Szegedi Tudományegyetem
Gazdaságtudományi Kar

Petres Tibor – Tóth László

STATISZTIKA

II. kötet

2001

Szerzők:

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

Tóth László
PhD-hallgató
Gazdaságtudományi Kar

Második kötet

Tartalomjegyzék

7. Statisztikai minták módszere

7.1. Általában a mintákról

Az 1.3. fejezetben már ismertettük, hogy milyen módszerekkel juthatunk statisztikai adatokhoz. Itt említettük meg azt is, hogy az adatgyűjtés (körét tekintve) lehet teljes vagy részleges, de ezekkel nem foglalkoztunk részletesen. A továbbiakban azonban ennek a témának több figyelmet szentelünk.

Teljes körű megfigyelés

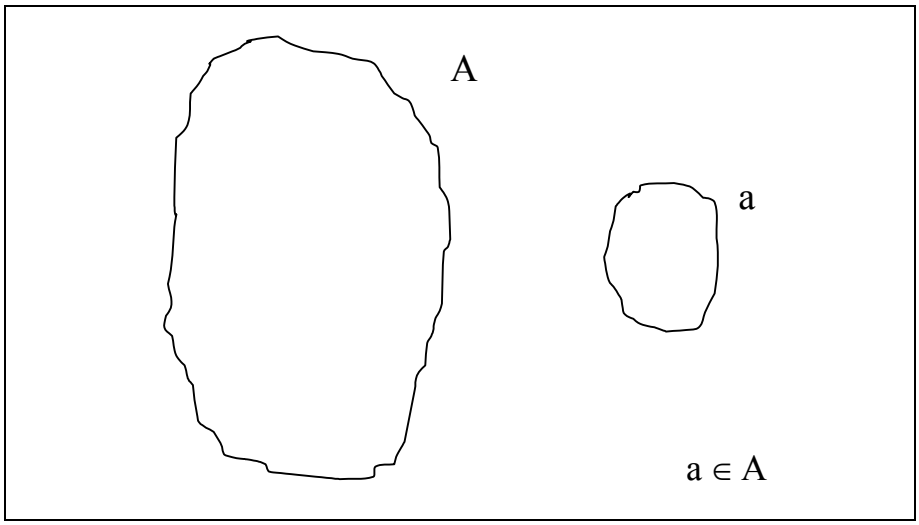
A teljes körű adatfelvétel klasszikus példája a népszámlálás. Népszámlálást már a Római Birodalomban is végeztek. A cenzus szó a népszámlálás szinonimájává vált, és azóta is minden ország statisztikai hivatalának legkomolyabb (legtöbb erőforrást igénylő) feladata. Magyarországon a nemzetközi gyakorlatnak megfelelően általában 10 évenként tartanak népszámlálást. (Megjegyzés: a népszámlálások közötti időszakban egy ún. **mikrocenzust** is lebonyolítanak. Ez azonban nem teljes körű.) Legutóbb 2001-ben volt hazánkban ilyen összeírás. A több milliárd forintba kerülő adatfelvételt a Központi Statisztikai Hivatal (KSH) 2001. február elején kezdte meg. A három hétig tartó munkában megközelítőleg 40 000 számlálóbiztos vett részt. A válaszadás állampolgári kötelesség, az adatszolgáltatás megtagadása pénzbírsággal büntethető. A népszámlálással kapcsolatban a parlament külön törvényt alkot.

Részleges megfigyelés

A népszámlálás példáján világossá vált, hogy egyes gazdasági, társadalmi jelenségek teljes megfigyelésen alapuló vizsgálata nagyon költséges, esetleg lehetetlen. A gyakorlat egyre gyakrabban alkalmazza a részleges adatgyűjtést, különösképpen annak egyik módját, a **reprezentatív megfigyelést**. A reprezentatív adatgyűjtés célja, hogy a sokaság egy részének megfigyeléséből következtessünk annak egészére.

Azt a sokaságot, amelyre a reprezentatív megfigyelés segítségével következtetünk **alapsokaságnak** vagy sokaságnak (jelöljük pl. A-val), az alapsokaság azon részét, amelyet megfigyelünk **mintasokaságnak** vagy **mintának** (jelöljük pl. a-val) nevezzük. Ennek megfelelő illusztráció a 27. ábrán látható.

A mintavétel grafikus modellje



27. ábra

Az alapsokaság lehet véges vagy végtelen, de a mintasokaság mindig véges elemszámú.

Mintavételi és nemmintavételi hiba

A minta alapján a sokasági jellemzők, a nem teljes körű megfigyelés miatt, csak bizonyos hibával közelíthetőek. Fontos azonban megkülönböztetnünk ezt a részlegességből adódó hibát a többi hibalehetőségtől, ezért ezt **mintavételi hibának** fogjuk nevezni. Azokat a hibalehetőségeket, amelyek mind a teljes, mind a részleges megfigyelés során fennállnak **nemmintavételi hibáknak** nevezzük. Ezek (mint például a definíciós, válaszadási, végrehajtási hiba) a statisztikai munka minden fázisában előfordulhatnak.

A tervezés során **definíciós hiba** az, ha a kérdőív pontatlanul, hibásan van megszerkesztve, az adatgyűjtéssel kapcsolatos fogalmak nem tisztázottak, stb.

Az adatgyűjtés során történhetnek **válaszadási hibák**, amikor az adatszolgáltató szándékosan vagy önhibáján kívül a valóságnak nem megfelelő adatokat szolgáltat az adatfelvétel tárgyáról, a megfigyelési egységről.

Az adatfelvétel (a tervezetnek) nem megfelelő elvégzése **végrehajtási hibát** jelent.

Természetesen a feldolgozás fázisában is történhet pontatlanság, például adatrögzítési hiba.

A mintavétel megbízhatóságát a nemmintavételi és a mintavételi hiba nagysága együttesen jellemzi. A nemmintavételi hibák nagyságára csak előző tapasztalatok

alapján vagy szubjektív módón következtethetünk, míg a mintavételi hiba elméleti megfontolásokra támaszkodva matematikai-statisztikai eszközökkel becsülhető. Ezzel a továbbiakban majd külön is foglalkozunk.

A nemmintavételi hiba bemutatására ismertetünk két részleges adatgyűjtést.

Háztartás-statisztika

Az egyik legnagyobb elemszámú mintavételre példa a KSH háztartás-statisztikai felvétele. Évente körülbelül 10 ezer háztartást kérnek fel arra, hogy bevételeikről és kiadásaikról naplót vezessenek. A felvétel 0,2-0,3%-os mintájának statisztikai mutatói természetesen kisebb pontosságúak, mint a teljes körű népszámlálás vagy a 2%-os mintájú mikrocenzus adatai. A mintavételi hibán kívül további torzítást eredményez, hogy a háztartási költségvetési felvételek nem tartalmazzák a legjobb és legrosszabb életkörülmények között élők adatait. Ez a felvétel ugyanis önkéntes, így a leggazdagabb rétegek (nemzetközi tapasztalatok is ezt mutatják) általában elzárkóznak az adatszolgáltatástól. A lakcímmel nem rendelkező hajléktalanok szintén nem kerülnek bele a felmérésbe. A részvétel megtagadása mellett a másik legnagyobb torzító tényező a jövedelmek tendenciózus eltitkolása, általában a gazdagabb háztartásokban, de az alacsonyabb jövedelműek körében is. Az említett jellemzők miatt a háztartás-statisztikai közleményekben a valóságosnál kevesebb magas jövedelmű és több alacsony jövedelmű háztartás szerepel. Ezt szem előtt kell tartani az adatok felhasználása során.

Közvélemény-kutatás

A közvélemény- és piackutatással általában erre szakosodott intézetek foglalkoznak. Ezek adataikat szinte kizárólag mintavételes felvétel útján nyerik. Az egyik leggyakoribb közvélemény-kutatási téma az állampolgárok pártpreferenciájára vonatkozik. Ennek felmérésére általában havonta körülbelül 1000 főt kérdeznek meg személyes megkereséssel. A mintába kerülő személyeket a szavazásra jogosult állampolgárok közül teljes véletlent biztosító módszerrel választják ki úgy, hogy az alapsokaság és a megkérdezettek összetétele megegyezzen. A pártpreferenciák felmérése során több torzító tényező is előfordul, amely nemmintavételi hibát eredményez. Ilyen például az, hogy a szélsőséges pártok szimpatizánsai általában elhallgatják véleményüket, és bizonytalannak mondják magukat a szavazatukat illetően.

A következő példánál (ellentétben ez előző kettővel) a részleges megfigyelés már nem tartalmaz válaszadási hibát.

Gyógyszerek hatásosságának vizsgálata

Újonnan kifejlesztett gyógyszerek hatásosságának vizsgálatára is gyakran alkalmazzák a mintavétel módszereit. Egy adott betegségben szenvedők közül kiválasztanak néhányat, és kezelésnek vetik alá őket. Ezzel párhuzamosan megfigyelnek egy olyan csoportot (kontrollcsoport), amelynek tagjai hatóanyag nélküli gyógyszert, ún. placebót kapnak. Ilyen esetben a statisztika eszközeivel arra kereshetjük a választ, hogy a két csoport egészségi állapotában bekövetkezett változások között van-e statisztikailag jelentős, ún. **szignifikáns** különbség.

7.2. A véletlen mintavétel

Ahhoz, hogy a mintavételi hiba matematikai-statisztikai eszközökkel kezelhető legyen olyan mintát kell választani, amely valamilyen értelemben reprezentálja a sokaságot. Erre egy lehetséges eljárás a **véletlen mintavétel**. A továbbiakban törvényszerűségeket fogunk megfogalmazni olyan mintákra vonatkozóan, amelyek elemeit az alapsokaságból úgy választottuk ki, hogy minden sokasági elem előre adott valószínűséggel kerülhetett a mintába.

(Megjegyzés: a véletlen fogalmával most nem foglalkozunk részletesen, annak értelmezései a valószínűségszámításból ismertek; véletlenen valamilyen valószínűséggel bekövetkező eseményt értünk.)

Véletlen számok előállítása és alkalmazása

Ha a sokaság minden egyes tagjához egy sorszámot rendelünk, akkor a mintavétel véletlenszerűségének biztosításához egy olyan számsort kell megadnunk, amelynek elemei egyenlő valószínűséggel kerültek kiválasztásra. Ilyen számsort háromféleképpen is kaphatunk.

- Sorsolás: például cédulákra felírt sorszámokat húzunk ki egy urnából, amelyet előtte jól megkevertünk.

- **Véletlen számok táblázata**: léteznek olyan táblázatok, amelyek ún. pszeudo-véletlen számsorozatokat tartalmaznak. (Ezeket a számsorozatokat matematikai képletekkel állították elő.) Úgy használjuk őket, hogy kisorsoljuk valamely sorát és oszlopát, és az ott található számtól kezdve folyamatosan kiolvassuk a táblázatban szereplő számokat. Ha a táblázatban szereplő számok közül olyanhoz érünk, amelyik nagyobb a sokaság elemszámánál, akkor azt átugorjuk.

- Gépi sorsolás: a számológépek legtöbbjében van beépített véletlenszám-generátor. Ennek többszöri meghívásával készíthetjük el a mintába kerülő elemek sorszámainak sorozatát. Véletlen számokat az Excel segítségével is kaphatunk. A VÉL() paraméter nélküli függvény meghívásával 0-nál nagyobb vagy egyenlő és 1-nél kisebb egyenletes eloszlású véletlen számot kapunk. (Ezt fel kell szoroznunk a sokaság elemszámával és hozzá kell adnunk egyet, ahhoz hogy sorszámot kapjunk.)

Ennél összetettebb és több beállítási lehetőséget tartalmaz az **Eszközök** menü **Adatelemzés…** almenüjében a Véletlenszám-generálási panel. Itt egy egész tartományt tölthetünk fel egymástól független véletlen számokkal. Az ezt megelőzően ismertetett eljárások egyenletes eloszlású véletlen számokat adnak, mert a leggyakrabban ezt használjuk. A véletlenszám-generálás párbeszéd-paneljében azonban mód van többféle eloszlás beállítására és azok paramétereinek megadására.

A mintajellemzők, mint valószínűségi változók

Egy adott sokaságból egy véletlenszerűen kiválasztott egyed ismérvértéke (a priori) véletlennek tekinthető. Ezt a véletlentől függő ismérvértéket ezért mint valószínűségi változót fogjuk tekinteni. Egy többelemű minta valamilyen jellemző adata szintén valószínűségi változó. Egy adott elemszámú (azonos módon végrehajtott) mintavétel nagyon sokféle mintajellemzőt eredményezhet, a minták statisztikai jellemzői mintáról mintára változhatnak, attól függően, hogy mely sokasági elemek kerültek a mintába. A véletlen mintavétel eredményeként kapott részsokaságot **valószínűségi mintának** is nevezzük.

A fentiekkel való összhang érdekében azt fogjuk feltételezni, hogy diszkrét sokaságaink valószínűségeloszlással, míg folytonos sokaságaink eloszlásfüggvényükkel adottak.

(Megjegyzés: az eddigiekben inkább azt a megközelítést követtük, hogy a sokaságaink elemeik felsorolásával adottak. Ez természetesen csak véges sokaság esetén lehetséges. Igaz persze, hogy a gyakorlatban szinte kizárólag véges sokaságokkal találkozunk, ám a statisztika tárgyából adódóan ezek nagy elemszámú sokaságok, gyakorlatilag végtelennek tekinthetőek. Ezzel szemben a mintát mindig elemeinek felsorolásával adjuk meg, mert az mindig véges.)

Mintaelemek kiválasztása visszatevéssel vagy visszatevés nélkül

A mintavétel során a mintaelemek kiválasztásánál két eltérő módszer létezik. Az egyik szerint a már kihúzott elemeket azonnal visszahelyezzük az alapsokaságba, így ugyanazon elem többször is beválogatható a mintába. Ezt a módszert **visszatevéses**

mintavételnek (leggyakrabban FAE[6]-nek) nevezzük. A másik módszer szerint a kiválasztásra került mintaelemeket nem rakjuk vissza, így minden sokasági egység csak egyszer kerülhet az adott mintába. Ezt a módszert **visszatevés nélküli mintavételnek** (leggyakrabban EV[7]-nek) nevezzük.

Egy N elemszámú sokaságból visszatevéses mintavétellel n elemet

$$k_{\mathrm{FAE}} = N^n \tag{152}$$

féleképpen választhatunk ki.

Egy N elemszámú sokaságból visszatevés nélküli mintavétellel n elemet

$$k_{\mathrm{EV}} = \binom{N}{n} \tag{153}$$

féleképpen választhatunk ki.

58. példa

A 7.1. fejezetben említett háztartás-statisztikai felvétel esetén mennyi a lehetséges minták száma, ha az ország megközelítően 3,8 millió háztartásából veszünk 10 ezres elemszámú mintát?

Legyen $N = 3{,}8 \cdot 10^6$ és $n = 10^4$.

Az összes lehetséges FAE minták száma (152) szerint:

$$k_{\mathrm{FAE}} = \left(3{,}8 \cdot 10^6\right)^{10^4} = \left(3{,}8\right)^{10^4} \cdot \left(10^6\right)^{10^4} = \left(3{,}8^{100}\right)^{100} \cdot 10^{6 \cdot 10^4}.$$

A megfelelő műveletek elvégzése után a következő eredményt kapjuk:

$$k_{\mathrm{FAE}} \approx 6{,}9 \cdot 10^{65\,797}.$$

[6] Az FAE rövidítés arra utal, hogy a visszatevéses mintavétel esetén a mintaelemek független és azonos eloszlású valószínűségi változók, hiszen a mintaelemeket egymástól függetlenül választjuk ki és mindig ugyanabból a sokaságból, az alapsokaságból.

[7] Az EV rövidítés a visszatevés nélküli módszert használó mintavételi terv elnevezésére, az egyszerű véletlen mintavételre utal.

Az összes lehetséges EV minták száma (153) szerint:

$$k_{\mathrm{EV}} = \binom{3{,}8\cdot 10^6}{10^4} = \frac{\left(3{,}8\cdot 10^6\right)!}{\left(10^4\right)!\cdot\left(3{,}8\cdot 10^6 - 10^4\right)!}.$$

Ennek kiszámításához felhasználjuk az ún. **STIRLING-féle összefüggést**:

$$n! = \sqrt{2n\pi}\cdot n^n\cdot e^{-n}\cdot\left(1+\frac{1}{12n}+\frac{1}{288n^2}+\ldots\right),$$

ahol $n>10$ értékekre a zárójelben levő kifejezés elhanyagolható.

Ezt felhasználva:

$$k_{\mathrm{EV}} \approx \frac{\sqrt{2\pi}\cdot\sqrt{3{,}8\cdot 10^6}\cdot\left(3{,}8\cdot 10^6\right)^{3{,}8\cdot 10^6}\cdot e^{-3{,}8\cdot 10^6}}{\sqrt{2\pi}\cdot\sqrt{10^4}\cdot\left(10^4\right)^{10^4}\cdot e^{-10^4}\cdot\sqrt{2\pi}\cdot\sqrt{3{,}79\cdot 10^6}\cdot\left(3{,}79\cdot 10^6\right)^{3{,}79\cdot 10^6}\cdot e^{-3{,}79\cdot 10^6}}.$$

A megfelelő műveletek elvégzése után a következő eredményt kapjuk:

$$k_{\mathrm{EV}} \approx 4{,}6\cdot 10^{30132}.$$

Megjegyzés: a kapott eredmények nagyságrendjének érzékeltetése végett, összevetésül megemlítjük, hogy a Világegyetemünk tömege megközelítőleg „csak” 10^{56} gramm! (Paul Davies: Az utolsó három perc, Kulturtrade Kiadó Kft, Bp., 1994.)

Adott alapsokaság esetén az Excel segítségével is ki tudunk választani véletlen mintát. Vigyük be az alapsokaságunk adatait egy munkatartományba, majd az **Eszközök** menü **Adatelemzés...** almenüjében hívjuk meg a **Mintavétel** menüpontot. A Bemeneti tartomány mezőben adjuk meg az alapsokaságot tartalmazó munkatartományt. Két mintavételi módszer közül választhatunk: A Periodikus időszak: választókapcsoló segítségével szisztematikus kiválasztást (ezt a 7.4. fejezetben részletesebben ismertetjük) végezhetünk, míg a A Véletlen minták száma: választókapcsolóval ismétléses véletlen mintát kapunk.

Az előbbi esetben meg kell adnunk a lépésközt. Ha a program az alapsokaság végére ér, akkor befejezi a mintavételt.

(Megjegyzés: ez a mintavételi módszer csak bizonyos esetekben tekinthető véletlen mintavételi módszernek.)

A Véletlen mintavételi módszert alkalmazva azt tudjuk megadni, hogy a program hány véletlenszerűen kiválasztott cella adatát másolja a Kimeneti tartomány mezőbe.

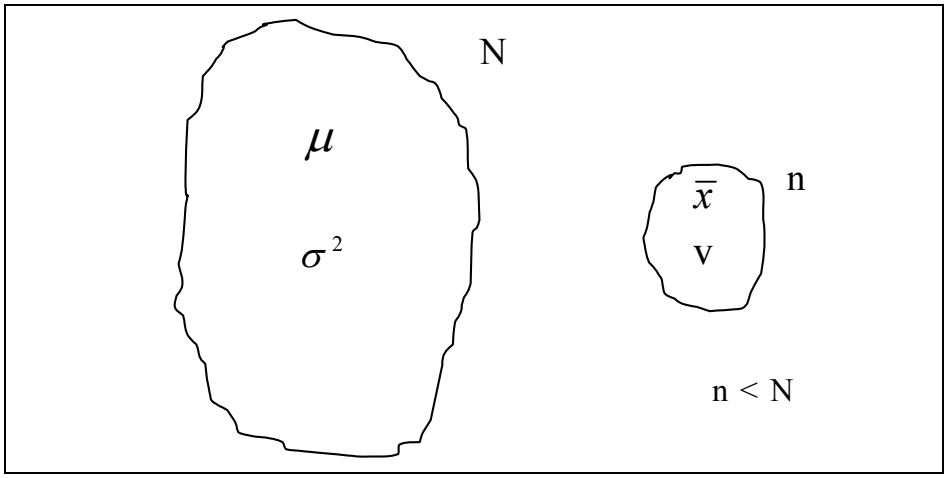
7.3. A mintajellemzők és a sokasági jellemzők kapcsolata

A mintákból a sokaságra vonatkozó következtetések levonását nevezzük **statisztikai indukciónak**. Ezzel a **statisztikai következtetéselmélet** foglalkozik. A továbbiakban azt fogjuk megvizsgálni, hogy melyek azok a törvényszerűségek, amelyek feljogosítanak minket arra, hogy az alapsokaság egy megfelelő módon kiválasztott részsokasága alapján az alapsokaságra vonatkozó állításokat fogalmazzunk meg.

Elemezzük egy adott sokaság esetén az (ebből azonos módon kiválasztható) *n* elemű minták összességét. Ha minden egyes mintára kiszámítjuk valamelyik mintajellemzőt, akkor az adott jellemző eloszlását kaphatjuk meg. A mintajellemzők eloszlását **mintavételi eloszlásnak** nevezzük. Vizsgáljuk most meg, hogy milyen tulajdonságokkal rendelkezik az egyik legfontosabb mintajellemző, a mintából számított átlag (az ún. **mintaátlag)**.

Használjuk a következő jelöléseket: a sokaság elemszáma legyen *N*, várható értéke μ, szórásnégyzete σ^2. A minta elemszáma legyen *n*, a mintaátlag $\bar{x}$, szórásnégyzete pedig v. Ennek megfelelő illusztráció a 28. ábrán látható. (Megjegyzés: ebben a fejezetben tehát v nem a relatív szórást jelöli!)

A sokaság és a minta fontosabb jellemzői



28. ábra

Van-e valamilyen kapcsolat a 28. ábrán feltüntetett (sokasági és minta-) jellemzők között? A (154)-(156) képletek definiálják ezeket a fontos összefüggéseket.

A mintaátlagok mintavételi eloszlása

A 28. ábrán látható minta csak egy az összes lehetséges minta közül. A mintavételi módszertől függően ezek száma (152)-(153) szerint adott. Természetesen mindegyiknek megvan a saját mintajellemzője. Az összes lehetséges mintaátlag gyakorisági sorát az 52. táblázat tartalmazza.

Az összes lehetséges minták átlagainak eloszlása

52. táblázat

Mintaátlagok	Gyakoriságok
$\bar{x}_1$	f_1
$\bar{x}_2$	f_2
$\vdots$	$\vdots$
$\bar{x}_k$	f_k
Összesen	k_{FAE} vagy k_{EV}

A fenti eloszlásnak kitüntetett szerepe van a statisztikában, mert ez az összekötő kapocs a minták és a sokaság között. Mint minden gyakorisági sornak, ennek is van átlaga és szórása. Megkülönböztetésül jelöljük ezeket a következő szimbólumokkal: $\mu_{\bar{x}}$, illetve $\sigma_{\bar{x}}$.

Az összes lehetséges n elemű visszatevéses minták esetén a mintabeli átlagok eloszlásának várható értéke:

$$E(\bar{x}) = \mu_{\bar{x}} = \mu \tag{154}$$

és szórása:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{155}$$

A visszatevés nélküli mintákra fennáll a következő két összefüggés:

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

és

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} . \qquad (156)$$

A mintajellemzők szórásával a mintavételi hibát tudjuk jellemezni, amely szórásnak a statisztikában külön elnevezése van: ezt nevezzük a mintajellemző **standard hibájának**[8]. A standard hiba négyzetét **mintavételi szórásnégyzetnek** nevezzük.

A mintaátlagok eloszlásával kapcsolatban megemlítünk néhány fontos tényt.

– A mintaátlagok eloszlása függ az alapsokaság eloszlásától. Ha az alapsokaság normális eloszlású, akkor a mintabeli átlagok is normális eloszlást követnek.

– Ha $n \geq 30$, akkor az alapsokaság eloszlásától függetlenül a mintaátlagok közelítőleg normális eloszlásúak lesznek $\mu_{\bar{x}}$ várható értékkel (ez a valószínűségszámításból ismert központi határeloszlás tételének következménye) és $\sigma_{\bar{x}}$ szórással. Emiatt a továbbiakban a 30 elemszámúnál nem kisebb mintákat **nagy mintáknak**, a 30-nál kevesebb elemet tartalmazó mintákat pedig **kis mintáknak** fogjuk nevezni.

  A mintaátlagok eloszlása annál jobban közelíti a normális eloszlást minél nagyobb a minta elemszáma. Az ilyen típusú eloszlásokat **aszimptotikusan normális eloszlásoknak** nevezzük.

A normális eloszlás

Az egyik nagyon fontos folytonos eloszlás az ún. **normális eloszlás**, vagy **GAUSS-féle eloszlás**. Ennek két paramétere van, amelyeket μ-vel és σ-val jelölünk. Az eloszlás sűrűségfüggvénye:

[8] A statisztikában fontos szerepe miatt kiemeljük, hogy a standard hiba egy közönséges szórás, csak nem akármelyik eloszlás szórása, hanem a mintavételi eloszlás szórása!

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \qquad (157)$$

A (157) grafikus ábrája az ún. **GAUSS-görbe**.

A normális eloszlást jellemző fontosabb momentumokat és mutatószámokat az 53. táblázat tartalmazza.

A normális eloszlás jellemzői

53. táblázat

várható érték	μ
szórás	σ
ferdeség-mutató (α_3)	0
csúcsosság-mutató (α_4)	3

(157) rövidebb jelölése:

$$x \sim N(\mu, \sigma^2).$$

Megjegyzés: egy normális eloszlású valószínűségi változó a $(-\infty, \infty)$ intervallumban bármilyen értéket felvehet. A gyakorlatban (gazdasági, társadalmi jelenségek vizsgálatánál) ilyen természetesen sohasem fordul elő, de gyakran találkozunk jó közelítéssel normális eloszlásúnak tekinthető sokaságokkal. Például az emberek magasságának, testtömegének, értelmi szintjének, stb. gyakorisági görbéje megközelítőleg GAUSS-görbe alakú. Általában minden olyan jelenség megközelítőleg normális eloszlású, amelyet befolyásoló tényezőkre jellemzőek az alábbiak:

- a tényezők száma nagy és
- egymástól függetlenek,
- egyenkénti hatásuk az összhatáshoz képest kicsi,
- különböző irányúak és intenzitásúak.

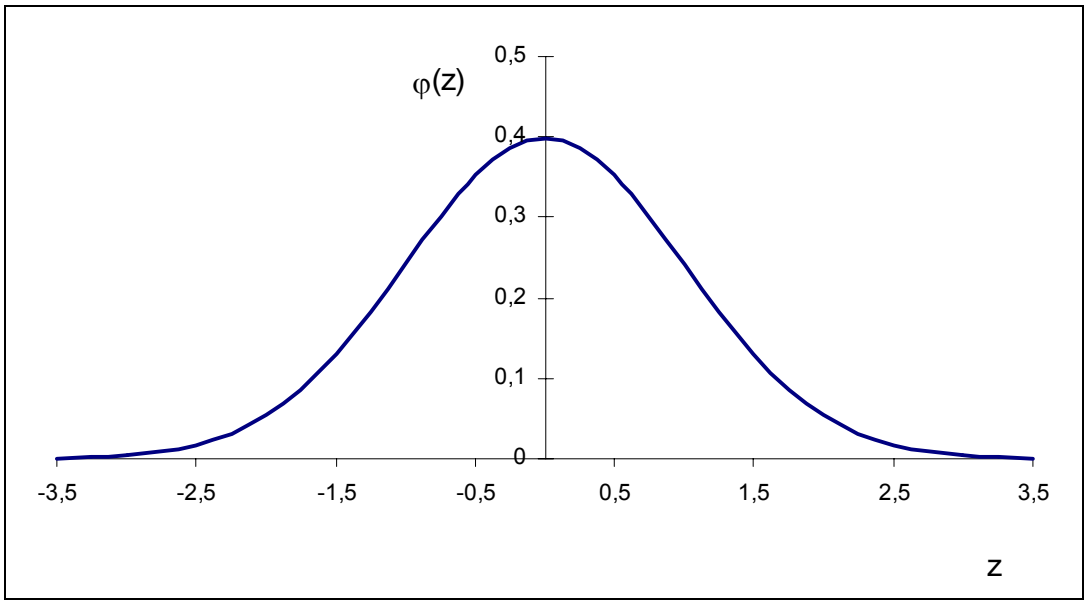
Ha normális eloszlású valószínűségi változónkat (55) szerint standardizáljuk, akkor a transzformált változó **standard normális eloszlású** lesz. (Megjegyzés: az ilyen változókat a statisztikában gyakran *z*-vel vagy *u*-val jelöljük.) Ennek sűrűségfüggvénye:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad (158)$$

grafikonja a 29. ábrán látható.

Megjegyzés: fontossága miatt kiemeljük a $z = 0$ értékhez tartozó valószínűséget. A $\varphi(0) = 0{,}39897 \approx 0{,}4$ minden átlagos (normális eloszlású) tulajdonság előfordulásának valószínűségét mutatja. Mivel (az előzőek alapján) az összes lehetséges mintaátlag is normális eloszlású, a sokaság várható értékével egyenlő mintaátlag előfordulásának van a legnagyobb valószínűsége, körülbelül 40%. A sokaság várható értékétől jelentősen eltérő mintaátlagok előfordulásának valószínűsége ennél jóval kisebb.

A standard normális eloszlás sűrűségfüggvényének grafikonja



29. ábra

A *z* standardizált változó 0 várható értékű és 1 szórású normális eloszlású valószínűségi változó, azaz

$$z \sim N(0,1) .$$

A standardizált változó univerzálisan használható (mivel mértékegység nélküli), azaz különböző típusú sokaságok esetén is alkalmazható összehasonlítás céljára.

A normális eloszlás egyik fontos tulajdonsága a következő:

$$\mu \mp z \cdot \sigma \qquad (159)$$

intervallumban található ($z = 1, 2, 3$ esetén) az összes (29. ábrán látható) görbe alatti terület 68,27; 95,45 és 99,73%-a.

Gyakran azonban szükség van standard normális eloszlású változó eloszlásfüggvényének értékeire akkor is, ha z nem egész szám. Ezekre az esetekre táblázatokat szoktunk használni. Lásd az I. táblázatot! Ebben a különböző z értékek az első tizedes jegyig az első oszlopban szerepelnek, míg a második tizedes az első sorban van. A táblázat belseje tartalmazza az eloszlásfüggvény értékeinek törtrészét. Ebből a táblázatból visszafelé is tudunk keresni: ha a lefedett terület nagysága adott, akkor meg tudjuk mondani az intervallumhoz tartozó z értéket.

A statisztikai irodalomban a (159) szerinti táblázatot legtöbbször nem közlik. Ez azzal magyarázható, hogy az eloszlásfüggvény (definíciójából adódóan) nem a (159) szerint, hanem a $(-\infty, z)$ intervallumban adja meg a 29. ábrán látható görbe alatti területet. Ennek megfelelő értéket a II. táblázat tartalmazza. Mi az összefüggés a két táblázatban közölt adatok között?

Az összefüggés felírása végett, a (159) szerinti valószínűségre vezessük be az $(1-\alpha)$ jelölést. Ebből következik, hogy a kiegészítő valószínűség α-val egyenlő. Például $z = 2$ esetén a valószínűség $100 \cdot (1-\alpha) = 95{,}45\%$; azaz $\alpha = 1 - 0{,}9545 = 0{,}0455$; tehát 4,55%. Figyelembe véve a fentieket, az I. táblázat közvetlenül $(1-\alpha)$-ra, a II. táblázat pedig $\left(1 - \frac{\alpha}{2}\right)$-re adja meg a (159) képlethez szükséges megfelelő z értéket.

Az I. és a II. táblázat értékeit az Excel segítségével számítottuk ki. A statisztikai függvények közül a STNORMELOSZL(z) függvény standard normális eloszlású

változó eloszlásfüggvényének értékeit adja, míg inverzét az INVERZ.STNORM(*valószínűség*) függvény segítségével határozhatjuk meg.

59. Példa

Milyen z értékre lesz a (159) által adott intervallumhoz tartozó terület az összterület legalább 90%-a?

A $z = 1{,}96$ értékhez hány százalékos részterület tartozik?

Az I. táblázatban közölt elméleti értékek alapján mindkét kérdés megválaszolható. Keressük meg a táblázatban a 90%-nak (illetve táblázatunk pontossága szerint 0,90000-nek) megfelelő értéket. (Lásd a 30. ábrát.)

Az I. táblázat része

z	0	...	4	5	6	...	9
⋮							
1,5	86639		87644	87886	88124		88817
1,6	89040		89899	**90106**	90309		90897
1,7	91087		91814	91988	92159		92655
⋮							

30. ábra

Legalább 90%-nak megfelelő terület a vastagon szedett 0,90106. Ebben a sorban z-nek megfelelő szám 1,6; függőlegesen pedig 5; ezért z értéke 1,65 ($z = 1{,}6 + 0{,}05 = 1{,}65$).

A táblázatban közölt adatok alapján a 90%-nak megfelelő pontosabb értéket nem tudunk megállapítani, de az Excel INVERZ.STNORM(0,95) függvényhívás segítségével ez könnyen meghatározható: $z = 1{,}6448530$.

Megjegyzés: az említett Excel függvény paraméterénél figyelembe kell venni azt, hogy *valószínűség* $= (1-\alpha)$ helyett *valószínűség* $= (1-\frac{\alpha}{2})$ -t kell venni, ahol $\alpha = 1-0{,}9$.

A $z = 1{,}96$ értékhez tartozó terület nagyságát szintén meg tudjuk határozni az I. táblázatból és az Excel segítségével is. A táblázatban a 31. ábrán látható módon (vastagon szedett 1,9 és 6 számoknál) keressük a megfelelő értéket.

A keresett érték tehát 0,95000; vagyis $z = 1{,}96$ -hoz 95%-os terület tartozik.

Az I. táblázat része

z	0	...	5	**6**	7	...	9
⋮							
1,8	92814		93569	93711	93852		94124
1,9	94257		94882	95000	95116		95341
2,0	95450		95964	96060	96155		96338
⋮							

31. ábra

Mint már említettük, az összes lehetséges minták átlagai normális eloszlásúak, ezért felírható a következő összefüggés:

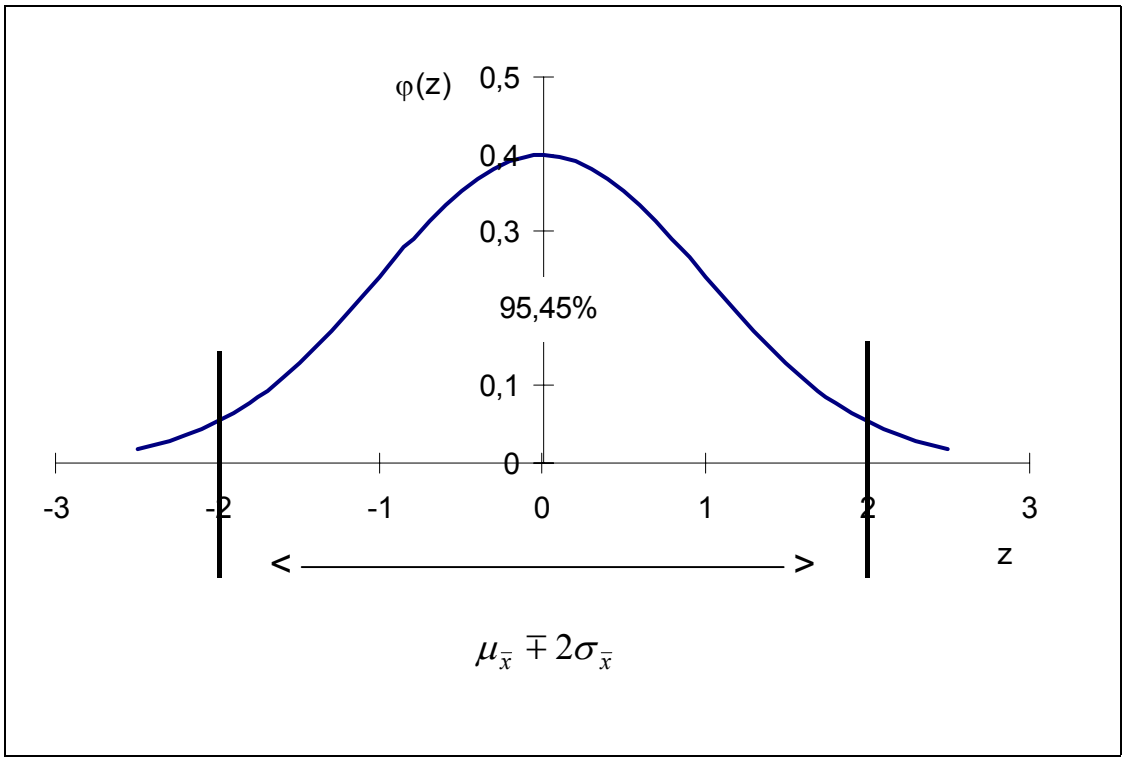
$$\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2). \qquad (160)$$

Ezek szerint, a normális eloszlásra vonatkozó (eddig említett) tulajdonságok a mintaátlagokra is érvényesek. A (159) alapján, igaz a következő összefüggés:

$$\mu_{\bar{x}} \mp z \cdot \sigma_{\bar{x}}. \qquad (161)$$

A 32. ábra a $z = 2$ értékhez tartozó területet illusztrálja.

A mintaátlagok (161) szerinti ábrázolása



32. ábra

60. Példa

Az összes lehetséges mintaátlag hány százaléka található a $\mu_{\bar{x}} \mp 2{,}58 \cdot \sigma_{\bar{x}}$ intervallumban; illetve melyik az az intervallum, amely ezeknek 99,5%-át tartalmazza?

Az I. táblázatban a 2,58 értéknek (2,5 és 8 számok kereszteződésében) 0,99012 vagy 99,012%-os valószínűség felel meg. Tehát (a mintavételi módszertől függően) $0{,}99012 \cdot k_{\mathrm{FAE}}$ vagy $0{,}99012 \cdot k_{\mathrm{EV}}$ mintaátlag található a vizsgált tartományban.

Az I. táblázatban a 99,5%-nál nem kisebb legközelebbi érték 0,99505. Ehhez $z = 2{,}81$ tartozik. A keresett intervallum: $\mu_{\bar{x}} \mp 2{,}81 \cdot \sigma_{\bar{x}}$.

Megjegyzés: az összes lehetséges mintaátlag 100%-át elméletileg a $z = \infty$ értékkel adott intervallum tartalmazza.

7.4. Véletlen mintavételi tervek

Független, azonos eloszlású minta (FAE)

Egyenlő valószínűséggel vett visszatevéses minta esetén **független, azonos eloszlású mintát** (FAE) kapunk. Végtelen sokaságból vett visszatevés nélküli minta is FAE mintának tekinthető, hiszen ebben az esetben a kiválasztott elemek nem befolyásolják a megmaradó sokaság eloszlását. A gyakorlatban a nagy elemszámú sokaságok is (jó közelítésben) végtelennek tekinthetőek. Az empirikus elemzéseknél (a nagy elemszámú sokaságból vett) visszatevés nélküli mintavételi módszert alkalmazzuk leggyakrabban.

Egyszerű véletlen minta (EV)

Ha homogén, véges elemszámú sokaságból visszatevés nélküli kiválasztást alkalmazunk, akkor **egyszerű véletlen mintát** (EV) kapunk.

Egyszerű véletlen minta kiválasztásához gyakran alkalmazzák az ún. **szisztematikus kiválasztást**. Ennek lényege az, hogyha rendelkezünk egy listával a sokaság elemeiről, akkor minden *k*-adik elemet kiválasztva véletlen mintához jutunk, amennyiben a lista sorba rendezésének alapjául szolgáló és a vizsgálni kívánt ismérv független egymástól. A k lépésköz értékét a $k = \left[\frac{N}{n} \right]$ képlettel határozhatjuk meg. A kiválasztás kiindulópontját véletlenszerűen jelöljük ki, majd ettől kezdve minden *k*-adikat kiválasztjuk. Ha a lista végére érünk, akkor folytatjuk a lista elejéről folyamatosan. Ennek a módszernek az előnye egyszerűségében van.

Rétegzett minta (R)

Minden mintavételi tervnél felmerül a következő kérdés: hogyan lehetne olyan módon kiválasztani a mintát, hogy az minél jobban reprezentálja a sokaságot. A 4.1. fejezetben már láttuk, hogy a heterogén sokaságok (valamilyen megfelelően megválasztott csoportképző ismérv szerint) gyakran megközelítőleg homogén részsokaságokra bonthatóak. Ezt használjuk ki a **rétegzett mintavétel** esetén, amelynek végrehajtása a következőképpen történik: először a sokaságot minél homogénebb (a vizsgált ismérv szempontjából kisebb szórású) részsokaságokra (átfedésmentesen és hézagmentesen)

bontjuk szét. Ezeket a részsokaságokat nevezzük **rétegeknek** vagy **sztrátumoknak**. A rétegeken belül ezután egyszerű véletlen mintavételt hajtunk végre.

Heterogén sokaságok esetén a rétegzett mintavétel (ugyanakkora nagyságú mintát feltételezve) általában kisebb mintavételi hibát eredményez, mint az EV vagy FAE minta. Az R minta hatásossága azon múlik, hogy sikerül-e megfelelően homogén rétegeket kialakítani.

A rétegzett mintavétel tárgyalásához a következőkben ismertetett jelölésrendszert alkalmazzuk.

A rétegek számát jelölje M, elemszámaikat pedig rendre:

$$N_1, N_2, ..., N_M;$$

míg a rétegekből kiválasztott elemek száma legyen

$$n_1, n_2, ..., n_M.$$

Ezek alapján a vizsgált sokaság elemszáma:

$$\sum_{j=1}^{M} N_j = N,$$

míg a mintanagyság:

$$\sum_{j=1}^{M} n_j = n.$$

A sztrátumok és a rétegekből vett minták más jellemzőire is indexeléssel utalunk.

A rétegzett mintavételnél döntenünk kell, hogy hogyan osztjuk szét a minta teljes elemszámát (n) a rétegek között. Erre többféle elosztási terv létezik.

– **Egyenletes elosztás:** az egyes rétegekből azonos számú elemet választunk a mintába. A *j*-edik sztratumból kiválasztott minta elemszáma:

$$n_j = \frac{n}{M} \qquad j = 1, 2, ..., M. \tag{162}$$

– **Arányos elosztás:** a rétegek elemszámának sokaságbeli arányát figyelembe véve történik a kiválasztás. A *j*-edik rétegből kiválasztott minta elemszáma:

$$n_j = n\frac{N_j}{\sum_{j=1}^{M} N_j} = n\frac{N_j}{N}. \tag{163}$$

Az arányos elosztás több hasznos tulajdonsággal rendelkezik, ezért a gyakorlatban gyakran alkalmazzák. Ez a mintavételi terv az egyenletes elosztáshoz hasonlóan szintén egyszerű, itt a sokaságban és a mintában ugyanazok a súlyarányok szerepelnek. Ennek következményeként belátható, hogy az arányos elosztással nyert mintából számított főátlag hibája (a rétegezéstől függetlenül) nem lehet nagyobb, mint EV minta esetén.

– **NEYMAN-féle optimális elosztás**: ha ismerjük az egyes részsokaságok vizsgált ismérv szerinti szórását, vagyis az egyes rétegek heterogenitásának mértékét, akkor ezt fel tudjuk használni arra, hogy a sokaságot jobban reprezentáló mintát válasszunk ki. A NEYMAN-féle optimális elosztás esetén a kisebb szórású rétegekből kisebb, míg a nagyobb szórású rétegekből nagyobb mintát veszünk. A *j*-edik rétegből kiválasztott minta elemszáma:

$$n_j = n\frac{N_j\sigma_j}{\sum_{j=1}^{M} N_j\sigma_j}. \tag{164}$$

Ez a mintavétel a főátlagot a legkisebb mintavételi hibával közelíti, de a gyakorlatban mégis ritkán alkalmazzuk, mert a rétegenkénti szórások általában ismeretlenek.

Csoportos minta (CS)

Az eddigi mintavételi terveknél feltételeztük, hogy rendelkezésünkre áll a sokaság összes egyedét tartalmazó lista, ami alapján a kiválasztás elvégezhető. A gyakorlatban ilyennel általában nem rendelkezünk, és elkészítése is nagyon költséges esetleg lehetetlen lenne. Ilyenkor a sokaságot nagyobb összetartozó egységekre bontjuk szét, amelyeknél a lista könnyebben beszerezhető. Ha ezen összetartozó csoportok (pl. területileg) koncentráltan helyezkednek el, akkor egy csoport teljes körű megfigyelése olcsóbb lehet, mint a más tervek szerint kiválasztott nem koncentráltan elhelyezkedő mintaelemek megfigyelése. A **csoportos mintavétel** esetén tehát a homogén sokaságot csoportokra bontjuk szét (általában természetesen adódó módon), és a csoportok halmazából választunk EV mintát, majd a kiválasztott csoportokat teljes körűen megfigyeljük. A csoportos mintavétel általában egyszerűbbé és olcsóbbá teszi a felvételt. Pontossága a csoportokon belüli homogenitástól függ. A csoportos mintavétel esetén a rétegzettel ellentétben az ad hatásosabb becslést, ha a csoportok heterogének, hiszen minden elemüket megfigyeljük, így homogén csoportok esetén ez redundáns és rontja a hatásosságot.

Fontossága miatt még egyszer kiemeljük, hogy a rétegzett mintavétel akkor hatásos, ha (a megfigyelt ismérv szempontjából) a sokaság heterogén és a rétegek homogének, míg a csoportos mintavétel akkor hatásos, ha a sokaság homogén és a csoportok heterogének.

Többlépcsős minta (TL)

A **többlépcsős mintavételt** hasonló esetekben alkalmazzuk, mint a csoportos mintavételt. Ennél a mintavételi tervnél több lépésben jutunk el a megfigyelési egységekhez. A leggyakoribb a kétlépcsős mintavétel, amelynek során (a csoportos mintához hasonlóan) csoportokat (elsődleges megfigyelési egység) választunk ki a sokaságból, de nem figyeljük meg ezeket teljes körűen, hanem újabb mintavételt alkalmazunk a csoportokon belül. A többlépcsős mintavétel előnye, hogy az elsődleges megfigyelési egység homogenitása esetén csökkenti a megfigyelés redundanciáját, így növeli a hatásosságot. A TL minta elosztásának kérdése bonyolultabb az egylépcsős mintákénál, általában arra törekszünk, hogy a végső minta a sokasági arányoknak megfelelő legyen.

Az említett mintavételi terveken kívül még számos más is ismeretes, de könyvünkben ezekkel nem foglalkozunk.

A következő két fejezetben csak az FAE, EV és R minták alkalmazásával foglalkozunk.

8. Minta alapján történő becslések

8.1. Becslőfüggvények és tulajdonságaik

Ahogy azt a 7. fejezetben már megállapítottuk, célunk az, hogy minta alapján következtessünk az alapsokaságra, illetve annak valamelyik jellemzőjére. Ebben a fejezetben olyan módszerekkel foglalkozunk, amelyek segítségével egy sokaság valamely jellemzőjét vagy eloszlását, illetve egy statisztikai modell valamilyen paraméterét tudjuk közelítőleg meghatározni.

A becslésünk tárgyát képező sokasági jellemzőt a továbbiakban Θ -val jelöljük.
A sokasági jellemző mintából történő közelítő meghatározására szolgáló statisztikát **becslőfüggvénynek** nevezzük. Az $x_1, x_2, ..., x_n$ mintaelemekhez tartozó becslőfüggvényre a következő jelöléssel hivatkozunk:

$$\hat{\Theta}(x_1, x_2, ..., x_n) = \hat{\Theta}_n = \hat{\Theta}.$$

A becslőfüggvény tehát olyan statisztika, amely a sokasági jellemzőt a mintajellemzők valamilyen függvényével közelíti, és mivel értéke a mintaelemektől függ, vagyis mintáról mintára változik, ez is valószínűségi változónak tekinthető. (A mintavétel végrehajtása után természetesen mind a minta, mind a becslőfüggvény értékei realizálódnak, tehát a posteriori módon már nem tekinthetőek valószínűségi változóknak.)

Először a pontbecsléssel, majd az intervallumbecsléssel foglalkozunk. **Pontbecslés** esetén (a becslőfüggvényünk segítségével) a mintához egyetlen számszerű értéket rendelünk, és ezt tekintjük a becsülni kívánt paraméter értékének. **Intervallumbecslés** esetén azonban egy olyan intervallumot határozunk meg, amely előre adott nagy valószínűséggel tartalmazza a becsülni kívánt paramétert.

Egy sokasági jellemző becslésére természetesen többféle becslőfüggvény is készíthető. A kérdés az, hogy hogyan lehet ezeket a statisztikákat összehasonlítani, és kiválasztani közülük a legjobbat. A becslőfüggvényeket, mint minden más valószínűségi változót, kézenfekvő eloszlásukkal, várható értékükkel és varianciájukkal jellemezni.

Torzítatlanság

A legalapvetőbb kritérium a becslőfüggvényekkel szemben, hogy értékük (a különböző mintákon) a sokasági jellemző körül ingadozzon. **Torzítatlannak** nevezünk egy becslőfüggvényt, ha annak várható értéke a becsülni kívánt sokasági jellemzővel egyenlő. Vagyis:

$$E(\hat{\Theta}) = \Theta . \tag{165}$$

A torzítás mértékét a

$$Bs(\hat{\Theta}) = \Theta - E(\hat{\Theta}) \tag{166}$$

mérőszámmal szoktuk kifejezni.[9)]

Bizonyos statisztikáknál előfordul, hogy a torzítás mértéke függ a mintanagyságtól. Ha a mintanagyság minden határon túl történő növelésekor a becslőfüggvény torzítatlanná válik, vagyis

$$\lim_{n\to\infty} Bs(\hat{\Theta}_n) = 0 ,$$

akkor azt mondjuk, hogy **aszimptotikusan torzítatlan**. A torzítatlan becslőfüggvények természetesen szintén aszimptotikusan torzítatlanok.

Azt már láttuk, hogy az FAE és az EV mintából számított mintaátlag a sokasági várható érték torzítatlan becslése, mivel (154) szerint:

$$E(\bar{x}) = \mu .$$

A 3. fejezetben taglaltak szerint, az átlag, illetve a várható érték mellett a sokaságok másik legfontosabb jellemzője a szórás, illetve annak négyzete a variancia. A mintából számított szórásnégyzet, amelyet **tapasztalati szórásnégyzetnek** nevezünk, torzítottan becsüli a sokasági varianciát. A torzítás mértéke FAE minta esetén:

$$Bs(\mathrm{v}) = \frac{\sigma^2}{n} .$$

[9)] A 'torzított' szó angol megfelelője: biased.

Ha képezzük az

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \tag{167}$$

illetve

$$s^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1} \tag{168}$$

becslőfüggvényt, akkor a sokasági variancia torzítatlan becslését kapjuk.

$$E(s^2) = \sigma^2 \tag{169}$$

A (167)-(168) segítségével definiált mintajellemzőt **korrigált tapasztalati szórásnégyzetnek**, négyzetgyökét korrigált tapasztalati szórásnak nevezzük.

EV minta esetén s négyzetét (170) szerint még egy korrekciós tényezővel kell szoroznunk, hogy torzítatlan becslőfüggvényt kapjunk.

$$E\left(s^2 \cdot \frac{N-1}{N}\right) = \sigma^2 \tag{170}$$

61. példa

A 22. példánál a 11. táblázat a kötelező gépjármű-biztosítással foglalkozó társaságok díjbevételeinek adatait tartalmazza 1999 első negyedévére. Ugyanezeket az adatokat tartalmazza az 54. táblázat is, de most nem ezer, hanem millió Ft-ban.

Megjegyzés: ezt a példát csak szemléltető igazolás céljából tárgyaljuk, a valóságban ilyen kis elemszámú sokaságnál mindig teljes körű felmérést alkalmazunk (nem pedig mintavételt)!

1999 első negyedévének díjbevételei

54. táblázat

Biztosítók	Díjbevételek (millió Ft)
Argosz	428
Axa Colonia	479
ÁB-Aegon	1 986
Generali-Providencia	3 456
Hungária	8 138
Közlekedési Biztosító Egyesület	100
OTP-Garancia	1 155
Összesen	15 742

Forrás: ÁBIF

Az adott sokaságból származó összes lehetséges minta alapján vizsgáljuk meg, hogy torzítatlan becslőfüggvény-e az $\bar{x}$, a v, az s, az s^2 és az $s^2 \cdot \frac{N-1}{N}$!

A sokaság 7 elemű: $N = 7$.

A sokaság elemei: 428, 479, 1986, 3456, 8138, 100, 1155.

A sokasági átlag: $\overline{X} = 2248{,}86$.

A sokasági szórás: $\sigma = 2631{,}41$; a variancia: $\sigma^2 = 6\ 924\ 330{,}98$.

Számításainkhoz vegyünk pl. kételemű mintákat!

Tekintsük először az FAE mintákat.

Az összes lehetséges kételemű FAE minták száma a (152) képlet szerint:

$$k_{\text{FAE}} = 7^2 = 49.$$

Ezeket a mintákat és a mintákból kiszámított mutatókat az 55. táblázat tartalmazza (ahol $i = 1{,}2{,}...{,}49$).

Az összes lehetséges kételemű FAE minta és néhány jellemzője

55. táblázat

Mintaelemek	$\bar{x}_i$	v_i	s_i^2	s_i
428 , 428	428,00	0,00	0,00	0,00
428 , 479	453,50	650,25	1 300,50	36,06
428 , 1986	1 207,00	606 841,00	1 213 682,00	1 101,67
428 , 3456	1 942,00	2 292 196,00	4 584 392,00	2 141,12
428 , 8138	4 283,00	14 861 025,00	29 722 050,00	5 451,79
428 , 100	264,00	26 896,00	53 792,00	231,93
428 , 1155	791,50	132 132,25	264 264,50	514,07
479 , 428	453,50	650,25	1 300,50	36,06
479 , 479	479,00	0,00	0,00	0,00
479 , 1986	1 232,50	567 762,25	1 135 524,50	1 065,61
479 , 3456	1 967,50	2 215 632,25	4 431 264,50	2 105,06
479 , 8138	4 308,50	14 665 070,25	29 330 140,50	5 415,73
479 , 100	289,50	35 910,25	71 820,50	267,99
479 , 1155	817,00	114 244,00	228 488,00	478,00
1986 , 428	1 207,00	606 841,00	1 213 682,00	1 101,67
1986 , 479	1 232,50	567 762,25	1 135 524,50	1 065,61
1986 , 1986	1 986,00	0,00	0,00	0,00
1986 , 3456	2 721,00	540 225,00	1 080 450,00	1 039,45
1986 , 8138	5 062,00	9 461 776,00	18 923 552,00	4 350,12
1986 , 100	1 043,00	889 249,00	1 778 498,00	1 333,60
1986 , 1155	1 570,50	172 640,25	345 280,50	587,61
3456 , 428	1 942,00	2 292 196,00	4 584 392,00	2 141,12
3456 , 479	1 967,50	2 215 632,25	4 431 264,50	2 105,06
3456 , 1986	2 721,00	540 225,00	1 080 450,00	1 039,45
3456 , 3456	3 456,00	0,00	0,00	0,00
3456 , 8138	5 797,00	5 480 281,00	10 960 562,00	3 310,67
3456 , 100	1 778,00	2 815 684,00	5 631 368,00	2 373,05
3456 , 1155	2 305,50	1 323 650,25	2 647 300,50	1 627,05
8138 , 428	4 283,00	14 861 025,00	29 722 050,00	5 451,79
8138 , 479	4 308,50	14 665 070,25	29 330 140,50	5 415,73
8138 , 1986	5 062,00	9 461 776,00	18 923 552,00	4 350,12
8138 , 3456	5 797,00	5 480 281,00	10 960 562,00	3 310,67
8138 , 8138	8 138,00	0,00	0,00	0,00
8138 , 100	4 119,00	16 152 361,00	32 304 722,00	5 683,72
8138 , 1155	4 646,50	12 190 572,25	24 381 144,50	4 937,73
100 , 428	264,00	26 896,00	53 792,00	231,93
100 , 479	289,50	35 910,25	71 820,50	267,99
100 , 1986	1 043,00	889 249,00	1 778 498,00	1 333,60
100 , 3456	1 778,00	2 815 684,00	5 631 368,00	2 373,05
100 , 8138	4 119,00	16 152 361,00	32 304 722,00	5 683,72
100 , 100	100,00	0,00	0,00	0,00
100 , 1155	627,50	278 256,25	556 512,50	746,00
1155 , 428	791,50	132 132,25	264 264,50	514,07
1155 , 479	817,00	114 244,00	228 488,00	478,00
1155 , 1986	1 570,50	172 640,25	345 280,50	587,61
1155 , 3456	2 305,50	1 323 650,25	2 647 300,50	1 627,05
1155 , 8138	4 646,50	12 190 572,25	24 381 144,50	4 937,73
1155 , 100	627,50	278 256,25	556 512,50	746,00
1155 , 1155	1 155,00	0,00	0,00	0,00
Átlag:	2 248,86	3 462 165,49	6 924 330,98	1 828,49

Vizsgáljuk meg, hogy melyik becslőfüggvény torzítatlan, vagyis melyiknek a várható értéke egyezik meg a becsülni kívánt sokasági jellemzővel.

$$E(\bar{x}) = \frac{1}{49} \cdot 428 + ... + \frac{1}{49} \cdot 1155 = 2248{,}86 = \overline{X}$$

A vártnak megfelelően a mintaátlag torzítatlanul becsüli a sokasági várható értéket.

$$E(\mathrm{v}) = \frac{1}{49} \cdot 0{,}00 + \frac{1}{49} \cdot 650{,}25 + ... + \frac{1}{49} \cdot 0{,}00 = 3\ 462\ 165{,}49 \neq \sigma^2 = 6\,924\,330{,}98$$

$$E(\mathrm{s}^2) = \frac{1}{49} \cdot 0{,}00 + \frac{1}{49} \cdot 1300{,}50 + ... + \frac{1}{49} \cdot 0{,}00 = 6\,924\,330{,}98 = \sigma^2 = 6\,924\,330{,}98$$

$$E(s) = \frac{1}{49} \cdot 0{,}00 + \frac{1}{49} \cdot 36{,}06 + ... + \frac{1}{49} \cdot 0{,}00 = 1828{,}49 \neq \sigma = 2631{,}41$$

Ez alapján azt látjuk, hogy a (nem korrigált) tapasztalati szórásnégyzet (v) torzítottan, míg a korrigált tapasztalati szórásnégyzet (s^2) torzítatlanul becsüli a sokasági szórásnégyzetet. Fontos összefüggés azonban, hogy a sokasági szórást a korrigált tapasztalati szórás is torzítottan becsüli, tehát

$$E(s) \neq \sigma\ .$$

Tekintsük most az EV mintákat.
Az összes lehetséges kételemű EV minták száma a (153) képlet szerint:

$$k_{\mathrm{EV}} = \binom{7}{2} = 21\,.$$

Ezeket a mintákat és a mintákból kiszámított mutatókat az 56. táblázat tartalmazza (ahol $i = 1,2,...,21$).

Az összes lehetséges kételemű EV minta és néhány jellemzője

56. táblázat

Mintaelemek	$\bar{x}_i$	$s_i^2 \cdot \frac{N-1}{N}$
428 , 479	453,50	1 114,71
428 , 1986	1 207,00	1 040 298,86
428 , 3456	1 942,00	3 929 478,86
428 , 8138	4 283,00	25 476 042,86
428 , 100	264,00	46 107,43
428 , 1155	791,50	226 512,43
479 , 1986	1 232,50	973 306,71
479 , 3456	1 967,50	3 798 226,71
479 , 8138	4 308,50	25 140 120,43
479 , 100	289,50	61 560,43
479 , 1155	817,00	195 846,86
1986 , 3456	2 721,00	926 100,00
1986 , 8138	5 062,00	16 220 187,43
1986 , 100	1 043,00	1 524 426,86
1986 , 1155	1 570,50	295 954,71
3456 , 8138	5 797,00	9 394 767,43
3456 , 100	1 778,00	4 826 886,86
3456 , 1155	2 305,50	2 269 114,71
8138 , 100	4 119,00	27 689 761,71
8138 , 1155	4 646,50	20 898 123,86
100 , 1155	627,50	477 010,71
Átlag:	2 248,86	6 924 330,98

$$E(\bar{x}) = \frac{1}{21} \cdot 453{,}5 + \ldots + \frac{1}{21} \cdot 627{,}5 = 2248{,}86 = \bar{X}$$

$$E\left(s^2 \cdot \frac{N-1}{N} \right) = \frac{1}{21} \cdot 1114{,}71 + \ldots + \frac{1}{21} \cdot 477010{,}71 = 6\,924\,330{,}98 \; = \sigma^2 = 6\,924\,330{,}98$$

Hatásosság

Egy torzítatlan becslőfüggvénynek lehet olyan nagy szóródása, hogy ez használhatatlanná teszi. A becslőfüggvény szórása a véletlen tényező okozta hiba mérőszámának tekinthető. Ezt a szórást a becslőfüggvény, illetve a becslés standard

hibájának nevezzük. A becslőfüggvénnyel szembeni további elvárt tulajdonság tehát, hogy szórása a lehető legkisebb legyen.

A 7.3. fejezetben említettekhez hasonlóan, a becslőfüggvény összes lehetséges mintán felvett értékeinek szórásnégyzetét mintavételi szórásnégyzetnek nevezzük. Jelölése: $\mathrm{var}(\hat{\Theta})$. A mintavételi szórásnégyzet négyzetgyöke a becslés standard hibája. Jelölése: $Se(\hat{\Theta})$ [10].

$$Se(\hat{\Theta}) = \sqrt{\mathrm{var}(\hat{\Theta})} .$$

A torzítatlan becslőfüggvényeket hatásosság szempontjából szórásnégyzetükkel vagy szórásukkal hasonlítjuk össze, a kisebb szórású statisztikát **hatásosabbnak** (efficiensebbnek) nevezzük.

Vegyük például a következő esetet: legyen a sokasági várható érték becslőfüggvénye a mindenkori minta első eleme, azaz $\hat{\Theta} = x_1$. A mintaátlaghoz hasonlóan ez a statisztika is torzítatlanul becsüli a várható értéket, de ennek standard hibája például FAE minta esetén $Se(x_1) = \sigma$, míg a mintaátlagé a (155) szerint $Se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$. Ebből következik, hogy az utóbbi hatásosabb becslése a várható értéknek.

Bizonyos esetekben létezik olyan torzítatlan becslőfüggvény, amelynél kisebb szórásnégyzetű statisztika nem készíthető. Az ilyen becslőfüggvényeket **minimális szórásnégyzetű torzítatlan** vagy **(abszolút) hatásos torzítatlan becslőfüggvényeknek** nevezzük.

Az aszimptotikusan torzítatlan becslőfüggvény fogalmához hasonlóan használjuk az **aszimptotikusan hatásos** becslőfüggvény elnevezést.

A $\hat{\Theta}_n$ statisztika aszimptotikusan hatásos, ha

$$\lim_{n\to\infty} Se(\hat{\Theta}_n) = 0 .$$

[10] A standard hiba angolul: standard error.

Bizonyos esetekben szükség lehet olyan becslőfüggvények hatásosságának összehasonlítására, amelyek közül legalább az egyik nem torzítatlan. Az **átlagos négyzetes hiba** (Mse[11)]) olyan mutatószám, amely a torzítást és a szórásnégyzetet is figyelembe veszi. Definícióját a (171) képlet tartalmazza.

$$Mse(\hat{\Theta}) = Bs^2(\hat{\Theta}) + Se^2(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2 \qquad (171)$$

Több torzított vagy legalább egy torzítatlan és több torzított becslőfüggvény közül azt tekintjük kedvezőbbnek, amelyiknek az átlagos négyzetes hibája kisebb.

Konzisztencia

Egy becslőfüggvényt **konzisztensnek** nevezünk, ha aszimptotikusan torzítatlan és aszimptotikusan hatásos.

(Megjegyzés: a szakirodalomban, a fenti definíció mellett, a konzisztenciának más tartalmú definíciói is léteznek.)

Például a sokasági várható értéknek a mintaátlag konzisztens becslőfüggvénye, hiszen:

$$Bs(\bar{x}) = \mu - E(\bar{x}) = 0 \qquad \text{és} \qquad \lim_{n\to\infty} Se(\bar{x}) = \lim_{n\to\infty} \frac{\sigma}{\sqrt{n}} = 0\,.$$

Robosztusság

Akkor mondjuk, hogy egy becslőfüggvény (illetve becslési eljárás) **robosztus**, ha az érzéketlen a kiinduló feltételekre. Ha a sokasági eloszlást nem ismerjük, akkor a becslésre robosztus becslőfüggvényt használunk. A robosztussággal, mint tulajdonsággal általánosságban nem foglalkozunk.

[11)] Az átlagos négyzetes hiba angolul: mean square error.

8.2. Pontbecslés

Ahogy azt már említettük, egy paraméter becslésére sokféle becslőfüggvény készíthető. Mi az eddigiekben az analógia elvét használtuk, amikor a sokasági várható értéket a mintaátlaggal becsültük. A továbbiakban olyan eljárásokat ismertetünk, amelyek segítségével becslőfüggvényeket készíthetünk.

A legkisebb négyzetek módszere (LNM)

Ezzel a módszerrel az első kötetben, a regressziószámítás tárgyalásakor már találkoztunk. A legkisebb négyzetek módszerét alkalmaztuk egy statisztikai modell paramétereinek meghatározására, becslésére. Az LNM mindig feltételezi egy modell létezését, vagyis azt, hogy egy jelenség leírása valamilyen összefüggés alapján lehetséges. Előnye, hogy a sokasági eloszlás ismerete nem kell az alkalmazásához.

Az LNM szerint úgy határozzuk meg a becsült paramétereket, hogy az ezeket használó modell alapján kapott értékek és a tényleges értékek eltéréseinek négyzetösszege minimális legyen.

62. példa

Határozzuk meg a sokasági várható érték becslőfüggvényét az LNM alapján!

Keressük tehát azt a $\hat{\mu}$ értéket, amelyre:

$$\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \to \min .$$

Deriválás után

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

adódik.

A maximum likelihood módszer (MLM)

A **maximum likelihood módszer** már feltételezi egy sokasági eloszlás ismeretét, és arra alkalmas, hogy annak valamely jellemzőjére becslőfüggvényt adjon. Alapgondolata az, hogy adott sokasági eloszlást feltételezve felírhatunk egy függvényt, amely az ismeretlen sokasági paraméter (vagy paraméterek) különböző lehetséges értékei mellett meghatározza annak valószínűségét, hogy éppen a rendelkezésünkre álló minta adódjon egy mintavétel eredményeképpen. Ezt a függvényt nevezzük **likelihood függvénynek.** Másképpen fogalmazva az MLM azt feltételezi, hogy egy esemény azért következik be, mert annak van a legnagyobb esélye a realizálódásra. Az MLM alapján a sokasági paramétert azzal az értékkel becsüljük, amelyik paraméterértékre a likelihood függvény felveszi maximumát, vagyis amelyik paraméter mellett a legnagyobb annak az esélye, hogy a megvalósult mintát kapjuk egy mintavétel alkalmával.

Ha (egy ismeretlen paramétert feltételezve) felírjuk a mintaelemek együttes bekövetkezésének valószínűségét, akkor a likelihood függvény a következőképpen adható meg:

$$L(x_1, x_2, ..., x_n, \Theta) = \prod_{i=1}^{n} f(x_i, \Theta).$$

Megjegyzés: *f* a feltételezett sokasági eloszlás sűrűségfüggvénye.

Az MLM segítségével konzisztens becslőfüggvényeket kapunk, és ha létezik minimális szórásnégyzetű torzítatlan becslőfüggvény, akkor a módszer ezt adja.

63. példa

Határozzuk meg a sokasági várható érték becslőfüggvényét az MLM alapján, normális eloszlású sokaságot feltételezve!

Írjuk fel a likelihood függvényt:

$$L(x_1, x_2, ..., x_n, \hat{\mu}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \hat{\mu}}{\sigma}\right)^2} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \hat{\mu}}{\sigma}\right)^2}.$$

A likelihood függvény helyett, a számítások egyszerűsítése érdekében, gyakran annak logaritmusát az ún. log-likelihood függvényt használjuk.

Ebben az esetben a log-likelihood maximumát keressük deriválással. Természetes alapú logaritmust véve:

$$\frac{d \ln L}{d \hat{\mu}} = \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$$

egyenlőséget kapjuk, innen becslőfüggvénynek $\hat{\mu} = \bar{x}$ adódik.

A momentumok módszere

A **momentumok módszerét** is ismert eloszlású sokaságok esetén tudjuk használni. Segítségével ismert eloszlástípus paramétereire adhatunk becslőfüggvényt. Olyan sokasági paraméterek becslésére alkalmas, amelyek momentumokkal felírhatóak. Lényege, hogy az elméleti momentumokat a mintából számított megfelelő empirikus momentumokkal tesszük egyenlővé, ami általában könnyen megoldható egyenletre vagy egyenletrendszerre vezet. Ez a módszer is konzisztens becslőfüggvényt eredményez, de erősen aszimmetrikus eloszlások esetén kevésbé hatékony.

64. példa

Határozzuk meg a normális eloszlású sokaság paramétereinek becslését a momentumok módszere alapján!

A normális eloszlásnak két paramétere van. Ezek felírhatóak momentumok segítségével:

$$\mu = M_1 \qquad \text{és} \qquad \sigma = \sqrt{M_2(\mu)}\,.$$

A minta első momentuma és második centrális momentuma:

$$m_1 = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{és} \qquad m_2(\bar{x}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}\,.$$

Innen:

$$\hat{\mu} = \bar{x} \qquad \text{és} \qquad \hat{\sigma}^2 = \mathrm{v}\,.$$

Megjegyzés: mint tudjuk, v csak aszimptotikusan torzítatlan becslése a sokasági szórásnégyzetnek, azaz nem torzítatlan a becslés:

$$E(\mathrm{v}) \neq \sigma^2.$$

Ezért az empirikus elemzéseknél nem v- vel, hanem s^2 -tel számolunk!

8.3. Intervallumbecslés

A pontbecslés során egyetlen olyan értéket határoztunk meg, amelyet valamilyen sokasági jellemző vagy statisztikai modell paramétere becslésének tekintettünk. Nem határoztuk meg, hogy mennyire megbízható a becslésünk, vagyis hogy hány százalék annak a valószínűsége, hogy a becsülni kívánt paraméter értéke a pontbecslés által adott számadattal lesz egyenlő. Ez egyébként nem is lehetséges, mert (folytonos esetben) egy valószínűségi változó egyetlen konkrét értéket 0% valószínűséggel vesz fel. A továbbiakban ezért egy intervallumot fogunk meghatározni, amelyről azt állíthatjuk, hogy előre adott nagy valószínűséggel tartalmazza a becsült paraméter tényleges értékét. Ezt az intervallumot **konfidencia intervallumnak** fogjuk nevezni, utalva arra, hogy bízhatunk abban, hogy a becslésünk helyes.
A konfidencia intervallum általános alakja az alábbi:

$$Pr\left(\hat{\Theta}_{a(\alpha)} < \Theta < \hat{\Theta}_{f(\alpha)} \right) = 1 - \alpha \,. \qquad (172)$$

A fenti egyenletben *Pr* az argumentum valószínűségének értékét jelöli. Olyan intervallumot akarunk meghatározni, amelyben a becsült sokasági jellemző $100 \cdot (1-\alpha)\%$ valószínűséggel található. Az intervallum alsó és felső határát ezért α értékét figyelembe véve kell meghatározni. Ezt az előre adott α értéket a becslésünk **megbízhatósági** vagy **konfidencia paraméterének** nevezzük. Ez általában 0-hoz közeli érték (pl. 0,01 azaz 1%), mert így $(1-\alpha)$ már 1-hez közeli, nagy valószínűség lesz.

8.4. Intervallumbecslés FAE minta esetén

Sokasági várható érték becslése

Normális eloszlású, ismert szórású sokaság esetén

Azt már tudjuk, hogy ha a sokaság normális eloszlású, akkor a minta is az. Sőt a mintaátlagok is normális eloszlásúak. Pontosabban:

$$X \sim N(\mu,\sigma^2) \;\Rightarrow\; \bar{x} \sim N(\mu,\frac{\sigma^2}{n}).$$

A szórás ismeretében elvégezhetjük a normális eloszlású mintaátlag standardizálását; a Z így standard normális eloszlású valószínűségi változó lesz.

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Ehhez az előző fejezetben leírtak szerint tudunk szimmetrikus intervallumot rendelni:

$$Pr\left(-z < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z\right) = 1-\alpha\,.$$

Feladatunk most nem az, hogy adott határok esetén keressünk valószínűséget, hanem éppen fordítva: adott valószínűség mellet keressük a megfelelő z értéket. A fenti egyenletet átrendezve:

$$Pr\left(\bar{x} - z_{(p)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{(p)}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha\,, \qquad (173)$$

ahol: $z_{(p)}$ az I. táblázat szerint az $(1-\alpha)$-hoz, míg a II. táblázat szerint az $(1-\frac{\alpha}{2})$-höz tartozó érték.

A $\Delta = \frac{\hat{\Theta}_{f(\alpha)} - \hat{\Theta}_{a(\alpha)}}{2}$ értéket **hibahatárnak** is szoktuk nevezni.

Ebben az esetben ez:

$$\Delta = z_{(\mathrm{p})} \frac{\sigma}{\sqrt{n}}. \tag{174}$$

A konfidencia intervallum a következőképpen is felírható:

$$\bar{x} \mp z_{(\mathrm{p})} \frac{\sigma}{\sqrt{n}} = \bar{x} \mp \Delta.$$

A mintavételi terv elkészítésénél lehetséges, hogy adott a hibahatár, vagyis, hogy milyen pontossággal akarjuk meghatározni a sokasági jellemzőt vagy paramétert. Ekkor a (175) képlet segítségével tudjuk megadni a szükséges mintanagyságot.

$$n = \frac{\left(z_{(\mathrm{p})}\sigma\right)^2}{\Delta^2} \tag{175}$$

Normális eloszlású, ismeretlen szórású sokaság esetén

A mintaátlagok ebben az esetben is normális eloszlásúak, de a standardizálás végrehajtásához a sokasági szórás nem áll rendelkezésre. A sokasági szórásnégyzetet a korrigált tapasztalati szórásnégyzet segítségével becsüljük, hiszen ez torzítatlan becslést ad. Bár a sokasági szórást a korrigált tapasztalati szórás nem becsüli torzítatlanul, mi mégis ezt fogjuk használni. A standardizált változónk a következő lesz:

$$T = \frac{\bar{x} - \mu}{s / \sqrt{n}}.$$

Ez nem normális eloszlású, hanem ***t*- (STUDENT-féle) eloszlású** változó $\nu = n - 1$ **szabadságfokkal**.

Megjegyzés: a statisztikában egy adott megfigyelési értékhalmaz szabadságfoka egyenlő a rendszeren belül szabadon (önkényesen) megválasztható értékek számával. Például az átlagnál $(n-1)$ adatot önkényesen választhatunk meg, de az n-edik elemet már nem, az már az előző adatok által meghatározott.

A normális eloszlású, ismeretlen szórású sokaság esetén a várható érték konfidencia intervalluma a (176) egyenlettel adott.

$$Pr\left(\bar{x} - t_{(p)}(\nu)\cdot\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(p)}(\nu)\cdot\frac{s}{\sqrt{n}}\right) = 1-\alpha\ , \qquad (176)$$

ahol: $t_{(p)}(\nu)$ a III. táblázat szerint az $(1-\alpha)$-hoz, míg a IV. táblázat szerint az $(1-\frac{\alpha}{2})$-höz tartozó érték.

A STUDENT-féle eloszlás vagy t-eloszlás

Ezt az eloszlástípust megalkotójáról W. S. GOSSETTről nevezték el, ő ugyanis STUDENT álnéven jelentette meg munkáit.

A STUDENT-féle eloszlás sűrűségfüggvénye a következő:

$$f(t) = \frac{Y_0}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}},$$

ahol Y_0 ν-től függő konstans, amelynek értékét úgy választjuk meg, hogy a sűrűségfüggvény görbe alatti területe 1 legyen.

A *t*-eloszlás sűrűségfüggvénye a 33. ábrán látható.[12)]

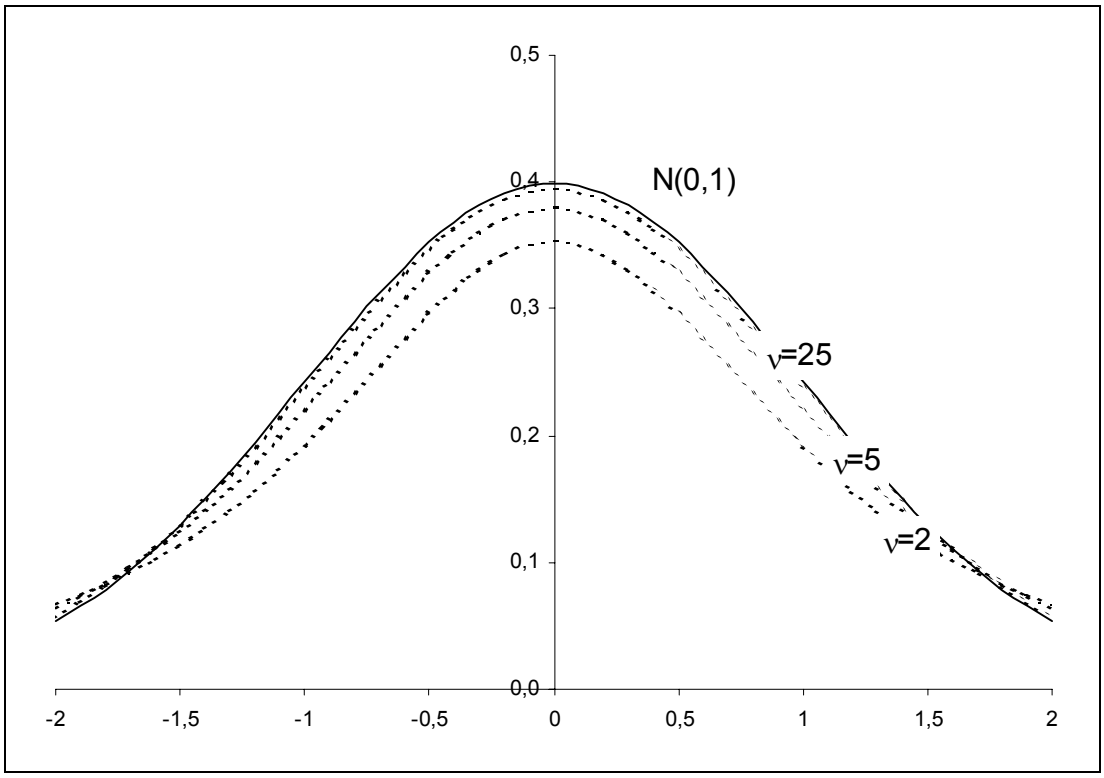
A *t*-eloszlás fontos tulajdonsága, hogy aszimptotikusan standard normális eloszlás, vagyis a szabadságfokát minden határon túl növelve közelít a standard normális eloszláshoz:

$$\lim_{\nu\to\infty} t_{(p)}(\nu) = z_{(p)}\ .$$

(Lásd a 33. ábrát.)

[12)] A fent közölt STUDENT-féle eloszlás számlálójában szereplő Y_0 érték meghatározása az Excel GAMMALN(*x*) függvény segítségével történt. (Ezt az eljárást nem részletezzük, mert nem része a tananyagnak!)
A statisztikában leggyakrabban alkalmazott eloszlásokról bővebben: [Denkinger, 1997], [Meszéna–Ziermann, 1981], [Spiegel,1995].

A *t*-eloszlás sűrűségfüggvényének grafikonja



33. ábra

A gyakorlatban $n \geq 30$ esetén a közelítés olyan mértékű, hogy ekkor már a standard normális eloszlás értékeivel számolunk.

A *t*-eloszláshoz tartozó értékeket a standard normális eloszláshoz hasonlóan táblázatok segítségével is meg tudjuk határozni. Erre a III. vagy a IV. táblázatot használhatjuk. A standard normális eloszlás táblázatával szemben ezek a táblázatok nem a *t* érték függvényében adják meg az eloszlásfüggyvény értékét, hanem a *t*-eloszlás kvantilis értékeit tartalmazzák.

Az Excelben a *t*-eloszlás kvantilis értékeit az INVERZ.T(*valószínűség*;*szabadságfok*) statisztikai függvény segítségével kaphatjuk meg. Itt a (176) szerinti konfidencia intervallum meghatározásához a *valószínűség* $= \alpha$ paraméterértéket kell megadnunk.

Szimmetrikus eloszlású, ismert szórású sokaság esetén

Nagy elemszámú minta esetén a központi határeloszlás tétele miatt a mintaátlag közelítőleg normális eloszlású lesz, így a standard normális eloszlással számolhatunk. A

kismintás esetben a konfidencia intervallum meghatározásához a valószínűségszámításból ismert **GAUSS-féle egyenlőtlenséget** alkalmazhatjuk. A mi jelölésrendszerünknek megfelelően:

$$Pr\left(\bar{x}-k\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+k\frac{\sigma}{\sqrt{n}}\right)\geq 1-\frac{4}{9k^2}=1-\alpha\,. \qquad (177)$$

Itt a *k* érték meghatározásához nem kell táblázatot használnunk. Annak értékét egyszerűen ki tudjuk számítani α segítségével:

$$k=\frac{2}{3}\cdot\frac{1}{\sqrt{\alpha}}=\frac{2\sqrt{\alpha}}{3\alpha}.$$

Ismeretlen eloszlású, ismert szórású sokaság esetén

A problémának ebben az esetben is csak kis minták alkalmazásakor van jelentősége, hiszen egyébként a normális eloszlás alkalmazható. Most is egy valószínűségszámításból ismert összefüggést alkalmazunk, a **CSEBISEV-egyenlőtlenséget**.

$$Pr\left(\bar{x}-k\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+k\frac{\sigma}{\sqrt{n}}\right)\geq 1-\frac{1}{k^2}=1-\alpha \qquad (178)$$

A *k* értéke ebben az esetben:

$$k=\frac{1}{\sqrt{\alpha}}=\frac{\sqrt{\alpha}}{\alpha}.$$

Sokasági értékösszeg becslése

A sokasági értékösszeg és a várható érték könnyen kapcsolatba hozható egymással, mert például diszkrét típusú változó esetén:

$$S=\sum_{i=1}^{N}X_i=N\cdot\bar{X}\,.$$

Egy valószínűségi változó konstanssal való szorzása esetén a változó eloszlástípusa

nem módosul,

$$E(N \cdot \bar{x}) = N \cdot E(\bar{x})$$

és

$$\sqrt{\text{var}(N \cdot \bar{x})} = N \cdot \sqrt{\text{var}(\bar{x})} ,$$

ha $\bar{x}$ valószínűségi változó és N konstans. Sokasági értékösszeg becslését ezért úgy végezzük, hogy először meghatározzuk a várható érték konfidencia intervallumát, majd a határokat megszorozzuk a sokaság nagyságával.

Sokasági arány becslése

Sokasági arány megállapítására alternatív ismérv esetén van lehetőség. Ekkor ismérvünknek két ismérvváltozata van, így BERNOULLI-féle valószínűségi változónak tekinthető. Ennek megfelelően végezzünk skálatranszformációt az ismérvértékeken és kódoljuk azokat 1 illetve 0 értékkel.

A sokasági arányt P-vel, míg a mintabeli arányt p-vel fogjuk jelölni.

A minta abszolút és relatív gyakorisági sorát az 57. táblázat tartalmazza.

Az alternatív ismérvek abszolút és relatív gyakorisági sora

57. táblázat

Ismérvváltozat (x)	Gyakoriság	Relatív gyakoriság
1	f_1	$p = \frac{f_1}{n}$
0	f_2	$q = 1 - p = \frac{f_2}{n}$
Összesen	n	1

Ezek alapján könnyen kiszámíthatjuk a minta átlagát

$$\bar{x} = \frac{f_1 \cdot 1 + f_2 \cdot 0}{n} = \frac{f_1}{n} = p .$$

A mintabeli arány tehát átlagként is értelmezhető.

Az (52) képlet alapján a minta szórásnégyzete:

$$\mathrm{v} = \frac{f_1 \cdot (1-p)^2 + f_2 \cdot (0-p)^2}{n} = \frac{1^2 \cdot np + 0^2 \cdot nq}{n} - p^2 = p(1-p) = pq\,.$$

(Megjegyzés: a 7. fejezethez hasonlóan, v ebben a fejezetben sem a relatív szórást jelöli!)

(154)-(156) szerint belátható, hogy

$$E(p) = P$$

és visszatevéses minta esetén

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{PQ}{n}}\,,$$

illetve visszatevés nélküli minta esetén

$$\sigma_p = \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}\,.$$

FAE minta esetén a standard hibát a következőképpen becsüljük:

$$s_p = \sqrt{\frac{pq}{n-1}}\,, \tag{179}$$

EV minta esetén pedig:

$$s_p = \sqrt{\frac{pq}{n-1} \cdot \frac{N-n}{N-1}}\,. \tag{180}$$

Visszatevéses minta esetén (vagy nagyon nagy alapsokaságból nem visszatevéses

mintánál) a p valószínűségi változó binomiális eloszlású[13)].

A binomiális eloszlás azonban közelíthető normális eloszlással, ha p és q nem 0-hoz közeli értékű és n elég nagy. Ezt a feltételt egzaktabban a következőképpen szokták megfogalmazni:

$$\min\{np, nq\} \geq 10 .$$

Ha tehát a fenti egyenlőtlenség fennáll, akkor a

$$Z = \frac{p - P}{\sqrt{\frac{pq}{n-1}}}$$

valószínűségi változót standard normális eloszlásúnak tekinthetjük. Ha diszkrét eloszlást közelítünk normális eloszlással, akkor használni szoktuk az ún. **folytonossági korrekciót** és a p arány helyett a $p \mp \frac{1}{2n}$ értéket használjuk, ennek azonban csak kis minták esetén van jelentősége.

Az elmondottak alapján a sokasági arány becslésére vonatkozó konfidencia intervallumot a (181) egyenlőség alapján tudjuk meghatározni.

$$Pr\left(p - z_{(p)} \cdot \sqrt{\frac{pq}{n-1}} < P < p + z_{(p)} \cdot \sqrt{\frac{pq}{n-1}} \right) = 1 - \alpha \qquad (181)$$

65. példa

Egy üzemben termoszokat gyártanak. A termékek minőségvizsgálata során egy 20 elemű (FAE) mintát vettek. Ellenőrizték, hogy a termoszok mennyi ideig tarják melegen a belejük helyezett adott hőmérsékletű vizet. A következő eredményeket kapták (órában):

7,8; 7,9; 8,8; 6,9; 7,5; 8,3; 8,4; 8,7; 7,8; 7,8;
8,1; 8,0; 8,2; 8,5; 7,6; 8,5; 8,6; 8,2; 8,1; 8,3.

[13)] Nem visszatevéses minta esetén a p valószínűségi változó hipergeometrikus eloszlású!

Készítsünk intervallumbecslést a hőtartás várható idejére 95,45%-os megbízhatósággal

1. ha előzetes felmérések alapján tudjuk, hogy a termoszok hőtartási ideje megközelítőleg normális eloszlású 0,4 óra szórással;
2. ha az eloszlás normális, de a szórás nem ismert;
3. ha az eloszlás típusa nem ismert csak a szórás, ami 0,4 óra;
4. ha az eloszlásról azt tudjuk, hogy szimmetrikus és a szórás 0,4 óra!
5. Határozzuk meg a 8,2 óránál kevesebb hőtartási jellemzővel rendelkező termoszok arányát (95,45%-os megbízhatósági szinten)!

1. A konfidencia intervallum nagyságának meghatározásához a (173) képletet használjuk. Becslőfüggvényünk a mintaátlag, ennek az adott mintán felvett értéke:

$$\bar{x} = 8{,}10 \text{ óra.}$$

A szükséges $z_{(p)}$ értéket az I. vagy a II. táblázat, illetve az Excel segítségével is megkaphatjuk. A hibahatár a (174) szerint behelyettesítés után:

$$\Delta = 2 \cdot \frac{0{,}40}{\sqrt{20}} = 0{,}18 \text{ óra.}$$

Ez alapján a konfidencia intervallum: $8{,}10 \mp 0{,}18$.

Azt mondhatjuk tehát, hogy az esetek átlagosan 95,45%-ban igaz, hogy a (7,92 óra; 8,28 óra) intervallumban található a termoszok tényleges hőtartási ideje.

2. Ekkor a (176) képletet alkalmazzuk. Mivel a sokasági szórás nem ismert, ezt a minta alapján becsüljük. A korrigált tapasztalati szórás:

$$s = 0{,}46 \text{ óra.}$$

A (176) képlethez szükséges pontos *t*-értéket az Excel segítségével tudjuk meghatározni INVERZ.T(1-0,9545;20-1) függvényhívással, azaz

$$t_{(p)}(19) = 2{,}1405.$$

Megjegyzés: a III., illetve a IV. táblázatból ezt a *t*-értéket pontosan nem tudjuk kiolvasni.

Így a konfidencia intervallum: $8{,}10 \mp 2{,}1405 \cdot \frac{0{,}46}{\sqrt{20}} = 8{,}10 \mp 0{,}22$.

3. Ebben az esetben robosztus becslést végzünk a (178) segítségével. Ehhez szükségünk van *k* meghatározására:

$$k = \frac{1}{\sqrt{1 - 0{,}9545}} = 4{,}69 .$$

Így a konfidencia intervallum: $8{,}10 \mp 4{,}69 \cdot \frac{0{,}40}{\sqrt{20}} = 8{,}10 \mp 0{,}42$.

4. Itt alkalmazhatjuk a (177) összefüggést.

$$k = \frac{2}{3} \cdot \frac{1}{\sqrt{1 - 0{,}.9545}} = 3{,}13$$

Így a konfidencia intervallum: $8{,}10 \mp 3{,}13 \cdot \frac{0{,}40}{\sqrt{20}} = 8{,}10 \mp 0{,}28$.

5. A minta alapján

$$p = \frac{10}{20} = 0{,}5000 \text{ vagy } 50{,}00\%;$$

$$s_p = \sqrt{\frac{0{,}5000 \cdot 0{,}5000}{20 - 1}} = 0{,}1147 \text{ vagy } 11{,}47\%.$$

Mivel $20 \cdot 0{,}5 \geq 10$, a sokasági arány becsléséhez a (181) képletet használhatjuk.

Így a konfidencia intervallum: $0{,}5000 \mp 2 \cdot 0{,}1147 = 0{,}5000 \mp 0{,}2294$.

Azt mondhatjuk tehát (95,45%-os megbízhatósági szint mellett), hogy a gyártott termoszok között azok aránya, amelyek 8,2 óránál kevesebb hőtartással rendelkeznek 27,06%–72,94% intervallumban található.

Megjegyzés: a kis elemszámú minta miatt (is) lett ilyen bizonytalan a becslésünk!

Sokasági szórásnégyzet becslése

Normális eloszlású sokaság esetén

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

valószínűségi változó $\nu = n-1$ szabadságfokú χ^2 eloszlást követ. Ez alapján a konfidencia intervallum:

$$Pr\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(\nu)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(\nu)} \right) = 1-\alpha \,. \qquad (182)$$

A χ^2 - eloszlás

A χ^2 ***(khi-négyzet)* - eloszlás** sűrűségfüggvénye a következő:

$$f(\chi^2) = Y_0 \cdot \left(\chi^2\right)^{\frac{\nu}{2}-1} \cdot e^{-\frac{1}{2}\chi^2},$$

ahol Y_0 ν-től függő konstans, amelynek értékét úgy választjuk meg, hogy a sűrűségfüggvény görbe alatti területe 1 legyen.

A χ^2-eloszlás sűrűségfüggvénye a 34. ábrán látható.[14)]

Ennek az eloszlásnak a gyakorisági görbéje baloldali aszimmetriát mutat a normális eloszlás gyakorisági görbéjéhez képest, ezért a (182) segítségével meghatározható konfidencia intervallum nem lesz szimmetrikus a pontbecslésre.

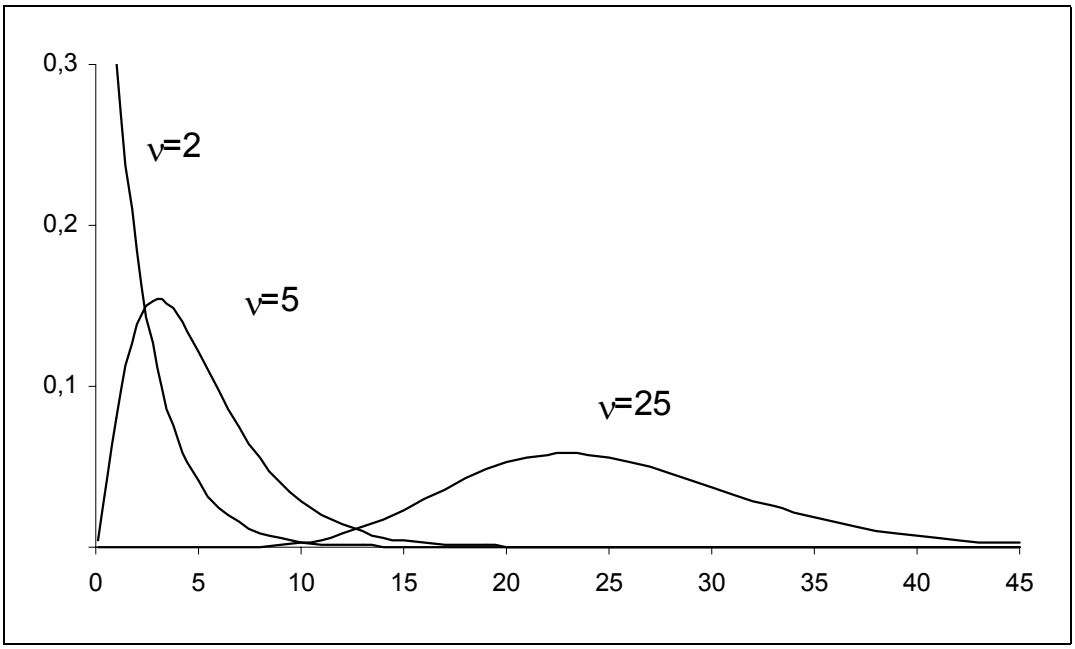
A χ^2 eloszlásfüggvényének értékeihez tartozó kvantiliseket az V. táblázat tartalmazza.

Az Excelben a χ^2-eloszlás (182) képletnek megfelelő kvantilis értékeit az INVERZ.KHI(*valószínűség*;*szabadságfok*) statisztikai függvény segítségével kaphatjuk

[14)] Lásd a [12)] lábjegyzetet!

meg. Itt a konfidencia intervallum meghatározásához a *valószínűség* $= \frac{\alpha}{2}$, illetve a *valószínűség* $= 1 - \frac{\alpha}{2}$ paraméterértéket kell megadnunk.

A χ^2-eloszlás sűrűségfüggvényének grafikonja



34. ábra

A χ^2 eloszlás aszimptotikusan normális eloszlás, vagyis a szabadságfokát minden határon túl növelve közelít a normális eloszláshoz. Ezért χ^2 táblázati értékét $n > 100$ esetén (adott α mellett) a következő összefüggések valamelyikével is megkaphatjuk:

$$\chi^2 \approx \nu\left(1 - \frac{2}{9\nu} + z\sqrt{\frac{2}{9\nu}}\right)^3,$$

illetve

$$\chi^2 \approx \frac{1}{2}\left(z + \sqrt{2\nu - 1}\right)^2,$$

ahol a z a standard normális eloszlású változó (α-nak) megfelelő táblázati értéke.

(Megjegyzés: a köbös összefüggés jelentősen pontosabb közelítést ad χ^2-re.)

66. példa

Egy mezőgazdasági Rt. 3000 hektáron búzatermesztéssel is foglalkozik. A termőterületükből véletlenszerűen (visszatevéses módszerrel) kiválasztott 300 db 1 hektáros terület alapján vizsgálták az átlaghozamot. Az adatokat az 58. táblázat tartalmazza. Becsüljük meg a 3000 hektár búzával bevetett terület átlaghozamának szórását 95%-os megbízhatósági szint mellett.

Az Rt 300 hektár búzával bevetett területének átlaghozamai

58. táblázat

Hozam (kg/ha)	Gyakoriság
– 2000	16
2001 – 4000	61
4001 – 6000	150
6001 – 8000	59
8001 –	14
Összesen	300

(Megjegyzés: az átlaghozamokat kilogrammos pontossággal mérték.)

Az 58. táblázat adatai alapján a mintaátlag $\bar{x} = 4960$ kg/ha; a korrigált tapasztalati szórás: $s = 1791$ kg/ha; az aszimmetria mérőszáma $\hat{\alpha}_3 = -0{,}017$; a csúcsosság mérőszáma pedig $\hat{\alpha}_4 = 3{,}103$. A minta mediánja $\hat{Me} = 4974$ kg/ha; a módusza $\hat{Mo} = 4990$ kg/ha.

A fenti adatok és a 3. fejezetben említett törvényszerűségek alapján, a búza átlaghozamának megközelítőleg normális eloszlása feltételezhető.

A konfidencia intervallum meghatározásához a (182) képletet használjuk. Az ehhez szükséges táblázati értékeket az Excel segítségével számíthatjuk ki:

$$\chi^2_{1-\frac{0,05}{2}}(300-1) = \text{INVERZ.KHI}(0,05/2;300\text{-}1) = 348,794$$

és

$$\chi^2_{\frac{0,05}{2}}(300-1) = \text{INVERZ.KHI}(1\text{-}0,05/2;300\text{-}1) = 252,993.$$

Megjegyzés: a statisztikai táblázatunkból ezeket az értékeket nem tudjuk kiolvasni, de Excel nélkül is meghatározhatjuk az említett két közelítő összefüggéssel. Például a köbös közelítő képlet alapján $\chi^2_{1-\frac{0,05}{2}}(300-1) = 348,797$; míg az egyszerűbb közelítő összefüggés szerint $\chi^2_{1-\frac{0,05}{2}}(300-1) = 348,311$.

A rendelkezésünkre álló adatok alapján a sokaság szórásnégyzetére (95%-os megbízhatósági szinten)

$$2\,479\,782 < \sigma^2 < 3\,791\,046$$

a szórására pedig

$$1\,658 < \sigma < 1\,947$$

becslést adhatjuk.

8.5. Intervallumbecslés EV minta esetén

Sokasági várható érték becslése

EV minta esetén a várható érték becslésének standard hibájánál figyelembe kell vennünk a sokaság elemszámát is.

$$\text{var}(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Az $\frac{N-n}{N-1}$ értéket **véges sokasági szorzónak** nevezzük.

Az EV mintából származó adatokra

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

valószínűségi változó standard normális eloszlású. Ezek alapján a sokasági várható értékre vonatkozó konfidencia intervallumot a (183) képlet alapján tudjuk meghatározni.

$$Pr\left(\bar{x} - z_{(p)} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} < \mu < \bar{x} + z_{(p)} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \right) = 1 - \alpha \qquad (183)$$

A véges sokasági szorzó értéke 0 és 1 között van, ezért EV minta esetén a hibahatár kisebb lesz, mint az FAE minta alkalmazásakor, tehát pontosabb becslést kapunk. Ennek az az oka, hogy az EV minta alapján történő becslés hatásosabb, mint az FAE minta alapján történő, hiszen ebben az esetben minden sokasági egység csak egyszer kerülhet a mintába.

Adott hibahatár esetén az EV mintához szükséges mintanagyságot a (175) helyett a (184) képlet segítségével határozhatjuk meg.

$$n = \frac{\left(z_{(p)}\sigma\right)^2}{\frac{\left(z_{(p)}\sigma\right)^2}{N} + \Delta^2} \qquad (184)$$

Ha a sokasági szórásnégyzet nem áll rendelkezésre, akkor ezt is a mintából kell becsülnünk. A 61. példában, a (170) képletnek megfelelően, már láttuk, hogy EV minta esetén

$$E\left(s^2 \cdot \frac{N-1}{N}\right) = \sigma^2,$$

illetve

$$E\left(s^2\right) = \frac{N \cdot \sigma^2}{N-1}.$$

Ebben az esetben az átlag standard hibájának becslése ($s_{\bar{x}}$) a (185) alapján történik.

$$s_{\bar{x}}^2 = \frac{s^2}{n} \cdot \left(1 - \frac{n}{N}\right) \tag{185}$$

Ez torzítatlan becslése a mintavételi szórásnégyzetnek:

$$E(s_{\bar{x}}^2) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \sigma_{\bar{x}}^2. \tag{186}$$

A (185) képlet négyzetgyöke:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}. \tag{187}$$

Sokasági értékösszeg becslése

Ebben az esetben is közvetlenül a sokasági várható érték becsléséből kaphatjuk meg a sokasági értékösszegre vonatkozó becslést, ha a konfidencia intervallum határait megszorozzuk a sokaság elemszámával, *N*-nel.

67. példa

Egy kistermelő 100 (azonos fajtájú) tehenet tart. Az egy tehénre jutó tejtermelés

meghatározása végett véletlenszerűen (ismétlés nélkül) kiválasztott 10-et, és a következő adatokat kapta (liter/év):

$$4512,\ 4923,\ 5810,\ 5167,\ 5216,\ 5342,\ 4985,\ 5098,\ 5156 \text{ és } 5512.$$

Határozza meg az egy tehénre jutó tejtermelés konfidencia intervallumát 95%-os megbízhatóság mellett, és a kistermelő által értékesíthető összes tejmennyiség intervallumát!

Mivel ismétlés nélküli a minta és a populáció szórása ismeretlen, a mintaátlagok standard hibájának kiszámításához a (187) képletet kell alkalmaznunk, ehhez pedig ismernünk kell a minta átlagát és korrigált tapasztalati szórását.

A kapott eredmények: $\bar{x} = 5172{,}1$; $s = 348{,}3$ és $s_{\bar{x}} = 104{,}5$ liter/év.

Figyelembe véve a (176) összefüggést, az egy tehénre jutó tejtermelés konfidencia intervalluma 95%-os megbízhatósági szinten (a III. táblázatot használva):

$$5172{,}1 \mp 2{,}2622 \cdot 104{,}5 = 5172{,}1 \mp 236{,}4 \text{ liter/év};$$

az egy év alatt (összesen) értékesíthető tej mennyisége pedig $4935{,}7 \cdot 100$ és $5408{,}5 \cdot 100$ liter között van.

Sokasági arány becslése

A sokasági arány EV mintán alapuló becslésekor a (180) szerint definiált standard hibát kell figyelembe venni.

Sokasági szórásnégyzet becslése

Ezzel az esettel könyvünkben részletesen nem foglalkozunk.

8.6. Intervallumbecslés R minta esetén

Sokasági várható érték és értékösszeg becslése

A rétegzett mintavétel esetén a viszonylag homogén sztrátumok mindegyikéből veszünk visszatevés nélküli (EV) mintát. A rétegek elemszámával súlyozott mintaátlag ebben az esetben is torzítatlanul becsüli a sokasági várható értéket.

A 7. fejezetben említettük, hogy rétegzett minta esetén több fajta elosztás is létezik. Ezek közül legtöbbször az arányos elosztást alkalmazzuk.

Arányos elosztás esetén az egyes sztrátumokból vett minták nagyságának aránya megegyezik a rétegek elemszámainak arányával. Ezért:

$$E(\bar{x}) = \mu ,$$

ahol (75) alapján

$$\frac{\sum_{j=1}^{M} n_j \bar{x}_j}{n} = \bar{x} .$$

Az átlag standard hibája:

$$\sigma_{\bar{x}} = \sqrt{\sum_{j=1}^{M} \frac{N_j^2}{N^2} \cdot \frac{\sigma_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j - 1}} , \qquad (188)$$

ahol σ_j^2 az alapsokaság j-edik rétegének szórásnégyzete.

Az empirikus elemzéseknél a véges sokasági szorzó értéke legtöbbször 1-hez közeli szám, ezért a továbbiakban ennek használatától eltekintünk.

Figyelembe véve a (163) összefüggést:

$$\frac{N_j^2}{N^2} = \frac{n_j^2}{n^2} .$$

Így a (188) képlet felírható a következő alakban is:

$$\sigma_{\bar{x}} = \sqrt{\sum_{j=1}^{M} \frac{n_j^2}{n^2} \cdot \frac{\sigma_j^2}{n_j}} .$$

A belső szórás (82) szerinti képlete alapján az átlag standard hibájára a (189) összefüggés adódik.

$$\sigma_{\bar{x}} = \frac{\sigma_B}{\sqrt{n}} \tag{189}$$

Az alapsokaság egyes rétegeinek szórásaira vonatkozóan általában nem rendelkezünk pontos információval, ezért helyettük a mintából (167) szerint kiszámított becsléseikkel dolgozunk. Ennek figyelembevételével felírható a (190) képlet.

$$s_{\bar{x}} = \frac{\sqrt{\sum_{j=1}^{M} n_j s_j^2}}{n} \tag{190}$$

Mivel minden rétegből vettünk mintát, a standard hiba csak a belső szórástól függ. Ez alapján megállapíthatjuk, hogy a rétegzett mintavétel akkor ad pontosabb becslést, vagyis akkor hatékonyabb a többi mintavételi módszernél, ha a sztrátumok megfelelően homogének, azaz a sokasági szórásnégyzet minél nagyobb részét a külső szórásnégyzet teszi ki. Ha a belső szórásnégyzet a sokasági szórásnégyzet nagyobb részét adja, akkor a rétegzett minta alkalmazása nem annyira hatékony, és ezért a sokaság (adott rétegképző ismérv szerinti) csoportosítása nem volt célszerű.

Ha a sokasági belső szórás nem ismert, akkor ezt a minta alapján a rétegek részszórásnégyzeteinek segítségével tudjuk becsülni. Mivel a gyakorlatban nagy mintákat használunk, a becsléshez használt statisztikánk standard normális eloszlásúnak tekinthető.

Az értékösszeg becslését ezúttal is a várható érték konfidencia intervallumának N konstanssal való szorzása révén tehetjük meg.

68. példa

A 66. példánál homogénnek tekintettük a sokaságot. Ha figyelembe vesszük azt a tényt, hogy nem azonos, hanem három fajta (megoszlásuk: 50% A, 20% B és 30% C típusú) búzával vetették be a 3000 hektárt, akkor milyen konfidencia intervallumot kapunk azonos megbízhatósági szint (95%) mellett, ha véletlenszerű kiválasztással és arányos elosztású rétegzett mintával dolgozunk?

A minta eredményeit az 59. táblázat tartalmazza.

Az Rt 300 hektáros (arányos elosztású) mintájának adatai

59. táblázat

Fajta	n_j	$\bar{x}_j$ (t/ha)	s_j (t/ha)
A	150	3,8	1,2
B	60	4,3	1,3
C	90	4,1	1,1

Figyelembe véve a (75) és (190) képleteket:

$$\bar{\bar{x}} = \frac{150 \cdot 3{,}8 + 60 \cdot 4{,}3 + 90 \cdot 4{,}1}{300} = 3{,}990 \text{ t/ha}$$

és

$$s_{\bar{x}} = \frac{\sqrt{150 \cdot 1{,}2^2 + 60 \cdot 1{,}3^2 + 90 \cdot 1{,}1^2}}{300} = 0{,}069 \text{ t/ha.}$$

Ezek alapján kiszámítható a keresett konfidencia intervallum: $3{,}990 \mp 1{,}96 \cdot 0{,}069 \approx 4{,}0 \mp 0{,}1$ t/ha.

Milyen konfidencia intervallumot kapnánk ha a 3000 ha búzával bevetett területből 300 hektárnyi FAE, illetve EV mintát vennénk?

9. Hipotézisek vizsgálata

9.1. Alapfogalmak

A gyakorlatban sokszor előfordul, hogy egy sokaság valamely paraméterére vonatkozóan van egy feltételezett érték, és csak azt szeretnénk eldönteni, hogy ez megfelel-e a valóságnak. Ha a sokaság teljes körű megfigyelésére nincs módunk, akkor a mintavétel módszeréhez folyamodhatunk. Ilyenkor egy véletlen minta alapján a fejezetben ismertetett módszerek szerint azt fogjuk megvizsgálni, hogy a mintánk támogatja-e a hipotézisünket, vagy szignifikánsan ellentmond neki. Így bizonyos megbízhatósággal állíthatjuk majd, hogy hipotézisünk igaz vagy hamis.

A felállított hipotézisek helyességének véletlen mintákra alapozott vizsgálatát **hipotézisvizsgálatnak** nevezzük. Az ennek során alkalmazott eljárások a **statisztikai próbák** vagy **tesztek**.

A hipotézisvizsgálat elemei

A hipotézisvizsgálat első fázisa a tesztelni kívánt feltételezés matematikai megfogalmazása. Ezt **nullhipotézisnek** nevezzük (jele: H_0). Az ezzel szemben álló feltételezés az **alternatív hipotézis** (jele: H_1). A fenti két állítás megfogalmazására egyszerre kerül sor, oly módon, hogy egymás komplementerei legyenek (a kettő közül pontosan egy igaz). A nullhipotézis helyessége egyúttal az alternatív hipotézis hamis voltát jelenti és fordítva.

Megkülönböztetünk **egyszerű** és **összetett** hipotéziseket. Egyszerű egy hipotézis, ha ebben azt feltételezzük, hogy az ismeretlen sokasági jellemző megegyezik egy adott értékkel. Például:

$$H: \Theta = \Theta_0 .$$

Az összetett hipotézisek esetében az ismeretlen sokasági jellemző értékére egy tartományt jelölünk ki. Például:

$$H: \Theta > \Theta_0 \quad \text{vagy} \quad H: \Theta \neq \Theta_0 .$$

A statisztikai próbák elvégzéséhez (a becslésekhez hasonlóan) mintaelemek egy függvényét használjuk. Olyan statisztikát konstruálunk, amelynek mintaelemeken felvett értéke alapján döntést tudunk hozni arra vonatkozóan, hogy a minta alátámasztja-e a nullhipotézisben megfogalmazott feltételezésünket.[15)] Ezt a függvényt **próbafüggvénynek** nevezzük. A próbafüggvény értéke is mintáról mintára változik, ezért a priori módon valószínűségi változónak tekinthető.

A próbafüggvénynek olyannak kell lennie, hogy valószínűségeloszlása egyértelműen meghatározható legyen a

- nullhipotézis helyességének feltételezése,
- a sokaságról rendelkezésre álló információk és
- a mintavétel módja alapján.

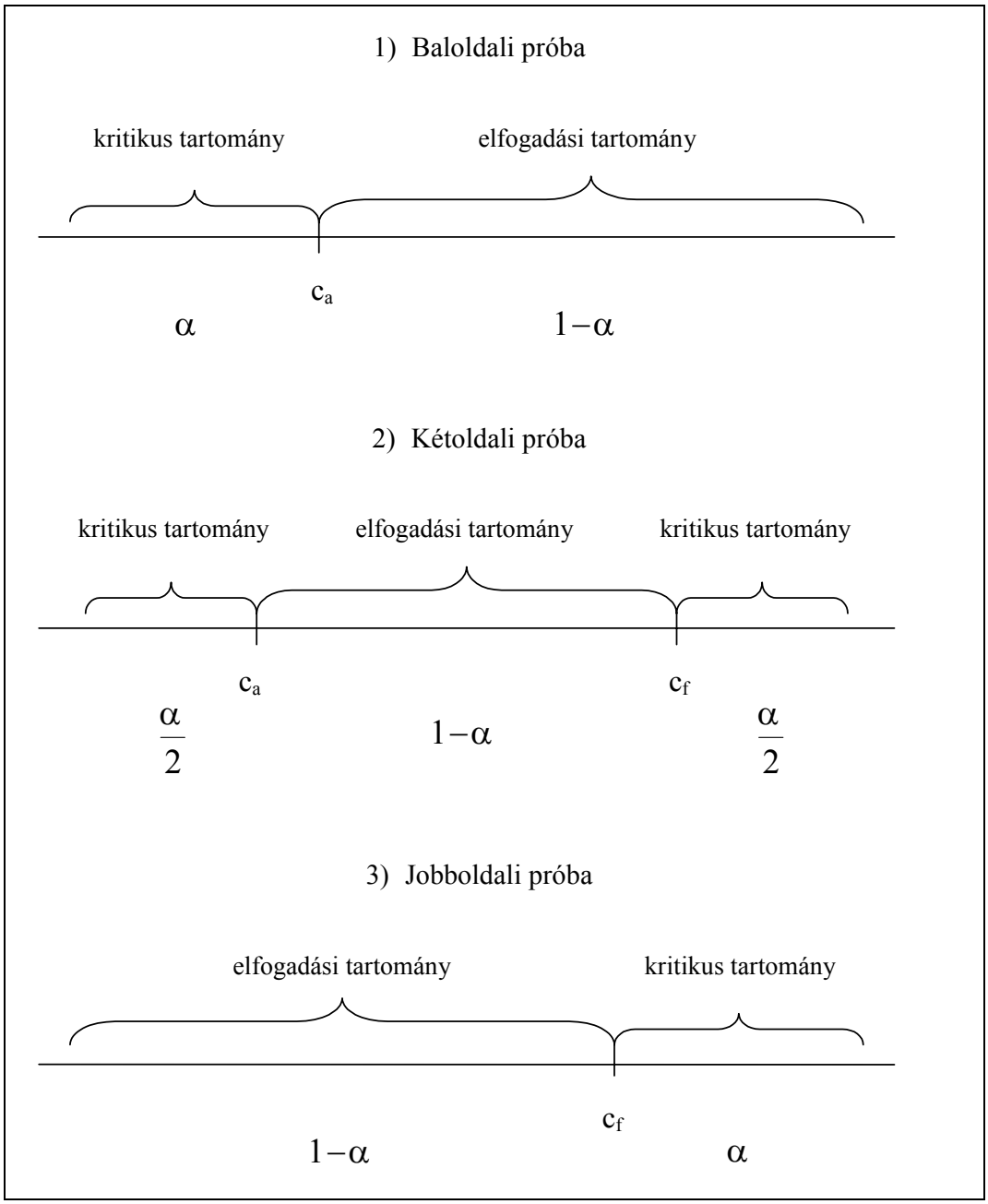
Azokat az információkat, kikötéseket, amelyek a próbafüggvény eloszlására hatással vannak, de a próba során helyességüket nem vizsgáljuk, a **próba alkalmazási feltételeinek** nevezzük.

A hipotézisvizsgálat során döntéseinket tehát a próbafüggvény mintán felvett értéke alapján hozzuk. Ehhez a próbafüggvény értékkészletét általában két – átfedésmentes és hézagmentes – tartományra bontjuk. Ezeket **elfogadási** illetve **kritikus (visszautasítási) tartománynak** nevezzük. A tartományok határait úgy határozzuk meg, hogy a nullhipotézis helyessége esetén a próbafüggvény értéke adott valószínűséggel az elfogadási tartományba essen. Ezt az előre adott valószínűséget **a próba megbízhatósági szintjének** nevezzük és $(1-\alpha)$-val jelöljük. Ekkor az ismeretlen sokasági paraméter mintából becsült értéke és a feltételezett érték eltérése a reprezentatív megfigyelés miatt fennálló véletlen mintavételi hibának tudható be. Ha a próbafüggvény értéke a kritikus tartományba esik, akkor azt mondhatjuk, hogy az ismeretlen sokasági jellemzőre vonatkozó feltételezésünk, valamint a minta alapján kapott becslésünk szignifikáns mértékben különbözik. Annak valószínűsége, hogy a nullhipotézis helyessége esetén a próbafüggvény értéke a kritikus tartományba essen α-val egyenlő. Ezt a valószínűséget nevezzük **szignifikancia-szintnek**.

[15)] Hipotézisek vizsgálatánál arra törekszünk, hogy a nullhipotézis egyszerű legyen, mert ekkor lehet legkönnyebben (a neki megfelelő) próbafüggvényt definiálni. Ha ez nem lehetséges, akkor ún. **technikai hipotézist** alkalmazunk. Könyvünkben ezek alkalmazásával nem foglalkozunk.

Az elfogadási és a kritikus tartomány egymáshoz viszonyított elhelyezkedése háromféle lehet. Ezeket az eseteket a 35. ábra szemlélteti.

Az elfogadási és a kritikus tartomány egymáshoz viszonyított elhelyezkedéseinek esetei



35. ábra

A **baloldali** és a **jobboldali próba** nem **kétoldali próba**, hanem ún. **egyoldali próba**.

Az eddigiek során már megismerkedtünk a fontosabb alapfogalmakkal, így fel tudjuk írni a hipotézisvizsgálat lépéseit.

1. A tesztelni kívánt, nullhipotézisnek nevezett, feltételezés megfogalmazása. Ezzel szemben mindig van egy alternatív hipotézis.
2. A nullhipotézist és a rendelkezésre álló információkat figyelembe véve a próbafüggvény kiválasztása.
3. A 0-hoz közeli α szignifikancia-szint kiválasztása, és a próbafüggvény értékkészletének elfogadási és kritikus tartományra bontása.
4. A próbafüggvény mintán felvett értékének megállapítása.
5. Döntés a nullhipotézis helyességének elfogadásáról-elvetéséről.

A hipotézisvizsgálat során elkövethető hibák

A hipotézisvizsgálat során is minta alapján következtetünk a sokaságra, így itt is számolnunk kell a reprezentatív megfigyelésből eredő véletlen mintavételi hibával. Ha a megfigyelésünk nem teljes körű, akkor teljes bizonyossággal nem tudunk dönteni a nullhipotézis helyességéről. Állásfoglalásunk kialakításakor alapvetően kétféle hibát követhetünk el:

- **elsőfajú hiba**: elvetjük a nullhipotézist, noha az megfelel a valóságnak,
- **másodfajú hiba**: elfogadjuk a nullhipotézist, noha az nem felel meg a valóságnak.

Az elsőfajú hiba elkövetésének valószínűsége a szignifikancia-szint definíciójából adódóan α. A másodfajú hiba elkövetésének valószínűségét β-val fogjuk jelölni.

A nullhipotézissel kapcsolatos döntésünk és a valóságban fennálló tényállás lehetséges eseteit és valószínűségüket a 60. táblázat tartalmazza.

Az elsőfajú hibával már érintőlegesen foglalkoztunk a szignifikancia-szint kapcsán, ám a másodfajú hiba nem került szóba a hipotézisvizsgálat lépéseinek tárgyalásánál. Ez azért van, mert a hipotézisvizsgálat alkalmazója csak az elsőfajú hiba nagyságát tudja befolyásolni (a szignifikancia-szint megadásával), de a másodfajú hibáét nem (ehhez tudnunk kellene, hogy mi felel meg a valóságnak).

Az elsőfajú hiba és a másodfajú hiba valószínűsége egymással ellentétesen alakul.

Általában úgy járunk el, hogy meghatározunk egy α szignifikancia-szintet és keressük azt a próbafüggvényt, amelyhez ekkor a legkisebb β tartozik adott mintanagyság mellett.

A hipotézisvizsgálat során elkövethető hibák és a helyes döntések valószínűségei

60. táblázat

H_0-t	H_0 megfelel a valóságnak	
	igaz	hamis
elfogadjuk	helyes döntés $(1-\alpha)$	másodfajú hiba (β)
elvetjük	elsőfajú hiba (α)	helyes döntés $(1-\beta)$

A könyvünkben bemutatott mintavételi tervek mindegyikét alkalmazhatnánk hipotézisvizsgálat céljából, de a továbbiakban mindig (a legegyszerűbb esetet) az FAE mintát feltételezzük.

Attól függően, hogy hány minta információi alapján történik a hipotézis tesztelése, könyvünkben megkülönböztetjük a következő eseteket:

- egymintás,
- két (egymástól független) mintás és
- több (egymástól független) mintás próba.

A hipotézisvizsgálatnál megkülönböztetünk **paraméteres** és **nemparaméteres** próbákat. Az előbbiek alkalmazási feltételei között szükségszerűen szerepelnek a vizsgált sokaság eloszlásának típusára vagy paramétereire vonatkozó feltételek, míg az utóbbiaknál ezekre nincs szükség.

A továbbiakban részletesebben bemutatjuk a gyakorlatban legtöbbször alkalmazott paraméteres próbákat.

9.2. Egymintás próbák

Az egymintás próbákat egy sokaság valamely jellemzőjére vagy paraméterére vonatkozó feltételezések helyességének vizsgálatára használjuk.

Sokasági várható értékre irányuló próba

Egy sokaság valamely jellemzőjének várható értékére vonatkozó nullhipotézishez háromféleképpen fogalmazhatunk meg alternatív hipotézist. Ezeket az eseteket tartalmazza a 61. táblázat.

Sokasági várható értékre irányuló próbák esetei

61. táblázat

Próba	Nullhipotézis	Alternatív hipotézis
baloldali	$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$
kétoldali		$H_1 : \mu \neq \mu_0$
jobboldali		$H_1 : \mu > \mu_0$

A sokasági várható értékre irányuló próba (a becsléshez hasonlóan) függ a sokaságra vonatkozó a priori információktól, kikötésektől. Ezeket neveztük a próba alkalmazási feltételeinek. Mi három esettel fogunk foglalkozni.

z-próba

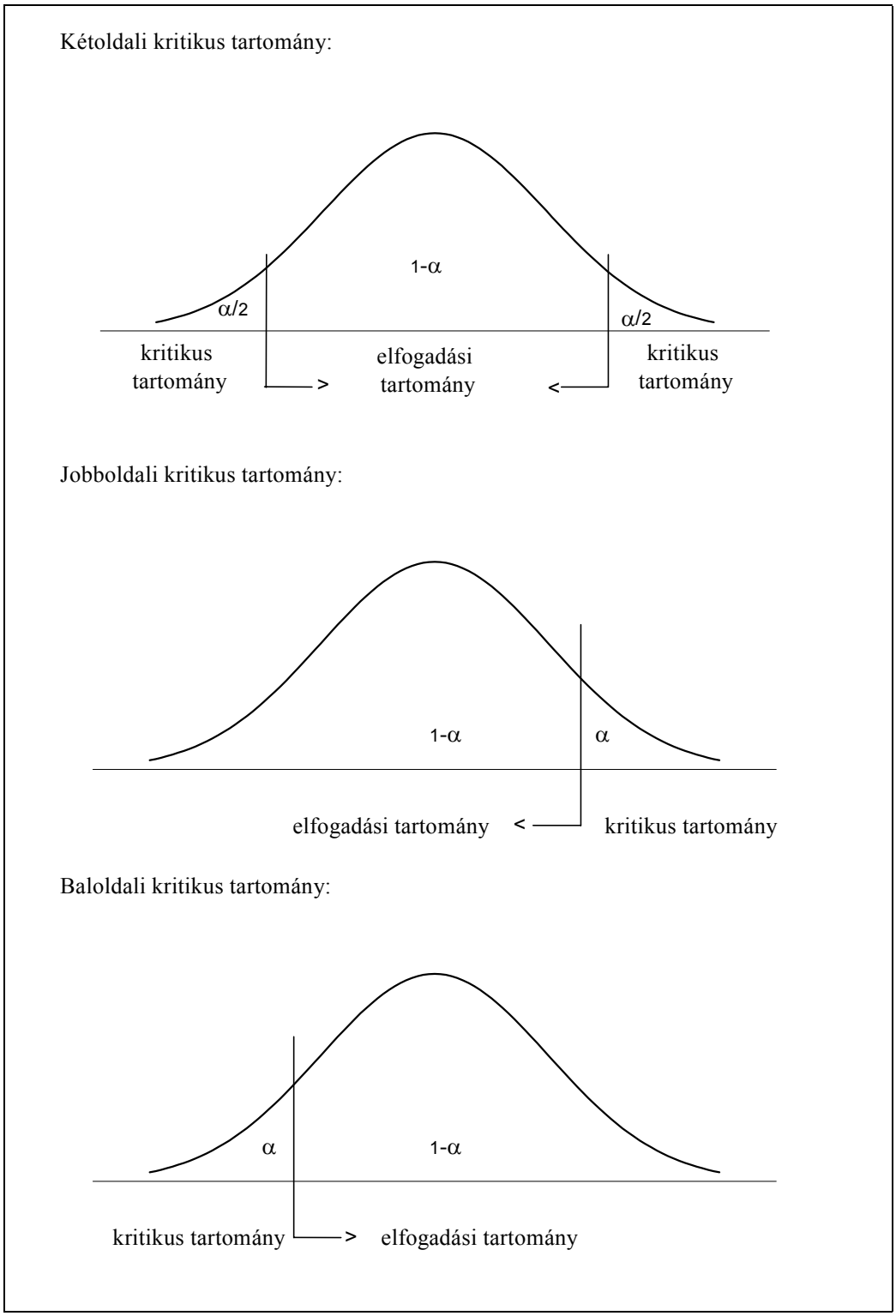
A **z-próba** alkalmazásának feltétele, hogy a mintánk ismert szórású (σ) normális eloszlású sokaságból származzon. Ebben az esetben a (191) szerint definiált próbafüggvényt használjuk.

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (191)$$

Ez a próbafüggvény standard normális eloszlású valószínűségi változó. Attól függően, hogy jobboldali, baloldali vagy kétoldali próbáról van-e szó, adott α szignifikancia-

szint mellett, a 36. ábrán szemléltetett módon tudjuk felosztani a próbafüggvény értékkészletét elfogadási és kritikus tartományra.

A döntéshozatal grafikus modellje



36. ábra

Ennek megfelelően, a II. táblázat szempontjából, a 62. táblázatban feltüntetett próbák és elfogadási tartományok adódhatnak. (Ezzel egyidejűleg adottak az alternatív hipotézisek és kritikus tartományok is.)

Várható értékre irányuló próbák és az ezekhez tartozó elfogadási tartományok ismert szórású normális eloszlású sokaság esetén

62. táblázat

Próba	Elfogadási tartomány
baloldali	$[-z_{1-\alpha}\ , \infty)$
kétoldali	$\left[-z_{1-\frac{\alpha}{2}}\ ,\ z_{1-\frac{\alpha}{2}}\right]$
jobboldali	$(-\infty\ ,\ z_{1-\alpha}]$

A kétoldali próba kritikus tartományának meghatározásához az I. táblázatot használhatjuk, míg az egyoldali próbákhoz a II. táblázatban egyszerűbb a megfelelő eloszlásfüggvény kvantilis értékének kikeresése. Mindhárom esetben használhatjuk természetesen az Excel INVERZ.STNORM(*valószínűség*) statisztikai függvényt is.

t-próba

A **t-próbát** akkor alkalmazhatjuk, ha a vizsgált sokaság (ismeretlen szórással) normális eloszlású. Ebben az esetben a (192) szerint definiált próbafüggvényt használjuk.

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}, \tag{192}$$

ahol s a mintából számított korrigált tapasztalati szórás. Ez a próbafüggvény $\nu = n - 1$ szabadságfokú STUDENT-féle eloszlást követ.

Ennek megfelelően, a IV. táblázat szempontjából, a 63. táblázatban közölt próbák és elfogadási tartományok adódhatnak.

Várható értékre irányuló próbák és az ezekhez tartozó elfogadási tartományok ismeretlen szórású normális eloszlású sokaság esetén

63. táblázat

Próbák	Elfogadási tartomány
baloldali	$[-t_{1-\alpha}(\nu), \infty)$
kétoldali	$\left[-t_{1-\frac{\alpha}{2}}(\nu),\ t_{1-\frac{\alpha}{2}}(\nu)\right]$
jobboldali	$(-\infty, t_{1-\alpha}(\nu)]$

A kétoldali próba kritikus tartományának meghatározásához legegyszerűbben az III. táblázatot használhatjuk, míg az egyoldali próbákhoz a IV. táblázatot. Mindhárom esetben itt is használhatjuk az Excel megfelelő statisztikai függvényét.

Aszimptotikus z-próba

Ha nagy minta áll rendelkezésünkre, akkor a sokasági jellemzőre tett egyéb ismeretek és feltételek nélkül[16)] is alkalmazhatjuk az **aszimptotikus z-próbát**, mert a (193) alapján definiált próbafüggvény (a központi határeloszlás tétele miatt) megközelítőleg standard normális eloszlású lesz.

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{193}$$

Ebben az esetben is a 62. táblázatnak megfelelő elfogadási tartományokat használjuk.

[16)] Véges szórás feltételezett ugyan, de ez az empirikus vizsgálatoknál teljesül is.

Sokasági arányra irányuló próba

Ennek vizsgálatát csak arra az esetre tárgyaljuk, amikor a minta olyan nagy, hogy $\mathrm{H}_0 : P = P_0$ nullhipotézis esetén eleget tesz az alábbi feltételnek:

$$\min\{nP_0, nQ_0\} \geq 10,$$

ahol $Q_0 = 1 - P_0$.

Ehhez hasonló feltétellel már a 8. fejezetben is találkoztunk a sokasági arány intervallumbecslésekor. A fenti feltételnek a teljesülése biztosítja számunkra, hogy a binomiális eloszlás helyett jó közelítéssel normális eloszlással dolgozzunk.

Sokasági arányra vonatkozó hipotézisek tesztelésére a (194) próbafüggvényt használjuk.

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \qquad (194)$$

Megjegyzés: itt is alkalmaznunk kellene a folytonossági korrekciót ($\mp \frac{1}{2n}$), de nagy minták esetén ennek értéke általában elhanyagolható, a döntést nem befolyásolja.

A sokasági arányra vonatkozó nullhipotézishez háromféleképpen fogalmazhatunk meg alternatív hipotézist. Ezeket az eseteket tartalmazza a 64. táblázat.

Sokasági arányra irányuló próbák esetei

64. táblázat

Próba	Nullhipotézis	Alternatív hipotézis
baloldali	$\mathrm{H}_0 : P = P_0$	$\mathrm{H}_1 : P < P_0$
kétoldali		$\mathrm{H}_1 : P \neq P_0$
jobboldali		$\mathrm{H}_1 : P > P_0$

Ezekhez a próbákhoz tartozó elfogadási tartományok (nagy minták esetén) megegyeznek a 62. táblázatban közöltekkel.

69. példa

Egy nagykereskedelmi vállalat 1 millió égőt vásárolt. A gyártó szerződésben vállalta, hogy a hibás égők részaránya 1%-nál nem lesz több. A vállalat ellenőrzés végett véletlenszerű kiválasztással ezer égőt vett a mintába, amelyben 12 hibás égőt találtak.

Elfogadható-e az a hipotézis (5%-os szignifikancia-szint mellett), hogy a szállítmányban a hibás égők részaránya nem több 1%-nál, azaz a gyártó eleget tett-e a szerződésben vállalat kötelezettségének?

A feladat szerint ismertek a következő adatok: $N = 10^6$; $n = 10^3$; $p = \frac{12}{1000} = 0{,}012$ vagy 1,2%; $\alpha = 0{,}05$.

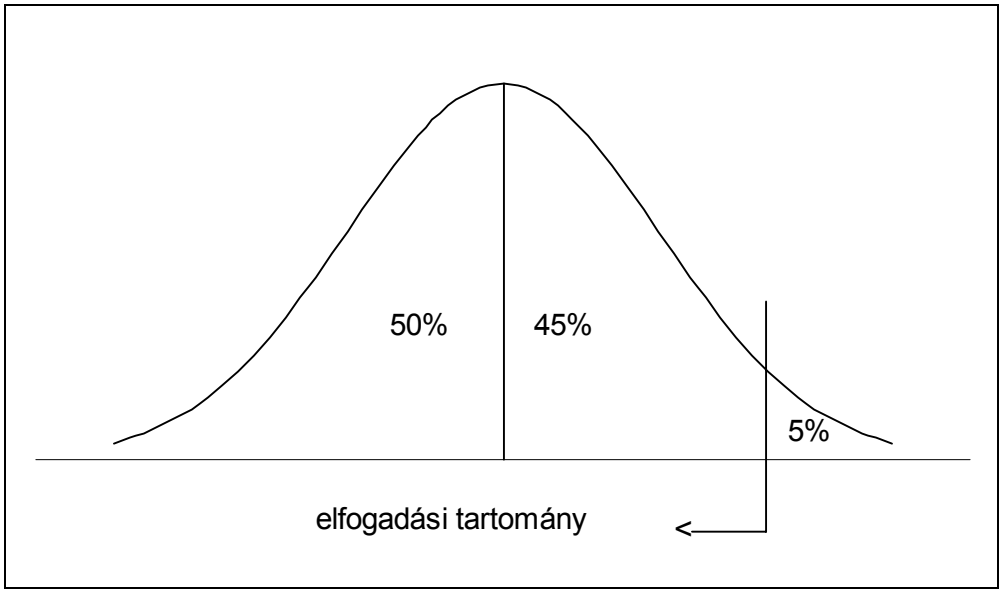
A feladatnak megfelelő nullhipotézis:

$$H_0 : P \le 0{,}01;$$

az alternatív hipotézis pedig:

$$H_1 : P > 0{,}01.$$

A feladatnak megfelelő grafikus modell



37. ábra

Mivel a minta nagysága az alapsokaság nagyságának csupán 1 ezreléke, és nagy mintáról van szó ($1000 \cdot 0{,}01 \geq 10$), azaz FAE mintát feltételezhetünk, a teszteléshez a (194) szerinti próbafüggvényt használhatjuk:

$$Z = \frac{0{,}012 - 0{,}010}{\sqrt{\frac{0{,}010 \cdot 0{,}990}{1000}}} = 0{,}64 .$$

A 37. ábra szerint jobboldali próbáról van szó, az ennek megfelelő elfogadási tartomány a 64. táblázat alapján: $(-\infty ; 1{,}65]$.

Mivel a kiszámított érték (0,64) az elfogadási tartományba esik, nullhipotézisünket 5%-os szignifikancia-szint mellett elfogadjuk, azaz a szerződés szerinti 1% és a minta alapján kiszámított 1,2% közötti különbség statisztikailag nem jelentős.

Függetlenségvizsgálat

Az eddigiek során olyan próbákkal foglalkoztunk, amelyek egy sokasági jellemzőre vonatkozó feltételezések ellenőrzését tették lehetővé. Most két sokasági jellemző között fennálló kapcsolatra vonatkozó hipotézisekkel foglalkozunk. A 4.2. fejezetben már tárgyaltuk azokat az eszközöket, amelyekkel a sokaság teljes körű ismerete esetén két ismérv kapcsolatát elemezhetjük. Ha azonban csak egy reprezentatív megfigyelés adatai állnak rendelkezésre, akkor a továbbiakban ismertetett módszert alkalmazzuk annak eldöntésére, hogy a vizsgált két ismérv függetlennek tekinthető-e.

Nullhipotézisünk: az adott sokaságon belüli két ismérv független egymástól, alternatív hipotézisünk: a két vizsgált ismérv között sztochasztikus vagy determinisztikus kapcsolat van.

Függetlenségvizsgálat χ^2-teszttel

A névleges mérési szintű adatok közötti kapcsolat vizsgálatánál már beszéltünk a χ^2 alapú mutatókról. Ott azt vizsgáltuk, hogy egy adott ($r \cdot c$ méretű) kombinációs tábla gyakoriságai mennyire különböznek egy (a két ismérv függetlensége esetén fennálló) gyakorisági eloszlástól.

Megjegyzés: a 4.2. fejezetben a χ^2 alapú mutatókat asszociációs kapcsolatoknál használtuk, de természetesen mennyiségi ismérveknél is alkalmazható, hiszen (osztályközöket képezve) ezeket is kombinációs táblába tudjuk rendezni.

A χ^2 statisztikát most mint próbafüggvényt alkalmazzuk. A függetlenségvizsgálat nullhipotézisét χ^2-teszt esetén az alábbi módon írhatjuk fel.

$$H_0 : Pr(C_{ij}) = P_{i.}P_{.j} \qquad i = 1,2,...,r\,; \qquad j = 1,2,...,c$$

$$H_1 : Pr(C_{ij}) \neq P_{i.}P_{.j} \qquad \textit{valamelyik i-re, illetve j-re}$$

A fenti megfogalmazás azt jelenti, hogy annak valószínűsége, hogy egy sokasági egység a kombinációs tábla (lásd a 4. táblázatot) adott C_{ij} cellájába esik, megegyezik a függetlenséget feltételezve kiszámított $P_{i.}P_{.j}$ valószínűséggel, ahol $P_{i.}$ és $P_{.j}$ a peremvalószínűségeket jelöli. Egy sokasági egység kombinációs tábla adott cellájába esésének valószínűségére pedig a minta relatív feltételes eloszlása (g_{ij}) alapján következtethetünk, ezért a (195) szerint definiált próbafüggvényt használjuk:

$$\chi^2 = n \cdot \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(g_{ij} - p_{i.} \cdot p_{.j}\right)^2}{p_{i.} \cdot p_{.j}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}, \qquad (195)$$

ahol *p*-k a *P* peremvalószínűségek mintából becsült értékei és $f_{ij}^* = n \cdot p_{i.} \cdot p_{.j}$.

A (195) szerint definiált statisztika χ^2-eloszlású valószínűségi változó, $\nu = (r-1)(c-1)$ szabadságfokkal, ha a kombinációs tábla $r \cdot c$ méretű.

Mivel a χ^2 mutató az eltérés mértékét számszerűsíti, a kis értékei megerősítik, míg nagy értékei cáfolják a nullhipotézist, tehát ezt minden esetben jobboldali próbaként hajtjuk végre. A χ^2-teszt alkalmazási feltételei között szerepel, hogy

legalább $n \cdot p_{i.} \cdot p_{.j} \geq 5$, de inkább $n \cdot p_{i.} \cdot p_{.j} \geq 10$ *minden i-re, illetve j-re*

fennálljon. Ezt az egyes osztályközök megfelelő kialakításával tudjuk biztosítani.

70. példa

A mérnök-munkanélküliek területi egységek (1999. június 30.) szerinti megoszlását a 65. táblázat tartalmazza.

Elfogadható-e az a hipotézis, hogy a munkanélküli mérnökök szakterületenkénti és lakóhelyük szerinti eloszlása között nincs szignifikáns összefüggés?

A mérnök-munkanélküliek megoszlása szakterületük és lakóhelyük szerint

65. táblázat

Szakterület	Lakóhely területi egységek szerint						
	KM	KD	NyD	DD	ÉM	ÉA	DA
1. Bánya-, kohó-, földmérnök	8	12	3	10	24	5	9
2. Gépészmérnök	66	43	32	26	53	72	71
3. Villamosmérnök	57	14	11	18	22	13	13
4. Építész-, építőmérnök	29	15	13	17	9	35	21
5. Mezőgazdasági, kertész-, faipari mérnök	59	39	64	76	87	127	98
6. Egyéb mérnöki végzettség	44	27	23	25	34	53	33

Forrás: OMK

Jelmagyarázat: KM: Közép-Magyarország, KD: Közép-Dunántúl, NyD: Nyugat-Dunántúl, DD: Dél-Dunántúl, ÉM: Észak-Magyarország, ÉA: Észak-Alföld, DA: Dél-Alföld.

A feladat megoldható a (195) képletben definiált próbafüggvénnyel. Ehhez szükségünk van a vizsgált két ismérv közötti kapcsolat függetlensége esetén fennálló elméleti eloszlásra, amelyet a 66. táblázat tartalmaz.

Ebben a táblázatban szereplő adatok eleget tesznek a χ^2-teszt alkalmazási feltételeinek, mert minden cellában 5-nél nem kisebb szám szerepel (ráadásul, 3 kivételével, még a szigorúbb feltételnek is megfelelnek, azaz 10-nél nem kisebbek az elméleti gyakoriságok).

Két ismérv függetlensége esetén fennálló eloszlás

66. táblázat

	KM	KD	NyD	DD	ÉM	ÉA	DA	Össz.
1	12	7	7	8	11	14	12	71
2	63	36	35	41	55	73	59	363
3	26	15	14	17	22	30	24	148
4	24	14	13	16	21	28	23	139
5	96	55	53	63	83	111	89	550
6	42	24	23	27	36	48	39	239
Össz.	263	150	146	172	229	305	245	1510

A 65. és 66. táblázat adatainak felhasználásával a (195) képlet jobboldala alapján a χ^2 próbafüggvény meghatározható:

$$\chi^2 = \frac{(8-12)^2}{12} + \frac{(12-7)^2}{7} + \ldots + \frac{(33-39)^2}{39} = 132{,}9\,.$$

A χ^2 elméleti értékét a $\nu = (6-1)(7-1) = 30$ szabadságfok figyelembevételével kell meghatároznunk. Ez az V. táblázatban minden szignifikancia-szinten alacsonyabb 132,9-nél; ezért azt mondhatjuk, hogy a minta nem támasztja alá a nullhipotézisünket, azaz a két ismérv nem tekinthető függetlennek.

Illeszkedésvizsgálat

Gyakran szükség van arra, hogy egy empirikus eloszlásnál megvizsgáljuk, hogy az megközelítően egyezik-e egy nevezetes eloszlással. Azt az egymintás próbát, amelynek során egy valószínűségi változó feltételezett eloszlására vonatkozó hipotézist tesztelünk **illeszkedésvizsgálatnak** nevezzük. Amennyiben a feltételezett eloszlás a normális eloszlás, akkor **normalitásvizsgálatról** beszélünk.

Ha a nullhipotézis meghatározza a feltételezett eloszlás minden paraméterét, akkor **tiszta illeszkedésvizsgálatról**, ellenkező esetben **becsléses illeszkedésvizsgálatról** van

szó. Az utóbbi esetben a feltételezett eloszlást leíró paramétereket ugyanis valamilyen pontbecsléssel határozzuk meg a minta alapján.

Nullhipotézisünk tehát az, hogy a minta egy adott elméleti eloszlásból származik. Ezt a következőképpen fogalmazhatjuk meg:

$$\mathrm{H}_0 : F_n(\mathrm{x}) = F_0(\mathrm{x}) .$$

Többféle próba létezik arra, hogy egy *n* elemű minta alapján teszteljük a hipotetikus $F_0(x)$ eloszlásfüggvényhez való illeszkedést.

Illeszkedésvizsgálat momentumok segítségével

Ahogy azt már láttuk a (néhány sokasági jellemzőre vonatkozó) hipotézisek tesztjeinél, a próba alkalmazási feltételei között gyakran szerepel az alapsokaság eloszlására tett kikötés. Természetesen ilyen esetben is illeszkedésvizsgálatot kell végeznünk. A 66. példában tulajdonképpen ezt tettük, amikor a minta momentumaiból következtettünk arra, hogy (az adott mezőgazdasági Rt-nél) a búza átlaghozama GAUSS-féle eloszlásúnak tekinthető-e. Ha a mintából becsült $\hat{\alpha}_3$ mutató 0 körüli, míg az $\hat{\alpha}_4$ mutató 3 körüli értéket vesz fel, akkor azt állíthatjuk, hogy a minta nem mond ellent az alapsokaság normalitására vonatkozó feltételezésnek.

Illeszkedésvizsgálat χ^2-teszttel

Az itt alkalmazott módszer lényegében megegyezik a függetlenségvizsgálatnál bemutatott χ^2-teszttel, de most két gyakorisági sor (lásd a 3. táblázatot) számpárosai közötti különbség statisztikai jelentőségét fogjuk vizsgálni. (A gyakorisági sor természetesen egy speciális kombinációs táblának is tekinthető.) Az illeszkedésvizsgálat nullhipotézisét χ^2-teszt estén az alábbi módon írhatjuk fel.

$$\mathrm{H}_0 : Pr(\mathrm{C_i}) = P_\mathrm{i} \qquad i = 1,2,...,k$$

$$\mathrm{H}_1 : Pr(\mathrm{C_i}) \neq P_\mathrm{i} \qquad \textit{valamelyik i-re}$$

A nullhipotézisunk tehát a következő: egy sokasági egység adott osztályközbe esésének hipotetikus és empirikus eloszlás szerinti valószínűsége megegyezik. Egy sokasági

egység adott osztályközbe esésének valószínűségére pedig a relatív gyakoriságok alapján következtethetünk, ezért a következő próbafüggvényt használhatjuk:

$$\chi^2 = n \cdot \left(\sum_{i=1}^{k} \frac{(g_i - P_i)^2}{P_i} \right) = \sum_{i=1}^{k} \frac{(f_i - f_i^*)^2}{f_i^*}, \qquad (196)$$

ahol $f_i^* = nP_i$.

A (196) szerint definiált statisztika χ^2-eloszlású valószínűségi változó, $\nu = k - 1 - b$ szabadságfokkal, ahol k a gyakorisági sor osztályközeinek száma, b pedig a mintából becsült paraméterek száma (tiszta illeszkedésvizsgálat esetén $b = 0$).

A függetlenségvizsgálat χ^2-tesztjéhez hasonlóan ez is jobboldali próba, és alkalmazási feltétele, hogy

$$\textit{legalább } nP_i \geq 5, \textit{ de inkább } nP_i \geq 10 \textit{ minden i-re}$$

fennálljon.

Megjegyzés: ha a fenti feltétel nem teljesül (ez leggyakrabban az első, illetve az utolsó osztályok valamelyikére igaz), akkor ezeket mindaddig összegezzük, amíg nem kapunk legalább 5-nél nagyobb f_i^* gyakoriságot. A szabadságfok meghatározásánál a k értékét ilyenkor az összevont osztályok figyelembevételével (és nem az eredeti osztályok száma alapján) határozzuk meg.

71. példa

Vizsgáljuk meg a 66. példa adatai alapján azt, hogy (a mezőgazdasági Rt-nél) a búza átlaghozama megközelítőleg normális eloszlásúnak tekinthető-e. Régebbi tapasztalatok alapján tudjuk, hogy az átlagtermés várható értéke 4950 kg/ha. Legyen a szignifikancia-szint 1%.

A normális eloszlásnak két paramétere van, de nekünk csak a várható érték adott. A szórásnégyzetet a mintából számított korrigált tapasztalati szórásnégyzet segítségével határozzuk meg. Ez alapján nullhipotézisünk az, hogy az átlaghozam (megközelítőleg)

normális eloszlást követ, 4950 [kg/ha] várható értékkel és (figyelembe véve a 66. példa részeredményét) $1791^2 = 3\,209\,097$ [kg^2/ ha^2] szórásnégyzettel.

Mivel csak a standard normális eloszlású valószínűségi változó eloszlásfüggvényének táblázati értékeivel rendelkezünk, ezért először az 58. táblázat adatait standardizáljuk. A transzformált változó értékeit a 67. táblázat tartalmazza.

Megjegyzés: a feladat szerint folytonos valószínűségi változó eloszlásáról van szó, ezért a standardizáláskor (a hézagmentesség biztosítása végett) a valódi (és nem a közölt) határok felső értékeit kell figyelembe venni.

A normalitásvizsgálathoz szükséges számítások

67. táblázat

Valódi osztályhatárok felső értékei	f_i	$\frac{X_{i,1}-4950}{1791}$	$\Phi\left(\frac{X_{i,1}-4950}{1791}\right)$	P_i	f_i^*
2000,5	16	-1,6471	0,0498	0,0498	14,9
4000,5	61	-0,5304	0,2979	0,2481	74,4
6000,5	150	0,5863	0,7212	0,4232	127,0
8000,5	59	1,7030	0,9557	0,2346	70,4
∞	14	∞	1,0000	0,0443	13,3
Összesen	300	–	–	1,0000	300,0

A P_i valószínűségeket az alábbi módon határoztuk meg:

$$P_i = \Phi\left(\frac{X_{i,1}-4950}{1791}\right) - \Phi\left(\frac{X_{i-1,1}-4950}{1791}\right).$$

A táblázat utolsó oszlopában szereplő elméleti gyakoriságok sorra mind nagyobbak 10-nél, ezért osztályközök összevonására nincs szükség.

A próbafüggvényünk értékét a 67. táblázat adatainak a (196) képlet jobboldalába helyettesítésével kaphatjuk meg.

$$\chi^2 = \frac{(16-14{,}9)^2}{14{,}9} + \frac{(61-74{,}4)^2}{74{,}4} + \ldots + \frac{(14-13{,}3)^2}{13{,}3} = 8{,}55\,.$$

A feladat szerint csak egy paramétert kellett becsülnünk a mintából ($b=1$) és az osztályközök száma $k=5$, így a χ^2 próbafüggvény szabadságfoka $\nu = 5-1-1=3$. Az 1%-os szignifikancia-szinthez tartozó elméleti érték az V. táblázat szerint 11,345.

Mivel $8{,}55 < 11{,}345$; a búza átlaghozamának normális eloszlására tett hipotézist 1%-os szignifikancia-szint mellett elfogadjuk.

9.3. Két független mintát igénylő próbák

Az előző fejezetben mindig egy sokaságból származó minta alapján következtettünk a sokaság valamely jellemzőjére. A továbbiakban azt vizsgáljuk, hogy két sokaság (azonos fajta) jellemzője eltér-e egymástól. A sokaságok összehasonlítására két mintát használunk, amelyek az egyes sokaságok reprezentatív megfigyeléséből származnak. A kétmintás vizsgálatok között megkülönböztetjük a **páros mintákat** és a független mintákat. Az előbbi esetben az egyik minta elemének kiválasztása maga után vonja a másik minta egy elemének kiválasztását. Ezek a minták ezért bizonyos értelemben egymintás próbának is tekinthetőek. Ezzel a speciális esettel azonban mi nem foglalkozunk.

A továbbiakban áttekintjük a két, egymástól függetlenül kiválasztott, mintán alapuló próbák legfontosabb eseteit. A két sokaság és a minták jellemzőire indexeléssel utalunk. Például a két sokaság várható értékét jelölje μ_1 és μ_2, a mintaátlagokat $\bar{x}_1$ és $\bar{x}_2$.

Várható értékek egyezőségére irányuló próbák

Két sokaság várható értéke egyenlőségére vonatkozó próbák nullhipotézisét és az alternatív hipotéziseit a 68. táblázatban feltüntetett módon fogalmazhatjuk meg.

Két sokaság várható értéke egyenlőségére irányuló próbák esetei

68. táblázat

Próba	Nullhipotézis	Alternatív hipotézis
baloldali	$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 < \mu_2$
kétoldali		$H_1 : \mu_1 \neq \mu_2$
jobboldali		$H_1 : \mu_1 > \mu_2$

Ezeknél a teszteknél is többféle próbafüggvényt használhatunk, attól függően, hogy melyik teszt alkalmazási feltételei állnak fenn. Most is három esettel fogunk foglalkozni.

Kétmintás z-próba

A **kétmintás z-próba** alkalmazásának feltétele, hogy mindkét mintánk ismert szórású normális eloszlású sokaságokból származzon. Ebben az esetben a (197) szerint definiált próbafüggvényt használjuk.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{197}$$

Ez a próbafüggvény standard normális eloszlású valószínűségi változó, így a próbát az eddigiekben ismertetett módon hajthatjuk végre.

A kétmintás *z*-próbát az Excelben is elvégezhetjük. Vigyük be az adatokat egy cellatartományba, majd hívjuk meg az **Eszközök** menü **Adatelemzés...** almenüjét és válasszuk ki a felkínált lehetőségek közül a **Kétmintás z-próba a várható értékre** menüpontot. A megjelenő párbeszédablakba bevihetjük a változótartományokat, a nullhipotézist, az ismert szórásnégyzeteket és a szignifikancia-szintet.

Kétmintás t-próba

A **kétmintás *t*-próbát** akkor alkalmazhatjuk, ha a két sokaság normális eloszlású és szórásaik ugyan ismeretlenek, de az feltételezhető, hogy egyformák (**homoszkedasztikus** sokaságok). Ekkor a (198) szerint definiált próbafüggvényt használjuk.

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{198}$$

ahol s_c a két sokaság egyforma szórásának a két minta alapján történő becslése. Ezt a minták adataiból többféleképpen is kiszámíthatjuk:

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} x_{1i}^2 - n_1\bar{x}_1^2 + \sum_{j=1}^{n_2} x_{2j}^2 - n_2\bar{x}_2^2}{n_1 + n_2 - 2}. \tag{199}$$

A (198) próbafüggvény $\nu = n_1 + n_2 - 2$ szabadságfokú STUDENT-féle eloszlást követ.

A homoszkedasztikus *t*-próba az Excelben az **Eszközök** menü **Adatelemzés...** almenüjében a **Kétmintás t-próba egyenlő szórásnégyzeteknél** menüponttal hívható meg.

Kétmintás aszimptotikus z-próba

Ha mindkét mintánk nagy, akkor a sokaságokra tett egyéb ismeretek és feltételek[17] nélkül is alkalmazhatjuk a kétmintás aszimptotikus *z*-próbát, mert a (200) alapján definiált próbafüggvény megközelítőleg standard normális eloszlású lesz.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (200)$$

A 68. táblázatban ismertetett próbákhoz tartozó elfogadási tartományok megegyeznek a 62., illetve 63. táblázatban közöltekkel.

72. példa

A 67. példában említett kistermelő újabb teheneket szeretne vásárolni. Egy kollégája másfajta teheneket tart. Annak eldöntésére, hogy az eddigi fajtából vásároljon-e vagy a kollégája által tartottakból, az utóbbi fajtából 8 elemű (ismétléses) mintát vettek. A mintában a tehenenkénti tejhozamok (liter/év) a következők:

5656, 4918, 5650, 5720, 4999, 5672, 5506, 5023.

Hogyan dönt a kistermelő 5%-os szignifikancia-szint mellett?

A feladat alapján felírható (lásd a 68. táblázatot) az alábbi két hipotézis.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

[17] A szórások végessége most is feltételezett.

Figyelembe véve azt a tényt, hogy kis mintákról van szó és a szórásnégyzetek is ismeretlenek, a kérdés megválaszolásához a (198) próbafüggvényt használhatjuk, amelynek egyik alkalmazási feltétele a szórásnégyzetek azonossága. Ennek ellenőrzése végett számítsuk ki a 8 elemű minta átlagát és korrigált tapasztalati szórásnégyzetét. (Emlékeztetőül megismételjük a 67. példa részeredményeit: $\bar{x} = 5172{,}1$ és $s = 348{,}3$.)

A rendelkezésünkre állnak a következő adatok:

$$n_1 = 10\,;\ \ \bar{x}_1 = 5172{,}1\,;\ \ s_1^2 = 121312{,}9\,;$$

$$n_2 = 8\,;\ \ \bar{x}_2 = 5393{,}0\,;\ \ s_2^2 = 121502{,}6\,.$$

Ezek szerint a próbafüggvény alkalmazásának említett feltétele biztosított, hiszen $s_1^2 \approx s_2^2$. (A tejhozamok megközelítőleg normális eloszlását feltételezzük.)

A (199) szerint:

$$s_c^2 = \frac{9 \cdot 121312{,}9 + 7 \cdot 121502{,}6}{10 + 8 - 2} = 121395{,}9\,.$$

A (198) szerint:

$$T = \frac{5172{,}1 - 5393{,}0}{\sqrt{121395{,}9} \cdot \sqrt{\frac{1}{10} + \frac{1}{8}}} = -1{,}3366\,.$$

A próbafüggvény empirikus és elméleti értékét a 68. és a 63. táblázatban közöltek szerint kell összehasonlítani.

A III. táblázatban a $\nu = 10 + 8 - 2 = 16$ szabadságfokhoz és $\alpha = 0{,}05$ szignifikancia-szinthez tartozó elméleti érték: 2,1199.

Mivel a próbafüggvény abszolút értéke (1,3366) kisebb a táblázati értéknél (2,1199), a nullhipotézist 5%-os szignifikancia-szint mellett elfogadjuk. Ez azt jelenti, hogy a két átlag közötti különbség (220,9 liter/év) statisztikailag nem jelentős (azaz a véletlennel magyarázható), ezért a tejhozam szempontjából nem indokolt a fajtaváltás.

A feladatot megoldhatjuk az Excel segítségével is az említett **Kétmintás t-próba egyenlő szórásnégyzeteknél** menüpont segítségével. A megfelelő adatok bevitele után kapott kimeneti eredményeket a 38. ábrán láthatjuk.

Az Excel outputja

Kétmintás t-próba egyenlő szórásnégyzeteknél

	Változó 1	*Változó 2*
Várható érték	5172,1	5393
Variancia	121311,8778	121502,571
Megfigyelések	10	8
Súlyozott variancia	121395,3063	
Feltételezett átlagos eltérés	0	
df	16	
t érték	-1,336606317	
P(T<=t) egyszélű	0,100024092	
t kritikus egyszélű	1,745884219	
P(T<=t) kétszélű	0,200048185	
t kritikus kétszélű	2,119904821	

38. ábra

Megjegyzés: az általunk közölt részeredményekben mutatkozó különbségek a kerekített adatainknak a következménye.

Sokasági arányok egyezőségére irányuló próba

Ennek vizsgálatát csak arra az esetre tárgyaljuk, amikor nagy minták állnak rendelkezésünkre, ekkor ugyanis a binomiális eloszlás helyett jó közelítéssel normális eloszlással dolgozhatunk.

Két sokasági arány egyenlőségére vonatkozó lehetséges nullhipotézist és az alternatív hipotéziseket a 69. táblázat tartalmazza.

Két sokasági arány egyenlőségére irányuló próbák esetei

69. táblázat

Próba	Nullhipotézis	Alternatív hipotézis
baloldali	$H_0 : P_1 = P_2$	$H_1 : P_1 < P_2$
kétoldali		$H_1 : P_1 \neq P_2$
jobboldali		$H_1 : P_1 > P_2$

A tesztelésére a (201) próbafüggvényt használjuk.

$$Z = \frac{p_1 - p_2}{\sqrt{\overline{p}\,\overline{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{201}$$

ahol

$$\overline{p} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2},$$

$$\overline{q} = \frac{q_1 n_1 + q_2 n_2}{n_1 + n_2}.$$

Természetesen $\overline{p} + \overline{q} = 1$.

A 69. táblázatban ismertetett próbákhoz tartozó elfogadási tartományok megegyeznek a 62. táblázatban közöltekkel.

9.4. Több független mintát igénylő próbák

Kettőnél több (M számú) sokaságból (külön-külön és egymástól függetlenül) vett minták alapján végezhető teszteket nevezzük többmintás próbáknak. Mi csak a várható értékek egyezőségére vonatkozó próbát tárgyaljuk.

Variancia-analízis

A **variancia-analízis** segítségével, nevével ellentétben, több (normális eloszlású és azonos szórásnégyzetű) sokaság várható értékének egyezősége tesztelhető. A nullhipotézisünket és az ehhez tartozó alternatív hipotézist az alábbiak szerint fogalmazhatjuk meg.

$$H_0 : \mu_j = \mu \qquad j = 1,2,...,M$$

$$H_1 : \mu_j \neq \mu \qquad \textit{valamelyik } j\textit{-re}$$

A fenti nullhipotézis helyességének ellenőrzésére a (202) szerint definiált próbafüggvényt használjuk.

$$F = \frac{SSK/(M-1)}{SSB/(n-M)} = \frac{s_K^2}{s_B^2}, \tag{202}$$

ahol M számú sokaságból M számú minta áll rendelkezésre, $n = \sum_{j=1}^{M} n_j$. Az SSK és az SSB a (77) képlet alapján értelmezett eltérés-négyzetösszegek.

A (202) próbafüggvény F eloszlást követ, a számláló szabadságfoka $\nu_1 = M - 1$ és a nevező szabadságfoka $\nu_2 = n - M$.

A variancia-analízis végrehajtását és eredményeit egy táblázatban szoktuk rögzíteni, amelyet leggyakrabban **ANOVA**[18] **táblázatnak** nevezünk. Ennek általános rendezési formáját a 70. táblázat tartalmazza.

[18] Analysis of Variance

Az ANOVA táblázat vázlata

70. táblázat

A szóródás oka	Eltérések négyzetösszege	Szabadságfok	Szórásnégyzet becslése	F
Tényező	*SSK*	$M-1$	s_K^2	$\frac{s_K^2}{s_B^2}$
Hiba	*SSB*	$n-M$	s_B^2	
Összesen	*SST*	$n-1$	–	

Az ANOVA táblázatban szereplő tapasztalati *F* értéket kell összevetnünk a megfelelő elméleti értékkel. Ez is jobboldali próba, tehát ha a tapasztalati *F* érték nagyobb az elméleti értéknél, akkor a várható értékek egyezőségére vonatkozó nullhipotézist (az adott szignifikancia-szint mellett) elutasítjuk és ezzel egyidejűleg a felállított alternatív hipotézist elfogadjuk.

A FISHER-féle F-eloszlás

Az ***F*-eloszlás** sűrűségfüggvénye a következő:

$$f(F)=\frac{Y_0 F^{(\nu_1/2)-1}}{(\nu_1 F+\nu_2)^{(\nu_1+\nu_2)/2}},$$

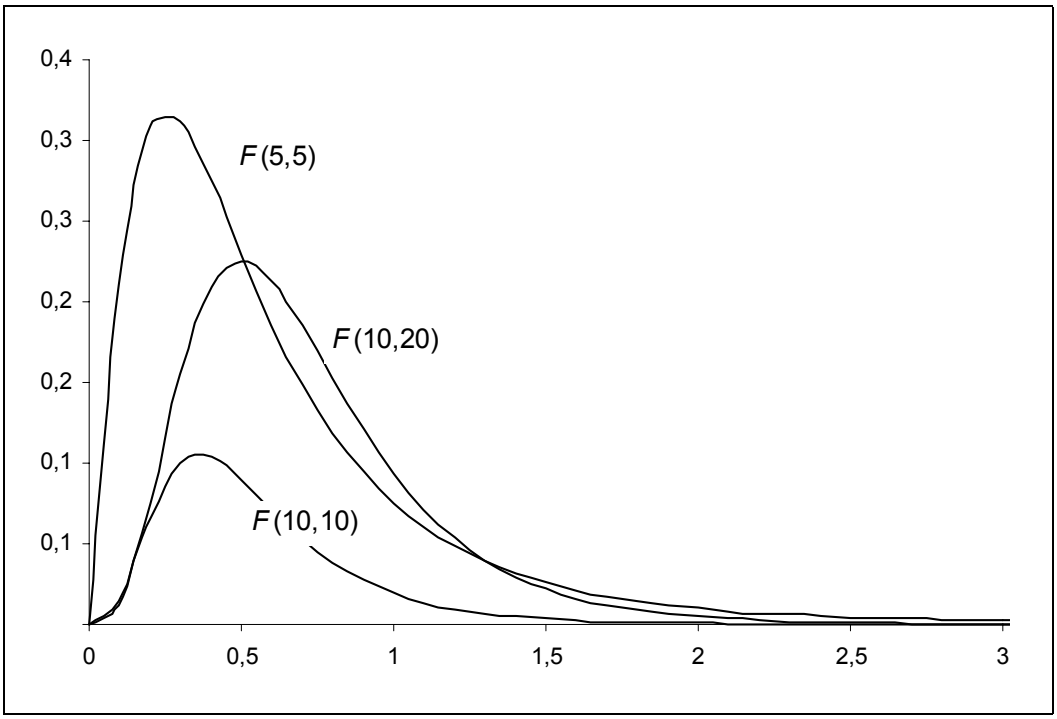
ahol Y_0 konstans a ν_1 és a ν_2 értékektől függ, amelyet úgy kell megválasztani, hogy a sűrűségfüggvény görbe alatti területe 1 legyen.

Az *F*-eloszlás sűrűségfüggvénye a 39. ábrán látható.[19)]

Az *F*-eloszláshoz tartozó értékeket a standard normális eloszláshoz hasonlóan táblázatok segítségével is meg tudjuk határozni. Erre a VI. vagy a VII. táblázatot használhatjuk.

[19)] Lásd a [12)] lábjegyzetet.

Az *F*-eloszlás sűrűségfüggvényének grafikonja



39. ábra

Az Excelben az *F*-eloszlás kvantilis értékeit az INVERZ.F(*valószínűség*; *szabadságfok1*;*szabadságfok2*) statisztikai függvény segítségével kaphatjuk meg. Itt a *valószínűség* $=\alpha$ paraméterértéket kell megadnunk a variancia-analízishez szükséges elméleti érték meghatározásához.

A *t*- eloszlás (IV. táblázat szerinti) értékeire és az *F*-eloszlás értékeire fennáll:

$$t^2_{1-\frac{\alpha}{2}}(\nu) = F_{1-\alpha}(1,\nu).$$

73. példa

Három kukoricafajta átlaghozamának összehasonlítása végett véletlenszerű kiválasztással (egymástól független) mintákat vettünk, és az alábbiakban ismertetett adatokhoz (t/ha) jutottunk.

Első fajta: 5,0; 5,1; 5,1; 5,3; 5,3; 5,3; 5,3; 5,4; 5,4; 5,4; 5,5; 5,5.
Második fajta: 5,2; 5,3; 5,4; 5,4; 5,5; 5,6; 5,6; 5,6; 5,7.
Harmadik fajta: 5,1; 5,2; 5,2; 5,2; 5,4; 5,4; 5,4; 5,6.

Az adatok alapján, 5%-os szignifikancia-szinten, elfogadhatjuk-e azt a hipotézist, hogy a három kukoricafajta átlaghozama megegyezik? (A hozamok megközelítőleg normális eloszlását feltételezzük.)

A feladatnak megfelelő nullhipotézis és alternatív hipotézis:

$$H_0 : \mu_j = \mu \qquad j = 1, 2, 3;$$

$$H_1 : \mu_j \neq \mu \qquad \textit{valamelyik } j\textit{-re.}$$

Az egyes fajtákra az alábbiakban feltüntetett mintajellemzőket számíthatjuk ki.

Első fajta: $n_1 = 12$; $\bar{x}_1 = 5{,}30$; $s_1^2 = 0{,}023$.

Második fajta: $n_2 = 9$; $\bar{x}_2 = 5{,}48$; $s_2^2 = 0{,}024$.

Harmadik fajta: $n_3 = 8$; $\bar{x}_3 = 5{,}31$; $s_3^2 = 0{,}024$.

Ezek alapján a variancia-analízis azonos szórásnégyzetekre vonatkozó feltételét az adataink kielégítik, így alkalmazhatjuk a (202) szerint definiált *F* próbafüggvényt.

Először határozzuk meg az eltérés-négyzetösszegeket a (77) összefüggésnek megfelelően.

$$SST = SSK + SSB$$

$$0{,}8403 = 0{,}2269 + 0{,}6134$$

Készítsük el az ANOVA táblázatot!

A kukoricahozamok ANOVA táblázata

71. táblázat

A szóródás oka	Eltérések négyzetösszege	Szabadságfok	Szórásnégyzet becslése	F
Fajta	0,2269	2	0,1135	4,809
Hiba	0,6134	26	0,0236	
Összesen	0,8403	28	–	

A kritikus érték 5 %-os szignifikancia-szinten és $\nu_1 = 2$, $\nu_2 = 26$ esetén a VI. táblázat

szerint (mint legközelebbi felhasználható érték) $F_{1-0,05}(2,25) = 3,385$. A pontos értéket az Excel megfelelő függvényének meghívásával kapjuk: INVERZ.F(0,05;2;26) = 3,369.

Mivel jobboldali próbáról van szó és a próbafüggvény aktuális értéke nagyobb a kritikus értéknél, a nullhipotézist elutasítjuk, tehát a minták 5%-os szignifikancia-szinten nem támasztják alá azt a feltételezést, hogy az egyes kukoricafajták átlaghozamai között nincs jelentős eltérés.

Megjegyzés: 1%-os szignifikancia-szinten, azaz az előbbinél kisebb elsőfajú hiba esetén, a nullhipotézist már elfogadnánk, mert az $F_{1-0,01}(2,26) = 5,526$ elméleti érték nagyobb a kiszámított $F = 4,809$ értéknél.

10. Dinamikus elemzés

Az eddigiek során leginkább egy vizsgált jelenség állapotával, illetve több jelenség közötti kapcsolat feltárásával foglalkoztunk. A jelenségek időbeli változásának nem tulajdonítottunk fontos szerepet, csupán a különböző időpontokban statikusan vizsgált jelenségek összehasonlítását végeztük. Ebben a fejezetben azonban minden jelenséget az idő függvényében vizsgálunk, megpróbáljuk leírni időbeli lefolyásukat. A dinamikus elemzéseknek három megközelítése ismert.

- **Sztochasztikus idősorelemzés**: azt feltételezi, hogy minden idősor alakulását saját korábbi állapota és a véletlen tényező befolyásolja. Az idősort sztochasztikus folyamatként fogja fel és rövid távú hatásait vizsgálja.
- **Spektrálanalízis**: Az idősorok adatait többfrekvenciás hullám eredőjeként fogja fel. Akkor használható, ha korlátlan számú kísérlet végezhető azonos feltételek mellett.
- **Determinisztikus idősorelemzés**: azt feltételezi, hogy az idősorokban hosszú távon érvényesülő törvényszerűségek, trendek vannak, amelyek matematikailag kezelhetőek.

Mi csak a legutóbbi megközelítéssel fogunk foglalkozni, de előbb tekintsük át az idősorok elemzésére szolgáló egyszerűbb módszereket.

10.1. Egyszerű elemzési módszerek

A dinamikus elemzések forrásai az idősorok. A 2.2. fejezetben már megismerkedtünk az idősor fogalmával és két fajtájával: az állapotidősorral (stock típusú) és a tartamidősorral (flow típusú). A 2.3. fejezetben részletesebben tárgyaltuk a dinamikus viszonyszámokat, amelyeket azonos sokaság két (időben különböző) adatának összehasonlításával kaptunk. A 2.4. fejezetben pedig az idősorok ábrázolásával is foglalkoztunk.

Idősor adatainak átlaga

Az idősorok egyszerű jellemzésére szolgál, ha egy nagyobb időintervallumban meghatározzuk az abban megfigyelt értékek átlagát. Ezt az átlagot, mint időtartamhoz tartozó adatot, az időszak közepéhez igazítjuk. Ennek megfelelően különböző módon

átlagoljuk a stock és a flow típusú idősorok adatait. Tartamidősor esetén számtani átlagot használunk:

$$\bar{x} = \frac{\sum_{t=1}^{n} x_t}{n},$$

ahol x_t a *t*-edik időszakhoz tartozó megfigyelt érték, *n* a megfigyelések száma.

Megjegyzés: a fenti képlet **ekvidisztáns** (azonos hosszúságú) időszakok megfigyeléseit feltételezi. Ha a megfigyelések időben nem egyenlő távolságra esnek, akkor súlyozott képletet kell alkalmaznunk. A továbbiakban azonban az idősorok ekvidisztáns jellegét mindig feltételezzük.

Állapotidősor esetén az idősor átlaga is állományi adat kell hogy legyen, ezért először meg kell határoznunk a megfigyelt időpontok közötti időszakokra eső átlagos állományokat, majd ezeket kell átlagolnunk. Ezt a (203) szerint számított mutatót **kronologikus átlagnak** nevezzük.

$$\bar{x}_k = \frac{\frac{x_1 + x_2}{2} + \frac{x_2 + x_3}{2} + \ldots + \frac{x_{n-1} + x_n}{2}}{n-1} = \frac{\frac{x_1}{2} + \sum_{t=2}^{n-1} x_t + \frac{x_n}{2}}{n-1} \qquad (203)$$

74. példa

Egy kft forgalmi és létszámadatait a 72. táblázat tartalmazza.

A kft fontosabb adatai

72. táblázat

Év	Forgalom (millió Ft)	Létszám az év elején
1994	56	460
1995	60	590
1996	80	720
1997	102	990
1998	140	1350

Számítsuk ki a kft átlagos forgalmát az adott időszakban és a foglalkoztatottak évi átlagos nagyságát, ha tudjuk hogy a kft 1999 elején 1340 főt foglalkoztatott!

A forgalomra vonatkozó idősor flow típusú, azaz a 72. táblázat első adatsora tartamidősor. Az átlagos forgalmat ezért a következőképpen tudjuk kiszámítani:

$$\bar{x} = \frac{56 + ... + 140}{5} = 87{,}6\,.$$

A létszám idősora azonban stock típusú, ezért itt a kronologikus átlagot használjuk:

$$\bar{x}_k = \frac{\frac{460}{2} + 590 + ... + 1350 + \frac{1340}{2}}{6-1} = 910\,.$$

Ezek alapján a kft-nek 1994. január 1. és 1998. december 31. között évente átlagosan 87,6 millió Ft forgalma volt; és e közben évente átlagosan 910 főt foglalkoztatott.

A változás intenzitásának egyszerű mutatószámai

Ha az egyik időpontról (vagy időszakról) a másikra történő változások nagysága a vizsgált időintervallumban bizonyos állandóságot mutat, tehát a szomszédos időpontok (vagy időszakok) adatainak különbsége nagyjából egyenlő, akkor a változás intenzitását jól jellemzi a (204) szerint definiált **növekedés átlagos mértéke**.

$$\bar{d} = \frac{\sum_{t=2}^{n}(x_t - x_{t-1})}{n-1} = \frac{x_n - x_1}{n-1} \qquad (204)$$

Ha a szomszédos időpontokhoz (vagy időszakokhoz) tartozó adatok hányadosai tekinthetőek állandónak, akkor a vizsgált időintervallumban a változás intenzitását a **növekedés átlagos üteme** jellemzi jól. Ezt (35) szerint definiáljuk:

$$\bar{l} = \sqrt[n-1]{\prod_{t=2}^{n}\frac{x_t}{x_{t-1}}} = \sqrt[n-1]{\frac{x_n}{x_1}}\,.$$

A fenti két mutató az idősornak csak az első és utolsó adatára támaszkodik, ezért csak akkor alkalmazható, ha az idősorban (abszolút vagy relatív módón) egyenletesen érvényesülő növekvő vagy csökkenő tendencia figyelhető meg.

Az idősorok összetevői

A determinisztikus idősorelemzés leggyakrabban alkalmazott modellje a **dekompozíciós idősormodell**. Ez azt feltételezi, hogy az idősorok alakulását négy fő összetevő befolyásolja.

- A legfontosabb összetevő a hosszabb időszakon át tartósan meglevő tendenciát (átlagos mozgásirányt) kifejező **trend**. Ez az alapirányzat, amelyet a vizsgált jelenségre ható alapvető gazdasági, társadalmi tényezők alakítanak ki.

- Az idősorok vizsgálatakor gyakran figyelhető meg szabályos ingadozás (a trendhez képest), amely rendszeresen ismétlődő hullámzást jelent. Ezt az összetevőt nevezzük **szezonális komponensnek**. A szezonalitás általában egy éven belül jelentkezik, természeti tényezőkkel, társadalmi szokásokkal magyarázható. Ez megfigyelhető például a mezőgazdaságban, az idegenforgalomban, a házasságkötések számának alakulásában, stb.

- A hosszabb idősorok vizsgálatánál megfigyelhetőek olyan periodikus ingadozások, amelyek nem olyan szabályosak és hosszúságuk több év. Ezek alkotják a **ciklikus komponenst**. Ilyenek például a gazdaságban kimutatható konjunktúrális ciklusok (lásd például KONDRATYEV-féle ciklus, sertésciklus).

- Az eddigi összetevőkkel nem magyarázható szabálytalan ingadozásokat a **véletlen tényezőnek** tulajdonítjuk. Ez okozza a megfigyelt értékeknek a trend, illetve a periodikus összetevők által meghatározott idősor görbéje körüli sztochasztikus ingadozását. Ezt a komponenst valószínűségi változónak tekinthetjük, éppúgy mint az idősor adatait, hiszen ezek sok, egyenként számba nem vehető tényező alakulásának függvényei.

A fentiekből következik, hogy egy idősor bármelyik tagja az említett tényezőknek a függvénye, ezért a továbbiakban nem x –szel jelöljük, hanem (utalva a függőségére) y-nal.

Arra vonatkozóan, hogy a fent ismertetett négy összetevő hogyan kapcsolódik egymáshoz, a statisztikai irodalomban alapvetően kétféle modell ismeretes. Az **additív**

modell szerint az összetevők összege adja azok eredőjét, míg a **multiplikatív modell** szerint az idősor a komponensek szorzataként képződik.

A továbbiakban szimbólumok segítségével fogjuk felírni e két modellt.

Additív modell:

$$y = T^a + S^a + C^a + \varepsilon .$$

Multiplikatív modell:

$$y = T^m \cdot S^m \cdot C^m \cdot \eta .$$

A két egyenletben T a trend, S a szezonális, C a ciklikus komponenst, míg ε és η a véletlen tényezőt jelöli.

Az additív modell esetén elvárjuk, hogy a szezonális komponensek összege 0 legyen, hiszen szabályos amplitúdót feltételeztünk. A véletlen tényező várható értékét szintén 0-nak feltételezzük. Multiplikatív modell esetén ezek logaritmusairól mondhatjuk el ugyanezt.

A dekompozíciós idősormodellek esetében célunk az, hogy ezeket az összetevőket elkülönítsük és számszerűsítsük. Mi a továbbiakban az alaptendenciát leíró trenddel és a szezonális komponenssel foglalkozunk részletesebben, míg a ciklikus tényező vizsgálatát nem tárgyaljuk.

Az általunk használt additív modell legyen:

$$y_{ij} = T_{ij}^a + S_j^a + e_{ij},$$

a multiplikatív modell pedig:

$$y_{ij} = T_{ij}^m \cdot S_j^m \cdot u_{ij},$$

ahol i=1,2,...,$\frac{n}{p}$ a periódusok sorszáma, j=1,2,..., p pedig a perióduson belüli időszak sorszáma.

10.2. Mozgó átlagok módszere

A trendszámítás az alaptendencia meghatározását, az idősor „kisimítását” jelenti. Célja a múltban megfigyelt átlagos mozgásirány jövőbe való kivetítése, amit **extrapolációnak** nevezünk, ellentétben az **interpolációval**, ami a vizsgált időszakra vonatkozó visszatekintést jelenti.

Megjegyzés: az idősorok empirikus elemzésénél extrapoláláskor abból a feltételezésből indulunk ki, hogy a vizsgált jelenség múltbeli átlagos mozgásiránya a jövőben is fennmarad. Ezért nem ajánlatos trendek segítségével túl távoli időintervallumokra következtetni.

A trendszámításnak két fő módszere ismeretes: a mozgó átlagok módszere és az analitikus trendszámítás.

Mozgó átlagok módszere

A **mozgó átlagok módszere** alkalmazásakor a trendet az idősor dinamikus átlagolásával határozzuk meg úgy, hogy az idősor minden eleméhez kiszámítjuk annak (valamekkora) környezetében levő elemek átlagát.

A mozgó átlagok módszerét mi csak additív modellt feltételezve tárgyaljuk és ekkor számtani átlagformát alkalmazunk. A multiplikatív modell esetén a módszer hasonlóan hajtható végre, csak mértani átlagokat kell használnunk.

A mozgó átlagok módszere azon alapszik, hogy additív esetben a szezonális tényező várható értéke 0 minden periódusban, ezért ha a periódus hosszának megfelelően választjuk meg annak a környezetnek a nagyságát, amelyben levő elemeket átlagoljuk, akkor megközelítőleg a trendértékekhez jutunk (amennyiben a trend megközelítőleg lineáris). Az átlagolással kiküszöböljük a szezonális komponenst és csökkentjük a véletlen tényező szerepét.

Fontos tehát a **mozgó átlagolás tagszámának**, vagyis az átlagolandó adatok számának a helyes meghatározása. Amennyiben ez nem egyenlő a periodikus komponens hullámhosszának egész számú többszörösével, akkor a szezonális összetevőt nem

tudjuk kiküszöbölni, és esetleg az eredeti idősornál is nagyobb hullámzást mutató trendet kapunk.

A simítás némiképpen különbözik, ha a szezonális komponens periódusának hullámhossza páratlan és páros. A páratlan tagszámú mozgó átlagolással kisimított trendet a (206) képlet segítségével kaphatjuk meg.

$$\hat{y}_t = \frac{y_{t-k} + y_{t-k+1} + \ldots + y_t + \ldots + y_{t+k}}{2k+1}, \tag{205}$$

ahol $2k+1$ a szezonális komponens periódusának hullámhossza.

Ha a periódus páros számú megfigyelésből áll, akkor a mozgó átlag nem rendelhető egész sorszámú időponthoz vagy időszakhoz. Például 4 tagú mozgóátlagokat számítva az idősor első 4 adatának átlaga a második és a harmadik megfigyelés „közötti időponthoz" tartozik, hiszen az e körüli környezetben levő adatokat átlagoltuk. Ilyenkor a kiszámított adatokat még középre kell igazítani. Ezt az utóbbi eljárást nevezzük **centrírozásnak**. Ennek során a mozgó átlagolással kapott idősoron újra elvégezzük a módszert kéttagú mozgó átlagokat alkalmazva. A centrírozás után kapott idősort közvetlenül az eredeti adatokból a következőképpen írhatjuk fel:

$$\hat{y}_t = \frac{\frac{y_{t-k}}{2} + y_{t-k+1} + \ldots + y_t + \ldots + y_{t+k-1} + \frac{y_{t+k}}{2}}{2k}. \tag{206}$$

A fenti képletek alkalmazásával a mozgó átlagolású trendet csak a $k+1 \le t \le n-k$ sorszámú adatokra tudjuk meghatározni, ezért az idősor elején és végén k számú időponthoz vagy időszakhoz nem számítható trendérték. Ezt nevezzük a trend mozgó átlagolásból adódó **rövidülésének**.

Megjegyzés: az előzőekben ismertetett módszer megközelítőleg lineáris alapirányzat esetén alkalmas a trendértékek elkülönítésére. Nemlineáris esetben más módszert kell alkalmazni (pl. **SPENCER-féle súlyozott mozgó átlagok**).

75. példa

Az élelmiszerek fogyasztói árindexeit (havi bontásban) 1995 és 1998 között a 73. táblázat tartalmazza.

Készítsük el az idősor mozgó átlagolású kisimítását!

A 40. ábra alapján megállapíthatjuk, hogy az élelmiszerek havi fogyasztói árindexeinek idősorában évenkénti periodicitás figyelhető meg, ezért a kisimításhoz 12 (vagy ennek egész számú többszöröse) tagszámú mozgó átlagolást használhatunk. Páros tagszám esetén alkalmaznunk kell a középre igazítást is. Az eredményeket a 74. táblázat tartalmazza.

Élelmiszerek fogyasztói árindexe 1995-1998 között

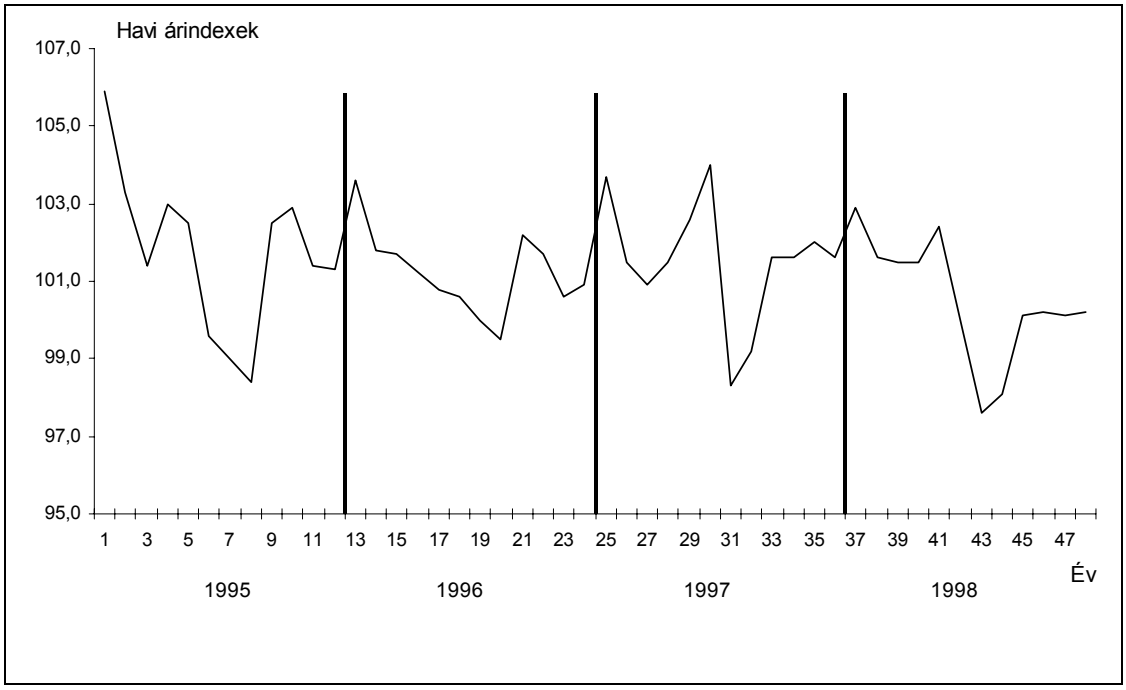
73. táblázat

Hónap	1995	1996	1997	1998
Január	105,9	103,6	103,7	102,9
Február	103,3	101,8	101,5	101,6
Március	101,4	101,7	100,9	101,5
Április	103,0	101,2	101,5	101,5
Május	102,5	100,8	102,6	102,4
Június	99,6	100,6	104,0	99,9
Július	99,0	100,0	98,3	97,6
Augusztus	98,4	99,5	99,2	98,1
Szeptember	102,5	102,2	101,6	100,1
Október	102,9	101,7	101,6	100,2
November	101,4	100,6	102	100,1
December	101,3	100,9	101,6	100,2

Forrás: Fogyasztói Árindex Füzetek, KSH, Bp., 1997-1999.

Először ábrázoljuk az adatokat vonaldiagram segítségével. (Lásd a 40. ábrát.)

Az élelmiszerek fogyasztói árindexének alakulása 1995-1998 között



40. ábra

A 74. táblázat elkészítésénél használhatjuk az Excelt is. Hívjuk meg az **Eszközök** menü **Adatelemzés...** almenüjét és válasszuk ki a felkínált lehetőségek közül a **Mozgóátlag** menüpontot. Az ekkor megjelenő párbeszédpanel segítségével adjuk meg a Bemeneti tartományt. Az Intervallum mezőbe kell beírnunk a mozgó átlagok tagszámát. A Diagramkimenet jelölőnégyzetet bekapcsolva grafikus ábrát is kaphatunk.

Az említett opciókon kívül az Excel még más lehetőségeket is felkínál, de ezekkel mi nem foglalkozunk.

Megjegyzés: az Excel által használt eljárás nem alkalmazza a (206) szerinti centrírozást! Ezt nekünk kell utólag elvégezni.

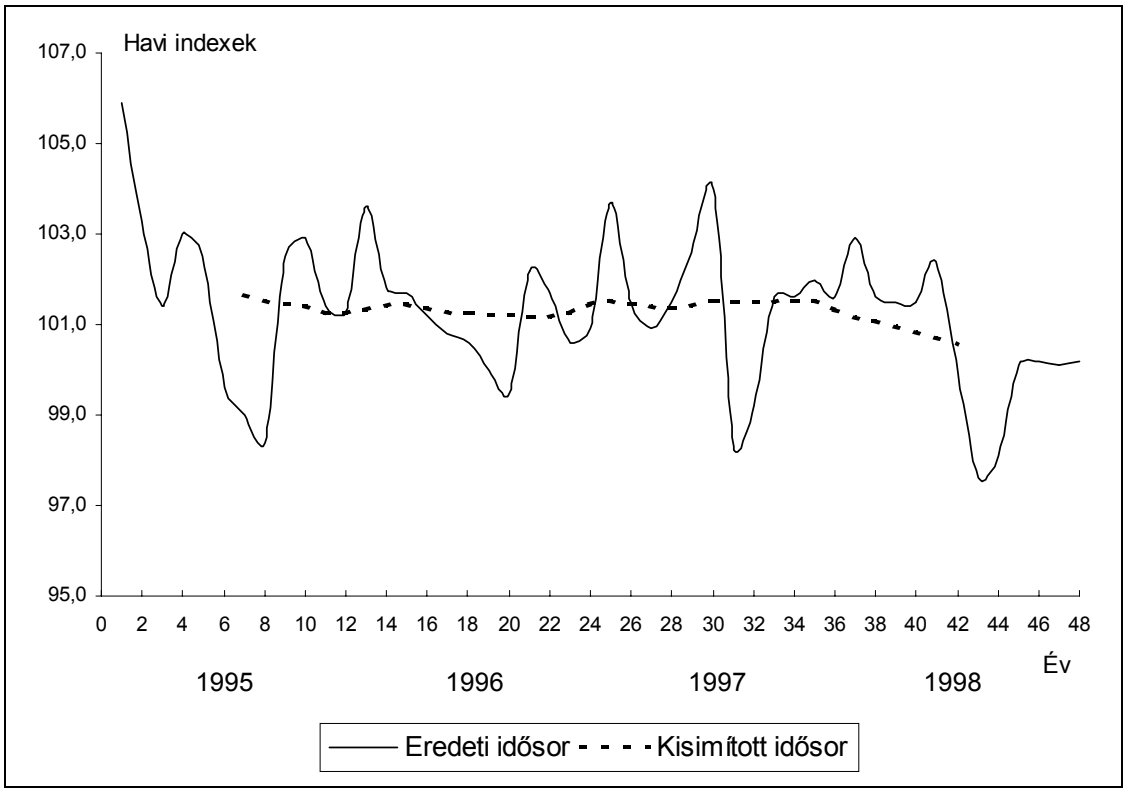
Élelmiszerek fogyasztói árindexeinek mozgó átlagolással kisimított idősora

74. táblázat

Év	Hónap	Árindex	Mozgó átlag	Centrírozás
1995	Január	105,9		–
	Február	103,3		–
	Március	101,4		–
	Április	103,0		–
	Május	102,5		–
	Június	99,6	101,77	–
	Július	99,0	101,58	101,67
	Augusztus	98,4	101,45	101,51
	Szeptember	102,5	101,48	101,46
	Október	102,9	101,33	101,40
	November	101,4	101,18	101,25
	December	101,3	101,27	101,23
1996	Január	103,6	101,35	101,31
	Február	101,8	101,44	101,40
⋮	⋮	⋮	⋮	⋮
1998	Január	102,9	101,18	101,15
	Február	101,6	101,12	101,07
	Március	101,5	101,03	100,96
	Április	101,5	100,90	100,84
	Május	102,4	100,78	100,70
	Június	99,9	100,63	100,57
	Július	97,6	100,51	–
	Augusztus	98,1		–
	Szeptember	100,1		–
	Október	100,2		–
	November	100,1		–
	December	100,2		–

Az eredeti és a kisimított idősort a 41. ábrán láthatjuk.

Élelmiszerek havi fogyasztói árindexeinek mozgó átlagolással kisimított idősora



41. ábra

A következő fejezetben egy másik (nagyon gyakran alkalmazott) eljárást ismertetünk, amely segítségével az idősor alapirányzata szintén számszerűsíthető.

10.3. Analitikus trendszámítás

Az **analitikus trendszámítás** során a vizsgált jelenség alapirányzatát analitikus függvény megadásával írjuk le. (Megjegyzés: a mozgó átlagok módszere nem eredményezett ilyen analitikusan felírható trendfüggvényt.) Ez a módszer a regressziószámítás egy speciális esetének is tekinthető. Ilyenkor a vizsgált jelenség adatait (y_i) az idő (x_i) függvényeként kezelhetjük, és ennek megfelelően végezhetjük el a görbeillesztést. A 6. fejezethez hasonlóan, most is az LNM-t használjuk. Megjegyzés: a regressziószámítással ellentétben, ahol az adatpárok sorrendje lényegtelen, az idősor esetén ugyanez már fontos szerepet játszik!

Az analitikus trendszámítás során is az az első feladatunk, hogy eldöntsük milyen típusú függvény illeszkedne legjobban az idősorra. A megfelelő függvénytípus kiválasztásánál most is használhatjuk az idősor grafikus ábráját.

Lineáris trend

Ha az idősor tartós tendenciáját lineáris függvénnyel modellezzük (**lineáris trend**), akkor felírhatjuk a következő összefüggést:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

A fenti modellben szereplő (számunkra ismeretlen) paraméterek becslése végett különböző időpontokra vagy időszakokra vonatkozó adatokat veszünk (ami egy mintának tekinthető). Ennek a mintának a segítségével (rendszerint az LNM alkalmazásával) határozzuk meg a becsült paramétereket, azaz a $\hat{\beta}_0$ -t, illetve a $\hat{\beta}_1$ -t.

Ha az LNM-t használjuk, a becsült paramétereket a (134)-(135) egyenletrendszer szerint számíthatjuk ki.

Így a (133) egyenletnek megfelelő összefüggéshez jutunk:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i .$$

A normálegyenletek egyszerűsítése végett, dinamikus elemzésnél, gyakran alkalmazunk lineáris transzformációt. Az eredeti időváltozót úgy transzformáljuk, hogy az így kapott

új változó (amelyet a továbbiakban t_i -vel jelölünk) értékeinek összege 0 legyen, azaz

$$\sum_{i=1}^{n} t_i = 0 \qquad (207)$$

teljesül.

A (207) összefüggés mindig biztosítható a 75. és a 76. táblázatban szereplő algoritmus szerint, amelynél a $t = 0$ értéket az idősor közepéhez rendeljük.

Megjegyzés: az analitikus trendszámítás alkalmazásakor mindig ekvidisztáns idősorokat feltételezünk!

Egy jelenség 1995-1999 közötti adatainak lehetséges kódolása
(páratlan számú megfigyelés)

75. táblázat

Év	1995	1996	1997	1998	1999	$\sum_{i=1}^{5} t_i = 0$
t_i	-2	-1	0	1	2	

Egy jelenség 1996-1999 közötti adatainak lehetséges kódolásai
(páros számú megfigyelés)

76. táblázat

Év	1996	1997	1998	1999	$\sum_{i=1}^{4} t_i = 0$
t_{1i}	-1,5	-0,5	0,5	1,5	
t_{2i}	-3	-1	1	3	

Ha az eredeti időváltozót transzformáltuk, akkor a trendegyenlet felírásakor kötelezően meg kell adnunk a kiindulópontot (a $t = 0$ értékhez tartózó időpontot), illetve az egyes tengelyeken használt egységeket.

Megjegyzés: a kiindulópont megadásánál mindenféleképpen figyelembe kell vennünk: az idősor típusát és azt, hogy adataink melyik időponthoz tartoznak.

Az új változó bevezetésével, figyelembe véve a (207) összefüggést, az eredeti normálegyenletek alkalmazása helyett, a becsült paramétereket a (208)-(209) képletek

segítségével számíthatjuk ki.

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} \tag{208}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} t_i \cdot y_i}{\sum_{i=1}^{n} t_i^2} \tag{209}$$

A $\hat{\beta}_0$ becsült paraméter a $t = 0$ időponthoz (ami az idősorunk közepén van) tartozó becslés, így ez (a regressziószámítással ellentétben) mindig értelmezhető. A konstans paraméter tartamidősor esetén az idősor átlagos értékének tekinthető.

A $\hat{\beta}_1$ becsült paraméter azt mutatja meg, hogy az adott időszakban a vizsgált jelenség időegységenként átlagosan hány egységnyivel változott.

Megjegyzés: a (204) szerint definiált mutatót ugyanígy értelmezhetjük. Azonban ez a két mutató általában nem egyenlő, mert a $\overline{d}$ meghatározásakor csak az idősor első és utolsó adatát, míg $\hat{\beta}_1$ kiszámításakor az idősor összes megfigyelési értékét figyelembe vesszük.

Az idősorok empirikus elemzésekor gyakran nem csak az éves adatokra van szükség, hanem a negyedéves, illetve havi adatokra is. Az éves lineáris trend ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) segítségével ezeket ki tudjuk számítani a 77. táblázatban közölt összefüggések szerint.

Megjegyzés: mivel különböző időegységek szerepelhetnek a trendfüggvényben, mindig fel kell tüntetni a kiindulópontot és az időtengelyen felvett egységet, ami a leggyakrabban év, negyedév, illetve hónap szokott lenni.

Negyedéves és havi trendértékek kiszámítása

77. táblázat

A trend fajtája	Az idősor típusa	
	tartamidősor	állapotidősor
negyedévi	$\hat{y} = \frac{\hat{\beta}_0}{4} + \frac{\hat{\beta}_1}{16} x$	$\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{4} x$
havi	$\hat{y} = \frac{\hat{\beta}_0}{12} + \frac{\hat{\beta}_1}{144} x$	$\hat{y} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{12} x$

Exponenciális trend

Ha az idősor folyamán az időegységenkénti relatív változás mutatkozik megközelítőleg állandónak, akkor **exponenciális trendegyenlettel** közelítjük a megfigyelési értékeket. Ennek felírása a (145) képletnek megfelelő. Ezt (a 6.2. fejezetben ismertetett módon) logaritmizálva, a lineáris esethez hasonló normálegyenletekhez jutunk. Ha a $t = 0$ értéket most is az idősor közepéhez igazítjuk, akkor a (210)-(211) szerinti képletek segítségével határozhatjuk meg a becsült paramétereket.

$$\log \hat{\beta}_0 = \frac{\sum_{i=1}^{n} \log y_i}{n} \tag{210}$$

$$\log \hat{\beta}_1 = \frac{\sum_{i=1}^{n} t_i \cdot \log y_i}{\sum_{i=1}^{n} t_i^2} \tag{211}$$

A paraméterek eredeti értékét a fentiek (logaritmus alapjának megfelelő) hatványozásával kaphatjuk meg.

A $\hat{\beta}_0$ becsült paraméter most is a $t = 0$ időponthoz tartozó becslés.

A $\hat{\beta}_1$ becsült paraméter az időegységenkénti átlagos változás relatív mértékét (p) és irányát adja meg a vizsgált időtartam alatt, ahol p százalékban kifejezve:

$$p = (\hat{\beta}_1 - 1) \cdot 100 .$$

A $\hat{\beta}_1$ (illetve a p) jelentését tekintve megegyezik a (35) szerint definiált növekedés átlagos ütemével ($\bar{l}$). Ez a két mutató sem mindig egyezik meg, mert az utóbbi (a $\bar{d}$ -hoz hasonlóan) a növekedés átlagos ütemének becslésére csak az idősor első és utolsó adatát használja, míg $\hat{\beta}_1$ most is figyelembe veszi az idősor összes megfigyelési értékét.

76. példa

A személyi jövedelemadó helyi önkormányzatoknál maradó részarányának tartamidősorát a 78. táblázat tartalmazza.

Az önkormányzatok részesedése az SZJA-ból

78. táblázat

Év	SZJA részesedés mértéke (%)
1991	50
1992	50
1993	30
1994	30
1995	30
1996	25
1997	22
1998	20
1999	15
2000	5

Forrás: Pénzügyminisztérium

Illesszünk exponenciális trendet az adott tartamidősorra!

A trendegyenlet meghatározásához szükséges mellékszámításokat a 79. táblázat tartalmazza.

Az exponenciális trendfüggvény illesztéséhez szükséges adatok

79. táblázat

Év	t_i	$\lg y_i$	$t_i \cdot \lg y_i$	t_i^2
1991	-4,5	1,6990	-7,645	20,25
1992	-3,5	1,6990	-5,946	12,25
1993	-2,5	1,4771	-3,693	6,25
1994	-1,5	1,4771	-2,216	2,25
1995	-0,5	1,4771	-0,739	0,25
1996	0,5	1,3979	0,699	0,25
1997	1,5	1,3424	2,014	2,25
1998	2,5	1,3010	3,253	6,25
1999	3,5	1,1761	4,116	12,25
2000	4,5	0,6990	3,145	20,25
Összesen	0,0	13,7458	-7,012	82,50

A táblázat utolsó sorának adatait a (210)-(211) képletekbe helyettesítve a következő eredményeket kapjuk:

$$\lg \hat{\beta}_0 = 1{,}3746\,; \quad \text{illetve} \quad \lg \hat{\beta}_1 = \text{-}0{,}0850\,.$$

Innen:

$$\hat{\beta}_0 = 23{,}6906\,; \quad \text{illetve} \quad \hat{\beta}_1 = 0{,}8223\,.$$

Az exponenciális trendegyenlet az alábbi.

$$\hat{y}_i = 23{,}6906 \cdot 0{,}8223^{t_i}$$

Kiindulópont: 1995. december 31.

A *t* tengelyen 1 egység:1 év.

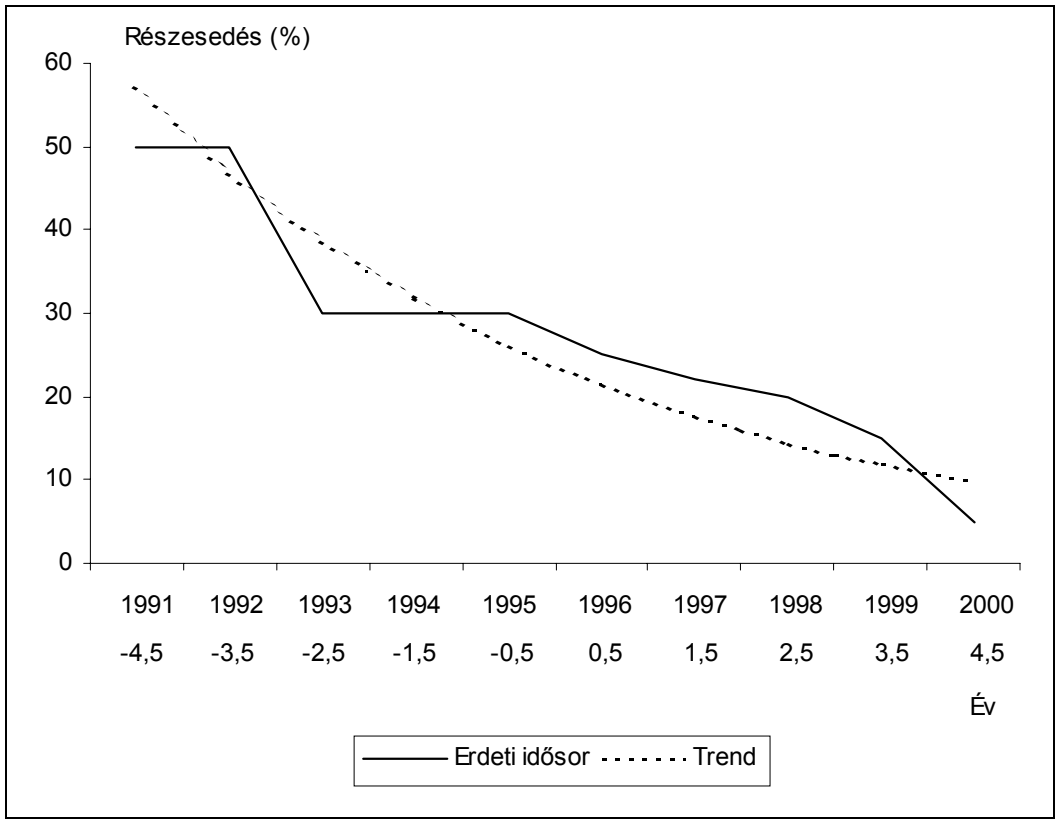
Az *y* tengelyen 1 egység:1 %.

A $\hat{\beta}_1 = 0{,}8223$ azt jelenti, hogy az önkormányzatok SZJA részesedésének mértéke (a vizsgált időszakban) évente átlagosan 0,8223 szorosára változik.

Mivel $p = (0{,}8223 - 1) \cdot 100$; az átlagos éves csökkenés 17,77%.

Az eredeti idősort és az illesztett trendet a 42. ábra mutatja.

Az önkormányzatok részesedése az SZJA-ból 1991-2000 között



42. ábra

Parabolikus trend

A (**másodfokú**) **parabola trendegyenletét** (147) képlethez hasonlóan definiálhatjuk. Ezt alkalmazva a 6.2. fejezetben ismertetett (parabolikus függvényhez tartozó) normálegyenletekből álló egyenletrendszert kell megoldanunk.

Ha a $t = 0$ értéket most is az idősor közepéhez igazítjuk, azaz a $\sum_{i=1}^{n} t_i = 0$ teljesül, akkor az egyenletrendszerünk a (212)-(214) összefüggésekkel is felírható.

Ez az (eredetinél egyszerűbb) egyenletrendszer, (213) szerint, közvetlenül adja a β_1 ismeretlen paraméter becsült értékét.

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_2 \sum_{i=1}^{n} t_i^2 \tag{212}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i^2} \tag{213}$$

$$\sum_{i=1}^{n} t_i^2 y_i = \hat{\beta}_0 \sum_{i=1}^{n} t_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} t_i^4 \tag{214}$$

Megjegyzés: a regressziószámításhoz hasonlóan, a trendszámításnál sem tudjuk közvetlenül értelmezni a $\hat{\beta}_1$ és $\hat{\beta}_2$ paramétereket. A $\hat{\beta}_0$ paraméter a kiindulóponthoz tartozó trendértéket adja, tehát ugyanúgy értelmezhető, mint a lineáris és az exponenciális trendfüggvények esetében.

77. példa

A táppénzre jogosultak átlagos napi létszámára vonatkozó adatokat a 80. táblázat tartalmazza.

A táppénzre jogosultak számának alakulása 1950-1995 között

80. táblázat

Év	Jogosultak napi átlagos létszáma (ezer fő)
1950	1 867
1955	2 594
1960	2 985
1965	3 417
1970	3 949
1975	4 219
1980	4 230
1985	4 164
1990	4 540
1995	3 827

Forrás: Országos Egészségbiztosítási Pénztár

Illesszünk (másodfokú) parabolát az adott tartamidősorra és számítsuk ki a 2005. évhez tartozó trendértéket!

A (212)-(214) összefüggések alkalmazásával a feladat megoldható. Ezekhez szükséges számításokat a 81. táblázat tartalmazza.

A parabolikus trendfüggvény illesztéséhez szükséges adatok

81. táblázat

t_i	y_i	$t_i \cdot y_i$	$t_i^2 \cdot y_i$	t_i^2	t_i^4
-4,5	1 867	-8 401,5	37 806,75	20,25	410,0625
-3,5	2 594	-9 079,0	31 776,50	12,25	150,0625
-2,5	2 985	-7 462,5	18 656,25	6,25	39,0625
-1,5	3 417	-5 125,5	7 688,25	2,25	5,0625
-0,5	3 949	-1 974,5	987,25	0,25	0,0625
0,5	4 219	2 109,5	1 054,75	0,25	0,0625
1,5	4 230	6 345,0	9 517,50	2,25	5,0625
2,5	4 164	10 410,0	26 025,00	6,25	39,0625
3,5	4 540	15 890,0	55 615,00	12,25	150,0625
4,5	3 827	17 221,5	77 496,75	20,25	410,0625
Összesen	35 792	19 933,0	266 624,00	82,50	1 208,6250

A parabolikus trend egyenlete az alábbi.

$$\hat{y}_i = 4027{,}0125 + 241{,}6121 \cdot t_i - 54{,}2803 \cdot t_i^2$$

Kiindulópont: 1972. december 31.

A *t* tengelyen 1 egység:5 év.

Az *y* tengelyen 1 egység: ezer fő.

Az eredeti idősort és az illesztett trendet a 43. ábra mutatja.

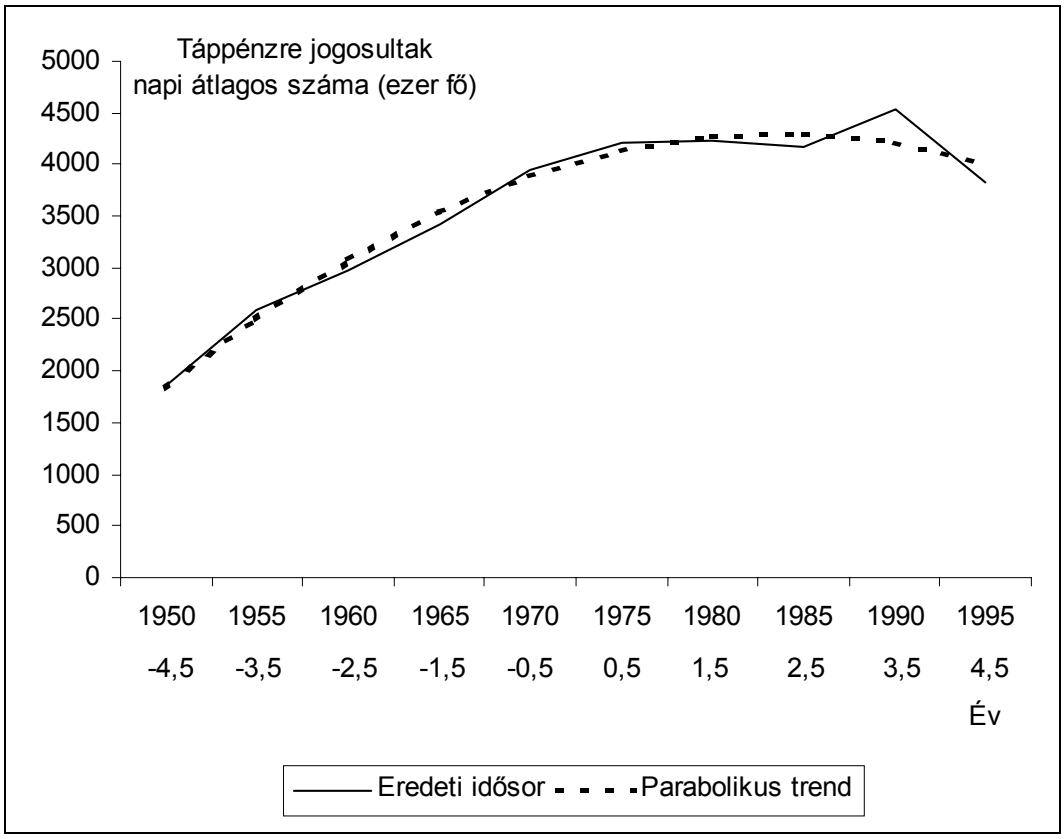
A 2005. évhez tartozó trendértéket a $t = 6{,}5$ helyettesítéssel kapjuk:

$$\hat{y}(t = 6{,}5) = 4027{,}0125 + 241{,}6121 \cdot 6{,}5 - 54{,}2803 \cdot 6{,}5^2 = 3304{,}149\,.$$

Ezek szerint, ha a vizsgált idősorban levő átlagos mozgásirány a 2005. évig változatlan

maradna, a táppénzre jogosultak átlagos napi létszáma 2005-ben 3 304 149 fő lenne.

A táppénzre jogosultak számának alakulása 1950-1995 között



43. ábra

Logisztikus trend

A hosszú idősorok vizsgálatánál a grafikus ábrán gyakran megkülönböztethetünk három szakaszt. Az első szakaszra a lassú növekedés jellemző, míg a másodikban ez felgyorsul, majd a harmadikban a növekedési ütem ismét lassúvá válik, és az adatok egy állandó érték felé tartanak. Ilyenkor célszerű (nyújtott) S alakú görbét illeszteni az idősorra. Ezt a függvénytípust nevezzük **logisztikus trendfüggvénynek**.

Ilyen típusú függvényt leggyakrabban a népességstatisztikában, (tartós fogyasztási) termékek keresleténél használhatunk. Az utóbbi esetben az említett S alakú görbe a termék életgörbéje, és szakaszai megfelelnek a termékbevezetés, a tömegszerűvé válás és a telítődés szakaszának.

A logisztikus görbék közül mi a (215) képlettel definiált becslőfüggvényt fogjuk használni.

$$\hat{y}_i = \frac{\hat{y}_{\max}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i}} \tag{215}$$

Az $\hat{y}_{\max}$ paraméter a telítődési szint, a (215) függvény felső (vízszintes) aszimptotája.

A logisztikus trend paramétereinek meghatározása a legkisebb négyzetek módszere szerint jóval bonyolultabb, mint az eddig ismertetett modellek esetében, ezért először egy egyszerűbb (kevésbé egzakt) megoldást ismertetünk: a **három kiválasztott pont módszerét**.

Első lépésként, az említett három szakaszra jellemző helyen, válasszunk ki három pontot. Ezek (kötelezően) egymástól egyenlő távolságra legyenek. Jelölésükre vezessük be a következő szimbólumokat: x_0, $x_0 + m$, $x_0 + 2m$, ahol m a kiválasztott pontok egymástól való (azonos) időbeli távolságát jelöli és $x_0 = 0$.

Második lépésként meghatározzuk az így kiválasztott időpontok környezetéhez tartozó átlagos adatot ($\overline{Y}_{x_0}$, $\overline{Y}_{x_0+m}$, $\overline{Y}_{x_0+2m}$).

Harmadik lépésként kiszámítjuk a (215) függvény paramétereit a (216)-(218) összefüggések segítségével.

$$\hat{y}_{\max} = \frac{2 \cdot \overline{Y}_{x_0} \cdot \overline{Y}_{x_0+m} \cdot \overline{Y}_{x_0+2m} - \overline{Y}^2_{x_0+m} \cdot (\overline{Y}_{x_0} + \overline{Y}_{x_0+2m})}{\overline{Y}_{x_0} \cdot \overline{Y}_{x_0+2m} - \overline{Y}^2_{x_0+m}} \tag{216}$$

$$\hat{\beta}_0 = \ln\left(\frac{\hat{y}_{\max} - \overline{Y}_{x_0}}{\overline{Y}_{x_0}}\right) \tag{217}$$

$$\hat{\beta}_1 = \frac{1}{m} \ln\left(\frac{\overline{Y}_{x_0} \cdot (\hat{y}_{\max} - \overline{Y}_{x_0+m})}{\overline{Y}_{x_0+m} \cdot (\hat{y}_{\max} - \overline{Y}_{x_0})}\right) \tag{218}$$

78. példa

Hazánk személygépkocsi-állományát az 1956-1997 közötti időszakra a 82. táblázat tartalmazza.

A személygépkocsi-állomány 1956-1997 között (az év végén, ezer db)

82. táblázat

Év	Szgk. száma	Év	Szgk. száma	Év	Szgk. száma
1956	11	1970	239	1984	1344
1957	13	1971	284	1985	1436
1958	18	1972	333	1986	1539
1959	25	1973	400	1987	1660
1960	31	1974	481	1988	1790
1961	40	1975	568	1989	1732
1962	53	1976	641	1990	1945
1963	71	1977	720	1991	2015
1964	86	1978	820	1992	2058
1965	99	1979	934	1993	2092
1966	117	1980	1013	1994	2177
1967	144	1981	1105	1995	2245
1968	162	1982	1182	1996	2264
1969	191	1983	1258	1997	2297

Forrás: Magyar Statisztikai Zsebkönyvek ’58-’98, KSH, Bp.

Illesszünk logisztikus trendfüggvényt az adott állapotidősorhoz a három kiválasztott pont módszerének alkalmazásával, és ábrázoljuk az empirikus és az elméleti adatokat!

A módszer lényege az, hogy (első lépésként) önkényesen kiválasztunk három, a szakaszokat jól jellemző pontot. Legyenek ezek 1962., 1977. és 1992. december 31.

A következő lépésben az adott pontok (önkényesen kiválasztott nagyságú) környezetében kiszámítjuk a kronologikus átlagokat a (203) képlet alapján a 83. táblázatban közöltek szerint.

A 83. táblázatban szereplő adatokat a (216)-(218) képletekbe helyettesítve a következő eredményeket kapjuk: $\hat{y}_{\max} = 2540{,}1$; $\hat{\beta}_0 = 3{,}8248$; $\hat{\beta}_1 = -0{,}1938$.

A logisztikus trendfüggvény meghatározásához szükséges részeredmények

83. táblázat

A kiválasztott három időpont	Az idősor tagjainak új jelölése	A 2 éves környezet kronologikus átlagai
1962. dec. 31.	$x_0 = 0$	$\overline{Y}_0 = \frac{\frac{40}{2} + 53 + \frac{71}{2}}{2} = 54{,}25$
1977. dec. 31.	$x_0 + m = 15$	$\overline{Y}_{15} = \frac{\frac{641}{2} + 720 + \frac{820}{2}}{2} = 725{,}25$
1992. dec. 31.	$x_0 + 2m = 30$	$\overline{Y}_{30} = \frac{\frac{2015}{2} + 2058 + \frac{2092}{2}}{2} = 2055{,}75$

Ezek szerint a logisztikus trendfüggvény az alábbi.

$$\hat{y}_i = \frac{2540{,}1}{1 + e^{3{,}8248 - 01938 \cdot x_i}}$$

Kiindulópont: 1962. december 31.

A *x* tengelyen 1 egység:1 év.

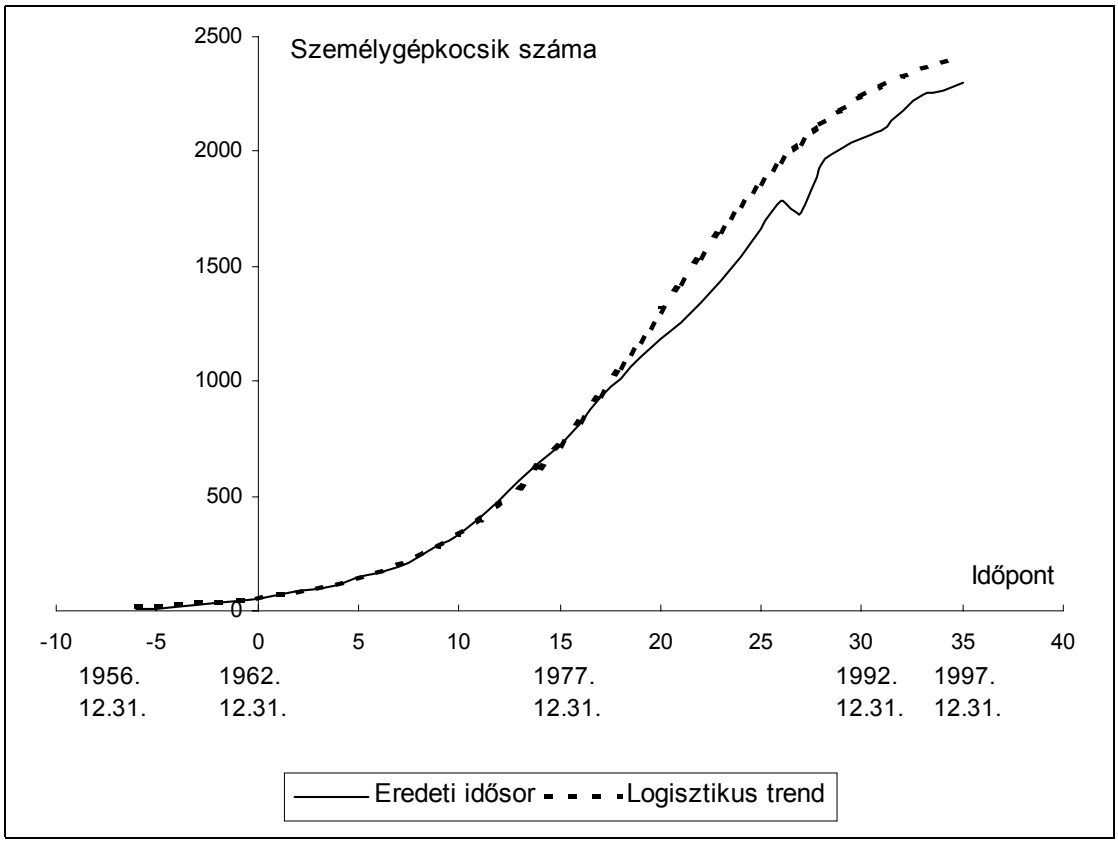
Az *y* tengelyen 1 egység: ezer db.

Az empirikus és a fenti függvény szerinti adatokat a 44. ábra mutatja.

Megjegyzés: az ismertetett módszer egyik hátránya, hogy az idősor harmadik szakaszában általában felülbecsüli a vizsgált adatsort. Ez a 44. ábrán is jól látható.

A logisztikus trend illesztésére most egy összetettebb, de önkényes elemeket nem tartalmazó módszert ismertetünk. Ennek az a lényege, hogy előbb (219) alapján megbecsüljük az idősor telítődési szintjét, és ennek ismeretében linearizáljuk a (215) trendfüggvényt.

A személygépkocsi-állomány alakulása (az év végén, ezer db)



44. ábra

A szaturációs szint becslése végett a következő differenciaegyenletből indulunk ki:

$$y_{i+1} = (1-\beta_1)y_i + \frac{\beta_1}{y_{\max}} y_i^2 .$$

Vezessük be az alábbi helyettesítéseket.

$$u_i = y_{i+1}$$
$$b = (1-\beta_1)$$
$$c = \frac{\beta_1}{y_{\max}}$$

Ezek szerint az eredeti differenciaegyenlet felírható a következő módon is:

$$u_i = b \cdot y_i + c \cdot y_i^2 \qquad i = 1,2,...,n-1.$$

Ez nem más, mint egy másodfokú parabola regressziófüggvénye. Megjegyzés: a vizsgált függvény nem azonos a (147) alatt ismertetett regressziófüggvénnyel, mert a konstans tag itt nem szerepel!

A legkisebb négyzetek módszerét alkalmazva megkapjuk a b és a c becsült értékét, amelyek segítségével az $y_{\max}$ szintén becsülhető. (Megjegyzés: a β_1 becsült értékét nem a b paraméter ismeretében számítjuk ki!)

Figyelembe véve a fentieket, a szaturációs szint becslésére felírható az alábbi explicit összefüggés.

$$\hat{y}_{\max} = \frac{\sum_{i=1}^{n-1} y_i^4 \cdot \sum_{i=1}^{n-1} y_i^2 - \left(\sum_{i=1}^{n-1} y_i^3\right)^2 - \sum_{i=1}^{n-1} y_i y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^4 + \sum_{i=1}^{n-1} y_i^2 y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^3}{\sum_{i=1}^{n-1} y_i^2 y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^2 - \sum_{i=1}^{n-1} y_i y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^3} \qquad (219)$$

A (219) segítségével kiszámított $\hat{y}_{\max}$ értéket tekintjük a (215) trendfüggvény (számlálójában szereplő) paraméterének.

A (215) egyenlet (átalakítások után) az alábbi alakra hozható:

$$\hat{z}_i = \left(\ln\left(\frac{\hat{y}_{\max} - y_i}{y_i}\right)\right)^{\wedge} = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad (220)$$

ahol:

$$z_i = \ln\left(\frac{\hat{y}_{\max} - y_i}{y_i}\right).$$

Megjegyzés: az előző egyenlet helyett a következő lineáris trendegyenletet is leírhattuk volna:

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

A paraméterek becslését (z_i megfelelő helyettesítésével) a (208)-(209) képletek

alkalmazásával kaptuk.

79. példa

A 78. példa adatai alapján, ezzel a módszerrel is határozzuk meg a személygépkocsi-állomány idősorához illesztett trendfüggvényt!

Először (a 84. táblázat adatai alapján) ki kell számítanunk a telítődési szint becslését. A (219) képletbe behelyettesítve:

$$\hat{y}_{\max} = \frac{\text{-1,08588}E20}{\text{-4,43270}E16} = 2\,449{,}71\,.$$

Most már felírhatjuk a linearizált egyenletet:

$$z_i = \ln\left(\frac{2\,449{,}71 - y_i}{y_i}\right) = \beta_0 + \beta_1 x_i + \varepsilon_i\,.$$

A logisztikus trendfüggvény telítődési szintjének becsléséhez szükséges részeredmények

84. táblázat

y_i	y_{i+1}	y_i^2	y_i^3
11	13	121	1 331
13	18	169	2 197
18	25	324	5 832
25	31	625	15 625
⋮			
2 092	2 177	4 376 464	91 55 562 688
2 177	2 245	4 739 329	10 317 519 233
2 245	2 264	5 040 025	11 314 856 125
2 264	2 297	5 125 696	11 604 575 744
35 336	37 622	55 346 172	99 671 436 704

A logisztikus trendfüggvény telítődési szintjének becsléséhez szükséges részeredmények (folytatás)

84. táblázat

y_i^4	$y_i \cdot y_{i+1}$	$y_i^2 \cdot y_{i+1}$
1,46410E+04	1,43000E+02	1,57300E+03
2,85610E+04	2,34000E+02	3,04200E+03
1,04976E+05	4,50000E+02	8,10000E+03
3,90625E+05	7,75000E+02	1,93750E+04
⋮		
1,91534E+13	4,55428E+06	9,52756E+09
2,24612E+13	4,88737E+06	1,06398E+10
2,54019E+13	5,08268E+06	1,14106E+10
2,62728E+13	5,20041E+06	1,17737E+10
1,90867E+14	5,78754E+07	1,03425E+11

A $\hat{\beta}_0$ és $\hat{\beta}_1$ kiszámításához szükséges részeredményeket a 85. táblázat tartalmazza.

A logisztikus trendfüggvény illesztéséhez szükséges részeredmények

85. táblázat

Év	t_i	y_i	z_i	$z_i \cdot t_i$	t_i^2	$\hat{y}_i$
1956	-20,5	11	5,40133	-110,72728	420,25	16,7
1957	-19,5	13	5,23346	-102,05240	380,25	20,1
1958	-18,5	18	4,90598	-90,76063	342,25	24,2
1959	-17,5	25	4,57459	-80,05538	306,25	29,1
⋮						
1994	17,5	2177	-2,07728	-36,35239	306,25	2195,1
1995	18,5	2245	-2,39485	-44,30467	342,25	2234,9
1996	19,5	2264	-2,50068	-48,76327	380,25	2269,0
1997	20,5	2297	-2,71079	-55,57124	420,25	2298,0
Össz.	0,0	37633	47,57055	-1159,02408	6170,50	37351,3

A (208)-(209) képletek figyelembevételével kiszámíthatjuk a (215) trendfüggvény még nem ismert paramétereit.
Ezek:

$$\hat{\beta}_0 = \frac{47{,}57055}{42} = 1{,}13263\,;$$

illetve:

$$\hat{\beta}_1 = \frac{-1159{,}02408}{6170{,}50} = -0{,}18783\,.$$

Ezek szerint a logisztikus trendfüggvény az alábbi.

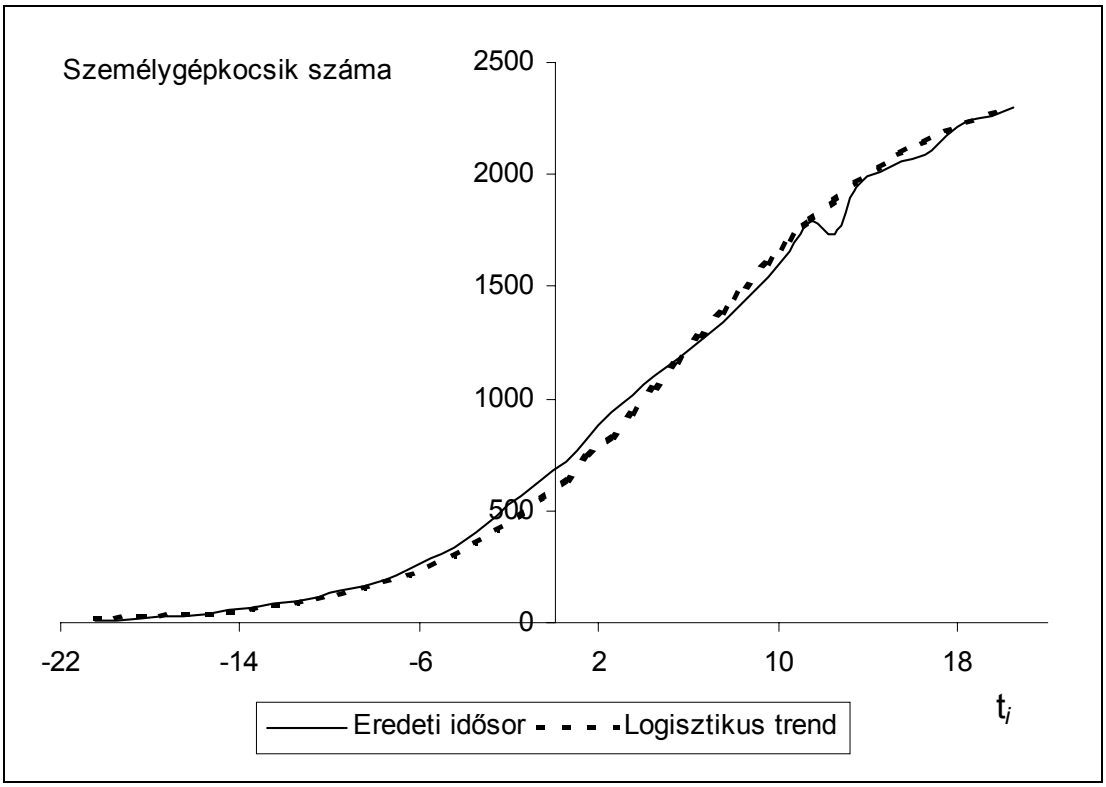
$$\hat{y}_i = \frac{2449{,}7}{1 + e^{1{,}13263 - 0{,}18783 \cdot t_i}}$$

Kiindulópont: 1977. június 30.
A *t* tengelyen 1 egység:1 év.
Az *y* tengelyen 1 egység: ezer db.

Az empirikus és a fenti függvény szerinti adatokat a 45. ábra mutatja.

A 44. és a 45. ábra összehasonlításával jól látható, hogy a második módszer jóval pontosabb (de összetettebb is) az elsőnél. Erre utal a becsült értékek összege is, ami a második módszer szerint 37351,3; az első módszer szerint 40572,8; míg az eredeti adatok összege 37633 ezer db.

A személygépkocsi-állomány alakulása (év végi adatok, ezer db)



45. ábra

A trendhatás mellett, az idősorok adatait a szezonális tényező is befolyásolhatja. A következő fejezetben ezen tényezők számszerűsítésének módszereit ismertetjük.

10.4. Szezonális ingadozások elemzése

Ahogy azt már említettük, a szezonális komponens (*S*) az idősorban rendszeresen ismétlődő, azonos periódusú és szabályos amplitúdójú ingadozásokat mutatja. Ezek az empirikus vizsgálatokban leggyakrabban havi vagy negyedéves ingadozások. Most azt fogjuk megvizsgálni, hogy az *S* komponens értékét hogyan tudjuk becsülni egy megfigyelt idősorból. Arra keressük tehát a választ, hogy a szezonális hatás az egyes periódusokban milyen mértékben (additív modell), illetve milyen arányban (multiplikatív modell) téríti el az idősor adatait az alapirányzattól. A szezonális hatás kimutatását úgy végezzük, hogy kiszűrjük az idősorból a másik két tényező hatását (a trendet most már *y*-nal helyettesítve).

Additív modell esetén:

$$y_{ij} = y_{ij}^a + S_j^a + e_{ij},$$

ezért a trendhatást az ismertetett eljárások alapján kiszámítva, és a megfigyelt értékekből levonva, majd a kapott értékeket átlagolva jutunk a becsült **nyers szezonális eltérésekhez**.

Ha a trendet a mozgó átlagok segítségével számítottuk ki, akkor:

$$s_j^a = \frac{\sum_{i=1}^{n/p}\left(y_{ij} - \hat{y}_{ij}^a\right)}{n/p - 1} \qquad j = 1,2,...,p; \tag{221}$$

ha pedig analitikus trendszámítást alkalmaztunk, akkor:

$$s_j^a = \frac{\sum_{i=1}^{n/p}\left(y_{ij} - \hat{y}_{ij}^a\right)}{n/p}. \tag{222}$$

Mivel a szezonális hatások egy perióduson belül kiegyenlítik egymást, ezt a becsült szezonális eltérésektől is elvárjuk. Ennek biztosítására a nyers szezonális eltérésekből kiszámítjuk a **korrigált szezonális eltéréseket**.

$$\tilde{s}_j^a = s_j^a - \bar{s}_j^a, \tag{223}$$

ahol:

$$\bar{s}_j^a = \frac{\sum_{j=1}^{p} s_j^a}{p}.$$

A becsült korrigált szezonális eltérésekre:

$$\sum_{j=1}^{p} \tilde{s}_j^a = 0.$$

A fenti módszerrel kapott becsült szezonális eltérések azt fejezik ki, hogy az idősor megfigyelt értékei átlagosan mennyivel térnek el a trendértéktől a szezonális hatás következtében.

Multiplikatív modell esetén

$$y_{ij} = y_{ij}^m \cdot S_j^m \cdot u_{ij}.$$

Itt az additív modellhez hasonló módon tudjuk kimutatni a szezonális hatást. A becsült **nyers szezonindexeket** is kétféleképpen lehet kiszámítani.

Ha a trendet a mozgó átlagok segítségével számítottuk ki, akkor:

$$s_j^m = \frac{\sum_{i=1}^{n/p} \frac{y_{ij}}{\hat{y}_{ij}^m}}{n/p - 1}, \tag{224}$$

ha pedig analitikus trendszámítást alkalmaztunk, akkor:

$$s_j^m = \frac{\sum_{i=1}^{n/p} \frac{y_{ij}}{\hat{y}_{ij}^m}}{n/p}. \tag{225}$$

A **korrigált szezonindexek**:

$$\tilde{s}_j^m = \frac{s_j^m}{\bar{s}_j^m}, \tag{226}$$

ahol:

$$\bar{s}_j^m = \frac{\sum_{j=1}^{p} s_j^m}{p}.$$

A becsült korrigált szezonindexekre:

$$\sum_{j=1}^{p} \tilde{s}_j^m = p \quad \text{vagy} \quad 100p\%.$$

Megjegyzés: havi adatok esetén a fenti összeg 12-vel vagy 1200 százalékkal egyenlő.

Az alkalmazott módszerrel kapott becsült szezonindexek azt fejezik ki, hogy az idősor megfigyelt értékei, a szezonális hatás következtében, átlagosan hányszorosai a trendértéknek.

80. példa

A 75. példa 73. táblázata az élelmiszerek fogyasztói árindexeit tartalmazza (havi bontásban) 1995 és 1998 között. Elemezzük az árindexek időbeli alakulását, számszerűsítsük a szezonális komponenst!

Ebben az esetben, az idősor alapirányzatát jellemző trend meghatározására, használjunk analitikus trendillesztést. A 41. ábra alapján lineáris modellt feltételezhetünk.
A (208)-(209) képletek alkalmazásával az alábbi eredményre juthatunk.

$$y_i = 101{,}258 - 0{,}039 \cdot t_i$$

Kiindulópont: 1996. december 31.
A t tengelyen 1 egység:1 hónap.
Az y tengelyen 1 egység:1 %.

Számítsuk most ki az eredeti adatok lineáris trendtől való különbségeit, illetve hányadosait.

A megfigyelt értékek és a trend értékeinek különbségei ($y_{ij} - \hat{y}_{ij}^{a}$)

86. táblázat

Hónap	1995	1996	1997	1998	Átlag
Jan.	3,732	1,897	2,461	2,125	2,554
Febr.	1,171	0,135	0,300	0,864	0,618
Márc.	-0,690	0,074	-0,262	0,803	-0,019
Ápr.	0,948	-0,387	0,377	0,842	0,445
Máj.	0,487	-0,749	1,516	1,780	0,759
Jún.	-2,374	-0,910	2,955	-0,681	-0,253
Júl.	-2,936	-1,471	-2,707	-2,942	-2,514
Aug.	-3,497	-1,932	-1,768	-2,404	-2,400
Szept.	0,642	0,806	0,671	-0,365	0,438
Okt.	1,081	0,345	0,709	-0,226	0,477
Nov.	-0,381	-0,716	1,148	-0,288	-0,059
Dec.	-0,442	-0,378	0,787	-0,149	-0,045
Összesen:					0,000

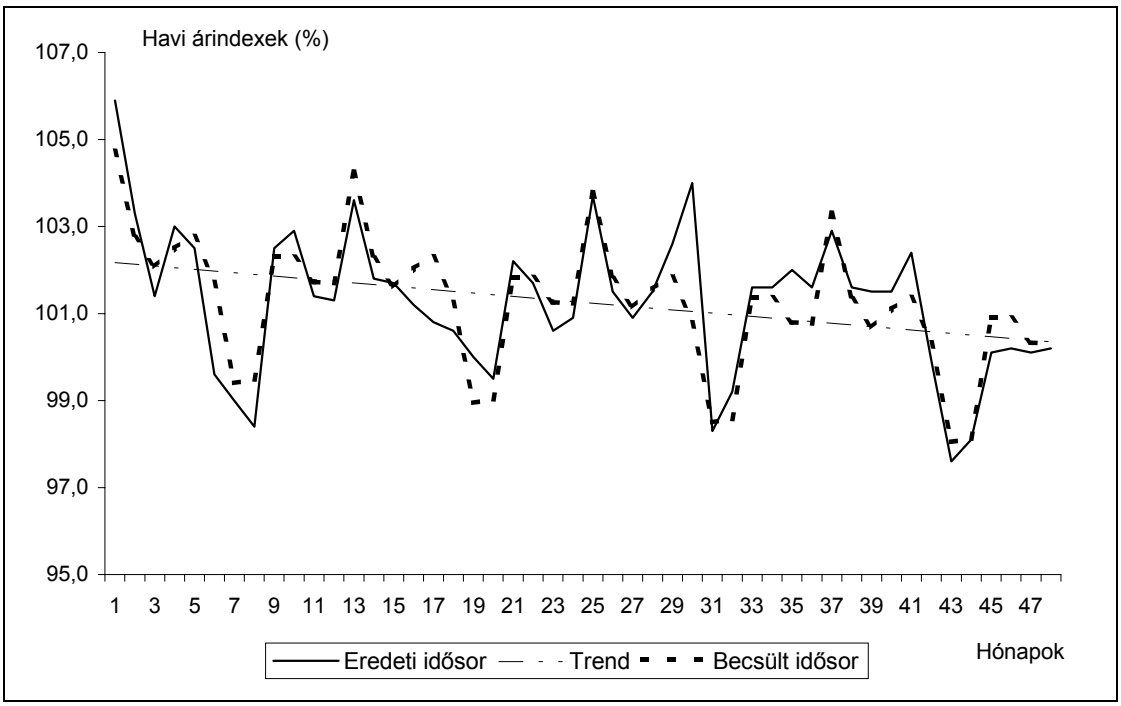
A megfigyelt értékek és a trend értékeinek hányadosai ($y_{ij} / \hat{y}_{ij}^{m}$)

87. táblázat

Hónap	1995	1996	1997	1998	Átlag
Jan.	1,037	1,019	1,024	1,021	1,025
Febr.	1,011	1,001	1,003	1,009	1,006
Márc.	0,993	1,001	0,997	1,008	1,000
Ápr.	1,009	0,996	1,004	1,008	1,004
Máj.	1,005	0,993	1,015	1,018	1,008
Jún.	0,977	0,991	1,029	0,993	0,998
Júl.	0,971	0,986	0,973	0,971	0,975
Aug.	0,966	0,981	0,982	0,976	0,976
Szept.	1,006	1,008	1,007	0,996	1,004
Okt.	1,011	1,003	1,007	0,998	1,005
Nov.	0,996	0,993	1,011	0,997	0,999
Dec.	0,996	0,996	1,008	0,999	1,000
Összesen:					12,000

A szezonindexek állandóbbak, mint a szezonális eltérések, ezért a továbbiakban a multiplikatív modell használata indokolt. Mivel a szezonindexek összege 12-vel egyenlő, ezért nincs szükség a (226) szerinti korrigálásra.

A szezonális hatás ábrázolása



46. ábra

A fejezet végén megemlítjük a **szezonális kiigazítás** fogalmát. Ezalatt azt értjük, hogy a megfigyelt idősort megtisztítjuk a szezonális hatásoktól.

A szezonális kiigazítás eredményeként ún. **szezonálisan kiigazított idősort** kapunk, amely gyakran szerepel a különböző statisztikai kiadványokban.

11. Többváltozós regresszió- és korrelációszámítás

11.1. Többváltozós regressziószámítás

A 6. fejezetben már részletesebben tárgyaltuk a kétváltozós regressziós modellt, amelyben egyetlen magyarázóváltozót szerepeltettünk. A gyakorlatban azonban egy jelenség alakulását általában nem egy, hanem több szignifikáns tényező határozza meg. A regressziós modell javítása érdekében ezért minden releváns tényező szerepeltetése célszerű.

A változók száma mellett fontos szerepe van a regressziós modellben alkalmazott függvény típusának is, amely egyszerűbb esetekben lineáris, de az empirikus elemzéseknél gyakran nemlineáris.

Az előzőek alapján a regressziós modellek négy esetét különböztethetjük meg.

A regressziós modellek esetei

88. táblázat

A regresszió-függvény típusától függően a modell lehet	A változók számától függően a modell lehet	
	kétváltozós lineáris	többváltozós lineáris
	kétváltozós nemlineáris	többváltozós nemlineáris

Empirikus elemzéseknél az első lépések egyikeként el kell dönteni, hogy a fenti esetek közül melyikkel dolgozunk. Ennek kiválasztását és a későbbiekben ismertetett egyéb feltételrendszer meghatározását nevezzük a **modell specifikációjának**.

A standard lineáris regressziós modell

A 88. táblázatban közölt esetek közül könyvünkben csak a lineáris, illetve lineáris alakra hozható kétváltozós vagy többváltozós regressziófüggvényekkel foglalkozunk. Ezek általános alakja, (132)-höz hasonlóan, (*n* elemű mintát feltételezve) felírható (227) szerint is.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_m x_{im} + e_i \quad i = 1,2,...,n \quad m+1 < n < N \quad (227)$$

A továbbiakban gyakran fogjuk alkalmazni a regressziós modell mátrixalgebrai jelölésmódját. A következő jelöléseket fogjuk használni:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & & x_{2m} \\ \vdots & & & \\ 1 & x_{n1} & & x_{nm} \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \qquad (228)$$

ahol m a magyarázóváltozók száma és $\mathbf{X}$ első oszlopa mindig egy összegező vektor.

A modell specifikációjának fontos részét alkotják még az alábbiakban ismertetett feltételek is.

– A változók között fennállnak a következő összefüggések:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}, \text{ illetve } \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

– A magyarázóváltozók nem sztochasztikusak (mérési hibát nem tartalmaznak), valamint lineárisan függetlenek (tehát nem redundánsak). Ez utóbbi azt jelenti, hogy az $\mathbf{X}$ mátrix rangja az oszlopainak számával egyenlő:

$$\rho(\mathbf{X}) = m + 1.$$

– A hibatagok nulla várható értékű, konstans varianciájú (σ^2), korrelálatlan valószínűségi változók, amelyek együttes eloszlása n-dimenziós normális eloszlás:

$$\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

ahol $\mathbf{I}$ az egységmátrix.

Az összes eddig ismertetett feltételeknek eleget tevő modelleket nevezzük **standard lineáris regressziós modelleknek**.

A regressziószámítás gyakorlati alkalmazásakor ügyelnünk kell arra, hogy a fenti modellt ne használjuk, ha valamelyik feltétele szignifikánsan nem teljesül!

Közgazdasági elemzéseknél ennek leggyakrabban három oka lehet:

- **multikollinearitás**: a magyarázóváltozók lineáris függetlenségének hiánya,
- **autokorreláció**: a hibatagok lineárisan nem függetlenek,
- **heteroszkedaszticitás**: a hibatag szórásnégyzete nem állandó.

Ezekkel a jelenségekkel részletesebben majd a 11.3. fejezetben foglalkozunk.

A modellünk funkcionális operátorának meghatározásakor olyan hipersíkot keresünk, amely a legközelebb van az *n*-dimenziós pontfelhőhöz. Ha a $\boldsymbol{\beta}$ paramétervektor becslésére most is a legkisebb négyzetek módszerét alkalmazzuk, akkor a (142) szerinti mátrixegyenlethez juthatunk.

A **GAUSS–MARKOV–tétel**: a legkisebb négyzetek módszere **BLUE** (best linear unbiased estimator) tulajdonságú $\hat{\boldsymbol{\beta}}$ vektort ad, vagyis a becslőfüggvény torzítatlan és (a lineáris modellek közül) a legkisebb szórásnégyzetű (**efficiens**).

A becsült paraméterek értelmezése

A $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_m$ becsült regressziós paramétereket a következőképpen értelmezhetjük: a $\hat{\beta}_j$ azt mutatja meg, hogy az x_j magyarázóváltozó egységnyi növekedése az eredményváltozó átlagosan mekkora változásával jár együtt, ha a többi magyarázóváltozó értéke nem változik. A $\hat{\beta}_j$ együtthatókat, emiatt a ceteris paribus értelmezés miatt, **parciális regressziós együtthatóknak** nevezzük.

A regressziós modell illeszkedésének jósága

Definiáljuk az alábbi eltérés-négyzetösszegeket.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{229}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{230}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{231}$$

Amennyiben a modellünk tartalmaz konstans paramétert, tehát $\beta_0 \neq 0$, akkor a (229)-(231) szerint definiált eltérés-négyzetösszegekre fennáll a következő összefüggés:

$$SST = SSR + SSE . \tag{232}$$

Ezek alapján a (150) szerint definiált lineáris determinációs együttható felírható a (233) képlettel is.

$$r^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \tag{233}$$

Ennek részletesebb ismertetésére majd a 11.2. fejezetben kerül sor. Egy modell illeszkedésének mértéke természetesen azzal definiálható, hogy a teljes eltérés-négyzetösszegnek mekkora részét teszi ki a regresszió által megmagyarázott és a hibataggal kapcsolatos négyzetösszeg.

A modell illeszkedésének jóságát variancia-analízis segítségével tesztelhetjük, amit a többváltozós regressziószámításban **globális *F*-próbának** is nevezünk. Nullhipotézisünk és alternatív hipotézisünk az alábbi módon fogalmazható meg.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$$
$$H_1 : \beta_j \neq 0 \qquad \textit{valamelyik } j\textit{-re} \qquad j = 1, 2, ..., m$$

A fenti nullhipotézis helyességének ellenőrzésére a (234) szerint definiált próbafüggvényt használjuk.

$$F = \frac{SSR / m}{SSE /(n - m - 1)} = \frac{MSR}{MSE} \tag{234}$$

A (234) próbafüggvény *F*-eloszlást követ, a számláló szabadságfoka $\nu_1 = m$, a nevező szabadságfoka $\nu_2 = n - m - 1$.

A variancia-analízis végrehajtását és eredményeit most is ANOVA táblázatban rögzítjük. Ennek általános rendezési formáját a 89. táblázat tartalmazza.

Az ANOVA táblázatban szereplő tapasztalati *F* értéket kell összevetnünk a megfelelő elméleti értékkel. A variancia-analízis (mint tudjuk) jobboldali próba, tehát ha a

tapasztalati F érték kisebb az elméleti értéknél, akkor a nullhipotézist (az adott szignifikancia-szint mellett) elfogadjuk, ami azt jelenti, hogy a vizsgált modell nem alkalmas a megfigyelt jelenség elemzésére. A nullhipotézis elutasítása azonban nem jelenti automatikusan a modell illeszkedésének jóságát!

Az ANOVA táblázat vázlata

89. táblázat

A szóródás oka	Eltérések négyzetösszege	Szabadságfok	Szórásnégyzet becslése	F
Regresszió	SSR	m	MSR	$\frac{MSR}{MSE}$
Hiba	SSE	$n-m-1$	MSE	
Összesen	SST	$n-1$	–	

Paraméterek tesztelése

Az előzőekben az egész modell illeszkedését vizsgáltuk, most egyetlen magyarázóváltozó fontosságát, magyarázó erejét fogjuk tesztelni. Nullhipotézisünk az lesz, hogy az adott x_j magyarázóváltozó nincs szignifikáns kapcsolatban az eredményváltozóval.

$$H_0 : \beta_j = 0 \qquad j = 1, 2, ..., m$$
$$H_1 : \beta_j \neq 0$$

A tesztelésre a következő próbafüggvényt használjuk:

$$F = \frac{\hat{\beta}_j^2}{\mathrm{var}(\hat{\beta}_j)}, \tag{235}$$

ahol $\mathrm{var}(\hat{\beta}_j)$ a

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \frac{\mathbf{e}'\mathbf{e}}{n-m-1} \cdot (\mathbf{X}'\mathbf{X})^{-1} = s_e^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} \tag{236}$$

variancia-kovarianciamátrix (lásd a következő fejezetet) főátlójában szereplő j-edik

elem.

Ez a statisztika $\nu_1 = 1$, $\nu_2 = n - m - 1$ szabadságfokú *F*-eloszlást követ. Ezt a tesztelést **parciális F-próbának** nevezzük.

Mivel a 9.4. fejezetben említett *t* (IV. táblázat szerinti) és *F* értékek közötti összefüggés most így is felírható:

$$t^2_{1-\frac{\alpha}{2}}(n-m-1) = F_{1-\alpha}(1, n-m-1),$$

ezért *t*-eloszlást is alkalmazhatunk. Ekkor a próbafüggvény:

$$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}. \qquad (237)$$

A *t*-próbához tartozó (IV. táblázat szerinti) elméleti érték α szignifikancia-szinten: $t_{1-\frac{\alpha}{2}}(n-m-1)$. Ha az empirikus *t*-érték abszolút értéke kisebb az elméleti értéknél, akkor a H_0-t elfogadjuk, ami azt jelenti, hogy a vizsgált magyarázóváltozó szignifikánsan nem befolyásolja az eredményváltozót, ezért nem célszerű szerepeltetnünk a modellben.

Megjegyzés: a standard lineáris regressziós modellnél a becslések varianciáját eredetileg nem a (236) szerint kell kiszámítani, hanem:

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$$

összefüggés szerint, ahol σ^2 a hibatagok számunkra ismeretlen szórásnégyzete. Az s_e^2, az ún. **reziduális szórásnégyzet**, ennek torzítatlan becslése.

11.2. Többváltozós korrelációszámítás

Korrelációs együtthatók

A 4. és a 6. fejezetben már tárgyaltuk a lineáris korrelációs együtthatót és a lineáris determinációs együtthatót kétváltozós esetre. A többváltozós modellben a lineáris korrelációs együtthatót a változók összes lehetséges párosításában ki tudjuk számítani. Két-két változó közötti kapcsolat szorosságát és irányát mérő lineáris korrelációs együtthatókat a többváltozós modellben **páronkénti korrelációs együtthatóknak** nevezzük. Ezek értékeit az ún. **korrelációs mátrixba** rendezzük, amely a (238) szerint definiált.

$$\mathbf{R} = \begin{bmatrix} 1 & r_{yx_1} & \cdots & r_{yx_m} \\ r_{x_1y} & 1 & & r_{x_1x_m} \\ \vdots & & & \\ r_{x_my} & r_{x_mx_1} & & 1 \end{bmatrix} \tag{238}$$

A lineáris korrelációs együttható szimmetriatulajdonságai miatt az **R** mátrix szimmetrikus, és a főátlójában levő elemek értéke 1. Első sorában (illetve oszlopában) az egyes magyarázóváltozók és az eredményváltozó közötti kapcsolatot jellemző együtthatók állnak, amelyek a regressziós modell magyarázóváltozóinak kiválasztásánál adhatnak segítséget.

Gyakran használjuk a kapcsolat természetének jellemzésére a kovarianciát is. A változók közötti kovarianciát a **variancia-kovarianciamátrixba** rendezzük.

$$\mathbf{C} = \begin{bmatrix} \sigma_y^2 & C_{yx_1} & \cdots & C_{yx_m} \\ C_{x_1y} & \sigma_{x_1}^2 & & C_{x_1x_m} \\ \vdots & & & \\ C_{x_my} & C_{x_mx_1} & & \sigma_{x_m}^2 \end{bmatrix} \tag{239}$$

A variancia-kovarianciamátrix szintén szimmetrikus, főátlójában az egyes változók

varianciája található.[20)]

Megjegyzés: ha a változók eredeti értékei helyett azok standardizált értékeivel dolgozunk, akkor a (238) és a (239) alatti mátrix megegyezik. Ez az összefüggés az empirikus elemzéseknél egyszerűsíti a számításokat.

Az említett **R** és **C** mátrixokat az Excel segítségével is ki tudjuk számítani. Hívjuk meg az **Eszközök** menü **Adatelemzés...** almenüjét és válasszuk ki a felkínált lehetőségek közül a **Korrelációanalízis**, illetve a **Kovarianciaanalízis** menüpontot. Az ekkor megjelenő párbeszédpanellel vigyük be a Bemeneti tartományba az adatainkat tartalmazó megfelelő cellahivatkozásokat. Ha bekapcsoljuk a Feliratok az első sorban (oszlopban) jelölőnégyzetet, akkor a (238)-(239) mátrixok elemei mellett még a hozzájuk tartozó változók megnevezéseit is láthatjuk. (Ezzel a megoldással áttekinthetőbbé válnak az adatok.)

A páronkénti korrelációs együtthatók számításánál a többi változón keresztül gyakorolt közvetett hatást is kimutattuk. Ha a kapcsolat természetét a többi magyarázóváltozót kiszűrve akarjuk kimutatni, akkor **parciális korrelációs együtthatóra** van szükségünk. Ennek kiszámításához fel kell használnunk a korrelációs mátrix inverzét.

$$r_{yx_j \,.\, x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m} = -\frac{\mathbf{R}^{-1}_{yx_j}}{\sqrt{\mathbf{R}^{-1}_{yy} \cdot \mathbf{R}^{-1}_{x_j x_j}}} \qquad (240)$$

A parciális korrelációs együttható indexében először a vizsgálat tárgyát képező változókat tüntetjük fel, majd egy pont után azokat, amelyeknek a hatását kiszűrtük.

A parciális korrelációs együttható négyzetét **parciális determinációs együtthatónak** nevezzük.

[20)] A korrelációs mátrix és a variancia-kovarianciamátrix között, elméleti esetet feltételezve, felírható a következő összefüggés:

$$\mathbf{R} = \left(\sigma^2\mathbf{I}\right)^{-1}\mathbf{C}\left(\sigma^2\mathbf{I}\right)^{-1},$$

ahol $\sigma^2\mathbf{I}$ a hibatag variancia-kovarianciamátrixa, azaz $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$.

A lineáris determinációs együtthatót a többváltozós modellben is többféleképpen kiszámíthatjuk, mi a (241) képletet fogjuk alkalmazni.

$$r^2_{y.x_1,x_2,\ldots,x_m} = 1 - \frac{1}{\mathbf{R}^{-1}_{yy}} \tag{241}$$

Ez az ún. **többszörös determinációs együttható**, amelynek négyzetgyökét **többszörös korrelációs együtthatónak** nevezzük.

A többszörös determinációs együttható azt mutatja meg, hogy az eredményváltozó szórásnégyzetének hány százalékát tudjuk megmagyarázni (együttesen) az összes független változóval.

Lineáris korrelációs együttható tesztelése

Empirikus elemzéseknél mintából szoktuk kiszámítani a lineáris korrelációs együttható (r) értékét, amely általában nullától különböző és a populáció azonos mutatójának (ρ) becslését adja. Az r értékének ismeretében lehetséges annak tesztelése, hogy a lineáris korrelációs együttható szignifikánsan különbözik-e 0-tól. Ennek eldöntésére a (242) szerint definiált próbafüggvényt használjuk, ha a hipotéziseinket az alábbi módon fogalmazzuk meg.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0 .$$

A próbafüggvényünk:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} . \tag{242}$$

Ez a statisztika $\nu = n-2$ szabadságfokú t-eloszlást követ. Kétoldali próbaként hajtjuk végre (azaz közvetlenül használhatjuk a III. táblázatot).

11.3. Multikollinearitás, autokorreláció, heteroszkedaszticitás

Multikollinearitás

A standard lineáris regressziós modell feltételezi, hogy a magyarázóváltozók egymástól lineárisan függetlenek. Ha valamelyik magyarázóváltozó kifejezhető a többi tényezőváltozó lineáris kombinációjaként, vagyis függvényszerű kapcsolatban áll a többi tényezőváltozóval, akkor **teljes** vagy **extrém multikollinearitásról** beszélünk. Ekkor $\mathbf{X}$ rangja nem egyenlő oszlopai számával és az $\mathbf{X}'\mathbf{X}$ mátrix szinguláris, ezért nem invertálható. A teljes multikollinearitás felismerése könnyű, és egyszerűen megoldható az adott magyarázóváltozó elhagyásával. Az empirikus vizsgálatoknál azonban a magyarázóváltozók között inkább sztochasztikus kapcsolat jelentkezik.

A multikollinearitás következményei

Ha a magyarázóváltozók egymástól lineárisan nem függetlenek, akkor az LNM közvetlen alkalmazásával kapott becslések fontosabb tulajdonságai az alábbiak.

- A becslés és az előrejelzés torzítatlan marad.
- A regressziós együtthatók standard hibái nőnek.
- Bizonytalanná, instabillá válnak (a továbbra is torzítatlan) becsléseink.
- Az egyes magyarázóváltozók hatásainak szeparált vizsgálata nem lehetséges, illetve a parciális regressziós együtthatók helyes értelmezése lehetetlenné válik.

A fentiek miatt a magyarázóváltozók kölcsönös függőségének mértékét mindig ellenőriznünk kell.

A multikollinearitás mérése

Ha egy új magyarázóváltozót kapcsolunk be a modellbe, akkor a többszörös determinációs együttható vagy növekszik, vagy egyáltalán nem változik. Minden magyarázóváltozóra kiszámítva, hogy a modellbe utolsó változóként bevonva mennyivel növeli a determinációs együtthatót, ellenőrizhető a multikollinearitás. Ha az említett hatásoknak az összege egyenlő a többszörös determinációs együtthatóval, akkor azt mondhatjuk, hogy a magyarázóváltozók lineárisan függetlenek. Ellenkező esetben az eredményváltozó szórásnégyzetének van olyan része, amit együttesen magyaráz több

változó. A multikollineritás nagyságát ezzel az együttesen magyarázott résszel a (243) módon mérhetjük.

$$M = r^2_{y.x_1,x_2,\ldots,x_m} - \sum_{j=1}^{m}\left(r^2_{y.x_1,x_2,\ldots,x_m} - r^2_{y.x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m} \right) \qquad (243)$$

Minél nagyobb az *M* mutató értéke, annál jelentősebb a multikollinearitás, és ennek következtében a modell paramétereinek becslése mindinkább instabillá válik.
Megjegyzés: a (243) szerinti *M* mutató negatív értéket is felvehet.

Egy adott parciális ($\hat{\beta}_{yx_j . x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}$)[21] és a neki megfelelő kétváltozós regressziós együttható ($\hat{\beta}_{yx_j}$) összevetésével, az *M* mutató kiszámítása nélkül is, következtethetünk a szignifikáns multikollinearitás létére. Ugyanis, szignifikáns multikollinearitás esetén, az említett együtthatók között általában nem csak nagyságbeli, hanem még előjelbeli különbség is előfordulhat! Az említett kétfajta regressziós együttható részletesebb összefüggéseivel az **út-elemzési módszerek** foglalkoznak.

Út-elemzési módszerek

Ha egy modell magyarázóváltozói egymással is kapcsolatban vannak, akkor az eredményváltozóra nem csak direkt, hanem (közvetlen és közvetett) indirekt módon is hatnak. Ezeknek a hatásoknak a szemléltetésére használjuk az **út-diagramot**, amely (*n* elemű mintát feltételezve) a 47. ábrán látható.

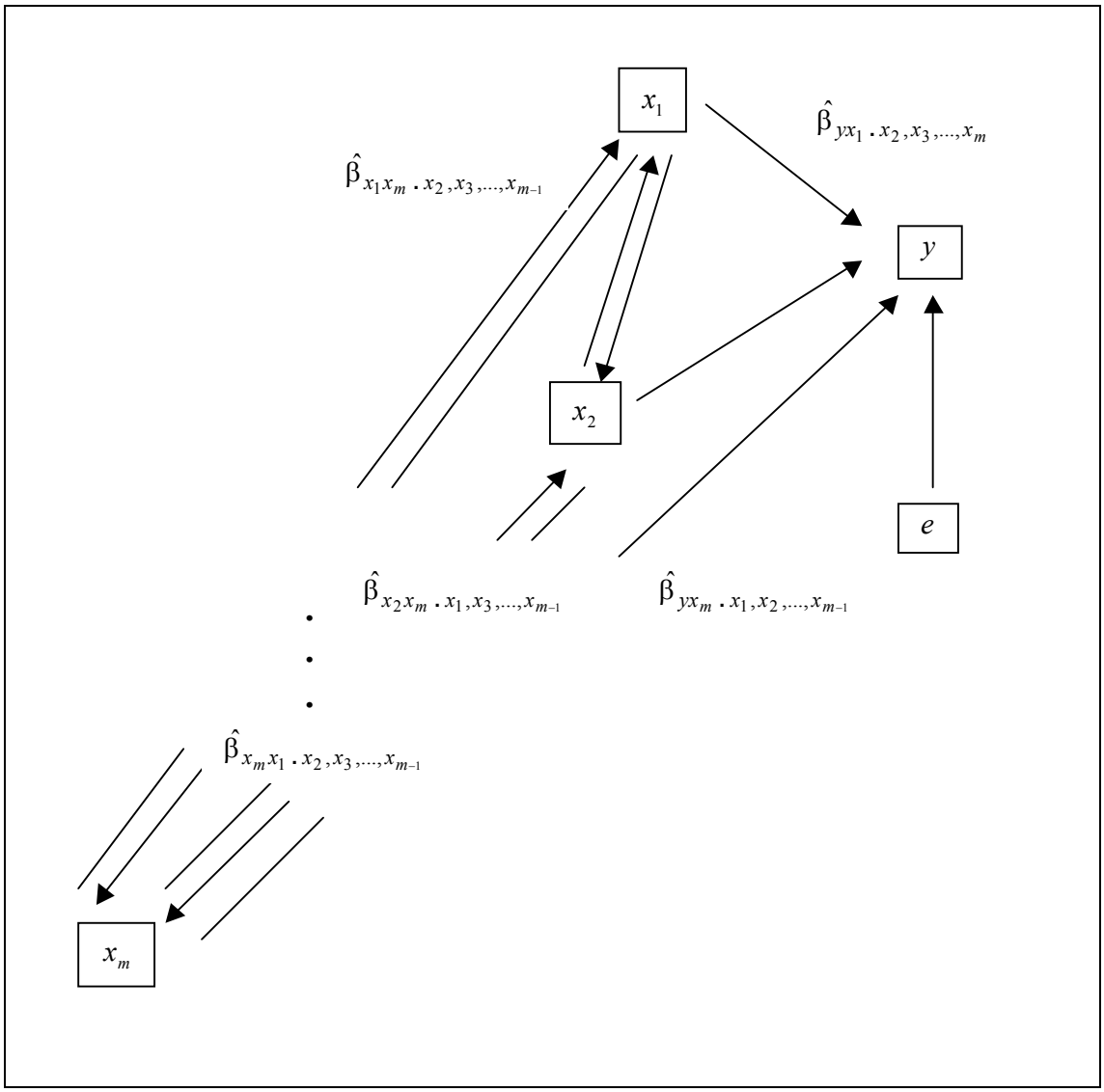
Négyváltozós modell esetén, például a második magyarázóváltozó teljes hatása az eredményváltozóra az alábbi.

	Hatások:
$\hat{\beta}_{yx_2} =$	teljes
$= \hat{\beta}_{yx_2 . x_1,x_3} +$	direkt
$+ \hat{\beta}_{x_1x_2 . x_3} \cdot \hat{\beta}_{yx_1 . x_2,x_3} + \hat{\beta}_{x_3x_2 . x_1} \cdot \hat{\beta}_{yx_3 . x_1,x_2} +$	közvetlen indirekt
$+ \hat{\beta}_{x_3x_2} \cdot \hat{\beta}_{x_1x_3 . x_2} \cdot \hat{\beta}_{yx_1 . x_2,x_3} + \hat{\beta}_{x_1x_2} \cdot \hat{\beta}_{x_3x_1 . x_2} \cdot \hat{\beta}_{yx_3 . x_1,x_2}$	közvetett indirekt

[21] Az eddigiektől eltérően, a könnyebb érthetőség végett, ebben a fejezetben az összetettebb jelölésmódot használjuk. A *j*-edik parciális együtthatót eddig $\hat{\beta}_j$, míg most $\hat{\beta}_{yx_j . x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}$ jelöli.

Megjegyzés: többváltozós modelleknél az áttételesebb indirekt hatások általában elhanyagolhatóak.

Út-diagram m magyarázóváltozót tartalmazó modell esetén



47. ábra

81. példa

A 90. táblázat a magyarországi állattenyésztés alakulását mutatja.

Számszerűsítsük a sertésállomány (közvetlen és közvetett) hatását a vágóállat-termelésre!

Állattenyésztés hazánkban 1974-1998 között

90. táblázat

Év	Vágóállat-termelés (ezer tonna)	Szarvasmarha-állomány (ezer db)	Sertésállomány (ezer db)	Baromfi-állomány (ezer db)
1974	1727	2017	8293	33154
1975	1898	1904	6953	38667
1976	1786	1887	7854	43449
1977	1958	1949	7850	43260
1978	2010	1966	8011	43294
1979	2032	1925	8355	41240
1980	2066	1918	8330	42764
1981	2079	1945	8296	42787
1982	2201	1922	9035	45397
1983	2319	1907	9844	41267
1984	2418	1901	9237	40962
1985	2307	1766	8280	38376
1986	2245	1725	8687	37176
1987	2339	1664	8216	36222
1988	2311	1690	8327	35607
1989	2260	1598	7660	34190
1990	2210	1571	8000	31121
1991	1976	1420	5993	28912
1992	1726	1159	5364	30535
1993	1513	999	5001	26542
1994	1405	910	4356	29847
1995	1402	928	5032	27549
1996	1499	909	5289	21062
1997	1394	871	4931	23419
1998	1428	873	5479	24082

Forrás: Magyar Statisztikai Évkönyv ’98, KSH, Bp., 1999.

Legyen a szarvasmarha- x_1, sertés- x_2 és a baromfiállomány x_3, a vágóállat-termelés pedig y. A feladat szerint meg kell határoznunk $\hat{\beta}_{yx_2}$ összetevőit az előbbiekben ismertetett módon. Ehhez még 5 regressziós modell paramétereit kell külön-külön kiszámítani.

A kapott eredmények vázlatos áttekintése az alábbi.

	Hatások:
$0{,}18809 =$	teljes
$= 0{,}23090 +$	direkt
$+ 0{,}13206 \cdot (-0{,}17157) + (-0{,}11769) \cdot (-0{,}00042) +$	közvetlen indirekt
$+ 3{,}71067 \cdot 0{,}02919 \cdot (-0{,}17157) + 0{,}24038 \cdot 15{,}92597 \cdot (-0{,}00042)$	közvetett indirekt

Ezek szerint a teljes hatáson belül a direkt hatásnak van a legnagyobb súlya, míg a közvetlen (-0,02261) és a közvetett (-0,02019) indirekt hatásoknak jóval kisebb.

A multikollinearitás következményeinek csökkentése, kiküszöbölése

- Ha célunk az előrejelzés és nem az együtthatók parciális vizsgálata, akkor a magyarázóváltozók lineáris függetlenségének hiánya nem okoz gondot.
- Nem teljes multikollinearitás esetén is megoldás lehet (néhány) magyarázóváltozó elhagyása a modellből, ha a közöttük fennálló kapcsolatok rendszere nem bonyolult.
- A modell újrafogalmazása, például TOBIN által alkalmazott módszer szerint.[22)]
- **Ridge-regresszió** alkalmazása.[23)]
- Főkomponens analízis alkalmazása. (Lásd a 11.5. fejezetet.)

Autokorreláció

Idősoros adatok vizsgálatánál a hibatagok egymást követő értékei gyakran korrelálnak. Ennek több oka lehet, általában specifikációs hibára vezethető vissza. Például, ha egy szignifikáns változót (amely értékei a statisztikai sorban egymástól nem függetlenek) figyelmen kívül hagyunk, akkor könnyen autokorrelált hibataghoz juthatunk.

---

22) A módszer lényege: a jövedelmi elaszticitások becslését keresztmetszeti, míg az árrugalmassági együtthatókat idősoros adatok alapján kapjuk.

23) A módszer az ismeretlen paraméterek becslésére (142) helyett az alábbi összefüggést alkalmazza:

$$\hat{\boldsymbol{\beta}}_a = (\mathbf{X}'\mathbf{X} + a\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}\,,$$

ahol az a önkényesen választott skalár (torzítási tényező). A módszer előnye, hogy szignifikáns multikollinearitás esetén is közvetlenül alkalmazható. Torzított becslést eredményez. A (0,1) intervallumban megfelelően választott a esetén azonban a becslés stabillá válik, és a (171) szerinti átlagos négyzetes hiba csökkenthető.

Az autokorreláció különböző rendű lehet, attól függően, hogy a hibatag i-edik értéke melyik értékkel van kapcsolatban. Ha a hibatag i-edik értéke az $(i-1)$-edik értékkel (tehát a közvetlenül előtte levő értékkel) áll korrelációs kapcsolatban, akkor **elsőrendű autokorrelációról**[24] beszélünk. (Könyvünkben csak ezzel az esettel foglalkozunk.)

Az elsőrendű autokorrelációnak megfelelő modell a következő:

$$\varepsilon_i = \rho \cdot \varepsilon_{i-1} + \eta_i ,$$

ahol ρ az **autokorrelációs együttható**.

Az η valószínűségi változóra igazak az alábbiak.

$$E(\eta_i) = 0$$

$$E(\boldsymbol{\eta}\boldsymbol{\eta}') = \mathrm{var}(\eta) \cdot \mathbf{I}$$

$$\mathrm{var}(\varepsilon_i) = \frac{\mathrm{var}(\eta)}{1-\rho^2}$$

Megjegyzés: az ismertetett modell éves idősorok alapján történő elemzéseknél általában jól alkalmazható.

Az autokorreláció következményei

Ha a hibatagok között szignifikáns lineáris kapcsolat van, akkor az LNM közvetlen alkalmazásával kapott becslések fontosabb tulajdonságai az alábbiak.

- A becslés és az előrejelzés torzítatlan marad.
- A regressziós együtthatók becslése nem efficiens.
- A reziduális szórásnégyzet a hibatag szórásnégyzetének torzított becslését adja, ezért az F-próbák nem alkalmazhatóak.

[24] A szakirodalomban ezekre gyakran AR(1) jelöléssel hivatkozunk, ahol az AR az autoregresszióra utal. AR(2) a másodrendű autokorrelációt jelöli, stb.

Az elsőrendű autokorreláció tesztelése

Az elsőrendű autokorreláció tesztelésére a **DURBIN-WATSON-féle próbát** fogjuk alkalmazni. Ennek próbafüggvénye a (244) képlet szerint definiált.

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \qquad (244)$$

ahol az e_i az LNM alkalmazásával kapott reziduumok, amelyeket a hibatagok becslésének tekinthetünk.

A ρ autokorrelációs együttható értékét, (98) figyelembevételével, az alábbiak szerint becsüljük.

$$\hat{\rho} = \frac{\sum_{i=2}^{n} e_i \cdot e_{i-1}}{\sqrt{\sum_{i=2}^{n} e_i^2} \cdot \sqrt{\sum_{i=2}^{n} e_{i-1}^2}}$$

Mivel $\sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2$, a megfelelő műveletek elvégzése után, (244) az alábbi alakra hozható.

$$d \approx 2(1 - \hat{\rho}) \qquad (245)$$

Az elsőrendű autokorreláció tesztelésekor, a (245) szerinti összefüggést figyelembe véve, a 91. táblázatban feltüntetett relációk alapján döntünk.

Nullhipotézisünk tehát az elsőrendű autokorreláció hiánya ($H_0: \rho = 0$). Amennyiben a próbafüggvényünk értéke 2-nél nagyobb, akkor alternatív hipotézisünk a negatív autokorreláció ($H_1: \rho < 0$), amennyiben 2-nél kisebb, akkor a pozitív autokorreláció ($H_1: \rho > 0$).

A kritikus értékek meghatározásához szükséges alsó (d_L) és felső (d_U) értékeket a VIII. és IX. táblázat tartalmazza (a megfigyelések száma és a magyarázóváltozók számának függvényében).

Megjegyzés: a megfelelő táblázati értékek forrása Savin, N. E. – White, K. J.: The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes of Many Regressors, Econometrica, 45, Nov. 1977.

DURBIN-WATSON-féle teszt döntési táblája

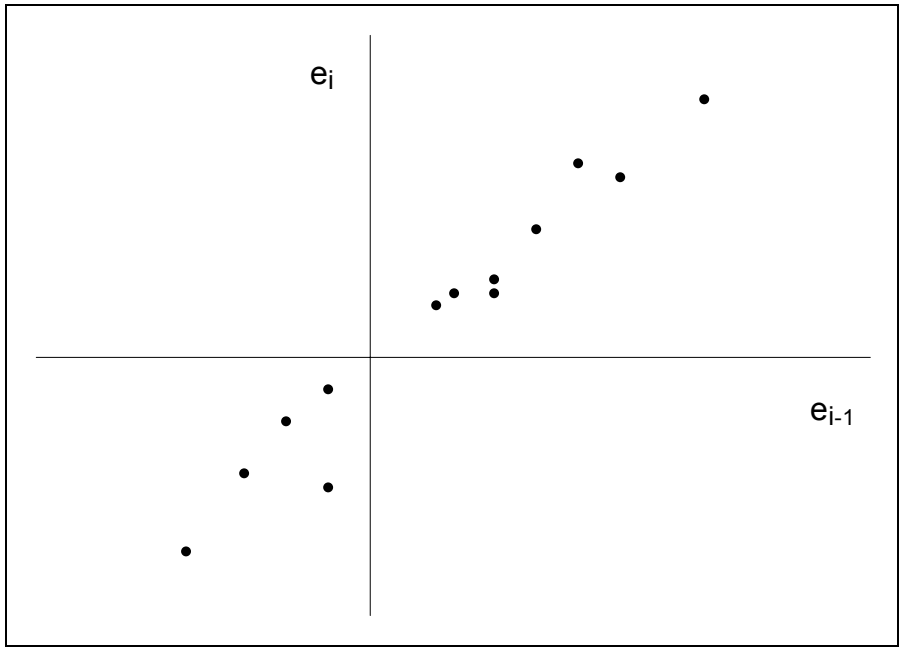
91. táblázat

Alternatív hipotézis	$H_0: \rho = 0$		
	Elfogadjuk	Elvetjük	Nincs döntés
$\rho > 0$	$d > d_U$	$d < d_L$	$d_L \le d \le d_U$
$\rho < 0$	$d < 4 - d_U$	$d > 4 - d_L$	$4 - d_U \le d \le 4 - d_L$

Abban az esetben, ha az autokorreláltságra vonatkozóan a teszt alapján nem tudunk döntést hozni, akkor a modell paramétereinek becslését újból el kell végezni, de most már több megfigyelést tartalmazó minta alapján!

Megjegyzés: empirikus elemzések alkalmával hasznos grafikusan ábrázolni az egymást követő reziduumok értékeit egy olyan grafikonon, amelynél az abszcissza-tengelyen az e_{i-1}, míg az ordináta-tengelyen az e_i értékeket tüntetjük fel, ahogy az például a 48. ábrán látható. A kapott pontdiagram alapján általában már következtetni tudunk az esetleges autokorreláció jellegére.

A reziduumok grafikus ábrázolása



48. ábra

Az autokorreláció kezelése

- A regressziós modell funkcionális operátorának megváltoztatása.
- Az általánosított legkisebb négyzetek módszerének alkalmazása. (Lásd a 11.4. fejezetet.)
- Általánosabb dinamikus modell megadása. (Könyvünkben ezekkel nem foglalkozunk.)

Heteroszkedaszticitás

Míg az idősoros adatoknál az autokorreláció okoz legtöbbször gondot, a keresztmetszeti adatok esetében gyakran a hibatagok varianciái (a standard lineáris regressziós modell feltételrendszerétől eltérően) nem állandóak. Ennek általában az az oka, hogy a hibatag nagysága függ valamelyik változótól.

A heteroszkedaszticitás következményei

Ha a hibatagok varianciái nem állandóak, akkor az LNM közvetlen alkalmazásával kapott becslések fontosabb tulajdonságai az alábbiak.

- A becslés és az előrejelzés torzítatlan marad.
- A regressziós együtthatók becslése nem efficiens.
- Az F-próbák nem alkalmazhatóak.

A heteroszkedaszticitás tesztelése

Empirikus elemzéseknél azt kell megvizsgálnunk, hogy milyen szoros a kapcsolat az egyes változók és a hibatagok (a gyakorlatban a reziduumok) abszolút értékei között.

Ha a minta n elemű, akkor a feltételezésünknek megfelelő modell az alábbi.

$$E(e_i^2) = \text{var}(e_i) \cdot x_{ij}^2$$

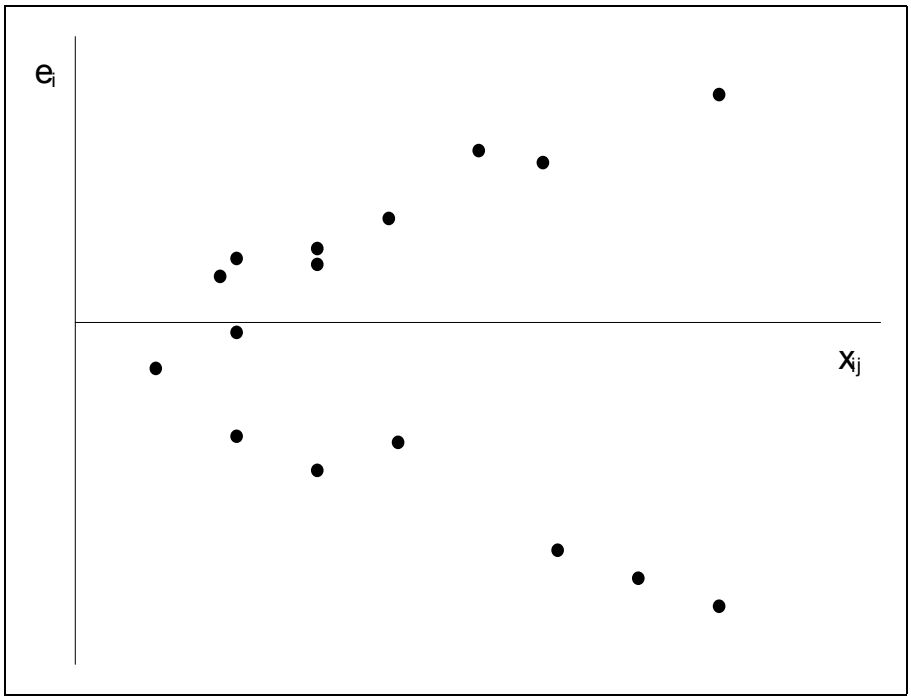
A heteroszkedaszticitás teszteléséhez a (242) próbafüggvényt használjuk. Külön-külön kiszámítjuk az egyes magyarázóváltozóknak, illetve a becsült eredményváltozónak a reziduumok abszolút értékeivel való szorosságát jellemző lineáris korrelációs együtthatót, és ezek közül a legnagyobb abszolút értékű együtthatót teszteljük. Amennyiben a nullhipotézist ($r = 0$) elvetjük, a modell heteroszkedasztikusnak tekinthető.

Az autokorrelációhoz hasonlóan, az esetleges heteroszkedaszticitás vizsgálatakor is célszerű a grafikus ábrázolás. A vizsgált változó rendelkezésünkre álló adatait felvisszük az abszcissza-tengelyre, a reziduumok érétkeit pedig az ordináta-tengelyre. Heteroszkedaszticitás esetén a pontdiagramon összetartó vagy széttartó pontfelhőt kapunk, ahogy az például a 49. ábrán látható.

A heteroszkedaszticitás kezelése

- Az általánosított legkisebb négyzetek módszere ebben az esetben is alkalmazható. (Lásd a 11.4. fejezetet.)

A heteroszkedasztikus reziduumok grafikus ábrázolása a j-edik magyarázóváltozó függvényében



49. ábra

A továbbiakban bemutatjuk az eddig ismertetett regresszió- és korrelációszámítással kapcsolatos elméleti összefüggéseket egy, az eddigiektől némileg összetettebb, valós példán keresztül.

82. példa

A szennyvízcsatorna- és az ivóvízvezeték-hálózat területi egységenkénti adatait 1998. évre vonatkozóan a 92. táblázat tartalmazza.

Az adatok jelölésére vezessük be a következő szimbólumokat:

- y_i : szennyvízcsatorna-hálózat hossza (m/lakos),
- x_{i1} : ivóvízvezeték-hálózat hossza (m/lakos),
- x_{i2} : száz lakásra jutó lakosok száma.

Lineáris modellt feltételezve, ellenőrizzük a standard regressziós modell feltételeinek teljesülését! Értelmezzük a kapott eredményeket! Vizsgáljuk a modellünk illeszkedésének jóságát, valamint értelmezzük és teszteljük a parciális regressziós

együtthatókat!

A szennyvízcsatorna- és az ivóvízvezeték-hálózat területi egységenként, 1998

92. táblázat

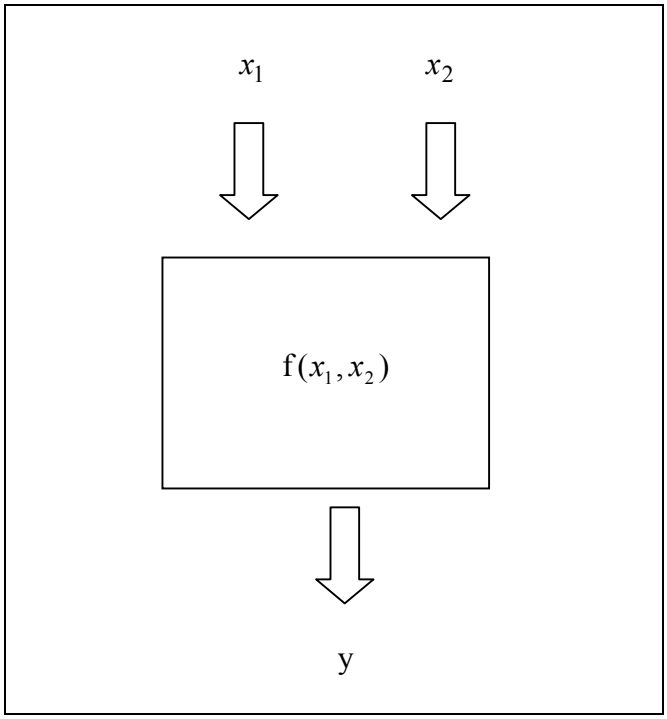
Megye	Szennyvíz-csatorna-hálózat hossza (m/lakos)	Ivóvízvezeték-hálózat hossza (m/lakos)	Száz lakásra jutó lakosok száma (fő)
	y_i	x_{i1}	x_{i2}
Bács-Kiskun	1,073	5,865	236
Baranya	2,303	7,308	258
Békés	1,501	7,871	237
Borsod-Abaúj-Zemplén	1,735	6,518	261
Csongrád	1,355	5,452	230
Fejér	2,136	6,577	269
Győr-Moson-Sopron	3,512	6,163	265
Hajdú	1,289	5,007	258
Heves	1,981	6,485	245
Jász-Nagykun-Szolnok	2,205	7,118	246
Komárom-Esztergom	2,765	5,897	261
Nógrád	1,248	9,587	246
Pest	2,529	7,038	273
Somogy	2,217	9,943	251
Szabolcs-Szatmár-Bereg	1,762	6,684	275
Tolna	1,649	6,967	252
Vas	2,067	6,858	261
Veszprém	2,675	9,288	260
Zala	2,618	7,358	254

Forrás: Magyar Statisztikai Zsebkönyv ’98, KSH, Bp., 1999.

Első lépésként az 50. ábrán megadjuk a bemeneti (okok) és a kimeneti adatok (okozat) grafikus modelljét. Az ezeket összekötő funkcionális operátor identifikálása végett alkalmazzuk az LNM-t a (227) alatt definiált modellünkre. A feladatnak megfelelő becslőfüggvény alapján

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + e_i \qquad i = 1,2,...,19\,.$$

A regressziós modell grafikus ábrája



50. ábra

Mielőtt elvégeznénk a modell paramétereinek becslését, vizsgáljuk meg, hogy teljesül-e a standard lineáris regressziós modell feltételrendszere.

Mindenekelőtt ellenőrizzük a magyarázóváltozók (egymástól való) lineáris függetlenségét. Számítsuk ki a (238) alatt definiált korrelációs mátrixot, amelynél a páronkénti korrelációs együtthatókhoz a (98) szerint juthatunk. Az Excel segítségével azonban, a korábbiakban már ismertetett módon, közvetlenül megkaphatjuk a mátrixot.

$$\mathbf{R} = \begin{bmatrix} 1{,}000 & 0{,}110 & 0{,}538 \\ 0{,}110 & 1{,}000 & -0{,}034 \\ 0{,}538 & -0{,}034 & 1{,}000 \end{bmatrix}$$

Mivel a mátrix főátlón kívüli elemei nagyrészt 0-hoz közeli értékek, nem következtetünk szignifikáns multikollinearitásra. Ezt a sejtésünket kétféleképpen ellenőrizzük.

A (243) képlet szerinti M mutató kiszámításához, mivel most háromdimenziós

modellről van szó, a többszörös determinációs együttható mellett a megfelelő páronkénti lineáris korrelációs együtthatókra van szükség.

Ezeket a korrelációs mátrix tartalmazza.

$$r_{yx_1} = 0{,}110$$

$$r_{yx_2} = 0{,}538$$

A többszörös determinációs együtthatót a (241) képlet szerint az $\mathbf{R}^{-1}$ mátrix segítségével tudjuk kiszámítani.

$$\mathbf{R}^{-1} = \begin{bmatrix} 1{,}441 & -0{,}185 & -0{,}782 \\ -0{,}185 & 1{,}025 & 0{,}135 \\ -0{,}782 & 0{,}135 & 1{,}425 \end{bmatrix}$$

A többszörös determinációs együttható értéke:

$$r^2_{y.x_1,x_2} = 1 - \frac{1}{1{,}441} = 0{,}306\,.$$

Ez azt jelenti, hogy az eredményváltozó szórásnégyzetének 30,6 százalékát tudjuk megmagyarázni az x_1, x_2 magyarázóváltozókkal.
A megfelelő adatok behelyettesítésével:

$$M = 0{,}306 - ((0{,}306 - 0{,}110^2) + (0{,}306 - 0{,}538^2)) = -0{,}0045\,.$$

Az M mérőszám 0-hoz közeli értéke is alátámasztja a magyarázóváltozók lineáris függetlenségét.

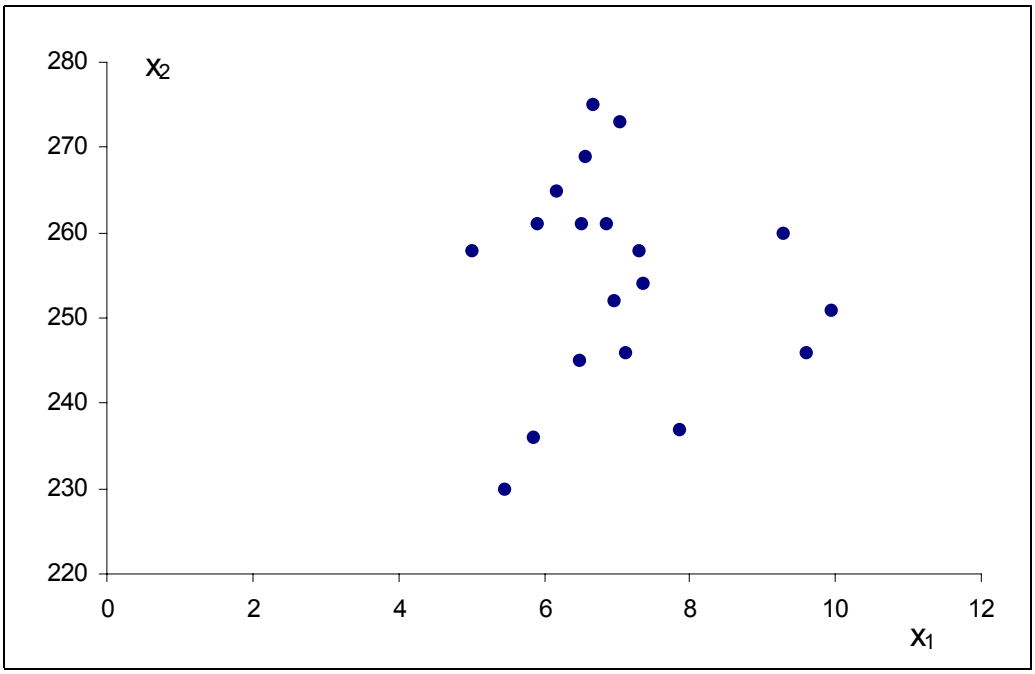
A két magyarázóváltozó kapcsolatának szorosságát tesztelhetjük a (242) próbafüggvény segítségével is.

$$t = \frac{-0{,}034\sqrt{17}}{\sqrt{1 - 0{,}0012}} = -0{,}140\,.$$

Kétoldali próbához ($\alpha = 0{,}05$ és $\nu = 17$ esetén) az elméleti t érték a III. táblázat szerint $2{,}1098$. Az empirikus $t = -0{,}140$ abszolút értéke kisebb az elméleti értéknél, ezért a nullhipotézist 5%-os szignifikancia-szinten elfogadjuk, ami a magyarázóváltozók lineáris függetlenségére utal.

Ugyanerre a következtetésre juthatunk a két magyarázóváltozó grafikus ábrázolásával is. Az 51. ábrán látható, hogy a pontok elrendeződése véletlenszerű.

A magyarázóváltozók pontdiagramja



51. ábra

Megjegyzés: elméletileg minden olyan esetben, amikor két magyarázóváltozó (például x_1 és x_2) lineárisan független egymástól, akkor az $x_1(x_2)$ és az $x_2(x_1)$ kétváltozós lineáris regressziós egyenesek (ugyanazon a diagramon ábrázolva) derékszögben metszik egymást.

A multikollinearitás után teszteljük az autokorrelációra vonatkozó nullhipotézisünket. Ehhez szükségünk van a reziduumokra.

Ha a mátrixalgebrai jelölésmódot alkalmazzuk, akkor felírhatjuk a következő összefüggést:

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e},$$

illetve, figyelembe véve a 92. táblázatban közölt adatokat és a (228) szerinti jelölésmódot, a következő mátrixegyenletet kapjuk:

$$\begin{bmatrix} 1{,}073 \\ 2{,}303 \\ \vdots \\ 2{,}618 \end{bmatrix} = \begin{bmatrix} 1 & 5{,}865 & 236 \\ 1 & 7{,}308 & 258 \\ \vdots & & \\ 1 & 7{,}358 & 254 \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{19} \end{bmatrix}.$$

Az ismeretlen $\boldsymbol{\beta}$ oszlopvektorának (142) szerinti becsléséhez szükségünk van a következő számításokra:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 5{,}865 & 7{,}308 & & 7{,}358 \\ 236 & 258 & & 254 \end{bmatrix} \cdot \begin{bmatrix} 1 & 5{,}865 & 236 \\ 1 & 7{,}308 & 258 \\ \vdots & & \\ 1 & 7{,}358 & 254 \end{bmatrix} =$$

$$= \begin{bmatrix} 19{,}000 & 133{,}984 & 4838{,}000 \\ 133{,}984 & 976{,}733 & 34106{,}376 \\ 4838{,}00 & 34106{,}376 & 1234674{,}000 \end{bmatrix};$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 25{,}4925 & -0{,}2507 & -0{,}0930 \\ -0{,}2507 & 0{,}0314 & 0{,}0001 \\ -0{,}0930 & 0{,}0001 & 0{,}0004 \end{bmatrix};$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 5{,}865 & 7{,}308 & & 7{,}358 \\ 236 & 258 & & 254 \end{bmatrix} \cdot \begin{bmatrix} 1{,}073 \\ 2{,}303 \\ \vdots \\ 2{,}618 \end{bmatrix} = \begin{bmatrix} 38{,}620 \\ 273{,}985 \\ 9908{,}839 \end{bmatrix};$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 25{,}4925 & -0{,}2507 & -0{,}0930 \\ -0{,}2507 & 0{,}0314 & 0{,}0001 \\ -0{,}0930 & 0{,}0001 & 0{,}0004 \end{bmatrix} \cdot \begin{bmatrix} 38{,}620 \\ 273{,}985 \\ 9908{,}839 \end{bmatrix} = \begin{bmatrix} -5{,}359 \\ 0{,}060 \\ 0{,}027 \end{bmatrix}.$$

A fenti mátrixműveletek könnyen elvégezhetőek az Excel segítségével a következő függvények alkalmazásával: TRANSZPONÁLÁS(tömb), MSZORZAT(tömb1;tömb2), INVERZ.MÁTRIX(tömb). Ezek eredménye tömb lesz, ezért ki kell jelölnünk egy megfelelő nagyságú cellatartományt (ahova az eredménytömböt várjuk), majd a függvény beillesztése után a szerkesztőlécre állva a SHIFT, a CTRL és az ENTER billentyűk együttes lenyomása után a kijelölt cellatartományban megkapjuk a keresett mátrixot.

A becsült paraméterek oszlopvektora segítségével, (141) szerint, a szennyvízcsatorna-hálózat hosszának becsült értékeire felírhatjuk a következő mátrixegyenletet:

$$\begin{bmatrix} 1{,}452 \\ 2{,}140 \\ \vdots \\ 2{,}034 \end{bmatrix} = \begin{bmatrix} 1 & 5{,}865 & 236 \\ 1 & 7{,}308 & 258 \\ \vdots & & \\ 1 & 7{,}358 & 254 \end{bmatrix} \cdot \begin{bmatrix} -5{,}349 \\ 0{,}060 \\ 0{,}027 \end{bmatrix}.$$

Az autokorreláció teszteléséhez szükséges adatokat a 93. táblázat tartalmazza.

A (244) képlet szerinti próbafüggvény:

$$d = \frac{11{,}680}{4{,}871} = 2{,}398\,.$$

A (245) képlet alapján az autokorrelációs együttható becslése:

$$\hat{\rho} \approx 1 - \frac{d}{2} = -0{,}199.$$

A kapott eredmények alapján az alternatív hipotézisünk a negatív autokorreláció.

A VIII. táblázat szerint 5%-os szignifikancia-szint mellett $d_U = 1{,}536$ és

$d = 2{,}398 < 4 - d_U = 2{,}464$; ezért a DURBIN-WATSON-féle próba nullhipotézisét elfogadjuk, tehát a hibatagok nem autokorreláltak.

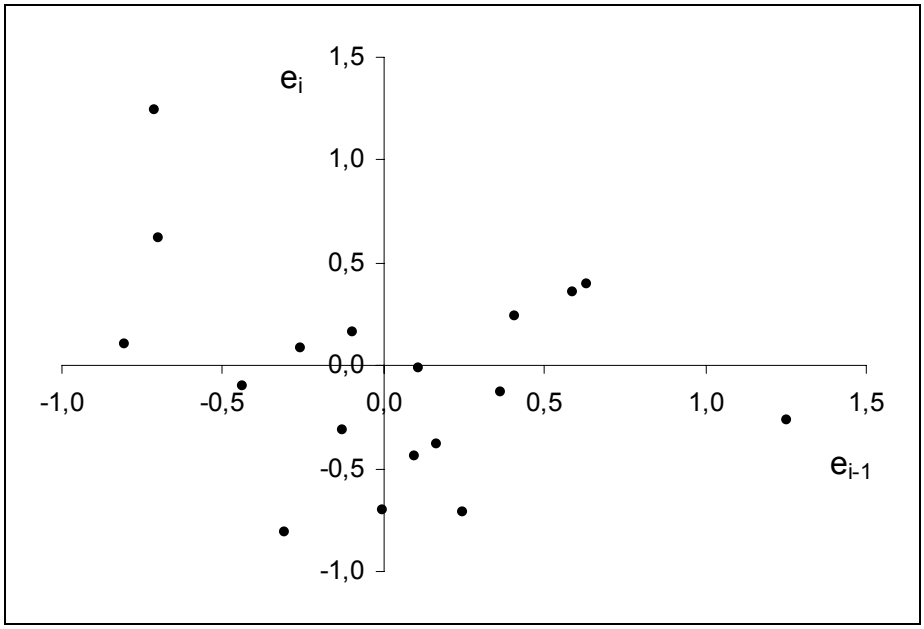
A regressziófüggvény becsült értékei és a reziduumok

93. táblázat

Megye	y_i	$\hat{y}_i$	e_i	e_i^2	e_{i-1}	$(e_i - e_{i-1})^2$	$\lvert e_i \rvert$
Bács-K.	1,073	1,452	-0,379	0,144	–	–	0,379
Baranya	2,303	2,140	0,163	0,027	-0,379	0,294	0,163
Békés	1,501	1,600	-0,099	0,010	0,163	0,069	0,099
BAZ	1,735	2,174	-0,439	0,193	-0,099	0,116	0,439
Csongrád	1,355	1,263	0,092	0,008	-0,439	0,282	0,092
Fejér	2,136	2,397	-0,261	0,068	0,092	0,124	0,261
GYMS	3,512	2,262	1,250	1,562	-0,261	2,281	1,250
Hajdú	1,289	2,001	-0,712	0,508	1,250	3,850	0,712
Heves	1,981	1,735	0,246	0,060	-0,712	0,918	0,246
JNSZ	2,205	1,801	0,404	0,163	0,246	0,025	0,404
KE	2,765	2,137	0,628	0,394	0,404	0,050	0,628
Nógrád	1,248	1,950	-0,702	0,492	0,628	1,768	0,702
Pest	2,529	2,534	-0,005	0,000	-0,702	0,486	0,005
Somogy	2,217	2,108	0,109	0,012	-0,005	0,013	0,109
SZSZB	1,762	2,567	-0,805	0,648	0,109	0,836	0,805
Tolna	1,649	1,956	-0,307	0,094	-0,805	0,248	0,307
Vas	2,067	2,195	-0,128	0,016	-0,307	0,032	0,128
Veszp.	2,675	2,314	0,361	0,130	-0,128	0,239	0,361
Zala	2,618	2,034	0,584	0,341	0,361	0,050	0,584
Összesen	38,620	38,620	0,000	4,871	-0,584	11,680	–

Megjegyzés: ugyanerre a következtetésre juthatunk a reziduumok és a késleltetett reziduumok grafikus ábrázolásával is. Az 52. ábrán látható, hogy a pontok elrendeződése véletlenszerű.

A reziduumok grafikus ábrázolása



52. ábra

A heteroszkedaszticitás vizsgálatához a reziduumok abszolút értékei és az egyes változók értékei közötti lineáris korrelációs együtthatót számítjuk ki.

$$r_{|e|\hat{y}} = 0{,}249$$

$$r_{|e|x_1} = -0{,}200$$

$$r_{|e|x_2} = 0{,}302$$

Ezek közül a legnagyobb abszolút értékű az $r_{|e|x_2} = 0{,}302$. Annak tesztelését kell elvégeznünk, hogy ez szignifikánsan különbözik-e 0-tól.

A (242) próbafüggvényt használjuk:

$$t = \frac{0{,}302\sqrt{17}}{\sqrt{1-0{,}091}} = 1{,}306.$$

Kétoldali próbához ($\alpha = 0{,}05$ és $\nu = 17$ esetén) az elméleti t érték a III. táblázat szerint $2{,}1098$. Az empirikus $t = 1{,}306$ érték az elfogadási tartományba esik, ezért a

nullhipotézist 5%-os szignifikancia-szinten elfogadjuk, ami a hibatagok homoszkedaszticitására utal.

Megjegyzés: ugyanerre a következtetésre juthatunk az egyes változók és a reziduumok grafikus ábrázolásával is. Az 54. ábrán látható, hogy a pontok elrendeződése véletlenszerű.

Az eddigi elemzések eredményeinek figyelembevételével megállapíthatjuk, hogy a standard lineáris regressziós modell alkalmazható.

A lineáris háromváltozós regressziófüggvény tehát:

$$\hat{y}_i = -5{,}349 + 0{,}060 \cdot x_{i1} + 0{,}027 \cdot x_{i2}\,.$$

A parciális regressziós együtthatókat a következőképpen értelmezhetjük:

$\hat{\beta}_1 = 0{,}060$ azt jelenti, hogy az ivóvízvezeték-hálózat egy lakosra jutó hosszának 1 méterrel történő növekedése a szennyvízcsatorna-hálózat egy lakosra jutó hosszának átlagosan 0,060 méteres növekedésével jár együtt, ha a száz lakásra jutó lakosok száma nem változik.

$\hat{\beta}_2 = 0{,}027$ azt jelenti, hogy a száz lakásra jutó lakosok számának 1 fővel történő növekedése a szennyvízcsatorna-hálózat egy lakosra jutó hosszának átlagosan 0,027 méteres növekedésével jár együtt, ha az ivóvízvezeték-hálózat egy lakosra jutó hossza nem változik.

Empirikus elemzéseknél, a trendfüggvény megadásához hasonlóan, nem elegendő pusztán a funkcionális operátor közlése, hanem e mellett még a következő adatokat is ajánlatos feltüntetni: a többszörös determinációs együttható értéke, a globális F-próba értéke, a regressziós paraméterek standard hibájának értékei, a parciális F-próba értékei, az autokorreláció tesztelésénél alkalmazott *d* statisztika értéke, a heteroszkedaszticitás teszteléséhez szükséges (legnagyobb) lineáris korrelációs együttható értéke és a korrelációs mátrix.

A többszörös determinációs együttható $r^2_{y.x_1,x_2} = 0{,}306$ értéke arra utal, hogy modellünk nem jól illeszkedik az empirikus adatokra. Az objektív következtetéshez alkalmazzuk a globális F-próbát.

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \beta_j \neq 0 \qquad \textit{valamelyik } j\textit{-re} \qquad j = 1, 2$$

A fenti nullhipotézis helyességének ellenőrzésére a (234) szerint definiált próbafüggvényt használjuk. Eredményeinket ANOVA táblázatba foglaljuk.

Az ANOVA táblázat

94. táblázat

A szóródás oka	Eltérések négyzetösszege	Szabadságfok	Szórásnégyzet becslése	F
Regresszió	2,147	2	1,074	3,527
Hiba	4,871	16	0,304	
Összesen	7,018	18	–	

5%-os szignifikancia-szint mellett az elméleti F érték: $F_{0,95}(2,16) = 3{,}634$. Mivel a próbafüggvény értéke kisebb ennél, a nullhipotézist nem vethetjük el.

A regressziós paraméterek teszteléséhez szükségünk van a paraméterek standard hibáira. Ennek kiszámítása a (236) képlet szerint történhet. (A reziduumok értékeit, illetve négyzetösszegüket a 93. táblázat tartalmazza.)

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \frac{4{,}871}{16} \cdot \begin{bmatrix} 25{,}4925 & -0{,}2507 & -0{,}0930 \\ -0{,}2507 & 0{,}0314 & 0{,}0001 \\ -0{,}0930 & 0{,}0001 & 0{,}0004 \end{bmatrix} = \begin{bmatrix} 7{,}76070 & -0{,}07632 & -0{,}02830 \\ -0{,}07632 & 0{,}00955 & 0{,}00004 \\ -0{,}02830 & 0{,}00004 & 0{,}00011 \end{bmatrix}$$

Innen a főátlóban levő elemek négyzetgyökei adják a keresett standard hibákat.

$$s_{\hat{\beta}_0} = 2{,}786$$

$$s_{\hat{\beta}_1} = 0{,}098$$

$$s_{\hat{\beta}_2} = 0{,}010$$

A parciális F-teszt próbafüggvényének (237) szerinti értékei:

$$t_{\hat{\beta}_0} = -1{,}920;$$

$$t_{\hat{\beta}_1} = 0{,}617;$$

$$t_{\hat{\beta}_2} = 2{,}603.$$

Kétoldali próbához ($\alpha = 0{,}05$ és $\nu = 16$ esetén) az elméleti t érték a III. táblázat szerint $2{,}1199$. Mivel $\left|t_{\hat{\beta}_1}\right| = 0{,}617 < 2{,}1199$, ez azt jelenti, hogy x_1 szignifikánsan nem befolyásolja az eredményváltozót.

A $\left|t_{\hat{\beta}_2}\right| = 2{,}603 > 2{,}1199$; így az x_2 magyarázóváltozót (a száz lakásra jutó lakosok számát) célszerű a modellben szerepeltetni.

Az egy lakosra jutó szennyvízcsatorna-hálózat hosszát számszerűsítő statisztikai modellt az alábbi formában közölhetjük.

$$\begin{array}{lllll} \hat{y}_i = -5{,}349 & + & 0{,}060 \cdot x_{i1} + & 0{,}027 \cdot x_{i2} & r^2_{y.x_1,x_2} = 0{,}306 \\ \quad (2{,}786) & & (0{,}098) & (0{,}010) & F = 3{,}527 \\ t = -1{,}920 & & t = 0{,}617 & t = 2{,}603 & \end{array}$$

$$\begin{array}{ll} r_{x_1x_2} = -0{,}034 & M = -0{,}0045 \\ d = 2{,}398 & 4 - d_U = 2{,}464 \\ r_{|e|x_2} = 0{,}302 & t = 1{,}306 \end{array}$$

Megjegyzés: regressziószámítás esetén, a modell becsült paraméterei mellett, célszerű közölni (a fentiekhez hasonlóan) az elemzés többi eredményét is.

A kapott eredmények nagy részét az Excel segítségével is kiszámíthatjuk a 6.1. fejezetben ismertetett módon.

Az eredményeket az 53. és az 54. ábrán láthatjuk.

Az Excel outputja

ÖSSZESÍTŐ TÁBLA

Regressziós statisztika	
r értéke	0,553
r-négyzet	0,306
Korrigált r-négyzet	0,219
Standard hiba	0,552
Megfigyelések	19

VARIANCIAANALÍZIS

	df	*SS*	*MS*	*F*	*F szignifikanciája*
Regresszió	2	2,147	1,074	3,527	0,054
Maradék	16	4,871	0,304		
Összesen	18	7,018			

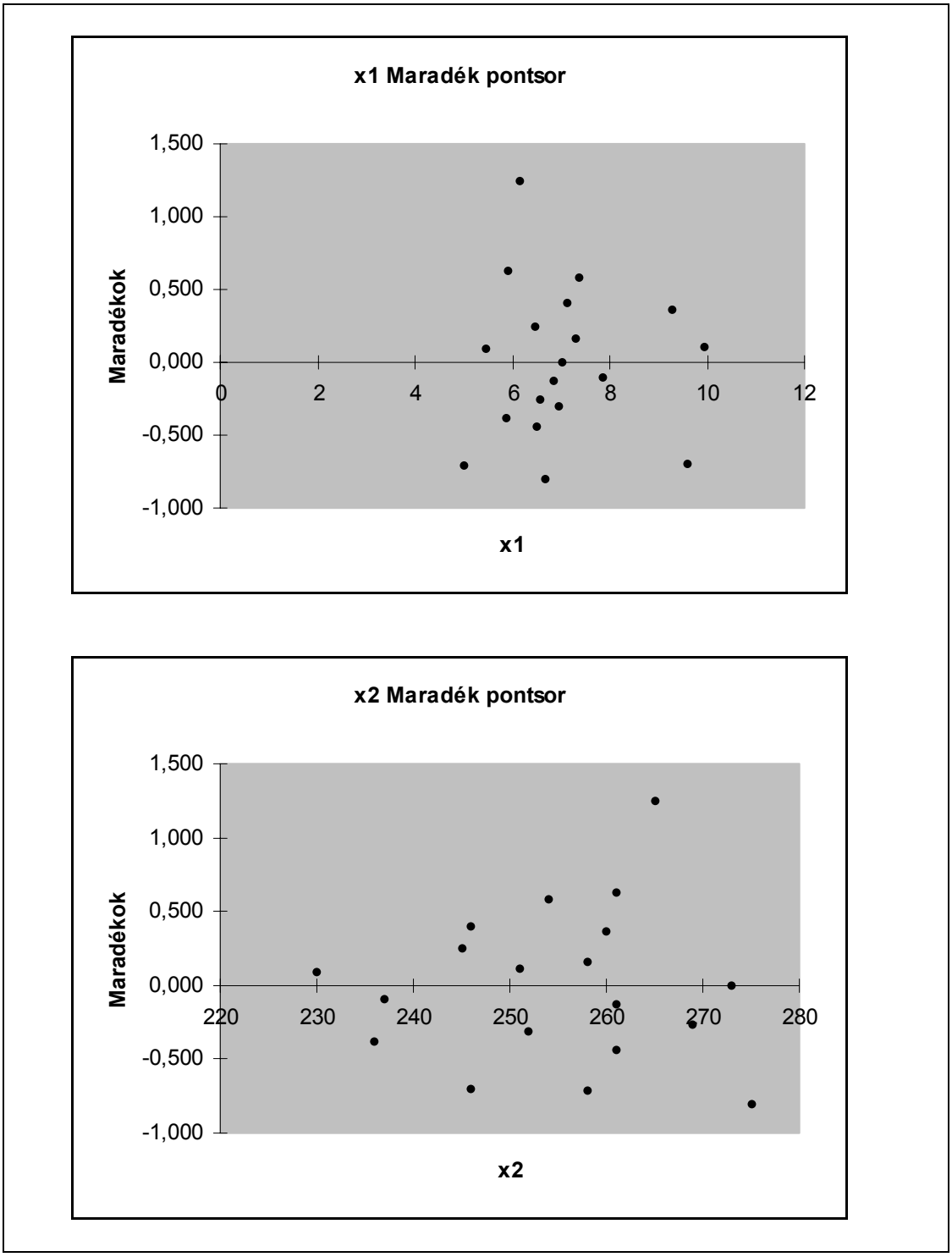
	Koefficiensek	*Standard hiba*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*	*Alsó 95,0%*	*Felső 95,0%*
Tengely metszet	-5,349	2,786	-1,920	0,073	-11,255	0,556	-11,255	0,556
x1	0,060	0,098	0,617	0,546	-0,147	0,267	-0,147	0,267
x2	0,027	0,010	2,603	0,019	0,005	0,050	0,005	0,050

MARADÉK TÁBLA

Megfigyelés	*Becsült y*	*Maradékok*
1	1,452	-0,379
2	2,140	0,163
3	1,600	-0,099
4	2,174	-0,439
5	1,263	0,092
6	2,397	-0,261
7	2,262	1,250
8	2,001	-0,712
9	1,735	0,246
10	1,801	0,404
11	2,137	0,628
12	1,950	-0,702
13	2,534	-0,005
14	2,108	0,109
15	2,567	-0,805
16	1,956	-0,307
17	2,195	-0,128
18	2,314	0,361
19	2,034	0,584

53. ábra

Az Excel outputja (folytatás)



54. ábra

Fontossága miatt még egyszer kiemeljük, hogy az empirikus elemzéseknél a (142) képletet nem szabad automatikusan alkalmazni, illetve a kapott eredményeket a standard lineáris regressziós modell feltételrendszerére vonatkozó ellenőrzések nélkül felhasználni!

A lehetséges hibák elkerülése végett a következő algoritmust célszerű követni:

- először a korrelációs mátrix segítségével ellenőrizzük a magyarázóváltozók lineáris függetlenségét. Így (esetleges) szignifikáns multikollinearitás esetén dönthetünk a modellbe vett magyarázóváltozók szerepeltetéséről;
- az eredményváltozó empirikus és becsült értékei segítségével teszteljük a reziduumok lineáris függetlenségét. Így (esetleges) szignifikáns (elsőrendű) autokorreláció esetén dönthetünk az adott modell alkalmazhatóságáról;
- ellenőrizzük a reziduumok szórásnégyzetének állandóságára vonatkozó feltevést. Így (esetleges) szignifikáns heteroszkedaszticitás esetén szintén dönthetünk az adott modell alkalmazhatóságáról.

Mivel az ivóvízvezeték-hálózat egy lakosra jutó hosszának (x_1 változó) magyarázó ereje nem bizonyult szignifikánsnak, ezért a modellünkből elhagyjuk, és csak a száz lakásra jutó lakosok számát (x_2 változó) hagyjuk az új modellben, amely becslése (általánosan) a következő alakban is felírható:

$$\hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot x_{i2} \qquad i = 1,2,...,19 .$$

A 92. táblázat y_i és x_{i2} adatai alapján a fenti kétváltozós lineáris modell becsült paramétereit a 6.1. fejezetben ismertetett módon tudjuk kiszámítani, vagy a (142) képlet alkalmazásával, vagy az Excel segítségével.

A szennyvízcsatorna-hálózat egy lakosra jutó hossza (y_i) és a száz lakásra jutó lakosok száma (x_{i2}) közötti összefüggést számszerűsítő lineáris regressziós modell becsült paraméterei:

$$\hat{\gamma}_0 = -4{,}868 ;$$

$$\hat{\gamma}_1 = 0{,}027 .$$

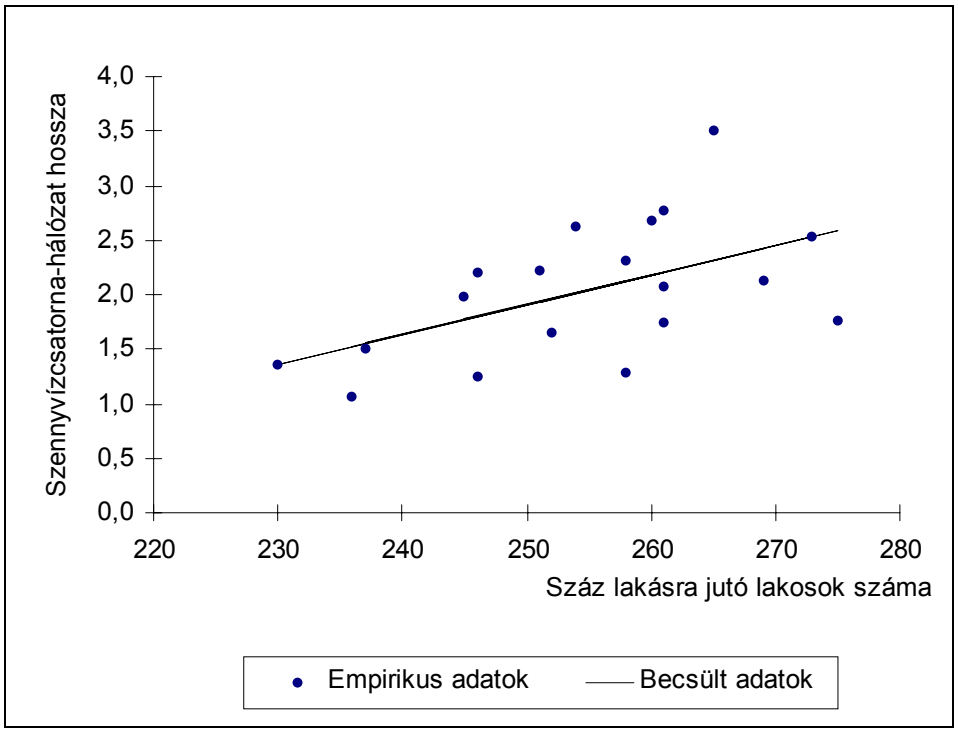
Az empirikus elemzés eredményeit most is a már említett (ajánlott) formában közöljük.

$$\begin{array}{lll} \hat{y}_i = -4{,}868 + 0{,}027 \cdot x_{i2} & & r^2 = 0{,}289 \\ \quad (2{,}625) \quad (0{,}010) & & F = 6{,}926 \\ \quad t = -1{,}854 \quad t = 2{,}632 & & \end{array}$$

$$\begin{array}{ll} d = 2{,}396 & 4 - d_U = 2{,}599 \\ r_{|e|x_2} = 0{,}357 & t = 1{,}576 \end{array}$$

5%-os szignifikancia-szint mellett az elméleti F érték: $F_{0,95}(1,17) = 4{,}451$. Mivel a próbafüggvény értéke $F = 6{,}926$ nagyobb az elméletinél, a nullhipotézist ($H_0 : \gamma_1 = 0$) elvetjük, ami azt jelenti, hogy szignifikáns (igaz, nagyon gyenge) összefüggés van a magyarázó- és az eredményváltozó között. (Lásd az 55. ábrát.)

A lineáris regressziófüggvény illesztése



55. ábra

Megjegyzés: mivel az eredeti modellben a két magyarázóváltozót egymástól gyakorlatilag (lineárisan) függetlennek tekinthetjük ($r_{x_1x_2} = -0{,}034$), a γ_1 becsült értéke nagyon kis mértékben különbözik a β_2 becsült értékétől (három tizedesig egyformák).

A kétváltozós modell reziduumai is lényegében homoszkedasztikusak és nem áll fenn közöttük statisztikailag jelentős elsőrendű autokorreláció.

11.4. Az általánosított legkisebb négyzetek módszere

Ahogy azt a 11.1. fejezetben láttuk, a standard lineáris regressziós modell feltételrendszere szerint a hibatagok nulla várható értékű, konstans varianciájú, korrelálatlan valószínűségi változók. Ekkor, mint tudjuk, a hibatag variancia-kovarianciamátrixa az alábbi.

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & & \\ 0 & & & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Ha a hibatag fent említett tulajdonságai nem teljesülnek, akkor az $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ mátrix főátlójában levő elemek nem egyenlőek, és a főátlón kívüli elemek nem mindegyike lesz 0. Ekkor a fenti mátrix felírható a (246) szerint.

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{\Omega}. \tag{246}$$

Ennek viszont az a következménye, hogy az LNM segítségével kapott képleteink már nem alkalmazhatóak. Ha az $\boldsymbol{\Omega}$ mátrix pozitív definit, akkor a (142) helyett $\boldsymbol{\beta}$ paramétervektor becslőfüggvénye

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}, \tag{247}$$

a $\hat{\boldsymbol{\beta}}$ paraméterek variancia-kovarianciamátrixa

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}, \tag{248}$$

a σ^2 becslése pedig

$$s_e^2 = \frac{\mathbf{e}'\boldsymbol{\Omega}^{-1}\mathbf{e}}{n-m-1}. \tag{249}$$

A (247)-(249) képletek az LNM általánosításai, amelyre az **általánosított legkisebb négyzetek módszereként** hivatkozunk.

Mivel a standard lineáris regressziós modellnek megfelelő esetben:

$$\mathbf{\Omega} = \mathbf{I},$$

a klasszikus legkisebb négyzetek módszere (LNM) az általánosított legkisebb négyzetek módszere egy speciális esetének tekinthető.[25)]

AITKEN-tétel: az általánosított legkisebb négyzetek módszere BLUE tulajdonágú becslést ad.

Megjegyzés: a GAUSS-MARKOV-tétel az AITKEN-tétel egy speciális esete.

Ahhoz, hogy a (247)-(249) képleteket alkalmazni tudjuk ismernünk kellene az $\mathbf{\Omega}$ mátrixot. Mivel ez az empirikus vizsgálatoknál ismeretlen, becsülnünk kell. Egy n elemű minta alapján azonban ezen mátrix $\frac{n(n+1)}{2}$ elemére nem következtethetünk, ezért az $\mathbf{\Omega} = \mathbf{\Omega}(\mathbf{\Theta})$ szerkezetére vonatkozó feltételezésből indulunk ki, és általában arra törekszünk, hogy minél kevesebb paramétert tartalmazzon. Ha $\mathbf{\Theta}$ paramétervektort legalább aszimptotikusan torzítatlanul tudjuk becsülni, akkor $\hat{\boldsymbol{\beta}}$ konzisztens lesz.

Becslés szignifikáns autokorreláció mellett

A 11.3. fejezetben ismertetett elsőrendű (lineáris) autokorrelációs modell (ahol $\rho^2 < 1$) esetén az $\mathbf{\Omega}$ mátrix a (250) szerinti.

$$\mathbf{\Omega} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \rho^{n-3} \\ \vdots & & & & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & & 1 \end{bmatrix} \qquad (250)$$

Innen

[25)] A klasszikus legkisebb négyzetek módszerére gyakran az OLS (Ordinary Least Squares), míg az általánosított legkisebb négyzetek módszerére a GLS (Generalized Least Squares) betűszóval hivatkozunk.

$$\boldsymbol{\Omega}^{-1} = \frac{1}{1-\rho^2}\cdot\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & & -\rho & 1 \end{bmatrix}. \qquad (251)$$

Ekkor csak egy paramétert, a ρ -t kell becsülnünk, például (252) szerint.

$$\hat{\rho} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_i^2} \qquad (252)$$

Az általánosított legkisebb négyzetek módszere helyett alkalmazhatjuk a **COCHRANE-ORCUTT iteratív módszert** is. Ez az alábbi lépésekből áll.

1) Az LNM alkalmazása és az autokorreláció tesztelése.
2) Az alternatív hipotézis elfogadása esetén a 3) lépés következik, különben megkaptuk a modell becslését.
3) Elvégezzük az alábbi transzformációkat.[26)]

$$y_i^* = y_i - \hat{\rho}\cdot y_{i-1}$$

$$x_{ij}^* = x_{ij} - \hat{\rho}\cdot x_{i-1,j} \qquad i = 2,3,...,n \qquad j = 1,2,...,m$$

4) Végrehajtjuk az 1) lépést.

Az eljárás egyszerű, ezért gyakran alkalmazzuk.

[26)] A 3) lépés az eredeti modell $\mathbf{T}$ transzformációs mátrixszal való beszorzásának következménye.

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad /\cdot\mathbf{T}$$
$$\mathbf{Ty} = \mathbf{TX\beta} + \mathbf{T\varepsilon}$$

Olyan $\mathbf{T}$-re van szükségünk, amelyre: $E(\mathbf{T\varepsilon\varepsilon'T'}) = \sigma_\eta^2\mathbf{I}$. Ha $\boldsymbol{\Omega}$ (250) szerinti, akkor (246) figyelembevételével,

$$\mathbf{T} = \begin{bmatrix} -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & & 1 \end{bmatrix}$$

$(n-1)\cdot n$ elemű mátrixra $\frac{1}{1-\rho^2}\cdot\mathbf{T'T} \approx \boldsymbol{\Omega}^{-1}$.

83. példa

A 90. táblázat harmadik és negyedik oszlopa a magyarországi szarvasmarha- és sertésállomány alakulását mutatja.
Ha a magyarázóváltozó a sertésállomány, lineáris modellt feltételezve, számítsuk ki a regressziós egyenes egyenletét!

Vizsgáljuk meg mindenekelőtt a standard modell feltételeinek teljesülését. Teszteljük a heteroszkedaszticitást és az autokorrelációt. Ehhez alkalmazzuk az LNM-et.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -183{,}5099 \\ 0{,}2404 \end{bmatrix}$$

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 29738{,}41550 & -3{,}89052 \\ -3{,}89052 & 0{,}00053 \end{bmatrix}$$

A kapott becslés alapján, a heteroszkedaszticitás teszteléséhez, szükségünk van az

$$r_{|e|x} = 0{,}1927$$

értékre. A (242) próbafüggvény értéke ($t = 0{,}9620$) alapján a modell homoszkedasztikusnak tekinthető.

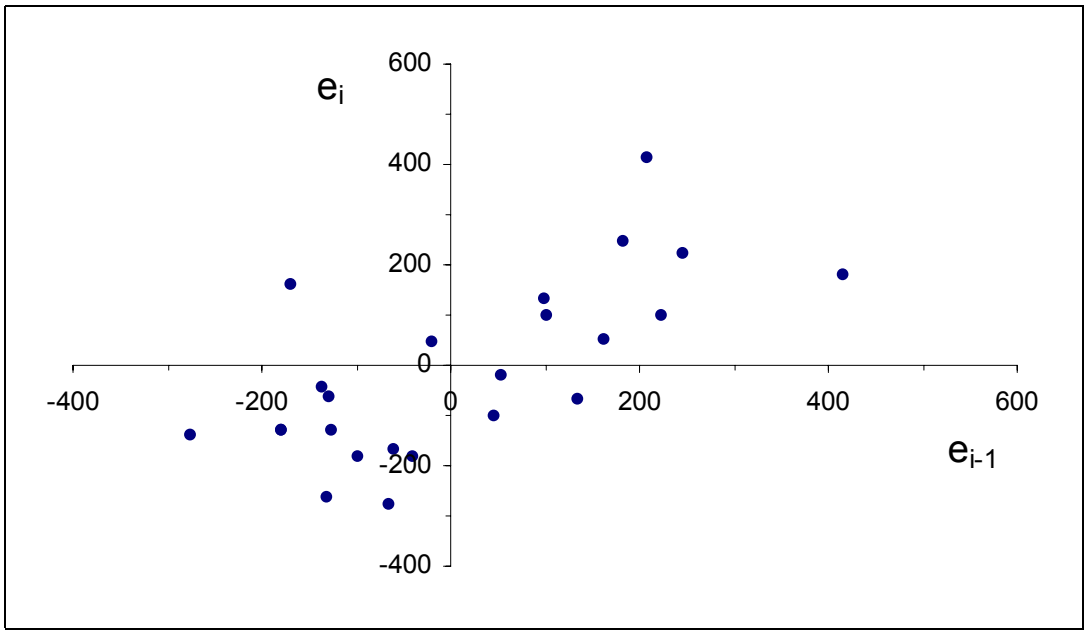
A (244) próbafüggvény értéke ($d = 0{,}5953$) alapján azonban a modell szignifikáns elsőrendű pozitív autokorrelációjára következtetünk ($\alpha = 0{,}01$ esetén $d_L = 1{,}055$).

A reziduumok grafikus ábrázolása (lásd az 56. ábrát) is a hibatagok közötti lineáris függőségre utal.

A szignifikáns autokorreláció miatt, a regressziós együtthatókat nem becsülhetjük az LNM segítségével, hanem az általánosított legkisebb négyzetek módszerét kell alkalmaznunk!

Az 56. ábra alapján a hibatagokra vonatkozó lineáris (elsőrendű) autokorrelációs modell feltételezhető, ezért az $\boldsymbol{\Omega}$ mátrix (250) szerinti szerkezete alkalmazható.

A reziduumok grafikus ábrázolása



56. ábra

Az autokorrelációs együttható becslése (252) szerint:

$$\hat{\rho} = \frac{473985{,}1620}{710815{,}3399} = 0{,}6668\,.$$

Így (251) mátrix a következő:

$$\boldsymbol{\Omega}^{-1} = \frac{1}{1-0{,}6668^2} \cdot \begin{bmatrix} 1 & -0{,}6668 & 0 & \cdots & 0 & 0 \\ -0{,}6668 & 1{,}4446 & -0{,}6668 & & 0 & 0 \\ 0 & -0{,}6668 & 1{,}4446 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & 1{,}4446 & -0{,}6668 \\ 0 & 0 & 0 & & -0{,}6668 & 1 \end{bmatrix}.$$

A (247)-(249) szerint, a megfelelő mátrixműveletek elvégzése után:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 357{,}9295 \\ 0{,}1652 \end{bmatrix},$$

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \frac{572001{,}2914}{23} \cdot \begin{bmatrix} 2{,}0121560561 & -0{,}0002537878 \\ -0{,}0002537878 & 0{,}0000000350 \end{bmatrix} =$$

$$= \begin{bmatrix} 50041{,}5592 & -6{,}3116 \\ -6{,}3116 & 0{,}0009 \end{bmatrix}.$$

Az ismertetett eljárás helyett alkalmazhatjuk a COCHRANE-ORCUTT iteratív módszert is. Ennek eredményeit a 95. táblázat tartalmazza.

A COCHRANE-ORCUTT iteratív módszer szerinti eredmények

95. táblázat

Az LNM alkalmazásának	eredménye					
sorszáma	n	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	d	d_L (1%)	$\hat{\rho}$
1.	25	0,2404	0,0231	0,5953	1,055	0,6668
2.	24	0,1518	0,0279	1,2146	1,037	0,3017

1%-os szignifikancia-szintet feltételezve, már az LNM második alkalmazása után elfogadhatjuk az autokorrelációra vonatkozó nullhipotézist.

Becslés szignifikáns heteroszkedaszticitás mellett

A 11.3. fejezetben ismertetett heteroszkedasztikus modell esetén az $\boldsymbol{\Omega}$ mátrix diagonális, és főátlójában levő ismeretlen elemek nem mind egyenlőek. Becslésük n elemű minta alapján történik, mint láttuk, a következő összefüggés feltételezése szerint:

$$E(e_i^2) = \operatorname{var}(e_i) \cdot x_{ij}^2 .$$

Ekkor a

$$\mathbf{P} = \begin{bmatrix} \frac{1}{x_{1j}} & 0 & \cdots & 0 \\ 0 & \frac{1}{x_{2j}} & & 0 \\ \vdots & & & \\ 0 & 0 & & \frac{1}{x_{nj}} \end{bmatrix} \tag{253}$$

mátrixra igaz az

$$\mathbf{\Omega}^{-1} = \mathbf{P}'\mathbf{P} = \mathbf{P}^2 \tag{254}$$

összefüggés.[27)]

A (253)-(254) segítségével már alkalmazhatjuk a (247)-(249) becslőfüggvényeket.

84. példa

A 96. táblázat az egy főre jutó bruttó hazai termék és a közműellátásra vonatkozó adatokat tartalmazza területi egységenként.

Ha a magyarázóváltozó az egy főre jutó GDP, lineáris modellt feltételezve, számítsuk ki a regressziós egyenes egyenletét!

Vizsgáljuk meg mindenekelőtt a standard modell feltételeinek teljesülését. Teszteljük az autokorrelációt és a heteroszkedaszticitást. Ehhez alkalmazzuk az LNM-et.

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -129{,}1844 \\ 0{,}5756 \end{bmatrix}$$

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 3340{,}2147 & -4{,}1387 \\ -4{,}1387 & 0{,}0057 \end{bmatrix}$$

[27)] Az eredeti modell (253) szerinti **P** transzformációs mátrixszal való beszorzásából adódik (254).

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad / \cdot \mathbf{P} \\ \mathbf{P}\mathbf{y} &= \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \\ E(\mathbf{P}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{P}') &= \sigma^2\mathbf{I} \\ \mathbf{P}\mathbf{\Omega}\mathbf{P}' &= \mathbf{I} \end{aligned}$$

A bruttó hazai termék és a szennyvízcsatorna-hálózat adatai területi egységenként 1997-ben

96. táblázat

Területi egység	Egy km vízvezeték-hálózatra jutó szennyvízcsatorna-hálózat (m)	Egy főre jutó bruttó hazai termék (ezer Ft)
Budapest	919,6	1575
Pest	290,0	653
Fejér	285,6	985
Komárom-Esztergom	409,4	724
Veszprém	256,0	675
Győr-Moson-Sopron	291,0	920
Vas	301,2	960
Zala	334,3	767
Baranya	287,9	672
Somogy	223,1	590
Tolna	233,6	708
Borsod-Abaúj-Zemplén	241,7	584
Heves	257,6	607
Nógrád	115,4	443
Hajdú-Bihar	239,0	642
Jász-Nagykun-Szolnok	300,8	632
Szabolcs-Szatmár-Bereg	242,8	487
Bács-Kiskun	183,9	615
Békés	173,6	603
Csongrád	232,4	755

Forrás: Magyar Statisztikai Évkönyv ’97, ’98, KSH, Bp., 1998-99.

A kapott becslés alapján, az autokorreláció teszteléséhez, szükségünk van a (244) próbafüggvény értékére.

$$d = 1{,}7990$$

Mivel 5%-os szignifikancia-szint mellett a megfelelő $d_U = 1{,}411$; a hibatagok függetlenségére vonatkozó nullhipotézist elfogadjuk.

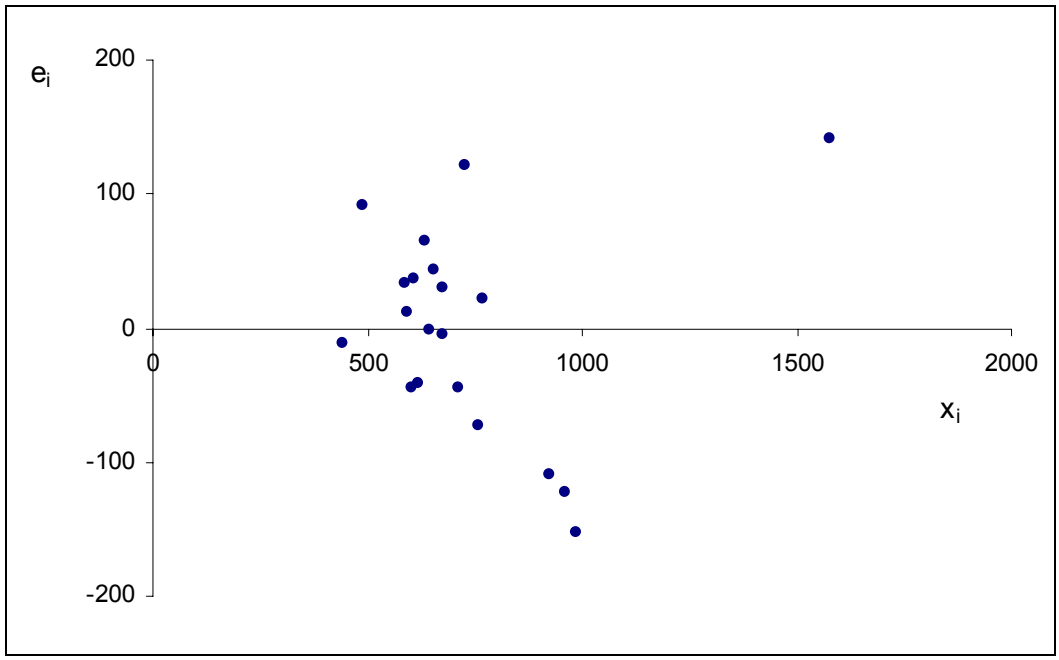
A heteroszkedaszticitás teszteléséhez szükségünk van az

$$r_{|e|x} = 0{,}6851$$

lineáris korrelációs együtthatóra. Ekkor a (242) próbafüggvény értéke $t = 3{,}9905$. Mivel a III. táblázat szerint $t_{0,95}(18) = 2{,}1009$; a modell heteroszkedasztikusnak tekinthető. Erre következtethetünk az 57. ábra alapján is.

A heteroszkedaszticitás miatt, a regressziós együtthatókat nem becsülhetjük az LNM segítségével, hanem az általánosított legkisebb négyzetek módszerét kell alkalmaznunk!

A reziduumok grafikus ábrázolása



57. ábra

Az 57. ábra alapján a reziduumok szórásnégyzetére vonatkozó $E(e_i^2) = \mathrm{var}(e_i) \cdot x_{ij}^2$ modell feltételezhető, ezért (254) mátrix a következő:

$$\boldsymbol{\Omega}^{-1} = \begin{bmatrix} \frac{1}{1575^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{653^2} & & 0 \\ \vdots & & & \\ 0 & 0 & & \frac{1}{755^2} \end{bmatrix}.$$

A (247)-(249) szerint, a megfelelő mátrixműveletek elvégzése után:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -46{,}0485 \\ 0{,}4582 \end{bmatrix},$$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \frac{0{,}1582}{18} \cdot \begin{bmatrix} 403336{,}0727 & -594{,}7470 \\ -594{,}7470 & 0{,}9270 \end{bmatrix} =$$

$$= \begin{bmatrix} 3545{,}0766 & -5{,}2275 \\ -5{,}2275 & 0{,}0081 \end{bmatrix}.$$

Az empirikus elemzéseknél az autokorreláció és a heteroszkedaszticitás mellett (amelyek negatív hatását az általánosított legkisebb négyzetek módszerével kezelni tudjuk) majdnem mindig jelentkezik a multikollinearitás is, de ennek következményeit a (247)-(249) képletekkel már nem tudjuk kiküszöbölni.

Szignifikáns multikollinearitás esetén hatékonyan alkalmazható eljárás a főkomponens analízis. Ezzel foglalkozik a 11.5. fejezet.

11.5. Főkomponens analízis

A standard regressziós modell feltételezi, hogy a magyarázóváltozók lineárisan függetlenek. Társadalmi, gazdasági adatok empirikus elemzésénél azonban, a változók között valamilyen mértékű sztochasztikus összefüggés szinte mindig előfordul. Ahhoz, hogy a 11.1. fejezetben ismertetett modellt alkalmazni tudjuk más módszerre van szükségünk, amellyel az eredeti magyarázóváltozókból olyan új változókat képezhetünk, amelyek teljesítik a standard modell feltételeit és megtartják a magyarázóváltozókban rejlő információkat. Az eredeti magyarázóváltozók transzformálásával kapott új változókat fogjuk főkomponenseknek nevezni.

A **főkomponens analízis** során a megfigyelések m dimenziós terét egy olyan új (derékszögű) koordináta-rendszerbe transzformáljuk, amelyben a transzformált változók varianciái rendre csökkennek. A főkomponens analízis során előállított új, mesterséges változók egymástól már függetlenek. A magyarázóváltozók multikollinearitása azt jelenti, hogy azok redundáns módon tartalmaznak információt. Például teljes multikollinearitás esetén a magyarázóváltozók mátrixának egy vagy több oszlopa elhagyható. Látni fogjuk, hogy a főkomponenseket úgy lehet előállítani, hogy az első néhánnyal már meg tudjuk magyarázni az eredményváltozó szórásnégyzetének igen nagy hányadát.

Főkomponensváltozók

Mivel különböző mértékegységű változókból fogunk új, mesterséges változókat előállítani, a mértékegységeket ki kell küszöbölnünk. Ehhez a standardizálás műveletét alkalmazzuk. A (31) képlet figyelembevételével:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad i = 1,2,...,n \qquad j = 1,2,...,m; \tag{255}$$

ahol s_j a j-edik magyarázóváltozó (167)-(168) szerinti korrigált tapasztalati szórását jelöli.

A főkomponensanalízis formális modellje a következő:

$$\mathbf{C} = \widetilde{\mathbf{X}}\mathbf{U}\,, \tag{256}$$

ahol **U** olyan lineáris transzformáció mátrixa, amely az $\tilde{\mathbf{x}}$ vektorváltozókat **c** korrelálatlan vektorváltozókba transzformálja. A **C** mátrix oszlopvektorait **főkomponensvektoroknak** vagy **főkomponenseknek** nevezzük.

Feladatunk tehát az **U** mátrix u_{kl} $(k,l = 1,2,...,m)$ elemeinek a meghatározása. Ezeket az $\tilde{x}_j$ standardizált változók variancia-kovarianciamátrixának $\mathbf{u}_l$ ortonormált sajátvektorai adják. Mivel a standardizált változók variancia-kovarianciamátrixa az eredeti változók korrelációs mátrixával (**R**) azonos, így eleve ebből a mátrixból indulhatunk ki.
Legyen **R** (önadjungált mátrix) spektrálfelbontása a következő:

$$\mathbf{R} = \mathbf{U}\Lambda\mathbf{U}'\,,$$

ahol **Ë** diagonális mátrix, amelynek főátlójában a $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ sajátértékek állnak, az **U** oszlopvektorai pedig a megfelelő sajátvektorok.

A sajátértékek összege a magyarázóváltozók számával egyenlő: $\sum_{j=1}^{m} \lambda_j = m$.

A főkomponensek **C** és a magyarázóváltozók $\widetilde{\mathbf{X}}$ mátrixa ugyanolyan alakú, azaz mindkét mátrix dimenziója $n \cdot m$.

A (256) figyelembevételével, a főkomponensek és a standardizált magyarázóváltozók között felírható a következő két összefüggés:[28)]

$$c_{ij} = u_{1j}\tilde{x}_{i1} + u_{2j}\tilde{x}_{i2} + \ldots + u_{mj}\tilde{x}_{im}\,, \tag{257}$$

[28)] Mivel **U** ortogonális, fennáll $\mathbf{U}^{-1} = \mathbf{U}'$.

$$\begin{aligned} \mathbf{C} &= \widetilde{\mathbf{X}}\mathbf{U} \qquad / \cdot \mathbf{U}^{-1} \\ \mathbf{C}\mathbf{U}^{-1} &= \widetilde{\mathbf{X}} \\ \widetilde{\mathbf{X}} &= \mathbf{C}\mathbf{U}' \end{aligned}$$

illetve

$$\tilde{x}_{ij} = u_{j1}c_{i1} + u_{j2}c_{i2} + \ldots + u_{jm}c_{im} . \qquad (258)$$

Megjegyzés: az eddigiekből következik, hogy a főkomponensek korrelálatlanok és $\mathbf{c}_j$ főkomponens szórásnégyzete a megfelelő λ_j sajátértékkel egyenlő.

A főkomponenssúlyok

A **főkomponenssúlyok** (**loading változók**) a sajátvektorok komponenseinek és a megfelelő sajátértékek négyzetgyökének a szorzatai:

$$a_{kl} = u_{kl}\sqrt{\lambda_l} \qquad k,l = 1,2,...,m . \qquad (259)$$

A főkomponenssúlyokat tartalmazó **A** mátrix az ún. **főkomponenssúly-mátrix**, dimenziója $m \cdot m$, és az alábbi tulajdonságokkal rendelkezik.

- A főkomponenssúlyok abszolút értékei 1-nél nem nagyobbak.
- Az oszloponkénti négyzetösszegük λ_j, a soronkénti négyzetösszegük 1.
- Oszloppáronkénti szorzatuk 0, sorpáronkénti szorzatuk a megfelelő két magyarázóváltozó lineáris korrelációs együtthatója.
- A főkomponenssúlyok megadják a magyarázóváltozók és a főkomponensváltozók közötti lineáris korrelációs együtthatót.

$$a_{kl} = r_{\tilde{x}_k c_l} = r_{x_k c_l} \qquad (260)$$

Kommunalitások

Ha az **A** mátrix i-edik sora első w darab elemeinek négyzeteit kumuláljuk, akkor az i-edik magyarázóváltozó $h_i^{(w)}$ **kommunalitásához** jutunk.

$$h_k^{(w)} = \sum_{l=1}^{w} a_{kl}^2 \qquad 1 \le w \le m \qquad (261)$$

A kumulált főkomponenssúly-négyzetek azt fejezik ki, hogy az egyes főkomponenseknek milyen jelentősége, súlya van a magyarázóváltozók varianciájában,

azaz az első w darab főkomponens milyen mértékben járul hozzá az $\tilde{x}_k$ magyarázóváltozó szórásnégyzetéhez. Például $h_4^{(3)} = a_{41}^2 + a_{42}^2 + a_{43}^2$ azt mutatja, hogy a negyedik magyarázóváltozó szórásnégyzetének az első három főkomponens $100 \cdot h_4^{(3)}$ százaléknyi hányadát értelmezi. Nyilvánvalóan $h_k^{(m)} = 1$, illetve 100%.

Mivel általában néhány főkomponens már jól jellemzi a mintában rejlő információt, a többi elhanyagolható, számuk csökkenthető.

Az eddigiekben a magyarázóváltozók szórásnégyzeteinek értelmezett hányadáról volt szó, de fontos tudni azt is, hogy az eredményváltozó szórásnégyzetének túlnyomó részét hány főkomoponenssel tudjuk értelmezni. Szignifikáns multikollinearitás esetén azokat a főkomponenseket, amelyekhez tartozó sajátérték 1-nél kisebb (vagyis nem éri el az átlagot) általában már nem vesszük figyelembe.

85. példa

Vizsgáljuk meg, hogy a 90. táblázat utolsó három oszlopában szereplő három magyarázóváltozót hány főkomponenssel lehetne helyettesíteni!

Először ellenőrizzük a magyarázóváltozók lineáris függetlenségét! Ehhez szükségünk van a magyarázóváltozókra vonatkozó korrelációs mátrixra.

$$\mathbf{R} = \begin{bmatrix} 1{,}0000 & 0{,}9084 & 0{,}9083 \\ 0{,}9084 & 1{,}0000 & 0{,}8206 \\ 0{,}9083 & 0{,}8206 & 1{,}0000 \end{bmatrix}$$

Már a korrelációs mátrix elemei alapján is következtethetünk arra, hogy szignifikáns, igen nagy mértékű multikollinearitás jellemző az adatokra. Erre utal az $M = 0{,}57$ érték is. A magyarázóváltozók közötti erős sztochasztikus kapcsolat miatt nem ajánlatos az LNM alkalmazása, hanem a főkomponens analízis végrehajtása volna célszerű.

Első lépésként (255) szerint standardizáljuk a magyarázóváltozókat. Az eredmény a 97. táblázatban található.

Standardizált adatok

97. táblázat

Év	Szarvasmarha-állomány	Sertésállomány	Baromfiállomány
	$\tilde{x}_{i1}$	$\tilde{x}_{i2}$	$\tilde{x}_{i3}$
1974	1,0478	0,6158	-0,2874
1975	0,7812	-0,2210	0,4739
1976	0,7411	0,3416	1,1343
1977	0,8874	0,3391	1,1082
1978	0,9275	0,4397	1,1129
1979	0,8307	0,6545	0,8292
1980	0,8142	0,6389	1,0397
1981	0,8779	0,6177	1,0428
1982	0,8237	1,0791	1,4033
1983	0,7883	1,5843	0,8329
1984	0,7741	1,2053	0,7908
1985	0,4555	0,6077	0,4337
1986	0,3588	0,8618	0,2680
1987	0,2148	0,5677	0,1363
1988	0,2762	0,6370	0,0513
1989	0,0591	0,2205	-0,1443
1990	-0,0046	0,4328	-0,5681
1991	-0,3610	-0,8205	-0,8732
1992	-0,9769	-1,2133	-0,6491
1993	-1,3544	-1,4400	-1,2005
1994	-1,5644	-1,8428	-0,7441
1995	-1,5220	-1,4206	-1,0614
1996	-1,5668	-1,2601	-1,9572
1997	-1,6565	-1,4837	-1,6317
1998	-1,6518	-1,1415	-1,5402

Végezzük el az eredeti magyarázóváltozók korrelációs mátrixának (**R**) spektrálfelbontását!

Ehhez az **R** sajátértékeire van szükségünk. Ezeket az Excel segítségével is meg tudjuk határozni, például a „célérték-keresés" felhasználásával. Az **Eszközök** menü **Adatelemzés...** almenüjében levő **Korrelációanalízis** menüpont segítségével számítsuk ki az eredeti magyarázóváltozók korrelációs mátrixát (vagy a **Kovarianciaanalízis** segítségével a standardizált magyarázóváltozók variancia-kovarianciamátrixát)! Készítsük el az $[\mathbf{R} - \lambda\mathbf{I}]$ mátrixot mondjuk a B6:D8 cellatartományban, úgy hogy λ például az F6 cellába kerüljön. Az F6 kezdőértéke legyen a változók száma, tehát 3. A B10 mezőben az MDETERM(tömb) függvénnyel számíttassuk ki a mátrixunk determinánsát: =MDETERM(B6;D8). Most hívjuk meg az **Eszközök** menü **Célérték-keresés...** almenüjét. A Célcella legyen B10, a Célérték 0, a Módosuló cella F6. Ekkor az F6 cellában megkapjuk a 3-hoz legközelebbi, tehát a legnagyobb sajátértéket ($\lambda_1 = 2{,}7589$). Most írjuk át az F6 értékét $3 - \lambda_1 = 0{,}2411$ értékre; majd újra végezzünk célérték-keresést az előző módon. A harmadik sajátértéket az első kettő segítségével már ki tudjuk számítani: $\lambda_3 = 3 - \lambda_1 - \lambda_2$.

A keresett három sajátérték az alábbi.

$$\begin{aligned} \lambda_1 &= 2{,}758835 \\ \lambda_2 &= 0{,}179400 \\ \lambda_3 &= 0{,}061765 \\ \hline & 3{,}000000 \end{aligned}$$

Az Excel mátrixokkal kapcsolatos műveleteit felhasználva oldjuk meg mind a három λ-ra az alábbihoz hasonló (u_{i2}-nek és u_{i3}-nak megfelelő) homogén lineáris egyenletrendszert, ahol az együtthatók az **R** mátrix elemei.

$$\begin{aligned} (1-\lambda)\cdot u_{11} + 0{,}9084\cdot u_{21} + 0{,}9083\cdot u_{31} &= 0 \\ 0{,}9084\cdot u_{11} + (1-\lambda)\cdot u_{21} + 0{,}8206\cdot u_{31} &= 0 \\ 0{,}9083\cdot u_{11} + 0{,}8206\cdot u_{21} + (1-\lambda)\cdot u_{31} &= 0 \end{aligned}$$

A normált sajátvektorokat és a hozzájuk tartozó sajátértékeket a 98. táblázat tartalmazza.

Az **R** mátrixból kiszámított sajátértékek és sajátvektorok

98. táblázat

Változók	u_{i1}	u_{i2}	u_{i3}
Szarvasmarha-állomány	0,5898	-0,0001	-0,8075
Sertés-állomány	0,5710	-0,7070	0,4172
Baromfi-állomány	0,5710	0,7072	0,4170
Sajátértékek	2,7588	0,1794	0,0618

A (259) figyelembevételével kiszámíthatjuk a főkomponenssúly-négyzeteket.

A főkomponenssúly-négyzetek

99. táblázat

Változók	a_{i1}^2	a_{i2}^2	a_{i3}^2
Szarvasmarha-állomány	0,9597	0,0000	0,0403
Sertés-állomány	0,8995	0,0897	0,0107
Baromfi-állomány	0,8995	0,0897	0,0107
Összesen (sajátértékek)	2,7588	0,1794	0,0618

Az első, a második és a harmadik magyarázóváltozó szórásnégyzetének rendre (megközelítőleg) 96; 90 és 90%-át lehet az első főkomponenssel értelmezni.

A 99. táblázat adatai és a (261) segítségével ki lehet számítani a három magyarázóváltozóhoz tartozó $h_k^{(w)}$ kommunalitási mutatókat. Például

$h_3^{(2)} = 0{,}8995 + 0{,}0897 = 0{,}9892$. Ez azt jelenti, hogy a harmadik magyarázóváltozó szórásnégyzetének 98,92%-át tudjuk az első két főkomponenssel megmagyarázni.

A (256) vagy a (257) alapján kiszámított főkomponenseket a 100. táblázat tartalmazza.

A főkomponensek

100. táblázat

Év	Szarvasmarha-állomány	Sertésállomány	Baromfiállomány
	c_{i1}	c_{i2}	c_{i3}
1974	0,8055	-0,6387	-0,7091
1975	0,6052	0,4913	-0,5254
1976	1,2799	0,5605	0,0170
1977	1,3498	0,5438	-0,1131
1978	1,4336	0,4760	-0,1015
1979	1,3372	0,1235	-0,0521
1980	1,4387	0,2834	0,0425
1981	1,4660	0,3007	-0,0165
1982	1,9033	0,2293	0,3701
1983	1,8452	-0,5313	0,3717
1984	1,5964	-0,2930	0,2074
1985	0,8633	-0,1230	0,0665
1986	0,8568	-0,4199	0,1815
1987	0,5287	-0,3051	0,1202
1988	0,5560	-0,4141	0,0641
1989	0,0783	-0,2580	-0,0159
1990	-0,0800	-0,7078	-0,0526
1991	-1,1800	-0,0373	-0,4149
1992	-1,6396	0,3990	0,0121
1993	-2,3066	0,1694	-0,0075
1994	-2,3999	0,7769	0,1844
1995	-2,3150	0,2540	0,1938
1996	-2,7613	-0,4929	-0,0765
1997	-2,7560	-0,1047	0,0383
1998	-2,5055	-0,2819	0,2155

Ellenőrzés végett számítsuk ki a főkomponensek variancia-kovarianciamátrixát. Ez diagonális mátrix, amelynek főátlójában a sajátértékek állnak.

$$\mathbf{C}_c = \begin{bmatrix} 2{,}7588 & 0{,}0000 & 0{,}0000 \\ 0{,}0000 & 0{,}1794 & 0{,}0000 \\ 0{,}0000 & 0{,}0000 & 0{,}0617 \end{bmatrix}$$

A kiszámított főkomponensek valóban korrelálatlanok és a főátlóban is (kerekítési hibával) a sajátértékek állnak.

Az említetteken kívül, a főkomponenselemzésnek van egy másik alkalmazási lehetősége is. Ez vagy a megfigyelések, vagy a magyarázóváltozók grafikus ábrázolásából áll. Olyan grafikonokról van szó, amelyeknél a vízszintes tengelyen az első főkomponens, míg a függőleges tengelyen a második főkomponens található.[29)]

Az ilyen grafikonoknál gyakran fordul elő az az eset, hogy az ábrázolt pontok egy része nagyon közel esik egymáshoz, azaz koordinátáik megközelítőleg azonosak. Ezeket a csoportosulásokat (általában több van belőlük) **clustereknek** nevezzük, amelyek mögött rendszerint valamilyen közös tényező, ún. **háttérváltozó** (**faktorváltozó**) áll. Ezeknek a háttérváltozóknak a részletes elemzése a **faktoranalízis** tárgya, de mi ezzel nem foglalkozunk.

A fentiekből következik, hogy kevés számú magyarázóváltozót tartalmazó modelleknél nincs értelme az esetleges háttérváltozók keresésének, ezért a 86. példa hat magyarázóváltozóból indul ki.

86. példa

Számítsuk ki a 101. táblázatban szereplő adatok alapján a főkomponenssúly-mátrixot és ábrázoljuk az első két oszlopát!
Jelölje rendre $\mathbf{x}_j$ ($j = 1,2,...,6$) a táblázat utolsó hat vektorát.

[29)] Elvileg háromdimenziós grafikus ábrát is alkalmazhatnánk, de szignifikáns multikollinearitás esetén (általában) a harmadik főkomponens szerepeltetése nem célszerű, mert a pontok elrendeződése a harmadik tengely mentén nagyon keskeny lenne, és nem nyújtana vizuálisan lényeges többletinformációt.

Hazánk ipari termelésének néhány fontosabb adata

101. táblázat

Év	Villamos-energia (millió kWh)	Kőolaj (1000 t)	Bauxit (1000 t)	Autóbusz (db)	Televízió-készülék (1000 db)	Műanyag-alapanyag (1000 t)
1969	14069	1754	1934	4774	345	39
1970	14542	1937	2022	5956	364	56
⋮						
1997	35305	1360	743	1951	963	855
1998	37023	1258	909	1232	1703	883

Forrás: Magyar Statisztikai Évkönyv ’98, KSH, Bp., 1999.

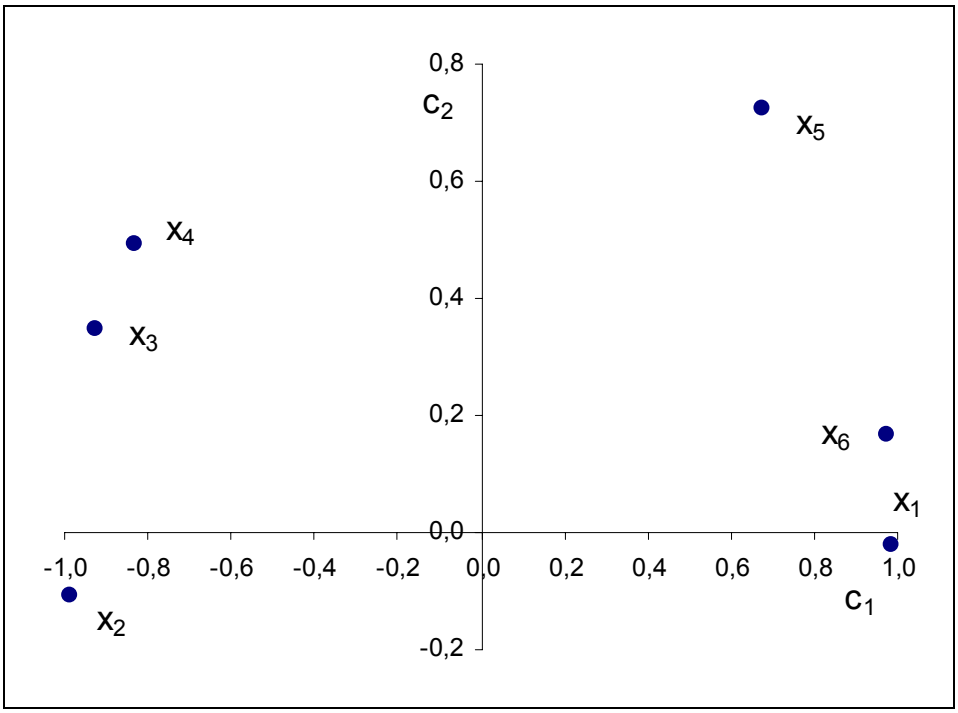
A főkomponensek meghatározása után a (259) szerinti mátrix az alábbi.

Főkomopensek

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 & \mathbf{c}_5 & \mathbf{c}_6 \end{matrix} \\ \begin{bmatrix} 0{,}9854 & -0{,}0204 & 0{,}1026 & -0{,}0995 & 0{,}0826 & 0{,}0358 \\ -0{,}9876 & -0{,}1059 & -0{,}0292 & 0{,}0713 & 0{,}0455 & 0{,}0739 \\ -0{,}9263 & 0{,}3478 & -0{,}0906 & -0{,}0310 & 0{,}0994 & -0{,}0408 \\ -0{,}8364 & 0{,}4939 & 0{,}2357 & -0{,}0080 & -0{,}0284 & 0{,}0037 \\ 0{,}6718 & 0{,}7267 & -0{,}1359 & -0{,}0044 & -0{,}0320 & 0{,}0334 \\ 0{,}9711 & 0{,}1675 & 0{,}0769 & 0{,}1401 & 0{,}0549 & -0{,}0199 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{Villanyáram} \\ \text{Kőolaj} \\ \text{Bauxit} \\ \text{Autóbusz} \\ \text{TV} \\ \text{Műanyag} \end{array}$$

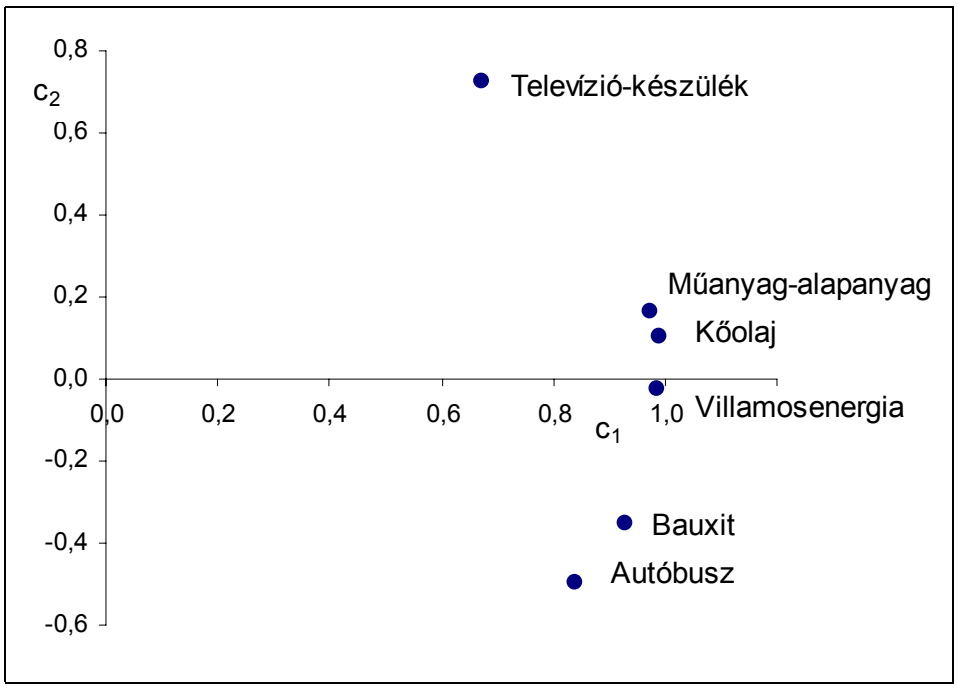
Az 58. ábrán az **A** mátrix első két oszlopa szerinti pontokat ábrázoltuk. Mivel most csak a korrelációs kapcsolat erőssége érdekel bennünket (és az iránya nem), a pontok esetleges csoportosulásának szemléltetése végett tükrözzük a második és a harmadik negyedbe eső pontokat az origóra. Az áttükrözés utáni kép az 59. ábrán látható. Ez alapján három pontcsoportosulást, azaz clustert különböztethetünk meg. Egyikbe tartozhat az autóbusz- és a bauxit-, egy másikba a kőolaj-, a műanyag-alapanyag- és a villanyáram-, egy újabba a televíziókészülék termelése. Ezek mögött álló háttérváltozók egy értelmezése lehetne a vizsgált termékek külkereskedelme. Az autóbusz és a bauxit tipikus kiviteli, míg a második cluster három eleme tipikusan behozatali termékünk.

A főkomponenssúlyok ábrázolása



58. ábra

A főkomponenssúlyok áttükrözés utáni ábrázolása



59. ábra

Tesztkérdések

I. Tesztkérdések válaszokkal

A következő két részben 15-15 tesztfeladatot talál, amelyek mindegyikében 4 állítást kell minősíteni aszerint, hogy azt igaznak vagy hamisnak ítéli meg.
Válaszát egyértelműen jelölje I vagy H betűvel!
Megjegyzés: ezeknél a feladatoknál mellékszámítást nem kell bemutatni.

1. Egy sokaság lehet:
   A. mozgó;
   B. lineáris;
   C. aggregált;
   D. diszkrét.

2. A következő mutatók a kvantilisekhez tartoznak:
   A. kvintilis;
   B. percentilis;
   C. módusz;
   D. medián.

3. Nagyság szempontjából, egyazon adatállományt vizsgálva, milyen összefüggés van az átlagos abszolút eltérés és a szórás között?
   A. Mindig a szórás kisebb;
   B. mindig a szórás nagyobb;
   C. általában a szórás nagyobb;
   D. nincsen szabály.

4. Adva van egy 5 tagból álló mennyiségi sor, amelyre vonatkozóan a számított átlagok: $\bar{x}_h = 3{,}9437$; $\bar{x}_g = 4{,}4737$; $\bar{x} = 5{,}0000$ és $\bar{x}_q = 5{,}4590$. Ezen adatok alapján leírhatjuk a következő egyenlőségeket:

   A. $M_{-1} = 0{,}2536$;

   B. $M_2 - M_1^2 = 4{,}8007$;

   C. $v = 0{,}4382$;

   D. v-t nem lehet kiszámítani.

5. A teljes szórásnégyzet a belső és a külső szórásnégyzet összege. Azt állíthatjuk, hogy:

   A. a belső szórás a részszórások súlyozott számtani átlaga;

   B. a belső szórás a csoporton belüli szórások súlyozott négyzetes átlaga;

   C. a belső szórásnégyzet a részszórások súlyozott négyzetes átlaga;

   D. a belső szórásnégyzet a részvarianciák súlyozott számtani átlaga.

6. Nagyon sok megfigyelésből álló gyakorisági sor (becsült) középértékei között, baloldali aszimmetria esetén, (általában) fennállnak a következő összefüggések:

   A. $\bar{x} < \hat{M}e < \hat{M}o$;

   B. $\hat{M}o < \hat{M}e < \bar{x}$;

   C. $\bar{x} < \hat{M}o < \hat{M}e$;

   D. $\hat{M}e < \hat{M}o < \bar{x}$.

7. Standardizálásnál ismert a következő összefüggés: $I = I' \cdot I''$. Azt állíthatjuk, hogy:

   A. az I'' azt mutatja, hogy a részviszonyszámok változása hogyan hatott a vizsgált összetett (intenzitási) viszonyszám változására;

   B. az I'' index csupán az összetételváltozás tényét fejezi ki;

   C. az I'' azt mutatja, hogy az összetételváltozás hogyan hatott a vizsgált összetett (intenzitási) viszonyszám változására;

   D. az I' indexet összetételhatás-indexnek nevezzük.

8. Az indexpróbák az indexekkel kapcsolatos követelményeket fejezik ki. Az alábbiak közül ezek tartoznak az indexpróbákhoz:
   A. függetlenségi próba;
   B. összemérhetőségi próba;
   C. négyzetes próba;
   D. tényezőpróba.

9. Homogén, véges elemszámú sokaság esetén a következő típusú mintákat szokás alkalmazni:
   A. egyenletes elosztású rétegzett minta;
   B. arányos elosztású rétegzett minta;
   C. csoportos minta;
   D. egyszerű véletlen minta.

10. Becslőfüggvényekkel kapcsolatosak a következő állítások:
   A. egy torzítatlan és egy torzított becslőfüggvényt hatásosság szempontjából nem tudunk összehasonlítani;
   B. ha egy becslőfüggvény konzisztens, akkor torzítatlan is;
   C. ha egy becslőfüggvény torzítatlan, akkor efficiens is;
   D. egy torzított becslőfüggvény lehet efficiens is.

11. A statisztikában használt nevezetes elméleti eloszlásokkal kapcsolatosak az alábbi összefüggések.
   A. Véges szabadságfok mellett a χ^2-eloszlásnak baloldali aszimmetriája van.
   B. Véges szabadságfok mellett az F-eloszlásnak jobboldali aszimmetriája van.
   C. Véges szabadságfok mellett a t-eloszlásnak jobboldali aszimmetriája van.
   D. A normális eloszlás néha aszimmetrikus is lehet.

12. A standard lineáris regressziós modelleknek megfelelő feltételek a következőek:

A. ekvidisztans megfigyelések kellenek;

B. homoszkedaszticitás;

C. a magyarázóváltozók között lehet szignifikáns lineáris kapcsolat;

D. autokorreláció.

13. Adva van két lineáris regressziófüggvény: $y(x)$ és $x(y)$, amelyeknél a két változó (X és Y) konkrét jelentése most irreleváns. A következő regressziós paraméterek párosai közül statisztikailag lehetségesek:

A. $y(x)$: 0,5 és $x(y)$: 1,5;

B. $y(x)$:-0.5 és $x(y)$:-1.5;

C. $y(x)$:-0,5 és $x(y)$: 1,5;

D. $y(x)$: 0,5 és $x(y)$: 2,3.

14. Autokorreláció tesztelésekor a d-statisztika nagyságát a DURBIN-WATSON-féle táblázat kritikus értékeivel szoktuk összehasonlítani. Ismertek a következő adatok: $n = 25$; $m = 3$ és $d = 3{,}8$. Ezek ismeretében, elsőrendű autokorrelációt feltételezve, az adatokból ($\alpha = 0{,}01$ esetén) az következik, hogy:

A. a reziduumok egymástól lineárisan függetlenek;

B. pozitív autokorrelációról van szó;

C. negatív autokorrelációról van szó;

D. elsőrendű autokorrelációnál a fenti adatok nem lehetségesek.

15. Három- vagy többváltozós regressziós elemzésnél a multikollinearitás majdnem mindig jelentkezik. Következményeihez az alábbiak tartoznak:

A. a becsült regressziós együtthatók nem torzítatlanok;

B. a becsült regressziós együtthatók szórását csökkenti;

C. instabillá teszi a becsléseket;

D. nem lehet kiszámítani a korrelációs mátrixot.

Válaszok

1.	A) **I**	B) **H**	C) **I**	D) **I**
2.	A) **I**	B) **I**	C) **H**	D) **I**
3.	A) **H**	B) **H**	C) **I**	D) **H**
4.	A) **I**	B) **I**	C) **I**	D) **H**
5.	A) **H**	B) **I**	C) **H**	D) **I**
6.	A) **H**	B) **I**	C) **H**	D) **H**
7.	A) **H**	B) **H**	C) **I**	D) **H**
8.	A) **H**	B) **I**	C) **H**	D) **I**
9.	A) **H**	B) **H**	C) **I**	D) **I**
10.	A) **H**	B) **H**	C) **H**	D) **H**
11.	A) **I**	B) **H**	C) **H**	D) **H**
12.	A) **H**	B) **I**	C) **H**	D) **H**
13.	A) **I**	B) **I**	C) **H**	D) **H**
14.	A) **H**	B) **H**	C) **I**	D) **H**
15.	A) **H**	B) **H**	C) **I**	D) **H**

II. Tesztkérdések válaszok nélkül

1. A momentumokkal kapcsolatos összefüggések:
   A. a nulladik momentum mindig 0-val egyenlő;
   B. a nulladik momentum mindig 1-gyel egyenlő;
   C. a nulladik centrális momentum mindig 0-val egyenlő;
   D. a nulladik centrális momentum mindig 1-gyel egyenlő.

2. A hatványkitevős regressziófüggvény becsült regressziós együtthatójának ($\hat{\beta}_1$) értelmezése:
   A. ha a magyarázóváltozó értékét (bármilyen szintről) 1 egységnyivel növeljük, akkor az eredményváltozó értéke átlagosan, megközelítő pontossággal $\hat{\beta}_1$ százalékkal változik;
   B. ha a magyarázóváltozó értékét (bármilyen szintről) 1 százalékkal növeljük, akkor az eredményváltozó értéke átlagosan, megközelítő pontossággal $\hat{\beta}_1$ egységnyivel változik;
   C. ha a magyarázóváltozó értékét (bármilyen szintről) 1 százalékkal növeljük, akkor az eredményváltozó értéke átlagosan, megközelítő pontossággal $p = (\hat{\beta}_1 - 1) \cdot 100$ százalékkal változik;
   D. ha a magyarázóváltozó értékét (bármilyen szintről) $\hat{\beta}_1$ százalékkal növeljük, akkor az eredményváltozó értéke átlagosan, megközelítő pontossággal 1 egységnyivel változik.

3. A szóródás mérőszámaira ismertek a következő összefüggések:
   A. a szórás a második centrális momentum ;
   B. a variancia a második momentum négyzete;
   C. a relatív szórás nem lehet negatív előjelű;
   D. a standardizált változó átlaga negatív is lehet.

4. Mintavétellel kapcsolatosan ismertek a következő állítások:
   A. csoportos mintavétel esetén az egyes részsokaságok homogenitása előnyös;
   B. csoportos mintavétel esetén az egyes részsokaságok homogenitása nem előnyös;
   C. rétegzett mintavétel esetén az egyes sztrátumok homogenitása előnyös;
   D. rétegzett mintavétel esetén az egyes sztrátumok homogenitása nem előnyös.

5. A középértékekre vonatkoznak a következő állítások:
   A. az egyes adatok számtani átlaguktól mért eltéréseinek összege minimális;
   B. az egyes adatok számtani átlaguktól mért eltérései négyzeteinek összege minimális;
   C. az egyes adatok mediánjuktól mért eltéréseinek összege minimális
   D. az egyes adatok mediánjuktól mért eltérései négyzeteinek összege minimális.

6. Három- vagy többváltozós regressziós elemzésekkel kapcsolatban ismertek az alábbiak:
   A. teljes multikollinearitás esetén az $\mathbf{X'X}$ mátrix szinguláris;
   B. teljes multikollinearitás esetén a korrelációs mátrix szinguláris;
   C. a heteroszkedaszticitás általában az idősor alapján történő becsléseknél fordul elő;
   D. az autokorreláció általában a keresztmetszeti adatok alapján történő becsléseknél fordul elő.

7. Az indexekkel kapcsolatosan ismertek a következő összefüggések:
   A. a LASPEYRES-féle volumenindex mindig nagyobb a PAASCHE-féle volumenindexnél;
   B. a PAASCHE- és a LASPEYRES-féle volumenindexek hányadosa különbözhet a PAASCHE- és a LASPEYRES-féle árindexek hányadosától;
   C. az egyedi ár- és volumenindexek közötti lineáris korrelációs együttható nem lehet pozitív előjelű;
   D. a PAASCHE- és a LASPEYRES-féle indexek hányadosa általában egynél kisebb.

8. Ismertek az FAE mintával kapcsolatos összefüggések:
   A. a tapasztalati szórás a populáció szórásának torzítatlan becslése;
   B. a tapasztalati szórásnégyzet a populáció varianciájának torzítatlan becslése;
   C. a korrigált tapasztalati szórás a sokaság szórásának torzítatlan becslése;
   D. a korrigált tapasztalati szórásnégyzet a populáció szórásnégyzetének torzítatlan becslése.

9. Az éves exponenciális (analitikus) trendfüggvény $\hat{\beta}_1$ becsült paraméterének értelmezése:
   A. a vizsgált jelenség évente átlagosan $\hat{\beta}_1$ egységnyivel változik;
   B. a vizsgált jelenség évente átlagosan $\hat{\beta}_1$–szeresére változik;
   C. a vizsgált jelenség évente átlagosan $p = (\hat{\beta}_1 - 1) \cdot 100$ százalékkal változik;
   D. a vizsgált jelenség évente átlagosan $p = (1 - \hat{\beta}_1) \cdot 100$ százalékkal változik.

10. Jobboldali aszimmetria esetén a középértékek között (általában) fennállnak a következő összefüggések:
    A. a számtani átlag a módusznál kisebb;
    B. a számtani átlag a módusznál nagyobb;
    C. a medián a módusznál kisebb;
    D. a medián a módusznál nagyobb.

11. Egy 60 tagú statisztikai adatállomány csoportosításánál az osztályok (k) ideális számára vonatkozóan állíthatjuk, hogy:
    A. homogén adatok esetén k ideális értéke 6;
    B. heterogén adatok esetén k ideális értékét nem lehet meghatározni;
    C. heterogén adatok esetén k ideális értéke 6;
    D. k értékének meghatározásához semmilyen támpont sem ismert.

12. A felfelé és lefelé kumulált gyakoriságokra vonatkozóan igazak az alábbi összefüggések:
    A. az első lefelé kumulált gyakoriság az utolsó abszolút gyakorisággal egyenlő;
    B. az utolsó lefelé kumulált gyakoriság az utolsó abszolút gyakorisággal egyenlő;
    C. az első felfelé kumulált gyakoriság az első abszolút gyakorisággal egyenlő;
    D. a felfelé és a lefelé kumulált gyakoriságok között nem létezik semmilyen nevezetes összefüggés.

13. A mennyiségi sorokkal kapcsolatban tudjuk, hogy:
    A. az ogiva a relatív gyakorisági sorok grafikus ábrája;
    B. az ogiva a felfelé kumulált gyakoriságok grafikus ábrája;
    C. a gyakorisági görbe a gyakorisági poligon határesete;
    D. a hisztogram a gyakorisági sor kördiagramja.

14. Két ismérv közötti összefüggés számszerűsítésével kapcsolatban azt állíthatjuk, hogy:
    A. egy területi és egy mennyiségi ismérv között korrelációs kapcsolatról beszélünk;
    B. egy minőségi és egy alternatív ismérv között vegyes kapcsolatról beszélünk;
    C. két mennyiségi ismérv között rangkorrelációs kapcsolatról beszélünk;
    D. egy területi és egy minőségi ismérv között asszociációs kapcsolatról beszélünk.

15. Ugyanazon adatok számított átlagaira vonatkozóan ismertek a következő összefüggések:
    A. a mértani átlag a számtani átlagnál mindig kisebb;
    B. a harmonikus átlag a kvadratikus átlagnál mindig kisebb;
    C. néha egy kiszámított átlag kisebb is lehet az adatállomány legkisebb adatánál;
    D. bármilyen adatállomány esetén: $\bar{x}_h < \bar{x}_g$.

Tárgymutató

Képletgyűjtemény

7. Statisztikai minták módszere

(152) $k_{\mathrm{FAE}} = N^n$

(153) $k_{\mathrm{EV}} = \binom{N}{n}$

(154) $E(\bar{x}) = \mu_{\bar{x}} = \mu$

(155) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

(156) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

(157) $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

(158) $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$

(159) $\mu \mp z \cdot \sigma$

(160) $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$

(161) $\mu_{\bar{\mathrm{x}}} \mp z \cdot \sigma_{\bar{x}}$

(162) $n_j = \frac{n}{M}$ j=1,2,...,M

(163) $$n_j = n\frac{N_j}{\sum_{j=1}^{M} N_j} = n\frac{N_j}{N}$$

(164) $$n_j = n\frac{N_j\sigma_j}{\sum_{j=1}^{M} N_j\sigma_j}$$

8. Minta alapján történő becslések

(165) $$E(\hat{\Theta}) = \Theta$$

(166) $$Bs(\hat{\Theta}) = \Theta - E(\hat{\Theta})$$

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

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

(169) $$E(s^2) = \sigma^2$$

(170) $$E\left(s^2 \cdot \frac{N-1}{N}\right) = \sigma^2$$

(171) $$Mse(\hat{\Theta}) = Bs^2(\hat{\Theta}) + Se^2(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2$$

(172) $$Pr\left(\hat{\Theta}_{a(\alpha)} < \Theta < \hat{\Theta}_{f(\alpha)}\right) = 1 - \alpha$$

(173) $$Pr\left(\bar{x} - z_{(p)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{(p)}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

(174) $$\Delta = z_{(\mathrm{p})} \frac{\sigma}{\sqrt{n}}$$

(175) $$n = \frac{\left(z_{(\mathrm{p})}\sigma\right)^2}{\Delta^2}$$

(176) $$Pr\left(\bar{x} - t_{(\mathrm{p})}(\nu) \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\mathrm{p})}(\nu) \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

(177) $$Pr\left(\bar{x} - k\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{4}{9k^2} = 1 - \alpha$$

(178) $$Pr\left(\bar{x} - k\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2} = 1 - \alpha$$

(179) $$s_p = \sqrt{\frac{pq}{n-1}}$$

(180) $$s_p = \sqrt{\frac{pq}{n-1} \cdot \frac{N-n}{N-1}}$$

(181) $$Pr\left(p - z_{(\mathrm{p})} \cdot \sqrt{\frac{pq}{n-1}} < P < p + z_{(\mathrm{p})} \cdot \sqrt{\frac{pq}{n-1}}\right) = 1 - \alpha$$

(182) $$Pr\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(\nu)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(\nu)}\right) = 1 - \alpha$$

(183) $$Pr\left(\bar{x} - z_{(\mathrm{p})}\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} < \mu < \bar{x} + z_{(\mathrm{p})}\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha$$

(184) $$n = \frac{\left(z_{(\mathrm{p})}\sigma\right)^2}{\frac{\left(z_{(\mathrm{p})}\sigma\right)^2}{N} + \Delta^2}$$

(185) $$s_{\bar{x}}^2 = \frac{s^2}{n} \cdot \left(1 - \frac{n}{N}\right)$$

(186) $$E(s_{\bar{x}}^2) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \sigma_{\bar{x}}^2$$

(187) $$s_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

(188) $$\sigma_{\bar{x}} = \sqrt{\sum_{j=1}^{M} \frac{N_j^2}{N^2} \cdot \frac{\sigma_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j - 1}}$$

(189) $$\sigma_{\bar{x}} = \frac{\sigma_B}{\sqrt{n}}$$

(190) $$s_{\bar{x}} = \frac{\sqrt{\sum_{j=1}^{M} n_j s_j^2}}{n}$$

9. Hipotézisek vizsgálata

(191) $$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

(192) $$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

(193) $$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

(194) $$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}$$

(195) $$\chi^2 = n \cdot \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(g_{ij} - p_{i.} \cdot p_{.j}\right)^2}{p_{i.} \cdot p_{.j}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

(196) $$\chi^2 = n \cdot \left(\sum_{i=1}^{k} \frac{\left(g_i - P_i\right)^2}{P_i} \right) = \sum_{i=1}^{k} \frac{\left(f_i - f_i^*\right)^2}{f_i^*}$$

(197) $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(198) $$T = \frac{\bar{x}_1 - \bar{x}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

(199) $$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} x_{1i}^2 - n_1 \bar{x}_1^2 + \sum_{j=1}^{n_2} x_{2j}^2 - n_2 \bar{x}_2^2}{n_1 + n_2 - 2}$$

(200) $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(201) $$Z = \frac{p_1 - p_2}{\sqrt{\bar{p}\,\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(202) $$F = \frac{SSK/(M-1)}{SSB/(n-M)} = \frac{s_K^2}{s_B^2}$$

10. Dinamikus elemzés

(203) $$\bar{x}_k = \frac{\frac{x_1}{2} + \sum_{t=2}^{n-1} x_t + \frac{x_n}{2}}{n-1}$$

(204) $$\bar{d} = \frac{x_n - x_1}{n-1}$$

(205) $$\hat{y}_t = \frac{y_{t-k} + y_{t-k+1} + ... + y_t + ... + y_{t+k}}{2k+1}$$

(206) $$\hat{y}_t = \frac{\frac{y_{t-k}}{2} + y_{t-k+1} + ... + y_t + ... + y_{t+k-1} + \frac{y_{t+k}}{2}}{2k}$$

(207) $$\sum_{i=1}^{n} t_i = 0$$

(208) $$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$$

(209) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} t_i \cdot y_i}{\sum_{i=1}^{n} t_i^2}$$

(210) $$\log \hat{\beta}_0 = \frac{\sum_{i=1}^{n} \log y_i}{n}$$

(211) $$\log \hat{\beta}_1 = \frac{\sum_{i=1}^{n} t_i \cdot \log y_i}{\sum_{i=1}^{n} t_i^2}$$

(212) $$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_2 \sum_{i=1}^{n} t_i^2$$

(213) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i^2}$$

(214) $$\sum_{i=1}^{n} t_i^2 y_i = \hat{\beta}_0 \sum_{i=1}^{n} t_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} t_i^4$$

(215) $$\hat{y}_i = \frac{\hat{y}_{\max}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i}}$$

(216) $$\hat{y}_{\max} = \frac{2 \cdot \overline{Y}_{x_0} \cdot \overline{Y}_{x_0+m} \cdot \overline{Y}_{x_0+2m} - \overline{Y}_{x_0+m}^2 \cdot (\overline{Y}_{x_0} + \overline{Y}_{x_0+2m})}{\overline{Y}_{x_0} \cdot \overline{Y}_{x_0+2m} - \overline{Y}_{x_0+m}^2}$$

(217) $$\hat{\beta}_0 = \ln\left(\frac{\hat{y}_{\max} - \overline{Y}_{x_0}}{\overline{Y}_{x_0}}\right)$$

(218) $$\hat{\beta}_1 = \frac{1}{m} \ln\left(\frac{\overline{Y}_{x_0} \cdot (\hat{y}_{\max} - \overline{Y}_{x_0+m})}{\overline{Y}_{x_0+m} \cdot (\hat{y}_{\max} - \overline{Y}_{x_0})}\right)$$

(219) $$\hat{y}_{\max} = \frac{\sum_{i=1}^{n-1} y_i^4 \cdot \sum_{i=1}^{n-1} y_i^2 - \left(\sum_{i=1}^{n-1} y_i^3\right)^2 - \sum_{i=1}^{n-1} y_i y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^4 + \sum_{i=1}^{n-1} y_i^2 y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^3}{\sum_{i=1}^{n-1} y_i^2 y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^2 - \sum_{i=1}^{n-1} y_i y_{i+1} \cdot \sum_{i=1}^{n-1} y_i^3}$$

(220) $$\hat{z}_i = \left(\ln\left(\frac{\hat{y}_{\max} - y_i}{y_i} \right) \right)^{\wedge} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

(221) $$s_j^a = \frac{\sum_{i=1}^{n/p} \left(y_{ij} - \hat{y}_{ij}^a \right)}{n/p - 1} \qquad j = 1,2,...,p$$

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

(223) $$\tilde{s}_j^a = s_j^a - \bar{s}_j^a$$

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

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

(226) $$\tilde{s}_j^m = \frac{s_j^m}{\bar{s}_j^m}$$

11. Többváltozós regresszió- és korrelációszámítás

(227) $$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_m x_{im} + e_i$$

$$i = 1,2,...,n \qquad m+1 < n < N$$

(228) $$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & & x_{2m} \\ \vdots & & & \\ 1 & x_{n1} & & x_{nm} \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

(229) $$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

(230) $$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

(231) $$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

(232) $$SST = SSR + SSE$$

(233) $$r^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

(234) $$F = \frac{SSR / m}{SSE / (n - m - 1)}$$

(235) $$F = \frac{\hat{\beta}_j^2}{\mathrm{var}(\hat{\beta}_j)} \qquad j = 1,2,...,m$$

(236) $$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \frac{\mathbf{e}'\mathbf{e}}{n - m - 1} \cdot (\mathbf{X}'\mathbf{X})^{-1} = s_e^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$$

(237) $$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

(238) $$\mathbf{R} = \begin{bmatrix} 1 & r_{yx_1} & \cdots & r_{yx_m} \\ r_{x_1y} & 1 & & r_{x_1x_m} \\ \vdots & & & \\ r_{x_my} & r_{x_mx_1} & & 1 \end{bmatrix}$$

(239) $$\mathbf{C} = \begin{bmatrix} \sigma_y^2 & C_{yx_1} & \cdots & C_{yx_m} \\ C_{x_1y} & \sigma_{x_1}^2 & & C_{x_1x_m} \\ \vdots & & & \\ C_{x_my} & C_{x_mx_1} & & \sigma_{x_m}^2 \end{bmatrix}$$

(240) $$r_{yx_j \bullet x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m} = -\frac{\mathbf{R}^{-1}_{yx_j}}{\sqrt{\mathbf{R}^{-1}_{yy} \cdot \mathbf{R}^{-1}_{x_jx_j}}}$$

(241) $$r^2_{y \bullet x_1,x_2,\ldots,x_m} = 1 - \frac{1}{\mathbf{R}^{-1}_{yy}}$$

(242) $$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

(243) $$M = r^2_{y \bullet x_1,x_2,\ldots,x_m} - \sum_{j=1}^{m} \left(r^2_{y \bullet x_1,x_2,\ldots,x_m} - r^2_{y \bullet x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m} \right)$$

(244) $$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

(245) $$d \approx 2(1 - \hat{\rho})$$

(246) $$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{\Omega}$$

(247) $$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$$

(248) $$\operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

(249) $$s_e^2 = \frac{\mathbf{e}'\boldsymbol{\Omega}^{-1}\mathbf{e}}{n-m-1}$$

(250) $$\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \rho^{n-3} \\ \vdots & & & & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & & 1 \end{bmatrix}$$

(251) $$\boldsymbol{\Omega}^{-1} = \frac{1}{1-\rho^2} \cdot \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & & -\rho & 1 \end{bmatrix}$$

(252) $$\hat{\rho} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_i^2}$$

(253) $$\mathbf{P} = \begin{bmatrix} \frac{1}{x_{1j}} & 0 & \cdots & 0 \\ 0 & \frac{1}{x_{2j}} & & 0 \\ \vdots & & & \\ 0 & 0 & & \frac{1}{x_{nj}} \end{bmatrix}$$

(254) $\mathbf{\Omega}^{-1} = \mathbf{P}'\mathbf{P} = \mathbf{P}^2$

(255) $\widetilde{x}_{ij} = \dfrac{x_{ij} - \overline{x}_j}{s_j} \qquad i = 1,2,...,n \qquad j = 1,2,...,m$

(256) $\mathbf{C} = \widetilde{\mathbf{X}}\mathbf{U}$

(257) $c_{ij} = u_{1j}\widetilde{x}_{i1} + u_{2j}\widetilde{x}_{i2} + \ldots + u_{mj}\widetilde{x}_{im}$

(258) $\widetilde{x}_{ij} = u_{j1}c_{i1} + u_{j2}c_{i2} + \ldots + u_{jm}c_{im}$

(259) $a_{kl} = u_{kl}\sqrt{\lambda_l} \qquad k,l = 1,2,...,m$

(260) $a_{kl} = r_{\widetilde{x}_k c_l} = r_{x_k c_l}$

(261) $h_k^{(w)} = \sum_{l=1}^{w} a_{kl}^2 \qquad 1 \le w \le m$

Statisztikai táblázatok

I. TÁBLÁZAT

Standard normális eloszlású változó eloszlásfüggvényének értékei
(kétoldali próbákhoz)

z	0	1	2	3	4	5	6	7	8	9
1,0	68269	68750	69227	69699	70166	70628	71086	71538	71986	72429
1,1	72867	73300	73729	74152	74571	74986	75395	75800	76200	76595
1,2	76986	77372	77753	78130	78502	78870	79233	79592	79945	80295
1,3	80640	80980	81316	81648	81975	82298	82617	82931	83241	83547
1,4	83849	84146	84439	84728	85013	85294	85571	85844	86113	86378
1,5	86639	86896	87149	87398	87644	87886	88124	88358	88589	88817
1,6	89040	89260	89477	89690	89899	90106	90309	90508	90704	90897
1,7	91087	91273	91457	91637	91814	91988	92159	92327	92492	92655
1,8	92814	92970	93124	93275	93423	93569	93711	93852	93989	94124
1,9	94257	94387	94514	94639	94762	94882	95000	95116	95230	95341
2,0	95450	95557	95662	95764	95865	95964	96060	96155	96247	96338
2,1	96427	96514	96599	96683	96765	96844	96923	96999	97074	97148
2,2	97219	97289	97358	97425	97491	97555	97618	97679	97739	97798
2,3	97855	97911	97966	98019	98072	98123	98173	98221	98269	98315
2,4	98360	98405	98448	98490	98531	98571	98611	98649	98686	98723
2,5	98758	98793	98826	98859	98891	98923	98953	98983	99012	99040
2,6	99068	99095	99121	99146	99171	99195	99219	99241	99264	99285
2,7	99307	99327	99347	99367	99386	99404	99422	99439	99456	99473
2,8	99489	99505	99520	99535	99549	99563	99576	99590	99602	99615
2,9	99627	99639	99650	99661	99672	99682	99692	99702	99712	99721
3,0	99730	99739	99747	99755	99763	99771	99779	99786	99793	99800
3,1	99806	99813	99819	99825	99831	99837	99842	99848	99853	99858
3,2	99863	99867	99872	99876	99880	99885	99889	99892	99896	99900
3,3	99903	99907	99910	99913	99916	99919	99922	99925	99928	99930
3,4	99933	99935	99937	99940	99942	99944	99946	99948	99950	99952

Megjegyzés: a táblázatban szereplő számok törtrészek (mindegyik előtt ’0,’ áll).

II. TÁBLÁZAT

Standard normális eloszlású változó eloszlásfüggvényének értékei
(egyoldali próbákhoz)

z	0	1	2	3	4	5	6	7	8	9
1,0	84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1,1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,3	90320	90490	90658	90824	90988	91149	91308	91466	91621	91774
1,4	91924	92073	92220	92364	92507	92647	92785	92922	93056	93189
1,5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1,6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,7	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1,8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2,0	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3,0	99865	99869	99874	99878	99882	99886	99889	99893	99896	99900
3,1	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,2	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,3	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,4	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976

Megjegyzés: a táblázatban szereplő számok törtrészek (mindegyik előtt ’0,’ áll).

III. TÁBLÁZAT

A STUDENT-féle t-eloszlású változó eloszlásának kvantilis értékei
(kétoldali próbákhoz)

ν	0,9	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
5	2,0150	2,0978	2,1910	2,2974	2,4216	2,5706	2,7565	3,0029	3,3649	4,0321
6	1,9432	2,0192	2,1043	2,2011	2,3133	2,4469	2,6122	2,8289	3,1427	3,7074
7	1,8946	1,9662	2,0460	2,1365	2,2409	2,3646	2,5168	2,7146	2,9979	3,4995
8	1,8595	1,9280	2,0042	2,0902	2,1892	2,3060	2,4490	2,6338	2,8965	3,3554
9	1,8331	1,8992	1,9727	2,0554	2,1504	2,2622	2,3984	2,5738	2,8214	3,2498
10	1,8125	1,8768	1,9481	2,0283	2,1202	2,2281	2,3593	2,5275	2,7638	3,1693
11	1,7959	1,8588	1,9284	2,0067	2,0961	2,2010	2,3281	2,4907	2,7181	3,1058
12	1,7823	1,8440	1,9123	1,9889	2,0764	2,1788	2,3027	2,4607	2,6810	3,0545
13	1,7709	1,8317	1,8989	1,9742	2,0600	2,1604	2,2816	2,4358	2,6503	3,0123
14	1,7613	1,8213	1,8875	1,9617	2,0462	2,1448	2,2638	2,4149	2,6245	2,9768
15	1,7531	1,8123	1,8777	1,9509	2,0343	2,1315	2,2485	2,3970	2,6025	2,9467
16	1,7459	1,8046	1,8693	1,9417	2,0240	2,1199	2,2354	2,3815	2,5835	2,9208
17	1,7396	1,7978	1,8619	1,9335	2,0150	2,1098	2,2238	2,3681	2,5669	2,8982
18	1,7341	1,7918	1,8553	1,9264	2,0071	2,1009	2,2137	2,3562	2,5524	2,8784
19	1,7291	1,7864	1,8495	1,9200	2,0000	2,0930	2,2047	2,3457	2,5395	2,8609
20	1,7247	1,7816	1,8443	1,9143	1,9937	2,0860	2,1967	2,3362	2,5280	2,8453
21	1,7207	1,7773	1,8397	1,9092	1,9880	2,0796	2,1894	2,3278	2,5176	2,8314
22	1,7171	1,7734	1,8354	1,9045	1,9829	2,0739	2,1829	2,3202	2,5083	2,8188
23	1,7139	1,7699	1,8316	1,9003	1,9783	2,0687	2,1770	2,3132	2,4999	2,8073
24	1,7109	1,7667	1,8281	1,8965	1,9740	2,0639	2,1715	2,3069	2,4922	2,7970
25	1,7081	1,7637	1,8248	1,8929	1,9701	2,0595	2,1666	2,3011	2,4851	2,7874
26	1,7056	1,7610	1,8219	1,8897	1,9665	2,0555	2,1620	2,2958	2,4786	2,7787
27	1,7033	1,7585	1,8191	1,8867	1,9632	2,0518	2,1578	2,2909	2,4727	2,7707
28	1,7011	1,7561	1,8166	1,8839	1,9601	2,0484	2,1539	2,2864	2,4671	2,7633
29	1,6991	1,7540	1,8142	1,8813	1,9573	2,0452	2,1503	2,2822	2,4620	2,7564

IV. TÁBLÁZAT

A STUDENT-féle t-eloszlású változó eloszlásának kvantilis értékei
(egyoldali próbákhoz)

ν	0,9	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
5	1,4759	1,5579	1,6493	1,7529	1,8727	2,0150	2,1910	2,4216	2,7565	3,3649
6	1,4398	1,5172	1,6033	1,7002	1,8117	1,9432	2,1043	2,3133	2,6122	3,1427
7	1,4149	1,4894	1,5718	1,6643	1,7702	1,8946	2,0460	2,2409	2,5168	2,9979
8	1,3968	1,4691	1,5489	1,6383	1,7402	1,8595	2,0042	2,1892	2,4490	2,8965
9	1,3830	1,4537	1,5315	1,6185	1,7176	1,8331	1,9727	2,1504	2,3984	2,8214
10	1,3722	1,4416	1,5179	1,6031	1,6998	1,8125	1,9481	2,1202	2,3593	2,7638
11	1,3634	1,4318	1,5069	1,5906	1,6856	1,7959	1,9284	2,0961	2,3281	2,7181
12	1,3562	1,4237	1,4979	1,5804	1,6739	1,7823	1,9123	2,0764	2,3027	2,6810
13	1,3502	1,4170	1,4903	1,5718	1,6641	1,7709	1,8989	2,0600	2,2816	2,6503
14	1,3450	1,4113	1,4839	1,5646	1,6558	1,7613	1,8875	2,0462	2,2638	2,6245
15	1,3406	1,4063	1,4784	1,5583	1,6487	1,7531	1,8777	2,0343	2,2485	2,6025
16	1,3368	1,4021	1,4736	1,5529	1,6425	1,7459	1,8693	2,0240	2,2354	2,5835
17	1,3334	1,3983	1,4694	1,5482	1,6370	1,7396	1,8619	2,0150	2,2238	2,5669
18	1,3304	1,3950	1,4656	1,5439	1,6322	1,7341	1,8553	2,0071	2,2137	2,5524
19	1,3277	1,3920	1,4623	1,5402	1,6280	1,7291	1,8495	2,0000	2,2047	2,5395
20	1,3253	1,3894	1,4593	1,5369	1,6242	1,7247	1,8443	1,9937	2,1967	2,5280
21	1,3232	1,3870	1,4567	1,5338	1,6207	1,7207	1,8397	1,9880	2,1894	2,5176
22	1,3212	1,3848	1,4542	1,5311	1,6176	1,7171	1,8354	1,9829	2,1829	2,5083
23	1,3195	1,3828	1,4520	1,5286	1,6148	1,7139	1,8316	1,9783	2,1770	2,4999
24	1,3178	1,3810	1,4500	1,5263	1,6122	1,7109	1,8281	1,9740	2,1715	2,4922
25	1,3163	1,3794	1,4482	1,5242	1,6098	1,7081	1,8248	1,9701	2,1666	2,4851
26	1,3150	1,3778	1,4464	1,5223	1,6076	1,7056	1,8219	1,9665	2,1620	2,4786
27	1,3137	1,3764	1,4449	1,5205	1,6056	1,7033	1,8191	1,9632	2,1578	2,4727
28	1,3125	1,3751	1,4434	1,5189	1,6037	1,7011	1,8166	1,9601	2,1539	2,4671
29	1,3114	1,3739	1,4421	1,5174	1,6020	1,6991	1,8142	1,9573	2,1503	2,4620

V. TÁBLÁZAT

A χ^2-eloszlású változó eloszlásának kvantilis értékei

ν	0,005	0,01	0,02	0,025	0,5	0,95	0,975	0,98	0,99	0,995
2	0,010	0,020	0,040	0,051	1,386	5,991	7,378	7,824	9,210	10,597
3	0,072	0,115	0,185	0,216	2,366	7,815	9,348	9,837	11,345	12,838
4	0,207	0,297	0,429	0,484	3,357	9,488	11,143	11,668	13,277	14,860
5	0,412	0,554	0,752	0,831	4,351	11,070	12,832	13,388	15,086	16,750
6	0,676	0,872	1,134	1,237	5,348	12,592	14,449	15,033	16,812	18,548
7	0,989	1,239	1,564	1,690	6,346	14,067	16,013	16,622	18,475	20,278
8	1,344	1,647	2,032	2,180	7,344	15,507	17,535	18,168	20,090	21,955
9	1,735	2,088	2,532	2,700	8,343	16,919	19,023	19,679	21,666	23,589
10	2,156	2,558	3,059	3,247	9,342	18,307	20,483	21,161	23,209	25,188
11	2,603	3,053	3,609	3,816	10,341	19,675	21,920	22,618	24,725	26,757
12	3,074	3,571	4,178	4,404	11,340	21,026	23,337	24,054	26,217	28,300
13	3,565	4,107	4,765	5,009	12,340	22,362	24,736	25,471	27,688	29,819
14	4,075	4,660	5,368	5,629	13,339	23,685	26,119	26,873	29,141	31,319
15	4,601	5,229	5,985	6,262	14,339	24,996	27,488	28,259	30,578	32,801
16	5,142	5,812	6,614	6,908	15,338	26,296	28,845	29,633	32,000	34,267
17	5,697	6,408	7,255	7,564	16,338	27,587	30,191	30,995	33,409	35,718
18	6,265	7,015	7,906	8,231	17,338	28,869	31,526	32,346	34,805	37,156
19	6,844	7,633	8,567	8,907	18,338	30,144	32,852	33,687	36,191	38,582
20	7,434	8,260	9,237	9,591	19,337	31,410	34,170	35,020	37,566	39,997
21	8,034	8,897	9,915	10,283	20,337	32,671	35,479	36,343	38,932	41,401
22	8,643	9,542	10,600	10,982	21,337	33,924	36,781	37,659	40,289	42,796
23	9,260	10,196	11,293	11,689	22,337	35,172	38,076	38,968	41,638	44,181
24	9,886	10,856	11,992	12,401	23,337	36,415	39,364	40,270	42,980	45,558

V. TÁBLÁZAT (folytatás)

A χ^2-eloszlású változó eloszlásának kvantilis értékei

ν	0,005	0,01	0,02	0,025	0,5	0,95	0,975	0,98	0,99	0,995
25	10,52	11,52	12,70	13,12	24,34	37,65	40,65	41,57	44,31	46,93
26	11,16	12,20	13,41	13,84	25,34	38,89	41,92	42,86	45,64	48,29
27	11,81	12,88	14,13	14,57	26,34	40,11	43,19	44,14	46,96	49,65
28	12,46	13,56	14,85	15,31	27,34	41,34	44,46	45,42	48,28	50,99
29	13,12	14,26	15,57	16,05	28,34	42,56	45,72	46,69	49,59	52,34
30	13,79	14,95	16,31	16,79	29,34	43,77	46,98	47,96	50,89	53,67
35	17,19	18,51	20,03	20,57	34,34	49,80	53,20	54,24	57,34	60,27
40	20,71	22,16	23,84	24,43	39,34	55,76	59,34	60,44	63,69	66,77
45	24,31	25,90	27,72	28,37	44,34	61,66	65,41	66,56	69,96	73,17
50	27,99	29,71	31,66	32,36	49,33	67,50	71,42	72,61	76,15	79,49
55	31,73	33,57	35,66	36,40	54,33	73,31	77,38	78,62	82,29	85,75
60	35,53	37,48	39,70	40,48	59,33	79,08	83,30	84,58	88,38	91,95
65	39,38	41,44	43,78	44,60	64,33	84,82	89,18	90,50	94,42	98,10
70	43,28	45,44	47,89	48,76	69,33	90,53	95,02	96,39	100,43	104,21
75	47,21	49,48	52,04	52,94	74,33	96,22	100,84	102,24	106,39	110,29
80	51,17	53,54	56,21	57,15	79,33	101,88	106,63	108,07	112,33	116,32
85	55,17	57,63	60,41	61,39	84,33	107,52	112,39	113,87	118,24	122,32
90	59,20	61,75	64,63	65,65	89,33	113,15	118,14	119,65	124,12	128,30
95	63,25	65,90	68,88	69,92	94,33	118,75	123,86	125,40	129,97	134,25
100	67,33	70,06	73,14	74,22	99,33	124,34	129,56	131,14	135,81	140,17

VI. TÁBLÁZAT

Az F-eloszlású változó eloszlásának kvantilis értékei

$\alpha = 0{,}05$

ν_2	ν_1 = 1	2	3	4	5	6	7	8	9
1	161,446	199,499	215,707	224,583	230,160	233,988	236,767	238,884	240,543
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211
35	4,121	3,267	2,874	2,641	2,485	2,372	2,285	2,217	2,161
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124
45	4,057	3,204	2,812	2,579	2,422	2,308	2,221	2,152	2,096
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073

VI. TÁBLÁZAT (folytatás)

Az F-eloszlású változó eloszlásának kvantilis értékei

$\alpha = 0{,}05$

ν_2 \ ν_1	10	15	20	25	30	35	40	45	50
1	241,882	245,949	248,016	249,260	250,096	250,693	251,144	251,493	251,774
2	19,396	19,429	19,446	19,456	19,463	19,467	19,471	19,473	19,476
3	8,785	8,703	8,660	8,634	8,617	8,604	8,594	8,587	8,581
4	5,964	5,858	5,803	5,769	5,746	5,729	5,717	5,707	5,699
5	4,735	4,619	4,558	4,521	4,496	4,478	4,464	4,453	4,444
6	4,060	3,938	3,874	3,835	3,808	3,789	3,774	3,763	3,754
7	3,637	3,511	3,445	3,404	3,376	3,356	3,340	3,328	3,319
8	3,347	3,218	3,150	3,108	3,079	3,059	3,043	3,030	3,020
9	3,137	3,006	2,936	2,893	2,864	2,842	2,826	2,813	2,803
10	2,978	2,845	2,774	2,730	2,700	2,678	2,661	2,648	2,637
11	2,854	2,719	2,646	2,601	2,570	2,548	2,531	2,517	2,507
12	2,753	2,617	2,544	2,498	2,466	2,443	2,426	2,412	2,401
13	2,671	2,533	2,459	2,412	2,380	2,357	2,339	2,325	2,314
14	2,602	2,463	2,388	2,341	2,308	2,284	2,266	2,252	2,241
15	2,544	2,403	2,328	2,280	2,247	2,223	2,204	2,190	2,178
16	2,494	2,352	2,276	2,227	2,194	2,169	2,151	2,136	2,124
17	2,450	2,308	2,230	2,181	2,148	2,123	2,104	2,089	2,077
18	2,412	2,269	2,191	2,141	2,107	2,082	2,063	2,048	2,035
19	2,378	2,234	2,155	2,106	2,071	2,046	2,026	2,011	1,999
20	2,348	2,203	2,124	2,074	2,039	2,013	1,994	1,978	1,966
25	2,236	2,089	2,007	1,955	1,919	1,892	1,872	1,855	1,842
30	2,165	2,015	1,932	1,878	1,841	1,813	1,792	1,775	1,761
35	2,114	1,963	1,878	1,824	1,786	1,757	1,735	1,718	1,703
40	2,077	1,924	1,839	1,783	1,744	1,715	1,693	1,675	1,660
45	2,049	1,895	1,808	1,752	1,713	1,683	1,660	1,642	1,626
50	2,026	1,871	1,784	1,727	1,687	1,657	1,634	1,615	1,599

VII. TÁBLÁZAT

Az F-eloszlású változó eloszlásának kvantilis értékei

$\alpha = 0{,}01$

	ν_1								
ν_2	1	2	3	4	5	6	7	8	9
2	98,502	99,000	99,164	99,251	99,302	99,331	99,357	99,375	99,390
3	34,116	30,816	29,457	28,710	28,237	27,911	27,671	27,489	27,345
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911
9	10,562	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780
17	8,400	6,112	5,185	4,669	4,336	4,101	3,927	3,791	3,682
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067
35	7,419	5,268	4,396	3,908	3,592	3,368	3,200	3,069	2,963
40	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,888
45	7,234	5,110	4,249	3,767	3,454	3,232	3,066	2,935	2,830
50	7,171	5,057	4,199	3,720	3,408	3,186	3,020	2,890	2,785

VII. TÁBLÁZAT (folytatás)

Az F-eloszlású változó eloszlásának kvantilis értékei

$\alpha = 0{,}01$

	ν_1								
ν_2	10	15	20	25	30	35	40	45	50
2	99,397	99,433	99,448	99,459	99,466	99,470	99,477	99,477	99,477
3	27,228	26,872	26,690	26,579	26,504	26,451	26,411	26,379	26,354
4	14,546	14,198	14,019	13,911	13,838	13,785	13,745	13,714	13,690
5	10,051	9,722	9,553	9,449	9,379	9,329	9,291	9,262	9,238
6	7,874	7,559	7,396	7,296	7,229	7,180	7,143	7,115	7,091
7	6,620	6,314	6,155	6,058	5,992	5,944	5,908	5,880	5,858
8	5,814	5,515	5,359	5,263	5,198	5,151	5,116	5,088	5,065
9	5,257	4,962	4,808	4,713	4,649	4,602	4,567	4,539	4,517
10	4,849	4,558	4,405	4,311	4,247	4,201	4,165	4,138	4,115
11	4,539	4,251	4,099	4,005	3,941	3,895	3,860	3,832	3,810
12	4,296	4,010	3,858	3,765	3,701	3,654	3,619	3,592	3,569
13	4,100	3,815	3,665	3,571	3,507	3,461	3,425	3,398	3,375
14	3,939	3,656	3,505	3,412	3,348	3,301	3,266	3,238	3,215
15	3,805	3,522	3,372	3,278	3,214	3,167	3,132	3,104	3,081
16	3,691	3,409	3,259	3,165	3,101	3,054	3,018	2,990	2,967
17	3,593	3,312	3,162	3,068	3,003	2,956	2,920	2,892	2,869
18	3,508	3,227	3,077	2,983	2,919	2,871	2,835	2,807	2,784
19	3,434	3,153	3,003	2,909	2,844	2,797	2,761	2,732	2,709
20	3,368	3,088	2,938	2,843	2,778	2,731	2,695	2,666	2,643
25	3,129	2,850	2,699	2,604	2,538	2,490	2,453	2,424	2,400
30	2,979	2,700	2,549	2,453	2,386	2,337	2,299	2,269	2,245
35	2,876	2,597	2,445	2,348	2,281	2,231	2,193	2,162	2,137
40	2,801	2,522	2,369	2,271	2,203	2,153	2,114	2,083	2,058
45	2,743	2,464	2,311	2,213	2,144	2,093	2,054	2,023	1,997
50	2,698	2,419	2,265	2,167	2,098	2,046	2,007	1,975	1,949

VIII. TÁBLÁZAT

DURBIN-WATSON-féle próba jobboldali kritikus értékei

$\alpha = 0{,}05$

	$m=1$		$m=2$		$m=3$		$m=4$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900
18	1,158	1,391	1,046	1,535	0,933	1,690	0,820	1,872
19	1,180	1,401	1,074	1,536	0,967	1,685	0,859	1,848
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828
21	1,221	1,420	1,125	1,538	1,026	1,669	0,927	1,812
22	1,239	1,429	1,147	1,541	1,053	1,664	0,958	1,797
23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,785
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727
65	1,567	1,629	1,536	1,662	1,503	1,698	1,471	1,731
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735
75	1,598	1,652	1,571	1,680	1,543	1,709	1,515	1,739
80	1,611	1,662	1,586	1,688	1,560	1,715	1,534	1,743

Forrás: Econometrica, 45, Nov. 1977.

IX. TÁBLÁZAT

DURBIN-WATSON-féle próba jobboldali kritikus értékei

$\alpha = 0{,}01$

	$m=1$		$m=2$		$m=3$		$m=4$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0,811	1,070	0,700	1,252	0,591	1,464	0,488	1,704
16	0,844	1,086	0,737	1,252	0,633	1,446	0,532	1,663
17	0,874	1,102	0,772	1,255	0,672	1,432	0,574	1,630
18	0,902	1,118	0,805	1,259	0,708	1,422	0,613	1,604
19	0,928	1,132	0,835	1,265	0,742	1,415	0,650	1,584
20	0,952	1,147	0,863	1,271	0,773	1,411	0,685	1,567
21	0,975	1,161	0,890	1,277	0,803	1,408	0,718	1,554
22	0,997	1,174	0,914	1,284	0,831	1,407	0,748	1,543
23	1,018	1,187	0,936	1,291	0,858	1,407	0,777	1,534
24	1,037	1,199	0,960	1,298	0,882	1,407	0,805	1,528
25	1,055	1,211	0,981	1,305	0,906	1,409	0,831	1,523
26	1,072	1,222	1,001	1,312	0,928	1,411	0,855	1,518
27	1,089	1,233	1,019	1,319	0,949	1,413	0,878	1,515
28	1,104	1,244	1,037	1,325	0,969	1,415	0,900	1,513
29	1,119	1,254	1,054	1,332	0,988	1,418	0,921	1,512
30	1,133	1,263	1,070	1,339	1,006	1,421	0,941	1,511
35	1,195	1,307	1,140	1,370	1,085	1,439	1,028	1,512
40	1,246	1,344	1,198	1,398	1,148	1,457	1,098	1,518
45	1,288	1,376	1,245	1,423	1,201	1,474	1,156	1,528
50	1,324	1,403	1,285	1,446	1,245	1,491	1,205	1,538
55	1,356	1,427	1,320	1,466	1,284	1,506	1,247	1,548
60	1,383	1,449	1,350	1,484	1,317	1,520	1,283	1,558
65	1,407	1,468	1,377	1,500	1,346	1,534	1,315	1,568
70	1,429	1,485	1,400	1,515	1,372	1,546	1,343	1,578
75	1,448	1,501	1,422	1,529	1,395	1,557	1,368	1,587
80	1,466	1,515	1,441	1,541	1,416	1,556	1,390	1,595

Irodalom

Denkinger G.: Valószínűségszámítás, Nemzeti Tankönyvkiadó, Budapest, 1997.

Éltető Ö.-Meszéna Gy.-Ziermann M.: Sztochasztikus módszerek és modellek, Közgazdasági és Jogi Könyvkiadó, Budapest, 1982.

Greene, W.H.: Econometric Analysis, Macmillan Publishing Company, New York, 1993.

Hunyadi L.-Mundruczó Gy.-Vita L.: Statisztika, Aula Kiadó, Budapest, 1996.

Kerékgyártó Gy.-Mundruczó Gy.: Statisztikai módszerek a gazdasági elemzésben, Aula Kiadó, Budapest, 1994.

Köves P.–Párniczky G.: Általános Statisztika, Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

Lukács O.: Matematikai statisztika, Műszaki Könyvkiadó, Budapest, 1987.

Meszéna Gy.-Ziermann M.: Valószínűségelmélet és matematikai statisztika, Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

Mundruczó Gy.: Alkalmazott regressziószámítás, Akadémiai Kiadó, Budapest, 1981.

Ramanathan, R.: Introductory Econometrics (with applications), Harcourt Brace, Orlando, 1995.

Spiegel, M. R.: Statisztika (elmélet és gyakorlat), Panem-McGraw-Hill, Budapest, 1995.

Sváb J.: Többváltozós módszerek a biometriában, Mezőgazdasági Könyvkiadó, Budapest, 1979.