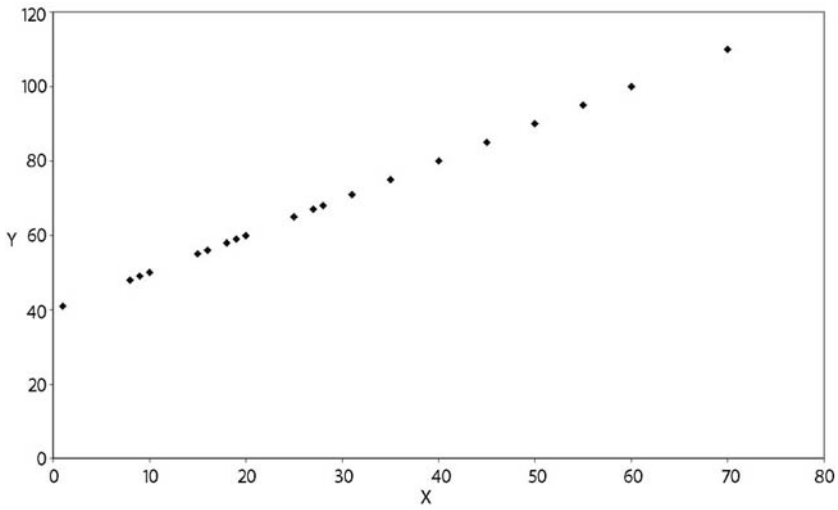
A továbbiakban elsősorban egy tényező- és egy eredményváltozó közötti kapcsolat mérését mutatjuk be, mindezzel csupán egy vázlatos bepillantást adunk a korrelációs kapcsolatok vizsgálatának gazdag módszertanába.

Két mennyiségi ismérv között meglévő kapcsolatot jól ábrázolhatjuk a derékszögű koordinátarendszerben, ún. pontdiagram segítségével. A kapcsolattípusok az alábbiak lehetnek:

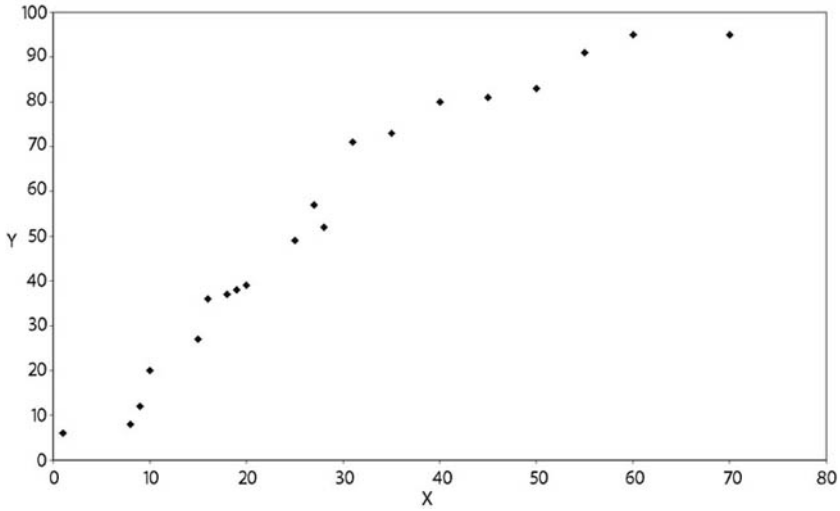
16.1. ábra - Korrelálatlanság (függetlenség)



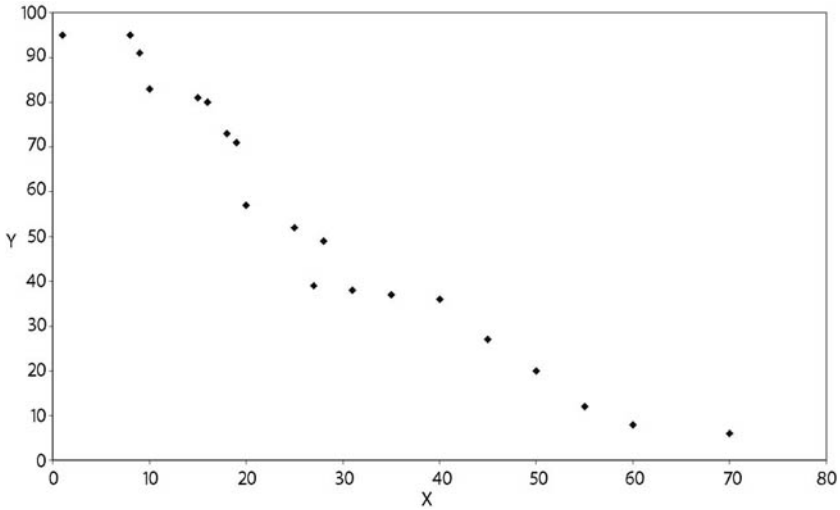
16.2. ábra - Determinisztikus kapcsolat



16.3. ábra - Pozitív korreláció



16.4. ábra - Negatív korreláció



A kétváltozós korrelációs kapcsolat lehet lineáris és görbevonalú.

A korrelációs kapcsolat mérésének legelterjedtebb mutatószáma a **lineáris korrelációs együttható** (jele**: r**), amelynek alkalmazása során feltételezzük a változók közötti lineáris kapcsolatot (ami a valóságban nem mindig teljesül). Abban az esetben azonban, ha a linearitás feltevése nem áll távol a vizsgált problémától, első megközelítésben hasznos mérőeszköze lehet a korrelációs kapcsolatnak. A korrelációs együttható kiszámítását az alábbi algoritmus segítségével végezhetjük el:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \times \sigma_x \times \sigma_y}$$

ahol: σ_x és σ_y a változók szórásai.

Az alábbi példa a korrelációs kapcsolat szorosságának mérését hivatott bemutatni.

Egy egyesület sportlövészklubjában a női skeetlövők körében elemezték a heti edzésidő és az egy adott napon elért eredmény közötti kapcsolatot. A sztochasztikus kapcsolatban (itt korrelációs kapcsolatban!) a tényezőváltozó (X) szerepét a heti edzésidő, míg az eredményváltozó (Y) szerepét az elért pontszám töltötte be. A 10 versenyzőre vonatkozó alapadatokat és a számítások részeredményeit az alábbi táblázat tartalmazza:

16.9. táblázat - A korrelációs együttható számításának

munkatáblája

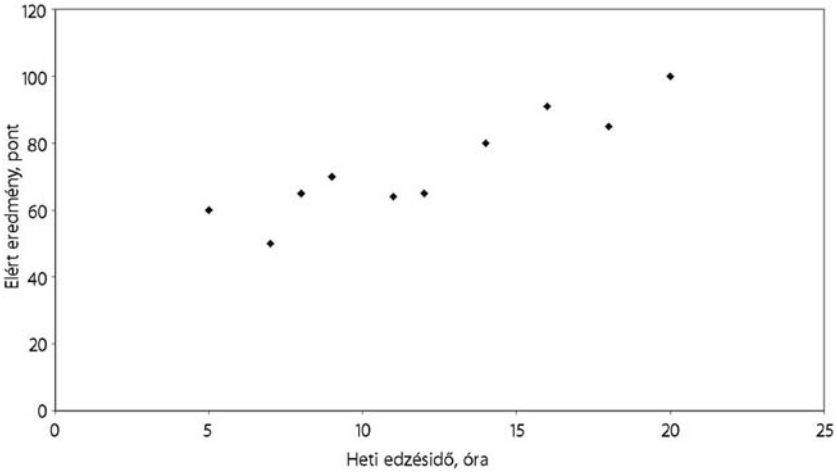
Sorszám	Heti edzésidő (óra)	Elért teljesítmény (pont)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$			
1.	5	60	–7	–13	49	169	91			
2.	7	50	–5	–23	25	529	115			
3.	8	65	–4	–8	16	64	32			
4.	9	70	–3	–3	9	9	9			
5.	11	64	–1	–9	1	81	9			
6.	12	65	0	–8	0	64	0			
7.	14	80	2	7	4	49	14			
8.	16	91	4	18	16	324	72			
9.	18	85	6	12	36	144	72			
10.	20	100	8	27	64	729	216			
Összesen:	120	730	0	0	220	2162	630			

Forrás: saját számítás

Határozzuk meg az edzésidő és az elért eredmény közötti kapcsolatot mérő lineáris korrelációs együtthatót!

A mért adatokat pontdiagrammal ábrázoltuk az alábbi 15. ábrán:

16.5. ábra - Az eredmények grafikus megjelenítése



Forrás: Saját szerkesztés

A 15. ábrából megállapítható, hogy a pontok egy egyenes mentén szóródnak.

A tábla összeállítása során fel kellett használni az egyes változók átlagait:

$$\bar{x} = \frac{120}{10} = 12 \text{ óra} \quad \bar{y} = \frac{730}{10} = 73 \text{ pont}$$

A fenti tábla adatai lehetőséget adnak a változók szórásának meghatározására:

$$\sigma_x = \sqrt{\frac{220}{10}} = 4{,}69 \text{ óra} \quad \sigma_y = \sqrt{\frac{2162}{10}} = 14{,}7 \text{ pont}$$

A lineáris korrelációs együttható:

$$r = \frac{630}{10 \times 4{,}69 \times 14{,}7} = 0{,}9135$$

A számítások azt igazolják, hogy igen erős az edzésidő és az elért eredmények közötti korrelációs kapcsolat. Az edzésidő növelése nagy valószínűséggel a teljesített pontszámok növekedését vonja maga után, a változók pozitív kapcsolatban vannak egymással.

Természetesen a mennyiségi ismérvek lehetőséget adnak arra is, hogy az elemzési eszköztárunkat kibővítsük, ne csak a kapcsolat szorosságára koncentráljunk. Közvetlenül adódik annak a lehetősége, hogy a változók közötti kapcsolat intenzitásán túl az összefüggés természetét is modellezzük, matematikailag kezelhető formába öntsük.

A kapcsolat törvényszerűségét az ún**. regresszióanalízis**sel elemezhetjük, és közvetlenül a **regresszió függvény**ek segítségével írhatjuk le.

Mint már szóltunk róla, a változók közötti kapcsolat a gyakorlatban sokszor nem lineáris. Ilyen esetben mind a szorosság mérésének, mind a kapcsolat törvényszerűségét felíró matematikai modellnek a felépítése viszonylag bonyolult matematikai–statisztikai eljárásokat igényel.

Amennyiben a változók között lineáris sztochasztikus kapcsolatot tételezünk fel, egy viszonylag egyszerű matematikai modellel, egy lineáris függvény (egyenes) paramétereinek meghatározásával jól felhasználható regresszió függvényt számszerűsíthetünk. Az egyenes konstans paramétereinek becslését az ún. legkisebb négyzetek módszerével[2] végezhetjük el. Itt jegyezzük meg, hogy nagyon sok olyan számítástechnikai szoftver ismert, amely a regressziós paraméterek meghatározását gyorsan és pontosan elvégzi, és csak az elemző munka vár a felhasználóra.

A két paramétert egyszerűen meghatározhatjuk az alábbi képletek segítségével:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

A tényezőváltozó paraméterének **(b_1)** kiemelt a szerepe, **regressziós együttható**nak nevezi a statisztika, segítségével közelebb kerülhetünk a kapcsolat törvényszerűségének megértéséhez.

Előző példánk adatai alapján nézzük a regresszió függvény kiszámítását!

A számításhoz szükséges adatok korábbról ismertek:

$$b_1 = \frac{630}{220} = 2{,}86$$

$b_0 = 73 - 2{,}86 \times 12 = 38{,}68.$

A regressziós egyenes egyenlete:

$\hat{y} = 38{,}68 + 2{,}86x.$

A regressziós együttható ismeretében azt mondhatjuk, hogy a heti edzésidő egységnyi (egy órányi) növekedése (többlete) várhatóan átlagosan 2,86 ponttal növeli a sportolók elért eredményét.

A regresszió függvény paramétereinek ismeretében természetesen egy adott, tetszőleges x érték függvényében meghatározhatjuk az y érték várható nagyságát, aminek bekövetkezési esélye annál nagyobb, minél erősebb a változók közötti korrelációs kapcsolat.

[2] A becslési módszer elvi leírásától tananyagunkban eltekintünk, csupán a módszer alkalmazásával nyert megoldóképleteket használjuk fel.

Becsüljük meg egy heti 10 órás edzési időt teljesítő versenyző modellünk szerint elérhető pontszámát!

$\hat{y}$ = 38,68 + 2,86 x 10 = 67,28 pont.

A regressziós függvény segítségével meghatározhatjuk a regresszió értékeit, amelyek a megfigyelt X értékhez rendelhető becsült Y értékek.

40. táblázat:

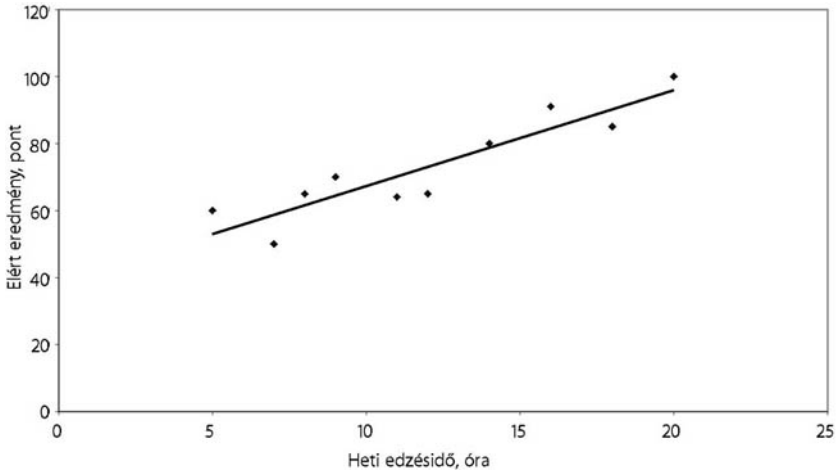
16.10. táblázat - A tényleges és a regresszióval becsült pontszámok

Heti edzésidő óra	Pontszám (Tényleges)	Pontszám (Becsült)
5	60	52,95
7	50	58,68
8	65	61,55
9	70	64,41
11	64	70,14
12	65	73,00
14	80	78,73
16	91	84,45
18	85	90,18
20	100	95,91

Forrás: saját számítás

A tényleges adatokat és a becsült regresszióértékeket ábrázoljuk a 16. ábrán.

16.6. ábra - A tényleges és a regresszióértékek ábrája



Forrás: Saját szerkesztés

A regressziós együttható ismerete lehetővé teszi, hogy lineáris összefüggés esetén is kvantifikáljuk az **elaszticitás**t (a **rugalmasság**ot), amely a változás relatív (százalékban kifejezett) mértékét fejezi ki. Az átlagos elaszticitás a változók átlagai segítségével az alábbi módon határozható meg:

$$El = b_1 \frac{\bar{x}}{\bar{y}}$$

Általános szabályként elmondhatjuk, hogy az elaszticitási mutatószám 1-nél nagyobb

értéke a változók közötti kapcsolat rugalmasságára utal, míg az 1-nél kisebb érték a rugalmatlanságnak a jelzője.

Az előbbi példában is kiszámíthatjuk az átlagok környezetében az elaszticitás mérőszámát:

$$El = 2{,}86\frac{12}{73} = 0{,}47$$

A heti edzésidő 1%-os növekedése átlagosan az elért eredmények (pontszámok) 0,47%-os növekedését vonja maga után. Belátható, hogy itt elég rugalmatlan kapcsolatot számszerűsítettünk.

A korrelációs kapcsolat – a mennyiségi ismérvek természete miatt – lehetőséget ad a kapcsolat összetett jellegének a vizsgálatához, további tényezőváltozók szerepeltetését is lehetővé teszi.

Több változó együttes hatását is mérhetjük az ún. többszörös korrelációs együttható segítségével. Egy vizsgálatba további változó (faktor) bevonása természetesen a kapcsolat szorosságának erősödésében is kifejezésre jut. Ugyancsak gyakran használják a jelenségek komplex elemzése érdekében a többváltozós regressziós modelleket, amelyekben egy eredményváltozót több tényezőváltozó segítségével magyarázunk. A többváltozós analízis egy árnyaltabb kép megrajzolását teszi lehetővé.

4. Ellenőrző feladatok, gyakorló példák a fejezethez

- A következő táblázat a súlyemelés súlycsoportonkénti világcsúcsait tartalmazza (2006. 02. 27-én).
  - Határozza meg a férfiak összetett eredményének és a súlycsoportoknak a kapcsolatát!
  - Számítsa ki, milyen szoros a kapcsolat köztük!
- A következő táblázat egy tényleges kutatás adatait tartalmazza, amelyben azt kérdezték, hogy a megkérdezett sportol-e. Az összegyüjtött eredményeket a következő táblázat közli.
  - Milyen erős kapcsolat van a sportolás és a lakóhely között?

17. fejezet - Idősorok elemzése

A különféle jelenségek és folyamatok számszerűsíthető értékei igen kedvező lehetőséget teremtenek az időbeni összehasonlításokra, összemérésekre, az időbeli változások vizsgálatára. Mindezek módot teremtenek arra is, hogy az elmúlt időszakok tendenciájának, összefüggéseinek feltárásával jobban megismerjük a jelenségek, folyamatok természetét és ezek egyben alapul szolgálnak a jövő várható folyamatainak előrelátásához.

Az idősor fogalmával már korábban megismerkedtünk. Általános sémája az alábbi:

17.1. táblázat - Az idősorok általános sémája

Idő (t^a)	Idősor értéke (y_t)
1	y_1
2	y_2
3	y_3
.	.
.	.
.	.
n	y_n

[a]A t a latin tempus (idő) szó kezdőbetűje.

Forrás: saját szerkesztés

Az idősor értékei tapasztalati adatokból épülnek fel, amelyek egyetlen realizációi egy általánosan definiálható elméleti idősornak.

Elméleti idősornak tekinthetjük például egy sportoló súlyának növekedését. Amennyiben a súlyt hetente megmérjük, és az adatokat feljegyezzük, konkrét tapasztalati idősort kapunk, amely alkalmas további számítások elvégzésére.

Az idősorban fellelhető változást, fejlődést több tényező együttes hatása alakítja ki. Az idősorok elemzésének egyik feladata, hogy az egyes összetevők (tényezők, komponensek) hatását elkülönítve számszerűsítse.

A klasszikus idősorelemzés abból a feltételezésből indul ki, hogy az idősort egy tartós, **hosszú távú tendencia (trend)**, szabályos **hullámmozgások, periodikus ingadozások (szezonalitás)** határozzák meg és ezektől eseti, egyenként nem jelentős eltérítő hatást vált ki a **véletlen ingadozás**.

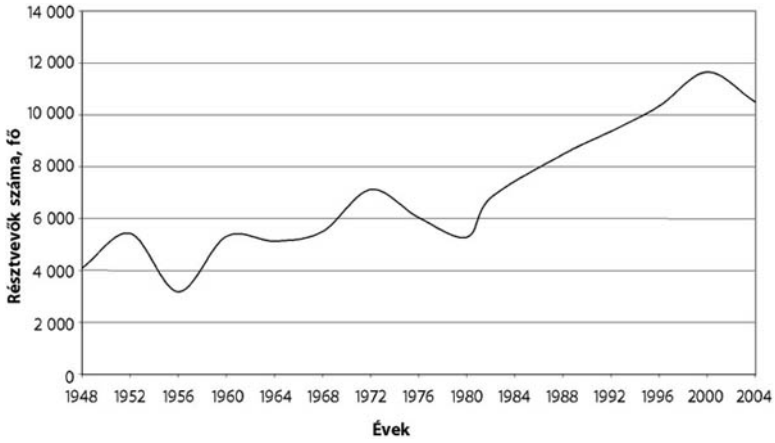
1. Az idősorelemzés egyszerűbb eszközei

A tapasztalati idősor már első ránézésre is alkalmas bizonyos főbb következtetések levonására.

A grafikus ábrázolás (az idősorok megjelenítése) lehetővé teszi a fő tendenciák, vonások felismerését, ezáltal is segíti az elemzés folyamatát.

A második világháború után megrendezett nyári olimpiai játékokon részt vevő versenyzőkről tájékozódhatunk a 17. ábra segítségével.

17.1. ábra - Az olimpiákon (1948–2004) indult versenyzők száma (fő)



Forrás: Saját szerkesztés

A grafikus ábrázolás szemlélteti az idősort, így megkönnyíti számunkra az elemzést. Leolvashatjuk az ábráról, hogy tendenciáját tekintve növekvő részt vevő szám mellett, két esetben, 1956-ban Melbourne-ben és 1980-ban Moszkvában volt jelentősebb visszaesés az olimpiai résztvevők létszámában. Láthatjuk, hogy Atlanta (1996) és Sydney (2000) esetében nagyon magas értéket ért el a játékon részt vevő sportolók száma, majd 2004-ben, Athénben kissé stagnáló jelleget öltött.

A **bázis- és láncviszonyszámok**, mint egyszerűen meghatározható mutatószámok, jól alkalmazhatók az idősorok gyors, előzetes elemzésére.

Az egyszerűbb elemzési eszközök fontos csoportját alkotják a **speciális átlagok**.

Ezek közül igen fontos információkat szolgáltat a **tartamidősorok** gyors vizsgálata során a **számtani átlag**.

$$\bar{y}_t = \frac{\sum_{t=1}^{n} y_t}{n}$$

17.2. táblázat - A nyári olimpiákon részt vevő versenyzők száma

Helyszín	Év	Résztvevők száma, fő
London	1948	4 092
Helsinki	1952	5 429
Melbourne	1956	3 178
Róma	1960	5 313
Tokió	1964	5 133
Mexikóváros	1968	5 498
München	1972	7 121
Montreál	1976	6 043
Moszkva	1980	5 283
Los Angeles	1982	6 802
Szöul	1988	8 473
Barcelona	1992	9 368
Atlanta	1996	10 332
Sidney	2000	11 651

Helyszín	Év	Résztvevők száma, fő
Athén	2004	10 500

Forrás: http://www.szalax.freeweb.hu

Az átlagos részvételi létszám:

$$\bar{y}_t = \frac{104216}{15} = 6947{,}73 \approx 6948 \text{ fő.}$$

A tartamidősorok[1] adatainak egyszerű összegzésével is viszonylag jó áttekintést kapunk az egész idősorról, így az átlagszám értelmezése is kézenfekvő.

Az **állapotidősor** átlagos értékének meghatározáshoz azonban egy sajátos átlagot, az ún. **kronologikus átlagot** kell használni.

Az idősorok értékei általában gyakran vagy folyamatosan változnak. Az állapotidősorok értékei egy-egy időpontra vonatkoznak. Természetesen az idősorról annál pontosabb képet kaphatnánk, minél gyakrabban ismétlődnek a megfigyelések. Ennek viszont gyakran korlátot szabnak szervezési, költségtényezők. Így sok esetben meg kell elégednünk a hosszabb időszakonként mért értékek ismeretével. Jellemző példák erre a készletek, valamint a létszám-adatok idősorai.

Az elmondottakból következik, hogy egy adott időszak (például egy év) korrekt jellemzéséhez a vizsgált időszakon (éven) kívüli megfigyelés is szükséges, de az első és utolsó megfigyelés – természetéből adódóan – csak fél súllyal szerepel. Mindezek alapján a kronologikus átlag képlete:

$$\bar{y}_t^{\,k} = \frac{\frac{y_1}{2} + y_2 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1}$$

Egy cég alkalmazottainak létszámadatait szemlélteti a 17-2. tábla.

43. táblázat: Egy sportáruház alkalmazottainak minden hó első napján mért létszámadata egy adott évben

Forrás: saját szerkesztés

Forrás: Saját szerkesztés

Az éves átlaglétszám:

$$\bar{y}_t^{\,k} = \frac{\frac{50}{2} + 55 + \ldots + 51 + \frac{48}{2}}{12} = \frac{640}{12} = 53{,}3333 \approx 53 \text{ fő.}$$

A havi átlaglétszámok az adatok átlagolási logikájának megfeleltetve:

január: (50 + 55)/2, február: (55 + 62)/2, március: (62 + 48)/2 stb. Amennyiben ezeket az értékeket átlagoljuk, algebrai egyszerűsítésre nyílik mód, ami a kronologikus átlag képletéhez vezet.

2. Az idősorok összetevői

Az idősorban rejlő változásokat, a jelenségek és folyamatok időbeli alakulását többféle tényező határozza meg. Ezek közül az összetevők közül a statisztikai elemzés általában három tényezőt szokott elkülöníteni.

- **Trend** vagy **alapirányzat**, amely főbb hatások eredményeként egy határozottan jelentkező tendencia, az idősor alakulásának fő iránya.

[1] Tartamidősornak tekinthetjük az olimpiai versenyeken résztvevő versenyzők számsorait, ahol a létszám kötődik a játékok időtartamához.

- **Periodikus ingadozás,** egy rendszeresen ismétlődő hullámmozgás. Amennyiben a hullámzás, a ritmikus mozgás állandó periódushosszúságú és a periódushossz egy év vagy annál rövidebb, **szezonális ingadozás**ról beszélünk (pl. kereskedelmi forgalom, idegenforgalom alakulása, elektromos energia fogyasztása). Az elnevezés eredetileg negyedéves hullámzásokat jelöl, azonban héten, vagy napon belüli hullámmozgásokra is érvényes. A negyedéves szezonalitás az évszakok változásaival és az ehhez kapcsolódó társadalmi szokásokkal van kapcsolatban. A változó periódushosszú ingadozások egyik jellemző fajtája a **konjukturális ingadozás**.

- A **véletlen ingadozás** az idősorban fellelhető olyan szabálytalan mozgás, amely nem mutat semmilyen szisztematikusságot. A véletlen felfogható sok, egyenként nem jelentős, egymás hatását erősítő vagy gyengítő tényezők végső eredőjeként.

Az idősor elemzésének hagyományos feladata az egyes komponensek (trend, szezonalitás és véletlen) hatásának elkülönítése. A komponensekre bontást az ún. dekompozíciós módszerek segítségével végezhetjük el.

Az egyes komponensek kapcsolódását alapvetően kétféle modell segítségével képzelhetjük el.

2.1. Additív kapcsolat

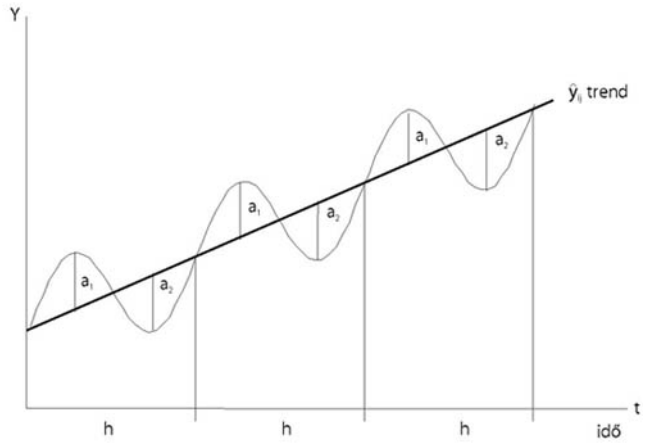
A különböző tényezők összegszerűen kapcsolódnak:

$y_{ij} = \hat{y}_{ij} + s^*_j + v^*_{ij}$

ahol: y_{ij} – a megfigyelt idősor értéke, $\hat{y}_{ij}$ – a trendérték, s^*_j – a szezonális eltérés, v^*_{ij} – a véletlen hatás, i – a periódusok (pl. évek) száma (1, 2, ..., n), j – a perióduson belüli időszakok, azaz a szezonok (pl. hónapok, negyedévek) száma (1, 2, ..., m).

18. ábra:

17.2. ábra - Az additív modell



Forrás: saját szerkesztés

ahol: h – a hullámhossz, a_1 – a periódusonkénti legnagyobb értéknél mért amplitúdó, a_2 – a periódusonkénti legkisebb értéknél mért amplitúdó.

Az egyes periódusokban mért amplitúdók egymással megegyeznek.

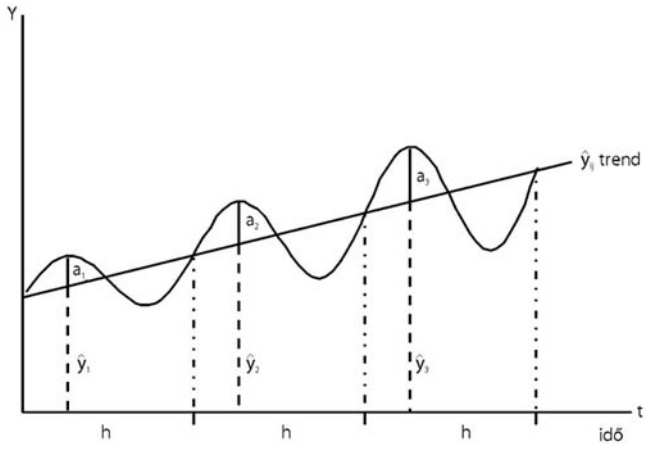
2.2. Multiplikatív kapcsolat

A tényleges idősori érték a három komponens szorzataként jön létre.

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

ahol a már ismert jelölések mellett s_j – a j-edik szezonhoz tartozó szezonális komponens, a szezonindex.

17.3. ábra - A multiplikatív modell



Forrás: Saját szerkesztés

Az 19 ábrából megállapíthatjuk, hogy a multiplikatív modell esetén az amplitúdók (a_1, a_2, a_3) periódusonként egymástól eltérnek, nem azonosak.

Összefoglalóan megállapítható, hogy a szezonalitás eltérítő hatása a megfelelő szezonoknál additív modellben *abszolút* állandóságot, multiplikatív modellben a trendhez mért *relatív* állandóságot mutat.

Gyakorlati tapasztalatok azt mutatják, hogy a társadalmi és gazdasági jelenségek idősorainál a komponensek legtöbbször multiplikatív módon kapcsolódnak egymáshoz.

A továbbiakban az egyes komponensek számszerűsítését mutatjuk be.

3. Trendelemzés

Az idősorban folyamatosan érvényesülő tendencia, a trend meghatározása – a dekompozíciós eljárás szerint – igényli, hogy a szezonális tényező és a véletlen komponens hatását kiszűrjük. A trend meghatározása az idősor kisimítását jelenti. A trendszámításnak két fő módszere terjedt el: a **mozgó átlagok módszere** és az **analitikus trendszámítás**.

3.1. A mozgó átlagok módszere

A mozgó átlagok módszere a trendet az idősor speciális, dinamikus átlagaként állítja elő. A véletlen tényező tompítását az átlagolás segítségével lehet megvalósítani, ugyanakkor a szezonális hatás kiszűrését is megoldja az átlagszámítás a tagszám megfelelő megválasztásával.

A számítás során meghatározzuk az átlagolandó értékek tagszámát (k) és vesszük az idősor első k értékének átlagát. Ez az érték lesz a trendadat, amelyet az átlagolt időszak közepéhez rendelünk. A következő lépésben az előzőeket oly módon ismételjük meg, hogy az első figyelembe vett adatot elhagyjuk, és helyette vesszük az idősor következő értékét. A műveleteket addig ismételjük, amíg az idősor utolsó értékét is felhasználjuk. A mozgóátlagolás során az idősor elejéhez és végéhez nem képződnek átlagok, így a trendadatok száma kevesebb lesz, mint az idősor adatainak száma.[2]

Amennyiben a k szám páros, a mozgóátlagolással kapott értékek az idősor két-két

[2]Annak ellenére, hogy nagyobb k alkalmazása az idősor rövidülését is jelenti, hosszabb idősor esetén célszerű nagyobb tagszámmal számolni, mert az idősor alaptendenciája biztosabban kerül kimutatásra.

időszaka, illetve időpontja közé kerülnek. Ilyen esetben egy ismételt $k = 2$ tagú mozgóátlagolást kell elvégezni, hogy a trendértékek megfeleltethetőek legyenek az idősor adatainak. Ezt középre igazításnak, ún. *centrírozásnak* nevezzük.

A szezonalitást tartalmazó idősorok esetében a mozgóátlagolás k tagszámának megválasztása során figyelembe kell venni azt, hogy a tagszám megegyezzen a perióduson belüli részidőszakok, a szezonok számával (m-mel), vagy annak egész számú többszöröse legyen. Így biztosíthatjuk, hogy valamennyi trendadat előállításában minden szezon hatása megjelenjen.

A 44. táblázatban lévő idősor adatait használjuk fel a mozgóátlagolás módszerének bemutatására.

17.3. táblázat - A kereskedelmi szálláshelyek vendégforgalma Baranya megyében 2001 és 2003 között

Év	Negyedév	Vendégek száma, fő
2001	I.	36 523
	II.	91 210
	III.	92 678
	IV.	54 152
2002	I.	43 609
	II.	96 982
	III.	91 183
	IV.	50 133
2003	I.	37 842
	II.	94 691
	III.	87 981
	IV.	54 416

Forrás: Baranya Megyei Statisztikai Évkönyvek

17.4. táblázat - A mozgóátlagolás munkatáblája

Év	Negyedév	Vendégek száma, fő	4 tagú mozóátlag	Centrírozott értékek
2001	I.	36 523,0		
	II.	91 210,0	68 640,8	
	III.	92 678,0	70 412,3	69 526,5
	IV.	54 152,0	71 855,3	71 133,8
2002	I.	43 609,0	71 481,5	71 668,4
	II.	96 982,0	70 476,8	70 979,1
	III.	91 183,0	69 035,0	69 755,9
	IV.	50 133,0	68 462,4	68 748,6
2003	I.	37 842,0	67 661,8	68 062,0
	II.	94 691,0	68 732,5	68 197,1
	III.	87 981,0		
	IV.	54 416,0		

Év	Negyedév	Vendégek száma, fő	4 tagú mozóátlag	Centrírozott értékek
			79 029,3	

Forrás: Saját számítás

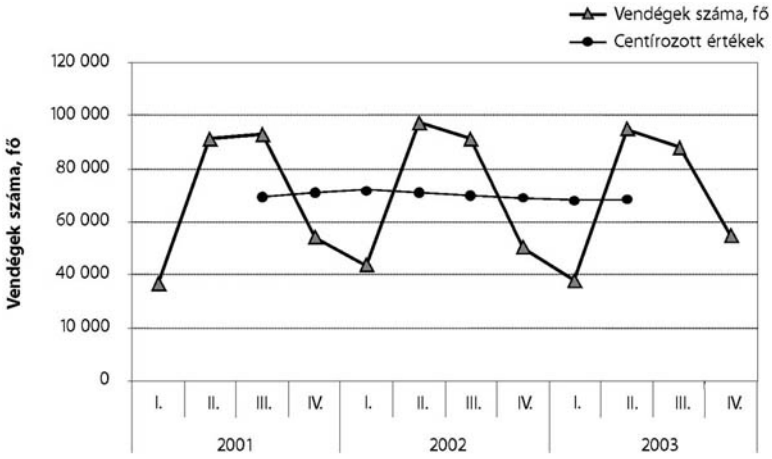
Az első 4 tagú átlag: (36 523 + 91 210 + 92 678 + 54 152)/4 = 68 640,8

A második 4 tagú átlag: (91 210 + 92 678 + 54 152 + 43 609)/4 = 70 412,3

Az első centrírozott (trend) érték: (68 640,8 + 70 412,3)/2 = 69 526,5 efő.

Az idősor tényleges adatait és a mozgóátlagolású trend adatait szemlélteti a 17-4. ábra, ahol a trendértékek a munkanélküliség tendenciájának kismértékű csökkenését jelzik.

17.4. ábra - A mozgóátlagolás grafikus ábrája



Forrás: Saját szerkesztés

3.2. Analitikus trendszámítás

Az analitikus trendszámítás módszerével az idősorban lévő alapirányzatot valamilyen ismert matematikai függvénnyel fejezzük ki. Itt a vizsgált jelenség, folyamat megfigyelt értékei (y_t) és az idő hatását kifejező (t) természetes számokból álló kapcsolatot modellezzük.

Ennél a módszernél elsőként azt kell megállapítani, hogy az alapirányzatot milyen függvény (egyenes vagy görbe) jellemzi a legjobban, majd ezután kerülhet sor a függvény paramétereinek megállapítására, becslésére.

A trend leírására általában a következő függvénytípusokat szokták a gyakorlatban alkalmazni:

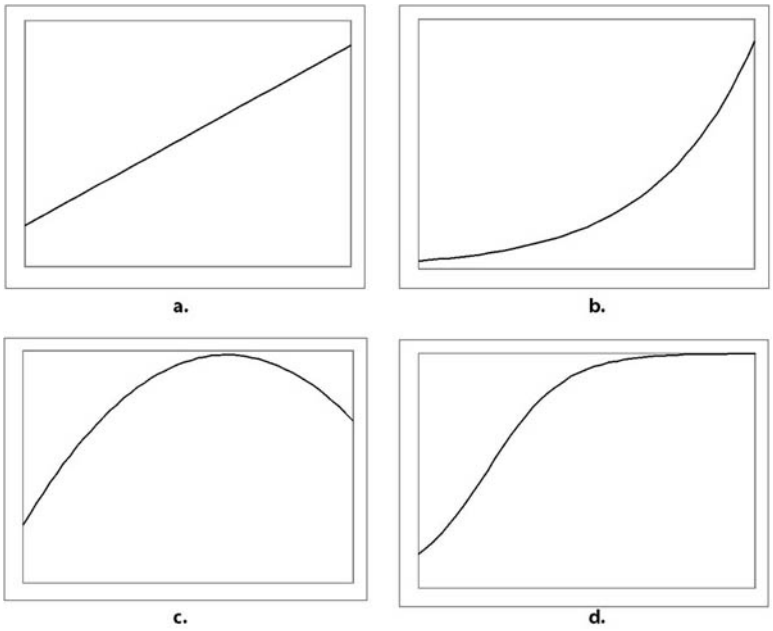
a. lineáris (egyenes) függvény,

b. exponenciális függvény,

c. másodfokú polinom,

d. logisztikus görbe.

17.5. ábra - A függvények képe sematikusan



Forrás: saját szerkesztés

A függvény típusát az idősor természete, illetve az idősor grafikus ábrája alapján lehet megválasztani.

Amennyiben az idősorban a szomszédos időszakok közötti változás (y_t – y_t – 1) – növekedés vagy csökkenés – állandóságot mutat, a **lineáris függvény**t alkalmazzuk.

Ha a változás relatív nagyságánál, üteménél az (y_t/y_t – 1) hányadosoknál tapasztalunk állandóságot – azaz a folyamat relatív (százalékos) változása állandó – **exponenciális függvény**t alkalmazunk.

Ha a folyamat változásának iránya megváltozik (pl. növekedés–stagnálás–csökkenés), illetve a változás nagysága nem állandó és a jelenség nagyságával sem arányos, a gyakorlatban általában **másodfokú polinom**ot használnak.

A **logisztikus függvény** (növekedési görbe) alkalmazása során az idősorban három fejlődési szakaszt különböztethetünk meg: az eleinte lassú növekedést erőteljes növekedés követi, amely később ismét lelassul.

Tananyagunkban csupán a lineáris trend meghatározásával foglalkozunk röviden.

A linearitás az idősorban azt jelenti, hogy egységnyi idő alatt a jelenség, folyamat azonos mértékben növekszik vagy csökken, tehát az abszolút változás állandó.

A lineáris trendfüggvény:

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

A trendfüggvény meghatározása a b0 és b1 paraméterek becslését jelenti az idősor adataiból. Erre a célra az ún. **legkisebb négyzetek módszere** révén nyerünk megoldást. Bizonyítás nélkül közöljük itt is a paraméterek meghatározására szolgáló képleteket:

$$b_1 = \frac{\sum (t - \bar{t})(y_t - \bar{y})}{\sum (t - \bar{t})^2} = \frac{\sum d_t d_y}{\sum d_t^2} \qquad b_0 = \bar{y} - b_1 \bar{t}$$

ahol: – a t időtényező (t = 1, 2, 3, ..., n) átlaga, $\hat{y}$ – az idősor értékeinek számtani átlaga.

A lineáris trendfüggvény b_1 paramétere az időszakonkénti állandó abszolút változást, másként az idősor átlagos abszolút változását mutatja. Értelmezése az egyenes

meredekségének felel meg.

Szemléltessük az elmondottakat az olimpiai résztvevők adatainak segítségével!

46. táblázat: Munkatábla a lineáris trend meghatározásához

17.5. táblázat - A nyári olimpiákon részt vevő versenyzők száma

Helyszín	Év	Résztvevők száma, fő	t	dt	dy	dt^2	dtdy
London	1948	4 092	1	–7,00	–2 855,73	49	19 990,13
Helsinki	1952	5 429	2	–6,00	–1 518,73	36	9 112,40
Melbourne	1956	3 178	3	–5,00	–3 769,73	25	18 848,67
Róma	1960	5 313	4	–4,00	–1 634,73	16	6 538,93
Tokió	1964	5 133	5	–3,00	–1 814,73	9	5 444,20
Mexikóváros	1968	5 498	6	–2,00	–1 449,73	4	2 899,47
München	1972	7 121	7	–1,00	173,27	1	–173,27
Montreál	1976	6 043	8	0,00	–904,73	0	0,00
Moszkva	1980	5 283	9	1,00	–1 664,73	1	–1 664,73
Los Angeles	1982	6 802	10	2,00	–145,73	4	–291,47
Szöul	1988	8 473	11	3,00	1 525,27	9	4 575,80
Barcelona	1992	9 368	12	4,00	2 420,27	16	9 681,07
Atlanta	1996	10 332	13	5,00	3 384,27	25	16 921,33
Sidney	2000	11 651	14	6,00	4 703,27	36	28 219,60
Athén	2004	10 500	15	7,00	3 552,27	49	24 865,87
Összesen:		104 216	120	0	0	280	144 968,00
Átlag:		6 947,73	8,00				

Forrás: Saját számítás

Az átlagok:

$$\bar{y} = \frac{104216}{15} = 6947{,}73$$

$$\bar{t} = \frac{120}{15} = 8$$

A paraméterek:

$$b_1 = \frac{144968}{280} = 517{,}74$$

$$b_0 = 6947{,}73 - 517{,}74 \times 8 = 2805{,}8$$

A lineáris trend egyenlete:

$\hat{y}_t$ = 2805,8 + 517,74 x t

A trendfüggvénybe behelyettesítve a *t* értékeket, megkapjuk a becsült trendértékeket:

$\hat{y}_1$ = 2805,8 + 517,74 x 1 = 3323,53

$\hat{y}_2$ = 2805,8 + 517,74 x 2 = 3841,28

.

.

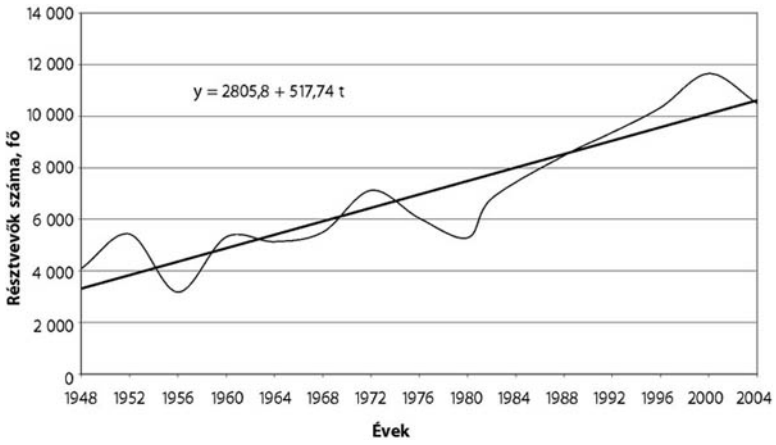
.

$\hat{y}_{15}$ = 2805,8 + 517,74 x 15 = 10 571,93.

Az idősor és a trend alakulását szemlélteti a 21. ábra.

21. ábra: A tényleges és a trendértékek ábrája

17.6. ábra - A függvények képe sematikusan



Forrás: Saját számítás

A legkisebb négyzetek módszerével illesztett trendfüggvény – az idősor tendenciájának pontos megismerésén túl – lehetőséget teremt a vizsgált időszakon túli előrebecslések elkészítésére is. Megfelelő *t* érték behelyettesítésével becsülhetjük a kívánt értéket.

A példa alapján a 2008. évi pekingi olimpia versenyzőinek várható száma:

$\hat{y}_{2008}$ = 2805,8 + 517,74 x 15 = 10 942,74 fő.

3.3. A szezonális hullámzás mérése

Az idősorok második fontos összetevője a periodikus ingadozás. Ez általában szabályos hullámzás, ezért elméletileg és gyakorlatilag is igen fontos jelenlétének és mértékének kimutatása, vizsgálata.

A szezonális hullámzás (idényszerű ingadozás) állandó periódushosszúságú ingadozás, ahol a periódus hossza egy év vagy annál rövidebb időszak.

A szezonalitás alapvetően bizonyos természeti jelenségekkel (pl. Föld forgása, meghatározott körforgása a Nap körül) magyarázható, amelyek napi, havi, évszakonkénti változásokban öltenek testet. Fontos tényezői a szezonalitásnak a társadalmi szokások, hagyományok, ünnepek és divat, amelyek jelentősebb nagyságú hullámzásokat okozhatnak életünk sok területén (pl. közlekedés, idegenforgalom, hírközlés, kereskedelem, távközlés).

A szezonalitás felismerése és megismerése kialakíthatja az emberekben a csillapítás (a szezonális hatások mérséklése) és az alkalmazkodás stratégiáját.

A szezonalitás megismerése és számszerűsítése során informálódunk arról, hogy a szezonalitás a periódus egyes szakaszaiban milyen *mértékben* vagy *arányban* téríti el az idősor értékét az alapirányzattól, a trendtől.

A szezontényező meghatározása feltételezi a korábban ismertetett trendszámítási módszerek segítségével a trendértékek előzetes számszerűsítését. A trendértékek birtokában lehetőség nyílik a

- trendhatás leválasztására és

• a véletlen hatás kiszűrésére.

A szezonalitás hatásának számszerűsítésére kétféle – a gyakorlatban igen elterjedt – módszert alkalmazhatunk.

3.4. Szezonális eltérés számítása

Szezonális eltérést additív modell feltételezése mellett használunk olyan esetben, amikor feltételezzük, hogy a szezonális hatás abszolút nagysága, a hullámzás amplitúdója állandó. Erre legkönnyebben az idősor grafikus ábrájából következtethetünk.

Korábban már láttuk, hogy additív kapcsolat esetén az idősor komponenseinek kapcsolódása:

$y_{ij} = \hat{y}_{ij} + s^*_j + v^*_{ij}$

A trendhatás kiszűrésének módja:

$y_{ij} = \hat{y}_{ij} + s^*_j + v^*_{ij}$

Az utóbbi összefüggés jobb oldalán a trendhatástól megtisztított idősort találjuk, amelyben a szezonalitás mellett csupán a véletlen hatás szerepel. A véletlen hatást oly módon szűrhetjük ki, ha a megfelelő szezonokra vonatkoztatva a trendhatástól már megtisztított elemeket átlagoljuk:[3]

$$s_j^* = \frac{1}{n}\sum_{i=1}^{n} y_{ij} - \hat{y}_{ij}$$

A kapott értékeket **nyers szezonális eltérés**eknek nevezzük, mivel nem minden esetben teljesül, hogy a szezonális eltérések összege, illetve átlaga nulla legyen. Ha nem teljesül, az eltéréseket korrigálni kell. A korrekció során a nyers szezonális eltérések átlagát képezzük, majd ezt az átlagot rendre levonjuk az egyes nyers szezonális eltérésekből. A **korrigált szezonális eltérés**:

$$s_j^* - \frac{1}{m}\sum_{j=1}^{m} s_j^*$$

A fentieket jól szemléltethetjük az alábbi példával:

17.6. táblázat - Egy városban ismerik a sportrendezvények látogatóinak számát 2001és 2004 között, negyedéves bontásban (ezer főben)

Év	I. negyedév	II. negyedév	III. negyedév	IV. negyedév
2001	87	85	110	84
2002	74	72	102	82
2003	71	67	99	78
2004	70	65	90	71

Forrás: Saját számítás

A cég trendértékeit 4 tagú mozgóátlagolás segítségével állítottuk elő.

17.7. táblázat - Mozgóátlagolás munkatáblája

[3]Itt az *n* a megfelelő szezonokban lévő elemek számát jelenti.

Sorszám	Megfigyelés	4 tagú átlag	Centrírozott érték
1.	87		–
2.	85	91,5	–
3.	110	88,25	89,875
4.	84	85	86,625
5.	74	.	.
6.	72	.	.
.	.	.	.
.	.	.	.

Forrás: Saját számítás

Az első 4 tagú átlag: (87 + 85 + 110 + 84)/4 = 91,5

A második 4 tagú átlag: (85 + 110 + 84 + 74)/4 = 88,25

Centrírozott érték: (91,5 + 88,25)/2 = 89,875

A mozgóátlagolással kiszámított trendértékek:

17.8. táblázat - A trendértékek

Év	I. negyedév	II. negyedév	III. negyedév	IV. negyedév
2001	–	–	89,875	86,625
2002	84,000	82,750	82,125	81,125
2003	80,125	79,250	78,625	78,250
2004	76,875	74,875	–	–

Forrás: Saját számítás

A tényleges- és a trendértékek különbségeként képezhetjük a trendhatástól megtisztított értékeket:

17.9. táblázat - A trendhatástól megtisztított értékek

Header 1	I.	II.	III.	IV.
	negyedév			
2001	–	–	20,125	–2,625
2002	–10,000	–10,750	19,875	0,875
2003	–9,125	–12,250	20,375	–0,250
2004	–6,875	–9,875	–	–
Átlag	–8,667	–10,959	20,125	–0,667
s_j*	–8,625	–10,917	20,167	–0,625

Forrás: Saját számítás

A nyers szezonális eltérések átlaga: –0,042. Ezt rendre levonva a negyedéves átlagos értékekből a korrigált szezonális eltéréseket kapjuk **(s_j*)**.

A szezonális eltérések jól jellemzik az idősort. Mivel a sportrendezvények erőteljesen kötődnek az idényhez, időjáráshoz, ezért nem meglepő, hogy az első negyedévben átlagosan 8625 alacsonyabb, míg a harmadik negyedévben 20 167 fővel nagyobb volt a sportrendezvényekre kilátógatók száma az alapirányzathoz képest.

3.5. A szezonindex számítása

Multiplikatív modell esetén a szezonális hullámzás a vizsgált jelenség, folyamat nagyságával arányos, a hullámzás amplitúdója nem állandó, csak relatív stabilitást mutat. Az ilyen típusú szezonalitás mérésére a **szezonindex**et használjuk.[4]A multiplikatív modell:

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

A trendhatástól megtisztított idősor:

$$\frac{y_{ij}}{\hat{y}_{ij}} = s_j \times v_{ij}$$

A szezonindexek számszerűsítése az alábbi módon történik.

$$s_j = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{ij}}{\hat{y}_{ij}}$$

Az így képzett szezonindexek ún. **nyers szezonindex**ek, mivel a módszer közvetlenül nem garantálja, hogy átlaguk 1 legyen, ezért a **tisztított szezonindex**ekhez úgy jutunk, ha a nyers szezonindexeket rendre elosztjuk saját átlagukkal.

$$s_j \Big/ \left(\frac{\sum_{j=1}^{m} s_j}{m} \right)$$

A szezonindex számítását is a sportrendezvények látogatottságának példáján mutatjuk be.

Az eredeti adatokat a trendértékekkel elosztva megkapjuk az alapirányzattól megtisztított értékeket, amelyek a szezonindex-számítás alapját képzik.

17.10. táblázat - A trendhatástól megtisztított értékek

Header 1	I.	II.	III.	IV.
	negyedév			
2001	–	–	1,2239	0,9696
2002	0,8809	0,8700	1,2420	1,0107
2003	0,8861	0,8454	1,2591	0,9968
2004	0,9105	0,8681	–	–
Átlag	0,8925	0,8611	1,2416	0,9923
s_j*	0,8953	0,8638	1,2455	0,9954

Forrás: Saját számítás

Az átlag sorban lévő ún. nyers szezonindexeket korrigáljuk (elosztjuk) a saját átlagukkal (0,9968-cal), így nyerjük a korrigált szezonindexeket.

A táblából megállapíthatjuk, hogy az első negyedévben a szezonhatás 10,47%-kal téríti el az idősor értékét a trendtől lefelé. A második negyedévben az elmaradás 13,62%-os,

[4]A gyakorlatban a leginkább elterjedt a szezonindex számítása.

míg a harmadik negyedév kompenzálja kissé az elmaradásokat, mivel a trendnél 24,55%-kal nagyobb látógatói számot számszerűsíthetünk. A negyedik negyedévben az alapirányzattól mért elmaradás igen kis mértékű, csupán 0,46%-os.

Az idősorok (általában analitikus trendszámításon alapuló) előrejelzése során természetesen a szezonális hatásokat is figyelembe vesszük.

4. Ellenőrző feladatok, gyakorló példák a fejezethez

- Vizsgáljuk a sportturizmus alakulását. A következő tábla a sportcélú autóbuszforgalmat szemlélteti.
  - Alkalmas mozgóátlagolásos trendszámítás segítségével jellemezze az idősor alapirányzatát!
  - Számítsa ki és értelmezze a szezonális eltéréseket!

17.11. táblázat - A Magyarországra belépett autóbuszok száma 2000 és 2004 között, negyedéves bontásban (ezer db)

- Vizsgáljuk a sportorvosok ténykedését az elmúlt években. A következő tábla a ténylegesen megvizsgált sportolók számát szemlélteti 1994 és 2004 között.
  - Határozza meg a 2007-re várható megvizsgált sportolók számát!
- A következő táblázat a magyar sportolók összlétszámát mutatja az eddigi olimpiákon.
  - Készítsen idősorelemzést a 4 tagú mozgóátlagolás módszerével, ezt követően az analitikus trendszámítással jelezze előre a 2008-as olimpián részt vevő sportolóink számát!

18. fejezet - Indexszámítás

Általánosságban az indexszám valamilyen szempontból összetartozó változók időbeli vagy térbeli összehasonlítását segítő mérőszám, egy összetett, összehasonlító viszonyszám. Az alábbi fejezetben az ún. **klasszikus indexszámítás** néhány kérdését villantjuk fel.

Kiindulásként fogalmazzuk meg azt a mindenki előtt jól ismert azonosságot, amely a gazdasági jelenségek, folyamatok elemzése során kiemelkedő fontossággal bír:

Bevétel = Egységár Mennyiség

Amennyiben az árat **p**-vel, a mennyiséget **q**-val és szorzatukat – amelyet értéknek nevez az irodalom – **v**-vel jelöljük,[1] a fenti összefüggés:

v = p x q

A fenti azonosság bármely elemének dinamikus változását egyszerű dinamikus viszonyszám segítségével jellemezhetjük. Ezeket a dinamikus viszonyszámokat az indexszámítás fogalomkörében **egyedi indexek**nek nevezzük (az összehasonlítandó időszakokat, a bázis-, illetve tárgyidőszakot, a szokásos 0 és 1 jelöléssel különböztetjük meg).

Egyedi árindex:

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

egyedi volumenindex:

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

illetve az **egyedi értékindex**:

$$i_v = \frac{v_1}{v_0} = \frac{p_1 q_1}{p_0 q_0}$$

Egy-egy termék, szolgáltatás, fogyasztási cikk stb. értékének, árának, volumenének változását az egyedi indexek segítségével jól jellemezhetjük. Arra a kérdésre, hogy miként változott több termék, fogyasztási cikk ára együttesen, célszerűbb egyetlen kifejező számértékkel válaszolni, a termékek áralakulásának felsorolása helyett. A különböző piaci árszínvonalak, a fogyasztási árak, vagy például különböző országok fogyasztásának, exportjának összehasonlítása egy-egy összetett indexszám segítségével oldható meg eredményesen. Az indexszámítás kérdését időbeli összehasonlítás példáján keresztül tárgyaljuk, de a módszert területi összehasonlításokra is ki lehet terjeszteni.

Elsőként érdemes kiszámítani a különféle termékek, szolgáltatások pénzben kifejezett értékét[2] (az egységárak és a mennyiség szorzataként), majd ezeket összegezve, összesített értékadatokat, ún. aggregátumokat[3] hozhatunk létre. A két aggregátum hányadosaként egy összesített indexet, ún. **értékindex**et kapunk. Felhasználva az általános jelöléseket az értékindex:

$$I_v = \frac{\sum q_1 p_1}{\sum q_0 p_0}$$

ahol az összegzés a különböző árucikkekre, szolgáltatásokra vonatkozik. Az alábbiakban az indexszámítás kérdéskörét egy nagyon leegyszerűsített példa segítségével mutatjuk be.

[1] A *p* a latin pretium (ár), a *q* a latin quantum (mennyiség) és a *v* a latin valor (érték).
[2] Könnyű belátni, hogy a különféle termékek, árucikkek, szolgáltatások, néha eltérő mértékegységű adatait, az árak, mint egyenértékes segítségével hozhatjuk közös nevezőre.
[3] Az aggregálás értékben való összesítést jelent, amelynek eredménye az aggregátum.

Tételezzük fel, hogy egy sportegyesület labdarúgó-stadionjának bevételét csak a nézők által vásárolt jegyek ára és a reklámfelületek bérleti díja alkotja. Az egyszerűség kedvéért csak egyféle belépőjegyárral és bérleti díjjal számolunk. A bevételre vonatkozó legfontosabb adatok két egymást követő hónapban az alábbiak:

18.1. táblázat - A stadion legfontosabb bevételi adatai szeptember és október hónapban

Megnevezés	Szeptember		Október	
Egységár (Ft) p_0	Mennyiség q_0	Egységár (Ft) p_1	Mennyiség q_1	
Belépőjegy (db)	800	5 000	880	4 000
Reklámfelület (m^2)	8 000	1 000	7 200	120

Forrás: Saját számítás

Az értékesítés bevételének (az árbevételnek) a változását értelemszerűen a fent megismert képlet segítségével tudjuk számszerűsíteni:

$$I_v = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{4000 \times 880 + 120 \times 7200}{5000 \times 800 + 100 \times 8000} = \frac{4.384.000}{4.800.000} = 0{,}9133$$

(azaz: 91,33%)

A stadion árbevétele szeptemberről októberre együttesen 8,67%-kal csökkent.

A fenti számítás során tulajdonképpen különböző szolgáltatások, termékek együttes értékváltozását határoztuk meg.

Az értékindex számításához szükséges adatok – a folyóáras értékadatok – általában minden gazdasági szinten közvetlenül rendelkezésre állnak. Az index segítségével a termelés értékének, a forgalomnak, a fogyasztásnak, illetve különféle egyéb gazdasági folyamatoknak (pl. export, import) változását nyomon követhetjük. Az értékindex számítási algoritmusa azonban arra is felhívja a figyelmet, hogy az árösszeg vagy -érték változását alapvetően két tényező idézi elő:

- az árak változása,
- a volumen változása.

A két tényező hatásának számszerűsítésére szolgál az árindex és a volumenindex.

Az **árindex** *több termék, szolgáltatás együttes, átlagos árváltozását fejezi ki.*

Ahhoz, hogy az árak együttes, átlagos változását számszerűsítsük, az értékindexből indulunk ki. Rögzítsük a *q* mennyiségi adatokat! Alapvetően két megoldás közül választhatunk:

a. a tárgyidőszaki volumenadatokat (q_1),

b. a bázisidőszaki volumenadatokat (q_0)

tekintjük állandónak.

Ezek alapján az árindex két alapvető típusát különböztethetjük meg:

$$I_p^{(1)} = \frac{\sum q_1 p_1}{\sum q_1 p_0} \qquad I_p^{(0)} = \frac{\sum q_0 p_1}{\sum q_0 p_0}$$

Az első indexet **tárgyidőszaki súlyozású,** ún. **Paasche-árindex**nek, a másodikat **bázissúlyozású,** ún. **Laspeyres-árindex**nek nevezzük.

Az indexek kiszámítását segíti az alábbi munkatábla:

18.2. táblázat - Munkatábla az indexek számításához

Megnevezés	q_0p_0	q_1p_1	q_0p_1	q_1p_0
Belépőjegy	4 000 000	3 520 000	4 400 000	3 200 000
Reklámfelület	800 000	864 000	720 000	960 000
Összesen:	4 800 000	4 384 000	5 120 000	4 160 000

Forrás: Saját számítás

Paasche-árindex:

$$I_p^{(1)} = \frac{4.384.000}{4.160.000} = 1{,}0538,$$

azaz 105,38%.

Laspeyres-árindex:

$$I_p^{(0)} = \frac{5.120.000}{4.800.000} = 1{,}0667$$

azaz 106,67%.

Arra a kérdésre, hogy átlagosan együttesen miként változtak együttesen az árak, egyértelmű választ nem tudunk adni. Azt mondhatjuk, hogy amennyiben a tárgyidőszaki (októberi) értékesítési volumenek valósultak volna meg szeptember hónapban is, 5,38%-kal nőttek volna együttesen és átlagosan az árak; míg ha a bázisidőszaki (szeptemberi) mennyiségi arányok érvényesültek volna mindkét hónapban, az árszint 6,67%-kal nőtt volna.

Az árváltozás mértékét forintban is kifejezhetjük az indexek számlálói és nevezői különbségeként:

4 384 000 – 4 160 000 = 224 000 Ft

illetve

5 120 000 – 4 800 000 = 320 000 Ft

A kétféle szemléletű árindex egyaránt az árszínvonal változását jellemzi, értékük azonban általában eltérő. A Laspeyres- és Paasche-index fentiekben bemutatott aggregát formái között csak abban találunk különbséget, hogy más-más időszak mennyiségi adatai (q) szerepelnek súlyként, tehát eltérő a súlyozásuk, így fejezve ki a különféle szolgáltatások, termékek időszakonként eltérő fontosságát.

A kétféle súlyozású index lényegében egyenrangú, a valóságot azonban kissé egyoldalúan próbálja megközelíteni. Ha az árak egyik időszakról a másikra jelentősen és eltérően változnak, a két számítás eredménye erősen eltérhet. A különböző súlyozású indexek értékének nagyobb eltérése esetén jó eredményt ad a két alapforma *mértani átlagaként előállított keresztezett formula*, **a Fisher-féle árindex:**

$$I_p^F = \sqrt{I_p^{(1)} \times I_p^{(0)}}$$

A kétféle súlyozású árindex alapján kiszámítható Fisher-féle árindex:

$$I_p^F = \sqrt{1{,}0538 \times 1{,}0667} = 1{,}0602$$

(százalékban: 106,02%)

Az árszínvonal 6,02%-kal nőtt egyik hónapról a másikra. Az áremelés tehát együttesen a

bevétel növelésének irányába hatott.

A **volumenindex** *a termékek bizonyos körére vonatkozóan, több, nem egynemű több termék, szolgáltatás együttes, átlagos volumenváltozását fejezi ki.*

A volumenindex is kétféle szemléletben írható fel:

Tárgyidőszaki súlyozású, **Paasche-volumenindex:**

$$I_q^{(1)} = \frac{\sum q_1 p_1}{\sum q_0 p_1}$$

Bázissúlyozású, **Laspeyres-volumenindex:**

$$I_q^{(0)} = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

Természetesen a kétféle szemléletű, súlyozású volumenindex is eltérő eredményt ad, így itt is kézenfekvő a keresztezett **Fisher-volumenindex** kiszámítása.

A munkatábla adatai alapján könnyen meghatározhatjuk a kétféle volumenindexet:

Paache-volumenindex:

$$I_q^{(1)} = \frac{4.384.000}{5.120.000} = 0{,}8563$$

(85,63%)

Laspeyres-volumenindex:

$$I_q^{(0)} = \frac{4.160.000}{4.800.000} = 0{,}8667$$

(86,67%)

Szeptemberről októberre a szolgáltatási volumen együttesen, átlagosan 14,37%-kal, illetve 13,33%-kal csökkent.

A Fisher-volumenindex:

$$I_q^F = \sqrt{0{,}8563 \times 0{,}8667} = 0{,}8614$$

(százalékban: 86,14)

Egyes szolgáltatásokra, termékekre vonatkozóan az ár és a mennyiség szorzataként a bevételt, értéket kapjuk eredményül. Ugyanilyen szorzatszerű összefüggés áll fenn az egyedi árindexek esetében is:

$i_p \times i_q = i_v$

A termékek, árucikkek meghatározott körére a fenti *multiplikatív összefüggés az ellentétes súlyozású indexek, illetve a Fisher-indexek között áll fenn.*

a. $i_p^{(1)} \times i_q^{(0)} = i_v$

b. $i_p^{(0)} \times i_q^{(1)} = i_v$

c. $i_p^F \times i_q^F = i_v$

A hazai és nemzetközi gyakorlat a vizsgálat természetétől függően számít bázis, vagy tárgyi súlyozású árindexeket, azonban ha jelentős eltérés várható, Fisher-indexeket számol.

Példánkban az indexek összefüggése:

0,9133 = 1,0538 x 0,8667 = 1,0667 x 0,8563 = 1,0602 x 0,8614.

Mivel a különböző súlyozású indexek jelentősen eltérnek egymástól, az együttes

elemzést a Fisher-indexek segítségével célszerű elvégezni. Ezek szerint megállapíthatjuk, hogy a bevétel együttesen és átlagosan 8,67%-kal csökkent szeptemberről októberre, amiben az árszínvonal 6,02%-os növekedése mellett az értékesítés volumenének 13,86%-os csökkenése döntő szerepet játszott. Látható, hogy – egyéb feltételezett tényezők (pl. a csapat rosszabb szereplése) mellett – a jegyárak emelése nem hatott kedvezően az együttes bevételre.

A kétféle súlyozású ár- és volumenindexet vizsgálva felvetődik a kérdés, hogy miért térnek el az indexek egymástól? Általánosságban a súlyozás különbözőségével magyarázhatjuk az eltérést. Az eltérés irányának vizsgálata azonban további információkat szolgáltat. Az egyes termékek ár- és volumenváltozása nem független egymástól, így *az egyedi ár- és volumenindexek között sztochasztikus kapcsolatot fedezhetünk fel*.

Az egyedi ár- és volumenindexeket (százalékban kifejezve) az 53. táblázat tartalmazza:

18.3. táblázat - Egyedi árindexek (%)

Megnevezés	$\mathbf{i_p}$	$\mathbf{i_p}$
Belépőjegy	110	80
Reklámfelület	90	120

Forrás: Saját számítás

Az egyedi ár- és volumenindexek között ellentétes, negatív hatás érvényesült. Ez kifejeződik a kétféle súlyozású index között felírható relációban is (pl. $i_p^{(1)} = 1{,}0538 < i_p^{(0)} = 1{,}0667$).

Amennyiben egy árucikk, termék, szolgáltatás jelentős árnövekedése a volumen csökkenésével jár együtt, az árak és volumenek változása között ellentétes irányú, **negatív sztochasztikus (korrelációs) kapcsolat** van. Ilyen esetben bebizonyítható, hogy *a bázissúlyozású index értéke nagyobb, mint a tárgyidőszaki súlyozású indexé*. Mindez az együttes ár- és volumenindexre egyaránt érvényes ($i_p^{(0)} > i_p^{(1)}$ illetve $i_q^{(0)} > i_q^{(1)}$). A negatív irányú kapcsolat általában a piacgazdaság viszonyai között általános.

Amennyiben az árak és volumenek változása között *pozitív irányú a kapcsolat, a különféle súlyozású indexek nagyságrendi relációja fordított* ($i_p^{(0)} < i_p^{(1)}$ illetve $i_q^{(0)} < i_q^{(1)}$). Az ilyen jellegű kapcsolatra általában hiánygazdaság, illetve erősen monopolisztikus piac esetén számíthatunk.

Az eddigiek során az ár- és volumenindex ún. aggregát formáját használtuk fel. Meg kell jegyeznünk, hogy a gyakorlatban sokszor alkalmazzák az indexek összefüggését. Az értékindex és az árindex ismeretében egyszerű osztás segítségével határozzák meg a volumenindexet. Ezt az eljárást – az árváltozások hatásának kiszűrését – **deflálás**nak is nevezik.

Az árindexet (és értelemszerűen a volumenindexet is) nemcsak aggregát forma segítségével lehet kiszámítani. Könnyen belátható, hogy a tárgyidőszaki vagy a bázisidőszaki árösszegek, valamint az egyedi indexek ismeretében az indexek ún. átlagformával is meghatározhatók:

$$I_p^{(1)} = \frac{\sum q_1 p_1}{\sum q_1 p_1 \div i_p} \quad \text{vagy} \quad I_p^{(0)} = \frac{\sum q_0 p_0 \times i_p}{\sum q_0 p_0}$$

Az első index a **Paasche-árindex harmonikus átlag formája**, míg a második index a **Laspeyres-árindex számtani átlag formája**. Az egyedi indexek ismeretében a fenti formákat szokta a gyakorlat előnyben részesíteni, mivel a folyóáras (q_1p_1 vagy q_0p_0) adatok közvetlenül rendelkezésre állnak.

Példánkban:

$$I_p^{(1)} = \frac{3.520.000 + 864.000}{\frac{3.520.000}{1,1} + \frac{864.000}{0,9}} = 1,0538$$

és

$$I_p^{(0)} = \frac{4.000.000 \times 1,1 + 800.000 \times 0,9}{4.000.000 + 800.000} = 1,0667$$

Számos esetben az árindex a termékek, árucikkek sokfélesége, a választék, az eladási helyek és ármozgások különbözősége miatt teljeskörűen nem határozható meg. Ilyen esetekben ún. **reprezentatív árindex**et számítanak, amely reprezentatív mintaadatokból épül fel. Az ilyen típusú árindex tipikus példája a fogyasztói árindex, amely alapvetően egy bázissúlyozású árindex. A **fogyasztói árindex** a lakosság által vásárolt fogyasztási cikkek, szolgáltatások árainak átlagos változását méri. A fogyasztói árindex alapján az inflatorikus tendenciákat számszerűsíteni lehet, az árindex felfogható az infláció közelítő mérőszámaként. A reálkeresetek, reáljövedelmek vizsgálata során fontos szerepet tölt be a fogyasztói árindex.

Természetesen az indexszámítás módszertana nemcsak időbeli, hanem területi vizsgálatokra is alkalmas. Tipikus felhasználása a **területi árindex**nek a nemzetközi összehasonlításban a **valuták vásárlóerő-arányai**nak kvantifikálása.

Az árindexeknek továbbá fontos felhasználási területe az **árarányok változásának vizsgálata.** Különböző, de egymással összefüggő területek árindexeinek összehasonlításával ún. **árollók** számíthatók. Ezek egyik fontos fajtája az ún. **agrárolló**, amely a mezőgazdaság inputjainak (vásárolt ipari termékeknek) árindexét az output (értékesítések) árindexeivel méri össze.

Szólni kell még a **külkereskedelmi cserearány** (terms of trade) mutatójáról, amely az export árindexet az import árindexszel hasonlítja össze. Egy nemzetgazdaság számára kedvező, ha értéke meghaladja a 100%-ot.

1. Ellenőrző feladatok, gyakorló példák a fejezethez

- Egy áruház sportosztályán a gyermektornacipő forgalmára vonatkozó adatokat tartalmazza a következő táblázat.
  - Elemezze a gyermektornacipő érték-, ár- és volumenalakulását 2004-ről 2005-re!
  - Vizsgálja meg, hogyan alakult az átlagár!

19. fejezet - Bevezetés a következtetéses statisztikába

1. A normális eloszlás és alkalmazása

A társadalmi és gazdasági jelenségek, valamint a sportteljesítmények jelentős köréről tudjuk vagy feltesszük, hogy folytonos, normális eloszlású valószínűségi változóként viselkednek. A folytonos valószínűségi változók egy adott intervallumban végtelen számú értéket vehetnek fel, és annak valószínűsége, hogy egy X változó pontosan x értékét veszi fel, zérus. A valószínűségi eloszlások fontos jellemzője, mintegy „azonosítója” a várható érték (μ) és a variancia, szórásnégyzet (σ^2).[1] A normális eloszlás könnyen azonosítható a várható érték és a szórás segítségével, jele: N(μ, σ).

A normalitás feltételezésével élünk pl. a súly, a térfogat, magasság, hosszúság, és a teljesítmények esetében.

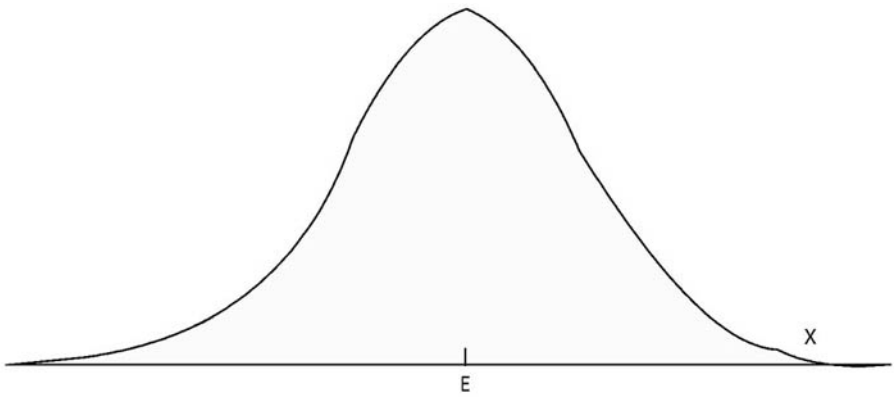
A várható értékek és a szórások, az elemzés tárgyától függően, igen sokféle értéket vehetnek fel, ami a munkát sokszor megnehezíti, hiszen nagyságuk a változók dimenziójától függ. Egy viszonylag egyszerű transzformáció segítségével azonban ez a probléma megoldható. Amennyiben a várható értéket kivonjuk a valószínűségi változó értékéből, és a különbséget elosztjuk a szórással, vagyis a változót **standardizáljuk**, a **standard normális eloszlású valószínűségi változó**t (jele: z) kapunk eredményül. Képletben:

$$z = \frac{x - \mu}{\sigma}$$

A standardizálás eredményeként kapott standard normális eloszlású valószínűségi változó várható értéke zérus, szórása egységnyi, azaz N(0, 1).

Mind a normális, mind a standard normális eloszlású valószínűségi változó sűrűségfüggvénye ún. haranggörbével, **Gauss-görbével** jellemezhető.

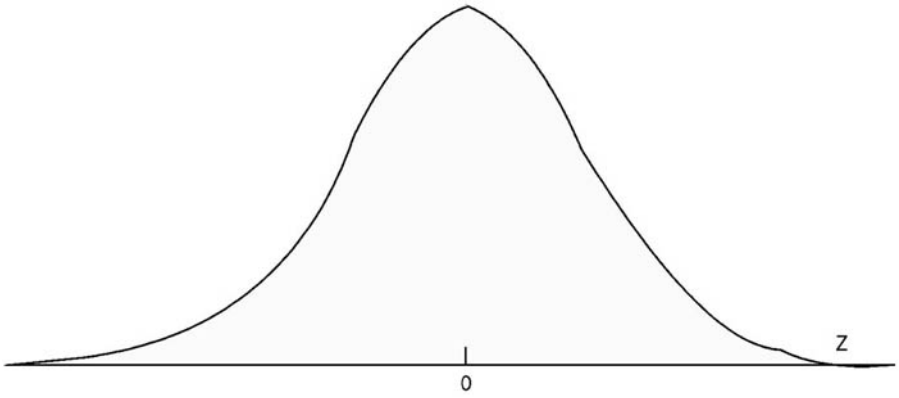
19.1. ábra - A normális eloszlás ábrája



Forrás: Saját számítás

19.2. ábra - A standard normális eloszlás

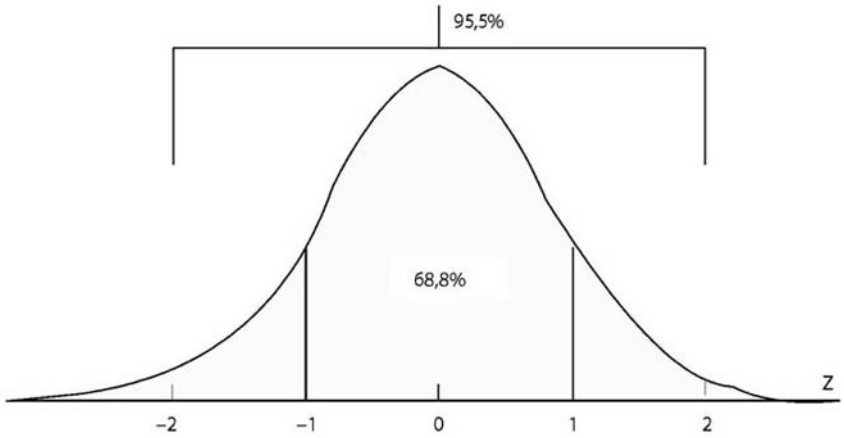
[1] A várható értéket korábban az E(X), a varianciát a Var(X) szimbólummal jelöltük. Az új jelöléseket praktikus okokkal magyarázhatjuk.



Forrás: Saját számítás

Standard normális eloszlás esetén mind a valószínűségi változók, mind a hozzájuk rendelhető valószínűségek táblázatba foglalhatók. Az így kapott értékek könnyen felhasználhatók gyakorlati problémák megoldására.

19.3. ábra - Néhány fontosabb valószínűség z függvényében



Forrás: Saját számítás

A várható értéktől egységnyi szórással eltérő intervallum – és ez nemcsak a standard, hanem az általános normális eloszlás esetére érvényes – és a valószínűségi görbe által bezárt terület 68,8%-os valószínűséget reprezentál. A kétszeres szórás által meghatározható intervallumhoz tartozó valószínűség 95,5%; míg a háromszoros szórással lefedhetjük a vízszintes tengely és a görbe által meghatározható teljes területet, szinte a teljes valószínűséget (99,9%). Ezt a felismerést ún. háromszigma-szabálynak hívják a statisztika használói.

Természetesen a fenti ábrában bemutatott eseteknél részletesebb adatokat szolgáltat a standard normális eloszlás táblázata. Gyakorlati elterjedtségét azzal magyarázhatjuk, hogy a nulla várható értékű és egységnyi szórású valószínűségi változók és hozzájuk tartozó valószínűségek egyszerű táblázatba rendezhetőek. Bármely normális eloszlású ismert várható értékű és szórású valószínűségi változó pedig a standardizálással könnyen transzformálható, és így a táblázatot használni lehet. Itt kell megemlíteni, hogy a standard normális valószínűségi változó sűrűségfüggvénye szimmetrikus, így elegendő a 0 és a pozitív végtelen közé eső számokhoz tartozó valószínűségi értékek meghatározása, mivel a negatív oldal már könnyen számszerűsíthető.

Tételezzük fel, hogy egy sportágban a sportolók súlya normális eloszlású valószínűségi változóként viselkedik. Az sportolók súlyának várható értéke (amit például egy korábbi számításból ismerünk) 80 kg, szórása pedig 10kg.

Számítsuk ki azon sportolók várható számát egy 160 fős versenyen, akiknek súlya meghaladja a 90 kg-ot!

Elsőként standardizáljuk a kritikus határt jelentő 90 kg-ot, mint a normális valószínűségi változó egy valós értékét!

$$u = \frac{90-80}{10} = 1$$

Az 1-hez tartozó valószínűségi számérték a **KH001** táblázatban: 0,159.

Ennek megfelelően, annak valószínűsége, hogy egy sportoló súlya nagyobb mint 90 kg:

Pr (x>90) = 0,159, azaz 15,9%.

A 160 résztvevő esetén a 90 kg-nál nagyobb súlyú sportolók várható száma:

160 x 0,159 ≈ 25 fő.

A következő példa a normális eloszlás gyakorlati felhasználásának újabb lehetőségére hívja fel a figyelmet.

Egy sportorvosi rendelő forgalmát felmérve megállapították, hogy a sportolókkal való foglalkozás időtartama normális eloszlást követ. Ismert, hogy a rendelőben egy sportolóra átlagosan negyed órát fordítanak, a vizsgálati idő szórása pedig 5 perc.

a. Határozza meg annak valószínűségét, hogy egy sportoló 20 percnél rövidebb időt tölt a rendelőben!

b. Mi a valószínűsége annak, hogy egy sportoló 20 percnél több időt tölt a rendelőben?

c. Mi a valószínűsége annak, hogy egy sportoló legalább 10 percet, de legfeljebb 18 percet tölt a rendelőben?

d. Napi 8 óra munkaidővel számolva, 96,4%-os valószínűség mellett állapítsa meg, hogy minimum hány fő fordul meg naponta a sportorvosi rendelőben!

a) $u = \frac{20-15}{5} = 1 \qquad Pr(1 \le u) = 0,159$

$Pr(X\langle 20) = 1 - 0,159 = 0,841 \quad \text{azaz } 84,1\%$

b) $Pr(X\rangle 20) = 0,159 \quad \text{azaz} \quad 15,9\%$

c) $u = \frac{10-15}{5} = -1 \qquad u = \frac{18-15}{5} = 0,6$

$Pr(10\langle X\langle 18) = 1 - (0,159 + 0,274) = 0,567 \quad \text{azaz } 56,7\%$

d) 1 - 0,964 = 0,036 $\qquad 0,036 \Rightarrow u = 1,8 \quad \frac{X-15}{5} = 1,8 \Rightarrow X = 24$

$\frac{480}{24} = 20$

azaz 20 fő.

A fentihez hasonló kérdések megválaszolását teszi lehetővé a következő példa.

Tételezzük fel, hogy a súlylökők dobási teljesítménye normális eloszlást követ. A dobások várható értéke 17 m, szórása 3 m. Válaszoljunk az alábbi kérdésekre:

a. Mi a valószínűsége a 17 méternél kisebb dobásnak?

b. Mi a valószínűsége a 24 méternél nagyobb dobásnak?

c. Milyen valószínűséggel várhatjuk, hogy a versenyzők dobása 20 és közé essen?

d. 15 méternél nagyobb értékű dobásokra milyen valószínűséggel számíthatunk?

a) $u = \frac{17-17}{3} = 0 \quad Pr(X\langle 17) = 0,5$

b) $u = \frac{23-17}{3} = 2 \quad Pr(X\rangle 24) = 1 - 0,023 = 0,977$

c) $u = \frac{20-17}{3} = 1 \quad u = \frac{22-17}{3} = 1,67$

$Pr(20\langle X\langle 22) = 0,159 - 0,047 = 0,112$

d) $u = \frac{15-17}{3} = -0,67 \quad Pr(X\rangle 15) = 1 - 0,251 = 0,749$

A statisztikai középértékek – különösen a számtani átlag – kiemelt fontossággal bírnak a következtetéses statisztikában is. Közvetlenül adódik annak igénye, hogy a reprezentatív módon kiválasztott minták átlagai és szórásai, valamint az alapsokaság átlaga és szórása között valamilyen összefüggést keressünk. Hangsúlyozni kell, hogy a **centrális határeloszlás** tétele értelmében bármilyen eloszlással rendelkező alapsokaságból egyszerű véletlen mintavétel segítségével nyert minta átlaga valószínűségi változó, mivel értéke mintáról mintára ingadozik, ugyanakkor az **átlagok normális eloszlású valószínűségi változók**. Mindez természetesen fokozottan aláhúzza a normális eloszlás gyakorlati hasznosíthatóságát, elterjedtségét.

Az alábbi sematikus példa segítségével mutatjuk be a mintaátlagok és az alapsokaság fontos paraméterei közötti összefüggéseket.

Csupán didaktikai okokból tételezzük fel – mivel példánk esetére a gyakorlatban nem találunk ésszerű magyarázatot –, hogy egy alapsokaság csak 5 elemből áll, de mégis mintavétellel kívánunk számszerű megállapításokat tenni. Öt birkózó súlya az alábbi legyen (kg): 90, 120, 130, 150, 160.

Vegyünk 2 elemű mintát egyszerű véletlen módon a fenti alapsokaságból!

Amennyiben egyszerű véletlen módszerrel, visszatevés nélkül választjuk ki a 2 elemű mintákat, tulajdonképpen a lehetséges esetek az ismétlés nélküli kombinációk számának felelnek meg. Tehát

$\binom{N}{n}$ azaz $\binom{5}{2} = 10$

féleképpen tudunk 5 elemből 2 elemet kiválasztani. Szimuláljuk az összes lehetséges mintát! (Ezt most az alapsokaság jelképes nagysága miatt könnyen megtehetjük.)

19.1. táblázat - A mintaelemek

Sorszám	Kiválasztott mintaelemek		Header 4
J	x_1	x_2	x_j
1.	90	120	105
2.	90	130	110
3.	90	150	120
4.	90	160	125
5.	120	130	125
6.	120	150	135
7.	120	160	140
8.	130	150	140
9.	130	160	145
10.	150	160	155

Forrás: Saját számítás

A kiindulásként feltüntetett alapsokaság csupán 5 elemű, ezért mind az átlagát, mind a

varianciáját könnyen meghatározhatjuk:

$$\overline{X} = \frac{90+120+130+150+160}{5} = 130\ \text{kg}$$

$$\sigma^2 = \frac{(90-130)^2+(120-130)^2+(130-130)^2+(150-130)^2+(160-130)^2}{5} = 600$$

amiből a szórás:

$$\sigma = \sqrt{600} = 24{,}5\ \text{kg}.$$

Látjuk, hogy a 10 különféle minta átlaga különbözik az alapsokasági átlagtól, aminek értékét becsülni hivatott. Néhol az eltérések jelentősek lehetnek. A valóságban az alapsokaság átlagára vonatkozóan nem rendelkezünk információkkal. Törekedni kell azonban arra, hogy becslésünk csak kismértékben térjen el a becsülni kívánt paraméterektől.

A mintavétel egyik legalapvetőbb formája az **egyszerű véletlen mintavétel.** Amennyiben az alapsokaságból a mintaelemeket véletlenszerűen, visszatevés nélkül választjuk ki, egyszerű véletlen mintavételről van szó.

A következtetéses statisztika igényli különböző összefüggések felismerését a mintaátlagok, azok szórása és az alapsokasági átlag és szórás között. Az alábbiakban ezeket az összefüggéseket empirikus módon mutatjuk be.

Könnyen belátható, hogy amennyiben ismerjük valamennyi minta átlagát – ez most a sematikus példánkban így van –, a minták átlagából képzett átlag megegyezik az alapsokasági átlaggal.

$$\overline{\overline{X}} = \frac{105+110+\ldots+145+155}{10} = 130\ \text{kg}$$

A mintaátlagok szórása azonban eltér az alapsokaság szórásától:

$$\sigma_{\overline{x}} = \sqrt{\frac{(105-130)^2+\ldots+(155-130)^2}{10}} = \sqrt{225} = 15\ \text{kg}$$

(Emlékezzünk rá, hogy az alapsokasági szórás 24,5 kg volt!)

Létezik azonban – bizonyítás nélkül közöljük – egy olyan összefüggés, amelynek segítségével közvetlen kapcsolat írható fel az alapsokasági szórás (szórásnégyzet) és a mintaátlagok szórása (szórásnégyzete) között:

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]$$

ahol: *n* a mintaelemek száma és *N* az alapsokaság elemeinek száma.

Itt jegyezzük meg, hogy a kifejezés második tagját, az

$$\left[\frac{N-n}{N-1}\right]$$

tényezőt korrekciós tényezőnek vagy véges szorzónak hívja az irodalom. A visszatevés nélküli kiválasztás[2] esetén játszik fontos szerepet, visszatevéses mintavétel alkalmazása során nem szerepel a képletben. Itt kell szólni arról, hogy a korrekciós tényezőt elhagyhatjuk visszatevés nélküli kiválasztás, azaz egyszerű véletlen mintavétel esetén is, amennyiben az alapsokaság (N) nagysága jelentősen eltér a minta (n) nagyságától, mivel ilyen esetekben a tényező 1-hez közeli értékkel bír.

A mintaátlag szórásnégyzete $\sigma^2_{\overline{x}}$, egy olyan átlagos négyzetes hiba, amelyet akkor követünk el, amikor következtetéseink során a sokasági várható értéket mindig a mintaátlaggal helyettesítünk. A statisztikai módszerek között kiemelkedő fontossággal bír a mintaátlag szórása ($\sigma_{\overline{x}}$), amit a mintaátlag **standard hibájának** neveznek.

Az előző képletben felírt összefüggést nézzük meg számpéldánkban!

$$\sigma^2_{\overline{x}} = \frac{600}{2}\left[\frac{5-2}{5-1}\right] = 225$$

[2] A visszatevés nélküli mintavétel (pl. egyszerű véletlen mintavétel) a gyakorlatban igen népszerű, mivel alkalmazása nem jár információveszteséggel.

Amiből a mintaátlag szórása, azaz standard hibája:
$\sigma_{\bar{x}} = \sqrt{225} = 15\,\text{kg}.$

Az alapsokaság szórásának ismeretében tehát könnyen kiszámítható a mintaátlagok szórása.

A véletlen minta elemei véletlen változók, ezért bármely transzformációjuk, így a belőlük számított számtani átlag is, véletlen változó lesz. Ha a sokasági eloszlás normális, akkor a mintaátlag is normális eloszlású, függetlenül a minta elemszámától. A mintaátlagokról azonban azt is tudjuk, hogy nagy minta esetén – erre utal a nagy számok törvénye és a központi (centrális) határeloszlás tétele – a mintát egyszerű véletlen módon, bármilyen alapeloszlású sokaságból kiválasztva, a mintaátlagok normális eloszlást fognak követni. Ezt figyelembe véve, a mintaátlagok is standardizálhatók a

$$z = \frac{x - \mu}{\sigma}$$

képlet alapján. Ezeket a megállapításokat felhasználva bővítsük ki a korábban megismert repülőtársasági példánkat!

Tételezzük fel, hogy a repülőtársaság arra kíváncsi, hogy milyen valószínűséggel várható egy-egy repülés alkalmával az, hogy a gép utasainak átlagos súlya (csomagokkal együtt) nem éri el a 78 kg-ot. Mindezt egy 100 elemű egyszerű véletlen módon vett minta alapján kívánják eldönteni. Korábbi elemzéskeből ismert, hogy az alapsokaság (az összes utas) átlagos súlya 80 kg, szórása 10 kg.

Szükségünk van a számításokhoz a mintaátlagok szórására:
$\sigma_{\bar{x}}^2 = \frac{10^2}{100} = 1 \quad \text{azaz} \quad \sigma_{\bar{x}} = 1\,\text{kg}$

Az átlagot, mint változót standardizálva
$z = \frac{78 - 80}{1} = -2 \quad \text{tehát} \quad \Pr(\bar{X}\langle 78) = 0{,}023$

Tehát 2,3% annak a valószínűsége, hogy az utasok átlagos súlya kisebb mint 78 kg.

2. Becslési módszerek

Az alapsokaság adott értékétől a becsült értékek gyakorta különböznek (elhanyagolhatóan kicsi annak a valószínűsége, hogy a két érték megegyezik). Láttuk azonban, hogy az alapsokaság átlaga, valamint a mintaátlagok között közvetlen, a szórás és a mintaátlagok szórása között is jól kifejezhető összefüggés írható fel. Különösen fontos szerepet tölt be a standard hiba, a mintaátlagok szórása. Ez a szóródási mérőszám lehetőséget ad arra, hogy a becslésünket egy olyan intervallummal adjuk meg, aminek a bekövetkezése, adott valószínűségi szinten, garantálható.

A
$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]$$

képletet megvizsgálva azonban szembetűnik az, hogy a standard hiba meghatározása igényli az alapsokaság szórásának ismeretét. Amennyiben az alapsokasági szórás rendelkezésünkre áll, többnyire nincs szükség a paraméterek becslésére. A gyakorlatban azonban legtöbbször csupán egy minta adatai állnak rendelkezésünkre, így ebből a minta alkotta adatbázisból kell a szórást is meghatároznunk. Matematikai–statisztikai módszerekkel bebizonyítható, hogy a mintabeli variancia (szórásnégyzet) is véletlen változó, és kis módosítással jól becsülhető segítségével az alapsokaság szórása. Az alapsokasági szórás becslésére az ún. **korrigált mintabeli szórás**t (jele: s) használhatjuk fel:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

A korrigált mintabeli szórás segítségével felírható a gyakorlatban jól használható standard hiba négyzetének képlete:

$$\sigma_{\bar{x}}^2 = \frac{s^2}{n}\left[\frac{N-n}{N-1}\right] \cong \frac{s^2}{n}\left(1-\frac{n}{N}\right)$$

Mivel a véges szorzót csak akkor indokolt használni, ha a minta nagysága az alapsokaság nagyságának 5%-át meghaladja, a standard hiba gyakran használt képlete:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Hangsúlyoznunk kell, hogy a fenti standardhibaképletek csupán az átlagok szóródását jellemzik. Más paraméterekre, pl. értékösszeg, arány is felírhatók a megfelelő szórások, más néven standard hibák.[3]

Egy egyszerű véletlen módon kiválasztott minta adatai alapján a mintasokaság értékeiből egy alkalmas képlet[4] (pl. számtani átlag) segítségével **közelítő érték**ét adhatjuk az ismeretlen sokaság paraméterének. Ezt **pontbecslés**nek hívjuk. A becslés során elkövethető véletlen hiba átlagos nagyságát a standard hiba szolgáltatja. A gyakorlatban jól felhasználható információt nyerünk azonban akkor, ha egy **intervallumbecslés**t szerkesztünk. Az intervallumbecslés során felhasználjuk azt, hogy a mintaparaméterek valamilyen ismert eloszlású valószínűségi változók, és így az adott eloszlás értékének felhasználásával egy **adott megbizhatósági szinten** állapíthatunk meg egy intervallumot. Ezt az intervallumot **konfidenciaintervallum**nak hívjuk. Átlagbecslés esetén a konfidenciaintervallum:

$$\bar{x} \pm z \times \sigma_{\bar{x}}$$

ahol: *u* a standard normális eloszlás adott értéke.

Korábbról ismerjük, hogy az **KH002** z értéke 95%-os megbízhatóság mellett 1,96; 95,5%-os megbízhatóságon 2; míg 90%-os szinten 1,645.

Módosítsuk kissé a repülőtársasággal kapcsolatos példánkat! Tegyük fel, hogy 100 000 utas közül egyszerű véletlen módszerrel választottak ki 100 utast. Megmérték az utasok súlyát (csomagjaikkal együtt) és 80 kg-os átlagos értéket kaptak. Ugyanakkor kiszámították a 100 elemű minta szórását, ami 20 kg volt.

Becsüljük meg 95%-os megbízhatóság mellett a légitársaság által szállított utasok átlagos súlyát!

A véges szorzót nem kell alkalmazni, mivel a kiválasztásai arány 100/100 000 kisebb mint 5%.

$$\sigma_{\bar{x}} = \frac{20}{\sqrt{100}} = 2$$

Az utasok súlyának konfidenciaintervalluma:

80 ± 1,96 x 2

tehát 95,5%-os megbízhatóság mellett állíthatjuk, hogy a légitársaság utasainak átlagos súlya legalább 76,08, legfeljebb 83,92 kg.

Az átlagbecslés ismeretében igen egyszerűen meghatározható az értékösszeg. Az **értékösszegbecslés** pontbecslése és standard hibája az átlagra vonatkozó összefüggésekből könnyen levezethető:

$$x' = N \times \bar{x} \text{ és } \sigma_{x'} = N \times \sigma_{\bar{x}}$$

Az értékösszegbecslésre jó példa egy adott repülőgépen utazó utasok összes súlyának meghatározása.

Tételezzük fel, hogy a légitársaság szeretné megtudni 95%-os megbízhatósági szinten, hogy egy 160 személyes repülőgépén mekkora lesz az utasok várható összes súlya.

[3] Ezeket a standard hiba képleteket a továbbiakban egy-egy adott probléma kapcsán ismertetjük.
[4] A becslés során felhasznált képletet becslőfüggvénynek is nevezi az irodalom.

A súlyra a pontbecslés az alábbi módon adható meg:

x' = 160 x 80 = 12 800 kg.

A standard hiba értéke:

$\sigma_{x'}$ = 160 x 2 = 320 kg.

A konfidenciaintervallum:

12 800 ± 1,96 x 320.

Tehát 95%-os megbízhatóság mellet állíthatjuk, hogy a gépen az utasok súlya nem lesz kevesebb, mint 12 172,8 kg és nem lesz több, mint 13 427,2 kg.

A fent tárgyalt átlagbecslésre nézzünk egy másik példát!

Egy közvélemény-kutató intézet a sportolással töltött időt vizsgálta. A vizsgálat céljából egy véletlenszerűen kiválasztott 1000 fős mintát vettek, amiből megtudták, hogy a sportolással töltött idő átlagosan 6,2 óra hetente, amelynek szórása 0,7 óra.

Becsüljük meg 95%-os megbízhatósági szinten a heti sportolási időt!

$$6,2 \pm 1,96 \times \frac{0,7}{\sqrt{1000}}$$

6,2 ± 0,04

Tehát 95%-os megbízhatóság mellett várhatjuk azt, hogy egy fő hetente átlagosan 6,16 óránál több, de 6,24 óránál kevesebb időt tölt sportolással.

Az átlagbecsléséhez hasonlóan becsülhető az alapsokaság valamilyen ismérv szerinti aránya (megoszlási viszonyszáma) is. Valamely tulajdonsággal bíró egyed arányát jelöljük az alapsokaságban P-vel. A P arány pontbecslése:

$$p = \frac{k}{n}$$

ahol: *k* az adott tulajdonsággal bíró egyedek száma.

A mintabeli aránynak a mintából számítható standard hibája:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)}$$

Nagy minta (ahol n ≥ 30) esetén joggal feltételezzük, hogy p eloszlása közelíthető a normális eloszlással, ezért a konfidenciaintervallum szerkesztéséhez felhasználhatjuk a standard normális eloszlás értékeit.

A konfidenciaintervallum:

p ± z x σ_p

A fentieket szemléltessük egy példával!

1000 fős mintát véletlenszerűen választottak ki egy város népességéből. A megkérdezettektől arról tudakozódtak, hogy szoktak-e rendszeresen mozogni, sétálni, sportolni? Az 1000 válaszadóból 450 igennel válaszolt a kérdésre.

Határozzuk meg 95%-os megbízhatósági szinten, hogy az adott térség népességének hány százaléka él az aktív pihenés, mozgás lehetőségével!

$$p = \frac{450}{1000} = 0,45$$

$$0,45 \pm 1,96 \times \sqrt{\frac{0,45 \times 0,55}{1000}}$$

$$0,45 \pm 0,03$$

42% 48%

Tehát 95%-os megbízhatósági szinten állíthatjuk, hogy a lakosságnak legalább 42%-a és legfeljebb 48%-a mozog rendszeresen.

Az arány becslés másfajta felhasználását mutatja be a következő példa.

Egy országban nyáron (átigazolási időszakban) az egyesületet kereső 1000 sportoló köréből egyszerű véletlen mintavétellel 100 elemű mintát vettek és ebből megállapították, hogy 10 fő 4 héten belül új egyesületet talált.

a. Állapítsuk meg 95%-os megbízhatósági szinten, hogy az egyesületet kereső sportolók hány százaléka talál 4 héten belül új egyesületet!

b. A fenti megbízhatósági szinten legalább hány sportoló talál egyesületet 4 héten belül?

a)

$$0,1 \pm 1,96 \times \sqrt{\frac{0,9 \times 0,1}{100} \cdot \left(1 - \frac{100}{1000}\right)}$$

0,1 ± 0,056

4,4% és 15,6%.

Tehát az egyesületet kereső sportolónak legalább 4,1, legfeljebb 15,9%-ának lesz egyesülete négy héten belül.

b)

1000 x 0,044 = 44 fő.

Az ezerből mintegy 44 sportoló helyezkedik el négy héten belül.

A fentiekből láttuk, hogy az egyszerű véletlen mintavétel miatt fontos véges szorzót a mintavételi arány függvényében alkalmaztuk. Megjegyezzük azonban, hogy megnyugtató, ha mindig „korrigálunk” ezzel a szorzószámmal.

3. Hipotézis-ellenőrzési módszerek

A véletlenszerűen kiválasztott minta nemcsak az alapsokaság valamely ismeretlen paraméterének megközelítő pontosságú becslését teszi lehetővé, hanem olyan következtetések elvégzését is, amely az alapsokasági paraméterre vagy alapsokasági jellemzőre megfogalmazott állítás helyességét hivatott eldönteni. A feltevések vizsgálata **statisztikai hipotézis-ellenőrzés**sel végezhető el. A feltevések, amiket **hipotézis**eknek nevezhetünk, egy-egy sokaság jellemzőjét (átlagát, arányát stb.), eloszlási paraméterét (pl. várható érték), az alapsokaság eloszlását (pl. normális eloszlás) tartalmazzák többnyire *egzakt matematikai–statisztikai formában.* Így lehetővé válik az, hogy a hipotéziseket a matematikai–statisztika eszközeivel, meghatározott valószínűség figyelembevétele mellett ellenőrizzük; és végezetül a feltevést elfogadjuk vagy elvessük.

A hipotézisek felállításának általános menete az alábbiakban foglalható össze:

Jelöljük Θ-val (theta) az ismeretlen alapsokasági értéket és Θ_0-val a feltételezett értéket! Kiinduló **nullhipotézis**ünket az alábbi módon írhatjuk fel:

H_0:Θ = Θ_0.

Természetesen ez a kifejezés önmagában még nem értelmezhető, meg kell fogalmazni ellentétpárját, azaz az **alternatív hipotézis**t. Az alternatív hipotézis lehet **kétoldalú**:

$H_1:\Theta \neq \Theta_0$,

illetve **egyoldalú**:

$H_1:\Theta < \Theta_0$

vagy

$H_1:\Theta > \Theta_0$.

A fent megfogalmazott hipotézisek ellenőrzését matematikai függvények, ún. **próbafüggvények** segítségével végezhetjük el. A próbafüggvény tulajdonképpen – leegyszerűsítve – a vizsgált paraméter változó eloszlásának megfelelő valószínűségi változó kiszámítását előíró algoritmus. A függvény lehetővé teszi az ismert statisztikai eloszlástípusoknak megfelelő elméleti értékkel való összevetést. Egy adott valószínűségi szint, ún. **szignifikanciaszint** mellett a számított értéket az elméleti értékkel összehasonlítva a hipotézist vagy elvetjük, vagy elfogadjuk; ezáltal teszteljük az adott alapsokaságra megfogalmazott állításunkat.

A hipotézis-ellenőrzés gondolatmenetének megértését segíti az alábbi okfejtés. Gondoljunk arra, hogy egy sokaságra vonatkozó feltevést (pl. számértéket) ellenőrizhetünk a teljes sokaság ismerete alapján. A gyakorlatban azonban a teljes sokaságot nem mindig ismerjük, így egy véletlenszerű minta alapján kell ítéletet alkotnunk. Tudjuk azt, hogy a véletlen mintából számított értékek mintáról mintára ingadoznak, tehát egy adott érték megegyezése vagy eltérése a hipotetikus értéktől nem jelenti egyben annak valódiságát vagy valótlanságát. Ha a mintából számított érték hipotetikus értéktől való eltérése meghaladja a véletlenek által befolyásolt, de még elfogadható szintet, akkor gondolhatunk olyan szisztematikus hatásra, amely a valóságban (teljes sokaságban) is érvényesül.

A gyakran használt egymintás próba az alábbi próbafüggvénnyel végezhető el:

$$z = \frac{\overline{X} - m_0}{\frac{s}{\sqrt{n}}}$$

Az alábbi példa a hipotézis-ellenőrzés metodikájának megértését segíti.

Egy nagyszabású nemzetközi verseny előkészítése során a diszkoszvetés limitszintjének megállapításához kiegészítő információkat is felhasználnak. Hosszú évek tapasztalataiból ismerik, hogy a diszkoszvetés eredményeinek átlagos értéke 60 méter. Az előkészítés során véletlen módszerrel kiválasztottak 100 versenyzői eredményt, ahol 64 méteres átlagos értéket és 20 méteres szórást állapítottak meg.

Elfogadhatjuk-e azt a feltevést, hogy a diszkoszvetés átlagos értéke a 60 métert nem haladja meg, tehát ez alatt az érték alatt kell a dobási minimumszintet megállapítani?

$H_0 : \overline{X} = 60$

$H_1 : \overline{X})60$

A nullhipotézis szerint feltesszük, hogy várható dobások átlaga megegyezik a várható értékkel, míg az alternatív hipotézisben azt fogalmazzuk meg, hogy ez az átlagos érték nagyobb lehet 60 méternél.

A hipotézis ellenőrzését a normális eloszlásra alapozhatjuk, mivel tudjuk, hogy a mintaátlagok normális eloszlású valószínűségi változókként viselkednek. Ennek alapján standardizálva a változót (a mintaátlagot) a standard normális eloszlás megfelelő elméleti értékéhez viszonyíthatjuk. A próbafüggvény tulajdonképpen a standardizálás elvén alapul:

$$z = \frac{64 - 60}{\frac{20}{\sqrt{100}}} = \frac{4}{2} = 2$$

Mivel hipotézisünk egyoldalú (amely az alternatív hipotézisben fogalmazódik meg), a standard normális eloszlás sűrűségfüggvényének elegendő csupán a pozitív oldalát tekinteni. Amennyiben 5%-os szignifikanciaszintet elegendőnek tartunk – ami a gyakorlatban egy elfogadott szint – a **KH002** *z* változó táblabeli értéke: számított érték a táblabeli értéket jelentősen meghaladja, ezért a nullhipotézist *(x = 60 m)* 5%-os szignifikanciaszinten elvetjük és az alternatív hipotézist fogadjuk el. Tehát várhatóan nagyobb lesz a dobások távolságának az átlaga a versenyen mint 60 méter. Azt is mondhatjuk, hogy a hipotetikus érték (60 méter) és a mintabeli átlag (a 64 méter) közötti 4 méter nagyságrendű eltérése nemcsak véletlen tényezőkkel, hanem szisztematikus okokkal magyarázható.

A számítási eljárás megegyezik az ún. kétoldalú hipotézis-ellenőrzés esetén is, de az értékelés eltérő. Ilyen esetben az alternatív hipotézis:

$H_1: X \neq 60$

A fenti alternatív hipotézis esetén a sűrűségfüggvény mindkét oldalát figyelembe kell venni, így a kritikus érték (5%-os szignifikanciaszinten): ± 1,96. Ehhez viszonyítva is el kell vetni a 60 méterre vonatkozó hipotézist.

Mindkét megoldás arra hívja fel a figyelmet, hogy a limitszintet érdemes 60 méter felett meghatározni.

A gyakorlatban sok esetben nincs lehetőség arra, hogy nagyobb elemű minta segítségével ellenőrizzük a hipotéziseket. Amennyiben a minta elemszáma nem éri el a 30-at, akkor ún. **kis mintával** kell dolgoznunk. Kis minta esetén a standard normális eloszlás nem alkalmazható, ilyenkor a **KH003 Student-féle t-eloszlás**t és ennek az eloszlásnak a táblázatát kell alkalmaznunk. A t-eloszlás alkalmazása során figyelembe kell venni az ún. **szabadságfok**ot, amely a minta elemszámának 1-gyel csökentett értéke.

Módosítsuk előző példánkat!

Tételezzük fel, hogy csak 16 sportoló eredményét ismerjük. Az átlagos érték a mintában 64 méter, a korrigált szórás azonban csak 10 méter. Mivel a mintánk kis minta – nem haladja meg a 30-at –, a t-eloszlást kell alkalmazni. Az alkalmazáshoz azonban előfeltételként rögzíteni kell, hogy a diszkoszdobások általában normális eloszlást követnek. A próbafüggvény lényegesen nem tér el a korábban megismerttől.

$$H_0 : \overline{X} = 60$$

$$H_1 : \overline{X}) 60$$

$$t = \frac{64 - 60}{\frac{10}{\sqrt{16}}} = 1{,}6$$

A **KH003** t-eloszlás kritikus értéke 5%-os szignifikanciaszinten 15 szabadságfok mellett 1,753.

Mivel a számított érték kisebb mint a táblabeli, ezért nincs okunk arra, hogy a nullhipotézist elvessük. A mintabeli és az elvárt érték közötti eltérést – 5%-os szignifikanciaszinten – a véletlen okozhatta. Elfogadhatjuk azt a feltevést, amely szerint a 60 méter körüli érték alkalmas a szinthatár megállapítására.

Hasonlóan kell eljárnunk, ha nem az alapsokasági átlagra, hanem az **alapsokasági arányra** vonatkozóan fogalmazunk meg feltevést. Itt is meg kell jegyeznünk, hogy csak **nagy minta** esetén használható a tesztelésre a standard normális eloszlás.

$$z = \frac{p - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$$

Egy választókörzetben azt szeretnék tudni, hogy a következő választáson megjelenik-e majd a szavazásra jogosultak 40%-a. A vizsgálat érdekében 200 főt kérdeztek meg egy egyszerű véletlen kiválasztás alapján, akik közül 68 fő igennel válaszolt a kérdésre:

„Részt vesz-e a választáson?”

Vizsgáljuk meg, hogy elvárható-e a következő választáson a 40%-os részvétel!

H_0:P = 0,4

H_1:P < 0,4

Az alternatív hipotézisben azt a feltevést fogalmaztuk meg, amely szerint a választópolgárok 40%-ánál kisebb lesz a részvételi arány.

$$p = \frac{68}{200} = 0,34$$

$$z = \frac{0,4 - 0,34}{\sqrt{\frac{0,4 \times (1 - 0,4)}{200}}} = -1,732$$

A táblabeli érték **KH002** (a negatív oldalt figyelembe véve) –1,645.

Abszolút értékben a számított érték nagyobb mint a táblabeli érték, ezért nincs okunk a nullhipotézist 5%-os szignifikanciaszinten elfogadni, tehát a részvételi arány feltehetően nem fogja elérni a 40%-ot.

Ugyancsak hipotézisellenőrzés segítségével vizsgálható, ha feltevésünket két alapsokaság pl. várható értékének azonosságára fogalmazzuk meg.

Két alapsokaság esetén a próbafüggvény természetesen módosul, amit az alábbi példa segítségével mutatunk be. Itt kell megjegyeznünk, hogy az alkalmazható eloszlás a z-eloszlás, ha nagy mintánk van, illetve ha kis minta esetén ismerjük az alapsokasági szórásokat és feltételezzük a normalitást. A próbát az alábbi függvény segítségével végezhetjük el:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Az óriás lesiklás pályájának kijelölése során értékelték a pálya felső és alsó szakaszán elért eredményeket. 50-50 elemű megfigyelés (lesiklás) alapján megállapították, hogy a pálya felső részét átlagosan 21 másodperc, míg az alsó szakaszát 19 másodperc alatt teljesítik a versenyzők. A felső szakaszon általában a részidők szórása 6 másodperc, az alsón 7 másodperc.

Ellenőrizzük, hogy szinginifkánsan gyorsabb-e az alsó pályaszakasz mint a felső!

$$H_0 : \overline{X}_1 = \overline{X}_2$$

$$H_1 : \overline{X}_1 > \overline{X}_2$$

$$z = \frac{21 - 19}{\sqrt{\frac{6^2}{50} + \frac{7^2}{50}}} = 1,53$$

5%-os szignifikanciaszinten – mint már láttuk – a kritikus érték **KH002**, z = 1,645. Mivel a számított érték nem haladja meg a kritikus értéket, a fenti szignifikanciaszint mellett elfogadhatjuk a nullhipotézist, tehát nem lasúbb a felső pályaszakasz.

Az alapsokasági varianciák ismeretének hiánya más megoldást igényel. Ilyenkor fel kell tételezni a két sokaság szórásának azonosságát, ami szintén hipotézis-ellenőrzés segítségével ellenőrizhető.[5] Az ilyen típusú problémákra általában ún. kis minták esetén (az elemszám kisebb mint 30) kerül sor a gyakorlatban, ezért mi is az ilyenkor használható **kétmintás t-próbá**t mutatjuk be, ami a **KH003** t-eloszlásra épít.

Az alkalmazható próbafüggvény:

[5]A szórások azonosságára vonatkozó próbák ismertetésétől tananyagunkban eltekintünk.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

A szabadságfok: $n_1 + n_2 - 2$

Ebben az ún. közös szórás (s_p) négyzetének képlete:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}$$

A fentieket világítsuk meg az alábbi példa segítségével!

Két hasonló sportágban szeretnénk vizsgálni a még nem doppingnak minősülő szerek felhasználását, és ebből a szempontból a sportágak különbözőségét. Köztudott, hogy teljes körű adatfelvételekre ilyen esetekben nincs mód, csupán a következtetéses statisztika eszközeivel lehet körvonalazni a vizsgált problémát. Kiinduló feltétel, hogy a két sportágban azonos a szerek használatának mértéke. A feltevés ellenőrzése céljából két független véletlen mintát vettünk. Az „A” sportágat 25 elemű minta reprezentálja, amelyben a havi átlagos kiegészítő szerfogyasztás sportolóként átlagosan 2200 Ft/hó, 400 Ft/hó szórás mellett. A „B” sportág esetében a 20 elemű mintában az átlag: 2400 Ft/hó, a szórás 300 Ft/hó.

Azonosnak tekinthető-e a két sportág a kiegészítő szerek fogyasztása szempontjából? (Feltételezzük, hogy a két alapsokaságban a szórások azonosak!)

$$H_0: \overline{X}_1 = \overline{X}_2$$

$$H_1: \overline{X}_1 < \overline{X}_2$$

$$s_p^2 = \frac{(25-1) \times 0{,}4^2 + (20-1) \times 0{,}3^2}{25+20-2} = 0{,}129$$

$$s_p = \sqrt{0{,}129} = 0{,}359$$

$$t = \frac{2{,}2 - 2{,}4}{0{,}359 \times \sqrt{\frac{1}{25} + \frac{1}{20}}} = -1{,}86$$

A **KH003** t-eloszlás táblabeli értéke 43-as szabadságfok esetén,[6] mivel egyoldalú a hipotézisünk, 1,684.

Mivel a t-eloszlásról tudjuk, hogy szimmetrikus, a számított és táblabeli abszolút értékek összevetése alapján azt mondhatjuk, hogy 5%-os szignifikanciaszinten nem azonos a két alapsokasági átlag, vagyis a két sportágban a kiegészítő szerek fogyasztása különbözik; a „B” sportágban a sportolók általában többet költenek a szerek fogyasztására.

Említettük, hogy a fent ismertetett kétmintás t-próbát általában kis minták esetében alkalmazzák. Nem követünk el nagy hibát azonban, ha nagyobb minta esetén is ezt a próbát[7] használjuk.

Már az eddigiek során gyakran hivatkoztunk az eloszlások típusára, amelynek ismerete egy-egy matematikai–statisztikai eljárás alkalmazásának egyik előfeltétele. A hipotézis-ellenőrzés módszerei lehetőséget adnak arra is, hogy az eloszlásokat azonosítsuk. Az eloszlások illeszkedésének vizsgálata, valamint a további próbák bemutatása azonban már túlmutat könyvünk keretein.

4. Ellenőrző feladatok, gyakorló példák a fejezethez

- A labdarúgó-bajnokság megkezdése előtt egy közvélemény-kutató cég szimpátiavizsgálatot végez a Ferencváros megbízásából. Korábbi tapasztalatok

[6] Mivel a táblázatban KH003 a 43-as szabadságfok értékét nem találjuk, a hozzá legközelebb eső 40-es szabadságfok kritikus értékét használjuk fel.

[7] Amennyiben a t-eloszlás nagyobb szabadságfokú értékeit a standard normális eloszlás hasonló adataival összevetjük, szembetűnő a hasonlóság.

alapján a kutatást kor szerint rétegzett minta alapján kívánják elvégezni. Ismeretes a válaszadók megoszlása (24%-a 18 és 30 év közötti; 25%-a 31 és 45 év közötti; 27%-a 46 és 60 év közötti). Az 1200 elemű, kor szerint rétegzett mintavétel fontosabb eredményeit tartalmazza a következő táblázat:

  - Becsülje meg 95%-os megbízhatósággal a Ferencvárosra szavazók arányát!

- Teniszezők első és második szerváinak eredményességére vonatkozik a vizsgálatunk. 1000-1000 megfigyelés alapján az első szerva 68%-ban, a második 62%-ban volt eredményes.
  - Azonos-e a szervák hatékonysága? Számolja ki a leggyakrabban előforduló értéket!
  - Van-e 10%-os hatékonyságkülönbség az első és a második szerva között?

20. fejezet - Táblázatok

20.1. táblázat - Standard normális eloszlásfüggvény valószínűségi (szignifikancia-) értékei

u	0	1	2	3	4	5	6	7	8	9
0,0	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
0,1	0,460	0,456	0,452	0,448	0,444	0,440	0,436	0,433	0,429	0,425
0,2	0,421	0,417	0,413	0,409	0,405	0,401	0,397	0,394	0,390	0,386
0,3	0,382	0,378	0,374	0,371	0,367	0,363	0,359	0,356	0,352	0,348
0,4	0,345	0,341	0,337	0,334	0,330	0,326	0,323	0,319	0,316	0,312
0,5	0,309	0,305	0,302	0,298	0,295	0,291	0,288	0,284	0,281	0,278
0,6	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
0,7	0,242	0,239	0,236	0,233	0,230	0,227	0,224	0,221	0,218	0,215
0,8	0,212	0,209	0,206	0,203	0,200	0,198	0,195	0,192	0,189	0,187
0,9	0,184	0,181	0,179	0,176	0,174	0,171	0,169	0,166	0,164	0,161
1,0	0,159	0,156	0,154	0,152	0,149	0,147	0,145	0,142	0,140	0,138
1,1	0,136	0,133	0,131	0,129	0,127	0,125	0,123	0,121	0,119	0,117
1,2	0,115	0,113	0,111	0,109	0,107	0,106	0,104	0,102	0,100	0,099
1,3	0,097	0,095	0,093	0,092	0,090	0,089	0,087	0,085	0,084	0,082
1,4	0,081	0,079	0,078	0,076	0,075	0,074	0,072	0,071	0,069	0,068
1,5	0,067	0,066	0,064	0,063	0,062	0,061	0,059	0,058	0,057	0,056
1,6	0,055	0,054	0,053	0,052	0,051	0,049	0,048	0,047	0,046	0,046
1,7	0,045	0,044	0,043	0,042	0,041	0,040	0,039	0,038	0,038	0,037
1,8	0,036	0,035	0,034	0,034	0,033	0,032	0,031	0,031	0,030	0,029
1,9	0,029	0,028	0,027	0,027	0,026	0,026	0,025	0,024	0,024	0,023
2,0	0,023	0,022	0,022	0,021	0,021	0,020	0,020	0,019	0,019	0,018
2,1	0,018	0,017	0,017	0,017	0,016	0,016	0,015	0,015	0,015	0,014
2,2	0,014	0,014	0,013	0,013	0,013	0,012	0,012	0,012	0,011	0,011
2,3	0,011	0,010	0,010	0,010	0,010	0,009	0,009	0,009	0,009	0,008
2,4	0,008	0,008	0,008	0,008	0,007	0,007	0,007	0,007	0,007	0,006
2,5	0,006	0,006	0,006	0,006	0,006	0,005	0,005	0,005	0,005	0,005
2,6	0,005	0,005	0,004	0,004	0,004	0,004	0,004	0,004	0,004	0,004
2,7	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
2,8	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
2,9	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,001	0,001	0,001
3,0	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001

20.2. táblázat - A z-statisztika fontosabb értékei

Szignifikancia-szint (α)						
Egyoldalú	0,1000	0,0500	0,0250	0,0225	0,0100	0,0050
Kétoldalú	0,2000	0,1000	0,0500	0,0450	0,0200	0,0100
z	1,28	1,64	1,96	2,00	2,33	2,58

20.3. táblázat - A Student-féle t-eloszlás kritikus értékei adott szignifikancia szinten

Szabadságfok	Szignifikancia szint				
0,1	**0,05**	**0,025**	**0,01**	**0,005**	
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	3,132	2,776	3,747	4,604
5	1,476	20,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,769	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,771	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
70	1,294	1,667	1,994	2,381	2,648
80	1,292	1,664	1,990	2,374	2,639
90	1,291	1,662	1,987	2,368	2,632
100	1,290	1,660	1,984	2,364	2,626

Szabadságfok	Szignifikancia szint				
150	1,287	1,665	1,976	2,351	2,609
200	1,286	1,653	1,972	2,345	2,601

Irodalom

ÁCS PONGRÁC: A sport területi koncentrációja, Pécsi Tudományegyetem, Sport és a Tudomány napja konferenciakötet, Pécs, 2006.

Ács Pongrác: A sportolói migráció és annak lehetőségei az EU-csatlakozásunk tükrében, Pécsi Tudományegyetem Közgazdaságtudományi Kar, Regionális Politika és Doktori Iskola Évkönyv 2004–2005 I. kötet, Pécs, 2005.

ÁCS PONGRÁC: A területi egyenlőtlenségek feltérképezése során leggyakrabban alkalmazott mérőszámok bemutatása a sporttehetségek területi elhelyezkedésének példáján, Egy életpálya három dimenziója – Tanulmánykötet Pintér József emlékére (ISBN 978-963-642-195-3), Pécsi Tudományegyetem Közgazdaságtudományi Kar, Pécs, 10–22. o.

ÁCS PONGRÁC: Nemzetközi és hazai sportgazdasági trendek, Sportszakember-továbbképzési konferenciasorozat kiadványa (ISBN: 978-963-88695-0-0), Nemzeti Sportszövetség, Budapest, 25–33. o.

BYRKIT, D. R.: Statistics today, The Benjamin/Cummings Publishing Company, Inc. , Menlo Park, California, 1987. 850 o.

Dr. Frenkl Róbert: Sportélettan, Magyar Testnevelési Egyetem, Budapest 1995.

Életminőség és egészség, Központi Statisztikai Hivatal, Budapest, 2002.

HAJDU OTTÓ–PINTÉR JÓZSEF–RAPPAI GÁBOR–RÉDEY KATALIN: Statisztika I., Janus Panonius Tudományegyetem, Pécs, 1994.

HEALEY, J. F.: Statistics, a tool for social research, Wadsworth Publishing Company, Belmont, California, 1984, 351 o.

HERMAN SÁNDOR–PINTÉR JÓZSEF–RAPPAI GÁBOR–RÉDEY KATALIN: Statisztika II., Janus Pannonius Tudományegyetem, Pécs, 1994.

HUNYADI L.–MUNDRUCZÓ GY.–VITA L.: Statisztika, AULA Kiadó, Budapest, 1996.

HUNYADI L.–VITA L.: Statisztika közgazdászoknak, KSH. Budapest, 2002.

KERÉKGYÁRTÓ GYÖRGYNÉ–MUNDRUCZÓ GYÖRGY–SUGÁR ANDRÁS: Statisztikai módszerek és alkalmazásuk a gazdasági, üzleti elemzésekben, Aula Kiadó, Budapest, 2001.

KORINEK LÁSZLÓ–PINTÉR JÓZSEF–SZŰCS ANDRÁSNÉ: Statisztika I. rész, Tankönyvkiadó, Budapest, 1976.

KÖVES PÁL–PÁRNICZKI GÁBOR: Általános statisztika I-II., Tankönyvkiadó, Budapest, 1981.

PINTÉR JÓZSEF: Bevezetés a statisztika módszereibe, Pécsi Tudományegyetem, Pécs, 2000.

RAPPAI GÁBOR: Üzleti statisztika Excellel, KSH, Budapest, 2001.

STATISZTIKAI ÉVKÖNYVEK

STATISZTIKAI HAVI KÖZLEMÉNYEK

VAN MATRE, J. G.–GILBREATH, G. H.: Statistics for Business and Economics, Business Publications, Inc. Plano, Texas, 1983.

www.mob.hu

www.nemzetisport.hu

A. függelék - Név- és tárgymutató