

Statisztika II.

Szerkesztette:
Makszim Györgyné dr. Nagy Tímea

A tananyag elkészítését a „NYE-DUÁL- Új utakon a duális felsőoktatással a Nyíregyházi Egyetemen, az Északkelet-Magyarországi térség felemelkedéséért” az EFOP-3.5.1-16-2017-00017 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

SZÉCHENYI 2020



MAGYARORSZÁG
KORMÁNYA

Európai Unió
Európai Szociális
Alap



BEFEKTETÉS A JÖVŐBE

Szerkesztő:

Makszim Györgyné dr. Nagy Tímea

Szerző:

Makszim Györgyné dr. Nagy Tímea

Lektor:

Kézirat lezárva: 2019. 09. 20.

ISBN 978-615-6032-27-0

Kiadja.....

TARTALOMJEGYZÉK

TARTALOMJEGYZÉK	3
TÁBLÁZATOK JEGYZÉKE	5
ÁBRÁK JEGYZÉKE	6
FELHASZNÁLT SZAKIRODALOM.....	7
1 STATISZTIKAI MINTAVÉTEL.....	8
1.1 A minta kiválasztása	9
1.2 A véletlen kiválasztás módszerei.....	11
1.2.1 Egyszerű véletlen kiválasztás.....	11
1.2.2 Rétegzett kiválasztás.....	12
1.2.3 Lépcsőzetes kiválasztás.....	12
1.3 Kiválasztási arány	12
1.4 Önellőrző kérdések	13
2 STATISZTIKAI BECSLÉS.....	14
2.1 Várható értékre vonatkozó becslés.....	20
2.1.1 A rétegzett minta értékelése	24
2.2 Értékösszeg becslése.....	25
2.3 Arány vagy valószínűség becslése.....	25
2.4 Mintafeladatok a statisztikai becslés témaköréhez	27
2.5 Önellőrző kérdések	36
3 HIPOTÉZISVIZSGÁLAT	38
3.1 Paraméterre vonatkozó próbák	42
3.1.1 Várható értékre vonatkozó egymintás próbák.....	42
3.1.2 Várható értékre vonatkozó kétmintás próbák	44
3.2 Eloszlásokra vonatkozó próbák	46
3.2.1 Illeszkedésvizsgálat.....	46
3.2.2 Homogenitásvizsgálat	48
3.2.3 Függetlenségvizsgálat.....	49
3.3 Mintafeladatok a hipotézisvizsgálat témakörhöz.....	50
3.4 Önellőrző kérdések	58
4 ÖSSZEFÜGGÉS-VIZSGÁLAT.....	59
4.1 A mennyiségi ismérvek közötti kapcsolatok vizsgálata	60
4.1.1 A lineáris kétváltozós korrelációs kapcsolat mérőszámai.....	62
4.1.2 Lineáris kétváltozós regresszió analízis.....	67
4.1.3 Nem lineáris kétváltozós regresszió analízis	73
4.1.4 Többváltozós korrelációs kapcsolat	74
4.2 A minőségi ismérvek közötti kapcsolatok vizsgálata	75
4.2.1 A Yule-féle asszociációs együttható.....	76
4.2.2 A Csuprov-féle asszociációs együttható	77
4.3 A vegyes jellegű kapcsolatok vizsgálata	78
4.4 Mintafeladatok az összefüggés-vizsgálat témakörhöz	80
4.5 Önellőrző kérdések	91
5 AZ IDŐSOROK ÖSSZETEVŐI	93
5.1 Az alapirányzat vagy trend meghatározása	95
5.1.1 Mozgóátlagolás.....	96

5.1.2	<i>Az analitikus trendszámítás</i>	98
5.2	<i>A periodikus ingadozások vizsgálata</i>	103
5.2.1	<i>A szezonálítás vizsgálata.....</i>	103
5.2.2	<i>A konjunkturális ingadozások vizsgálata.....</i>	104
5.3	<i>Mintafeladatok az idősor témakörhöz</i>	105
5.4	<i>Önellenőrző kérdések</i>	112
6	GYAKORLÓ VIZSGAFELADATOK	113

TÁBLÁZATOK JEGYZÉKE

1. A hipotézisvizsgálat során elkövethető hibák
2. A háromtagú mozgóátlag számítása

ÁBRÁK JEGYZÉKE

1. Gauss-görbe
2. A Gauss-görbe jellegzetes pontjai
3. Regressziós transzformáció

FELHASZNÁLT SZAKIRODALOM

1. HAJDU Ottó: *Többváltozós statisztikai számítások*. Budapest: Központi Statisztikai Hivatal Akadémiai Nyomda, 2003 –ISBN 963 215 600 5
2. HUNYADI László, MUNDRUCZÓ György, VITA László: *Statisztika*. Budapest: Aula Kiadó, 2001., IV. kiadás –ISBN 963 921 556 2
3. SZARVAS Beatrix, SUGÁR András: *Példatár a statisztika című tankönyvhöz*. Budapest: Aula Kiadó, 2001., IV. kiadás
4. MADDALA Gangadharrao Soundaryarao: *Bevezetés az ökonometriába*. Budapest: Nemzeti Tankönyvkiadó, 2004 –ISBN 963 194 111 6

1 STATISZTIKAI MINTAVÉTEL

A Statisztika I. tantárgy keretében áttekintettük a sokaság jellemzőit, az azok elemzésére szolgáló főbb módszereket. Kevés szó esett ugyanakkor arról, hogy hogyan juthatunk el az alapadatokhoz, vagy hogy milyen szabályszerűsége van a statisztikai megfigyelés folyamatának. Jelen fejezet éppen ezért a mintavétel kérdéseivel foglalkozik, ami a statisztikai következtetéselmélet kiindulópontjának is tekinthető.

A statisztikai elemzésekhez felhasználható adatok különböző forrásokból származhatnak. Ezek a források általánosan:

- statisztikai célú megfigyelések, adatfelvételek,
- adminisztrációs információforrások (adatátvétel),
- kontrollált kísérletek.

Közülük a gazdaságtudományok területén a legfontosabb adatszerzési mód a **statisztikai megfigyelés**, amely a statisztikai munka legfontosabb mozzanata, de elválaszthatatlan annak minden egyes szakaszától. A megfigyelés sikerességét alapvetően meghatározza a megelőző **programkészítés**, amely a vizsgálati cél megjelölését, az eredmények közzétételét, a megfigyelt adatok feldolgozását, az adatgyűjtés módjának a konkretizálását és az adatfelvétellel kapcsolatos szervezési munkálatokat veszi sorra. Fontos tervezési szempont az adatgyűjtés megtervezése, különösen az adatgyűjtési körre majd a megfigyelési módra vonatkozó elhatározás.

A megfigyelés köre szerint elsődlegesen **teljes körű** és **részleges** megfigyelést különböztetünk meg.

A teljes körű megfigyelés a vizsgálni kívánt sokaság minden egyes egyedére kiterjed (pl. népszámlálás), míg a részleges megfigyelés a vizsgált sokaságnak csak egy kiválasztott részét öleli fel. A statisztikai következtetéselmélet alapvető célja, hogy a részleges megfigyelések eredményeiből következtessen a teljes sokaságra.

Teljes körű adatfelvétel csak véges sokaság esetén alkalmazható. Mivel igen idő- és költségigényes, ezért csak ritkán, vagy kis elemszámú sokaság esetén kerül rá sor. Éppen ezért legtöbbször a részleges megfigyelés módszerét alkalmazzuk, ahol az alapsokaságból vett minták számbavételére kerül sor, melynek eredményeként mintasokaságot kapunk.

Mintának nevezzük a sokaságnak azt a részét, amelyre az adatgyűjtés kiterjed. A minta mindig véges, és elemszámát n -nel jelöljük. A mintát mindig elemeinek felsorolásával adjuk meg. A mintával

kapcsolatban fontos fogalom a kis ($n \leq 30$) és a nagy ($n > 30$) minta. A mintaelemek változók, hiszen értékük mintáról mintára változik. A mintaelemek mellett a mintajellemzők, így a minta átlaga, szórásnégyzete, stb. is változók. A különböző mintákból számított mintaátlagokkal kapcsolatban a következő általánosítható következtetéseket vonhatjuk le:

- Az egyes mintákból számított mutatók (átlagok) értéke mintáról mintára változik;
- A mintákból számított mutatók (átlagok a sokasági mutató (átlag) körül szóródnak;
- Ez a szóródás a nagyobb minták esetén kisebb, a kisebb minták esetén kisebb.

A statisztikai megfigyelés mindig a sokaság egységeinek számba vételére irányul. A megfigyelési egységek adatait általában **kérdőíveken** gyűjtjük be. A kérdőív egy előzetesen elkészített „nyomtatvány”, amely kérdéseket, lehetséges válaszokat, az adatfeldolgozást segítő jelzéseket tartalmaz, kiegészítve az adatgyűjtő hivatalos jellegét igazoló tartalmi elemekkel és útmutatókkal. A kérdőív lehet egy egységre vonatkoztatott **egyéni kérdőív** és több megfigyelési egységre kiterjedő **lajstrom**. A begyűjtött adatok hasznosíthatóságának előfeltétele a jól szerkesztett kérdőív, ezért ezzel a témakörrel külön tudományterületek foglalkoznak.

A kérdőíveket vagy **közvetlen megfigyeléssel** vagy **kikérdezéssel** módszerrel a számlálóbiztosok tölthetik ki, vagy **önszámlálás útján** maga az adatszolgáltató tölti ki.

Az utóbbi időben elterjedt a telefonos kikérdezés és a postai levelezés útján történő adatgyűjtés, amelynek a hatékonysága eléggé alacsony.

Az adminisztrációs információforrások általában teljes körű megfigyelési adatokat szolgáltatnak (pl. a demográfiai véleményekről és folyamatokról), a kontrollált kísérlet pedig reprezentatív sokaságot biztosít az értékeléshez.

1.1 A minta kiválasztása

A mintavétel két fő lépésre osztható:

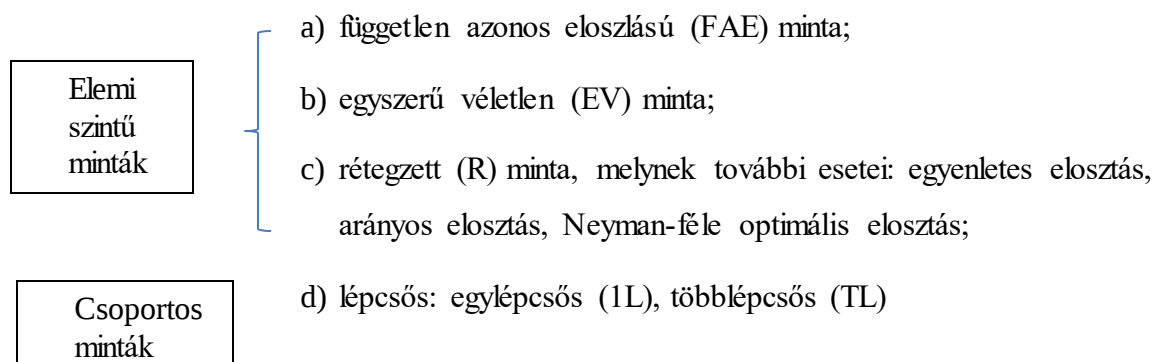
1. a mintavétel **megtervezése** és
2. a kijelölt **mintaelemek megfigyelése**.

- **Nem reprezentatív:** nem lehet belőle általánosítani, a következtetések kizárólag a megfigyelt egyedekre vonatkoznak.
- **Reprezentatív:** alkalmas a sokaságra vonatkozó általános következtetések levonására. A reprezentatív megfigyeléssel nyert minta képezi az alapadathalmazát a következtető statisztikának.

A gyakorlatban számtalan olyan részleges megfigyelési móddal is találkozunk, amelyek nem biztosítanak reprezentatív mintát, így értékelésük eredményei sem általánosíthatók.

A mintavételi eljárásoknak három csoportját különböztetjük meg:

1. véletlen mintavételi tervek/eljárások:



2. nem véletlen mintavételi tervek/eljárások:

- a) szisztematikus,
- b) kvótás,
- c) koncentrált,
- d) hólabda,
- e) önkényes (szubjektív) kiválasztás.

3. Kombinált és mesterséges minták

Fontos kiemelni, hogy a reprezentatív mintavétel főként véletlen kiválasztáson alapul. A **véletlen kiválasztás**, vagy **véletlen mintavétel** azt jelenti, hogy a minta kiválasztásakor biztosítjuk, hogy minden sokasági elem előre meghatározott valószínűséggel kerüljön be a mintába. A véletlen mintát más néven valószínűségi mintának nevezzük. A valószínűségi minta elemei és jellemzői valószínűségi változók lesznek.

A mintavétel lehet továbbá:

- **Ismétléses:** a sokaság egyes elemei ismételten szerepelhetnek a mintában.
- **Ismétlés nélküli:** a már kiválasztott egységek kirekesztése a mintavétel további folyamatából. (Ez az eljárás azonban megsérti a függetlenség követelményét, ennek ellenére a gyakorlatban mégis sokszor kerül sor az ismétlés nélküli kiválasztásra.)

A kiválasztást természetesen az is befolyásolja, hogy a sokaság valószínűségi, vagy nem valószínűségi, illetve, hogy véges, vagy végtelen. (A gyakorlatban a valószínűségi és véges sokaság, a reprezentatív megfigyelés, valamint legtöbbször az ismétlés nélküli kiválasztás esetét alkalmazzuk.)

1.2 A véletlen kiválasztás módszerei

A következőkben a véletlen kiválasztáson alapuló mintavételi tervek fontosabb jellemzőit szeretném bemutatni.

1.2.1 Egyszerű véletlen kiválasztás

- Mikroszintű;
- Homogén és végtelen vagy nagyon nagy elemszámú sokaság esetén alkalmazható;
- A minta kiválasztása történhet visszatevéssel vagy visszatevés nélkül;
- Minden lehetséges minta kiválasztásának valószínűsége azonos.
- FAE mintához vezet az is, amikor véges sokaságból egyenlő valószínűséggel veszünk visszatevéssel mintát.

A kiválasztás technikája is különböző lehet. Leggyakoribb:

- Sorsolás;
- Véletlen számtáblázat módszere;
- Véletlen számok gerjesztése, generálása;

- Véletlen koordináták módszere. (Az alapsokaság területéhez kötődik. A véletlen hiba elkövetésének valószínűsége nagy.)

1.2.2 Rétegzett kiválasztás

Legfontosabb jellemzői:

- Mikroszintű;
- Heterogén sokaság esetén alkalmazható;
- Először a heterogén főszokaságot valamilyen ismerv szerint átfedés-mentesen homogén részekre, rétegekre osztjuk, majd az egyes rétegeken belül egymástól függetlenül egyszerű véletlen mintát veszünk;
- Megkülönböztetünk arányos, illetve nem arányos rétegzést. (Általában az arányos rétegzést alkalmazzuk.)

1.2.3 Lépcsőzetes kiválasztás

A lépcsőzetes kiválasztás makroszintű kiválasztást jelent. Lépcsőzetes megfigyelés esetén több lépcsőben jutunk el a végső megfigyelési egységhez. Ezután ennél a végső megfigyelési egységnél teljes körű megfigyelést végzünk.

Lehet: egy-, vagy többlépcsős minta.

1.3 Kiválasztási arány

A kiválasztás módszerének megválasztása mellett a minta fontos jellemzője a **kiválasztási arány** meghatározása is, azaz, hogy az alapsokaság hányad része kerül a mintába.

Kiválasztási arány jele: f

Meghatározása: $f=n/N$

ahol: n –mintaelemek száma

N –az alapsokaság elemszáma

5%-nál nagyobb kiválasztási arány igen ritka.

1.4 Önellenző kérdések

1. Rendszerezze a megfigyelés módszereit a megfigyelés köre alapján!
2. Mit nevezünk mintának!
3. Mit nevezünk egyéni kérdőívnek, illetve lajstromnak!
4. A statisztikai megfigyelés lépései!
5. Mikor beszélünk reprezentatív, valamint nem reprezentatív mintáról!
6. Mit értünk véletlen kiválasztás alatt!
7. Rendszerezze a mintavételi módokat!
8. Jellemezze az egyszerű véletlen kiválasztás módszerét!
9. Jellemezze a rétegzett kiválasztás módszerét!
10. Mit értünk lépcsőzetes kiválasztás alatt!
11. Mit nevezünk kiválasztási aránynak!

2 STATISZTIKAI BECSLÉS

A véletlen kiválasztással nyert minták értékelésének egyik fontos módszere a statisztika becslés.

A statisztikai sokaság valamely jellemzőjét **paraméternek** nevezzük. Ilyen jellemző lehet pl. a várható érték, a szórás, stb.

Az alapsokaságból nyert minták számított mutatói alapján történő sokasági jellemzők meghatározását **statisztikai becslésnek** nevezzük. Ennek eszköze a mintából származó megfigyelések valamely függvénye, a **becslőfüggvény**. Azt mutatja meg, hogy a mintákból nyert információk alapján hogyan határozható meg a sokasági jellemző:

- a mintából számított érték a **becsült érték (mintaátlag)**, tapasztalati statisztikai jellemző,
- az összes lehetséges minta tapasztalati statisztikai jellemzőinek átlaga a **várható érték**. Az **átlagoknak az átlaga megegyezik a sokasági átlaggal**. Ez azt is jelenti, hogy a mintaátlagok a sokasági átlag körül ingadoznak, ingadozásuk középpontjában pedig a sokasági átlag áll. Minél nagyobb a sokasági szórás, annál nagyobb az ingadozás, és minél nagyobb a minta elemszáma, annál kisebb ez az ingadozás, azaz annál nagyobb lesz a pontosság. A mintavételi ingadozások mérőszáma a mintavételi variancia vagy ennek négyzetgyöke, a mintavételi szórás, amit **standard hibának** nevezünk.

A becslőfüggvénnyel szemben támasztott követelmények:

- **Torzítatlanság:** torzítatlannak nevezzük a becslőfüggvényt, ha annak várható értéke megegyezik a becsülni kívánt sokasági jellemzővel.
- **Hatásosság:** két becslő formula közül azt tekintjük hatásosabbnak, amelynek kisebb a szórása (standard hibája).
- **Konzisztencia:** a konzisztencia követelménye azt írja elő, hogy a becslés torzítatlan legyen, és a mintanagyság minden határon túli növelése esetén szórásnégyzete a 0-hoz tartson.

A becslés célja lehet:

- **Pontbecslés:** egy konkrét mintából kapott érték.

- **Intervallumbecslés:** megjelölünk egy olyan sávot, amely nagy valószínűséggel közrefogja a sokasági paramétert. A határok által kijelölt sávot konfidencia intervallumnak nevezzük.

A pontbecslés módszerei:

- Legkisebb négyzetek módszere
- Momentumok elve
- Maximum likelihood módszer

A statisztikai adatok pontossága:

A statisztikai adatok fontos információkat hordoznak, ezért elvárjuk tőlük, hogy:

- pontosak legyenek, a megfigyelési céloknak megfelelő, valódi értéket közvetítsenek,
- időben álljanak rendelkezésre, mivel a késedelem az információ aktualitását csökkenti,
- az adatszerzés és az információ feltárás gazdaságos legyen, amihez a pontosság, az időbeliség és az olcsóság együttes optimumát kell biztosítanunk.

Az elvárások ellenére azt kell mondanunk – és a statisztikai elemzések etikájához az hozzátartozó -, hogy a statisztikai adatok **korlátozott pontosságú** értékek. Ez a megállapításunk mind a teljes körű, mind pedig a részleges megfigyelésekre igaz, hiszen az adatokat mérés, számlálás útján nyerjük, miközben tévedhetünk, vagyis **szubjektív hibát** követhetünk el. Ez a hiba elméletileg teljesen kiküszöbölhető lenne, de igazából, közvetlenül meg sem tudjuk állapítani, hogy elkövettük. Utólagosan tudjuk becsülni a nagyságát, ami idővesztéssel és többletköltséggel járna. Ezért az adataink pontosságának jellemzése céljából bevezetjük a **szignifikáns számjegyek** fogalmát. Ennek megállapítása a vizsgált sokaság nagyságától, annak jellegétől és a vizsgálati területen szerzett előzetes tapasztalatoktól függ. A pontosságot úgy jellemezzük, hogy **megadjuk az adat utolsó szignifikáns számjegyének a helyi értékét**, majd az utolsó szignifikáns helyi értéken található számjegyet az előtte álló számjegy figyelembe vételével kerekítjük. A nem szignifikáns számjegyek helyét nullákkal helyettesítjük, vagy a mértékegységet választjuk meg úgy, hogy a közölt adat csupa szignifikáns számjegyből álljon.

Pl. egy régióban 2019. május 31-én 167632 db szarvasmarhát számláltak. Az utolsó szignifikáns helyi értékül az ezrest választjuk, így a közlésre kerülő adat:

168000 db,

illetve 168 ezer db

Azáltal, hogy rögzítettük az utolsó szignifikáns helyi értéket, egyúttal kijelöltük az adatunk megállapításánál elkövetett hiba felső határát is, amit **hibahatárnak** nevezünk. Ennek nagysága a szignifikáns helyi érték fele, példánkban ± 500 . Így a hibahatárokkal megadott adatunk:

168000 ± 500 db

Általánosságban tehát, ha az adatunk tényleges értékét A -val, az azt közelítő adatot pedig $\hat{A}$ -val jelöljük, akkor az elkövetett hiba abszolút nagysága:

$$\hat{a} = |A - \hat{A}|,$$

amit a pontos értéket közelítő érték **abszolút hibájának** nevezünk. A gyakorlatban A pontos értéke nem ismert, ezért megadjuk azt a legkisebb pozitív a értéket, amelynél az abszolút hiba biztosan nem nagyobb, amelyre nézve a

$$a = |A - \hat{A}| \leq a$$

teljesül. Az a az $\hat{A}$ abszolút hibahatára. A pontos adatunk tehát $\hat{A} - a$ és az $\hat{A} + a$ intervallumban helyezkedik el, amely intervallumot **konfidencia intervallumnak** nevezünk, és az alábbiak szerint értelmezzük:

$$\hat{A} - a \leq A \leq \hat{A} + a$$

Az abszolút hibahatár mellett szokásos még a relatív hibahatárt is megadni. Ezt α -val jelölve:

$$\alpha = \frac{a}{\hat{A}},$$

a pontos adatunk tehát:

$$\hat{A}(1 - \alpha) \leq A \leq \hat{A}(1 + \alpha)$$

intervallumban helyezkedik el.

Korlátozott pontosságú adatokkal, ha különböző műveleteket végzünk, akkor a műveletek eredményeinek a hibakorlátjait is kell jelölnünk. Jelen keretek között azonban ezzel mi nem

foglalkozunk. Megjegyezzük, hogy a mintasokaságok esetében a most értelmezett hibát mintavételi, vagy objektív hibának nevezzük, melyet elvileg soha nem tudunk megszüntetni, de nagysága becsülhető és mérsékelhető.

A minta elemeiből becslt értékek tehát a legtöbbször különböznek az alapsokaság paramétereitől, ami azt jelenti, hogy a becslésünk hibával terhelt.

A becslési hiba lehet:

- egyrészt **szubjektív/nem véletlen hiba**, a statisztikai megfigyelés rendszeres hibái, tehát nem mintavételi hiba (akár meg is szüntethetők, nagyságát nem tudjuk meghatározni),
- másrészt **objektív/véletlen hiba**, amit az idéz elő, hogy az alapsokaságnak csak egy részét figyeltük meg, tehát mintavételi hiba (elvileg soha nem tudjuk megszüntetni, de nagysága becsülhető és mérsékelhető).

A becslés vonatkozhat:

- átlagra (várható értékre),
- értékösszegre,
- arányra,
- szórásra, stb.

A mintából való következtetés elvi alapjai:

A mintából való következtetéshez elvi alapokat a valószínűség számítás törvényszerűségei, mégpedig a véletlen tömegjelenségek törvényei adnak:

- a nagy számok törvénye,
- a központi határeloszlás tétele.

A **nagy számok törvénye** szerint a megfigyelésben résztvevő egyedek számát fokozatosan növelve, a bizonyosság felé közeledik annak valószínűsége, hogy a relatív gyakoriság és a matematikai valószínűség csak egy általunk tetszőlegesen és előre meghatározható mértékben tér el

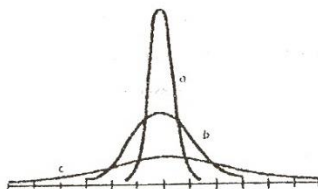
egymástól. (A nagy számok törvényét J. Bernoulli francia matematikus fogalmazta meg a XVIII. században.)

Statisztikailag fontos **Csebisev** (orosz matematikus) **tétele**, mely a nagy számok törvényének egyik továbbfejlesztett változata. E tétel szerint a megfigyelésben résztvevő egyedek számának növelésével a bizonyossághoz közeledik annak valószínűsége, hogy a megfigyelésben résztvevő, kiválasztott egyedek vizsgált tulajdonságának átlaga és az egész statisztikai sokaságban érvényesülő átlag közötti különbség tetszőlegesen kicsiny legyen. **Csebisev tételét az átlagszámok törvényének is nevezzük.**

Azon ismérvek, amelyeknek számszerű értékei a véletlen tényezők hatására ingadoznak **valószínűségi változóknak** tekinthetők. A valószínűségi változók eloszlásának különböző típusai ismeretesek. Közülük legfontosabb:

- A **binomiális eloszlás**, amelyet a diszkrét értékek követik. A binomiális eloszlás már viszonylag kis minta elemszám esetén is jól közelíthető a normális eloszlással.
- A **normális eloszlás**, amely a folytonos változók egyik eloszlástípusa. Létrejöttének feltételeit a központi határeloszlás tétele határozza meg. A **központi határeloszlás tétele** - amit Ljapunov általánosított - a következőkben összegezhető: ha a statisztikai sokaság egyedeire sok befolyásoló tényező van hatással, és egy-egy tényező hatása a többitől független, önmagában csak kis mértékben járul hozzá az összehatáshoz, akkor a valószínűségi eloszlás jó megközelítéssel normális eloszlást követ.

A normális eloszlás valószínűségi sűrűség-függvényét ábrázolva jellegzetes harang alakot kapunk, amelyet a normális eloszlás görbéjének vagy **Gauss-görbének** nevezünk. A szimmetrikus eloszlás alakja a szóródás mértékétől függően lehet laposabb és hegyesebb.

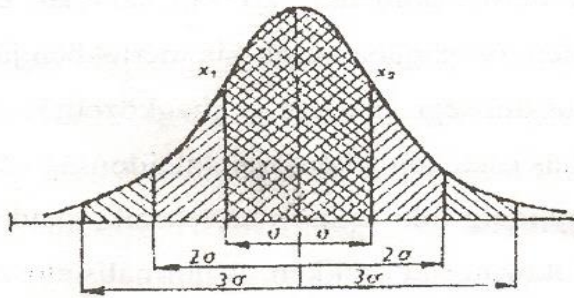


1. ábra: Gauss-görbe

Forrás: Saját szerkesztés

A normális eloszlás görbéje a következő sajátosságokat mutatja: a görbének egy csúcspontja van, mely a várható értéknek felel meg. A várható érték a tapasztalati átlag elméleti megfelelője. A harang-

görbét a csúcspontból az alapra húzott merőleges két szimmetrikus részre osztja. Ez azt jelenti, hogy a várható értéktől azonos távolságra lévő értékek valószínűsége is azonos. (Lásd. a következő ábrát.)



2. ábra: A Gauss-görbe jellegzetes pontjai

Forrás: Saját szerkesztés

A görbén három jellegzetes pontot találunk: a csúcspontot és attól balra, valamint jobbra az inflexiós pontokat (x_1 és x_2).

A szimmetria tengely és az x_1 és x_2 pontokból az alapra bocsátott merőlegesek közötti távolság nem más, mint a szórás (σ).

A normális eloszlás jellegzetességéből következik, hogy ha a valószínűségi változók normális vagy közelítőleg normális eloszlást követnek, akkor a szórás ismeretében meghatározható, hogy az átlagtól felfelé és lefelé tetszős szerint választott határok között az ismerv értékeinek hány százaléka helyezkedik el.

Eszerint az összes értéknek átlagosan 68,3%-a a várható értéktől nem tér el nagyobb mértékben, mint a szórás (σ) nagysága. Az értékek 95%-a kétszeres szórás távolságra, 99,7%-a pedig háromszoros szórás távolságra helyezkedik el a várható értéktől.

A statisztikai becslés folyamatának logikai rendje a következő:

1. A feladat értelmezése;
2. A kiválasztási arány meghatározása;
3. Döntés a kiválasztás módszeréről;

4. A mintavétel végrehajtása;
5. A mintaelemekből statisztikai sorok, táblák szerkesztése;
6. A mintavétel eredményeinek elemzése, értékelése.

Az eredmények értékelésének is van egy általános logikai sorrendje, amely a következő lépéseket foglalja magában:

1. A mintasokaság átlagának meghatározása ($\bar{x}_m$);
2. A mintasokaság szórásának meghatározása (σ_m);
3. A standard hiba kiszámítása (S_x);
4. A becslés megbízhatósági szintjének megállapítása (t vagy z);
5. A hibahatár megállapítása (Δ);
6. A megbízhatósági értékköz vagy konfidencia intervallum kijelölése ($\bar{x}_m \pm \Delta$);
7. A végeredmény értékelése.

2.1 Várható értékre vonatkozó becslés

Becslőfüggvény: a mintaátlag

Az eljárás lépései:

1. A mintasokaság átlagának meghatározása:

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

x_i -mintaelem

n-mintaelemek száma

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

f_i -gyakoriság

x_i -mintaelem

$$\bar{x}_m = a + \frac{\sum_{i=1}^k f_i d_i}{\sum_{i=1}^n f_i} \cdot i$$

a —a tetszőlegesen választott osztályköz közepe

f_i —gyakoriság

i —intervallum hossza

d_i —...-2, -1, 0, 1, 2...

Az összes lehetséges n elemű minta átlagainak gyakorisága normális eloszlást követ, átlaguk az alapsokaság átlagát adja.

2. A mintasokaság szórásának meghatározása:

$$\sigma_m = \sqrt{\frac{\sum (x_i - \bar{x}_m)^2}{n}}$$

$$\sigma_m = \sqrt{\frac{\sum f_i (x_i - \bar{x}_m)^2}{\sum f_i}}$$

$$\sigma_m = i \cdot \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i} \right)^2}$$

A mintabeli szórás értéke eltér az alapsokaság szórásától→véletlen hiba.

Korrigált szórás meghatározása:

$$S = \sqrt{\frac{\sum (x_i - \bar{x}_m)^2}{n-1}}$$

A mutatóban az (n-1) arra utal, hogy a szórás szabadságfoka (n-1). A szabadságfokban pedig az 1-es érték az előzetesen felhasznált mutatók számát jelenti. Ennek az (n-1)-es korrekciónak elsősorban akkor van jelentősége, ha a mintaelemek száma kicsi.

3. A standard hiba kiszámítása:

A mintaátlagok ingadozásának mértékét standard hibának nevezzük, amely a mintából számított jellemző átlagos négyzetes eltérését adja meg. A standard hiba a mintavételi hiba általánosan alkalmazott mérőszáma.

➤ *A standard hiba meghatározása ismétléses kiválasztás esetén:*

$$S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Mivel a σ nem ismert, ezért az S-sel helyettesítjük.

A standard hiba tehát:
$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

➤ *A standard hiba meghatározása egyszerű véletlen mintavétel, vagyis ismétlés nélküli kiválasztás esetén:*

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

ahol $1 - \frac{n}{N} \rightarrow$ korrekciós tényező $\frac{n}{N} :$ kiválasztási arány

n: a minta elemszáma

N: alapsokaság elemszáma

Ha a kiválasztási arány 5%-nál kisebb, a korrekciós tényező elhagyható.

(Belátható, hogy a standard hiba értéke akkor kisebb, ha a kiválasztás ismétlés nélküli.)

4. A becslés megbízhatósági szintjének megállapítása:

A megbízhatósági szinthez tartozó 'z' vagy 't' értéket a megfelelő eloszlás táblázatokból keressük ki. (z: standard normális eloszlású valószínűségi változó; t: Student-féle t-eloszlású valószínűségi változó) A 'z' értéket a standard normális eloszlás táblázatból, a 't' értéket pedig a Student-féle t-táblázatból keressük ki. Leggyakrabban a 95%-os megbízhatósági szintet alkalmazzuk, ebben az esetben $z=1,96$ és $t=2,0$. A 95%-os megbízhatósági szint egyébként annyit jelent, hogy azonos feltételek mellett sokszor megismételt mintavétel esetén 100 mintából átlag 95 esetben a konfidencia intervallum közrefogja a várható értéket, 5 esetben pedig kedvezőtlen eredménnyel jár.

5. A hibahatár, vagy más néven a maximális hiba megállapítása:

$$\Delta = t \cdot S_x^- \text{ vagy } \Delta = z \cdot S_x^-$$

6. A megbízhatósági értékköz vagy konfidencia intervallum kijelölése:

$$\bar{x}_m - \Delta < \bar{x} < \bar{x}_m + \Delta$$

7. A végeredmény értékelése:

Amennyiben nem vagyunk elégedettek a becslés eredményével, mert a konfidencia intervallum túlságosan nagy:

- Egyrészt növelhetjük a megbízhatósági szintet.

A hibahatárok közé eső minták valószínűsége:

$$t = \frac{\Delta}{S_x^-}$$

- Másrészt csökkenthetjük a hibahatárt (Δ).
- A harmadik lehetőség, hogy növeljük a megfigyelések számát.

A mintaelem számának meghatározása:

Ismétléses kiválasztás esetén

Ismétlés nélküli kiválasztás esetén

$$n = \left(\frac{t \cdot s}{\Delta} \right)^2$$

$$n = \left(\frac{t \cdot s}{\Delta} \right)^2, \text{ majd } n' = \frac{n}{1 + \frac{n}{N}}$$

2.1.1 A rétegzett minta értékelése

Becslőfüggvény: a főátlag

A számítás lépései:

1. Meghatározzuk a rétegenkénti mintaátlagokat (részátlagokat: $\bar{x}_1; \bar{x}_2 \dots$), majd azok súlyozott átlagát (főátlagot: $\bar{x}$).
2. Meghatározzuk a rétegenkénti szórásnégyzetet, majd azok súlyozott átlagaként az ún. belső szórásnégyzetet. $\rightarrow \sigma_B^2 \rightarrow S_B^2$
3. Kiszámítjuk a standard hiba nagyságát:

$$S_{\bar{x}} = \frac{S_B}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

4. A becslés megbízhatósági szintjének megállapítása.
5. A hibahatár, vagy más néven a maximális hiba megállapítása:

$$\Delta = t \cdot S_{\bar{x}} \text{ vagy } \Delta = z \cdot S_{\bar{x}}$$

6. A megbízhatósági értékköz vagy konfidencia intervallum kijelölése:

$$\bar{x} \pm \Delta$$

7. A végeredmény értékelése.

A mintaelem számának meghatározása a következő képlettel történik:

$$n = \frac{\sum_{j=1}^r n_j S_j^2}{N^2 S_{\bar{x}}^2 + \sum_{j=1}^r n_j S_j^2}$$

2.2 Értékösszeg becslése

Becslőfüggvény: a mintaátlag

Lépései:

1. Az alapsokaság értékösszegének meghatározása:

$$N \cdot \bar{x}_m \text{ ahol } N\text{-az alapsokaság elemszáma}$$

2. Az értékösszeg standard hibájának meghatározása:

$$N \cdot S_x$$

3. A hibahatár kijelölése:

$$\Delta = t \cdot N \cdot S_x$$

4. A konfidencia intervallum megállapítása:

$$N \cdot \bar{x}_m \pm \Delta$$

2.3 Arány vagy valószínűség becslése

Ennek lényege, hogy a mintasokaságon belül egyetlen ismerv szerint két csoportba soroljuk az elemeket, azaz tulajdonképpen alternatív ismérvek becsléséről van szó. Ha a sokasági arányt p -vel jelöljük, akkor ennek becslőfüggvénye $p=k/n$ relatív gyakoriság.

Lépései:

1. A relatív gyakoriság meghatározása:

$$p = \frac{k}{n}$$

p : relatív gyakoriság, ez lesz az arány becslésének becslőfüggvénye

k : 'kedvező' esetek száma, vagyis a mintában hány olyan elem van, amelyek arányát a sokaságon belül vizsgálni akarjuk

n : mintaelemek száma

2. A szórás meghatározása:

Alternatív ismérvek szórása nem más, mint a relatív gyakoriságok mértani átlaga.

$$S = \sqrt{p \cdot q}, \text{ ahol } q=1-p$$

3. A standard hiba kiszámítása:

Ismétléses kiválasztás esetén

$$S_p = \sqrt{\frac{pq}{n}}$$

Ismétlés nélküli kiválasztás esetén

$$S_p = \sqrt{\frac{p \cdot q}{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

4. A becslés megbízhatósági szintjének megállapítása: ('z' vagy 't')

5. A hibahatár, vagy más néven a maximális hiba megállapítása:

$$\Delta = t \cdot S_p, \text{ vagy } \Delta = z \cdot S_p$$

6. A megbízhatósági értékköz vagy konfidencia intervallum kijelölése:

$$p - \Delta < P < p + \Delta$$

7. A végeredmény értékelése.

2.4 Mintafeladatok a statisztikai becslés témaköréhez

1. feladat

Egy tehenészeti telepen 600 darab tehenet tartanak, amelynek bizonyos %-a öt évente selejtezésre kerül. Az egyik ilyen selejtezés alkalmával az állomány 20%-a került lesejtezésre, amelyből 10%-os kiválasztási arány mellett véletlen mintát vettek. A minta esetében megvizsgálták a tejtermelés színvonalát, melynek eredményei a következők voltak (liter/nap):

12; 6; 8; 10; 5; 12; 14; 10; 4; 6; 11; 5

Feladat: Határozza meg 95,5%-os megbízhatósággal, hogy milyen határok között van:

- egy tehén átlagos tejtermelési színvonala;
- a lesejtezésre kerülő teljes tehénállomány tejtermelési színvonala!
- Mekkora minta-elemszámra lenne szükség ahhoz, hogy a maximális hiba 1 liter legyen! (95,5%)

Megoldás:

A sokaság elemszáma $\rightarrow N=600$ db

A selejtezésre kerülő állomány száma $\rightarrow N_1=120$ db

A minta elemszáma (10%-os kiválasztási arány mellett) $\rightarrow n=12$ db

Mivel a megbízhatósági szint 95,5% $\rightarrow t=2$

Az a) kérdésben átlagra vonatkozó becslést kell végezni!

$$\text{Mintaátlag: } \bar{x}_m = \frac{\sum x_i}{n} = \frac{103}{12} = 8,58 \text{ liter}$$

$$\text{Mintabeli szórás: } \sigma_m = \sqrt{\frac{(x_i - \bar{x})^2}{n}} = \sqrt{\frac{(12 - 8,58)^2 + (6 - 8,58)^2 + \dots + (5 - 8,58)^2}{12}} = 3,2$$

$$\text{Korrigált szórás: } S = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(12 - 8,58)^2 + (6 - 8,58)^2 + \dots + (5 - 8,58)^2}{12-1}} = 3,34$$

Standard hiba (ismétléses kiválasztás): $S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{3,34}{\sqrt{12}} = 0,96$

Standard hiba (ismétlés nélküli kiválasztás): $S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = \frac{3,34}{\sqrt{12}} \cdot \sqrt{1 - \frac{12}{120}} = 0,91$

Maximális hiba (ismétléses kiválasztás): $\Delta = t \cdot S_{\bar{x}} = 2 \cdot 0,96 = 1,92$

Maximális hiba (ismétlés nélküli kiválasztás): $\Delta = t \cdot S_{\bar{x}} = 2 \cdot 0,91 = 1,82$

Így a konfidencia intervallum ismétléses kiválasztás esetén: $\bar{x}_m \pm \Delta = 8,58 \pm 1,92 = [6,66; 10,5]$

Tehát egy (selejtezésre kerülő) tehén átlagosan 6,66 liter, illetve 10,5 liter közötti tejet ad naponta.

A konfidencia intervallum ismétlés nélküli kiválasztás esetén: $\bar{x}_m \pm \Delta = 8,58 \pm 1,82 = [6,76; 10,4]$

Tehát egy (selejtezésre kerülő) tehén átlagosan 6,76 liter, illetve 10,4 liter közötti tejet ad naponta.

A b) kérdésben értékösszegre vonatkozó becslést kell végezni!

A vizsgált 120 db-os tehénállomány értékösszegének meghatározása:

$$N \cdot \bar{x}_m = 120 \cdot 8,58 = 1029,6 \text{ liter}$$

Az értékösszeg standard hibájának meghatározása:

$$N \cdot S_{\bar{x}} = 120 \cdot 0,96 = 125 \text{ liter}$$

A hibahatár kijelölése:

$$\Delta = t \cdot N \cdot S_{\bar{x}} = 2 \cdot 120 \cdot 0,96 = 230,4 \text{ liter}$$

A konfidencia intervallum megállapítása:

$$N \cdot \bar{x}_m \pm \Delta = 1029,6 \pm 230,4 = [799,2; 1260] \quad N \cdot \bar{x}_m \pm \Delta$$

A leselejtezésre kerülő 120 darab tehén összes tejtermelése 799,2 és 1260 liter közé esik naponta.

Végül a c) kérdésben a minta-elemszám meghatározásához a következő képletet használjuk:

Ismétléses kiválasztás esetén: $n_1 = \left(\frac{t \cdot S}{\Delta} \right)^2 = \left(\frac{2 \cdot 3,34}{1} \right)^2 = 44,62 \approx 45$ darab

Tehát 95,5%-os megbízhatósággal, valamint ismétléses kiválasztás mellett 45 elemű mintát kell vinni ahhoz, hogy a maximális hiba 1 liter legyen.

Ismétlés nélküli kiválasztás esetén: $n_1' = \frac{n_1}{1 + \frac{n}{N}} = \frac{45}{1 + \frac{12}{120}} = 40,9 \approx 41$ darab

95,5%-os valószínűségi szint és ismétlés nélküli kiválasztás esetén 41 elemű mintára van szükség ahhoz, hogy a hibahatár 1 legyen.

2. feladat

Egy multinacionális vállalatnál felmérést végeztek arra vonatkozóan, hogy a dolgozók évente milyen összeget költenek üdülésre. A cég 1450 dolgozója közül egyszerű véletlen mintavétellel (ismétlés nélküli kiválasztással) kiválasztottak 255 főt. A vizsgálat eredményét a következő táblázat tartalmazza:

Üdülésre fordított összeg (eFt)	Létszám (fő)	i	u _i	d _i	f _i d _i	f _i d _i ²
-50,0	142	50	25	-1,25	-177,5	221,875
50,1-100,0	68	50	75	0,75	-51,0	38,25
100,1-200,0	31	100	150	0	0	0
200,1-500,0	10	300	350	2,00	20,0	40,0
500,1-750,0	4	250	625	4,75	19,0	90,25
Σ	255	-	-	-	-189,5	390,375

Feladat:

- a) Becsülje meg 95,5%-os valószínűséggel, hogy egy dolgozó átlagosan milyen összeget költött üdülésre!
- b) Határozza meg, hogy a dolgozók 2005-ben mennyi pénzt fordítottak összesen üdülésre! (95,5%)
- c) Becsülje meg, hogy a dolgozók hány %-a fordított 100eFt-nál kevesebbet üdülésre! (95,5%)
- d) Becsülje meg azoknak a dolgozóknak az arányát, akik 200eFt és 500eFt közötti összeget költenek üdülésre! (99,5%)

Megoldás:

$N=1450$ fő

$n=255$ fő

Mivel ismétlés nélküli a kiválasztás és $f \geq 5\%$, ezért szükséges a korrekciós tényező alkalmazása!

Kiválasztási arány: $f = \frac{n}{N} = \frac{255}{1450} = 0,176 \rightarrow 17,6\%$

Korrekciós tényező: $k = \sqrt{1 - \frac{n}{N}} = \sqrt{1 - \frac{255}{1450}} = 0,908$

Mivel a megbízhatósági szint: $\pi = 95,5\% \rightarrow t=2$

Az a) kérdésben átlagra vonatkozó becslést kell végezni!

$$\bar{x}_m = a + \frac{\sum f_i d_i}{\sum f_i} \cdot i$$

A tetszőlegesen választott osztályközép: $u_i=150$

A tetszőlegesen választott osztályközép hossza: $i=100$

Ennek megfelelően:

$$d_i = \frac{u_{ij} - u_i}{i} = \frac{75 - 150}{100} = 0,75$$

$$d_i = \frac{u_{ij} - u_i}{i} = \frac{350 - 150}{100} = 2,00$$

$$d_i = \frac{u_{ij} - u_i}{i} = \frac{25 - 150}{100} = -1,25$$

$$d_i = \frac{u_{ij} - u_i}{i} = \frac{625 - 150}{100} = 4,75$$

$$\bar{x}_m = a + \frac{\sum f_i d_i}{\sum f_i} \cdot i = 150 + \frac{-189,5}{255} \cdot 100 = 75,7 \text{eFt}$$

$$S = \sigma_m = i \cdot \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i} \right)^2} = 100 \cdot \sqrt{\frac{390,375}{255} - \left(\frac{189,5}{255} \right)^2} = 98,9 \text{eFt}$$

$S = \sigma_m$, tehát a korrigált szórás egyenlő a mintabeli szórással, mivel nagymintáról van szó (n=255), így az (n-1)-es korrekciónak nincs jelentősége!

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = \frac{98,9}{\sqrt{255}} \cdot \sqrt{1 - \frac{255}{1450}} = 5,6$$

$$\Delta = t \cdot S_{\bar{x}} = 2 \cdot 5,6 = 11,2$$

$$\text{Konfidencia intervallum: } \bar{x}_m \pm \Delta = 75,7 \pm 11,2 = [64,5; 86,9]$$

Tehát 95,5%-os valószínűséggel egy dolgozó átlagosan 64,5eFt és 86,9eFt közötti összeget költött üdülésre.

A b) kérdésben értékösszegre vonatkozó becslést kell végezni!

$$N \cdot \bar{x}_m = 1450 \cdot 75,7 = 109765 \text{eFt} \rightarrow 109,765 \text{MFt}$$

$$N \cdot S_{\bar{x}} = 1450 \cdot 5,6 = 8120$$

$$\Delta = t \cdot N \cdot S_{\bar{x}} = 2 \cdot 1450 \cdot 5,6 = 16240$$

$$N \cdot \bar{x}_m \pm \Delta = 109765 \pm 16240 = [93525; 126005]$$

A dolgozók 2005-ben összesen 93525eFt és 126005eFt közötti összeget fordítottak üdülésre 95,5%-os megbízhatósági szinten.

A c) kérdésben valószínűsége/arányra vonatkozó becslést kell végezni!

Az eredeti feladatot átalakítjuk úgy, hogy alternatív ismérveket hozunk létre. A tulajdonság meglétét x_1 -gyel, a hiányát x_0 -lal jelöljük el.

x_i	f_i
x_1 (100eFt-nál kevesebbet költők száma)	$f_1=210$ (142+68)
x_0 (100eFt-ot vagy ennél többet költők száma)	$f_0=45$ (31+10+4)
Σ	255

$$p = \frac{f_1 \cdot 1 + f_0 \cdot 0}{f_1 + f_0} = \frac{f_1}{f_1 + f_0} = \frac{210}{255} = 0,8235$$

$$p = \frac{k}{n} = \frac{210}{255} = 0,8235, \text{ ahol}$$

k: 'kedvező' esetek száma, vagyis a mintában hány olyan elem van, amelyek arányát a sokaságon belül vizsgálni akarjuk

n: mintaelemek száma

Mivel $p+q=1$, így

$$q = 1 - p = 1 - 0,8235 = 0,1765$$

$$S = \sqrt{p \cdot q} = \sqrt{0,8235 \cdot 0,1765} = 0,38$$

$$S_p = \sqrt{\frac{p \cdot q}{n}} \cdot \sqrt{1 - \frac{n}{N}} = \sqrt{\frac{0,8235 \cdot 0,1765}{255}} \cdot \sqrt{1 - \frac{255}{1450}} = 0,0217$$

$$\Delta = t \cdot S_p = 2 \cdot 0,0217 = 0,0434$$

Konfidencia intervallum: $P = p \pm \Delta = 0,8235 \pm 0,0434 = [0,7801; 0,8669] \rightarrow [78,01\%; 86,69\%]$

95,5%-os megbízhatósági szinten 78,01% és 86,69% között van azoknak a dolgozóknak az aránya, akik 100eFt-nál kevesebbet költenek üdülésre.

A d) kérdésben szintén valószínűsége/arányra vonatkozó becslést kell végezni!

$$p = \frac{k}{n} = \frac{10}{255} = 0,0392$$

Mivel $p+q=1$, így

$$q = 1 - p = 1 - 0,0392 = 0,9608$$

$$S = \sqrt{p \cdot q} = \sqrt{0,0392 \cdot 0,9608} = 0,19$$

$$S_p = \sqrt{\frac{p \cdot q}{n}} \cdot \sqrt{1 - \frac{n}{N}} = \sqrt{\frac{0,0392 \cdot 0,9608}{255}} \cdot \sqrt{1 - \frac{255}{1450}} = 0,011$$

$$\Delta = t \cdot S_p = 2 \cdot 0,011 = 0,022$$

Konfidencia intervallum: $P = p \pm \Delta = 0,0392 \pm 0,022 = [0,0172; 0,0612] \rightarrow [1,72\%; 6,12\%]$

95,5%-os megbízhatósági szinten 1,72% és 6,12% között van azoknak a dolgozóknak az aránya, akik 200 és 500eFt közötti összeget költenek üdülésre.

3. feladat

Nyíregyháza vonzáskörzetéből 15500 munkavállaló ingázik nap mint nap dolgozni a megyeszékhelyre. Az ingázók 40%-a vonattal, 25%-a autóbusszal, 35%-a pedig személygépkocsival közelíti meg munkahelyét. Arányosan rétegzett mintavétellel, 5%-os kiválasztási arány mellett mintát vettünk és megvizsgáltuk, hogy a dolgozók milyen távolságból utaznak be Nyíregyházára. A felmérés eredményeit a következő táblázat tartalmazza:

Utazási távolság (km)	Ingázók száma (fő)			d_j	$f_j d_j$			$f_j d_j^2$		
	v.	a.	szgk.		v.	a.	szgk.	v.	a.	szgk.
-10,0	45	74	120	-1	-45	-74	-120	45	74	120
10,1-20,0	69	80	81	0	0	0	0	0	0	0
20,1-30,0	96	30	50	1	96	30	50	96	30	50
30,1-40,0	75	7	10	2	150	14	20	300	28	40
40,1-50,0	25	3	10	3	75	9	30	225	27	90
Σ	310	194	271	-	276	-21	-20	666	159	300

v.: vonat

a.: autóbusz

szgk.: személygépkocsi

Feladat: Becsülje meg, hogy a dolgozók átlagosan milyen távolságból ingáznak 95,5%-os valószínűségi szint mellett!

Megoldás:

Átlagbecslést kell végezni arányosan rétegzett mintavétel esetén!

$N=15500$ fő

$$f = 5\% = \frac{n}{N} \cdot 100 = \frac{n}{15500} \cdot 100 \Rightarrow n = 775$$

Tehát 775 főből álló mintát vettünk.

Mivel arányos rétegzéssel vettük a mintát, így:

Vonattal utazók száma: $n_v = 775 \cdot 0,4 = 310$ fő

Autóbusszal utazók száma: $n_a = 775 \cdot 0,25 = 194$ fő

Személygépkocsival utazók száma: $n_{szgk} = 775 \cdot 0,35 = 271$ fő

Meghatározzuk a rétegenkénti mintaátlagokat (részátlagokat: $\bar{x}_v; \bar{x}_a; \bar{x}_{szgk}$), majd azok súlyozott átlagát (főátlagot: $\bar{x}$).

$$\bar{x}_v = a + \frac{\sum f_i d_i}{\sum f_i} \cdot i = 15 + \frac{276}{310} \cdot 10 = 23,9km$$

$$\bar{x}_a = a + \frac{\sum f_i d_i}{\sum f_i} \cdot i = 15 + \frac{-21}{194} \cdot 10 = 13,9km$$

$$\bar{x}_{szgk} = a + \frac{\sum f_i d_i}{\sum f_i} \cdot i = 15 + \frac{-20}{271} \cdot 10 = 14,3km$$

$$\bar{x} = \frac{\sum n_j \cdot \bar{x}_j}{\sum n_j} = \frac{310 \cdot 23,9 + 194 \cdot 13,9 + 271 \cdot 14,3}{775} = 18,0km$$

Meghatározzuk a rétegenkénti szórásnégyzetet, majd azok súlyozott átlagaként az ún. belső szórásnégyzetet. $\rightarrow \sigma_B^2 \rightarrow S_B^2$

$$S_v = \sigma_v = i \cdot \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i} \right)^2} = 10 \cdot \sqrt{\frac{666}{310} - \left(\frac{276}{310} \right)^2} = 11,64km$$

$$S_a = \sigma_a = i \cdot \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i} \right)^2} = 10 \cdot \sqrt{\frac{159}{194} - \left(\frac{-21}{194} \right)^2} = 8,99km$$

$$S_{szgk} = \sigma_{szgk} = i \cdot \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i} \right)^2} = 10 \cdot \sqrt{\frac{300}{271} - \left(\frac{-20}{271} \right)^2} = 10,5km$$

$$S_B^2 = \sigma_B^2 = \frac{\sum n_j \cdot S_j^2}{\sum n_j} = \frac{310 \cdot 11,64^2 + 194 \cdot 8,99^2 + 271 \cdot 10,5^2}{775} = 112,98$$

$$S_B = \sqrt{112,98} = 10,63$$

$S = \sigma_m$, tehát a korrigált szórással egyenlő a mintabeli szórással, mivel nagymintáról van szó ($n \geq 100$), így az $(n-1)$ -es korrekciónak nincs jelentősége!

Kiszámítjuk a standard hiba nagyságát:

$$S_{\bar{x}} = \frac{S_B}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = \frac{10,63}{\sqrt{775}} \cdot \sqrt{1 - \frac{755}{15500}} = 0,37$$

A becslés megbízhatósági szintjének megállapítása:

$$\Delta = t \cdot S_{\bar{x}} = 2 \cdot 0,37 = 0,74$$

Konfidencia intervallum: $\bar{x} \pm \Delta = 18 \pm 0,74 = [17,26; 18,74]$

Tehát 95,5%-os megbízhatósági szinten 17,26, illetve 18,74 km között van az a távolság, amit a dolgozók átlagosan ingázásra fordítanak.

2.5 Önellenző kérdések

Értelmezze a paraméter fogalmát!

Mit értünk statisztikai becslés alatt!

Definiálja a becslőfüggvényt!

Mit nevezünk becsült értéknek!

Mit nevezünk várható értéknek!

Mit értünk torzítatlanság alatt!

Mit jelent a becslőfüggvény hatásossága!

Mit jelent a konzisztencia követelménye!

Sorolja fel a pontbecslés módszereit!

Mit értünk pontbecslés, illetve intervallumbecslés alatt!

Rendszerezze és értelmezze a becslés során elkövethető hibákat!

Ismertesse a nagy számok törvényét!

Ismertesse Csebisev tételét!

Mit nevezünk valószínűségi változónak!

Ismertesse Ljapunov tételét (központi határeloszlás tétele)!

Értelmezze a Gauss-görbe három jellegzetes pontját!

Milyen logikai lépései vannak a statisztikai becslésnek!

Mit értünk 95%-os megbízhatósági szint alatt!

3 HIPOTÉZISVIZSGÁLAT

A sokaságra vonatkozó állítást vagy feltételezést **hipotézisnek** nevezzük. A hipotézis vonatkozhat:

- paraméterre,
- eloszlásra.

Azt az eljárást, amely segítségével eldöntjük, hogy az adott állítás elfogadható-e, vagy sem, **hipotézisvizsgálatnak** nevezzük. A hipotézis helyességét a véletlen mintából származó adatok segítségével ellenőrizzük. A minta elemeit **valószínűségi változóknak** nevezzük.

A véletlen minta jellemzői:

- Az elemek mintába kerülésének esélye (valószínűsége) azonos.
- A sokaság összetétele a kiválasztás során nem változik.
- Az egymás után kiválasztott mintaelemek függetlenek.

Azt a hipotézist, amelynek elfogadásáról, vagy elutasításáról dönteni kívánunk **nullhipotézisnek** (induló hipotézisnek), a vele szemben álló feltevést **alternatív hipotézisnek** (ellenhipotézisnek) nevezzük. A nullhipotézist H_0 , az alternatív hipotézist H_1 szimbólummal jelöljük.

A hipotézis lehet:

- **Egyszerű:** azt állítjuk, hogy az ismeretlen paraméter egyenlő egy meghatározott értékkel. Ez a hipotézis egyértelműen specifikálja a sokasági paraméter értékét.
- **Összetett:** a paraméter egy meghatározott tartományban helyezkedik el.

A következőkben mi csak az egyszerű (kétoldalú) hipotézisekkel foglalkozunk.

A hipotézisvizsgálat lépései:

1. A feladat értelmezése

2. A nullhipotézis felállítása

A nullhipotézisben azt állítjuk, hogy az alapsokaság paramétere és annak feltételezett értéke, vagy a tényleges és feltételezett alapeloszlás között nincs különbség, vagyis az összehasonlítandó jellemzők különbségének várható értéke nulla.

3. Próbafüggvény szerkesztése

A hipotézisvizsgálat eszköze a próbafüggvény, amely a véletlen minta elemeinek, vagyis a valószínűségi változók függvénye.

4. A szignifikancia szint megállapítása (α)

Mivel döntésünket véletlen mintára alapozzuk, az eredmények csak bizonyos valószínűséggel fogadhatók el, vagyis döntésünk során objektív hiba léphet fel.

A szignifikancia szint azt jelenti, hogy a mintavételt sokszor megismételve 100 eset közül átlag hányszor fordul elő az az esemény, hogy a próbafüggvény mintából számított értéke nem az elfogadási tartományba kerül. (Az α értéke többnyire 10%-nál kisebb. A leggyakrabban használt szignifikancia szint az 5%, illetve az 1%.)

5. A próbafüggvény tapasztalati értékének meghatározása

Behelyettesítünk a próbafüggvénybe. A konkrét esetre vonatkozó behelyettesítést statisztikai próbának nevezzük.

6. A próbafüggvény elméleti, vagy kritikus értékének meghatározása a kiválasztott szignifikancia szint mellett

A próbafüggvény eloszlásának ismeretében kijelölhető egy olyan tartomány, intervallum, amely tetszőlegesen nagy valószínűséggel foglalja magában a próbafüggvény értékeit. Az intervallum két végpontját kritikus értékek, magát az intervallumot elfogadási tartománynak nevezzük. Az értékek kijelölésekor a statisztikai biztonságból kell kiindulni.

Normális eloszlásnál már tudjuk, hogy a szórás többszörös által megadott határok közé esés valószínűsége a statisztikai biztonság.

$PL: \pm 1,96\sigma$ intervallumon belül található az ismértékek 95%-a, vagyis 95% a statisztikai biztonsága annak, hogy az egyes értékek nem lépik túl a megadott intervallumot, és csak 5% a valószínűsége annak, hogy a kritikus értékeken kívül esnek. Példánkban a 95% a statisztikai biztonság, az 5% a szignifikancia szint, amit P-vel, vagy α -val jelölünk.

A különböző típusú nullhipotézisek vizsgálatához, kidolgozott próbafüggvények állnak rendelkezésre, és hozzájuk táblázatos formában adottak a különböző szignifikancia szintekhez tartozó kritikus értékek.

7. Döntés a nullhipotézis kérdésében

Döntés során összevetjük a próbafüggvény számított és elméleti értékét, majd értelmezzük az eredményt.

Ha a próbafüggvény értéke az elfogadási tartományba esik, akkor a nullhipotézist adott szignifikancia szint mellett elfogadjuk, következésképpen nincs lényeges eltérés a feltevésünk és a valóság között. (Az elfogadási tartományba esés valószínűsége $1-\alpha$.)

Ellenkező esetben, ha tehát a próbafüggvény értéke az elfogadási tartományon kívül, vagyis a kritikus tartományba esik, elvetjük a nullhipotézist az alternatív javára. Tehát bizonyítottan nagy az eltérés a feltevésünk és a valóság között. (A kritikus tartományba esés valószínűsége α .)

Vizsgálataink - és azon belül a hibaszámítás - szempontjából lényeges az **alapeloszlás típusa**. Legfontosabb szerepet a normális eloszlás, illetve az abból származtatható eloszlások töltenek be. Ezek:

- standard normális eloszlás (z-táblázat),
- az $n-1$ szabadságfokú Student-féle t-eloszlás (t-táblázat),

- a χ^2 (khi-négyzet) eloszlás (χ^2 -táblázat),
- az n_1 és n_2 szabadságfokú F-eloszlás (F-táblázat).

Az ismertett eloszlások fontos jellemző paramétere az ún. **szabadságfok**, amely nem más, mint az adott valószínűségi változó képzésénél alapul vett független összeadandók száma.

1. táblázat: A hipotézisvizsgálat során elkövethető hibák

	Döntés	
Nullhipotézis	Elfogadjuk	Elvetjük
Igaz	Helyes $1-\alpha$	Helytelen α Elsőfajú hiba
Hamis	Helytelen β Másodfajú hiba	Helyes $1-\beta$ A próba ereje

Forrás: Saját szerkesztés

Döntésünk során kétféle hibát követhetünk el:

1. Elvetjük a nullhipotézist, bár az igaz. Az elkövetett hiba az ún. elsőfajú hiba. Ez akkor következik be, ha a nullhipotézis helyessége ellenére a próbafüggvény értéke a kritikus tartományba esik. Ennek valószínűsége $P\%$, vagyis a szignifikancia szinttel egyenlő. Mivel P megválasztása tőlünk függ, ez azt jelenti, hogy az elsőfajú hiba elkövetésének valószínűsége tetszés szerint csökkenthető.
2. Elfogadjuk a nullhipotézist - mert a próbafüggvény értéke az elfogadási tartományba esik - bár az nem igaz. Az elkövetett hiba a másodfajú hiba, amelynek valószínűsége β . Másodfajú hiba akkor léphet fel, ha a tényleges tapasztalati eloszlás nem egyezik meg a hipotézisben felállított eloszlással. A másodfajú hiba elkövetésének valószínűségét nem tudjuk

korlátozni, mert β a valóságban ténylegesen fennálló - általában nem ismert – helyzettől függ.

A két hiba elkövetésének valószínűsége fordított arányban áll egymással, tehát csakis egymás rovására változtatható. A szignifikancia szint rögzítése előtt ezért mérlegelni kell, hogy számunkra a kétféle hiba elkövetése közül melyik a veszélyesebb. A gyakorlatban arra kell törekedni, hogy egyik se legyen túl nagy. Tapasztalatok szerint ez az 5%-os szignifikancia szinten teljesül.

3.1 Paraméterre vonatkozó próbák

Már korábban definiáltuk, hogy a paraméter nem más, mint a sokasági eloszlás valamely jellemzője. Ilyen paraméter pl.: a várható érték, a szórás, az arány. Ezek közül a következőkben mi csak a várható értékre (átlagra) vonatkozó hipotézisekkel foglalkozunk.

- Alapfeltétel: a vizsgálatban felhasznált véletlen minta vagy minták normális eloszlású alapsokaságból származnak. Ezen feltétel teljesülését - a később tárgyalásra kerülő χ^2 -próba segítségével ellenőrizhetjük.
- Az alapfeltétel mellett figyelembe kell venni azt is, hogy az alapsokaság szórása ismert-e, vagy sem.

3.1.1 Várható értékre vonatkozó egymintás próbák

A várható értékre vonatkozó próba kétféle lehet: z, vagy t-próba.

- **Ha az alapsokaság szórása (σ) ismert: z-próbát alkalmazunk.**

Segítségével az a hipotézis ellenőrizhető, hogy egy normális alapeloszlású sokaság várható értéke egy feltételezett értékkel megegyezik.

– A nullhipotézis: $(\bar{x} - \bar{x}_m) = 0$

Az alkalmazás feltétele az alapeloszlás szórásának az ismerete. A nullhipotézis eldöntéséhez az alapsokaság feltételezett várható értékének ($\bar{x}$) és az ismert szórású (σ) normális eloszlású alapsokaságból vett n elemű véletlen minta átlaga különbségének a szignifikanciáját kell megállapítani.

– A próbafüggvény:

$$z = \frac{\bar{x} - \bar{x}_m}{\frac{\sigma}{\sqrt{n}}}, \text{ amely standard normális eloszlást követ.}$$

– A kritikus érték ($z_{\alpha\%}$) a z-táblázatból kereshető ki ($1 - \frac{\alpha}{2}$ -nél).

– Amennyiben:

- $z > z_{\alpha\%}$ a nullhipotézist elvetjük, vagyis a várható érték és a feltételezett érték között szignifikáns különbség van;
- $z \leq z_{\alpha\%}$ a nullhipotézist elfogadjuk, vagyis a várható és feltételezett érték különbsége véletlenszerű, statisztikailag nem igazolható (nem szignifikáns).

➤ **Ha az alapsokaság szórása nem ismert: t-próbát alkalmazunk.**

– A nullhipotézis szintén: $(\bar{x} - \bar{x}_m) = 0$

– Ennek próbafüggvénye:

$$t = \frac{\bar{x} - \bar{x}_m}{\frac{S}{\sqrt{n}}}, \quad \text{amely } n-1 \text{ szabadságfokú Student-féle}$$

t-eloszlást követ, ahol S a σ -nak a mintából nyert becslése.

– A kritikus értéket ($t_{\alpha\%}$) n-1-es szabadságfoknál a t-eloszlás táblázatból keressük ki.

– Amennyiben:

- $t_{\alpha\%} < t$ a nullhipotézist elvetjük,
- $t_{\alpha\%} \geq t$ a nullhipotézist elfogadjuk.

3.1.2 Várható értékre vonatkozó kétmintás próbák

A várható értékre vonatkozó kétmintás próbák közül a **két független mintát igénylő próba** alkalmazását emeljük ki, melynek két fajtája van: a z-próba és a t-próba.

➤ z-próba

A z-próbát két mintaátlag közötti különbség szignifikancia vizsgálatára használjuk, feltéve, hogy az alapsokaság eloszlása normális, továbbá a két szórás azonos és ismert.

- Nullhipotézis: $(\bar{x}_1 - \bar{x}_2) = 0$
- Próbafüggvény: $z = \frac{\bar{x}_1 - \bar{x}_2}{S_d}$, ahol $S_d = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}$, amely standard normális eloszlást követ.

S_d : a véletlen hiba átlagos nagysága

A hipotézisvizsgálat további menete megegyezik az egymintás z-próbánál leírtakkal.

➤ t-próba

A t-próbával ellenőrizhetjük két ismeretlen várható értékű és ismeretlen, de azonos szórású normális eloszlású sokaságból származó mintaátlagok különbségének szignifikanciáját.

$$\underline{\text{Ha } n_1=n_2:} \quad t = \frac{\bar{x}_1 - \bar{x}_2}{S_d}, \text{ ahol } S_d = \sqrt{\frac{S_1^2 + S_2^2}{n}}, \text{ amely}$$

Student-féle t-eloszlást követ $2(n-1)$ szabadságfokkal.

S_d : a standard hiba átlagos nagysága

Ha $n_1 \neq n_2$:

Feltétel: a szórások egyezőségének vizsgálata $\rightarrow S_1 = S_2 \rightarrow$ F-próbával

– A nullhipotézis arra vonatkozik, hogy két normális eloszlású ismeretlen várható értékű sokaság szórása azonos. Ha ez igaz, akkor a két sokaságból vett n_1 és n_2 elemű minta szórásnégyzeteinek (S_1^2, S_2^2) hányadosa F-eloszlást követ, n_1-1 illetve n_2-1 szabadságfokkal:

$$\text{– F-próba: } \frac{S_1^2}{S_2^2}$$

S_1^2 mindig a nagyobb szórásnégyzet n_1-1 szabadságfokkal,

S_2^2 mindig a kisebb szórásnégyzet n_2-1 szabadságfokkal.

– A kritikus értéket az F-eloszlás táblázatból keressük ki, mégpedig a számláló és a nevező szabadságfoka szerint, a megadott szignifikancia szintnél.

– Amennyiben:

– $F \leq F_{p\%}$ a nullhipotézist fenntartjuk,

– $F > F_{p\%}$ a nullhipotézist elvetjük.

- Ha a nullhipotézis igaz, tehát a szórások egyezősége fennáll, elvégezhetjük a t-próbát.

$$\underline{S_1 = S_2}$$

Nullhipotézis: $(\bar{x}_1 - \bar{x}_2) = 0$

$$\text{A próbafüggvény: } t = \frac{\bar{x}_1 - \bar{x}_2}{S_d}, \text{ ahol } S_d = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}$$

A t valószínűségi változó Student-féle t-eloszlást követ, $n_1 + n_2 - 2$ szabadságfokkal.

- Ha a nullhipotézis nem igaz, vagyis a szórások nem egyeznek meg, a Welch-féle közelítő eljárást alkalmazzuk.

$$S_1 \neq S_2$$

$$\text{Nullhipotézis: } (\bar{x}_1 - \bar{x}_2) = 0$$

$$\text{A próbafüggvény: } t = \frac{\bar{x}_1 - \bar{x}_2}{S_d}, \text{ ahol } S_d = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \text{ ez csak egy közelítő}$$

megoldás, ezért:

$$S_{zf} = \frac{1}{\frac{(1-c)^2}{n_1-1} + \frac{c^2}{n_2-1}}, \text{ ahol } c = \frac{\frac{S_2^2}{n_2}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \text{ amely Student-féle}$$

t-eloszlást követ.

A Welch-féle próbának elsősorban elméleti jelentősége van.

3.2 Eloszlásokra vonatkozó próbák

Az eloszlások paramétereire vonatkozó hipotéziseink ellenőrzése során feltételeztük, hogy a véletlen minta normális eloszlású alapsokaságból származik.

Ezen feltételezésünkön túl olyan elemzési feladatok is adódnak, amikor különböző sokaságok valamilyen ismerv szerinti eloszlását kell összehasonlítani. Az eloszlásokra vonatkozó nullhipotéziseink ellenőrzését a χ^2 -próba segítségével végezhetjük el.

3.2.1 Illeszkedésvizsgálat

A sokaság egészének eloszlására vonatkozó hipotézisek vizsgálatát illeszkedésvizsgálatnak nevezzük. A kérdés tehát az, hogy az alapsokaság eloszlása valóban tekinthető-e normálisnak.

A nullhipotézis szerint azon alapsokaság eloszlása, amelyből a minta származik, megegyezik valamilyen elméleti következtetéssel nyert eloszlással (pl. normális eloszlással). A χ^2 -próba ebben az esetben egy tapasztalati eloszlás (minta) gyakorisága és egy ismert alapsokaság eloszlása alapján várható gyakoriságok különbségeinek a szignifikanciáját vizsgálja. A két eloszlásra vonatkozó nullhipotézis ellenőrzésére alkalmas próbafüggvény:

$$\chi^2 = \sum_{i=1}^r \frac{(f_i - f_i^*)^2}{f_i^*}$$

ahol: f_i = az i -edik ismérvváltozat (osztályköz) megfigyelt gyakorisága,

f_i^* = az i -edik ismérvváltozathoz (osztályközhöz) tartozó elméleti eloszlás.

A próbafüggvény alkalmazásának feltételei:

- $\sum_{i=1}^r = n \geq 50$ és
- $f_i \geq 5$, vagyis bármely osztályközhöz tartozó gyakoriság legalább 5.

Az illeszkedésvizsgálatnak két fajtáját különböztetjük meg:

– **Tiszta illeszkedésvizsgálat**

Ha az alapsokaságra vonatkozó paraméterek (várható érték és a szórás) ismertek, tiszta illeszkedés vizsgálatról beszélünk. Ekkor a χ^2 szabadságfoka $r-1$, ahol r az osztályközők száma.

– **Becsléses illeszkedésvizsgálat**

Ha az alapsokaságra vonatkozó várható érték és a szórás nem ismert, tehát csak a minta jellemzőivel dolgozhatunk, becsléses illeszkedésvizsgálatról beszélünk. A számítás menete a tiszta illeszkedésvizsgálatéval azonos, csak a χ^2 szabadságfoka $r-1-b$, ahol b a becsült (minta alapján számított és felhasznált) paraméterek száma.

A kritikus értéket a χ^2 -eloszlás táblázatból, $r-1$, illetve $r-1-b$ szabadságfok és a kiválasztott szignifikancia szint mellett keressük ki.

Amennyiben:

- $\chi^2 > \chi^2_{p\%}$ a nullhipotézist elvetjük, ez azt jelenti, hogy a minta eloszlása eltér az elméleti eloszlástól,
- $\chi^2 \leq \chi^2_{p\%}$ a nullhipotézis fennáll, tehát a mintasokaság eloszlása megfelel az elméleti eloszlásnak, a különbségek csak a véletlennek tulajdoníthatók.

3.2.2 Homogenitásvizsgálat

Két eloszlás egyezésének vizsgálatát homogenitásvizsgálatnak nevezzük. A nullhipotézisben azt állítjuk, hogy valamely változó két sokaságon belüli eloszlása azonos. A kérdés úgy vetődik fel, hogy a két minta származhat-e azonos alapsokaságból.

Jelöljük a két minta megoszlását f_{1i} és f_{2i} -vel, amikor:

$$\sum_{i=1}^r f_{1i} = n_1 \quad \text{és} \quad \sum_{i=1}^r f_{2i} = n_2$$

A próbafüggvény, amennyiben n_1 és n_2 kellően nagy:

$$\chi^2 = n_1 \cdot n_2 \sum_{i=1}^r \frac{\left(\frac{f_{1i} - f_{2i}}{n_1 + n_2} \right)^2}{\frac{f_{1i} + f_{2i}}{n_1 + n_2}}, \text{ amely}$$

$r-1$ szabadságfokú χ^2 -eloszlást követ, ahol r az osztályközök számát jelenti.

Amennyiben:

- $\chi^2 \leq \chi^2_{p\%}$ a nullhipotézis fennáll, vagyis a két minta eloszlása közötti eltérés véletlenszerű, így azonos sokaságból származik, tehát $100-p\%$ szinten homogénnek tekinthető.
- $\chi^2 > \chi^2_{p\%}$, elvetjük a nullhipotézist, tehát a két minta eloszlása nem tekinthető homogénnek, így nem származhatnak azonos alapsokaságból.

3.2.3 Függetlenségvizsgálat

A mintaelemek egyszerre két tulajdonság szerint is vizsgálhatók. A függetlenségvizsgálat során azt vizsgáljuk, hogy a két tulajdonság (ismérv) az adott sokaságon belül egymástól független-e. A kérdés tehát úgy vetődik fel, hogy a tulajdonságok függetleneknek tekinthetők-e. Ennek eldöntéséhez szintén felhasználható a χ^2 -próba.

A próbafüggvény:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}, \text{ amely } (r-1)(k-1) \text{ szabadságfokkal } \chi^2\text{-eloszlást követ, ahol:}$$

$i=1, 2, \dots, r$ – sorok száma,

$j=1, 2, \dots, k$ – oszlopok száma

A mintaelemeket két tulajdonság szerint rendezve ún. kontingencia táblázatba foglaljuk. A két ismérv közötti kapcsolat léte a táblamezőkben lévő tényleges gyakoriságok (f_{ij}) és az elméleti gyakoriságok (f_{ij}^*) egybevetése útján mutatható ki. Az elméleti vagy fiktív gyakoriságokat a megfelelő peremgyakoriságok szorzatának és a minta elemszámának a hányadosaként számítjuk ki.

Amennyiben:

- $\chi^2 \leq \chi^2_{p\%}$, elfogadjuk a nullhipotézist, tehát a két tulajdonság független egymástól.
- $\chi^2 > \chi^2_{p\%}$, elvetjük a nullhipotézist, tehát van összefüggés, van kapcsolat a két tulajdonság között.

3.3 Mintafeladatok a hipotézisvizsgálat témakörhöz

1. feladat

Egy szarvasmarha telepen a napi átlagos tejtermelési szint 18 liter/db, a sokasági szórás pedig 2,5 liter (feltételezzük, hogy az alapsokasági eloszlás normálisnak tekinthető). Kísérletképpen egyszerű véletlen mintavétellel kiválasztottak 25 db tehenet és új takarmányadagot kezdtek el etetni velük. A kísérleti időt követően ismét megvizsgálták a tejtermelés színvonalát, melynek átlagos értéke a minta esetében: 19,5 liter/db lett.

Feladat: Vizsgálja meg, hogy az új takarmány etetésének van-e statisztikailag kimutatható hatása a tejtermelésre. ($\alpha = 5\%$)

Megoldás: Hipotézisvizsgálattal kell ellenőrizni, hogy a sokasági átlag és a mintabeli átlag között szignifikáns-e a különbség.

Várható értékre vonatkozó egymintás próbát végzünk, még hozzá úgy, hogy az alapsokasági szórás ismert. Mindebből az következik, hogy z-próba függvényt alkalmazunk.

$$\bar{x} = 18 \text{ liter/db}$$

$$\sigma = 2,5 \text{ liter}$$

$$n = 25 \text{ db}$$

$$\bar{x}_m = 19,5 \text{ liter/db}$$

$$H_0: \bar{x} = \bar{x}_m, \text{ vagy } H_0: \bar{x} - \bar{x}_m = 0,$$

$$H_1: \bar{x} \neq \bar{x}_m,$$

$$\text{Próbafüggvény: } z = \frac{\bar{x} - \bar{x}_m}{\frac{\sigma}{\sqrt{n}}} = \frac{18 - 19,5}{\frac{2,5}{\sqrt{25}}} = 3$$

A kritikus érték ($z_{\alpha\%}$) a z-táblázatból kereshető ki ($1 - \frac{\alpha}{2}$ -nél), amely 5%-os szignifikancia szint mellett: $z_{\alpha 5\%} = 1,96$

Mivel $z > z_{\alpha\%}$ ($z = 3 > z_{\alpha 5\%} = 1,96$) a nullhipotézist elvetjük, vagyis az alapsokasági és a mintabeli átlag között szignifikáns különbség van. Tehát 5%-os szignifikancia szinten az új takarmány etetésének statisztikailag kimutatható hatása van a tejtermelés alakulására.

2. feladat

Egyik nap egy pékség két mintaboltjában ellenőrizték a kiszállított kenyerek súlyát:

Ssz.	Kenyerék súlya (g)	
	1. mintabolt	2. mintabolt
1	498	485
2	505	500
3	493	495
4	485	490
5	502	487
6	509	502
7	497	505
8	487	-
9	500	-

Feladat: Vizsgálja meg, hogy származhatnak-e a két mintaboltban lévő kenyerek egy pékségből. Tehát a két boltban lévő kenyerek tömege tekinthető-e azonosnak? ($\alpha = 5\%$)

Megoldás: kétmintás t-próbát végzünk (mivel a σ nem ismert), figyelembe véve, hogy $n_1 \neq n_2$

$$n_1 = 9$$

$$n_2 = 7$$

$$\bar{x}_1 = \frac{\sum x_i}{n} = \frac{4476}{9} = 497,3$$

$$\bar{x}_2 = \frac{\sum x_i}{n} = \frac{3464}{7} = 494,9$$

$$S_1 = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}} = 7,92$$

$$S_2 = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}} = 7,78$$

Feltétel: $H_0: S_1=S_2$

A nullhipotézist F-próbával ellenőrizzük!

$$F = \frac{S_1^2}{S_2^2} = \frac{7,92^2}{7,78^2} = 1,04$$

A kritikus értéket az F-eloszlás táblázatból keressük ki, mégpedig a számláló és a nevező szabadságfoka szerint, a megadott szignifikancia szintnél.

A számláló szabadságfoka: $n_1 - 1 = 8$

A nevező szabadságfoka: $n_2 - 1 = 6$

5%-os szignifikancia szint mellett $F_{\alpha 5\%} = 4,15$

Mivel: $F \leq F_{p\%}$ a nullhipotézist fenntartjuk, tehát a szórások egyezősége fennáll, elvégezhetjük a t-próbát.

$$H_0 : (\bar{x}_1 - \bar{x}_2) = 0$$

$$\text{A próbafüggvény: } t = \frac{\bar{x}_1 - \bar{x}_2}{S_d} = \frac{497,3 - 494,9}{4,23} = 0,57$$

$$S_d = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}} = \sqrt{\frac{9 \cdot 7,92^2 + 7 \cdot 7,78^2}{9 + 7 - 2} \cdot \frac{9 + 7}{9 \cdot 7}} = 4,23$$

A t valószínűségi változó Student-féle t-eloszlást követ, n_1+n_2-2 szabadságfokkal.

Szabadságfok: $Szf. = n_1 + n_2 - 2 = 9 + 7 - 2 = 14$

$$t_{p5\%} = 2,15$$

Mivel $t < t_{p5\%} \rightarrow$ a nullhipotézis igaz, tehát a két boltban lévő kenyerek átlagsúlya megegyezik.

3. feladat

Egy állattenyésztő telepen azt vizsgálták, hogy a hízóbikák napi testtömeg szerinti gyarapodása megfelel-e a normális eloszlásnak. A 600 darab marhából egy 60 darabból álló egyszerű véletlen mintát vettek. A mintavétel eredményeit a következő táblázat tartalmazza:

Napi súlygyarapodás (g)	Hízóbikák (db)	Transzformált osztályköz határa		p _i	f	f _i -f _i '	$\frac{(f_i - f_i')^2}{f_i'}$
		alsó	felső				
751-850	7	-1,95	...	0,0537	3	4	5,33
851-950	10	-1,29	...	0,1729	11	-1	0,09
951-1050	15	-0,63	...	0,3172	19	-4	0,84
1051-1150	12	0,03	...	0,2902	17	-5	1,47
1151-1250	9	0,70	...	0,1324	8	1	0,13
1251-1350	7	1,36	...	0,0336	2	5	12,5
Σ	60	-	-	1,00	60	0	20,36

Feladat: Ellenőrizze, hogy a hízóbikák testtömeg szerinti eloszlása tekinthető-e normálisnak!

Megoldás: Az ellenőrzéshez illeszkedés-vizsgálatot hajtunk végre, és mivel nem ismertek a paraméterek (átlag, szórás), ezért becsléses illeszkedés-vizsgálatot végzünk.

Meghatározzuk a mintabeli átlagot és szórást:

$$\bar{x}_m = a + \frac{\sum f_i d_i}{\sum f_i} \cdot i = 1000 + \frac{27}{60} \cdot 100 = 1045$$

$$S = \sigma_m = i \cdot \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i} \right)^2} = 100 \cdot \sqrt{\frac{149}{60} - \left(\frac{27}{60} \right)^2} = 151,024$$

A transzformált osztályköz alsó határát a következő képlet segítségével határozzuk meg:

$$x' = \frac{x_i - \bar{x}_m}{S} = \frac{700 - 1045}{151,024} = -1,95$$

p_i : az osztályközbe esés valószínűsége

$$f'_i = n \cdot p_i$$

$$\chi^2 = \sum_{i=1}^r \frac{(f_i - f'_i)^2}{f'_i} = 20,36$$

$$Szf. = r - 1 - k = 6 - 1 - 2 = 3$$

$$\alpha = 5\%$$

$$\chi^2_{p5\%} = 7,815 \text{ (}\chi^2\text{-eloszlás táblázatból)}$$

Mivel $\chi^2 > \chi^2_{p5\%}$, ezért a nullhipotézist elvetem, tehát a hízóbikák testtömeg szerinti eloszlása nem tekinthető normális eloszlásúnak.

4. feladat

Egy csomagológép két tételben végzi a gyártást:

Csomagok súlya (g)	1. tétel	2. tétel	$g_{1i} = \frac{f_{1i}}{\sum f_{1i}}$	$g_{2i} = \frac{f_{2i}}{\sum f_{2i}}$	$\frac{(g_{1i} - g_{2i})^2}{f_{1i} + f_{2i}}$
	f_{1i}	f_{2i}			
461-470	9	4	0,067	0,043	
471-480	10	5	0,074	0,054	
481-490	9	8	0,067	0,087	
491-500	20	14	0,148	0,152	
501-510	36	25	0,267	0,272	
511-520	23	18	0,170	0,196	
521-530	18	13	0,133	0,141	
531-540	10	5	0,074	0,054	
Σ	135	92	1,00	1,00	$14,080 \cdot 10^{-5}$

Feladat: Vizsgálja meg, hogy a két minta eloszlása tekinthető-e azonosnak!

Megoldás: Homogenitás-vizsgálatot kell végezni!

Mindenek előtt ellenőrizzük, hogy teljesül-e a feltétel, miszerint $n \geq 50$, illetve $f_i \geq 5$!

Megállapíthatjuk, hogy mindkét mintára teljesülnek a feltételek, így a homogenitás-vizsgálat elvégezhető.

$$\sum_{i=1}^r f_{1i} = n_1 = 135$$

$$\sum_{i=1}^r f_{2i} = n_2 = 92$$

$$\lambda^2 = n_1 \cdot n_2 \cdot \sum_{i=1}^r \frac{(g_{1i} - g_{2i})^2}{f_{1i} + f_{2i}} = 135 \cdot 92 \cdot 14,08 \cdot 10^{-5} = 1,75$$

$$\lambda^2_{\alpha=5\%; szf.=7} = 14,067$$

Mivel $\lambda^2 < \lambda^2_{p5\%}$, ezért a nullhipotézist elfogadjuk, tehát a két minta eloszlása azonosnak, vagy homogénnek tekinthető.

5. feladat

Egy térség vállalkozásairól a következő információk állnak rendelkezésre:

Nyereségi szint Vállalati forma	veszteséges	alacsony	átlagos	magas	Σ
Egyéni vállalkozás	25 32 1,96	37 38 0,03	29 24 0,86	15 12 0,6	106
Bt.	19 18 0,05	28 30 0,14	22 21 0,05	11 11 0	80
Kft.	38 32 0,95	54 51 0,17	41 47 0,88	22 25 0,41	155
Σ	82	119	92	48	341

Feladat: Vizsgálja meg, hogy a vállalati forma és a gazdálkodás jövedelmezősége összefüggésben áll-e egymással, vagy a kettő független!

Megoldás: Függetlenség-vizsgálatot kell végezni!

Hipotézis: H_0 = független a két ismerv

H_1 = nem független

Jelölések:

f_{ij} - tényleges gyakoriság

Bal felső sarok: f_{ij}^* - elméleti gyakoriság, amely nem más, mint a tényleges gyakorisághoz tartozó peremgyakoriságok szorzata és az összes elem hányadosa.

$$\text{Pl.: } f_{11} = \frac{82 \cdot 106}{341} = 25$$

$$\text{Jobb alsó sarok: } \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

$$\text{Pl.: } \frac{(32 - 25)^2}{25}$$

$$\text{Próbafüggvény: } \chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*} = 6,1$$

$$\text{Szf.} = (r-1)(k-1) = (3-1)(4-1) = 6$$

$$\chi^2_{p5\%} = 12,592$$

Mivel $\chi^2 < \chi^2_{p5\%} \rightarrow$ elfogadom a nullhipotézist

A H_0 hipotézis igaz, vagyis a vállalati forma és a gazdálkodás eredményessége független egymástól. A különbség csupán a véletlen hatásának tulajdonítható.

3.4 Önellenző kérdések

1. Mit nevezünk hipotézisnek!
2. Definiálja a hipotézisvizsgálatot!
3. Mit nevezünk nullhipotézisnek, illetve alternatív hipotézisnek!
4. Mit értünk egyszerű és összetett hipotézis alatt!
5. Sorolja fel a hipotézisvizsgálat lépéseit!
6. Mit jelent a szignifikancia szint!
7. Milyen normális eloszlásból származtatott eloszlásokat ismer!
8. Rendszerezze a hipotézisvizsgálat során elkövethető hibákat (elsőfajú, másodfajú hiba)!
9. Mit értünk illeszkedésvizsgálat alatt!
10. Értelmezze a homogenitásvizsgálatot!
11. Mit nevezünk függetlenségvizsgálatnak!
12. Mikor beszélünk tiszta, illetve becsléses illeszkedésvizsgálatról!

4 ÖSSZEFÜGGÉS-VIZSGÁLAT

A jelenségek egymáshoz való viszonyulását:

1. Függatlenség vagy
2. Függőség jellemezheti. Ez utóbbi lehet:
 - a) Oksági, amikor egy vagy több ok hatására jön létre valamely okozat, de nem mindig azonosan érvényesülő tendencia szerint:
 - *Egzakt, függvényszerű* a kapcsolat, ha egyértelmű és kölcsönös a meghatározottság, vagyis X változó adott értékeihez Y meghatározott értéke tartozik (bekövetkezési valószínűsége 1).
 - *Sztokasztikus, valószínűségi jellegű* a kapcsolat, ha feltételes, átlagosan érvényesülő tendenciáról van szó, tehát X változó adott értékeihez Y változó több lehetséges értéke tartozhat.
 - b) Látszólagos, amikor a kapcsolat véletlenszerűen alakul, mert:
 - az ismérvek között kölcsönhatás áll fenn,
 - harmadik ismérv játssza a közös ok szerepét,
 - együtt haladás fedezhető fel, tehát az idő a közvetítő változó.

A statisztika a **sztokasztikus kapcsolatokat** vizsgálja, azokat az eseteket, amikor az egymással összefüggésben álló jelenségek nem határozzák meg egyértelműen egymás alakulását, a köztük levő kapcsolat csak átlagosan érvényesülő tendenciaként, a valószínűségek törvényszerűségére támaszkodva állapítható meg.

A sztokasztikus kapcsolat két, vagy több változó között érvényesülő valószínűségi jellegű kapcsolat.

A sztokasztikus kapcsolat fajtái:

- a mennyiségi ismérvek közötti kapcsolatot **korrelációnak**,
- a minőségi/területi ismérvek közötti kapcsolatot **asszociációnak** (előjelét nem mindig értelmezzük, mérőszáma: Yule, illetve Csuprov-féle asszociációs együtthatók),

- a mennyiségi és nem mennyiségi ismérvek kapcsolatát pedig **vegyes kapcsolatnak** nevezzük (mérőszáma: szóráshányados).

4.1 A mennyiségi ismérvek közötti kapcsolatok vizsgálata

A mennyiségi ismérvek közötti sztochasztikus kapcsolatot tehát **korrelációs kapcsolatnak** nevezzük. Az ok szerepét betöltő ismérvet **független, vagy más néven tényező (magyarázó) változónak** (X_i), az okozat szerepét játszó ismérvet pedig **függő, vagy eredményváltozónak** (Y_i) nevezzük.

A mennyiségi ismérvek kapcsolatainak vizsgálatával a korrelációs számítás foglalkozik. Vizsgálatának területei:

- a **korrelációelemzés**, azaz a kapcsolat szorosságának elemzése, valamint
- a **regresszió analízis**, vagyis kapcsolat törvényszerűségének feltárása.

Az alkalmazásra kerülő módszerek megválasztása attól függ, hogy:

- a) Az okozat alakulásában hány ok szerepének a vizsgálatát tartjuk fontosnak, vagyis hány változós az összefüggés-vizsgálatunk:
 - kétváltozós, ha egy függő és egy független változó szerepel az elemzésben,
 - többváltozós korreláció és regresszió analízis, ha egy függő és több független változó szerepel az elemzésben.
- b) A független változó értelmezési tartományán belül hogyan alakul a függő változó meghatározottsága, vagyis az okra hogyan reagál az okozat:
 - közel azonos törvényszerűség szerint → lineáris jellegű kapcsolat,
 - változóan → nem lineáris kapcsolat.

Mindkét kérdésre megközelítő választ adhatunk az adatok:

- kombinációs táblákba foglalása,

- korrelogrammokon történő ábrázolása útján.

Különbség a matematikai és a sztochasztikus függvények között:

- Matematikai függvények: feltétlen érvényűek, helytől és időtől függetleníthetők.
- Sztochasztikus függvények: feltételes érvényűek, helyhez és időhöz kötötten érvényesülnek.

A korreláció fontosabb típusai:

- A korreláció hiányát a függvény úgy jelzi, hogy bármely X helyen azonos, illetve közel azonos értéket vesz fel, tehát a függvény képe vízszintes vonal (korrelációs ellipszis) lesz. A jelenség azzal magyarázható, hogy Y független X -től, vagyis X nem befolyásolja Y értékét.
- A korreláció hiányának logikai ellentéte a függvényszerű kapcsolat, amikor egy adott X értékhez csupán egyetlen Y érték tartozik, azaz a regressziós görbe körül nincs szóródás.
- Korrelációról akkor beszélünk, ha a regressziós görbe körül van szóródás. Minél kisebb intervallumban szóródnak az adatok, annál szorosabb a kapcsolat. A korrelációs kapcsolat lehet lineáris jellegű, vagy görbe vonalú.
 - Lineáris jellegű a kapcsolat, ha a pontok egy képzeletbeli egyenes mentén szóródnak. Ezen belül megkülönböztetünk pozitív, illetve negatív irányú lineáris korrelációt.

Pozitív lineáris korreláció esetén a nagyobb X értékekhez általában nagyobb Y értékek tartoznak, tehát a pontok egy pozitív meredekségű képzeletbeli egyenes mentén helyezkednek el. *Negatív lineáris korreláció* esetén nagyobb X értékekhez kisebb Y értékek tartoznak, vagyis a független változó növelése a függő változó csökkenését eredményezi.
 - A görbe vonalú (nemlineáris) kapcsolatok egy részénél továbbra is van értelme pozitív, illetve negatív irányról beszélni, de csak abban az esetben, ha a görbe az értelmezési tartományon belül monoton növekvő vagy csökkenő. Nem lehet azonban pozitív vagy negatív irányról beszélni, ha a regresszió irányt változtat.

4.1.1 A lineáris kétváltozós korrelációs kapcsolat mérőszámai

A korrelogram pontjai egy képzeletbeli egyenes - regressziós függvény - mentén kisebb vagy nagyobb mértékben szóródnak. Vizsgálataink során egyetlen mutatószámba szeretnénk sűríteni a pontthalmaz és a regressziós függvény kölcsönös helyzetét. Ez a mutatószám a korrelációs együttható, amely meghatározható:

- közelítő pontossággal és
- egzakt módon.

1. Közelítő kapcsolatszorossági mutatók

- **Előjel korrelációs együttható:** megmutatja, hogy az egyes értékek az átlagtól kisebbek vagy nagyobbak. Az átlagtól kisebb értékek negatív előjelet, a nagyobb értékek pozitív előjelet kapnak. Ezt követően képezzük a kapcsolatban összetartozó előjelek szorzatát, majd megállapítjuk a pozitív (p) és negatív (q) előjelpár szorzatok számát, és az alábbi képlettel állapítjuk meg a szorosság mértékét. A mutató negatívuma, hogy nem veszi figyelembe, hogy az egyes értékek mennyivel térnek el az átlagtól.

$$r_e = \frac{p - q}{n}, \text{ ahol}$$

p: pozitív előjel szorzatok száma

q: negatív előjel szorzatok száma

n: vizsgált elemek száma (n=p+q)

$$-1 \leq r_e \leq +1$$

- **Rangkorrelációs együttható:** meghatározásához mindkét ismerv értékeit azonos irányból kiindulva rangsorszámokkal látjuk el. Meghatározzuk az összetartozó rangsorszámok különbségeit (D), majd négyzetösszegét és az alábbi képletbe behelyettesítve kapjuk a mérőszámot.

$$r_r = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}, \text{ ahol}$$

d_i a két sorszám különbségét jelöli, vagyis $d_i = x_i - y_i$.

A mutatószáma értéke -1 és +1 között helyezkedik el. $-1 \leq r_r \leq +1$

A rangsorszámok összege mindkét ismérvnél azonos kell, hogy legyen, így ha vannak azonos ismérvértékek, azok rangsorszámái megegyezők (az egymást követő rangsorszámok átlaga lesz). Az azonos rangsorszámú ismérvértékeket kapcsolt rangoknak nevezzük.

A mutató hátránya: nem veszi figyelembe a nagyságrendi értékeket, csak a sorrendiséget.

Jelentősége elsősorban szociológiai vizsgálatoknál van.

– **Sperman-féle rangkorrelációs együttható:**

$$\frac{\frac{n^3 - n}{6} - T_x - T_y - \sum D_i^2}{\sqrt{\left(\frac{n^3 - n}{6} - 2T_x\right)\left(\frac{n^3 - n}{6} - 2T_y\right)}}, \text{ ahol}$$

$$T = \sum_{j=1}^r \frac{t_j^3 - t_j}{12}, \quad r: \text{ a kapcsolt rangcsoportok száma,}$$

t_j : az azonos rangsorszámmal rendelkező elemek száma az adott kapcsolt rangcsoporton belül.

Kapcsolt rang: ahol azonos rangsorszámokat osztottunk ki.

Előjele lehet pozitív vagy negatív:

- Pozitív: ha az egyik ismerv növekedése a másik ismerv növekedését vonja maga után, illetve az egyik ismerv csökkenése a másik ismerv csökkenését idézi elő, tehát azonos irányú a kapcsolat.
- Negatív: ha az egyik ismerv növekedése a másik ismerv csökkenését idézi elő és fordítva, tehát ellentétes irányú a kapcsolat.

Ez a forma akkor alkalmazható, ha két ismerv között vizsgáljuk a kapcsolatot szorosságát.

- **Kendall-féle egyetértési (konkordencia) együttható:** azt vizsgálja, hogy van-e összhang a bírálók bírálatai között. A mutató lényege, hogy a véleménykülönbségeket (D_i^2) a teljes összhang esetéhez viszonyítja.

$$W = \frac{12 \sum D_i^2}{m^2 (n^3 - n)}, \text{ ahol}$$

$D_i = d_i - \bar{d}$ (d_i : összeadjuk a helyezési sorszámokat függőlegesen)

$$\bar{d} = \frac{\sum d_i}{n}$$

m: bírálók száma

n: a verseny résztvevőinek száma

Értelmezése: ha a mutató értéke 1-hez közeli, akkor az egyének értékítéleteiben nagy az összhang.

2. Egzakt mutatók

- A legkisebb négyzetek módszerének elvén alapuló **korrelációs koefficiens** (r), amelynek meghatározása a számtani átlag négyzetes minimum tételére vezethető vissza. A mutató adatok konkrét számszerű értékeire támaszkodik, értékét a vizsgált változók szóródási viszonyai határozzák meg.

Meghatározásához kiinduló a két változó együttes szóródását jellemző mutató, az ún. **kovariancia** (C) értékének a meghatározása. Ez nem más, mint a két változó átlagtól vett eltéréseinek szorzatából számított számtani átlag.

$$C = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

A kovariancia tulajdonságai:

- $C=0$, ha a két ismerv egymástól független. Ilyenkor a regresszióvonal helyzete vízszintes (X és Y változók korrelálatlanok).
- $C_{\max.} = \sigma_x \sigma_y$, X és Y változók kovarianciája akkor maximális, ha a két adatsor között lineáris és függvényyszerű kapcsolat áll fenn, és ekkor a kovariancia egyenlő az X és Y változók szórásainak szorzatával.
- Ha $C < \sigma_x \sigma_y$, a kapcsolat sztochasztikus.
- A kovariancia előjele a kapcsolat irányát mutatja.
- A kovariancia abszolút mértékének nincs határozott felső korlátja. Csupán annyit állapíthatunk meg, hogy a nullához közeli abszolút érték laza kapcsolatra utal, egyébként a korreláció szorosságának növekedésével a kovariancia is emelkedő tendenciát mutat.

A kovariancia erőssége: a lineáris korrelációs együttható:

$$r = \frac{C}{\sigma_x \cdot \sigma_y},$$

ami azt mutatja meg, hogy a kovariancia milyen mértékben közelíti meg a függvényyszerű kapcsolatot (a maximális értéket).

A korrelációs együttható tulajdonságai:

- A mutató értéke 0 ± 1 közötti:
 - 0,00-0,30 → laza
 - 0,30-0,70 → közepes
 - 0,70-0,90 → szoros
 - 0,90-1,00 → igen szoros összefüggést mutat.
- Előjele a kapcsolat irányát jelzi. Előjele pozitív, ha a változás a két változónál egyirányú, és negatív, ha a változás ellentétes irányú.
- A kapcsolat hiányát (korrelálatlanság) az $r=0$ érték jelzi.

- Tökéletes függvényszerű lineáris kapcsolat esetén – az iránytól függően – $r=+1$, illetve $r=-1$.

A számítások során alkalmazott formulák:

$$r = \frac{C}{\sigma_x \sigma_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

A korreláció megbízhatóságának ellenőrzése t-próbával történik:

- H_0 : r értéke véletlenül tér el nullától,
- Próbafüggvény: $t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$
- Szignifikancia szint: minimum 5%
- Szabadságfok: $n-2$, ahol $2 \rightarrow$ az előzetesen becsült mutatók száma

$$(\bar{x}, \sigma)$$

- Kritikus érték: $t_{p5\%}$,
- Döntés: $t > t_{p5\%} \rightarrow$ nullhipotézist elvetjük, tehát a korrelációnak 0-tól

való eltérése szignifikáns.

$t \leq t_{p5\%} \rightarrow$ nullhipotézist elfogadjuk, tehát r értéke csak véletlenül tér el 0-tól.

Nagy számú adatpár esetén (legalább 100 vagy azt meghaladó adatpár) r értékét a következő módon számoljuk:

$$r = \frac{\sum f_i (x - \bar{x})(y - \bar{y})}{\sqrt{\sum f_i (x - \bar{x})^2 \sum f_i (y - \bar{y})^2}}$$

Osztályközös gyakorisági sorból a számtani átlagnál és a szórás számításnál tanult segédszámításokat vezetjük be:

$$r = \frac{C}{\sigma_x \cdot \sigma_y}$$

ahol:

$$C = i_x i_y \left[\frac{\sum \sum f_{ij} (d_{xi} d_{yj})}{\sum \sum f_{ij}} - \left(\frac{\sum f_{xi} d_{xi}}{\sum f_{xi}} \cdot \frac{\sum f_{yj} d_{yj}}{\sum f_{yj}} \right) \right]$$

$$\sigma_x = i_x \sqrt{\frac{\sum f_{xi} \cdot d_{xi}^2}{\sum f_{xi}} - \left(\frac{\sum f_{xi} \cdot d_{xi}}{\sum f_{xi}} \right)^2}$$

$$\sigma_y = i_y \sqrt{\frac{\sum f_{yj} \cdot d_{yj}^2}{\sum f_{yj}} - \left(\frac{\sum f_{yj} \cdot d_{yj}}{\sum f_{yj}} \right)^2}$$

- Az r érték négyzetét (r^2) **determinációs együtthatónak** nevezzük, amely %-os formában mutatja a független változó eredményváltozóra gyakorolt hatását.

4.1.2 Lineáris kétváltozós regresszió analízis

Analitikus regressziószámítás során azt vizsgáljuk, hogy a tényezőváltozó milyen módon gyakorol hatást az eredményváltozóra. Lineáris kétváltozós esetben a két változó kapcsolatát a következő egyenessel írhatjuk fel:

$$\hat{y} = a + bx, \text{ ahol}$$

a: állandó, azt mutatja, hogy a regressziós egyenes az y tengelyt hol metszi, tehát $x=0$ helyen a függvény milyen értéket vesz fel,

b: regressziós együttható, az egyenes iránytangense, amely azt fejezi ki, hogy ha a független változót 1 egységgel változtatjuk, várhatóan mennyivel változik a függő változó.

Egy pontthalmazhoz sokféle egyenes illeszthető. A lineáris regressziós függvény meghatározása során arra törekszünk, hogy olyan egyenessel jellemezzük a változók közötti kapcsolatot, amelyik a legjobban illeszkedik a megfigyelésből származó X, Y adatpárokhoz. Azt az egyenest célszerű választani, amelyik esetében a pontoknak a regressziós egyenestől mért átlagos távolsága a legkisebb. Erre pedig a legalkalmasabb a legkisebb négyzetek módszere:

$$\Sigma(y - \hat{y})^2 \rightarrow \min.$$

y: tapasztalati (megfigyelt) értékek,

$\hat{y}$: számított értékek, melyet úgy számolunk ki, hogy a megfigyelt x_i adatokat rendre behelyettesítjük a regressziós függvénybe.

Mivel az egyenes helyzetét a koordináta-rendszeren belül az a és b paraméterek értéke határozza meg, így nyilvánvaló, hogy a minimum összege is ettől függ.

A minimum követelménybe behelyettesítve:

$$\Sigma(y - a - bx)^2 \rightarrow \min.$$

Az a és b meghatározásához két egyenletre van szükség, amit a többváltozós függvények feltételes szélsőértékének meghatározási szabályait alkalmazva vezethetünk le.

Változók: (a,b)

$$f(a;b) = \Sigma(y - a - bx)^2 \rightarrow \min.$$

$$F_a'(a;b) = \Sigma -2(y - a - bx)$$

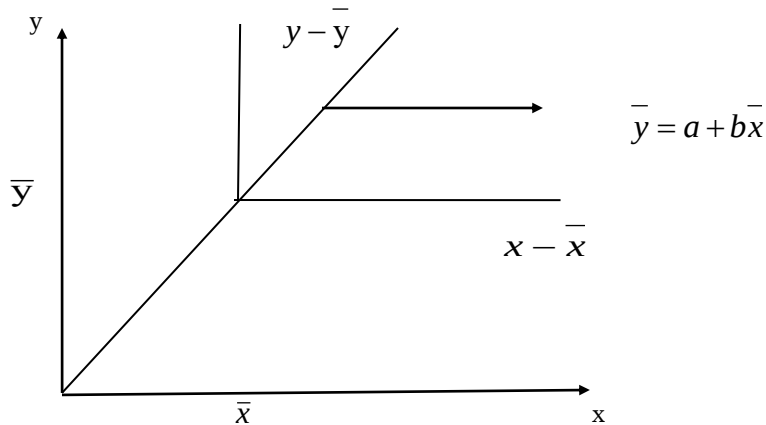
$$F_b'(a;b) = \Sigma -2x(y - a - bx)$$

Ennek eredménye: Gauss-féle normálegyenletek:

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

A normálegyenletek megoldása céljából célszerű egy transzformáció. Mind az X, mind az Y változó helyett ezek átlagtól számított eltéréseit szerepeltetjük.



3. ábra: Regressziós transzformáció

Forrás: Saját szerkesztés

Az új koordináta rendszerben $(x - \bar{x})$ és $(y - \bar{y})$ összefüggéseit vizsgáljuk, a normálegyenletek pedig:

$$\Sigma(y - \bar{y}) = na + b\Sigma(x - \bar{x})$$

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2$$

innen: $a=0$ (A regressziós egyenes a függőleges tengelyt 0-ban metszi.)

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}, \quad b_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x}$$

(A számláló előjele határozza meg a korreláció irányát. A korreláció és a regresszió iránya megegyezik.)

Az $\bar{y} = a + b\bar{x}$ egyenletből viszont:

$$a = \bar{y} - b\bar{x}$$

A függvényillesztés hibája:

A sztochasztikus jellegű kapcsolat hibával terhelt.

Az egyenes illeszkedésének pontosságát a függvényillesztés:

– standard hibájának (reziduális szórás) $\rightarrow S_{\bar{y}} = \sqrt{\frac{(y - \hat{y})^2}{n}}$, $S_{\bar{y}} = S_y \sqrt{1 - r^2}$ és

– relatív hibájának $\rightarrow S_{y\%} = \frac{S_{\bar{y}}}{\bar{y}} \cdot 100$

meghatározásával vizsgáljuk. Az a jó, ha $S_{y\%} \rightarrow 10\%$ alatti. 20% fölött nem fogadjuk el, ebben az esetben a megfigyelt értékek már nem lineáris függvénnyel közelíthetők.

$S_{\bar{y}}$: reziduális szórás, ami a belső szórásnégyzet szerepét játssza. A tapasztalati értékek és a regressziófüggvény szerinti értékek eltérésnégyzetösszege.

$S_{y\%}$: a reziduális szórás és a függő változó átlagának aránya.

Bármely megfigyelt érték (y_i) felírható ugyanezen ponthoz számított regressziós pont ($\hat{y}_i$) és e két pont távolságának ($y_i - \hat{y}_i$) összegeként.

Tehát: $y_i = \hat{y}_i + (y_i - \hat{y}_i)$, amiből adódik, hogy az eltérések összetevőkre bontása a megfelelő négyzetösszegekre is érvényes:

$$S_y^2 = S_{\hat{y}}^2 + S_{(y-\hat{y})}^2, \text{ ahol}$$

S_y : a megfigyelt y értékek átlagtól számított eltérésnégyzetösszege,

$S_{\hat{y}}$: a regresszió által magyarázott eltérésnégyzetösszeg, a számított értékek ($\hat{y}$) átlagtól ($\bar{y}$) vett eltérésnégyzetösszege,

$S_{(y-\hat{y})}$: reziduális eltérésnégyzetösszeg

Rezidium (reziduális változó): maradvány érték, amely a megfigyelt érték (y) és a regressziós függvény szerinti érték ($\hat{y}$) különbsége. Tehát: $e_i = (y_i - \hat{y}_i)$

$$\sum_{i=1}^n e_i = 0$$

Az $S_y^2 = S_{\hat{y}}^2 + S_{(y-\hat{y})}^2 / S_y^2$, felbontjuk komponenseire, így megoszlási viszonyszámokat kapunk:

$$1 = \frac{S_{\hat{y}}^2}{S_y^2} + \frac{S_{(y-\hat{y})}^2}{S_y^2}, \text{ amiből}$$

$$r^2 = \frac{S_{\hat{y}}^2}{S_y^2}, \text{ így} \quad r = \sqrt{\frac{S_{\hat{y}}^2}{S_y^2}}, \text{ vagy}$$

$$r = \sqrt{1 - \frac{S_{(y-\hat{y})}^2}{S_y^2}} \quad (\text{Ezzel a képlettel nem lineáris és többváltozós esetben is}$$

számítható a kapcsolatszorosság. Nem lineáris esetben korrelációs indexnek (I_r), többváltozós esetben totális korrelációnak (R) nevezzük.)

Inverz regresszió:

Az okozati összefüggések egy része szakmai szempontokat figyelembe véve megfordítható. Ilyenkor az inverz regresszió is meghatározható. Az inverz regressziós egyenes:

$$\hat{x} = a + b \cdot y$$

$$a_{x/y} = \bar{x} - b_{x/y} \cdot \bar{y} \quad b_{x/y} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$S_x^- = S_x \sqrt{1 - r^2} \quad S_{x\%}^- = \frac{S_x^-}{\bar{x}} \cdot 100$$

(az eredeti jelölhető $a_{y/x}$

$b_{y/x}$ formában)

- A két egyenes ($\hat{y} = a + b \cdot x$, illetve $\hat{x} = a + b \cdot y$) metszéspontja $\bar{x}, \bar{y}$.
- A két egyenes által bezárt szög utal a kapcsolat szorosságára.
- Ha a két egyenes merőleges egymásra, nincs kapcsolat.
- Minél hegyesebb az egyenesek által bezárt szög, annál erősebb a kapcsolat.

A regressziós együtthatók ismeretében számítható az r^2 :

$$r^2 = b_{y/x} \cdot b_{x/y}$$

illetve r ismeretében a regressziós együtthatók:

$$b_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x}, \text{ illetve } b_{x/y} = r \cdot \frac{\sigma_x}{\sigma_y}$$

Elaszticítás:

Sok esetben nem az abszolút, hanem a relatív változások összefüggései érdekelnek bennünket. Felvetődő kérdéseinkre az elaszticitás vizsgálat adhat választ, vagyis azt tudhatjuk meg, hogy y relatív változása hányszorosa x relatív változásának.

A változást Δ -val jelölve az elaszticitás:

$$E = \frac{\Delta y}{y} : \frac{\Delta x}{x}$$

Deriválható regressziós függvény esetén:

$$E_{y/x} = \frac{dy}{dx} : \frac{y}{x}$$

$$E = \frac{dy}{dx} \cdot \frac{x}{y}$$

Az elaszticitás értéke lehet:

- $0 \rightarrow$ nincs korreláció,
- pozitív $\rightarrow x$ növelése y -t is növeli, x csökkentése y -t is csökkenti,
- negatív $\rightarrow x$ növelése y -t csökkenti, x csökkentése y -t növeli.
- Az elaszticitásnak nincs határozott felső korlátja.

A lineáris regresszió elaszticitása:

$$E(x, y) = b \frac{x_i}{a + bx_i}$$

4.1.3 Nem lineáris kétfváltozós regresszió analízis

Ha x értelmezési tartományában az y -ra gyakorolt hatás mértéke és iránya változó a kapcsolat görbevonallú.

Leggyakoribb típusai:

- Exponenciális: $y = a \cdot b^x$ (Elsősorban idősoroknál!)

Multiplikatív jellegű, így logaritmizálással könnyen linearizálható.

Lineáris forma: $\log y = \log a + x \log b$

- Hatványkitevős: $y = a \cdot x^b$

Multiplikatív jellegű, így linearizálható.

Lineáris alak: $\log y = \log a + b \log x$, ahol a b egyúttal az elaszticitás kifejezője. (A relatív változás állandó.)

Normálegyenletek:

$$\sum \log y = na + b \sum \log x$$

$$\sum \log xy = a \sum \log x + b \sum \log x^2$$

- Hiperbolikus: $y = a + b \frac{1}{x}$ - ez önmagában is lineáris alak

Reciprok transzformáció: $y = \frac{1}{a + bx}$

- Parabolikus: $y = a + bx + cx^2$ (linearizálható)

A másodfokú parabola 3 paraméterének meghatározásához 3 normálegyenlet szükséges:

$$\sum y = na + b \sum x + c \sum x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Magasabb fokszám esetén a normálegyenletek a fenti logika szerint bővíthetők.

Az illeszkedés pontosságának vizsgálata a korábban hivatkozott standard és relatív hiba meghatározásával történik.

A nem lineáris kapcsolatok szorosságát korrelációs index méri. Meghatározása a szórásnégyzet felbontás módszerére épül:

$$\sigma_y^2 = \sigma_y^2 + \sigma_{y-\hat{y}}^2, \text{ innen}$$

$$I_r = \sqrt{1 - \frac{\sigma_{y-\hat{y}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}}$$

Értelmezése és megbízhatóságának vizsgálata a lineáris korrelációs koefficiensével megegyező.

4.1.4 Többváltozós korrelációs kapcsolat

A társadalmi-gazdasági életben sokkal gyakoribb az az eset, hogy egy jelenség alakulására több más tényező fejt ki hatását, vagyis a többváltozós kapcsolatok előfordulása a gyakoribb. Kezelése nehezebb, elektronikus számítógépet igényel, ezért most csak a legszükségesebb tételek tárgyalására szorítkozunk.

- A vizsgálat a regresszió-analízissel kezdődhet, amely a kapcsolatok iránya alapján szintén lehet lineáris és nem lineáris jellegű.

A többváltozós lineáris regressziós függvény egyenlete:

$$\hat{y} = a + bx_1 + cx_2$$

Az illesztés módszere a legkisebb négyzetek elve, amiből levezetve egy háromváltozós függvény paramétereinek meghatározására az alábbi egyenletrendszer szolgál:

$$\Sigma y = n a + b \Sigma x_1 + c \Sigma x_2$$

$$\Sigma x_1 y = a \Sigma x_1 + b \Sigma x_1^2 + c \Sigma x_1 x_2$$

$$\Sigma x_2 y = a \Sigma x_2 + b \Sigma x_1 x_2 + c \Sigma x_2^2$$

A független változók számának növekedésével az egyenletrendszer tovább bővíthető.

Az a paraméter = konstans,

b és c = parciális regressziós együtthatók.

A függvényillesztés hibáját a standard és a relatív hiba nagysága méri.

- A kapcsolat szorosságát többféleképpen is értelmezhetjük.

Vizsgálhatjuk ugyanis:

- a változók együttes kapcsolatának erősségét → **többszörös korrelációs együtthatóval (R)**,
- az eredmény és a tényezőváltozók közötti **kapcsolat szorosságát páronként** → a lineáris kétváltozós esethez hasonlóan,
- a páronkénti kapcsolatszorosságot úgy, hogy a vizsgálatba csak egy tényezőváltozót vonunk be, a többi hatását pedig kiszűrjük → **parciális korrelációs együttható**.

A többszörös korreláció korrelációs indexként is meghatározható, de a páronkénti kapcsolatszorossági mérőszámok korrelációs matrixából is meghatározható, csakúgy, mint a parciális korrelációs együtthatók.

- Célszerű a többváltozós összefüggés-vizsgálatot kiterjeszteni az elaszticitás vizsgálatára is. Természetesen parciális jellegű elemzéseket végzünk. Együtthatójának általános képlete:

$$E_1 = b \cdot \frac{x_1}{a + bx_1 + cx_2} \quad E_2 = c \cdot \frac{x_2}{a + bx_1 + cx_2}$$

4.2 A minőségi ismérvek közötti kapcsolatok vizsgálata

A két minőségi vagy területi ismerv között érvényesülő sztochasztikus kapcsolatot asszociációnak nevezzük. A kapcsolat szorosságát egyetlen számmal, számadattal kívánjuk kifejezni, amely azt méri

meg, hogy a kapcsolat mennyire közelíti meg egyfelől a kapcsolat teljes hiányát, másfelől a függvényszerű kapcsolatot. A mérésre alkalmas mutatószámot asszociációs együtthatónak nevezzük.

Az asszociáció különféle módon fejezhető ki. Mi most kétféle mérőszám meghatározását tekintjük át:

- az alternatív ismérvek közötti kapcsolat szorosságának mérésére alkalmas **Yule-féle együtthatót**,
- a kettőnél több ismerv kapcsolatának szorosságát mérő **Csuprov-féle asszociációs együtthatót**.

4.2.1 A Yule-féle asszociációs együttható

Az alternatív ismérvek mennyiségi ismérvekké történő átalakítását úgy hajtottuk végre, hogy az ismerv egyik változatának 1, a másiknak 0 értéket adtunk.

Az előforduló gyakoriságok ennek megfelelően:

$$\begin{vmatrix} f_{11} & f_{10} \\ f_{01} & f_{00} \end{vmatrix}$$

Függetlenség esetén az:

$$\frac{f_{10}}{f_{11}} = \frac{f_{00}}{f_{01}}$$

vagyis a két koordinációs részviszonyszám egyenlő. Átalakítva:

$$f_{10} \cdot f_{01} = f_{11} \cdot f_{00}$$

ahonnan:

$$f_{11} \cdot f_{00} - f_{10} \cdot f_{01} = 0$$

Amennyiben van kapcsolat a két alternatív ismerv között, a két szorzat különbsége nem nulla. A különbséget ugyanezen szorzatok összegével elosztjuk, a Yule-féle asszociációs együtthatóhoz jutunk:

$$Y = \frac{f_{11} \cdot f_{00} - f_{10} \cdot f_{01}}{f_{11} \cdot f_{00} + f_{10} \cdot f_{01}}$$

Értéke -1 és +1 között helyezkedhet el. Minél közelebb van az együttható abszolút értéke 0-hoz, annál lazább, minél közelebb van 1-hez, annál szorosabb a sztochasztikus kapcsolat. Ha az együttható értéke 0,5 körüli, akkor közepes erősségű kapcsolatról beszélhetünk. Mivel sokszor önkényesen döntjük el, hogy az átalakítás során melyik ismerváltozatnak adunk 1, illetve 0 értéket, az előjelet mindig a konkrét esetben értelmezzük.

4.2.2 A Csuprov-féle asszociációs együttható

Ha a két minőségi ismerv legalább egyike nem alternatív, akkor a kapcsolat szorosságát a Csuprov-féle asszociációs együtthatóval mérhetjük. Számítása a függetlenség-vizsgálatnál tárgyalt χ^2 próbafüggvény alkalmazásával történik.

Képlete tehát:

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

ahol n az egyik, m pedig a másik ismerv változatainak a száma.

A χ^2 értékét felhasználva a Csuprov-féle együttható értékének négyzete:

$$T^2 = \frac{\lambda^2}{N \cdot \sqrt{n-1} \cdot \sqrt{m-1}}$$

ahol:

$$N = \sum_{i=1}^n \sum_{j=1}^m f_{ij}$$

A T^2 értéke olyan szám, amelyik függetlenség esetén 0, függvényszerű kapcsolat esetén pedig 1 értéket vesz fel. Ennek négyzetgyöke a Csuprov-féle asszociációs együttható.

A mutatóval kapcsolatosan megjegyezzük, hogy értékét csak pozitív előjellel értelmezzük. Alkalmazása nem kizárt alternatív ismérvek kapcsolatszorosságának mérésére sem.

Már korábban is utaltunk rá, hogy a minőségi ismérvek kapcsolatának szorosságát másféleképpen is mérhetjük. Alkalmas erre a célra például a rangkorrelációs együttható, abban az esetben, ha a minőségi ismerv változatai között ki tudunk alakítani valamilyen sorrendiséget.

4.3 A vegyes jellegű kapcsolatok vizsgálata

Egy mennyiségi és egy nem mennyiségi ismerv kapcsolatát **vegyes jellegű összefüggésnek** nevezzük. A nem mennyiségi ismerv az esetek döntő részében minőségi vagy területi ismerv, amely szerint csoportosítottuk a megfigyelt adatainkat. A vizsgálat szempontjából azonban nemcsak a csoportképző ismerv változataihoz tartozó - átlag jellegű - értékeknek van jelentősége, hanem a csoporton belüli értékek is szerephez jutnak.

A kapcsolat vizsgálatának elméleti alapját heterogén sokaságok szóródási viszonyainak tanulmányozása képezi. Vizsgálataink során tehát a heterogén sokaságok átlag- és szóródási viszonyait kell tanulmányoznunk, mégpedig arra vagyunk kíváncsiak, hogy a szóródási viszonyok alakulásában milyen szerepe van annak, hogy a sokaság homogén részekre tagolódik.

A Statisztika I. jegyzetben részletesen foglalkoztunk a heterogén sokaságok szóródási viszonyainak tanulmányozásával. Ott megismertük, hogy a heterogén sokaság j -edik csoportjának i -edik értékére vonatkozóan 3 félé eltérés értelmezhető.

Ezek:

$$x_{ji} - \bar{x} = \text{teljes eltérés}$$

$$x_{ji} - \bar{x}_j = \text{belső eltérés}$$

$$\bar{x}_j - \bar{x} = \text{külső eltérés}$$

Összefüggésük:

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

vagyis a teljes eltérés a belső és a külső eltérés összegeként értelmezhető.

Ugyanez az összefüggés az eltérések négyzetére, négyzetösszegére illetve szórásnégyzetére vonatkozóan is fennáll:

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

ahol:

$$\sigma_B^2 = \frac{\sum_{j=1}^m n_j \cdot \sigma_j^2}{\sum_{j=1}^m n_j}, \quad \sigma_K^2 = \frac{\sum_{j=1}^m n_j (\bar{x}_j - \bar{x})^2}{\sum_{j=1}^m n_j}$$

A kettő arányát megoszlási viszonzszám formájában határozhatjuk meg, amely:

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

Az összefüggésből a külső szórásnégyzet arányát szórásnégyzet-hányadosnak nevezzük, értékét pedig felhasználjuk a vegyes jellegű kapcsolatszorossági mérőszám meghatározásához. A szórásnégyzethányados tehát:

$$H^2 = \frac{\sigma_K^2}{\sigma^2}$$

de ugyanazt az értéket kapjuk:

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

A belőle vont négyzetgyök érték a szóráshányados, a vegyes kapcsolatok szorosságának mérőszáma:

$$H = \sqrt{\frac{\sigma_K^2}{\sigma^2}}$$

H értéke: $0 \leq H \leq 1$

0-hoz közel eső érték laza jellegű kapcsolatra utal, míg 1-hez közelítve erősödik az összefüggés. Ha a két ismerv között nincs kapcsolat, a mutató értéke 0. Ez azt is jelenti, hogy a részátlagok nem szóródnak. Ebben az esetben a külső szórás 0, a belső szórás pedig egyenlő a teljes szórással. Az ellenkező szélső esetben az egyes csoportokon belül egyeznek meg az értékek, vagyis a belső szórás értéke 0, illetve a külső szórás a teljes szórással egyezik meg.

4.4 Mintafeladatok az összefüggés-vizsgálat témakörhöz

1. feladat

Hét őstermelő esetében vizsgálták a kiadott tápanyag mennyisége és a termés mennyisége közötti összefüggést. A vizsgálat eredményei:

Őstermelők	Tápanyag menny. (kg)	Termés menny. (t/ha)	Rangsorszám		D_i	D_i^2
			x_i	y_i		
1.	50	3,2	1	1	0	0
2.	100	3,8	2	3,5	-1,5	2,25
3.	1250	3,6	3	2	1	1
4.	150	3,9	4	5	-1	1
5.	175	4,1	5	6	-1	1
6.	200	3,8	6	3,5	2,5	6,25
7.	225	4,2	7	7	0	0
Σ			-	-	0	11,5

Feladat: határozza meg, hogy van-e összefüggés a kiadott tápanyag mennyisége és a termésátlag között! Határozza meg a kapcsolat szorosságát a tanult közelítő és egzakt módszerek segítségével!

Megoldás:

Mindkét ismerv értékeit azonos irányból kiindulva rangsorszámokkal látjuk el. Ha vannak azonos ismervértékek, azok rangsorszámai megegyeznek, méghozzá az egymást követő rangsorszámok átlaga lesz. Ezt követően meghatározzuk az összetartozó rangsorszámok különbségeit (D_i), majd négyzetösszegét és az alábbi képletbe behelyettesítve kapjuk a mérőszámot:

$$r_r = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 11,5}{7(7^2 - 1)} = 0,79$$

Szöveges értékelés: Szoros összefüggés van a kiadott tápanyag mennyisége és a termés mennyisége között!

2. feladat

Kilenc vállalkozás esetében megvizsgálták, hogy van-e kimutatható kapcsolat a vállalatvezető iskolai végzettsége és a vállalat nyereségtermelő képessége között.

Ssz.	Vállalatvezető iskolai végzettsége	Vállalat nyereségtermelő képessége	5		D_i	D_i^2
			x_i	y_i		
1.	alapfok	Közepes	2	5	-3	9
2.	OKJ	Veszteséges	4	1	3	9
3.	Főiskola	Alacsony	7	2,5	4,5	20,25
4.	FSZ	Alacsony	5	2,5	2,5	6,25
5.	alapfok	Közepes	2	5	-3	9
6.	egyetem	Jó	9	7	2	4
7.	alapfok	Közepes	2	5	-3	9
8.	főiskola	Magas	7	8,5	-1,5	2,25
9.	főiskola	magas	7	8,5	-1,5	2,25
Σ	-	-	-	-	-	71

Feladat: Állapítsa meg van-e statisztikailag kimutatható kapcsolat a vállalat vezető iskolai végzettsége és a vállalat nyereségtermelő képessége között!

Megoldás:

Minőségi ismérvek esetén is számolhatunk rangkorrelációs együtthatót abban az esetben, ha az ismérvek rangsorba állíthatók.

Rangsor: a: 1

veszteséges: 1

OKJ: 2

alacsony: 2

FSZ: 3

közepes: 3

Főiskola: 4

jó: 4

Egyetem: 5

magas: 5

Kapcsolt rang: ahol azonos rangsorszámokat osztunk ki!

$$r_r = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 71}{9 \cdot (81 - 1)} = 0,408$$

Közepesen erős kapcsolat van az iskolai végzettség és a vállalat nyereségtermelő képessége között.

3. feladat

Egy borversenyen a négy tagú zsűri a legjobb tíz bor minőségét 1-10-ig terjedő skálán rangsorolta (egy sorszám csak egyszer adható ki):

Sorsz. Bírálok	1	2	3	4	5	6	7	8	9	10	Σ
A	2	3	1	6	4	5	1 0	9	8	7	55
B	1	2	3	7	4	8	5	10	6	9	55
C	2	4	1	3	5	7	6	9	8	10	55

D	1	2	3	6	4	5	1	8	7	9	55
d_i	6	11	8	2 2	1 7	2 5	3 1	36	2 9	35	220
$D_i = d_i - \bar{d} $	16	11	14	0	5	3	9	14	7	13	-
D_i^2	25 6	12 1	19 6	0	2 5	9	8 1	19 6	4 9	16 9	110 2

Feladat: Állapítsa meg van-e összhang a bírálók bírálatai között!

Megoldás: Az összhang mértékét a Kendall-féle egyetértési (konkordencia) együtthatóval vizsgáljuk meg.

d_i : összeadjuk a helyezési sorszámokat függőlegesen;

$$\bar{d} = \frac{\sum d_i}{n} = \frac{220}{10} = 22$$

$$m=4$$

$$n=10$$

$$W = \frac{12 \cdot \sum D_i^2}{m^2 \cdot (n^3 - n)} = \frac{12 \cdot 1102}{4^2 \cdot (10^3 - 10)} = 0,83 \text{ A bírálók értékítéleteiben nagy az összhang.}$$

4. feladat

Hét vállalat méretéről és nyereségtermelő képességéről a következő adatok állnak rendelkezésre:

Ssz.	Vállalat mérete (Mó Ft)	Vállalat nyeresége (Mó Ft)	Átlagtól vett eltérés előjele		Előjelpárok szorzata
1.	25	3	-	-	+
2.	32	5	+	-	-
3.	28	7	-	+	-
4.	41	4	+	-	-
5.	20	3	-	-	+
6.	37	7	+	+	+
7.	40	10	+	+	+
Σ	223	39	-	-	-

Feladat: Állapítsa meg van-e összefüggés a vállalat mérete és nyeresége között, és milyen szoros a kapcsolat a két ismerv között! Olyan közelítő eljárást alkalmazzon, amely részben figyelembe veszi a nagyságrendet is, nemcsak a sorrendiséget!

Megoldás: Előjel-korrelációs együtthatót számolunk!

$$\bar{x} = \frac{\sum x_i}{n} = \frac{223}{7} = 31,9$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{39}{7} = 5,6$$

$$r_e = \frac{p-q}{n} = \frac{4-3}{7} = 0,14$$

Laza jellegű kapcsolat van a vállalat mérete és nyereségtermelő képessége között.

5. feladat

Egy vállalat hét dolgozójának szolgálati idejére és havi átlagkeresetére vonatkozó adatok a következők:

Ssz.	Szolgálati idő (év)	Havi átlagkereset (eFt)	x^2	y^2	xy
1.	12	122	144	14884	1464
2.	7	96	49	9216	672
3.	9	104	81	10816	936
4.	15	110	225	12100	1650
5.	8	125	64	15625	1000
6.	22	96	484	9216	2112
7.	20	148	400	21904	2960
			1447	93761	10794

Feladat: Állapítsa meg, hogy a szolgálati idő és a havi átlagkereset kapcsolatban áll-e egymással és ha igen, milyen szoros a kapcsolat!

Megoldás: A legkisebb négyzetek elvén alapuló korrelációs koefficiens számolunk.

$$\sum x = 93$$

$$\sum y = 801$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{93}{7} = 13,3$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{801}{7} = 114,4$$

$$\sum x^2 = 1447$$

$$\sum y^2 = 93761$$

$$\sum xy = 10794$$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}} = \frac{10794 - 7 \cdot 13,3 \cdot 114,4}{\sqrt{(1447 - 7 \cdot 13,3^2)(93761 - 7 \cdot 114,4^2)}} = 0,214$$

Laza jellegű kapcsolat van a szolgálati idő és a havi átlagkereset között.

Hipotézis vizsgálattal ellenőrizzük, hogy a kapcsolat valóban fenn áll:

H_0 : r értéke a véletlen hatására tér el 0-tól

H_1 : r értéke nem a véletlen hatására tér el 0-tól

Az ellenőrzés próbafüggvénye: t-próba

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} = \frac{0,214}{\sqrt{1-0,214^2}} \sqrt{7-2} = 0,49$$

$$t_{p5\%} = 2,57$$

Mivel $t < t_{p5\%}$, ezért elfogadjuk a nullhipotézist, tehát a korrelációs együttható csak a véletlen hatására tér el nullától. Tehát nem áll fenn korrelációs kapcsolat a két ismerv között.

6. feladat

Egy térség 275 vállalkozásának méret és nyereség szerinti megoszlása 2018-ban a következő volt:

Nyere ség x_i	-	5,1	10,	15,	20,		d	f_{xi}	$f_{xi}d$	$f_{xi}(d_{xi}$
Méret y_i	5,0	-10,0	1-15,0	1-20,0-	1-	f_{xi}	x_i	d_{xi}	x_i^2	d_{yj})
-25,0	⁴ 26 104	² 14 28	⁰ 50	-	-	4 5	- 2	- 90	180	132
25,1- 50,0	² 10 20	¹ 26 26	⁰ 14 0	⁻¹ 5 5	-	5 5	- 1	- 55	55	41
50,1- 75,0	-	⁰ 80	⁰ 22 0	⁰ 14 0	⁰ 60	5 0	0	0	0	0
75,1- 100,0	-	-	⁰ 26 0	¹ 24 24	² 15 30	6 5	1	65	65	54
100,1-	-	-	⁰ 18 0	² 22 44	⁴ 20 80	6 0	2	12 0	240	124
f_{yj}	36	48	85	65	41	2 75	-	40	540	351
d_{yj}	-2	-1	0	1	2	-				
$f_{yj}d_{yj}$	-72	-48	0	65	82	2 7				
$f_{yj}d_{yj}^2$	144	48	0	65	16 4	4 21				
$f_{yj}(d_{yj}d_{yj})$	124	54	0	63	11 0	3 51				

Feladat:

- Határozza meg a kapcsolat szorosságát a két ismerv között!
- Milyen a kapcsolatban rejlő törvényszerűség!
- Határozza meg az inverz regressziós függvény egyenletét!
- Határozza meg, illetve értelmezze a determinációs együtthatót!

Megoldás:

a) A kapcsolat szorosságát a korrelációs együttható fejezi ki:

$$r = \frac{C}{\sigma_x \sigma_y}$$

$$d_{xi} / d_{yj} = \frac{u_i - a}{i}$$

Meghatározzuk mindkét változó átlagát ($\bar{x}, \bar{y}$), szórását (σ_x, σ_y), majd a kovarianciát (C).

$$\bar{x} = a + \frac{\sum f_{xi} d_{xi}}{\sum f_{xi}} \cdot i = 62,5 + \frac{40}{275} \cdot 25 = 66,14$$

$$\bar{y} = a + \frac{\sum f_{yj} d_{yj}}{\sum f_{yj}} \cdot i = 12,5 + \frac{27}{275} \cdot 5 = 12,99$$

$$\sum_{i=1}^n f_{xi} = \sum_{j=1}^m f_{yj} = \sum_{i=1}^n \sum_{j=1}^m f_{ij} = n = 275$$

$$\sigma_x = i \sqrt{\frac{\sum f_{xi} d_{xi}^2}{\sum f_{xi}} - \left(\frac{\sum f_{xi} d_{xi}}{\sum f_{xi}} \right)^2} = 25 \cdot \sqrt{\frac{540}{275} - \left(\frac{40}{275} \right)^2} = 34,84$$

$$\sigma_y = i \sqrt{\frac{\sum f_{yj} d_{yj}^2}{\sum f_{yj}} - \left(\frac{\sum f_{yj} d_{yj}}{\sum f_{yj}} \right)^2} = 5 \cdot \sqrt{\frac{421}{275} - \left(\frac{27}{275} \right)^2} = 6,12$$

$$C = i_{x \cdot y} \left[\frac{\sum \sum f_{ij} (d_{xi} d_{yj})}{\sum \sum f_{ij}} - \left(\frac{\sum f_{xi} d_{xi}}{\sum f_{xi}} \cdot \frac{\sum f_{yj} d_{yj}}{\sum f_{yj}} \right) \right] = 25 \cdot 5 \cdot \left[\frac{351}{275} - \left(\frac{40}{275} \cdot \frac{27}{275} \right) \right] = 157,76$$

Behelyettesítünk a korrelációs együttható képletébe:

$$r = \frac{C}{\sigma_x \sigma_y} = \frac{257,76}{34,8 \cdot 6,12} = 0,74$$

Tehát pozitív irányú, erős kapcsolat van a vállalati méret és a nyereség között.

Hipotézis vizsgálattal ellenőrizzük, hogy a kapcsolat valóban fenn áll:

H_0 : r értéke a véletlen hatására tér el 0-tól

Az ellenőrzés próbafüggvénye: t-próba

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} = \frac{0,74}{\sqrt{1-0,74^2}} \sqrt{275-2} = 47,03$$

$$t_{p5\%}=1,97$$

Mivel $t > t_{p5\%}$, ezért elvetjük a nullhipotézist, tehát a korrelációs együttható nullától való eltérése szignifikáns.

b) Meghatározzuk a lineáris regressziós függvény egyenletét! (A kapcsolatban rejlő törvényszerűséget.)

A lineáris kétváltozós regressziós függvény általános egyenlete:

$$\hat{y} = a + bx$$

Az egyenes felírásához meghatározzuk az a, illetve b regressziós együtthatók értékét.

a=?, b=?

$$b_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x} = 0,74 \cdot \frac{6,12}{34,84} = 0,13$$

Értelmezés: Ha 1 MFt-tal nő a vállalat mérete, várhatóan 0,13 MFt-tal nő a nyereség.

$$a = \bar{y} - b_{y/x} \cdot \bar{x} = 12,99 - 0,13 \cdot 66,14 = 4,39$$

Az a állandót nem értelmezzük, mivel $x=0$ -nál nincs értelme a függvénynek.

A lineáris regressziós függvény egyenlete: $\hat{y} = 4,39 + 0,13x$

A függvényillesztés hibája:

A függvényillesztés standard hibája (reziduális szórás):

$$S_y^- = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

Azonban nagyszámú elem esetén ($n \geq 100$) a következő képlettel célszerű számolni:

$$S_y^- = \sigma_y \sqrt{1 - r^2} = 6,12 \sqrt{1 - 0,74^2} = 4,12$$

Ezt az eredményt viszonyítjuk az $\bar{y}$ -hoz, így megkapjuk a függvényillesztés relatív hibáját:

$$S_{y\%}^- = \frac{S_y^-}{\bar{y}} \cdot 100 = \frac{4,12}{12,99} \cdot 100 = 31,7\%$$

Mivel 20%-nál nagyobb a hiba, ezért nem fogadjuk el, tehát lineáris függvénnyel nem közelíthető a feladat. (Az a jó, ha 10%-nál kisebb a hiba!)

Egy másik megoldás lehet a hiba csökkentése érdekében, hogy növeljük a megfigyelt elemek számát.

c) Az inverz regressziós függvény meghatározása: $\hat{x} = a + by$

$$b_{x/y} = r \cdot \frac{\sigma_x}{\sigma_y} = 0,74 \cdot \frac{34,84}{6,12} = 4,21$$

$$a = \bar{x} - b_{x/y} \cdot \bar{y} = 66,14 - 4,21 \cdot 12,99 = 11,45$$

$$\hat{x} = 11,45 + 4,21y$$

$$S_x^- = \sigma_x \sqrt{1 - r^2} = 34,84 \sqrt{1 - 0,74^2} = 34,06$$

$$S_{x\%}^- = \frac{S_x^-}{\bar{x}} \cdot 100 = \frac{34,06}{66,14} \cdot 100 = 51,5\% \text{ Mivel a hiba igen nagy ezért a lineáris korrelációt nem}$$

fogadom el.

d) A determinációs együttható meghatározása:

$$r^2 = b_{y/x} \cdot b_{x/y} = 0,13 \cdot 4,21 = 0,547$$

Tehát a vállalat mérete átlagosan 54,7%-ban határozza meg a vállalat nyereségének alakulását.

4.5 Önellenőrző kérdések

1. Rendszerezze a függőségi kapcsolatokat!
2. Mit nevezünk sztochasztikus kapcsolatnak!
3. Mit értünk regressziós összefüggés alatt!
4. Mit értünk asszociáció alatt!
5. Mit nevezünk vegyes kapcsolatnak!
6. Mit tekintünk független változónak!

7. Értelmezze a függő változó fogalmát!
8. Mi a korrelációelemzés lényege!
9. Mit értünk regresszioanalízis alatt!
10. Mi a különbség a két-, illetve többváltozós korreláció és regresszioanalízis között!
11. Jellemezze a sztochasztikus függvényeket!
12. Rendszerezze a korreláció fontosabb típusait!
13. Milyen közelítő kapcsolatszorossági mérőszámokat ismer!
14. Mit mutat meg az előjel korrelációs együttható!
15. Hogyan határozzuk meg a rangkorrelációs együtthatót!
16. Hogyan határozzuk meg a korrelációs koefficiens!
17. Mit nevezünk kovarianciának!
18. Sorolja fel a korrelációs együttható tulajdonságait!
19. Mit nevezünk determinációs együtthatónak!
20. Rendszerezze az ún. egzakt kapcsolatszorossági mutatókat!
21. Mit vizsgálunk analitikus regressziószámítás során!
22. Írja fel a kétváltozós regressziós egyenes általános egyenletét!
23. Mit jelent a regressziós egyenes a, illetve b paramétere!
24. Hogyan határozzuk meg a függvényillesztés standard-, illetve relatív hibáját!
25. Mi a reziduális szórás!
26. Mit nevezünk rezidiumnak!
27. Mit értünk inverz regresszió alatt!
28. Mit értünk regressziós elaszticitás alatt!
29. Mikor alkalmazzuk a Yule-féle asszociációs együtthatót!
30. Hogyan határozzuk meg a Csuprov-féle asszociációs együtthatót!
31. Mit értünk heterogén sokaság alatt!
32. Hogyan határozzuk meg a szóráshányados mutatót!

5 AZ IDŐSOROK ÖSSZETEVŐI

Az idősorok az idő függvényében mutatják a jelenségek, folyamatok időbeli változását, alakulását. Az idősorok a valóság egyetlen független relációi. Az idősorok adatai az időben véletlenszerűen lejátszódó, vagyis sztochasztikus folyamatok empirikus adatai.

Sajátossága:

- egyetlen megfigyelés áll rendelkezésre,
- nem lehet megismételni a folyamatot.

A statisztikai sorok közül az idősorokra jellemző leginkább a függvénytípus.

Idősorok elemzésének módszerei:

- ábrák,
- viszonyszámok (dinamikus),
- átlagok:
 - tartam idősorok esetén számtani átlag,
 - időbeli változás: - a fejlődés átlagos mértéke ($\bar{d}$),
- a fejlődés átlagos üteme ($\bar{I}$),
 - állapot idősorok esetén kronologikus átlag,
- Indexek, indexsorok (összetett dinamikus viszonyszámok),
- Idősorok analitikus vizsgálata (ez is összefüggés-vizsgálat, csak itt az idő jelenti a független változót).

Az eddigiekben tanult elemzési lehetőségeken túl az idősorok grafikus képének felbontása, és az összetevők megismerése a célunk. A legösszetettebb idősor grafikus képének hullámzó alakulását 3 tényező idézheti elő:

- **Alapirányzat vagy trend:** az idősorban érvényesülő tartós tendencia, a fejlődés legfontosabb komponense. Azokat az idősorokat, amelyekből a trend hiányzik, illetve amelyek értékei valamilyen állandó érték körül ingadoznak, stacionárius idősoroknak nevezzük.
- **Periodikus ingadozás:** szabályos vagy kevésbé szabályos hullámmás, ami lehet:
 - **Idényszerű vagy szezonális hullámmás:** állandó periódushosszúságú hullámmás, ritmikus ingadozás, általában olyan idősorok esetén állapíthatjuk meg jelenlétét, amelynek adatai egy évnél rövidebb időszakra vonatkoznak. Az idényszerűség ugyanis legtöbbször az évszakok változásával kapcsolatos. De szezonálitást idézhetnek elő a szokások, a tradíciók is.
 - **Konjunkturális ingadozás/hullámmás:** olyan periodikus ingadozás, amely kevésbé szabályos, jelenlétét csak hosszabb idősorok alapján lehet felfedni és tanulmányozni.
- **Véletlen ingadozás:** véletlennek igen sok, egyenként nem jelentős tényező együttes hatását nevezzük, amely zezugossá teszi az idősor vonalát.

Természetesen egy-egy konkrét idősorban nem mindig tapasztalhatjuk valamennyi komponens. A fenti komponensek együttesen is előfordulhatnak, de az is lehet, hogy egyik vagy másik hiányzik az idősorból. Általában csak alapirányzattal és véletlennel kell számolni.

Az elemzés célja: a komponensek megismerése.

Elemzés végrehajtása: az elemzés során arra törekszünk, hogy a vizsgálat tényét lássuk tisztán. Egy tényezőt vizsgálunk, a komponensek közül pedig kiküszöböljük azokat, amelyek hatását nem kívánjuk mérni. Ha például a fejlődés alapirányzatát kívánjuk megismerni, akkor eltekintünk a többi összetevőtől. Az idősort tehát mintegy kisimítjuk, megtisztítjuk a szezonális, a konjunkturális és a véletlen tényezőtől.

Az idősor komponenseinek elemzési módszerei:

- alapirányzat meghatározás→trendanalízissel,
- szezonálitás kimutatása→szezonindex, szezonális eltérés számításával,
- konjunktúra vizsgálat→grafikus elemzéssel,
- véletlen tényező kimutatása→reziduális eltérés-számításával.

Az alkalmazási módszert befolyásolja a komponensek kapcsolódási módja, amely lehet:

- **additív:** összegszerűen egymásra épülő, amikor növekvő vagy csökkenő trend mellett a periodikus ingadozás hullámai közel állandóak,
- **multiplikatív:** szorzatszerűen egymásra épülő, amikor növekvő vagy csökkenő trend mellett a periodikus ingadozás hullámai növekvők vagy csökkenők.

Idősorok vizsgálata esetén kellően hosszú adatsornak kell rendelkezésre állni ahhoz, hogy megalapozott következtetéseket tudjunk levonni.

A következőkben az idősorok egyes komponenseinek részletes elemzése következik.

5.1 Az alapirányzat vagy trend meghatározása

Azt kell felmérni, hogy milyen alapirányzatot mutat az idősor. A trend ugyanis lehet:

- lineáris,
- exponenciális,
- hiperbolikus,
- parabolikus,
- logisztikus.

A lineáris trend meghatározása történhet:

- **valószínű trendvonal becslésszerű megrajzolásával**, úgy, hogy a grafikonon az eredeti értékek pontjait összekötő vonalak és a berajzolt trendvonal által határolt területek összege a trendvonal felett és alatt közelítőleg egyenlő legyen;
- **mozgó átlagolással**, azaz a trendet az eredeti adatok dinamikus átlagaként állítjuk elő;
- **analitikus trendszámítással**, amikor függvényt illesztünk a pontok halmazához.

5.1.1 Mozgóátlagolás

A mozgó átlagolás alapgondolata az, hogy a trendet az eredeti sor dinamikus átlagaként állítjuk elő. A számítás első mozzanata annak eldöntése, hogy hány tagú mozgó átlagot kívánunk számítani, vagyis hány adatot használunk fel egy-egy átlag kiszámításához.

A számítás menete:

- Legyen a mozgóátlag tagszáma: k .
- Kiszámítjuk az idősor első k adatának egyszerű számítani átlagát. Ez az első trendérték, amelyet az érintett időszak közepéhez – vagyis a $(k+1)/2$ -edik időszakhoz – rendelünk.
- Ezután elhagyjuk az első adatot, és helyett vesszük a következő $(k+1)$ -ediket. Ismét átlagot számítva nyerjük a következő mozgó átlagot, vagyis trendértéket, amelyet a megfelelő időszakhoz rendelünk.
- Így haladunk, amíg az utolsó adatot is felhasználjuk. Az eredményül kapott trendértékek sorozata a kiegyenlített idősor.

Speciális háromtagú mozgó átlag esetén a számítást a következő táblázat szemlélteti:

3. táblázat: A háromtagú mozgóátlag számítása

Időszak (Időpont)	Idősor adata y_t	Mozgó összeg $\bar{o}_t$	Mozgó átlag $\hat{y}_t$
1	y_1	-	-
2	y_2	$\bar{o}_2 = y_1 + y_2 + y_3$	$\hat{y}_2 = \bar{o}_2 / 3$
3	y_3	$\bar{o}_3 = y_2 + y_3 + y_4$	$\hat{y}_3 = \bar{o}_3 / 3$
.	.	.	.
.	.	.	.
.	.	.	.

n-2	y_{n-2}	$\ddot{o}_{n-2} = y_{n-3} + y_{n-2} + y_{n-1}$	$\hat{y}_{n-2} = \ddot{o}_{n-2} / 3$
n-1	y_{n-1}	$\ddot{o}_{n-1} = y_{n-2} + y_{n-1} + y_n$	$\hat{y}_{n-1} = \ddot{o}_{n-1} / 3$
n	y_n	-	-

Látható, hogy a háromtagú mozgó átlagok alkalmazása esetén az első és az utolsó időszakhoz (időponthoz) nem tartozik trendérték. Ezáltal eredetileg n adatot tartalmazó idősorunk a kiegyenlítés során $n-2$ tagra rövidült. A mozgó átlag tagszámának növelésével a sor elején és végén „üresen maradó” időszakok száma is növekszik. Pontosan k tagú mozgó átlag használatánál:

- $(k-1)$ adattal rövidül a sor, ha k páratlan,
- k adattal, ha k páros szám. (A páros tagszám esetén követendő eljárást később ismertetjük.)

Páratlan k tagszám esetén az y_t ($t=1,2,\dots,n$) idősorból számított k tagú mozgó átlagok sorozata a $t=j+1$ -edik időszaktól a $t=n-j$ -edik időszakig tart, ahol $j=(k-1)/2$. A t -edik időszakhoz rendelt mozgó átlag:

$$\hat{y}_t^{(k)} = \frac{1}{k} \sum_{i=t-j}^{t+j} y_i$$

A trendtől azt vártuk, hogy megmutassa az idősor alapirányzatát, miközben eltekintünk a többi komponenstől. A számítás módszerének megismerése után meg kell vizsgálnunk, hogyan biztosítja a mozgó átlagolás a kitűzött feladat megvalósítását:

- Az átlagolás kiszűri a véletlen hatását. Nyilvánvaló, hogy a véletlen kikapcsolása annál tökéletesebb, minél nagyobb tagszámú mozgó átlagokat számítunk.
- A mozgóátlag tagszámának helyes megválasztása kiküszöböli a szezonális hatást. Szezonális ingadozásnál ügyeljünk arra, hogy minden egyes mozgó átlag átfogjon egy (vagy több) teljes idényciklust. Ha tehát a jelenség fejlődéséről negyedéves adataink vannak, ezekből számíthatunk négytagú vagy nyolctagú mozgó átlagokat, de pl. öt- vagy hattagú stb. nem megfelelő. A mozgó átlag tagszámát tehát úgy választhatjuk meg, hogy egy-egy ciklushoz tartozó adatok számával egyenlő vagy annak egész számú többszöröse legyen.

Korábban azt mondtuk, hogy a trendértékeket a megfelelő szakasz középső időszakához (időpontjához) rendeljük. Pl. az első háromtagú mozgó átlagnál ez a 2., öttagúnál a 3. időszakhoz stb.

Páros tagszám esetén viszont az az időszak, amelyet a mozgó átlag jellemez, mindig két, eredetileg megadott időszak közé esik, így pl. négy-tagúnál a 2. és a 3. időszak közé. Ezen a helyzeten egy újabb művelet, az ún. középre igazítás, vagy centrírozás beiktatásával segítünk. A középre igazítás úgy történik, hogy a kiszámított mozgó átlagokat páronként rendre átlagozzuk, vagyis újabb ezúttal kéttagú mozgó átlagok sorozatát számítjuk ki. Ezek a trendértékek már a megadott időszakra vonatkoznak.

Amennyiben a trenden kívül csak a véletlen hatása érvényesül az idősorban, az előbbi szempontok nem játszanak szerepet. Itt az idősor hossza számít (hosszabb idősnél inkább lehet nagyobb tagszámot alkalmazni), továbbá előnyös páratlan tagszámot választani. Minél nagyobb a mozgótágok száma, annál jobban kiszűri a véletlen hatását.

A mozgóátlagolás előnye: alkalmas a periodikus és a véletlen tényező kiszűrésére is.

Hátránya: lerövidíti az idősort.

5.1.2 Az analitikus trendszámítás

Alkalmazásának lényege abban áll, hogy az idősor tartós irányzatát valamely analitikusan meghatározott függvénytípussal fejezzük ki. Az alkalmazásra kerülő módszer a regresszió-számítás speciális esete, amikor X az időt és Y valamely időben lejátszódó (sztochasztikus) folyamat adatait jelenti. Így az analitikus trend – a legkisebb négyzetek módszerével becsülve – az a konkrét függvény lesz, amelynek értéke ($\hat{y}_i$) és az ugyanazon időponthoz tartozó tényleges értékek (tapasztalati y_i) közötti eltérések négyzeteinek összege minimális.

Az idő speciális volta a szóba jöhető függvények körét is érinti. Az időskála ugyanis tipikusan különbségskála, így a különböző függvénytípusok közül ki kell ejtenünk a hatványkitevős függvényt. Az időskála logaritmikus transzformációjának ugyanis nem tulajdoníthatunk értelmet.

A leggyakrabban előforduló függvénytípusok:

- Lineáris: $y = a + bx$
- Exponenciális: $y = a \cdot b^x$
- Hiperbolikus: $y = \frac{1}{a + bx}$ illetve $y = a + b \frac{1}{x}$
- Parabolikus: $y = a + bx + cx^2$
- Logisztikus: $y = \frac{k}{1 + m}$

• Lineáris trend:

Amennyiben a vizsgált jelenség egységnyi idő alatt közel állandó értékkel változik az alapirányzat lineáris trenddel fejezhető ki. Általános egyenlete:

$$\hat{y} = a + bx$$

A paraméterek meghatározásához szükséges normálegyenletek:

$$\text{legjobb illesztés: } \Sigma(y - \hat{y})^2 \rightarrow \min.$$

feltételből:

$$\Sigma y = na + b \Sigma x$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2$$

Bizonyos esetekben célszerű az x (idő) értékeket úgy megválasztani, hogy érvényesüljön a $\Sigma x = 0$ feltétel.

Ez úgy érhető el, ha az idősor tagjait középről kiindulva rangsorszámokkal látjuk el:

- páros tagszám esetén: két középső időszak -1 és +1,
felfelé páratlan negatív (-) értékek,
lefelé páratlan pozitív (+) értékek,
- páratlan tagszám esetén: középső időszak lesz 0,
felfelé negatív egész értékek,
lefelé pozitív egész értékek.

A normálegyenletekből így:

$$a = \frac{\sum y}{n} \text{ az idősor átlaga,}$$

$$b = \frac{\sum xy}{\sum x^2} \text{ az átlagos időbeli változás kifejezője:}$$

A lineáris trendfüggvény legfőbb mondanivalóját az iránytangens, a b paraméter képviseli. Ez az idősorra nézve megadja az időszakonkénti átlagos abszolút változás nagyságát. (Erre a célra korábban a fejlődés átlagos mértékét, a $\bar{d}$ -ot használtuk.)

Az a paraméter értelmezése: az időegységre jutó átlagos értéket mutatja.

A b paraméter értelmezése:

- páratlan tagszám esetén az időegységről időegységre bekövetkező átlagos változást mutatja,
- páros tagszám esetén 2b fejezi ki az időegységre jutó átlagos változást.

A függvényillesztés pontosságának vizsgálata a regressziónál ismertetettel azonos (S_y és $S_{y\%}$).

Előrejelzés:

Az $\hat{y} = a + b \cdot x$ képletben az x értékeként az általunk előre jelezni kívánt időszakhoz tartozó értéket helyettesítjük be.

- **Exponenciális trend:**

Akkor alkalmazzuk, ha az idősor egyenletesen gyorsulva vagy lassulva változik, az idősorra a relatív változás viszonylagos állandósága jellemző.

A trendfüggvény általános alakja:

$$\hat{y} = a \cdot b^x$$

Logaritmizált változat:

$$\log \hat{y} = \log a + x \log b$$

A függvény paraméterei:

$$a = \frac{\sum \log y}{n} \quad \text{az idősor mértani átlaga,}$$

$$b = \frac{\sum x \log y}{\sum x^2} \quad \text{az átlagos relatív változás kifejezője:}$$

A b paraméter értelmezése:

- páratlan tagszám – b,
- páros tagszám – b².

• **Hiperbolikus trend:**

Amennyiben az idősoroknál fokozatos lassuló csökkenést, majd stagnálást tapasztalunk a tendenciát hiperbolával írjuk le.

Alkalmazott egyenletek:

$$\hat{y} = \frac{1}{a + bx} \quad \text{illetve} \quad \hat{y} = a + b \frac{1}{x}$$

Linearizálva:

$$\frac{1}{y} = a + bx$$

A paraméterek:

$$a = \frac{\sum \frac{1}{y}}{n} \quad a = \frac{\sum y}{n}$$

$$b = \frac{\sum x \frac{1}{y}}{\sum x^2} \quad b = \frac{\sum \frac{1}{x} y}{\sum \left(\frac{1}{x}\right)^2}$$

• **Parabolikus trend:**

Akkor alkalmazzuk, ha az alapirányzatban növekedések és csökkenések (esetleg stagnálások) váltogatják egymást. A változás fordulópontjainak számától függ az alkalmazandó polinom.

A másodfokú parabola egyenlete:

$$\hat{y} = a + bx + cx^2$$

A $\Sigma(y - \hat{y})^2 \rightarrow \min.$ feltételből levezetett normálegyenletek:

$$\Sigma y = n a + b \Sigma x + c \Sigma x^2$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2 + c \Sigma x^3$$

$$\Sigma x^2 y = a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4$$

Ha $\Sigma x = 0 \rightarrow \Sigma x^3 = 0$

$$\Sigma y = n a + c \Sigma x^2$$

$$\Sigma x^2 y = a \Sigma x^2 + c \Sigma x^4$$

és

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

Ahhoz, hogy mennyi legyen a parabola fokszáma támpontot nyújt a szukcesszív differenciák módszere. A paraméterek becsléséhez pedig felhasználható az ortogonális polinomok módszere is.

• **Logisztikus trend:**

Amennyiben az idősor grafikus képe elnyújtott S alakot mutat az alapirányzatot logisztikus trenddel írhatjuk le. A görbének három jellegzetes szakasza különíthető el:

- kis intenzitással (lanyhán) növekedő,
- nagy intenzitással növekedő és
- telítődő, stagnáló szakasz.

A tendenciát leíró függvény egyenlete:

$$\hat{y} = \frac{k}{1 + m} \quad \text{ahol } m = e^{a+bx}$$

A függvényillesztés módszere a három kiválasztott pont módszer, amelynek alkalmazása során a görbe 3 egymástól egyenlő távolságra lévő (de jellegzetes) pontját választjuk ki, és azt a görbét határozzuk meg, amely áthalad a kiválasztott pontokon.

Az x értékeit úgy választjuk meg, hogy az első kiválasztott ponthoz 0 értéket rendelünk, azt megelőzően a negatív egész, azt követően pedig a pozitív egész értékeket soroljuk fel.

A kiválasztott pontok távolsága n ,

jelölése A , B és C (amelyeket átlagolással a véletlentől megtisztítunk)

A paraméterek számításának képletei:

$$k = \frac{2ABC - B^2(A + C)}{AC - B^2}$$

$$a = \ln \frac{k - A}{A}$$

$$b = \frac{1}{n} \ln \frac{A(k - B)}{B(k - A)}$$

5.2 A periodikus ingadozások vizsgálata

Az elemzés során azt kell bemutatni, hogy a periodicitást előidéző ok a periódus különböző szakaszaiban milyen irányban és milyen mértékben téríti el az idősor adatait a tartós tendenciától.

5.2.1 A szezonálitás vizsgálata

Az idősorok komponenseinek kapcsolódási módjától függően:

- additív kapcsolódás esetén szezonális eltéréseket,
- multiplikatív kapcsolódás esetén pedig szezonindexeket számolunk.

Meghatározásuk menete:

Szezonális eltérés

Szezonindex

1. analitikus vagy mozgóátlagolású trendszámítás

2. trendredukció (trendhatás kiszűrése):

- kivonással ($y - \hat{y}$)

- osztással $\left(\frac{y}{\hat{y}} \right)$

3. a trendre redukált adatokat szezononként átlagoljuk:

- nyers szezonális eltérés

- nyers szezonindex

4. követelmény, hogy a nyers mutatók átlaga:

- 0 legyen

- 1 legyen

5. a feltételek nem teljesülése esetén korrekció:

- nyers eltérés átlaga

- nyers indexek átlaga

6. korrigált értékek, amelyek a szezonális eltérést mutatják.

5.2.2 A konjunkturális ingadozások vizsgálata

Az eddigiekben tárgyalt periodikus ingadozásoknak az volt a jellemzője, hogy az előforduló periódusok hosszában csak viszonylag kis mértékű eltolódások léptek fel, többnyire azonosak voltak. A gyakorlatban azonban olyan idősorok is előfordulnak, amelyekben az ingadozások periódusának a hosszúsága nem állandó. Ezeknek okai:

- részben természetiek (pl. meteorológiai ciklusok),
- másrészt pedig gazdasági jellegűek (pl. gazdasági-konjunktúra ciklusok) lehetnek.

Közgazdaságtanból, gazdaságtörténetből ismeretes a gazdasági és társadalmi fejlődésének ciklikus jellege, amelynek alakulását a fellendülések és válságok változása idézi elő.

A statisztikai elemzésnek az a feladata, hogy mutassa ki a gazdasági ciklust. Ez úgy lehetséges, ha a többi komponensétől megtisztítjuk az idősort.

5.3 Mintafeladatok az idősor témakörhöz

1. feladat

Egy vállalat nyereségére vonatkozó adatok a következőképpen alakultak 2012 és 2018 között:

Év	Nyeresség (MóFt)	Mozgó összeg	Mozgó átlag
2012	12,5	-	-
2013	10,7	37,5	12,5
2014	14,3	41,8	13,93
2015	16,8	48,3	16,1
2016	17,2	49,5	16,5
2017	15,5	50,7	16,9
2018	18,0	-	-

Feladat: Határozza meg mozgóátlagolással a trend értékeit! (3-tagú)!

2. feladat

Egy vállalat 2016 és 2018 közötti forgalmi adatai negyedéves bontásban:

Év	Negyedév	Piaci forgalom (eFt)	4-tagú mozgóösszeg	4-tagú mozgóátlag	Centrírozás (2-tagú)
2016	I.	128	-	-	-
	II.	152	-	-	-
	III.	187	614	153,5	154,875
	IV.	147	625	156,25	159,125
2017	I.	139	648	162	163,625
	II.	175	661	165,25	165,75
	III.	200	665	166,25	166,5
	IV.	151	667	166,75	167,375
2018	I.	141	672	168	175,5
	II.	180	732	183	186,625
	III.	260	761	190,25	-
	IV.	180	-	-	-

Feladat: Mozgóátlagolással határozza meg az idősorban érvényesülő trend értékeit!

Megoldás:

Mivel az idősorban szezonális is érvényesül, ezért lényeges a mozgó átlag tagszámának helyes megválasztása. Szezonális ingadozásnál ügyelnünk kell arra, hogy minden egyes mozgó átlag átfogjon egy (vagy több) teljes idényciklust. Jelen esetben a teljes idényciklus egy év, ami négy

negyedéből áll, így 4-tagú mozgó átlagot kell alkalmazni. Mivel azonban páros a mozgó átlag tagszáma ezért nem szabad elfeledkezni a centírozásról sem.

3. feladat

Egy vállalat 2011 és 2018 közötti nyeresége a következőképpen alakult:

Év	Nyereség (Mó Ft)	x	xy	x ²	$\hat{y}$	$(y - \hat{y})^2$
2011	12,5	-7	-87,5	49	11,747	0,567
2012	10,7	-5	-53,5	25	12,855	4,644
2013	14,3	-3	-42,9	9	13,963	0,114
2014	16,8	-1	-16,8	1	15,071	2,989
2015	17,2	1	17,2	1	16,179	1,042
2016	15,5	3	46,5	9	17,287	3,193
2017	18,0	5	90	25	18,395	0,156
2018	20,0	7	140	49	19,503	0,247
Σ	125	0	93	168	125	12,95

Feladat: Határozza meg és értelmezze a lineáris trend paramétereit!

Megoldás: Ez az idősor tartalmaz alapirányzatot és véletlen hatást!

Először ki kell mutatni a trendhatást (miközben a másik tényező hatását figyelmen kívül hagyjuk)!

A lineáris trend általános egyenlete: $\hat{y}_i = a + b \cdot x_i$

A lineáris trend paramétereit:

$$a = \frac{\sum y_i}{n} = \frac{125}{8} = 15,625$$

Tehát a vizsgált időszakban 15,625 MFt volt az átlagos nyereség.

$$b = \frac{\sum x_i \cdot y_i}{\sum x_i^2} = \frac{93}{168} = 0,554$$

A vizsgált időszakban évente átlagosan 1,108 MFt-tal nőtt a nyereség.

A lineáris trend egyenlete tehát: $\hat{y} = 15,625 + 0,554 \cdot x_i$

Mivel véletlen hatás is érvényesül az idősorban, ezért illeszkedési hiba léphet fel. Ezt is meg kell vizsgálni standard, illetve relatív hiba számításával!

$$S_{\bar{y}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} = \sqrt{\frac{12,95}{6}} = 1,469$$

$$S_{\bar{y}\%} = \frac{S_{\bar{y}}}{\bar{y}} \cdot 100 = \frac{1,469}{15,625} \cdot 100 = 9,4\%$$

Mivel az illeszkedési hiba 10% alatt van, ezért elfogadom a lineáris illesztést, tehát a vizsgált idősor adatai jól illeszkednek a trend egyeneséhez!

4. feladat

Egy vállalkozás bevételi pénzforgalma 2015 és 2018 között negyedéves szezonalitást mutat:

Év	Negyedév	Bevételi forgalom (Mó Ft)	x_i	$x_i \cdot y_i$	$\hat{y}_i$	$(y - \hat{y})^2$
2015	I.	5,5	-15	-82,5	7,3	3,24
	II.	7,0	-13	-91	7,8	0,64
	III.	12,0	-11	-132	8,2	14,44
	IV.	7,5	-9	-67,5	8,6	1,21
2016	I.	6,5	-7	-45,5	9,0	6,25
	II.	9,5	-5	-47,5	9,5	0
	III.	16,0	-3	-48	9,9	37,21
	IV.	10,0	-1	-10	10,3	0,09
2017	I.	8,0	1	8	10,7	7,29
	II.	10,5	3	31,5	11,2	0,49
	III.	14,5	5	72,5	11,6	8,41
	IV.	9,5	7	66,5	12,0	6,25
2018	I.	10,0	9	90	12,5	6,25
	II.	12,0	11	132	12,9	0,81
	III.	18,0	13	234	13,3	22,09
	IV.	12,0	15	180	13,7	2,89
Σ	-	168,5	0	290,5	168,5	117,56

Feladat:

- a) Határozza meg és értelmezze a lineáris trend paramétereit!
- b) Számítsa ki és értelmezze a megfelelő szezonális mutatókat!
- c) Készítsen előrejelzést 2006. III. és IV. negyedévére!

Megoldás:

Ez az idősor tartalmaz alapirányzatot, szezonálítást és véletlen hatást is!

Először ki kell mutatni a trendhatást (miközben a többi tényező hatását figyelmen kívül hagyjuk)!

A lineáris trend általános egyenlete: $\hat{y}_i = a + b \cdot x_i$

A lineáris trend paramétereit:

$$a = \frac{\sum y_i}{n} = \frac{168,5}{16} = 10,53$$

Tehát a vizsgált időszakban 15,625 MFt volt az átlagos bevételi forgalom.

$$b = \frac{\sum x_i \cdot y_i}{\sum x_i^2} = \frac{290,5}{1360} = 0,2136$$

A vizsgált időszakban évente átlagosan 0,42 MFt-tal nőtt a bevételi forgalom.

A lineáris trend egyenlete tehát: $\hat{y} = 10,53 + 0,2136 \cdot x_i$

Most a véletlen hatását küszöböljük ki:

Mivel véletlen hatás is érvényesül az idősorban, ezért illeszkedési hiba léphet fel. Ezt is meg kell vizsgálni standard, illetve relatív hiba számításával!

$$S_{\hat{y}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} = \sqrt{\frac{117,56}{14}} = 2,9$$

$$S_{\hat{y}\%} = \frac{S_{\hat{y}}}{\bar{y}} \cdot 100 = \frac{2,9}{10,53} \cdot 100 = 27,54\%$$

Mivel az illeszkedési hiba 10% felett van, ezért nem fogadom el, tehát lineáris függvénnyel nem lehet jól közelíteni az adatokat!

Végül ki kell mutatni a szezonális ingadozást!

Év	Negyedév			
	I.	II.	III.	IV.
	$y - \hat{y}$			
2015	-1,8	-0,8	3,8	-1,1
2016	-2,5	0	6,1	-0,3
2017	-2,7	-0,7	2,9	-2,5
2018	-2,5	-0,9	4,7	-1,7
Σ	-9,5	-2,4	17,5	-5,6
átlag	-2,375	-0,6	4,375	-1,4

Mivel a nyers szezonális eltérések (átlagok) összege 0, ezért nincs szükség korrekcióra, tehát ezek lesznek egyben a korrigált mutatók is!

Előrejelzés:

2019. I. negyedév: $(10,53 + 0,2136 \cdot 17) - 2,375 = 11,79$ Tehát szezonálitást is figyelembe véve
2019. I. negyedévében várhatóan 11,79 MFt lesz a bevételi forgalom.

2019. III. negyedév: $(10,53 + 0,2136 \cdot 21) + 4,375 = 19,39$ Tehát szezonálitást is figyelembe véve
2019. III. negyedévében várhatóan 19,39 MFt lesz a bevételi forgalom.

5.4 Önellenző kérdések

1. Mit nevezünk idősornak!
2. Rendszerezze az idősorok elemzési módszereit!
3. Mit nevezünk alapirányzatnak/trendnek!
4. Mit nevezünk periodikus ingadozásnak!
5. Értelmezze a periodikus ingadozás típusait!
6. Mit értünk véletlen alatt az idősorok elemzésében!
7. Hogyan hajtjuk végre az idősorok komponenseinek elemzését!
8. Rendszerezze az idősorok komponenseinek elemzési módszereit!
9. Mit értünk additív kapcsolódás alatt!
10. Mit értünk multiplikatív kapcsolódás alatt!
11. Milyen módszerekkel határozhatjuk meg a lineáris trendet!
12. Mi a mozgóátlagolás lényege!
13. Mit nevezünk centírozásnak, mikor alkalmazzuk, mi a lényege!
14. Mi a mozgóátlagolás előnye, illetve hátránya!
15. Írja fel a lineáris trend egyenes általános egyenletét!
16. Mikor alkalmazzuk a szezonális eltéréseket, illetve a szezonindexeket!
17. Mi az oka a konjunkturális ingadozásoknak!

6 GYAKORLÓ VIZSGAFELADATOK

1. feladat

Fogyasztói bejelentés érkezett, hogy a „Pékáru” sütődeből kiszállított 500g szabvány tömegű kenyerekkel sorozatosan megkárosítják a vásárlókat. Egyszerű véletlen mintavétellel ($n=15$) ellenőrizzük, hogy a bejelentés valóban megalapozott-e!

A mintába került kenyerek mért tömege (gramm):

487, 502, 476, 489, 907, 500, 492, 507, 502, 478, 481, 497, 500, 478, 498.

FELADAT: Válaszoljon a bejelentőnek!

Megoldás:

H_0 :

$Szf.=$

H_1 :

$\bar{x}_m =$

$S_m =$

$t =$

$t_{5\%} = 2,15$

Válasz, értékelés:

2. feladat

Szabolcs-Szatmár-Bereg megyében 8 mezőgazdasági vállalkozásnál vizsgálták az egy hektárra jutó költség és jövedelem adatokat:

Sorszám	Költség (eFt)	Jövedelem (eFt)
1.	160	50
2.	220	100
3.	180	80
4.	70	35
5.	100	45
6.	150	50
7.	130	40
8.	200	90

- a) Állapítsa meg, hogy a költség és a jövedelem alakulása között van-e kapcsolat, és ha igen, milyen szoros ez a kapcsolat (C , r , r^2)!
- b) Készítsen összefoglaló szöveges értékelést!

3. feladat

Szabolcs-Szatmár-Bereg megyében 7 kis- és középvállalkozásnál vizsgálták az egy hektárra jutó költség és jövedelem adatokat:

Sorszám	Költség (eFt)	Jövedelem (eFt)
1.	160	50
2.	220	100
3.	180	80
4.	70	35
5.	100	45
6.	150	50
7.	130	40

Határozza meg, milyen a kapcsolatban rejlő törvényszerűség! (Illesszen az alapadatokra kétváltozós lineáris regressziós függvényt!) Értelmezze a b paramétert!

4. feladat

Egy vállalat 8 dolgozója esetén megvizsgálták a havi nettó jövedelem és a kultúrára fordított kiadások összegét, melynek eredményei a következők:

Sorszám	Jövedelem (ezer Ft)	Kultúrára fordított kiadás (ezer Ft)
1.	206	4,4
2.	185	3,8
3.	215	4,6
4.	230	5,1
5.	190	4,3
6.	180	2,9
7.	250	4,1
8.	220	2,7

Kérdések:

- Állapítsa meg a jövedelem és a kultúrára fordított kiadás kapcsolatának szorosságát (C, r, r^2)!
- Készítsen szöveges értékelést!

5. feladat

Egy vállalat 7 dolgozója esetén megvizsgálták a havi nettó jövedelem és a kultúrára fordított kiadások összegét, melynek eredményei a következők:

Sorszám	Jövedelem (ezer Ft)	Kultúrára fordított kiadás (ezer Ft)
1.	206	4,4
2.	185	3,8
3.	215	4,6
4.	230	5,1
5.	190	4,3
6.	250	4,1
7.	220	2,7

Kérdések:

- Határozza meg az összefüggést leíró lineáris regressziós függvény egyenletét!
- Értelmezze a b paramétert!

6. feladat

Két élelmiszer boltban egyszerű véletlen mintavétellel 7-7 elemű mintát vettek a szabvány szerint 1000 g tömegű kenyerekből. Megmérték a kenyerek tömegét, melyek eredményei a következők lettek (g):

1. bolt: 882; 965; 1040; 1005; 1065; 995; 980;

2. bolt: 1020; 890; 1012; 994; 895; 1025; 1040;

Feladat: Állapítsa meg, hogy van-e statisztikailag kimutatható különbség a két élelmiszer boltba kiszállított kenyerek átlagtömegében! ($\alpha=5\%$)

Megoldás: H_0 :

H_1 :

$\bar{x}_1 =$

$\bar{x}_2 =$

$S_1 =$

$S_2 =$

$t =$

$t_{5\%} = 1,99$

Szf. =

Döntés, következtetés:

7. feladat

Egy 250 g-os kávé csomagoló gép működésének ellenőrzéséhez 100 elemű egyszerű véletlen mintát vettek egy véletlenszerűen kiválasztott napon:

A csomagok töltési tömege (g)	A csomagok száma (db)
-235,0	8
235,1-240	22
240,1-245,0	32
245,1-250	28
250,1-	10
Összesen	100

Határozza meg 95%-os megbízhatósággal, hogy milyen határok között van az átlagos nettó töltő súly!

$$z=1,96$$

8. feladat

Egy 250 g-os kávé csomagoló gép működésének ellenőrzéséhez 100 elemű egyszerű véletlen mintát vettek egy véletlenszerűen kiválasztott napon:

A csomagok töltési tömege (g)	A csomagok száma (db)
-235,0	8
235,1-240	22
240,1-245,0	32
245,1-250	28
250,1-	10
Összesen	100

- a)** Becsülje meg a legfeljebb 245 g töltő súlyú üvegek arányát 95%-os megbízhatósági szint mellett!

$$z=1,96$$

9. feladat

Egy külkereskedelmi vállalat 2000 üvegből álló gyümölcskonzerv szállítmányt exportál. (Az alapsokaság eloszlása normálisnak tekinthető.) A nettó töltő súly ellenőrzése céljából egy 110 elemű egyszerű véletlen mintát vettek. A mérések eredményei:

Nettó töltő súly (g)	Üvegek száma (db)
470-480	9
480-490	17
490-500	34
500-510	30
510-520	15
520-530	5
Σ	110

Feladat:

- a) Becsülje meg a 2000 üveg összes nettó töltő súlyát 95%-os megbízhatósági szint mellett!

$$t=1,98!$$

10. feladat

Egy véletlen mintavételen alapuló felmérés szerint a megfigyelt vállalkozások dolgozóinak képzettségi szint és termelés-hatékonysági színvonal szerinti megoszlása:

Hatékonysági színvonal	Képzettségi szint				Σ
	8 általános	szakmunkás	középiskola	főiskola, egyetem	
alacsony	64	56	24	12	
közepes	22	40	64	36	
magas	5	16	44	64	
Σ					

Feladat: Állapítsa meg, hogy a vizsgált két tulajdonság (hatékonysági színvonal és képzettségi szint) összefüggésben áll-e egymással, vagy a kettő független! Amennyiben van kapcsolat, határozza meg a kapcsolat szorosságának mértékét!

Megoldás:

Nullhipotézis:

$$\chi^2 =$$

$$\chi^2_{p5\%} =$$

$$\text{Szf.} =$$

$$T =$$

Döntés, következtetés

11. feladat

Egy közvélemény kutatás során az egyik gazdasági témájú TV műsorról az alábbi kép alakult ki a diplomások körében:

Nyilatkozó foglalkozása	A műsor megítélése			Összesen
	Jó	Megfelelő	Rossz	
Közgazdász	100	200	100	
Jogász	100	60	40	
Egyéb diplomás	100	60	40	
Összesen				

Feladat: Tesztelje 5%-os szignifikancia szinten a foglalkozás jellege és a TV műsor minősítése közötti kapcsolatot!

$$\lambda^2=9,49$$

Határozza meg a kapcsolat szorosságát!

12. feladat

Szabolcs-Szatmár-Bereg megyében 8 ipari vállalkozásnál vizsgálták az egy termékre jutó költség és bevétel adatokat:

Sorszám	Költség (eFt)	Bevétel (eFt)
1.	190	50
2.	230	100
3.	180	80
4.	80	35
5.	100	45
6.	150	50
7.	130	40
8.	200	90

- Állapítsa meg, hogy a költség és a jövedelem alakulása között van-e kapcsolat, és ha igen, milyen szoros ez a kapcsolat!
- Határozza meg, milyen a kapcsolatban rejlő törvényszerűség! (Illesszen az alapadatokra kétváltozós lineáris regressziós függvényt!)
- Ellenőrizze a függvényillesztés hibáját (standard, illetve relatív hiba)!
- Határozza meg, és értékelje a determinációs együtthatót (r^2)!
- Készítsen összefoglaló szöveges értékelést!

13. feladat

Egy térségben – egyszerű véletlen mintavétellel – kiválasztott 7 korlátolt felelősségű társaság 2018. évi vagyonállománya és mérleg szerinti eredménye:

Sorszám	Vagyon	Nyereség
	Millió Ft	
1.	32	6,5
2.	39	8,0
3.	47	10,0
4.	52	9,5
5.	56	12,0
6.	61	13,5
7.	70	15,5

Kérdések:

- Állapítsa meg a vagyon és nyereség kapcsolatának szorosságát (C, r)!
- Határozza meg az összefüggést leíró lineáris regressziós függvény egyenletét!
- Állapítsa meg a függvényillesztés standard, illetve relatív hibáját ($S_y, S_{y\%}$)!

14. feladat

Egy kistérségben megvizsgálták a jogi és nem jogi személyiségű vállalkozások nyereségét (mérleg szerinti eredményét). (Előzetes felmérés alapján tudjuk, hogy a mintába került vállalkozások normális eloszlású alapsokaságból származnak.) Ebből a célból 50 nem jogi és 50 jogi személyiségű vállalkozást választottak ki. Az adataikból számolt jellemzők:

Mérleg szerinti eredmény (MFt)	Nem jogi	Jogi
-5,00	24	13
5,01-10,00	15	17
10,01-15,00	8	10
15,01-20,00	2	7
20,01-25,00	1	3
Σ	50	50

Feladat: Állapítsa meg, hogy van-e statisztikailag kimutatható különbség a nem jogi, illetve a jogi személyiségű vállalkozások átlagos mérleg szerinti eredménye között! ($\alpha=5\%$)

Megoldás: H_0 :

H_1 :

$\bar{x}_1 =$

$\bar{x}_2 =$

$S_1 =$

$S_2 =$

$t =$

$t_{\alpha 5\%} =$

Szf. =

Döntés, következtetés:

15. feladat

Egy almáskertben 5000 db termő almafa található, amelynek 50 %-a Jonathán (J), 30 % Starking (S) és 20 % az Idared (I). Az értékesítés ütemezéséhez termésbecslést végeztek, amihez 300 almafa termését takarították be és mérlegelték. A 300 almafa kiválasztása arányos rétegzéssel, illetve ismétlés nélküli kiválasztással történt. A mérlegelés adatai:

Fánkenti termés (kg/fa)	J	S	I
-140	10	2	6
141-160	15	2	5
161-180	40	13	12
181-200	55	26	15
201-220	25	20	17
221-	5	27	5
Σ	150	90	60

Feladat: 95,5 %-os valószínűségi szinten becsülje meg:

- a fajtánként várható átlagtermést és összes termést,
- az együttesen várható átlagtermést és összes termést!

16. feladat

Egy Kft. legnépszerűbb termékének termelésére vonatkozó adatok a következők:

Év	Termelés mennyisége (ezer tonna)
2011	800
2012	920
2013	1000
2014	1100
2015	1180
2016	1250
2017	1270
2018	1200

Feladat: Számítsa ki és értelmezze a lineáris trend paramétereit! Írja fel a lineáris trend egyenes egyenletét! Készítsen előrejelzést 2020-ra!

17. feladat

Egy mikro vállalkozás nyereségre vonatkozó adatai 2016 és 2018 között:

Év	Negyedév	Nyereség (eFt)
2016	I.	5500
	II.	5580
	III.	5640
	IV.	5530
2017	I.	5540
	II.	5630
	III.	5685
	IV.	5605
2018	I.	5545
	II.	5620
	III.	5670
	IV.	5570

Feladat:

- Határozza meg a lineáris trend egyenletét és értelmezze a paramétereit!
- Határozza meg a szezonális eltéréseket és értelmezze a IV. negyedévit!
- Készítsen előrejelzést 2019. I., II. és III. negyedévére!