Elméleti összefoglalók

dr. Kovács Péter

Ebben az anyagban az alkalmazott módszerek rövid leírása található. Elsősorban azokat az elemeket fejtettük ki jobban, amelyek nem részei az alapstatisztikai kurzusoknak. A leírások során törekedtem arra, hogy túl sok képletet, összefüggést ne közöljek.

1. Adatállományok létrehozása, kezelése

Statisztikai elemzéseknél az adatok tárolása úgynevezett adatmátrixokban történik. Ez gyakorlatilag egy olyan táblázatot jelent, ahol mindegyik változót egy oszlop reprezentálja, míg az egyes rekordokban (sorokban) az egyes megfigyelések találhatóak.
A statisztikai szoftverek használata esetén is lehetőségünk van adatállomány megnyitására, módosítására, importálására, mentésére, illetve az adatok begépelésére.
Ha szöveges állományból szeretnénk egy adatállományt beolvasni, akkor egy varázslón kell végig haladnunk, melynek során meg kell adnunk a szeparátor jelet (tagoló karaktert), a változók típusait, stb.

Amennyiben új változót hozunk létre, meg kell adnunk a **változó fontosabb tulajdonságai**t, így például:

- a változó nevét, amivel azonosítani tudjuk. Ez lehet rövidítés is. Ez egyfajta munkanévnek fogható fel.
- A változó címkéjét. Ez a név fog az elemzések kimenetében, végeredményében szerepelni.
- A változó típusát. Ez nagyon sokféle lehet, például szöveg, szám, pénznem, stb. A változó típusa meghatározza a változóval elvégezhető műveleteket.
- A változó mérési szintjét. A mérési szint határozza meg azt, hogy a változót milyen elemzésekbe, és ezeken belül milyen szerepkörbe lehet bevonni.
- A hiányzó adatok helyettesítési körét. Ha egy megfigyelés során hiányzik egy adatunk, akkor ezt egy általunk meghatározható kóddal lehet helyettesíteni.
- Az esetek kódolását, ha van. Például, ha az elemzésekben szerepel változóként a nem, akkor beállíthatjuk azt, hogy a 0 szám(jegy) a nőket, míg az 1 szám(jegy) a férfiakat jelentse.
- A változó formai beállításait. Például hány tizedes jegy pontosságú értékekkel dolgozunk stb.

Néhány elemzési módszer esetében nem az eredeti adatokkal, hanem ezek egy változatával dolgozunk. A legfontosabbak adat-átalakítások az alábbiak.

- **Transzformálás**. Ekkor a változó értékeit képletek és függvények segítségével új értékekké alakítjuk át.
- **Standardizálás**. Ez a transzformálás speciális esete. Ennek lényege az, hogy az adatokat mértékegységtől függetlenné tesszük.
- **Átkódolás**. Ekkor vagy megváltoztathatjuk az egyes értékeket, például a férfiak kódját egyről kettőre változtatjuk, vagy pedig csoportokat készíthetünk valamely adattartományból.

2. Leíró statisztikai eljárások

A leíró statisztika az információtömörítés legegyszerűbb formájának tekinthető. Gyakorlatilag ide tartozik a megfigyelt egyedek egy változó (ismérv) szerinti eloszlásának jellemzése: diagramok, táblázatok készítése, az átlagos tendenciák, azaz a középértékek és a szóródás jellemzése.

Statisztikai megfigyelések típusai

Teljes körű adatgyűjtések során az egész vizsgált sokaságot megfigyelik. A teljes körű megfigyelések általában nagyon költségesek, sokszor lehetetlen megvalósítani.
A teljes körű adatgyűjtés tipikus példája a népszámlálás. A nemzetközi gyakorlat szerint általában 10 évenként tartanak népszámlálást. Legutóbb 2001-ben volt hazánkban ilyen összeírás.

A teljes körű megfigyelés helyett általában a részleges megfigyeléseket használjuk a gyakorlatban. **Részleges megfigyelések** során csak a sokaság egy részét, néhány egyedét figyeljük meg. A részleges megfigyelések főbb típusa a monográfia, a reprezentatív megfigyelés, illetve egyéb részleges megfigyelések.

A **monográfia** a sokaság néhány kiemelt, fontos egyedének a vizsgálatát jelenti.

A **reprezentatív** megfigyelés során a megfigyelt egyedek kiválasztása különböző kritériumok alapján történik, úgy, hogy a megfigyelt egyedek tulajdonságai tükrözik az alapsokaság tulajdonságait. Ekkor a vizsgált sokaságot **alapsokaságnak**, míg a megfigyelt részsokaságot pedig **mintának** nevezzük. A minta csak véges elemszámú lehet.

A reprezentatív mintavétel megvalósításának módja **a véletlen mintavétel**. Ez azt jelenti, hogy az alapsokaság mindegyik egyede valamilyen valószínűséggel, eséllyel kerülhet a mintába.

Ha a mintába kerülő elemeket visszatevéssel választjuk ki, akkor az alapsokaság mindegyik egyede ugyanakkora valószínűséggel kerülhet be a mintába. Ekkor **független, azonos eloszlású mintát** (**FAE-mintát**) kapunk. Ekkor az alapsokaság egyedei akár többször is bekerülhetnek a mintába. Ez néha problémát okozhat akkor, ha valamilyen szélsőséges elem többször bekerül a mintába.

Ha a mintába kerülő elemeket visszatevés nélkül választjuk ki, akkor **egyszerű véletlen mintát** (**EV-mintát**) kapunk. Ekkor az alapsokaság egyedei csak egyszer kerülhetnek a mintába. Ezért az EV-minta jobbnak tekinthető az FAE-mintánál. Egy vizsgált alapsokaságból vehető, adott elemszámú összes lehetséges FAE-minták száma nagyobb az EV-minták számánál.

Az előző két véletlen mintához viszonyítva az alapsokaság jobb reprezentációját kapjuk a rétegezett minta alkalmazásával. Amennyiben egy heterogén sokaságot megközelítőleg homogén részsokaságokra tudunk bontani, akkor alkalmazhatjuk a rétegzett mintavételt. A **rétegzett mintát** (**R-mintát**) úgy kapjuk meg, hogy

minden rétegből (részsokaságból) EV-mintát veszünk. Az egyes rétegekből vett EV-minták elemszámainak meghatározására két módszert említek meg.
Egyenletes elosztás esetén mindegyik rétegből ugyanannyi elemet válogatunk a mintába.

Arányos elosztás esetén a rétegek elemszámának sokaságbeli arányát figyelembe véve történik a kiválasztás.

Csoportos (CS) mintavétel esetén az alapsokaságot heterogén csoportokra bontjuk szét. Ezután a csoportok közül veszünk EV-mintát. A kiválasztott csoportokat pedig teljes körűen megfigyeljük.

A **többlépcsős (TL) mintavétel** az előző eljárások kombinálását jelenti. Például egy kétlépcsős mintavétel esetén először csoportos mintavételt alkalmazunk, majd a kiválasztott csoportokat nem teljes körűen figyeljük meg, hanem ezekből EV-mintákat veszünk.

Példa
Egy áruházlánc közérzetjavító intézkedéseket szeretne végrehajtani. Ehhez meg szeretnék kérdezni a dolgozóik véleményét is.
Amennyiben minden dolgozó véleményét megkérdezik, akkor teljes körű megfigyelésről van szó.
Amennyiben a dolgozók közül reprezentatív véletlen mintát vesznek akkor az előbbi eljárásokra az alábbi példa adható.
EV-minta: véletlenszerűen választunk ki néhány dolgozót.
FAE-minta: véletlenszerűen választunk ki néhány dolgozót.
R-minta: az alkalmazottakat beosztásuk szerint csoportosítását tekintve (pl. pénztáros, eladó, osztályvezető, stb.) minden egyes csoportból választunk elemeket a mintába. Egyenletes elosztás esetén minden csoportból ugyanannyi főt kérdezünk meg, míg arányos elosztás esetén a mintában ugyanannyi lesz minden csoport aránya, mint az alapsokaságban.
CS-minta: véletlenszerűen kiválasztunk néhány áruházat, majd az ott dolgozók mindegyikét megkérdezzük.
TL-minta: véletlenszerűen kiválasztunk néhány áruházat, majd az ott dolgozókból EV-mintát veszünk.

Az adatgyűjtések, megfigyelések hibákkal járnak.
A **nemmintavételi hibák** azok a hibák, amelyek mind a teljes, mind a részleges megfigyeléseknél felléphetnek. Ezek matematikai eszközökkel nem kezelhetőek. Ilyenek például a definíciós hiba (rossz kérdőívszerkesztés), a válaszadási hiba (téves adat közlése), a végrehajtási hiba (rossz lekérdezés), az adatrögzítési hiba.
A **mintavételi** hiba a részleges megfigyelésből fakadó hiba. Ez a típus matematikailag kezelhető.

A számítógépes statisztikai eljárások többsége feltételezi, hogy **minta esetleg minták alapján** dolgozunk. Ekkor gyakorlatilag bármi, amit leíró statisztikai eszközökkel állítunk kizárólag az adott mintára vonatkozik. A gyakorlatban viszont sokszor nem pusztán az adott minta, hanem az egész alapsokaság jellemzői érdekelnek bennünket. Tehát – egy minta alapján – a vizsgált sokaság valamely jellemzőjét kell meghatározni, jellemezni. Erre két lehetőség van.

Pontbecslés esetén a mintához egyetlen számszerű értéket rendelünk, és ezt tekintjük a becsülni kívánt paraméter értékének, azaz a mintából kiszámított értéket tekintjük a sokasági jellemző becsült értékének.
Intervallumbecslés esetén azonban egy olyan intervallumot határozunk meg, amely előre adott nagy valószínűséggel tartalmazza a becsülni kívánt paramétert. Ezt az intervallumot konfidenciaintervallumnak nevezzük.
Gondoljunk bele, hogy ha vennénk 100 különböző mintát, akkor mindegyiknek a mintaátlaga más és más érték lehet. Éppen ezért – például – a 95 százalékos konfidenciaintervallum azt jelenti, hogy összes lehetséges – adott elemszámú – mintát véve, az esetek 95 százalékában a sokasági átlag bele esik a konfidenciaintervallumba.
Az intervallum megadása a szoftverek többségében az intervallum alsó és felső határának megadását jelenti. Az intervallum középpontja a vizsgált mintajellemző lesz.

Leíró statisztikai elemzés a statisztikai szoftverek használatakor általában az alábbi mutatók számszerűsítését és kiírását jelenti.

- *N:* a megfigyelés elemszáma.
- Sum, összeg: a változó értékeinek összege.
- Mean, átlag, várható érték: a mintaátlag.
- Median, medián: a medián.
- Modus, módusz: a módusz. A számítógépes szoftverek többmóduszú eloszlás esetén csak egy móduszt írnak ki.
- Minimum, maximum: a változó legkisebb és a legnagyobb értéke.
- Std. Deviation, szórás: a változó szórásának kiszámítása. Ez programonként változó, hiszen van, ahol a „közönséges” szórást, de van ahol a korrigált tapasztalati szórást kapjuk eredményül.
- Kurtosis, csúcsosság: egy eloszlás csúcsosságának megállapítása az azonos szórású normális eloszláshoz viszonyítva. Az alapértelmezésként használt mutató pozitív értéke csúcsosabb, míg negatív értéke lapultabb eloszlást jelez.
- Skewness, ferdeség: egy eloszlás aszimmetriájának megállapítása az azonos szórású normális eloszláshoz viszonyítva. Az alapértelmezésként használt mutató pozitív értéke baloldali aszimmetriát, azaz jobbra (pozitív irányban) hosszan elnyúló eloszlást, míg negatív értéke jobboldali aszimmetriát jelez. Leegyszerűsítve, például a baloldali aszimmetria úgy képzelhető el, hogy az ismérvértékek többsége átlag alatti.
- Konfidenciaintervallum: egy általunk megadott megbízhatósági szintű konfidenciaintervallum megállapítása a sokasági várható értékre.

A felsoroltakon kívül opcionálisan általában kvartiliseket, grafikus ábrákat kérhetünk.

3. Várható értékek (átlagok) vizsgálatára irányuló próbák

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 azt fogjuk megvizsgálni, hogy a mintánk támogatja-e a

hipotézisünket, vagy szignifikánsan ellentmond neki. Így bizonyos megbízhatósággal állíthatjuk majd, hogy hipotézisünk teljesül vagy sem.
A felállított hipotézisek helyességének véletlen mintákra alapozott vizsgálatát **hipotézisvizsgálatnak** nevezzük. Az ennek során alkalmazott eljárások a statisztikai próbák vagy tesztek.
A hipotézisvizsgálat során a **hipotézisvizsgálat lépéseit** kell végrehajtani.
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. Ez gyakorlatilag egy statisztika (képlet) kiválasztását jelenti. Ez a pórbafüggvény gyakorlatilag a mintából kiszámámított érték(ek)et hasonlítja össze a tesztelt sokasági paraméter feltételezett értékével. Ezt a szoftverek a megfelelő elemzési eljárások kiválasztásával – többnyire – elvégzik helyettünk, a számítógépes kimeneteken ezeknek a vizsgált mintán felvett értékeit is megtalálhatjuk.
3. A 0-hoz közeli α szignifikanciaszint kiválasztása, és a próbafüggvény értékkészletének elfogadási és kritikus tartományra bontása. Gyakorlatilag meghatározzuk egy olyan eltérés tartományt, amire azt mondjuk, hogy a tesztelt sokasági paraméter mintából kiszámított értéke és várt értéke közötti eltérés nem szignifikáns. A szoftverek általában csak egy úgynevezett kritikus értéket közölnek a tartományok helyett.
Gyakorlatban általában ötszázalékos szignifikanciaszint mellett dolgozunk, de előfordul a 10, illetve minőségügyi vizsgálatok során az 1 százalékos szint alkalmazása is. Például az ötszázalékos szignifikanciaszint azt jelenti, hogy ha százszor elvégeznénk ugyanazt a vizsgálatot, akkor 95 esetben helyes döntést hozunk.
4. A próbafüggvény mintán felvett értékének megállapítása.
5. Döntés a nullhipotézis helyességének elfogadásáról-elvetéséről. Ez gyakorlatilag a próbafüggvény mintán felvett értékének összehasonlítását jelenti a kritikus értékkel. A nullhipotézis elvetése maga után vonja az alternatív hipotézis elfogadását.
Amikor döntést hozunk a nullhipotézis elfogadásáról, illetve elvetéséről, akkor természetesen előfordulhat, hogy rossz döntést hozunk. Abban az esetben, ha a nullhipotézist elvetjük, noha az a valóságnak megfelel, akkor **elsőfajú hibát** követünk el. Ennek elkövetési valószínűsége megegyezik az α szignifikanciaszinttel. Abban az esetben, ha a nullhipotézist elfogadjuk, noha az a valóságnak nem felel meg, akkor **másodfajú hibát** követünk el. Ennek valószínűségét csak becsülni tudjuk. Annyit érdemes megjegyezni, hogy az elsőfajú hiba elkövetésének valószínűségét nem célszerű túlságosan kicsinek választani, mert csökkentése növeli a másodfajú hiba elkövetésének valószínűségét.
Mivel a statisztikai szoftverek jelentős hányada, az elemzések egy részénél azt feltételezi, hogy az adatállományunk egy minta jellemzői, ezért az elemzések során többnyire hipotézisvizsgálatot is folytatunk. A szoftverek segítségével általában semmit sem kell számolnunk, így a statisztikai szoftverek segítségével végrehajtott hipotézisvizsgálat felhasználói szempontból egyszerűbb, azonban, egyrészt tudnunk kell, hogy az adott hipotézisvizsgálatnak mi a nullhipotézise, illetve az alternatív hipotézise, másrészt pedig, az eredményeket helyesen kell értelmezni. A hipotézisvizsgálat számítógépes végrehajtása az alábbi lépések végrehajtását jelenti:

- Közöljük a vizsgáltunk nullhipotézisét. A hipotézisek bármely szakirodalomban, illetve a programok súgójában is megtalálhatóak.

- A számítógépes kimenten kiválasztjuk a megfelelő értékeket, amely alapján döntést hozunk.
- Eredményeinket, következtetéseinket „hétköznapi” nyelven tálaljuk. Az eredmények helyes értelmezéséhez, azaz a döntéshez pedig az alábbiakat kell tudnunk.
  A statisztikai szoftverek hipotézisvizsgálat során, az outputon megadnak egy értéket, a *p*-érték, *p*-value, Sig. jelölések valamelyikével. Ez az érték az első fajú hiba elkövetésének valószínűségét jelenti. Ha vizsgálataink során például 5 százalékos szignifikanciaszintet használunk, akkor amennyiben a kapott érték 0,05-nál kisebb, akkor a nullhipotézist – ötszázalékos szignifikanciaszint mellett – elvetjük. Tehát akkor fogadjuk el a nullhipotézist, ha az elsőfajú hiba elkövetésének valószínűsége legfeljebb ötszázalék, azaz 95 százalékos valószínűséggel helyes döntést hozunk.

A várható értékek összehasonlítására számos lehetőség közül választhatunk a változók típusa és a minták száma alapján. Mi csak a metrikus változók várható értékeinek összehasonlításával foglalkozunk.

Egymintás próbák esetén meg kell adnunk, hogy a várható értéket milyen konkrét értékkel szeretnénk összehasonlítani. Kétmintás próbák esetében, vagy csak a vizsgált két változót kell megadnunk, vagy pedig a várt eltérést is. Ennek az az oka, hogy nem csak a két várható érték egyezőségét, hanem adott értékkel való különbözőségét is lehet tesztelni.
Ezeknél a próbáknál a nullhipotézis mindig a várható értékek egy adott értékkel történő, egymással való megegyezését, vagy különbségük adott értékkel való egyezőségét jelenti.
Két és több mintás esetekben a várható értékek összehasonlítása azzal a kérdéssel ekvivalens, hogy szignifikáns különbség van-e a minták között az adott változóban. Ekkor a vizsgálat nullhipotézise a várható értékek egyezősége, azaz az, hogy nincs szignifikáns különbség a minták között.

Ezeknél a vizsgálatoknál az elsőfajú hiba elkövetése azt jelenti, hogy elvetjük a helyes nullhipotézist, azaz szignifikáns különbséget mutatunk ki ott, ahol nincs is.

A várható értékek hipotézisvizsgálattal történő összehasonlításának lehetőségei

Várható értékek összehasonlítása	**Sorrendi mérési szintű változó (ordinális skála)**	**Metrikus változó (intervallumskála, arányskála)**
	Medián	**Átlag**
Várható érték összehasonlítása egy adott értékkel	WILCOXON-próba	Egymintás *z*-próba (ha az alapsokaság szórása ismert). Egymintás *t*-próba (ha az alapsokaság szórása nem ismert).
Két várható érték összehasonlítása egymással	MANN-WHITNEY próba	Kétmintás *z*-próba (ha az alapsokaságok szórása ismert). Kétmintás *t*-próba (ha az alapsokaságok szórása nem ismert. Ekkor külön tesztelnünk kell *F*-próbával a varianciák egyezőségét.)
Több várható érték összehasonlítása egymással	KRUSKALL-WALLIS próba (független minták esetén) FRIEDMANN próba (nem független minták esetén)	Varianciaanalízis (varianciahomogenitás és normális eloszlás esetén).

Ebben a fejezetben, a továbbiakban a kétmintás *t*-próbákkal foglalkozunk. Ennek alkalmazására általában az alábbi három esetben kerül sor.

1. **Kétmintás, páros *t*-próba**. Ekkor az egyes megfigyelések egymással párba állíthatók (nem függetlenek a minták). Tipikusan erről az esetről van szó, ha egy csoportot megfigyelünk valamilyen kísérlet előtt és után. Például szignifikáns különbség van-e a kezdő fizetés és a jelenlegi fizetés között.
2. **Kétmintás *t*-próba egyenlő varianciákkal**. A próba előtt, *F*-próbával meg kell vizsgálnunk a varianciák egyezőségét.
3. **Kétmintás *t*-próba nem egyenlő varianciákkal**. A próba előtt, *F*-próbával meg kell vizsgálnunk a varianciák egyezőségét.

A kétmintás t-próbák esetén beszélhetünk egyoldali és kétoldali próbákról is. Például, ha azt a nullhipotézist vizsgáljuk, hogy a férfiak és a nők várható élettartama megegyezik, azaz szignifikánsan nem különbözik, akkor legegyszerűbb alternatív hipotézisként azt fogalmazhatjuk meg, hogy a férfiak és a nők várható élettartama nem egyezik meg, azaz szignifikánsan különbözik. Ennél azonban többet is állíthatunk, például azt, hogy a nők várható élettartama nagyobb (kisebb), mint a férfiaké. Az első esetben úgynevezett **kétoldali próbát**, míg az utóbbi esetben **egyoldali próbát** hajtunk végre. Az egyoldali próbák esetén beszélhetünk bal-, illetve jobboldali próbáról is, az alternatív hipotézis függvényében. A statisztikai programcsomagok általában csak kétoldali próbát hajtanak végre, mivel néhány paraméterből következtethetünk az

egyoldali próba eredményére is. Például, ha szignifikáns különbséget találunk a férfiak és a nők várható élettartama között, akkor például a mintaátlagok nagyságából megállapíthatjuk, hogy melyik nem várható élettartama szignifikánsan nagyobb.

4. Eloszlások vizsgálata

Illeszkedésvizsgálat során egy változó feltételezett eloszlását ellenőrizzük a hipotézisvizsgálat segítségével. A próba nullhipotézise szerint a sokaság eloszlása megközelítőleg a feltételezett eloszlást követi. Amennyiben azt vizsgáljuk, hogy egy adott változó normális eloszlású-e, akkor **normalitásvizsgálat**ról beszélünk.
A statisztikai elemzéseknél alkalmazott egyes módszerek feltételezik a modell valamely adott változójának normális eloszlását. Noha a legtöbb módszer elég robosztus erre a feltételezésre, ajánlatos ellenőrizni, hogy egy változó megközelítőleg normális eloszlást követ-e.

A normális (GAUSS-féle) eloszlásnak két paramétere van: a μ várható érték és a σ szórás. Általában, ha ezek a paraméterek ismeretlenek, akkor a hipotézisvizsgálat előtt, előbb meg kell becsülnünk az értéküket.

Fontos kiemelni, hogy empirikus eloszlások csak megközelítőleg lehetnek normális eloszlásúak; a normális eloszlás csak elméletben létezik, empirikus adatok tapasztalati eloszlásaként nem. Egy normális eloszlású valószínűségi változó ugyanis a $(-\infty, \infty)$ intervallumban bármilyen értéket felvehet, ami a gyakorlatban (gazdasági, társadalmi jelenségek vizsgálatánál) természetesen sohasem fordul elő. Gyakran találkozunk azonban (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 a normális eloszlású valószínűségi változónkat standardizáljuk, akkor a transzformált változó standard normális eloszlású lesz. Az ilyen változókat a statisztikában gyakran *z*-vel vagy *u*-val jelöljük.
A standardizált változó – mivel mértékegység nélküli – univerzálisan használható, azaz különböző típusú – megközelítőleg normális eloszlású – sokaságok esetén is alkalmazható összehasonlítás céljára.

A továbbiakban a normalitás tesztelésére mutatunk be néhány lehetőséget.

Normalitásvizsgálat momentumok segítségével
A normalitás-tesztek legegyszerűbb típusa arra épül, hogy a normális eloszlású változó bizonyos momentumainak hányadosai konstans értékűek.

- $\frac{\delta}{\sigma} = \sqrt{\frac{2}{\pi}} \approx 0{,}8$ (GEARY-féle próba)
- $\alpha_3 = \frac{M_3(\mu)}{\sigma^3} = 0$
- $\alpha_4 = \frac{M_4(\mu)}{\sigma^4} = 3$

Grafikus Q-Q teszt

A grafikus tesztek valójában nem statisztikai próbák, hanem vizualizációs eszközök, ezért objektív döntési szabály nem rendelhető melléjük. Alkalmazásuk népszerűségét gyorsaságuk, egyszerűségük és kényelmes kezelhetőségük adja. A *Q-Q* (quantile-quantile) teszt alapötlete az, hogy ha a nullhipotézisünk igaz, vagyis a tesztelt változónk $F(x)$ elméleti eloszlású, akkor

$$u = F^{-1}\left(F_n(x)\right) = x\,.$$

Így az (x, u) pontokat, vagyis az eredeti megfigyeléseket és azok oda-vissza transzformált értékeit ábrázolva 45 fokos egyenest kapunk. Ha $F(x)$ a normális eloszlás, akkor a változónk standardizálása után (a paramétereket például momentum módszerrel becsülve)

$$u = \Psi^{-1}\left(\Psi\left(\frac{x-\mu}{\sigma}\right)\right) = \frac{x}{\sigma} - \frac{\mu}{\sigma}\,.$$

Ekkor a kapott egyenes meredeksége $1/\sigma$, a tengelymetszet $-\mu/\sigma$. Ha más pozíciójú egyenest kapnánk, akkor a normalitást elfogadhatjuk, csak a paraméterek becslését kell felülvizsgálni.

KOLMOGOROV-SZMIRNOV-féle próba

A próba lényege, hogy előállítjuk a kumulált gyakorisági eloszlást és azt vizsgáljuk, hogy ennek görbéje maximálisan mennyire távolodik el az elméleti eloszlásfüggvénytől.

A KOLMOGOROV-SZMIRNOV-féle próba más adott folytonos eloszlás tesztelésére is használható, de az esetleges ismeretlen paraméterek becslését előbb meg kell tenni, például momentumok módszerével. A teszt az eloszlás középső részén érzékenyebb, mint a szélein.

SHAPIRO-WILK-féle próba

Ez a próba SHAPIRO és WILK 1965-ben definiált *W* statisztikáján alapul.

5. Összefüggés-vizsgálat

Az adatok több szempontú rendezése céljából, célszerű ezeket táblázatokba foglalni.

Két minőségi, vagy területi ismérv esetén – kétdimenziós – kombinációs táblát készíthetünk. Ekkor mindegyik megfigyelést egyidejűleg két ismérv szerint osztályozzuk.

A kombinációs tábla alapján lehetőségünk van a két vizsgált változó közötti kapcsolat szignifikanciájának vizsgálatára. Ezt a vizsgálatot **függetlenségvizsgálat**nak nevezzük. A próba nullhipotézise szerint a két ismérv (változó) egymástól független. A nullhipotézis elvetése csupán azt jelenti,

hogy a két változó nem tekinthető egymástól függetlennek. Ettől a köztük lévő kapcsolat erőssége – gyakorlati szempontból – még jelentéktelen is lehet.

Empirikus vizsgálatok során kutatási problémaként gyakran kerülhetünk szembe a „miért", a „mi ennek az oka", „milyen összefüggés van a válktozók között" kérdésekkel. Ezen kérdések kezelésére mind kvalitatív, mind kvantitatív eljárások ismertek. Ám a kvalitatív vizsgálatokkal szemben, a kvantitatív vizsgálatok eredményei számszerűsíthetőek. A kérdéskör statisztikai vizsgálatára a statisztikai magyarázó modelleket is alkalmazhatjuk. Az alkalmazandó magyarázó modell kiválasztása előtt számos kérdést kell tisztázni. Először is meg kell állapítani a vizsgálatba bevont függő változó és a magyarázatként lehetséges tényezők közötti függőségi kapcsolat típusát. Ehhez a változók típusait kell felismernünk.

- Amennyiben mind a független, mind a függő változó metrikus akkor korreláció- és regresszióanalízist végezhetünk. Például befolyásolja-e egy termék értékesítését a termék ára. Ezzel a vizsgálati típussal a későbbi fejezetekben foglalkozunk.
- Amennyiben a független változó metrikus, a függő változó kategorizált akkor diszkriminancianalízist, illetve logisztikus regressziót célszerű alkalmaznunk. Például befolyásolja-e egy termék ára azt, hogy megvesszük, vagy sem. Ezzel a vizsgálati típussal nem foglalkozunk.
- Amennyiben a független változó kategorizált, a függő változó pedig metrikus akkor varianciaanalízist végezhetünk. Például befolyásolja-e egy termék értékesítését a termék minősége. Ezzel a vizsgálati típussal a későbbi fejezetekben foglalkozunk, az alapjaival az I. félévben foglalkoztunk a vegyes kapcsolatok vizsgálatánál...
- Amennyiben mind a független, mind a függő változó kategorizált akkor kereszttáblaelemzést végezhetünk. Például befolyásolja-e az áruházakban a termék elhelyezését a termék minősége. Ennek alapjaival találkoztunk már az I. félévben az asszociációs kapcsolat vizsgálatakor. Ebben a fejezetben ezen ismereteinket mélyítjük tovább.

Változók közötti összefüggések vizsgálatának típusai

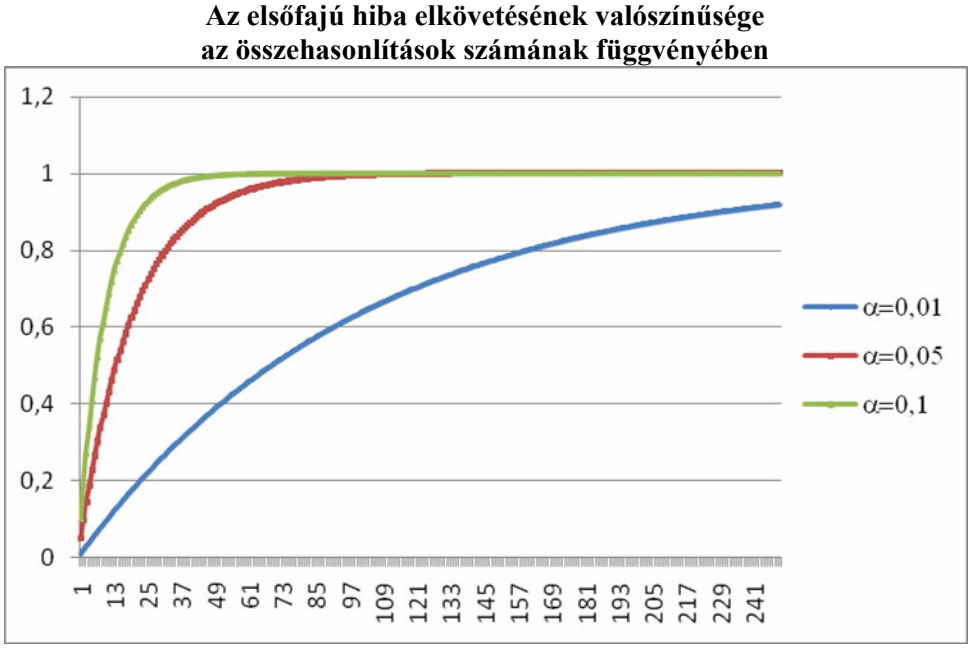
	Metrikus független	**Nem metrikus független**
Metrikus függő	Korrelációszámítás (kapcsolatvizsgálat) Regressziószámítás (ok-okozati vizsgálat)	Varianciaanalízis (ANOVA)
Nem metrikus függő	diszkriminanciaanalízis	kereszttáblaelemzés

Két minőségi vagy területi ismérv esetén – kétdimenziós – kombinációs táblát készíthetünk. Ekkor mindegyik megfigyelést egyidejűleg két ismérv szerint osztályozzuk. Ezt az Excelben a kimutatáskészítés segítségével végezhetjük el.
A kombinációs tábla alapján lehetőségünk van a két vizsgált változó közötti kapcsolat szignifikanciájának vizsgálatára. Ezt a vizsgálatot **függetlenségvizsgálat**nak nevezzük. A próba nullhipotézise szerint a két ismérv (változó) egymástól független. A nullhipotézis elvetése csupán azt jelenti, hogy a két változó nem tekinthető egymástól függetlennek. Ettől a köztük lévő kapcsolat erőssége – gyakorlati szempontból – még jelentéktelen is lehet. Tehát

először meg kellene vizsgálnunk, hogy szignifikáns kapcsolat van-e két változó között, és csak azután beszélhetünk a kapcsolat erejéről.

6. Varianciaanalízis

A varianciaanalízis két alkalmazását érdemes megemlíteni: az első alkalmazás segítségével több sokaság várható értékének egyezősége tesztelhető, a másik alkalmazás segítségével pedig regressziós modellek illeszkedése, illetve a többszörös korrelációs együttható tesztelhető. Ebben a fejezetben az első alkalmazást taglaljuk.
Az első alkalmazás segítségével több sokaság várható értékének egyezősége tesztelhető. A két mintás t-próbák általánosításának tekinthető **varianciaanalízis,** több, egyenlő szórású, normális eloszlású csoport/sokaság várható értékének összehasonlítására alkalmas statisztikai módszer, melyet angol elnevezésének kezdőbetűiből adódóan ANOVA-ként (ANalysis Of VAriance) is emlegetnek. Tehát legalább egy csoportosító ismérv szerint részekre bontott sokaság valamely, legalább intervallumskálán mért ismérvének és a csoportosító változók a kapcsolatát vizsgáljuk. Arra keressük a választ, hogy a csoportok statisztikailag szignifikánsan különböznek-e a metrikus változóban. A vizsgálat eredménye az **ANOVA táblázat**ból olvasható ki. A próba nullhipotézise szerint a csoportok várható értékei megegyeznek, azaz a csoportosító ismérv befolyásolja a metrikus válttozót. Míg az alternatív hipotézis ennek tagadása. Tehát az alternatív hipotézis nem azt jelenti, hogy mindegyik csoport várható értéke különbözik, hanem csak azt, hogy nem tekinthető mindegyik azonosnak. Ha ennél több információra van szükségünk, azaz a várható értékeket külön-külön is szeretnénk összehasonlítani, akkor úgynevezett **Post Hoc teszt**et kell végrehajtanunk.



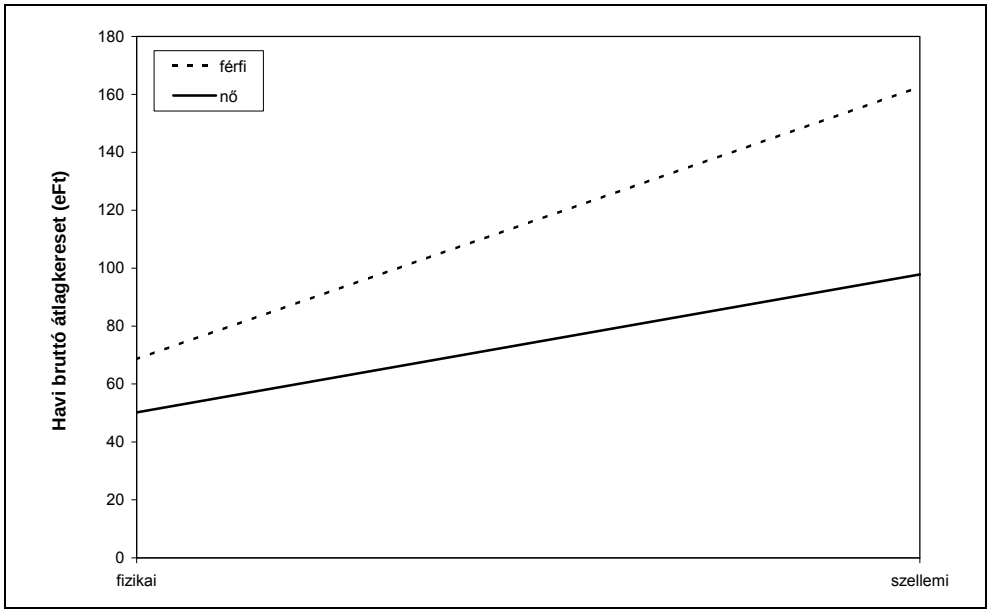
Kettőnél több minta alapján történő várható értékek összehasonlítására is elvileg működnek a kétmintás próbák, a minták összes lehetséges páronkénti

összehasonlításával. Azonban, ez az eljárás nem ajánlott, ugyanis az összehasonlítások számának növekedésével drasztikusan emelkedik az elsőfajú hiba elkövetésének valószínűsége. A varianciaanalízis alkalmazásakor az elsőfajú hiba elkövetése gyakorlatilag azt jelenti, hogy kapcsolatot mutatunk ki ott, ahol nincs is.

A varianciaanalízis alkalmazásának két feltétele van. Az egyik a sokaság normális eloszlása, a másik pedig a varianciák egyezősége. Az első feltétel nem teljesülése nem okoz jelentős torzulást a végeredményekben. Az Excel Adatelemzés modulja nem ellenőrzi le a feltételek teljesülését.

Attól függően, hogy hány csoportosító ismérv (hatótényező vagy faktor) hatását vizsgáljuk, beszélünk egyszempontos (egyutas), kétszempontos (kétutas), illetve többszempontos (többutas) varianciaanalízisről.
Már az egyszerűbb jelenségek vizsgálatakor is felmerül, hogy nem csak egy tényező befolyásolja a vizsgált metrikus változónk értékét. Több kategoriális változót bevonva azonban gyorsan nő modellünk bonyolultsága. Két hatótényező esetén már számolnunk kell az azok közötti függőségi viszonnyal is, azaz azzal a hatással, amelyet a két változó együttese gyakorol a vizsgált függő változónkra. Ezt **interakciós hatás**nak nevezzük. (A faktorok által külön-külön magyarázott részt főhatásoknak nevezzük.) Az interakció két hatótényező esetén azt jelenti, hogy rögzítve az egyik értékét (ismérváltozatát) a másik független változó különböző ismérváltozatai mentén a függő változó másként viselkedik, mint az első változó más rögzített értékei mellett. Az alábbi ábrán láthatunk erre egy példát: a férfiak átlagkeresete másként alakul a fizikai-szellemi változó mentén mozogva, mint a nőké.

Alkalmazásban állók havi bruttó átlagkeresete nemenként 2000-ben



A főhatások elemzése csak akkor végezhető el, ha nincs interakciós hatás, ellenkező esetben az egyik faktor hatása függ a másik adott értékétől, és fordítva.

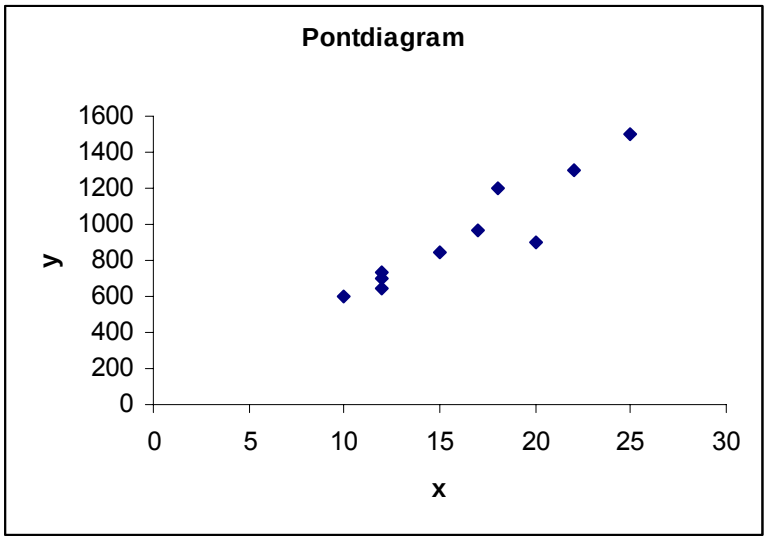
Érdekességként megjegyezzük, hogy ha modellünkben a kategoriális hatótényezők mellett metrikus független változót is szerepeltetünk, akkor **ANCOVA** módszert alkalmazunk. A nemmetrikus független változókat **faktor**oknak, a metrikus független változókat pedig **kovariáns**oknak nevezzük. A függő változó és a kovariánsok között (többszörös) determinációs együttható számítható. Ezzel kiszámíthatjuk a függő változó heterogenitásának azt a részét, amit a kovariáns magyaráz. A fennmaradó eltérés-négyzetösszeg egy további hányada magyarázható a faktorokkal. Természetesen itt is felléphetnek interakciós hatások pusztán faktorok, pusztán kovariánsok között és persze faktorok és kovariánsok között is.
A független változók (csoportosító ismérvek) mellett növelhetjük a függő változók számát is. Ezt **többváltozós szórásnégyzetelemzés**nek (**MANOVA**) nevezzük. Természetesen a MANOVA eljárásnak is létezik többutas változata, illetve a faktorok mellett kovariánsokat is szerepeltethetünk. Ez utóbbi modelleket **MANCOVA** modelleknek nevezzük.

A varianciaanalízis második alkalmazására regressziós modellek használatakor kerül sor. Ekkor a feltételezett modell illeszkedése jóságának vizsgálatára használjuk. Ezt a regressziószámításnál fogjuk tárgyalni.

7. Korreláció- és regressziószámítás

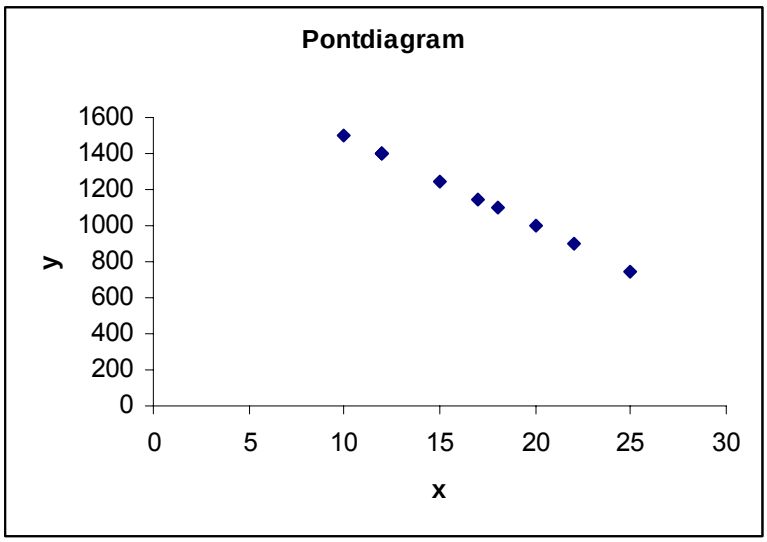
Korrelációszámítás esetén az elemzésbe vont metrikus változók közötti kapcsolatot vizsgáljuk. Két metrikus változó (x,y) közötti kapcsolat vizsgálatának első fázisában pontdiagramot készíthetünk az x-y változópár alapján. A pontdiagram alapján megállapíthatjuk a változópár közötti kapcsolat típusát: lineáris, vagy nem lineáris a kapcsolat. Lineáris kapcsolat esetén a pontok egy képzeletbeli egyenes, nem lineáris kapcsolat esetén egy szabályos görbe körül szóródnak. Mivel a gyakorlatban nagyon gyakran élünk a linearitás feltételezésével, így a továbbiakban erre koncentrálunk. A pontoknak a képzeletbeli egyenes körüli szóródásából következtethetünk arra, hogy milyen szoros kapcsolat van a két változó között. Az egyenes meredekségéből pedig következtethetünk a kapcsolat irányára, ami pozitív, vagy negatív lehet. A pozitív irányú kapcsolat azt jelenti, hogy a két változó azonos irányba változik. Mivel a pontdiagram nem egzakt megoldása a korrelációszámításnak, ezért a kapcsolat erősségének jellemzésére mérőszámokat használunk. Lineáris kapcsolat esetén az úgynevezett **lineáris korrelációs együtthatót**, míg nem lineáris kapcsolatok esetén az úgynevezett **korrelációs indexet** használjuk. A *r* lineáris korrelációs együttható értéke [-1;+1] tartományba esik. Előjele megadja a két változó közötti kapcsolat irányát, míg abszolút értéke a kapcsolat erősségét. A nullához közeli érték gyenge, az egyhez közeli érték erős kapcsolatot jelent. A korrelációs index értéke [0;+1] tartományba esik, és kizárólag a változópár közötti kapcsolat erősségét adja meg.

Például, mit állapíthatunk meg az alábbi pontdiagramok alapján?
A)



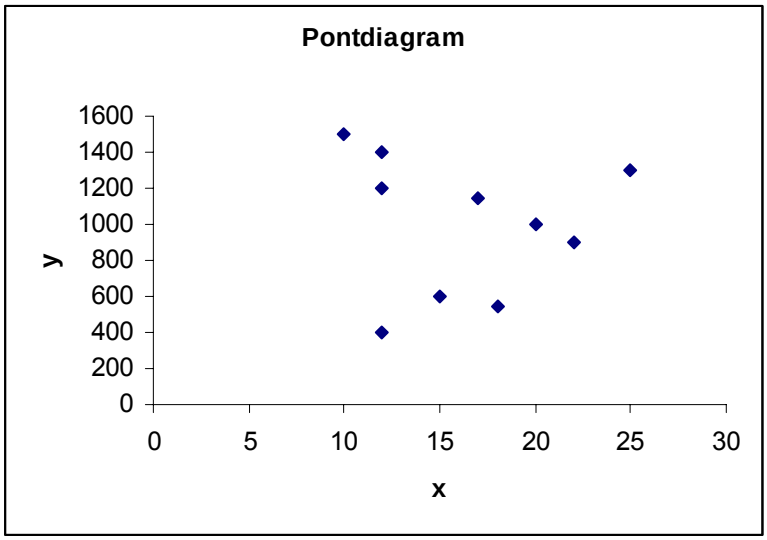
Mivel a pontok nagyon kis mértékben szóródnak egy képzeletbeli, pozitív meredekségű egyenes körül, ezért a két mennyiségi ismérv között pozitív irányú, erős lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke egyhez közeli.

B)



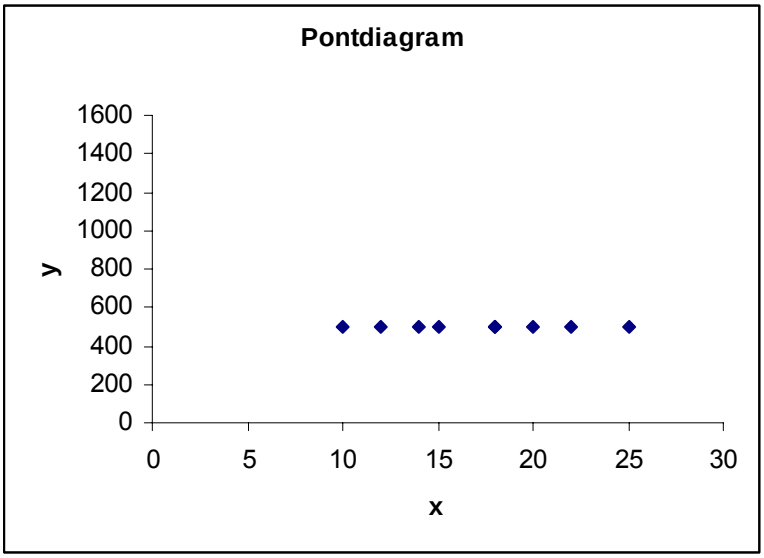
Mivel a pontok kis mértékben szóródnak egy képzeletbeli, negatív meredekségű egyenes körül, ezért a két mennyiségi ismérv között negatív irányú, nagyon erős lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke mínusz egyhez közeli.

C)



Mivel a pontok mindkét dimenzióban nagyon szóródnak, így nem lehetséges egy képzeletbeli egyenes rájuk illesztése, ezért a két mennyiségi ismérv között nagyon gyenge, szinte elhanyagolható lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke nullához közeli.

D)



Mivel a pontok kis mértékben szóródnak egy képzeletbeli, zéró meredekségű egyenes körül, azaz az *x* értékétől függetlenül *y* megközelítőleg konstans, ezért a két mennyiségi ismérv között elhanyagolhatóan gyenge lineáris korrelációs kapcsolat van. A lineáris korrelációs együttható értéke nullához közeli.

Abban az esetben, ha több változó közötti kapcsolat vizsgálunk, akkor egyrészt beszélhetünk a változópárok közötti kapcsolatról. Ekkor minden egyes változópárra kiszámíthatjuk a lineáris korrelációs együttható értékét. Ekkor ezeket egy mátrixba rendezve adjuk meg, melyet **korrelációs mátrixnak** nevezünk. Másrészt változók csoportjai között is vizsgálhatunk kapcsolatot. A **kanonikus korrelációanalízis** változók csoportjai közötti kapcsolatot vizsgálja. Ezzel részletesen nem foglalkozunk.

Lineáris korreláció esetén érdekes kérdés a kapcsolat szignifikaciájának vizsgálata. A próba nullhipotézise szerint a vizsgált két változó egymástól lineáris független, tehát a korrelációs együttható értéke szignifikánsan nem különbözik nullától. A korrelációs mátrix mellett általában megtalálhatjuk a lineáris korrelációs kapcsolatok szignifikanciáját, azonban az Excel ezt nem vizsgálja meg, ezért hasznos lehet az alábbi táblázat. A táblázatban a megfigyelés számának függvényében megtalálható a lineáris korrelációs együttható kritikus értéke. Ez azt jelenti, hogy, ha a vizsgálatainkban kiszámított korrelációs együttható abszolút értéke nagyobb a kritikus értéknél, akkor a két változó közötti kapcsolat szignifikáns.

Meg kell jegyezni, hogy a kapcsolatvizsgálat pusztán matematikailag vizsgálja a változók együttmozgását, ami nem feltétlen jelent valós kapcsolatot.

A lineáris korrelációs együttható kritikus értéke

Mintaelemszám	$r_{kritikus}$	Mintaelemszám	$r_{kritikus}$
3	0,997	28	0,374
4	0,950	29	0,367
5	0,878	30	0,361
6	0,811	35	0,334
7	0,754	40	0,312
8	0,707	45	0,294
9	0,666	50	0,279
10	0,632	55	0,266
11	0,602	60	0,254
12	0,576	65	0,244
13	0,553	70	0,235
14	0,532	75	0,227
15	0,514	80	0,220
16	0,497	85	0,213
17	0,482	90	0,207
18	0,468	95	0,202
19	0,456	100	0,197
20	0,444	200	0,139
21	0,433	300	0,113
22	0,423	400	0,098
23	0,413	500	0,088
24	0,404	1000	0,062
25	0,396	1500	0,051
26	0,388	2000	0,044
27	0,381		

Statisztikai elemzéseknél gyakran vetődik fel az a kérdés, hogy sztochasztikus kapcsolat esetén az egyik ismérv (vagy több ismérv) által hordozott információt hogyan tudnánk felhasználni a másik ismérv értékeinek meghatározására. Az összefüggéseket ok(x)-okozati(y) kapcsolattal leíró egyik ilyen módszert **regressziószámítás**nak nevezzük. Ekkor egy egyenlettel adott kapcsolatot létesítünk a változók között, melynek segítségével a magyarázóváltozók (x) alapján becslést adhatunk az eredményváltozóra (y). Például, ha jégkrémet

árulunk, akkor egy adott napra az eladott mennyiséget előre bevsülhetjük pusztán hasraütés szerűen is, de akár azt is mondhatjuk, hogy az értékesítést befolyásolja a külső hőmérséklet, és ez alapján becsüljük meg az értékesítést.
A regressziós modellben alkalmazott függvény típusának fontos szerepe van; ez egyszerűbb esetekben lineáris, de az empirikus elemzéseknél gyakran nemlineáris.

A regressziószámítás outputján látható, hogy milyen változók és milyen szerepkörben szerepelnek a modellben.

Egy regressziós modellbe többféle eljárással válogathatunk be magyarázóváltozókat. Ezeket a számítógépes programcsomagok rendszerint felkínálják, amelyek közül néhányat az alábbiakban ismertetek.
Forward módszer: a modellbe egyesével léptetjük be a magyarázóváltozókat. Először az eredményváltozóval (PEARSON-féle lineáris korrelációs együttható alapján) legerősebb kapcsolatban levőt vonjuk be. Majd a parciális korrelációs együttható alapján azt, amelyik a legnagyobb mértékben növeli a magyarázott hányadot, ezzel biztosítva, hogy a már modellben szereplő változókon felül a legnagyobb többletinformációt adó változót vonjuk be. Addig vonunk be újabb változót, amíg az általa magyarázott rész (egy előre meghatározott szignifikanciaszint mellett) még jelentős.
Backward módszer: kezdetben minden magyarázóváltozó benne van a modellben, majd minden becsült regressziós együtthatóhoz elvégezzük a parciális *F*-próbát. Előre meghatározott szignifikanciaszinthez tartozó *F* (illetve *t*) érték felettiek közül az a magyarázóváltozó kerül ki a modellből, amelyiknek a legkisebb az *F* (illetve *t*) értéke.
Stepwise módszer: a magyarázóváltozók itt is egyesével lépnek be a modellbe, de ha az újabb belépők hatására (egy adott lépésben) a már bent levő változók közül valamelyikhez tartozó *t* érték adott szint alá csökken, akkor azok kilépnek a modellből.

Regressziószámítás során feltétlenül meg kell vizsgálnunk

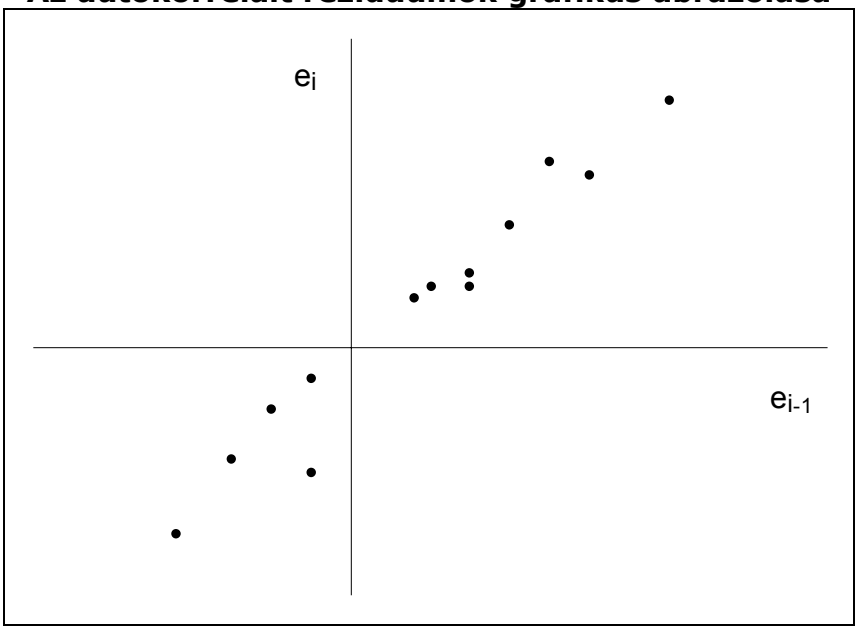
- a **többszörös determinációs együttható**t (*R*-square), ami a modell magyarázó erejét adja meg, azaz azt, hogy a magyarázóváltozók együttesen milyen mértékben magyarázzák az eredményváltozó értékeinek különbözőségét. Egy adott modell magyarázóereje általában 80 százalék felett tekinthető elfogadhatónak. Ha egy adott modellbe újabb változókat csatolunk, akkor a modell magyarázóereje biztosan nem csökken. Ez által a magyarázó erő akár 100% közelébe is felhúzható, azonban ettől mindenkit óva intenék, ugyanis egyrészt túl bonyolultá teszi a modellt, másrészt sok negatív következménnyel jár;
- a **többszörös korrelációs együttható**t (*R*), ami a többszörös determinációs együttható gyöke. Megmutatja, hogy a magyarázóváltozók együttese (mint változók halmaza) milyen szoros kapcsolatban áll (mennyire mozog együtt) az eredményváltozóval (y);
- a **reziduális szórás**t (standard error of the estimates), ami az eredményváltozó tényleges és a modell alapján becsült értékeinek az átlagos eltérését adja meg;
- a **regressziós modell illeszkedésének jóságát**. Ez a vizsgálat egy variancianalízis végrehajtását jelenti. A próba nullhipotézise szerint a modell illeszkedése nem megfelelő, azaz azt hogy a többszörös korrelációs

együttható értéke szignifikánsan nem különbözik nullától. A vizsgálat eredményét egy ANOVA táblázatban kapjuk meg. Fontos megjegyeznünk, hogy a nullhipotézis elvetése csak azt jelenti, hogy a modell alkalmazható a probléma vizsgálatára.

- Végre kell hajtanunk a **paraméterek tesztelését**. Ekkor mindegyik magyarázóváltozó fontosságát, magyarázó erejét fogjuk tesztelni külön-külön. A próba nullhipotézise szerint a vizsgált magyarázóváltozó szignifikánsan nem befolyásolja az eredményváltozót. Azokat a változókat, amik nincsenek szignifikáns hatással az eredményváltozóra nem érdemes szerepeltetnünk a modellben. Az is – különösen szignifikáns multikollinearitás esetén – lehetséges, hogy egy magyarázóváltozó ereje önmagában nem csekély, de – a modellben levő többi magyarázóváltozó által hordozott megmagyarázott hányadon felül – további többletinformációval nem rendelkezik. Fontos megjegyezni, hogy szignifikáns multikollinearitás esetén a teszt eredményei nem értelmezhetőek!
- Meg kell vizsgálnunk a **multikollinearitás** jelenlétét, azaz a magyarázóváltozók függetlenségének hiányát. A multikollinearitás mértékének meghatározására a számítógépes szoftverek általában (az egynél nem nagyobb, nemnegatív értékű) tolerancia-mutatót és ennek reciprokát, a *VIF* mutatót használják. Ezek a mutatók minden magyarázóváltozó esetében kiszámításra kerülnek. Minél jobban távolodik a mutatók értéke egytől, annál nagyobb a multikollinearitás mértéke.
- Miután megállapítottuk, hogy megfelelő a modell magyarázó ereje, a modellben csak megfelelő változók szerepelnek, felírhatjuk a **regressziós modell egyenletét, majd értelmezzük ennek paramétereit**. A regreesziós paraméterek megadják, hogy az adott magyarázóváltozó változására – ceteris paribus – átlagosan hogyan változik az eredményváltozó értéke. A pontos értelmezés a modell típusától függ (lásd az Excel alkalmazása részben).

A fentieken kívül még ajánlatos megvizsgálni az **autokorreláció** jelenlétét. 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.
Az autokorreláció különböző rendű lehet. Ha a hibatag a közvetlenül előtte levő értékkel áll lineáris korrelációs kapcsolatban, akkor elsőrendű autokorrelációról beszélünk. Az elsőrendű autokorreláció tesztelésére számos próba létezik. A leggyakrabban a DURBIN-WATSON-féle próbát alkalmazzuk.
Empirikus elemzések alkalmával hasznos grafikusan ábrázolni az egymást követő reziduumok értékeit egy olyan grafikonon, amelynél az abszcisszatengelyen az e_{i-1}, míg az ordinátatengelyen az e_i értékeket tüntetjük fel, ahogy az például az alábbi ábrán látható. A kapott pontdiagram alapján általában már következtetni tudunk az esetleges autokorreláció jellegére.

Az autokorrelált reziduumok grafikus ábrázolása



Míg az idősoros adatoknál az autokorreláció okoz legtöbbször gondot, addig a keresztmetszeti adatok esetében a **heteroszkedaszticitás**. Ez azt jelenti, hogy a hibatagok varianciái nem állandóak. Ennek általában az az oka, hogy a hibatag nagysága függ valamelyik változótól.
A homoszkedaszticitás vizsgálatára több módszer ismeretes, melyek közül a BARTLETT-féle próbát és LEVENE-féle tesztet említjük meg.
Mindkettő feltételezi a normális eloszlást, de a BARTLETT-féle próba sokkal érzékenyebb rá, míg a LEVENE-féle teszt robosztusabb erre a feltételezésre.
Mindkét próba nullhipotézise szerint a hibatagok homoszkedasztikusak, azaz a modell nem heteroszkedasztikus.

Az adatállományunkban szereplő **extrém érték**ek (**outlier**s) a magyarázóváltozó szempontjából és az eredményváltozó szempontjából is lehetnek kilógóak. A regressziós modell paraméterei általában az eredményváltozó szempontjából kilógó adatokra érzékenyebbek. Az extrém értékek elég nagy mértékben befolyásolják a regressziószámítás eredményeit.
Ennek vizsgálatára az egyik lehetséges próba a GRUBBS-féle teszt. Ez feltételezi a tesztelt változó normális eloszlását. A nullhipotézis az outlierek hiánya, míg az alternatív hipotézis szerint létezik legalább egy outlier a sokaságban.

8. Klaszteranalízis

A klaszteranalízis arra a problémára keresi a megoldást, hogy hogyan rendezhetjük megfigyeléseinket – azok hasonlósága, illetve különbözősége alapján – valamilyen struktúrába úgy, hogy ezzel egy csoportosítást hajtsunk végre. Mivel osztályozásról van szó, ezért mindegyik objektum pontosan egy **klaszter**be (csoportba) kerülhet. Az osztályozásnak stabilnak és optimálisnak kell lennie.
A hasonlóság mértéke az objektumok páronkénti távolsága. A leggyakrabban alkalmazott távolságmetrika az EUKLIDESZI távolság vagy annak négyzete. A különböző távolságmértékek használata eltérő klasztermegoldásokhoz vezet.

Mivel a változók mértékegységei nagyon eltérhetnek egymástól, ezért érdemes standardizálni az adatokat. Ez egy természetes súlyozása a dimenzióknak. Az egyes dimenziók azonban az elemzés szempontjából különböző fontosságúak lehetnek, ezért a standardizáláson kívül a gyakorlatban gyakran súlyozzuk a változókat valamilyen – szakmai becslés szerinti – súlyokkal.

A gyakorlatban több klaszterezési eljárás ismeretes, amelyek elsősorban az alkalmazott metrikában és az alkalmazott klaszterezési módszerben különböznek egymástól.

Hierarchikus (összevonó, felosztó) klaszterezés
Ez a módszer külön-külön mindegyik objektumot egy-egy klaszternek tekinti, majd ezeket összevonja. Megkeresi a két legközelebbit és először azokat egyesíti egy klaszterré, majd keresi újra a két legközelebbit. Egy lépésben két klasztert von össze mindaddig, amíg nincs egyetlen klaszterben az összes objektum. A módszer azért hierarchikus, mert egy összevonással összekerült objektumok (klaszterek) ezután végig együtt maradnak, és az összevonásban a már előző összevonások eredményeképpen előálló klaszterek szerepelnek.
Például, ha van *N* objektumunk, azaz klaszterünk, akkor első lépésben a két legközelebbit egyesítjük. Így *N*–1 darab klaszterhez jutunk. A második lépésben ezen *N*–1 darab klaszter közül választjuk ki a két legközelebbit és egyesítjük egy klaszterré. Az eljárás végén minden objektum 1 klaszterben lesz.

A módszer visszafelé is alkalmazható, amikor az összes objektumot egy klaszternek vesszük és azt bontjuk részekre mindaddig, amíg nincs minden objektum egyedül egy klaszterben. Ezt nevezzük **felosztó klaszterezés**nek. Az összevonó típusú módszerek az elterjedtebbek.

A hierarchikus klaszterezés eredményét **dendrogram**mal szemléltetjük. Ez olyan fastruktúra (gráf), amelyben két objektum olyan szinten érhető el egymásból, amilyen szinten egy klaszterbe kerültek.
A klaszterek távolsága az alkalmazott módszertől függ. A legelterjedtebb távolságértelmezések: a legközelebbi szomszéd elve, a legtávolabbi szomszéd elve, az átlagmódszer, a centroidmódszer, a WARD-féle variancia-módszer.

Nemhierarchikus (K-közép) klaszterezés
Ennél a módszernél adott a kialakítandó klaszterek száma *k*. A klaszterek számát az elemzés kezdetén meg kell adnunk. A módszer először valamilyen egyszerű eljárással kialakítja a kezdeti klaszter-magpontokat (pl. az adatbázis első *k* rekordja alapján). Ezután iteratív módszerrel keresi a végső klaszter-középpontokat és az azokhoz tartozó objektumokat. Minden lépésben besorolja az objektumokat aszerint, hogy melyik klaszter-középponthoz van a legközelebb a *k* közül, majd újraszámítja a klaszter-középpontokat. Az eljárás akkor ér véget, ha egy lépésben már nem változnak a klaszter-középpontok.

A kétféle módszer közötti választást általában az adatbázis mérete befolyásolja. A hierarchikus klaszterezésnél ugyanis mindegyik lehetséges klaszter-pár közötti távolságot ki kell számítani, és ez már például 200 rekordnál is sokáig tarthat, nem beszélve arról, hogy a klaszterezés eredményeképpen előálló dendrogram áttekinthetetlen lesz. A *K*-közép módszer problémája, hogy előre meg kell adni a

végső klaszterek számát, ezért a gyakorlatban több megoldást is érdemes kipróbálni. Igaz ugyan, hogy a hierarchikus klaszterezés sem adja meg a klaszterek számát, de ez a dendrogram alapján eldönthető. Tulajdonképpen nincs olyan teszt vagy eljárás, amely megadná, hogy egy adathalmazt hány csoportba érdemes klaszterezni.

9. Többdimenziós skálázás (MDS)

A többdimenziós skálázás egy olyan elemzési technika, amellyel egy halmaz objektumainak páronkénti távolságait ismerve, azok geometriai reprezentációját készítjük el. Az objektumokat egy, az eredeti adatok dimenziójánál alacsonyabb dimenziós térben képzeljük el úgy, hogy ezek a lehető legjobban reprezentálják az objektumok közötti ismert különbözőségeket. Két hasonló objektumot ebben a térben két egymáshoz közeli pont, míg két jelentősen különböző objektumot két távoli pont reprezentál.
Az MDS segítségével lehetséges a jelentéssel bíró háttérdimenziók meghatározása, illetve az objektumhalmaz struktúrájának feltérképezése.

Az adatok mérési szintje és az adatmátrix formája alapján többféle eljárás – például metrikus MDS, nemmetrikus MDS – különböztethető meg.

Az MDS outputja egy mátrix, ami mindegyik objektum koordinátáit tartalmazza az alacsonyabb dimenzióban. A koordináták értelmezését az ábrázolt pontok helyzetének vizsgálatával végezzük. Megpróbáljuk meghatározni a tengelyek jelentését, irányát. A tengelyek beskálázásával minden objektumhoz skálaértéket kapunk a tengelyeknek megfelelő dimenziók mentén.

Az MDS „jóságának” mérésére több mutató ismeretes. Ezek közül a legelterjedtebb az ***S*-stress**. A mutató az objektumok eredeti és a modell által meghatározott térben létrejött pontok távolságainak eltérését viszonyítja az eredeti különbözőségekhez. Minél kisebb az *S*-stress értéke, annál jobban illeszkedik a modell. A mutató értékeinek jelentését az alábbi táblázat tartalmazza.

Az *S*-stress értelmezése

S-stress	**Minősítés**
[0,00 ; 0,05)	Kiváló, valószínűleg minden releváns információt tartalmaz
[0,05 ; 0,10)	Jó
[0,10 ; 0,20)	Elfogadható
[0,20 ;1,00]	Meg kell próbálni eggyel nagyobb dimenziószámú modellt alkalmazni.

10. Főkomponensanalízis

Többváltozós elemzések esetén gyakran jelent problémát a vizsgált változók korreláltsága. A főkomponenselemzés segítségével ezen változók lineáris transzformációjával olyan – a magyarázóváltozók *m* számánál kevesebb – mesterséges (hipotetikus), egymástól független változók (főkomponensek) állíthatóak elő, amelyek lényeges információveszteség nélkül biztosítják a változók függetlenségét.
A főkomponensanalízis során előállított új, mesterséges változók egymástól már függetlenek, és – különösen erős multikollinearitás esetén – az első néhánnyal már meg tudjuk magyarázni az eredeti változók szórásnégyzetének igen nagy hányadát. A magyarázóváltozók sorrendjétől függetlenül, a főkomponensek szórása, azaz az információtartalma rendre csökken. Az egyes főkomponensek információtartalmát megkaphatjuk a magyarázóváltozók korrelációs mátrixából kiszámítható sajátértékként.
Mivel általában néhány főkomponens már jól jellemzi a mintában rejlő információt, a többi elhanyagolható, számuk csökkenthető.
Szignifikáns multikollinearitás esetén azokat a főkomponenseket, amelyekhez tartozó sajátérték 1-nél kisebb (vagyis nem éri el az átlagot) általában már nem vesszük figyelembe.

Az elemzés során egyrészt megkapjuk a **főkomponenssúly**okat (**loading változó**kat) is, amelyek megadják a vizsgált változók és a főkomponensek közötti lineáris korrelációs együtthatót, másrészt pedig a kommunalitásokat.
A kumulált főkomponenssúly-négyzetek azt fejezik ki, hogy az egyes főkomponenseknek milyen jelentősége, súlya van a megfigyelt változók varianciájában, azaz az első *w* darab főkomponens milyen mértékben járul hozzá a *k*-adik standardizált változó szórásnégyzetéhez. $(k, \mathrm{w} \le m)$

Például $h_4^{(3)} = a_{41}^2 + a_{42}^2 + a_{43}^2$ azt mutatja, hogy az első három főkomponens a negyedik standardizált változó szórásnégyzetének $100 \cdot h_4^{(3)}$ százalékát értelmezi. Nyilvánvalóan $h_k^{(m)} = 1$, illetve 100%.

11. Faktoranalízis

A faktoranalízis a főkomponensanalízis speciális esetének tekinthető.
Többváltozós elemzéseknél, ahol a változók között komplex kapcsolat van, hasznos módszer a faktoranalízis, amelynek segítségével csökkenthető a változók száma, azáltal, hogy a megfigyelt változók információtartalmát néhány hipotetikus (faktor-) változóba vonjuk össze. A faktorváltozók gyakran értelmezhetőek, konkrét jelentéssel bírnak, de közvetlenül nem figyelhetőek meg, létezésüket a változók sztochasztikus kapcsolatai alapján feltételezzük és értékeit ezen keresztül becsüljük.
A célunk az, hogy a nagyszámú korrelált változó közötti összefüggéseket megmagyarázzuk, ennél kevesebb korrelálatlan látens faktor segítségével, a megfigyelt változók közös és egyedi tulajdonságainak szétválasztásával. Ezáltal feltérképezhető a változók közötti kapcsolat, illetve segítségével csökkenthető a változók száma az adathalmaz további elemzésében.

Egy faktormodell megoldásaként kapott faktorok azonosítása, interpretálása gyakran nehézkes. Ekkor hasznos lehet a faktorsúlyok ortogonális transzformációja, amely a koordináta-rendszer rotációját jelenti. Ez megkönnyíti a faktorok felismerését. A transzformáció sokféleképpen elvégezhető, különböző kritériumoknak eleget tevő módszerek léteznek, legáltalánosabban alkalmazott a varimax, amely a kvadratikus faktorsúlyok szórásnégyzetét maximalizálja.

A faktoranalízis alkalmazásának akkor van értelme, ha a változók redundáns módon tartalmaznak információt, és mögöttük egy látens struktúra húzódik meg. Az előbbit a BARTLETT-féle teszt, az utóbbit a **KAISER-MEYER-OLKIN (KMO) mutató** segítségével vizsgálhatjuk.
A BARTLETT-féle próba azt vizsgálja, hogy a változóink korrelációs mátrixa mennyire hasonlít egy egységmátrixhoz, vagyis változóink páronként korrelálatlanok-e. A teszt egy χ^2-próba, aminek nullhipotézise a korrelációs mátrix és az egységmátrix egyezősége.

A mögöttes struktúra létét akkor feltételezhetjük, ha nem csak változópárok vannak egymással kapcsolatban, hanem sztochasztikus kapcsolatok rendszere jelentkezik. A KMO mutató egy 0 és 1 közé eső számot ad. A KMO mutató értelmezése az alábbi táblázatban található.

Útmutató a KMO értelmezéséhez

KMO	Minősítés
[0,0 ; 0,5)	elfogadhatatlan
[0,5 ; 0,6)	szánalmas
[0,6 ; 0,7)	mérsékelt
[0,7 ; 0,8)	közepes
[0,8 ; 0,9)	dicséretes
[0,9 ; 1,0]	csodálatos

A KMO magas értékei tehát azt jelzik, hogy a faktorelemzés alkalmazása eredményes lehet.

12. Idősorok vizsgálatának alapjai

Idősorok vizsgálatakor valamilyen jelenség, sokaság időbeli változását, alakulását vizsgáljuk. Ennek alapjaival, azaz a bázis- és a láncviszonyszámokkal már korábban találkoztunk.

A viszonyszámokon túl lehetőség van az időbeli változás **modellezésére** is.
Az egyik legegyszerűbb vizsgálati módszer a determinisztikus idősorelemzés. E szerint az idősorban matematikailag jól kezelhető, hosszú távú trendek vannak. A determinisztikus idősorelemzés leggyakrabban alkalmazott modellje a dekompozíciós idősormodell. Ez azt feltételezi, hogy az idősorok alakulását négy fő összetevőre bonthatjuk.

1. **Trend**: hosszabb időszakon át, tartósan meglevő tendencia (átlagos mozgásirány). Ez az alapirányzat, amit a vizsgált jelenségre ható alapvető

gazdasági, társadalmi tényezők alakítanak ki. Az idősorok legfontosabb összetevői.
A trend meghatározása történhet a **mozgóátlagok módszere** alapján, illetve **analitikus trendszámítás** segítségével. Trendszámítás során célunk az idősor kisimítása.

2. **Szezonális hatás**: szabályos ingadozás a trend körül, amely rendszeresen ismétlődő hullámzást jelent. Általában egy éven belül jelentkezik, természeti tényezőkkel, társadalmi szokásokkal magyarázható.
3. **Ciklikus komponens**: kevésbé szabályos, hosszú ingadozások a trend körül. Ilyenek például a konjunktúra-ciklusok. Ennek a komponensnek a vizsgálatával nem foglalkozunk.
4. **Véletlen tényező**: az eddigi összetevőkkel nem magyarázható szabálytalan ingadozások.

Aszerint, hogy a négy összetevő hogyan tevődik össze, megkülönböztetünk **additív** és **multiplikatív** modelleket. Az additív modelleknél a tényezők összeadódnak, míg multiplikatív modellek esetében összeszorzódnak.
Az egyszerűség kedvéért, a továbbiakban feltételezzük, hogy a véletlen tényező várható értéke nulla. Az idősorok – statisztikai szoftverekkel történő – alapszintű elemzése során először a **trend** vizsgálata történik. Ez vagy a mozgóátlagok módszerével, vagy pedig analitikus úton történhet. Ezután következhet a **szezonális hatás** vizsgálata. Bármelyik módszert is választjuk, menetközben lehetőségünk van előrejelzések készítésére. Természetesen nagyon fontos az adatok **grafikus megjelenítés**e is.

A mozgóátlagok módszere

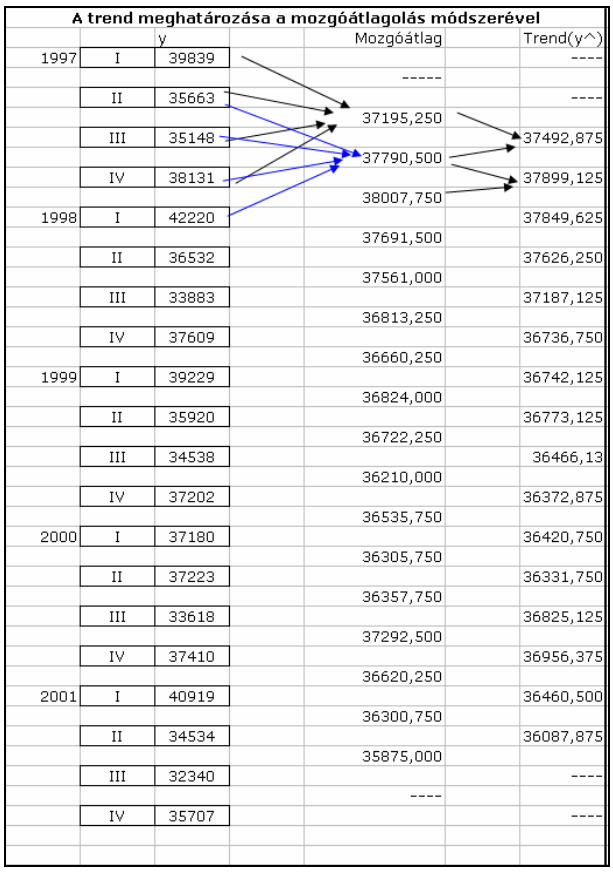
Ennél a módszernél a trendet átlagszámítás segítségével határozzuk meg. Minden egyes időszakban (általában egy időszak= egy negyedév, vagy egy hónap) kiszámítjuk valamekkora körzetben az adatok átlagát. Ezért hívjuk a módszert mozgóátlagolásnak. Nagyon fontos az, hogy hogyan határozzuk meg a mozgóátlag tagszámát. Mivel a módszer arra épül, hogy egy perióduson (általában egy periódus = egy év) belül a szezonális hatás kinullázza magát, ezért negyedéves bontású idősor esetében a mozgóátlag tagszáma 4, vagy ennek egészszámú többszöröse, havi bontású idősor esetében a mozgóátlag tagszáma 12, vagy ennek egészszámú többszöröse lehet.
Páros tagszámú mozgóátlag esetén a kapott mozgóátlagokat még **centrírozni** kell, azaz az időszakokhoz kell ezeket igazítani.
Minél nagyobbra választjuk a mozgóátlag tagszámát, az idősor annál jobban rövidülni fog. Ez azt jelenti, hogy az idősor elején és végén néhány időszakhoz nem tudjuk a trendértéket meghatározni.

Példa
A halálozások számának alakulását negyedéves bontásban az alábbi táblázat tartalmazza. Határozzuk meg a trendet a mozgóátlagok módszere alapján.

Negyedév	Év				
	1997	1998	1999	2000	2001
I	39839	42220	39229	37180	40919

II	35663 36532 35920 37223 34534
III	35148 33883 34538 33618 32340
IV	38131 37609 37202 37410 35707

Első lépésben a mozgóátlag tagszámát kell meghatároznunk. Mivel az adatok negyedéves bontásban vannak megadva, így a mozgóátlag tagszáma vagy négy, vagy ennek egészszámú többszöröse lehet. Minél nagyobbnak választjuk ezt, annál nagyobb mértékben fog rövidülni az idősor. Ezért a mozgóátlag tagszámát négynek választjuk. Ezután kiszámítjuk a mozgóátlagokat, azaz minden „adat-négyesnek" vesszük a számtani átlagát. Mivel a mozgóátlag páros tagszámú, így ezek centrírozásával, azaz a szomszédos két mozgóátlag átlagolásával kapjuk meg a keresett trendértékeket.



A trend meghatározása a mozgóátlagolás módszerével

		y	Mozgóátlag	Trend(y^)
1997	I	39839		----

	II	35663		----
			37195,250	
	III	35148		37492,875
			37790,500	
	IV	38131		37899,125
			38007,750	
1998	I	42220		37849,625
			37691,500	
	II	36532		37626,250
			37561,000	
	III	33883		37187,125
			36813,250	
	IV	37609		36736,750
			36660,250	
1999	I	39229		36742,125
			36824,000	
	II	35920		36773,125
			36722,250	
	III	34538		36466,13
			36210,000	
	IV	37202		36372,875
			36535,750	
2000	I	37180		36420,750
			36305,750	
	II	37223		36331,750
			36357,750	
	III	33618		36825,125
			37292,500	
	IV	37410		36956,375
			36620,250	
2001	I	40919		36460,500
			36300,750	
	II	34534		36087,875
			35875,000	
	III	32340		----

	IV	35707		----

Az analitikus trendszámítás

Ennél a módszernél a trendet, mint regressziófüggvényt határozzuk meg. A magyarázóváltozók vagy az évszámok lesznek, vagy egy mesterséges változót vezetünk be az időre.

A gyakorlatban analitikus trendszámításkor többféle megoldási utat lehet használni. Ez azzal a következménnyel jár, hogy ugyanazt a trendvonalat többféle egyenlettel is meg lehet adni. Ez pedig azt jelenti, hogy a trendegyenletek paramétereinek más a jelentése. Ezért a lehetséges félreértések elkerülése végett – a trendegyenlet mellett – kötelezően meg kell adni azt is, hogy ennek kiszámításakor mit vettünk kiindulópontnak, valamint, hogy a tengelyeken 1 egység mit jelent.

Idősorok vizsgálatakor mindkét „regressziós paraméter” értelmezhető. Például lineáris trend esetén az egyenes meredeksége megadja a vizsgált jelenségnek az időegységenkénti átlagos változást a vizsgált időszakban, míg a tengelymetszet a kiinduláskor adja meg az idősor értékét.
Ennek bemutatása végett tekintsük az alábbi –már korábban is használt – idősort!

Év	Kivitel (t)
2000	100
2001	105
2002	109
2003	116
2004	120
2005	125

1. megoldás
Az eredeti adatok alapján lineáris regressziószámítást végzünk úgy, hogy az évszám lesz a magyarázóváltozó, míg a kivitel az eredményváltozó.

Ekkor a tengelymetszet azt jelenti, hogy az adott kivitel elméletileg mennyi volt időszámításunk kezdetekor. Ennek természetesen nincsen gazdasági értelme.

A többi megoldási módnál nem az eredeti évszámokkal dolgozunk, hanem egy új, mesterséges – *t*-vel jelölt, lineáris transzformációval kapott – változót vezetünk be az idő jelölésére, így ezeket a megoldási módokat tetszőleges bontású idősor esetén is alkalmazhatjuk.
A lineáris transzformáció leegyszerűsítve azt jelenti, hogy kiválasztunk egy tetszőleges kezdőpontot (*t*=0), majd a *t* változó értékét időegységenként ugyanannyival változtatjuk.
Pontosan ebben különböznek az alábbi megoldási módszerek.

2. megoldás
Ebben a megoldásban a *t*=1 értéket az idősor kezdőértékéhez rendeljük.

Év	y_i	t_i
2000	100	1
2001	105	2
2002	109	3
2003	116	4
2004	120	5
2005	125	6

Ekkor a tengelymetszet azt jelenti, hogy az adott kivitel trendszerinti értéke mennyi volt 1999-ben.

Az utolsó két megoldási mód esetén a *t*=0 értéket az idősor közepéhez rendeljük. Ez azzal az előnnyel jár, hogy a normálegyenlet-rendszer leegyszerűsödik, mivel ekkor a *t* változó összege nullával egyenlő. Ezért, ezt a

megoldási módot $\sum_{i=1}^{n} t_i = 0$ jelöléssel illetik. Mint látni fogjuk, ez a jelölés hiányos, mivel a *t* értékek időegységenkénti változtatására nem utal.

3. megoldás
Ebben a megoldásban időegységenként a *t* értéket eggyel változtatjuk.

Év	y_i	t_i
2000	100	–2,5
2001	105	–1,5
2002	109	–0,5
2003	116	0,5
2004	120	1,5
2005	125	2,5

Ekkor a tengelymetszet azt jelenti, hogy mennyi volt az adott kivitel trendszerinti értéke 2002 végén.

4. megoldás
Ebben a megoldásban időegységenként a *t* értéket kettővel változtatjuk.

Év	y_i	t_i
2000	100	–5
2001	105	–3
2002	109	–1
2003	116	1
2004	120	3
2005	125	5

Ekkor a tengelymetszet azt jelenti, hogy az adott kivitel trendszerinti értéke mennyi volt 2002 évvégén.
Az egyenes meredeksége viszont most azt mutatja meg, hogy a kivitel a vizsgált időszakban félévente átlagosan mennyivel változott.

A szezonális hatás vizsgálata

Jelen képzési szintben elégedjünk meg azzal, hogy – a mozgóátlagokhoz hasonlóan – a szezonális hatás vizsgálatának csak akkor van értelme, ha az adatok bontása egy évnél „sűrűbb”, azaz például havi, negyedéves, harmadéves, stb. adatok állnak a rendelkezésünkre.
Mivel a ciklikus komponenstől eltekintünk a vizsgálataink során, ezért a szezonális hatás vizsgálata az alábbiak szerint történik.
Először azt kell eldöntenünk, hogy additív, vagy multiplikatív modellt vizsgálunk. Additív modell esetén **szezonális eltéréseket**, míg multiplikatív modell esetén szezonális indexeket (**szezonindexeket**) vizsgálunk.
A szezonális eltérések megadják, hogy a vizsgált változó tényleges értékei átlagosan mennyivel térnek el a trendtől a vizsgált időszakban.
A szezonindexek megadják, hogy a vizsgált változó tényleges értékei átlagosan hány százalékkal térnek el a trendtől a vizsgált időszakban.

Mivel additív modell esetén

Vizsgált változó értéke= Trend+szezonális eltérés;

multiplikatív modell esetén

Vizsgált változó értéke= Trend*szezonindex;

ezért, egyrészt a

vizsgált változó értéke – Trend;

illetve

a vizsgált változó értéke / Trend

változókat kell meghatározni.

Megjegyzem, analitikus trendszámítás esetén az előbbit a számítógépes kimeneten a *Maradék Tábla Maradékok* oszlopa adja. Az utóbbit a vizsgált változó tényleges és becsült értékeinek hányadosaként számíthatjuk ki.
Amelyik változat nagyobb állandóságot mutat, olyan típusúnak tekintjük a modellt. Ennek eldöntése meglehetősen szubjektív. Az idősor grafikus ábrázolásából is sejtéseket fogalmazhatunk meg a modell típusára vonatkozóan. Amennyiben az ingadozás mértéke állandónak tekinthető a trendhez képest, akkor additív, amennyiben a trenddel arányosnak tekinthető a trend körül, akkor multiplikatív modelltípusra következtethetünk.

Ha eldöntöttük, hogy milyen típusú a modellünk, akkor következhet a szezonális hatás tényleges vizsgálata. Minden időszakban a szezonális hatást egyetlen számszerű értékkel kell jellemeznünk, azaz például negyedéves bontás esetén egy-egy számot kapunk az első, a második, a harmadik, a negyedik negyedévre.

Additív modell esetén a kiszámított különbségeknek időszakonként kiszámítjuk a számtani átlagát, így minden negyedévre vonatkozóan megkapjuk a **nyers szezonális eltéréseket**. Mivel a szezonalítás egy perióduson belül kiegyenlítődik, ezért a szezonális eltérések összegének nullának kell lennie. Amennyiben a nyers szezonális eltérések összege nulla, akkor ezek az értékek lesznek a **tényleges szezonális eltérések**. Amennyiben ezek összege nem nulla, akkor korrigálnunk kell, még pedig minden egyes értéket ugyanazzal a **korrekciós tényezővel**. A korrekciós tényező a nyers szezonális eltérések számtani átlaga lesz. Ezt az értéket egyszerűen kivonjuk a nyers szezonális eltérésekből. A kapott számok lesznek a tényleges szezonális eltérések.

Multiplikatív modell esetén a kiszámított hányadosoknak időszakonként kiszámítjuk a mértani átlagát, így minden negyedévre vonatkozóan megkapjuk a **nyers szezonális indexeket**. Mivel a szezonalítás egy perióduson belül kiegyenlítődik, ezért a szezonális indexek szorzatának egynek kell lennie. Amennyiben a nyers szezonális eltérések szorzata egy, akkor ezek az értékek lesznek a **tényleges szezonindexek**. Amennyiben ezek szorzata nem egy, akkor korrigálnunk kell, még pedig mindenegyes értéket ugyanazzal a

korrekciós tényezővel. A korrekciós tényező a nyers szezonális indexek mértani átlaga lesz. Ezzel az értékkel egyszerűen leosztjuk a nyers szezonális indexeket. A kapott számok lesznek a tényleges szezonindexek.

A trend és a szezonális hatás meghatározása után lehetőségünk van előre jelzések készítésére, feltételezve, hogy a vizsgált jelenség mozgásiránya nem változik meg.