KVANTITATÍV ELEMZÉSI MÓDSZEREK A GAZDÁLKODÁS- ÉS SZERVEZÉSTUDOMÁNYOK TERÜLETÉN

DEBRECENI EGYETEM

Debrecen, 2025

DEBRECENI EGYETEM
GAZDASÁGTUDOMÁNYI KAR

KVANTITATÍV ELEMZÉSI MÓDSZEREK A GAZDÁLKODÁS- ÉS SZERVEZÉSTUDOMÁNYOK TERÜLETÉN

Debrecen
2025

Kvantitatív elemzési módszerek a gazdálkodás- és szervezéstudományok területén

Szerkesztette:
Prof. Dr. Balogh Péter

A fejezetek szerzői:
Prof. Dr. Balogh Péter
Dr. Huzsvai László
Dr. Lengyel Péter
Dr. Szenderák János

Lektorok:
Prof. Dr. Jámbor Attila (Budapesti Corvinus Egyetem)
Dr. Kovács Péter (Szegedi Egyetem)

A könyv a Magyar Nemzeti Bank támogatásával készült.

ISBN (pdf): 978 963 490 727 5

Felelős kiadó: Debreceni Egyetem Gazdaságtudományi Kar
Kiadásért felelős személy: **Prof. Dr. Fenyves Veronika** dékán

Debrecen
2025

Tartalomjegyzék

1. Bevezetés

A kvantitatívkutatás eredményeként nem csak információkat kapunk a fogyasztók viselkedéséről illetve a vállalkozások tevékenységeiről, hanem ezek alapján olyan következtetések levonása is lehetséges lesz, amelyet előzetesen – ezen adatok ismeretek nélkül – nem tudtunk volna megalapozni (MALHOTRA, 2010). Napjainkra az információstechnológia elterjedésével – számos statisztikai programcsomag segítségével – rövid idő alatt el tudjuk végezni az összegyűjtött adataink elemzését, felgyorsítva ezzel a kutatási folyamatot. Ugyanakkor a technológia még nem képes helyettesíteni a szakmai tudást, amivel az eredményeinket szakszerűen értelmezni tudjuk.

A kutatási folyamat 4 fő lépése (KOTLER – ARMSTRONG, 2020) közül ebben a könyvben a 3. azaz az **Adatelemzés** rész kerül bemutatásra.
A statisztikai adatelemző módszerek helyes alkalmazásának feltétele a megszerzett információk szakszerű értelmezése, amihez szükséges a marketingkutatási és statisztikai alapfogalmak pontos ismerete is. Az eltérő statisztikai programcsomagok (pl. IBM SPSS, R program különböző csomagjai) lehetővé teszik számunkra a jelenségek gyors és sokoldalú vizsgálatát, de ha az általunk begyűjtött információk nem fedik le teljesen a valóságot – esetleg tévesek – abban az esetben az ezekből kiszámított eredmények és a levonható következtetések hamisak lesznek (MALHOTRA – SIMON, 2016; KOTLER – KELLER, 2017).
Az információ jellegzetessége szerint beszélhetünk kvantitatív, mennyiségi és kvalitatív, minőségi jellegű kutatásról. A kvantitatív (mennyiségi) módszerekhez a számszerű és számszerűsíthető információk, illetve ezek megszerzésének módszerei tartoznak.
A kvantitatív mennyiségi kutatási eljárások során, reprezentatív minta alapján az egész sokaságra általánosítható eredményekhez juthatunk statisztikai elemzések segítségével.

A kvantitatív kutatás az adatok jellege szerint mennyiségi, a minta elemszáma szerint nagy elemű (reprezentatív) és a célja szerint az alapsokaságra általánosítható, számszerűsíthető eredményeket vár el. A kutatási eszköz általában a standardizált kérdőív, az eredmények feldolgozása matematikai, statisztikai eljárások alkalmazásával történik, ugyanakkor a kiszámított eredmények hasznosításával üzleti döntési javaslatokat, alternatívákat állíthatunk fel (TAMUSné, 2009).
Leggyakrabban kérdőíveket használunk ahhoz, hogy egy adatbázist létrehozzunk. A kérdőíves felmérésben mind a három féle elem (feltáró, leíró és ok-okozati) is jelen lehet (SAJTOS – MITEV, 2007).
A kérdőív különböző kérdéseit külön-külön és egymással összevetve is értékelhetjük. Ez azt jelenti, hogy az alkalmazott elemzési technika függ attól, hogy mennyi kérdést (változót) viszonyítunk egymáshoz.
Egy kérdőívben nem lehet – vagy nem érdemes – egyetlen direkt kérdéssel mérni olyan összetett dolgokat, mint a környezeti attitűd, környezetvédelemmel kapcsolatos viselkedés, lakóhelyhez vagy egy bizonyos termékhez való ragaszkodás. A kérdőív kérdéseire kapott válaszokat a marketing mérésként kezeli, és elfogadhatónak azokat a méréseket tekinti,

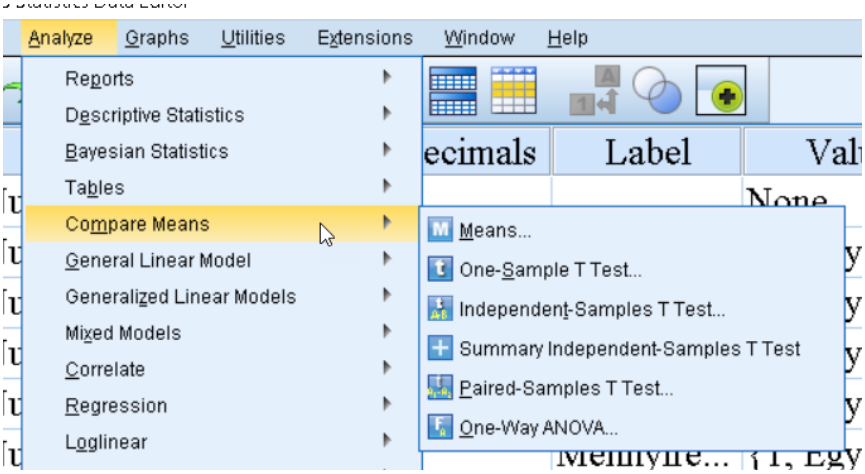
amelyek érvényesek és megbízhatóak. Megbízhatónak akkor tekinthető egy mérés, ha megismételve ugyanazt az eredményt kapjuk. Ezért megállapíthatjuk, hogy ha egy összetett dolgot nem egyetlen, hanem több kérdéssel mérünk, fokozzuk (*fokozhatjuk*) a mérés érvényességét (SZÉKELYI – BARNA, 2002). Azonban a több kérdéssel mérő kérdések megbízhatósága nem minden esetben ennyire egyértelmű.

A kvantitatív elemzési módszerek részletes bemutatása a könyv további fejezeteiben kerül ismertetésre.

2. A kvantitatív kutatási módszerek csoportosítása

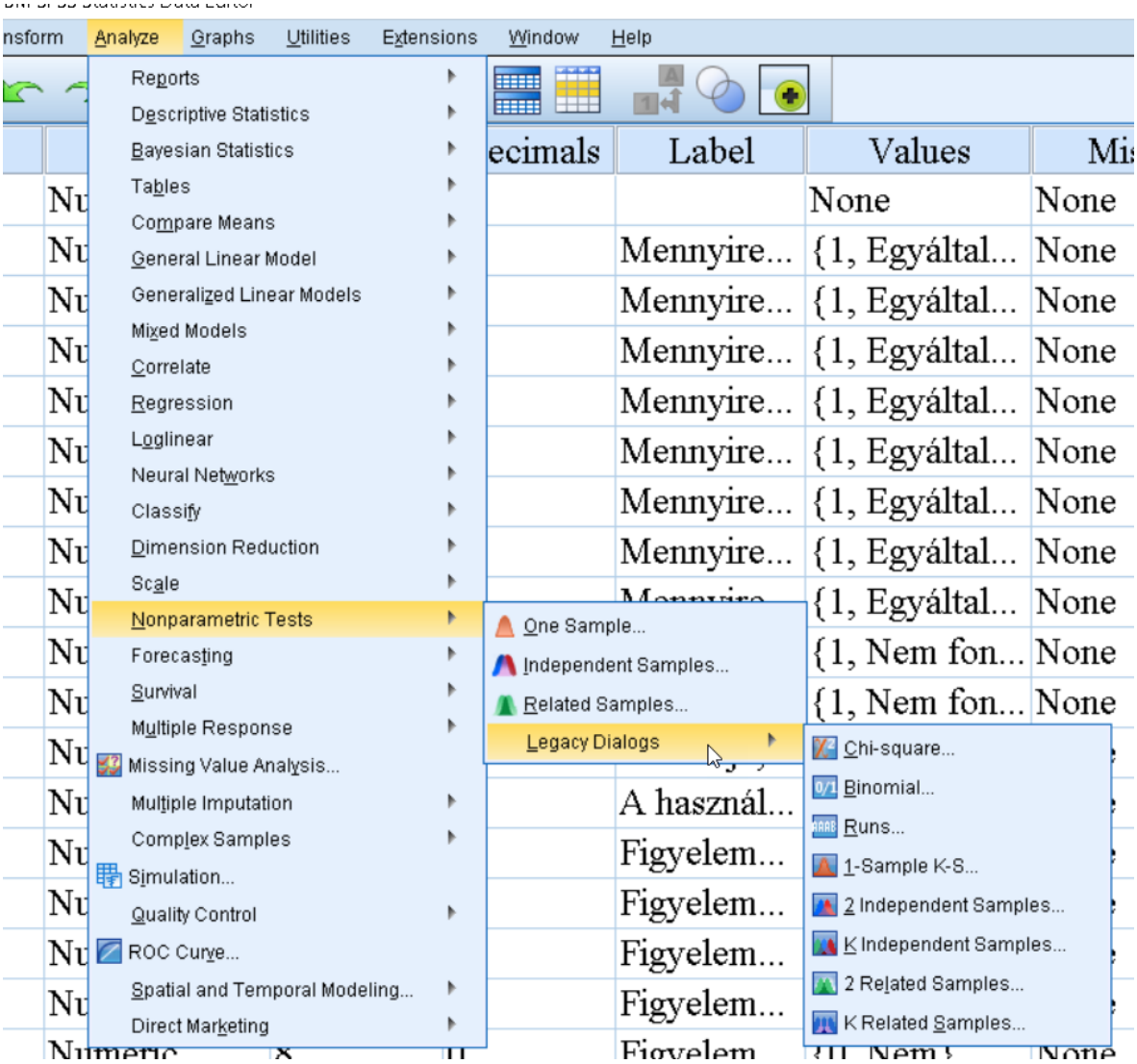
Míg a kvalitatív kutatás eredményei nem számszerűsíthetők, nem lehet belőlük általános következtetéseket levonni, a kvantitatív (mennyiségi) kutatás eredményei számszerűsíthetők, és megfelelő mintavétel esetén általánosíthatók a vizsgált sokaságra. A kutatás kezdetekor a szekunder információk jó kiindulópontot jelentenek. De ezek önmagukban nem mindig biztosítják a kutató illetve vállalatvezető számára a szükséges ismereteket. Ezért szükséges a primer adatok összegyűjtése és elemzése is (KOTLER – KELLER, 2017; MALHOTRA, 2010).
A kvantitatív módszerek között megkülönböztethetünk függőségi és kölcsönös függőségi kapcsolatokat (SAJTOS – MITEV, 2007). Az ezek közötti különbséget az jelenti, hogy az egyik esetében meg tudjuk különböztetni a függő és független változókat, a másik esetben pedig nem tudunk ilyen jellegű megállapítást tenni.
A ***függőségi kapcsolat*** esetében fontos az, hogy a függő és független változók milyen skálán mértek, s ez befolyásolja azt, hogy milyen jellegű elemzést tudunk alkalmazni, amelyek a következők lehetnek:

Paraméteres próbák:
egy mintás t-próba,
két független mintás t-próbák,
párosított mintás t-próba,
egy és több szempontos varianciaelemzés,



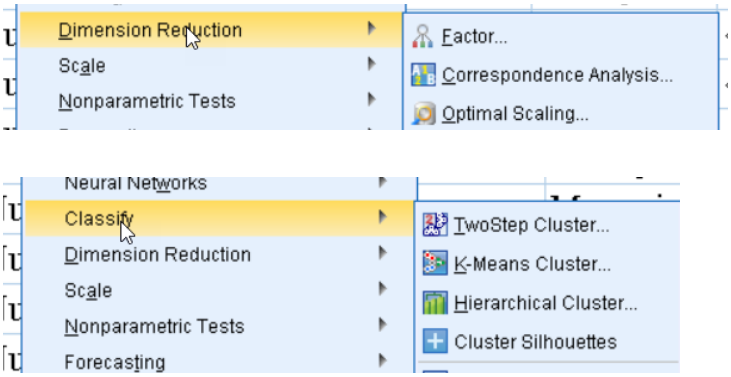
2.1. ábra: **A paraméteres próbák beállításának lehetősége az SPSS programban**

Nem paraméteres próbák:
Kereszttáblaelemzés Khi-négyzet próbával,
Fisher-féle egzakt teszt,
Mann-Whitney próba,
Kruskal-Wallis próba.



2.2. ábra: **A nemparaméteres próbák beállításának lehetősége az SPSS programban**

A ***kölcsönös függőségi kapcsolatok*** között a faktor- (változókat vizsgálunk) és klaszterelemzés (eseteket vizsgálunk).

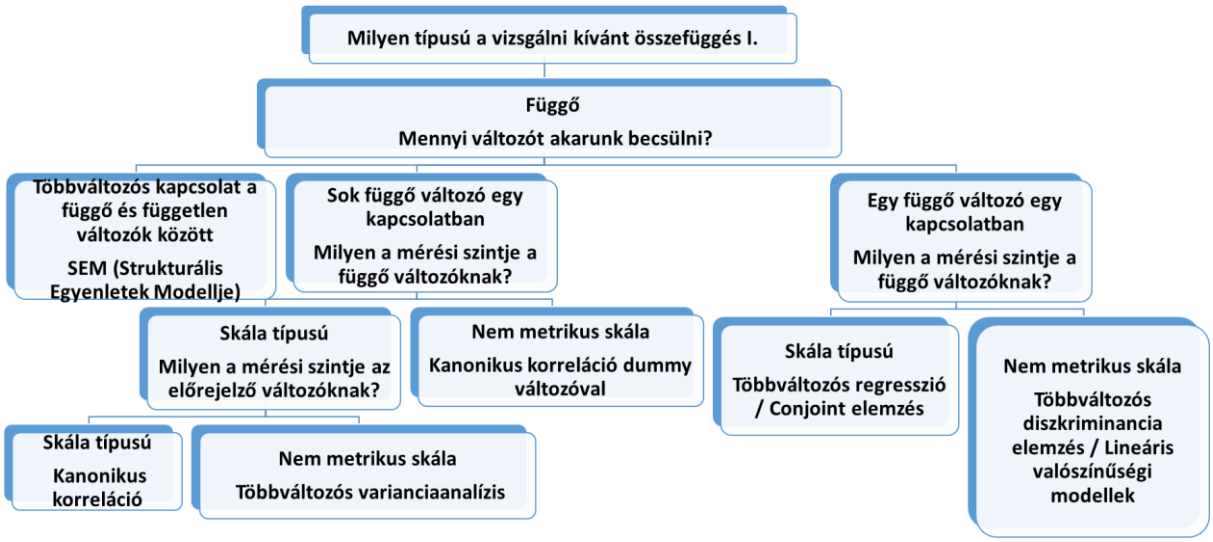


2.3. ábra: **A kölcsönös függőségi kapcsolatok beállításának lehetősége az SPSS programban**

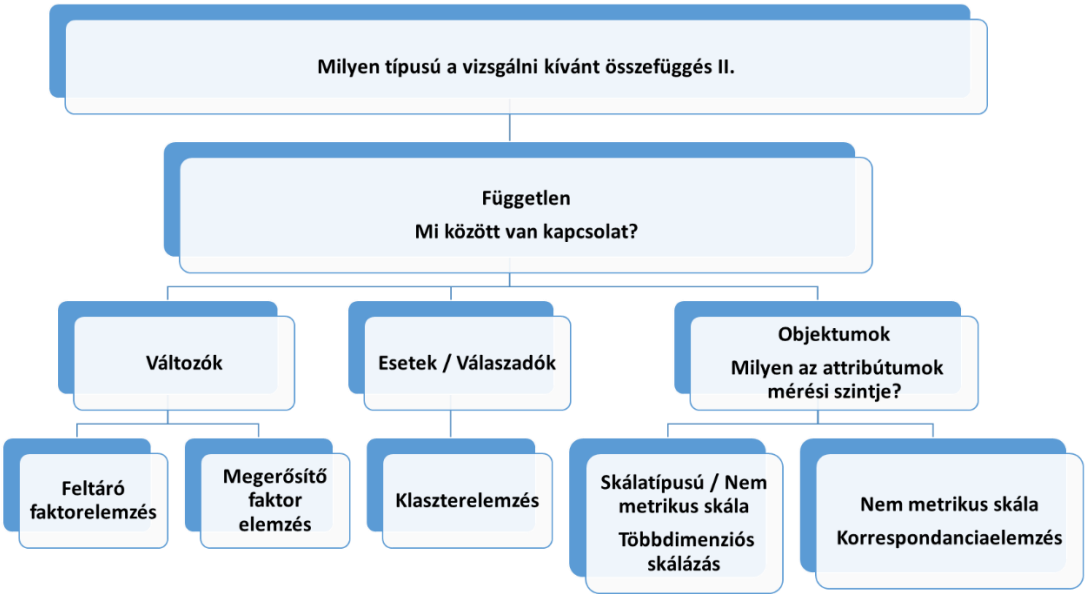
A következő ábrákon bemutatjuk, hogy milyen kérdések alapján lehet eldönteni, hogy melyik többváltozós elemzést alkalmazzuk (HAIR et al., 2014).

- Valamilyen elmélet alapján elkülöníthetők-e a változóink függő és független változókra?
- Ha elkülöníthetők a változók, akkor mennyi a függő változók száma egy elemzésben?
- Milyen mérési szintűek a függő és független változók?

A döntést az előzőekben ismertetett kérdésekre adott válaszok alapján hozhatjuk meg.

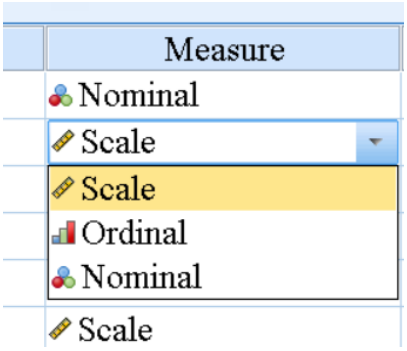


2.4. ábra: **A többváltozós elemzési technikák kiválasztásának szempontjai I.**
Forrás: HAIR et al., (2014) alapján saját módosítás



2.5. ábra: **A többváltozós elemzési technikák kiválasztásának szempontjai II.**
Forrás: HAIR et al., (2014) alapján saját módosítás

A mérési skálák egyik típusát a nem metrikus skálák jelentik, melyek egy tulajdonság meglétét vagy hiányát mutatják. A nem metrikus skálák két típusa a nominális (névleges) és az ordinális (sorrendi) skála. A mérési skálák másik típusát a metrikus skálák jelentik, melyek egyik típusa az intervallum (különbségi), másik típusa pedig az arányskála.



2.6. ábra: **A mérési skálák beállításának lehetősége az SPSS programban**

3. A hipotézisvizsgálatok és jelentőségük az elemzésekben

A hipotézisvizsgálatnak a marketingkutatáson belüli használata nagyon elterjedt és széles körűen alkalmazott eljárás. Alkalmazása arra irányul, hogy egy vagy több sokaságra vonatkozó olyan feltevések – ún. hipotézisek – helyességét vizsgálja, ellenőrizze mintavételi eredményekre támaszkodva, melyek fennállásában nem vagyunk biztosak. A hipotézisek a vizsgált sokaság eloszlására vagy az adott eloszlás egy vagy több paraméterére vonatkozhatnak. A hipotézisek helyességének ellenőrzésére különféle teszteket, próbákat használunk (VITA, 2011). A terjedelmi korlátok miatt most csak röviden teszünk említést az elméleti alapokról. További részletes információ található a különböző statisztikai tankönyvekben (pl. HUNYADI, 2001; RAMANATHAN, 2003; HUZSVAI, 2012; HUZSVAI – BALOGH, 2015).

3.1. A hipotézis vizsgálat menete

A hipotézisvizsgálatra a gyakorlatban minden olyan esetben szükség van, amikor valamely sokaság, illetve eloszlás jellemzőivel kapcsolatban bizonyos feltevéseink vagy elvárásaink vannak, s azok teljesülését nem teljes körű adatfelvételből, hanem csak a sokaságból vett mintából nyert információkra támaszkodva tudjuk vizsgálni (VITA, 2011). A próbák fontos szerepet játszanak a különféle marketingkutatás segítségével elemzett jelenségek leírására törekvő statisztikai modellek építése és használata során.

3.1.1 A szakmai kérdés megfogalmazása, statisztikai formája

Minden hipotézisvizsgálat két egymásnak ellentmondó feltevés: egy H_0 -lal jelölt nullhipotézis és egy H_1 -gyel jelölt ellenhipotézis – más néven alternatív hipotézis – megfogalmazásával kezdődik. A két hipotézisnek olyannak kell lennie, hogy azok kizárják egymást és bármelyik alapján meg tudjuk válaszolni a bennünket érdeklő kérdést.

3.1.2 A szignifikanciaszint meghatározása

Az $1-\alpha$ értéket (pl. $1-\alpha=0,95$) a próba megbízhatósági szintjének, az α (pl. $\alpha=0,05$) értéket pedig szignifikanciaszintnek nevezzük. Ezeket az értékeket százalékos formában szokták közölni.

3.1.3 A próbastatisztika helyes megválasztása, számítása

Ennek során kiszámítjuk a próbafüggvényt, és meghatározzuk a próbastatisztika értékét. Ez a statisztikai teszt kiválasztását és alkalmazását jelenti. Lényeges információ, hogy csak az alkalmazhatósági feltételeknek megfelelő teszte(ke)t alkalmazhatjuk. A próbafüggvény kiszámított értéke az elméleti eloszlás értéke. Ez normáleloszlás esetén a z-érték, míg t-eloszlás esetében a t-érték. Ezeknek az értékeknek az ismeretében megadható egy p valószínűség, amely megmutatja, hogy milyen valószínűséggel vehet fel a próbafüggvény a kiszámítottal azonos vagy nagyobb értéket, ha a H_0 igaz. Ha ez a valószínűség kisebb, mint a szignifikancia-szint,

akkor elutasíthatjuk a nullhipotézist, mert a hiba elkövetésének a valószínűsége kisebb, mint az előre választott megengedhető maximális érték (HUZSVAI, 2012).

3.1.4. A kritikus érték meghatározása, a statisztikai döntés H_0 elfogadásáról

Ha a próbafüggvénynek a minta adataiból számított értéke a visszautasítási tartományba esik, akkor elvetjük H_0 -t, ellenkező esetben megtartjuk azt az adott szignifikanciaszinten.

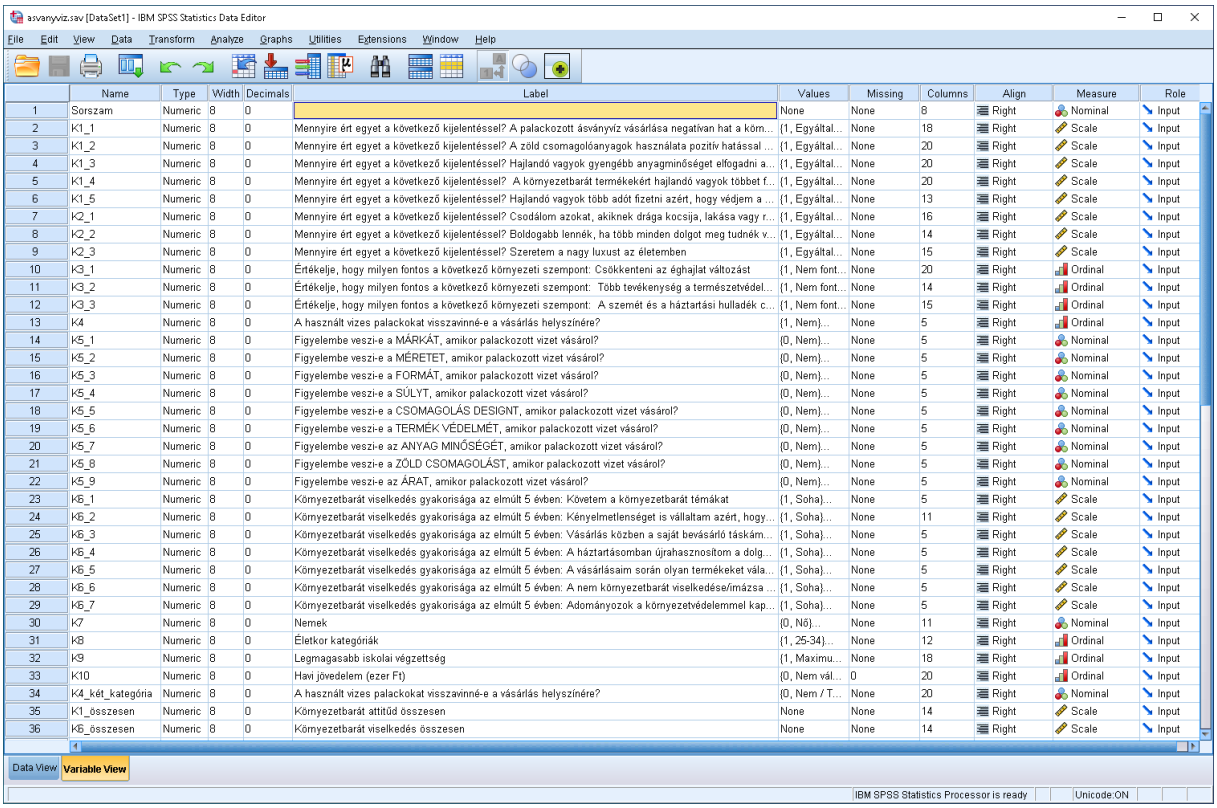
3.1.5. Szakmai következtetés

A hipotézisvizsgálat kapott eredményeinek ismeretében szakmailag kell meghoznunk a döntést arról, hogy milyen következtetést tudunk levonni.

3.2. Paraméteres próbák

Ha az eloszlás jellege ismert, és a nullhipotézisünk az eloszlás valamely paraméterére vonatkozik, akkor beszélünk paraméteres próbáról. Az ilyen jellegű vizsgálatoknál a próbák alkalmazása nominális és ordinális mérési szintű változók esetében nem ajánlott. A középértékek összehasonlítására leggyakrabban alkalmazott próbák az egy és két mintás z-próba és t-próba attól függően, hogy az alapsokaság szórása ismert-e.

Ebben az anyagban a vizsgálataink során egy közös mintaadatbázist (*asvanyviz.sav*) fogunk használni 36 változóval, melyen keresztül bemutatjuk a különböző kvantitatív kutatásokban alkalmazott elemzési technikák kivitelezését és az eredmények értelmezéseit (3.1. ábra). A tisztított adatbázisban eredetileg 33 változó volt (plusz 4 új változót már mi állítottunk elő), míg a kérdőívet helyesen kitöltők száma 906 fő volt. A vizsgálat során számos kérdésre kellett a résztvevőknek válaszolnia az ásványvízzel kapcsolatos fogyasztói szokásokkal kapcsolatban.



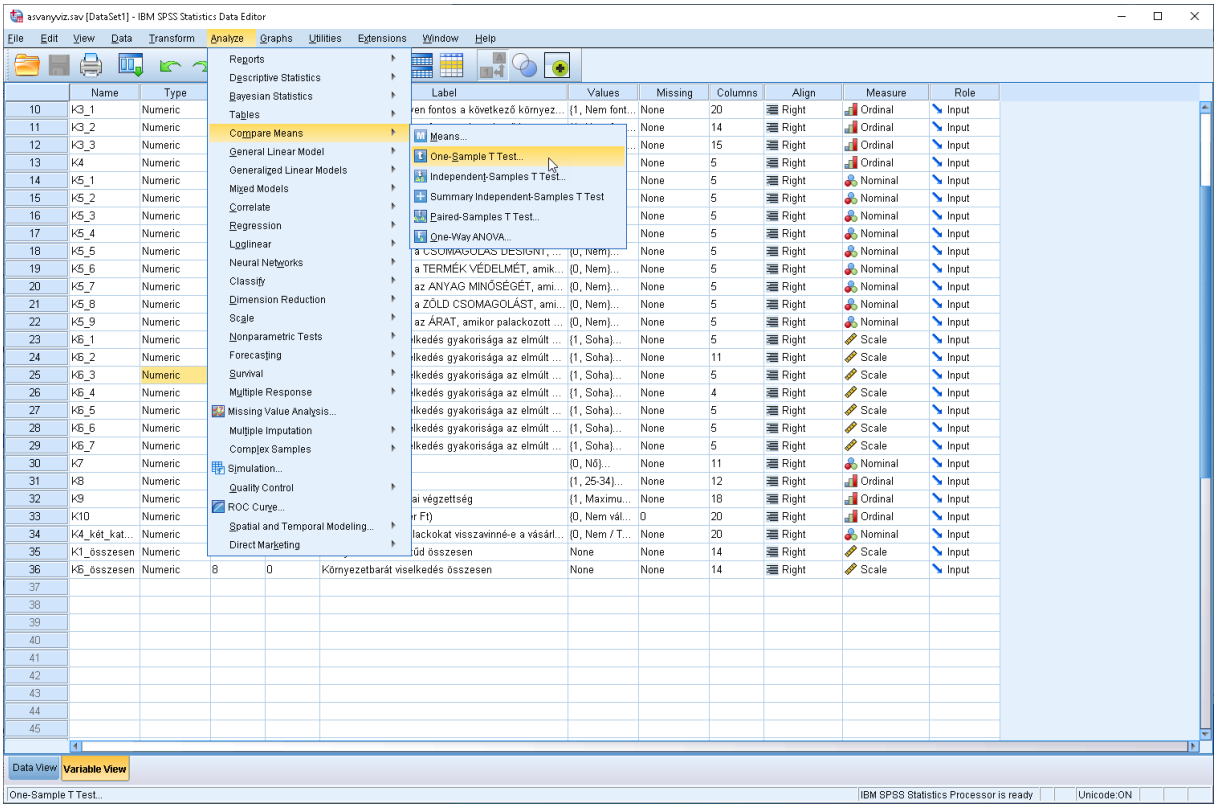
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Sorszam	Numeric	8	0		None	None	8	Right	Nominal	Input
2	K1_1	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a körn...	{1, Egyáltal...	None	18	Right	Scale	Input
3	K1_2	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással ...	{1, Egyáltal...	None	20	Right	Scale	Input
4	K1_3	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni a...	{1, Egyáltal...	None	20	Right	Scale	Input
5	K1_4	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet f...	{1, Egyáltal...	None	20	Right	Scale	Input
6	K1_5	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a ...	{1, Egyáltal...	None	13	Right	Scale	Input
7	K2_1	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy r...	{1, Egyáltal...	None	16	Right	Scale	Input
8	K2_2	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék v...	{1, Egyáltal...	None	14	Right	Scale	Input
9	K2_3	Numeric	8	0	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	{1, Egyáltal...	None	15	Right	Scale	Input
10	K3_1	Numeric	8	0	Értékelje, hogy milyen fontos a következő környezeti szempont: Csökkenteni az éghajlat változást	{1, Nem font...	None	20	Right	Ordinal	Input
11	K3_2	Numeric	8	0	Értékelje, hogy milyen fontos a következő környezeti szempont: Több tevékenység a természetvédel...	{1, Nem font...	None	14	Right	Ordinal	Input
12	K3_3	Numeric	8	0	Értékelje, hogy milyen fontos a következő környezeti szempont: A szemét és a háztartási hulladék c...	{1, Nem font...	None	15	Right	Ordinal	Input
13	K4	Numeric	8	0	A használt vizes palackokat visszavinné-e a vásárlás helyszínére?	{1, Nem}...	None	5	Right	Ordinal	Input
14	K5_1	Numeric	8	0	Figyelembe veszi-e a MÁRKÁT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
15	K5_2	Numeric	8	0	Figyelembe veszi-e a MÉRETET, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
16	K5_3	Numeric	8	0	Figyelembe veszi-e a FORMÁT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
17	K5_4	Numeric	8	0	Figyelembe veszi-e a SÚLYT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
18	K5_5	Numeric	8	0	Figyelembe veszi-e a CSOMAGOLÁS DESIGNT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
19	K5_6	Numeric	8	0	Figyelembe veszi-e a TERMÉK VÉDELMÉT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
20	K5_7	Numeric	8	0	Figyelembe veszi-e az ANYAG MINŐSÉGÉT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
21	K5_8	Numeric	8	0	Figyelembe veszi-e a ZÖLD CSOMAGOLÁST, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
22	K5_9	Numeric	8	0	Figyelembe veszi-e az ÁRAT, amikor palackozott vizet vásárol?	{0, Nem}...	None	5	Right	Nominal	Input
23	K6_1	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: Követem a környezetbarát témákat	{1, Soha}...	None	5	Right	Scale	Input
24	K6_2	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: Kényelmetlenséget is vállaltam azért, hogy...	{1, Soha}...	None	11	Right	Scale	Input
25	K6_3	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: Vásárlás közben a saját bevásárló táskám...	{1, Soha}...	None	5	Right	Scale	Input
26	K6_4	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: A háztartásomban újrahasznosítom a dolg...	{1, Soha}...	None	5	Right	Scale	Input
27	K6_5	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: A vásárlásaim során olyan termékeket vála...	{1, Soha}...	None	5	Right	Scale	Input
28	K6_6	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: A nem környezetbarát viselkedése/imázsa ...	{1, Soha}...	None	5	Right	Scale	Input
29	K6_7	Numeric	8	0	Környezetbarát viselkedés gyakorisága az elmúlt 5 évben: Adományozok a környezetvédelemmel kap...	{1, Soha}...	None	5	Right	Scale	Input
30	K7	Numeric	8	0	Nemek	{0, Nő}...	None	11	Right	Nominal	Input
31	K8	Numeric	8	0	Életkor kategóriák	{1, 25-34}...	None	12	Right	Ordinal	Input
32	K9	Numeric	8	0	Legmagasabb iskolai végzettség	{1, Maximu...	None	18	Right	Ordinal	Input
33	K10	Numeric	8	0	Havi jövedelem (ezer Ft)	{0, Nem vál...	0	20	Right	Ordinal	Input
34	K4_két_kategória	Numeric	8	0	A használt vizes palackokat visszavinné-e a vásárlás helyszínére?	{0, Nem / T...	None	20	Right	Nominal	Input
35	K1_összesen	Numeric	8	0	Környezetbarát attitűd összesen	None	None	14	Right	Scale	Input
36	K6_összesen	Numeric	8	0	Környezetbarát viselkedés összesen	None	None	14	Right	Scale	Input

3.1. ábra: **Az SPSS adatbázis bemutatása**

A továbbiakban bemutatjuk azokat az elemzési lehetőségeket, amelyekkel számos szakmai kérdésekre kapjuk meg a válaszokat.

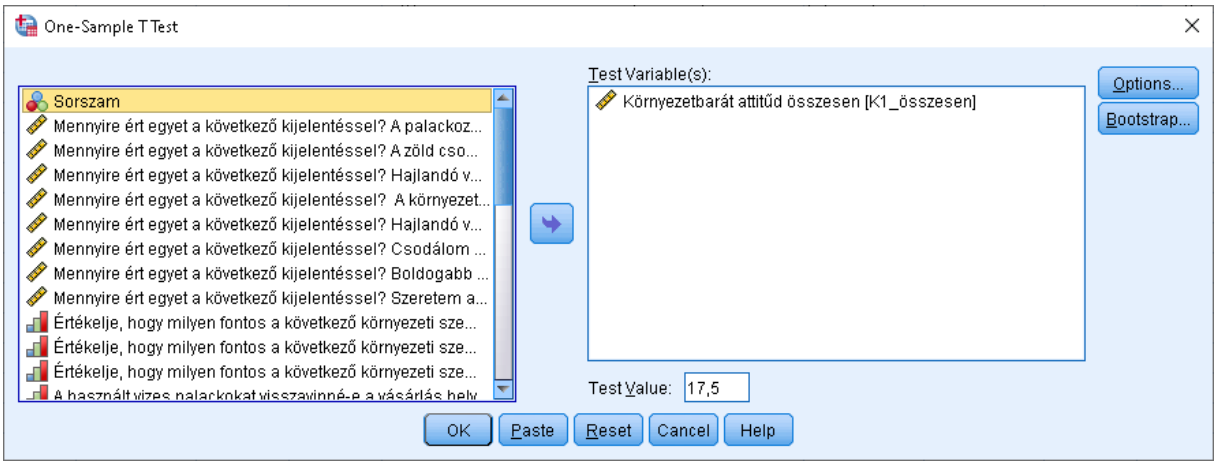
3.2.1 Egymintás t-próba

Vizsgáljuk meg, hogy a válaszadók un. környezet iránti attitűdjei együttesen milyenek és az átlaguk megegyezik-e egy már lefolytatott másik kutatás átlagával (17,5-el)!
Ehhez előzetesen a „Mennyire ért egyet a következő kijelentéssel?” kérdések közül az első ötöt (K1_1 – K1_5) összegeztük. Az így létrehozott új változó neve a „Környezetbarát attitűd összesen” (K1_összesen) volt.
Ha választ akarunk adni a fenti kérdésre, ahhoz az egy mintás t-próbát kell kiszámítanunk. A számítást az SPSS ANALYZE / COMPARE MEANS / ONE–SAMPLE T TEST... menüpontjával tudjuk elvégezni (3.2. ábra).



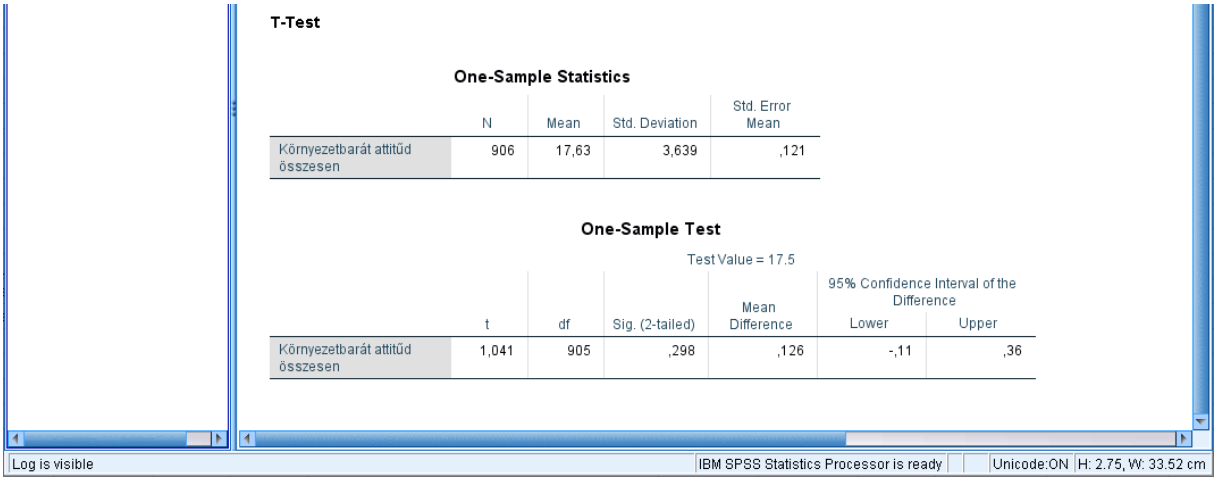
3.2. ábra: **Az egy mintás t-próba indítása**

A párbeszédablakban a bal oldali változólistából vigyük át az általunk vizsgálni kívánt változót (K1_összesen). Ezt követően a „Test Value” cellába írjuk be az általunk már ismert 17,5-es átlagértéket (3.3. ábra). Meg kell említeni, hogy egyszerre több változót is elemezhetünk az SPSS segítségével, ha ugyan azt az átlagértéket szeretnénk mindegyik esetében tesztelni. Ha készen vagyunk, kattintsunk az „OK”-ra. Az eredmények az SPSS Viewer (vagy output-) ablakban jelennek meg.



3.3. ábra: **Az egy mintás t-próba beállításai**

Az outputablakban megjelenő két táblázatot a 3.4. ábra mutatja. Az első táblázat „One-Sample Statistics" a változó leíró statisztikai jellemzőit mutatja. Az első oszlopban a változó un. hosszú neve (label: Környezetbarát attitűd összesen) szerepel. Ezt követi az átlag („Mean"), szórás („Std. Deviation") és az átlag hibája („Std. Error Mean"). Ezekből leolvashatjuk, hogy a válaszadók átlaga 17,63 volt, amit a program a második táblázatban fog összehasonlítani az általunk megadott értékkel (17,5). A második táblázat „One-Sample Test" tartalmazza az egy mintás t-próba eredményeit. A számított t-érték („t") kicsi (1,041), a szabadság fok („df") 905 és a szignifikancia-szint („Sig. (2-tailed)") 0,298, ami nagyobb, mint 0,05, így azt mondhatjuk, hogy a mintánk átlaga (17,63) és az általunk megadott előzetes érték („Test Value = 17.5") közötti különbség nem szignifikáns. *Figyeljünk arra, hogy az SPSS program tizedes vessző helyett tizedes pontot jelez ki!* A két érték közötti különbséget a („Mean Difference") jelzi, de három tizedes pontossággal (0,126 ~ 0,13). Az output tábla két utolsó oszlopa – az általunk megadott szinten (95%) – a különbség konfidenciaintervallum alsó és felső határát („95% Confidence Interval of the Difference") is megmutatja nekünk (-0,11 – 0,36). Ha ez a tartomány tartalmazza a nullát, az azt jelenti, hogy a két átlag közötti különbség lehet nulla, azaz a mintánk átlaga nem különbözik a tesztelt értéktől.



T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Környezetbarát attitűd összesen	906	17,63	3,639	,121

One-Sample Test

	Test Value = 17.5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Környezetbarát attitűd összesen	1,041	905	,298	,126	-,11	,36

Log is visible | IBM SPSS Statistics Processor is ready | Unicode:ON | H: 2.75, W: 33.52 cm

3.4. ábra: **Az egy mintás t-próba eredményei**

3.2.2 Két független mintás t-próba

A kétmintás t-teszttel megvizsgálhatjuk, hogy származhat-e a két független megfigyelés, minta azonos középértékű populációból? Azonosnak tekinthető-e a két populáció középértéke, amelyekből a minták származnak?
A középértékek összehasonlítására szolgáló statisztikai próbák eltérőek attól függően, hogy az alappopulációk szórása egyenlőnek tekinthető-e. Amennyiben a szórások megegyeznek az alábbi próbastatisztikát használjuk, az eloszlás t-eloszlású, DF = $n_1 + n_2 - 2$ szabadságfokkal.
A nevezőben az s_p a két minta összevont varianciájának (pooled variance) négyzetgyökét jelenti, melyet a két minta összevont szórásának nevezzük.

Alkalmazhatósági feltételek:

- Két független minta,
- Normális eloszlású sokaságok,
- A szórások ismeretlenek, de azonosak

Az ismeretlen közös szórást a mintákból számított szórásnégyzetekből becsülhetjük meg.
A két populáció középértéke, amelyekből a minták származnak, abban az esetben tekinthetők azonosnak, ha:

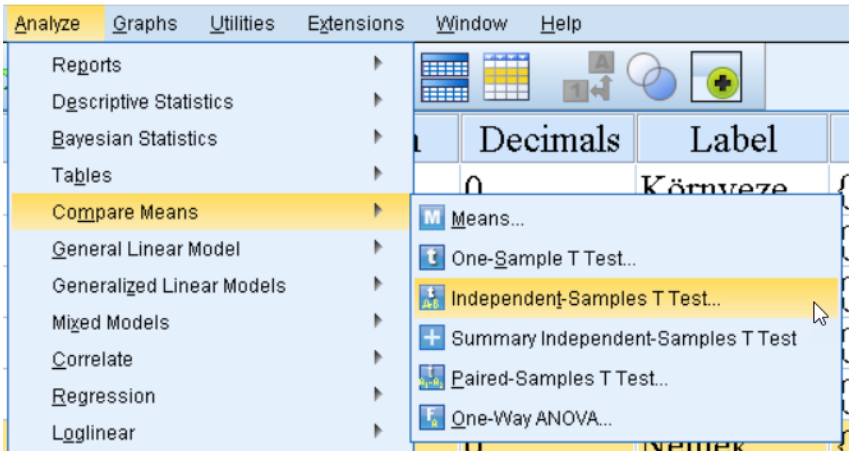
$$|t| \leq t^*$$

A próbastatisztika kritikus t–értékét statisztikai táblázatból kell meghatározni. Ha a két populáció ismeretlen szórásnégyzete korábbi ismeretek, ill. a mintákból számított szórásnégyzetek alapján nem tekinthető azonosnak, akkor a t–próba helyett a Welch-próbát kell alkalmazni, mely igen hasonló a t-próbához, a különbség a szabadságfokok meghatározásában van.

A t-teszt alkalmazásakor előre tudni kell, hogy a két csoport szórása megegyezik-e, tehát tesztelni kell a csoportok szórását (F-próba). Amennyiben a szórások egyenlők, akkor a vizsgálatba vont összes csoportból kell a varianciát becsülni (pooled variance). A próba valószínűségi változója t-eloszlású, így a középértékek különbségének szignifikanciája a kritikus t-érték alapján állapítható meg.
Amennyiben a két csoport szórása szignifikánsan különbözik, ilyenkor a két összehasonlítandó csoport varianciáját súlyozni kell a variancia becsléséhez (separate variancia). A próba valószínűségi változója ebben az esetben nem t-eloszlású, ezért a szabadságfokokat Bonferroni módszerével korrigálni kell, és ezt kell használni a középértékek különbözőségének elbírálásakor (a szabadságfokok korrekciója (Bonferroni teszttel).

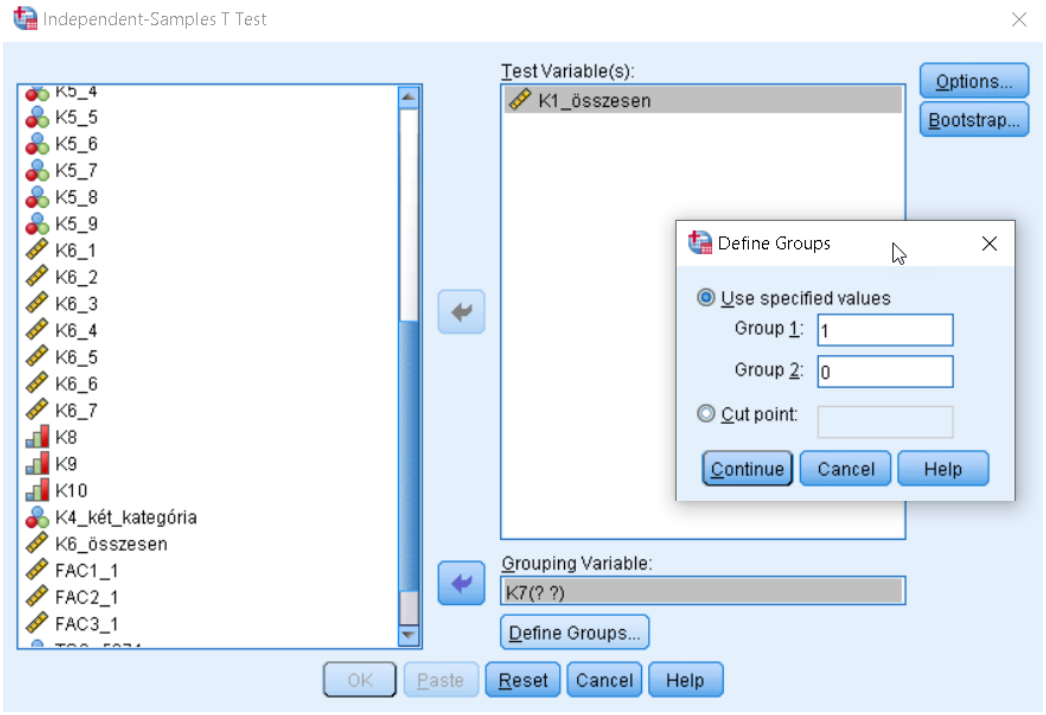
Az előzőekben már vizsgált változó esetében ki tudunk-e mutatni különbséget a férfiak és a nők átlagai között?
Ehhez a „Környezetbarát attitűd összesen” (K1_összesen) és a „Nemek” (K7) változókat vizsgáljuk meg.

Ha választ akarunk adni a fenti kérdésre, ahhoz a két független mintás t-próbát kell kiszámítanunk. A számítást az SPSS ANALYZE / COMPARE MEANS / INDEPENDENT-SAMPLES T TEST... menüpontjával tudjuk elvégezni (3.5. ábra).



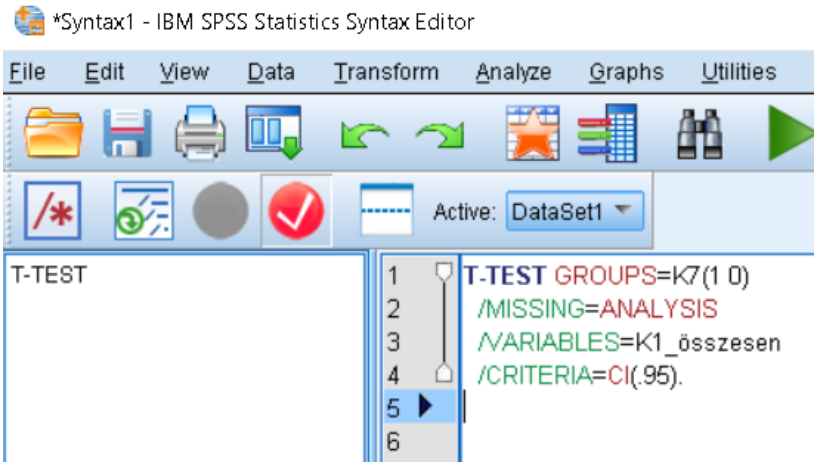
3.5. ábra: **A két független mintás t-próba indítása**

A párbeszédablakban a bal oldali változólistából vigyük át az általunk vizsgálni kívánt változót (K1_összesen) a „Test Variable(s)" ablakba. Ezt követően a „Grouping Variable" részbe mozgassuk át a Nemek (K7) változót. A „Define Groups" gombot megnyomva tudjuk a két csoport számkódjait beállítani. A férfiak esetében az 1-et, a nők esetében a 0-át (3.6. ábra). Lehetőség van arra is, hogy egyszerre több skálatípusú változó átlagát hasonlítsunk össze a nemek szerint. Ebben az esetben a „Test Variable(s)" részbe minden skálatípusú változót mozgassunk bele. Ha készen vagyunk, kattintsunk az „OK"-ra. Az eredmények az SPSS Viewer (vagy output-) ablakban jelennek meg.

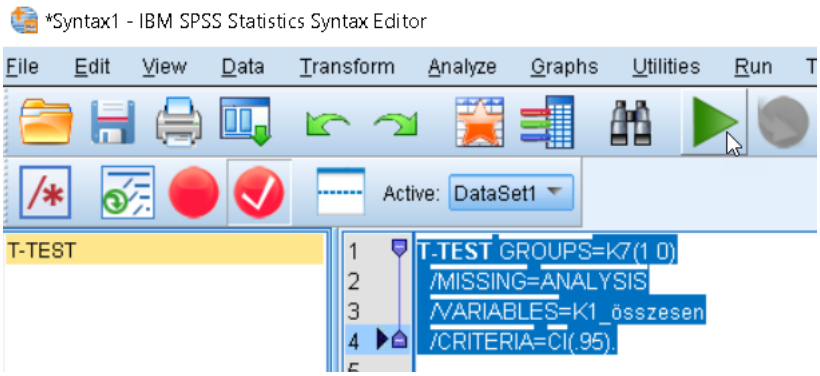


3.6. ábra: **A két független mintás t-próba beállításai**

Az SPSS program esetében is tudunk un. programozást végezni, ha nem a főmenüből választjuk ki a módszereket, hanem „Scripteket” írunk vagy már meglévő kódokat másolunk be. Ehhez meg kell nyitnunk az SPSS programban a „Syntax Editor”-t és ide kell beírnunk a parancskódokat (3.7. ábra). Ezután ki kell jelölnünk a kódokat és a zöld „Run” gomb segítségével le tudjuk futtatni az elemzést (3.8. ábra).



3.7. ábra: **A két független mintás t-próba „Srcipt”-je**



3.8. ábra: **A két független mintás t-próba parancssorainak lefuttatása a „Syntax Editor”-ban**

A két független mintás t-próba paarancskódja az SPSS program output-jában:

```
T-TEST GROUPS=K7(1 0)
  /MISSING=ANALYSIS
  /VARIABLES=K1_összesen
  /CRITERIA=CI(.95).
```

Az outputablakban megjelenő két táblázatot a 3.9. ábra mutatja. Az első táblázat „Group Statistics” a két csoportra vonatkozó leíró statisztikai jellemzőket mutatja. Az első oszlopban a változó un. hosszú neve (label: Környezetbarát attitűd összesen) szerepel. Ezután a csoportok nevei következnek (Férfi és Nő), majd a csoportok elemszámai (N). Ezt követi az átlag („Mean”), szórás („Std. Deviation”) és az átlag hibája („Std. Error Mean”). Ezekből leolvashatjuk, hogy a női válaszadók átlaga 17,95 nagyobb volt, mint a férfiak 17,21-es átlaga.

A program a második „Independent Samples Test” táblázatban statisztikai tesztekkel összehasonlítja, hogy a két csoport szórása azonos- e és van-e különbség a két csoport átlagai között. A „Levene’s Test” részben leolvashatjuk, hogy jelentős különbség van a csoportok szórásai között, mivel a „Sig.” oszlopban szereplő érték kisebb, mint 0,05 (p=0,003). Ezért a táblázat következő részében „t-test for Equality of Means” a második sorban szereplő értékek alapján döntünk a két csoport átlagi közötti különbségről.

A számított t-érték („t”) nagy (-2,958), a szabadság fok („df”) 750,809 és a szignifikancia-szint („Sig. (2-tailed)”) 0,003, ami jóval kisebb, mint 0,05, így azt mondhatjuk, hogy a két csoport (férfiak és nők) átlagai között a különbség „Mean Difference” (-0,737) szignifikáns. Az output tábla két utolsó oszlopa – az általunk megadott szinten (95%) – a különbség konfidenciaintervallum alsó és felső határát („95% Confidence Interval of the Difference”) is megmutatja nekünk (-1,226 – -0,248). Mivel ez a tartomány nem tartalmazza a nullát, az azt jelenti, hogy a két csoport átlaga közötti különbség nem lehet nulla, azaz a mintánk átlagai szignifikánsan különböznek egymástól.

Group Statistics

	Nemek	N	Mean	Std. Deviation	Std. Error Mean
Környezetbarát attitűd összesen	Férfi	393	17,21	4,003	,202
	Nő	513	17,95	3,301	,146

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Környezetbarát attitűd összesen	Equal variances assumed	8,893	,003	-3,034	904	,002	-,737	,243	-1,213	-,260
	Equal variances not assumed			-2,958	750,809	,003	-,737	,249	-1,226	-,248

3.9. ábra: **A két független mintás t-próba eredményei**

3.2.3 Párosított mintás t-próba

Párosított t-próbát akkor használunk, ha a két minta elemei páronként összefüggnek, pl. ugyanazon egyeden két különböző időpontban mérünk egy tulajdonságot, vagy valamilyen csoportképző tulajdonság alapján párokat tudunk képezni.
A két minta középértékének azonossága helyett a párosított minták *d* különbségének (előjeles) várható értékére fogalmazzuk meg a H_0 hipotézist.
Az előző eljárásokhoz hasonlóan itt is z-, ill. t-próbát alkalmazhatunk attól függően, hogy ismert-e a *d* különbségek eloszlása és szórása, illetve mekkora a minta elemszáma?

Alkalmazhatósági feltételek:

- A *d* különbségek eloszlása normális
- *d* ismeretlen (a mintából számított)

A próba t-eloszlású, DF=n-1 szabadságfokú. A képletben a párosított minták különbségének szórása szerepel, amelyet a minta alapján becsülünk.
A t-próba ereje
A statisztikai próba ereje a korábban definiáltak szerint: a valódi d különbség kimutatásának valószínűsége. Ezt 1-β-val jelöltük. Annál erősebb egy statisztikai próba, minél nagyobb valószínűséggel mutatja ki a valódi hatást. A t-próbánál a t-érték valójában egy standardizált hatás (Es, standardised effect), amelynél két csoport átlagának különbségét osztjuk a különbség várható értékének szórásával.

A standardizált hatás nagysága alapján:

- kicsi 0,2
- közepes 0,5
- nagy hatás 0,8 feletti

Amennyiben a két csoport várható értéke megegyezik, a különbségük várható értéke nulla körül mozog. Tehát, ha H_0 igaz, t várható értéke nulla.
A d valódi különbség létezésekor a hatás kimutatásának valószínűsége függ:

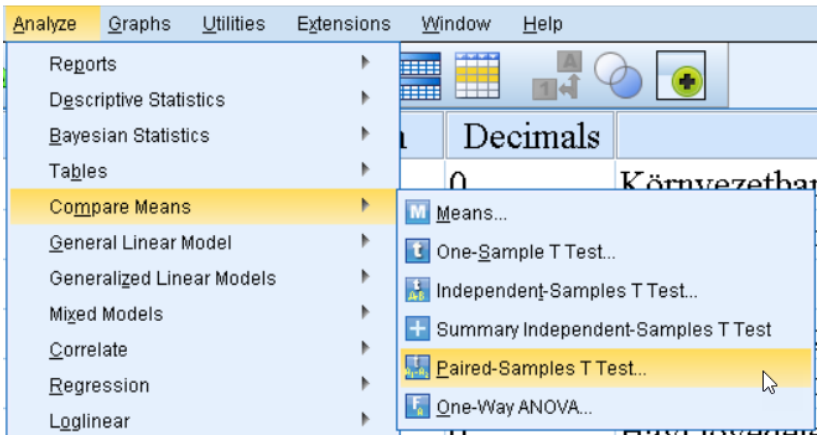
- a minta elemszámától
- a d nagyságától
- a szórástól
- az elsőfajú hiba nagyságától, azaz a szignifikancia-szinttől
- a t-próba típusától (egymintás, kétmintás, párosított)
- alternatív hipotézistől (egyoldali vagy kétoldali)

A fenti tényezők zömét az analízis megkezdése előtt tudjuk beállítani. Ilyen a minta elemszáma, szignifikancia-szint, t-próba típusa, alternatív hipotézis. A szórás a vizsgált jelenség tulajdonsága, ezt csak megbecsülni lehet. A valódi különbség nagyságáról csak előzetes információink lehetnek, pl. korábbi szakirodalmi adatok alapján.

Vizsgálataink során nagyon fontos előre tudni, hogy adott különbséget mekkora valószínűséggel lehet kimutatni. Ekkor tervezzük meg a kísérletet, felmérést, a minimális mintaelemszámot. Ebben a fázisban kell eldönteni, hogy egyáltalán érdemes-e hozzáfogni a vizsgálathoz. Ehhez először a másodfajú hiba nagyságát kell meghatározni. Hogyan? Meg kell határozni a d középértékű t-eloszlásnál egy adott értéknél kisebb értékek előfordulási valószínűségét. Mit jelent az adott érték? A kritikus t-értéket. Ha a kritikus t-értéknél kisebb számított t-értéket kapunk, a nullhipotézist kell elfogadni akkor is, ha a d különbség valóban létezik. Azt mondhatjuk, hogy a nullhipotézis erősebb.

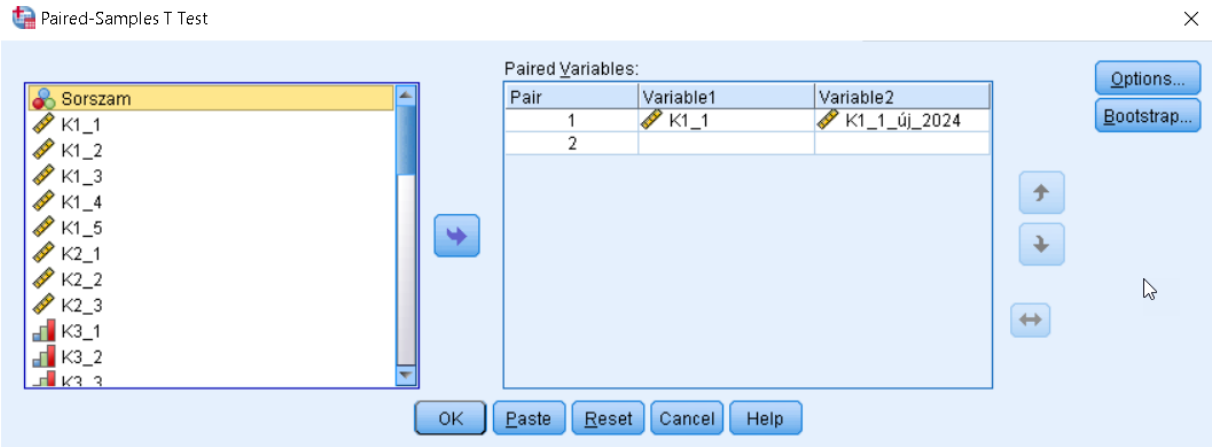
Párosított mintás t-próba gyakorlati példájaként vizsgáljuk meg, hogy változott-e a válaszadók megítélése a „A palackozott ásványvíz vásárlása negatívan hat a környezetre" kérdés esetében, amióta hazánkban bevezették az un. PET-palack visszaváltási rendszert.
Magyarországon 2024. január 1-től működik a visszaváltási rendszer, amely kiterjed a PET-palackokra, üvegekre és fém italos dobozokra. A visszaváltás az erre kijelölt gyűjtőhelyeken, azaz REpontokon történik. A műanyagpalackok visszaváltási díja jelenleg 50 forint Magyarországon. Ez azt jelenti, hogy minden PET-palackért, üvegért és fémdobozért, amelyre a visszaváltási rendszerben szereplő logó van, 50 forintot kaphatunk vissza a megfelelő gyűjtőhelyeken. A visszaváltási díj célja, hogy ösztönözze a fogyasztókat a palackok szelektív gyűjtésére és a hulladék mennyiségének csökkentésére. A rendszer keretében a gyártóknak is kötelezettségeik vannak a termékek forgalomba hozatalával és a hulladékkezeléssel kapcsolatban.

Ehhez előzetesen a „Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre" (K1_1) kérdést és a 2024 után ismételten megkérdezett „Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre" (K1_1_új_2024) kérdést hasonlítjuk össze.
Ha választ akarunk adni a fenti kérdésre, ahhoz a párosított mintás t-próbát kell kiszámítanunk. A számítást az SPSS ANALYZE / COMPARE MEANS / PAIRED–SAMPLE T TEST... menüpontjával tudjuk elvégezni (3.10. ábra).



3.10. ábra: **A párosított mintás t-próba indítása**

A párbeszédablakban (3.11. ábra) a bal oldali változólistából először vigyük át a ”Paired Variables” részbe azt a változót „Variable1”, amelyikhez szeretnénk hasonlítani a változást (K1_1), majd ezt követően válasszuk ki a második változót is azt, amelyik tartalmazza a hatás utáni eredményeket „Variable2”, azaz a PET-palack visszaváltási rendszer bevezetése után begyűjtött válaszokat (K1_1_új_2024). Lehetőség van arra is, hogy egyszerre több skálatípusú változót hasonlítsunk össze. Ebben az esetben a „Paired Variables” részbe minden skálatípusú változót páronként mozgassunk át. Ha készen vagyunk, kattintsunk az „OK”-ra. Az eredmények az SPSS Viewer (vagy output-) ablakban jelennek meg.



3.11. ábra: **A párosított mintás t-próba beállításai**

Az output ablakban megjelenő két táblázatot a következő két ábra mutatja. Az első táblázat „Paired Samples Statistics” a két skála típusú változóra vonatkozó leíró statisztikai jellemzőket mutatja (3.12. ábra). Az első oszlopban a változók un. hosszú nevei vannak feltüntetve. Ezt követi az átlag („Mean”) az elemszámmal (N), majd a szórás („Std. Deviation”) és az átlag hibája („Std. Error Mean”). Jól látható, hogy a 2024-ben bevezetett szabályzás hatására a válaszadók véleménye kevésbé volt negatív a palackozott ásványvizek megítélésével kapcsolatban, mivel csökkent 3,66-ról 3,50-re az átlag érték *(a kérdőívben az 5-ös érték képviselte a negatív hozzáállást, míg az 1-es érték a pozitív hozzáállást)*.

A program a második „Paired Samples Test” táblázatban (3.13. ábra) statisztikai teszttel összehasonlítja, hogy a két időszak átlaga között van-e különbség. A számított t-érték („t”) nagy (11,975), a szabadság fok („df”) 905 és a szignifikancia-szint („Sig. (2-tailed)”) 0,000, ami jelentősen kisebb, mint 0,05. Ezek alapján megállapíthatjuk, hogy a szabályzásnak statisztikailag kimutatható hatása volt arra, hogy a válaszadók, hogyan ítélik meg a palackozott ásványvizek környezetre gyakorolt negatív hatását. Ez a negatív attitűd szignifikánsan csökkent 0,161-el 2024 után. ***Azonban fel kell hívni a figyelmet arra, hogy az átlag még mindig 3 feletti, tehát a válaszadók összességében negatívan állnak hozzá a palackozott ásványvizek környezetre gyakorolt hatásaihoz.***

A különbség konfidenciaintervallum alsó és felső határa („95% Confidence Interval of the Difference”) pozitív (0,135 – 0,188). Mivel ez a tartomány nem tartalmazza a nullát, az azt jelenti, hogy a két időszak átlaga közötti különbség nem lehet nulla, azaz a válaszok átlaga szignifikánsan csökkent 2024 után.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	3,66	906	1,028	,034
	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	3,50	906	1,189	,039

3.12. ábra: **A párosított mintás t-próba eredményei (leíró statisztikák)**

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre - Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	,161	,405	,013	,135	,188	11,975	905	,000

3.13. ábra: **A párosított mintás t-próba eredményei (teszt statisztikák)**

3.2.4 Egy és több szempontos varianciaelemzések

A variancia-analízis a t-próba kiterjesztése kettőnél több minta esetére. Tehát három vagy több mintával rendelkezünk. Mindegyik minta egy csoportképző ismérv egy-egy szintjét reprezentálja. Pl. különböző kefir márkák. A márkákon belül egy skála típusú változó várható értékét vizsgálhatjuk. Megvizsgálhatjuk, hogy a különböző kefirek várható eladási árai, zsírtartalmai, fehérjetartalmai, stb. megegyeznek-e. Ez azt jelenti, hogy a skála típusú változó nem függ a nominális, csoportképző változótól. A függővváltozó a variancia-analízis modellben mindig valamilyen skála típusú, a független változó(k) nominális mérési szintűek. Amennyiben az árak, stb. nem egyeznek meg a különböző kefireknél, akkor összefüggés van közöttük, és a márkákkal részben magyarázhatjuk a különbségeket. A magyarázat a függő változó teljes heterogenitásának két részre bontását jelenti. A teljes heterogenitás egyik része az, amelynek „okai" a független változók, a másik heterogenitás-rész pedig az, amelynek „okait" az egyéb, általunk nem vizsgált tényezők tartalmazzák. Ez utóbbit sokszor a véletlen hatásaként is emlegetik.

A heterogenitás mérésére korábban többféle mérőszámot ismertettünk, ismétlésként a legfontosabbak az alábbiak:
(1) terjedelem (range); a legnagyobb és legkisebb érték közötti távolság;
(2) átlagos eltérés;
(3) szórás;
(4) variancia- vagy szórásnégyzet.

Alapfogalmak
Nézzük át azokat az alapfogalmakat, amelyeket a variancia-analízis során használunk.
a) Faktor: Faktornak nevezzük a vizsgálatba bevont független változókat, pl. különböző kezeléseket, tényezőket, ilyen a kefir márka. Kísérletekben inkább kezeléseknek hívjuk.
b) Faktor szint: A faktor értékkészletének az eleme, mely beállítása mellett vizsgálhatjuk meg a függő változónkat. A kezelések szintjei, pl. kefir márkán belül Danone, Milli, Müller, stb. Kísérletben pl. műtrágyaadagok.
c) Kvalitatív és kvantitatív faktorok: Ha a faktorszintek nem numerikusak vagy intervallum skálájúak, akkor kvalitatív, ellenkező esetben kvantitatív faktorokról beszélünk.
d) Cellák: Egyfaktoros modellekben a cellák megfelelnek a faktorok szintjeinek, többfaktoros esetben a figyelembe vett faktorok szintjeiből előálló kombinációk a cellák. Pl. amikor a 2 faktor műtrágyaadagok és öntözési módok, akkor a cellák a (műtrágyaadagok, öntözési módok) összes lehetséges kombinációjából állnak.
e) Interakció: Két független változó kapcsolatában akkor áll fenn interakció (kölcsönhatás), ha változó hatása függ az változó szintjétől és fordítva.
f) Egyszempontos variancia-analízis: Variancia-analízis, ahol csak egy faktor van. Egyutasnak is nevezik
g) Többszempontos variancia-analízis: Variancia-analízis, ahol kettő vagy több faktor van.
h) Egyváltozós variancia-analízis: amelyben csak egy függő változót vizsgálunk.
i) Többváltozós variancia-analízis: amelyben kettő vagy több függő változót elemzünk.

A lineáris modell

Alkossuk meg az egytényezős variancia-analízis matematikai modelljét. Egy kísérletben k számú kezeléssel vagy populációval és r számú ismétléssel rendelkezünk. Az adataink száma tehát n=k*r. Minden mért adat y_{ij} felbontható három összetevőre, amelyek: a kísérlet főátlaga (m), a kezeléshatás (A_i), és a meg nem magyarázott rész, a maradék (e_{ij}). A maradéktagokat hibának is nevezik (error).

Az egytényezős lineáris modell:

Valójában a kezeléshatás az m és A_i összege. Ez a kettő adja a lineáris modellel becsült értéket, azaz a modellezett értéket. Az A_i a kísérlet főátlagtól vett eltérést jelenti (kezeléshatás-főátlag). A korábban tanultak szerint az alapadatok átlagtól vett eltéréseinek összege nulla. Ez a lineáris modellre is igaz, a kezeléshatások összege nulla, vagyis a kezelések szimmetrikusak a főátlagra. Az e_{ij} maradéktagok tulajdonságai nagyon fontosak, amelyek egyben megegyeznek a variancia-analízis alkalmazhatósági feltételeivel, melyeket a következőkben ismertetünk.

A variancia-analízis alkalmazásának feltételei

Az alkalmazhatósági feltételek a maradéktagokra vonatkoznak:

(a) Az egyes kezelésekhez tartozó maradékoknak függetleneknek kell lenniük a blokk, a kezeléshatástól és a függő változótól. Ezt leginkább a kísérleti elrendezéssel, randomizálással biztosíthatjuk. A függetlenség azt jelenti, hogy a maradékok nagyságát nem befolyásolhatja a kezelés. Amennyiben hatással van rá, akkor ez keveredhet a kezeléshatással, és torz becslést kapunk, helytelen becsült értékeket fogunk előállítani.

(b) A maradék normális eloszlású, nulla várható értékű valószínűségi változó. Attól, hogy egy normál-eloszlású mintához egy konstans értéket hozzáadunk, vagy abból levonunk, az eloszlás és a minta szórása nem változik. A normalitást korábban ismertetett módszerek valamelyikével ellenőrizhetjük. (Megjegyezzük, hogy a matematikai-statisztikai kézikönyvek az ANOVA-t robusztus eljárásnak tekintik, s azt állítják, hogy a függő változónak nem kell normális eloszlásúnak lennie). Ha matematikailag korrekt módon akarjuk az ANOVA-t használni, akkor a függő változót normális eloszlásúvá transzformálhatjuk. Azért kell normáliseloszlásúnak lennie, mert a hatások megítélésekor a normál-eloszlás tulajdonságait használjuk fel, az eloszlás nevezetes értékeit.

(c) A maradékok szórásnégyzetei a kezeléskombinációkon belül azonosak, azaz homoszkedasztikus a modell. (Az SPSS programban ezt a homogenitást a Levene teszt alapján tesztelhetjük.)

A variancia-analízis alkalmazásának lépései

1. A variancia-analízis modell felállítása
2. Szignifikancia-szint megválasztása
3. A variancia-analízis kiszámítása, az F-próba
4. A modell érvényességének ellenőrzése
5. Amennyiben az F-próba szignifikáns, középértékek többszörös összehasonlítása

A középértékre vonatkozó hipotézisek a következők:

- azoknak a populációknak a középértékei, amelyekből a minták származnak azonosak.

- létezik legalább egy olyan középérték pár, ahol a középértékek nem tekinthetők azonosnak, legalább egyszer.

Az analízis megkezdése előtt ábrázolni kell az alapadatokat. Olyan ábrát érdemes készíteni, amelyben a várhatóérték mellett a középérték hibáját (se) is ábrázoljuk. Erre azért van szükség, mert ha csak az átlagokat tüntetjük fel az y-tengely léptékétől függően nagyon kicsi különbségeket is fel lehet nagyítani, és a jelentős különbségeket is el lehet tüntetni. A standardizált hatások, amit az angol szakirodalomban „standard effect" néven emlegetnek, nem más mint a kezeléshatás osztva a szórással, ingadozással. Ez azt mutatja, hogy a kezeléshatás, hogyan aránylik a szóráshoz, azaz a véletlen ingadozáshoz.
Az így meghatározott standard hatások nagyságát a korábban ismertetett módon ítélhetjük meg.

1. A variancia-analízis modell felállítása
A módszer alapgondolata szerint a modellben a mérési, megfigyelési értékeket összegként képzeljük el. Az n megfigyelés mindegyikére a korábban ismertetett modellegyenlet írható fel, amelynek alapján a mintaelemeken mért, ill. megfigyelt yij értékek felbonthatók a modell által meghatározott részekre és a hibára. A modell által meghatározott rész a szisztematikus hatásokat tartalmazza, a hibakomponens pedig a véletlen hatást jelenti.
A variancia-analízis legegyszerűbb modelljében a vizsgálatban szereplő k számú populációból egyszerűen r elemű véletlen mintát veszünk, majd a mintánkénti középértékeket hasonlítjuk össze, ezt nevezzük egyszempontos variancia-analízisnek (kísérlet esetén teljesen véletlen elrendezésnek). Az elrendezés modellegyenlete, ahol X_{ij} az i-edik minta j-edik eleme; a kísérlet vagy minta főátlaga; A_i az i-edik mintához tartozó populáció hatása (növelheti vagy csökkentheti a főátlagot); e_{ij} véletlen hatás. Ebben a modellben a modell által meghatározott rész, csak az i-edik mintához tartozó populáció várható értékét tartalmazza, tehát szisztematikus különbséget csak a populációk várható értékei között tételezhetünk fel. A véletlen okozta hatásokat a hibakomponens tartalmazza. Amennyiben teljesülnek a variancia-analízis alkalmazásának feltételei, akkor A_i összege nulla, és e_{ij} normális eloszlású nulla várhatóértékű sokaság, és független a blokk és kezeléshatástól, valamint a modell homoszkedasztikus.

2. Szignifikancia-szint megválasztása
A szignifikancia-szint nagyságát leggyakrabban 5%-nak választják. Ez az érték szerepel legtöbb statisztikai programban is kezdeti értékként. Amennyiben túl szigorúnak ítéljük ezt, választhatunk 10%-os szintet is. Ebben az esetben a kezelés okozta valódi hatások kimutatásának nagyobb a valószínűsége. Természetesen az elsőfajú hiba ilyenkor 5-ről 10%-ra nő. A szignifikancia-szintet választhatjuk 1 vagy 0,1%-nak is. Ezek már nagyon szigorú feltételek, alig követünk el elsőfajú hibát, de annál nagyobb a valószínűsége a másodfajú hibának. Elméletileg bármilyen szignifikancia-szintet választhatunk, ha szakmailag meg tudjuk indokolni. Amennyiben eldöntöttük az elsőfajú hiba nagyságát, meg tudjuk határozni a kritikus F-értéket. A kritikus F-érték az a legnagyobb érték, amelyet a véletlen ingadozás mellett kaphatunk. Ennél kisebb érték esetén a H_0-t kell elfogadni.

3. A variancia-analízis kiszámítása, az F-próba
Az SPSS program eredménytáblázatában az alábbi fogalmakkal találkozunk:
Tényezők: a variancia okai
Eltérés-négyzetösszegek (SS)
Csoportok között: kezelésátlagok eltérés-négyzetösszegei * r.
Csoporton belüli: kezeléseken belül az eltérés-négyzetösszegek összege
Összes: alapadatok eltérés-négyzetösszegei

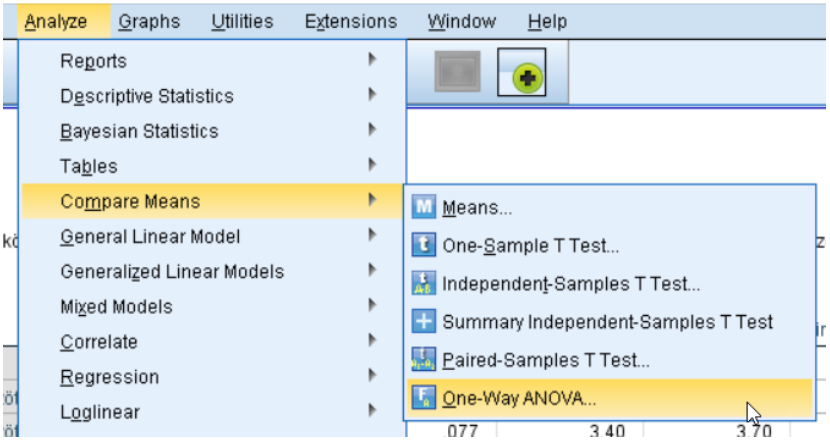
Szabadságfokok:
Csoportok között: k-1
Csoporton belül: n-k
Összesen: n-1

Varianciák: eltérés-négyzetösszegek osztva a szabadságfokokkal.

F-próba

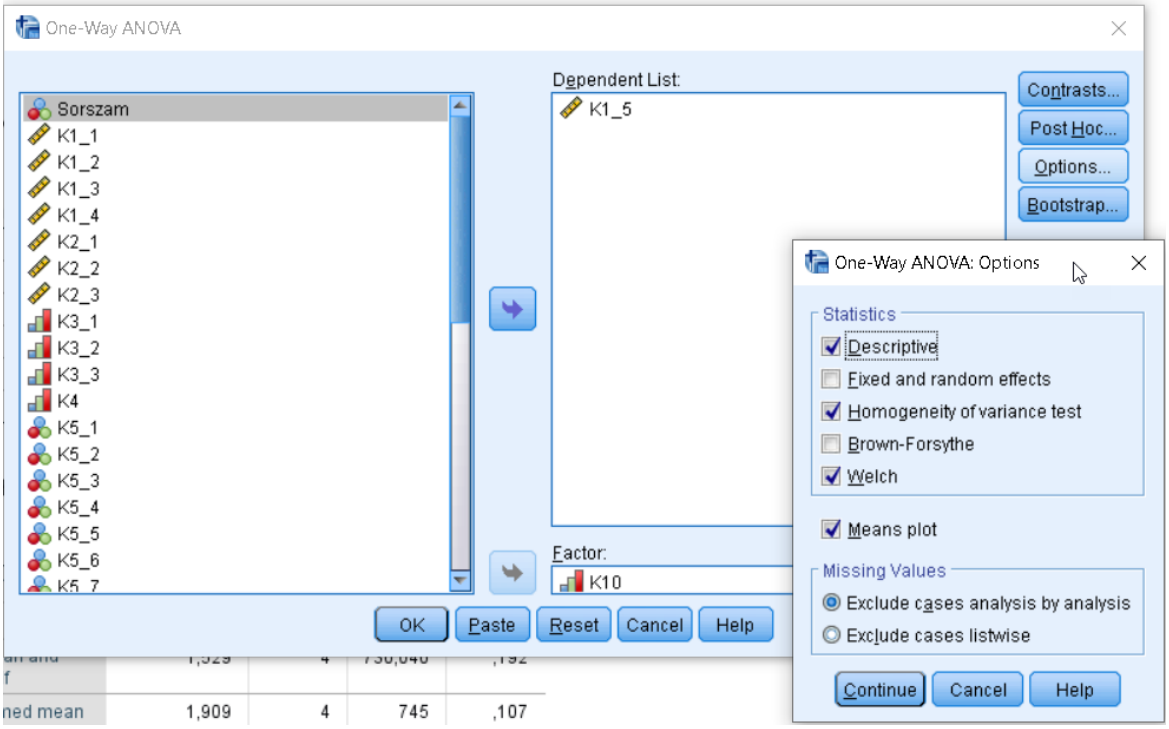
Az F-eloszlás sűrűségfüggvénye: Az x-tengelyen az F-értékek, az y-tengelyen a valószínűségek láthatók. Általában egy függőleges vonal mutatja a kritikus F-értéket. Ezt a szignifikancia-szint és a két szabadságfok ismeretében tudjuk meghatározni. Korábban említettük, ha ennél kisebb a számított F, akkor a nullhipotézist kell elfogadni. Ha a számított F-érték nagyobb, mint a kritikus, akkor már nem tekinthető a véletlen ingadozás hatásának, a nullhipotézist vissza kell utasítani.
Mikor szignifikáns az F-próba? Amennyiben szakmailag teljesen korrektek akarunk lenni, akkor azt kell válaszolni, ha létezik legalább egy szignifikáns kontraszt a csoportok között. A kontraszt egy lineáris összehasonlító függvény. A függvény együtthatóinak összege nulla.

Példa:
Ki tudunk-e mutatni különbséget az eltérő jövedelmi helyzetben levők és a környezetvédelem miatti adófizetési hajlandóság között?
Ehhez a „Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét” (K1_5) és a „Havi jövedelem (ezer Ft)” (K10) változókat vizsgáljuk meg.
Ha választ akarunk adni a fenti kérdésre, ahhoz a két független mintás t-próbát kell kiszámítanunk. A számítást az SPSS ANALYZE / COMPARE MEANS / ONE-WAY ANOVA... menüpontjával tudjuk elvégezni (3.14. ábra).

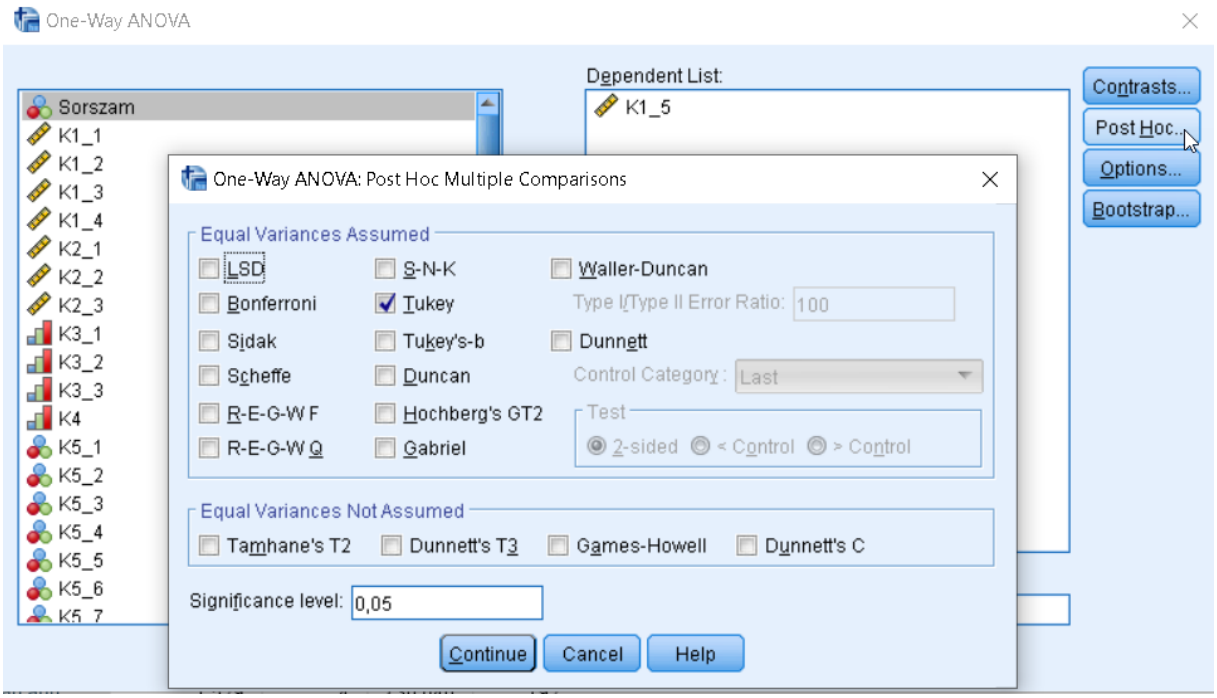


3.14. ábra: **Az egy szempontos varianciaelemzés indítása**

A párbeszédablakban a bal oldali változólistából vigyük át az általunk vizsgálni kívánt változót (K1_5) a „Dependent List” ablakba. Ezt követően a „Factor” részbe mozgassuk át a Jövedelem (K10) változót. (3.15. ábra). Ennél a tesztnél is lehetőség van arra, hogy egyszerre több skálatípusú változó átlagát hasonlítsunk össze a jövedelmek szerint. Ebben az esetben a „Dependent List” részbe minden skálatípusú változót mozgassunk bele. A következő lépésben az „Options…” gombra kattintva lehetőségünk lesz különböző „statisztikákat” lekérni, ezért végezzük el a 3.15. ábra szerinti beállításokat. A 3.16. ábrán láthatjuk, hogy a „Post Hoc” gomb segítségével lehetőségünk van többszörös összehasonlítási tesztek lefuttatására is. ***A gyakorlatban ezt a lépést csak azután végezzük el, ha a varianciák összehasonlításának tesztjét lefuttattuk és tudjuk, hogy a csoportvarianciák között jelentős-e az eltérés.***
Ha készen vagyunk, kattintsunk az „OK”-ra. Az eredmények az SPSS Viewer (vagy output-) ablakban jelennek meg.



3.15. ábra **Az egy szempontos varianciaelemzés beállításai**



3.16. ábra **Az egy szempontos varianciaelemzés „Post Hoc Test" beállítása**

Az output ablakban megjelenik 6 táblázat és egy ábra. A 3.17. ábrán a leíró statisztikák eredménye jelenik meg a jövedelem csoportok szerint. A legnagyobb átlag (3,92) a legtöbb jövedelemmel rendelkezőknél figyelhető meg, míg a legszegényebbek válaszainak átlaga (3,10) volt a legkisebb. A 95%-os konfidencia intervallumok alapján feltételezhető, hogy van különbség az egyes jövedelmi csoportok átlagai között (pl. 2,94-3,25 és 3,55-4,30).

Descriptives

Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
150 ezer Ft alatt	168	3,10	,998	,077	2,94	3,25	1	5
150 - 250 ezer Ft között	270	3,38	1,073	,065	3,25	3,51	1	5
250 - 350 ezer Ft között	162	3,55	,985	,077	3,40	3,70	1	5
350 - 450 ezer Ft között	111	3,69	1,007	,096	3,50	3,88	1	5
450 ezer Ft fölött	39	3,92	1,156	,185	3,55	4,30	1	5
Total	750	3,43	1,055	,039	3,35	3,50	1	5

3.17. ábra: **Az egy szempontos varianciaelemzés eredményei (leíró statisztikák)**

A varianciák homogenitásának Levene tesztje esetében a „Based on Mean" sorban a „Sig." érték 0,05-nél nagyobb, ami azt jelenti, hogy a jövedelem csoportok közötti varianciák/szórások nem különböznek jelentősen egymástól (3.18. ábra).

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	Based on Mean	1,615	4	745	,169
	Based on Median	1,529	4	745	,192
	Based on Median and with adjusted df	1,529	4	730,040	,192
	Based on trimmed mean	1,909	4	745	,107

3.18. ábra: **Az egy szempontos varianciaelemzés Levene teszt eredményei**

A számított F-érték („F”) nagy (9,742) és a szignifikancia-szint („Sig.”) 0,000, ami jóval kisebb, mint 0,05, így azt mondhatjuk, hogy a jövedelem csoportok között az adófizetéssel történő egyetértések alapján számolt átlagok között van különbség (3.19. ábra). Arra viszont nem tudunk választ adni, hogy pontosan mely csoportok átlagai különböznek egymástól jelentősen.

ANOVA

Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	38,969	4	9,742	9,134	,000
Within Groups	794,643	745	1,067		
Total	833,612	749			

3.19. ábra: **Az egy szempontos varianciaelemzés ANOVA táblázata**

Meg kell jegyezni, hogy abban az esetben, ha a varianciák homogenitásának Levene tesztje (3.18. ábra) szignifikáns lenne, azaz a „Sig.” érték 0,05-nél kisebb lett volna, akkor az un. Welch próba eredményeit kellett volna figyelembe venni (3.20. ábra).

Robust Tests of Equality of Means

Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét

	Statistic[a]	df1	df2	Sig.
Welch	9,035	4	202,715	,000

a. Asymptotically F distributed.

3.20. ábra: **Az egy szempontos varianciaelemzés Welch próba táblázata**

Mivel az ANOVA táblázatban szignifikáns eredményt kaptunk, ezért érdemes elvégezni a páronkénti többszörös összehasonlítások tesztjeit. A példánkban a csoportok varianciái megegyeztek, ezért a „Post Hoc Tests” menüben (3.16. ábra) az „Equal Variances Assumed” részből választunk ki egy tesztet. Mi most a „Tukey” tesztet választottuk ki, de lehetett volna más teszteket is kiválasztani innen. A 3.21. ábrán megfigyelhető, hogy az eltérő jövedelemmel rendelkező csoportokban a válaszok átlagai jelentősen eltérnek egymástól (ott ahol a „Sig.” oszlopban 0,05 vagy kisebb az érték). Ezt jelzi az is, ha csillag van a „Mean Difference” oszlopban feltüntetett különbségek mellett (pl. a legszegényebb csoport átlaga minden más

csoporttól jelentősen különbözött). Ugyanakkor a legnagyobb jövedelemmel rendelkezők átlaga (450 ezer Ft fölött) az előző két kategória átlagától nem különbözött („Sig.” oszlop értéke 0,253 és 0,755, ami jóval magasabb volt, mint a 0,05).

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem k

Tukey HSD

(I) Havi jövedelem (ezer Ft)	(J) Havi jövedelem (ezer Ft)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
150 ezer Ft alatt	150 - 250 ezer Ft között	-,286*	,101	,039	-,56	-,01
	250 - 350 ezer Ft között	-,454*	,114	,001	-,77	-,14
	350 - 450 ezer Ft között	-,598*	,126	,000	-,94	-,25
	450 ezer Ft fölött	-,828*	,184	,000	-1,33	-,33
150 - 250 ezer Ft között	150 ezer Ft alatt	,286*	,101	,039	,01	,56
	250 - 350 ezer Ft között	-,168	,103	,475	-,45	,11
	350 - 450 ezer Ft között	-,312	,116	,058	-,63	,01
	450 ezer Ft fölött	-,542*	,177	,019	-1,03	-,06
250 - 350 ezer Ft között	150 ezer Ft alatt	,454*	,114	,001	,14	,77
	150 - 250 ezer Ft között	,168	,103	,475	-,11	,45
	350 - 450 ezer Ft között	-,144	,127	,788	-,49	,20
	450 ezer Ft fölött	-,374	,184	,253	-,88	,13
350 - 450 ezer Ft között	150 ezer Ft alatt	,598*	,126	,000	,25	,94
	150 - 250 ezer Ft között	,312	,116	,058	-,01	,63
	250 - 350 ezer Ft között	,144	,127	,788	-,20	,49
	450 ezer Ft fölött	-,229	,192	,755	-,76	,30
450 ezer Ft fölött	150 ezer Ft alatt	,828*	,184	,000	,33	1,33
	150 - 250 ezer Ft között	,542*	,177	,019	,06	1,03
	250 - 350 ezer Ft között	,374	,184	,253	-,13	,88
	350 - 450 ezer Ft között	,229	,192	,755	-,30	,76

*. The mean difference is significant at the 0.05 level.

3.21. ábra: **Az egy szempontos varianciaelemzés többszörös összehasonlítási tesztjeinek (Post Hoc Test) táblázata**

A post-hoc tesztek közül azért választottuk a Tukey tesztet, mert ennél lehetőségünk van sorba rendezni a csoportokat az átlaguk szerint és a szignifikánsan eltérő átlagok különböző oszlopokba kerülnek (3.22. ábra).

Homogeneous Subsets

Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét

Tukey HSD[a,b]

		Subset for alpha = 0.05		
Havi jövedelem (ezer Ft)	N	1	2	3
150 ezer Ft alatt	168	3,10		
150 - 250 ezer Ft között	270	3,38	3,38	
250 - 350 ezer Ft között	162		3,55	3,55
350 - 450 ezer Ft között	111		3,69	3,69
450 ezer Ft fölött	39			3,92
Sig.		,292	,209	,082

Means for groups in homogeneous subsets are displayed.

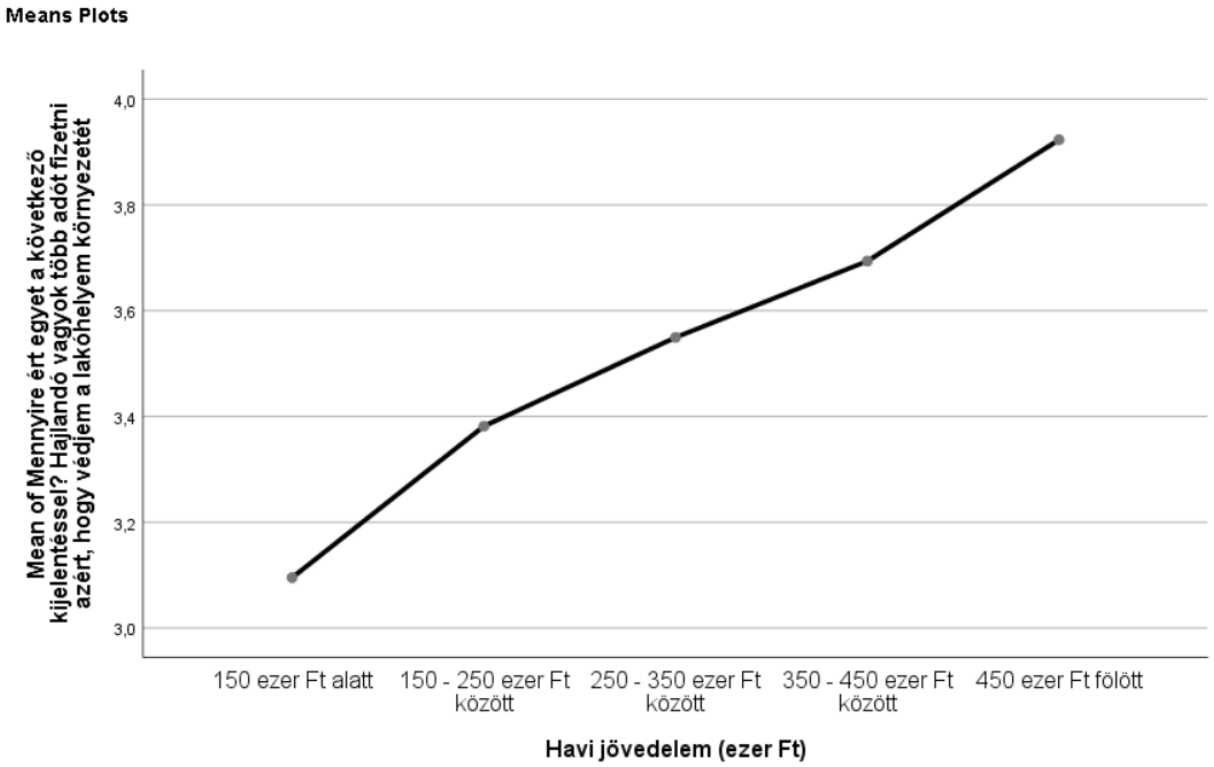
a. Uses Harmonic Mean Sample Size = 99,051.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

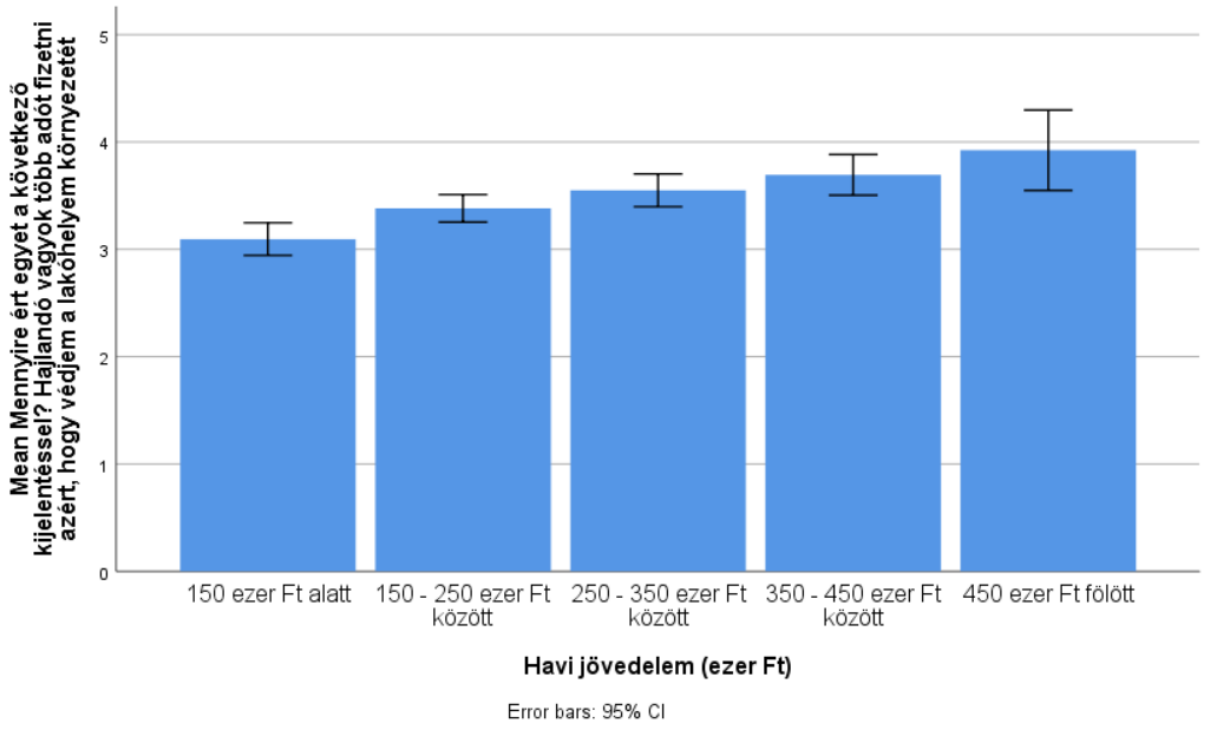
3.22. ábra: **A Tukey HSD többszörös összehasonlítási teszt eredményeinek összefoglaló táblázata**

Abban az esetben, ha egy szám több oszlopban is szerepel, az az érték nem különbözik az adott oszlopban feltüntetett többi értéktől (pl. 3,38 az első és a második oszlopban is benne van), tehát a 150-250 ezer Ft-os jövedelemmel rendelkezők véleménye nem különbözik statisztikailag a legszegényebb (150 ezer Ft alattiak) és a következő két jövedelmi csoport (250-350 ezer Ft és a 350-450 ezer Ft-ba sorolt válaszadók) véleményétől.

Az elemzés során a program előállít egy speciális ábrát (vonaldiagram), amely az eltérő jövedelmekhez tartozó átlagokat mutatja (3.23. ábra). Ebből az ábrából nem minden esetben lehet egyértelműen következtetni arra, hogy melyik csoport átlaga különbözik a másik csoport átlagától. Ezért érdemes a 3.24. ábrán bemutatott oszlopdiagramot elkészíteni, mivel erről könnyebben leolvasható, hogy melyik konfidenciaintervallum fedi át a másik intervallumot. Ahol nincs átfedés, ott valószínűsíthető az átlagok közötti különbség.



3.23. ábra: **Az egy szempontos varianciaelemzés eredményének ábrája**



3.24. ábra: **Az egy szempontos varianciaelemzés alapadatainak oszlopdiagramja**

4. A koncentráció mérése

Az erőforrások (pénz, természeti illetve humán erőforrások, stb.) tömörülésének, koncentrációjának ismerete és jellemzése fontos a közgazdaság területén. A történelem során először a termelésre értelmezték a koncentráció fogalmát, majd később más területekre is. A XVIII. század második felétől kezdik el tömegesen használni. Innentől vált általános közgazdasági fogalommá, amely napjainkban a gazdaságban lévő tömörülések, összpontosulások fogalmát takarja.

A statisztikában is megjelenik a koncentráció fogalma, Itt mennyiségi ismérv szerint vizsgálva egy jelenséget, arra vagyunk kíváncsiak, hogy mennyire összpontosul az értékösszeg a sokaság egységeire. Koncentrációról akkor beszélünk, ha az értékösszeg jelentő része a sokaság kevés vagy relatíve kevés egységére összpontosul.

A társadalom és gazdaság területén az erőforrások bizonyos fokú koncentrációja szükséges a normális gazdasági élet fenntartásához. Ha mindenki egyenlő arányban részesülne az erőforrásokból, akkor azok elaprózódnának, és nem hasznosulnának. A nagy társadalmi programok megvalósításához szükség van a koncentrációra. Enélkül nem lehetne hidakat, autópályákat, stadionokat, stb. építeni. Az űrprogram megvalósítása lehetetlen lenne a tőke nagyfokú koncentrálódása nélkül.

A nagyon-nagyon fokú koncentráció viszont káros is tud lenni, gondoljunk csak a monopóliumokra, ahol a tisztességes szabadpiaci verseny nem tud megvalósulni. A világon mindenütt felügyelik ezt a jelenséget, és különböző mutatószámokkal mérik a káros mértékű fúziókat, és ezzel próbálják megakadályozni a „gigacégek" létrejöttét és a tisztességes szabadpiaci verseny kereteinek sérülését.

Megkülönböztetünk abszolút és relatív koncentrációt. Az **abszolút koncentráció** esetén az értékösszeg kevés egység között oszlik el. Ilyen a monopólium, duopólium és az oligopólium. **Relatív koncentráció** esetén az értékösszeg egyenetlenül oszlik el a sokaság egységeire, jelentős része a sokaság kis hányadához tartozik. A kétféle megjelenési forma között nincs merev elhatárolódás.

4.1. A koncentráció jellemzése

Számos koncentrációt mérő mutatószám létezik, mi ezek közül az alábbiakat fogjuk tárgyalni:

- Koncentrációs táblázat
- Kvantilis (kvartilis, decilis, percentilis)
- Lorenz-görbe
- Gini-koncentrációs együttható
- Herfindahl-Hirschman index

4.2. Koncentrációs táblázat

A koncentrációs táblázat bemutatásához a magyarországi pénzintézetek eszközállományát fogjuk felhasználni. Ezek az adatok megtalálhatók a Magyar Nemzeti Bank éves jelentéseiben, az úgynevezett Aranykönyvben. (https://statisztika.mnb.hu/publikacios-temak/felugyeleti-statisztikak/aranykonyv/aranykonyv)

4.1. táblázat: **A magyarországi első tíz legnagyobb bank eszközállománya (MFt)**

Bankok	Eszközök (mFt)	Bankok rel.gyak.	Eszközök rel.gyak.	Bankok kum.gyak.	Eszközök kum.gyak.
1	6 213 397	10%	25%	10%	25%
2	3 213 379	10%	13%	20%	37%
3	2 948 517	10%	12%	30%	49%
4	2 749 837	10%	11%	40%	60%
5	2 482 860	10%	10%	50%	69%
6	2 400 580	10%	9%	60%	79%
7	1 675 031	10%	7%	70%	86%
8	1 566 193	10%	6%	80%	92%
9	1 189 217	10%	5%	90%	96%
10	900 025	10%	4%	100%	100%
Összesen:	25 339 036	100%	100%		

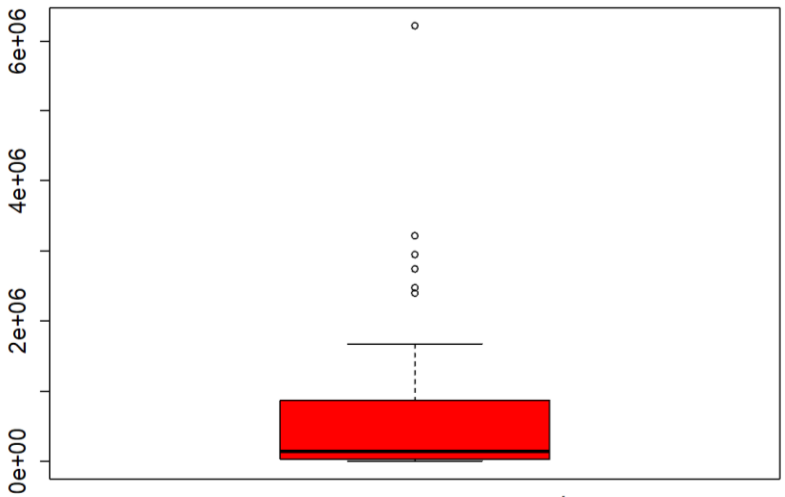
Forrás: PSZÁF, 2010. Aranykönyv

A 4.1. táblázat első oszlopa a sorszámot, a második az eszközállomány forintban kifejezett értékét mutatja. A koncentrációs táblázathoz a bankok és eszközök relatív gyakoriságát kell meghatározni. Az eszközök relatív gyakorisága valójában a megoszlási viszonyszámnak felel meg. Ezt úgy kapjuk meg, hogy az összes eszközállománnyal osztjuk az egyes bankok eszközállományát és százalékos formátumban fejezzük ki. Az utolsó két oszlop a kumulatív értékeket mutatja. A koncentráció teljes hiánya esetén a két oszlop tökéletesen megegyezne. Minél nagyobb az eltérés a két oszlop értékei között, annál nagyon fokú az koncentráció. A 2010. évi adatok alapján az első pénzintézet az eszközök 25%-t, azaz a negyedét uralja. A táblázat készítésekor figyeljünk oda, hogy az utolsó cella értéke mindig 100%-t adjon.

Szerencsésebb, ha a táblázatok helyett ábrákat készítünk az adatokból.

4.3. Kvartilis ábra

A kvartilis ábra (4.1. ábra) szemléletesen használható az adatok eloszlásának bemutatására. Ezt az ábrát több névvel is szokták említeni: doboz ábra, box plot. Ez egy eloszlásfüggetlen ábra. Nincs semmilyen feltétel az eloszlással szemben, hiszen az ábrán helyzeti jellemzőket ábrázolunk. A kicsitől a nagy értékek felé rendezett adatsort négy egyenlő részre bonjuk. Minden rész az adatok 22-25%-t fogja tartalmazni.



Forrás: PSZÁF, 2010. Aranykönyv

4.1. ábra **Box-plot ábra**

Az ábra legalsó vonala a minimális, a második az alsó kvartilis (Q_1), a harmadik a középső kvartilis (Q_2), a negyedik a felső kvartilis (Q_3) értékét mutatja. A legfelső vonal az adatok maximális értékét jelölik. A piros doboz magassága a Q_3 és Q_1 különbsége, amit interkvartilis terjedelemnek (IQR) hívunk. Ez tartalmazza az adatok legjellemzőbb 50%-t. A dobozon belül található a Q_2, amely egyben a medián is, amely a legfontosabb helyzeti középérték. Ő felezi meg a nagyság szerint sorbarendezett adatokat. A medián csak akkor helyezkedik el a doboz közepén, ha az adatok eloszlása szimmetrikus. Esetünkben ez nem így van, hiszen a gazdasági életben a jelenségek nem szoktak szimmetrikusak lenni. Itt nagyon ritka a normális eloszlású jelenség. A maximális értéken túli üres karikák a kiugró értékeket jelölik. Ezek az 1,5*IQR-nél nagyobb adatokat jelölik. Ez azok a pénzintézetek, amelyek a legnagyobb eszközállománnyal rendelkeznek. Itt tehát látszólagos kiugró értékek szerepelnek, hiszen ezek valóban a bankok tényleges adatait jelölik.

A 40 magyarországi bank kvartilis értékei (MFt):

Minimum	Q_1	Median	Q_3	Maximum
3 031	38 796	137 802	858 910	6 213 397

IQR=820 114 MFt.

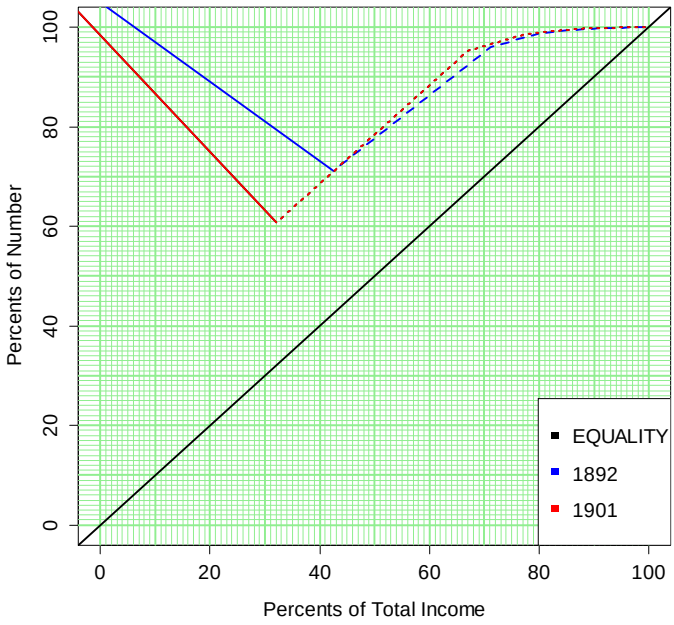
Mivel több látszólagos kiugró érték is van, ezért az első öt pénzintézet uralja az eszközök jelentős hányadát.

4.4. Lorenz görbe

Max Otto Lorenz amerikai közgazdász 1905-ben doktorandusz hallgatóként tanulmányozta a poroszországi jövedelmek eloszlását 1892. és 1901. évben. Arra volt kíváncsi, hogy az idő múlásával a jövedelmek eloszlása igazságosabb lesz-e. Fejlődik-e a társadalmi igazságosság? A cikk alapján ennek az ellenkezője igazolódott, és a jövedelmek koncentrációjának

növekedése ment végbe. A szegények szegényebbek, a gazdagok még gazdagabbak lettek. A tanulmány az American Statistical Association 1905. júniusi számában jelent meg.

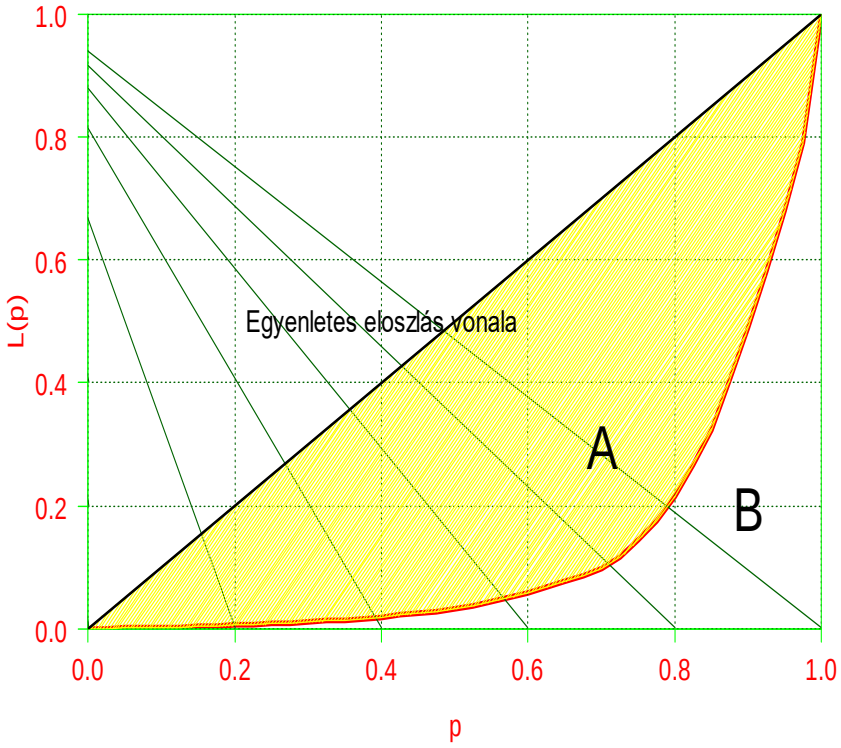
4.2. ábra: **Max Otto Lorenz amerikai közgazdász (1876-1959)**



4.3. ábra: **Lorenz eredeti ábrája 1905-ben**

A Lorenz-görbe alkalmazása:

- Relatív koncentráció szemléltetése
- Interpoláció
- Koncentráció idő és térbeli összehasonlítása
- Több, eltérő mértékegységű jelenség koncentrációjának összehasonlítása



4.4. ábra: **Lorenz-görbe mai formája**

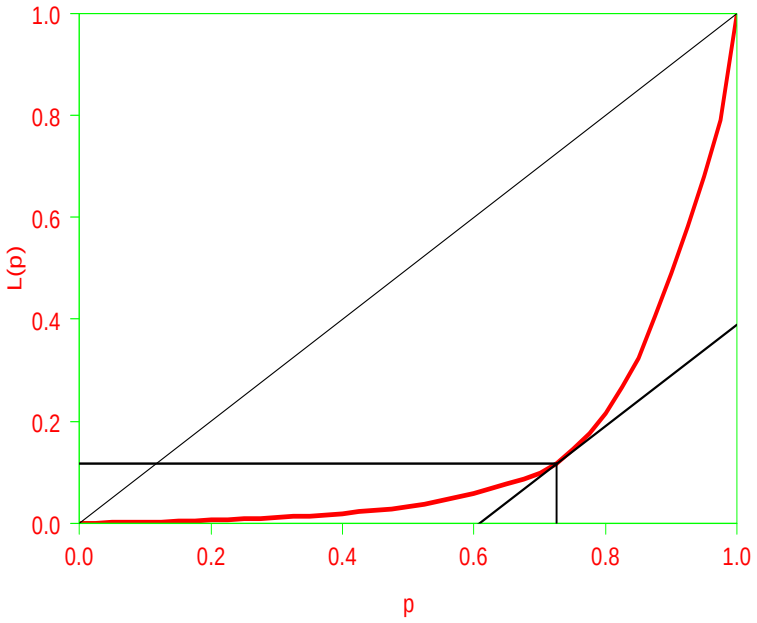
Az *A* terület nagysága mutatja a koncentráció fokát, ezt nevezik koncentrációs területnek. Minél nagyobb, annál nagyobb a koncentráció foka.

4.4.1. Átlagpont

Az átlagpont (4.5. ábra) a görbe azon része, ahol az átlóval párhuzamos egyenes érinti a görbét. Az egyenes meredeksége ekkor *tg(alfa)=1*. Ez megmutatja, hogy a piaci résztvevők hány százaléka rendelkezik átlag alatti eszközállománnyal. Ezt a x-tengelyre vetítve olvashatjuk le. Esetünkben ez 72,5%.

Az y-tengelyre vetített értéke megmutatja, hogy az átlag alatti piaci résztvevők az összes eszközállomány hány százalékát birtokolják. A példánkban ez 11,6%.

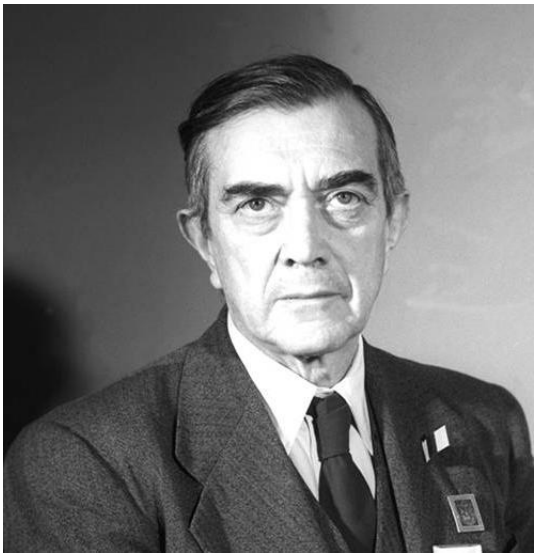
Fordítva is értelmezhetjük a kapott eredményeket. 27,5%-a bankoknak átlagnál nagyobb eszközállománnyal rendelkezik, és ezek az összes eszközállomány 88,4%-val rendelkeznek. Ez elég koncentráltnak tűnik.



4.5. ábra: **Átlagpont**

4.5. Gini-index

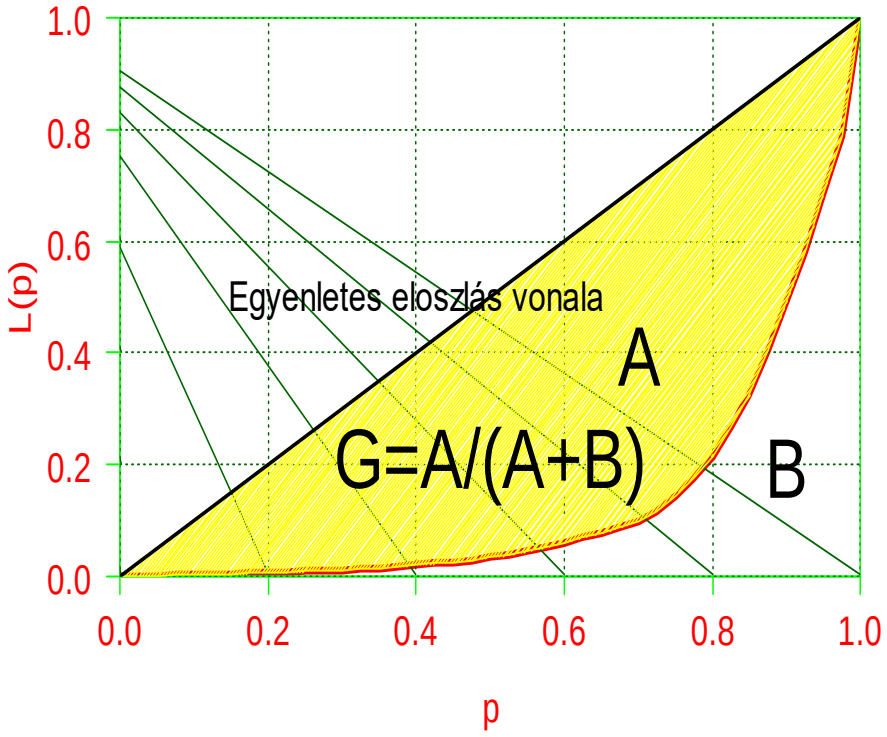
A Gini-index a koncentrációs terület (*A*) nagyságának meghatározása és osztva kettővel. Ez gyakorlatilag a *2A* területnek felel meg. A görbe alatti terület meghatározásához egy határozott integrál megközelítést fogunk használni. Ezt fogjuk alkalmazni az alapadatokon és a megoszlási viszonyszámokon is. A kétféle megközelítés pontosan ugyanazt az eredményt fogja adni.



4.6. ábra: **Corrado Gini olasz statisztikus és szociológus (1884-1965)**

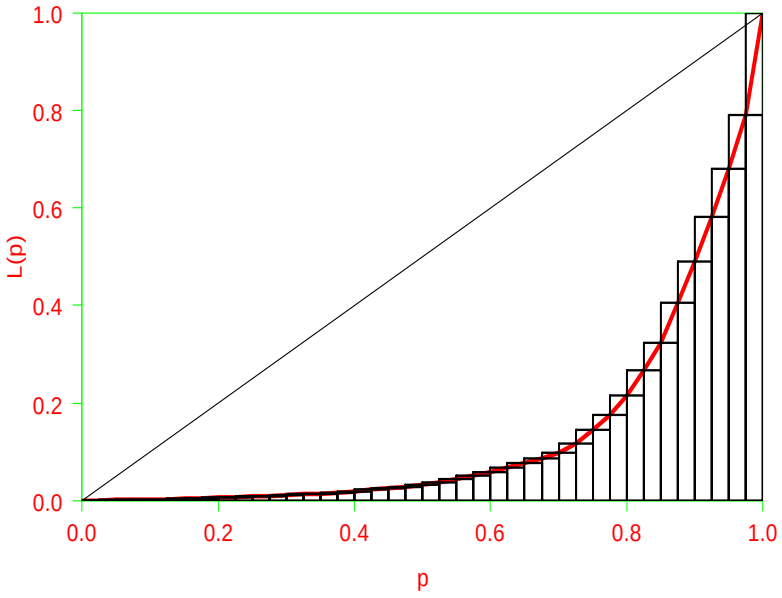
A 4.7. ábra mutatja a Gini-index értelmezését. Mivel az ábra egy 1x1-es négyzet, ezért a területe egyenlő eggyel. A koncentráció teljes hiányának vonala megfelezi ezt a területet. Az *A* terület tehát a koncentrációs terület. Minél nagyobb, annál nagyobb a koncentráció foka. Monopólium

esetén az *A=0,5*, ezért a Gini-index értéke 1. A koncentráció teljes hiánya esetén az index értéke 0.



4.7. ábra: **A Gini-index grafikus értelmezése**

A 4.8. ábra a határozott integrál közelítő eljárását mutatja. Hasonlót alkalmazott annak idején Gauss is. Közelítő eljárással meghatározzuk az alsó és felső oszlopok területét és egyszerű számtani átlaggal becsüljük a tényleges területet.



4.8. ábra: **A görbe alatti terület becslése határozott integrállal**

A Gini-index tehát *G=2A*. Ezt különböző algoritmusokkal tudjuk meghatározni. Lehet az alapadatokból és a megoszlási viszonyszámokból is számítani.

Gini-index számítása alapadatokból:

$$G = \frac{2\sum_{i=1}^{n} iy_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}$$

ahol:

n: alapadatok száma, esetünkben 40

i: adatok sorszáma, esetünkben 1…40

y_i: alapadatok, esetünkben az eszközállományok Ft-ban

A szumma y_i az összes eszközállományt jelenti. Az y_i a kicsitől a nagy felé rendezett alapadatot jelenti. Ez nagyon fontos, máskülönben rossz eredményt kapunk.

Lássuk a számítást:

$$G = \frac{2 * 1\,038\,710\,831}{40 * 29\,613\,973} - \frac{41}{40} = 0{,}7288$$

Sajátos szóródási mutatóként is felfogható a Gini-index. Itt nem a számtani átlagtól vett eltérések átlagaként határozzuk meg a szórást, hanem minden adat minden adattól vett eltéréseként. A két szumma jel a sor és oszlop irányú összegzést jelenti. Gyakorlatilag a mátrix elemeinek abszolút értékeit kell összeadni.

$$G = \frac{\sum_i \sum_j |y_i - y_j|}{2n \sum_{i=1}^{n} y_i}$$

A számítás eredménye:

$$G = \frac{1\,726\,497\,538}{2 * 40 * 29\,613\,973} = 0{,}7288$$

Gini-index számítása megoszlási viszonyszámokból:

$$G = 2 \sum_{i=1}^{n} \frac{i}{n} y_i - \frac{n+1}{n}$$

ahol:

n: alapadatok száma, esetünkben 40

i: adatok sorszáma, esetünkben 1…40

y_i: megoszlási viszonyszámok (nem százalékos formátumban)

Az *i/n* a sorszámok kumulatív relatív gyakoriságát jelentik (0,025; 0,05 … 1).

A számítás eredménye:

$$G = 2 * 0{,}8769 - \frac{41}{40} = 0{,}7288$$

Ebben az esetben is szóródási mutatóként is számíthatjuk a Gini-index.

$$G = \frac{\sum_i \sum_j |y_i - y_j|}{2n}$$

A számítás eredménye:

$$G = \frac{58{,}3}{80} = 0{,}7288$$

Amint látjuk, mind a négy képlet tökéletesen ugyanazt az eredményt adta. A Gini-index alkalmas az abszolút és relatív koncentráció jellemzésére.

4.6. Herfindahl-Hirschman-index

A Herfindahl-Hirschman-indexet a közgazdaságtanban a piaci koncentráció jellemzésére használjuk. Jelölése: HHI.

Egy adott ágazat, gazdasági szektor HHI-e a piacon található vállalatok, egységek részesedésének, megoszlási viszonyszámainak (V_{mi}) négyzetösszege.

A Herfindahl-Hirschman-index képlete

$$HHI = \sum_{i=1}^{n} V_{mi}^2$$

A HHI értéke *1/n* és *1* között van. *1/n* értéket akkor vesz fel a mutató, ha a gazdasági szereplők egyenlő piaci részesedéssel rendelkeznek. Amennyiben sok, egyenként kicsi piaci részesedéssel rendelkező szereplő van, akkor a HHI értéke nullához közelít. Egyhez közeli érték esetén egy szereplő kezében koncentrálódik a piaci részesedés jelentős része, ekkor beszélünk monopóliumról. Ebben az esetben a szabad piaci verseny veszélybe kerülhet. A HHI indexet ezért használják a különböző állami felügyeleti szervek.

4.6.1. A HHI matematika elmélete

A piaci szereplők száma legyen *n*. A részesedésüket jelöljük V_{m1}, V_{m2}, ..., V_{mn}-nel. A piaci részesedések megegyeznek a korábban tárgyalt megoszlási viszonyszámokkal, ezért jelöltük *V*-vel. Itt is a részsokaságot viszonyítottuk az egészhez. A megoszlási viszonyszámok átlaga pedig *1/n* volt. Ezt az összefüggést a HHI meghatározásakor is hasznosítani fogjuk. Természetesen a piaci részesedések egyenkénti összege egyenlő 1-gyel, amit az alábbi módon írhatunk fel:

$$\sum_{i=1}^{n} V_{mi} = 1$$

Általánosságban a szóródási mutatók az átlagtól mért eltéréseket jellemzik, amit a távolságok négyzetes átlagával becslünk. Határozzuk meg ebben az esetben is a piaci részesedések átlagtól vett eltérés-négyzetösszegét. Ezt használjuk a variancia, ill. a szórás meghatározásakor is az első lépésben.

$$\sum \left(V_{mi} - \overline{V}_{mi}\right)^2 = \sum_{i=1}^{n} \left(V_{mi} - \frac{1}{n}\right)^2$$

Ez az összeg akkor egyelő nullával, ha minden szereplő piaci részesedése az átlaggal egyenlő. Amennyiben nem, akkor nullánál nagyobb értéket eredményez. Erős koncentrációnál közelíteni fog egyhez.

Végezzük el a négyzetre emelést és vizsgáljuk meg a három tagot.

$$\sum \left(V_{mi}^2 + \frac{1}{n^2} - \frac{2V_{mi}}{n}\right)$$

Az első tag = HHI, mivel az index a piaci részesedések négyzetösszege.

A második tag: $\frac{n}{n^2} = \frac{1}{n}$, mivel n-szer kell összeadni egy konstanst.

A harmadik tag: $\frac{2}{n}\sum V_{mi}$, mivel $\sum V_{mi} = 1$, ezért ez a kifejezés egyenlő $-\frac{2}{n}$ -vel.

A koncentráció teljes hiánya esetén a képletünk az alábbi módon alakul:

$$HHI + \frac{1}{n} - \frac{2}{n} = 0$$

$$HHI - \frac{1}{n} = 0$$

$$HHI = \frac{1}{n}$$

A fenti levezetés tehát igazolja, hogy a HHI minimális értéke $\frac{1}{n}$, maximuma 1 lehet.

4.2. táblázat: **Magyarországi bankok összes eszközállománya MFt**

Megnevezés	Eszközök összesen	Megnevezés	Eszközök összesen	Megnevezés	Eszközök összesen	Megnevezés	Eszközök összesen
OTP Bank Nyrt.	6 213 397	FHB Jelzálogbank Nyrt.	845 205	UniCredit Jelzálogbank Zrt.	136 925	DRB Dél-Dunántúli Regionális Bank Zrt.	38 335
Kereskedelmi és Hitelbank Zrt.	3 213 379	Magyarországi Volksbank Zrt.	503 582	SOPRON BANK BURGENLAND Zrt.	97 129	Kinizsi Bank Zrt.	35 545
ERSTE BANK HUNGARY Zrt.	2 948 517	Magyar Takarékszövetkezeti Bank Zrt.	379 938	Magyar Cetelem Bank Zrt.	85 895	Mohácsi Takarék Bank Zrt.	33 959
MKB Bank Zrt.	2 749 837	Merkantil Váltó és Vagyonbefektető Bank Zrt.	277 388	Allianz Bank Zrt.	77 534	Garantiqa Hitelgarancia Zrt.*	32 325
CIB Bank Zrt.	2 482 860	FHB Kereskedelmi Bank Zrt.	267 742	Deutsche Bank Zrt.	76 208	Banif Plus Bank Zrt.	29 395
Raiffeisen Bank Zrt.	2 400 580	Commerzbank Zrt.	262 298	Központi Elszámolóház és Értéktár (Budapest) Zrt.	69 437	Credigen Bank Zrt.	20 862
OTP Jelzálogbank Zrt.	1 675 031	Fundamenta-Lakáskassza Lakás-takarékpénztár Zrt.	254 718	MagNet Magyar Közösségi Bank Zrt.	56 246	GRÁNIT Bank Zrt.	13 081
UniCredit Bank Hungary Zrt.	1 566 193	Magyar Export-Import Bank Zrt.	194 696	Banco Popolare Hungary Bank Zrt.	48 975	Hanwha Bank Magyarország Zrt.	11 648
MFB Magyar Fejlesztési Bank Zrt.	1 189 217	OTP Lakástakarékpénztár Zrt.	192 610	Porsche Bank Hungaria Zrt.	48 475	Széchenyi Kereskedelmi Bank Zrt.	4 126
BUDAPEST Hitel- és Fejlesztési Bank Nyrt.	900 025	KDB Bank (Magyarország) Zrt.	138 679	Bank of China (Hungária) Hitelintézet Zrt.	38 950	MV-Magyar Vállalkozásfinanszírozási Zrt.*	3 031

Forrás: PSZÁF, 2010.

A bankok összes eszközállománya: 29 613 973 millió Ft.

Példa:

Határozzuk meg a magyarországi bankok 2010. évi piaci részesedését, az összes eszközállomány figyelembevételével.

4.3. táblázat**: A magyarországi bankok piaci részesedése nagyság szerint rendezve**

Megnevezés	Piaci részesedés	Megnevezés	Piaci részesedés	Megnevezés	Piaci részesedés	Megnevezés	Piaci részesedés
OTP Bank Nyrt.	20,98%	FHB Jelzálogbank Nyrt.	2,85%	UniCredit Jelzálogbank Zrt.	0,46%	DRB Dél-Dunántúli Regionális Bank Zrt.	0,13%
Kereskedelmi és Hitelbank Zrt.	10,85%	Magyarországi Volksbank Zrt.	1,70%	SOPRON BANK BURGENLAND Zrt.	0,33%	Kinizsi Bank Zrt.	0,12%
ERSTE BANK HUNGARY Zrt.	9,96%	Magyar Takarékszövetkezeti Bank Zrt.	1,28%	Magyar Cetelem Bank Zrt.	0,29%	Mohácsi Takarék Bank Zrt.	0,11%
MKB Bank Zrt.	9,29%	Merkantil Váltó és Vagyonbefektető Bank Zrt.	0,94%	Allianz Bank Zrt.	0,26%	Garantiqa Hitelgarancia Zrt.*	0,11%
CIB Bank Zrt.	8,38%	FHB Kereskedelmi Bank Zrt.	0,90%	Deutsche Bank Zrt.	0,26%	Banif Plus Bank Zrt.	0,10%
Raiffeisen Bank Zrt.	8,11%	Commerzbank Zrt.	0,89%	Központi Elszámolóház és Értéktár (Budapest) Zrt.	0,23%	Credigen Bank Zrt.	0,07%
OTP Jelzálogbank Zrt.	5,66%	Fundamenta-Lakáskassza Lakás-takarékpénztár Zrt.	0,86%	MagNet Magyar Közösségi Bank Zrt.	0,19%	GRÁNIT Bank Zrt.	0,04%
UniCredit Bank Hungary Zrt.	5,29%	Magyar Export-Import Bank Zrt.	0,66%	Banco Popolare Hungary Bank Zrt.	0,17%	Hanwha Bank Magyarország Zrt.	0,04%
MFB Magyar Fejlesztési Bank Zrt.	4,02%	OTP Lakástakarékpénztár Zrt.	0,65%	Porsche Bank Hungaria Zrt.	0,16%	Széchenyi Kereskedelmi Bank Zrt.	0,01%
BUDAPEST Hitel- és Fejlesztési Bank Nyrt.	3,04%	KDB Bank (Magyarország) Zrt.	0,47%	Bank of China (Hungária) Hitelintézet Zrt.	0,13%	MV-Magyar Vállalkozásfinanszírozási Zrt.*	0,01%

Határozzuk meg a HHI értékét, képezzük a piaci részesedés négyzetösszegét.

Ennek az értéke: 0,0982

Most el kell dönteni, hogy ez az érték a koncentráció milyen fokát mutatja. A HHI minimális értéke *1/n.* A példában 40 bank szerepel, ezért a minimális érték 0,0250. Ez akkor lenne, ha minden bank azonos piaci részesedéssel bírna. A számított érték ennél nagyobb, azonban nem éri el a 0,1-es lélektani határt. A gyakorlatban a 0,1-es érték alatt a koncentráció hiányáról beszélhetünk. A HHI kritikus értékeit a gyakorlati alkalmazás részben ismertetjük.

4.6.2. A HHI különböző változatai

Sokszor a piaci részesedést nem 0 és 1 közötti számmal jellemzik, hanem százalékos formában. Ekkor a HHI értéke nem 0 és 1 között, hanem 0 és 10 000 között alakul. A 10 000 mutatja a teljes koncentráció mértékét (100*100).

Példa:

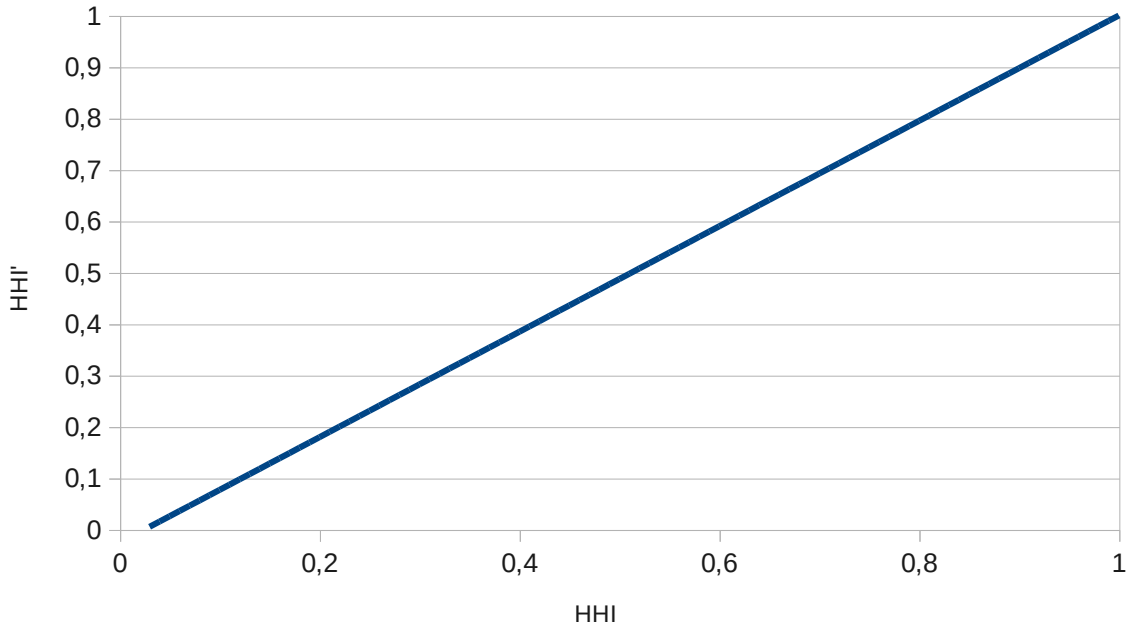
Határozzuk meg az előbb kiszámított HHI-t a százalékos adatok felhasználásával. Amennyiben jól számoltunk, 982-t kell kapni.

Normalizált Herfindahl-Hirschman-index

A normalizált index legkisebb értéke 0, legnagyobb értéke pedig 1 lehet. Ezt egy skálaeltolással és transzformációval tudjuk előállítani. A normalizált indexet jelöljük HHI'-vel.

$$HHI' = \frac{HHI - \frac{1}{n}}{1 - \frac{1}{n}}$$

Az így előállított normalizált HHI értéke tehát 0 és 1 között van, ellentétben a hagyományos HHI-vel szemben, aminek a minimális értéke *1/n*.



4.9. ábra. **A HHI és normalizált HHI közötti összefüggés, n=40**

Példa:

Határozzuk meg a normalizált HHI-t.

$$HHI' = \frac{0{,}0982 - \frac{1}{40}}{1 - \frac{1}{40}}$$

Ennek az értéke: 0,0751

4.6.3. Gyakorlati alkalmazás

A modern piacgazdaságokban az állam egyik legfontosabb gazdasági feladata az, hogy őrködjön a piaci verseny szabadsága fölött. Ennek egyik eszköze az, hogy az állam felügyeleti jogkörével élve visszaszorítja a túlzott piaci fölény megszerzésére irányuló törekvéseket. A különböző állami felügyeleti szervek gyakran használják a HHI-t annak objektív mérésére, hogy egy adott piaci szektor, vagy egy esetleges cégfúzió után létrejövő piaci helyzet nem túlzottan koncentrált-e.

Az Egyesült Államok Igazságügyi Minisztériumának Versenyhivatala (Antitrust Division of the US Department of Justice) például a Herfindahl–Hirschman-index segítségével hoz döntést arról, hogy jóváhagyjon-e cégegyesüléseket. Ha a tervezett cégfúziót követően a kérdéses piaci szektorban a HHI 0,1 alatt marad, akkor a Versenyhivatal nem tekinti aggályosnak az egyesülést. Másrészt, ha a fúzió utáni HHI 0,18 fölött van, és a HHI a cégegyesülés hatására

több mint 0,01-dal növekszik, akkor az egyesülni kívánó cégeknek igazolniuk kell, hogy egyéb okok miatt nem várható, hogy a fúzió nyomán tisztességtelen előnyhöz jutnának (Horizontal Merger Guidelines: Concentration and Market Shares U.S. Department of Justice and the Federal Trade Commission).

A HHI értékének kritikus értéke tehát 0,1. Ez csak akkor következhet be, ha tíznél több piaci szereplő van a szektorban.

Magyarországon a PSZÁF és elődszervezetei valamint a Magyar Nemzeti Bank a 90-es évek óta figyelemmel kísérik a bankrendszer Herfindahl–Hirschman-indexét, amely az 1991-es 0,1565-ről 2002-re a 0,0986-os értékig csökkent (VÁRHEGYI, 2004).

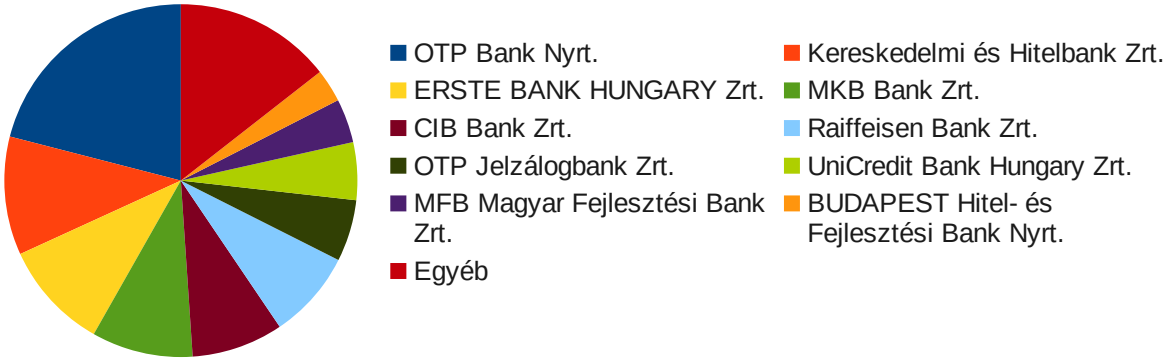
Példa:

Ábrázoljuk az első 10 legjelentősebb bank piaci részesedését kördiagramon. Képezzünk egy egyéb kategóriát is.

Becsüljük meg a HHI értékét a fenti adatok birtokában. Képzeljük el, hogy egy folyóiratból csak ezek az adatok állnak rendelkezésünkre. Az első tíz bank piaci részesedése a 25. táblázatban látható, az egyéb kategória 14,44%-t képvisel. Milyen pontosan lehet megbecsülni a HHI-t? Mivel becslésről van szó, egy alsó és felső értéket kell meghatároznunk, ami közé fog esni a valódi HHI.

Az alsó érték becslése a 10 első bank piaci részesedése alapján számított HHI, ennek az értéke: 0,0964.

A felső érték becsléséhez képzeljük el, hogy az egyéb kategória 14,44%-os részesedését n számú bank adja. Ezen bankok piaci részesedésének négyzetösszege maximum $0{,}1444^2/n$ lehet. Mivel az egyéb kategóriába tartozó bankok mindegyikének kisebb a piaci részesedése, mint 3,04%, ezért n legkisebb értéke $\frac{14{,}44}{3{,}04}$ felfelé kerekítve 5 lehet. Ezek értelmében az egyéb kategóriába tartozó bankok piaci részesedésének maximális négyzetösszege $0{,}1444^2/5=0{,}0042$ lehet. A felső érték ezek szerint 0,0964+0,0042=0,1006. A valódi HHI 0,0964 és 0,1006 között van. (a tényleges érték 0,0982) A becslésünk tehát pontos volt.



4.10. ábra. **A tíz legjelentősebb bank piaci részesedése**

Példa:

Vizsgáljuk meg, hogyan változik a HHI az első két bank fúziója után. Képzeljük el, hogy az OTP és a Kereskedelmi Bank egyesül (ez csak fikció). Az így létrejött fúzió után a HHI értéke: 0,1438. Ez már aggályos mértékű, mert meghaladja a 0,1-es értéket, és a növekedés mértéke is nagyobb, mint 0,01. Ezt az egyesülést az USA-ban alaposan indokolni kellene, mert felmerülhet a tisztességtelen piaci előny megszerzése.

4.6.4. A variációs együttható és a HHI közötti összefüggés

A variációs koefficiens (CV) és a HHI hasonló tulajdonságot jellemez, a kettő egymásba átszámítható. Mivel a CV a relatív változékonyságot mutatja, ezért a HHI alkalmas a relatív koncentráció mérésére. Amennyiben ismerjük az egyik értékét, a másik meghatározható belőle.

Az alapösszefüggés:

$$HHI = \frac{CV^2 + 1}{n}$$

Matematikai elmélet

Már tudjuk, hogy a HHI a piaci részesedések négyzetösszege.

$$HHI = \sum_{i=1}^{n} V_{mi}^2$$

A variációs koefficiens a szórás és a számtani átlag hányadosa.

$$CV = \frac{S}{\overline{x}}$$

A megoszlási viszonyszámok szórásának képlete:

$$S = \sqrt{\frac{\Sigma\left(V_{mi} - \frac{1}{n}\right)^2}{n}}$$

Mivel a piaci részesedés átlaga:

$$\overline{x} = \frac{1}{n}$$

A fentiek ismeretében írjuk fel a piaci részesedés variációs együtthatóját:

$$CV = \frac{\sqrt{\frac{\Sigma\left(V_{mi} - \frac{1}{n}\right)^2}{n}}}{\frac{1}{n}}$$

Végezzük el az alábbi számításokat.

$$S = n\sqrt{\frac{\Sigma\left(V_{mi} - \frac{1}{n}\right)^2}{n}}$$

$$CV^2 = \frac{n^2 \sum \left(V_{mi} - \frac{1}{n}\right)^2}{n}$$

$$CV^2 = n \sum \left(V_{mi} - \frac{1}{n}\right)^2$$

A szummás kifejezésről már korábban bebizonyítottuk, hogy egyenlő HHI-1/n-nel, ezért:

$$CV^2 = n\left(HHI - \frac{1}{n}\right)$$

$$CV^2_{\square} = n * HHI - 1$$

$$CV^2_{\square} + 1 = n * HHI$$

Az utolsó lépesben megkapjuk a HHI és CV közötti összefüggés képletét.

$$HHI = \frac{CV^2 + 1}{n}$$

Fontos megjegyzés a számításokhoz. Csak abban az esetben kapunk helyes eredményt, ha sokasági szórást használunk a CV meghatározásakor, azaz n-nel osztunk. Ebben az esetben a CV maximális értéke:

$$CV_{max} = \sqrt{n - 1}$$

Példa:

Számítsuk ki a fenti adatok felhasználásával a CV értékét, és ebből határozzuk meg a HHI-t. Először a piaci részesedés átlagát és szórását kell meghatározni. A piaci részesedés átlaga egyszerű, mert 1/40 azaz 0,025. A szórása: 0,0428. Még egyszer hangsúlyozzuk, hogy a sokasági szórást kell meghatározni, tehát *n*-nel kell osztani. A variációs koefficiens tehát 1,7116, százalékban kifejezve 171%. A további számításokat ne a százalékos értékkel végezzük.

$$HHI = \frac{1{,}7116^2 + 1}{40} = 0{,}0982$$

Az eredmény tökéletesen megegyezik a korábban kiszámolt értékkel.

5. A lineáris regressziós modellezés alapjai

A regressziós elemzés a közgazdaságtan és pénzügytan során leggyakrabban használt módszer, amely egyúttal más, komplikáltabb módszerek alapját is nyújtja. Az ökonometria eszköztára az 1960-as évek óta jelentős megnőtt, ennek ellenére a legtöbb módszer még mindig a regressziós modellezésen alapszik, vagy ahhoz szorosan kapcsolódik.

5.1. A kétváltozós lineáris regresszió alapjai

A legegyszerűbb esetben mindössze két változóval rendelkezünk, amelyeket y_i és x_i jelöl. A modellezés során y_i jelöli az úgynevezett függő változót, amely a kutatás középpontjában áll, míg x_i a magyarázó változót. A megfigyeléseket i jelzi, amelyek1-től n-ig terjednek, ahol utóbbi a mintanagyság. Mivel két változóval dolgozunk, ezért kétváltozós lineáris regresszióról beszélünk, de rendkívül egyszerű a modellt realisztikusabbá és flexibilisebbé tenni több x_i változó és transzformációk alkalmazásával (ekkor már többváltozós lineáris regresszióról beszélünk). A regresszió a feltételes várható értéket modellezi, azaz azt, hogy y_i átlagos értéke hogyan változik x_i értékének függvényében. A megfogalmazás itt szándékosan összefüggésre utal (y_i értékének változása akkor különböző x_i értékek esetén), és nem oksági kapcsolatra (mi x_i hatása y_i változóra). Bizonyos esetekben le lehet vonni oksági következtetéseket is, megfelelő elővigyázatosság mellett.

A kétváltozós lineáris regressziót az alábbiak szerint írhatjuk fel:

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

Ebben az esetben y_i a függő változó, x_i a magyarázó változó, β_0 és β_1 a becsült paraméterek, amelyek a függő és a magyarázó változó közötti összefüggést mutatják. A kapcsolatot egy hibatag (maradéktag, rezidum) befolyásolja, amelyet ε_i jelöl, amely magában foglal minden olyan hatást, amelyet nem modelleztünk. A későbbiekben számos feltételt teszünk a hibatagra, amelyeket teljesítenie kell ahhoz, hogy a modell „elfogadható” legyen. A hibatag nem figyelhető meg közvetlenül, csak a becslés után látjuk, hogy „mekkorát tévedtünk”. Amennyiben közvetlenül megfigyelhető lenne, úgy változóként a modellbe illesztve y_i ingadozását tökéletesen meg tudnánk magyarázni. Mivel nem vagyunk képesek olyan modellt építeni, amely hiba nélkül meg tudna magyarázni bármilyen jelenséget, ezért a hibatag mindig a regressziós modell része. Úgy is fogalmazhatunk, hogy ide kerül minden olyan jelenség, amelyet a kutató ignorál. Itt fel kell arra hívni a figyelmet, hogy két dolog miatt ignorálhatunk

egy hatást: vagy azért mert olyan jelentéktelen és véletlenszerű, hogy nincs értelme modellezni, vagy azért mert szakértelem vagy megfelelő adatok hiányában a modell felépítése nem megfelelő.

5.2. A paraméterek becslése

A paramétereket a megfigyelt adatok segítségével becsüljük meg. A paramétereket az úgynevezett Legkisebb Négyzetek Módszerével becsüljük (Ordinary Least Squares, OLS), amely a maradék négyzetösszeg minimalizálására épül. Mivel az OLS magyar nyelvterületen is gyakran használt, ezért a továbbiakban OLS becslésként hivatkozunk a paraméter becslésekre. Tekintsük meg újra a következő modellt:

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

A fenti egyenletben y_i és x_i adatok segítségével becsüljük meg β_0 és β_1 paraméterek értékét. Ezek a paraméterek az elméleti (populációs) paramétereket jelzik, amelyeket egy adott mintán keresztül tudunk megbecsülni. Ebben az esetben a mintából becsült paramétereket $\hat{\beta}_0$ és $\hat{\beta}_1$ (kalap) jelöléssel látjuk el. A becslési probléma a következő:

$$min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Azaz a paramétereket szeretnénk úgy megválasztani, hogy a hibatag négyzetes összege a lehető legkisebb legyen. A hibatag nem más, mint a függő változó és annak becsült értéke közötti eltérés $(y_i - \hat{y}_i)$. Ezt az egyenletet behelyettesítve, paraméterek szerint deriválva, majd azokat nullával egyenlővé téve megoldhatjuk a normálegyenleteket és kifejezhetjük a paraméterek becslőfüggvényeit. Ezek az alábbiak szerint alakulnak:

$$\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_i)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ahol $\bar{y}$ és $\bar{x}$ a függő és a magyarázó változó átlagát jelentik. Ezek segítségével kiszámítható a konstans becsült paramétere, $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

A maradéktagok négyzetes változatára részben azért van szükség, mert az OLS becslés eredményeként a maradékok nullára összegződnek:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

A becsült y_i értékeket felírhatjuk végül az alábbi formában:

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

Érdemes még egy összefüggést megjegyezni a becsült maradéktaggal kapcsolatban még, amely a matematikai levezetések során hasznos lehet:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \, x_i = 0 \text{ és } \sum_{i=1}^{n} x_i = n\bar{x}$$

5.3. A becsült paraméterek értelmezése és hipotézis vizsgálat

5.3.1. Paraméter értelmezés

A β_0 és β_1 a ***konstans*** és a ***meredekség*** paraméter (az angol elnevezésben intercept és slope). A becsült paraméterek interpretációján keresztül érthető meg a függő és a magyarázó változók közötti összefüggés. Az interpretáció a következő:

- **konstans, $\boldsymbol{\beta_0}$:** y_i átlagos értékét mutatja, abban az esetben ha x_i értéke nulla. Mivel több magyarázó változó esetén az egyenletnek legtöbbször nincs értelme, ha minden magyarázó változót nulla értékre állítunk, ezért a gyakorlatban a konstans nem kap különösebb figyelmet.
- **meredekség, $\boldsymbol{\beta_1}$:** y_i értéke β_1 értékével magasabb átlagosan azon megfigyelések esetében, ahol x_i egységnyi értékkel magasabb. Tehát x_i nem „okozza” a változást, pusztán azokat a megfigyeléseket összehasonlítva, amely esetében x_i értéke egységnyivel különbözik, azt várjuk, hogy y_i értéke β_1 értékével lesz magasabb

átlagosan, az egységnyi értékkel magasabb x_i esetében. Az egyszerűség kedvéért gyakran úgy is fogalmazhatunk, hogy egységnyi értékkel magasabb x_i esetében y_i értéke átlagosan β_1 értékkel változik, minden mást változatlannak tekintve. Itt viszont figyelni kell arra, hogy x_i értéke nem „időben" növekszik egységnyi értékkel, hanem két csoportot hasonlítunk össze, ahol az egyik csoport értéke x, míg a másiké $x + 1$.

Egy egyszerű példa segíti a megértést, amely során az egyének közötti jövedelem ingadozást az iskolai végzettséggel magyarázzuk:

$$jövedelem_i = \beta_0 + \beta_1 végzettség_i + \varepsilon_i$$

Itt legyen y_i egy adott személy jövedelme (HUF), míg x_i az adott személy iskolai végzettsége (oktatásban eltöltött évek száma). A β_1 ebben az esetben az oktatás hatását mutatja meg (azaz két csoportot összehasonlítva, ahol az egyik csoport iskolai végzettsége egy évvel magasabb, ott az átlagos jövedelem β_1 értékével lesz várhatóan magasabb). A becsült β_1 paraméter esetében azt várjuk, hogy pozitív értékű lesz, és a legtöbb gyakorlati eredmény (szerencsére) ezt alá is támasztja. A konstans β_0 pedig az átlagos jövedelmet mutatja meg egy végzettséggel nem rendelkező egyén esetében (azaz ha $x = 0$). Ez a kapcsolat matematikailag és intuitív szempontból is egyszerű, bár azonnal felmerül néhány kérdés. Más változókra is szükségünk van, hiszen nem csak az oktatási szint befolyásolja a jövedelmet. Ezen felül, a jövedelem változása nem feltétlenül ugyanakkora az oktatás különböző szintjein: a felsőoktatási végzettséggel és az alapfokú végzettséggel rendelkezők bére közötti különbség várhatóan sokkal nagyobb, mint egy első- és egy másoddiplomás bére között. Ezen felül olyan változóknak is lehet hatása, amelyet nem tudunk (megfelelően) mérni. Ilyen az intelligencia szint (amelyre az IQ pontok számítása csak egy közelítés), vagy az adott személye tehetsége. Míg az intelligenciát mérjük, de a mérés nem tökéletes, addig a tehetséget nem is tudjuk mérni. Ez jól szemlélteti azt, hogy a két változó közötti komplex kapcsolatot nem lehet ilyen egyszerűen modellezni. A jó hír viszont az, hogy megfelelő adatokkal és szaktudással gyakorlatilag bármilyen kapcsolat megfelelően modellezhető a későbbiekben.

A „minden mást változatlannak tekintve" kifejezés a *ceteris paribus* elvre utal, azaz, a regresszió parciális hatást mér. Az előző példát egészítsük a munkaerőpiaci tapasztalat bevonásával is, amelynek hatását β_2 mutatja:

$$jövedelem_i = \beta_0 + \beta_1 végzettség_i + \beta_2 tapasztalat_i + \varepsilon_i$$

A becsült β_1 paraméter azt mutatja, hogy mekkora az átlagos bérkülönbség két olyan megfigyelés között, akik csak egy évnyi oktatásban különböznek (x vs. $x + 1$), *és* azonos munkaerőpiaci tapasztalattal rendelkeznek. Ilyenkor gyakran úgy fogalmazunk, hogy kontrolláltunk a munkaerőpiaci tapasztalatra a modellben. A változók vizsgálatakor csak a végzettség, illetve csak a munkaerőpiaci tapasztalat hatására vagyunk kíváncsiak külön-külön. Tehát a végzettség és az átlagos bérváltozás összefüggésére, miközben olyan megfigyeléseket hasonlítunk össze, amelyek munkaerőpaci tapasztalata azonos. Matematikailag ezt egyszerű deriválással lehet kifejezni. Vegyük például a legegyszerűbb többváltozós regressziót, azaz:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Ilyenkor:

$$\frac{\partial y}{\partial x_1} = \beta_1 \text{ és } \frac{\partial y}{\partial x_2} = \beta_2$$

Amennyiben nem lineáris tagokat is szerepeltetünk a modellben, változik a deriválás eredménye. Például ha x_2 helyett x_1 négyzetes tagját vonjuk be magyarázó változóként (x_1^2), akkor:

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2\beta_2 x_1$$

Ebben az esetben a hatás már nem konstans a megfigyelések minden szintjén, hanem x_1 aktuális értékétől is függ! Az interpretáció során különösen óvatosan kell eljárni, ha speciális tagokat is használunk.

Mielőtt meghatározzuk a változók kapcsolatát, több feltételt is teszünk. Az egyik azt mondja ki, hogy a hibatag várható értéke (átlaga) nullával egyenlő ($E(.)$ a várható érték operátor):

$$E(\varepsilon_i) = 0$$

Ez a feltétel azt mondja, hogy a maradéktagba került hatások átlagos értéke nulla. Ez a feltétel nem mond semmit y_i és x_i kapcsolatáról. Ezért egy, a modellezés szempontjából erősebb feltételt teszünk:

$$E(\varepsilon_i|x_i) = 0$$

Ez azt jelenti, hogy a hibatag értéke nem függ x_i aktuális értékétől. Ezekre a feltételekre azért van szükség, hogy a regressziós modell megfelelően értelmezhető legyen.

Egyes esetekben x_i nem folytonos, hanem bináris (amit gyakran dummy vagy kétértékű változónak is neveznek). Ebben az esetben x_i értéke két szintű, amelyet legtöbbször 0 és 1 jelez. Bármilyen alkalmas, kategorikus változót jelölhetünk bináris változókkal, például nemi különbségek esetén lehet a férfi = 0, nő = 1, vagy cégelemzés során a hazai tulajdonú cégek = 0, külföldi tulajdonú cégek = 1. A paraméter interpretáció ilyenkor némileg változik:

- **meredekség bináris változó esetén, β_1:** a β_1 paraméter y_i átlagos különbségét mutatja a két csoport között (0 és 1). Azaz, y_i átlagos értéke közötti különbséget mutatja azon csoportok esetében, amelyek az 1-essel jelölt, illetve a 0-val jelölt csoportba tartoznak.

A 0-val jelölt csoport lesz mindig a referencia szint, amelyet könnyű megérteni az alábbi példa alapján. Legyen x_i egy bináris változó, amely 0 és 1 értéket vesz fel. Ebben az esetben a feltételes várható értékek a következő képpen alakulnak:

$$E(y_i|x_i = 0) = \beta_0 + \beta_1 * 0 = \beta_0$$

Illetve:

$$E(y_i|x_i = 1) = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1$$

A két feltételes várható érték közötti különbség β_1. Azaz β_1 pontosan azt mutatja amit vártunk: mi a különbség az átlagos y_i értékben akkor ha $x_i = 0$ és ha $x_i = 1$. A várható érték behelyettesíthető az átlaggal a gyakorlatban, míg a feltételes várható érték olyan átlagra vonatkozik, ahol a csoportra valamilyen feltételt tettünk. (Például ha testsúly adataink vannak, akkor az átlagos testsúly ad becslést a várható értékre a mintában, míg a feltételes várható érték elképzelhető úgy, ha az átlagos testsúly értékét férfiak és nők szerinti csoportosításban nézzük).

5.3.2. Hipotézis vizsgálat

A becsült paraméterek értéke a minta függvényében ingadozik, de mi csak egyetlen mintával rendelkezünk, ezért szeretnénk tudni, hogy a becsült paraméterek statisztikai értelemben különböznek-e nullától. Az egyéni paraméterek szignifikancia tesztelése az alábbi módon történik:

$$t = \frac{becsült\ paraméter}{becsült\ paraméter\ standard\ hibája}$$

A nullhipotézis szerint $H_o: regressziós\ paraméter = 0$, míg az alternatív hipotézis szerint $H_1: regressziós\ paraméter \neq 0$. Itt nem teszünk arra megkötést, hogy a hatás csak negatív vagy pozitív lehet. Amennyiben a p-érték kisebb mint a választott szignifikancia szint (legtöbbször 0.05), abban az esetben elutasítjuk ezt a nullhipotézist. Ezzel azt mutatjuk meg, hogy amennyiben a fennmaradó magyarázó változók hatását kiszűrtük, úgy a tesztelni kívánt változónak nincs hatása a függő változó várható értékére. A nulla hatás azért fontos, mert ha a becsült paraméter értéke statisztikai értelemben nem nulla, akkor van valamilyen „hatása" a függő változóra. Itt fontos kihangsúlyozni, hogy a becsült paraméter numerikus értéke elképzelhető, hogy egy nem nulla érték, de statisztikai értelemben a becsült paraméter mégis nullának tekinthető, ha olyan ingadozással rendelkezik, amely nem teszi lehetővé, hogy ennél többet mondjunk. Bármilyen becslés során a becsült paraméter értéke sosem lesz pontosan nulla, ezért inkább az a kérdés, hogy milyen messze van a nullától. Ennek teszteléséhez viszont figyelembe kell venni azt, hogy a becsült paraméter értékét mintavételi hiba terheli, amelyet össze kell vetni a becsült paraméter értékével. A becsült paraméter standard hibája gyakorlatilag annak szórása, ezért a t-próba azt vizsgálja, hogy hány szórásnyi távolságra van a becsült paraméter a nullától. Amennyiben elég nagy távolságra, úgy elutasítjuk a nullhipotézist. Az, hogy mit tekintünk elég nagy távolságnak, az az alternatív hipotézistől és a választott szignifikancia szinttől függ. Fontos kiemelni, hogy a hipotézis vizsgálat a valós regressziós paraméter értékére vonatkozik (β), amely ismeretlen, nem pedig a becsült paraméter értékére ($\hat{\beta}$), mivel nem egy bizonyos mintára vonatkoztatjuk a tesztelést, hanem általánosságban szeretnénk következtetéseke levonni. Egyes esetekben azt szeretnénk vizsgálni, hogy több paraméter értéke együttesen különbözik-e nullától, ilyenkor F-próbát használunk. Ennek kiszámításához a maradék négyzetösszeget fogjuk használni, amely a következő módon van definiálva:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

ahol $\hat{y}_i$ a becsült függő változó értéke. Az OLS becslés minimalizálja az SSR értékét, ezért ha elhagyunk változókat, akkor az SSR értéke mindig nőni fog. Egy lehetséges teszt alapját adhatja ezért az, ha összehasonlítjuk azt a modellt, amelyre nem tettünk restrikciókat (ur, unrestricted) azzal, amelyből elhagytuk az adott változókat (r, restricted). Az ezekhez tartozó SSR értékeket jelöljük SSR_{ur} és SSR_r elnevezéssel. A fentiekből következik, hogy $SSR_r > SSR_{ur}$. Így az F-statisztikai alakja a következő:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

Ahol k a magyarázó változók száma az eredet modellben (tehát $(k+1)$ az összes becsült paraméter száma), q a restrikciók száma, míg n a minta nagyság. A fenti tesztstatisztika gyakorlatilag csak az SSR értékek közötti relatív növekedést számolja ki a két modell esetében. Ez az érték a gyakorlatban szinte mindig pozitív. Amennyiben a teszt azt mutatja, hogy a modell „teljesítménye" a restrikciók (tehát például bizonyos változók elhagyása) után sem változott jelentősen, akkor azt mondjuk, hogy a teszt statisztikai értelemben nem szignifikáns. A gyakorlatban könnyen előfordulhat az, hogy több változó paramétere egyénileg nem szignifikáns, de az együttes próba mégis azt mondja, hogy szükség van rájuk a modellben. Különösen akkor igaz ez, ha a restrikcióval érintett változók között magas a korreláció, amely a standard hibájuk növekedéséhez vezet.

Fontos kiemelni azt, hogy hibás gyakorlat „üldözni" a szignifikáns eredményeket, a modell építését mindig valamilyen elméletre kell alapozni. Másrészt különbséget kell tenni *elméleti* és *gyakorlati* szignifikancia között. A t-statisztika szignifikáns lehet abban az esetben ha a becsült paraméter értéke nagy, de akkor is, ha a standard hibája kicsi (ami nagy minta esetében gyakran teljesül). Ezért pusztán az, hogy a becsült paraméter statisztikai értelemben szignifikánsan különbözik nullától, még nem jelenti azt, hogy valóban fontos a hatása.

5.4. A modell illeszkedés mutatói

5.4.1. Hagyományos $\mathbf{R^2}$

A regressziós modellek legismertebb teljesítmény mutatója az R^2. Ennek a mutatónak az értelmezése rendkívül intuitív, könnyű kiszámolni és mindig 0 és 1 közötti értéket vesz fel. Részben ez okozta a rendkívüli népszerűségét, számos hibája ellenére. Az eddigiek alapján definiálhatunk három fajta négyzetösszeget (amelyből egy már ismert), amelyek hasznosak lesznek a regressziós egyenlet értelmezésében. Ez a Teljes négyzetösszeg (SST), a Magyarázott négyzetösszeg (SSE) és a Maradék négyzetösszeg (SSR):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

A teljes négyzetösszeget a magyarázott és a maradék négyzetösszeg összeadásával is megkapjuk, azaz *SST = SSE + SSR*. Ez abból fakad, hogy két komponensre bonthatjuk a magyarázó változó ingadozását: arra a hányadra, amire kontrollálunk (amit „magyarázunk") és arra, amire nem. Ezt a kapcsolatot végig osztva *SST* értékével és a következőt kapjuk:

$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSE}{SST}$$

$$1 = \frac{SSE}{SST} + \frac{SSE}{SST}$$

Majd ebből kifejezve:

$$\frac{SSE}{SST} = R^2 = 1 - \frac{SSE}{SST}$$

Ezt az értéket az OLS maximalizálja, mivel a becslés az *SSE* összeget minimalizálja. Ez a mutató a „magyarázott" ingadozás arányát mutatja a teljes ingadozáson belül. Mivel a magyarázott ingadozás nem lehet nagyobb a teljes ingadozásnál, ezért a mutató felső határa

egy. Így ha ebből kivonjuk a „nem magyarázott” hányadot, azaz a maradék ingadozás arányát, szintén megkapjuk az R^2 mutatót.

Az SSE legtöbbször csökken (vagy változatlan marad), amennyiben új magyarázó változót adunk a regressziós egyenlethez. Ezért az R^2 legtöbbször növekedni fog új változók bevonásával (legalábbis biztosan nem csökken). Ez egy nagyon kedvezőtlen tulajdonság, ugyanis teljesen irreleváns változók is bekerülhetnek a modellbe.

5.4.2. Korrigált $\mathbf{R^2}$

Az előbbiek korrekcióját elvégezve egy új mutatót kapunk, amely a korrigált R^2 néven ismert (adjusted R^2). Az úgynevezett szabadságfok korrekció következő módon alakítja a mutatót:

$$korrigált\ R^2 = 1 - \left[\frac{n-1}{n-k}\right](1 - R^2)$$

Ahol n a megfigyelések száma, míg k a magyarázó változóké. Ez a mutató korrigálva van a magyarázó változók számával, amely kompromisszumot keres az R^2 növekedése és a becsült paraméterek számának növekedése között. Ezzel kiküszöböli az R^2 hibáját. A mintaelemszám növekedésével az $\left[\frac{n-1}{n-k}\right]$ tag 1-hez közelít, a korrigált R^2 értéke pedig a hagyományos R^2 értékéhez. Rendkívül rossz illeszkedés esetében a korrigált R^2 értéke lehet negatív is! A legtöbb modell specifikációs mutató egyébként hasonló elven fog működni, azaz kompromisszumot keresnek majd a modell magyarázó ereje és a becsült paraméterek száma között.

Habár az R^2 a modell illeszkedését méri, jól specifikált modell esetében is lehet alacsony, és rosszul specifikált modell esetében is lehet magas (sőt, idősoros elemzésekben a rendkívül magas R^2 gyakran hibás specifikációra utal). Ezért nagyon rossz gyakorlat a modellezés során a legmagasabb R^2 mutatóval rendelkező modellt keresni!

5.5. Az OLS becslések varianciája

Érdemes levezetés nélkül felírni az OLS becslések varianciáját, amely a következő formát ölti a többváltozós lineáris regresszió esetében:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1 - R_j^2)}$$

Itt σ^2 a hibatag varianciáját jelenti, j az adott magyarázó változót jelenti, azaz $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ a j-ik változóhoz tartozó eltérés négyzetösszeg. Az utolsó tag több magyarázatot igényel. A $(1 - R_j^2)$ tag a j-ik magyarázó változónak a többi magyarázó változón és egy konstanson végzett regressziójából származó R^2 (tehát egy $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$ modell esetén, x_1 változó x_2 változón és egy konstanson történő regressziójából). Az egyenletből következik, hogy a becsült paraméterek varianciája az alábbi tényezők függvénye:

- **Az első komponens $\boldsymbol{\sigma^2}$.** Azaz, ha a hibatag varianciája magasabb (tehát nagyobb a függő változó ingadozásának az a hányada, amit nem tudtunk „megmagyarázni"), akkor a becslés is kevésbé lesz precíz. A maradéktag becslőfüggvénye az alábbiak szerint épül fel:

$$Var(\hat{\varepsilon}_i) = \sigma^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - (k+1)} = \frac{SSR}{n - (k+1)}$$

  A nevezőben lévő $n - (k + 1)$ tag neve szabadságfok (degrees of freedom, df), jelentése pedig egyszerű: a megfigyelések száma mínusz a becsült paraméterek száma (k magyarázó változó és egy konstans). Az első komponens csökkentésének egyetlen lehetséges módja, hogy több releváns magyarázó változót adunk a modellhez.
- **A második komponens $\boldsymbol{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$**, amely a j-ik magyarázó változó teljes ingadozását mutatja. Tehát ha a teljes ingadozás nagyobb, akkor a $Var(\hat{\beta}_j)$ hányados kisebb, azaz precízebb a becslés. Ennek az intuitív magyarázata az, hogy amennyiben nagyobb az ingadozás, úgy hatékonyabban tudjuk megfigyelni a függő és a magyarázó változó kapcsolatát. Abban az esetben lehet növelni ezt a tagot, ha növeljük a mintaelemszámot is. Ez a tag elméletben lehet zéró, ha x_{ij} konstans, viszont a modellfeltételek között azt a kikötést tesszük, hogy egy magyarázó változó sem lehet konstans.
- **A harmadik tag $\boldsymbol{(1 - R_j^2)}$** a kétváltozós lineáris regresszióban nem szerepel, mivel csak egy magyarázó változó van. A magas R_j^2 a magyarázó változók közötti korrelációt mutatja. Három eset releváns itt:
  1) Az $R_j^2 = 0$ eset a magyarázó változók közötti korreláció teljes hiányára utal, amely a gyakorlatban szinte sosem érvényesül.

2) Az $R_j^2 = 1$ a magyarázó változók tökéletes lineáris kombinációjára utal, amelyet a modellépítés feltételei zárnak ki.
3) Az $0 < R_j^2 < 1$ eset az, amellyel a gyakorlatban találkozunk. A magyarázó változók közötti korreláció növekedésével (a multikollinearitás megjelenésével) R_j^2 egyre közelebb kerül 1-hez, míg $(1 - R_j^2)$ zéróhoz. Így a magyarázó változók közötti magas korreláció, azaz a multikollinearitás, a becslés varianciájának növekedéséhez vezet.

Összeségében tehát két rendkívül fontos eleme van a modellépítés során használt adatoknak: mennyire precízen mérik az adott jelenséget és mennyire sok van belőlük. A jó minőségű és nagy mennyiségű minta ma már a regressziós modellezés során alapelvárás.

5.6. A klasszikus lineáris modell (Classical Linear Model, CLM) feltételek

Ahhoz, hogy a modellt „elfogadhatónak" ítéljük, teljesítenie kell néhány feltételt. Amennyiben ezek a feltételek megváltoznak, úgy egy újfajta becslési környezet áll elő, amelyben az OLS már nem feltétlenül lesz az optimális becslőfüggvény. A becslés során ezért ezeket a feltételeket ellenőrizzük, és amennyiben eltérést találunk, azt legjobb tudásunk szerint kezeljük.

A CLM feltételek

1) **CLM 1. A modell lineáris a paramétereiben. A függő változó y, a magyarázó változó x, és a hibatag ε az alábbi egyenlet szerint kapcsolódik egymáshoz**

$$y = \beta_0 + \beta_1 x + \varepsilon$$

2) **CLM 2. Random mintával rendelkezünk, amelynek nagysága n.**
3) **CLM 2. A multikollinearitás hiánya. Egy magyarázó változó sem konstans és nincs tökéletes lineáris kapcsolat a magyarázó változók között.**
4) **CLM 4. A hibatag magyarázó változók szerinti feltételes várható értéke 0.**

$$E(\varepsilon|x_1, x_2, \dots x_k) = 0$$

5) **CLM 5. Homoszkedaszticitás. A hibatag varianciája a magyarázó változók értékétől független.**

$$Var(\varepsilon|x_1, x_2, \dots x_k) = \sigma^2$$

6) **CLM 6. Normalitás*. A hibatag ε független a magyarázó változóktól, és normális eloszlást követ 0 várható értékkel és σ^2varianciával.**

$$\varepsilon \sim N(0, \sigma^2)$$

Az első öt feltételt Gauss – Markov feltételeknek is nevezzük. A 6) számú feltétel nem kritikus, mivel megfelelően nagy minta esetében elhagyható (de kis minta esetében fontos, ami már egyre ritkább gazdasági, pénzügyi területen).

Ezek a feltételek rendkívül fontosak két tényező igazolásához. Egyrészt, szeretnénk, hogy a becslőfüggvény torzítatlan legyen. Mivel számos torzítatlan becslőfüggvényt lehet konstruálni, ezért szeretnénk továbbá, hogy a becslőfüggvény varianciája is a legkisebb legyen (azaz legyen a leghatásosabb). Az 1-4. feltételek teljesítése esetén torzítatlan a becslőfüggvény. Az 1-5. feltétel teljesítése esetén az OLS becslés rendelkezik a legkisebb varianciával a torzítatlan becslőfüggvények között, azaz a Legjobb Lineáris Torzítatlan Becslés. Ezt gyakran az angol BLUE mozaikszóval jellemezzük (Best Linear Unbiased Estimator, BLUE).

A következőkben kimondjuk a Gauss – Markov tételt.

Gauss – Markov tétel

A CLM 1. – CLM 5. feltételek teljesülése esetén a $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}, \dots \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ OLS becslések adják a $\boldsymbol{\beta}_{\boldsymbol{0}}, \boldsymbol{\beta}_{\boldsymbol{1}}, \dots \boldsymbol{\beta}_{\boldsymbol{k}}$ populációs paraméterek legjobb lineáris torzítatlan becsléseit (BLUE).

Ez a tétel azt mondja, hogy a lineáris és torzítatlan becslőfüggvények körében, az OLS becslés varianciája a legkisebb. Ez a tétel rendkívül erős állítást fogalmaz meg, hiszen azt mondja, hogy ha ezek a feltételek teljesülnek, akkor egy becslőfüggvény sem fog jobban teljesíteni, mint az OLS.

5.7. Hogyan lehet megsérteni a Gauss-Markov feltételeket?

1. feltétel: A regressziós egyenlet alapján három főbb esetet írhatunk fel az 1. feltétel megsértésére. Az egyenlet leírja a változók közötti feltételezett kapcsolatot, ezért a releváns változók kihagyása vagy a felesleges változók bevonása sérti az első feltételt. Ugyanez történik, ha nem lineáris a valódi kapcsolat a változók között. Végül feltételeztük, hogy a paraméterek konstansok, azaz minden megfigyelés esetében ugyanaz β_0 és β_1 értéke. Amennyiben a paraméterek változnak az adatgyűjtés ideje alatt, úgy ez a feltétel is sérül.

A releváns változók kihagyása a becsült paraméterek torzított becsléséhez vezet. Ezt csak abban az esetben lehet kikerülni, amennyiben a kihagyott változó megfigyelései nem korrelálnak a

modell változóinak megfigyeléseivel (ami a gyakorlatban rendkívül ritka). Ebben az esetben a meredekség paraméterei torzítatlanok lesznek (a konstans viszont így is torzított marad, kivéve akkor, ha a kihagyott változó átlaga zéró). Tekintsünk erre egy példát, amely során jelöljük a valódi modellt a következők szerint:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Feltételezzük, hogy x_2 változóra nincsenek adataink, ezért jobb híján a következő egyenletet becsüljük meg:

$$\hat{y} = \beta_0 + \beta_1 \hat{x}_1 + \underbrace{(\beta_2 x_2 + \varepsilon)}_{v}$$

Ebben a modellben v egy összetett hibatag, amely az eredeti ε hibatagon kívül magába foglalja a kihagyott változó hatását is. Így viszont már az alábbi eset áll fenn:

$$E(v|\hat{x}_1) \neq 0$$

Ennek oka, hogy a magyarázó változók legtöbbször korrelálnak egymással. Abban az esetben marad a β_1 paraméter torzítatlan, ha a két magyarázó változó közötti korreláció zéró. Viszont ebben az esetben is fennáll, hogy $E(v) \neq 0$ (kivéve ha x_2 átlaga zéró). A β_1 paraméter varianciája viszont még a korrelálatlan magyarázó változók esetén is felfelé torzított lesz, amely az inferencia során okoz problémát. Valamivel kevesebb problémát okoz a felesleges változó bevonása, de a becslés kevésbé lesz hatékony (kivéve, ha szintén korrelálatlanok a magyarázó változók).

Ezt a problémát csak úgy lehet kiküszöbölni, ha a modell építés során a közgazdasági elméletre hagyatkozunk, és azokat a változókat vonjuk be a modellbe, amelyekre az elmélet szerint szükségünk van. Természetesen a valóságban ez legtöbbször közel sem ilyen egyszerű.

A nemlineáris kapcsolatok kezelésének hatékony módja lehet a különböző (például logaritmikus) transzformációk használata, vagy az ha nem lineáris modelleket használunk. Ezeknek a modelleknek viszont gyakran nehezebb a becslése, a paraméterek interpretációja és az OLS számos kedvező tulajdonságát sem viszik tovább.

2. feltétel: Amennyiben a minta nem random, hanem valamilyen módon szisztematikusan torzított, úgy sérül a mögöttes matematikai levezetés is, amely ezen a feltételen alapszik. Gyakorlati szempontból azért jelent ez problémát, mivel a minta torzított lesz bizonyos értékek irányába. Például ha egy kérdőíves adatgyűjtést csak az ismerőseink körében végzünk, úgy nagyobb eséllyel kerülnek be olyan személyek a mintába, akik hasonló szocio-demográfiai háttérrel rendelkeznek, mint mi (mivel valószínűleg azonos helyen élnek, hasonló oktatásban vettek részt stb.).

3. feltétel: Amennyiben az adatsorban megfelelő ingadozás van, úgy hatékonyabban lehet „lekövetni" a függő és a magyarázó változó közötti kapcsolatot. Amennyiben elképzelünk egy konstans értékű magyarázó változót, úgy könnyen látható, hogy bármilyen y érték mellett x értéke ugyanaz lesz! Így tehát semmilyen hasznos információt nem közöl a modellező számára. Matematikai oka is van annak, hogy miért nem lehet konstans változót használni. A $\hat{\beta}_1$ becslőfüggvényét az alábbiak szerint írtuk fel:

$$\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_i)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

A $\sum_{i=1}^{n}(x_i - \bar{x})$ tag érdemel ez esetben külön figyelmet, mivel ennek a tagnak mindig nulla az összege. Ez egyszerűen bizonyítható:

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

Az első tag $\sum_{i=1}^{n} x_i$ átírható $n\bar{x}$ formába. A második tag, $\sum_{i=1}^{n} \bar{x}$ pedig csak a számtani átlag (amely értéke konstans minden megfigyelésre) összegzése n esetben, azaz szintén $n\bar{x}$. Így a következőre jutunk:

$$n\bar{x} - n\bar{x} = 0$$

Azaz konstans x esetében nem csak $\sum_{i=1}^{n}(x_i - \bar{x})$ tag, de a $\sum_{i=1}^{n}(x_i - \bar{x})^2$ tag is nulla lesz, így $\hat{\beta}_1$ nem becsülhető. Azaz, hogy ha x konstans, akkor a varianciája (és így a szórása) is nulla kell, hogy legyen. Ez a feltétel tehát pusztán azért is szükséges, hogy $\hat{\beta}_1$ értékét kiszámíthassuk.

Érdemes megemlíteni egy kapcsolódó problémát is, mivel hasonló helyzet alakul ki akkor, ha kétértékű változókat helytelenül használunk. Például ha egy bináris változóval jelöljük a válaszadók nemét, ahol két változót kreálunk, jelöljük őket x_1 és x_2 elnevezéssel:

$$x_1 = \begin{cases} 0, & ha\ x = nő \\ 1, & ha\ x = férfi \end{cases}$$

Illetve:

$$x_2 = \begin{cases} 0, & ha\ x = férfi \\ 1, & ha\ x = nő \end{cases}$$

Amennyiben mind a két változót szeretnénk szerepeltetni, akkor x_1 és x_2 multikollinearitást okoz a modellben, mivel $x_1 + x_2 = 1$ és $x_2 = 1 - x_1$ azaz egymás tökéletes lineáris kombinációi. Úgy is mondhatjuk, hogy a két változó ugyanazt az információt közvetíti, így az egyik redundáns. Ezt a jelenséget ’dummy-csapda’ néven is illetik. Ezért egy kategória mindig a referencia szintként fog szolgálni, amelyhez a többi szintet viszonyítjuk. Az ökonometriai szoftverek ezt a problémát könnyen észlelik és legtöbbször egy hibaüzenet kíséretében az egyik változó szint elhagyásával történik meg a becslés.

A gyakorlatban nem feltétlenül a tökéletes lineáris kombináció, hanem a magyarázó változók közötti magas korreláció okoz problémát (ami egy közelítése a tökéletes lineáris kombinációnak). Ezért valójában ez a probléma nem multikollinearitás, de a továbbiakban a szoros kapcsolódás miatt így utalunk rá továbbra is. A multikollinearitás esetén az OLS becslés BLUE marad, azaz nem csak hogy torzítatlan, de továbbra is a legkisebb varianciával rendelkező. Viszont a paraméterek OLS becslésének variancia viszonylag nagy lesz azoknak a változóknak az esetében, amelyeket érint a multikollinearitás. Ez kevésbé precíz becsléseket és a hipotézis vizsgálatok alacsonyabb erejét eredményezi. Éppen ezért a multikollinearitás kimutatására az úgynevezett Variancia Inflációs Mutató (Variance Inflation Factor, VIF) használható:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Az R_j^2 a magyarázó változók egymáson történő regressziójából származó R^2 a j változó esetében. A VIF_j pontosan az a faktor, amennyivel a magyarázó változók korrelációja miatt a β varianciája nő a multikollinearitás teljes hiányához képest. Problémát jelent az, hogy nem tudjuk, hogy mekkora korreláció számít túl magasnak két változó között, hiszen ezekre nincsenek egzakt szabályok. Általában egy önkényesen megválasztott küszöbhatár fölött feltételezzük problémásnak a multikollinearitás jelenlétét, amely a gyakorlatban 2 vagy 5 szokott lenni. A becsült paraméterek varianciája viszont nem csak a magyarázó változók közötti korrelációtól függ, így a VIF_j szerepe inkább indikatív. Mivel a VIF_j egy szorzófaktor, ezért a minimum értéke 1. Az egyik legegyszerűbb megoldás a multikollinearitás kezelésére az, ha elhagyjuk azt a változót, amelyik a problémát okozta (mivel definíció szerint létezik a modellben egy másik változó, amivel erős korrelációt mutat, ezért a modell magyarázó erejét várhatóan nem igazán befolyásolja). Olyan eset is előfordulhat, amikor számos változó segítségével mérjük ugyanazt, vagy hasonló hatásokat. Például a környezetbarát viselkedés felmérésekor megkérdezésre kerül, hogy az adott válaszadó mennyire tartja fontosnak a szelektív hulladékgyűjtést, az energiatakarékosságot és így tovább. Ilyen esetben várható, hogy az egyes tényezőkre adott értékelés szoros összefüggést fog mutatni, mivel egy mögöttes, úgynevezett látens változót követnek, és egy modellbe vonva multikollinearitás lép fel közöttük. Ilyenkor az is megoldás lehet, ha megfelelő módszerek segítségével egy kombinált változót készítünk, amely jól reprezentálja a mögöttes látens változót.

Fontos megemlíteni, hogy a magyarázó változók közötti magas korreláció nem sérti egyik modell feltételt sem, ezért nem is jól definiált a probléma, mivel a becslések varianciájára kifejtett hatása más tényezőktől is függ.

4. feltétel. Ennek a feltételnek a megsértése komoly következményekkel jár, mivel a paraméter becslések értéke torzított lesz. A $E(\varepsilon|x) = 0$ feltétel azt mondja, hogy x realizált értékétől függetlenül, a hibatag értéke átlagosan zéró marad. Azaz, gyakorlatilag nincs korreláció a hibatag értéke és a magyarázó változók között. A $E(\varepsilon|x) = 0$ kitételből következik, hogy $E(\varepsilon) = 0$ és $cov(x, \varepsilon) = 0$.

5. feltétel. Ez a feltétel azt mondja, hogy a hibatag varianciája nem függ x realizált értékeitől, azaz érték konstans minden megfigyelésre. Ilyenkor azt mondjuk, hogy a hibatag homoszkedasztikus (konstans szórással rendelkező folyamat). Amennyiben ez nem áll fenn, a heteroszkedaszticitás jelenségével szembesülünk.

Amennyiben a modell heteroszkedasztikus, az OLS becslés még mindig torzítatlan marad, de a becslések varianciája még a mintaelemszám növekedése mellett is torzított lesz. Ez azt jelenti, hogy a konfidencia intervallum és a hipotézis vizsgálat eredménye is helytelen. Ezen felül a becslés nem lesz továbbá hatásos, így a BLUE tulajdonság sem teljesül, tehát lesz más, olyan lineáris és torzítatlan becslőfüggvény, amelynek kisebb a varianciája. A probléma megoldása viszonylag egyszerűen orvosolható úgynevezett robosztus standard hibák számításával. Ez manapság olyan gyakori, hogy minden tudományos folyóirat a robosztus standard hibákat közli.

5.8. Egyéb témakörök

5.8.1. Transzformációk

Bizonyos esetekben érdekes transzformálni a függő és a magyarázó változókat. Mivel a közgazdaságtani és a pénzügyi területen a logaritmikus transzformáció nem csak hogy elterjedt, de gyakran a kizárólagos transzformáció, ezért csak ezt a típust részletezzük itt.

A logaritmikus transzformáció esetén a természetes alapú logaritmust használjuk, amelynek az Euler féle szám, $e = 2.71 \ldots$, az alapja, amelynek számos tulajdonsága van. A modellezés során több féle indoklással találkozunk a szakirodalomban. Gyakori az, hogy a logaritmikus transzformáció erősíti a normalitás feltételeit. Habár ez bizonyos típusú adatok esetében igaz, de azon a téves feltételezésen alapszik, hogy a regresszió függő és magyarázó változóinak normális eloszlásúnak kell lennie. Ez nem igaz, hiszen a Gauss – Markov feltételek között nem szerepel ilyen, az OLS becslőfüggvény ennek ellenére is BLUE. A linearitás feltételét viszont erősítheti, mivel a logaritmikus transzformáció egyik előnyös tulajdonsága, a következő:

$$\log(ab) = \log(a) + \log(b)$$

illetve

$$\log(a^b) = b\log(a)$$

Ezen felül a logaritmikus transzformáció „összehúzza” a számokat, így az adatok szóródását csökkenti, a jelentős numerikus különbségek eltüntetése pedig gyakran algoritmikus szempontból is kedvező. A regressziós modell becsült paramétereinek interpretációja annak függvényében változik, hogy csak a függő, csak a magyarázó, vagy mind a két oldal logaritmizált. Ebben az esetben négy típus különböztetünk meg:

Level – Level modell

Ez a modell a hagyományos regressziót jelenti, ahol y és x transzformáció nélkül lépnek a modellbe. A becsült modell a következő:

$$y = \beta_0 + \beta_1 x$$

A paraméter interpretáció ebben az esetben értelemszerűen ugyanaz marad.

Log – Level modell

Ebben a modellben csak y van logaritmizálva.

$$\log(y) = \beta_0 + \beta_1 x$$

Viszonylag kis változások esetén a $100 * \beta_1$ jó közelítése a százalékos változásnak, de a pontos kapcsolatot a $100 * (e^{\beta_1} - 1)$ adja meg. Az alábbi felírás segít a megértésben, ahol y^* arra a modellre vonatkozik, ahol a magyarázó változó egységnyi értékkel nagyobb:

$$\log(y^*) = \beta_0 + \beta_1(x + 1) = \underbrace{\beta_0 + \beta_1 x}_{\log(y)} + \beta_1 = \beta_1$$

Azaz:

$$\log(y^*) - \log(y) = \beta_1$$

Ebből következik, hogy:

$$\frac{y^*}{y} = e^{\beta_1}$$

Végül mind a két oldalt szorozva 100-al és kivonva 1-et a következőt kapjuk:

$$100 * \left(\frac{y^*}{y} - 1\right) = 100 * (e^{\beta_1} - 1)$$

A bal oldalon lévő tag csupán a százalékos változást mutatja y^* és y^* között. Így a $100 * (e^{\beta_1} - 1)$ megadja y százalékos változását abban az esetben, ha egységnyi értékkel nagyobb x értéket veszünk alapul.

Level – Log modell

Ebben a modellben csak x van logaritmizálva (bármelyik, vagy akár az összes magyarázó változó logaritmusát is vehetjük a megfelelő feltételek teljesülése esetén).

$$y = \beta_0 + \beta_1 \log(x)$$

Az értelmezés szerint x 1%-os változása a függő változó értékének $\beta_1/100$ egységnyi változásával jár együtt átlagosan, minden mást változatlannak tekintve.

Log – Log modell

Ebben a modellben y és x is logaritmizálva van.

$$\log(y) = \beta_0 + \beta_1 \log(x)$$

A paraméter β_1 ebben az esetben elaszticitást mér. Azaz x 1%-os változása a függő változó β_1%-os változásával jár együtt, minden mást változatlannak tekintve.

A logaritmikus transzformáció segíthet olyan esetekben, amikor a modell nem-lineáris formában van felírva. A közgazdaságtanban jól ismert Cobb – Douglas függvény formája például a következő:

$$Y = c * K^{\alpha} * L^{\beta}$$

Ahol Y a kibocsátás, c konstans, K és L a termelési tényezők mennyiségét jelzi, míg α és β értéke 0 és 1 között van. Ez egy nem-lineáris modell, de elvégezhető az alábbi transzformáció:

$$\log(Y) = \log(c * K^{\alpha} * L^{\beta})$$

Amely az alábbi formára írható át:

$$\log(Y) = \log(c) + \alpha \log(K) + \beta \log(L)$$

Ez a forma már a jól ismert lineáris formát mutatja. A logaritmikus transzformáció fő hátránya, hogy nem értelmezett a 0 és a negatív értékek tartományában. Ebben az esetben gyakori a

$\log(1+Y)$ vagy a $\log(Y+A)$ transzformáció, ahol A adott értékű konstans szám. Ezek célja, hogy a megfelelő értékek segítségével az adatokat eltolják a pozitív tartományba, de egyúttal számos egyéb problémát idéznek elő, így használatuk nem minden esetben ajánlott[1].

5.8.2. Mérési hiba a változókban

Egyes esetekben a függő vagy magyarázó változók mérési hibákkal terheltek. Tegyük fel, hogy a valódi kapcsolat felírható az alábbi módon:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Viszont x értékét hiba mellett mérjük, azaz az általunk rögzített változó az $x^* = x + u$, ahol u a 0 várható értékű mérési hiba. Ebben az esetben $x = x^* - u$, és y és x kapcsolata az alábbiak szerint írható fel:

$$y = \beta_0 + \beta_1(x^* - u) + \varepsilon$$

azaz

$$y = \beta_0 + \beta_1 x^* + (\varepsilon - \beta_1 u)$$

Mivel u a mérési hibával terhelt magyarázó változó és az új hibatag része is, ezért a magyarázó változó és a hiba korreál egymással. Ennek eredményeként meg lehet mutatni, hogy a mérési hiba β_1 értékét a zéró felé torzítja átlagosan, a becslés pedig egyúttal nem konzisztens. Hozzá kell tenni, hogy a többváltozós esetben ezek a következtetések nem ilyen egyértelműek. Amennyiben y mérését terheli hiba (ahol azt a feltételt tesszük, hogy nem korrelál a magyarázó változókkal), úgy a hibatag és az OLS becslőfüggvények varianciája nagyobb lesz, de a becslőfüggvények torzítatlanok és konzisztensek, a hagyományos t és F statisztikák pedig érvényesek maradnak. A gyakorlatban nehéz megmondani, hogy mely változót mekkora mérési hiba terheli, egyúttal a torzítás mérete és iránya is nehezen meghatározható. Az alapvető megoldás ebben az esetben a jobb minőségű adatok használata.

5.9. Gyakorlati példa

5.9.1. A változók jellemzése

A regressziós modellezés bemutatásához egy 1000 vállalkozásból álló (generált) adatbázist használunk, a következő változókkal:

[1] Például CHEN – ROTH (2024)

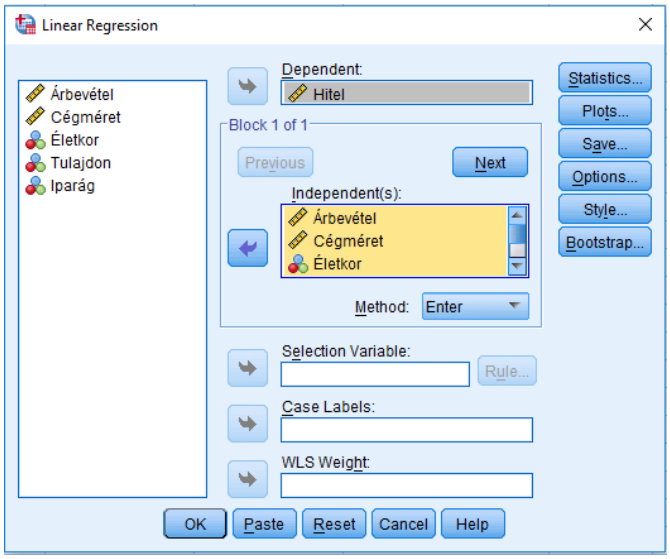
Függő változó

- **Hitel:** A hitelfelvétel nagysága, millió HUF

Magyarázó változók

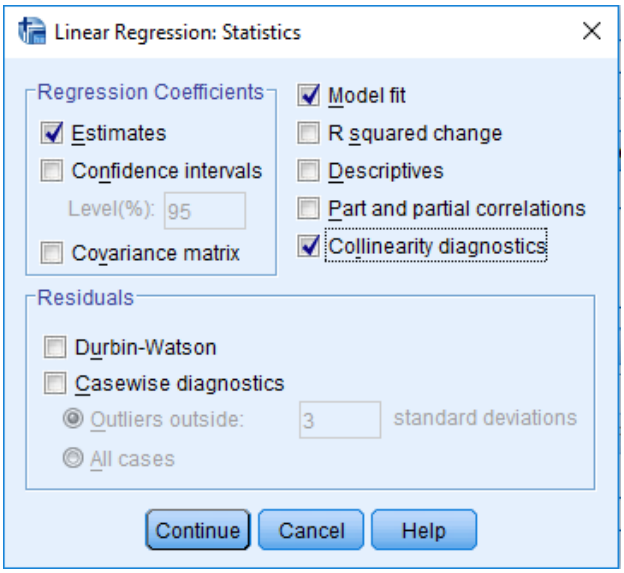
- **Árbevétel:** A cég nettó árbevétele, millió HUF
- **Cégméret:** a beosztottak száma, fő
- **Életkor:** a cég életkora, év
- **Tulajdoni státusz:** bináris változó, amely 1, ha külföldi és 0, ha hazai cég
- **Iparág:** 1, ha magas növekedésű iparágban működik a vállalkozás, 0, ha más jellegű iparágban

Az SPSS programban az regressziós modellezés elérhető az *Analyze – Regression – Linear* funkciókkal. A ’Dependent’ rész a függő változóra utal, míg az ’Independent(s)’ a magyarázó változókra.



5.1. ábra: **A függő és független változók a regresszió elemzésben**

Szeretnénk látni a becsléseket (’Estimates’), a modell illeszkedését (’Model fit’) és a VIF értékeket (’Collinearity diagnostics’).



5.2. ábra: **A „Linear Regression: Statistics" menü beállításai**

A következőkben az SPSS becslés eredményeit mutatjuk be. Az első táblázat mindössze azt mutatja, hogy minden változót egyszerre illesztettünk a modellben (Enter módszer). Elképzelhető az, amikor lépésenként (Stepwise módszer) választunk ki változókat egyéni szignifikancia alapján, ez viszont könnyen belátható, hogy rossz gyakorlat (ennek ellenére több programban szerepel).

5.9.2. SPSS becslés eredménye

5.1. táblázat: **Az output eredményekben közölt függő és független változók**

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Iparág, Méret, Árbevétel, Tulajdon, Életkor[b]	.	Enter

a. Dependent Variable: Hitel

b. All requested variables entered.

Az R^2 megmutatja, hogy a magyarázó változók ingadozása a függő változó ingadozásának közel 70%-át magyarázták. Az „adjusted" a korrigált R^2 értékét mutatja. Mivel ebben a modellben nincs irreleváns magyarázó változó (hiszen az adatokat szimuláltuk, ezért tudjuk a „valós" modellt), a két mutató értéke között nincs jelentős különbség. Az alábbiak szerint ki is számolható az értéke:

$$korrigált\ R^2 = 1 - \left[\frac{1000-1}{1000-5}\right](1-0.691) = 0.689$$

Az R a változók közötti többváltozós korreláció, míg a ’Std. Error of the Estimate’ pedig a maradéktag szórását mutatja.

5.2. táblázat: **Az output eredményekben közölt modell illeszkedési mutatók**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,831[a]	,691	,689	12,878

a. Predictors: (Constant), Iparág, Méret, Árbevétel, Tulajdon, Életkor

A regressziós modell ANOVA táblája egy F-statisztika eredményét mutatja meg, amely a paraméterek együttes szignifikanciáját teszteli. A nullhipotézis itt azt feltételezi, hogy a magyarázó változók hatása nulla. Azaz nincs változó, amelynek hatása szignifikánsan különbözik nullától. Ez azt jelentené, hogy a regressziós egyenes nem létezik. Ennek elutasítása pedig azt mutatja meg, hogy van legalább egy olyan változó, amely hatása szignifikánsan különbözik nullától. Elképzelhető, hogy csak egy változónak van nullától szignifikánsan különböző hatása, de az is lehet, hogy az összesre igaz ez a megállapítás. Ezt a későbbiekben a változók egyéni szignifikancia próbája mutatja meg. Az F-statisztika értékét a következő módon számíthatjuk ki:

$$F = \frac{73715{,}922}{165{,}840} = 444{,}501$$

A Sum of Squares oszlop a különböző négyzetösszeg értékekre utal. A Regression sor a Magyarázott négyzetösszeg (SSE), míg a Residual sor a Maradék négyzetösszeg (SSR) értékét mutatja. Ezek alapján:

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 368579{,}611$$

$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = 164844{,}724$$

A teljes négyzetösszeg (SST) pedig a kettő összege:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSE + SSR = 533424{,}335$$

A Mean Square értékek az adott sorhoz tartozó négyzetösszeg értékek elosztva a szabadságfokkal (df). Ebben az értelemben mind a két érték variancia becslésnek számít, amely magyarázatot ad az F- eloszlásra. Tehát:

$$Regression\ Mean\ Square = \frac{368579{,}611}{5} = 73715{,}922$$
$$Residual\ Mean\ Square = \frac{164844{,}724}{994} = 165{,}840$$

A gyakorlatban az utolsó oszlop a legfontosabb, amely az F-értékhez tartozó p-értéket mutatja meg. Mivel ez az érték kisebb, mint a választott szignifikancia szint (legtöbb esetben 0,05), így elutasítjuk a nullhipotézist. Azaz, van legalább egy változó, amely hatása szignifikánsan különbözik nullától. Ez azt jelenti, hogy áttérhetünk a becsült paraméterek értelmezésére.

5.3. táblázat: **A regressziós modell ANOVA eredményei**

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	368579,611	5	73715,922	444,501	,000[b]
	Residual	164844,724	994	165,840		
	Total	533424,335	999			

a. Dependent Variable: Hitel

b. Predictors: (Constant), Iparág, Méret, Árbevétel, Tulajdon, Életkor

Az alábbiakban a B oszlopra fogunk koncentrálni, amely a változóhoz tartozó becsült paramétereket mutatja meg. Az egyes változók paraméterét az alábbiak szerint értelmezhetjük:

- **(Constant):** Ez a konstans paraméter. Ez a érték a hitelfelvétel átlagos értékét mutatja meg, abban az esetben, ha minden magyarázó változó értékét nullára állítjuk. Ebben a modellben a konstans értéke azt mutatná meg, hogy mi volt az átlagos hitelfelvétel mértéke egy olyan cég esetében, amelynek nincs árbevétele, dolgozója, nulla éves, hazai tulajdonban van és nem gyorsan növekedő iparágban tevékenykedik. A legtöbb modellezési probléma során nem értelmezhető a konstans értéke (ahogy itt sem), így habár matematikai szempontból szükséges, az értelmezésével nem foglalkozunk.

- **Árbevétel:** Az árbevétel becsült paramétere negatív. Ez arra utal, hogy két olyan vállalkozást összehasonlítva, amelyek árbevétele egységnyi értékkel különbözik, de egyébként minden más szempontból azonosak, azon vállalkozás esetében, ahol az árbevétel egységnyi értékkel magasabb, ott az átlagos hitelfelvétel értéke -0,411 millió HUF értékkel kisebb volt. A körülményes megfogalmazást kicsit leegyszerűsíthetjük, és mondhatjuk azt, hogy egyforma méretű, életkorú, tulajdoni státusszal rendelkező vállalkozások esetében, amelyek azonos iparágban tevékenykednek, az 1 millió HUF értékkel (egységnyi értékkel) magasabb árbevétel átlagosan körülbelül 400 ezer HUF értékkel kevesebb hitelfelvétellel járt együtt. Ezt skálázhatjuk is, mivel lineáris a hatás, azaz azt is mondhatjuk, hogy a 10 millió HUF értékkel magasabb árbevétel átlagosan 4 millió HUF értékkel alacsonyabb hitelfelvétellel járt együtt, minden mást változatlannak tekintve. Mit jelent ez a gyakorlatban? Elképzelhető, hogy azok a vállalkozások, amelyek árbevétele magasabb, nem szorulnak hitelfelvételre, ezért a negatív kapcsolat.
- **Méret:** A cégméretet a dolgozók számával közelítettük. Hasonló értelmezés szerint, amennyiben a dolgozók száma 1 fővel magasabb volt, úgy az átlagos hitelfelvétel körülbelül 100 ezer HUF értékkel volt alacsonyabb, minden mást változatlannak tekintve. Azaz két céget összehasonlítva, amelyek dolgozói létszáma egy fővel különbözik, egyébként minden másban azonosak, ott az egy fővel magasabb dolgozói létszámmal rendelkező vállalkozás átlagos hitelfelvétele 100 ezer HUF értékkel alacsonyabb volt. A „minden mást változatlannak” tekintve arra utal, hogy a méreten kívül minden egyéb tulajdonságban azonos vállalkozásokat vizsgálunk, azaz csak a méretváltozás parciális hatását vizsgáljuk. A negatív kapcsolat itt szintén arra utalhat, hogy a nagyobb vállalkozások tőkeerősebbek lehetnek, így ritkábban szorulnak hitelfelvételre.
- **Életkor:** Habár numerikusan ez az érték nem nulla, de statisztikai értelemben igen, mivel a változóhoz tartozó egyéni szignifikancia 0,784. Tehát tekinthetjük úgy, hogy *ebben a mintában,* az életkor változásának nincs hatása a hitelfelvétel változására.
- **Tulajdon és iparág:** Mind a két változó bináris (dummy, kétértékű) változó, így az értelmezés változik. A paraméterek alapján a külföldi tulajdonban lévő vállalkozások átlagos hitelfelvétele több mint 12 millió HUF értékkel kevesebb volt, mint a hazai vállalkozások értéke (referencia kategória), míg azon vállalkozások átlagos hitelfelvétele, amelyek gyors növekedéssel jellemezhető iparágban tevékenykedtek, több mint 25 millió HUF értékkel volt magasabb (azokhoz képest, akik nem ebben az iparában tevékenykednek). Ez arra utal, hogy a külföldi tulajdonú vállalkozások

hiteligénye átlagosan alacsonyabb volt. Az iparág esetében elképzelhető, hogy azok a vállalkozások, amelyek gyorsan fejlődő iparágban tevékenykednek, nem tudják a működési és fejlesztési kiadásaikat finanszírozni, ezért több hitel felvételére kényszerültek. Ilyen eset lehet például a Startup vállalkozások helyzete: rendkívül ígéretes iparágban tevékenykednek, gyors fejlődési rátával, de mivel az életciklusuk első szakaszában vannak, ezért számos fejlesztést csak külső finanszírozással tudnak megvalósítani.

5.4. táblázat: **A regressziós modell paramétereinek illeszkedési mutatói**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	**(Constant)**	306,309	7,011		43,689	,000		
	Árbevétel	-,411	,014	-,529	-29,950	,000	,997	1,003
	Méret	-,101	,006	-,305	-17,291	,000	,996	1,004
	Életkor	-,025	,090	-,005	-,274	,784	,995	1,005
	Tulajdon	-12,393	,817	-,268	-15,177	,000	,995	1,005
	Iparág	24,574	,817	,531	30,087	,000	,996	1,004

a. Dependent Variable: Hitel

A változókhoz tartozó t és p-érték azt mutatja meg, hogy az adott változó hatása szignifikánsan különbözik nullától. A t-érték kiszámítása egyszerű, hiszen csak a becsült paraméter és az ahhoz tartozó standard hiba hányadosa. Például a **Tulajdon** esetében (kisebb eltérések pusztán kerekítési hibák):

$$t_{Tulajdon} = \frac{-12{,}393}{0{,}817} = -15{,}177$$

Végül az utolsó oszlop a VIF értékeket tartalmazza, míg a Tolerance az $1/VIF$ értéket jelenti. Hogyan kapjuk meg például a **Méret** VIF értékét?

- Futtatunk egy regressziót, ahol a Méret lesz a függő változó míg a fennmaradó változók (kivéve a Hitelfelvételt) a magyarázó változók. Ebből a regresszióból elmentjük az R^2 értékét (amely ebben az esetben R^2 = 0,003847).
- Kiszámítjuk a VIF értékét az alábbiak szerint:

$$VIF_{Méret} = \frac{1}{1 - R^2_{Méret}} = \frac{1}{1 - 0{,}003847} = 1{,}004$$

Természetesen eltérő módszerekkel is ki lehet számítani a VIF értékét, de ebben az esetben jól szemléltethető, hogy amennyiben magasabb a korreláció a magyarázó változók között (és ez megmutatkozik a magasabb R^2 értékben), úgy a VIF értéke is növekszik.

Az SPSS esetében körülményes kiszámítani a heteroszkedaszticitás robosztus standard hibákat, ezért itt csak feltüntetjük őket a hagyományos standard hibák mellett. Ebben az esetben nincs jelentős eltérés a kettő között, azonban elég súlyos heteroszkedaszticitás esetén az eltérés drasztikus lehet. A robosztus standard hibák a legtöbb esetben valamivel nagyobbak mint a hagyományos OLS standard hibák, amely egy konzervatívabb hipotézis tesztelési folyamathoz vezet, azaz több bizonyítékra lesz szükségünk a nullhipotézis elutasításához.

5.5. táblázat: **A hagyományos és a robosztus standard hibák közötti különbség**

Változók neve	B	Hagyományos standard hibák	Robosztus standard hibák
(Constant)	306,309	7,011	6,708
Árbevétel	-,411	,014	,013
Méret	-,101	,006	,006
Életkor	-,025	,090	,085
Tulajdon	-12,393	,817	,820
Iparág	24,574	,817	,820

6. Az idősoros ökonometria alapjai

6.1. Az idősor, mint modellezési probléma

Az idősoros adatok jellemzője, hogy a megfigyelések ugyanarra az egységre vonatkoznak, de különböző időpontokban, ezért sorrendjük kötött. Gazdasági és pénzügyi területen jellemző a GDP, a részvényárak, a munkanélküliség vagy az infláció alakulásának vizsgálata. Minden esetben egy egységet vizsgálunk (pl. Microsoft részvényárak), de különböző időpontokban mérve (pl. napi gyakorisággal). Az idősoros adatokat két tényezővel jelölhetjük, amely a változó elnevezéséből és a időszakot jelző indexből tevődik össze:

$$y_t, t = 1, 2, \dots . . T$$

Itt y a megfigyelést jelenti, míg t a „time", azaz idő rövidítése, T időszakig terjed, amely az utolsó időszakot jelenti. Tehát az idősor olyan megfigyelések együttese, ahol minden megfigyelést egy adott időponttal jelölünk. Az idősoros ökonometria során arra keressük a választ, hogy hogyan viselkednek az adatok időben egyénileg vagy rendszerben tekintve. Ebben a fejezetben nem térünk ki a klasszikus trend és szezonalitás felbontásra, de a szezonalitás (és annak szűrése), a kiugró értékek és törések kezelése fontos része az idősoros szakirodalomnak. A fejezetben $E(.)$ a várható érték operátor, $Var(.)$ a variancia operátor, $Cov(.)$ pedig a kovarianciát jelenti.

6.2. Stacionaritás

Az idősorok különbözőségét részben az adja, hogy milyen időbeli „stabilitást" mutatnak. Az adatok időbeli „stabilitása" alapján két típusú viselkedést különböztethetünk meg. Az úgynevezett stacioner folyamat időben stabil viselkedést mutat, míg a nem-stacioner folyamat esetén hiányzik ez a stabilitás. Mit is jelent a stabilitás ebben az esetben? Amennyiben az idősor átlaga és autokovarianciája nem függ időtől, úgy y_t folyamatot gyengén stacioner folyamatnak nevezzük. Ezek megkülönböztetése és a modellezés során történő kezelése azért fontos, mert a nem-stacioner folyamatok „hamis" képet nyújthatnak a modellezés során. Például ha két idősor, y_t és x_t közötti kapcsolatra vagyunk kíváncsiak, akkor modellezhetjük őket az alábbiak szerint:

$$y_t = \alpha + \beta x_t$$

Azaz y_t viselkedését x_t függvényének tekintjük. A gazdasági idősorok azonban gyakran tartalmaznak trendet, azaz időben növekszik az értékük. Ebben az esetben β még akkor is

pozitív lesz, ha y_t és x_t a valóságban semmilyen kapcsolat nincs. Ennek az az oka, hogy a korábbi időszakokban mind a két idősor értéke alacsonyabb volt, míg a későbbiekben mindig magasabb (hiszen trendálnak). A szakirodalomban ezt hamis korrelációnak (spurious correlation) nevezik. Ezt figyelembe kell venni bármilyen idősoros modellezés során és a megfelelő módszerekkel kezelni kell azt.

Mind a két típusú folyamatnak megvannak a maga modellezési gyakorlatai, habár a stacioner folyamatok vizsgálata lényegesen egyszerűbb. Ezért legtöbbször olyan transzformációt hajtunk végre, amely eltünteti a trendet és stacioner folyamatot eredményez. A leggyakoribb ilyen eset, amikor az időszakok közötti változás alakulását vizsgáljuk, azaz differenciáljuk az idősort. A differenciált idősort nevezzük hozam idősornak:

$$\Delta y_t = y_t - y_{t-1}$$

Érdemes megjegyezni, hogy ez az úgynevezett differencia stacioner folyamatoknál eredményez stacioner folyamatot (azaz, amikor a differenciálás után a folyamat stacionerré válik). Trend stacioner folyamatok esetén detrendálás, azaz a trend eltüntetése szükséges. Gyakran a relatív százalékos változások vizsgálata is szükséges lehet:

$$\%y_t = 100 * \left(\frac{y_t - y_{t-1}}{y_{t-1}}\right)$$

A relatív százalékos változásokat viszont legtöbbször logaritmikus differenciákkal közelítjük:

$$\Delta \ln(y_t) = \ln(y_t) - \ln(y_{t-1})$$

A legtöbb idősoros modell felírható az alábbi formában:

$$y_t = \underbrace{f(y_{t-1}, \dots, y_1)}_{jel} + \underbrace{\varepsilon_t}_{zaj}$$

A jel szisztematikus, a múltbéli adatok függvénye, így azokból előre jelezhető, míg a zaj független tag, amelyet nem lehet a saját múltja alapján előre jelezni. Legtöbbször azt feltételezzük, hogy az egyenlet valamilyen lineáris formát ölt majd.

6.3. Késleltetési operátor

Az idősorok elemzéséhez bevezetünk egy hasznos eszközt, a késleltetési operátort (’lag’ vagy ’backshift’ operátor), amelyhez *L* jelölést használunk és a következőképpen definiáljuk:

$$Ly_t = y_{t-1}$$

Az operátort többször is lehet alkalmazni:

$$L(Ly_t) = y_{t-2}$$

Ennek általánosítása k késleltetésre:

$$L^k y_t = y_{t-k}$$

Ahol $k = 1,2,3$ ... Ezen felül $L^0 y_t = y_t$ és $L^{-k} y_t = y_{t+k}$, (ez utóbbi egyfajta előre hozás). Ezen felül az operátorokat kombinálhatjuk is:

$$L^m(L^n y_t) = L^{m+n} y_{t-(m+n)}$$

Miért hasznos ez? Az idősoros modellezés esetén rendkívül gyakran használjuk a modellezni kívánt idősor vagy más idősorok korábbi értékeit (késleltetéseit), amely felírásához a késleltetési operátort rendkívül nagy segítséget nyújt. Például az első differencia:

$$\Delta y_t = y_t - y_{t-1} = y_t - Ly_t = (1 - L)y_t$$

6.4. Erős és gyenge stacionaritás

A stacionaritás az idősoros modellezés egyik alap koncepciója, amely az elemezni kívánt idősor statisztikai tulajdonságaira vonatkozik. Az erős stacionaritás a gyakorlatban azt jelenti, hogy az idősor hasonló statisztikai tulajdonságokkal rendelkezik időben, amely a konstans valószínűségi eloszlást is magában foglalja. Az erős stacionaritás folyamatát az alábbiak szerint definiálhatjuk, ahol F az együttes eloszlást jelenti:

$$F(y_1, \dots, y_T) = F(y_{t+k}, \dots, y_{T+k})$$

Azaz az erősen stacioner folyamat együttes eloszlása időben konstans minden t és k esetében. Gyakran elégséges viszont az úgynevezett gyenge stacionaritás fogalmát használni, amely a következő:

$$E(y_t) = \mu$$
$$Var(y_t) = \sigma^2$$
$$Cov(y_t, y_{t+k}) = \gamma_k$$

Azaz az idősor várható értéke és varianciája konstans, nem függ az időtől, míg az autokovariancia csak a t és $t + k$ közötti távolság függvénye. Tehát a gyenge stacionaritás már nem tesz feltételt az együttes eloszlásra, csak az első két momentumra (várható érték és variancia). Gyakorlati szempontból vizsgálva ezt azt jelenti, hogy ha eltérő időszakokban vizsgáljuk meg az idősort, akkor ugyanazokat az általános viselkedési tulajdonságokat mutatja.

A fentiekből következően érdemes definiálni az autokorrelációt, amely hasznos eszköze az idősor elemzésnek:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{Cov(y_t, y_{t+k})}{Var(y_t)}$$

Ezt több késleltetésre megvizsgálva kapjuk az autokorrelációs függvény (ACF). Sok esetben a két idősor közötti korrelációt az okozza, hogy mind a kettő korrelál egy harmadik folyamattal. Idősoros modellezés esetében ez azt jelenti, hogy y_t és y_{t-k} között gyakran azért van korreláció, mivel más késleltetésekkel korrelálnak. Ennek kiszűrésére vezették be a parciális autokorrelációs függvényt (PACF), amelyet kiszűri a köztes dependencia hatását.

6.5. *Fontosabb idősoros folyamatok*

6.5.1. Fehér zaj folyamat

Az egyik legfontosabb idősoros folyamat az úgynevezett *fehér zaj folyamat* (white noise process):

$$y_t = \varepsilon_t$$

Az ε_t valószínűségi változók sorozat, amelyek nem korrelálnak időben és 0 várható értékkel és konstans varianciával rendelkeznek. Valamivel erősebb feltételt is tehetünk, miszerint ε_t értékei időben függetlenek egymástól. Ebben az esetben mivel a megfigyelések egymástól függetlenek, ezért korrelálatlanok is (fordítva ez nem mindig igaz). Ez az idősor pusztán egy zaj folyamat, amelyben nem találunk dinamikát. Végül ha a megfigyelések teljesítik az előző feltételeket és egyúttal normál eloszlást követnek 0 várható értékkel és konstans varianciával, azaz $N(0, \sigma^2)$, úgy *normál eloszlású fehér zaj folyamatról* beszélünk (Gaussian white noise process). A független és azonos eloszlással rendelkező folyamatok esetében a várható érték és a variancia konstans, míg az autokorrelációs függvény:

$$\gamma(k) = Cov(\varepsilon_t, \varepsilon_{t+k}) = 0$$

Minden $k = \pm 1,2,$ ...értékre. A fehérzaj folyamat tehát stacioner folyamatnak számít.

6.5.2. A Mozgóátlag (MA) folyamatok

Egyes esetekben előfordulhat, hogy a múltbéli sokkok még a rákövetkező periódusokban is érezhető hatást fejtenek ki. Ilyen esetben a mozgóátlag folyamatok jó statisztikai leírását adhatják a folyamatnak. Ebben a fejezetben a legegyszerűbb, elsőrendű MA folyamat tárgyaljuk.

Az MA(1) folyamat

Az elsőrendű mozgóátlag folyamat a fehérzaj megfigyelések lineáris kombinációja, és a leírása a következő:

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1 + \theta L)\varepsilon_t$$

Ahol ε_t definíciója a korábbiakkal megegyező. Innen viszonylag rövid számolással megmutatható, hogy $E(y_t) = 0$ és $Var(y_t) = (1 + \theta^2)\sigma_\varepsilon^2$, mivel:

$$E(y_t) = E(\varepsilon_t) + \theta E(\varepsilon_{t-1})$$
$$E(y_t) = 0 + \theta * 0 = 0$$

Míg a variancia esetében:

$$Var(y_t) = Var(\varepsilon_t) + \theta^2 Var(\varepsilon_{t-1})$$
$$Var(y_t) = \sigma_\varepsilon^2 + \theta^2\sigma_\varepsilon^2 = (1 + \theta^2)\sigma_\varepsilon^2$$

Az $MA(1)$ folyamat autokorrelációját az alábbi függvény adja:

$$\rho_0 = 1, \rho_1 = \frac{\theta}{1 + \theta^2}, \text{ és } \rho_k = 0, minden\ k \geq 2$$

Habár a stacionaritás igazolásához nem kell további feltételeket tennünk, érdemes egy további feltétellel élni θ esetében és megbizonyosodni, hogy invertibilis a folyamat. Az MA(1) folyamat csak akkor invertibilis, ha $|\theta| < 1$. Ennek magyarázatára érdemes két folyamatot bemutatni, amely gyakori példaként szokott szolgálni:

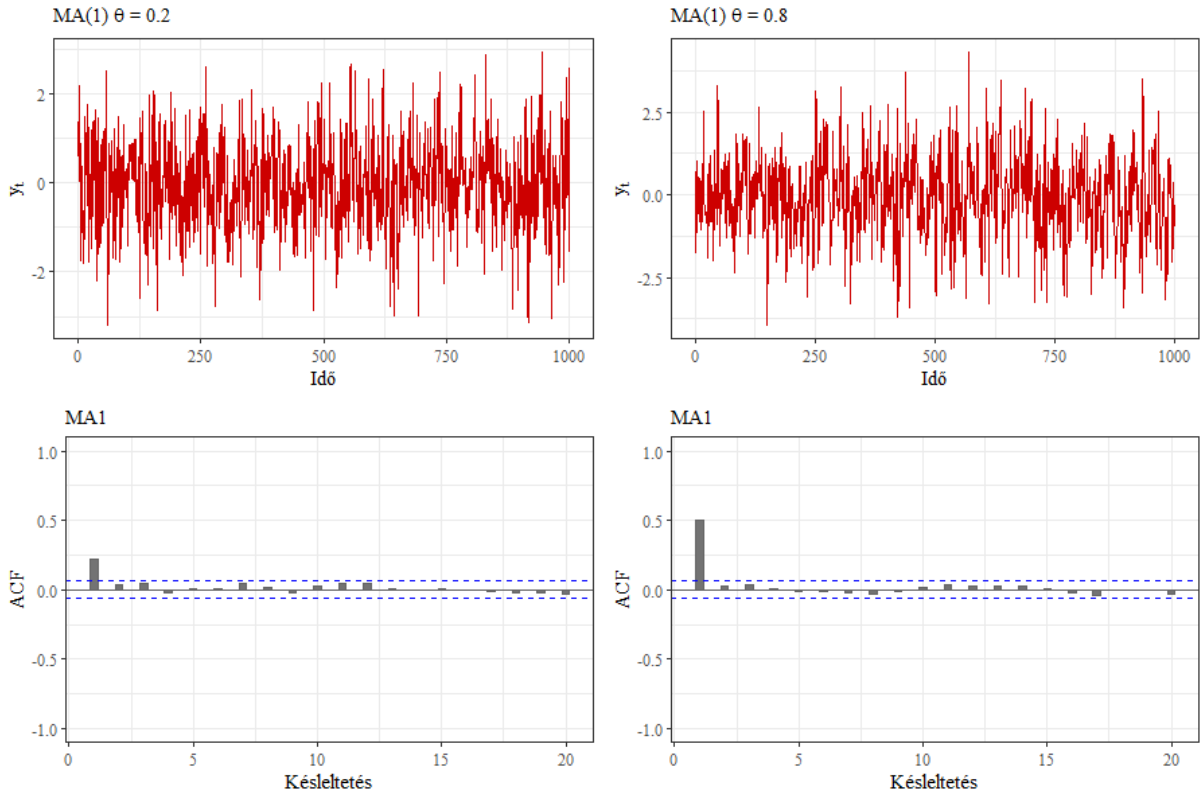
$$(1)\ y_t = \varepsilon_t + \theta\varepsilon_{t-1}$$
$$(2)\ y_t = \varepsilon_t + \frac{1}{\theta}\varepsilon_{t-1}$$

Annak ellenére, hogy ε_{t-1} paramétere különbözik, a két folyamat autokorrelációs függvénye ugyanaz! Azaz, nem derül ki melyik folyamatról van szó. Viszont ha y_t és késleltetései függvényében fejezzük ki a folyamatot, azaz:

$$(1)\ \varepsilon_t = y_t - \theta y_{t-1} + \theta^2 y_{t-2} - \cdots$$
$$(2)\ \varepsilon_t = y_t - \frac{1}{\theta}y_{t-1} + \frac{1}{\theta^2}y_{t-2} - \cdots$$

A két folyamat összehasonlításakor kiderül, hogy a $|\theta| < 1$ feltétel teljesülése esetén csak az (1) folyamat konvergens, a második nem, azaz (1) invertibilis, de (2) nem. Tehát ez a feltétel biztosítja, hogy egy adott autokorrelációs függvényhez egyedi MA folyamat társul.

A 6.1. ábra két generált MA folyamatot mutat, eltérő paraméterek mellett.



6.1. ábra: **Az MA folyamatok és ACF ábrájuk különböző paraméterek mellett**

Az általánosított MA folyamat jelölése $MA(q)$, ahol q a késleltetések száma:

$$y_t = \varepsilon_t + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j}$$

6.5.3. Az autoregresszív (AR) folyamatok

Az AR(1) folyamat

Az idősor modellezése történhet a saját, egy időszaki késleltetett értékével is, amelynek y_{t-1} a jelölése. Amennyiben az aktuális értéket csak a korábbi, megelőző érték befolyásolja, úgy a folyamatot első rendű *autoregresszív folyamatnak (AR)* nevezzük, amelynek jelölése AR(1):

$$y_t = \delta + \phi y_{t-1} + \varepsilon_t$$

Itt ϕ a késleltetett változó, y_{t-1} hatását mutatja a jelenlegi értékre, amelyet y_t jelöl, δ pedig egy egyszerű konstans, ami akár 0 is lehet. Ebben az esetben a folyamatot autokorreláció jellemzi, azaz az adott időszak értékét befolyásolja az, hogy hogyan alakult(ak) a korábbi érték(ek). Azaz, az autokorreláció az adott megfigyelés és ugyanannak a folyamatnak a korábbi megfigyelései közötti korrelációra utal. A becsült paraméter értéke felírható az alábbiak szerint is, ahol $Cor(.)$ a korreláció:

$$\phi = Cor(y_t, y_{t-1})$$

Azaz y_t és y_{t-1} közötti korrelációként. Fontos kritérium ez esetben, hogy $|\phi| < 1$, mivel csak így teljesül az időbeli stabilitás, azaz pusztán az autokorreláció megléte nem befolyásolja a stacionaritás tulajdonságát. Ez könnyen kiderül a várható érték és variancia kalkulációból:

$$E(y_t) = E(\delta) + E(\phi y_{t-1}) + E(\varepsilon_t)$$
$$\mu = \delta + \phi\mu + 0$$
$$\mu - \phi\mu = \delta$$
$$\mu(1 - \phi) = \delta$$
$$\mu = \frac{\delta}{(1 - \phi)}$$

Ennek értéke konstans, egyúttal ha $\delta = 0$, akkor a várható érték is zéró.

$$Var(y_t) = Var(\delta) + Var(\phi y_{t-1}) + Var(\varepsilon_t)$$
$$\sigma_y^2 = 0 + \phi^2\sigma_y^2 + \sigma_\varepsilon^2$$
$$\sigma_y^2 - \phi^2\sigma_y^2 = \sigma_\varepsilon^2$$

$$\sigma_y^2(1-\phi^2) = \sigma_\varepsilon^2$$

$$\sigma_y^2 = \frac{\sigma_\varepsilon^2}{(1-\phi^2)}$$

Ha $|\phi| = 1$, akkor a variancia végtelen. Az autokovariancia függvény az alábbiak szerint írható fel $k = 0, 1, 2, ...$ értékekre:

$$\gamma(k) = \sigma^2 \phi^k \frac{1}{1-\phi^2}$$

A variancia ebből következően:

$$\gamma(0) = \sigma^2 \frac{1}{1-\phi^2}$$

Végül az autokorrelációs (ACF) függvény:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \phi^k$$

Azaz az AR(1) folyamat ACF ábrája exponenciális csökkenést mutat. Mi történik akkor, ha $|\phi| \geq 1$? Ebben az esetben explozív folyamatról beszélünk, hiszen a hibatag ε_t nem tűnik el idővel, hanem felhalmozódik.

A következő esetben nézzünk egy AR(1) folyamatot konstans tag nélkül:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

Egymás utáni behelyettesítéssel a következőt kapjuk:

$$y_t = \phi \underbrace{(\phi y_{t-2} + \varepsilon_{t-1})}_{y_{t-1}} + \varepsilon_t = \phi^2 y_{t-2} + \varepsilon_t + \phi\varepsilon_{t-1}$$

$$y_t = \phi^2 \underbrace{(\phi y_{t-3} + \varepsilon_{t-2})}_{y_{t-2}} + \varepsilon_t + \phi\varepsilon_{t-1} = \phi^3 y_{t-3} + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2}$$

Végül az alábbi felírást kapjuk (ha $|\phi| < 1$, akkor a ϕ^k tag elhanyagolhatóan kicsit lesz k növekedésével):

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots$$

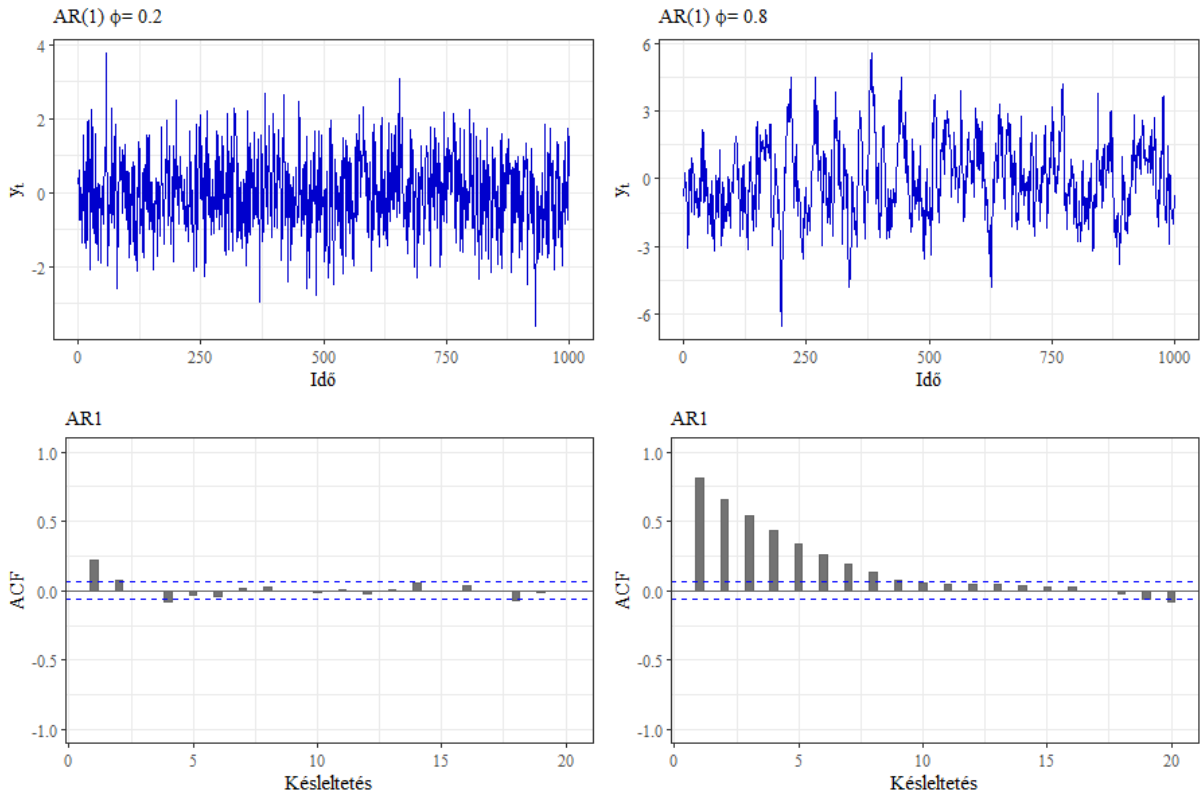
Azaz az AR(1) folyamat felírható egy végtelen MA folyamatként. Ez a felírás valamivel egyszerűbb a késleltetési operátor használatával:

$$y_t(1 - \phi L) = \varepsilon_t$$

$$y_t = \frac{\varepsilon_t}{(1 - \phi L)} = (1 + \phi L + \phi^2 L^2 + \cdots)\varepsilon_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots$$

Itt felhasználtuk, hogy $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$, ha $|x| < 1$. Ezekkel a reprezentációkkal szintén megmutatható a folyamat várható értéke és varianciája. Természetesen a modell kiegészíthető további késleltetésekkel, ezeket p-rendű folyamatoknak nevezzük, ahol p a késleltetések számát jelenti, a folyamatot pedig AR(p) jelöléssel látjuk el.

Az ábra két különböző paraméterrel rendelkező AR(1) folyamatot mutat be. Az első esetben $\phi = 0.2$, míg a másodikban $\phi = 0.8$. Tehát az első esetben a tegnapi érték hatása a mai értékre gyenge, míg a második esetben erősebb autoregresszív viselkedés figyelhető meg. Mindkét esetben érdemes észrevenni, hogy az idősor az átlaghoz való visszatérés jeleit mutatja (’mean reverting behavior’), azaz az átlagos érték körül ingadozik (6.2. ábra).



6.2. ábra: **Az AR folyamatok és ACF ábrájuk különböző paraméterek mellett**

Az általánosított AR folyamatot $AR(p)$ jelöléssel látjuk el, ahol p a késleltetések számát jelenti.

$$y_t = \delta + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \varepsilon_t$$

6.5.4. Véletlen bolyongás

Különleges helyet foglal el az idősoros ökonometriában a *véletlen bolyongás folyamat* (random walk), ahol az egyszerűség kedvéért nincs konstans tag:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

Itt $\phi = 1$, azaz a folyamat a következőre egyszerűsödik:

$$y_t = y_{t-1} + \varepsilon_t$$

A ε_t az előzőek során definiált fehér zaj folyamatot jelzi, y_t a mai érték, míg y_{t-1} a tegnapi érték. A véletlen bolyongás szerepe rendkívül nagy a valószínűségi alapokon nyugvó modellezésben. Érdemes úgy gondolni erre a folyamatra, mintha a késleltetett értékhez hozzáadódna ε_t, és ez utóbbi egy pénzfeldobás függvénye lenne. Amennyiben a dobás eredménye fej, ε_t pozitív értéket vesz fel, míg ha írás, akkor negatív értéket. Ebben az esetben a mai értékre (y_t) a tegnapi érték (y_{t-1}) adja a legjobb becslést.

A folyamat megegyezés szerint a $y_0 = 0$ értéktől indul és a

$$y_1 = \varepsilon_1$$

$$y_2 = y_1 + \varepsilon_2 = \varepsilon_1 + \varepsilon_2$$

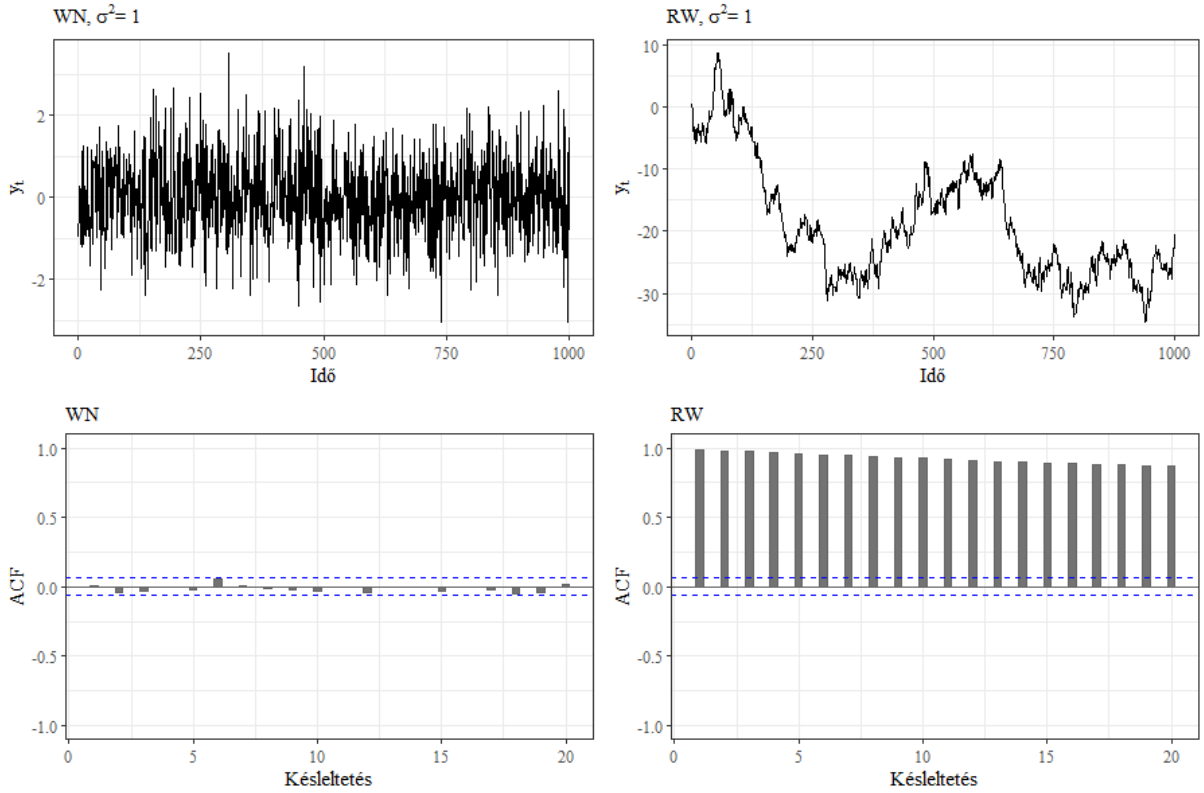
Végül:

$$y_t = \sum_{i=0}^{t} \varepsilon_i$$

Ebben az esetben a múltbéli sokkhatások összeadódnak és permanens hatásuk lesz. A várható érték operátort alkalmazva kapjuk, hogy $E(y_t) = t\mu$ és $Var(y_t) = t\sigma_\varepsilon^2$. Azaz a folyamat várható értéke és varianciája időben változik, azaz nem konstans, így a folyamat nem stacioner. Érdemes látni, hogy a Véletlen Bolyongás nem stacioner folyamat, de a differenciálás során azzá válik, hiszen:

$$\Delta y_t = y_t - y_{t-1} = (y_{t-1} + \varepsilon_t) - y_{t-1} = \varepsilon_t$$

Ahol ε_t a korábbiakban ismertetett folyamat. Az alábbi ábra egy tetszőleges fehér zaj és egy véletlen bolyongás folyamatot mutat be. Mivel a fehér zaj folyamat stacioner, a nulla várható érték körül ingadozik, míg a nem-stacioner véletlen bolyongás folyamat nevéhez hűen viselkedik.



6.3. ábra: **Tetszőleges fehér zaj és egy véletlen bolyongás folyamat bemutatása**

6.5.5. Autoregresszív mozgóátlag folyamatok

A két folyamatot kombinálhatjuk is. Ezt autoregresszív mozgóátlag folyamatnak nevezzük, amely az $ARMA(p,q)$ jelölést kapja, ahol p és q az autoregresszív és a mozgóátlag késleltetések száma.

$$y_t = \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j}$$

Ebben az esetben $\theta_0 = 1$. Az $MA(q)$ definíció szerint stacioner folyamat, de $ARMA(p,q)$ folyamat esetében más feltételeket is tenni kell a stacionaritás biztosítása érdekében. Ehhez egy úgynevezett késleltetési polinomot definiálunk:

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$$

és

$$\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_p z^p$$

Ezek segítségével az $ARMA(p,q)$ folyamat felírható, mint:

$$\phi(L)y_t = \theta(L)\varepsilon_t$$

A stacionaritás feltétele az, ha az autoregresszív polinom $\phi(L)$ gyökei az egységkörön kívülre esnek (a mozgóátlag polinom $\theta(L)$ esetében pedig a szokásos invertibilis feltétele él). Abban az esetben ha az idősorokat differenciálni is kell a stacionaritás eléréshez, akkor $ARIMA(p,d,q)$ modellről beszélünk, ahol d az integráció rendje.

A $\boldsymbol{\phi}$ és $\boldsymbol{\theta}$ rendjének meghatározása

Az idősoros ökonometria első időszakában a folyamatok rendjét az ACF és PACF ábrák segítségével határozták meg, de ez nem egzakt megoldás és jelentős gyakorlatot igényel. Sokkal hatékonyabb megoldás információs kritériumokra hagyatkozni, amelyek közül a legismertebbek a következők:

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T}$$

$$BIC = \ln(\hat{\sigma}^2) + \frac{k}{T}\ln(T)$$

$$HQIC = \ln(\hat{\sigma}^2) + \frac{2k}{T}\ln(\ln(T))$$

Ahol $\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T}\hat{\varepsilon}_t^2$, és $k = p + q$. Jellemző, hogy az AIC túl sok késleltetést választ, annak érdekében, hogy a maradéktagban semmilyen autokorreláció ne maradjon, míg a BIC túl keveset. Az információs kritériumok jellemzője, hogy kompromisszumot kötnek a modell magyarázó ereje és a becsült paraméterek száma között. Az ökonometriai szoftverek általában kiszabnak egy P és Q felső határt az autoregresszív és a mozgóátlag rend számára, majd szisztematikus teszteléssel vizsgálják meg az összes lehetséges kombinációt. Végül az a modell kerül kiválasztásra, amely minimalizálja az adott információs kritériumot.

Ez a fejezet csak az idősoros modellezés bevezetésének tekinthető, ezért érdemes felhívni néhány modellezési kérdésre a figyelmet. Habár a vizuális vizsgálat fontos része az ökonometriai adatelemzésnek, idősorok esetében lehetetlen megállapítani a valódi adatgeneráló folyamatot. A fejezetben tárgyalt egyszerű idősoros modellek jó alapot nyújtanak a komplexebb modellek felépítéséhez, mivel jó közelítéssel szolgálnak egyes bonyolultabb folyamatokhoz.

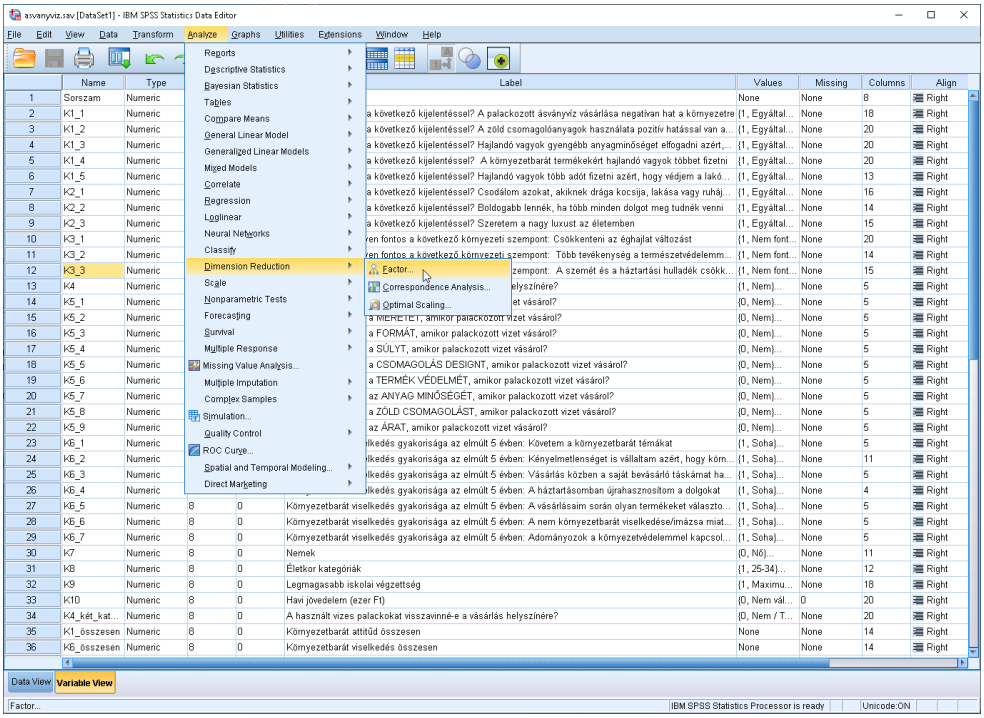
Ennek ellenére gyakori, hogy az idősor kevert folyamatú, azaz nem lehet pusztán egy kategóriába besorolni. Hasonlóan, a paraméterek vagy akár az adatgeneráló folyamat is megváltozhat (például egy szakpolitikai döntés miatt). Az idősor töréseket is tartalmazhat, amely lehet csak szint eltolódás, de akár trendváltozás is. Ezen felül elképzelhető, hogy egy adott folyamatra csak néhány adattal rendelkezünk és azok gyakorisága sem megfelelő. Éppen ezért a gyakorlatban fontos, hogy flexibilis modelleket alkalmazzunk, amelyek képesek szükség esetén ezeket a problémákat kezelni. A haladóbb modellezés már többváltozós környezetben vizsgálja az idősorokat és a közöttük lévő kapcsolatot, különösen a kointegrációt, amely a hosszú távú együttmozgást jelenti, a feltételes variancia változását, illetve a sokkhatások lecsapódását és átterjedését más folyamatokra. Ezek matematikája sokkal mélyebb, de ennek köszönhetően rendkívül jól le tudják írni a modellezni kívánt folyamatokat, így az idősoros modellezés rendkívül hasznos eszköze lehet a szakpolitikai javaslat tételnek.

7. A faktoranalízis számítása

A faktorelemzés során az általunk feltett kérdések mögött valamilyen látens struktúrát feltételezünk/keresünk. A faktorelemzés logikája az, hogy lehet néhány látens (nem mérhető / közvetlenül nem megfigyelhető) változó és ezek a látens elemek bizonyos kérdésekkel megragadhatók (vagy megpróbáljuk kifejezni ezeket a feltett kérdéseink segítségével). A gyakorlatban sokszor a kutatók a vizsgálni kívánt jelenségekhez kapcsolódó kérdéseket fogalmaznak meg és az elemzés során derül ki, hogy ezek a látens változók kimutathatóak-e (SZÉKELYI – BARNA, 2002). A nagyszámú sztochasztikusan összefüggő eredeti változó helyett, kisszámú un. faktorváltozót képezünk, melyek segítségével az adataink értelmezése és további elemzése leegyszerűsödik, mivel csökkenteni tudjuk a kiinduló változóink számát (ÁCS, 2009). A faktoranalízis egy struktúra feltáró módszer, mivel a függő és független változók nem előre meghatározottak és így a módszer fő célja a változók közötti összefüggések feltárása (SAJTOS – MITEV, 2007). Megállapítható, hogy a kialakított faktorok között minimális a korreláció nagysága. A faktorokkal további elemzéseket végezhetünk, helyettesítve velük az eredeti (nagyszámú) kérdéseket.

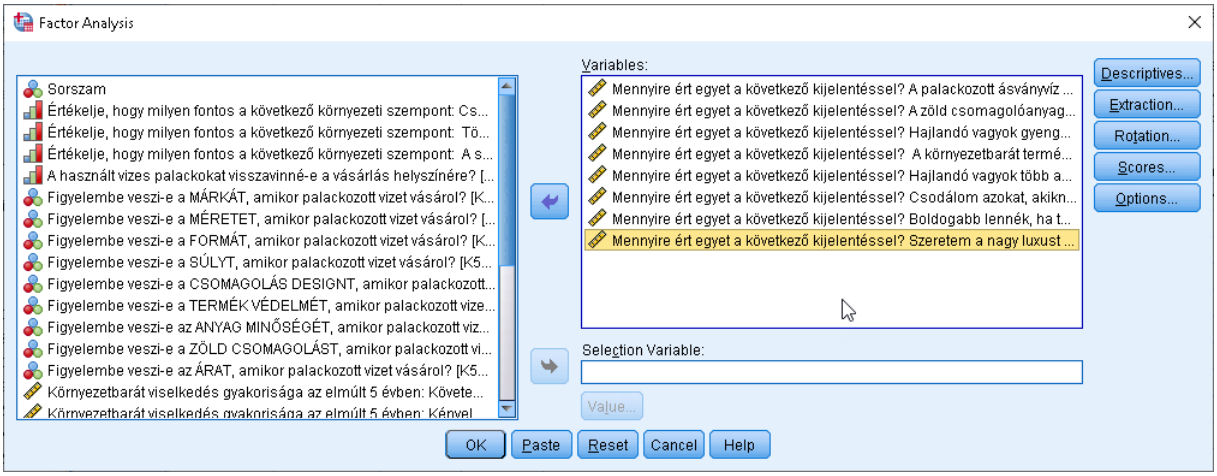
Vizsgáljuk meg – 8 kérdés (K1_1 – K2_3 változók) alapján – faktorelemzés segítségével, hogy a válaszadóknak a környezet iránti és a gazdagság iránti attitűdjei („Mennyire ért egyet a következő kijelentéssel?.....”) mögött van-e valamilyen közös látens struktúra!

A számítást az SPSS ANALYZE / DIMENSION REDUCTION / FACTOR… menüpontjával tudjuk elvégezni (7.1. ábra).



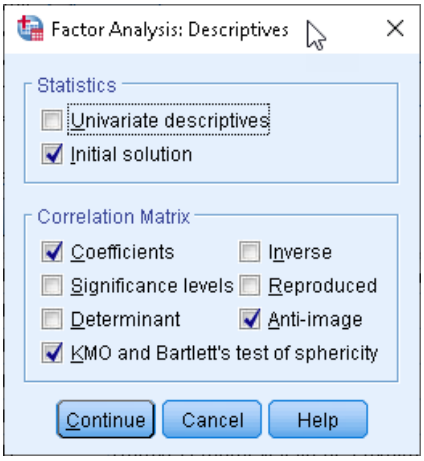
7.1. ábra. **A faktorelemzés indítása**

Ezután a megjelenő párbeszédablakban a bal oldali változólistából vigyük át az általunk vizsgálni kívánt 8 változót (K1_1 – K2_3) a jobb oldali változó („Variables”) listába.



7.2. ábra: **A faktorelemzés beállításai**

A „Descriptives” opciót választva egy új párbeszédablakban (7.3. ábra) lehet megvizsgálni, hogy a bevont változók/kérdések alkalmasak-e a faktoranalízis lefuttatására.



7.3. ábra: **A „Descriptive” menüpont lehetséges beállítása**

A „Statistics” menürészben egyváltozós leíró statisztikákat („Univariate descriptives”) kérhetünk a vizsgálatba vont változóinkról az alapbeállításon („Initial solution”) felül. A korrelációs mátrix („Correlation Matrix - Coefficients”) kiszámíttatható a programmal, mivel a változók közötti korreláció bizonyos szintje alapvető ahhoz, hogy faktorelemzést készíthessünk. Ezeket a korrelációs mátrixban szereplő kiszámított értékeket mutatja a 7.4. ábra.

Correlation Matrix

		Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben
Correlation	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	1,000	,597	,299	,216	,377	,157	,137	,049
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	,597	1,000	,240	,225	,395	,105	-,009	,040
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	,299	,240	1,000	,656	,445	-,068	-,003	-,115
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,216	,225	,656	1,000	,602	-,053	,013	-,017
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	,377	,395	,445	,602	1,000	,002	-,052	,014
	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	,157	,105	-,068	-,053	,002	1,000	,497	,617
	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	,137	-,009	-,003	,013	-,052	,497	1,000	,530
	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az	,049	,040	-,115	-,017	,014	,617	,530	1,000

7.4. ábra: **A faktorelemzésbe vont változók korrelációs mátrixa**

Ez alapján megállapíthatjuk, hogy van kapcsolat a változók között és kijelenthetjük, hogy a kiválasztott változók alkalmasak a faktorelemzés elvégzésére.

A „Descriptives” párbeszédablakban (7.3. ábra) a másik fontos előfeltétel teszteléséhez az „Anti-image” doboz kipipálása (kijelölése) szükséges. Mivel a változók szórásnégyzete felbontható két részre – a megmagyarázott és a nem megmagyarázott szórásnégyzetre –, amit az „Anti-image” kovariancia és korrelációs mátrixok mutatnak. Az „Anti-image” kovarianciamátrix átlótól különböző értékei a varianciának azt a részét fejezik ki, amely a többi változótól független. Ezért érdemes megvizsgálni ezeket az értékeket, és ha több mint háromnegyedük 0,09-nél kisebb az arra utal, hogy van mögöttes kapcsolat a vizsgálatba vont változóink között SAJTOS – MITEV, 2007). Az „Anti-image” korrelációs mátrix főátlójában szereplő értékei (0-1 közötti tartományban mozognak) az un. MSA „Measure of Sampling Adequacy” értékek, amelyek azt mutatják meg, hogy egy változó milyen szoros kapcsolatban van a többi – faktorelemzésben szereplő – változóval. Azokat a változókat, amelyeknek az MSA értéke 0,5 alatti, ki kell zárni a további elemzésből, mivel nem illeszkednek megfelelően a faktorszerkezetbe. A mi elemzésünkben a változóink MSA értékei 0,624 és 0,743 között voltak (7.5. ábra).

	életemben								
Anti-image Correlation	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	,640[a]	-,520	-,162	,102	-,176	-,087	-,170	,093
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	-,520	,663[a]	-,008	,002	-,176	-,054	,128	-,041
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	-,162	-,008	,694[a]	-,550	-,023	-,003	-,043	,136
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,102	,002	-,550	,624[a]	-,463	,071	-,063	-,055
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	-,176	-,176	-,023	-,463	,743[a]	-,001	,129	-,073
	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	-,087	-,054	-,003	,071	-,001	,690[a]	-,243	-,476
	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	-,170	,128	-,043	-,063	,129	-,243	,692[a]	-,344
	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	,093	-,041	,136	-,055	-,073	-,476	-,344	,640[a]

a. Measures of Sampling Adequacy(MSA)

7.5. ábra: **A faktorelemzésbe vont változók „Anti-image Correlation” mátrixa**

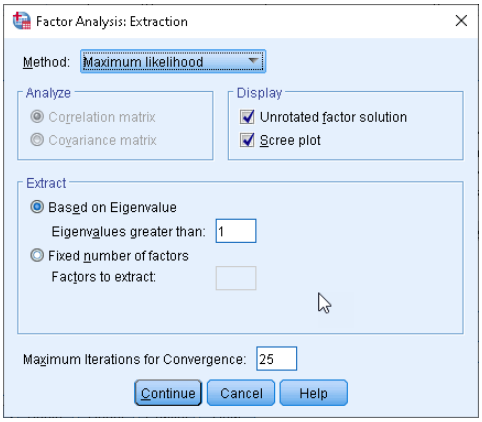
A 7.3. ábrán bemutatjuk, hogy a „Descriptives” opció választása után be lehet állítani a párbeszédablakban a „KMO and Bartlett’s test of sphericity” teszteket is. Ennek eredményeként ki tudjuk számítani az MSA értékek átlagát az összes változóra, aminek a Kaiser-Meyer-Olkin kritérium (KMO) a neve. 0,5 alatti érték esetén adataink alkalmatlanok a faktoranalízis elvégzésére. 0,9 feletti érték esetén tökéletesnek tekinthetők az adataink az elemzésre. A Bartlett-teszt nullhipotézise az, hogy a változóink nem korrelálnak egymással. Ami azt jelenti, hogy a vizsgálat során a korrelációs mátrix főátlón kívüli elemei nem térnek el szignifikánsan a nullától. Célunk, hogy ezt a nullhipotézist elvessük.
A következő ábrán (7.6. ábra) a „KMO and Bartlett’s Test” eredményeit tüntettük fel, amely alapján megállapítható, hogy a kiszámított KMO érték 0,672. Ez azt jelenti, hogy a változóink közepesen alkalmasak a faktoranalízisre. A Bartlett-próba nullhipotézisét is elvethetjük, mivel a teszt szignifikancia értéke („Sig.”) kisebb, mint 0,05, Ez azt jelenti, hogy a változóink alkalmasak a faktorelemzésre, mivel van korreláció közöttük.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,672
Bartlett's Test of Sphericity	Approx. Chi-Square	2435,466
	df	28
	Sig.	,000

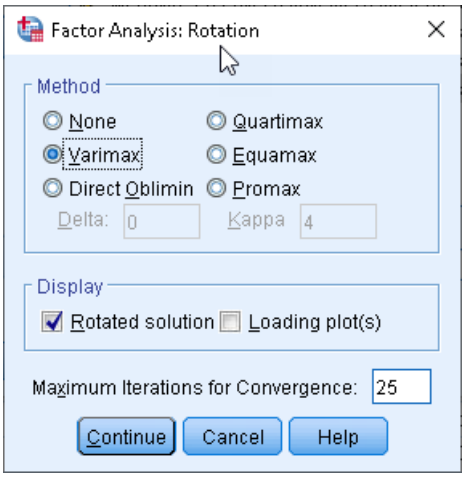
7.6. ábra: **A KMO kritérium és a Bartlett-próba értékei**

A következő lépésben a lehetséges faktorok számának meghatározását kell elvégeznünk és ehhez több eljárás közül választhatunk. Az „Extraction” gomb kiválasztása után lehetséges az eltérő faktorelemzési módszerek közül kiválasztanunk a számunkra legmegfelelőbbet (részletek SAJTOS – MITEV, 2007. 253-254. o.). A példánk esetében – mögöttes faktorstruktúrát feltételezve – a „Maximum likelihood” eljárást választottuk ki (7.7. ábra). Ez az eljárás a megfigyelt korrelációs mátrixból indul ki és olyan becslést ad, amely ezt a korrelációs mátrixot a legnagyobb valószínűség mellett alakíthatta ki (*normális eloszlást feltételez az alkalmazása!*) Ha a célunk az lett volna, hogy a sokaságban lévő legkevesebb információt veszítsük el, akkor érdemes lett volna a „Principal components” módszert kiválasztani. Mivel a főkomponens-elemzés a teljes varianciát figyelembe veszi.
Az „Extract” dobozban a faktorok számát lehet meghatározni. Ha a sajátérték „Eigenvalue” alapján akarunk dönteni (*Kaiser-kritérium*), akkor az alapbeállítást használjuk („Based on Eigenvalue”). Ebben az esetben a program abból indul ki, hogy egy faktornak több információt kell jelentenie, mint egy eredeti változó. Ha ismerjük a lehetséges faktorok számát választhatjuk a másik beállítást is („Fixed number of factors”). Ekkor elég a négyzetben megadnunk a kívánt faktorok számát.



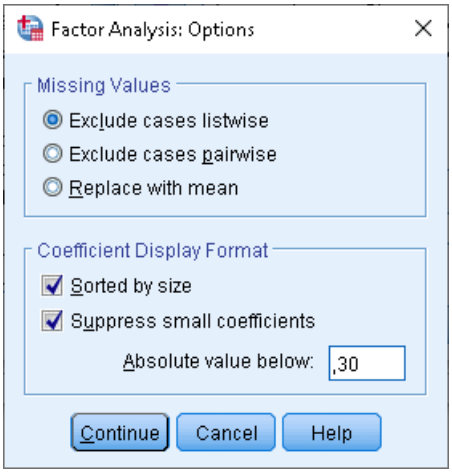
7.7. ábra: **A faktorelemzés módszerének kiválasztása és a lehetséges faktorok számának meghatározása**

A faktorok számának meghatározásához további lehetőséget jelent a „Scree plot” doboz kipipálása. A *„Könyökszabály”* ábra a sajátértékeket („Eigenvalue”) fogja megmutatni a lehetséges faktorok száma szerint. Ez alapján vizuálisan is könnyen eldönthetjük, hogy mennyi legyen a faktorok száma, mivel ahol a vonal meredeksége csökken és kezd ellaposodni, az adja meg a kutató számára a javasolt faktorszámot.
Ezután a „Rotation” almenüt (7.8. ábra) választhatjuk ki, amely lehetővé teszi számunkra, hogy a kiszámított faktorok tengelyeit elforgatva könnyebben értelmezhető eredményt kapjunk. A forgatás során a program a megmagyarázott varianciát egyenletesebben fogja elosztani a faktorok között. Javasolt a „Varimax” módszer kiválasztása, mivel így – az eredetihez képest – stabilabban lesznek szétválasztva – és jobban is magyarázhatók – a faktorok. Ez egy ortogonális forgatási eljárás és az így elkülönített faktorok egymással nem korrelálnak. A „Display” dobozban elég a „Rotated solution” feliratot választani, mivel most nem szükséges az elforgatott eredmények térbeli „Loading plot(s)” ábrázolása.



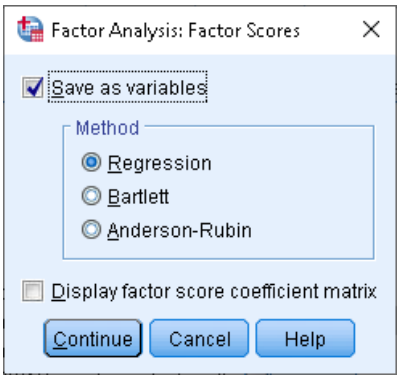
7.8. ábra: **A faktorok elforgatásának („Rotation”) beállítása**

Az „Options” almenü kiválasztásával meghatározhatjuk a közölt faktorok értelmezhetőségének kritériumait (7.9. ábra). A „Sorted by size” doboz kijelölésével a faktorsúly–mátrixban („Rotated Factor Matrix”) csökkenő sorrendben lesznek súlyok szerint a változóink. A „Suppress small coefficients” dobozt kiválasztva csak a megadott faktorsúly értéknél (0,3) nagyobb értékeket tünteti fel a program a két faktorsúly–mátrixban.



7.9. ábra: **A kialakított faktorok faktorsúly-mátrixának beállításai**

Ha az eredményeinket el szeretnénk menteni a „Scores” gomb kijelölésével kell folytatnunk. A „Save as variables” dobozt kipipálva a „Regression” módszer aktív lesz (7.10. ábra). Így a kialakított faktorszámnak megfelelő számú faktorváltozót hoz létre az SPSS az adatbázisunkban. Minden megfigyelt személyhez faktoronként külön értékek fognak tartozni. Ezeket az új változókat a későbbi elemzésekben (pl. regressziószámítása vagy klaszterezés) fel fogjuk tudni használni.



7.10. ábra: **A kialakított faktorok elmentése**

Ha végeztünk minden beállítással, az „OK” gomb megnyomásával tudjuk elindítani a faktorelemzést. Ezután az outputablakban számos táblázat fog megjelenni, melyek közül eddig már az első hármat elemeztük. A következő ábrán (7.11. ábra) a faktorelemzésben szereplő változók un. kommunalitás értékeit láthatjuk. A „Maximum likelihood” faktorelemzési eljárás esetében az „Initial” oszlopban a többszörös korrelációs együtthatók négyzetei (R^2) szerepelnek. Ezek az értékek mutatják meg azt, hogy egy változó szórását/varianciáját a többi elemzésbe vont változó mekkora mértékben magyarázza. Ha valamelyik kérdés esetében nagyon kicsi értéket kapunk, azt a kérdést a későbbiekben – nagy valószínűséggel – ki kell zárnunk a faktorelemzésből.

Fel kell hívnunk a figyelmet arra, hogy ha az „Extraction / Method” menüben (7.7. ábra) módszerként a főkomponens elemzés („Principal components”) módszert választottuk volna ki, minden változó esetében a kommunalitás „Initial” értéke 1 lenne.

Az „Extraction” oszlopban a végső kommunalitások szerepelnek, és ez azt mutatja meg, hogy a kialakított 3 faktor az egyes változók szórásának/varianciájának hány százalékát magyarázza meg. Ha valamelyik változó esetében 0,25 alatti érték szerepel, annak a változónak nincs elég magyarázó ereje és ezt ki kell hagyni a további faktorelemzésből. Az adatokat vizsgálva megállapíthatjuk, hogy mind a 8 elemzésbe vont változó megfelelő kommunalitás értékkel rendelkezik ahhoz, hogy a faktoranalízis további vizsgálataiban szerepeljen.

Communalities[a]

	Initial	Extraction
Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	,428	,666
Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	,401	,542
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	,471	,473
Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,562	,999
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	,458	,474
Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	,438	,590
Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	,362	,428
Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	,466	,665

Extraction Method: Maximum Likelihood.

7.11. ábra: **A változók kommunalitásának táblázata**

A következő 7.12. ábrán a teljes és a faktorok által magyarázott variancia értékei láthatóak a kezdeti, a faktoranalízis utáni és a forgatást követő értékek feltüntetésével. Mivel a főkomponens és a faktorelemzés az SPSS-ben közös pontból indul ki, ezért a maximum likelihood eljárásnál az első három oszlop („Initial Eigenvalues") azt a kiindulási helyzetet mutatja, mintha egy főkomponenselemzést hajtottunk volna végre (SZÉKELYI – BARNA, 2002). Ez alapján megállapítható, hogy a 8 bevont változóból képzett 8 főkomponens a teljes variancia 100%-át magyarázza és ebből az első három főkomponens sajátértéke lenne 1 felett (az általuk megmagyarázott variancia 73,77% lenne). Most mi a többi oszlop tartalmát fogjuk elemezni. Az „Extraction Sums of Squared Loadings" részben lévő adatok a faktorok által reprezentált információtartalmat (a „Total" oszlopokban van a sajátérték feltüntetve) mutatják. A 8 változó által képviselt lehetséges összes információ 8 egység, amiből az első faktor 1,896, a második 1,753 és a harmadik 1,188 egységnyit jelent. A „% of Variance" oszlopban az egyes faktorok által megmagyarázott variancia nagyságát olvashatjuk le (23,7%; 21,9% és 14,9%) csökkenő sorrendben. A „Cumulative %" oszlop összesítve mutatja az egyes faktorok összeadása utáni magyarázott varianciákat (pl. 23,7 + 21,9 = 45,6). Minket az érdekel, hogy a kialakított faktorok által megmagyarázott rész összesített % értéke 60% felett legyen. Ez azt jelenti, hogy sikerült olyan faktorstruktúrát találnunk, ami „*elég nagy részét"* magyarázza az eredeti 8 változó által képviselt összes információnak. A legtöbb esetben az így kapott eredeti faktorstruktúra nem igazán jól magyarázható szakmailag, ezért érdemes elforgatnunk a faktorok tengelyeit (Ennek a módszerét állítottuk be a „Rotation" almenüben „Varimax" eljárásként). Az ábrán megfigyelhetjük a „Rotation Sums of Squared Loadings" részben, hogy a forgatás

természetesen nincs hatással az elforgatott faktorok által megmagyarázott összes variancia nagyságára (60,456% maradt). Ismert az is, hogy a forgatás nem változtatja meg a modell illeszkedést és az egyes változók kommunalitását sem. Arra törekszik, hogy egyenletesebben ossza el a magyarázott varianciát a faktorok között. A varimax eljárással az egy változóhoz tartozó faktorsúlyok négyzetösszegeit maximalizáljuk. Ennek eredménye az, hogy – a forgatás előtti állapothoz képest – minden változót megpróbál még inkább egy faktorhoz hozzárendelni, amivel növeli a faktorok magyarázhatóságát.

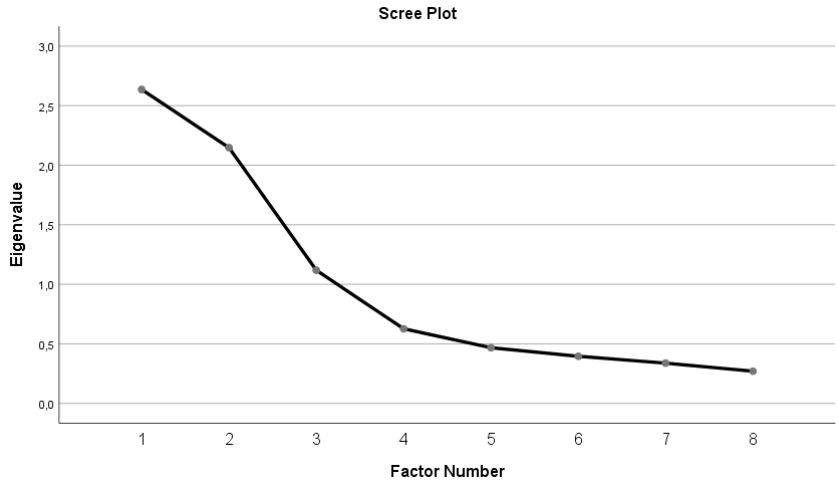
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,636	32,946	32,946	1,896	23,696	23,696	1,835	22,937	22,937
2	2,148	26,846	59,792	1,753	21,910	45,605	1,681	21,009	43,946
3	1,118	13,979	73,770	1,188	14,851	60,456	1,321	16,510	60,456
4	,627	7,832	81,602						
5	,468	5,844	87,447						
6	,396	4,947	92,394						
7	,338	4,225	96,619						
8	,270	3,381	100,000						

Extraction Method: Maximum Likelihood.

7.12. ábra: **A teljes és a faktorok által magyarázott variancia értékei**

A faktoranalízis beállításainál már említésre került, hogy a lehetséges faktorok számáról grafikus ábra („Scree-plot” vagy „Könyök-ábra”) segítségével is dönthetünk. A 7.13. ábráról könnyen leolvasható, hogy az ajánlott faktorok száma 3, mivel ez után már ellaposodik az ábránk. A „könyök-szabály” alapján ezért kijelenthető, hogy az ajánlott faktorszám maximálisan 3 lehet, mivel a negyedik faktor esetében a sajátérték már nagyon kicsi lenne (0,627).



7.13. ábra: **A javasolható faktorok számát mutató „Könyök-ábra”**

Az output file-ban a kiszámított eredmények között a következő táblázatok a forgatás nélküli („Factor Matrix”) és a forgatás utáni („Rotated Factor Matrix”) faktorsúlyokat bemutató táblázatok. *Ha főkomponens elemzést végzünk ezek a táblázatok „Component Matrix” illetve „Rotated Component Matrix” néven szerepelnek az outputban.*

A „Factor Matrix”-ban – előjellel ellátva – a faktorsúlyok vannak feltüntetve, amelyek azt jelzik, hogy az egyes változók milyen mértékben (milyen súllyal) vesznek részt a különböző faktorok kialakításában. A faktorsúly az eredeti változó és a faktor közötti korreláció szorosságát mutatja, aminek négyzetes értéke kifejezi, hogy a faktor a változó varianciájának hány százalékát magyarázza. Érdemes megemlíteni, hogy a nagyon kicsi (kisebb, mint 0,3) faktorsúly azt jelenti, hogy a változó nem kapcsolható össze az adott faktorral. Ha egyik faktor esetében sem éri el az adott változó faktorsúlya ezt a kritikus értéket, abban az esetben a változót ki kell zárni a további faktorelemzésből. Ha pedig több faktoron is nagyobb érték szerepel, mint 0,3 akkor a legnagyobb értéket kell összevetni a többi érték kétszeresével. Az egyes faktorokban feltüntetett értékek alapján megpróbálhatjuk a különböző faktorokat elnevezni. Felhívjuk a figyelmet arra, hogy ez nem minden esetben egyszerű. Általában a forgatás segít abban, hogy tisztábban lássuk a változóink és a faktorok közötti kapcsolatokat, ezzel megkönnyíti a kutató helyzetét abban, hogy jelentést tudjon adni az egyes elforgatott faktoroknak. A következő ábrán csak a 0,3 feletti faktorsúlyokat láthatjuk, mivel ezt az értéket állítottuk be az „Options” almenü „Suppress small coefficients” dobozában (7.9. ábra). Megfigyelhetjük, hogy az egyes faktorokon belül a változók súlyai csökkenő sorrendbe rendezettek, mivel előzetesen az „Options” menüben ezt a parancsot „Sorted by size” is beállítottuk. A faktorsúlyok abszolút értékben mutatják a változók értelmezhetőségét a faktorok viszonylatában. Példánkban az első faktornál a legjelentősebb hatása a „környezetbarát termékekért való magasabb fizetési hajlandóságnak” volt és ezt követte a másik két („anyagminőségre” és az „adófizetésre” vonatkozó) változó. A következő faktort a „szeretem a nagy luxust….”, a „csodálom azokat, akiknek drága kocsija, …..” és a „boldogabb lennék, ha több minden dolgot…..” változók alakították ki. A harmadik faktorba csak két kérdés a „palackozott ásványvíz vásárlása negatívan hat a környezetre” és a „zöld csomagolóanyagok használata pozitív…” tartozott.

Rotated Factor Matrix[a]

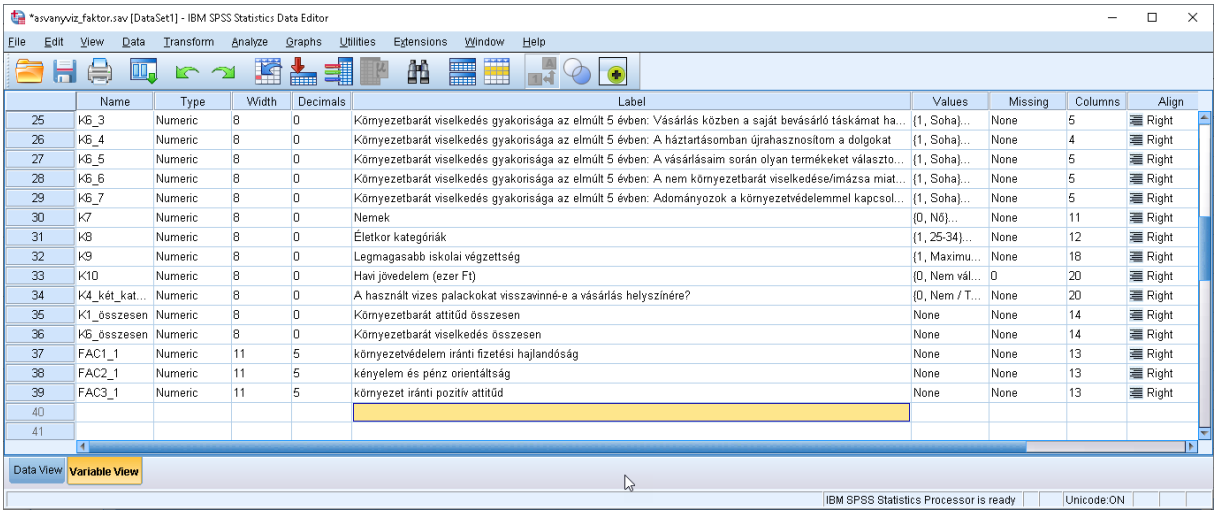
	Factor 1	2	3
Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,998		
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	,648		
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	,586		
Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben		,815	
Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van		,756	
Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni		,653	
Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre			,790
Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre			,710

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

7.14. ábra: **A rotálás utáni faktorsúlymátrix**

Mivel előzetesen a „Scores” gomb kijelölésével a „Save as variables” dobozt kipipáltuk az SPSS az eredményeinket (a 3 új faktort) az adatbázisan elmenti, azaz 3 új változót (FAC1_1 – FAC3_1) hozott létre. A további elemzések megkönnyítésére a „Data Editor”-ban a „Variable view” ablakban az új változók „Label” celláiba lehet beírni az elnevezéseket.
A rotált faktorsúlymátrix alapján megpróbálhatjuk elnevezni is ezt a három új változót:

- Környezetvédelem iránti fizetési hajlandóság (1. faktor)
- Kényelem és pénz orientáltság (2. faktor)
- Környezet iránti pozitív attitűd (3. faktor)



7.15. ábra: **Az új faktorok és elnevezéseik az adatbázisban**

8. Klaszterelemzési technikák

A marketingkutatás egyik leggyakoribb célja a fogyasztók, a megkérdezettek szegmentációja, vagyis olyan csoportok képzése, amelyek belül a válaszadók viszonylag homogének, ugyanakkor egymástól jól elkülönülők (SIMON, 2006). A leggyakrabban használt csoportképző ismérvek a fogyasztókat jellemző szocio-demográfiai jellemzők, de lehetséges a választott termékek tulajdonságai alapján is szegmentálni a fogyasztókat.
Általában a klaszterelemzés fő célja, hogy a megfigyelési egységeket viszonylag homogén csoportokba sorolja a kiválasztott változók alapján úgy, hogy az adott csoportba tartozó megfigyelési egységek hasonlítsanak egymásra, de különbözzenek más csoportok tagjaitól (FÜSTÖS et al., 2004; HAJDU, 2003). Ebben az anyagban a skálatípusú adatok elemzésére alkalmas klaszterezési eljárást mutatjuk be.
A klaszterelemzés menete:

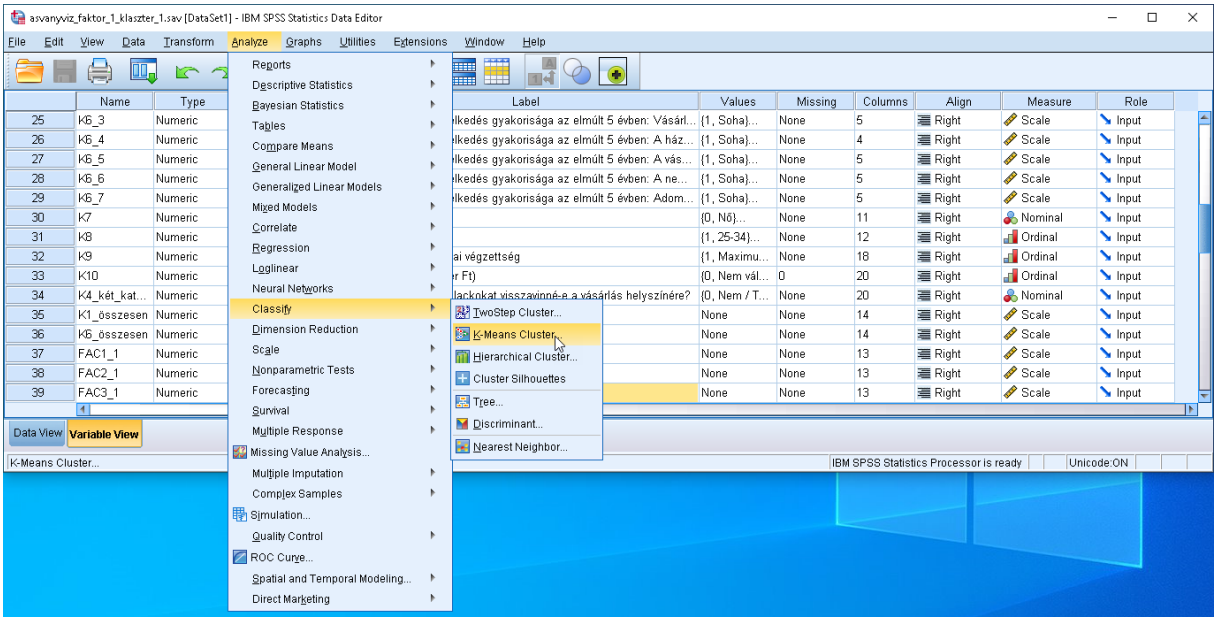
- A probléma megfogalmazása
- A távolság mérték kiválasztása
- A klasztermódszer kiválasztása
- Döntés a klaszterek számáról
- A klaszterek értelmezése és jellemzése
- A klaszterelemzés érvényességének ellenőrzése

A probléma megfogalmazása során kiválasztjuk a csoportképzés alapjául szolgáló változókat. A nem megfelelő változó bevonása ronthat a bevonás nélküli jó csoportosításon. A változók kiválasztása történhet korábbi kutatások alapján, elméleti megfontolások vagy a kutató saját döntése, intuíciója alapján (BACKHAUS et al., 2003).
A távolság mérték kiválasztása során az egységek közötti hasonlóságot azok közötti távolsággal mérjük. Különböző távolságmértékek használata (pl. Euklideszi, Csebisev, Manhattan, Pearson) eltérő megoldásokhoz vezethet, így célszerű különböző mértékeket használni, úgy elvégezni az elemzést, majd az eredményeket összehasonlítani.
A klasztermódszer kiválasztása esetében is több lehetőség adódik, mivel a klaszterezési eljárások lehetnek hierarchikusak és nem hierarchikusak. A hierarchikusak módszereket két csoportra bonthatjuk, mint agglomeratív (összevonó) és divizív (felosztó) eljárások. A klaszterek közötti távolság képzésének eljárásai a következők: egyszerű lánc módszer, teljes lánc módszer, centroid módszer, medián módszer, csoportátlag módszer és ward módszer.
A nem hierarchikus klaszterképzés eredményeként diszjunkt klasztereket állítunk elő. A klaszterek száma az egyes módszerek során alakul ki, más módszereknél előre meg kell adjuk ezt az értéket. A hierarchikus és nem hierarchikus módszerek között az az egyik lényeges különbség, hogy míg a hierarchikus eljárások esetén, ha két elem egy csoportba kerül, akkor a továbbiakban már együtt is marad, addig a nem hierarchikus eljárások esetében lehet, hogy később külön csoportba kerülnek át.
A klaszterelemzés érvényességének ellenőrzése során számos különböző lépést tehetünk. Ezek a következők lehetnek:

– Más távolság mértéket alkalmazunk és az így kapott eredményeket összehasonlítjuk.
– Különböző klaszterelj árásokkal dolgozunk.
– Az adatainkat véletlenszerűen két almintára bontjuk, és mindkettőre elvégezzük az elemzést.
– Véletlenszerűen elhagyunk változókat, és csökkentett változószámmal végezzük el újra az elemzést.
– Nem hierarchikus elemzéseknél futtassuk az elemzést az esetek különböző sorrendjével, mígnem stabilizálódik a megoldás.

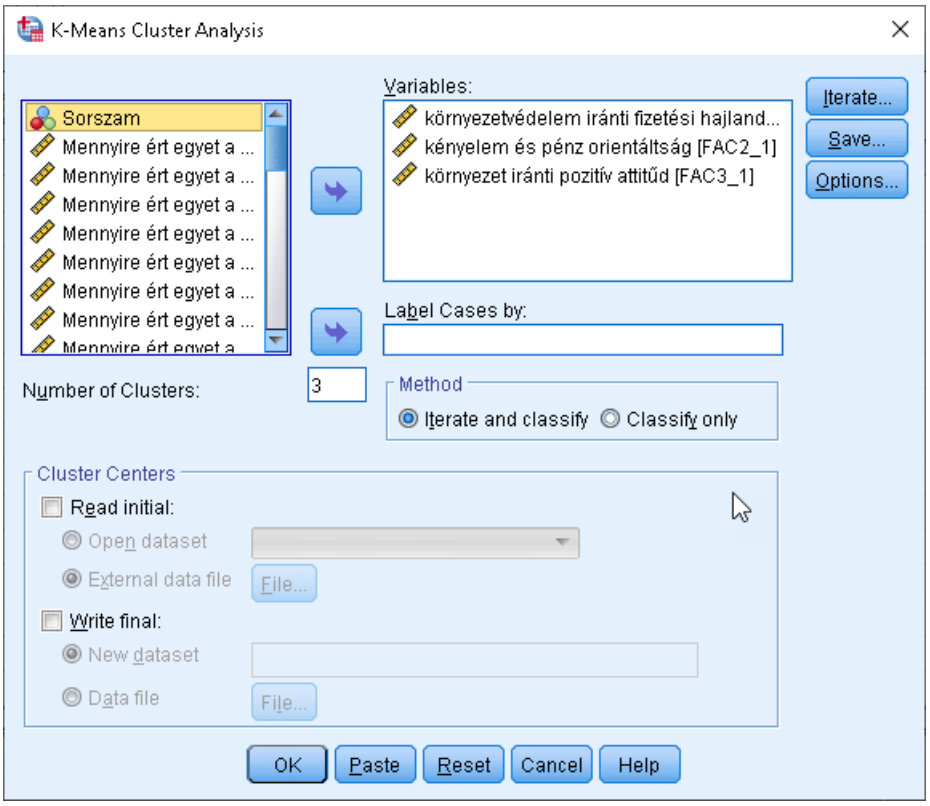
Vizsgáljuk meg – a környezetvédelemmel kapcsolatos 3 faktorváltozó (FAC1_1 – FAC3_1 változók) alapján – klaszterelemzés segítségével, hogy a válaszadók milyen egymástól jól elkülönülő csoportokba sorolhatók!

Az általunk bemutatott példa esetében a nem hierarchikus klaszterképzést fogjuk alkalmazni, mivel előzetes szakirodalmi adatok alapján általában három klaszterbe sorolhatók az ásványvizet vásárlók. A számítást az SPSS ANALYZE / CLASSIFY / K-MEANS CLUSTER… menüpontjával tudjuk elvégezni (8.1. ábra).



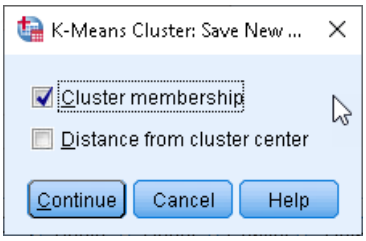
8.1. ábra: **A klaszteranalízis kiválasztása**

A K-közép klaszteranalízis kiválasztása után meg kell adnunk azokat a változókat, amelyek alapján a válaszadókat külön csoportokba szeretnénk rendezni (8.2. ábra). Ezért a baloldali ablakban ki kell jelölnünk a három faktorváltozót (FAC1_1 – FAC3_1) és át kell mozgatnunk a jobboldali változó (Variables) ablakba. A szükséges csoportok számát a „Number of Clusters:” dobozban állíthatjuk be. Mivel előzetesen volt információnk a lehetséges csoportok számáról, ezért 3-at írjuk be értékként.



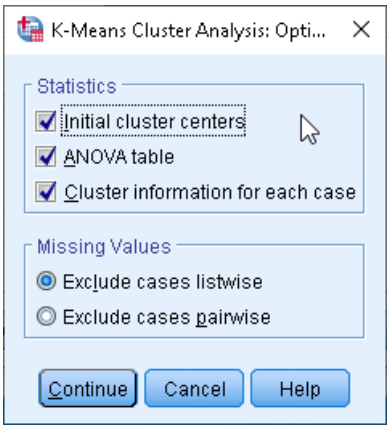
8.2. ábra: **A 3 faktorváltozó kiválasztása és a klaszterszám megadása**

Ezt követően lehetőségünk lesz arra, hogy a „Save” menü kiválasztásával beállítsuk azt, hogy a kiszámított klasztereket elakarjuk-e menteni („Cluster membership”) azért, hogy később elemzéseket futtassunk a kialakított klaszterekbe sorolt válaszadókkal.



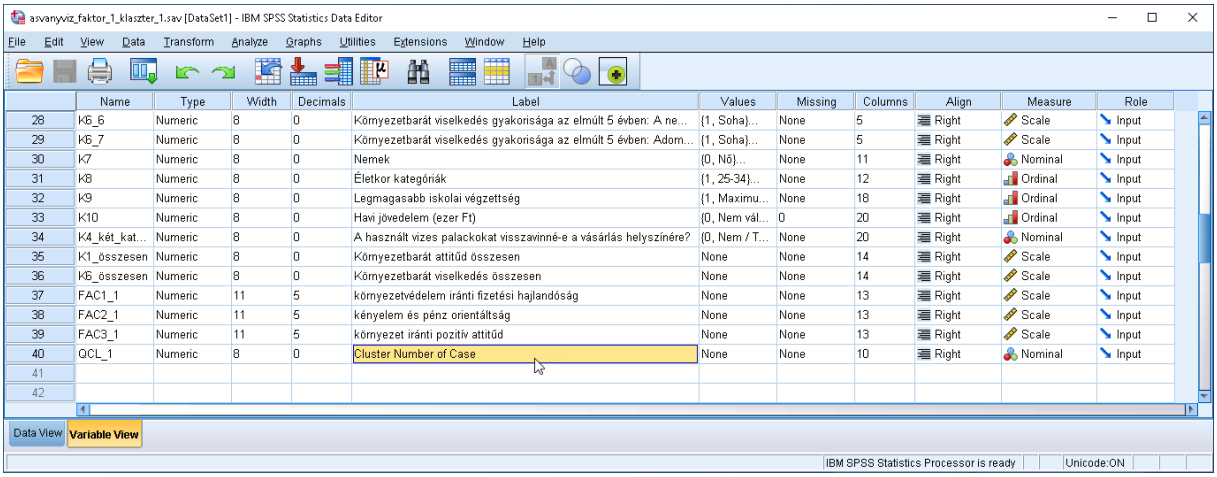
8.3. ábra: **A klasztervváltozó létrehozása az adatbázisban**

Az „Options” menüben be lehet állítani további eredményeket, ha bekattintjuk az alábbi cellákat.



8.4. ábra: **Az Option almenü beállításai**

Lefuttatva a fenti beállításokkal a klaszterezést, az SPSS változói között megjelenik egy új klaszterváltozó is.



8.5. ábra: **Az új klaszterek egyedi értékeit tartalmazó klasztervváltozó**

A kiindulási klaszterértékeket nagyon jól nyomon lehet követni a következő ábrán feltüntetett értékek alapján.

Initial Cluster Centers

	Cluster		
	1	2	3
környezetvédelem iránti fizetési hajlandóság	-1,71696	-2,94564	1,47346
kényelem és pénz orientáltság	-1,91199	1,46550	1,99534
környezet iránti pozitív attitűd	-,82473	1,53577	,24664

8.6. ábra: **A kiindulási klaszterek közötti távolságok értékei**

A program 8 iterációban alakította ki a klasztereket.

Iteration History[a]

Iteration	Change in Cluster Centers 1	2	3
1	1,956	1,915	1,813
2	,239	,371	,140
3	,282	,271	,129
4	,286	,258	,103
5	,095	,072	,014
6	,026	,023	,000
7	,000	,017	,019
8	,000	,000	,000

8.7. ábra: **Az iterációnként változások eredményei**

Meg lehet vizsgálni egyénenként azt, hogy ki melyik klaszterbe tartozik és milyen távol van a hozzá leghasonlóbbtól.

Cluster Membership

Case Number	Cluster	Distance
1	2	3,144
2	2	3,144
3	2	3,144
4	2	3,144
5	2	3,144
6	2	3,144
7	2	2,596
8	2	2,596
9	2	2,596
10	2	2,922
11	2	2,922
12	2	2,922
13	2	2,640
14	2	2,640

8.8. ábra: **A klaszterszám megadása és a közöttük lévő távolságok**

A végső klaszterközéppontokat érdemes elemezni szakmai következtetések levonása miatt is.

Final Cluster Centers

	Cluster 1	2	3
környezetvédelem iránti fizetési hajlandóság	,53263	-1,08863	,66831
kényelem és pénz orientáltság	-,88709	,12897	,75047
környezet iránti pozitív attitűd	,05718	,00750	-,06587

8.9. ábra: **A 3 faktorváltozó központjainak klaszterenkénti értékei**

A program számszerűsíteni tudja a végső klaszter-középpontok közötti távolságokat is.

Distances between Final Cluster Centers

Cluster	1	2	3
1		1,914	1,648
2	1,914		1,865
3	1,648	1,865	

8.10. ábra: **A klaszter-középpontok közötti távolság megadása**

Varianciaanalízissel össze tudjuk hasonlítani a különböző csoportátlagok közötti különbségeket. Az alábbi ábrából leolvasható, hogy az utolsó változó (környezet iránti pozitív attitűd) megítélése nem szignifikáns azaz a válaszadók kb. egységesen ítélik ezt meg.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
környezetvédelem iránti fizetési hajlandóság	297,351	2	,342	903	869,201	,000
kényelem és pénz orientáltság	199,909	2	,364	903	549,839	,000
környezet iránti pozitív attitűd	1,119	2	,759	903	1,474	,230

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

8.11. ábra: **Variancianalízis eredményei**

Fontos információ az is, hogy ez egyes klaszterekben hány válaszadó van, mivel nekik hasonló ízlésük és várhatóan hasonló döntéseik lesznek.

Number of Cases in each Cluster

Cluster	1	293,000
	2	322,000
	3	291,000
Valid		906,000
Missing		,000

8.12. ábra: **A 3 klaszter elemszámainak megadása**

9. Korrespondancia-analízis

A korrespondencia analízis egy olyan exploratív technika, mely az asszociációs kapcsolatok vizuális elemzése érdekében egy gyakorisági tábla adatait grafikus ábrává konvertálja. A tábla sorai mint pontok az oszlopok terében, az oszlopai pedig mint pontok a sorok terében kerülnek ábrázolásra egy redukált egy-, két-, vagy háromdimenziós altérben. A pontfelhők helyzetének az elemzése révén nyílik lehetőség a sorok és az oszlopok közötti asszociációk feltárására. Sor vagy oszlop kategóriák kombinációját is reprezentálhatja.
Egyszerű korrespondencia analízisről (CA) beszélünk, ha a sorok v.s. oszlopok pontfelhőt ábrázoljuk. Ez a helyzet például, mikor csak két változó alkotja a gyakorisági táblát. Ezzel szemben többszörös – multiple - korrespondencia analízist (MCA) alkalmazunk, ha kettőnél több változó szerepel elemzésünkben és valamennyi változó valamennyi kategóriáját önálló pontként ábrázoljuk (HAJDU, 2011).

Technikailag az MCA nem más, mint egy speciális gyakorisági táblán, az indikátor mátrixon végzett CA. Az indikátor mátrixban a megfigyelési egységek (háztartások, személyek) képezik a sorokat, és valamennyi változó valamennyi kategóriája egy-egy önálló oszlopot alkot: a megfelelő belső cella gyakorisága 1, ha az illető megfigyelés az illető oszlophoz, mint kategóriához tartozik, a gyakoriság egyébként zéró. Matematikailag a korrespondencia analízis az asszociáció Pearson-féle χ 2 mértékét bontja komponensekre hasonló módon, mint azt a főkomponens analízis a totális varianciával teszi. Az eljárás a sorokat (oszlopokat) a megoszlásaikból képzett, redukált dimenziójú térben, mesterséges CA koordináták alapján ábrázolja. Itt a tengelyeket úgy definiáljuk, hogy rendre csökkenő százalékos mértékben (sorrendben) járuljanak hozzá a χ 2 statisztikához. Mikor az első, vagy az első kettő mesterséges tengely a teljes asszociáció meghatározó (80-90% körüli vagy több) hányadát magyarázza, a gyakorisági tábla síkbeli ábrája a pontfelhő eredeti konfigurációját hűen, kevés információvesztéssel tükrözi.

Az asszociáció mértéke egy n számú megfigyelést csoportosító kontingencia táblában adott. A CA a korrespondencia mátrixot elemzi, melynek általános eleme az i sor és a j oszlop együttes $p_{ij}=f_{ij}/n$ relatív gyakoriság ahol f_{ij} az előfordulási gyakoriság.
Az s_i sorösszesenek és az o_j oszlopösszesenek szintén az n mintaméret százalékában kerülnek kifejezésre. A relatív gyakoriságok feltételes, adott soron belüli sorozatát sorszerkezetnek, adott oszlopon belüli sorozatát pedig oszlopszerkezetnek nevezzük. A sorszerkezeteket az oszlopok által kifeszített J-dimenziós tér, míg az oszlopszerkezeteket a sorok által kifeszített I-dimenziós tér pontjaiként kezeljük, két pontfelhőt definiálva. A peremmegoszlásokat, mint a megfelelő tengelyekhez rendelt súlyokat az s és az o vektorokba foglaljuk.
A sorszerkezetek és az oszlopszerkezetek a korrespondencia mátrix elemeihez a következő módon kapcsolódnak.
Megfigyelésenként összegezve mindkét oldalát, a tábla alapján látszik, hogy az s_i és o_j peremek a belső oszlop- és sorszerkezetek súlyozott átlagai, súlyként a peremek másik körét alkalmazva. Így az oj perem egyben a sorprofilok centroidja, míg az si perem az oszlopprofilok centroidja.

Az asszociáció teljes hiányát a sorok és az oszlopok között tehát úgy is definiálhatjuk, hogy adott pontfelhő valamennyi pontja egybeesik egymással, és így a saját centroidjával is. Az asszociáció természetére tehát a pontfelhők szóródásából következtethetünk. A szóródás mérésére a Pearson-$\chi 2$ statisztikát alkalmazzuk, melynek megszokott megnevezése a CA terminológiában: totális inercia. Ahol $s_i o_j$ az [i,j] cella függetlenség esetén várható relatív gyakorisága és a standardizált korrespondencia gyakoriság. Zérótól különböző g_{ij} előjele pozitív, vagy negatív asszociációt jelez az i sor és a j oszlop között. Az azonosság alapján a teljes inercia akár a sorfelhő, akár az oszlopfelhő pontjainak súlyozott, többváltozós varianciája. ahol $s_{ijc}=s_{ij} - o_j$ és $o_{ijc}=o_{ij} - s_i$ a centrált sor- és oszlopprofilok, melyek centroidja mindig az origó. A korrespondencia analízis a pontfelhőket mesterséges CA koordináták alapján ábrázolja. A koordináták értékét úgy határozzuk meg, hogy minden egyes egymást követő (k=1,2,...,K) koordináta-tengely egyre inkább csökkenő hányadban járul hozzá a teljes inerciához. Mikor például az első két tengely az inercia döntő hányadát magyarázza, akkor a pontfelhők konfigurációja a CA koordináták síkbeli ábrázolásában is valósághű marad. A centrált sorszerkezetek ábrázolására x, a centrált oszlopprofilok ábrázolására pedig y koordinátákat számítunk, melyeket az X(I,K) és az Y(J,K) mátrixokba foglalunk. A CA tengelyek lehetséges maximális száma K=min{I-1, J-1} mivel a relatív gyakoriságok összege adott profilon belül 1. Fontos mozzanat, hogy a sorkoordináta a standardizált oszlopkoordináták súlyozott átlaga, míg az oszlopkoordináta a standardizált sorkoordináták súlyozott átlaga, súlyként rendre a sor-, illetve oszlopprofilt alkalmazva. Az oszlopok és a sorok egymásba való átvitele lényegében a koordináták duális skálázását jelenti. Az x_{ik} és y_{jk} akkor van közel egymáshoz, mikor a j oszlop nagy s_{ij} súllyal szerepel az i sorprofilban vagy az i sor szerepel nagy o_{ij} súllyal a j oszlopprofilban. Ebben az esetben egy nagy sorkoordináta a k tengelyen szükségszerűen szintén nagy oszlopkoordinátát eredményez ugyanezen a tengelyen. Közös koordináta rendszerben ábrázolva a sorok és az oszlopok pontfelhőjét tehát, azon sorok és oszlopok kerülnek várhatóan közel egymáshoz, amelyek között szoros az asszociáció mértéke. Ez teszi lehetővé a kapcsolatok feltárását, mert pontok közötti távolságot csak pontfelhőn belül értelmezünk, pontfelhők között nem. A pontfelhők közötti korrespondenciát tehát a duális skálázás elve alapján ítéljük meg.

Az eljárás a Data Reduction főmenűben a Correspondence Analysis almenüjéből végezhető el, ahol először a sorváltozókat, majd az oszlopváltozókat kell megadnunk.

```
CORRESPONDENCE TABLE=K3_1(1 3) BY TSC_5674(1 3)
/DIMENSIONS=2
/MEASURE=CHISQ
/STANDARDIZE=RCMEAN
/NORMALIZATION=SYMMETRICAL
/PRINT=TABLE RPOINTS CPOINTS
/PLOT=NDIM(1,MAX) BIPLOT(20).
```

9.1. ábra: **A parancskódok bemutatása**

Ezután mindkét ismérvet definiálni kell, a benne szereplő ismérvváltozatok számának segítségével. A sorváltozót (kérdés kódok) 1-től 3-ig, míg az oszlopváltozót (klasztercsoportok száma) szintén 1-től 3-ig. A többi beállításon nem változtatva futtassuk le az elemzést. A keletkező eredmények a következő ábrákon követhetjük nyomon.

Correspondence Table

Értékelje, hogy milyen fontos a következő környezeti szempont: Csökkenteni az éghajlat változást	3 klaszter			
	1	2	3	Active Margin
Nem fontos	141	54	153	348
Közepesen fontos	114	93	132	339
Nagyon fontos	102	51	66	219
Active Margin	357	198	351	906

9.2. ábra: **A korrespondancia táblázat**

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,128	,016			,617	,617	,032	-,013
2	,101	,010			,383	1,000	,033	
Total		,026	23,907	,000[a]	1,000	1,000		

a. 4 degrees of freedom

9.3. ábra: **A statisztikai különbséget bemutató táblázat**

Overview Row Points[a]

Értékelje, hogy milyen fontos a következő környezeti szempont: Csökkenteni az éghajlat változást	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
Nem fontos	,384	,444	-,077	,010	,593	,022	,977	,023	1,000
Közepesen fontos	,374	-,347	-,271	,009	,353	,273	,675	,325	1,000
Nagyon fontos	,242	-,169	,541	,008	,054	,704	,110	,890	1,000
Active Total	1,000			,026	1,000	1,000			

a. Symmetrical normalization

9.4. ábra: **A sorokat kialakító változó értékeinek eredményei**

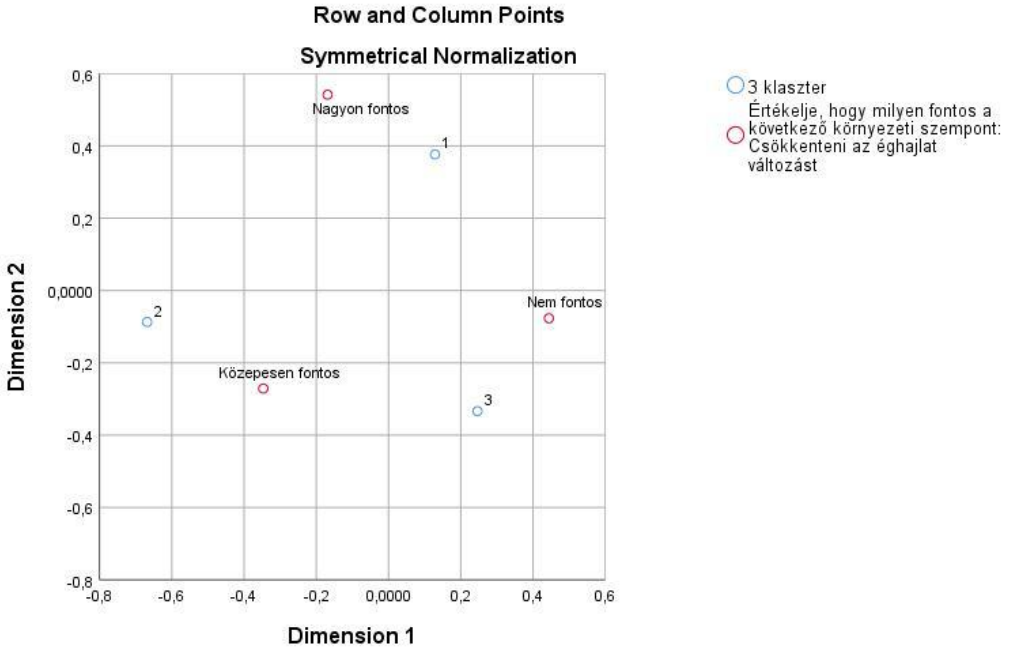
Overview Column Points[a]

3 klaszter	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution Of Point to Inertia of Dimension 1	Contribution Of Point to Inertia of Dimension 2	Contribution Of Dimension to Inertia of Point 1	Contribution Of Dimension to Inertia of Point 2	Contribution Total
1	,394	,129	,376	,006	,051	,555	,129	,871	1,000
2	,219	-,668	-,087	,013	,765	,017	,987	,013	1,000
3	,387	,246	-,334	,007	,184	,429	,409	,591	1,000
Active Total	1,000			,026	1,000	1,000			

a. Symmetrical normalization

9.5. ábra: Az oszlopokat kialakító változó értékeinek eredményei

A keletkező eredmények közül a grafikus ábrázolást vizsgálva, láthatóvá válnak az összetartozó értékek is.



9.6. ábra: **A korrespondancia-analízis grafikus megjelenítése**

A 9.6. ábra alapján jól elkülönülnek az egyes klaszterek és az éghajlatváltozás fontosságának szintjei. Megfigyelhetjük, hogy az első klaszter szorosan összefügg azokkal, akik azt választották, hogy számukra nagyon fontos az éghajlatváltozás, mint környezeti szempont. A második klaszterbe azok különültek el, akiknek közepesen fontos volt ez a szempont. Az utolsó klaszterbeliek számára elhanyagolható az éghajlatváltozás fontosságának figyelembevétele.

10. A diszkriminancia analízis bemutatása

A diszkriminanciaanalízis az az eljárás, ami a következő kérdésre ad választ: milyen csoportba tartoznak célcsoportunk alanyai? A csoportot itt igen tágan értelmezhetjük. Tulajdonképpen bármilyen csoportosítás lehetséges, amelyet értelmesen vizsgálni lehet. A diszkriminancia-analízis olyan adatelemzési módszer, amelyet kategóriába tartozás előrejelzésére lehet használni. Alkalmazásával alacsony mérési szintű függő változót magas mérési szintű független változók segítségével magyarázunk.
Azt vizsgáljuk, hogy a csoporthoz tartozás mekkora százalékban becsülhető a független változókkal (pl. azt, hogy valaki drog függő vagy nem, mekkora mértékben magyarázza az életkor, iskolázottság, stb.). Az eljárás során, akárcsak a lineáris regresszió esetében olyan egyenest keresünk, amely a legjobban szétválasztja az elemzendő csoportokat.
A diszkriminanciaanalízis során tehát azt a problémát járjuk körül, hogyan lehet az emberek egyes csoportjait valamilyen vizsgált jellemzők alapján szétválasztani, az egyes csoportokat azonosítani, valamint a csoporttagságokat az előbb említett vizsgált jellemzők alapján előrejelezni.
Példa: megfelelő-e egy jelölt az állásra vagy sem, hajlamos-e depresszióra vagy nem, visszaeső bűnöző lesz-e a személy vagy a büntetés után feladja a bűnözői karrierjét.
Hogy a csoporttagságokat előre tudjuk jelezni, valamilyen jellemző vagy képesség vizsgálata szükséges, amely vagy amelyek alapján a becslésünket meg tudjuk tenni. Például a jelölt alkalmasságának vizsgálatához tesztelhetjük a munkavégzés szempontjából kulcsfontosságú képességeit. A depresszióra való hajlam vizsgálatánál megnézhetjük a családbeli előfordulását a depressziónak, nézhetjük a személyt érő stressz mennyiségét, stb. Mindezek, és az ehhez hasonló vizsgálatok végrehajtására alkalmas eljárás a diszkriminancia-analízis.

Lineáris diszkriminancia-analízis a statisztikában, minta-felismerésben és gépi tanulásban használt módszer, amely a független változók olyan lineáris kombinációját képes megtalálni, amely a függő változó alapján kialakított csoportokat a lehető legjobban megkülönbözteti (diszkriminálja). A diszkriminancia-analízis szorosan kapcsolódik a varianciaanalízishez és a regresszióanalízishez, amelyek úgyszintén egy függő változót igyekeznek kifejezni más változók lineáris kombinációjaként. Azonban míg e két utóbbi eljárásnál a függő változó folytonos változó, addig a diszkriminancia-analízisnél ez kategorikus változó. Ellentétben a varianciaanalízissel, ahol kategorikus független változókkal magyarázzuk a folytonos függő változókat, a diszkriminancia-analízis esetében folytonos független változók mellett kategorikus függő változókat használunk. A regresszióanalízis pedig abban különbözik a diszkriminancia elemzéstől, hogy esetében magas mérési szintű folytonos függő és független változók szerepelhetnek az elemzésben. Kategorikus független változók esetén az ekvivalens eljárás a megfelelési diszkriminancia-analízis.

A diszkriminancia-analízis központi lépése a diszkrimináló függvény(ek) kiszámítása. A szükséges diszkrimináló függvények száma úgy számítható ki, hogy a függő változó lehetséges kimeneteleinek száma -1 és a független változók száma közül a kisebbet kell venni. Ha tehát

két csoportunk van és két folytonos független változók (ú.n. prediktor változók), akkor egy diszkrimináló függvényünk lesz. Ellenben négy csoport és két folytonos prediktor változó esetén kettő. A diszkrimináló függvény általános képlete az alábbi:
$D_j = d_{0j} + d_{1j}x_1 + d_{2j}x_2 + ... + D_{kj}x_k$,
ahol j az adott diszkrimináló függvény sorszáma, az x_i-k a mért független változók, d_0 konstans, a d_{ij} az x_i mért változók j-edik diszkrimináló függvényhez tartozó együtthatója. A függvény akkor optimális, ha a függő változó által meghatározott csoportok közötti külső négyzetösszeg és a csoportokon belüli négyzetösszeg hányadosa maximális. A négyzetösszeg a varianciaanalízisben használt heterogenitást kifejező átlagos négyzetes eltérést jelenti. A Wilks-féle lambda, amely a csoportokon belüli átlagos négyzetes eltérés és teljes átlagos eltérés aránya, megadja a diszkrimináló függvény jóságát. Értéke egy 0 és 1 közötti szám. 0-hoz közelítő értékek esetén a csoportokon belüli variabilitás kicsi, ami azt is jelzi, hogy függvényünk jól diszkriminál a csoportok között. Ezzel szemben az 1 közeli érték azt jelzi, hogy a csoporton belüli négyzetösszeg közel áll a teljes négyzetösszeghez, és így a csoportok közötti négyzetösszeg kicsi, ami azt mutatja, hogy a függvény kevéssé tudja a csoportokat jól megkülönböztetni.

A marketing területén a diszkriminancia-analízis felhasználható arra, hogy egy empirikusan összegyűjtött adatsor alapján meghatározzuk, mely faktorok különítik el a vásárlókat vagy termékeket két vagy több csoportra. Manapság erre a célra a logisztikus regresszió szélesebb körben használt eljárás. A diszkriminancia-analízis használata a marketingben az alábbiak szerint foglalható össze:

A kutatási kérdés megfogalmazása és az adatok összegyűjtése. Először meg kell határozni azokat a kitüntetett jellemzőket, amelyek alapján a vásárlók értékelik a terméket. Ezután kvantitatív marketing technikákkal (pl. kérdőíves felmérésekkel) a potenciális vásárlók egy csoportján fel kell mérni az adott termék minden lényeges sajátosságát. Ez az adatgyűjtési szakasz általában marketing szakemberekre hárul. A kérdőívben 1-től 5-ig (vagy 7-ig ill. 10-ig) kell értékelni a terméket több (átlagosan 5 és 20 közötti) sajátosság tekintetében. Ezek az alábbiak lehetnek: a használat egyszerűsége, súly, pontosság, tartósság, szín, ár vagy méret, stb... Az értékelési szempontok nagyban függenek a termék mibenlététől. Ugyanezen szempontok alapján értékelik a potenciális vásárlók a többi terméket. Az adatokat lekódolják, beviszik egy statisztikai programba, mint pl. R, SPSS vagy a SAS. (Eddig a lépésig a teendők megegyeznek a faktor-analízissel.)
A diszkrimináló függvények kiszámítása és statisztikai szignifikancia és validitás meghatározása. A diszkrimináló függvények olyan függvények, amelyek a legnagyobb különbséget produkálják a kategorikus függő változó által definiált csoportok között. Az első lépés a megfelelő diszkriminancia-analízis eljárás kiválasztása. A közvetlen módszer esetében a független változók egyszerre kerülnek be az eljárásba, és így számítjuk ki a diszkrimináló függvényeket. A lépcső-módszer esetében egymást követően kerülnek be a független változók a modellbe. A kétmodelles eljárás alkalmazandó abban az esetben, ha függő változónak két szintje van. A többszörös diszkriminancia-analízis pedig akkor szükséges, ha három vagy több függő változónk van. Az SPSS-ben a Wilks-féle lambda, míg a SAS programban az F-statisztika tájékoztat bennünket a szignifikanciáról. A diszkrimináló függvények alapján a

statisztikai programok lehetőséget biztosítanak arra, hogy a függő változó csoportjaiba tartozó személyeket vagy termékeket újra klasszifikáljuk, így tesztelve, hogy a diszkriminancia-analízisben kialakított magyarázótér mennyire hatékony.
Kétszintű függő változó esetén felrajzolhatóak az eredmények egy kétdimenziós diagramon. A termékek vagy csoportok távolsága jelzi, hogy azok mennyire különbözőek. A diagram tengelyeit a kutatóknak kell elnevezniük. Az ábra értelmezése sokszor erősen szubjektív.

```
DISCRIMINANT
/GROUPS=CLU3_1(1 3)
/VARIABLES=K1_1 K1_2 K1_3 K1_4 K1_5 K2_1 K2_2 K2_3
/ANALYSIS ALL
/SAVE=CLASS
/PRIORS EQUAL
/STATISTICS=MEAN STDDEV UNIVF BOXM CORR TABLE CROSSVALID
/PLOT=COMBINED
/CLASSIFY=NONMISSING POOLED.
```

10.1. ábra: **A Diszkriminancia analízis lefuttatásához szükséges parancssor bemutatása**

Az első táblázat „Group Statistics” az elemzésbe bevont összes változó csoportok szerinti és összesített átlagát, szórását, súlyát mutatja.

Group Statistics

3 klaszter (Hierarchikus)		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
1	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	4,47	,608	201	201,000
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	4,05	,904	201	201,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	3,93	,636	201	201,000
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	3,76	,430	201	201,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	3,95	1,021	201	201,000
	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	3,90	,837	201	201,000

	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	4,13	,723	201	201,000
	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	3,40	,996	201	201,000
2	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	3,43	,899	417	417,000
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	3,00	1,116	417	417,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	4,15	,593	417	417,000
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	4,28	,460	417	417,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	3,72	,932	417	417,000
	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	2,50	1,181	417	417,000
	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	3,32	1,194	417	417,000
	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	2,24	,969	417	417,000
3	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	3,42	1,145	288	288,000
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	2,69	,875	288	288,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	3,21	,969	288	288,000
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	2,58	,703	288	288,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	2,58	,915	288	288,000

	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	2,98	1,022	288	288,000
	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	3,55	1,137	288	288,000
	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	2,51	,875	288	288,000
Total	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	3,66	1,028	906	906,000
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	3,13	1,119	906	906,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	3,80	,849	906	906,000
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	3,62	,918	906	906,000
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	3,41	1,107	906	906,000
	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	2,96	1,192	906	906,000
	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	3,57	1,131	906	906,000
	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	2,59	1,048	906	906,000

10.2. ábra: **A változók leíró statisztikáját bemutató táblázat**

Az ezt követő 10.3. ábrán feltüntettük azt a táblázatot, amelyben meg tudjuk vizsgálni, hogy a független változók milyen mértékben járulnak hozzá a létrejövő függvényhez. A változók szignifikáns voltának tesztelésére az F-érték mellett, a Wilks'-Lambda statisztika is szerepel.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	,823	97,280	2	903	,000
Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	,794	116,818	2	903	,000
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	,761	142,100	2	903	,000
Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,351	836,640	2	903	,000
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	,731	166,162	2	903	,000
Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	,793	117,580	2	903	,000
Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	,923	37,718	2	903	,000
Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	,814	103,238	2	903	,000

10.3. ábra: **A független változók hozzájárulásainak eredményei**

Látható, hogy minden változónak szignifikáns hatása van. A Wilks'-Lambda értéke 0 és 1 közé eső értékek, melyek közül a mindig a nullához közeli értékekhez tartozó változóknak (*Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni*) van a legjelentősebb hatása a diszkriminancia függvényre.

A következő 10.4. ábrán egy alapfelvetés tesztelése történik meg. A Pooled Within-Groups Matrices táblázat eredményei segítségével a multikollinearitást teszteljük.

Pooled Within-Groups Matrices

		Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben
Correlation	Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	1,000	,510	,331	,332	,341	-,023	,031	-,153
	Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	,510	1,000	,191	,183	,300	-,074	-,133	-,160
	Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	,331	,191	1,000	,510	,278	-,023	,022	-,119
	Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,332	,183	,510	1,000	,438	,103	,113	,078

10.4. ábra: **A multikollinearitás tesztelésének adatai**

A következő fontos táblázat (Eigenvalues), mely során először kapunk információt a keletkező függvényről. Megfigyelhető, hogy két függvény keletkezett. A függvények számát megállapíthatjuk, ha a csoportok száma, illetve a független változók száma közül a kevesebbikből egyet kivonunk. A két függvény fontosságának megállapításában, a sajátérték segíti a kutatót. A táblázat sajátértékei és magyarázott variancia értékei alapján az első függvény lesz fontosabb számunkra. A kanonikus korreláció (0,834) azt jelenti, hogy az adott függvény igen számottevő részt magyaráz a teljes varianciából. A kapott értékek négyzete megmutatja, hogy a függő változó varianciájának, hány százalékát magyarázzák a független változók csoportja (70%).

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2,276[a]	76,6	76,6	,834
2	,694[a]	23,4	100,0	,640

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,180	1541,596	16	,000
2	,590	474,277	7	,000

10.6. ábra: **A sajátérték és a Wilks'-Lambda eredmények bemutatása**

A megjelenő Wilks'-Lambda táblázat a függvények szignifikanciájának tesztelését végzi. Láthatóan mindkét függvény szignifikáns, de az első hatása jelentősebb.

A korrelációs együttható mátrixa (Structure Matrix) hasonlóan értelmezendő, mint a faktoranalízisnél a Component Matrix, hiszen a független változók és a diszkriminancia függvények közti, csoportonkénti átlagolt (Pooled within groups) Pearson féle lineáris korrelációk.

Structure Matrix

	Function	
	1	2
Mennyire ért egyet a következő kijelentéssel? A környezetbarát termékekért hajlandó vagyok többet fizetni	,874*	,405
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek	,353*	,211
Mennyire ért egyet a következő kijelentéssel? A zöld csomagolóanyagok használata pozitív hatással van a környezetre	,036	,607*
Mennyire ért egyet a következő kijelentéssel? A palackozott ásványvíz vásárlása negatívan hat a környezetre	-,042	,552*
Mennyire ért egyet a következő kijelentéssel? Szeretem a nagy luxust az életemben	-,129	,524*
Mennyire ért egyet a következő kijelentéssel? Csodálom azokat, akiknek drága kocsija, lakása vagy ruhája van	-,176	,523*
Mennyire ért egyet a következő kijelentéssel? Hajlandó vagyok több adót fizetni azért, hogy védjem a lakóhelyem környezetét	,313	,458*
Mennyire ért egyet a következő kijelentéssel? Boldogabb lennék, ha több minden dolgot meg tudnék venni	-,089	,307*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

10.7. ábra: **A struktúra mátrix eredményeinek bemutatása**

Ez alapján az első függvény a „környezetbarát termékekért hajlandó vagyok többet fizetni" és a „hajlandó vagyok gyengébb anyagminőséget elfogadni azért, hogy környezetbarát legyek"

szempontokat, míg a második az összes többit foglalja magában, mely alapján a kutató el tudja nevezni a dimenziókat (hasonlóan a faktorelemzéshez).

A következő 10.8. ábrán a csoportok középpontértékeit mutatjuk be. Megállapíthatjuk, hogy második és a harmadik csoport (klaszter) magas értékekkel rendelkezik az első dimenzióban, míg az első klaszter magas értékei a második dimenzió mentén jelentkeznek. Ezeket a koordinátákat fogja a program felhasználni a grafikus megjelenítéskor is.

Functions at Group Centroids

	Function	
3 klaszter (Hierarchikus)	1	2
1	-,436	1,539
2	1,526	-,318
3	-1,905	-,614

Unstandardized canonical discriminant functions evaluated at group means

10.8. ábra: **A csoportok középpont értékei**

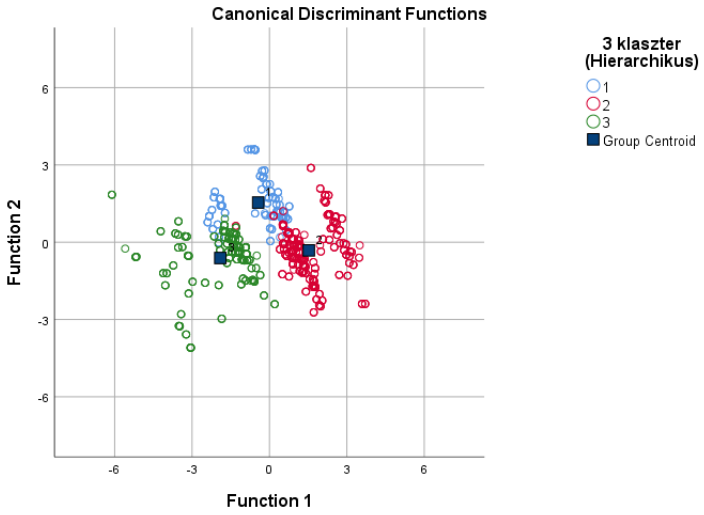
Az elemzés folytatásaként a program kiszámolja az un. klasszifikációs statisztikai mutatókat. Az egyik ilyen eredmény a következő ábrán látható. Látszik, hogy a csoportokba kerülés esélye 33,3 százalék volt.

Prior Probabilities for Groups

		Cases Used in Analysis	
3 klaszter (Hierarchikus)	Prior	Unweighted	Weighted
1	,333	201	201,000
2	,333	417	417,000
3	,333	288	288,000
Total	1,000	906	906,000

10.9. ábra: **A klasszifikációs statisztika eredményei**

A következőkben a grafikus ábrázolás történik, ahol a tengelyek maguk a függvények (dimenziók). A 10.10. ábra az analízisbe bevont egyedek értékeit és a centrumközéppontokat ábrázolja.



10.10. ábra: **A független változók hozzájárulásainak eredményei**

A helyesen kategorizált csoporttagságok arányát a klasszifikációs eredmények elnevezésű táblázatban (Classification Results) láthatjuk.

Classification Results[a,c]

			Predicted Group Membership			
		3 klaszter (Hierarchikus)	1	2	3	Total
Original	Count	1	173	7	21	201
		2	18	399	0	417
		3	4	3	281	288
	%	1	86,1	3,5	10,4	100,0
		2	4,3	95,7	,0	100,0
		3	1,4	1,0	97,6	100,0
Cross-validated[b]	Count	1	173	7	21	201
		2	18	399	0	417
		3	4	3	281	288
	%	1	86,1	3,5	10,4	100,0
		2	4,3	95,7	,0	100,0
		3	1,4	1,0	97,6	100,0

a. 94,2% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 94,2% of cross-validated grouped cases correctly classified.

10.11. ábra: **A klasszifikációs eredmények**

A táblázat alján láthatjuk, hogy a modell 94,2%-ban tudta helyesen kategorizálni a megadott független változó mentén.

11. Conjoint-analízis

A conjoint-analízist a marketingkutatás használja a leggyakrabban, és igen intenzíven alkalmazzák a ’70-es évek óta. Feladata a termékekkel, szolgáltatással kapcsolatos fogyasztási szokások, elvárások egységes skálán való megjelenítése. A szolgáltatás-, terméktervezés szempontjából olyan alapvető kérdésekre adja meg a választ, mint pl.:

- Egy adott terméknek vagy szolgáltatásnak milyen tulajdonságai (attribútumai) fontosak (illetve nem fontosak) a fogyasztók számára?
- A tulajdonságok egyes szintjeit (értékeit) hogyan preferálja a fogyasztó?
- A tulajdonságok kombinációit (a termékváltozatokat) hogyan értékelik a fogyasztók?
- Milyen piaci szegmenseket lehet elkülöníteni?
- Milyen lesz az adott szolgáltatás vagy termék várható piaci részesedése, ha változtatunk a tulajdonságain?

11.1. A conjoint-analízis lépései

1. termék vagy szolgáltatás tulajdonságainak, a tulajdonságok szintjeinek meghatározása,
2. tervezett termékek, szolgáltatások meghatározása a tulajdonságok kombinációja alapján, tervkártyák előállítása,
3. adatgyűjtési módszer megválasztása,
4. az adatgyűjtés (megkérdezés) végrehajtása,
5. conjoint-analízis,
6. a modell jóságának mérése (validálás),
7. az eredmények értelmezése, mérlegelése,
8. becslés a modell segítségével, „mi lenne, ha...” típusú kérdések megválaszolása.

Az analízis során a tulajdonságokat faktoroknak (angolul factor), ezek konkrét értékeit szinteknek (angolul level) és a megkérdezettek válaszait preferenciának (angolul preference) fogjuk nevezni. A conjoint-eljárás során a faktorszintekhez regressziós eljárással rendelünk értékeket úgy, hogy a faktorkombinációkhoz tartozó eredmények vagy hasznosságok a kialakult eredeti preferenciákhoz a legjobban illeszkedjenek. Gyakorlatilag a megfigyelt és becsült preferenciák közötti különbségek négyzetösszegét fogjuk minimalizálni. A preferenciákat függő, a faktorszinteket független változónak tekintjük.

Abban az esetben, ha csak nominális faktorok szerepelnek a vizsgálatban, többtényezős variancia-analízissel is meghatározhatjuk a hasznosságokat. Ilyenkor a variancia-analízis eredményei közül a hatásokat (angolul effect) kell figyelembe venni. A mostani, és a későbbiekben tárgyalt módszerek zöme egyetlen nagy családba foglalhatók össze az általános lineáris modellekbe (GLM). Az R programban a GLM függvény használható erre, természetesen a paraméterek pontos megadása után. A különböző eljárások főként az alapadatmátrix előállításában térnek el. Az R speciális csomagjai a tervezést, az alapadatmátrix előállítását, és az eredmények megjelenítését segítik, a lényegi számítást azonban a GLM végzi.

A faktorok értékeik jellegének megfelelően az alábbiak lehetnek:

- **Diszkrét:** az értékek kategorizált vagy ordinális skáláról származnak, és nincs előzetes információnk, a faktorszintek és a preferenciák kapcsolatáról.
- **Lineáris:** a faktorszintek és a preferenciák közötti kapcsolat lineáris, és az adatok legalább intervallum szintűek. Ekkor a faktorszinteket folyamatos változónak tekintjük.
- **Négyzetes:** másodfokú kapcsolatot feltételezünk a faktorszintek és preferenciák között. Ebben az esetben a faktornak legalább három szinttel kell rendelkeznie.

Diszkrét faktorértékeknél nincs semmilyen előzetes feltételezésünk a faktorszintek és preferencia közötti összefüggésre. Lineáris kapcsolatnál a faktorértékeket legalább ordinális skálán kell értelmezni, mert csak ekkor állíthatjuk, hogy egyenes vagy fordított lineáris összefüggés van a két változó között. Négyzetes faktorszinteknél kétféle elképzelésünk lehet. Az egyik, hogy a faktornak van egy ideális szintje, és ettől az ideális ponttól bármelyik irányba távolodva a preferencia csökken. Konkáv parabola. A másik, hogy van egy legrosszabb szintje a faktornak, és ettől a legrosszabb ponttól távolodva a preferencia nő. Konvex parabola.

A tervezett termékeket, szolgáltatásokat ebben a módszerben hagyományosan tervkártyáknak nevezzük, mivel az adatgyűjtés során ezeket kártyákra kinyomtatva adták oda a megkérdezetteknek. A conjoint-analízishez szükséges tervkártyákat, faktorkombinációkat sokszor ortogonális polinomok segítségével állítjuk elő. Az eljárás az összes lehetséges kombináció (full factorial design) helyett redukált faktoriális terveket (replikációk) készít. A tervezés során a faktorkombinációk számát és az óhatatlanul fellépő információveszteséget minimalizálják. E két kívánalom ellentétesen alakul, tehát a kombinációk számának csökkentésével az információveszteség nő. Szerencsére az összefüggés nem lineáris, az információveszteség lassabban nő, mint a kártyák számának csökkenése.

A tervkártyáknak három típusát különböztetjük meg:

- A **közönséges kártya** részt vesz a becslésben és a modell jóságának megítélésében. Szigorú értelemben ezt nevezzük tervkártyának, angolul „design card”.
- A **visszatartott kártya** (angolul holdout) nem vesz részt a becslésben csak a modell jóságának megítélésében. Ezek a kártyák szerepelnek a felmérésben, a megkérdezett véleményét tudjuk.
- **Szimulált kártya**. Ez nem szerepel a felmérésben, nem rendelkezünk a megkérdezettek véleményével, a modell paraméterei alapján becsüljük a hasznosságukat.

Az adatgyűjtés módja:

Szekvenciális adatgyűjtéskor a megkérdezett személyek sorba rendezést végeznek a tervkártyákkal. Ez a leggyakrabban alkalmazott módszer. A gyakorlatban a megkérdezett kiválasztja a kártyák közül a legkedveltebbet, a maradékból megint a legkedveltebbet és így tovább. Az adatfájlban a kártyák azonosítói szerepelnek. Az adatfájl változói: legkedveltebb kártya (first) … legkevésbé kedvelt kártya (last).

- **Rangszámok** alapján. Az adatfájlban minden adat egy rangszám. A változók: első kártya (RANK1)…utolsó kártya (RANKn).

- **Pontozásos**. Az adatfájl ebben az esetben pontszámokat tartalmaz. Például a megkérdezetteknek 1-től 100-ig terjedő skálán kell értékelni az adott termékeket. A magasabb pontszám nagyobb kedveltséget jelent. Az adatfájl változói: első kártya (SCORE1) … utolsó kártya (SCOREn)

A számítógépes statisztikai programok sokszor a rangszámok inverze alapján számolnak. Ilyenkor sorba rendezés történik, a legjobb kombináció kapja az 1 rangszámot, a legrosszabb az *n* rangszámot. Az adatfájl készítése során viszont megfordítjuk a sorrendet, és a legjobb kapja a legnagyobb, a legrosszabb az 1 rangszámot. Így a termék vagy szolgáltatás hasznossága a rangszámok inverzének növekedésével arányos. Ez az eredmények értelmezését egyértelművé teszi.

A conjoint-analízishez tehát minimum két adatfájlra lesz szükségünk.

- Az egyik a **tervfájl**, amely a tervkártyákat tartalmazza, és nem lehet benne üres cella. A faktoroknak minimum két szinttel kell rendelkezniük, maximális számuk 99 lehet.
- A másik a **preferenciafájl**, amely a megkérdezettek véleményét tartalmazza. Erre is érvényes, ha bármelyik preferenciaérték hiányzik (rangszám, súly, vagy kártyaszám), a megfigyelés ki lesz zárva az analízisből.

A modell jóságát a becsült és a mért preferenciák közötti korrelációval ellenőrizzük. Meghatározhatjuk a szorzatmomentum és rangkorrelációs mutatókat is. A Pearson-féle mutatót szakmailag helyesen csak akkor lehet meghatározni, ha a preferenciák skála típusú adatok. Esetleg akkor, ha pontozzák a termékeket, és erről azt feltételezzük, hogy közel skála típusú változó. A Likert-skáláról sokszor tévesen ezt feltételezik. Sokszor a preferencia mérőszámát nem tekinthetjük skála típusú adatnak, mivel a megkérdezettek legtöbbször csak sorba rendezik a termékeket vagy szolgáltatásokat. Ezért a modelljóság mérésére leggyakrabban a Kendall-féle rangkorrelációt kell alkalmazni.

A modell jóságának megítélését, a korrelációs együttható meghatározását úgy is elvégezhetjük, hogy a conjoint-analízisben részt nem vevő termékekre vagy szolgáltatásokra (visszatartott kártyák) becsüljük meg a preferenciákat, és ezt korreláltatjuk a mért preferenciákkal. Az így számított korreláció gyakorlatiasabb, hihetőbb eredményt ad. Természetes ez a modell jóságának megítélésére egy sokkal szigorúbb mérőszám.

Az elmélet után nézzünk egy gyakorlati példát.

Példa:

Tekintsük példának egy rizsfelmérést. A rizs három jellemző attribútuma (faktor):

- Származási hely: India, Kína, Magyarország
- Gazdálkodás módja: hagyományos, vegyszer nélküli, organikus
- Ár: 400, 500, 600 Ft

Amint látjuk minden tulajdonságnak három szintje (angolul level) van. A lehetséges kombinációk száma: $3x3x3 = 27$. Amennyiben minden kombinációt meg szeretnénk kérdezni 27 termékről kellene információt gyűjteni. Ez túlságosan sok, ezt nem várhatjuk el a

megkérdezettektől. Ezért a termékek (kártyák) számát csökkenteni kell. Az összes lehetséges kombinációból ortogonális polinomok segítségével célszerű kiválasztani a kérdőíveken szereplő termékeket. Ezek lesznek a tervkártyák (termékek), ezeket kell a megkérdezetteknek pontozniuk $1-n$-ig. A legkedveltebb termék n pontot, a legrosszabb 1 pontot kap. Ezeket így preferenciának is tekinthetjük. Amennyiben rangsorolják a termékeket, akkor a rangszámok inverzét kell képezni.

Az elemzést az R program conjoint csomagján keresztül mutatom be. Ezt a csomagot két lengyel kutató, Andrzej Bak és Tomasz Bartlomowicz készítette, akik a Wroclawi Közgazdasági Egyetem Ökonometria és Számítástudományi Tanszékén dolgoznak.

A teljes faktoriális terv elkészítése az R környezetben.

```
> library(conjoint)
> experiment = expand.grid(
+   Region = c("India","Kína","Magyar"),
+   Cultivation = c("Conv", "NoChem", "Organic"),
+   Price = c("400", "500", "600"))
```

Az R-ben az expand.grid() függvénnyel állíthatunk elő teljes faktoriális kísérleti terveket. A függvény paraméteriben meg kell adni a termék jellemzőit és szintjeit. Ezután a függvény előállítja az összes lehetséges kombinációt.

```
> experiment
   Region Cultivation Price
1   India        Conv   400
2    Kína        Conv   400
3  Magyar        Conv   400
4   India      NoChem   400
5    Kína      NoChem   400
6  Magyar      NoChem   400
7   India     Organic   400
8    Kína     Organic   400
9  Magyar     Organic   400
10  India        Conv   500
11   Kína        Conv   500
12 Magyar        Conv   500
13  India      NoChem   500
14   Kína      NoChem   500
15 Magyar      NoChem   500
16  India     Organic   500
17   Kína     Organic   500
18 Magyar     Organic   500
19  India        Conv   600
20   Kína        Conv   600
21 Magyar        Conv   600
22  India      NoChem   600
23   Kína      NoChem   600
24 Magyar      NoChem   600
25  India     Organic   600
26   Kína     Organic   600
27 Magyar     Organic   600
```

11.2. Tervkártyák készítése ortogonális módszerrel

Az ortogonalitás azt jelenti, hogy az attribútumok szintjei lineárisan függetlenek egymástól, azaz egyik sem állítható elő a másik vagy másikok lineáris kombinációjából. Ezt le is fogjuk ellenőrizni.

```
> design<-caFactorialDesign(data=experiment, type="orthogonal")
> print(design)
   Region Cultivation Price
2    Kína        Conv   400
6  Magyar      NoChem   400
7   India     Organic   400
12 Magyar        Conv   500
13  India      NoChem   500
17   Kína     Organic   500
19  India        Conv   600
27 Magyar     Organic   600
```

Az R-ben frakcionált kísérleti terveket a caFactorialDesign() függvénnyel állíthatunk elő. A type attribútummal szabályozhatjuk a replikációk készítésének módját. A type értéke lehet: full, fractional, ca, aca, orthogonal vagy, null.

Összesen 8 termékkombinációt kaptunk a 27-ből. Az első oszlopban látjuk, hogy melyik termékek kerültek a tervkártyák közé. Ennek ismeretében választhatjuk ki majd a szimulált kártyákat, melyek nem kerültek megkérdezésre. A tervkártyákat fogjuk felhasználni a felmérés során. Régen ezeket a kártyákat kinyomtatták, és sorba rendezést kértek a megkérdezettektől. Az értékelés során ilyenkor a rangszámok inverzét kell képezni. Érdemes ezt a kísérleti tervet elmenteni, mert az R újbóli futtatása során más megoldást is kaphatunk. Pl. ebben a példában a másik megoldással 9 tervkártyát is kaphatunk.

A termékek szövegesen vannak jellemezve, ez így nem alkalmas a számítási műveletek elvégzésére. Át kell kódolni numerikus mátrixszá.

```
> profile<-caEncodedDesign(design)
> print(profile)
   Region Cultivation Price
2       2           1     1
6       3           2     1
7       1           3     1
12      3           1     2
13      1           2     2
17      2           3     2
19      1           1     3
27      3           3     3
```

A caEncodeDesing() függvény kódolja át a tervet numerikus mátrixszá. Most már le tudjuk ellenőrizni, hogy teljesül-e a lineáris függetlenség, azaz ortogonalitás.

```
> print(cor(profile))
            Region Cultivation Price
Region           1           0     0
```

```
Cultivation     0         1     0
Price           0         0     1
```

A cor() függvénnyel állítjuk elő a korrelációs mátrixot. Csak a főátlóban vannak egyesek, a többi elem nulla, tehát teljesül a függetlenség.

A „profile” mátrixot is érdemes egy fájlba kimenteni, mert az R újabb futtatása esetén más megoldást is kaphatunk.

```
> write.csv2(profile,"profile_rizs.csv",row.names = FALSE)
```

A row.names paraméterrel letiltottam a sorneveket, így a mentett mátrixnak 8 sora és 3 oszlopa lesz.

Ezek után el kell végezni a felmérést. A felmérésünkben, száz fő töltötte ki a kérdőíveket, ezért a preferencia-mátrixnak 100 sora és 8 oszlopa van. Minden sor egy megkérdezett preferenciáját tartalmazza. A sorrend nagyon fontos, az első elem az első kártya pontszáma, a nyolcadik az utolsóé. A kártyák sorrendje a profile-ban látható.

A preferencia-mátrix részlete látható lent.

```
    p.1 p.2 p.3 p.4 p.5 p.6 p.7 p.8
1     2   5   3   7   1   4   8   6
2     6   4   5   2   1   8   3   7
3     2   5   3   7   4   1   6   8
4     7   4   2   1   6   5   8   3
5     2   4   1   6   7   8   3   5
6     3   7   1   6   4   8   2   5
7     5   1   8   6   4   7   3   2
8     3   5   1   4   6   7   2   8
9     4   5   3   6   2   1   7   8
10    7   1   8   6   2   3   5   4
.
.
.
98    1   6   4   7   5   8   3   2
99    1   2   6   8   7   3   4   5
100   3   4   8   5   6   2   7   1
```

Az elemzés során a leggyakoribb hiba, hogy a preferencia-mátrix nem olyan sorrendben tartalmazza a pontszámokat, mint a profile.

Még egy fájlra lesz szükségünk, hogy futtatni tudjuk az R-ben a conjoint-analízist, a szintek fájlra. Ez egy oszlopvektorban tartalmazza a tulajdonságok szintjeit, folyamatosan felsorolva, és csak az eredmények megjelenítéséhez szükséges.

```
> szintek=c("India","Kína","Magyar","Hagyományos", "Nincs_vegyszer",
 "Organikus","400", "500", "600")
```

Ezek a nevek fognak megjelenni az eredménylistában és az ábrákon. A három fájl birtokában most már futtathatjuk a conjoint-analízist. Az R függvény az alábbi:

```
> Conjoint(prefm,profile,szintek)
```

A függvény paraméterei sorrendben: a preferencia-mátrix, a tervkártyák és a jellemzők szintjei. Ez a függvény egy gyűjtőfüggvény, azaz több önálló függvényt futtat egymás után. Ezek a függvények az alábbiak: caPartUtilities, caUtilities és caImportance.

A példánkban minden attribútumnak három szintje (angolul level) van. Ezekből a program mesterséges (angolul dummy) változókat képez, és ezekkel becsüli meg a tulajdonságok jelentőségét (angolul importance), illetve a szintjeinek hasznosságát (angolul utility).

Gyakorlatilag egyszerű többszörös lineáris regresszió analízissel határozzuk meg a mesterséges változók együtthatóit. Függő változó (y) pontszám (angolul score), független változók a mesterséges változók. A becsült és megfigyelt pontszámok eltérés-négyzetösszegeit fogjuk minimalizálni. Az együtthatókat a főátlagra szimmetrikusan kapjuk meg. Ez azt jelenti, hogy az összegük nullát fog adni. A conjoint-analízis megoldása:

```
Call:
lm(formula = frml)

Residuals:
   Min     1Q Median     3Q    Max
-3,278 -1,798 -0,215  1,820  4,258

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4,28500    0,08987  47,682  < 2e-16 ***
factor(x$Region)1    0,40833    0,11602   3,520 0,000457 ***
factor(x$Region)2   -0,80333    0,14675  -5,474 5,90e-08 ***
factor(x$Cultivation)1  0,21833    0,11602   1,882 0,060212 .
factor(x$Cultivation)2 -0,19667    0,14675  -1,340 0,180580
factor(x$Price)1     0,57833    0,11602   4,985 7,61e-07 ***
factor(x$Price)2     0,28167    0,11602   2,428 0,015411 *
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 2,201 on 793 degrees of freedom
Multiple R-squared:  0.0664,    Adjusted R-squared:  0.05934
F-statistic:   9.4 on 6 and 793 DF,  p-value: 5,524e-10

[1] "Part worths (utilities) of levels (model parameters for whole sample):"
levnms    utls
1       intercept   4,285
2           India  0,4083
3            Kína -0,8033
4          Magyar   0,395
5     Hagyományos  0,2183
6  Nincs_vegyszer -0,1967
7       Organikus -0,0217
8             400  0,5783
9             500  0,2817
10            600   -0,86
[1] "Average importance of factors (attributes):"
[1] 56,68 24,25 19,07
```

[1] Sum of average importance: 100
[1] "Chart of average factors importance"

Az eredménylista első sora mutatja, hogy egyszerű többszörös lineáris regresszióval határoztuk meg az együtthatókat (lm, azaz lineáris modell). A következő részben a modell maradékainak (angpolul residuals) egyszerű leíró statisztikája, kvartilisei következnek. A modell akkor jó, ha a maradék normális eloszlású, várható értéke nulla, és homoszkedasztikus. Amennyiben a maradék normáliseloszlású, akkor a várható érték megegyezik a mediánnal. A mediánnak tehát nulla közelébe kell lennie.

A conjoint-analízis tényleges megoldását a Coefficients rész tartalmazza, ez mutatja meg a tulajdonságok szintjeinek együtthatóit. A megoldásban a tulajdonságok szintjei mindig eggyel kevesebb, mint a tényleges szintek száma, azért, hogy a modell ne legyen túlparaméterezett. Az Estimate oszlopban találjuk az együtthatókat, a Std. Error oszlopban az együttható standard hibáját, ez után a t-próba statisztikáját, és végül az elsőfajú hiba elkövetésének valószínűségét. A Cultivation2 kivételével minden együttható szignifikáns 10%-on.

A maradékok standard hibája 2,2, szabadságfoka 793. A többszörös R-négyzet: 0,0664, a korrigált értéke 0,05934. Ezek nagyon alacsony értékek, a lineáris modell nem illeszkedik jól a 100 válaszadó preferenciájára. A regresszió F-statisztikája 9,4 6 és 793 szabadságfok mellett a p-érték 5,524e-10. A regresszió tehát létezik, csak az illeszkedés nagyon rossz.

A [1] Part worths (utilities) of levels (model parameters for whole sample): alatt találjuk a tulajdonságok minden szintjének együtthatóját. Erre érvényes az, ha egy tulajdonság együtthatóit összeadjuk, nullát kapunk.

Melyik jellemzőnek van a legnagyobb befolyása a vásárlási hajlandóságra? Ezt az Importance fogja mutatni. A példánkban a származási hely 56,68%-ban, a termesztés módja 24,25%-ban és az ár csak 19,07%-ban fontos a megkérdezettek szerint. Természetesen ezek összege 100%-t ad.

Az R a számszerű megoldás mellett grafikonon is meg tudja jeleníti a legfontosabb eredményeket.

Olyan esetekben, amikor a jellemzők vagy változók hatásainak összege 100%-t eredményez sokkal látványosabb, ha kördiagramon ábrázoljuk a hatásokat az oszlopdiagram helyett. Gyakorlatilag megoszlási viszonyszámokat ábrázolunk ilyenkor. Az adatbázis kitalált, nem a magyar valóságot tükrözi. Sajnos, hazánkban az ilyen jellegű felmérések a magyar vásárlók árérzékenységét szokták kimutatni, és sokszor a termékek illetve szolgáltatások tényleges, döntő mértékben ható tulajdonságait nem sikerül feltérképezni. Mindegy hogy milyen a termék, csak olcsó legyen.

A tulajdonságok szintjeinek átlagos hasznossága pontszámban kifejezve. Pl. Kína átlagosan 0,8 ponttal csökkenti a rizs kedveltségét. A másik két ország átlagosan 0,4 ponttal növeli a preferenciát.

A hagyományos termesztés növeli a preferenciát, a másik kettő csökkenti. Úgy látszik a megkérdezettek hagyománykedvelők, nem hisznek a újabb alternatív termesztési eljárásokban, vagy nem ismerik őket. Ennek a megállapítása egy másik módszert igényel. A kérdőíven tehát olyan kérdéseket is érdemes feltenni burkoltan, hogy a kérdőívet kitöltők egyáltalán tudják-e, hogy a termék tulajdonságainak különböző szintjei mit jelentenek. Nem valószínű, hogy ismeretlen tulajdonságot keresnének a vásárlás során. Bár néha olyan eredményt is kaptak kérdőívezés során, hogy egy védjegyet senki sem ismert, de a későbbi kérdésekre a válaszolók

többsége úgy válaszolt, hogy a vásárlás során keresik az ilyen védjeggyel rendelkező terméket. Ezért a szavahihetőséget a kérdőíveken valahogy mérni kell.

Az ár preferenciát befolyásoló hatása a magyar valóságot mutatja. A legkisebb árak növelik, a legnagyobb csökkenti a preferenciát. A 600 Ft átlagosan 0,9 ponttal csökkenti a preferenciát.

Hogyan határoztuk meg a jellemzők átlagos fontosságát? Minden válaszadónál megbecsüljük a fontosságot, és ezután átlagoljuk őket. A fontosság az adott jellemző legkisebb és legnagyobb pontszámának különbsége, azaz a terjedelme. Példánkban három tulajdonság terjedelmét kell kiszámítani és a 100 válaszadó értékeit átlagolni. Az R-ben az alábbi függvény megadja a válaszadók hasznosságait.

```
> caPartUtilities(prefm,profile,szintek)
    intercept  India   Kína Magyar Hagyományos Nincs_vegyszer Organikus    400    500    60
0
[1,]     4.833 -1.167  2.667 -1.500      -0.500         1.333    -0.833  1.167  0.167 -1.333
[2,]     4.833 -1.500  2.667 -1.167      -0.833         1.333    -0.500  0.167  1.167 -1.333
[3,]     4.000 -0.667 -2.333  3.000      -0.333        -0.667     1.000  0.667  0.333 -1.000
.
.
.
[98,]    4.833 -1.500  2.667 -1.167      -0.833         1.333    -0.500  0.167  1.167 -1.333
[99,]    4.000 -0.667 -2.333  3.000      -0.333        -0.667     1.000  0.667  0.333 -1.000
[100,]    4.000  0.000 -3.000  3.000       0.667        -1.333     0.667  0.333  0.333 -0.667
```

Az első válaszadó fontosságának meghatározása. Először a tulajdonságok hasznosságainak terjedelmeit számítjuk ki. A legnagyobb értékből levonjuk a legkisebbet.

Származási hely=2,667-(-1,5)=4,167

Termesztés módja=1,333-(-0,833)=2,166

Ár=1,167-(-1,333)=2,5

Ezután a három érték megoszlási viszonyszáma százalékos formában (%):

47,18; 24,52; 28,31.

A fenti számításokat mind a 100 válaszadóra meg kell ismételni, és a kapott eredményeket átlagolni. Ennek az R-kódja az alábbi:

```
# Átlagos fontosság meghatározása **********
> imp=caPartUtilities(prefm,profile,szintek)
# Tulajdonságok terjedelmei
> orsz=apply(imp[,2:4],1,max)-apply(imp[,2:4],1,min)
> tec=apply(imp[,5:7],1,max)-apply(imp[,5:7],1,min)
> ar=apply(imp[,8:10],1,max)-apply(imp[,8:10],1,min)
# Mátrix
> tmp=cbind(orsz,tec,ar)
# Megoszlási viszonyszámok
> imp=prop.table(tmp,1)
```

```
# Átlagos fontosság
> apply(imp,2,mean)
```

Az eredmény:

```
  orsz      tec       ar
0.5668279 0.2424858 0.1906863
```

Ez tökéletesen megegyezik a korábbi számításaink átlagos fontosságával. Érdemes néhány szót mondani az R egyik leghatékonyabb függvényéről, az apply() függvényről. Ezt a függvényt mátrixokon használhatjuk, és minden sorra vagy oszlopra ugyanazt a számítást végezhetjük el vele, pl. kiszámíthatjuk az átlagát a soroknak vagy az oszlopoknak. A függvény első paramétere a mátrix neve, a második dönti el, hogy soronként (1) vagy oszloponként (2) számolunk, és a harmadik az algoritmus, pl. átlag (angolul mean).

A prop.table() függvény egy mátrix sorainak vagy oszlopainak megoszlási viszonyszámait adja eredményül. Itt is a második paraméter dönti el, hogy soronként (1) vagy oszloponként (2) számolunk.

Adott tulajdonság fontosságát gyakran a fizetési hajlandósággal fejezik ki. Mennyit hajlandó a leendő vásárló az adott tulajdonságért fizetni? Pénzben kifejezve, érték alapján sok minden összehasonlítható, így a jellemzők különböző szintjei is. Az ár hasznosságának megítélésével azonban ennél a módszernél óvatosan kell bánni. Az óvatosság a szigorú feltételek miatt indokolt.

A szigorú feltételek: a válaszadók minden tulajdonságot azonos pontossággal és alapossággal ítélnek meg. Ez azt jelenti, hogy például az ár tulajdonságnál ugyanúgy pontoznak, mint a származási helynél, így a pontok megfeleltethetők egymással. Ez nagyon fontos feltételezés. Ha nem teljesül, a fizetési hajlandóság becslése nagyon pontatlan lesz. Ráadásul a fizetési hajlandóságról kapott információk minden termékkombináció esetében ugyanazok. Az árak alakulását ugyanúgy ítéli meg a módszer egy kedvezőtlen kombinációban, mint egy kedvenc termék esetében. Pedig a kedvelt termékek esetében az árak növekedését jobban toleráljuk, mint egy kevésbé kedvelt termék esetén. Ez a módszer sajátossága, mivel a hagyományos conjoint-analízis az attribútumok függetlenségét feltételezi. Az ár hasznosságra gyakorolt hatása viszont nem független a többi jellemzőtől. Ezért sokan az ár kihagyását javasolják az egyszerű conjoint-analízis során.

Amennyiben a rezervációs árat is pontosan szeretnénk meghatározni, akkor a hagyományos conjoint-analízis továbbfejlesztett változatát kell használni (angolul choice based conjoint, CBC).

Térjünk vissza a példánkhoz. Láttuk, hogy a többszörös r-négyzet nagyon alacsony volt, 0,0664. Tehát a regressziós modell nem írja le jól a mért adatokat. Mi lehet ennek az oka? Legtöbbször az adatok heterogenitása. Esetünkben valószínűleg eltérő fogyasztói szokásokkal rendelkező csoportok töltötték ki a kérdőíveket. A hasonló értékítélettel, fogyasztói viselkedéssel bíró egyének csoportját nevezik piaci szegmensnek. Ezek elkülönítéséhez szakmai tudás, illetve előzetes ismeretek szükségesek. Amennyiben sikerül beazonosítani a szegmenseket, mindegyikkel külön-külön el kell végezni a conjoint-analízist, és meg kell határozni a termékjellemzők hasznosságait.

Milyen módszerrel tudjuk elkülöníteni a csoportokat? A legkézenfekvőbb eljárás a klaszter-analízis. Klaszterezzük a megkérdezetteket a nyolc termékkombinációra adott preferenciáik alapján. Az R conjoint-analízisében erre a caSegmentation() függvény szolgál. Paraméterei

sorrendben: preferencia-mátrix, tervkártyák, osztályok száma. Az osztályok száma alapbeállításban 3.

```
> caSegmentation(prefm,profile,3)
K-means clustering with 3 clusters of sizes 37, 24, 39

Cluster means:
     [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
1 4.238838 3.112514 7.000000 4.324324 6.536135 1.761162 7.463865 1.563162
2 7.500000 5.500000 3.500000 3.500000 5.500000 7.500000 1.500000 1.500000
3 2.333333 6.641026 4.666667 7.025641 2.666667 3.307692 2.384615 7.333333

Clustering vector:
[1] 2 2 3 3 3 1 1 1 2 2 3 3 3 1 1 1 2 2 3 3 3 1 1 1 2 2 3 3 3 1 1 1 2 2 3 3 3
[38] 1 1 1 2 2 3 3 3 1 1 1 2 2 3 3 3 1 1 1 2 2 3 3 3 1 1 1 2 2 3 3 3 1 1 1 2 2
[75] 3 3 3 1 1 1 3 3 1 1 1 2 2 3 3 3 1 3 3 1 1 1 2 2 3 3

Within cluster sum of squares by cluster:
[1] 112.59462  44.00002 196.15385
(between_SS / total_SS =  90.5 %)

Available components:

[1] "cluster"     "centers"     "totss"       "withinss"    "tot.withinss"
[6] "betweenss"   "size"        "iter"        "ifault"
```

Az eredmények alapján tehát a három osztályba 37, 24 és 39 fő tartozik. A három csoport ráadásul még nem is azonos nagyságú, tehát az adatbázis ebből a szempontból nem kiegyensúlyozott.

Az osztályátlagokat a Cluster means mutatja. Ebből nagyjából be lehet azonosítani a piaci szegmenseket. Ahol a legnagyobb az átlag, az jellemző a klaszterre. Ehhez elő kell venni a tervkártyákat, és beazonosítani az osztályokat.

	Region	Cultivation	Price
2	Kína	Conv	400
6	Magyar	NoChem	400
7	India	Organic	400
12	Magyar	Conv	500
13	India	NoChem	500
17	Kína	Organic	500
19	India	Conv	600
27	Magyar	Organic	600

Tekintsük az első klasztert. A legnagyobb értékek a 3., 5. és 7. oszlopban találhatók. A tervkártyákon ez az indiai rizs jelenti. Tehát az első osztály az indiai rizst keresők piaci szegmense. A második klaszterben a legnagyobb értékek a 1. és 6. oszlopban találhatók. Ez a kínai rizst jelenti. A második csoportba a kínai rizst kedvelők tartoznak. A harmadik klaszterbe pedig a magyar rizst fogyasztják szívesen. Megkapjuk a klasztervektort is, ezzel az eredeti adatbázist három részre tudjuk bontani. A következő részben a klaszterezés jóságáról kapunk információt. Az összes variancia a klaszterek között 90,5%-ban csoportosul, ez nagyon jó

elkülönítést jelent. Még egyszer felhívom a figyelmet, hogy az adatbázis kitalált, a módszer bemutatását szolgálja.

Természetesen egy kérdőívben nem csak a conjoint-analízishez szorosan illeszkedő kérdéseket tesszük fel, hanem egyéb, az árnyaltabb elemzéshez szükséges kérdéseket is. Ezek között mindig szerepelnek szociodemográfiai adatok. Ebben a példában csak egyetlen ilyen szerepel, az iskolai végzettség. Természetes az éles kérdőívben, az előzetes szakirodalmi kutatások alapján, a termék vagy szolgáltatás vásárlását döntő mértékben befolyásoló tényezőkre is rá kell kérdezni.

Az eredeti preferencia-mátrixhoz kapcsoljuk hozzá a klaszter-vektort, amely megmutatja, hogy a válaszadó melyik osztályhoz tartozik.

```
> clus=caSegmentation(prefm,profile,3)
> cbind(prefm,clus$cluster)
   r.1 r.2 r.3 r.4 r.5 r.6 r.7 r.8 clus$cluster
1    8   6   4   3   5   7   2   1            2
2    7   5   3   4   6   8   1   2            2
3    1   8   5   7   2   4   3   6            3
4    2   6   5   8   3   2   4   7            3
5    3   7   4   6   2   5   1   8            3
6    5   3   8   4   7   2   6   1            1
7    4   2   6   5   8   1   7   3            1
8    5   3   7   4   6   1   8   2            1
9    8   6   4   3   5   7   2   1            2
10   7   5   3   4   6   8   1   2            2
.
.
.
```

A klaszter-analízis eredményét egy clus nevű objektumban tároljuk. Az objektum változóira az objektum után írt $ jelt követően hivatkozhatunk. Most a clus$cluster vektorra van szükségünk. A cbind() függvény oszlopokat kapcsol össze, a preferencia-mátrixot kapcsoljuk össze a klaszter-vektorral. A módosított adatbázis első tíz sorát láthatjuk.

Az iskolai végzettséggel összehasonlítva a klasztereket be tudjuk azonosítani. Esetünkben az első klaszter az általános iskolai, a második a középiskolai, a harmadik az egyetemi végzettséggel rendelkező válaszadókat tartalmazza.

Elemezzük az általános iskolát végzettek adatait.

```
# Általános iskolai végzettség
> prefm=read.csv2(file="pref_matrix_rizs.csv",header = T)
> prefm=prefm[prefm$iskola==1,1:8]
> Conjoint(prefm,profile,szintek)
```

Olvassuk be újból az eredeti preferencia-mátrixot, és szűrjük le az általános iskolát végzettekre. Azért kell csak az első nyolc oszlop, mert ezek tartalmazzák a preferenciákat. A kilencedik oszlopban az iskolai végzettség kódjai találhatók.
Az eredménytáblázat:

```
Call:
lm(formula = frml)

Residuals:
Min     1Q Median     3Q    Max
-0,5   -0,5    0,0    0,5    0,5

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4,83333    0,04245  113,87   <2e-16 ***
factor(x$Region)1     -1,33333    0,05480  -24,33   <2e-16 ***
factor(x$Region)2      2,66667    0,06932   38,47   <2e-16 ***
factor(x$Cultivation)1 -0,66667    0,05480  -12,17   <2e-16 ***
factor(x$Cultivation)2  1,33333    0,06932   19,23   <2e-16 ***
factor(x$Price)1       0,66667    0,05480   12,17   <2e-16 ***
factor(x$Price)2       0,66667    0,05480   12,17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,5094 on 185 degrees of freedom
Multiple R-squared:  0.9524,    Adjusted R-squared:  0.9508
F-statistic: 616.7 on 6 and 185 DF,  p-value: < 2,2e-16

[1] "Part worths (utilities) of levels (model parameters for whole sample):"
levnms    utls
1       intercept  4,8333
2           India -1,3333
3            Kína  2,6667
4         Magyar -1,3333
5     Hagyományos -0,6667
6  Nincs_vegyszer  1,3333
7       Organikus -0,6667
8             400  0,6667
9             500  0,6667
10            600 -1,3333
[1] "Average importance of factors (attributes):"
[1] 47,17 24,53 28,30
[1] Sum of average importance:  100
```

Az ábrákat is megkapjuk, de ezeket most nem közöljük le. A maradékok nagyon szabályosak, nulla várható értékű, a kvartilisekben szimmetrikusan helyezkednek el. Minden regressziós együttható szignifikáns, a többszörös r-négyzet értéke nagyon magas, 0,9524. A modellünk nagyon jól leírja a preferenciákat. Ezt a jó illeszkedést hiányoltuk korábban. A rizs legfontosabb tulajdonsága a származási ország, 47,17%. Ez az összevont adatbázisban is így volt, ez nem változott az osztályok elkülönítésével. A második tulajdonság az ár, 28,3%. Ez a klaszter tehát már árérzékeny. Fontosság szerint a harmadik a termesztés módja, 24,53%.

Az első klaszter vásárlóinak „kedvenc terméke” tehát a kínai, vegyszer nélkül termesztett, 400 vagy 500 Ft-os rizs. Az áremelkedést negatív preferenciával jutalmazzák.

A középiskolai végzettségűek eredménye:

```
# Közép iskolai végzettség
> prefm=read.csv2(file="pref_matrix_rizs.csv",header = T)
> prefm=prefm[prefm$iskola==2,1:8]
> Conjoint(prefm,profile,szintek)

Call:
lm(formula = frml)

Residuals:
Min      1Q  Median      3Q     Max
-1,3846 -0,6667  0,1538  0,6667  1,6923

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4,11111    0,06130  67,069  < 2e-16 ***
factor(x$Region)1      -0,87179    0,07913 -11,017  < 2e-16 ***
factor(x$Region)2      -2,01709    0,10010 -20,152  < 2e-16 ***
factor(x$Cultivation)1 -0,19658    0,07913  -2,484   0,0135 *
factor(x$Cultivation)2 -0,79487    0,10010  -7,941 3,88e-14 ***
factor(x$Price)1        0,43590    0,07913   5,508 7,70e-08 ***
factor(x$Price)2        0,22222    0,07913   2,808   0,0053 **
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,9377 on 305 degrees of freedom
Multiple R-squared:  0.8274,    Adjusted R-squared:  0.824
F-statistic: 243.6 on 6 and 305 DF,  p-value: < 2,2e-16

[1] "Part worths (utilities) of levels (model parameters for whole sample):"
levnms    utls
1       intercept  4,1111
2           India -0,8718
3           Kína -2,0171
4         Magyar  2,8889
5     Hagyományos -0,1966
6  Nincs_vegyszer -0,7949
7       Organikus  0,9915
8             400  0,4359
9             500  0,2222
10            600 -0,6581
[1] "Average importance of factors (attributes):"
[1] 63,16 24,04 12,80
[1] Sum of average importance:  100
```

Az eredmények részletes értelmezését most már az olvasóra bízzuk. Az R-négyzet itt is magas, minden együttható szignifikáns. Ebben a klaszterben a „kedvenc termék” a magyar, organikusan termesztett, 400 Ft-os rizs. A vásárlás során a legfontosabb a származási hely. Csak magyar legyen, a többi tényező nem olyan fontos.

Végül az egyetemet végzettek vásárlási hajlandóságának eredménye:

```
# Egyetemet végzettek
> prefm=read.csv2(file="pref_matrix_rizs.csv",header = T)
> prefm=prefm[prefm$iskola==3,1:8]
> Conjoint(prefm,profile,szintek)

Call:
lm(formula = frml)

Residuals:
Min      1Q  Median      3Q     Max
-1,4640 -0,5360 -0,1126  0,5360  1,4640

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4,11261    0,05189  79,254  < 2e-16 ***
factor(x$Region)1     2,88739    0,06699  43,101  < 2e-16 ***
factor(x$Region)2    -1,77477    0,08474 -20,944  < 2e-16 ***
factor(x$Cultivation)1  1,22973    0,06699  18,356  < 2e-16 ***
factor(x$Cultivation)2 -0,55856    0,08474  -6,592 2,06e-10 ***
factor(x$Price)1      0,67117    0,06699  10,019  < 2e-16 ***
factor(x$Price)2      0,09459    0,06699   1,412    0,159
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,7732 on 289 degrees of freedom
Multiple R-squared:  0.8888,    Adjusted R-squared:  0.8865
F-statistic: 385.1 on 6 and 289 DF,  p-value: < 2,2e-16

[1] "Part worths (utilities) of levels (model parameters for whole sample):"
levnms    utls
1       intercept  4,1126
2           India  2,8874
3           Kína -1,7748
4         Magyar -1,1126
5     Hagyományos  1,2297
6  Nincs_vegyszer -0,5586
7       Organikus -0,6712
8             400  0,6712
9             500  0,0946
10            600 -0,7658
[1] "Average importance of factors (attributes):"
[1] 56,02 24,29 19,69
[1] Sum of average importance:  100
```

Az eredmények röviden összefoglalva. Az illeszkedés nagyon jó, az R-négyzet 0,8888. Az együtthatók szignifikánsak. A tulajdonságok fontossága sorrendben: származási hely, termesztéstechnológia, ár. A „kedvenc termék” az indiai, hagyományos technológiával termesztett 400 Ft-os rizs.

Mi a közös mindhárom klaszterben? A legfontosabb, hogy melyik országból származik a rizs, és a lehető legolcsóbb legyen.

Most már tudjuk, hogy három szegmens létezik és beazonosítottuk a klaszterek legfontosabb tulajdonságait is.

11.3. A piaci részesedés becslése

Meg kell nézni, hogy milyen rizst lehet kapni a boltokban, mivel a conjoint-analízisben hipotetikus rizs változatok szerepeltek. Információt kell gyűjteni, hogy a kapható termékeknek mennyi a tényleges piaci részesedése. Ez a modellünk validálásához kell majd. Amennyiben a modellünk valid, akkor becsléseket tehetünk egy új termék várható piaci részesedésére. A validálás során a becsült és tényleges piaci részesedést fogjuk összehasonlítani. Amennyiben kicsi lesz az eltérés, akkor a modellünk alkalmas az előrejelzésre, az új termékek piaci részesedésének becslésére. Nagy eltérés esetén a modell nem valid, nem alkalmas a jelenség leírására. Ilyenkor el kell gondolkodni, hogy tényleg a jellemző tulajdonságokat vontuk-e be az analízisbe. A megkérdezettek reprezentatívak voltak, a minta tükrözi a valóságot, stb.? Nem valid modell esetén a vizsgálatot meg kell ismételni, a tanulságok birtokában.

Tételezzük fel, hogy a magyar piacon csak négyféle rizst lehet kapni. Ezek az alábbiak:

11.1. táblázat: **A magyar piacon kapható rizs tulajdonságai**

Termék	Származási ország	Termesztési technológia	Ár (Ft/kg)	Piaci részesedés (%)
1.	India	hagyományos	500	40
2.	Kína	organikus	400	30
3.	Magyarország	hagyományos	600	20
4.	Magyarország	vegyszer nélküli	600	10

Az általunk forgalmazott rizs a 3. Határozzuk meg a várható piaci részesedéseket a piaci szegmensek figyelmen kívül hagyásával, azaz a 100 válaszadót együtt értékelve.

```
> prefm=prefm[,1:8]
> ShowAllSimulations(sim,prefm, profile)

  TotalUtility MaxUtility BTLmodel LogitModel
1         5,19         37    31,75      38,98
2         4,04         24    24,35      26,35
3         4,04         26    23,17      21,26
4         3,62         13    20,73      13,40
```

Hogyan jött ki a fenti eredmény? A lineáris modellbe behelyettesítettük a piacon kapható rizs tulajdonságait. Az első termék TotalUtility értéke, azaz a modell által becsült preferenciája:

4,285+0,4083+0,2183+0,2817=5,1933

A második termék:

4,285-0,8033-0,0217+0,5783=4,0383

És így tovább a harmadik és negyedik termék.

A számítások megkönnyítése érdekében idemásoltam a modell együtthatóit.

[1] "Part worths (utilities) of levels (model parameters for whole sample):"

	levnms	utls
1	intercept	4,285
2	India	0,4083
3	Kína	-0,8033
4	Magyar	0,395
5	Hagyományos	0,2183
6	Nincs_vegyszer	-0,1967
7	Organikus	-0,0217
8	400	0,5783
9	500	0,2817
10	600	-0,86

A következő három oszlop három különböző módszerrel becsli a piaci részesedést. Az értékek százalékos formában vannak megadva. Természetesen a három módszer eltérő eredményt ad. Ahány modell, annyiféle eredmény. Melyik modellt válasszuk? Természetesen azt, amelyik eredménye a legközelebb van a valósághoz. Hogyan lehet ezt eldönteni? Például egy illeszkedésvizsgálattal, khi-négyzet próbával. Miután kiválasztottuk a megfelelő modellt, becsléseket végezhetünk, megvizsgálhatjuk, hogy a rizs tulajdonságainak változtatásával hogyan változna a piaci részesedés. Példánkban a legjobb becslést a LogitModel adja. Az előrejelzésekre ezt érdemes használni.

A 100 válaszadó együttes eredményét csak akkor használhatjuk fel becslésre, ha reprezentatív a minta. Példánkban az iskolai végzettséget tudjuk. Vajon tényleg ilyen arányú a lakosság iskolai végzettsége Magyarországon? Ha nem, akkor a klaszterek alapján kell becsülni a piaci részesedést, és a tényleges iskolai végzettség arányával súlyozottan kell meghatározni a valós részesedést. Tételezzük fel, hogy a minta nem reprezentatív az iskolai végzettségre, ezért a három piaci szegmensre határozzuk meg a termékek várható piaci részesedését. Először az általános iskolát végzettek vásárlási hajlandóságából becsüljük a piaci viselkedésüket.

```
> prefm=read.csv2(file="pref_matrix_rizs.csv",header = T)
> prefm=prefm[prefm$iskola==1,1:8]
> ShowAllSimulations(sim,prefm, profile)
```

	TotalUtility	MaxUtility	BTLmodel	LogitModel
1	3,5	0	21,87	1,96
2	7,5	100	46,88	95,83
3	1,5	0	9,38	0,25
4	3,5	0	21,88	1,96

A három modell itt már nagyon eltérő becslést ad.

A középiskolai végzettséggel rendelkezők csoportján belül várható piaci részesedés:

```
> prefm=read.csv2(file="pref_matrix_rizs.csv",header = T)
> prefm=prefm[prefm$iskola==2,1:8]
> ShowAllSimulations(sim,prefm, profile)
```

TotalUtility	MaxUtility	BTLmodel	LogitModel
1	3,26	0,00	17,63	5,48
2	3,52	0,00	19,11	8,25

3	6,15	66,67	33,21	53,27
4	5,55	33,33	30,05	33,00

Az egyetemet végzettek csoportján belül a forgalmazott rizsek várható piaci részesedése.

```
> prefm=read.csv2(file="pref_matrix_rizs.csv",header = T)
> prefm=prefm[prefm$iskola==3,1:8]
> ShowAllSimulations(sim,prefm, profile)
```

	TotalUtility	MaxUtility	BTLmodel	LogitModel
1	8,32	100	53,03	98,30
2	2,34	0	15,26	0,37
3	3,46	0	21,53	1,15
4	1,68	0	10,18	0,17

A piaci szegmensek részarányát a KSH felméréséből tudhatjuk meg. A 2014. évi felmérés alapján:

általános iskola vagy annál kevesebb 22%

különböző középiskolai végzettség 60%

főiskola, egyetem 18%

A korábban kapott eredményeket tehát ezekkel az értékekkel kell súlyozni. A termékek tényleges piaci részesedése tehát a modellekkel kapott hányad szorozva a piaci szegmensek arányával.

A modellünk valid, tehát alkalmas a „mi lenne, ha” típusú kérdések megválaszolására. Változtassuk meg az általunk forgalmazott magyar rizs termesztés technológiáját hagyományosról organikusra. Vizsgáljuk meg, hogy a tulajdonság módosítása hogyan hat a piaci részesedésre?

```
> ShowAllSimulations(sim,prefm, profile)
```

	TotalUtility	MaxUtility	BTLmodel	LogitModel
1	5,19	37	34,10	38,96
2	4,04	24	24,53	24,17
3	3,80	39	20,84	31,08
4	3,62	0	20,53	5,80

Tovább folytathatnánk a „mi lenne, ha” típusú kérdések vizsgálatát, pl. egy új termék bevezetése a piacra, hogyan módosítaná a várható eladásokat? Az új terméket a szimulált kártyák közé kell felvenni, és megismételni a conjoint-analízist.

Mi történik, ha valamelyik együttható nem szignifikáns? Meg kell ismételni a vizsgálatot? Igen, a nem szignifikáns tényezők kihagyásával. Hiszen ezek a tényezők nem befolyásolják a vásárlási szokásokat, azaz semleges tényezők.

12. Hálózatelemzés: elmélet, módszerek és alkalmazások

12.1. Bevezetés

A hálózatelemzés napjaink egyik legdinamikusabban fejlődő tudományterülete, amely a komplex rendszerek különböző aspektusainak megértését célozza. A hálózatok mindenhol jelen vannak, legyen szó a természetről, a társadalomról vagy a technológiáról. Olyan jelenségek, mint a szociális kapcsolatok, az üzleti ökoszisztémák vagy az internetes infrastruktúra, mind-mind hálózati struktúrát mutatnak, amelyeket a hálózatelemzés rendszerezni, értelmezni és modellezni képes. E tudományterület alapjait a gráfelmélet biztosítja, amely matematikai eszközöket nyújt a csúcsok és az élek által alkotott rendszerek leírására. Azonban a hálózatelemzés nem csupán matematikai vagy informatikai problémákat vizsgál; interdiszciplináris jellege miatt a közgazdaságtan, a biológia, a szociológia és még a humán tudományok területeire is kiterjed.

A mai világban a komplex rendszerek megértése elengedhetetlen, hiszen a szociális hálózatok elemzésével megismerhetjük, hogyan terjednek az információk. A természeti rendszerek vizsgálatával feltárhatjuk az ökoszisztémák kölcsönhatásait, míg az üzleti hálózatok tanulmányozásával optimalizálhatók a beszerzési láncok vagy a pénzügyi rendszerek. Az olyan kérdések, mint a hálózati csomópontok szerepe vagy a kapcsolatok szerkezete, kulcsfontosságúak lehetnek a stratégiai döntéshozatalban. A hálózatelemzés gyökerei a 18. századig nyúlnak vissza, amikor Leonhard Euler megoldotta a königsbergi hidak problémáját, amely a gráfelmélet ősének tekinthető. A 20. század közepén a szociális hálózatok elemzése került előtérbe, amelyben olyan kutatók, mint Jacob Moreno és Stanley Milgram, alapvető újításokat hoztak. Az internet és a digitális technológiák megjelenésével a hálózatelemzés képes lett nagyméretű adathalmazokat kezelni és komplex struktúrákat feltárni.

A hálózatelemzés fontossága folyamatosan nő, ahogy új adatvezérelt ökoszisztémák jelennek meg, és egyre bonyolultabb rendszerek modellezésére van szükség. A következő fejezetben a hálózatelemzés alapfogalmait fogjuk részletesen megvizsgálni.

12.2. Kapcsolatháló elemzés

A kapcsolathálózat elemzés alapját a gráfok adják. A gráf csúcsokból és élekből áll. A csúcsok azoknak az adatoknak feleltethetők meg, melyek kapcsolatait vizsgáljuk. Az élek pedig akkor keletkeznek, ha valós kapcsolódást találunk a vizsgált adatok között. A gráfok használatának célja, hogy a különböző kapcsolatokat ábrázoljunk vele (BARABÁSI, 2013).

A gráfelmélethez kapcsolódó első probléma a XVIII. századból származik. Ekkoriban Königsberg lakosai vetették fel azt a kérdést, hogy a várost átszelő Pergel folyón átívelő hét hídon lehet-e olyan sétát tenni, hogy mind a hét hídon pontosan egyszer haladjanak át. A problémával a szentpétervári akadémia tanárához, Eulerhez fordultak, aki bebizonyította, hogy ilyen séta nem létezik. A bizonyításáról szóló dolgozat, mely 1730-ban jelent meg, a gráfelméleti munka alapkövének tekinthető. Az első tudományos színvonalú gráfelméleti könyv pedig 1936-ban jelent meg, Kőnig Dénestől, a Budapesti Műegyetem akkori magántanárától származik (ANDRÁSFAI, 1997).

A hálózatokkal kapcsolatos kutatások módszertani alapját a gráfelmélet adja, amely az ún. véletlen hálózatok elméletével próbált meg választ adni a hálózatokkal kapcsolatos kérdések egy részére (ERDŐS – RÉNYI, 1959; BOLLOBÁS, 2001). Hamar kiderült azonban, hogy a valódi világ hálózatai nem írhatók le teljes mértékben véletlen hálózatokkal, mivel jól azonosítható, specifikus struktúrákba rendeződnek. Először a szociológiai vizsgálatok mutattak

rá, hogy a társadalmi hálózatok jellegzetes szerveződési struktúrája nem felel meg a véletlenszerűség követelményének. Ezek a társadalmi hálózatokat ún. „kisvilágokként" írják le, ahol a szorosan összefüggő lokális csoportokat áthidaló kapcsolatok kötnek össze. Maga az elnevezés arra utal, hogy a csomópontok közötti átlagos távolság relatíve kicsi, miközben a lokális csoportok megőrzik viszonylag éles határvonalaikat.

Travers és Milgram (1969) a Harvard Egyetem ismeretségi hálózatát vizsgálva jutott arra a felismerésre, hogy az átlagos elérési út még egy ilyen kiterjedt kapcsolati hálózatban is meglepően rövid, mindössze 5 és fél lépés. A rövid átlagos távolságok gondolatát korábban már Karinthy Frigyes is felvetette egy írásában, ahol meglepően pontosan a későbbi tudományos eredményeket előre jelezve, ötlépéses távolságról ír (Karinthy, 1929). Referenciaműnek számít ebben a témakörben Granovetter (1973) tanulmánya is, aki a lokális csoportokat összekötő „gyenge" kapcsolatok jelentőségét emeli ki. A társadalmi kapcsolatrendszerek általa felvázolt struktúrája a kisvilágok reprezentációja. A kisvilágok intuitív elképzelését később Watts és Strogatz (1998) formalizálták.

Akárcsak a véletlen hálózatok, a kisvilágok is leírhatók egy reprezentatív csomóponttal, vagyis egy átlagos kapcsolati számmal. A valós hálózatok nem jellemezhetők ilyen tipikus szereplőkkel: néhány csomópont rendkívül nagyszámú, míg a többség kevés kapcsolattal bír (BARABÁSI et al, 2000). Az átlagos fokszám ugyan megadható, a hálózat struktúráját azonban döntően a nagyszámú kapcsolattal rendelkező elemek határozzák meg. Egy ilyen csomópont kiesése adott esetben a hálózat széteséséhez vezethet. Ezt a speciális struktúrát skálafüggetlen hálózatnak nevezzük. Barabási és kollégái arra a fontos felismerésre jutottak, hogy a valóságban előforduló hálózatok nagy része skálafüggetlen tulajdonságot mutat (BARABÁSI – ALBERT, 1999; BARABÁSI, 2002).

A skálafüggetlen topológia a valódi hálózatok örökké terjeszkedő természetének természetes következménye. Két összekötött pontból indulunk, és minden egyes mezőben egy új pontot adunk hozzá a hálózathoz. Amikor elhatározzuk, hogy hová kapcsolódjunk, az új pontok előnyben részesítik a jobban összekötött pontokat. A növekedésnek és a népszerűsítő kapcsolódásnak köszönhetően néhány sok kapcsolattal rendelkező középpont keletkezik (BARABÁSI, 2002).

A hálózati kapcsolatok és struktúrák elemzése elsősorban a szociológia területén vált népszerűvé, innen ered a társadalmi kapcsolatháló elemzés kifejezés is. E tudományág elsősorban gyakorlati szempontból közelít e kérdéshez, és viszonylag szűkebb matematikai hátteret ad. Bár a hálózatelemzés a gráfelmélet eredményeire építő, fontos matematikai apparátussal rendelkezik, a hálózati struktúrák leíró elemzésére használt mutatószámok erre viszonylag korlátozott mértékben támaszkodnak.

A kapcsolatháló definíció szerint társadalmi szereplők véges számú készletéből és a közöttük lévő kapcsolatokból áll. A módszer kiválóan alkalmas bonyolult társadalmi struktúrák komplex vizsgálatára és azok modellezésére (WASSERMAN – FAUST, 1994). A kapcsolatháló-elemzés az egyének viselkedését mikro-, az egyének közötti kapcsolatokat és a köztük lévő interakciókat makroszinten vizsgáló tudományterület (STOKMAN, 2005). A leggyakrabban vizsgált társas kapcsolatok a kommunikáció, tanácsadás, befolyásolás, barátság, bizalmi kapcsolatok.

A kapcsolati hálók elemzése egyre népszerűbbé válik a tudományos életben is (LIU et al., 2005). A tudományos együttműködés egy összetett kapcsolati háló (POPP et al., 2015). A 20. század közepén a tudományos kutatás magányos kutatók munkáját jelentette, de ez az utóbbi évtizedekben jelentősen megváltozott. Manapság már mind az élettudományok, mind a társadalomtudományok terén egyre inkább együttműködés jellemzi a kutatást. A tudományos együttműködéseket alapvetően a technológiai fejlesztések, a földrajzi közelség és a kutatási témák hasonlósága mozdítja elő. Az is látható, hogy a színvonalas cikkek publikálása elengedhetetlenül fontos az egyéni tudományos karrierhez (ACEDO et al., 2006). Amikor egy kutató társszerzővel közösen publikál, létrehoz egy egyéni társszerzői hálózatot. A társszerzők köre lefedi azokat a személyeket, akik érdemben közvetlenül hozzájárultak a cikk tartalmához. Több ilyen egyéni hálózat együttes ábrázolásával a szerzők és társszerzők közötti kapcsolat vizsgálható az egész mintában.

A tudományos hálózatok nagy része néhány tekintélyes személy köré épül, akik egyfajta központi szereplőként irányítják saját hálózatukat, tudományos csoportokat, ezzel klikkeket hozva létre. Ezek a tudományos klikkek exponenciálisan növelik saját publikációs teljesítményüket egyrészt közös publikációkkal, másrészt az egymásra való hivatkozásokkal (KATZ – MARTIN, 1997).

12.3. Hálózati mutatószámok

A hálózati mutatószámok kulcsfontosságú szerepet játszanak a hálózatok elemzésében, mivel lehetővé teszik a hálózat szerkezeti tulajdonságainak kvantifikálását. Ezek a mutatószámok segítenek megérteni a csomópontok szerepét, a kapcsolatok erősségét és az egész hálózat viselkedését. A mutatószámok lehetnek lokálisak (csomópontra vonatkozók) vagy globálisak (teljes hálózatra vonatkozók).

A normálizált mutatószámok különösen fontosak a különböző méretű és szerkezetű hálózatok összehasonlításához. A normalizációval a mutatóértékek egységes mértékskálára hozhatók, lehetővé téve az értelmezhető összehasonlítást.

A gráfelméleti megközelítést jól lehet alkalmazni a legfontosabb szereplő meghatározására. A fontos szereplők általában a kapcsolatháló stratégiai pontjaiban helyezkednek el, de a fontosság számítása több módon is megközelíthető, attól függően, hogy mi alapján tekintünk valakit fontosnak. Tekinthetjük azt központi személynek, aki a legnagyobb kapcsolati aktivitást mutatja, és akihez sokan kapcsolódnak, vagy aki sok szereplővel tart fenn szorosabb kapcsolatot; esetleg olyan szereplőket, akik hálózatmegszakító pozícióban vannak. A centralitás fogalmát általában nem irányított gráfoknál, míg a presztízst irányított gráfok esetén alkalmazzák. A centralitásnál elsősorban az a fontos, hogy a szereplő részt vesz kapcsolatokban, a presztízs esetén pedig azt vizsgáljuk, hány kapcsolat mutat az adott szereplő felé (KÜRTÖSI, 2024).

A **fok centralitás** (degree centrality, CD) esetén az egyes pontok kapcsolatainak számát viszonyítjuk az összes kapcsolathoz.

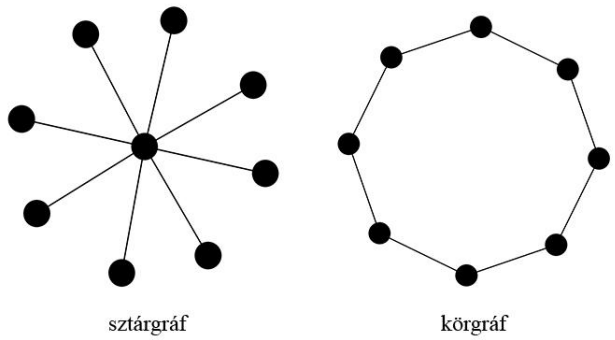
$$C_D(x_i) = \frac{d(x_i)}{n-1}$$

ahol, $C_D(x_i)$ az i. szereplő foka, $d(x_i)$ az i. szereplő fokszáma. A mutató értéke 0, ha a szereplőnek egyáltalán nincs kapcsolata a gráf többi pontjával (különálló pont), 1, ha az adott szereplő minden más ponttal kapcsolatban áll.

Különböző elemszámú hálózatok összehasonlítására a Freeman fokszám központiság mutatót alkalmazzuk.

$$C_D = \frac{\sum_{i=1}^{n}[C_D(n^*) - C_D(x_i)]}{(n-1)(n-2)}$$

ahol C_D a csoportszintű centralitás, $C_D(n^*)$ az előforduló legnagyobb fokszám, n a hálóban lévő szereplők száma. A mutató a maximális 1 értéket akkor éri el, ha egy tag minden más személlyel kapcsolatban van és a többiek csak vele vannak kapcsolatban (sztárgráf, 1. ábra). Az érték akkor 0, ha az egyes tagok központiságai között nincs differencia (pl. körgráf, 12. 1. ábra).



12.1. ábra: **Sztárgráf és körgráf**

Központiságot számíthatunk közelség centralitással (closeness centrality, C_C) is, eszerint egy személy akkor kerül központi helyzetbe, ha az összes szereplőt egyszerűen, rövid idő alatt eléri.

$$C_C(x_i) = \frac{1}{\sum_{j=1}^{n} d(x_i, x_j)}$$

ahol $j \neq i$ és $d(x_i, x_j)$ az i és j pontot összekapcsoló legrövidebb út hossza. Normalizáláshoz a mutatót szorozni kell $(n-1)$ értékkel.

A következő centralitás számítási mód a közöttiség centralitás (betweenness centrality, C_B), mely azon alapszik, hogy azok a szereplők a legbefolyásosabbak, akik sok másik szereplő között foglalnak helyet.

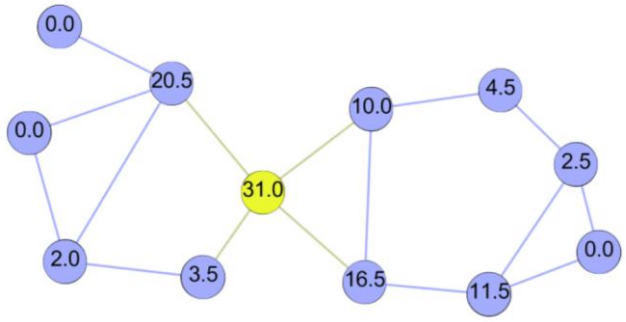
$$C_B(x_i) = \sum_{i \neq j \neq k} \frac{g_{jk}(x_i)}{g_{jk}}$$

ahol g_{jk} a j és k pont közötti legrövidebb utak száma, a $g_{jk}(x_i)$ pedig csak az i. ponton áthaladó *j* és *k* pont közötti utak száma. A mutató összeg tényezője 1, ha a szereplő rajta van mindegyik legrövidebb úton. Az érték pedig akkor 0, ha egyiken sem szerepel. Így a mutató maximális értéke:

$$\binom{n-1}{2} = (n-1)(n-2)/2$$

A normalizáláshoz a mutatót ezzel az értékkel kell osztani, mely az összes lehetséges pontpár száma, kivéve amelyikben az i. pont is szerepel.

A 2. ábra egy nem irányított gráfot szemléltet a szereplők közöttiség értékeivel. Jól látható, hogy a hálóban a sárga szereplő biztosítja a két részháló közötti kapcsolatot, ami kiemelkedő közvetítő szerepére utal (LENGYEL et al, 2018a)



12.2. ábra: **Közöttiség centralitás mutató értékei nem irányított gráf esetén**

A bemutatott hálózati mutatószámok alapvető eszközként szolgálnak a hálózatok szerkezetének és dinamikájának megértéséhez. A normálizált mutatók lehetővé teszik a különböző hálózatok összehasonlítását, míg a globális és lokális mutatók kombinált alkalmazása átfogó képet nyújt a hálózatok működéséről.

12.4. A tudományos hálózatelemzés korszerű eszközei: Gephi, VOSviewer és Bibliometrix

A mai tudományos kutatások egyik legfontosabb eszköze a bibliometriai elemzés, amely segíti a kutatókat a tudományos publikációk, hivatkozások és együttműködések feltérképezésében. A Gephi, a VOSviewer és a Bibliometrix három kiemelkedő szoftver, amelyek hatékonyan támogatják ezeket az elemzéseket. Mindhárom eszköz különböző megközelítésekkel és funkciókkal járul hozzá a kutatók munkájához:

- **Gephi**: Erős vizualizációs képességei révén lehetővé teszi a komplex hálózati kapcsolatok feltárását és elemzését.
- **VOSviewer**: Kifejezetten a tudományos publikációk és hivatkozási hálózatok vizualizációjára lett tervezve, egyszerű kezelőfelületével a bibliometriai elemzések alapvető eszköze.
- **Bibliometrix**: Egy R-alapú eszköz, amely rendkívül részletes elemzéseket kínál, és a hozzá tartozó Biblioshiny felülettel programozási ismeretek nélkül is könnyen használható.

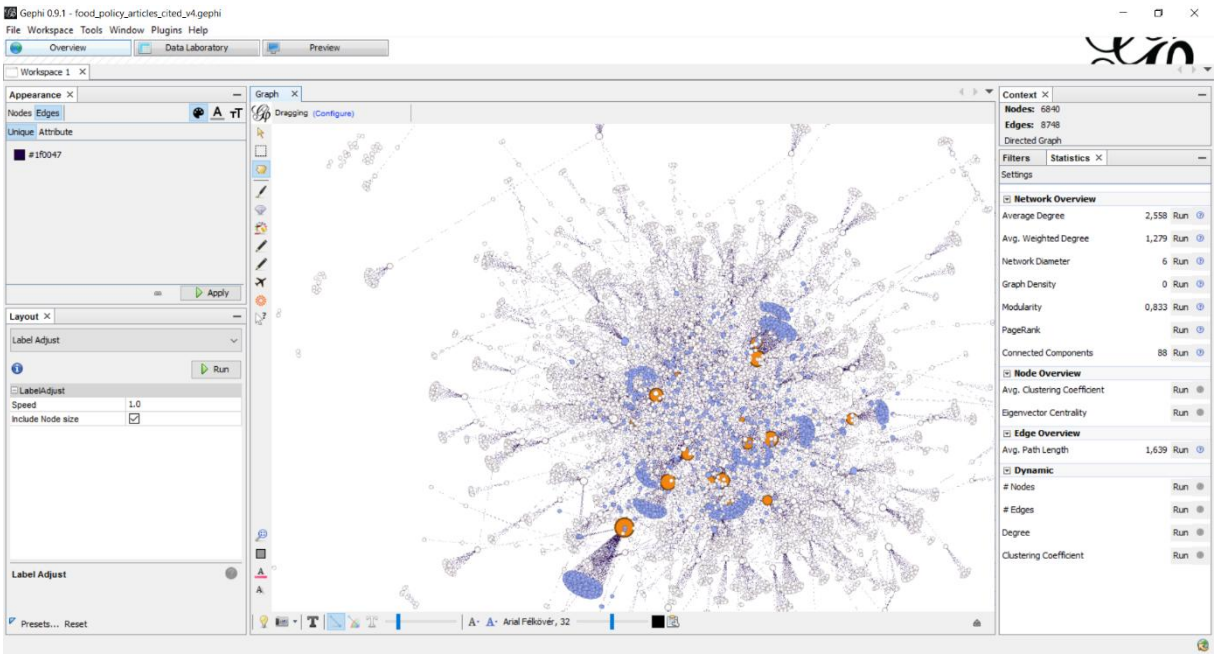
Ezek az eszközök nemcsak a kutatások tendenciáinak és mintázatainak feltárásában segítenek, hanem hozzájárulnak a tudományos együttműködések és teljesítmények értékeléséhez is. Az alábbiakban részletesen bemutatom mindegyik eszköz fő funkcióit és alkalmazási lehetőségeit.

12.4.1. Gephi - hálózatelemző vizualizációs alkalmazás

A Gephi egy nyílt forrású, interaktív vizualizációs szoftver, amelyet hálózati adatok elemzésére és megjelenítésére terveztek. A szoftver különösen alkalmas nagy mennyiségű adathalmazok feldolgozására, és lehetővé teszi, hogy a felhasználó vizuális eszközökkel fedezze fel a hálózatok struktúráját, kapcsolatokat és mintákat (LENGYEL et al., 2018b).

A Gephi többek között kutatási projektek, marketinganalízisek, közösségi média elemzések, valamint tudományos publikációk elemzésében használható. A szoftver élvonalbeli algoritmusokkal rendelkezik a hálózatok elemzésére, mint például a csoportosítási együttható kiszámítása vagy a centralitási mutatók meghatározása.

A Gephi (12.3. ábra) megnyitása után először két táblázatot kell importálni. Az élekhez (edges) a szerzőket és a hozzájuk rendelt cikkek azonosítóját tartalmazó Excel táblát kell feltölteni, a csúcsokhoz (nodes) pedig a szerzők azonosítóját tartalmazó táblát kell importálni az adatlaboratóriumban (*Data Labority*). Ekkor a program a betáplált adatok alapján megjeleníti a szükséges táblázatot, az *Overview* menüpont alatt pedig különböző témák közül választhatjuk ki, hogy milyen formában szeretnénk illusztrálni a kapcsolati hálót.



12.3. ábra. **A Gephi program felhasználói felülete**

A modularitás funkciót használva a program feltérképezi a hálózaton belüli csoportosulásokat és az ezen belül lévő kapcsolatok erősségét. Ahhoz, hogy a központiság mutatókat megkapjuk a *Statistics* menüpontot kell használni. A kalkuláció végeztével átváltva az adatlaboratóriumba, láthatóvá válnak a különböző szereplőkhöz rendelt mutatók értékei.

Gephi funkciói

- **Adatimportálás:** A Gephi támogatja különböző adatformátumok, például csv, GraphML fájlok importálását, amelyek a hálózat csomópontjait és éleit írják le.
- **Interaktív vizualizáció:** A hálózatok valós idejű manipulációja, amely lehetővé teszi a csomópontok mozgatását, az élek megjelenítésének módosítását és különböző elrendezési algoritmusok alkalmazását.
- **Hálózati mutatók számítása:** Központossági mutatók, sűrűség, átmérő és más alapvető hálózati metrikák egyszerű kiszámítása.
- **Dinamikus hálózatok elemzése:** Időbeli változások követése a hálózatban.

- **Közösségi detekció:** Klaszterek és közösségek azonosítása moduláris algoritmusok segítségével.

Gephi használata:

Hálózati adatok importálása

1. **Fájl betöltése:** Indítsd el a Gephi-t, majd a **File > Open** menüpont alatt válassz egy támogatott formátumú fájlt (pl. csv).
2. **Adattípusok meghatározása:** Az importálás során megadhatod, hogy a fájl csomópontokat (nodes) és éleket (edges) tartalmaz-e, valamint beállíthatod az élek irányítottságát (irányított vagy irányítatlan hálózat).
3. **Adatellenőrzés:** Ellenőrizd az adatszerkezetet, és győződj meg róla, hogy minden csomópont és él helyesen van definiálva.

Hálózat vizualizálása

1. **Elrendezési algoritmusok:**
    - Válassz elrendezési algoritmust a **Layout** panelen. Népszerű opciók:
        - **Force Atlas:** Klaszterek és közösségek azonosítására.
        - **Yifan Hu:** Nagyméretű hálózatok vizualizációjára optimalizálva.
2. **Stílusbeállítások:**
    - A **Nodes** és **Edges** panelen megadhatod a csomópontok méretét, színét, valamint az élek vastagságát és színét.
3. **Interaktív manipuláció:**
    - Mozgatható csomópontok, hogy jobban áttekinthetővé váljon a hálózat szerkezete.

Hálózati mutatók kiszámítása

1. **Statisztikai panel:** A **Statistics** modul segítségével számítsd ki a kívánt hálózati mutatókat, például:
    - **Átlagos fokszám:** Egy csomópont átlagos kapcsolódási száma.
    - **Központossági mutatók:** Betweenness centrality, closeness centrality, eigenvector centrality.
    - **Klaszterezettségi együttható:** A csomópontok lokális kapcsolódási sűrűsége.
2. **Eredmények értelmezése:** Az eredményeket megtekintheted táblázatos formában vagy vizualizációként.

Közösségi detekció

1. **Moduláris algoritmus:**

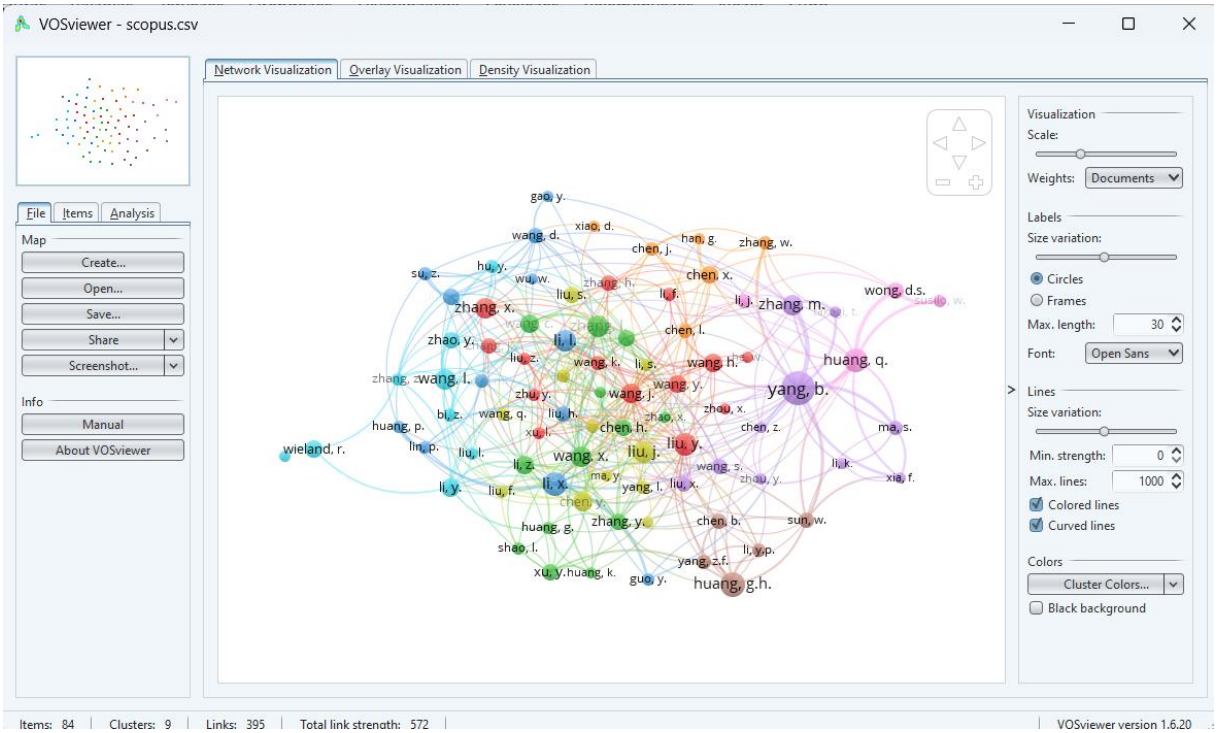
- A **Modularity** algoritmus azonosítja a hálózatban lévő közösségeket.
- Az eredmények színekkel megjeleníthetők a vizualizációban.

2. **Közösségek elemzése:** Elemezd a klaszterek méretét és a kapcsolódási mintázatokat.

A Gephi egy sokoldalú eszköz, amely megkönnyíti a hálózatok elemzését és vizualizációját. A különböző funkciók, például a statisztikai elemzések és a közösségi detekció, lehetővé teszik, hogy mélyreható betekintést nyerjünk a hálózatok szerkezetébe és dinamikájába. Az egyszerű használat és a vizualizációs lehetőségek miatt a Gephi az egyik legjobb választás a hálózatelemzési feladatokhoz.

12.4.2. A VOSviewer szoftver bemutatása és alkalmazási lehetőségei

A VOSviewer egy ingyenesen elérhető szoftver, amelyet tudományos hálózatok vizualizálására és elemzésére terveztek. Ez a program lehetővé teszi, hogy kutatási eredményeket és társzerzői kapcsolatokat átláthatóan és vizuálisan értelmezhető módon jelenítsünk meg. A VOSviewer képes támogatni a bibliometriai elemzéseket, beleértve a hivatkozási hálózatokat, a bibliográfiai kapcsolódásokat, a társzerzői kapcsolatokat, valamint a fogalmi hálózatok feltárását. Egyedi szövegbányászati funkciói segítségével következtetéseket vonhatunk le tudományos publikációkból, valamint kulcsfontosságú kifejezések kapcsolatát és gyakoriságát vizualizálhatjuk.



12.4. ábra: **A VOSviewer felhasználói felülete**

A VOSviewer nagy előnye, hogy számos különböző adatformátumot kezel, és különböző adatbázisokból, mint a Web of Science vagy a Scopus, közvetlenül importálhatók adatok. Az ilyen funkciók lehetővé teszik a kutatók számára, hogy hatékonyan elemezzék a tudományos kapcsolódásokat. Ezen kívül a szoftver felhasználható interdiszciplináris kutatások trendjeinek feltérképezésére is.

A VOSviewer használata

A szoftver telepítése egyszerű és gyors, Windows és Mac rendszerekre egyaránt elérhető. A működéshez Java 8 vagy magasabb verziójú futtatókörnyezet szükséges.

Adatforrások kiválasztása:

A VOSviewer olyan adatforrásokat támogat, mint a Web of Science, a Scopus, a PubMed, valamint referencia-kezelők, mint az EndNote és a Mendeley.

Az adatok letöltése során különböző formátumok (általában .txt vagy .csv) használhatók.

Az adatforrás kiválasztásának fontos szempontja, hogy az adott kutatási területre releváns publikációkat tartalmazzon.

Első lépések:

Indítsa el a programot, majd válassza ki az „Create Map” lehetőséget.

Válasszon az elérhető elemzési típusok közül: fogalomtérkép, kulcsszótérkép vagy társzerzői térkép.

A térképkészítés folyamata során a felhasználó számos paramétert beállíthat, például a minimális előfordulási küszöbértéket.

Elemzési lehetőségek

Fogalomtérkép létrehozása:

A fogalomtérképek vizualizálják a szöveges adatokban előforduló kulcsfontosságú kifejezések kapcsolatait. A közelebbi kifejezések nagyobb kapcsolódást jeleznek.

Társzerzői térkép:

Ez a térkép a kutatók közötti kollaborációt ábrázolja. Támogatja az egyéni kutatók, intézmények és országok szerinti elemzést. Az algoritmus alapján csoportokat (klasztereket) hoz létre a kapcsolatok erőssége alapján. Az ilyen térképek segítenek megérteni, hogy mely kutatók dolgoznak együtt a leggyakrabban, és hol vannak erős tudományos kapcsolatok.

Kulcsszótérkép:

Ez a térkép a publikációkban előforduló kulcsszavak együtt-előfordulását vizualizálja.

Lehetőség van az "Author Keywords" és "Keywords Plus" elemzésére is.

Az ilyen típusú térképek segítenek azonosítani a tudományos publikációkban használt legfontosabb kifejezéseket.

Kulcsszótérképek és fogalomtérképek együttes elemzése lehetővé teszi a kutatási területek precízebb megértését és az új irányok feltárását.

A VOSviewer hatékony eszközt nyújt a tudományos hálózatok és publikációs trendek vizualizálásához. Az eszköz nagy előnye, hogy ingyenesen elérhető, és többféle adattípust képes kezelni. Használata segítheti a kutatókat a tudományos trendek és kapcsolatok

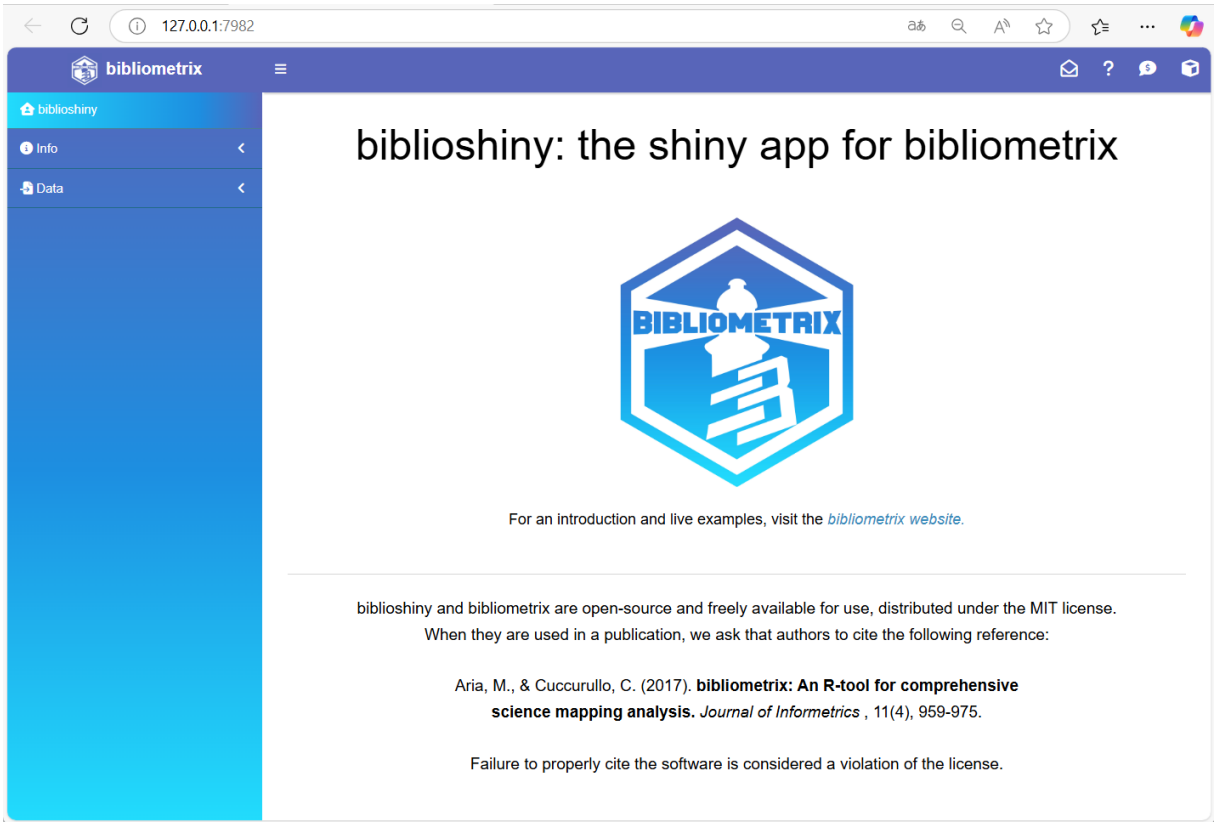
átláthatóbbá tételében. A vizualizációs térképek támogatják a tudományos publikációk értelmezését, és segítenek megérteni a kutatások közötti kapcsolatokat. Az ilyen elemzések hasznosak lehetnek a kutatási stratégiák kidolgozásában is.

Az eszköz különösen hasznos lehet doktori kutatásokhoz, intézményi teljesítménymméréshez és tudományos hálózatok összehasonlító elemzéséhez is.

12.4.3. A Bibliometrix és Biblioshiny bemutatása és alkalmazási lehetőségei

A Bibliometrix és a hozzá tartozó interaktív felület, a Biblioshiny, a bibliometriai elemzések elvégzéséhez nyújtanak hatékony és könnyen használható megoldásokat. A Bibliometrix egy R-alapú csomag, amely lehetővé teszi a kutatók számára a tudományos publikációk metaadatainak részletes elemzését és vizualizálását. A Biblioshiny pedig egy webalapú felület, amely a Bibliometrix funkcióit teszi elérhetővé programozási ismeretek nélkül is.

Ezek az eszközök különösen hasznosak lehetnek olyan kutatási területeken, ahol nagy mennyiségű tudományos publikáció bibliográfiai adatait kell elemezni, például a hivatkozási hálózatok, a kulcsszó-elemzések vagy a szerzői együttműködések területén.



12.5. ábra: **A Biblioshiny felhasználói felülete**

A Bibliometrix és Biblioshiny telepítése és használata

Bibliometrix telepítése:

- Nyissa meg az R vagy RStudio környezetet.

- Adja ki a következő parancsot a csomag telepítéséhez: *install.packages("bibliometrix")*
- Töltse be a csomagot az alábbi parancs segítségével: library(bibliometrix)

Biblioshiny indítása:

- A Biblioshiny felületet a következő paranccsal lehet elindítani az R környezetből: *biblioshiny()*
- Ez megnyit egy webalapú felületet, ahol az adatok feltöltése és az elemzések elvégzése interaktívan történhet.

Főbb funkciók és alkalmazási lehetőségek

Adatok importálása

A Bibliometrix és a Biblioshiny támogatja a különböző forrásokból származó bibliográfiai adatok importálását, például Web of Science, Scopus, PubMed vagy Dimensions adatbázisokból. Az importálás során a metaadatok, például a címek, absztraktok, kulcsszavak és hivatkozások automatikusan feldolgozásra kerülnek.

Elemzési lehetőségek:

- **Hivatkozási hálózatok elemzése:** Az eszköz lehetővé teszi a hivatkozások közötti kapcsolatok feltárását és vizualizálását.
- **Kulcsszó-elemzés:** Azonosíthatók a leggyakrabban használt kulcsszavak és azok kapcsolatai.
- **Társzerzői hálózatok:** A szerzők közötti együttműködések és csoportok elemzése.
- **Időbeli trendek vizsgálata:** Az eszköz lehetőséget nyújt a publikációs aktivitás időbeli változásának elemzésére.

Vizualizáció:

- A Bibliometrix és Biblioshiny interaktív grafikonokat és diagramokat kínál, amelyek segítségével a kutatók könnyen értelmezhetik az eredményeket.
- Példák: hálózati diagramok, tematikus térképek, időbeli trendek grafikonjai.

Gyakorlati alkalmazások

1. **Tudományos területek feltérképezése:** A kulcsszó-elemzések és a hivatkozási hálózatok segítségével azonosíthatók a kutatási területek közötti kapcsolatok és az új, feltörekvő témák.
2. **Intézményi teljesítmény értékelése:**
   - A szerzők és intézmények közötti kapcsolatok elemzése segíthet az együttműködések és az intézményi hatás azonosításában.

3. **Publikációs trendek vizsgálata:** Az időbeli elemzések segítségével megérthetők a kutatási prioritások változásai és a legnépszerűbb témák fejlődése.

A Bibliometrix és Biblioshiny eszközök egyaránt rugalmas és hatékony megoldást nyújtanak a bibliometriai elemzések elvégzéséhez. Az interaktív vizualizációk és a részletes elemzési lehetőségek révén ezek az eszközök hozzájárulnak a kutatók munkájának támogatásához és a tudományos teljesítmény mélyebb megértéséhez.

Felhasznált irodalom

Acedo, F. J., Barroso, C., Casanueva, C., & Galán, J. L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *Journal of Management Studies, 43*(5), 957-983. https://doi.org/10.1111/j.1467-6486.2006.00625.x

Ács P. (2009): Sporttudományi kutatások módszertana. Pécs 291. o. ISBN: 9789636422752

Allenby, G. M. Hardt, N., Rossi P. E. (2019). Chapter 3 - Economic foundations of conjoint analysis. Ed. J-P. Dubé, P.E. Rossi, Handbook of the Economics of Marketing, North-Holland, Volume 1. 151-192. ISBN 9780444637598.

Al-Omari B, Farhat J, Ershaid M. (2022). Conjoint Analysis: A Research Method to Study Patients' Preferences and Personalize Care. J Pers Med. 12(2):274. doi: 10.3390/jpm12020274.

Andrásfai, B. (1997). *A gráfelmélet elemei*. Akadémiai Kiadó.

Backhaus K., Erichson B., Plinke W., Weiber R. (2016): Multivariate Analysemethoden. 10. Auflage, Springer Gabler, Berlin, Heidelberg 647. o. ISBN 978-3-662-46075-7

Barabási, A.-L. (2002). *Linked: The new science of networks*. Perseus Publishing.

Barabási, A.-L. (2013). *Network science*. Cambridge University Press.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512. https://doi.org/10.1126/science.286.5439.509

Barabási, A.-L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications, 281*(1-4), 69-77. https://doi.org/10.1016/S0378-4371(00)00018-2

Baráth CS.-né - Ittzés A. - Ugrósdy GY.:1996. Biometria: módszertan és a MINITAB programcsomag alkalmazása. Mezőgazda Kiadó, Budapest

Bollobás, B. (2001). *Random graphs*. Cambridge University Press.

Chen, J. Roth J. (2024): Logs with Zeros? Some Problems and Solutions, *The Quarterly Journal of Economics*, 139(2), 891–936, https://doi.org/10.1093/qje/qjad054

Cochran, W. G., and G. M. Cox 1957. Experimental Designs. 2d. ed. New York: Wiley.

Dunn, O. J., and V. A. Clark. 1987. Applied Statistics: Analysis of Variance and Regression. 2d. ed. New York: Wiley.

Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press.

Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae, 6*(1), 290-297.

Fishbein, M. (1967): Attitude and the Prediction of Behaviour. In: Fishbein, M. (Ed.): Readings in Attitude Theory and Measurement. Wiley, New York

Freund, J. and Perles, B. "A New Look at Quartiles of Ungrouped Data." Ameri can Stat. 41, 200-203, 1987.

Füstös et al. (2004): Alakfelismerés. UMK. Budapest.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380. https://doi.org/10.1086/225469

Green, M. C. – Keegan, W. J. (2020): Global Marketing. Global Edition, 10th Edition, Pearson, Boston

Hair J.F., Anderson R.E., Tatham R.L., Grablowsky B.J. (2014): MULTIVARIATE DATA ANALYSIS: Pearson New International Edition. 7th edition, Pearson Publisher, UK, ISBN: 9781292021904

Hajdu O. (2003): Többváltozós statisztikai számítások. KSH. Budapest.

Hoaglin, D.; Mosteller, F.; and Tukey, J. (Ed.). Understanding Robust and Explo ratory Data Analysis. New York: Wiley, pp. 39, 54, 62, 223, 1983.

Hunyadi L., Vita L.: Statisztika I. Aula Kiadó, Budapest, 2008. 1-348. o.

Hunyadi L., Vita L.: Statisztikai képletek és táblázatok (oktatási segédlet), Aula Kiadó, Budapest, 2008. 1-51. o.

Huzsvai L. – Balogh, P. (2015): Lineáris modellek az R-ben. SENECA BOOKS, 150. o. ISBN 978-615-801172-0-6

Huzsvai L. (2012): STATISZTIKA gazdaságelemzők részére Excel és R alkalmazások. SENECA BOOKS, 173. o. ISBN 978-963-08-5016-2

J.P. Marques de Sá (2007): Applied Statistics, Using SPSS, STATISTICA, MATLAB and R. Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-71971-7.

John, P.W.M. 1971. Statistical Design and Analysis of Experiments. New York: MacMillan.

Karinthy, F. (1929). Láncszemek. *Minden másképp van*. Athenaeum.

Katona Tamás - Lengyel Imre (szerk.): Statisztikai ismerettár - fogalmak, képletek, módszerek Excel és SPSS alkalmazásokkal. JATEPress, Szeged, 1999. 121 oldal, (közgazdász, jogász, kísérletes és társadalomtudomány)

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kenney, J. F. and Keeping, E. S. "Quartiles." §3.3 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 35-37, 1962.

KIRK, R. E. 1982 Experimental Design. 2d ed. Monterey, CA: Brooks/Cole Publishing Co.

Kotler, P. – Armstrong, G (2020): Principles of Marketing. Global Edition, 18/E, Pearson, Harlow, etc.

Kotler, P. – Keller, K. L. (2017): Marketingmenedzsment. Akadémiai Kiadó, Budapest

Kőnig, D. (1936). *Theorie der endlichen und unendlichen Graphen*. Akadémiai Kiadó.

Kürtösi Zs.(2004): A társadalmi kapcsolatháló-elemzés módszertani alapjai. In Letényi László (szerk.): Településkutatás. Budapest, L'Harmattan, pp. 663-684.

Lengyel, P., Pancsira, J., & Füzesi, I. (2018a). Szerzői kapcsolatháló-elemzés. *International Journal of Engineering and Management Sciences*, 3(3), 76–84. https://doi.org/10.21791/IJEMS.2018.3.7.

Lengyel, P., Török, É., & Füzesi, I. (2018b). Szerzői kapcsolatháló-elemzés a gyöngyöző borokról szóló tudományos cikkek alapján. *Információs Társadalom*, 18(2), 98–113. https://doi.org/10.22503/inftars.XVIII.2018.2.6

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management, 41*(6), 1462-1480. https://doi.org/10.1016/j.ipm.2005.03.012

Lothar Sachs: 1985. Statisztikai módszerek. Mezőgazdasági Kiadó, Budapest

Malhotra, N. K. – Simon, J. (2016): Marketingkutatás. Akadémiai Kiadó, Budapest

Malhotra, N. K. (2010): Marketing Research – An Applied Orientation. 6th Edition, Prentice Hall, Boston etc.

Marshall D., Bridges J.F.P., Hauber B., Cameron R., Donnalley L., Fyie K., Johnson F.R. (2010). Conjoint Analysis Applications in Health—How are Studies being Designed and Reported? An Update on Current Practice in the Published Literature between 2005 and 2008. Patient. 3:249–256. doi: 10.2165/11539650

Mendenhall, W. and Sincich, T. L. Statistics for Engineering and the Sciences, 4th ed. Prentice-Hall, 1995.

Mérő L. (1992): A pszichológiai skálázás matematikai alapjai. Tankönyvkiadó, Budapest, 1992. 15. o.

Moksony Ferenc: Gondolatok és adatok: Társadalomtudományi elméletek empirikus ellenőrzése. Budapest, Osiris Kiadó, 1999.

Moore, D. S. and McCabe, G. P. Introduction to the Practice of Statistics, 4th ed. New York: W. H. Freeman, 2002.

Neter, J., W. Wassermann, and M. H. Kutner. 1985. Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. 2d ed. Homewood, Illinois.: Richard D.Irwin, Inc.

Newman, M. E. J. (2018). Networks: An Introduction. Oxford University Press.

PENG, K. C. 1967. The Design and Analysis of Scientific Experiments. Reading, MA: Addison-Wesley.

Petrovics, P. – Géczi-Papp, R. (2021): Kereszttábla elemzés az SPSS-ben. Oktatási segédanyag, Miskolci Egyetem, https://gtk.uni-miskolc.hu/files/12362/10_SPSS+keresztt%C3%A1bla.pdf (letöltés ideje: 2021.06.31.)

Popp, B., Wilson, S., Horstmann, N., & de Ruyter, K. (2015). Social media network engagement and firm value: The impact of social media activities on firm performance. *Journal of Interactive Marketing, 31*, 17-35. https://doi.org/10.1016/j.intmar.2015.05.004

Rao, V. R. (2025). Applied Conjoint Analysis - From Product and Service Design to Market and Pricing Strategies. Springer International Publishing AG.

SAGE (2019): Learn to Use the Eta Coefficient Test in R With Data From the NIOSH Quality of Worklife Survey (2014). SAGE Publications, Ltd., https://methods.sagepub.com/base/download/DatasetStudentGuide/eta-coefficient-niosh-qwl-2014-r (letöltés ideje: 2021.06.10.)

Sajtos, L – Mitev. A. (2007): SPSS kutatási és adatelemzési kézikönyv. Alinea Kiadó, Budapest

Simon J. (2006): A klaszterelemzés alkalmazási lehetőségei a marketingkutatásban. Statisztikai Szemle 85(7) 627-650. o.

Soekhai V., Whichello C., Levitan B., Veldwijk J., Pinto C.A., Donkers B., Huys I., van Overbeeke E., Juhaeri J., de Bekker-Grob E.W. (2019). Methods for exploring and eliciting patient preferences in the medical product lifecycle: A literature review. Drug Discov. Today. 24:1324–1331. doi: 10.1016/j.drudis.2019.05.001.

SPSS (2021): IBM SPSS Statistics honlapja: https://www.ibm.com/products/spss-statistics (letöltés ideje: 2021.06.18.)

Stokman, F. N. (2005). Network science: General overview and research examples. In P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 19-43). Cambridge University Press.

Sváb, J. 1981. Biometriai módszerek a kutatásban. Mezőgazdasági Kiadó. Budapest.

Székelyi M. – Barna I. (2002): Túlélőkészlet az SPSS-hez. Többváltozós elemzési technikákról társadalomkutatók számára. 4. kiadás TYPOTEX Elektronikus Kiadó Kft., ISBN:9789632790121

Szűcs István Szerk.: Alkalmazott statisztika. Agroinform Kiadó, 2002.

Tamus Antalné (2009): A marketing kutatás gyakorlata. Károly Róbert Kutató – Oktató Közhasznú Nonprofit Kft., Gyöngyös ISBN: 9789639941083

Taylor W.J. 82016). Pros and cons of conjoint analysis of discrete choice experiments to define classification and response criteria in rheumatology. Curr. Opin. Rheumatol. 28:117–121. doi: 10.1097/BOR.0000000000000259.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry, 32*(4), 425-443. https://doi.org/10.2307/2786545

Várhegyi É. 2004. „Bank Competition in Hungary".Acta Oeconomica 54(4), 403–424.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442. https://doi.org/10.1038/30918

Whittaker, E. T. and Robinson, G. The Calculus of Observations: A Treatise on Nu merical Mathematics, 4th ed. New York: Dover, pp. 184-186, 1967.

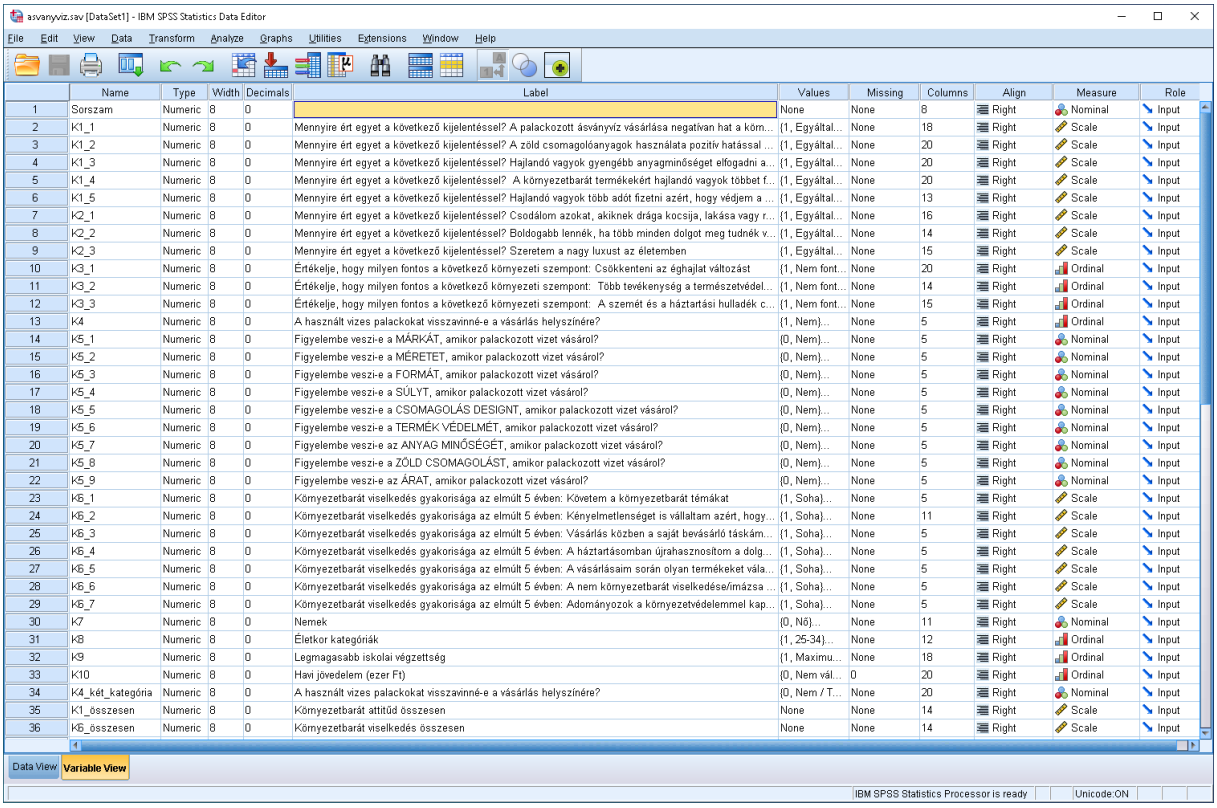
Winer, B. J. 1971. Statistical Principles in Experimental Design, 2d. ed. New York.

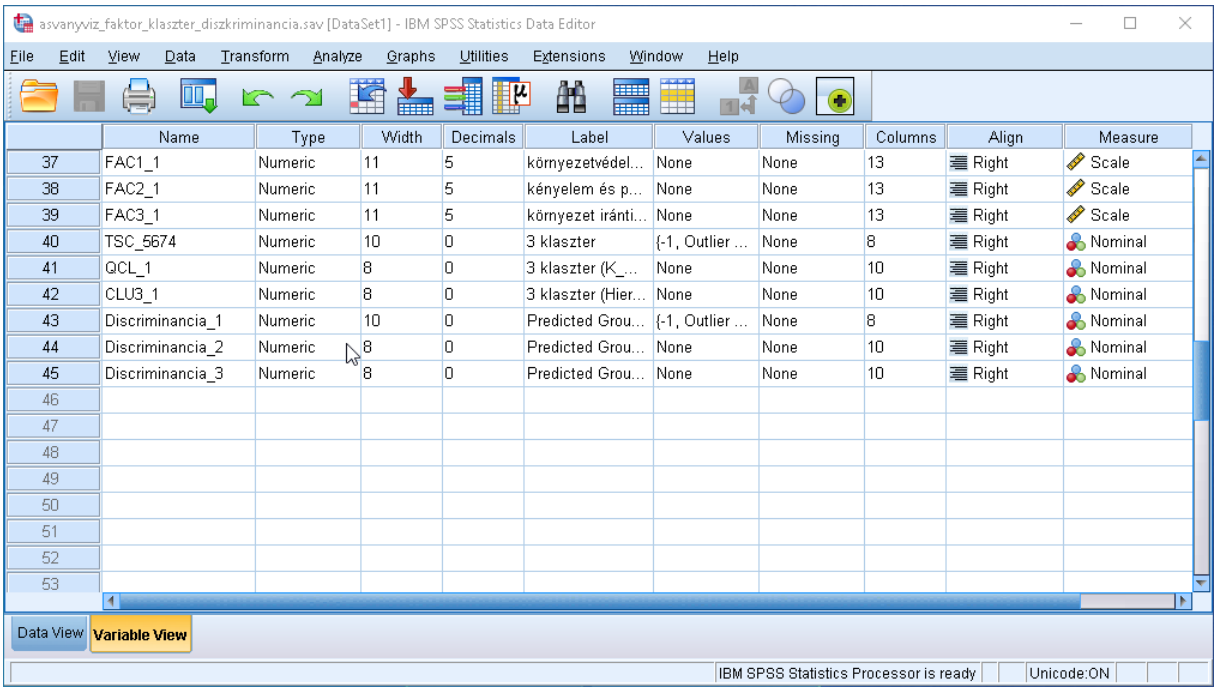
Köszönetnyilvánítás:

„A TKP2021-NKTA-32 számú projekt az Innovációs és Technológiai Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a TKP2021-NKTA pályázati program finanszírozásában valósult meg."

Mellékletek

1. sz. melléklet. Az ásványvíz fogyasztást és vásárlást elemző kérdőív kérdéseinek bemutatása





DEBRECENI
EGYETEM