SAPIENTIA
ERDÉLYI MAGYAR
TUDOMÁNYEGYETEM
CSÍKSZEREDAI KAR

jegyzet

BÁLINT GYÖNGYVÉR

STATISZTIKAI MÓDSZEREK ÉS ADATELEMZÉS A TÁRSADALOMTUDOMÁNYOKBAN IBM SPSS SEGÍTSÉGÉVEL

Scientia Kiadó | 2026

Bálint Gyöngyvér

Statisztikai módszerek és adatelemzés a társadalomtudományokban IBM SPSS segítségével

SAPIENTIA ERDÉLYI MAGYAR TUDOMÁNYEGYETEM
CSÍKSZEREDAI KAR
TÁRSADALOMTUDOMÁNYI TANSZÉK

Bálint Gyöngyvér

Statisztikai módszerek és adatelemzés a társadalomtudományokban IBM SPSS segítségével

Scientia Kiadó
Kolozsvár · 2026

Készült
a Magyar Kormány
Támogatásával

Miniszterelnökség
Nemzetpolitikai Államtitkárság

Bethlen Gábor
Alap

DOI: 10.47745/SAPVOL.2026.02

Felelős kiadó:
Sorbán Angella

Lektor:
Csata Zsombor (Marosvásárhely)

Borítóterv:
Tipotéka Kft.

Kiadói koordinátor:
Szabó Beáta

A szakmai felelősséget teljes mértékben a szerkesztők, illetve a szerzők vállalják.

Második, bővített kiadás

ISBN: 978-606-975-115-2

TARTALOMJEGYZÉK

CUPRINS

TABLE OF CONTENTS

ELŐSZÓ

A *Statisztikai módszerek és adatelemzés a társadalomtudományokban IBM SPSS segítségével* című egyetemi jegyzet elsősorban társadalomtudományi szakos hallgatók számára készült. Célja kettős: egyrészt bevezetést nyújt a tárgy elméleti alapjaiba, másrészt bemutatja annak gyakorlati alkalmazási lehetőségeit. A jegyzet szándéka, hogy egyszerű, lépésről lépésre építkező módon ismertesse meg az olvasót a társadalomtudományokban leggyakrabban használt, alapvető statisztikai technikákkal.

Elméleti részei nagyrészt a Hunyadi–Mundruczó–Vita szerzőhármas által jegyzett statisztikai tankönyvre támaszkodnak, míg az SPSS-alkalmazások a *European Values Survay Románia – magyar kisebbség kutatás* adatbázisát használják (az adatbázis és a kérdőív a regisztrációt követően ingyenesen letölthető a https://search.gesis.org/research_data/ZA7550?doi=10.4232/1.13562 oldalról).

A jegyzet 47, kézi számítással és/vagy SPSS segítségével megoldott példafeladat mentén kíséri végig az olvasót az elemzési lépéseken. Az első fejezet a társadalomstatisztika alapfogalmait (sokaság, változó, mérési szintek) és az adatbázisokkal kapcsolatos alapvető műveleteket mutatja be (adatbázis-létrehozás, változók címkézése, adatimport, adatállományok összekapcsolása, esetkiválasztás, változóátalakítás). Ezt követik az egyváltozós elemzések témái, beleértve a gyakorisági eloszlásokat, a középértékeket, a szóródási mutatókat és az eloszlás alakját jellemző statisztikákat. A kétváltozós elemzések előtt a harmadik fejezet rövid áttekintést ad a valószínűségszámítás és a valószínűségi mintavétel alapelveiről, érintve a standard hiba fogalmát, a nem valószínűségi mintavételi eljárásokat, valamint a társadalomtudományi alkalmazásban gyakran használt szignifikanciateszteket. A negyedik fejezet részletesen bemutatja a két minőségi változó, egy kategoriális és egy mennyiségi változó, továbbá két mennyiségi változó közötti kapcsolat elemzésére szolgáló eljárásokat (khí-négyzet próba és gamma, t-teszt és F-próba, korreláció). Az utolsó fejezet áttekintést nyújt a többváltozós elemzések logikájáról, majd lépésről lépésre vezeti végig az olvasót a többváltozós lineáris regresszió, a főkomponens-elemzés és a K-közép klaszterelemzés SPSS-beli megvalósításán. A melléklet az SPSS 22.0-ás verziójának menüparancsait foglalja össze, megkönnyítve a szoftverhasználat elsajátítását.

A jegyzet két alapvető gondolata már az első fejezetben megfogalmazódik. Egyrészt: a statisztikai ismeretek megértésének kulcsa a módszerek gyakorlati alkalmazása – az elmélet támpontul szolgál, de az érdemi adatelemzési készségek csak tényleges elemzői munkában fejleszthetők. Ebben a számítógépes programok pótolhatatlan segítséget nyújtanak. Másrészt: a matematikai eszközök mechanikus alkalmazása nem elegendő; a hatékony adatelemzéshez szaktudásra, társadalomtudományi gondolkodásra van szükség. A legösszetettebb módszerek

sem képesek korrigálni a kutatás megtervezése során elkövetett hibákat, és a kapott eredmények értelmezése is csak megfelelő szakmai háttérrel lehetséges.

Végezetül szeretném kifejezni hálámat Mezei Elemérnek, aki a jegyzet 2009-es kiadásához fűzött alapos észrevételeivel és építő javaslataival jelentős mértékben hozzájárult a kézirat szakmai továbbgondolásához és finomításához. Ugyancsak nagy köszönettel tartozom Csata Zsombornak és Nistor Laurának, akik a jelen, második – átdolgozott és bővített – kiadás lektorálását vállalták. Precíz megjegyzéseik, konstruktív kritikáik és lényeglátó tanácsaik nemcsak a szöveg pontosítását segítették, hanem érdemben hozzájárultak ahhoz is, hogy a jegyzet a társadalomtudományi kutatási gyakorlatban még inkább alkalmazható legyen. Őszinte köszönettel tartozom mindhármuknak szakmai hozzájárulásukért és támogatásukért. Köszönöm továbbá Ruzsa Istvánnak az igényes tördelést és az együttműködést a kézirat végleges formába öntésében.

Csíkszereda, 2026. január 15.

A szerző

1. FEJEZET

BEVEZETÉS A TÁRSADALOMSTATISZTIKÁBA

1.1. Mi a statisztika?

A statisztika (általános statisztika, matematikai statisztika) a valóság számszerű információinak megfigyelésére, összegzésére, elemzésére és modellezésére irányuló gyakorlati tevékenység és tudomány. A statisztika tömegjelenségekkel foglalkozik. Tehát módszeresen megfigyeli a tömegjelenségek tulajdonságait, begyűjti a jellemző információkat, és feldolgozza, értékeli, elemzi ezeket.

A statisztika legfőbb érdeme, hogy:

– információt szolgáltat a megfigyelt jelenségekről,

– lehetőséget ad a tudományos elemzésekhez,

– tájékoztat a fontosabb társadalmi-gazdasági folyamatokról (legfontosabb az állami vagy hivatalos statisztika).

A statisztika fogalmán az általános és az alkalmazási területhez kötődő módszertannak, valamint a gyakorlati tevékenységnek a szorosan összefüggő egységét értjük. A statisztika arra szolgál, hogy a valóság tényeinek valamely adott körét tömören, a számok nyelvén jellemezze.

A statisztika történeti kialakulása és fejlődése

A statisztika először mint gyakorlati, számbavételi tevékenység jelent meg az ókorban. A legkorábbi statisztikai adatok az ókori államokban végrehajtott népszámlálásból származnak. A középkorban a hűbérurak földbirtokával összefüggő leltározó jellegű összeírásokat végeztek, később, a polgári társadalmak kialakulásával pedig egyre nőtt az érdeklődés a különböző országok földrajzi, politikai és gazdasági viszonyai iránt. Mindezek az úgynevezett német leíró iskola kifejlődéséhez vezettek. Maga a statisztika szó is ebből az időből származik, a státus (állam) szóból ered.

A polgári társadalmak fejlődésével a leíró jellegű információk köre bővült, a közöttük lévő számszerű összefüggések ismeretének igénye pedig kikényszerítette az elemzések módszertani fejlesztését is. Ebben az időben az államszámtant átnevezték politikai aritmetikának – ez lett a tudományos elemző statisztika alapja.

A legnagyobb előrelépést az a tény képezte, hogy a 18–19. században meghatározták a valószínűség-számítás tételeit, és ezen tudományág fejlődésének hatására alakult ki a mai matematikai statisztika.

A statisztika ágazatai és kapcsolata más tudományokkal

Miként ez köztudott, a statisztikának a matematikához való kötődése a legerősebb, hiszen a matematika elmélete (főként a valószínűség-számítás elmélete, lásd a *3. Mintavétel* fejezetet) a szakmai összefüggések leírására megfelelő módszertani tárházat nyújt. A statisztika a matematika eredményeit (amelyek alkalmasak a tömegjelenségekben rejlő törvényszerűségek feltárására) és a szakmai jelenség természetét ismerve alakítja ki módszereit.

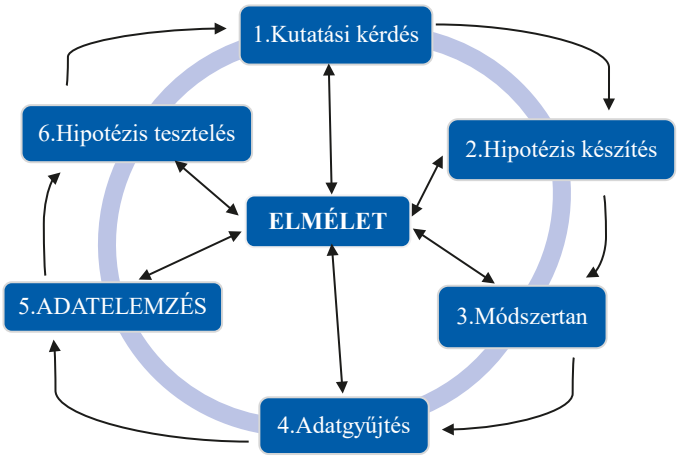
A statisztikai tevékenység sok irányba ágazik szét, így alakulnak ki a szakstatisztikák. A szakstatisztikák egy-egy terület szakmai összetevőit ismerve olyan matematikai módszert választanak, amely az ott előforduló jelenségeket szakmai szempontból is helyesen írja le. A szakstatisztika nem más, mint a társadalmi-gazdasági élet egy-egy területének statisztikai módszerekkel való vizsgálata (pl. gazdaságstatisztika, népességstatisztika stb.). A szakstatisztikán belül is további differenciálódás következik be, de egy szakterületen belül egységes alapelvek érvényesülnek.

A társadalomstatisztika

A társadalomstatisztika az általános statisztika egy sajátos változata. A társadalomstatisztika is az általános statisztikán alapul, de a vizsgált változók, mutatók és eljárások a társadalmi viszonyok sajátos mérési módjához vannak igazítva, így egyes számítások matematikai értelemben vett pontossága magyarázatra szorul (Mezei–Veres 2001). Ilyen gyakran előforduló, módszertanilag érzékeny eset például a Likert-skálás adatok kezelése. A társadalomtudományokban használt 5 fokú elégedettségi skálák, például a kollégákkal való elégedettség (1 = nagyon elégedetlen … 5 = nagyon elégedett) formálisan ordinális mérési szintűek: a kategóriák sorrendje adott, de a pontok közötti távolság nem bizonyítottan egyenlő. A gyakorlatban azonban ezeket az adatokat gyakran kvázi intervallumként kezelik, és átlagot, szórást vagy akár korrelációt és regressziót is számolnak rájuk. Ez a megoldás sokszor működik, de elméletileg vitatható. Például ha három válaszadó a kollégákkal való elégedettségre 1, 2 és 4 pontot ad, az átlag 2,33. De nem tudjuk, hogy az 1 és 2 közti különbség ugyanakkora jelentésű-e, mint a 2 és 4 közötti, pedig az átlag számítása ezt feltételezné.

A mérési szint meghatározása, a mérési hibák befolyása sajátos jelleggel bír a társadalomtudományokban. Megtörténik, hogy egy módszert olyan adatokra is alkalmaznak, amelyek nincsenek kellő pontossággal mérve (pl. faktorelemzést alkalmaznak ordinális mérési szintű változókon). A társadalomstatisztika ezekkel a problémákkal is meg kell birkózzon.

A statisztikai elemzés leginkább az adatelemzés lépcsőjéhez köthető (1. ábra). De a kutatás minden lépését a mögöttes elmélet határozza meg, és fordítva, minden lépés eredménye hatással lehet az elméletre.



1. ábra. *A társadalomtudományi kutatás lépései*

Ebből következnek a társadalomstatisztika legfontosabb korlátai:

– az elemzések eredménye erősen függ a vizsgálatba bevont szempontoktól, változóktól (elméleti kerettől),

– a bevont szempontok kiválasztásának mindig szakmai döntésre kell támaszkodnia,

– minden szakmailag releváns szempontot be kell vonni az elemzésbe,

– a matematikai eszközök mechanikusan nem alkalmazhatók, szükség van szaktudásra (társadalomtudományi ismeretekre).

Tehát a statisztikai módszerekkel kapott eredményeket csak megfelelő szakmai ismerettel lehet hatékonyan felhasználni, ugyanakkor a korszerű társadalomtudományi szakismeret elképzelhetetlen a mennyiségi összefüggések ismerete nélkül. Az eddigi átfogó értékelés helyett a különböző szakterületek igénye az értékelés mélysége felé mutat, amely a módszertani apparátus ismeretén túl igényesebb a vizsgált szakterület ismeretét (elméleti vonatkozásait) illetően.

A statisztikai ismeretek megértésének talán legjelentősebb összetevője a módszerek alkalmazásának gyakorlása. Az elméleti ismeretek segítik a gyakorlást, ám a készségek effektív munka során alakíthatók ki (ez utóbbi jelentősen visszahat az elméleti ismeretek elmélyítésére is), amelyben nagy segítséget nyújtanak a számítógépes programcsomagok. A statisztikai programcsomagok közül a szociológusok által leginkább használt IBM SPSS (Statistical Package for the Social Sciences) Windows alatt futó programjának 22.0-ás alkalmazását ismertetem (röviden SPSS).

1.2. Alapfogalmak

A szociológiában a társadalmi valóság tömör, számszerű jellemzéséhez az operacionalizálás révén jutunk el. Mindezt megelőzi a vizsgált területre vonatkozó szakismeret áttekintése, a kutatási kérdések és hipotézisek megfogalmazása és konceptualizálása (lásd társadalomtudományi kutatási módszerek és technikák tárgy). Ezeket a fázisokat követi maga az operacionalizálás, amely nem más, mint a vizsgált kutatási probléma különböző jellemzőinek megadása (kérdőíves adatfelvételek esetén a kérdőív kérdéseinek megfogalmazása képezi ezt a tevékenységet). Az operacionalizálás elképzelhetetlen a megfigyelési egységek definiálása (a vizsgált sokaság beazonosítása), valamint a mérési eljárások kialakítása (az ismérvek vagy változók megfogalmazása) nélkül.

A vizsgálat tárgyát képező egységek összességét, halmazát statisztikai sokaságnak, vagy rövidebben *sokaságnak*, esetleg *populációnak* nevezzük.

A statisztikai sokaság egyedei a statisztikai egységek. Ezek az egységek lehetnek élőlények: emberek, pl. a népszámlálás esetén; állatok, a mezőgazdasági összeírásoknál; tárgyak, pl. a személygépkocsi-állomány állapotának felmérésénél; szervezetek, pl. a vállalkozások IT-felszereltségének felmérésekor, események, pl. a kulturális rendezvények vizsgálata esetén, de lehetnek képzett egységek is, pl. a GDP alakulásának vizsgálatakor. Azt, hogy mit tekintünk a statisztikai vizsgálatnál sokaságnak, mindig a vizsgálat célja dönti el. Ha pl. a Sapientia egyetem hallgatóinak tévénézési szokásait szeretnénk vizsgálni, akkor az alapsokaság nem más, mint az abban az időpontban hallgatói jogviszonnyal rendelkező diákok sokasága. Mivel a valóságban legtöbbször nem áll módunkban a populáció egészéről adatfelvételt készíteni, ezért mintát veszünk, és az ilyen módon begyűjtött adatokon végzünk statisztikai elemzéseket.

A sokaság egységei különböző tulajdonságaik megadásával jellemezhetőek. Ezen tulajdonságok egy része a sokaság minden egyes egységére nézve közös, más részük azonban nem.

A sokaság tagjai, egységei a vizsgálat tárgyának ismeretében legtöbbször elég egyértelműen adódnak, de vannak olyan esetek is, amikor a sokaság egységei nem különülnek jól el egymástól, hanem csak önkényesen definiálhatóak (vagy a valóságban nem is léteznek).

Amikor a valóság jól elkülönülő egységekből áll (számolásnál), *diszkrét sokaságról* beszélünk, ilyen pl. egy adott településen élő lakosok száma. Amikor valóságos, de csak önkényesen elkülöníthető egységekből áll (két adott érték között elméletileg az összes értéket felveheti), akkor *folytonos sokaságról* beszélünk, mint pl. a Sapientia egyetem diákjai által egy nap elfogyasztott ásványvíz mennyisége.

Ha a sokaság elképzelt egységekből áll, *fiktív sokaságról* beszélünk (pl. Románia 2028. január 1-jei lakosainak száma).

Amikor a sokaság csak egy adott időpontra vonatkozóan értelmezhető, *álló sokaságnak* (pl. a lakosság száma 2025. január 1-jén), amikor pedig csak valamely adott időtartamra vonatkoztatva értelmezhető, *mozgó sokaságnak* nevezzük (pl. a Hargita megyei munkanélküliek száma a 2024-es év folyamán).

Ismérv vagy változó

Az ismérvek olyan vizsgálati szempontok, amelyek alapján egy sokaság egymást át nem fedő részekre bontható. A sokaság egyes egységeinek e felbontásban való elhelyezkedését az egységek adott szempont szerinti tulajdonságai határozzák meg. A valamely szempont szerint lehetséges tulajdonságokat ismérvváltozatoknak (attribútumnak) nevezzük. Ha az ismérv változatai számszerűek, akkor azokat ismérvértékeknek, magát az ismérvet pedig változónak (a logikailag egymáshoz tartozó attribútumok halmazának) nevezzük. A mindössze két változattal rendelkező ismérveket alternatív ismérveknek (dumy vagy dichotóm változónak) nevezzük.

Nézzük az alábbi példát (1. példa): kérdőíves kutatást készítettünk a Sapientia EMTE diákjai körében, amelynek néhány ismérve és ismérvváltozata a táblázatban található.

1. példa ▼

▶ *Az ismérvfajták által hordozott információk közötti különbségek*

Sokaság: a 2024/2025-ös tanévben az egyetemmel hallgatói jogviszonyban álló diákok	
Ismérvek:	**Ismérvváltozatok**:
Nem	férfi, nő
Születési év	2004, 2005 stb.
Állandó lakóhely (település neve)	Csíkszentgyörgy, Sepsiszentgyörgy stb.
C típusú nyelvvizsga	alapfokú, középfokú, felsőfokú
Internethasználat	igen, nem
Magasság (cm)	171, 168 stb.
Testsúly (kg)	48, 66 stb.
Fizikai állapotával való elégedettség	elégedetlen, igen is meg nem is, elégedett

Látható, hogy az 1. példában alkalmazott ismérvek nem ugyanolyan jellegű információt hordoznak. Az életkor, magasság és testsúly ismérvek ismérvváltozatai konkrét számértékek, amelyekkel akár műveleteket is végezhetünk (például annak megállapítására, hogy a diák hány éves lesz négy év múlva, vagy átlagosan milyen magasak a diákok). Ezzel szemben a nyelvvizsga foka, valamint a fizikai állapotával való elégedettség olyan ismérvek, amelyek ismérvváltozatai nem számértékek, de mégis fennáll valamiféle hierarchia az is-

mérvváltozatok között, hiszen tudjuk, hogy a középfokú nyelvtudás magasabb szintű, mint az alapfokú, stb. A nem, az internethasználat, illetve az állandó lakhely esetében azonban az ismérvváltozatok egyrészt nem számértékek, másrészt nem áll fenn semmiféle hierarchia sem az egyes ismérvváltozatok között, hiszen nem dönthető el, hogy Csíkszentgyörgyön lakni jobb vagy rosszabb, mint Sepsiszentgyörgyön, és az sem egyértelműen eldönthető, hogy nőnek vagy férfinek lenni jobb, stb. Ezenkívül a nem és az internethasználat ismérveknek csak két ismérvváltozata lehet, míg a lakóhelynek jóval több.

Összefoglalva tehát azt mondhatjuk, hogy mivel a statisztikai egységek tulajdonságainak észlelése és rögzítése adat formájában valamiféle mérésnek tekinthető, a különböző ismérveknek más-más mérhetőségi tulajdonságaik vannak. Mindez jelentősen befolyásolhatja a statisztikai vizsgálatot. Az ismérvek mérhetőségi tulajdonságainak egyik jellemzője a hozzájuk tartozó mérési szint vagy mérési skála.

Bizonyos szabályok betartása mellett egy eredetileg nem mennyiségi ismérv (valamilyen számlálás vagy mérés számszerű eredményeit rendeli hozzá a sokaság egységeihez) lehetséges változatai számértékké alakíthatóak, „kódolhatók". Ilyen módon bármely észlelt tulajdonság szám formájában történő rögzítése az egységek számokkal való jellemzésének, azaz mérésnek tekinthető. De miként a fenti példából is kitűnik, egyáltalán nem mindegy, hogy a sokaság egységeihez ilyen módon hozzárendelt számértékek mely tulajdonságai érvényesek a sokaság egységeinek a számértékekkel jellemezni kívánt tulajdonságaira is. Erről szólnak a mérési skálák vagy mérési szintek.

1.3. Mérési szintek

A szociológiában négy mérési skálát szokás használni:
1. nominális, megnevezéses vagy névleges mérési szint,
2. ordinális, rendezési vagy sorrendi mérési szint,
3. intervallum- vagy különbségi mérési szint,
4. arányskála.

Ebből az első két skálát szokás még minőségi, a második kettőt pedig mennyiségi mérési skáláknak nevezni.

A *nominális skála* a legegyszerűbb és legkevésbé informatív mérési fokozat. Csak az egységekhez rendelt számértékek egyező vagy különböző voltát engedi meg az egységeket ténylegesen is jellemző tulajdonságként elfogadni. Az egységekhez hozzátartozó számértékeknek nincs mértékegysége, tulajdonképpen csupán egy megkülönböztető címkéről beszélhetünk. A kódszámok közti különbségeknek, azok hányadosának vagy a nagyságrendjének nincsen semmi értelme, viszont az egységek csoportosítására kiválóan alkalmas. A fenti példánkban ilyen mérési szintű változó a nem, az állandó lakhely és az internethasználat.

Az *ordinális skála* esetében nemcsak a skálaértékek azonos vagy nem azonos volta, hanem azok sorrendisége is az egységek között fennálló valós viszonyokat írja le. Az egységekhez hozzárendelt számértékek sorrendje az adott egységek valamilyen szempontból vett sorrendjét mutatja (az egyes attribútumok a vizsgált tulajdonsággal relatíve kisebb vagy nagyobb mértékben rendelkeznek). A skálaértékek bármilyen, az egységek adott sorrendjét megtartó számértékek lehetnek, hiszen maguk a számértékek nem hordoznak információt, csakis azoknak a sorrendje. Akárcsak a nominális mérési szintű változók esetében, ezeknek a számértékeknek sincs mértékegysége, valamint a skálaértékek különbsége sem informatív, továbbá nincs értelme a skálaértékekkel végzett más műveleteknek sem. A fenti példánkban ilyen mérési szintű változó a nyelvvizsga, valamint a fizikai állapottal való elégedettség.

Az *intervallumskála* a szó szoros értelmében is mérést jelent, mivel a mennyivel nagyobb kérdésre is választ tudunk adni. A skálaértékek különbségei is valós információt nyújtanak a sokaság egységeiről, valamint e skálának már valamilyen mértékegység is a szerves tartozékát képezi. A skála kezdőpontja a 0 pont, azonban ez önkényes, illetve valamilyen konvención alapszik – ez lehetetlenné teszi a skálaértékek egymás közötti arányának meghatározását. A szociológiai adatfelvételekkor ritkán találkozunk intervallumskálával, a fenti példánk sem tartalmaz ilyen változót. A klasszikus példa intervallummérési szintű változóra a hőmérséklet, hiszen nincs abszolút 0 pont, a víz fagyáspontjának választása esetleges, függ az alapul vett hőmérsékleti skálától (2. példa).

2. példa ▼

▶ *Az intervallummérési szintű változók és az alapul vett mérési skála*

1. A 10 ^{0}C és 20 ^{0}C hőmérséklet közötti különbség Fahrenheit-skálán mérve is ugyanannyi, mint a –5 ^{0}C és 5 ^{0}C közötti különbség (a különbségnek valós értelme van).

$$F = 9 \cdot C/5 + 32$$

a) $10\ ^0C = 9 \cdot 10/5 + 32 = 50\ ^0F$

b) $20\ ^0C = 9 \cdot 20/5 + 32 = 68\ ^0F$

c) $-5\ ^0C = 9 \cdot (-5)/5 + 32 = 23\ ^0F$

d) $5\ ^0C = 9 \cdot 5/5 + 32 = 41\ ^0F$

$20\ ^0C - 10\ ^0C = 10\ ^0C$ $\qquad$ $68\ ^0F - 50\ ^0F = 18\ ^0F$

$5\ ^0C - (-5)\ ^0C = 10\ ^0C$ $\qquad$ $41\ ^0F - 23\ ^0F = 18\ ^0F$

2. A 20 ^{0}C és az 5 ^{0}C hőmérséklet egymáshoz viszonyított aránya nem független az alapul vett hőmérsékleti skálától (az arányoknak nincs értelme).

$20\ ^0C = 68\ ^0F$ (b.) $\qquad$ $5\ ^0C = 41\ ^0F$ (c.)

$68\ ^0F/41\ ^0F = 1{,}66$ $\qquad$ $20\ ^0C/5\ ^0C = 4$

Az *arányskála* a legtöbb információt nyújtó mérési szint. Már a kezdőpont is egyértelműen adott és rögzített, bármely két skálaérték egymáshoz viszonyított aránya is egyértelműen meghatározható, azaz információt hordoz. A fenti példánkban ilyen mérési szintű változó az életkor, magasság és testsúly változók.

A mérési szintek egymáshoz való viszonya

A mérési szintek bemutatott sorrendje a mérés egymást követő olyan fokozatainak tekinthetők, amelyek a mérés eredményeit kifejező számértékek egyre több tulajdonságának kihasználását teszi lehetővé. Ilyen értelemben a nominális mérési szint a legalacsonyabb, az arányskála pedig a legmagasabb mérési szint, ugyanakkor egy adott mérési szintű változó alacsonyabb szintűként is kezelhető.

Az ismérvfajták és mérési skálák egymástól való megkülönböztetése azért lényeges, mert más-más fajta elemzést tesznek lehetővé. Az ismérvek fajtája, illetve a mérés adott szintje mindig behatárolja az elemzés egy-egy adott esetben szóba jövő eszközeit, tehát különböző mérési szintű változók más-más típusú statisztikai elemzéseket tesznek vagy nem tesznek lehetővé.

A mérés adott szintje azonban kétféle értelemben is relatív:

1. sohasem függetleníthető el teljesen a vizsgálat célkitűzéseitől – a magas mérési szintek „alacsonyabbakká” válhatnak,

2. bizonyos elemzési technikák a megkívántnál alacsonyabb mérési szintű adatok elemzésére is jól használhatók (pl. faktorelemzés).

△ *Gyakorlófeladatok a mérési szintekhez*

Határozzuk meg a Mellékletben található EVS Románia – magyar kisebbség vizsgálat kutatás kérdőívében szereplő változók mérési szintjeit (pl. v1–v8: ordinális skála, v9–v31: nominális skála stb.)!

1.4. Adatbázisok létrehozása, címkézés

Az adatbázis (adatmátrix) nem más, mint a kutatás során a sokaság (vagy minta) elemeiről begyűjtött adatok halmaza. Az adatokat kódolt és rendszerezett formában szokás elektronikus formában rögzíteni, úgy, hogy minden egyes egységünk (esetünk, amely lehet egy megkérdezett személy, szervezet stb.) külön sorba, minden egyes változónk (ismérvünk, mért tulajdonságuk) pedig külön oszlopba kerüljön (2. ábra). Az adatbázisban minden egyes cellában egyetlen érték szerepelhet. Az operacionalizálás során nyert fogalmak, tulajdonságok a mérés eredményeként elvileg megfelelői lesznek a statisztikai adatbázist alkotó változóknak, de ez a megfelelés nem teljes. Vannak olyan tulajdonságok, amelyeknél a megfeleltetett kérdésből nem egy, hanem több változó is készül, pontosan azért, hogy a statisztikai feldolgozhatóság kedvéért egy cellában csak egyetlen adat szerepeljen.

Adatbázist több programban is létre lehet hozni, Excelben, dBase-ben, IBM SPSS-ben stb. A továbbiakban csak a szociológusok által leggyakrabban használt IBM SPSS programcsomag 22.0-ás verziójára fogok kitérni (röviden csak SPSS), amely menürendszerének leírása a Mellékletben található. A példákban és illusztrációkban használt adatbázis az *EVS Románia – magyar kisebbség vizsgálat* című szociológiai kutatás erdélyi adatbázisa (az adatbázis és kérdőív regisztráció után ingyenesen letölthető a https://search.gesis.org/research_data/ZA7550?-doi=10.4232/1.13562 oldalról (EVS, 2020), a kérdőív a jegyzet Mellékletében is megtalálható).

		Változó 1	Változó 2	Változó 3	Változó 4	Változó 5	Változó 6
		caseno	year	fw_start	fw_end	country	c_abrv
Eset 1	1	1	2019	201911	202003	642	RO
Eset 2	2	2	2019	201911	202003	642	RO
Eset 3	3	3	2019	201911	202003	642	RO
Eset 4	4	4	2019	201911	202003	642	RO
Eset 5	5	5	2019	201911	202003	642	RO
Eset 6	6	6	2019	201911	202003	642	RO
Eset 7	7	7	2019	201911	202003	642	RO
Eset 8	8	8	2019	201911	202003	642	RO
Eset 9	9	9	2019	201911	202003	642	RO
Eset 10	10	10	2019	201911	202003	642	RO

2. ábra. *Az adatbázis formája az SPSS-ben*

Az EVS Románia – magyar kisebbség vizsgálat az Európai Értékek Felmérése (European Values Study, EVS) nemzetközi kutatási program részét képezi, amelyet 1981 óta rendszeres, hozzávetőleg kilencéves ciklusokban hajtanak végre. A felmérés mintegy 300 kérdésből áll, amelyek célja az európai országok lakosságának értékrendjére, attitűdjeire és véleményére vonatkozó átfogó információk gyűjtése. Az adatfelvétel eredményei komplex képet nyújtanak többek között a családi élethez, a munkához, a környezetvédelemhez, a világnézeti orientációkhoz, a politikai és társadalmi viszonyokhoz, a vallási és erkölcsi normákhoz, valamint a nemzeti identitáshoz való viszonyulásról. A romániai magyar mintában az alapsokaságot a 18 éven felüli, Erdélyben élő (15 megye a 16-ból), magukat magyar nemzetiségűnek valló és a kérdőívre magyar nyelven válaszolni tudó személyek alkották. Ez a népesség mintegy 1,2 millió főt jelentett, amely a romániai magyarok közel 99%-át fedte le. A mintavétel három lépcsőben történt, több mintavételi technika (bővebben lásd a *3. Mintavétel* fejezetet) alkalmazásával. Első lépcsőben a települések több rétegből való véletlenszerű kiválasztása történt. A rétegek létrehozása három változó alapján történt: kistérségi hovatartozás, településnagyság és a magyar lakosság településen belüli aránya (a 2011-es településszintű népszámlálási adatok alapján). Második lépcsőben az első lépcsőben kiválasztott településeken véletlen kezdőpontú, lépésközös technikát alkalmaztak a háztartások kiválasztására. Végül pedig harmadik lépcsőben egy kvótala-

pot használtak annak eldöntésére, hogy a háztartáson belül melyik személyt kell megkérdezni. Az adatfelvétel személyes, papíralapú lekérdezéssel történt. A végső minta 1106 esetet tartalmazott. A romániai magyar kisebbség adatfelvétele 2019-ben zajlott, és az adatbázis 395 változót foglal magába (EVS 2020).

Adatbázis létrehozása SPSS-sel

Az SPSS-program több ablakot, felületet is tartalmaz, amelyek külön fájlként menthetők, kezelhetők. Alapbeállításnál az SPSS két ablakot nyit meg:

a) A *Data Editor* (Adatszerkesztő ablak) az adatokat tartalmazza, itt tudjuk az adatokat bevinni, módosítani (.sav fájlok).

b) Az *Output Viewer* (Eredménykijelző ablak) a feldolgozott statisztikai eredményeket jeleníti meg táblázatok, grafikonok formájában (.spv fájlok).

Az adatszerkesztő ablaknak (a. *Data Editor*) két nézete van, amelyek között az ablak bal alján levő fülekre kattintva válthatunk:

1. a tényleges Adattábla nézet *(Data View)*,
2. a változók különböző jellemzőinek beállításait lehetővé tevő Változó nézet *(Variable View)*.

Adatbázis létrehozásakor az SPSS-program indításakor válasszuk a *Type in Data* opciót, és kattintsunk az *Ok* gombra. Ha már fut a program, akkor a *FILE* főmenüpontban a *New* pontban válasszuk a *Data*-t. Miként a 2. ábrából is kitűnik, az SPSS adattáblája hasonlít az Excelére. Számozott sorok vannak, ahova az egyes esetek/megkérdezettek (*Cases*) adatai fognak kerülni, az oszlopokban (*Variables*) pedig a változók szerepelnek.

Első lépésben el kell neveznünk (definiálnunk) az egyes változókat és azok tulajdonságait. Ezt úgy kezdjük, hogy a bal oldali alsó sarokban az 1. *Data View* nézetről átváltunk a 2. *Variable View* nézetre, vagy a *DATA* főmenüpont *Define Variable Properties* menüpontjára megyünk, vagy duplán klikkelünk az első oszlop *var* (az első változó) mezőjére. Itt a *Name* pontnál nevet adunk a változónak, amely meg fog jelenni az adatbázis fejlécében. Érdemes olyan rövid nevet adni, amivel könnyen beazonosítható, hogy melyik kérdésről is van szó, továbbá figyelembe véve, hogy:

– minden változónév csak egyszer szerepelhet egy adatbázisban,
– max. 64 karakter hosszúságú lehet,
– nem kezdődhet számmal,
– nem tartalmazhat szóközt,
– a nagy- és kisbetűk bármilyen kombinációja használható,
– ékezetes betűket nem ajánlott használni,
– kerülni kell a jelek (pl. #, $ stb.) és jelentést hordozó szókombinációk használatát (pl. OR, NOT, AND, TO stb.).

A *Type* pontnál beállítjuk a változó formátumát a cella jobb oldalán aktívvá váló legördülő menüből (*Variable Type*). Legtöbb esetben numerikus adataink (számok) vannak, mivel a kódszámokat sokkal könnyebb bevezetni, mint a szö-

veget, így az SPSS is alapértelmezésben numerikus adatbevitelre van beállítva. Sokszor azonban előfordul, hogy pl. egy nyílt kérdést nem sikerült kódolni, és a szöveget szeretnénk bevezetni – ilyenkor a *Variable Type*-nál a *String* gombra kattintunk. Amikor idősoros adatokat akarunk bevezetni, a *Date* gombra klikkelünk, és a jobb oldali mezőben megjelenő időformátumok közül (pl. mm/dd/yyyy) kiválasztjuk az adatainknak megfelelő opciót.

A *Width* (kiterjesztés) oszlopnál beállítjuk a változóra felvett értékek/attribútumok maximális karakterszámát, vagy változatlanul hagyjuk a nyolc karakterszámon. A *Decimals* oszlopnál átállíthatjuk az alapbeállításban szereplő kéttizedes pontosságot (legtöbbször 0-ra állítjuk).

A *Label* pontnál felcímkézzük a változókat, vagyis azt a tetszőleges nevet adjuk a változóknak, amelyet a statisztikai feldolgozás mellett szeretnénk látni.

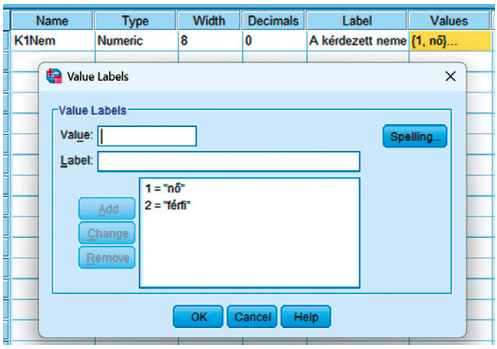
A *Values*-nál megadjuk a változóhoz tartozó egyes attribútumokat (minden egyes bevezetett címke után *Add*-et nyomunk). Amikor mindent beírtunk, akkor *Ok*-t nyomunk. A *Remove* gombbal törölhetjük, a *Change* gombbal módosíthatjuk a korábbiakat.

3. példa ▼

▶ *Változók definiálása az SPSS-ben*

Amennyiben a kérdezett nemét szeretnénk bevezetni az adatbázisba, a korábbiakban leírtak alapján a *Name*-nél a *K1Nem* nevet adjuk, a *Variable Type*-nál Numeric-en hagyjuk a beállítást, a *Width*-nél is 8 karakter hosszúságnál maradunk, a *Decimals*-t levesszük 0-ra, a *Label*-hez „A kérdezett neme” teljes változónevet írjuk. Az attribútumok felcímkézése esetében a *Values*-nál a következőképpen járunk el (3. ábra):

– *Value*: beírjuk a kódszámot (1),
– *Label*: beírjuk az attribútum nevét (nő),
– *Add*-re klikkelünk,
– *Value*: beírjuk a második kódot (2),
– *Label*: beírjuk az attribútum nevét (férfi), majd *Ok*.



3. ábra. *Címkézés az SPSS-ben (3. példa)*

Visszatérve a változók definiálásához, a *Missing* pontnál megadhatjuk, hogy milyen kóddal szereplő eseteket kezeljen az SPSS hiányzó adatként: pl. ha a 0 azt jelentette, hogy valaki *„nem tud válaszolni”*, és nem szeretnénk a számításainkba bevonni ezt az értéket, a *Missing Values*-nál a 0-t beírjuk a *Discrete missing values* pontnál, majd *Ok*-t nyomunk. A *Column* oszlopnál még beállítható az oszlopszélesség, az *Align*-nál, hogy merre rendezze az SPSS az értékeket, és a *Measure*-nél pedig az ismérv mérési szintje (nominális, ordinális vagy skála, azaz intervallum- vagy arányskálák).

Amikor több változónk ugyanazokkal az attribútumokkal rendelkezik (pl. megkérdeztük a háztartásban élő összes személy foglalkozását, vagy több olyan kérdésünk van, amelyekre igen/nem válaszokat lehet adni), két lehetőségünk van. Ha a változó már létezik egy másik, a gépünkön található SPSS-adatbázisban (1), a *DATA* főmenü *Copy Data Properties* almenüt használjuk. Itt a létező változó beállításait (értékcímkék, mérési szint stb.) csak akkor lehet átmásolni, ha az új változónknak is ugyanaz a neve (*Name*), típusa (*Type*) és karakterszáma (*Width*). Amennyiben a munkaadatbázisunkba éppen bevezetésre került egy ugyanolyan/hasonló jellemzőkkel bíró változó (2.), az *EDIT* főmenü másolási és beillesztési parancsait használjuk. Ilyen módon tudunk létrehozni változókat, el tudjuk őket nevezni. Miután megvan a keretfájlunk, nem marad más dolgunk, mint bevezetni az adatokat a kódutasítás (a *FILE* főmenü *Display Data File Information* pontja segítségével könnyen elkészíthető) szerint. Adatbázisunkat a többi Windows alatt futó programokhoz hasonlóan a *FILE* főmenü *Save* vagy *Save As…* menüpontjai segítségével menthetjük el.

△ *Gyakorlófeladatok adatbázis létrehozásához*

A Mellékletben szereplő kérdőív K1-K22-es kérdéseiből (2–5. oldalak) kialakítandó statisztikai változókat definiáljuk az SPSS-ben (*Variable View*), megadva a változók nevét, típusát, címkéit, hiányzó adatait és mérési szintjeit!

1.5. Az SPSS által kezelt adatállományok, adatbázisok összekapcsolása, esetek leválogatása

Az SPSS által kezelt adatállományok

Miként a többi ismert programban is, az adatállományok megnyitása a *FILE* főmenü *Open* almenüjéből, ezen belül a *Data* úton történik. Amennyiben már létező SPSS-adatbázist kívánunk megnyitni, a *Files of type* opciónál válasszuk a .sav kiterjesztést, és keressük meg a kívánt adatbázist.

Az SPSS több más, nem SPSS (.sav) formátumú adatbázist is be tud olvasni, mint Excel, Lotus, dBase, SAS, Stata, Text stb. Amikor nem .sav adatbázist

akarunk megnyitni, a *FILE, Open, Data, Files of type* menüpontokon végigmenve válasszuk a megfelelő formátumú fájltípust. Ez akkor igen hasznos, amikor nem saját adatbázisból szeretnénk dolgozni, ismerjük az SPSS-programcsomagot, viszont a feldolgozandó adatfájlunk nem SPSS-ben készült.

Az egyik leggyakoribb eset, amikor az adatokat egy Excel-file-ba vezették be. Ilyenkor gyors módszerrel (hátránya, hogy a változókon nem tudunk változtatni az átalakítás előtt, viszont az Excel-fájl fejléces sora alapján definiálhatjuk a változókat) a következő lépések után tudjuk megnyitni az Excel-adatbázist az SPSS-ben:

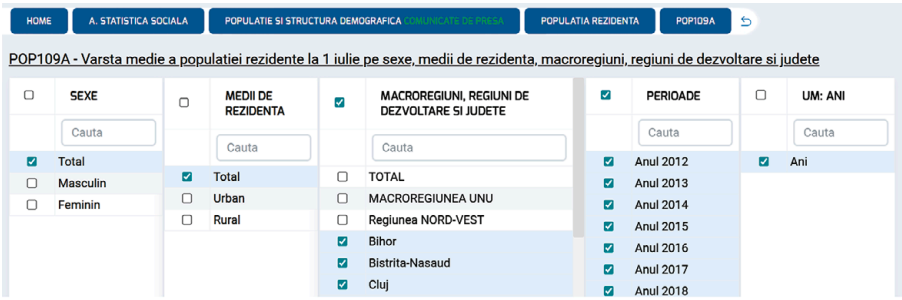
1. *FILE* főmenü
2. *Open*
3. *Data*
4. *Files of type*: .xlsx/.xls
5. Kikeressük a kívánt Excel-fájlt
6. *Open*
7. Beolvastatjuk változónévként az Excel első sorát, vagy kikapcsoljuk a *Read Variable Names from the first row of data* opciót
8. Megadjuk az importálni kívánt munkalapot (Worksheet)
9. *Ok*

Amikor változtatni akarunk a beolvasandó fájl változóin, a *FILE, Open Database, New Query* opciót válasszuk. Itt dBase, Excel és MS Access adatfájlokat olvashatunk be, ahol az Excel oszlopcímkék névvé (*Name*) alakulnak, a program felismeri a szöveges (*String*) változókat, és nagyobb oszlopszélességet, karakterszámot és nominális mérési szintet ad nekik. Minden numerikus változót magas mérési szintűvé tesz (felül kell vizsgálni).

4. példa ▼

▶ *Excel-fájlok gyors beolvasása az SPSS-be*

Feltételezzük, hogy elemezni szeretnénk a Románia különböző megyéiben élők átlagéletkorát különböző évekre. Látogassunk el a Romániai Statisztikai Hivatal honlapjára: *https://insse.ro/cms/*, itt a *Baze statistice* (Statisztikai adatbázisok) menü alatt keressük meg a *Baza de date TEMPO*-t (TEMPO idősoros adatbázis), ezen belül a *Statistică socială* (Társadalomstatisztika), *A1.1. Populaţia rezidentă* (Lakónépesség), *POP109A – Vârsta medie a populaţiei rezidente la 1 iulie pe sexe, medii de rezidentă, macroregiuni, regiuni de dezvoltare şi judeţe* (A lakónépesség átlagéletkora július 1-jén nemek, lakóhelytípusok, makrorégiók, fejlesztési régiók és megyék szerint) adatsorokat. Itt (4. ábra) a harmadik oszlopban válasszuk csak a megyéket, és jelöljük be az összes elérhető évet (2012–2023), majd *Caută*, és a kigenerált adatokat Excelben letöltjük.



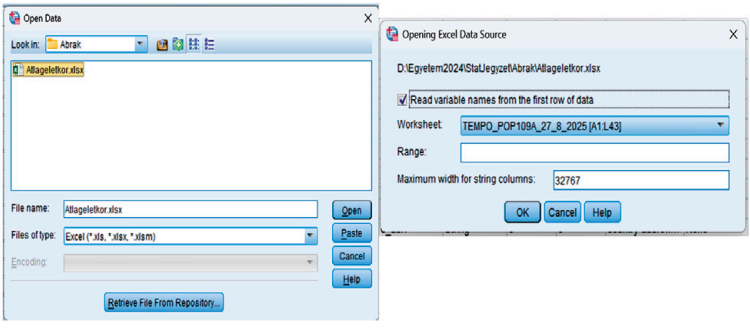
4. ábra. *Az átlagéletkorok lekérése az INS oldaláról (4. példa)*

Mielőtt a letöltött Excel-fájlt beolvasnánk az SPSS-be, először megfelelő formára hozzuk, hogy csak egy fejléces sor maradjon benne és a megfelelő adatsorok (töröljük az 1, 2, 4, 5, 48, 49 és 50 sorokat, illetve az A és B oszlopokat – 5. ábra), majd Excel-fájlként elmentjük (pl. *Atlageletkor.xlsx* néven).

A	B	C	D	E	F	G	H	I	J	K	L
	Anul 2012	Anul 2013	Anul 2014	Anul 2015	Anul 2016	Anul 2017	Anul 2018	Anul 2019	Anul 2020	Anul 2021	Anul 2023
Bihor	39,9	40,1	40,3	40,5	40,7	40,9	41,1	41,2	41,4	41,4	41,4
Bistrita-Nasaud	39,4	39,6	39,8	40	40,2	40,3	40,5	40,6	40,7	40,6	40,7
Cluj	40,5	40,8	41	41,1	41,3	41,5	41,6	41,7	41,8	42	42,1
Maramures	39,5	39,8	40,1	40,4	40,6	40,8	41,1	41,3	41,4	41,6	41,9
Satu Mare	39	39,3	39,6	39,8	40	40,2	40,4	40,6	40,7	40,8	41,2
Salaj	41	41,2	41,4	41,5	41,8	41,9	42,1	42,2	42,4	42,3	42,3
Alba	41,8	42,1	42,3	42,6	42,8	43	43,2	43,5	43,6	43,6	43,7
Brasov	40,5	40,7	40,9	41	41,2	41,4	41,5	41,6	41,6	41,8	42
Covasna	39,6	39,9	40,1	40,3	40,6	40,8	41	41,2	41,3	41,4	41,6
Harghita	39,8	40	40,3	40,5	40,8	41	41,2	41,4	41,6	41,7	41,8
Mures	40,4	40,6	40,8	41	41,2	41,3	41,5	41,6	41,7	41,8	42
Sibiu	39,5	39,7	39,9	40,1	40,3	40,5	40,6	40,8	40,9	41,1	41,5
Bacau	40,4	40,6	40,7	40,8	41	41,1	41,2	41,3	41,4	41,5	41,8
Botosani	40,2	40,4	40,6	40,8	41	41,2	41,4	41,5	41,5	41,4	41,4
Iasi	38	38,3	38,4	38,5	38,6	38,8	38,8	38,9	38,9	38,9	39
Neamt	41,8	42	42,1	42,3	42,4	42,6	42,7	42,8	42,8	43	43,4

5. ábra. *Az átlagéletkorokat tartalmazó megfelelő formára hozott Excel-fájl (4. példa)*

Az SPSS-ben a korábban leírt lépések szerint (*FILE, Open, Data, Files of type .xlsx, Atlageletkor.xlsx, Open, Read variable names from the first row of data, Ok)* beolvassuk az Excel-fájlt az SPSS-be (6. ábra).



6. ábra. *Az átlagéletkorokat tartalmazó Excel-fájl beolvasása SPSS-be (4. példa)*

△ *Gyakorlófeladatok Excel-fájlok SPSS-be való beolvasásához*

Adott az alábbi kérdőívrészlet:

K1. Kérjük, sorolja fel, hogy kik laknak Önnel egy háztartásban, és adjon meg róluk néhány adatot! Önnel kezdjük, és ha ötnél többen élnek együtt, csak a házastárs legidősebb tagjainak adatait vesszük fel.

Családi státusza a kérdezetthez viszonyítva (pl. férj, gyerek stb.)	**Neme** 1. férfi 2. nő	**Születési éve** 9999. NT/NV	**Iskolai végzettsége** 1. alapfokú 2. középfokú 3. felsőfokú 9. NT/NV	**Jelenleg mit csinál?** 1. gazdaságilag aktív 2. gazdaságilag inaktív
Kérdezett				
……..				
……..				
……..				
……..				

Feltételezzük, hogy a kérdőívet 100 személy töltötte ki.

1. A kérdőívrészletre adott fiktív válaszok (tetszőleges értékek, de vigyázzunk arra, hogy változónként ne mind egyforma válaszaink legyenek, és az értékek a reális tartományban maradjanak, pl. a *Neme* esetében csak 1-es és 2-es értékek legyenek) alapján hozzunk létre egy adatbázist Excelben, majd mentsük el *Haztartas.xlsx* néven! Az Excel-adatbázisban adjunk fejlécet az egyes változóknak!
2. Olvassuk be SPSS-be az adatbázist, majd definiáljuk a változókat a kérdőívrészlet alapján! Mentsük el *Haztartas.sav* néven!

Adatfájlok összekapcsolása az SPSS-ben

A program lehetőséget ad különböző SPSS-adatbázisok összekapcsolására. Adatmátrixról lévén szó, két lehetőségünk van:

1. Olyan adatbázisokat ragasztunk össze, amelyek ugyanazokat a változókat tartalmazzák, de más-más esetekre vonatkoznak (pl. egy kérdőíves felmérés kitöltött kérdőíveit több személy vezette be számítógépbe úgy, hogy X az A településen lekérdezetteket, Y pedig a B településen lekérdezetteket).

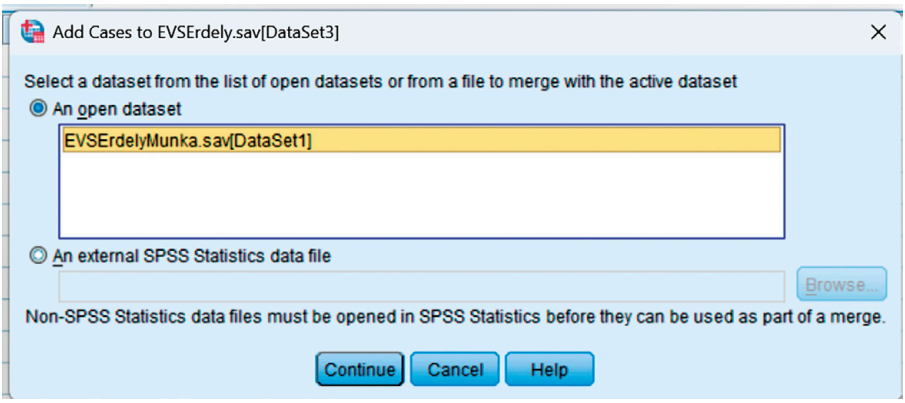
2. Olyan adatbázisokat ragasztunk össze, amelyeknél ugyanazok a megfigyelési egységek/esetek, de különböző változók szerepelnek (pl. egy kérdőíves felmérés kitöltött kérdőíveit több személy vezette be számítógépbe úgy, hogy X minden kérdőív első 20 kérdését, Y pedig minden kérdőív utolsó 10 kérdését).

Az első esetben a *DATA* főmenü *Merge Files*, *Add Cases* menüpontjával, a második esetben a *Merge Files*, *Add Variables* menüponttal dolgozunk. Mindkét esetben a megnyíló ablakban kiválasztjuk a megnyitott adatbázishoz kapcsolni kívánt fájlt, majd az *Open* gombra kattintunk. Mindkét esetben az SPSS lehetőséget ad arra, hogy ellenőrizzük az új, összeragasztott adatbázis változóit és módosítsunk rajta [a megnyitott adatbázisunk változóit (*)-gal, az importált adatbázis változóit pedig (+)-szal jelöli]. A második esetben (*Add Variables*) összekapcsolhatjuk a két adatállományt vakon (azaz semmi összekötő kulcs nélkül, csupán a sorok sorrendjére bízva azokat), és összeköthetjük azonosító kulcs (egy vagy több változó) segítségével. Ez utóbbi esetben a különböző soroknak különböző azonosító kulcsa kell legyen, ellenkező esetben véletlenszerű az összekapcsolás, és itt kötelezően a kulcsváltozó(k) szerinti sorrendbe kell rendeznünk mindkét adatállományunkat (a *DATA, Sort Cases* segítségével).

5. példa ▼

▶ *Adatbázisok egyesítése az SPSS-ben*

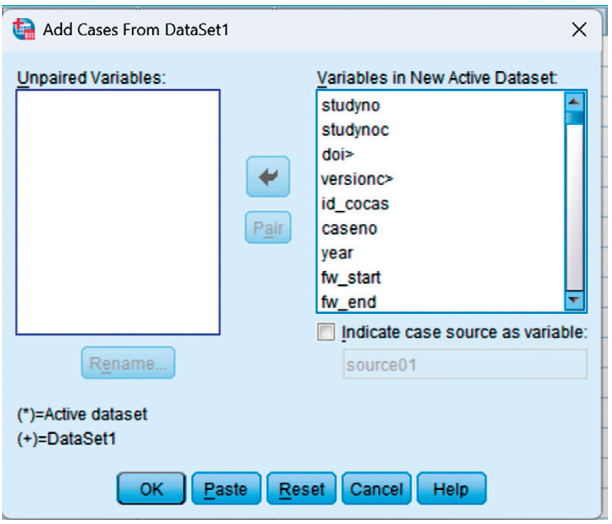
Feltételezzük, hogy meg szeretnénk duplázni az EVS erdélyi adatbázisában szereplő esetek számát. Ekkor vagy egy mappában másolatot készítünk az adatbázisról és más nevet adunk neki (pl. *EVSErdelyMunka.sav*), vagy a megnyitott adatbázis (*EVSErdely.sav*) adatait másoljuk egy új adatbázisba a *DATA, Copy Dataset* segítségével. A két adatbázis egységesítésére a *DATA, Merge Files, Add Cases* utat választjuk. Amennyiben a hozzáadandó adatbázis már meg van nyitva, az *An open dataset* mezőből választjuk ki a megfelelő fájlt, ha nincs megnyitva, akkor az *An external SPSS Statistics data file* mezőből, ahogyan ezt a 7. ábra mutatja.



7. ábra. *Az egyesíteni kívánt fájlok összekapcsolása az SPSS-ben (5. példa)*

A *Continue* után az SPSS lehetőséget ad arra, hogy ellenőrizzük az új, összeragasztott adatbázis változóit és módosítsunk rajta (8. ábra). Esetek egysé-

gesítésekor arra kell figyelnünk, hogy az *Unpaired Variables* mező üres legyen, vagyis ne legyen olyan változó, amelyik csak az egyik adatbázisban szerepel.



8. ábra. *Változók ellenőrzése az SPSS-ben az adatbázisok összekapcsolásakor (5. példa)*

Megfigyelések leválogatása az SPSS-ben

A megfigyelések/esetek szelektálása SPSS-sajátosság. Miként a neve is jelzi, olyankor használjuk, amikor nem a teljes adatbázissal, hanem csak annak egy részével kívánunk dolgozni.

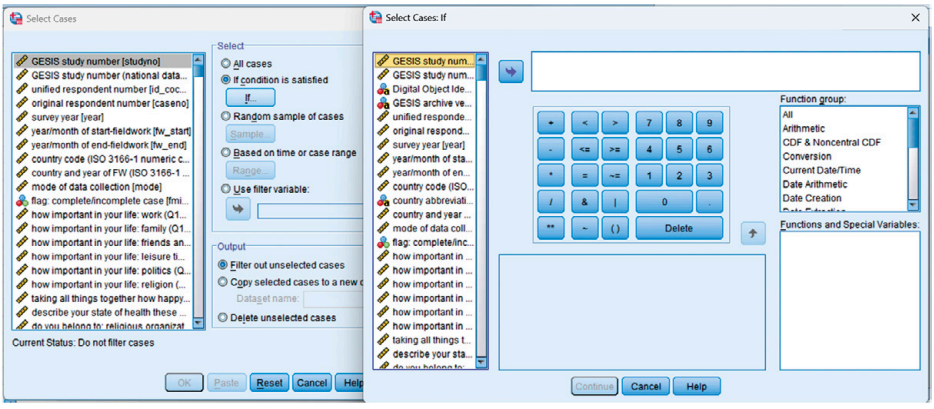
A leválogatásra több lehetőségünk is van a *DATA* főmenü *Select Cases* almenüjében.

Az *If condition is satisfied, If* mezőnél egy vagy több változó értékei szerinti feltételes leválogatást hajthatunk végre numerikus és logikai műveletek segítségével. Mint a legtöbb SPSS-főablakban, ebben is (bal oldalon) megtalálható az összes változó, amivel jelenleg dolgozunk. Jobb oldalon helyezkednek el (egy számológépre emlékeztető rész formájában) a különböző műveleti és numerikus gombok. Az ismerős műveleti jeleken kívül (**+**, **–**, *****, **/**) vannak olyanok is, amelyek az egyszerű számológépeken nem találhatók meg. Ilyen pl. az **&**, a **~** stb., ezek logikai műveletek elvégzését teszik lehetővé, amelyekről az 1. táblázat nyújt összefoglalót.

1. táblázat. *A különböző logikai műveletek jelentése*

Jel	Jelentése
<	„Kisebb, mint…”
>	„Nagyobb, mint…”
<=	„Kisebb vagy egyenlő, mint…”
>=	„Nagyobb vagy egyenlő, mint…”
=	„Egyenlő”
~=	„Egyenlőtlenség”
&	„És”
|	„Vagy”
~	„Nem”

A numerikus gombok mellett található még egy ablak, a *Functions,* amely előre elkészített utasításokat, függvényeket tartalmaz, egyszerűbbeket és bonyolultabbakat is (9. ábra).



9. ábra. *Az esetek leválogatása az SPSS-ben*

Az SPSS választási lehetőséget kínál, hogy miként kezelje a leválogatott eseteket. Alapértelmezésben a szűrőfeltételnek nem megfelelő, ideiglenesen kizárt esetek sorszáma át van húzva, illetve az adattábla végén egy új változó (*filter_$*) is jelzi, hogy mely esetek lettek szelektálva. A leválogatás eredményét háromféleképpen kezelhetjük a *Select Cases* menü jobb alsó részén, az *Output* opciónál:

1. alapértelmezésben a *Filter out unselected cases*, vagyis a (meg)szűrt esetek szerepelnek (fizikailag továbbra is minden adatunk az adatbázisban van) az adatbázisban,

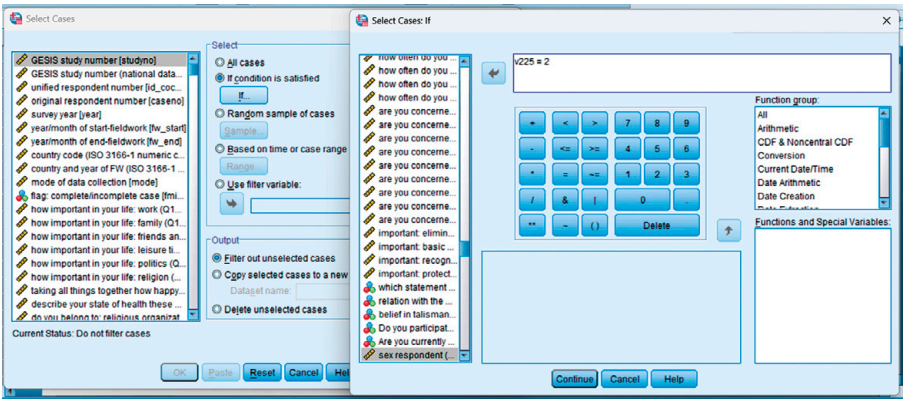
2. a kiválogatott eseteket egy új adatbázisban elmenthetjük (*Copy selected cases to a new dataset*),
3. kitöröltetünk minden olyan esetet, amelyikkel nem dolgozunk (*Delete unselected cases*) – ebben az esetben nagyon kell figyelni arra, hogy a teljes adatbázisunk még valahol meglegyen, mivel ennek létrehozása rendkívül időigényes munka.

Minden esetleválogatáskor nagyon figyeljünk arra (erre az SPSS *Data Editor* ablak jobb alsó sarkában levő *Filter on* jelzés is figyelmeztet), hogy amikor befejeztük a részsokaságunk elemzését, és újra a teljes adatbázissal szeretnénk dolgozni, mindig vegyük vissza a leválogatási feltételeinket (*DATA, Select Cases, All cases*, a *Select Cases* menü jobb felső része).

6. példa ▼

▶ *Esetleválogatás az SPSS-ben*

Feltételezzük, hogy egy sajátos kérdésben csak a nőkről szeretnénk részelemzést készíteni. Munkaadatbázisunkból (*EVSErdely.sav*) ezért a férfiakat „kiszűrjük". Ekkor a következőképpen járunk el: a *DATA* főmenü *Select Cases* almenüjében az *If condition is satisfied* almenü *If* mezőjére kattintva átvisszük a *sex respondent (Q63)* változót (a válaszadó neme, az erdélyi kérdőívben a K57-es kérdés, a *v225* – a *Variable View*-ban a 265. sorszámú változó az adatbázisban), majd megadjuk a leválogatás feltételét, vagyis hogy a változó értékei legyenek egyenlőek 2-vel (2. Nő). Miután megadtuk a leválogatás feltételét, tehát *v225* = 2, *Continue*-t klikkelünk, majd visszaérve a *Select Cases* almenübe az *Ok* gombra kattintunk (10. ábra). A leválogatás eredményeként az adatbázist leszűkítettük az 577 női válaszadóra.



10. ábra. *A nők leválogatása (6. példa)*

Miként már korábban említésre került, több változó szerint is lehet feltételes leválogatási parancsot adni. Ha tovább szeretnénk szűkíteni a kört, és

pl. csak 5000 fő alatti településeken élőket szeretnénk vizsgálni, akkor a következőképpen adjuk meg a parancsot: *v225* = 2 & *v276_r* = 1. A *v276_r* nem más, mint a megkérdezett lakóhelyének nagyságára vonatkozó változó (*size of town where interview was conducted Q106 (5 categories)*, vagyis az interjú helyszínéül szolgáló település nagysága 5 kategória szerint – ez az adatbázisban a 377. sorszámú változó, ami a magyar kérdőívben a K79-es kérdésből lett képezve). Az 1-es kód az 5000 lakos alatti településnagyságot jelöli. A logikai feltételek közül az „és" logikai feltételt alkalmazzuk, mivel azt szeretnénk, hogy feltételeink közül mindkettő teljesüljön. Természetesen ugyanazt a leválogatási feltételt többféleképpen meg lehet adni, a legfontosabb, hogy mindig ellenőrizzük a leválogatás eredményét. A leválogatás következményeként az adatbázist leszűkítettük 232 fő 5000 lakos alatti településen élő női válaszadóra.

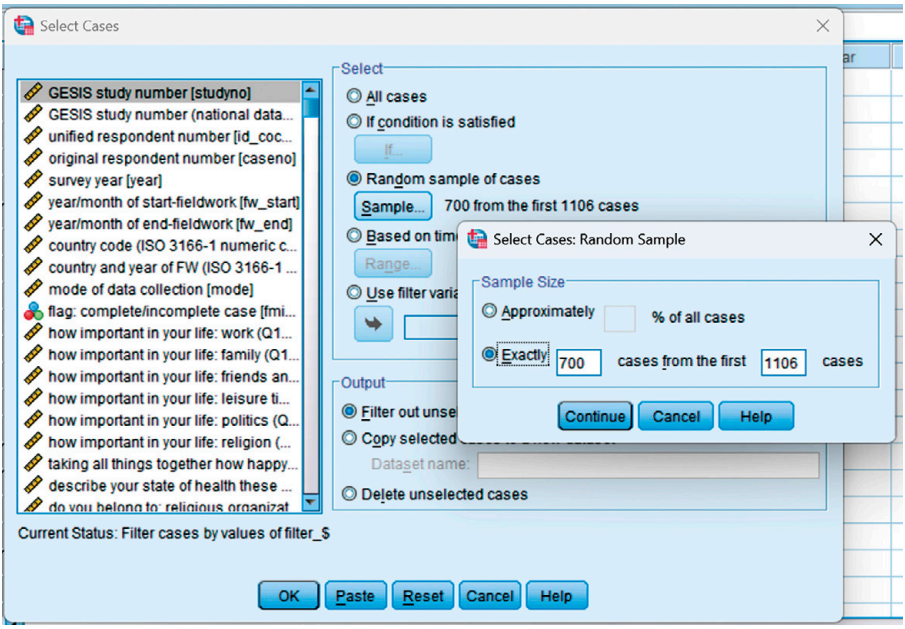
Mintavétel az SPSS-ben

A mintavétel is tulajdonképpen esetleválogatást jelent, hiszen akkor használjuk, amikor nem a teljes adatainkból, hanem azoknak csak egy véletlen halmazából kívánunk dolgozni (a mintavételről lásd bővebben a *3. Mintavétel* című fejezetet). Az SPSS-ben két lehetőségünk van a mintavételre: vagy arra utasítjuk a programot, hogy az esetek bizonyos százalékának megfelelően alkosson véletlen mintát, vagy megadjuk a kívánt mintánk pontos esetszámát. Mindezt szintén a *DATA* főmenü *Select Cases* almenüjében, a *Random sample of cases* segítségével lehet megvalósítani.

7. példa ▼

▶ *Véletlen mintavétel az SPSS-ben*

Ha például adatbázisunkból egy 700 fős véletlen mintát szeretnénk venni, a *DATA* főmenü *Select Cases* almenüjében a *Random sample os cases, Sample* mezőjére kattintunk, és utasítjuk az SPSS-t, hogy pontosan egy 700 fős véletlen mintát válasszon az első 1106 (az összes) eset közül, majd *Continue*-t és végül *Ok*-t kattintunk (11. ábra).



11. ábra. *Mintavétel az SPSS-ben (7. példa)*

Ennek a műveletnek nyilvánvalóan csak szemléltető szerepe van, hiszen az SPSS gyakorlatilag ugyanolyan gyorsan elemez 1106 esetet, mint 700-at. Erre az eljárásra olyan esetben van szükség, amikor van egy adatállományunk egy intézményen belül a személyekről (például a Sapientia EMTE diákjainak azonosító adataiból, ami alatt név, kar, szak, évfolyam, csoport értendő) vagy egy nagyváros háztartásairól (a villamosművek vezetősége a fogyasztókról óhajt véleménykutatást végezni/végeztetni), és szükségünk van egy egyszerű véletlen mintára, mivel a teljes sokaság igen nagyszámú esetből áll. Ilyenkor a mintautasítás eredményét lapra rendezve kinyomtatjuk és a kérdezőbiztosokhoz eljuttatjuk.

1.6. Változók átalakítása vagy transzformációja az SPSS-ben

Ahhoz, hogy az adatbázisunkban szereplő változóinkkal dolgozni tudjunk, legtöbb esetben módosítanunk, alakítanunk kell rajtuk. Elég, ha csak arra gondolunk, hogy minden elemzés előtt meg kell tisztítanunk adatainkat a nem releváns válaszoktól, össze kell vonnunk, csoportosítanunk kell adatainkat. Az SPSS-ben minden, a meglévő adatsokaságunk változtatásához (transzformációjához), új változók létrehozásához szükséges alkalmazás a *TRANSFORM* főmenüben ta-

lálható. A *TRANSFORM*-on belül megjelenő menüsor elemei közül a leggyakrabban használt négyet, vagyis a különböző számítások, matematikai műveletek elvégzésére használatos *Compute,* az egyes változóértékek többszöri előfordulása összegzésére használatos *Count Values within Cases*, az átkódolásra használt *Recode*, valamint a szöveges adataink kezelésére használható *Automatic Recode* alkalmazásokat ismertetem.

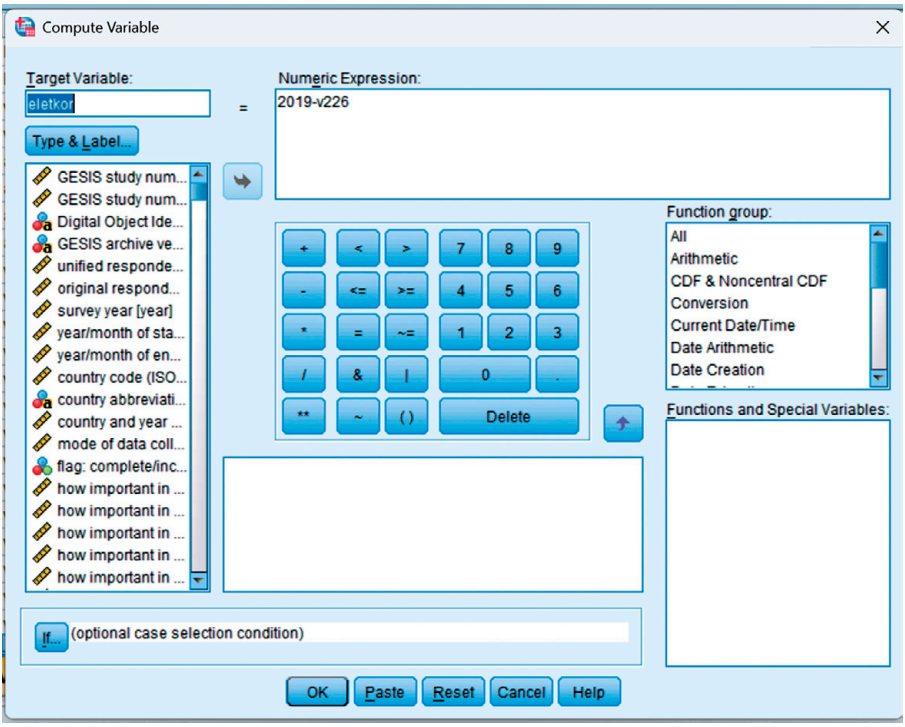
A *Compute* almenü

Mint a legtöbb SPSS-főablakban, ebben is (bal oldalon) megtalálható az összes változó, amivel dolgozunk. Az adatok különféle transzformációinál (pl. a *Recode*-ban) lehetőség van választani, hogy a változtatásokat ugyanabba a változóba (*Into Same Variables*) vagy egy újba (*Into Different Variables*) végezzük el. Jelen esetben a *Compute*-nál azonban erre nincs lehetőség. A program alapértelmezettnek veszi, hogy a változón/változókon a különböző matematikai műveleteket úgy akarjuk végrehajtani, hogy az eredeti változó/változók sértetlenek maradjanak, vagyis nevet kell adnunk az új változónak, mely a már transzformált adatokat fogja tartalmazni. Ezt az új nevet adhatjuk meg a *Target Variable* mezőben, közvetlenül a változók neveit tartalmazó ablak fölött. A *Numeric Expression* elnevezésű ablakban fognak megjelenni a kért változtatások algebrai alakjai, ahogyan ezt már az esetek leválogatásánál (*Select Cases, If)* megismertük. A változók alatt található egy *If…* feliratú gomb. Amennyiben szűkíteni akarjuk a változtatni kívánt adatok körét, ezt az *If…* -re kattintva megjelenő ablak segítségével megtehetjük (ahogyan a *Select Cases*-nél is).

8. példa ▼

▶ *Az életkor kiszámítása* ***Compute****-tal*

Adatbázisunkban szerepel a megkérdezettek születési éve, de mivel ez intervallummérési szintű változó, nagyon könnyen arányskálává tudjuk változtatni olyan módon, hogy életkorrá alakítjuk. Mivel adataink 2019-ből származnak, minket az érdekel, hogy a kérdezés időpontjában a megkérdezettek hány évesek voltak, így 2019-ből kivonjuk minden egyes megkérdezettünk (esetünk) születési évét. Ekkor a *TRANSFORM* főmenü *Compute* almenüjében nevet adunk a létrehozni kívánt új változónknak (*eletkor*), a *Numeric Expression* mezőbe beírjuk az algebrai műveletet: 2019 – v226. A születési év (*year of birth respondent (Q64)*, *v226,* az adatbázis 266. változója, az erdélyi kérdőívben a K58-as kérdés) változót átvisszük bal oldalról a *Numeric Expression* felületre és az *Ok*-ra kattintunk (12. ábra). Ekkor adatbázisunk végén meg fog jelenni az új *eletkor* nevű változónk, amelynek a korábban elmondottak szerint megadjuk a paramétereit. Az új változónkban olyan értékek fognak szerepelni, mint 17, 18, 19…82, tehát a megkérdezettek életkora a kérdezés időpontjában.



12. ábra. *A Compute almenü használata az SPSS-ben (8. példa)*

A *Count Values within Cases* almenü

A *Count Values within Cases* almenüt akkor használjuk, amikor olyan új változót kívánunk létrehozni, amelyben a kijelölt változók együttes előfordulásait szeretnénk regisztrálni. Itt is a *Target Variable* mezőnél nevet adunk az új változónknak, a *Target Label* mezőnél az új értékünk nevét adjuk meg, a *Variables* mezőbe átvisszük azokat a változókat, amelyeknek az együttes előfordulásait vizsgáljuk, majd a *Define Values*-nál megadjuk a vizsgált értéket/értékeket, amelyek érdekelnek. Az *If...* segítségével itt is szűkíthető a vizsgált esetek köre.

9. példa ▼

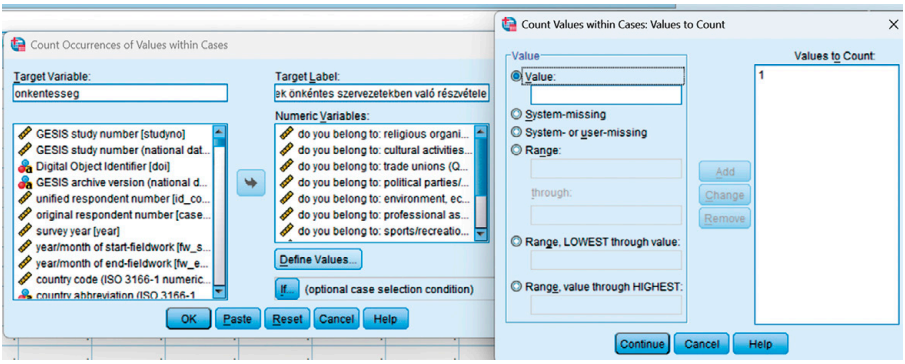
▶ *Az azonos válaszlehetőségek együttes előfordulása több változónál*

Adatbázisunkban a *v9–v19* változók a különböző önkéntes szervezetekben való tagságot regisztrálják a 2. táblázat szerint.

2. táblázat. *Önkéntes szervezetekben való tagság (kérdőívrészlet, K4-es kérdés)*

		Említette	Nem említette	NT	NV
v9	Vallási vagy egyházi szervezetek	1	2	8	9
v10	Oktatási, művészeti, zenei vagy kulturális tevékenység	1	2	8	9
v11	Szakszervezet	1	2	8	9
v12	Politikai párt vagy csoport	1	2	8	9
v13	Környezetvédelem, az élővilág megőrzése, állatok jogai	1	2	8	9
v14	Szakmai szervezet	1	2	8	9
v15	Sport vagy aktív pihenés	1	2	8	9
v16	Humanitárius vagy jótékonysági szervezet	1	2	8	9
v17	Fogyasztói érdekvédelmi szervezet	1	2	8	9
v18	Önsegítő csoport, kölcsönös segítségnyújtásra irányuló csoport	1	2	8	9
v19	Más csoportok	1	2	8	9

Az ebben a formában szereplő adatok esetében egy egyszerű relatív gyakoriság segítségével (lásd a *2.2. Gyakorisági eloszlások* című alfejezetet) rögtön megtudhatjuk, hogy a megkérdezettek hány százaléka tagja vallási, oktatási stb. szervezeteknek, viszont a különböző önkéntes szervezetekben való tagságok együttes előfordulásáról nincs információnk. Amennyiben pl. azt szeretnénk megtudni, hogy a mintánkban szereplő esetek hány százaléka tagja a felsorolt önkéntes szervezetek közül legalább kettőnek, a *Count Values within Cases* menühöz folyamodunk. A *Target Variable* mezőnél az *onkentesseg* nevet adjuk az új változónak, a *Target Label* mezőnél *A megkérdezettek önkéntes szervezetekben való részvétele* nevet adjuk, a *Variables* mezőbe átvisszük a *v9–v19* változókat (11 db), majd a *Define Values*-nál megadjuk az 1 (az „Említette” kategória kódja) értéket, mivel az érdekel, hogy az egyes megkérdezettek a maximális 11 típusú önkéntes szervezet közül hánynak tagjai. Ezt követően *Add*-et és *Continue*-t, majd visszatérve a főablakba *Ok*-t klikkelünk (13. ábra).



13. ábra. *A Count Values within Cases almenü használata (9. példa)*

Ilyen módon tehát létrehoztuk az *onkentesseg* nevű új változót, amelyben 0 és 11 közötti értékek szerepelnek. A 0 azt jelenti, hogy a 11 önkéntes szervezet közül egyiknek sem tagja a kérdezett, az 1, hogy a tizenegy szervezet közül egynek, a 2, hogy a tizenegy közül kettőnek tagja stb. Tehát most már a gyakoriságok alapján (lásd a *2.2. Gyakorisági eloszlások* című alfejezet) meg tudjuk mondani, hogy pontosan 252 fő (a megkérdezettek 22,9%-a) legalább két önkéntes szervezetnek is tagja a felsorolt tizenegyből. Ez és az ehhez hasonló együttes előfordulások sokkal árnyaltabb képet mutatnak az önkéntességről, mint ha csak azt mondjuk, hogy a megkérdezettek közül 129 fő kulturális szervezetekben önkénteskedik, vagy 57 személy szakszervezeti tag.

A *Recode* almenü

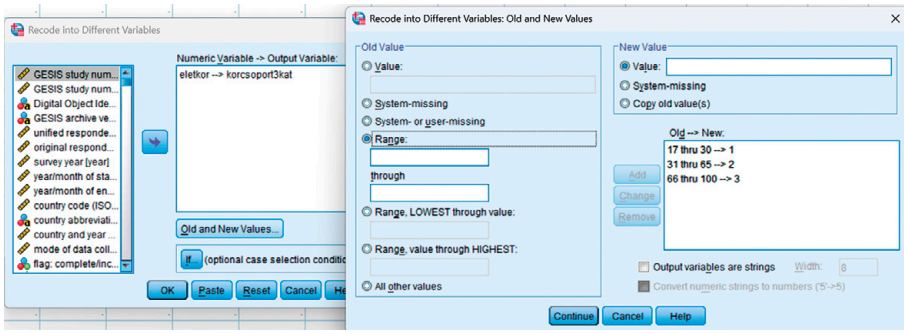
A *Recode* menü a változók legegyszerűbb átkódolására használatos menüpont. Két lehetőségünk van erre: 1. az *Into Same Variables*-szel a kért változtatásokat új változó képzése nélkül hajthatjuk végre (pl. adattisztításnál, amire a példákban használt adatbázis esetében nincs szükség, mivel ez már megtörtént, illetve a nem releváns válaszok már a *Missing* tartományban szerepelnek), és 2. az *Into Different Variables*-szel értelemszerűen a kért változásokat egy új változó létrehozásával végezzük el (pl. csoportosítások esetén). Mindkét esetben, ahogyan ezt már korábban is láttuk, bal oldalon lesz felsorolva az összes mért változó. A jobb oldali kis ablakba (*Variables*, illetve *Input Variable*) kell áttenni azt a változót/változókat, amelyeken az átkódolást végre akarjuk hajtani. Egy vagy több változót egyenként vagy egyszerre is át lehet tenni a jobb oldali ablakba úgy, hogy duplán kattintunk a változóra, vagy kijelöljük és a középen található nyílra kattintunk. Amikor új változóba kódolunk, az *Output Variable*-ben el kell nevezni az új változót, ahol már a képzett csoportok fognak szerepelni. A *Name* mezőbe kell megadni az új változó nevét, majd a *Change* gombra klikkelve aktiváljuk az új változó nevét. Ahogy ez megvan, az ablak legalján található *Old and New Values* mezőben az *Old Value* alatt található *Value* mezőbe kell beírni, hogy

mi a kiválasztott változó(k) eredeti értéke. Jobb oldalon van a *New Value* alatt a másik *Value* mező, ide kell beírni, hogy az eredeti értékből mi legyen. Ezután az *Add* gombra kattintunk, és az *Old-->New* ablakban (jobb oldalon) megjelenik a kért műveleti utasítás. Ugyanígy kell eljárni a tétel összes értékével. Fontos, hogy minden változtatás, amit végre szeretnénk hajtani, az *Old-->New* ablakban szerepeljen. Miután végeztünk, itt ellenőrizzük az utasításokat, mielőtt még a *Continue*-ra kattintanánk, majd a főablakban az *Ok*-ra klikkelünk.

10. példa ▼

▶ *Változók átkódolása*

A változó átkódolására a 8. példánál, a születési évből létrehozott *eletkor* változót használjuk. Azt szeretnénk, hogy a későbbi elemzésekben majd három életkorcsoportunk legyen: 1. a legtöbb 30 évesek, 2. a 31–65 évesek és 3. az 65 év felettiek korcsoportja. Mivel semmiképpen nem szeretnénk elveszíteni az arányskála mérési szintű *eletkor* változónkat, új változóba kódolunk. Úgy járunk el, hogy a *TRANSFORM, Recode Into Different Variables*-szel transzformálunk. Tehát átvisszük balról jobbra az *eletkor* változót, jobb oldalon a *Name* mezőbe adunk egy új nevet (pl. *korcsoport3kat*), majd *Change*. Az *Old and New Values* mezőnél a 17–30 eredeti értékekből (*Old Values, Range:*) 1-es kódszámút (*New Value, Value*) transzformálunk (ők lesznek a legtöbb 30 évesek) és *Add*, a 31–65 eredeti életkor értékekből ugyanígy 2-es kódszámú (31–65 évesek csoportja) értékeket gyártunk, majd *Add*, végül a 66–100 régi értékekből 3-as új értékeket (65 évnél idősebbek kategóriája) és *Add*-et kattintunk. Az egyes értékek egyenkénti bevitele helyett tanácsos tehát a *Range* (terjedelem) gombot használni ott, ahol több egymást követő értéknek azonos új kódja lesz. Miután megnéztük, hogy így akartuk-e kódolni, *Continue*-t kattintunk, visszatérve az előző ablakba pedig *Ok*-t klikkelünk (14. ábra).



14. ábra. *Új változóba való átkódolás (10. példa)*

Az új, átkódolt változónkat egy gyakorisággal leellenőrizzük (204 legtöbb 30 éves, 611 fő 31–65 éves és 291 fő 65 év feletti kell legyen), majd felcímkézzük, és megadjuk az új *korcsoport3kat* változó beállításait.

Az *Automatic Recode* almenü

Ez a menüpont a szöveges (stringes) változók könnyed kezelésében nyújt segítséget. Tulajdonképpen az történik, hogy az SPSS a változó szöveges értékeit azok rangszámaival cseréli fel, tehát minden egymástól különböző jelölés (szó, kifejezés, mondat) mellé egy (rang)számot rendel. Az automatikusan létrehozott új változó már nem a begépelt szövegeket, hanem az ezekhez tartozó kódszámokat tartalmazza, így lényegesen megkönnyítve a csoportosítást és a további elemzést. Akár a többi alkalmazásnál, itt is ki kell választani az adatbázisban szereplő változók közül azt, amelyiket át szeretnénk kódoltatni, majd a *Variable --> New Name* mezőnél új nevet adunk a változónak és *Ok*-t klikkelünk.

△ *Gyakorlófeladatok átkódolásra*

Nyissuk meg az *EVSErdely.sav* adatbázist!

1. A K1 kérdés *v1–v6* változóit kódoljuk át új kétértékű változókká úgy, hogy az 1. kategória mindenik esetében a fontos (régi 1 és 2 értékek), a 2. kategória pedig a nem fontos (régi 3 és 4 értékek) válaszlehetőségeket csoportosítsa! A nem releváns válaszokat kódoljuk hiányzó adattá (*System-missing*)! Címkézzük fel a 6 új változót a kérdőív megfelelő sorai és az új kódok jelentése szerint!
2. A K63-as kérdés a kérdezett legmagasabb befejezett iskolai végzettségére vonatkozik, ez az adatbázis 302. sorában szereplő *v243_RO_hu* változó (*highest level education respondent – Q81*). Kódoljuk át ezt a változót egy kétkategóriás *iskola2kat* új változóba úgy, hogy az 1. nem felsőfokú végzettségű (régi 0–18 kategóriák) és 2. felsőfokú végzettségű (régi 19–24 kategóriák) legyen! A nem releváns válaszokat kódoljuk hiányzó adattá (*System-missing*)! Címkézzük fel az új változót az új kódok jelentése szerint!

2. FEJEZET

EGYVÁLTOZÓS ELEMZÉSEK

2.1. Statisztikai alapműveletek, egyszerű elemzések

Statisztikai alapműveletek

A statisztikai alapműveletek, mint az összehasonlítás, csoportosítás, szinte minden statisztikai elemzés részét vagy kiindulópontját képezik. Ezek közül az egyik legfontosabb alapművelet *a sokaság nagyságának meghatározása*. Legfőbb előnye, hogy a valóságról nyújt igen tömör és lényeges számszerű információt (pl. népesség nagysága). Egy megfelelően meghatározott sokaság nagysága mindig valamilyen jelenségnek a valóságban való elterjedtségét, egyfajta fontosságát jellemzi (pl. öngyilkosok száma).

Diszkrét és véges sokaságok esetében ez a művelet egy egyszerű megszámlálást igényel, folytonos és véges sokaságok esetében a sokaság meghatározása valamilyen mérést igényel (pl. havi húsfogyasztás). Nyilvánvaló, hogy a végtelen sokaságok nagysága nem adható meg számszerűen.

Amikor két vagy több, azonos fajta egységekből álló sokaság nagyságát összeadjuk, általában egy nagyobb sokaság egységeihez jutunk (pl. különböző települések lakosságának összeadásával megkapjuk egy nagyobb térség lakosságát). Azt, hogy mit tekintünk tartalmilag homogénnek, összeadhatónak, nemcsak a vizsgált dolog vagy jelenség, hanem az értékelési szempont is befolyásolja. Amennyiben pl. vidéki gazdák mezőgazdasági tevékenységét vizsgáljuk, nem adjuk közvetlenül össze a megtermelt burgonya, répa stb. termékmennyiségeket, de a mezőgazdasági kistermelés nagysága szempontjából ezek értéke a mérvadó, és ekkor már összeadhatjuk.

Több sokaság nagyságát vagy más adatát nem csak összeadhatjuk, hanem egymással *összehasonlíthatjuk*, így szintén a sokaság egészét jellemző számszerű információt nyerünk. Az összehasonlítás vagy az adott jelenség időbeli alakulásáról, vagy területileg eltérő megnyilvánulásairól, vagy pedig egymáshoz valamilyen módon kapcsolódó jelenségek viszonyáról ad tömör, számszerű információt.

Az összehasonlítás többféle lehet: egyszerű felsorolás (idősor vagy területi sor, pl. a népesség száma két különböző évben vagy országban), különbség vagy hányados (viszonyszám) képzése.

Szemben az összeadással, ami kommutatív ($A + B = B + A$), a kivonás nem az ($A - B \neq B - A$), és az sem igaz, hogy ahol a kivonásnak van értelme, ott az

összeadásnak is van. Ha például egy ország lakosságából kivonjuk a városlakók számát, megkapjuk a vidéken élők számát, viszont ha összeadjuk a teljes népességet a városon élők számával, az eredménynek sok értelme nincs.

Különbséget csak akkor számíthatunk, ha az adatok mértékegysége azonos, viszont két adat hányadosa akkor is meghatározható, ha a két adat mértékegysége eltérő. Ilyen módon az osztás vagy hányados képzése az új adatok előállításának egyik legtermékenyebb módja (elég, ha csak a különböző relatív adatokra gondolunk). Az összehasonlító viszonyszámok és az indexszámok mértékegység nélküli, „tiszta" számok. A 3. táblázat több sokaság nagyság- vagy más adatainak összehasonlítását szemlélteti.

3. táblázat. *A sokaságok adatainak összehasonlítása*

A sokaságok jellege	**A sokaságok nagyság- vagy más adatainak**		**A hányados mértékegysége**
	felsorolására	**hányadosára**	
	használt elnevezés		
Időben és/vagy térben különböző sokaságok	Összehasonlító sor (idősor, területi sor)	**Összehasonlító viszonyszám, index** (dinamikus viszonyszám/ területi összehasonlító viszonyszám)	–, illetve %, ‰
Időben és/vagy térben azonos, de különböző fajta egységekből álló sokaságok	–	Intenzitási viszonyszám	a két adat mértékegységének a hányadosa

Forrás: Hunyadi–Mundruczó –Vita 2000. 39.

Az intenzitási viszonyszámok mértékegysége mindig a megfelelő tört mértékegységeinek hányadosa, az összehasonlító viszonyszámokat és indexszámokat leggyakrabban százalékként vagy ezrelékként adják meg.

11. példa ▼

▶ *Dinamikus viszonyszámok számítása*

Nézzük, az alábbi fiktív adatok (amely egy iskola két tanévre vonatkozó különböző adatait tartalmazza) alapján hogyan lehet dinamikus viszonyszámokat számolni (4. táblázat).

4. táblázat. *Egy iskola két tanévre vonatkozó fiktív adatai (11. példa)*

Ssz.	Megnevezés	Mérték-egység	2015	2025	Dinamikus viszonyszám, index (1991 = 100)
1	Diákok átlagos évi száma	Fő	1000	750	750·100/1000 = 75% 2015-höz képest a diákok átlagos évi száma 2025-re 25%-kal (100-75) csökkent.
2	Ebből I–VIII. osztályos	Fő	800	600	600·100/800 = 75% Az I–VIII. osztályosok száma is 25%-kal csökkent.
3	Megírt dolgozatok száma	db (1000)	56	40	40·100/56 = 71,4% A megírt dolgozatok száma 28,6%-kal csökkent.
4	10-es feleletek száma	db (1000)	47	35	35·100/47 = 74,5% A 10-es feleletek aránya közel 25%-kal csökkent.
5	Megtartott órák száma	db (1000)	92	71	71·100/92 = 77,2% A megtartott órák száma 22,8%-kal csökkent.
6	Alkalmazott tanárok száma	Fő	107	100	100·100/107 = 93,5% A tanárok aránya csak 6,5%-kal csökkent.

Ha intenzitási viszonyszámokat számolunk, megkaphatjuk pl. a 2015-ös tanévre az egy tanárra jutó megtartott órák számát: 92 000/107 = 860 óra/tanár. Ha ezt az adatot összevetjük a 2025-ös tanév adatával (71 000/100 = 710 óra/tanár), kiderül, hogy 10 év alatt 17,4%-kal (710·100/860 = 82,6%) csökkent az egy tanárra jutó megtartott órák száma.

△ *Gyakorlófeladatok összehasonlításra*

1. Adott az alábbi táblázat, amely a világ népességét tartalmazza kontinensek szerint 1950-ben, 2017-ben és 2050-ben (fő). Forrás: ENSZ World Population Prospects, 2015.

	1950	2017	2050
Afrika	228 901 723	1 246 504 865	2 477 536 324
Ázsia	1 394 017 757	4 478 315 164	5 266 848 432
Európa	549 089 107	739 207 742	706 792 824
Észak-Amerika	171 614 868	363 224 006	433 113 731
Közép- és Dél-Amerika	168 843 911	647 565 336	784 247 223
Ausztrália és Óceánia	12 681 946	40 467 040	56 609 460
Összesen	**2 525 149 312**	**7 515 284 153**	**9 725 147 994**

2. Adott az alábbi táblázat, amely Románia és Hargita megye lakosságát tartalmazza településtípusok szerinti bontásban 2002-ben, 2011-ben és 2021-ben (fő). Forrás: Nemzeti Statisztikai Hivatal, Népszámlálási adatok (https://www.recensamantromania.ro/).

		2002	**2011**	**2021**
Hargita megye	város	144 083	132 418	118 754
	vidék	182 139	178 449	173 196
Románia	város	11 435 080	10 858 790	9 939 102
	vidék	10 245 894	9 262 851	9 114 713

3. Adott az alábbi táblázat, amely Románia és Kovászna megye lakosságát tartalmazza településtípusok szerinti bontásban 1992-ben, 2002-ben és 2021-ben (fő). Forrás: Nemzeti Statisztikai Hivatal, Népszámlálási adatok (https://www.recensamantromania.ro/).

		1992	**2002**	**2021**
Kovászna megye	város	122 905	111 996	91 593
	vidék	110 351	110 453	108 449
Románia	város	12 391 819	11 435 080	9 939 102
	vidék	10 418 216	10 245 894	9 114 713

Számítsunk viszonyszámokat (%) kézi számítással, és értelmezzük a kapott adatokat mindhárom táblázat alapján! A számolt értékeket tartalmazó táblázatoknak adjunk címet is!

2.2. Gyakorisági eloszlások

Egy ismérv szerinti osztályozás

Egy további gyakran használt alapművelet a valamely adott sokaság egy vagy több ismérv szerinti tagolása, osztályozása. Az osztályozást gyakran csoportosításnak is szokás nevezni. Az osztályozás során egy sokaság *különböző ismérv(ek) szerinti szerkezetét* lehet megismerni, és leggyakoribb célja, hogy a sokaságot valamilyen szempontból homogénebb csoportokra bontsuk. Az osztályok számát nem célszerű túl nagyra választani, mivel további kezelésük nehézkessé válik.

Az osztályozás eredményeként kapott *sokaságrészeket osztályoknak*, az osztályok egymástól való elhatárolására használt ismérveket *csoportképző ismérveknek* nevezzük. Az osztályozás követelményei:

1. teljesség,
2. átfedésmentesség,
3. az eredmény homogén osztályok kialakítása legyen.

Az egy ismérv szerinti osztályozás eredménye csoportosító (gyakorisági) sor formájában adható meg. A csoportosító sor általános formáját az 5. táblázat szemlélteti.

5. táblázat. *A gyakorisági sor általános formája*

Osztály	Egységek száma
C_1	f_1
C_2	f_2
.	.
.	.
C_i	f_i
.	.
.	.
C_k	f_k
Összesen	N

ahol:

C_i – a csoportképző ismérv alapján képzett i-edik osztály azonosítója,
f_i – a sokaság C_i osztályába sorolt egységeinek száma, *gyakorisága*,
k – a kialakított osztályok száma,
N – a sokaság egységeinek a száma, a *sokaság nagysága*.
Nyilvánvaló, hogy:

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

vagyis a sokaság nagysága egyenlő a sokaság különböző osztályaiba sorolt egységei számának summájával (összegével). Az f_i gyakoriságok helyett/mellett *viszonyszámokat* (*relatív gyakoriságokat*) is használhatunk, például ha az előbbieket elosztjuk a sokaság egységeinek számával, *arányszámokat* kapunk, és ha ezeket 100-zal szorozzuk, *százalékos eloszlásokat*, 1000-rel szorozva *ezrelékes eloszlásokat* kapunk.

Az osztályokat definiáló jelölést (pl. C_1) osztályköznek nevezzük. Amennyiben az osztályköz egy intervallum (pl. 15–19 évesek), a végpontokat osztályközhatároknak (15 és 19 év), a köztük lévő távolságot pedig osztályközhosszúságnak (5 év) nevezzük. Amikor az osztályköznek nincs alsó vagy felső határa, nyitott osztályközről beszélünk.

△ *Gyakorlófeladatok relatív gyakoriságok kézi számítására*

1. Adott az alábbi csoportosító sor, amely 170 diák válaszait összesíti azzal kapcsolatosan, hogy mennyit tesznek meg egészségük megóvása érdekében:

	Gyakoriság (fő)
nagyon keveset	5
keveset	43
átlagosat	71
sokat	46
nagyon sokat	5

2. Adott az alábbi gyakorisági sor, amely az *alkoholfogyasztás gyakoriságára* vonatkozik:

	Alkoholfogyasztók (fő)
naponta	287
hetente többször	566
hetente egyszer-kétszer	847
havonta többször	513
ennél ritkábban	412
nem válaszolt	83

3. Adott az alábbi gyakorisági sor, amely a *Facebook-használat gyakoriságára* vonatkozik:

	Facebook-használók (fő)
naponta	1527
hetente többször	524
hetente egyszer-kétszer	321
havonta többször	52
ennél ritkábban	87
nem válaszolt	38

Számítsunk relatív gyakoriságokat, valódi relatív gyakoriságokat és kumulált gyakoriságokat a fenti három táblázat alapján, majd értelmezzük a kapott értékeket!

A gyakorisági eloszlások kiszámítása és ábrázolása az SPSS segítségével

Az elemezni kívánt változó eloszlásának, gyakorisági sorának megtekintése minden elemzés első lépését képezi. Kattintsunk az *ANALYZE* főmenü *Descriptive Statistics* almenüje *Frequencies* parancsára. Ebben a menüben általános információlekérdező parancsok találhatóak, amelyek segítségével a változók legfontosabb tulajdonságait (elemszám, terjedelem, középértékek stb.) tudjuk megtekinteni. A megnyíló ablakban, bal oldalon, minden változó szerepel, amelyek közül kiválaszthatjuk azt/azokat, amelyekre gyakoriságot akarunk kérni. A változók adatbázisban való gyors keresésére ajánlott az SPSS-ben alapbeállításként szereplő, felcímkézett változónevek megjelenítését (hosszú név, *Label*) megváltoztatni a változók rövid nevének (*Name*) megjelenítésére. Ez az *EDIT* főmenüben, az *Options* almenü *General* lehetőségnél, a *Display labels* helyett a *Display names* átállítással tehető meg. Mivel ezzel az átállítással könnyebb megtalálni a változót, viszont nem látszik a változó teljes neve, ezért a jegyzetben használt példákban megtartottuk a *Display labels* alapbeállítást. A változó kiválasztása után egyszerűen *Ok*-t klikkelünk, és az *Output* ablakban máris megjelenik a kért gyakorisági tábla.

Az SPSS-ben lehetőségünk van a gyakorisági sorunk grafikus megjelenítésére is. Ez leggyorsabban szintén az *ANALYZE* főmenü *Descriptive Statistics*, *Frequencies* almenüben oldható meg, az ablak alján, középen található *Chart* segítségével. Itt beállítható a kért diagram típusa (oszlop, kör vagy hisztogramm), valamint megadható, hogy az adatok abszolút vagy százalékos formában jelenjenek meg (gyakoribb a százalékos formában való ábrázolás). Mennyiségi változók esetében tanácsos hisztogramot, kategoriális változók esetében pedig kör- vagy oszlopdiagramot kérni. Kétszer az ábrára kattintva eljutunk az ábraszerkesztőbe (*Chart Editor*).

Az SPSS-ben az adatok grafikus megjelenítésére külön főmenü is van, a *GRAPHS*. Ennek különböző lehetőségeiről a Mellékletben található menüleírásból tájékozódhatunk. Egyváltozós gyakorisági sorok grafikus szemléltetésére tehát a legalkalmasabbak a kör (*Pie*) és oszlop (*Bar*) diagramok. Az általános elv a kettő közötti választáskor, hogy 4-5 válaszlehetőség felett oszlopdiagramot (*GRAPHS, Lagacy Dialogs, Bar, Simple, Define* úton), kevesebb válaszlehetőség esetén pedig kördiagramot (*GRAPHS, Lagacy Dialogs, Pie, Summaries for groups of cases, Define* segítségével) kérünk. Miután mindkét esetben eljutottunk az SPSS-ben a grafikon definiálásáig (*Define*), bal oldalról az ábrázolni kívánt változót át kell tenni oszlopdiagram esetén a *Category Axis*-ra, kördiagram esetén pedig a *Define slices by*-ra. Szintén fontos, hogy ha adatainkat százalékban szeretnénk megjeleníteni, akkor ezt ugyanitt a *Define*-nál a *Bars*/*Slices Represent*-nél *N of cases*-ről (alapbeállítás) tegyük át *% of cases*-re (középen felül). Kétszer az ábrára kattintva eljutunk itt is a *Chart Editor* ablakba, ahol kedvünkre szerkeszthetjük diagramunkat (3D, átszínezés, betűtípusok stb.). Fontos, hogy a számér-

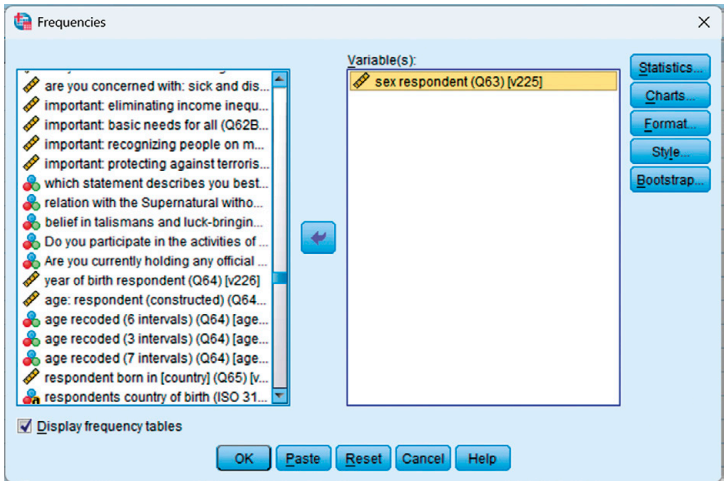
tékeket a *Chart Editor, Elements, Show Data Labels* alatt tudjuk megjeleníteni az ábránkon, illetve a tizedesek a *Properties, Number Format* alatt módosíthatóak (százalékos megoszlásoknál 0 vagy 1 tizedes ajánlott). Itt a változó nevén és értékcímkéin is változtathatunk (pl. magyarul is kiírhatjuk az eredetileg más nyelvű értékcímkéket).

Itt szükséges ugyanakkor megjegyezni, hogy a tanulmányokat szinte kizárólag Wordban írjuk, és az SPSS-ből átmásolt grafikonok a Wordban csak igen kis mértékben engednek meg módosításokat. Ezért ajánlatos a grafikonokat nem az SPSS-ben, hanem pl. Excelben készíteni.

12. példa ▼

▶ *Gyakoriságok lekérése és ábrázolása az SPSS-ben*

Adatbázisunkban a *v225*-ös változó (K57, az adatbázis 265. változója: *sex respondent (Q63)*) a megkérdezettek nemét jelöli. A gyorsabb változókeresés érdekében (ezt a jegyzetben nem tettük meg) ha a fentebb leírtak szerint rövid változónézetre állítjuk az adatbázist (*EDIT, Options, Display labels* helyett a *Display names)*, akkor a változó nevéből csak annyi látszik, hogy *v225*. Tehát a *v225*-ös nem változóra kérünk a fentiek szerint egy gyakorisági táblát (15. ábra).



15. ábra. *A Frequencies almenü használata (12. példa)*

Első lépésként meg kellene tisztítanunk az adatokat a nem releváns válaszoktól (a vizsgált adatbázis esetében ez az alap kezdőlépés nem szükséges, mivel a nem releváns válaszok már a *Missing* tartományban szerepelnek). A kért gyakoriságok a 16. ábrán szemléltetett formában jelennek meg. Az első táblázat azt mutatja, hogy a változóban hány érvényes adat (*Valid*) és hány hiányzó adat (*Missing*) szerepel. A tényleges gyakoriságok a második táblázat-

ban vannak feltüntetve. Az eredménykijelző ablakban szereplő angol kifejezéseket manuálisan magyarul is be lehet írni (mivel ezt minden esetben meg kellene tenni, ezért ettől most eltekintünk).

sex respondent (Q63)

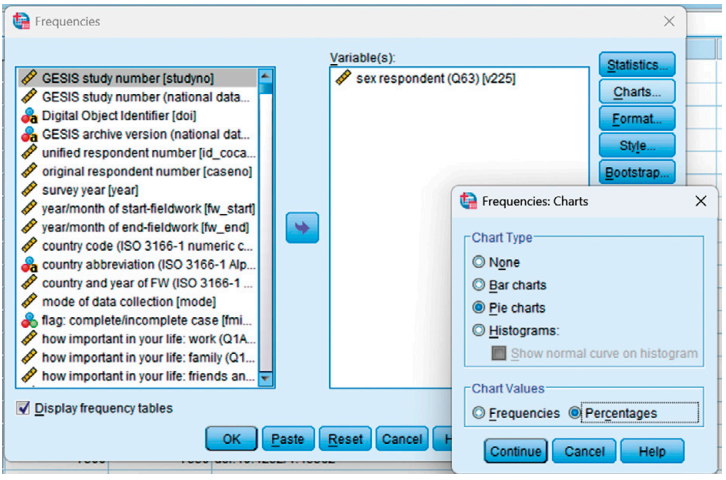
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	529	47.8	47.8	47.8
	female	577	52.2	52.2	100.0
	Total	1106	100.0	100.0	

16. ábra. *Gyakorisági tábla az SPSS-ben (12. példa)*

A gyakorisági tábla első oszlopában (16. ábra, második táblázat) a változó értékei jelennek meg, vagyis a male (férfi) és female (nő), a második oszlop az egyes változóértékek abszolút gyakoriságait mutatja (*Frequency*), a harmadik oszlopban a relatív gyakoriságok olvashatók (*Percent*), a negyedik oszlopban az érvényes relatív gyakoriságok (*Valid Percent*), az utolsó oszlopban pedig a kumulált százalékos gyakoriságok találhatóak (*Cumulative Percent*). Az érvényes relatív gyakoriság nem más, mint az egyes értékek előfordulásainak az érvényes adatokhoz való viszonyítása (amikor érvénytelen adataink is vannak, ezek nem kerülnek be az érvényes százalékok és a kumulált gyakoriságok számításába). A kumulált gyakoriság nem más, mint a valódi relatív gyakoriságok osztályonkénti összeadása.

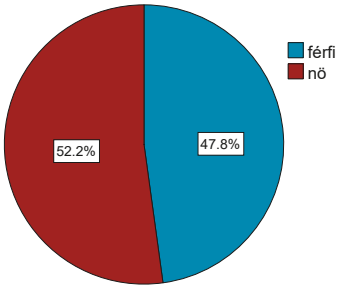
Értelmezvén a gyakorisági táblázatot elmondhatjuk, hogy egyetlen hiányzó adat sincs, az 1106 megkérdezett 47,8%-a (529 fő) férfi, 52,2%-a (577 fő) nő.

A gyakoriságok ábrázolásának legkézenfekvőbb útját (a *Frequencies*-ből) a korábbiakban leírtak szerint a 17. ábra szemlélteti.



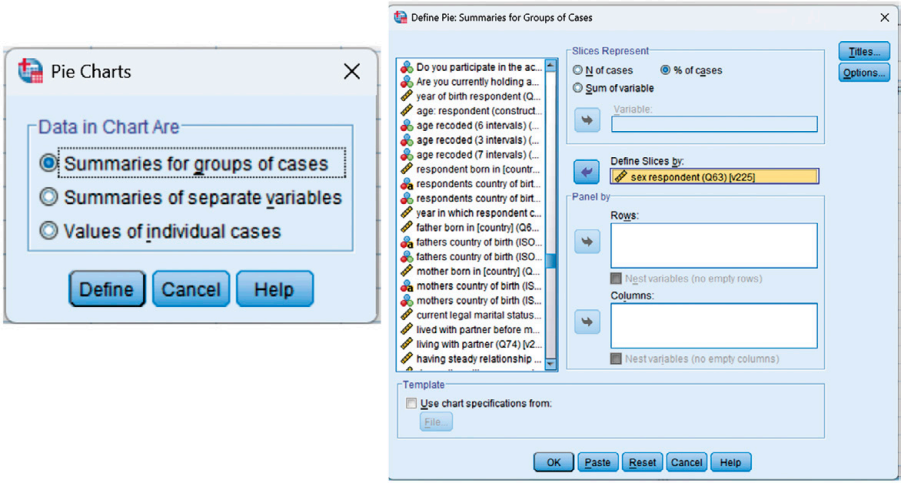
17. ábra. *A gyakorisági sor grafikus megjelenítése a Frequencies-ből (12. példa)*

A vizsgált változó esetében a korábbiakban ismertetett ajánlásnak megfelelően kördiagramot kértünk. Az ábrára kétszer kattintva a *Chart Editor* ablakban kedvünkre alakítottuk diagramunkat, míg a 18. ábrában látható formára hoztuk. Fontos, hogy a számértékeket egytizedes pontossággal jelenítettük meg százalékban, illetve az eredeti angol helyett a magyar attribútumokat adtuk meg (manuálisan át lehet írni).

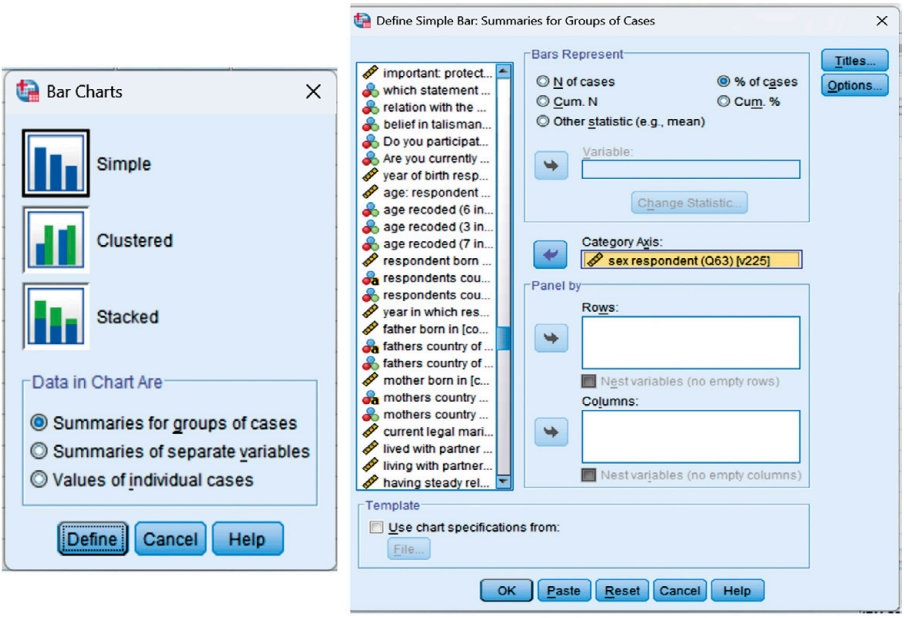


18. ábra. *A kérdezettek nemek szerinti megoszlása (százalékban, N = 1106) (12. példa)*

A *GRAPHS* főmenüből a korábbiakban leírt úton lekérve (19. ábra) is hasonló diagramhoz jutunk. Oszlopdiagramként szintén generálható (20. ábra).



19. ábra. *A kördiagram lekérése a GRAPS-ból (12. példa)*



20. ábra. *Az egyszerű oszlopdiagram lekérése a GRAPS-ból (12. példa)*

△ *Gyakorlófeladatok gyakoriságok SPSS-ben való lekérésére és ábrázolására*
Kérjünk gyakorisági eloszlásokat, értelmezzük és ábrázoljuk (a megfelelő beállításokkal) a K1-es kérdés változóit (v1–v6, a 15–20-as sorszámú változók az adatbázisban)!

Két ismérv szerinti osztályozás: a kereszttábla

A sokaság osztályozással kialakított részeit külön-külön is tovább lehet vizsgálni, ilyenkor az osztályokat részsokaságoknak nevezzük (pl. N1-gyel jelöljük), az egész sokaságot pedig fősokaságnak (N).

A sokaság több ismérv szerinti kombinatív osztályozása révén *kombinációs, kontingencia-* vagy *kereszttábla* elnevezést viselő csoportosítást nyerünk. A kereszttábla belső rovatait *celláknak (rovatoknak)*, az osztályozási ismérvek számát pedig *dimenziószámnak* nevezzük. A kontingenciatábla általános sémáját a 6. táblázat mutatja.

6. táblázat. *A kereszttábla általános formája*

X ismérv szerinti osztályok	Y ismérv szerinti osztályok						
	R_1	R_2	...	R_j	...	R_c	$\sum_j$
C_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1c}	$f_{1.}$
C_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2c}	$f_{2.}$
...	...	...	...	...	...	...	...
C_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{ic}	$f_{i.}$
...	...	...	...	...	...	...	...
C_r	f_{r1}	f_{r2}	...	f_{rj}	...	f_{rc}	$f_{r.}$
$\sum_i$	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.c}$	N

C_i – az X ismérv szerint képzett i-edik osztály azonosítója (i = 1, 2, ..., r),
R_j – az Y ismérv szerint képzett j-edik osztály azonosítója (j = 1, 2, ..., c),
f_{ij} – az a gyakoriság, amelynek egyedei X szerint az i-edik, Y szerint a j-edik osztályba tartoznak,
r – az X szerint képzett osztályok száma,
c – az Y szerint képzett osztályok száma,
$f_{i.}$, $f_{.j}$ – *peremgyakoriságok*.

Nyilvánvaló, hogy:

$$\sum_{j=1}^{c} f_{ij} = f_{i.} \qquad \sum_{i=1}^{r} f_{ij} = f_{.j} \quad \text{és}$$

$$\sum_{i=1}^{r} f_{i.} = \sum_{j=1}^{c} f_{.j} = \sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij} = N$$

13. példa ▼

▶ *A kereszttáblák értelmezése*

Nézzük az alábbi kereszttáblát, amely egy új törvény bevezetésével kapcsolatos véleményeket tartalmazza, nemek szerinti bontásban (7. táblázat).

7. táblázat. *A vélemények nemek szerinti bontásban, abszolút gyakoriságok (fiktív adatok) (13. példa)*

	Nő	**Férfi**	**Összesen**
Egyetért	30	80	**110**
Nem ért egyet	70	50	**120**
Összesen	**100**	**130**	**230**

A relatív gyakoriságokat úgy számoljuk ki, hogy a nők esetében a 30 egyetértő nőt viszonyítjuk az összes nő számához (30·100/100), a 70 nem egyetértő nő számát pedig szintén az összes nő számához (70·100/100). A férfiak esetében a 80 egyetértő férfit a 130 fő összes férfihez (80·100/130), az 50 nem egyetértő férfit pedig szintén a 130 fő összes férfi számához arányítjuk (50·100/130). Tehát kiszámolva a relatív gyakoriságokat a *Nem* változó szerint, a 8. táblázat adatait kapjuk.

8. táblázat. *A vélemények nemek szerinti bontásban (relatív gyakoriságok) (13. példa)*

	Nő	**Férfi**
Egyetért	30,0	61,5
Nem ért egyet	70,0	38,5
Összesen	**100%**	**100%**

A 8. táblázatban lévő kontingenciatábla alapján kijelenthetjük, hogy a nők 70%-a nem ért egyet, 30%-a egyetért, a férfiak 61,5%-a egyetért, 38,5%-a pedig nem ért egyet az új törvény bevezetésével. Viszont ez még mindig nem árul el semmit arról, hogy a nők vagy a férfiak értenek-e egyet nagyobb arányban, holott a kereszttábla lényege az alcsoportok összehasonlítása. A kereszttáblát mindig relatív gyakoriságok alapján olvassuk. Miként a fenti példában is, leggyakrabban a független változó szerint (a *Nem* változó szerint, mivel ez befolyásolhatja a kérdéssel való egyetértést, és nem fordítva) százalékolunk. A relatív gyakoriságok azt mutatják, hogy a férfiak nagyobb arányban értenek egyet a törvény bevezetésével, mint a nők (a nők nagyobb arányban nem értenek egyet a törvény bevezetésével, mint a férfiak), vagy a férfiak körében nagyobb az egyetértők aránya, mint a nők körében. Bár a fenti példánkban következtetésünk nyilvánvalónak tűnik az abszolút gyakoriságok alapján is, figyeljünk arra, hogy *mindig relatív gyakoriságok* alapján olvassuk a kereszttáblákat (egyáltalán nem mindegy, hogy jelen esetben hány férfi és hány nő törvény bevezetésével való egyetértését ismerjük).

△ *Gyakorlófeladatok kereszttáblák kézi számítására*

1. Adott az alábbi kereszttábla:

	Van gyereke	**Nincs gyereke**	**Összesen**
Jár színházba	251	311	**562**
Nem jár színházba	328	403	**731**
Összesen	**579**	**714**	**1293**

2. Adott az alábbi kereszttábla:

	35 évnél fiatalabb	**35 évnél idősebb**	**Összesen**
Sízik	532	618	**1150**
Nem sízik	401	814	**1215**
Összesen	**933**	**1432**	**2365**

3. Adott az alábbi kereszttábla:

	Férfi	**Nő**	**Összesen**
Volt színházban ebben a hónapban	28	70	**98**
Nem volt színházban ebben a hónapban	37	41	**78**
Összesen	**65**	**111**	**176**

Számítsunk relatív gyakoriságokat a magyarázó változó szerint mindhárom kereszttáblára, majd értelmezzük!

Kereszttábla készítése és ábrázolása az SPSS-sel

Akárcsak a gyakorisági táblákat, kontingenciatáblákat is az *ANALYZE* főmenü *Descriptive Statistics* almenüjében, viszont a *Crosstabs...* menüpontnál készíthetünk. A bal oldalon szereplő változók közül kiválasztjuk azt a kettőt (többet is lehet, de minél több dimenziós a kereszttáblánk, annál kevésbé áttekinthető), amelyikre kereszttáblát kérünk, majd a *Cells* gombnál beállítjuk, hogy sorra vagy oszlopra (ahol a magyarázó/független változó van) százalékoljon a program, és *Continue*-t, majd *Ok*-t kattintunk.

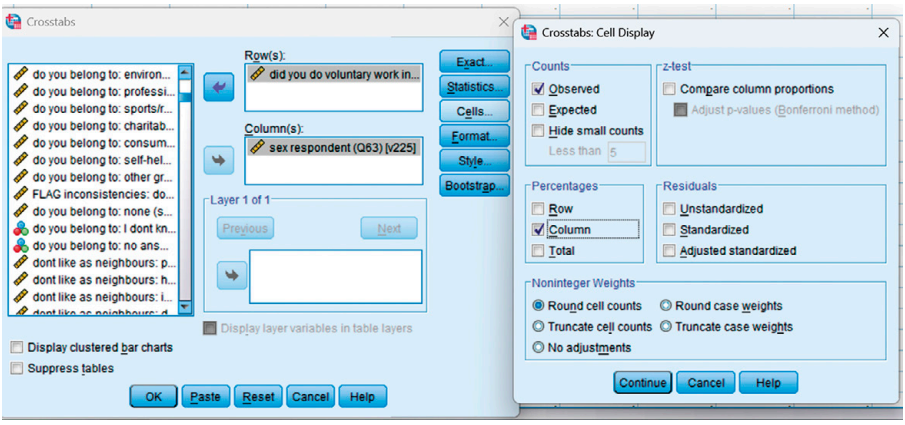
Kétdimenziós kereszttábla ábrázolására az SPSS-ben két lehetőségünk van a *GRAPHS* menüben (bár a *Crosstabs…* almenüben is van rá lehetőség, ez nem az optimális megoldás rá). Az 1. út a *GRAPHS, Lagacy Dialogs*, *Bar*, *Stacked*, ahol a *Category Axis*-nál kell szerepeljen a független/magyarázó változó, a *Define Stacks by*-nál pedig a függő változó (az adatok megjelenítését első körben a *N of cases*-en kell hagyni a *Define*-nál). Kétszer az ábrára kattintva lehet szerkeszteni a *Chart Editor* ablakban (százalékban ábrázolni: *Options, Scale to 100%*, illetve megjeleníteni az ábrán az értékeket az *Elements, Data Label Mode* segítségével). A 2. lehetőség a *GRAPHS, Bar, Clustered* út választása, itt a *Category Axis*-nál kell szerepeljen a függő változó, a *Define Clusters by*-nál pedig a magyarázó változó.

Ennél a második lehetőségnél fontos, hogy a *Define* mezőnél *% of cases*-re kell állítani az adatok megjelenítését. Kétszer az ábrára kattintva kell tovább szerkeszteni: megjeleníteni az értékeket (*Elements, Show Data Labels*) és átállítani a tizedeseket, ha szükséges.

14. példa ▼

▶ *Kétdimenziós kereszttábla készítése és ábrázolása az SPSS-ben*

Készítsünk egy kereszttáblát az adatbázisunkban szereplő *v225*-ös változó (K57, az adatbázis 265. változója: *sex respondent (Q63)*, vagyis a megkérdezettek nemét jelöli) és a *v21* (K5. az adatbázis 38. változója: *did you do voluntary work in the last 6 months (Q5)*, vagyis végzett-e a megkérdezett önkéntes munkát az elmúlt 6 hónapban) változók között, a fentiek szerint (a nemre százalékoltatva) (21. ábra).



21. ábra. *A Crosstabs almenü használata (14. példa)*

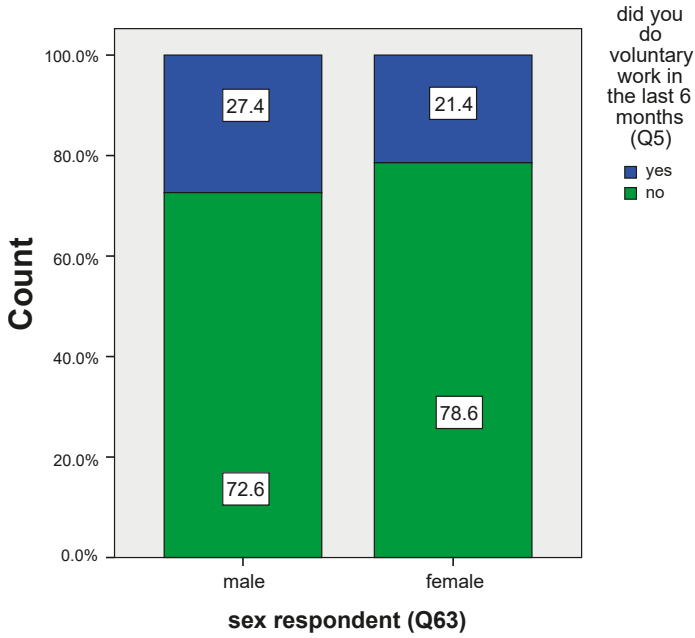
Az SPSS által generált, magyar nyelvűre átírt kereszttáblánkat a 22. ábra mutatja.

			Nem		
			férfi	nő	Összesen
Önkéntesség az elmúlt hat hónapban	igen	Gyakoriság	144	123	267
		% a nemenkénti válaszadók körében	27.4%	21.4%	24.3%
	nem	Gyakoriság	382	451	833
		% a nemenkénti válaszadók körében	72.6%	78.6%	75.7%
Összesen		Gyakoriság	526	574	1100
		% a nemenkénti válaszadók körében	100.0%	100.0%	100.0%

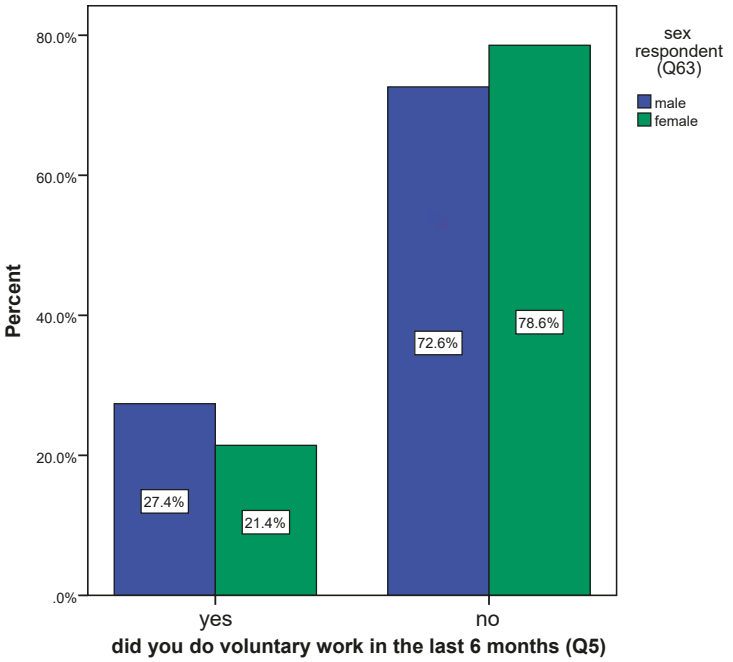
22. ábra. *A megkérdezettek nem és önkéntesség szerinti bontásban (14. példa)*

A kereszttábla adatai (22. ábra) alapján elmondhatjuk, hogy a férfiak 27,4%-a önkénteskedett az elmúlt 6 hónapban, 72,6%-a nem. A nők 21,4%-a önkénteskedett, 78,6%-a nem végzett ilyen tevékenységet. Tehát a megkérdezett férfiak egy kicsivel nagyobb arányban önkénteskedtek az elmúlt 6 hónapban, mint a nők.

Grafikus formában a 23. (zsákolt oszlopdiagram, 1. lehetőség) és 24. (klaszteres oszlopdiagram, 2. lehetőség) ábrák szerint néznek ki az adataink (a címkéket magyarra is át lehet írni, a lépések nyomonkövethetősége kedvéért most nem tettük meg).



23. ábra. *A zsákolt (Stacked) oszlopdiagram (14. példa)*



24. ábra. *A klaszteres (Clustered) oszlopdiagram (14. példa)*

△ *Gyakorlófeladatok kereszttáblák SPSS-ben való lekérésére és ábrázolására*

Kérjünk kereszttáblákat (9 db) a *Compute* (8. példa) és *Recode* (10. példa) almenüknél ismertetett módon létrehozott *korcsoport3kat* változó és a K6 kérdés *v22–v30* (az adatbázisban a 39–47. sorszámú) változói között! Értelmezzük a kereszttáblát! Ábrázoljuk a kereszttábláinkat kétféleképpen, megfelelő formára szerkesztve!

Rangsorok

Az ismérvértékek számszerű jellegében rejlő egyik legkézenfekvőbb lehetőség a sokaság egységeinek sorba rendezése a változó nagysága szerint. Ez akkor is igaz, ha a változó ordinális mérési szintű.

A változó értékeinek nagysága szerint növekvő vagy csökkenő sorba rendezhetjük a sokaságot, és ennek eredményét rangsornak nevezzük. Általában *monoton nem csökkenő* módon szokás rangsorolni.

Míg a sokaságnak egy diszkrét ismérv azonos értékeivel bíró egységei gyakorlatilag egyformák az adott ismérv szempontjából (ezért tetszőleges sorrendbe állíthatóak), addig a folytonos vagy folytonosként kezelt diszkrét ismérv azonos értékeivel jellemzett egységek nem feltétlenül egyformák (csak kényszerűségből, a mérés adott pontossága miatt állíthatóak egymás között tetszés szerinti sor-

rendbe). Ha például Románia megyéinek lakosságszámát vizsgáljuk, és adatainkat ezer főben adjuk meg (pl. 329,34), akkor egy elvileg diszkrét változót (amelynek értékei pozitív egész számok: 329 344) folytonosként kezelünk, hiszen a közölt formában a lakosok száma csak bizonyos pontosságra kerekítve adható meg. Ebben az esetben csak kényszerűségből rangsorolhatjuk adatainkat, hiszen nem tudhatjuk, hogy két 329,34 ezer fős lakosú megye közül melyik a népesebb.

A rangsor igen gyakran kizárólag azon célból készül, hogy megkönnyítse az osztályozást. Főként mennyiségi mérési szintű változók esetén használjuk.

2.3. A centrális tendenciák mutatói: átlag, medián, módusz

A középértékek vagy helyzetmutatók olyan mutatószámok, amelyek a sokaság egészét vagy a vizsgált gyakorisági eloszlás helyzetét egyetlen számértékkel jellemzik, így a sokaságok tulajdonságait a legtömörebb formában fejezik ki.

A középértékek legfőbb előnyei:

– közepes helyzetűek (a minimum és maximum értékek között helyezkednek el),
– tipikusak (viszonylag szűk környezetében az összes ismérvértéknek nagy hányada található),
– egyértelműen meghatározhatóak,
– könnyen értelmezhetőek,
– közérthetőek.

A középértékeket két nagy csoportba szokás sorolni: vannak *számított középértékek* (különböző átlagok) és *helyzeti középértékek* (medián és módusz).

Az átlagok matematikai számítások eredményei, az ismérvértékekkel matematikai, számszerű összefüggésben állnak, és értéküket nem befolyásolja az észlelési adatok sorrendje. A számított középértékek: számtani átlag (egyszerű, súlyozott), harmonikus átlag (egyszerű és súlyozott), mértani átlag, négyzetes átlag.

A helyzeti középértékek az értékek nagysága szerint rendezett statisztikai sorban, általában matematikai számítás nélkül jelölhetőek ki, és az ismérvértékek közötti elhelyezkedésüknél fogva jellemzik a sokaságot. A helyzeti középértékek: medián, módusz.

A számtani átlag

Az egyszerű számtani átlag (röviden: átlag) az észlelési adatok (X_i) összegének és az átlagolandó adatok előfordulási számának hányadosa (N), képlete:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Tehát egy mennyiségi változó átlaga a felvett összes érvényes érték számtani középarányosa. Az átlagot csupán *mennyiségi változókra* számítjuk ki (az SPSS-program bármilyen numerikus típusnak definiált változó esetén kiszámítja az átlagértéket, még akkor is, ha annak semmi értelme, pl. a *Nem* változóra is).

Az átlag legfontosabb tulajdonságai

Minden ismérvértéket a számtani átlaggal helyettesítve a sor összege változatlan marad, vagyis megegyezik az eredeti sor összegével. Ha minden ismérvértéket a számtani átlaggal helyettesítünk, akkor az is következik, hogy a helyettesítéssel elkövetett előjeles hibák pontosan kiegyenlítik egymást:

$$\sum_{i=1}^{N}(X_i - \overline{X}) = 0$$

Az ismérvértékek számtani átlaggal való helyettesítése minimálissá teszi a helyettesítéssel elkövetett hibák négyzetösszegét:

$$\sum_{i=1}^{N}(X_i - \overline{X})^2$$

Az átlag egyik legfontosabb sajátossága, hogy eltünteti az észlelt adatok értéknagyságbeli különbségét, viszont egyetlen értéknagyság változása megváltoztatja az átlag értékét (függ minden egyes értéktől).

15. példa ▼

▶ *Az egyszerű számtani átlag kézi kiszámítása*

Nézzük a következő szemléltető példát átlagszámításra. Adott az alábbi, monoton nem csökkenő módon rendezett értéksorunk:

0 ; 0 ; 0,5 ; 0,6 ; 0,8 ; 1 ; 1 ; 1 ; 3 ; 5 ; 10.

A számtani átlagot a következőképpen számoljuk ki:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{0+0+0{,}5+0{,}6+0{,}8+1+1+1+3+5+10}{11} = \frac{22{,}9}{11} = 2{,}08$$

Súlyozott átlag

A számtani átlagot nagyon gyakran nem az egyenként ismert alapadatokból számítjuk ki, hanem egy gyakorisági sor adataiból. Ekkor súlyozott számtani átlagról beszélünk.

A *súlyozott átlagot* úgy számoljuk ki, hogy az X ismérv szerint képzett C_i osztályok gyakoriságait (fi) szorozzuk a C_i osztály ismérvértékével, majd ezen szorzatokat összeadjuk:

$$\overline{X} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + \ldots + f_k \cdot X_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{i=1}^{k} f_i \cdot X_i}{\sum_{i=1}^{k} f_i}$$

Tehát egy súlyozott számtani átlag nagyságát mindig két tényező határozza meg: az átlagolandó értékek nagysága, azaz az X_i értékek sorozata, valamint az átlagolandó értékekhez tartozó f_i súlyszámok egymás közötti aránya, azaz relatív nagysága.

Amikor egy ismérvnek a megfigyelt sokaság egységeinél fellépő értékei egyenként ismertek, akkor a súlyozatlan esetet, ha pedig az ismérvnek a megfigyelt sokaság egységeinél fellépő értékei gyakorisági sorba rendezetten ismertek, akkor súlyozott esetet használunk. Súlyozott esetben az X ismérv szerint képzett osztályok gyakoriságait súlyoknak is nevezik. A súlyok összege mindig N.

16. példa ▼

▶ *A súlyozott számtani átlag kiszámítása*

Nézzünk két feladatot a súlyozott átlagszámításra.

1. Egy diák 4 tárgyból az alábbi jegyeket kapja: 8, 9, 7, 10. Azt is tudjuk, hogy amiből 8-as és 10-es osztályzatot kapott, az két 3 kredites tárgy, 7-est egy 5 kredites tárgyból, 9-est pedig egy 4 kredites tárgyból kapott. A kérdés, hogy hányas lesz a tanulmányi átlaga.

Miként már a *Bevezető*ben is említésre került, a társadalomstatisztikában sokszor előfordul, hogy egyes számítások *matematikai értelemben vett pontossága magyarázatra szorul*. Ebben a példánkban is egy ilyen esettel találkozunk, hiszen az iskolai osztályzat egy ordinális mérési szintű változó (nem tudjuk azt mondani, hogy aki 10-est kap, az kétszer annyit tud, mint aki 5-öst kap), és átlagot csak mennyiségi változókból számítunk. Viszont a mindennapi életben nagyon gyakran előfordul, hogy egyetlen számmal szükséges jellemezni egy személy teljesítményét, rangsort kell felállítanunk, és ilyenkor átlagot számolunk.

$$\overline{X} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + \ldots + f_k \cdot X_k}{f_1 + f_2 + \ldots + f_k} = \frac{3 \cdot 8 + 4 \cdot 9 + 5 \cdot 7 + 3 \cdot 10}{3 + 4 + 5 + 3} = \frac{125}{15} = 8{,}33$$

Értelmezés. A diák négy tantárgyra számított tanulmányi átlaga 8,33 (az iskolai szabályzat szerint ezt az értéket 2 tizedesjegyre kell csonkítani, és nem kerekíteni).

2. Egy iskolai osztályban a gyerekek közül 4-nek nincs testvére, 11-nek 1 testvére van, 5-nek 2 testvére, 1-nek pedig 4 testvére van. Akkor átlagosan hány testvére van az osztályban a gyerekeknek?

$$\overline{X} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + \ldots + f_k \cdot X_k}{f_1 + f_2 + \ldots + f_k} = \frac{4 \cdot 0 + 11 \cdot 1 + 5 \cdot 2 + 1 \cdot 4}{4 + 11 + 5 + 1} = \frac{29}{21} = 1{,}38$$

Értelmezés. Az osztályban a gyerekeknek átlagosan 1,38 testvérük van.

A folytonos változók (pl. jövedelem) sokféle, egymástól eltérő értéket vehetnek fel. Amennyiben az adatokat pontos értékükkel rögzítettük, az SPSS segítségével könnyedén kiszámíthatjuk az átlag pontos értékét. Néha azonban előfordul, hogy adatainkat csoportosított formában rögzítettük (pl. jövedelemkategóriákat adtunk meg a nagyobb válaszolási arány kedvéért), vagy mások által gyűjtött adatokon dolgozunk, ahol a folytonos adatok csoportosított formában szerepelnek. Ebben az esetben az átlagértéket pontosan nem tudjuk kiszámítani, csak jó becslést tudunk adni rá (nem tudjuk, hogy egy intervallumon belül a kisebb érték vagy a nagyobb érték köré tömörülnek az adatok). Alapvető, hogy adataink oly módon legyenek csoportosítva, hogy a változó legalább intervallum-mérési szintű legyen (nem feltétlenül egyenlő hosszúságú intervallumok). Ilyenkor az átlag kiszámításakor az osztályközepeket kell súlyozni. Az osztályközép nem más, mint az egy osztályba tartozó legkisebb és legnagyobb érték számtani átlaga: $(X_{min} + X_{max})/2$.

17. példa ▼

▶ *Átlagszámítás csoportosított adatokból*

A 9. táblázat 40 diák feladatmegoldási idejét tartalmazza, másodpercben kifejezve (3 diák 118–126 másodperc közötti időintervallumban oldotta meg a feladatot stb.).

9. táblázat. *Gyakorisági sor (17. példa)*

Idő (s)	Gy (f_i)	Számítás
118–126	3	$\overline{X} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + \ldots + f_k \cdot X_k}{f_1 + f_2 + \ldots + f_k} =$
127–135	5	
136–144	9	$= \frac{3 \cdot [(118+126)/2] + 5 \cdot [(127+135)/2] + \ldots + 2 \cdot [(172+180)/2]}{3+5+9+12+5+4+2} =$
145–153	12	
154–162	5	$= \frac{5879}{40} = 146{,}98$
163–171	4	A diákok átlagosan 147 másodperc alatt oldották meg a feladatot.
172–180	2	

A többi átlagfajtát a következő, 10. táblázat szemlélteti:

10. táblázat. *Az egyéb átlagfajták*

Elnevezés	Jelölés	Számítás	
		súlyozatlan	súlyozott
Harmonikus átlag	$\bar{X}_h$	$\frac{N}{\sum_{i=1}^{N}\frac{1}{X_i}}$	$\frac{N}{\sum_{i=1}^{k}\frac{f_i}{X_i}}$
Mértani (geometriai) átlag	$\bar{X}_g$	$\sqrt[N]{\prod_{i=1}^{N}X_i}$	$\sqrt[N]{\prod_{i=1}^{k}X_i^{f_i}}$
Négyzetes (kvadratikus) átlag	$\bar{X}_q$	$\sqrt{\frac{\sum_{i=1}^{N}X_i^2}{N}}$	$\sqrt{\frac{\sum_{i=1}^{k}f_iX_i^2}{N}}$

Forrás: Hunyadi–Mundruczó–Vita 2000. 107.

A harmonikus és mértani átlag általában olyan esetekben használható, amikor nem az ismérvértékek összegének, hanem az azok reciprokából képzett összegnek vagy azok szorzatának van valamilyen értelme. Ilyenkor közelítő értéket kapunk. Négyzetes átlagot akkor számolunk, amikor ki akarjuk küszöbölni az átlagolni kívánt érték előjelét.

A medián

A medián ordinális skálán mért adatokból is meghatározható. A medián vagy középső érték az ismérvértékek nagyság szerint rendezett adatsorának közepén elhelyezkedő számérték, amelynél ugyanannyi nagyobb, mint kisebb értékű esetünk van.

Ha N páratlan, akkor a medián értéke közvetlenül a középső érték lesz, amelynek a sorszáma az összes érték növekvő sorba rendezése esetében $(N + 1) / 2$ lesz. Ha N páros, akkor nincs egy pontosan beazonosítható középső eset. Ilyenkor konvenció szerint a medián értéke a két középső érték számtani átlaga lesz.

Az észlelési adatoknak bármely tetszőleges számtól számított (abszolút) eltéréseinek összege akkor minimális, ha az eltéréseket a mediántól vesszük. Ha a változó értékei közt nincsenek kirívóan kicsik vagy nagyok és eloszlásbeli aránytalanságok, a medián és az átlag közötti különbség általában nem nagy. Legfőbb előnye, hogy nem igényel számítást, ezért gyorsan meghatározható. A medián mint felezőérték nagyszámú megfigyelés esetén az értékek eloszlásának megítélésében játszik szerepet, közvetlenül nem függ az összes rendelkezésre álló ér-

téktől, de a szélsőséges értékektől sem. Ezért tekintik a legfontosabb pozicionális centrális mutatónak.

18. példa ▼

▶ *A medián meghatározása*

Nézzük az előző szemléltető példánkat. Adott az alábbi, monoton nem csökkenő módon rendezett értéksorunk:

0 ; 0 ; 0,5 ; 0,6 ; 0,8 ; *1* ; 1 ; 1 ; 3 ; 5 ; 10.

Nagyon fontos arra figyelnünk, hogy az adataink *monoton nem csökkenő módon legyenek rendezve* (ha nem ilyen formában szerepelnek, rendezzük sorba), hiszen pozicionális mutatót vizsgálunk. Ebben az esetben értéksorunk páratlan számú tagból áll, tehát a medián pontosan a középső érték, azaz a $(11 + 1)/2$-ik esetnek megfelelő érték, vagyis 1. Értelmezése, hogy a 11 esetünk fele 1 vagy ennél nagyobb értékű, fele 1 vagy 1-nél kisebb értéket vesz fel.

Abban az esetben, ha folytonos jellegű adatokból egyenlő hosszúságú intervallumokat hozunk létre, akkor számíthatunk mediánt, ha az eseteket úgy tekintjük, mintha az adott intervallumon belül egyenletesen oszlanak meg. Ilyenkor a mediánt az alábbi tapasztalati képlettel számítjuk ki:

$$M_e = L_1 + \left(\frac{\frac{N+1}{2} - \left(\sum f_1 \right)}{f_{M_e}} \right) \cdot c$$

ahol:

L_1 – a mediánt tartalmazó osztály valódi alsó határa,

$\sum f_1$ – a mediánt tartalmazó osztály előtt lévő osztályokhoz tartozó gyakoriságok összege (kumulált gyakoriság),

f_{Me} – a mediánt tartalmazó osztály gyakorisága,

c – osztályköz vagy osztályhosszúság.

Az eljárás a következő lépéseket tartalmazza: kiszámítjuk a kumulált gyakorisági értékeket, kijelöljük a középső esetet tartalmazó osztályt, meghatározzuk a mediánt tartalmazó osztály valódi alsó határát, kiszámítjuk az osztályhosszúságot, majd kiszámítjuk a mediánt.

19. példa ▼

▶ *A medián számítása egyenlő hosszúságú intervallumokból*

Nézzük a 40 diák feladatmegoldási idejét tartalmazó előző fiktív példánkat, átmásolva a 9. táblázatot, kiegészítve a kumulált gyakoriságokkal (11. táblázat).

11. táblázat. *Gyakorisági sor (19. példa)*

Idő (s)	Gy (f_i)	$\sum f_i$	Számítás
118–126	3	3	1. Kiszámoljuk a kumulált gyakoriságokat egy új oszlopba.
127–135	5	8	2. (N + 1)/2 = 20,5, tehát a medián a huszadik és huszonegyedik esetet tartalmazó osztályban van (az értéke 145 és 153 között kell legyen).
136–144	9	17	
145–153	*12*	*29*	3. A mediánt tartalmazó osztály valódi alsó határa (L1) 144,5 (mivel folytonos változónk van, az értékek tizedesek is lehetnek).
154–162	5	34	
163–171	4	38	4. Az osztályhosszúság (c) a valódi felső és alsó határok különbsége, azaz 9 másodperc (153,5 – 144,5).
172–180	2	40	

Behelyettesítve a képletbe, megkapjuk a medián értékét:

$$M_e = L_1 + \left(\frac{\frac{N+1}{2} - \left(\sum f_1\right)}{f_{M_e}} \right) \cdot c = 144{,}5 + \left(\frac{\frac{40+1}{2} - (3+5+9)}{12} \right) \cdot 9 = 147{,}13$$

Értelmezés. A 40 diák fele 147,1 másodpercnél kevesebb, fele pedig ennél több idő alatt oldotta meg a feladatot.

A módusz

A módusz a legnagyobb gyakoriságú (leggyakoribb, legvalószínűbb) érték az eloszlásban, csoportosított adatok esetében a legnagyobb gyakoriságú osztály osztályközepének értéke. A módusz megállapításához célszerű az adatokat gyakorisági sorba rendezni, így a módusz a sor legnagyobb gyakorisággal előforduló értéke. Vannak esetek, amikor többmóduszú gyakorisági sorokat észlelünk – ilyen esetekben akkor szokás használni, amikor értelmezhetőek az értékek. A módusz szabálytalanul növekvő adatsor esetében sem jellemzi a sokaságot. De mivel a ténylegesen leggyakrabban előforduló érték, sokszor a jelenség természetét jobban kifejezi, mint a többi középérték. További előnye, hogy nominális skálán mért alapadatokból is meghatározható.

20. példa ▼

▶ *A módusz meghatározása*

Az előző szemléltető példánk egy egyszerű értéksort tartalmaz.

0 ; 0 ; 0,5 ; 0,6 ; 0,8 ; *1 ; 1 ; 1* ; 3 ; 5 ; 10.

Ebből egyértelmű, hogy a módusz 1, hiszen ez a leggyakrabban előforduló érték.

Folytonos ismérven mért, intervallummérési szintű csoportosított adatokból az alábbi tapasztalati képlettel számítunk móduszt:

$$M_o = L_1 + \left(\frac{D_1}{D_1 + D_2} \right) \cdot c$$

ahol:

L_1 – a móduszt tartalmazó osztály valódi alsó határa,

D_1 – a móduszt tartalmazó és az előtte lévő osztály gyakoriságainak különbsége,

D_2 – a móduszt tartalmazó és az utána lévő osztály gyakoriságainak különbsége,

c – osztályköz vagy osztályhosszúság.

Az eljárás a következő lépéseket foglalja magába: kijelöljük a legtöbb esetet tartalmazó osztályt, meghatározzuk a móduszt tartalmazó osztály valódi alsó határát, kiszámítjuk a D_1 és a D_2 értékeit a gyakorisági sorból, kiszámítjuk az osztályhosszúságot, majd kiszámítjuk a mediánt.

21. példa ▼

▶ *A módusz kiszámítása egyenlő hosszúságú intervallumokból*

Nézzük újra a 40 diák feladatmegoldási idejét tartalmazó példánkat (12. táblázat).

12. táblázat. *Gyakorisági sor (21. példa)*

Idő (s)	Gy (f_i)	Számítás
118–126	3	1. A legtöbb eset a 12 diákot tömörítő 4. osztályban van, tehát a módusz értéke 145–153 között kell legyen.
127–135	5	
136–144	9	2. A móduszt tartalmazó osztály valódi alsó határa (L_1) 144,5 (mivel folytonos változónk van, az értékek tizedesek is lehetnek).
145–153	*12*	3. $D_1 = 12 - 9 = 3$
154–162	5	4. $D_2 = 12 - 5 = 7$
163–171	4	5. Az osztályhosszúság (c) a valódi felső és alsó határok különbsége, azaz 9 másodperc (153,5 – 144,5).
172–180	2	

$$M_o = L_1 + \left(\frac{D_1}{D_1 + D_2}\right) \cdot c = 144{,}5 + \left(\frac{(12-9)}{(12-9)+(12-5)}\right) \cdot 9 = 147{,}20$$

Értelmezés. A legtöbben a diákok közül 147 másodperc körül oldották meg a feladatot.

Választás a középértékek között

Gyakorlati szempontból a három legfontosabb középérték az átlag, a módusz és a medián. Annak eldöntése, hogy adott esetben melyiket használjuk, nem egyszerű kérdés. A középértékek közötti választást leggyakrabban motiváló szempontok a következők:

– az adott középérték mindig egyértelműen meghatározható-e,
– az összes rendelkezésre álló ismérvértéktől függ-e vagy nem,
– mennyire érzékeny a szélsőséges ismérvértékekre,
– mekkora és milyen módon értelmezhető hibával képes helyettesíteni az alapadatokat.

A döntéshez a 13. táblázat nyújt segítséget.

13. táblázat. *Választás a középértékek között*

Átlag	Módusz	Medián
Egyértelműen meghatározható	Nem mindig határozható meg egyértelműen	Mindig egyértelműen meghatározható
Függ az összes értéktől	Nem függ az összes értéktől	Nem függ az összes értéktől
Érzékeny a szélsőséges értékekre	Nem érzékeny a szélsőséges értékekre	Nem érzékeny a szélsőséges értékekre
Az előjeles hibák összességükben kiegyenlítik egymást, és minimálissá teszi a helyettesítéssel elkövetett hibák négyzetösszegét	Az ismérvértékek helyébe téve ritkán és csak kis hibát követünk el	A hibaösszeget minimalizálva helyettesíti az ismérvértékeket

22. példa ▼

▶ *Választás a középértékek között*

A diákok feladatmegoldó képességéről szóló példánkban a három középérték:

$$\overline{X} = 146{,}98 \quad M_e = 147{,}13 \quad M_o = 147{,}20$$

Látható, hogy mindhárom középérték egymáshoz nagyon közeli érték, így ebben az esetben mindhárom mutató jól jellemzi a 40 diákot. Jelentősebb különbségek esetén az elemzés céljának a függvényében kell eldöntenünk, hogy

melyik információ mond a legtöbbet a sokaságról. Ha pl. több diákcsoport teljesítményét szeretnénk összehasonlítani, akkor átlaggal jellemezzük a sokaságot, ha azt szeretnénk eldönteni, hogy mennyi idő alatt lehet egy ilyen típusú feladatot megoldani, akkor móduszt használunk stb.

△ *Gyakorlófeladatok a középértékek kézi számítására*

1. Adott az alábbi táblázat: Arad megye és Arad város lakossága 15 éves korcsoportok szerinti bontásban 2021-ben (fő). *Forrás: Nemzeti Statisztikai Hivatal, Népszámlálási adatok (https://www.recensamantromania.ro/).*

	Arad megye	**Arad város**
0–14 éves	64 841	20 904
15–29 éves	63 171	19 748
30–44 éves	84 359	31 651
45–59 éves	91 280	32 575
60–74 éves	77 041	29 410
75 év feletti	29 451	10 790

2. Adott az alábbi táblázat: Brassó megye és Brassó város lakossága 15 éves korcsoportok szerinti bontásban 2021-ben (fő). *Forrás: Nemzeti Statisztikai Hivatal, Népszámlálási adatok (https://www.recensamantromania.ro/).*

	Brassó megye	**Brassó város**
0–14 éves	95 045	33 319
15–29 éves	79 347	30 093
30–44 éves	117 034	53 468
45–59 éves	107 929	47 501
60–74 éves	108 100	53 662
75 év feletti	39 160	19 546

Számítsuk ki a számtani átlagot, a mediánt és a móduszt a két gyakorlófeladatban szereplő 4 területre (Arad megye, Arad város, Brassó megye és Brassó város)! A 75 év felettieket 82 éves osztályközéppel becsüljük! Értelmezzük a kapott adatokat!

A középértékek kiszámítása SPSS-sel

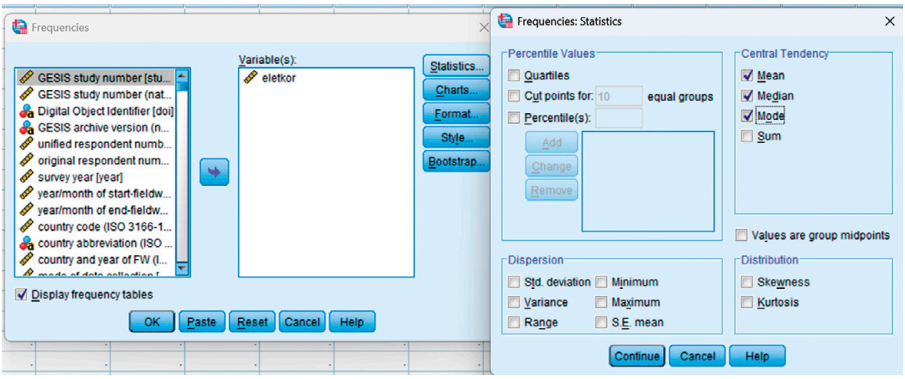
A centrális tendenciák kiszámítása nagyon egyszerű az SPSS-sel. Ahogyan már korábban is említésre került, legfőképpen arra kell figyelnünk, hogy a középértékekkel jellemezni kívánt változónk mérési szintje megengedi-e a számítást.

Akárcsak a gyakorisági tábla lekérése, a középértékek kiszámítása is az *ANALYZE* főmenü *Descriptive Statistics/Frequencies...* menüvel történik. Miután átvittük az elemezni kívánt változónkat/változóinkat, az ablak alsó részén található *Statistics* mezőre kattintunk, és bejelöljük a kért statisztikákat. A középértékek a *Central Tendency* ablakrészben találhatóak, ahol az átlagot a *Mean*, a mediánt *a Median*, a móduszt pedig a *Mode* mellett szereplő mezőkre klikkelve lehet lekérni.

23. példa ▼

▶ *Középértékek lekérése az SPSS-ben*

Adatbázisunkban a *Compute* almenü ismertetésekor létrehoztuk az *eletkor* változót. Tehát arányskálánk van, minden középérték kiszámítható és értelmezhető. Először azonban, a már ismert módon, kérjünk a változóra egy gyakoriságot, hogy ellenőrizzük le adatainkat (kell-e tisztítani, vannak-e nem releváns adataink). A gyakorisági tábla azt mutatja, hogy 1106 releváns válaszadónk van, és egyetlen értéktől sem kell megválnunk. Az adattisztítás minden egyes elemzés esetén elengedhetetlen, hiszen néhány rosszul bevitt vagy az elemzés szempontjából értelmetlen adat nagyon eltorzíthatja következtetéseinket. Például ha a mi esetünkben szerepelt volna egy 1010-es érték, és nem válunk meg tőle az elemzés előtt, teljesen hibás átlagéletkort számolunk a megkérdezettekre. A gyakorisági tábla szemrevételezése után az előzőek szerint lekérjük a középértékeket, majd *Continue*-t, és visszatérve az előző ablakba *Ok*-t kattintunk (25. ábra).



25. ábra. *A centrális tendenciák mutatószámainak lekérése (23. példa)*

Az *Output* ablakban rögtön megjelennek a kért statisztikák (26. ábra), amelyből kiolvasható, hogy 1106 érvényes válaszadónk van, ezek átlagéletkora 49,89 év, a válaszadók fele 50 évnél idősebb, fele pedig ennél fiatalabb, és a legtöbb válaszadó 67 éves.

Statistics

eletkor

N	Valid	1106
	Missing	0
Mean		49.8870
Median		50.0000
Mode		67.00

26. ábra. *Az Output-ban megjelenő középértékek (23. példa)*

△ *Gyakorlófeladatok a középértékek SPSS-ben való lekérésére*

1. Munkaadatbázisunkban a K61-es kérdésre vonatkozó *v240* változó [*number of people in household (Q78)*, a 289. sorszámú változó az adatbázisban] a háztartásban élők számát mutatja. Kérjük le rá a középértékeket és értelmezzük!
2. A K62-es kérdésre vonatkozó *v242* változó [*age completed education respondent (Q80)*, a 291. sorszámú változó az adatbázisban] a kérdezett azon életkorát mutatja, amikor befejezte tanulmányait. Kérjük le rá a középértékeket és értelmezzük!

2.4. Szórás és szóródás

Egy statisztikai sokaság elemei valamely tulajdonság értéknagysága tekintetében eltérnek egymástól, változatosak. Míg a középérték alkalmas arra, hogy e változatosság ellenére az adott tulajdonság értéknagyságát tömören, az egész sokaságra nézve kifejezze (a középérték a sokaság közös jellemzője), addig a szóródás a sokaság elemei valamely középértékhez vagy egymáshoz való viszonyulásának tömör jellemzője.

A szóródás egyes változók esetén nagyobb, a másiknál kisebb is lehet annak ellenére, hogy az átlaguk megegyezik. Ugyanakkor a szóródás nagyságának a kifejezésére a középérték megfelelő bázist nyújt, mivel az egyes értékek nemcsak egymástól, hanem a középértéktől is különböznek.

Az ismérvértékek egymás közötti különbségeiből számított szóródási mutatókat és a valamely kitüntetett értéktől számított eltéréseken alapuló mutatókat *abszolút szóródási mutatóknak* nevezik. Az abszolút szóródási mutatók mértékegysége mindig az ismérvértékek mértékegysége.

A szóródás *relatív mutatószámai* elvonatkoztatnak az ismérvérték eredeti mértékegységétől, és elsősorban összehasonlítási célokat szolgálnak.

A szóródás kifejezésére használatos mutatószámok:
– a szórás terjedelme,
– a kvartilis eltérés,
– átlagos különbség,
– a középeltérés,
– az abszolút átlageltérés,
– a négyzetes átlageltérés (szórás) és a variancia,
– szóródási együttható.

A szórás terjedelme (Range)

A szórás terjedelme annak a legkisebb intervallumnak a teljes hossza, amelyet az ismérvértékek kitöltenek.

$$i_S = X_{\max} - X_{\min}$$

Tehát a szóródás terjedelme az észlelési adatok közül a legnagyobb és a legkisebb értéknagyságú adat különbsége.

Mivel a két legszélsőségesebb ismérvértéktől függ, csak kevéssé jellemzi a vizsgált jelenség valódi természetét. Alkalmazása inkább homogén részsokaságoknál fejezi ki a szakmai szempontból elfogadható terjedelmet, osztályközös gyakorisági sorokból csak a két szélső kategória felezőpontjainak különbségéből becsülhető.

Egyértelmű hátránya tehát az, hogy az értékskála közbeeső értékeiről semmit sem tudunk meg, viszont nagyon egyszerűen előállítható és könnyen érthető adat. Például ha egy háztartási adatbázisban a legkisebb bevétel 50 lej, a legnagyobb pedig 32 000 lej, akkor a terjedelem 31 950 lej.

A kvartilis eltérés vagy interkvartilis félterjedelem

A kvartilis eltérés számítására akkor van szükség, ha a sokaság adatainak szélső értékei nagymértékben eltérnek a többi adattól. Használata olyan gyakorisági soroknál a legindokoltabb, ahol nyitott osztályközökkel indul és zárul a statisztikai sor (a szórás terjedelme nem becsülhető kiegészítő információk nélkül).

A nagyság szerint rendezett értéksort negyedelő értékek a kvartilisek. Három kvartilist szoktak megkülönböztetni:

– *alsó kvartilis (Q_1):* az az érték, amely alatt a sokaság egynegyede által felvett értékek találhatóak, az $n_{Q1} = \frac{n+1}{4}$-edik esetnek megfelelő érték,

– *középső kvartilis (Q_2):* az az érték, amely alatt a sokaság fele által felvett értékek találhatóak, az $n_{Q2} = \frac{n+1}{2}$-edik esetnek megfelelő érték, vagyis a *medián*,

– *felső kvartilis (Q_3):* az az érték, amely alatt a sokaság háromnegyede által felvett értékek találhatóak, az $n_{Q3} = \frac{3(n+1)}{4}$-edik esetnek megfelelő érték.

Akárcsak a medián esetében, intervallummérési szintű gyakorisági soroknál a kvartilisek értéknagyságát becsléssel lehet meghatározni:

$$Q_i = Q_{L_1} + \frac{n_{Qi} - \sum f_1}{f_{Q_i}} \cdot c$$

ahol:

Q_{L_1} – a kvartilis adat sorszámának megfelelő osztály alsó határa,

n_{Q_i} – az i-edik kvartilis sorszáma,

$\sum f_1$ – a kvartilis osztályig terjedő kumulált gyakoriságok összege,

f_{Q_i} – a kvartilist tartalmazó osztály gyakorisága,

c – osztályköz vagy osztályhosszúság.

Az *interkvartilis terjedelem* mérőszáma – a szélső értékektől függetlenül – azt a távolságot adja meg, amelyen belül az észlelési adatok 50%-a megtalálható.

$$i_q = Q_3 - Q_1$$

A *kvartilis eltérés* vagy *interkvartilis félterjedelem* a harmadik és az első negyedelő értékek különbségének a fele.

$$Q_e = \frac{Q_3 - Q_1}{2}$$

24. példa ▼

▶ *Interkvartilis terjedelem kiszámítása csoportosított adatokból*

Adott az alábbi fiktív adatsor (14. táblázat), amelyen az interkvartilis terjedelem kiszámítását mutatjuk be. A lépések hasonlóak a mediánnál leírtakkal.

14. táblázat. *Gyakorisági sor (24. példa)*

Család évi jövedelme (ezer lej)	Családok száma	Kumulált gyakoriság	Számítás
2–3,9	5	5	1. Kiszámítjuk a két kvartilis sorszámát: $n_{Q1} = \frac{n+1}{4} = \frac{95+1}{4} = 24$ $n_{Q3} = \frac{3(n+1)}{4} = \frac{3(95+1)}{4} = 72$ 2. Kiszámoljuk a kumulált gyakoriságokat egy új oszlopba. 3. Beazonosítjuk a kvartiliseket: az alsó kvartilis a 3., a felső pedig a 6. osztályban van. 4. Kiszámoljuk az osztályhosszúságot: 7,95–5,95 = 2.
4–5,9	13	18	
6–7,9	*18*	*36*	
8–9,9	17	53	
10–11,9	14	67	
12–13,9	*13*	*80*	
14–15,9	7	87	
16–17,9	4	91	
18–19,9	4	95	

$$Q_1 = Q_{L_1} + \frac{n_{Q_1} - \sum f_i}{f_{Q_1}} \cdot c = 5{,}95 + \frac{24 - (5+13)}{18} \cdot 2 = 6{,}62$$

$$Q_3 = Q_{L_1} + \frac{n_{Q_3} - \sum f_1}{f_{Q_3}} \cdot c = 11{,}95 + \frac{72 - (5+13+18+17+14)}{13} \cdot 2 = 12{,}72$$

$$i_q = Q_3 - Q_1 = 12{,}72 - 6{,}62 = 6{,}10$$

Az alsó kvartilis értéke 6,62, tehát a vizsgált családok egynegyedének 6620 lej alatt van az évi jövedelme, háromnegyedének pedig e felett. A felső kvartilis értéke 12,72, tehát a családok háromnegyede 12 720 lejnél kisebb, egynegyede pedig ennél nagyobb évi jövedelemmel rendelkezik. Az interkvartilis terjedelem értéke 6,10, így a családok fele 6620–12 720 lej közötti bevételre tesz szert évente.

Átlagos (abszolút) különbség

Ez a szóródási mutató minden lehetséges módon párba állított ismérvértékek különbségeinek abszolút értékéből számított átlag.

$$G = \frac{1}{N(N-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} f_i f_j \left| X_i - X_j \right|$$

A Gini-féle mutató azt mutatja, hogy az *X* ismérv értékei átlagosan mennyire különböznek egymástól. Ha minden ismérvérték egyforma, azaz nincs szóródás, akkor G = 0.

Az átlagos különbség számszerű meghatározása elég kényelmetlen, ezért a gyakorlatban ritkán használják. Jelentőségét a koncentrációhoz való szoros kapcsolódása adja.

A középeltérés

A középeltérés a sokaságelemek mediántól számított eltéréseinek az átlaga.

$$K_e = \frac{\sum_{i=1}^{n} \left| X_i - M_e \right|}{n}$$

Alkalmazása főként arra az esetre koncentrálódik, amikor a sokaság jellemzésére a medián a legalkalmasabb jellemző. Gyakorisági sorok esetében nem használható.

Az abszolút átlageltérés vagy átlagos eltérés

A számtani átlag körüli elhelyezkedés egyik mutatója. Mivel az értékek számtani átlagtól vett különbségeinek összege 0, ezért a különbségek abszolút értékeivel számolunk.

Az abszolút átlageltérés az ismérvértékek számított átlagtól való eltéréseinek számtani átlaga.

$$A_e = \frac{\sum_{i=1}^{n} \left| X_i - \overline{X} \right|}{n}$$

A gyakorlatban ritkán használják. Gyakorisági sorok esetén az X_i helyett az osztályközép kerül.

Szórás (négyzetes átlageltérés) és variancia

A szórás a szóródás legfontosabb mérőszáma. Nagyon hasonlít az abszolút átlageltéréshez, csak az abszolút eltérés helyett négyzetre emeléssel iktatja ki a különbségek előjelét. A négyzetre emelés az eltérések abszolút értelemben vett nagyságát is jobban kiemeli. Az utólagos gyökvonás a négyzetre emelés tompítását és az alapadatok eredeti mértékegységéhez való visszatérést is szolgálja.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{k} f_i \cdot (X_i - \overline{X})^2}$$

A szórás az átlagtól vett eltérések négyzetes átlaga.

A szórás azt mutatja, hogy az X_i ismérvértékek átlagosan mennyivel térnek el a számtani átlagtól. Számításmódjából adódóan a szórás olyan átlagos hibaként is felfogható, amit abban az esetben követünk el, ha minden alapadatot a számtani átlaggal helyettesítünk.

Sok esetben nem a szórás, hanem annak négyzete, a variancia vagy szórásnégyzet bír jelentőséggel.

25. példa ▼

▶ *A szórás kiszámítása és értelmezése*

Nézzük az előző példánkat, és számoljuk ki a szórást (15. táblázat).

15. táblázat. *Gyakorisági sor (25. példa)*

Család évi jövedelme (ezer lej)	Családok száma	Osztály-közép	Számítás
2–3,9	5	2,95	1. Első lépésként kiszámoljuk az osztályközepeket egy új oszlopba. 2. Kiszámítjuk az átlagot. 3. Kiszámítjuk a szórást.
4–5,9	13	4,95	
6–7,9	*18*	*6,95*	
8–9,9	17	8,95	
10–11,9	14	10,95	
12–13,9	*13*	*12,95*	
14–15,9	7	14,95	
16–17,9	4	16,95	
18–19,9	4	18,95	

$$\bar{X} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + \ldots + f_k \cdot X_k}{f_1 + f_2 + \ldots + f_k} = \frac{5 \cdot 2{,}95 + 13 \cdot 4{,}95 + \ldots\ldots 4 \cdot 18{,}95}{5 + 13 + 18 + \ldots + 4} = \frac{926{,}25}{95} = 9{,}75$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i \cdot (X_i - \bar{X})^2} = \sqrt{\frac{1}{95} \cdot \left[5 \cdot (2{,}95 - 9{,}75)^2 + 13 \cdot (4{,}95 - 9{,}75)^2 + \ldots + 4 \cdot (18{,}95 - 9{,}75)^2\right]} =$$

$$= \sqrt{\frac{1571{,}2}{95}} = \sqrt{16{,}539} = 4{,}07$$

Értelmezés. A szórás 4,70 lej, tehát a vizsgált családok évi jövedelme átlagosan 4 ezer lejjel tér el a 9,75 ezer lejes átlagjövedelemtől.

Szóródási együttható vagy relatív szórás

A szóródási együttható (variációs koefficiens) a különböző átlagú és eltérő tulajdonságú sokaságok szórásának összehasonlítását teszi lehetővé. Elsősorban különböző ismérvek összehasonlítására használják, és igazából csak az arányskálán mért ismérveknél van jelentősége.

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

A szóródási együttható az ismérvértékeknek az átlagtól vett átlagos relatív (százalékos) eltérését mutatja. A közgazdasági vizsgálatoknál általában a következő tapasztalati határokat tekintik mértékadónak:

– 0–10% állandóságot mutat,

– 10–20% közepes változékonyságot mutat,

– 20–30% erős változékonyságot mutat,
– 30%-on felüli együttható szélsőséges ingadozást fejez ki.

A közölt határok általános érvényűek és tájékoztató jellegűek. A vizsgálat céljának, a jelenség természetének és a számításban részt vevő elemek számának figyelembevételével lehet a szóródás nagyságát szakmai szempontból megítélni.

△ *Gyakorlófeladatok a szóródási mutatók kézi számítására*

1. Adott az alábbi táblázat: Arad megye és Arad város lakossága 15 éves korcsoportok szerinti bontásban 2021-ben (fő). *Forrás: Nemzeti Statisztikai Hivatal, Népszámlálási adatok (https://www.recensamantromania.ro/).*

	Arad megye	**Arad város**
0–14 éves	64 841	20 904
15–29 éves	63 171	19 748
30–44 éves	84 359	31 651
45–59 éves	91 280	32 575
60–74 éves	77 041	29 410
75 év feletti	29 451	10 790

2. Adott az alábbi táblázat: Brassó megye és Brassó város lakossága 15 éves korcsoportok szerinti bontásban 2021-ben (fő). *Forrás: Nemzeti Statisztikai Hivatal, Népszámlálási adatok (https://www.recensamantromania.ro/).*

	Brassó megye	**Brassó város**
0–14 éves	95 045	33 319
15–29 éves	79 347	30 093
30–44 éves	117 034	53 468
45–59 éves	107 929	47 501
60–74 éves	108 100	53 662
75 év feletti	39 160	19 546

Számítsuk ki az alsó és felső kvartilist, illetve a szórást a két gyakorlófeladatban szereplő 4 területre (Arad megye, Arad város, Brassó megye és Brassó város)! A 75 év felettieket 82 éves osztályközéppel becsüljük! Értelmezzük a kapott adatokat!

A kvartilisek és a szóródási mutatók kiszámítása az SPSS-sel

Miként már korábban is említésre került, kvartilisek és szóródás csak mennyiségi adatokból számítható. Akárcsak a többi egyváltozós statisztika lekérése, a kvartilisek és szóródási mutatók is az *ANALYZE* főmenü *Descriptive Statistics*, *Frequencies* parancsával számíttathatóak ki. Miután átvittük az elemezni kívánt változónkat/változóinkat, az ablak alsó részén található *Statistics* mezőre kattintunk, és bejelöljük a kért statisztikákat. A szóródási mutatók a *Dispersion* ablakrészben találhatóak, ahol a terjedelmet a *Range*, a szórást a *Std. Deviation*, a varianciát a *Variance* mellett szereplő mezőkre klikkelve lehet lekérni. A pozicionális mutatók a bal felső részben, a *Percentile Values* ablakrészben találhatóak, ahol a *Quartiles* mellett szereplő mezőkre klikkelve lehet őket lekérni.

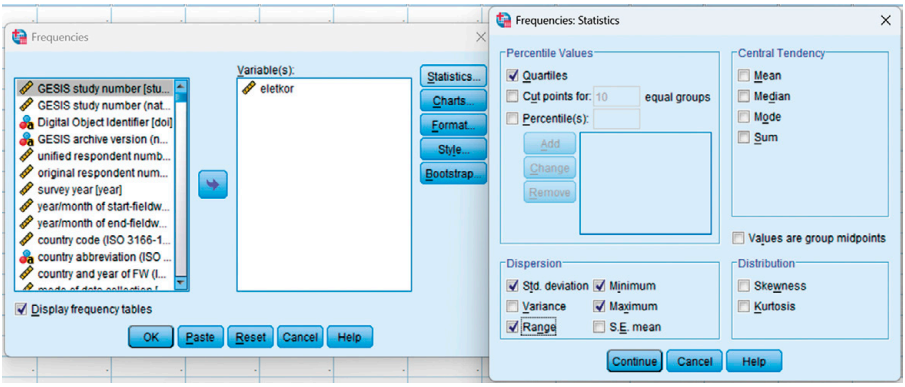
Az adatok eloszlása, a medián, a kvartilisek, az interkvartilis terjedelem és a kiugró értékek együttes, tömör szemléltetésére a *Boxplot* (dobozdiagram) a legalkalmasabb grafikus eszköz, amely vizuálisan is megerősíti a korábban ismertetett fogalmakat. A boxplot külön előnye, hogy a kiugró értékeket is jelöli – ezek megfelelő kezelése később, a többváltozós elemzések során szintén kiemelt szerepet kap, hiszen jelentősen befolyásolhatják az eredményeket és a modellek stabilitását. A *Boxplot* tehát a változók eseteinek elhelyezkedését szemlélteti oly módon, hogy az esetek túlnyomó többsége a doboz által kijelölt intevallumba esik. Fő elemei: a medián (a doboz közepe, egy vízszintes vastag vonal), interkvartilis terjedelem (a doboz hossza, a Q_1 az alsó szél és Q_3 a felső szél, az adatok középső 50%-a), „bajuszok” (*Whiskers*), azaz a Q_1 alatti (alsó bajusz) és Q_3 feletti (felső bajusz) értékek egy meghatározott tartományig (általában $Q_1 - 1{,}5 \times i_q$-ig és $Q_3 + 1{,}5 \times i_q$-ig, függőleges vonal), kiugró értékek (*Outliers*), vagyis a bajuszokon kívül eső pontok, külön körökkel jelölve.

A *Boxplot* a *GRAPHS* főmenü *Legacy Dialogs, Boxplot, Simple, Summaries for groups of cases* úton kérhető le az SPSS-ben. Itt a *Define* lehetőségnél a *Variable* mezőbe kell betenni a vizsgálni kívánt változót, illetve a *Category Axis*-ra egy olyan változót, ami egyetlen értéket tartalmaz (pl. a TRANSFORM, *Recode into Different Variables*-szel a vizsgálni kívánt változó minden értékét 1-gyé kódoltatjuk).

26. példa ▼

▶ *Szóródási mutatók lekérése az SPSS-ben*

Adatbázisunkban újra vizsgáljuk meg az *eletkor* változót, ezúttal a szóródás szempontjából. Tehát arányskálánk van, a szóródási mutatók kiszámíthatóak és értelmezhetőek. Az előzőekben leírtak szerint lekérjük a mutatókat, majd *Continue*-t, és visszatérve az előző ablakba *Ok*-t kattintunk (27. ábra).



27. ábra. *A szóródási mutatók lekérése (26. példa)*

Az *Output* ablakban megjelenő statisztikákat (28. ábra) értelmezzük.

Statistics

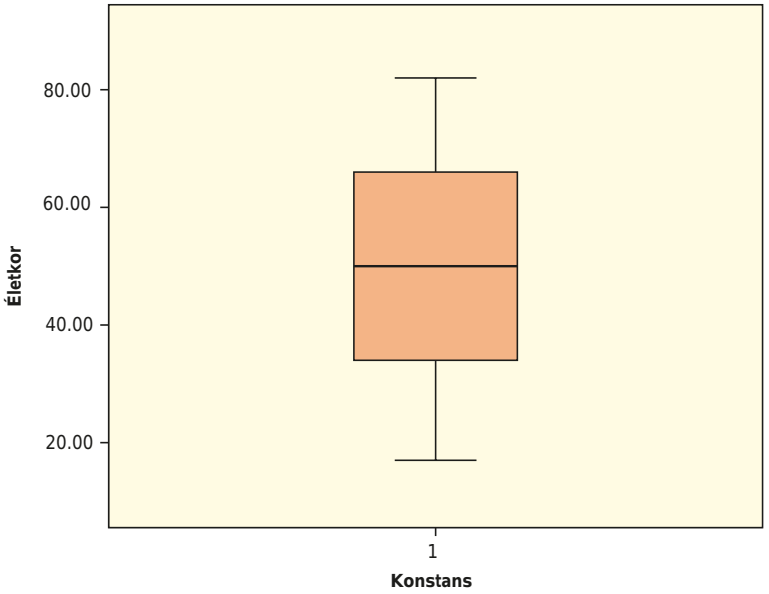
eletkor

N	Valid	1106
	Missing	0
Std. Deviation		18.43739
Range		65.00
Minimum		17.00
Maximum		82.00
Percentiles	25	34.0000
	50	50.0000
	75	66.0000

28. ábra. *Az SPSS által számolt szóródási mutatók (26. példa)*

A kérdésre 1106 személy adott érvényes választ, a legfiatalabb megkérdezett 17 éves, a legidősebb 82 éves. A terjedelem tehát 65 év, vagyis a legidősebb és a legfiatalabb válaszadó életkora között 65 év korkülönbség van. A szórás 18 év, tehát a kérdezettek életkora átlagosan 18 évvel tér el a válaszadók átlagéletkorától. A megkérdezettek egynegyede 34 évnél, fele 50 évnél és háromnegyede 66 évnél fiatalabb, egynegyede 66 évnél, fele 50 évnél és háromnegyede 34 évnél idősebb.

Végül kérjünk az *eletkor* változóra egy *Boxplot*-ot az előzőekben leírtak szerint (29. ábra), miután előzőleg létrehoztuk a *Konstans* nevű, csak 1-es értékeket tartalmazó változót.



29. ábra. *A dobozdiagram (26. példa)*

A dobozdiagram alapján az életkor mediánja kb. 50 év, a középső 50% pedig nagyjából 35 és 65 év között helyezkedik el. A bajuszok azt mutatják, hogy a legalacsonyabb életkor kb. 18, a legmagasabb pedig kb. 85 év. Kiugró érték nem látható, az alsó bajusz hosszabb volta pedig enyhe bal oldali ferdeségre utal (lásd a 2.5-ös alfejezetet).

△ *Gyakorlófeladatok a szóródási mutatók SPSS-ben való lekérésére*

1. Munkaadatbázisunkban a K61-es kérdésre vonatkozó *v240*-es változó [*number of people in household (Q78)*, a 289. sorszámú változó az adatbázisban] a háztartásban élők számát mutatja. Kérjük le rá a szóródási mutatókat és a dobozdiagramot, majd értelmezzük!
2. A K62-es kérdésre vonatkozó *v242*-es változó [*age completed education respondent (Q80)*, a 291. sorszámú változó az adatbázisban] a kérdezett azon életkorát mutatja, amikor befejezte tanulmányait. Kérjük le rá a szóródási mutatókat és a dobozdiagramot, majd értelmezzük!

2.5. Momentumok, ferdeség és csúcsosság

A momentumok

A momentumok a különféle átlagok és a szórás általánosításának tekinthetőek, mivel az $X_i - \overline{X}$ eltérések helyett az $X_i - A$ eltérések hatványait átlagolják (A egy tetszőleges állandó).

Súlyozatlan esetben a momentumokat az alábbi képlettel számoljuk,

$$M_r(A) = \frac{\sum_{i=1}^{n} (X_i - A)^r}{n}$$

súlyozott esetben pedig az alábbi képlet használatos:

$$M_r(A) = \frac{\sum_{i=1}^{k} f_i (X_i - A)^r}{n}$$

A képlettel meghatározott mennyiségeket az X ismérv vagy a gyakorisági eloszlás *A körüli r-edik momentumainak* nevezzük.

Az A = 0 speciális esetben az általános képletek r-edik momentumokat adnak, amelyekre az egyszerű M_r jelölést használjuk. Az $A = \overline{X}$ választás esetén az r-edik *centrális momentumokhoz* jutunk.

A momentumok több eddig megismert mutatószámot foglalnak egységes elméleti keretbe. Gyakorlati jelentőségüket a gyakorisági eloszlások alakjának jellemzésekor való felhasználásuk adja. A 16. táblázat néhány nevezetes momentumot foglal össze.

16. táblázat. *Nevezetes momentumok*

r (Hatvány)	$A = 0$		$A = \overline{X}$	
	Jelölés	Elnevezés	Jelölés/érték	Elnevezés
1	$\overline{X}$	számtani átlag	0	–
2	$\overline{X}_q^2$	négyzetes átlag négyzete	σ^2	variancia

Forrás: Hunyadi–Mundruczó–Vita 2000. 121.

27. példa ▼

▶ *Momentumok kiszámítása*

Adott az alábbi 5 esetből álló értéksorunk:
2 ; 3 ; 7 ; 8 ; 10.

Határozzuk meg az első, a második és a harmadik momentumot (A=0)!

$$M_1(0)=\frac{\sum_{i=1}^{n}(X_i-0)^1}{n}=\frac{\sum_{i=1}^{n}X_i}{n}=\overline{X}=\frac{2+3+7+8+10}{5}=\frac{30}{5}=6$$

$$M_2(0)=\frac{\sum_{i=1}^{n}(X_i-0)^2}{n}=\frac{\sum_{i=1}^{n}X_i^{\,2}}{n}=\overline{X}^2=\frac{2^2+3^2+7^2+8^2+10^2}{5}=\frac{226}{5}=45{,}2$$

$$M_3(0)=\frac{\sum_{i=1}^{n}(X_i-0)^3}{n}=\frac{\sum_{i=1}^{n}X_i^{\,3}}{n}=\frac{2^3+3^3+7^3+8^3+10^3}{5}=\frac{1890}{5}=378$$

Határozzuk meg az átlag körüli első és második (centrális) momentumot ($A=\overline{X}$)!

$$M_1(\overline{X})=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^1}{n}=\frac{(2-6)+(3-6)+(7-6)+(8-6)+(10-6)}{5}=0$$

$$M_2(\overline{X})=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}=\sigma^2=\frac{(2-6)^2+(3-6)^2+(7-6)^2+(8-6)^2+(10-6)^2}{5}=9{,}2$$

A koncentráció elemzése

A koncentráció a sokasághoz tartozó értékösszeg jelentős részének vagy egészének kevés egységre történő összpontosulása. A koncentráció foka vagy a sokaság nagyságának megadásával, vagy a relatív gyakoriságok és relatív értékösszegek egybevetésével jellemezhető. Ha a vizsgált sokaság mérete kicsi, *abszolút koncentrációról*, ha a sokaság nagy, *relatív koncentrációról* beszélünk.

Amikor a teljes értékösszeg egyetlen egységre jut, értelemszerűen a lehető legnagyobb koncentrációról van szó, amennyiben a teljes értékösszeg a sokaság egységei között egyenletesen oszlik meg, a koncentráció hiányáról van szó.

A koncentráció mértékét különböző mutatószámokkal jellemezzük, amelyek az alábbi csoportokba sorolhatók:

a) az abszolút koncentráció mutatószámai;

b) a relatív koncentráció mutatószámai.

Az abszolút koncentráció mutatószámai

Az abszolút koncentráció a vizsgált sokaság nagyságára és szerkezetére vonatkozó jellemzőket foglalja magába:

1. az egységek száma (n),
2. valamilyen értelemben vett átlagos nagysága ($\overline{X}$).

A relatív koncentráció mutatószámai

A relatív koncentráció az eloszlás egyenlőtlenségének mértékét fejezi ki. Ennek szemléltetésére szolgál a Lorenz-görbe, amely az egységek kumulált arányát és a hozzájuk tartozó értékek kumulált arányát ábrázolja, egyenes szakaszokkal összekötött vonaldiagram formájában. A koncentrációs terület (tc) a Lorenz-görbe és az egyenletes eloszlást jelképező átló által közrezárt terület. A koncentráció mértéke a koncentrációs terület és az átló alatti háromszög területének hányadosaként értelmezhető. E hányados a koncentrációs együttható (L).

$$L = \frac{t_c}{1/2} = 2t_c$$

A koncentrációs együttható meghatározható az átlagos abszolút különbség Gini-féle mutatójából is:

$$L = \frac{G}{2\overline{X}}$$

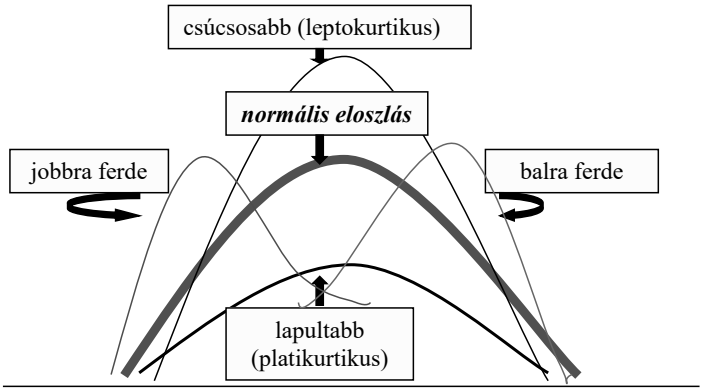
Az L koncentrációs együttható értéktartománya $0 \leq L \leq 1$, ahol $L = 0$ a teljes egyenlőséget, míg $L = 1$ a maximális koncentrációt jelenti: minél nagyobb az L értéke, annál erőteljesebb a vizsgált jelenség koncentráltsága, azaz az eloszlás egyenlőtlensége. Pl. a jövedelmek eloszlását vizsgálva a magas L koncentrációs együttható arra utal, hogy a vizsgált térségben az összjövedelem jelentős része viszonylag kevés háztartás kezében összpontosul.

Alakmutatók

A gyakorisági eloszlások alakmutatószámai azt jellemzik tömören, hogy milyen tekintetben és milyen mértékben térnek el a normális eloszlás gyakorisági görbéjétől (a Gauss-görbétől). Mivel a normális eloszlás egymóduszú, csak egymóduszú gyakorisági görbék körében van értelme.

A gyakorisági eloszlás grafikus ábrája kétféle tekintetben térhet el a normális eloszlás görbéjétől (30. ábra):

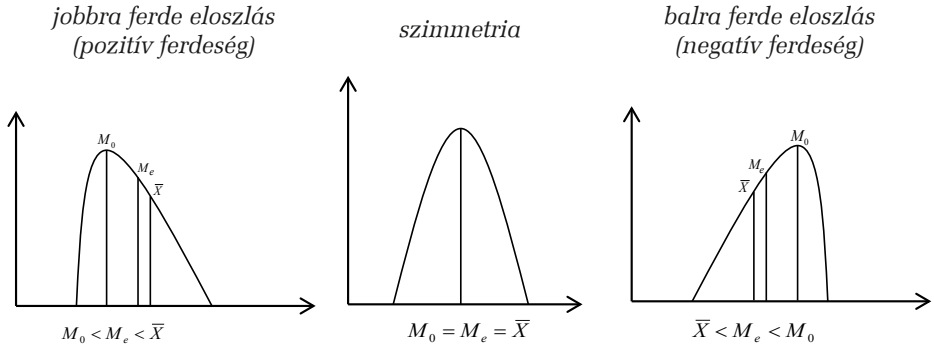
1. valamilyen irányban hosszabban elnyúlhat,
   ekkor *aszimmetria vagy ferdeség* áll fenn,
2. az ábra csúcsa alacsonyabban vagy magasabban lehet,
   ilyenkor *csúcsosságról* vagy *lapultságról* beszélünk.



30. ábra. *A gyakorisági eloszlások Gauss-görbétől való eltérései*

Aszimmetria: ferdeségi mutatók

Az egymóduszú gyakorisági eloszlások szimmetrikus vagy aszimmetrikus volta többféleképpen is megragadható az eddig megismert mutatószámok segítségével.



Az aszimmetria mutatószámai

A Pearson-féle mutatószám (rendszerint a -1 és 1 határok között mozog) arra a tapasztalati megállapításra alapoz, amely szerint mérsékelten aszimmetrikus eloszlás esetében a medián az átlagtól az átlag és a módusz közötti különbség mintegy egyharmadával balra vagy jobbra esik:

$$\overline{X} - \mathrm{Mo} \cong 3 \cdot (\overline{X} - \mathrm{Me})$$

A *Pearson-féle mutatószám* az alábbi képlettel számítható ki:

$$P = \frac{3 \cdot (\overline{X} - M_e)}{\sigma}$$

A két szélső kvartilis és a medián közötti eltéréseken alapul az *A aszimmetria mérőszám*. Alapja, hogy szimmetria esetén $Q_3 - M_e = M_e - Q_1$. Olyankor használjuk, ha a szóródást is a kvartilisek felhasználásával jellemeztük.

$$A = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)}$$

Az α_3 mutatószám a harmadik centrális momentum viselkedésén alapszik.

$$\alpha_3 = \frac{M_3(\overline{X})}{\sigma^3}$$

Szimmetria esetén $\alpha_3 = 0$, jobb oldali ferdeség esetén $\alpha_3 > 0$, bal oldali ferdeség esetén pedig $\alpha_3 < 0$. Az aszimmetria mértékének megítélését nem könnyíti meg egy alsó és felső határ, ugyanakkor elég érzékenyen reagál az eloszlás alakjának kismértékű változására is.

Az aszimmetria mindhárom mutatója szimmetrikus gyakorisági sorok esetén 0 vagy 0 körüli értéket vesz fel (sokszor becsüljük). *A jobb oldali ferdeséget a mutatók pozitív értékei, a bal oldali ferdeséget a mutatók negatív értékei jelzik.*

Csúcsosság: csúcsossági mutatók

A csúcsosság mértékének megállapítására a két legismertebb mutató a K és az α_4.

A *K mérőszám* alapja: minél csúcsosabb egy eloszlás, annál kisebb a felső és alsó kvartilis különbségének a fele a két szélső decilis különbségéhez viszonyítva.

$$K = \frac{Q_3 - Q_1}{2 \cdot (D_9 - D_1)}$$

Normális eloszlás esetében $K \approx 0{,}263$ (ehhez lehet viszonyítani a K értékét). Minél csúcsosabb az eloszlás, K értéke annál kisebb lesz.

Az α_4 mutatószám a negyedik centrális momentumhoz kötődik. Alapja: a 0 várható értékű és 1 szórású normális eloszlás negyedik centrális momentuma egyenlő 3-mal.

$$\alpha_4 = \frac{M_4(\overline{X})}{\sigma^4} - 3$$

A ferdeségi és csúcsossági mutatószámokat csak akkor ajánlatos használni, ha *a gyakorisági poligon a gyakorisági görbe elég jó közelítésének tekinthető*. A megfigyelt sokaság ehhez szükséges minimális nagysága *50–100* között van.

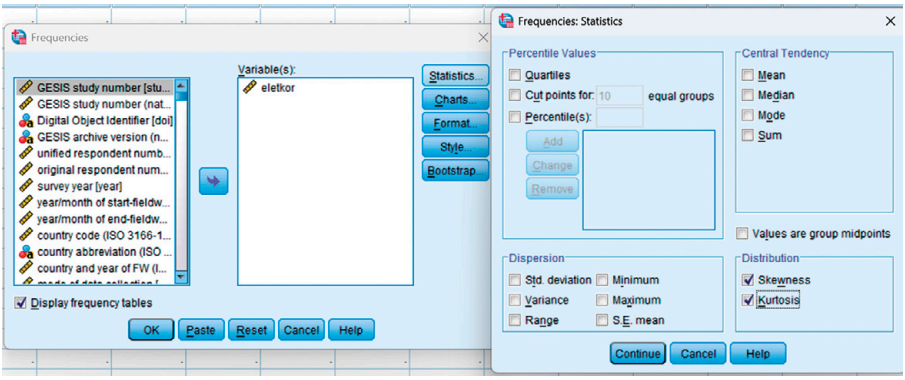
Alakmutatók és gyakorisági poligonok az SPSS-ben

Az alakmutatók is (akárcsak a többi egyváltozós statisztika) az *ANALYZE* főmenü *Descriptive Statistics*, *Frequencies* parancsával számíttathatóak ki. Miután átvittük az elemezni kívánt változónkat/változóinkat, az ablak alsó részén található *Statistics* mezőre kattintunk, és bejelöljük a kért statisztikákat. A szóródási mutatók a *Distribution* ablakrészben találhatóak, ahol a ferdeséget a *Skewness*, a csúcsosságot pedig a *Kurtosis* mellett szereplő mezőkre klikkelve lehet lekérni. A *Continue*-val visszatérve a *Frequencies* ablakba, a *Charts* opciónál le lehet kérni a gyakorisági poligonnak a normális eloszlás görbéjével együtt való ábrázolását (*Histograms with normal curve*).

28. példa ▼

▶ *Alakmutatók az SPSS-ben*

Adatbázisunkban újra vizsgáljuk meg az *eletkor* változót, ezúttal az alakmutatók szempontjából. Az előzőek szerint lekérjük a ferdeségi és csúcsossági mutatókat, majd a gyakorisági poligonra ábrát kérünk (31. ábra).



31. ábra. *Ferdeségi és csúcsossági mutatók lekérése (28. példa)*

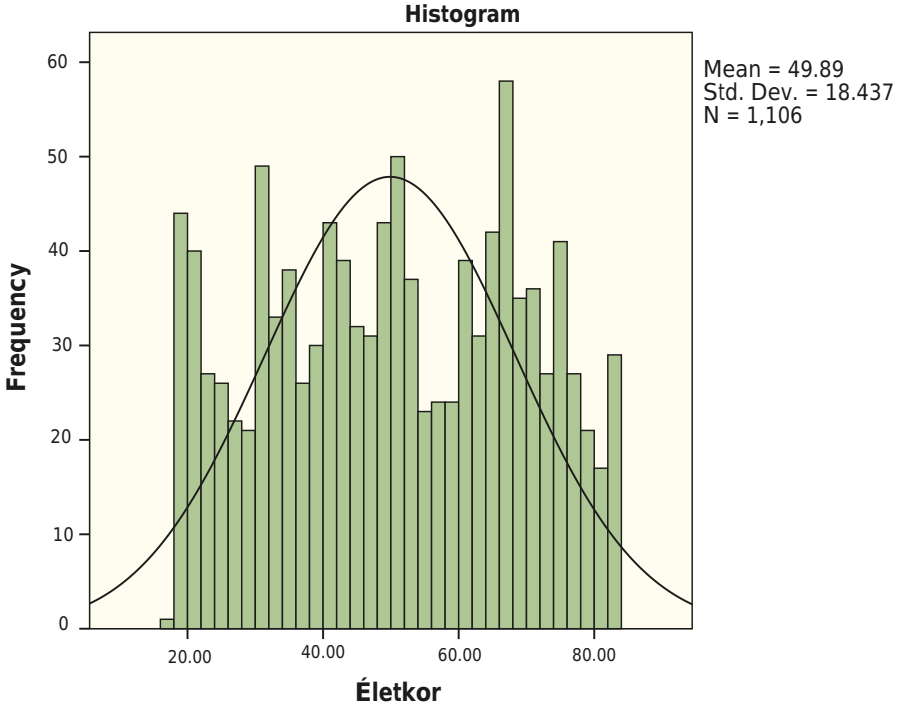
Alakmutatóink értékét az *Output* ablakban tekinthetjük meg (32. ábra).

eletkor

N	Valid	1106
	Missing	0
Skewness		-.035
Std. Error of Skewness		.074
Kurtosis		-1.141
Std. Error of Kurtosis		.147

32. ábra. *Alakmutatók (28. példa)*

A ferdeségi mutató -0,35, tehát kisebb, mint 0. Bár a negatív értékek bal oldali ferdeséget jeleznek (a normális eloszláshoz képest több a nagyobb érték), a hüvelykujjszabály szerint csak az 1-nél kisebb értékek utalnak olyan eloszlásra, amely szignifikánsan különbözik a normális eloszlástól. Ilyen módon a kapott értékünk alapján nem beszélünk bal oldali ferdeségről, vagyis nem mondhatjuk, hogy szignifikánsan több lenne az idősebb, mint a fiatalabb megkérdezett a mintában. A csúcsossági mutatónk –1,141, tehát egy elég alacsony negatív érték. Ekkor azt mondhatjuk, hogy a normális eloszláshoz képest az adataink egy kicsivel kisebb mértékben csoportosulnak a centrális értékek körül (egy kicsivel laposabb a görbénk), ahogyan ezt a 33. ábra is mutatja.



33. ábra. *A hisztogram és a normális eloszlás görbéje (28. példa)*

3. FEJEZET

MINTAVÉTEL

3.1. Elemi valószínűségelmélet, várható érték

A valószínűség definíciói: a klasszikus (eseményekre épülő) definíció

A *kísérlet* olyan jelenség, amely ugyanolyan körülmények közt akárhányszor ismételhető (a valóságban *nagyon hasonló körülmények között*, mert két kockadobás alatt pár molekula különbség beállhat a dobókocka anyagában, pár ezrednyi Celsius-fok különbség a hőmérsékletében stb.). A kísérlet egyszeri ismétlése a *próba*, mely során egyértelműen eldönthetjük, hogy valamely, a kísérlet kimenetelére tett kijelentésünk bekövetkezett-e vagy nem. Tehát eseménynek azt a kijelentést tekintjük, amelyről a próbák során egyértelműen eldönthető az, hogy bekövetkezett-e vagy nem (pl. „a 6-os szám megjelenése a kockán").

A próba lehetséges kimenetelei az *elemi események* (az egyetlen lehetséges esettel megvalósuló események), ezek sokasága pedig az *E* jelű halmaz. Minden egyes vizsgálat alkalmával bármely esemény megvalósulhat (bekövetkezik) vagy nem valósulhat meg (nem következik be), és minden esemény meghatározható a *kedvező esetek*, kimenetelek valamilyen halmazával, vagyis *E*-nek valamely részhalmazával.

Szélsőséges esetekben az esemény lehet biztos esemény és lehetetlen esemény: a *biztos esemény* minden vizsgálat során teljes bizonyossággal bekövetkezik, a *lehetetlen esemény* a kísérlet egyetlen ismétlésekor sem következhet be. Két vagy több esemény *egymást kizáró (inkompatibilis) esemény*, ha a kísérlet egyetlen ismétlése során sem valósulhatnak meg egyszerre. Például legyen egy kísérlet a játékkocka dobása. A kísérlet leírásához tartozik még a megfigyelt véletlen jelenség leírása: a felső lapon levő pöttyök száma. Egy próba előtt nem tudjuk biztosan, hogy hányast fogunk dobni, de abban biztosak lehetünk, hogy a felső lapon 1, 2, 3, 4, 5 vagy 6 pötty lesz. Az elemi események ekkor: a kocka felső lapján 1 pötty van, a kocka felső lapján 2 pötty van, …, a kocka felső lapján 6 pötty van. Az elemi eseményeket minél egyszerűbben szokták jelölni, ebben az esetben erre legalkalmasabb a pöttyök számát adó számjegy: 1, 2, …, 6. Az eseménytér ekkor az $E = \{1, 2, 3, 4, 5, 6\}$. Biztos esemény lehet ilyenkor az a kijelentés, hogy 7-nél kevesebb pötty van a kocka felső lapján, lehetetlen esemény pedig, hogy a kocka felső lapján 7 pötty van.

Az eseményekhez számszerű érték, az esemény valószínűsége rendelhető, és a valószínűségszámítás megmutatja, miként rendelhetünk hozzá eseményekhez valós számokat. Feltételezve, hogy egy tetszőleges A esemény *h*-féleképpen következhet be az összes, egyformán lehetséges *n* kimenetelből, akkor az esemény előfordulásának (kedvező kimenetelének) valószínűsége:

$$p = \Pr\{A\} = \frac{h}{n}$$

Annak a valószínűsége, hogy az esemény nem következik be (kedvezőtlen kimenetel):

$$q = \Pr\{nemA\} = \frac{n-h}{n} = 1 - \frac{h}{n} = 1 - p = 1 - \Pr\{A\}$$

Ilyen módon p + q = 1, azaz Pr{A} + Pr{nem A} = 1. *Egy esemény bekövetkezésének valószínűsége mindig egy 0 és 1 közötti szám.* Ha az esemény nem következhet be (lehetetlen esemény), akkor valószínűsége 0, ha az eseménynek be kell következnie (biztos esemény), akkor valószínűsége 1.

Ha egy esemény bekövetkezésének valószínűsége p, akkor p : q („p a q-hoz") annak az esélye, hogy bekövetkezik, és q : p annak az esélye, hogy nem következik be.

A valószínűség definíciói: a relatív gyakoriságra épülő definíció – statisztikai definíció

A valószínűség klasszikus definíciójának az a hátránya, hogy sok olyan kísérlet van, amelyben a lehetséges kimenetelek nem egyformán valószínűek, vagy nem vezethetők le olyan modellből, ahol a lehetséges kimenetelek egyformán valószínűek. Ilyenkor az események valószínűségének megfelelő becslésére a relatív gyakoriságok használhatók.

Nagyon nagy számú megfigyelés esetén egy esemény becsült vagy tapasztalati valószínűsége az esemény bekövetkezésének *relatív gyakorisága*. Ekkor maga a valószínűség a relatív gyakoriság határértéke, amikor a megfigyelések száma korlátlanul nő. Például ha egy érmét 1000-szer feldobunk, 529-szer fej lesz az eredmény, így a relatív gyakoriság 529/1000 = 0,529. Ha a következő 1000 dobás 493 fejet eredményez, akkor az összes 2000 dobásból a fej relatív gyakorisága (529 + 493)/2000 = 0,511. A statisztikai definíció szerint ilyen módon folytatva végül egyre közelebb jutunk ahhoz az értékhez, amely megmutatja, hogy mennyi a fej valószínűsége egy érme feldobása esetén.

Ez a statisztikai megközelítés a gyakorlatban hasznos, viszont matematikai szempontból problémás, mivel a tényleges határérték nem biztos, hogy létezik. Ezért a modern valószínűségelmélet axiomatikusan felépített, azaz a valószínűség fogalmát nem definiálja.

Feltételes valószínűség: független és nem független események

Ha A_1 és A_2 egy-egy esemény, akkor annak valószínűségét, hogy A_2 bekövetkezik, feltéve, hogy A_1 már bekövetkezett, A_2 A_1-re vonatkoztatott feltételes valószínűségének nevezzük.

$$\Pr\{A_2|A_1\} \quad \textit{vagy} \quad \Pr\{A_2 \text{ feltéve } A_1\}$$

Ha A_1 bekövetkezése vagy nem bekövetkezése nem befolyásolja A_2 bekövetkezésének valószínűségét, akkor A_1 és A_2 *független események*.

$$\Pr\{A_2|A_1\} = \Pr\{A_2\}$$

Ha A_1A_2-vel jelöljük azt az eseményt, hogy „mind A_1, mind A_2 bekövetkezik” (összetett esemény):

$$\Pr\{A1\ A2\} = \Pr\{A1\} \cdot \Pr\{A2|A1\} \quad \text{– függő eseményekre}$$
$$\Pr\{A1\ A2\} = \Pr\{A1\} \cdot \Pr\{A2\} \quad \text{– független eseményekre}$$

Három eseményre (A_1, A_2, A_3):

$$\Pr\{A_1A_2A_3\} = \Pr\{A_1\} \cdot \Pr\{A_2|A_1\} \cdot \Pr\{A_3|A_1A_2\} \text{ – függő eseményekre}$$
$$\Pr\{A_1A_2A_3\} = \Pr\{A_1\} \cdot \Pr\{A_2\} \cdot \Pr\{A_3\} \text{ – független eseményekre}$$

Általános esetben ha A_1, A_2, A_3, ..., A_n *n* számú független esemény, amelynek valószínűségei rendre p_1, p_2, p_3, ..., p_n, akkor A_1 és A_2 és A_3 és …A_n együttes bekövetkezésének valószínűsége $p_1 \cdot p_2 \cdot p_3 \cdot \ldots \cdot p_n$.

29. példa ▼

▶ *Függő és független események*

Nézzük az alábbi feladatot. Egy jól megkevert, 52 lapos kártyacsomagból 2 lapot húzunk ki. Határozzuk meg annak a valószínűségét, hogy mindkét lap ász lesz, ha:

A) az első lapot visszatesszük
B) az első lapot nem tesszük vissza

Az A) esetünkben két független eseményünk van:
A1 – az első lap ász
A2 – a második lap ász

$$\Pr\{A_1\ A_2\} = \Pr\{A_1\} \cdot \Pr\{A_2\} = \frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$$

A második esetben függő eseményekkel van dolgunk, hiszen a két esemény együttes bekövetkezése függ attól, hogy elsőként milyen lapot húztunk.

$$\Pr\{A_1\ A_2\} = \Pr\{A_1\} \cdot \Pr\{A_2|A_1\} = \frac{4}{52} \cdot \frac{3}{52} = \frac{1}{221}$$

Valószínűségeloszlások

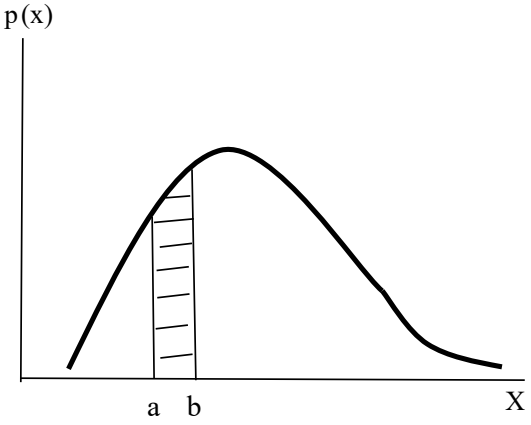
Diszkrét eloszlások

Ha egy X változó az X_1, X_2, ... X_k diszkrét értékeket veheti fel, rendre p_1, p_2, ... p_k valószínűségekkel, ahol $p_1 + p_2 + ... p_n = 1$, akkor ezzel X-hez egy *diszkrét valószínűségeloszlást* definiáltunk. A *p(X)* függvényt, amelynek értékei $X = X_1$, X_2, ...X_k-ra rendre a p_1, p_2,.....p_k értékek, *X valószínűségi vagy gyakorisági függvényének nevezzük.* Mivel X csak bizonyos értékeket vehet fel előre meghatározott valószínűségekkel, ezért *diszkrét véletlen változónak* szokták nevezni. A véletlen változót *sztochasztikus változónak* is szokták nevezni. A relatív gyakorisági eloszláshoz való hasonlósága miatt a *valószínűségeloszlások a relatív gyakoriságeloszlások ideális határértékeként* is felfoghatóak (amikor a megfigyelések száma nagyon nagy). Ilyen módon a valószínűségeloszlások *sokasági eloszlások*, a relatív gyakorisági eloszlások a sokaságból vett *minták eloszlásai*.

A valószínűségek egymás utáni összeadásával kumulált valószínűségeloszlásokat kapunk. A kumulált valószínűségeloszlás hasonló a kumulált relatív gyakorisági eloszláshoz, és a hozzá rendelt függvényt *eloszlásfüggvénynek* nevezik.

Folytonos eloszlások

A folytonos eloszlás arra az esetre vonatkozik, amikor X változó folytonos halmazon vehet fel értékeket. A minta relatív gyakorisági poligonja sokaságra folytonos görbe lesz, melynek egyenlete *Y = p(X)* (34. ábra).



34. ábra. *A sűrűségfüggvény*

A görbe alatti, az X tengely által határolt rész teljes területe 1. Az X = a és az X = b egyenesek által határolt görbe alatti terület annak a valószínűségét adja meg, hogy X az a és b érték közé esik (Pr{a < X < b}). A p(X) függvény neve *valószínűségi sűrűségfüggvény* vagy csak sűrűségfüggvény, és ezzel definiáljuk X folytonos valószínűségeloszlását. Ebben az esetben X folytonos véletlen változó.

Várható érték

Ha annak a valószínűsége, hogy valaki S összegű pénzt kap p, akkor a matematikai várható érték vagy várható érték $p \cdot S$. Ha X diszkrét valószínűségi változó X_k értékekkel és rendre p_k valószínűségekkel, akkor X várható értéke E(X):

$$E(X) = p_1 \cdot X_1 + p_2 \cdot X_2 + \ldots + p_k \cdot X_k = \sum_{j=1}^{k} p_j \cdot X_j$$

Amennyiben a p_j valószínűségeket f_j/n relatív gyakoriságokkal helyettesítjük ($n = \sum f_j$), akkor a várható érték:

$$E(X) = \frac{\sum_{j-1}^{k} f_j \cdot X_j}{n} = \overline{X}$$

Minél nagyobb az *n*, annál inkább közelítik a relatív gyakoriságok a valószínűségeket. Ilyen módon E(X)-et úgy is tekinthetjük, mint annak a sokaságnak az átlagát, amelyikből a mintát vettük. A várható érték folytonos valószínűségi változók esetén a matematikai analízis eszközeivel definiálható.

30. példa ▼

▶ *A várható érték kiszámítása*

A következő példánk egy üzleti vállalkozás helyzetét szemlélteti, amelyben egy szerződés megkötése 60%-os valószínűséggel 300 dollár nyereséget, 40%-os valószínűséggel pedig 100 dollár veszteséget fog hozni. A kérdés, hogy ebben a helyzetben érdemes-e megkötni az üzletet, vagyis mennyi a várható nyereség/veszteség összege.

$$E(X) = p_1 \cdot X_1 + p_2 \cdot X_2 + \ldots + p_k \cdot X_k = 0{,}6 \cdot 300 + 0{,}4 \cdot (-100) = 140$$

Tehát a szerződés várhatóan 140 dollár nyereséget fog hozni.

3.2. Elemi mintavételi elmélet, standard hiba

Bevezetés a mintavételbe

A mintavétel a társadalomstatisztikában az adatokhoz való hozzájutás fő módja. A mintavétel melletti legfontosabb érv az, hogy a sokaság igen nagyszámú egyedből áll, és ezek teljes körű lekérdezése egyrészt rendkívül idő- és energiaigényes, másrészt az ekkora adatfelvételből adódó hiba minden bizonnyal felülmúlná a mintavételi hibákat. A mintavétel célja olyan adatokat nyerni, amelyek segítségével megalapozott következtetéseket lehet levonni a sokaságra (populációra) vonatkozóan. Egy mintából akkor vonhatunk le használható következte-

téseket a sokaságra nézve, ha a mintának lényegében (a kutatás szempontjából lényeges változók tekintetében) ugyanolyan az összetétele, mint a sokaságnak (reprezentativitás).

Az adatfelvételek mindig tartalmaznak hibákat, viszont ezek egy részét a statisztika segítségével meg lehet becsülni, a lehetőségek keretei között lehet csökkenteni. Tehát az adatfelvételi hibák alapvetően kétfélék: nem mintavételi és *mintavételi hibák* (a hiba abból adódik, hogy nem a teljes sokaságot figyeltük meg). Bizonyos mintavételi tervek esetén *a mintavételi hiba nagysága előre becsülhető*, míg *a nem mintavételi hiba nagyságát sem előre, sem utólag nem lehet megadni*.

A mintavételi tervek alapvető kérdése az, hogy miként választjuk ki a mintát: *véletlenszerűen* – ekkor *valószínűségi mintavételről* beszélünk, vagy nem véletlenszerűen – ekkor *nem valószínűségi mintavétellel* van dolgunk.

A reprezentatív mintavétel főként véletlen kiválasztáson alapul (a sokaság minden egységének egyforma esélye van a mintába való bekerülésre: $p = 1/N$), ilyen módon a valószínűségelmélet segítségével meg tudjuk becsülni, hogy a minta mennyire pontosan írja le a sokaságot.

A mintavételi tervek fajtái:

a) véletlen mintavételi tervek

1. egyszerű véletlen minta (homogén, véges, visszatevés nélkül),
2. független, azonos eloszlású minta (homogén, végtelen, nagyon nagy VAGY véges, visszatevéses),
3. szisztematikus minta (homogén, véges, visszatevés nélküli, lépésköz alkalmazása),
4. rétegzett minta (homogén rétegekbe sorolás, majd egyszerű véletlen minta),
5. csoportos minta (homogén, véges, nagyobb összetartozó csoportokból mindenkit),
6. többlépcsős minta (több lépésben jutunk el a megfigyelt egységekhez).

b) nem véletlen mintavételi tervek

1. kvótás minta (előre megadott összetételű mintához való véletlen hozzájutás),
2. önkényes vagy szakértői minta,
3. hólabda minta,
4. egyszerűen elérhető alanyokra hagyatkozó minta.

Az egyes mintavételi tervek részletes kifejtésére ebben a jegyzetben nem térünk ki, mivel ez a *Társadalomkutatási módszerek és technikák* tárgy keretében történik.

A mintavétel elmélete

A jelenségeknél, ha azonos körülményeket biztosítunk, és ugyanarra a jelenségre nézve ugyanazt a vizsgálatot többször elvégezzük, akkor „n" számú megfigyelésnél az esemény „k" számú előfordulása (relatív gyakorisága) valószínűségi változóként kezelhető. *Bernoulli tétele* alapján a relatív gyakoriság eltérése a vizsgált jelenség előfordulási valószínűségétől tetszőleges valószínűséggel tetszőlegesen kismértékűvé tehető, ha a minta nagysága (n) minden határon túl növekszik (nagyszámok törvénye). A törvény szerint ha a mintaelemek számát fokozatosan növeljük, a bizonyosság felé közeledik annak a valószínűsége, hogy a relatív gyakoriság és a matematikai valószínűség csak az általunk tetszőlegesen és előre meghatározható mértékben tér el. Nyilvánvaló ugyanakkor az is, hogy a társadalmi élet területén a törvény érvényesülése korlátozott (a társadalmi jelenségek tulajdonságai változnak), de érvényes az a megállapítás, amely szerint *minél nagyobb a minta, annál pontosabb az ebből nyert becslés*.

A véletlen tömegjelenségeknél a tapasztalatok szerint a normális vagy arra visszavezethető eloszlás a leggyakoribb. A *köponti határeloszlás tétele* szerint (Markov és Ljapunov) minden véletlen esemény, amely sok egymástól független valószínűségi változó összegzéseként áll elő, és ezek értéke összegükhöz mérten igen kicsi, jó megközelítéssel normális eloszlású lesz.

A standard hiba

Amennyiben tehát a mintavételnél biztosítottuk az alapsokaság minden tagjának a mintába való bekerülését, akkor a központi határeloszlás tételének megfelelően egy adott változó esetében ennek a mintabeli átlagértéke, mint valószínűségi változó, erősen megközelít egy N(m,σ) paraméterű normális eloszlású változót, ahol m és σ a teljes sokaságbeli átlagérték és szórás.

A normális eloszlás jellegzetességeiből az következik, hogy ha a valószínűségi változók normális eloszlást mutatnak, akkor meghatározható, hogy a várható érték (az alapsokaság átlaga) bizonyos határok közötti elhelyezkedésének milyen a valószínűsége. A határok kijelölésénél a szórást (vagy annak többszöröseit) vehetjük figyelembe. A szórás által kijelölt határokat *valószínűségi határoknak*, a határok közé esés valószínűségét pedig *valószínűségi szintnek* nevezzük.

Az alapsokaságból nyerhető *lehetséges mintaátlagok szórása* vagy a mintaátlagok *standard hibája* egyenesen arányos az alapsokaság szórásával és fordítottan arányos a mintanagyság négyzetgyökével. Tehát minél nagyobb a minta nagysága, annál kisebb a lehetséges mintaátlagok szórása, a standard hiba. Ha a minta nagysága egyenlő a sokaság nagyságával, a standard hiba = 0.

Ez a standardhiba-meghatározás nyilvánvalóan a valószínűségszámítás elméletének arra az esetére vonatkozik, amikor a sokaságból nagyszámú véletlen mintát veszünk. Ha ismerjük a sokaság jellemzőit és nagyon nagy számú véletlen mintát veszünk, akkor meg lehet becsülni, hogy a mintákból számolt statisztikák

közül hány fog a sokaság átlaga körüli meghatározott nagyságú intervallumokba esni.

Azonban egy valós kutatásnál általában egészen más történik. Mivel általában azért végzünk kutatásokat, hogy a sokaság paraméterét megbecsüljük, ezt előzőleg nem ismerjük. Továbbá általában nem szokás nagyszámú mintát venni, csak egyet.

A gyakorlatban tehát legtöbbször nem ismerjük az alapsokaságra vonatkozó átlagot és szórást, ezért az egyetlen mintánkon mért adatainkból becsüljük meg az alapsokaságra vonatkozó értékeket.

Amennyiben az alapsokaság-beli átlagot akarjuk megbecsülni, az alábbi képlettel számoljuk a standard hibát:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

ahol n a minta nagysága, σ a minta szórása és $\sigma_{\bar{x}}$ a standard hiba.

A normális eloszlás a korábbiakban elmondottak alapján tehát lehetővé teszi, hogy megállapítsuk becslésünk megbízhatóságát, *valószínűségi szintjét* (a minta átlagától milyen valószínűséggel tér el az alapsokaság átlaga). Továbbá így a standard hiba egy tetszőleges t többszörösével megadhatjuk a becslésünk hibahatárát, konfidencia (megbízhatósági) intervallumát. Az átlag esetében ezt a

$$\bar{x} \pm t \cdot \sigma_{\bar{x}}$$

képlettel számoljuk ki.

A *t*-értékekhez tartozó leghasználatosabb valószínűségek a 17. táblázatban szerepelnek.

17. táblázat. *A leghasználatosabb valószínűségi szinteknek megfelelő t- és p- értékek (n > 120)*

t érték	Statisztikai biztonság (valószínűségi vagy konfidenciaszint)	Szignifikanciaszint (p-érték)
1,65	90%	0,10
1,96	95%	0,05
2,58	99%	0,01
3,29	99,9%	0,001

Dichotóm ismérvek esetén a standard hibát könnyebb megbecsülni a relatív gyakoriságok (vagy valószínűségek szorozva 100) segítségével:

$$\sigma_p = \sqrt{\frac{p \cdot q}{n}}$$

és ekkor a konfidenciaintervallumot az alábbi képletekkel számoljuk:

$$p \pm t \cdot \sigma_p \quad q \pm t \cdot \sigma_p$$

31. példa ▼

▶ *A standard hiba és a konfidenciaintervallum kiszámítása*

1. A repülőtéri utasokból egy 100 elemű véletlen mintát veszünk. A mintába bekerült utasok átlagos súlya 80 kg, a minta szórása 20 kg. Állapítsuk meg 95%-os valószínűséggel (t = 1,96) a repülőtéri utasok átlagos súlyát.
   Első lépésben kiszámítjuk a standard hibát:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{100}} = 2$$

   Második lépésben kiszámítjuk a két konfidenciaintervallumot:

$$\bar{X} \pm t \cdot \sigma_{\bar{x}} \quad 80 + 1{,}96 \cdot 2 = 83{,}92 \quad 80 - 1{,}96 \cdot 2 = 76{,}08$$

   Tehát 95%-os valószínűséggel (0,05-ös szignifikanciaszint mellett) a konfidenciaintervallum: (76,08–83,92). 95%-os valószínűséggel kijelenthetjük, hogy a repülőtéri utasok átlagos súlya 76,08 és 83,92 kg között van.
   99,7%-os valószínűségi szint mellett (t = 3) azt mondhatjuk, hogy a repülőtéri utasok átlagos súlya 74 és 86 kg között van (80 ± 3•2). Tehát nagyobb valószínűségi szint mellett szélesebb a megbízhatósági intervallum is.

2. X kisvárosban egy 1000 fős véletlen mintát vettek a 18 éven felüli lakosságból. A mintába bekerült személyek 45%-a A-t, 55%-a pedig B-t választaná polgármesternek. Számítsuk ki, hogy 95%-os valószínűséggel ki fog nyerni a választásokon.

$$\sigma_p = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{45 \cdot 55}{1000}} = \sqrt{2{,}475} = 1{,}57$$

$$p \pm t \cdot \sigma_p \quad 45 \pm 1{,}96 \cdot 1{,}57 \quad 45 + 3{,}08 = 48{,}08 \quad 45 - 3{,}08 = 41{,}92$$

   Tehát 95%-os valószínűséggel (p = 0,05-ös szignifikanciaszint mellett) a kisváros választópolgárainak 41,92–48,08%-a fog A-ra szavazni, így 95%-os valószínűséggel állíthatjuk, hogy B fogja megnyerni a választásokat.

A nem valószínűségi mintavételi eljárások és szignifikanciatesztek alkalmazása a társadalomtudományokban

A standard hibák, konfidenciaintervallumok és szignifikanciatesztek közös statisztikai alapja az, hogy a megfigyelt mintabeli mintázatok mögött feltételezhető valamely véletlen mintavételi ingadozás. Valószínűségi mintavétel esetén a standard hiba a mintavételi bizonytalanság formális mérőszáma, a konfidenciaintervallumok a populációs paraméterek becslésének megbízhatóságát fejezik ki, a szignifikanciatesztek pedig annak valószínűségét értékelik, hogy egy megfigyelt hatás pusztán a véletlen következtében állt elő. E három eszköz tehát akkor működik elméletileg hibátlanul, ha a mintavétel minden szakaszában teljesül a véletlenszerű kiválasztás feltétele.

A jegyzet során használt kutatás esetén olyan háromlépcsős mintavételi eljárást alkalmaztak, amelyben az első két lépés, azaz a települések rétegzett, véletlenszerű kiválasztása és a háztartások véletlen kezdőpontú, lépésközös kiválasztása valószínűségi logikát követett, míg a harmadik lépésben használt kvótás személykiválasztás már nem (részletes leírását az *1.4. Adatbázisok létrehozása, címkézés* alfejezetben ismertettük). Ennek következtében a teljes minta nem tekinthető teljes mértékben valószínűségi mintának, így az inferenciális statisztikai eljárások klasszikus előfeltevései csak részben teljesülnek.

Ez módszertani feszültséget eredményez az elméleti statisztikai követelmények és a társadalomtudományi adatfelvételi gyakorlat között. Mindazonáltal a szignifikanciatesztek, a standard hibák és a konfidenciaintervallumok kiszámítása továbbra is indokolt lehet: nem a szigorú populációs általánosítás céljával, hanem a mintán belüli bizonytalanság strukturált, összehasonlítható és értelmezhető leírása érdekében. Ezek az eszközök jelzik, hogy a megfigyelt összefüggések mennyiben tekinthetők stabilnak, illetve mennyire valószínű, hogy csupán a véletlenszerű mintabeli eltérésekből erednek.

Fontos azonban hangsúlyozni, hogy ilyen, részben nem valószínűségi minták esetében az eredmények értelmezése fokozott óvatosságot igényel. A szignifikanciaszintek és konfidenciaintervallumok nem szolgálhatnak erős populációs következtetések alapjául, hanem inkább tájékozódó jelzésként értelmezhetők, amelyek a mintabeli mintázatok stabilitására, robusztusságára utalnak. A korlátok egyértelmű és hangsúlyos bemutatása azért fontos, mert rávilágít arra, hogy a gyakorlati adatfelvétel sokszor eltér az elméleti ideáltípusoktól, ami szükségessé teszi az eredmények körültekintő értelmezését.

Ugyanakkor itt azt is fontos megjegyeznünk, hogy a szignifikanciatesztek alkalmazása a társadalomtudományokban sajátos értelmezési keretben nyeri el jelentőségét, amely gyakran túlmutat a pusztán technikai-statisztikai megközelítésen (erre korábban a bevezető *1.1. Mi a statisztika?* alfejezetben már utaltunk). Ezek az eljárások elsősorban annak valószínűségét hivatottak jelezni, hogy a megfigyelt eredmények a véletlenszerű ingadozás következményei-e, nem pedig

annak eldöntésére szolgálnak, hogy egy társadalmi jelenség létezik-e vagy sem. A p-értékek értelmezése azonban, ahogyan korábban láttuk, nagymértékben függ a minta nagyságától: nagy minták esetében egészen csekély, gyakorlati szempontból marginális különbségek is könnyen szignifikánsnak bizonyulhatnak, míg kis minták esetében akár számottevő hatások is rejtve maradhatnak a statisztikai bizonytalanság miatt, és így „nem szignifikáns" eredményként jelennek meg. Ez a sajátosság világossá teszi, hogy a szignifikancia önmagában nem elegendő a társadalmi jelenségek érdemi értelmezéséhez. A kutatói gyakorlatban ezért elengedhetetlen a p-érték mellé a hatásméret, az elemzett jelenség kontextusa, valamint az elméleti háttér együttes mérlegelése, mivel csak így biztosítható, hogy az eredmények valódi társadalomtudományi tartalommal bírjanak, és ne csupán statisztikai formalitásként jelenjenek meg.

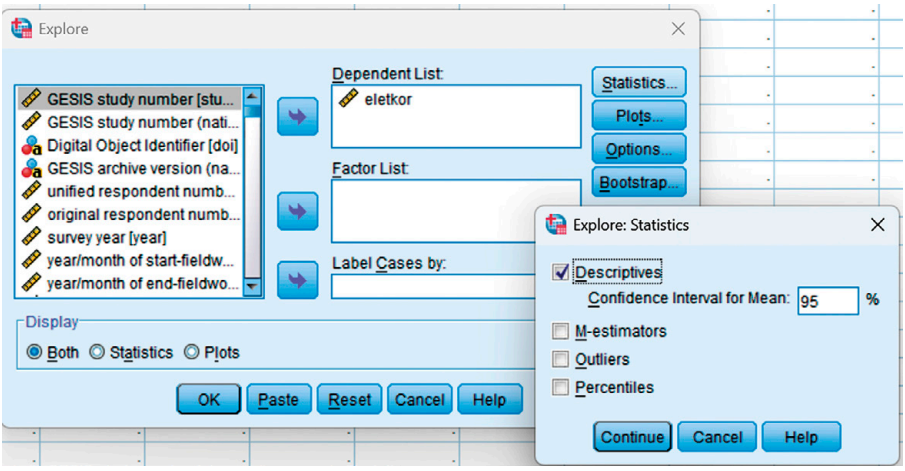
Konfidenciaintervallum kiszámítása az SPSS-sel

A megbízhatósági intervallumot SPSS-ben az *ANALYZE* főmenü *Descriptive Statistics, Explore* menüpontjánál lehet lekérni. Itt a program 95%-os megbízhatósági intervallumot számol az átlagra, de a valószínűségi szint a *Statistics* mezőben tetszőlegesre állítható.

32. példa ▼

▶ *Konfidenciaintervallum az SPSS-ben*

Nézzük ismét az *eletkor* változót. Az előzőekben leírtak szerint lekérjük a 95%-os valószínűségnek megfelelő konfidenciaintervallumot (35. ábra).



35. ábra. *A konfidenciaintervallum lekérése (32. példa)*

A kért adatok az *Output* ablakban olvashatóak (36. ábra).

Descriptives

			Statistic	Std. Error
eletkor	Mean		49.8870	.55440
	95% Confidence Interval for Mean	Lower Bound	48.7992	
		Upper Bound	50.9748	
	5% Trimmed Mean		49.8858	
	Median		50.0000	
	Variance		339.937	
	Std. Deviation		18.43739	
	Minimum		17.00	
	Maximum		82.00	
	Range		65.00	
	Interquartile Range		32.00	
	Skewness		-.035	.074
	Kurtosis		-1.141	.147

36. ábra. *A konfidenciaintervallumok megjelenítése az Output-ban (32. példa)*

Tehát 95%-os valószínűségi szint mellett (p = 0,05) állíthatjuk, hogy a megkérdezettek átlagos életkora 48,8–50,97 év között van. 99%-os megbízhatósági szint mellett (p = 0,01) a konfidenciaintervallum 48,46–51,32 év között helyezkedik el.

4. FEJEZET

KÉTVÁLTOZÓS ELEMZÉSEK

4.1. Változók közötti kapcsolatok

Az ismérvek közötti kapcsolatok vizsgálatának célja a valóság jelenségei között fennálló összefüggések tömör, számszerű jellemzése. Ez a terület a statisztikai módszertan kiemelkedő részét képezi.

Egy sokaság egységei különféle tulajdonságaik felsorolásával jellemezhetőek. A tulajdonságok egy része a sokaság minden egységére nézve közös, másik része azonban egységről egységre változik, azaz egyedi. Végső soron minden tulajdonság a vizsgált egységekre vonatkozó ismereteket pontosítja valamilyen módon. Ha a vizsgált sokaság egységeinek valamilyen nem közös tulajdonságát rögzítjük, akkor mindig egy *részsokasághoz* jutunk (leszűkül az egységek köre). Egy ismérv/változó vizsgálatára azért van szükség, mivel az egyes egységek különböző ismérvértékeket vesznek fel, tehát szóródó változókat elemzünk (a „szóródás" itt és a továbbiakban nagyon általánosan értendő: minőségi ismérvekre is vonatkozik).

Az ismérvek közötti kapcsolat

Egy sokaság (a továbbiakban *fősokaság*) egységeinek valamilyen ismérv (*X*) szerinti megoszlását *feltétel nélküli megoszlásnak* nevezzük (pl. *önkénteskedik-e*). A fősokaságból egy más ismérv (*Y*) alapján kijelölt részsokaságok (pl. *a kérdezett neme*) előző (*X*, *önkénteskedik-e*) ismérv szerinti megoszlását *feltételes megoszlásnak* nevezzük (pl. önkénteskedik, ha az illető nő). Míg a feltétel nélküli megoszlások mindig (másképp nem lenne értelme az elemzésnek), addig a feltételes megoszlások nem szükségképpen szóródóak (egy jó osztályozással néha el lehet érni, hogy egy-egy részsokaságba a vizsgált ismérv szempontjából azonos vagy közel azonos elemek kerüljenek). Amennyiben például az önkénteskedés hasonló a férfiak és nők között, a feltételes megoszlások is hasonlóak lesznek. Vagy pl. ha a kitűnő teljesítményt nyújtó sportolók jövedelemkülönbségeit vizsgáljuk, egy jó sportágakra alapuló csoportosítással el lehet érni, hogy egy-egy kategóriába nagyon hasonló jövedelmű sportolók kerüljenek.

A feltételes megoszlások szóródásának vizsgálata az ismérvek közötti kapcsolatra világít rá.

A feltételes megoszlásoknak a feltétel nélküli megoszláshoz való viszonyulása kétféle lehet.

1. *Minden feltételes megoszlás egyforma*, így megegyezik a feltétel nélküli megoszlással – ekkor *függetlenség* áll fenn. A részsokaságok képzésére használt csoportképző ismérvet (*Y*) és a részsokaságon belüli elemzésre használt ismérvet (*X*) egymástól függetlennek nevezzük, amikor az *Y* szerinti csoportba való tartozásának ismerete (pl. ha tudjuk a kérdezett nemét) nem ad semmiféle többletinformációt a részsokaságon belül használt valamely más ismérv, vagyis *X* (pl. *önkénteskedés*) szerinti hovatartozásáról, tulajdonságáról.

2. *Nem minden feltételes megoszlás egyforma – a két változó között összefüggés van:*

a) a feltételes megoszlásokon belül van szóródás
– *sztochasztikus* (statisztikus) *kapcsolat*,

b) a feltételes megoszlásokon belül nincs szóródás
– *determinisztikus, függvényszerű kapcsolat.*

Amikor a két változó között összefüggés van, biztosan tudjuk, hogy legalább egy feltételes megoszlás más, mint a feltétel nélküli megoszlás (pl. a férfiak és nők önkénteskedése eltér). Ilyen módon nem mindegy, hogy egy részsokaság (pl. csak férfiak vagy nők), vagy a teljes sokaság (minden megkérdezett) megoszlását vizsgáljuk, mivel a csoportosító ismérv (*Y*) nem független a másik ismérvtől (*X*), a kettő között összefüggés van.

Determinisztikus kapcsolat esetén a részsokaságon belüli ismérvértékek nem szóródnak, a csoportképző (*Y*) ismérv egyértelműen meghatározza a másik ismérv (*X*) nagyságát vagy értékét. Ebben az esetben a két ismérv függvényszerű kapcsolatban áll egymással: az *Y* értéke pontosan megadja *X*-ét. Például ha *a kérdezett neme* (*Y*) és az *önkénteskedés* (*X*) közötti összefüggést vizsgáljuk, determinisztikus kapcsolat esetén minden férfi önkénteskedik, és egyetlen nő sem végez önkéntes tevékenységet. Tehát ha ismerjük a személy nemét (az *Y* változóra felvett értékét), egyértelműen meghatározhatjuk, hogy önkénteskedik-e vagy nem (az *X* szerinti értékét). Nyilvánvalóan a determinisztikus kapcsolat a valóságban igen ritkán fordul elő, sokkal gyakoribbak a sztochasztikus kapcsolatok.

A sztochasztikus kapcsolat a függetlenség és a determinisztikus kapcsolat között helyezkedik el: az ismérvek nem függetlenek, de nincs is közöttük függvényszerű kapcsolat. Az egyik ismérv (*Y*) hatással van a másikra (*X*), de annak értékeit nem határozza meg egyértelműen. Sztochasztikus kapcsolat esetén az *Y* ismérv szerinti hovatartozás ismeretében levonható valamilyen következtetés az egységek *X* szerinti hovatartozásáról, de ez a következtetés nem teljesen egyértelmű. Az előző példánk esetében a kérdezett nemének ismeretében következtethetünk arra, hogy pl. a férfiak nagyobb arányban önkénteskednek, mint a nők, de ha tudjuk, hogy valaki férfi, az még nem jelenti egyértelműen, hogy önkénteskedik.

Az eddigiek könnyen általánosíthatóak kettőnél több ismérv esetére is. Több ismérv kapcsolatának vizsgálatakor az is elemezhető, hogy milyen természetű kapcsolat van két vagy több ismérv között (X_1, X_2 stb.) egy másik ismérv (*Y*) szerint kialakított részsokaságon belül (parciális kapcsolat).

Az ismérvek közötti kapcsolat fajtái

Amikor két vagy több ismérv közötti kapcsolatot vizsgálunk, először mindig meg kell néznünk, hogy van-e kapcsolat a vizsgált ismérvek között, s amennyiben van, milyen szoros (annál szorosabb, minél közelebb áll a determinisztikus kapcsolathoz), majd el kell döntenünk, hogyan lehet felhasználni a kapcsolat természetének ismeretét következtetések levonására. A kérdések megválaszolása függ az egyszerre vizsgált *ismérvek számától* és *mérési szintjétől.*

Ebben a fejezetben csak két ismérv kapcsolatát vizsgáljuk.

Az ismérvek jellege szerint a következő eseteket szokás megkülönböztetni:

- *minőségi változók közötti kapcsolat, asszociáció*: mindkét változó nominális mérési szintű (vagy egyik változónk nominális, a másik pedig ordinális mérési szintű), illetve mindkét változó ordinális mérési szintű,
- *vegyes kapcsolat, átlagértékek összehasonlítása:* egy nominális és egy intervallum- vagy arányskálán mért változó összefüggése,
- *mennyiségi változók közötti kapcsolat, korreláció*: két intervallum- vagy arányskálán mért változó közötti kapcsolat.

Ezt a három esetet kapcsolatfajtáknak nevezik.

A statisztika kizárólag az ismérvek együtt változásának számszerű jellemzésére képes (az együtt változás okát nem vizsgálja). Amikor az ismérvek között közvetlen okozati kapcsolat van, függő és független változókról beszélünk. A jegyzetben a csak feltételezett okozati kapcsolatok esetén is használjuk a függő és független változók közötti megkülönböztetést.

A kapcsolatvizsgálat általános eszközei

Ha a sokaság elég nagy, a két ismérv közötti kapcsolat vizsgálatának legegyszerűbb és legáltalánosabb eszköze a két ismérv szerinti kombinatív osztályozás, a *kontingenciatábla* vagy *kereszttábla* (18. táblázat).

18. táblázat. *A kereszttábla általános formája*

X ismérv szerinti osztályok	**Y ismérv szerinti osztályok**						
	$\mathbf{R_1}$	$\mathbf{R_2}$	...	$\mathbf{R_j}$	...	$\mathbf{R_c}$	$\mathbf{\sum_j}$
C_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1c}	$\mathbf{f_{1.}}$
C_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2c}	$\mathbf{f_{2.}}$
...	...	...	...	...	...	...	...
C_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{ic}	$\mathbf{f_{i.}}$
...	...	...	...	...	...	...	...
C_r	f_{r1}	f_{r2}	...	f_{rj}	...	f_{rc}	$\mathbf{f_{r.}}$
$\sum_i$	$\mathbf{f_{.1}}$	$\mathbf{f_{.2}}$	...	$\mathbf{f_{.j}}$	...	$\mathbf{f_{.c}}$	**N**

C_i – az *X* ismérv szerint képzett i-edik osztály azonosítója (i = 1, 2, ..., r); pl. *önkénteskedés* szerinti két osztály: önkénteskedik, nem önkénteskedik (r = 2),

R_j – az *Y* ismérv szerint képzett j-edik osztály azonosítója (j = 1, 2, ..., c); pl. *a kérdezett neme* szerinti két osztály: féri, nő (c = 2),

f_{ij} – az a gyakoriság, amelynek egyedei X szerint az i-edik, Y szerint a j-edik osztályba tartoznak (pl. önkénteskedő férfiak száma),

r – az *X* szerint képzett osztályok száma,

c – az *Y* szerint képzett osztályok száma,

$f_{i.}$, $f_{.j}$ – *peremgyakoriságok* (pl. nők összesen, férfiak összesen, önkénteskedők összesen, nem önkénteskedők összesen),

N – az összes eset.

A két ismérv közötti kapcsolat fennállása konkrétan a feltételes és feltétel nélküli *X* megoszlások összehasonlításával mutatható ki. Ha minden sorban azonos a megoszlás, *függetlenségről beszélünk.* Ha minden sor csak egy 0-tól különböző gyakoriságot tartalmaz, és ezek nem mind ugyanabban az oszlopban találhatók, akkor *függvényszerű kapcsolatról beszélünk.*

A fentiek alapján a két ismérv közötti kapcsolat léte legegyszerűbben vagy a *soronként számított megoszlási viszonyszámokból* ($f_{11}/f_{1.} = f_{21}/f_{2.}$ stb.), vagy az f_{ij} tényleges és f^*_{ij} feltételezett gyakoriságok összehasonlítása útján vizsgálható. A feltételezett gyakoriságokat a két ismérv függetlenségének feltételezése melletti gyakoriságoknak szokás nevezni. A feltételezett vagy elméleti gyakoriság egyenlő a két változó szerinti feltétel nélküli megoszlások (peremgyakoriságok) szorzatának és a sokaság nagyságának hányadosával:

$$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{N}$$

A *kapcsolat szorosságának mérésére* ez az eljárás csak bizonyos esetekben használható, az egyik ismérv szerinti hovatartozásból a másik ismérv szerinti hovatartozásra való *következtetésre* pedig egyáltalán.

A *PRE-eljárás* a függőség oldaláról közelít. *X* és *Y* között annál szorosabb a kapcsolat, minél nagyobb segítséget ad az egységek *Y* szerinti hovatartozásának ismerete (pl. *a kérdezett neme*) az adott egységek *X* szerinti hovatartozásának (pl. *önkéntesség*) kitalálásához, tehát a többletinformáció mennyiségét próbálja mérni. A PRE minden sztochasztikus kapcsolat szorosságának mérésére alkalmas, azonban a képletben szereplő hibák értelmezése és számítási módja mindig a következtetés konkrét módjától függ.

A PRE mutatószám mindig 0 és 1 közé esik és azt fejezi ki, hogy a vizsgált egységek *Y* szerinti hovatartozásának megtudása milyen mértékben csökkenti az egységek *X* szerinti hovatartozásával kapcsolatos bizonytalanságot. Ha PRE = 0,

egyáltalán nem csökkenti a bizonytalanságot, vagyis a két változó független, ha PRE = 1, akkor teljesen megszűnik a bizonytalanság, tehát a két változó függvényszerű kapcsolatban áll egymással.

A PRE meghatározása:

1. lépés: meghatározzuk, hogy összességében mekkora hibával járna, ha az *X* szerinti hovatartozást kizárólag az *X* szerinti feltétel nélküli megoszlásra alapozva próbálnánk meg kitalálni (E_1),

2. lépés: meghatározzuk az előző értelemben vett összes hibát azon feltevés mellett is, hogy ismerjük az *Y* szerinti hovatartozást is, és azok *X* szerinti hovatartozását mindig a megfelelő feltételes megoszlásra támaszkodva próbáljuk megadni (E_2),

3. lépés: meghatározzuk a hibacsökkenés relatív mértékét, amely a feltételes megoszlások ismeretének tulajdonítható:

$$PRE = \frac{E_1 - E_2}{E_1}$$

A mutatószám azt fejezi ki, hogy a vizsgált egységek *Y* szerinti hovatartozásának ismerete (pl. a kérdezett nemének ismerete) milyen mértékben csökkenti az egységek *X* (pl. önkéntesség) szerinti hovatartozásával kapcsolatos bizonytalanságot.

4.2. Minőségi változók közötti kapcsolat

A minőségi változók értékei között nincsenek egyértelmű mennyiségi különbségek, így a kapcsolatvizsgálat azt jelenti, hogy összehasonlítjuk a feltételes eloszlásokat, és ebből megállapítjuk, hogy van-e eltérés és az milyen jellegű. Ezt a típusú kapcsolatot asszociációnak nevezzük. Két változó között akkor van asszociáció, ha az egyik értékeinek eloszlása aszerint változik, hogy a másik változó különböző értékeket vesz fel.

Asszociációszámítás feltételezett gyakoriságok használatával

Az asszociációs kapcsolatot a feltételes és a feltétel nélküli megoszlások összehasonlítása révén vizsgáljuk.

Először az f_{ij} tényleges és az f^*_{ij} feltételezett gyakoriságok szembesítése útján végezzük (a két eljárás ekvivalens) az összefüggés-vizsgálatot.

A χ^2 (khí-négyzet) mutató az f_{ij} és f^*_{ij} összehasonlítására szolgáló igen nevezetes mennyiség. A χ^2-próba azt vizsgálja, hogy egy mintán két mért változó megfigyelt értékeinek feltételes gyakoriságai mennyire térnek el a függetlenség esetén várható elméleti gyakoriságoktól, azaz mekkora valószínűséggel fordulnak elő ekkora eltérések.

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

A χ^2 tulajdonságai:

– méri az f_{ij} és f^*_{ij} különbségét,

– az $(f_{ij} - f^*_{ij})^2$ különbség-négyzet f^*_{ij} –vel való osztása révén relatív értéket kapunk,

– érvényesül a $0 \leq \chi^2 \leq N \cdot min\{r - 1, c - 1\}$ egyenlőtlenség, ahol $min\{r–1, c–1\}$ az r (sorok száma) – 1 és c (oszlopok száma) – 1 számok *kisebbikét* jelöli.

Ha a $\chi^2 = 0$, akkor $f_{ij} = f^*_{ij}$ i és j minden értékére, ekkor X és Y független egymástól. A valószínűségszámításból azonban tudjuk, hogy a sztochasztikus összefüggésekre vonatkozó kijelentések csak *bizonyos valószínűséggel* igazak. Kézi számítások esetében mi választunk ki egy vagy több szignifikanciaszintet, és ehhez keressük a megfelelő értéket/értékeket. Általában p = 0,05-öt, azaz 95%-os valószínűségi szintet (vagy ennél kisebb szintet, p = 0,01, p = 0,001 stb.) szokás választani. Annak eldöntésére, hogy a χ^2-értékünk a választott valószínűség mellett szignifikáns összefüggést mutat-e, az úgynevezett *χ^2-eloszlás táblázatát* használjuk. Ebből a táblázatból egy szignifikanciaszintnek és egy szabadságfoknak (df, *degree of freedom*, df = (r–1)×(c–1), azaz „sorok száma mínusz egy szorozva oszlopok száma mínusz egy") egyetlen χ^2-érték olvasható le. Ezt az értéket *küszöbszámnak* tekintjük (jelöljük *k*-val), és ezzel hasonítjuk össze az általunk számított χ^2-értéket. Ha $\chi^2 < k$, akkor *X* és *Y* között nincs szignifikáns kapcsolat a választott szignifikanciaszinten (p = 0,05 esetében 95%-os valószínűséggel állítható). Ugyanakkor nagyon fontos megjegyezni, hogy a küszöbszám alatti értéknél kicsivel kisebb χ^2 inkább azt jelenti, hogy összefüggés van a két változó között, csupán a megvizsgált sokaság kicsi ahhoz, hogy ez a kapcsolat statisztikailag szignifikánsnak látszódjon (ahogyan ezt a *3.2. Elemi mintavételi elmélet, standard hiba* alfejezetben részletesen kifejtettük).

Ha tehát a számított χ^2 érték eléri vagy meghaladja a *küszöbszámot*, akkor elvetjük a nullhipotézist (H_0), amely szerint a vizsgált változók között nincs kapcsolat / a változók függetlenek, és statisztikailag szignifikáns kapcsolatot feltételezünk az X és Y változók között. A 95%-os konfidenciaszint alkalmazása a khí-négyzet próbában azt jelenti, hogy a nullhipotézis igazságát feltételezve legfeljebb 5% annak a valószínűsége, hogy a megfigyelttől legalább ilyen mértékben eltérő gyakorisági eloszlást kapjunk. Ha tehát a p-érték 0,05 alatt van (p < 0,05), a megfigyelt eltérés a függetlenség fennállása mellett túl valószínűtlennek tekinthető, ezért a H_0-t, a függetlenséget elvetjük.

A gyakorlatban a kutatási hipotéziseknél valamilyen összefüggést feltételezünk (H_1) két változó asszociációs kapcsolatáról. Ha a khí-négyzet p értéke p < 0,05, azt mondjuk, hogy a kapcsolat 95%-os valószínűséggel statisztikailag

szignifikáns, azaz a mintában megfigyelt összefüggés valószínűleg nem a véletlen eredménye, és ez statisztikai értelemben alátámasztja a kutatási hipotézist. Ez nem azt jelenti, hogy 95% valószínűséggel igaz a hipotézis, hanem azt, hogy a mintabeli összefüggés nem magyarázható pusztán a véletlennel a választott (legalább) 95%-os valószínűségi szinten.

A χ^2 próbával vizsgált összefüggés erősségét csak viszonylagosan tudjuk megállapítani. Minél nagyobb a χ^2 értéke a neki megfelelő táblázatbeli értéknél, annál erősebb a kapcsolat.

Az adatok számítógépes feldolgozásakor χ^2-eloszlás táblázat használatára nincs szükség, hiszen az SPSS automatikusan kiszámolja az adott értéknek megfelelő szignifikanciaszintet.

33. példa ▼

▶ *A χ^2 kiszámítása*

A χ^2 kiszámítására nézzük az alábbi fiktív példát. A keresztábla egy ezerfős véletlen minta nem és tévénézési szokások szerinti megoszlásának eredményeit tartalmazza (19. táblázat).

19. táblázat. *Tényleges abszolút gyakoriságok: f_{ij} (33. példa)*

Nem/Legtöbbet nézett tévéadó	Férfi	Nő	Összesen
Duna	200	350	550
Acasă	50	200	250
Eurosport	150	50	200
Összesen	**400**	**600**	**1000**

Először dolgozzunk relatív gyakoriságokkal.

Mivel feltételezzük, hogy a nem változó határozza meg a tévénézési szokásokat, és nem fordítva, a nemet tekintjük független változónak, és e szerint százalékolunk (20. táblázat).

20. táblázat. *Tényleges relatív gyakoriságok (33. példa)*

Nem/Legtöbbet nézett tévéadó	Férfi	Nő	Összesen
Duna	50	58,4	55%
Acasă	12,5	33,3	25%
Eurosport	37,5	8,3	20%
Összesen	**100%**	**100%**	**100%**

A 20. táblázatot úgy kaptuk, hogy az egyes cellagyakoriságokat elosztottuk a peremgyakoriságokkal és megszoroztuk százzal. Így a Duna tévét néző férfiak az összes férfi 50%-át jelentik (200 • 100/400 = 50,0%), az Acasă tévét

néző férfiak az összes férfi 12,5%-át (50 • 100/400 = 12,5%), a Duna tévét néző nők az összes nő 58,4%-át (350 • 100/600 = 58,4%), az összes Duna tévét néző a megkérdezettek 55%-át képezik (550 • 100/1000 = 55%) stb.

A soronként számított megoszlási viszonyszámok a két változó közti sztochasztikus kapcsolatot mutatják, hiszen függetlenség esetén a táblázatunk a 21. táblázat képét mutatná, determinisztikus (függvényszerű) kapcsolat esetén pedig a 22. táblázatét vagy egy ehhez hasonlót.

21. táblázat. *Elméleti relatív gyakoriságok függetlenség feltételezése esetén (33. példa)*

Nem/Legtöbbet nézett tévéadó	Férfi	Nő	Összesen
Duna	55%	55%	55%
Acasă	25%	25%	25%
Eurosport	20%	20%	20%
Összesen	**100%**	**100%**	**100%**

22. táblázat. *Elméleti abszolút gyakoriságok determinisztikus kapcsolat feltételezése esetén (33. példa)*

Nem/Legtöbbet nézett tévéadó	Férfi	Nő	Összesen
Duna	0	0	0
Acasă	0	600	600
Eurosport	400	0	400
Összesen	**400**	**600**	**1000**

Abszolút gyakoriságokban kifejezve, függetlenség esetén a kereszttáblánk a 23. táblázat formájában nézne ki.

23. táblázat. *Elméleti abszolút gyakoriságok függetlenség feltételezése esetén: f^*_{ij} (33. példa)*

Nem/Legtöbbet nézett tévéadó	Férfi	Nő	Összesen
Duna	220	330	550
Acasă	100	150	250
Eurosport	80	120	200
Összesen	**400**	**600**	**1000**

A 23. táblázatot az előző, függetlenség esetén várt relatív gyakoriságokat tartalmazó táblázatból kaptuk, úgy, hogy az egyes peremgyakoriságot megszoroztuk a független változó (nem) szerinti relatív gyakoriságokkal és visszaosztottuk 100-zal. Így függetlenség esetén 220 Duna tévét néző férfi (400 • 55/100 = 220), 330 Duna tévét néző nő (600 • 55/100 = 330), 100 Acasă tévét néző

férfi (400 • 25/100 = 100), 150 Acasă tévét néző nő (600 • 25/100 = 150) stb. kellene legyen.

Mivel tehát a tényleges (19. táblázat) és a függetlenség esetén feltételezett (elméleti) abszolút gyakorisági táblázat (23. táblázat) egyértelműen eltér egymástól (elméletileg, ha a tévénézést nem befolyásolná a nem, 100 férfi kellene nézze az Acasă tévét, ezzel szemben az adataink szerint csak 50 férfi nézi stb.), jó okunk van feltételezni, hogy a két változó között van sztochasztikus (statisztikus) kapcsolat.

Másodszor pedig mutassuk ki a kapcsolatot a χ^2 kiszámításával. Ehhez első lépésben kiszámítjuk a két ismérv függetlenségének feltételezése mellett a várható gyakoriságokat (f^*_{ij}).

$$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{N} \qquad f_{11}^* = \frac{400 \cdot 550}{1000} = 220 \qquad f_{12}^* = \frac{400 \cdot 250}{1000} = 100$$

$$f_{21}^* = \frac{600 \cdot 550}{1000} = 330 \qquad f_{22}^* = \frac{600 \cdot 250}{1000} = 150$$

Észrevehető, hogy mind a képlettel, mind a relatív gyakoriságok segítségével ugyanazokat az adatokat kaptuk (23. táblázat).

Ismervén az elméleti gyakoriságokat, a χ^2 képletébe behelyettesítjük őket és a tényleges gyakoriságokat, majd elvégezzük a számításokat.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*} = \frac{(200-220)^2}{220} + \frac{(50-100)^2}{100} + \ldots + \frac{(50-120)^2}{120} = 146{,}8$$

Ilyen módon látható, hogy χ^2 értéke 0-tól különböző, azaz *a két ismérv között valószínűleg van kapcsolat.* A keresztáblákból az is kitűnik, hogy a kapcsolat nem függvényszerű, hanem sztochasztikus.

Nézzük most a Mellékletben szereplő χ^2-eloszlás táblázatot. A szabadságfokunk (sorok száma mínusz egy szorozva oszlopok száma mínusz egy): df = (3–1) • (2–1) = 2, a választott valószínűségi szint 0,05. A χ^2-eloszlás táblázatból idevágó értékek a 24. táblázatban szerepelnek.

24. táblázat. *A szabadságfoknak és szignifikanciaszinteknek megfelelő χ^2 értékek (33. példa)*

Szabadságfok	Szignifikanciaszint		
	p = 0,05	p = 0,01	p = 0,001
2	5,991	9,210	13,815

A táblázatból kiolvashatjuk, hogy a választott szignifikanciaszintnek (95%-os statisztikai biztonság) megfelelő χ^2-érték 5,991. Az általunk számított

érték 146,8, így jóval nagyobb a küszöbértéknél (k), tehát az összefüggés szignifikáns (99,9%-os valószínűségi szint mellett is).

Ezek alapján elmondható, hogy igen jelentősen eltérnek a férfiak és nők tévénézési szokásai. A férfiak négyszer nagyobb arányban nézik a sportadót, mint a nők, akik viszont háromszorosnál nagyobb arányban a sorozatfilmeket sugárzó adót nevezik meg leginkább nézettnek. A Duna tévé kedveltsége nagyon hasonló arányt mutat a két nem esetében, fele, illetve kicsivel több mint fele a megkérdezett férfiaknak és nőknek ezt preferálja a többi adó ellenében.

Az asszociáció mérőszámai

A χ^2 mennyiséget valamilyen alkalmas viszonyítási alaphoz hasonlítva megkapjuk az asszociáció szorosságának különféle χ^2 *alapú mérőszámait.* A leghasználatosabb viszonyítási alap a χ^2 *felső határaként* definiált $N \times min\{r-1,\ c-1\}$ érték, ezt használva az asszociáció *Cramer-féle V asszociációs együtthatóját* kapjuk meg.

$$C^2 = \frac{\chi^2}{N \cdot \min\{r-1, c-1\}} \qquad C = \sqrt{C^2}$$

A C mutatószám 0 és 1 határok között helyezkedik el. C = 0, ha $\chi^2 = 0$, vagyis ha a két változó független, C = 1, ha a kapcsolat determinisztikus.

A gyakorlatban szintén gyakran használt asszociációs együttható a *Csuprov-féle asszociációs együttható.* Ez a mutató az $N \cdot \sqrt{(r-1)\cdot(c-1)}$ viszonyítási alapot használja, ahol a szabadságfok (df) az (r-1)•(c-1) szorzat:

$$T^2 = \frac{\chi^2}{N \cdot \sqrt{(r-1)\cdot(c-1)}} \qquad T = \sqrt{T^2}$$

Ha r ≠ c, akkor a T viszonyítási alapja nagyobb, mint a C viszonyítási alapja, ha r = c, akkor egyenlőek.

A Cramer-féle V és a Csuprov-féle T asszociációs együtthatón kívül még számos más χ^2 alapú asszociációs együttható létezik.

34. példa ▼

▶ χ^2 *alapú asszociációs mutatók kiszámítása*

Visszatérve az előző, 33. példánkhoz (a nem és tévénézési szokások összefüggése), számoljuk ki a C és a T értékeit.

$$C^2 = \frac{\chi^2}{N_{\min}\{(r-1),(c-1)\}} = \frac{146{,}8}{1000 \cdot (2-1)} = 0{,}1468$$

$$C = \sqrt{C^2} = \sqrt{0{,}1468} = 0{,}383$$

$$T^2 = \frac{\chi^2}{N \cdot \sqrt{(r-1)\cdot(c-1)}} = \frac{146{,}8}{1000 \cdot \sqrt{(2-1)\cdot(3-1)}} = \frac{146{,}8}{1000 \cdot 1{,}41} = 0{,}104$$

$$T = \sqrt{T^2} = \sqrt{0{,}104} = 0{,}322$$

Mindkét mutató azt jelzi, hogy a két változó közötti kapcsolat elég laza (közepesnél gyengébb).

Asszociációszámítás PRE (proportionate reduction of error) eljárással

A PRE-eljárás alkalmazásával szintén *többféle asszociációs együttható* képezhető. A továbbiakban az úgynevezett λ *mutatókkal* (lambda) foglalkozunk. A $\lambda_{Y/X}$ mutató azt méri, hogy az *Y* szerinti hovatartozás ismerete *hány százalékkal csökkenti az X szerinti hovatartozás becslésekor elkövetett hibát.*

Ha nem ismerjük az *Y* szerinti hovatartozást, csak az egységek *X* szerinti megoszlását, akkor minden egység *X* szerinti hovatartozását legkézenfekvőbb a *legnagyobb (modális) gyakoriságú X osztállyal becsülni.* Mivel ennek az osztálynak a gyakorisága $\max_j\{f_{.j}\}$, ilyen módon eljárva összesen $N - \max_j\{f_{.j}\}$ számú egység *X* szerinti besorolása esetén tévedünk, azaz hibázunk:

$$E_1 = N - \max_j\{f_{.j}\} \qquad \text{(j szerinti oszlop max. peremeloszlása)}$$

Egy olyan egység *X* szerinti hovatartozását, amelyről tudjuk, hogy *Y* szerint a C^Y_i osztályba tartozik, azzal az *X* osztállyal fogjuk becsülni, amelyre nézve f_{ij} az i-edik sorban j szerint maximális. Ilyen módon a C^Y_i osztályba tartozó egységek *X* szerinti besorolásakor $f_{i.} - \max_j\{f_{ij}\}$ számú esetben fogunk hibázni:

$$E_2 = \sum_i (f_{i.} - \max_j\{f_{ij}\}) = N - \sum_i \max_j\{f_{ij}\}$$

Ezek alapján kiszámítható a PRE mutató:

$$PRE = \frac{E_1 - E_2}{E_1} = \frac{\sum_i \max_j\{f_{ij}\} - \max_j\{f_{.j}\}}{N - \max_j\{f_{.j}\}} = \lambda_{Y|X}$$

Ha PRE = 0, nem feltétlenül függetlenség áll fenn. PRE = 0, ha mind az *Y* szerinti feltételes eloszlások, mind a feltétel nélküli eloszlások modális osztálya megegyezik, de az eloszlások egyébként eltérőek. A PRE vagy lambda (λ) azt mutatja, hogy az egységek *Y* szerinti hovatartozásának ismerete hány százalékkal csökkenti az azok *X* szerinti hovatartozását illető bizonytalanságot, ez az ismeret hogyan javítja az *X* szerinti hovatartozás becsülhetőségét. A fejezet elején használt példánk (*nem* és *önkéntesség* kapcsolata) esetében a λ azt mutatja meg, hogy a nem ismerete hány százalékkal csökkenti az önkéntességgel kapcsolatos bizonytalanságunkat.

35. példa ▼

▶ *A λ kiszámítása*

Térjünk vissza a 33. példánkhoz (a nem és tévénézési szokások összefüggése), és számítsuk ki a λ értékét.

Ha nem tudjuk a nemek szerinti megoszlást, csak azt ismerjük, hogy hányan nézik a különböző tévéadókat, akkor hibázunk a legkevesebbet, ha arra tippelünk, hogy mindenki a Duna tévét nézi, mivel ezt nézik legtöbben.

$$E_1 = N - \max_j \{f_{.j}\} = 1000 - 550 = 450$$

Ismerve a nemek szerinti megoszlást is, minden nőt és minden férfit Duna tévét nézőnek érdemes tippelni:

$$E_2 = \sum_i (f_{i.} - \max_j \{f_{ij}\}) = (400-200)+(600-350) = 450$$

$$E_2 = N - \sum_i \max_j \{f_{ij}\} = 1000 - (200+350) = 450$$

Ezek alapján kiszámítható a λ:

$$PRE = \lambda = \frac{E_1 - E_2}{E_1} = \frac{(450-450)}{450} = 0$$

Tehát a λ értéke 0, mivel mind a nők, mind a férfiak közül legtöbben a Duna tévét nézik, és nem azért, mert a két változó független lenne.

△ *Gyakorlófeladatok a khí-négyzet kiszámítására*

1. Adott az alábbi kereszttábla, amely 134 Sapientia EMTE egyetemi hallgató válaszait tartalmazza szakcsoportonként azzal kapcsolatosan, hogy megszervezi-e tudatosan a napjait vagy nem:

	Mérnök	**Közgazdász**	**TT**	**Humán**	**Összesen**
Megszervezi	28	30	8	3	69
Nem szervezi meg	20	30	10	5	65
Összesen	48	60	18	8	134

2. Adott az alábbi kereszttábla, amely 537 középiskolai tanuló válaszait tartalmazza iskolatípusonként azzal kapcsolatosan, hogy járt-e különórára vagy nem:

	Elméleti líceum	**Technikai szakközépiskola**	**Művészeti és teológiai középiskola**	**Összesen**
Járt különórára	76	55	21	152
Nem járt különórára	134	178	73	385
Összesen	210	233	94	537

Számítsuk ki a χ^2 mutatót mindkét feladatra és értelmezzük!

Asszociáció számítása az SPSS-sel

Ahogyan már a *2.2. Gyakorisági eloszlások* alfejezetben megismertük, kereszttáblákat az *ANALYZE* főmenü *Descriptive Statistics* almenüjében, a *Crosstabs* menüpontnál készíthetünk. A bal oldalon szereplő változók közül kiválasztjuk azt a kettőt (többet is lehet, de minél több dimenziós a kereszttáblánk, annál kevésbé áttekinthető), amelyikre kereszttáblát kérünk. A *Cells* gombnál beállítjuk, hogy sorra vagy oszlopra százalékoljon a program (*Percentages* ablakrész), opcionálisan a *Counts* ablakrészben az elméletileg várt gyakoriságok megjelenítését (*Expected Counts)* is kérhetjük (csak ellenőrzés esetén szükséges), majd *Continue*-t kattintunk.

Visszatérve a főablakba, a *Statistics* gombnál lekérjük a χ^2–et (*Chi-square*) és a *Nominal* ablakrészben feltüntetett asszociációs mutatókat:

- kontingencia együttható (*Contingency coefficient*): 0 és 1 értékek közötti χ^2 alapú mutató (általában nem szükséges lekérni, elég a *Phi és Cramer-féle V*),
- *Phi és Cramer-féle V*: 0 és 1 értékek közötti χ^2 alapú mutató,
- *lambda*: 0 és 1 érték közötti PRE-mutató,
- *bizonytalansági együttható* (*Uncertainty coefficient*): 0 és 1 érték közötti PRE mutató (akkor hasznos, ha a lambda értéke 0).

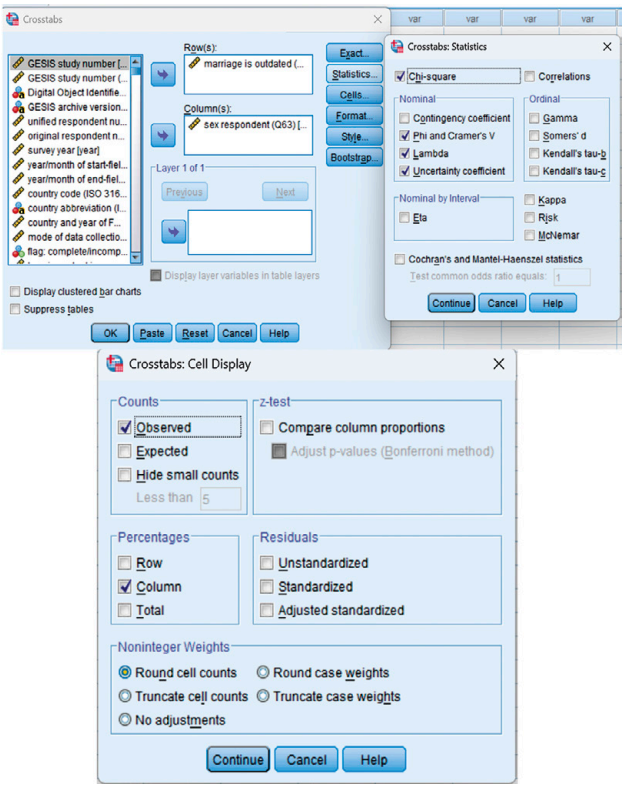
Kereszttáblánkat a *GRAPHS* főmenü alatt szintén a *2.2. Gyakorisági eloszlások* alfejezetben leírt módon ábrázolhatjuk.

36. példa ▼

▶ *χ^2 és asszociációs mutatók az SPSS-ben*

Adatbázisunkban a *v71* változó (*marriage is outdated (Q24)*, K23-as kérdés, az adatbázis 97. sorszámú változója) az arra a kérdésre adott válaszokat rögzíti, hogy a kérdezettek idejétmúlt intézménynek tartják-e a házasságot. Vizsgáljuk meg, hogy van-e összefüggés ezen változó és a *v225* (*a kérdezett neme* változó, K57, az adatbázis 265. sorszámú változója) között. Azt feltételezzük (az a hipotézisünk), hogy a férfiak nagyobb arányban gondolják úgy, hogy a házasság intézménye idejétmúlt, mint a nők.

Első lépésként gyakoriságot kellene kérnünk mindkét változóra, és megtisztítanunk az adatokat a nem releváns válaszoktól (a vizsgált adatbázis esetében ez az alap kezdőlépés nem szükséges, mivel a nem releváns válaszok már a *Missing* tartományban szerepelnek). Kérjünk egy kereszttáblát a *v225* és a *v71* változókra (a *Nem* változó szerint százalékoltassunk), lekérve az χ^2-et és az asszociációs együtthatókat (*Phi and Cramer's V, Lambda, Uncertainty coefficient*), ahogyan a 37. ábra mutatja.



37. ábra. *Asszociációs mutatók lekérése (36. példa)*

A kért statisztikák az *Output* ablakban tekinthetők meg. A 38. ábra a kért kereszttáblát mutatja (a változóneveket és -attribútumokat a könnyebb reprodukálhatóság kedvéért nem fordítottuk magyarra).

marriage is outdated (Q24) * sex respondent (Q63) Crosstabulation

			sex respondent (Q63)		
			male	female	Total
marriage is outdated (Q24)	agree	Count	78	70	148
		% within sex respondent (Q63)	15.3%	12.5%	13.8%
	disagree	Count	431	490	921
		% within sex respondent (Q63)	84.7%	87.5%	86.2%
Total		Count	509	560	1069
		% within sex respondent (Q63)	100.0%	100.0%	100.0%

38. ábra. *A kontingenciatábla (36. példa)*

A kereszttáblára pillantva a relatív gyakoriságok alapján azt látjuk, hogy a férfiak igen kevéssel nagyobb arányban értenek egyet (*agree*) azzal, hogy a házasság egy elavult intézmény, mint a nők. Teszt nélkül azonban nem tudhatjuk, hogy a véletlen mintánkban tapasztalt különbözőség mekkora valószínűséggel állhat elő egy olyan sokaságból, amelyben a férfiak és nők véleménye ebben a kérdésben azonos lenne.

A χ^2 tesztünk (39. ábra) nem mutat szignifikáns összefüggést, hiszen a Pearson-féle χ^2 1,783-as igen alacsony értékének megfelelő szignifikanciaszint, azaz a $p = 0{,}182$ (*Asymp. Sig. (2-sided)*). Ez a 0,182-es p-érték sokkal nagyobb, mint az általánosan elfogadott 95%-os megbízhatósági szintnek megfelelő $p < 0{,}05$-ös érték. Sok esetben a szignifikanciaszint 0,000-ként jelenik meg az SPSS-ben (erős szignifikáns összefüggés), de ez nem 0, nem azt jelenti, hogy abszolút biztos az összefüggés, csupán a program számította szignifikanciaszint kisebb, mint 0,0005, tehát 3 tizedesjegyre kerekítve íródik 0,000-nak. A pontosabb érték elérhető, ha a *Chi-Square Tests* táblázatra duplát kattintunk, és aztán duplát a .000 kijelzésre.

Összességében a példánk alapján azt mondhatjuk, hogy nincs lényeges (szignifikáns) különbség a nők és férfiak egyetértése között abban, hogy a házasság elavult intézmény lenne ($\chi^2 = 1{,}783$, $p < 0{,}05$). Az erdélyi magyar férfiak és nők hasonlóan magas arányban (84,7 – 87,5%) gondolják úgy, hogy a házasság nem egy idejétmúlt intézmény, tehát a hipotézisünk nem igazolódott be.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.783[a]	1	.182		
Continuity Correction[b]	1.554	1	.213		
Likelihood Ratio	1.781	1	.182		
Fisher's Exact Test				.185	.106
Linear-by-Linear Association	1.781	1	.182		
N of Valid Cases	1069				

a) 0 cells (0.0%) have expected count less than 5. The minimum expected count is 70.47.
b) Computed only for a 2x2 table

39. ábra. *A χ^2 statisztika (36. példa)*

Mivel a két változó között nincs statisztikailag szignifikáns összefüggés (a függetlenségre vonatkozó nullhipotézis nem vethető el), a többi asszociációs mutató értéke sem az. A lambda értéke 0,000 – mivel a példánkban a *marriage is outdated (Q24)* a függő (*Dependent*) változó, a 40. ábrában szereplő táblázat 2. értéksorát kell kiolvasnunk. Tehát a nem ismerete nem csökkenti (0%-kal csökkenti) a házasság elavult intézményként való értékelésével kapcsolatos

bizonytalanságot. Amennyiben a lambda egy statisztikailag szignifikáns 0 érték lenne (a 2. sor negyedik oszlopában a $p < 0{,}05$), a táblázatból a bizonytalansági együtthatót (*Uncertainty Coefficient*, másik PRE mutató) olvasnánk ki hasonló módon.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.012	.018	.658	.511
		marriage is outdated (Q24) Dependent	.000	.000	.[c]	.[c]
		sex respondent (Q63) Dependent	.016	.024	.658	.511
	Goodman and Kruskal tau	marriage is outdated (Q24) Dependent	.002	.002		.182[d]
		sex respondent (Q63) Dependent	.002	.002		.182[d]
	Uncertainty Coefficient	Symmetric	.002	.002	.667	.182[e]
		marriage is outdated (Q24) Dependent	.002	.003	.667	.182[e]
		sex respondent (Q63) Dependent	.001	.002	.667	.182[e]

a) Not assuming the null hypothesis.
b) Using the asymptotic standard error assuming the null hypothesis.
c) Cannot be computed because the asymptotic standard error equals zero.
d) Based on chi-square approximation
e) Likelihood ratio chi-square probability.

40. ábra. *A lambda és a bizonytalansági együttható (36. példa)*

Végül a Cramer-féle asszociációs együttható értéke 0,041, ami egy nagyon minimális erősségű kapcsolatot jelez (41. ábra). Ez az érték tehát nagyon közel van a függetlenséget jelölő 0 értékhez és nem szignifikáns ($p = 0{,}182$, vagyis nagyobb, mint a 0,05-ös küszöbérték).

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.041	.182
	Cramer's V	.041	.182
N of Valid Cases		1069	

41. ábra. *A χ^2 alapú asszociációs mutatók (36. példa)*

Összességében tehát azt mondhatjuk, hogy a nem és a házasság elavult intézményként való megítélése között nincs szignifikáns összefüggés ($p = 0{,}182$ vagy $p > 0{,}05$): a férfiak hasonlóan alacsony arányban tartják a házasságot elavult intézménynek (15,3%), mint a nők (12,5%).

△ *Gyakorlófeladatok χ^2 és asszociációs mutatók SPSS-ben való lekérésére*

Kérjünk χ^2 és asszociációs mutatókat a korábban létrehozott *korcsoport3kat* és a már ismert *v225* (*a kérdezett neme*) független változók és a *v57–v60* (a kérdőív K17-es kérdése, az adatbázisban a 83–86. sorszámú, *a hit különböző formáira* vonatkozó 4 változó) függő változók között! Értelmezzük a kapott adatokat (8 asszociációs vizsgálat) az elemzett változók kontextusában!

Két ordinális mérési szintű változó közötti kapcsolat

Arra az esetre vonatkozik, amikor *mindkét változó sorrendi* (ordinális) skálán mérhető. A továbbiakban a kapcsolat szorosságának mérésére használható leggyakrabban alkalmazott mutatóval, a gamma (γ) mérőszámmal foglalkozunk.

Akár csak a lambda, a gamma is azon alapul, hogy mennyire segíti az egyik változó szerinti hovatartozás ismerete a másik értékének becslését. Ilyen módon szintén a PRE-eljárás alapján dolgozunk.

Tudjuk, hogy az ordinális mérési szintű változók értékeinek csak a sorrendje jelent valamilyen információt, ezért nem a leggyakoribb értékre, hanem az értékek ordinális elrendezésére, sorrendjére tippelünk. Minden egyes esetpárnál azt tippeljük, hogy a két eset elrendezése az egyik változó szerint megfelel (pozitívan vagy negatívan) a másik változó szerinti elrendezésnek: az egyik változó szerint „nagyobb" eset a másik változó szerint is mindig „nagyobb", vagy pedig a másik változó szerint mindig „kisebb".

A gamma kiszámításánál két mennyiséget kell ismerni:

- azon esetpárok számát, amelyeknél egyforma a két változó szerinti nagyságviszony,
- azon esetpárok számát, ahol az egyik változó szerint az egyik eset a nagyobb, a másik változó szerint a másik eset a nagyobb.

Az egyező nagyságrendű számpárok kiszámítása: mindegyik cellában az elemek számát megszorozzuk az alatta és ugyanakkor tőle jobbra fekvő cellákban lévő elemek számának összegével, majd összeadjuk ezeket a szorzatokat.

Az ellentétes nagyságviszonyú számpárok kiszámítása: a keresztábla mindegyik cellájában az elemek számát megszorozzuk az alatta és egyben tőle balra fekvő cellákban lévő elemek számának összegével, majd összeadjuk a szorzatokat.

A gammát az egyező és az ellentétes rendezésű párok számából számítjuk ki:

$$\gamma = \frac{N_{egyezo} - N_{ellentetes}}{N_{egyezo} + N_{ellentetes}}$$

A γ értéke mindig –1 és 1 között van, így a kapcsolat szorosságán kívül annak irányát is megadja. Ha a gamma pozitív, akkor az egyik változó szerint „nagyobb" eset a másik változó szerint is „nagyobb" (konkordáns), ha a gamma negatív, akkor az egyik változó szerint „nagyobb" eset a másik változó szerint „kisebb" (diszkordáns).

37. példa ▼

▶ *A gamma mutató kézi számítása*

A 25. táblázat a saját munkaerőpiaci helyzet megítélését jelzi iskolai végzettség szerinti bontásban (fiktív adatok).

25. táblázat. *A saját munkaerőpiaci helyzet megítélése iskolai végzettségi szintek szerint (37. példa)*

Végzettség/Munkaerőpiaci helyzet	Rossz	Közepes	Jó	Összesen
Alapfokú	200	50	50	**300**
Középfokú	50	400	150	**600**
Felsőfokú	10	20	70	**100**
Összesen	**260**	**470**	**270**	**1000**

Számoljuk ki a γ értékét.

$$N_{egyezo} = 200 \cdot (400 + 150 + 20 + 70) + 50 \cdot (20 + 70) + 50 \cdot (150 + 70) + 400 \cdot (70) = 171\,500$$

$$N_{ellentetes} = 80 \cdot (400 + 80 + 10 + 20) + 150 \cdot (10 + 20) + 50 \cdot (50 + 10) + 400 \cdot (10) = 52\,300$$

$$\gamma = \frac{N_{egyezo} - N_{ellentetes}}{N_{egyezo} + N_{ellentetes}} = \frac{171500 - 52300}{171500 + 52300} = \frac{119200}{223800} = 0{,}532$$

A γ értéke egy közepes erősségű, pozitív kapcsolatot mutat a két változó között: a magasabb iskolai végzettségű személyek elégedettebbek a munkaerőpiaci helyzetükkel, míg az alacsony iskolai végzettségűek elégedetlenebbek saját munkaerőpiaci helyzetükkel.

Két ordinális változó kapcsolatának vizsgálata az SPSS-sel

Akárcsak az asszociációs együtthatókat, a gammát is az *ANALYZE* főmenü *Descriptive Statistics* almenüjében, a *Crosstabs* menüpontnál kérhetjük le. A bal oldalon szereplő változók közül kiválasztjuk azt a kettőt, amelyekre az összefüggést szeretnénk vizsgálni. A *Statistics* ablakban lekérjük az *Ordinal* ablakrészben feltüntetett mutatókat vagy azok egyikét, leggyakrabban az általánosan használt gammát:

– *gamma:* -1 és 1 értékek közé eső mutató, minden kereszttáblaméret esetén használható),
– *Somer's d:* a gamma kiterjesztése (az elemzésbe bevonja a független változóhoz nem kötődő esetpárokat is), értéke -1 és 1 közé esik,
– *Kendall's tau-b*: -1 és 1 értékek közé eső mutató, figyelembe veszi a kötődéseket, és csak szimmetrikus (azonos számú sorok és oszlopok) kereszttábláknál alkalmazható,
– *Kendall's tau-c:* -1 és 1 értékek közé eső mutató, nem veszi figyelembe a kötődéseket, és nem szimmetrikus kereszttábláknál használatos.

Visszatérve a főmenübe bejelöljük, hogy az eredménykijelző fájlban (*Output*) ne jelenjen meg a kereszttábla, mivel az összefüggés értelmezésére elég a gamma mutató. Ezt a változók alatt található *Suppress tables* opció bejelölésével tehetjük meg.

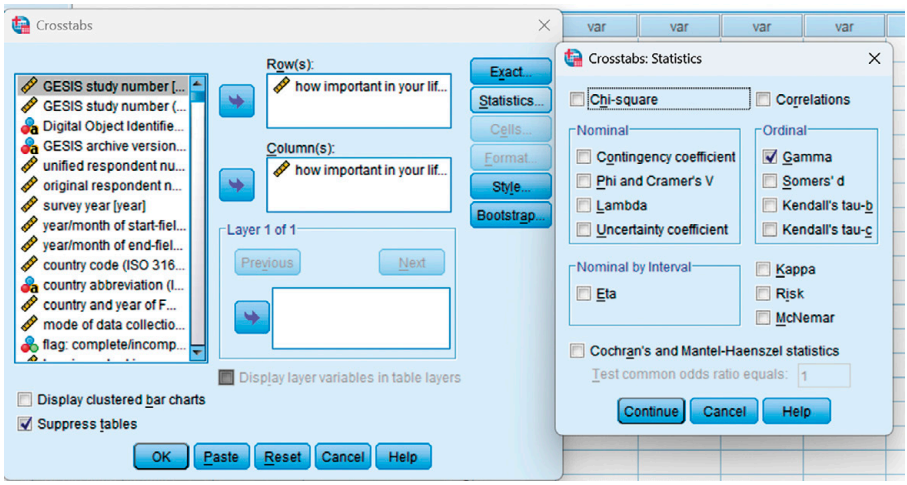
38. példa ▼

▶ *A gamma lekérése az SPSS-ben*

Az adatbázisban a *v1* és *v2* változók (az adatbázis 15. és 16. sorszámú változói) a kérdőív első két kérdésére (K1 első két sora) adott válaszokat rögzíti és arra vonatkozik, hogy mennyire fontos érték a kérdezettek életében a munka (*how important in your life: work [Q1A]*) és a család (*how important in your life: family [Q1B]*). Azt feltételezzük, hogy a két érték megítélése konkordáns viszonyban van: minél fontosabb a család, annál fontosabb a munka is (egy pozitív, szignifikáns gamma értéket várunk).

Vizsgáljuk meg a két ordinális mérési szintű változó közötti kapcsolatot.

Első lépésként itt is gyakoriságot kellene kérnünk mindkét változóra, és megtisztítanunk az adatokat a nem releváns válaszoktól (ebben az adatbázisban ez nem szükséges). Tehát az előzőekben leírtak szerint lekérjük a gammát (42. ábra) és bejelöljük a *Suppress tables* opciót. A két változónk esetében most teljesen mindegy, hogy melyiket tesszük sorba vagy oszlopba, hiszen nem tudjuk eldönteni, hogy melyik a függő és melyik a független változónk.



42. ábra. *A gamma mutató lekérése az SPSS-ben (38. példa)*

A kért gamma mutató értéke az *Output*-ban a 43. ábrán látható táblázat formájában jelenik meg.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.592	.058	6.895	.000
N of Valid Cases		1096			

a) Not assuming the null hypothesis.

b) Using the asymptotic standard error assuming the null hypothesis.

43. ábra. *A gamma értéke és szignifikanciaszintje (38. példa)*

A γ értéke 0,592 és az összefüggés szignifikáns ($p = 0{,}000$), tehát a két változó között egy statisztikailag szignifikáns, jó közepes erősségű, pozitív irányú kapcsolat van. Értelmezéskor vegyük figyelembe, hogy mindkét változó esetében az 1-es kód a nagyon fontos, a 4-es pedig az egyáltalán nem fontos opciót jelölte (előfordulhat, hogy ha a skálák terjedelme azonos is, az egyes értékek más opciókat jelöltek a két változó esetében). A gamma mutató értelme tehát, hogy minél fontosabb értéknek tartják a megkérdezettek a munkát, annál fontosabb értéknek tekintik a családot is ($\gamma = 0{,}592$, $p = 0{,}000$). Ebben az esetben a hipotézisünk beigazolódott.

△ *Gyakorlófeladatok a gamma SPSS-ben való lekérésére*

Kérjünk páronkénti gamma mutatókat a v32–v37 változók (a kérdőív K8-as kérdése, az adatbázisban a 53–58-as sorszámú, *a különböző csoportok-*

hoz tartozó emberekbe vetett bizalomra vonatkozó változók) között! Értelmezzük a kapott adatokat az elemzett változók kontextusában!

4.3. Vegyes kapcsolat

A vegyes kapcsolatot egy nominális és egy intervallum- vagy arányskálán mért változó közötti kapcsolat vizsgálatára használjuk. A kapcsolat szorosságának mérésére a PRE-eljárást használjuk fel.

Az egyszerre vizsgált két változó közül a mennyiségi ismérvet jelöljük *Y*-nal, a nominálisat *X*-szel. Ha az *Y* megoszláson kívül nem áll rendelkezésünkre semmilyen információ, akkor a sokaság valamelyik (pl. g-edik) egységének *Y* szerinti hovatartozását (Y_g-t) legcélszerűbb a feltétel nélküli *Y* megoszlás átlagával, $\overline{Y}$-nal becsülni. Ha az átlaggal becsüljük az Y_g-t, az ezzel összességében elkövetett hiba kisebb, mint ha bármely más értéket használnánk erre a célra:

$$E_1 = \sum_{g=1}^{N} (Y_g - \overline{Y})^2$$

Ha valamely egységről ismertté válik, hogy az *X* ismérv szerint a C^x_i osztályba tartozik, akkor az *Y* ismérv annál előforduló értékét az előbbieknek megfelelően a C^x_i osztályba tartozó egységek átlagos *Y* értékével, -gal (részátlaggal) célszerű becsülni:

$$E_2 = \sum_{i=1}^{r} \sum_{j=1}^{f_{i\cdot}} (Y_{ij} - \overline{Y}_i)^2$$

ahol Y_{ij} – a C^x_i osztály j-edik egyedének *Y* értéke.

Tehát a PRE mutató a következő lesz:

$$PRE = \frac{E_1 - E_2}{E_1} = \frac{\sum_{g=1}^{N} (Y_g - \overline{Y})^2 - \sum_{i=1}^{r} \sum_{j=1}^{f_{\cdot j}} (Y_{ij} - \overline{Y}_i)^2}{\sum_{g=1}^{N} (Y_g - \overline{Y})^2} = 1 - \frac{\sigma_B^2}{\sigma^2} = H^2$$

ahol:

H^2 – *varianciahányados*,

σ^2 – a sokaság szórásnégyzete, teljes varianciája,

σ^2_B – belső variancia (a fősokaság Y_{ij} értékei átlagosan mennyivel térnek el saját részátlaguktól).

A H^2 megadja, hogy az egységek *X* szerinti hovatartozásának ismerete hogyan javítja az *Y* szerinti hovatartozás becsülhetőségét, vagyis az *Y* ismérv szórásnégyzetének az *X* ismérv által megmagyarázott hányadát. A H^2 egy 0 és 1 közötti

érték: $0 \le H^2 \le 1$. Ha $H^2 = 0$, X és Y független (az X szerint képzett részátlagok mind egyformák), a feltételes és a feltétel nélküli gyakorisági eloszlások mind egyformák. Ha $H^2 = 1$, X és Y függvényszerű, determinisztikus kapcsolatban áll egymással (az X szerinti csoportokon belül Y nem szóródik), az X szerinti hovatartozás mindent elmond Y-ról.

A gyakorlatban szokták használni a $H = \sqrt{H^2}$ mutatót is, ez a szórásháyados. A H szintén 0 és 1 között mozgó érték. Ha H = 0, függetlenség áll fenn, ha H = 1, a két változó között függvényszerű kapcsolat van. H esetén kizárólag a 0-hoz, illetve 1-hez való közelségre alapozható a kapcsolat szorosságának megítélése, nem használható megoszlási viszonyszámként.

39. példa ▼

▶ *A varianciahányados kiszámítása*

Nézzük az alábbi szemléltető példát. A 26. táblázatban szereplő fiktív adatok egyedülálló nők (8) és férfiak (7) keresetét jelölik (1000 lejben).

26. táblázat. *15 személy jövedelme nemek szerinti bontásban (39. példa)*

Nem (X)	Jövedelem (1000 lejben, Y)	N	Σ
1. Férfi	1; 2; 2; 3; 5; 10; 12	7	35
2. Nő	1; 1; 1; 2; 2; 2; 3; 4	8	16
Összesen		**15**	**51**

Első lépésben kiszámoljuk a teljes sokaság átlagát, majd a férfiak és a nők jövedelmeinek átlagát (a részátlagokat).

$$\overline{Y} = \frac{1\cdot 1+2\cdot 2+1\cdot 3+1\cdot 5+1\cdot 10+1\cdot 12+3\cdot 1+3\cdot 2+1\cdot 3+1\cdot 4}{15} = \frac{51}{15} = 3{,}4$$

Most kiszámoljuk, hogy mekkora hibát követnénk el, ha nem ismernénk a jövedelmek nemek szerinti megoszlását (akkor tévednénk a legkevesebbet, ha a sokaság átlagával helyettesítenénk):

$$E_1 = \sum_{i=g}^{N} (Y_g - \overline{Y})^2 = 4\cdot(1-3{,}4)^2 + 5\cdot(2-3{,}4)^2 + 2\cdot(3-3{,}4)^2 +$$
$$+ (4-3{,}4)^2 + ... + (12-3{,}4)^2 = 153{,}96$$

Harmadik lépésben kiszámoljuk a férfiak és a nők jövedelmeinek átlagát (a részátlagokat):

$$\overline{Y}_1 = \frac{1\cdot 1+2\cdot 2+1\cdot 3+1\cdot 5+1\cdot 10+1\cdot 12}{7} = \frac{35}{7} = 5$$

$$\overline{Y}_2 = \frac{3\cdot 1+3\cdot 2+1\cdot 3+1\cdot 4}{8} = \frac{16}{5} = 2$$

Most, mivel ebben a lépésben már ismerjük a nemek szerinti jövedelemeloszlásokat is, kiszámítjuk mindkét részsokaságra, hogy mekkora hibát követnénk el, ha a részátlagokkal becsülnénk meg adatainkat:

$$E_2 = \sum_{i=1}^{r}\sum_{j=1}^{f_{i\cdot}}(Y_{ij} - \overline{Y}_i)^2 = (1-5)^2 + 2\cdot(2-5)^2 + (3-5)^2 + (5-5)^2 + (10-5)^2 + (12-5)^2 +$$
$$+ 3\cdot(1-2)^2 + 2\cdot(2-2)^2 + (3-2)^2 + (4-2)^2 = 120$$

Ezek után kiszámítható a varianciahányados:

$$H^2 = \frac{E_1 - E_2}{E_1} = \frac{153{,}96 - 120}{153{,}96} = 0{,}22$$

$$H \approx 0{,}47$$

Értelmezés szerint a két változó között közepes erősségű kapcsolat van ($H \approx 0{,}47$). A nem ismerete 22%-át magyarázza meg a jövedelmek szórásnégyzetének, vagyis a nem ismerete 22%-kal csökkenti a jövedelmek ismeretével kapcsolatos bizonytalanságot.

A t-teszt

A lényegesebb kapcsolatvizsgálat akkor kezdődik el, amikor nem ismerjük a sokaságbeli eloszlást, és arra a kérdésre keressük a választ, hogy a mintánk *két részsokaságában* az átlagok között tapasztalható eltérés annak tudható-e be, hogy az alsokaságokban is megvan a különbözőség, vagy a kimutatott különbség csak a véletlen műve. A fenti 39. példánk esetében azt akarjuk megtudni, hogy a nők és férfiak között kimutatott jövedelemkülönbség csak onnan adódik-e, hogy pont ezt a 15 embert kérdeztük meg, vagy a nők és férfiak körében ténylegesen létezik ez a különbség. A t-teszttel tehát arra kapunk választ, hogy a mintavétel során fellépő véletlen tényező mekkora valószínűséggel okoz különbözőségeket.

A t-eloszlás arra alapoz, hogy $n > 30$ elemszám vagy *egymástól szignifikánsan eltérő szórások* esetén, feltételezve, hogy a *kétértékű kategoriális változónknál* az átlagértékek a teljes sokaságban egyformák (a mintánkban kimutatható különbség csak a véletlen műve), a két mintaátlag különbsége normális eloszlást követ 0 várható értékkel és $\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$ mintaszórással (s^2 – mintavariancia helyett a korábbiaknak megfelelően σ^2-val jelölöm).

Tehát

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}}$$

A t-teszt esetében nem a t értéke, hanem a neki megfelelő *szignifikanciaszint* érdekel bennünket. Ha a t-értéknek megfelelő szignifikanciaszint kisebb,

mint 0,05 ($p < 0,05$), akkor 95%-os biztonsággal állíthatjuk, hogy a mintánkon (a megfigyelt adatainkon) számolt csoportátlagok közötti eltérések nem a véletlen művei.

Ha a mintánkon számolt *két részátlag szórása nem különbözik szignifikánsan* ($p > 0,05$), vagy kicsi a mintanagyságunk ($n < 30$), akkor a

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1 - 1)\cdot \sigma_1^{\,2} + (n_2 - 1)\cdot \sigma^{2}{}_2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

képlettel számolunk. Ebben az esetben is nem a *t* értéke, hanem a neki megfelelő szignifikanciaszint a fontos.

Kézi számításokkor a *t* értékét az úgynevezett *t-táblázat* vagy a *3. Mintavétel* című fejezetben már megadott t-értékek segítségével (17. táblázat) értékeljük. Tehát ha $n > 120$, szignifikáns összefüggés esetén a t értéke nagyobb vagy egyenlő kell legyen, mint 1,96.

Ha $n < 120$, a t értékét a *t-táblázat* (lásd a Mellékletben) segítségével értékeljük, és hasonlóan járunk el a khí-négyzet esetében leírtaknál: ha t értéke *kisebb*, mint a megfelelő szabadságfokoknál és valószínűségi szintnél szereplő táblázati érték, akkor a két változó között nincs szignifikáns kapcsolat a választott szignifikanciaszint mellett (nagyobb a valószínűsége annak, hogy az összefüggés a véletlen műve). Kétmintás t-próba esetén a t szabadságfoka:

$$df = n_1 + n_2 - 2$$

Az F-próba

Az F-próba azt mutatja meg, hogy két vagy több részmintában a szórások közti különbség mennyire a véletlen műve, és mennyire annak tudható be, hogy különbözik a populáció alsokaságaiban is. Dichotóm változók esetében az F értéke a két részsokaság szórásnégyzetének hányadosa:

$$F = \frac{\sigma_1^{\,2}}{\sigma_2^{\,2}}$$

Tehát két vagy több átlagértéket is össze lehetne hasonlítani F-próbával, de a kissé hosszadalmasabb számítással kapott F-érték éppen a *t* négyzete, és mindkettő ugyanazt a szignifikanciaszintet eredményezi (akárcsak a *khí-négyzet* vagy a *t* értéke esetében, itt is nem az *F* értéke, hanem a neki megfelelő szignifikanciaszint bír jelentőséggel). Ilyen módon kézi számításnál előnyösebb a *t* képletével számolni (ezért is alkalmazták gyakrabban). A számítógépes program gyakorlatilag ugyanannyi idő alatt szolgáltatja az eredményeket.

Kézi számításokkor az F értékét az úgynevezett *F-táblázat* segítségével értékeljük, és hasonlóan járunk el a khí-négyzet esetében részletesen leírtaknál (ha

F értéke kisebb, mint a megfelelő szabadságfokoknál szereplő táblázati érték, akkor a választott valószínűségi szint mellett az összefüggés nem szignifikáns).

Ha a magyarázó változó két kategóriából áll, akkor az átlagok közötti különbség vizsgálatára a t-próba és az F-próba egyaránt megfelelő, mivel ebben az esetben a két eljárás matematikailag ekvivalens. A t-próba közvetlen módon hasonlítja össze a két csoport átlagát, míg az F-próba egy általánosabb keret részeként teszi ugyanezt. Két kategória esetén az F-statisztika a t-statisztika négyzetével egyenlő ($F = t^2$), ezért mindkét módszer ugyanarra az eredményre és szignifikanciaszintre vezet. A gyakorlatban a t-próba az áttekinthetőbb és célszerűbb választás, míg az F-próba elsősorban akkor válik relevánssá, amikor a vizsgálat összetettebb elemzési keretbe illeszkedik (pl. ha több kategóriát tartalmazó csoportot akarunk összehasonlítani, többszempontos ANOVA-t szeretnénk alkalmazni stb.).

A t-próba és az F-próba (ANOVA) alkalmazhatósági feltételei

A paraméteres hipotézisvizsgálatok, mint a t-próba és az F-próba (ANOVA), a statisztikai következtetésalkotás alapvető eszközei. Azonban ahogyan ezt már korábban jeleztük, a módszerek érvényes alkalmazása csak akkor biztosítható, ha a vizsgált adatok megfelelnek bizonyos elméleti előfeltételeknek. A paraméteres átlag-összehasonlító eljárások közös alapját két kulcsfontosságú feltétel képezi:

1. a függő változó közel normális eloszlása a csoportokon belül, valamint
2. a csoportok varianciáinak homogenitása.

1. A függő változó közel normális eloszlása a csoportokon belül

A t-próba és az F-próba (ANOVA) egyaránt azon az alapfeltevésen alapul, hogy a függő változó eloszlása az összehasonlítani kívánt csoportokon belül legalább megközelítően normális. Ez a feltétel nem pusztán formai követelmény, hanem a hipotézisvizsgálati eljárások matematikai érvényességének egyik alapja, mivel a t- és F-statisztikák elméleti tulajdonságai a normális eloszlás sajátosságaira épülnek.

A normalitás vizsgálata történhet formális statisztikai próbákkal, illetve grafikus eszközökkel. A legelterjedtebb formális eljárás a *Shapiro-Wilk-próba*, amely kis és közepes mintanagyság ($n < 200$) esetén nagy érzékenységgel képes kimutatni a normálistól való eltéréseket. A *Kolmogorov–Smirnov-próba* a minta empirikus eloszlását hasonlítja össze a normális eloszlással, és a gyakorlatban Lilliefors-korrekcióval használják. Inkább nagyobb minták esetén alkalmazható, de kevésbé alkalmas a normalitás vizsgálatára, mert nagy mintákban rendkívül érzékenyen reagál, és már egészen apró, gyakorlati szempontból jelentéktelen eltéréseket is szignifikánsnak mutathat. Ilyenkor a teszt azt jelzi, hogy az eloszlás nem normális, holott az eltérés valójában nem befolyásolná érdemben az elemzést. Mindkét próba *p* értéke arról ad visszajelzést, hogy a minta szignifikánsan

eltér-e a normális eloszlástól. A formális teszteket továbbá célszerű a ferdeségi és csúcsossági mutatók vizsgálatával, illetve grafikus eszközökkel is kiegészíteni, így például *hisztogrammal, boxplottal és Q-Q plottal*, amelyek vizuálisan is megmutatják az eloszlás alakját, aszimmetriáját és a kiugró értékeket.

A t-próba elméleti alapját a *Student-féle t-eloszlás* (a kis mintákból számított átlag szórásának bizonytalanságát leíró eloszlás, amely a normális eloszláshoz hasonlít, de annál kissé szélesebb) adja, amely normális eloszlású változók mintavételi tulajdonságaiból vezethető le. Ugyanakkor nem szükséges, hogy a teljes sokaság maga normális legyen. A centrális határeloszlás tételének következtében már néhány elemből álló minta átlaga is olyan valószínűségi eloszlást követ, amely közel áll a normálishoz, és a t-eloszlás csak csekély mértékben tér el attól. A mintanagyság növekedésével a t-eloszlás fokozatosan közelít a normálishoz, száz fő körül pedig a két eloszlás gyakorlatilag egybeesik. Mindezek ellenére a t-próba kis minták esetén érzékenyen reagál a normalitástól való eltérésekre. Erősen ferde eloszlás, laposság vagy éppen csúcsosság, illetve kiugró értékek torzíthatják a szignifikanciát. Ilyen helyzetekben ajánlott nemparaméteres alternatívák például a *Mann-Whitney U-próba* vagy a *Wilcoxon-próba* használata. Nagyobb minták esetén azonban a t-próba már jelentősen robusztus, és mérsékelt eloszlási torzulások mellett is megbízhatóan működik.

Az F-próba normalitási feltételei hasonló módon értelmezhetők. Az ANOVA keretében használt F-statisztika elméleti eloszlása akkor érvényes, ha a függő változó eloszlása minden vizsgált csoportban közel normális. A klasszikus F-próba különösen érzékeny a csoportonkénti normalitás súlyos megsértésére: ha valamelyik csoport erősen aszimmetrikus vagy sok kiugró értéket tartalmaz, akkor az F-statisztika eloszlása torzulhat, ami a szignifikanciaszint megbízhatatlanságát eredményezi. Kis minták esetén ezért célszerű az ANOVA helyett nemparaméteres alternatívát alkalmazni, például a *Kruskal–Wallis*-próbát. Nagy mintanagyság esetén ugyanakkor az F-próba a centrális határeloszlás tételének köszönhetően mérsékelten robusztusnak tekinthető, így a normalitástól való közepes mértékű eltérések még nem teszik érvénytelenné az eredményt, feltéve, hogy nincsenek szélsőséges ferdeségek vagy kiugró értékek.

A társadalomtudományi kutatásokban a normalitási feltétel gyakran sérül, mivel sok társadalmi jelenség természetes módon nem követi a normális eloszlást. A jövedelem eloszlása rendszerint jobbra ferde, a politikai érdeklődés pedig torz, hiszen kevés válaszoló ad nagyon magas értéket. Ezek a sajátosságok különösen kis minták esetén torzíthatják a paraméteres próbák szignifikanciáját. Nagy minták esetén a centrális határeloszlás tétele enyhíti ezeket a problémákat, de a normalitás vizsgálata ilyenkor is indokolt.

2. A csoportok varianciáinak homogenitása

A t-próba és az F-próba (ANOVA) másik alapvető alkalmazhatósági feltétele, hogy az összehasonlítani kívánt csoportok varianciái ne különbözzenek lényege-

sen egymástól. Ezt a követelményt varianciahomogenitásnak nevezzük. A paraméteres próbák matematikai modellje abból indul ki, hogy a csoportokon belüli szóródások nagysága azonos vagy legalábbis közel azonos, mivel a tesztstatisztikák kiszámítása és az elméleti eloszlásokhoz való illeszkedése ezt feltételezi. Ha a varianciák jelentősen eltérnek, akkor a tesztek hibavalószínűsége torzulhat, és a szignifikanciaszint megbízhatatlanná válhat.

A varianciahomogenitást a gyakorlatban leggyakrabban a *Levene-próbával* vizsgáljuk. A Levene-próba azt teszteli, hogy a csoportok varianciái szignifikánsan különböznek-e egymástól. Ha a próba nem szignifikáns ($p > 0{,}05$), akkor a varianciák egyenlőnek tekinthetők, és a klasszikus t-próba vagy ANOVA használata megfelelő. Ha azonban a Levene-próba szignifikáns ($p \leq 0{,}05$), akkor a varianciák eltérnek, és a hagyományos eljárások eredménye torzulhat. Ilyen esetben célszerű robusztus alternatívát alkalmazni. A t-próbánál ez a *Welch-féle korrekció*, amely nem feltételezi a varianciák egyenlőségét. Az ANOVA esetében hasonló célt szolgál a *Welch ANOVA*, amely varianciahomogenitás hiányában is megbízható szignifikanciabecslést ad. A varianciák eltérése gyakran a csoportok természetes különbségeiből adódik, ezért diagnosztikai szempontból is fontos információt hordoz.

A társadalomtudományi kutatásokban a varianciahomogenitás sérülése gyakori jelenség. Bizonyos társadalmi csoportokban természetes módon nagyobb a szóródás, például az életkor szerinti csoportok fizikai aktivitásában, vagy az iskolai végzettség szerinti csoportok kulturális fogyasztásában. Ezért a varianciahomogenitás vizsgálata minden paraméteres elemzés esetén elengedhetetlen lépés, amely biztosítja, hogy a következtetések statisztikailag érvényesek és megbízhatóak legyenek.

40. példa ▼

▶ *A t értékének kézi számítása*

Nézzük a nők és férfiak kereseteit tartalmazó korábbi feladatunkat (39. példa), és számítsuk ki a t értékét. A kis elemszámra való tekintettel a normalitást az SPSS-ben szigorú módszerrel, *Shapiro-Wilk-teszttel* ellenőriztük (*ANALYZE* főmenü, *Descriptive Statistics*, *Explore*, *Plots*/*Normality plots with tests*) és a normalitás feltétele teljesült ($p > 0{,}085$), tehát a t-teszt alkalmazható.

A t-teszt kézi számításakor először ki kell számítanunk a két alcsoportunk szórását (a csoportátlagokat már kiszámoltuk).

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i \cdot (X_i - \bar{X})^2}$$

$$\sigma_1 = \sqrt{\frac{1}{7}\cdot\left[1\cdot(1-5)^2 + 2\cdot(2-5)^2 + 1\cdot(3-5)^2 + 1\cdot(5-5)^2 + 1\cdot(10-5)^2 + 1\cdot(12-5)^2\right]} = \sqrt{\frac{112}{7}} = 4$$

$$\sigma_2 = \sqrt{\frac{1}{8}\cdot\left[3\cdot(1-2)^2 + 3\cdot(2-2)^2 + 1\cdot(3-2)^2 + 1\cdot(4-2)^2\right]} = \sqrt{\frac{8}{8}} = 1$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)\cdot\sigma_1^{\,2} + (n_2-1)\cdot\sigma^2{}_2}{n_1+n_2-2}}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{5-2}{\sqrt{\frac{(7-1)\cdot 16 + (8-1)\cdot 1}{7+8-2}}\cdot\sqrt{\frac{1}{7}+\frac{1}{8}}} = 2{,}09$$

Kikeressük a t-eloszlás táblázatból az értéket, ha df = 7 + 8 – 2 = 13 (27. táblázat).

27. táblázat. *A t-eloszlás táblázatból idevágó értékek (40. példa)*

Szabadságfok	Szignifikanciaszint		
	p = 0,05	**p = 0,01**	**p = 0,001**
13	2,160	3,012	4,221

Összevetve értékünket (2,09) a küszöbértékekkel látjuk, hogy a két változó közötti mintánkon megfigyelt összefüggés 95,0%-os valószínűségi szint mellett sem szignifikáns. Mivel az értékek (számított és küszöbérték) közötti eltérés nagyon kicsi, azt mondhatjuk, hogy az alapsokaságra, vagyis a településen élő felsővezetők körére is elég nagy valószínűséggel igaz lehet, hogy a férfiak többet keresnek, mint a nők, csak az alacsony mintaelemszám miatt az összefüggés nem szignifikáns.

Vegyes kapcsolat kiszámítása az SPSS-sel

Az SPSS segítségével három egyszerűen kivitelezhető módszerrel vizsgálhatjuk meg egy minőségi és egy mennyiségi változó kapcsolatát (átlagok összehasonlítását). A többi, ritkábban használt t-próba (egymintás és páros mintás) és az egyutas ANOVA alkalmazásától most eltekintünk – erről egy leírást a *Melléklet* SPSS menüleírásában találunk.

1. A már ismert módon, az *ANALYZE* főmenü *Descriptive Statistics* almenüjének *Crosstabs* parancsával lekérjük a *Statistics* mezőnél, a *Nominal by Interval* ablakrésznél található *Eta* statisztikát. Ez a mutató a *H* mutatóhoz hasonlóan egy 0 és 1 közötti érték, amely a két változó összefüggésének erősségét mutatja, amikor a független változónk kategoriális mérési szintű, a függő változónk pedig mennyiségi skála. Ebben az esetben az SPSS nem számol szignifikanciaszintet, így ezt a módszert ritkán érdemes használni.

2. Az *ANALYZE* főmenü *Compare Means* almenüjénél az *Independent-Samples T Test…* (független mintás vagy kétmintás t-teszt) paranccsal lekérhetjük a t-eloszlást és az ennek megfelelő szignifikanciaszintet. Itt fontos még megjegyezni, hogy a mennyiségi változónk lesz a *Test Variable*, a dichotóm változónk

pedig a *Grouping Variable*. A kategoriális változónknál minden egyes t-próba lefuttatásakor be kell írni a két kategória kódszámát (*Group1* – az első csoport vagy osztály kódja, *Group2* – a második csoport vagy osztály kódja), még akkor is, ha biztosan nem fordul elő az adatállományban ennél a változónál kettőnél több érték. Utána *Continue*-t, majd *Ok*-t kattintunk.

3. Az *ANALYZE* főmenü *Compare Means, Means* almenüjénél, az *Options* ablakban, a *Statistics for Firs Layer* (bal alsó rész) ablakrészben, az *Anova table and Eta* bejelölésével lekérhető az F-próba. A változók átvitelénél figyeljünk arra, hogy a kategoriális változónk mindig a független, a mennyiségi változónk pedig a függő változó legyen. A kijelölés után *Continue*-t, majd *Ok*-t kattintunk.

Bár csak kétértékű kategoriális változókká alakított formában alkalmazható, mivel két átlagértéket hasonlítunk össze (ha több attribútummal rendelkezik egy ismérv, azt a *t teszt* előtt kétértékűvé kell kódolni), vegyes kapcsolatok elemzésekor leggyakrabban a t-tesztet szokás használni (a kézi számítása egyszerűbb, ezért elterjedtebb, ahogyan már a korábbiakban említésre került).

A normalitás és varianciahomogenitás tesztelése

Ahogyan már korábban jeleztük, mivel mind a t-próba, mind az F-próba érzékeny a normalitásra, ezért kis mintákon (30 eset alatt) szigorúan ellenőriznünk kell, hogy a függő változó csoportokon belüli eloszlása közelít-e a normálishoz. Ilyen helyzetben a *Shapiro-Wilk-próba* alkalmazása ajánlott. SPSS-ben ezt úgy kérhetjük le, hogy az *ANALYZE* főmenüben a *Descriptive Statistics* almenü alatt kiválasztjuk az *Explore* funkciót, a vizsgált függő változót a *Dependent List* mezőbe helyezzük, a kategoriális magyarázó változót a *Factor List*-hez, majd a *Plots* gombra kattintva bejelöljük a *Normality plots with tests* és *Histogram* lehetőséget. Az SPSS ekkor a normalitásteszteket és az ábrákat (*Boxplot, Q-Q plot*) is automatikusan megjeleníti. A *Shapiro-Wilk-próbát* úgy értelmezzük, hogy ha $p > 0{,}05$, akkor nincs szignifikáns eltérés a normális eloszlástól, vagyis a változó eloszlása közel normálisnak tekinthető. Ha $p \leq 0{,}05$, akkor szignifikáns az eltérés a normális eloszlástól, vagyis a változó nem tekinthető normális eloszlásúnak. A hisztogram (kétszer a hisztogramra kattintva az *Elements, Show Distribution Curve, Normal* úton megjeleníthető a normális eloszlás az ábrán) az eloszlás alakját, a Q-Q plot pedig azt mutatja meg, mennyire illeszkednek az értékek a normális eloszláshoz. Így ha a hisztogram közel szimmetrikus, és a Q-Q plot pontjai az egyenes mentén helyezkednek el, az adatok közel normális eloszlásúak.

Nagyobb minták esetén a normalitás ellenőrzése két módon történhet. Egyrészt alkalmazható a *Kolmogorov–Smirnov-próba* (1) (ami a túlérzékenysége miatt sokszor félrevezető eredményekhez vezet) ugyanazon az úton, mint a *Shapiro–Wilk-próba*. Másrészt elegendő lehet a normalitás kevésbé szigorú vizsgálata a ferdeség és csúcsosság, illetve a kiugró értékek ellenőrzésével a leíró statisztikák, hisztogram, boxplot és a Q-Q plot alapján, a *Shapiro–Wilk-próbánál* fentebb leírt SPSS menüútvonalon (2).

A jegyzetben a továbbiakban a (2) kevésbé szigorú normalitástesztelést használjuk, és csak az abszolút értékben 1-nél nagyobb ferdeségi és csúcsossági mutatóval rendelkező változókat tekintjük nem normális eloszlásúaknak, illetve grafikus módszerrel is ellenőrizzük a normalitást (hisztogram, boxplot és Q-Q plot: *ANALYZE, Descriptive Statistics, Explore (Dependent List – függő változó, Factor List –* kategoriális magyarázó változó*), Plots, Histogram* és *Normality plots with tests, Continue, OK*).

A varianciahomogenitás az SPSS-ben tehát a Levene-próbával vizsgálható, amely automatikusan megjelenik a korábban leírt t-próba (*ANALYZE, Compare Means, Independent-Samples T Test…*) és az ANOVA (*ANALYZE, Compare Means, One-Way ANOVA,* majd *Options, Homogeneity of variance test*) kimenetében, ahol a Levene-próba p értéke alapján dönthető el, hogy a csoportok varianciái szignifikánsan különböznek-e.

1. Ha a Levene-próba p értéke $p > 0{,}05$, akkor a varianciahomogenitás teljesül, és a hagyományos t-próba vagy ANOVA eredményei értelmezhetők. Ekkor a t-próbánál az *equal variances assumed* sort (felső sor) kell használni.
2. Ha a p érték $p \leq 0{,}05$, akkor a varianciák különböznek, ezért t-próbánál a *Welch-féle korrekciót,* vagyis az *equal variances not assumed* sort (alsó sor) kell használni, ANOVA esetén pedig robosztus alternatívát (pl. *Welch ANOVA*) kell választani. A Welch ANOVA SPSS-ben úgy kérhető le, hogy az *ANALYZE, Compare Means, One-Way ANOVA* útvonalat követjük, majd a függő változót (*Dependent List*) és a csoportképző változót (*Factor*) megadjuk. Ezután az *Options* gombra kattintva bejelöljük a *Welch* lehetőséget, majd *Continue, OK*. Az SPSS ekkor a robusztus Welch ANOVA eredményét is megjeleníti az outputban.

Ábrázolás

Végül pedig a csoportátlagokat a *GRAPHS* főmenüből, a *Legacy Dialogs, Bar, Simple* utasításokkal lehet ábrázolni. A *Bars represent, Other statistics, Mean* mezőbe kell kerülnie a függő változónak, a *Category Axis*-ra pedig a magyarázó/független változónak. Ahogyan korábban is, kétszer az ábrára kattintva, a *Chart Editor*-ban lehet megjeleníteni az értékeket és megadni azok pontosságát.

41. példa ▼

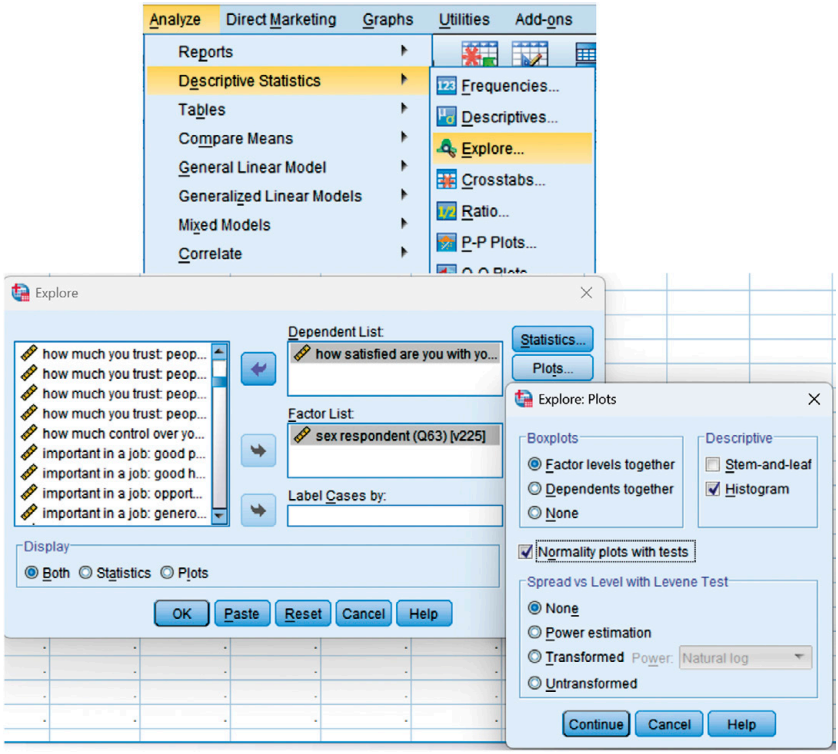
▶ *Átlagok összehasonlítása az SPSS-ben: t-teszt*

Az adatbázisban a már ismert *v225*-ös változó a megkérdezettek *nemét* mutatja, a *v39*-es változó (K10-es kérdés, *how satisfied are you with your life (Q10)*, az adatbázis 60. sorszámú változója) pedig a kérdezettek jelenlegi életükkel való elégedettségét. Bár az 1–10-fokú skálán mért élettel való elégedettség nem egy szigorúan vett skála mérési szintű változó, a gyakorlatban a t-tesztet és F-próbát (mivel a társadalomtudományokban kevés jelenséget

lehet magas mérési szintű változókkal mérni) sokszor szokták szélesebb skálán mért ordinális változókra is alkalmazni. Az a hipotézisünk, hogy a férfiak élettel való átlagos elégedettségi szintje nagyobb, mint a nőké.

Vizsgáljuk meg tehát, hogy van-e szignifikáns összefüggés a *nem* és *az élettel való elégedettség* között.

Első lépésben teszteljük a normalitást a (2) kevésbé szigorú módszerrel, az előzőekben leírt módon (44. ábra).



44. ábra. *A normalitás tesztelésének lekérése (41. példa)*

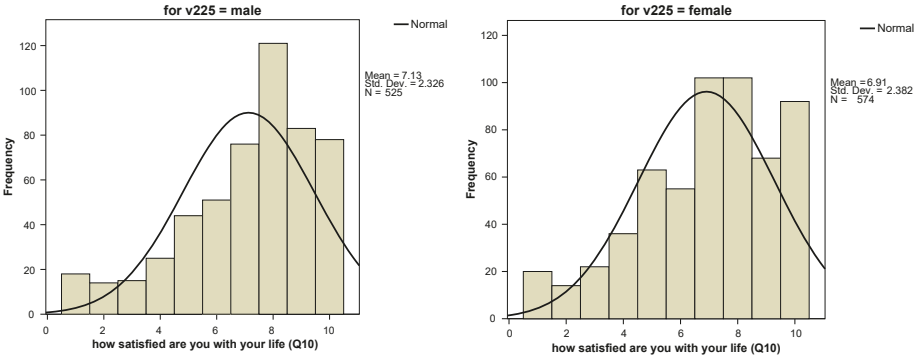
Elsőként a *Descriptives* táblázatból a ferdeségi (*Skewness*) és csúcsossági (*Kurtosis*) mutatókat vizsgáljuk meg a két csoportra (45. ábra). Mindkét csoport esetében az értékek normális közeli eloszlást mutatnak.

Descriptives

sex respondent (Q63)				Statistic	Std. Error
how satisfied are you with your life (Q10)	male	Mean		7.13	.102
		95% Confidence Interval for Mean	Lower Bound	6.93	
			Upper Bound	7.33	
		5% Trimmed Mean		7.30	
		Median		8.00	
		Variance		5.412	
		Std. Deviation		2.326	
		Minimum		1	
		Maximum		10	
		Range		9	
		Interquartile Range		3	
		Skewness		-.898	.107
		Kurtosis		.230	.213
	female	Mean		6.91	.099
		95% Confidence Interval for Mean	Lower Bound	6.71	
			Upper Bound	7.10	
		5% Trimmed Mean		7.05	
		Median		7.00	
		Variance		5.672	
		Std. Deviation		2.382	
		Minimum		1	
		Maximum		10	
		Range		9	
		Interquartile Range		4	
		Skewness		-.645	.102
		Kurtosis		-.225	.204

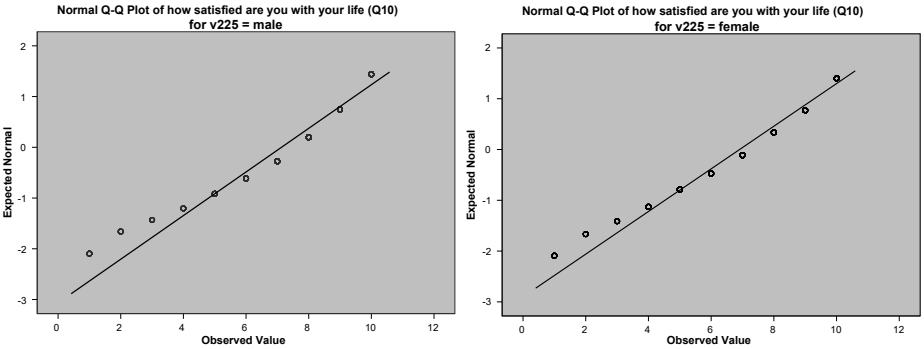
45. ábra. *A leíró statisztikákat tartalmazó táblázat (41. példa)*

A hisztogramok, miután feltüntettük rajtuk a normális eloszlás görbéjét (dupla klikk az ábrára, *Elements, Show Distribution Curve, Normal*), enyhén negatív, balra ferde (több magas érték), normális eloszláshoz közeli szóródást mutatnak (46. ábra).

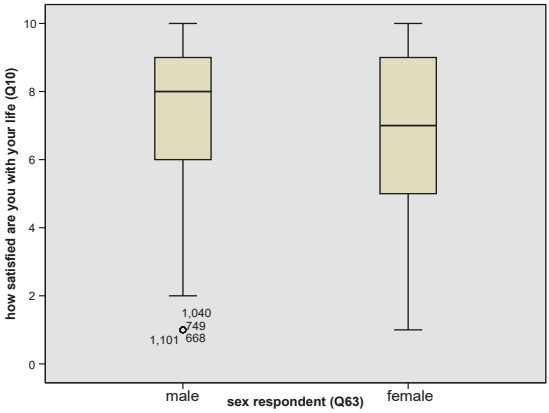


46. ábra. *A hisztogramok (41. példa)*

Ugyanez a kép rajzolódik ki a *Q-Q plotból* (47. ábra) is: a függő változónk csoportokon belüli eloszlása nem teljesen normális, de jól közelíti azt.



47. ábra. *A két Q-Q plot (41. példa)*



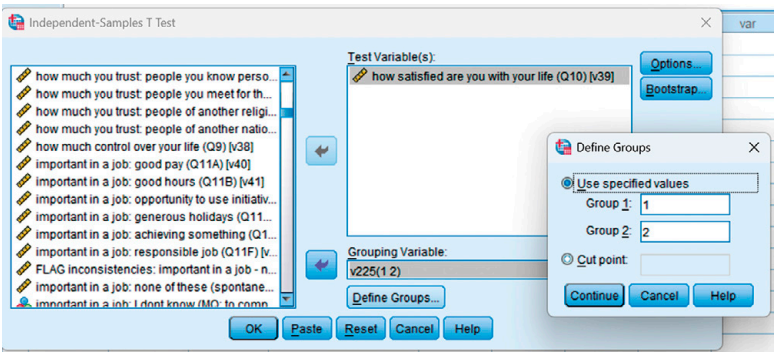
48. ábra. *A dobozdiagramok (41. példa)*

A *Boxplot*-ból (48. ábra) az is kiolvasható, hogy a férfiak életelégedettségének mediánja valamivel magasabb és kevésbé szórt, míg a nőknél nagyobb a szóródás, és emellett néhány kiugró érték, outlier is megjelenik a férfiaknál az alacsony értéktartományban.

Összességében tehát a kevésbé szigorú normalitásvizsgálat azt mutatja, hogy alkalmazhatjuk a t-tesztet és az F-próbát.

A kapcsolatvizsgálatkor az 1. eljárás (*Crosstabs*-ból való lekérés) bemutatására nem térek ki, hiszen az F-próbánál is megjelenik az Eta értéke.

Először nézzük a független mintás t-tesztet, az előzőekben leírtak szerint (49. ábra). Mivel a varianciahomogenitás tesztelése (Levene-próba) automatikusan megjelenik a t-próba lekérésénél, erre külön nem térünk ki.



49. ábra. *A t-teszt lekérése (41. példa)*

Az *Output* ablakban megjelenik a t-teszt (50. ábra) és a csoportstatisztikákat tartalmazó táblázat (51. ábra).

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
how satisfied are you with your life (Q10)	Equal variances assumed	.565	.452	1.573	1097	.116	.224	.142	-.055	.503
	Equal variances not assumed			1.575	1092.259	.116	.224	.142	-.055	.503

50. ábra. *A független mintás t-teszt (41. példa)*

Először a t-teszt táblázatát értelmezzük. Első lépésben megnézzük az F értékének a szignifikanciaszintjét, azaz a Leven-próba eredményét a csoporton belüli homogenitás ellenőrzésére. Mivel $p > 0{,}05$ (az F-re $p = 0{,}452$), 95%-os valószínűségi szint mellett nem utasítjuk el a nullhipotézist, vagyis nincs bizonyíték arra, hogy a két csoport varianciája különbözne. Ezért a t-teszt *Equal variances assumed* eredménye tekinthető érvényesnek, vagyis a felső értéksorban található t-érték szignifikanciaszintjét vizsgáljuk. Az itt szereplő t-érték ($t = 1{,}573$) szignifikanciaszintje ($p = 0{,}116$) azt mutatja, hogy a két alcsoport átlaga közötti különbség 95%-os valószínűségi szinten nem szignifikáns, vagyis az adatok alapján nem zárható ki, hogy a megfigyelt eltérés akár az 5%-os véletlen ingadozás következménye legyen. Ezért nincs lényeges eltérés a férfiak és a nők között az élettel való elégedettség átlagos megítélésében.

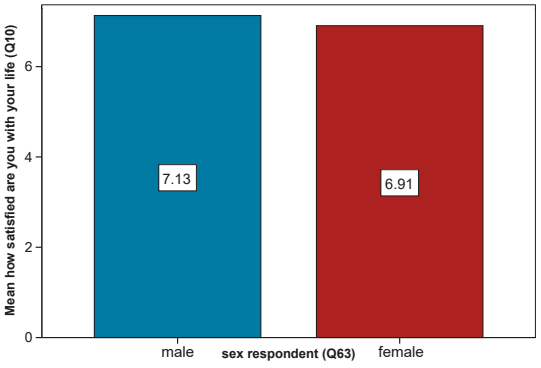
Group Statistics

	sex respondent (Q63)	N	Mean	Std. Deviation	Std. Error Mean
how satisfied are you with your life (Q10)	male	525	7.13	2.326	.102
	female	574	6.91	2.382	.099

51. ábra. *A csoportstatisztikák (41. példa)*

A csoportstatisztikákat szemléltető táblázat (51. ábra) alapján elmondhatjuk, hogy a nők átlagos élettel való elégedettségi szintje (6,91) nagyon hasonló a férfiakéhoz (7,13), az a kis csoportátlagok közötti eltérés csak a véletlen műve (abból adódik, hogy pont annak az 1099 személynek a válaszait elemeztük, aki a mintába bekerült és válaszolt a kérdésre). Tehát a hipotézisünk nem igazolódott be.

Végül pedig a már ismertetett módon lekérjük az átlagértékeket tartalmazó ábrát (52. ábra).



52. ábra. *Az élettel való elégedettség nemi bontásban (átlagértékek, N = 1099) (41. példa)*

42. példa ▼

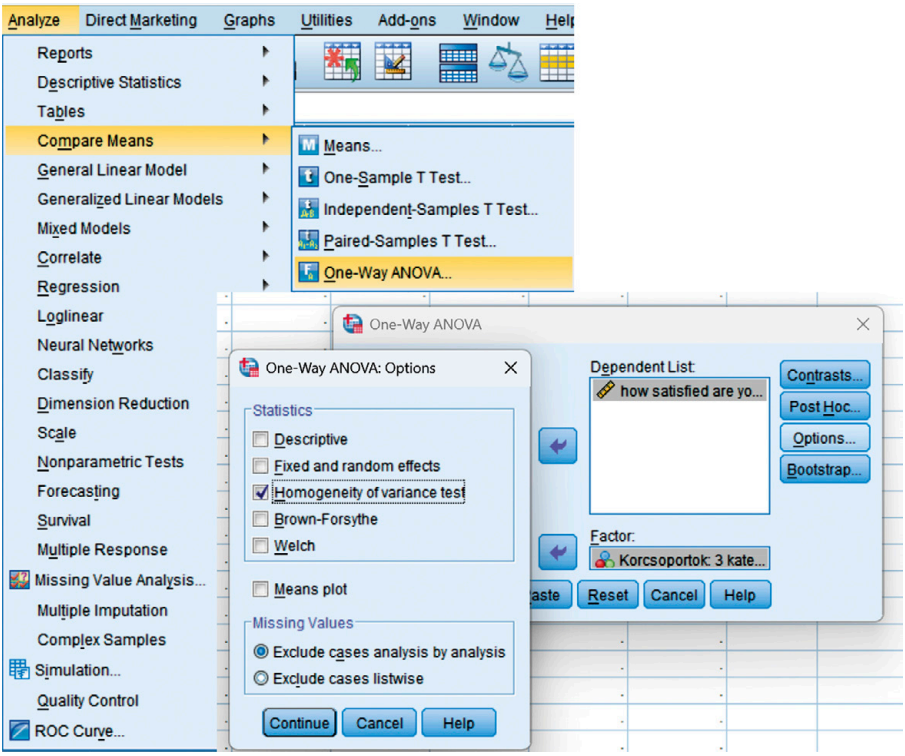
▶ *Átlagok összehasonlítása az SPSS-ben: F-próba*

Adatbázisunkban a már ismert *korcsoport3kat* változó a megkérdezettek 3 korcsoport szerinti bontását mutatja (a 8. és 10. példa szerint hoztuk létre), a *v39*-es változót az előző, 41. példánál ismertettük (*a kérdezettek jelenlegi életükkel való elégedettségét* méri). Ebben az esetben azt feltételezzük, hogy a fiatal korcsoport átlagos élettel való elégedettségi szintje magasabb, mint a két idősebb korcsoporté.

Vizsgáljuk meg tehát, hogy van-e szignifikáns összefüggés a korcsoportok és az élettel való elégedettség között.

A korábbiakban ismertetett módon első lépésben a csoportokon belüli normális eloszlást vizsgáljuk a kevésbé szigorú (2) módszerrel: *ANALYZE, Descriptive Statistics, Explore* (függő változó, a *v39* a *Dependent List* mezőbe, a kategoriális magyarázó változó, a *korcsoport3kat* a *Factor List*-hez kerül), *Plots, Normality plots with tests* és *Histogram*. A ferdeségi és csúcsossági mutatók abszolút értékben 1 alattiak, mindhárom korcsoportban normális közeli, enyhén balra (negatív) ferde eloszlás rajzolódik ki. A hisztogramok és Q-Q plotok is ezt igazolják, a dobozdiagram is csak néhány kiugró értéket jelez a két 65 év alatti korcsoportban, ezért úgy döntünk, hogy a három korcsoporton belül a normalitás feltétele teljesül (az ábrákat ezúttal nem másoltuk be, a t-tesztnél leírtakhoz hasonlóak).

A csoporton belüli homogenitás ellenőrzésére a Levene-tesztet ezúttal az egyutas ANOVA menüpontnál kérjük le: *ANALYZE, Compare Means, One-Way ANOVA*, majd *Options, Homogeneity of variance test* (53. ábra).



53. ábra. *A Levene-teszt lekérése (42. példa)*

A *Levene-próba* eredménye (54. ábra) szerint p = 0,653, ami jóval nagyobb a 0,05-nél, tehát nem utasítjuk el a nullhipotézist, vagyis a csoportok varianciája homogén, nincs szignifikáns különbség a szórások között. Tehát az alkalmazhatósági feltételek teljesültek, alkalmazható az F-próba.

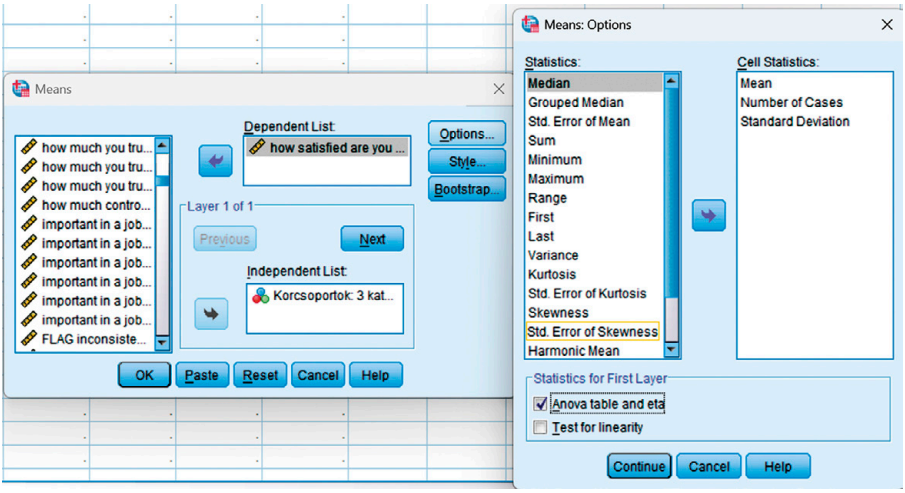
Test of Homogeneity of Variances

how satisfied are you with your life (Q10)

Levene Statistic	df1	df2	Sig.
.427	2	1096	.653

54. ábra. *A Levene-teszt eredménye (42. példa)*

A korábbiakban ismertetett módon kérjük le az F-próbát (55. ábra).



55. ábra. *Az F-próba (ANOVA) lekérése (42. példa)*

Az ANOVA (*Analyze of Variance*) táblázatunk (56. ábra) azt mutatja, hogy a két változó közötti összefüggés szignifikáns (F = 7,268, p = 0,001), vagyis a különböző korcsoportok élettel való átlagos elégedettsége lényegesen eltér.

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
how satisfied are you with your life (Q10) * Korcsoportok: 3 kategória	Between Groups	(Combined)	79.846	2	39.923	7.268	.001
	Within Groups		6019.921	1096	5.493		
	Total		6099.767	1098			

56. ábra. *Az F-próba (ANOVA) eredménye (42. példa)*

A csoportátlagokat a *Report* elnevezésű, *Output*-ban megjelenő táblázatból olvassuk ki (57. ábra).

Report

how satisfied are you with your life (Q10)

Korcsoportok: 3 kategória	Mean	N	Std. Deviation
legtöbb 30 éves	7.49	204	2.339
31-65 éves	7.02	609	2.329
65 év feletti	6.67	286	2.378
Total	7.01	1099	2.357

57. ábra. *Csoportátlagok (42. példa)*

A csoportátlagok azt mutatják, hogy a 30 év alattiak szignifikánsan elégedettebbek az életükkel (7,49-es átlagérték), mint a 31–65 évesek (7,02-es átlagérték), ők pedig, mint a 65 év felettiek (6,67-es csoportátlag). Tehát a hipotézisünk beigazolódott.

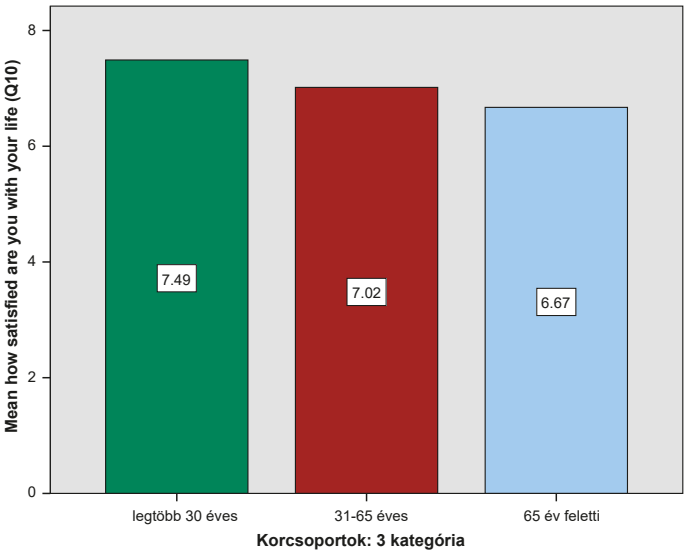
Az Eta és Eta-négyzet statisztikák (58. ábra) azt jelzik, hogy a két változó közötti kapcsolat gyenge (Eta = 0,114), és a korcsoport ismerete csak 1,3%-ban segít megbecsülni, hogy ki mennyire elégedett az életével.

Measures of Association

	Eta	Eta Squared
how satisfied are you with your life (Q10) * Korcsoportok: 3 kategória	.114	.013

58. ábra. *Eta és Eta-négyzet (42. példa)*

Végül pedig a már ismertetett módon lekérjük a csoportátlagokat szemléltető ábrát (59. ábra).



59. ábra. *Az élettel való elégedettség korcsoportos bontásban (átlagértékek, N = 1099) – 42. példa*

△ *Gyakorlófeladatok a t-teszt és F-próba SPSS-ben való lekérésére*

1. Kérjünk független mintás t-teszteket (9 db.) a v225 (nem) és a v133–v141 változók (a kérdőív K34-es kérdése, az adatbázisban a 166–174-es sorszámú, a demokrácia különböző jellemzőinek fontosságára vonatko-

zó változók) között! Értelmezzük lépésről lépésre a kapott adatokat az elemzett változók kontextusában! Ábrázoljuk is!

2. Kérjünk F-próbákat (9 db) a *korcsoport3kat* és a *v32–v37* változók (a kérdőív K34-es kérdése, az adatbázisban a 166–174-es sorszámú, *a demokrácia különböző jellemzőinek fontosságára* vonatkozó változók) között! Értelmezzük lépésről lépésre a kapott adatokat az elemzett változók kontextusában! Ábrázoljuk is!

4.4. Két mennyiségi változó közötti kapcsolat: korreláció

A korreláció arra az esetre vonatkozik, amikor mindkét változó mennyiségi (intervallum- vagy arányskálán mérhető). Akárcsak a vegyes kapcsolat esetén, itt is megtehető, hogy az egyik ismérvet (*Y*) csak osztályozásra használjuk, a másikat pedig átlag- és varianciaszámítás segítségével vizsgáljuk. Két mennyiségi ismérv esetében azonban két vonatkozásban tehetünk ennél többet:

1. kihasználhatjuk azt, hogy az *X* ismérv szerint képzett osztályok az *X* változó nagysága szerint egyértelműen sorrendbe állíthatóak,
2. nemcsak *Y*, hanem *X* szerint is osztályozhatjuk a sokaságot, és ekkor *Y*-t vizsgáljuk varianciaanalízis segítségével.

Az *X* és *Y* szerint képzett osztályok egyértelmű rendezhetősége az ismérvek közötti kapcsolat irányának értelmezését teszi lehetővé (akárcsak γ esetében):

c) ha *X* növekedésével párhuzamosan *Y* is növekszik, a kapcsolat *pozitív irányú*,
d) ha *X* növekedésével párhuzamosan *Y* csökken, a kapcsolat *negatív irányú*.

A kapcsolat iránya csak akkor értelmezhető, ha a két ismérv közötti kapcsolat monoton természetű.

Az *Y* szerint képzett osztályokhoz hozzárendelt X_i részátlagok sorozatát az *X* változó *Y* változóra vonatkozó (*Y* szerinti) empirikus regressziófüggvényének nevezzük. Az empirikus regressziófüggvény nemcsak annak jelzésére szolgál, hogy van-e kapcsolat a két változó között, hanem a kapcsolat természetének tömör kifejezésére is. A kapcsolat létét itt is az jelzi, hogy az egyes *Y* osztályokhoz különböző X_i részátlagok tartoznak, ellenkező esetben az *Y* ismerete nem adna semmiféle többletinformációt az *X* szerinti hovatartozás becsléséhez.

Az empirikus regressziófüggvény grafikusan is ábrázolható az (X_i, Y_i) pontokat összekötő vonaldiagram formájában, ahol Y_i vagy egyedi ismérvérték, vagy az *Y* szerint képzett osztályköz osztályközepe, vagy az adott osztályközbe tartozó *Y*-értékek átlaga. Az empirikus regressziófüggvény önmagában nem mutatja meg, hogy a két változó közötti kapcsolat függvényszerű-e vagy nem, mert nem derül ki belőle, hogy az X_i részátlagok körül van-e szóródás, ezért célszerűbb a pontdiagrammal közös ábrát használni.

Az eddig tárgyalt esethez rendelhető varianciahányadosnak külön neve és jelölése van: az *Y* szerinti osztályokhoz rendelt részátlagok sorozatából számítható varianciahányados *X*-nek az *Y*-ra vonatkozó determinációs hányadosa, jelölése $\eta^2_{x/y}$.

$$\eta^2_{x|y} = 1 - \frac{\sigma^2_B(X)}{\sigma^2(X)} \qquad \eta_{x|y} = \sqrt{\eta^2_{x|y}}$$

Ekkor a $\eta_{x/y}$ a korrelációs hányados.

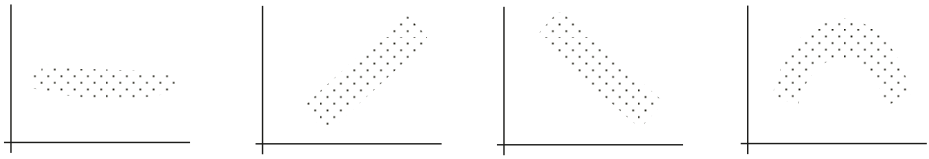
Teljesen hasonlóan értelmezhető *Y*-nak az *X*-re vonatkozó empirikus regressziófüggvénye és az ehhez tartozó determinációs hányados és korrelációs hányados. Ha az *X* és *Y* közötti kapcsolat sztochasztikus, általában $\eta^2_{x/y} = \eta^2_{y/x}$.

Tapasztalati regressziófüggvényt és determinációs hányadost csak akkor ajánlott használni, ha a megfigyelt sokaság elég nagy ahhoz, hogy az osztályokba *1-nél több* egység tartozzon. Ha minden osztályban csak egy egység van, egyik osztályon belül sincs szóródás, és így $\eta^2 = 1$, ami megtévesztő. A determinációs hányados értéke mindig nagyon függ a számításhoz használt osztályozás konkrét módjától. A korrelációs hányados nem értelmezhető százalékként.

Ha azonban áttérünk a sokaság egységeinél *együttesen fellépő* (X_i, Y_j) *értékpárok vizsgálatára*, akkor továbbmehetünk a két mennyiségi változó közötti kapcsolat elemzésében. Ebben az esetben az a kérdés, hogy az az információ, hogy a sokaság valamely egységénél az *X* ismérv értéke éppen X_i, felhasználható-e valahogyan az adott egységnél előforduló Y_i becslésére. E kérdés megválaszolása a *regressziószámítás* feladata, amelynek célja az *X* és *Y* közötti sztochasztikus kapcsolat természetének egy *f(X)* függvénnyel való leírása. Az *f(X)* függvényt az empirikus regressziófüggvénytől való megkülönböztetés céljából *analitikus regressziófüggvénynek* szokás nevezni, és elsősorban arra használjuk, hogy annak X_i helyen vett $f(X_i)$ helyettesítési értékével megbecsüljük az *Y* változónak az X_i értékével együtt előforduló értékét.

Arról, hogy egy ilyen *f(X)* függvény létezésére lehet-e számítani, a *pontdiagram* nyújt segítséget. Ha a pontdiagram pontjai *nem véletlenszerűen szóródnak, biztosak lehetünk az f(X) létezésében.* A pontdiagram nemcsak a változók közötti kapcsolat létéről, hanem *a kapcsolat jellegéről* is informál. Leghasznosabb a pontdiagram és az empirikus regressziófüggvény közös ábrázolása, mivel csak egy ilyen ábra segítségével lehet különbséget tenni a sztochasztikus és függvényszerű kapcsolat között, és az empirikus regressziófüggvény a pontdiagram lényegét is megjeleníti.

A 60. ábra néhány jellegzetes pontdiagramsémát szemléltet.



a) X és Y független *b) pozitív kapcsolat* *c) negatív kapcsolat* *d) nem-lineáris kapcsolat*

60. ábra. *Néhány jellegzetes pontdiagram*

Forrás: Hunyadi–Mundruczó–Vita 2000. 181.

Amennyiben már ismert az *f(X)* függvény típusa, a következő lépés a *paraméterek meghatározása, becslése* a megfigyelt (Xi, Yi) értékpárok alapján {f(X) = aX + b}.

A paraméterek meghatározása után a regressziófüggvény felhasználásával megadható az *Y* változónak az *X* változó X_i értékével együtt előforduló értékére az $\hat{Y}_i = f(X_i)$.

A következő lépésben *alkalmazzuk a PRE-eljárást* az *X* és *Y* közötti korrelációs kapcsolat szorosságának mérésére, feltételezve, hogy a két változó közötti sztochasztikus kapcsolat természetét leíró analitikus regressziófüggvény lineáris.

E_1 esetén nem ismerjük az *X* szerinti hovatartozást, így az $\hat{Y}_i$-t nyilvánvalóan az $\overline{Y}$ -nal becsüljük. Ha ismerjük az *X* szerinti hovatartozást, $\hat{Y}_i$-t az $f(X_i)$ felhasználásával becsüljük (E_2).

$$E_1 = \sum \left(Y_i - \overline{Y}\right)^2 = \sum d_y{}^2 = N \cdot \sigma^2{}_y$$

$$E_2 = \sum \left[Y_i - f(x_i)\right]^2 = 1 - r^2) \cdot \sum d^2{}_y = (1 - r^2) \cdot N \cdot \sigma^2{}_y$$

ahol:

r – lineáris *korrelációs együttható*

r^2 – *determinációs együttható*, PRE mutató

$$d_x = X_i - \overline{X}, \quad d_y = Y_i - \overline{Y}$$

$$r = \frac{\sum d_x \cdot d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} \qquad PRE = \frac{E_1 - E_2}{E_1} = \frac{N\sigma_y^2 - (1 - r^2) \cdot N\sigma_y^2}{N\sigma_y^2} = r^2$$

A determinációs együttható (r^2) azt mutatja, hogy az *X* változó egyes egységeknél előforduló X_i értékeinek ismerete hány százalékkal csökkenti az *Y* változó azokhoz tartozó Y_i értékeinek becslésekor elkövetett hibát, ha a becslés a lineáris analitikus regressziófüggvény segítségével történik.

A korrelációs együttható (r) kifejezhető a *kovariancia* segítségével is, amely bár nem PRE mutató, mégis alkalmas a két változó együtt ingadozásának mérésére:

$$r = \frac{C}{\sigma_x \cdot \sigma_y} \qquad C = \frac{\sum d_x \cdot d_y}{N}$$

Ha C = 0, *X* és *Y* között nincs kapcsolat, ha C > 0, a két változó közti kapcsolat pozitív, ha C < 0, a két változó közti kapcsolat negatív irányú. A C önmagában nem alkalmas a kapcsolat szorosságának jellemzésére (a szorosság függ a szóródástól is). Az *r* korrelációs együttható kiküszöböli a kovariancia e hátrányát (osztja a két változó szóródásával). Az *r* vagy a *Pearson-féle korrelációs együttható* egy [-1; 1] intervallumba eső érték, mérőszám. Ha r = 1 vagy r = -1, a két változó függvényszerű lineáris kapcsolatban áll egymással. Az r értéke a kapcsolat szorosságát méri, és minél nagyobb, annál szorosabb kapcsolatot jelez.

Ha a nullhipotézisünk az, hogy a teljes sokaságban az *X* és *Y* változók függetlenek (r = 0), akkor az n elemű összes lehetséges minták sokaságán a

$$t = r_{xy} \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

valószínűségi változó n–2 paraméterű t-eloszlás (*Student-féle t-eloszlás* n–2 szabadságfokkal), ami elég nagy n esetén (n > 120) n (0,1) paraméterű normális eloszlás. Így, ha az esetek száma nagy, a p = 0,05, a p = 0,01 és a p = 0,001 szignifikanciaszinteknek megfelelő t-érték 1,96, 2,58 és 3,29. Ha viszont az esetek száma kevesebb 100-nál, szükségünk van egy t-eloszlás táblázatra (lásd a *Mellékletet*).

43. példa ▼

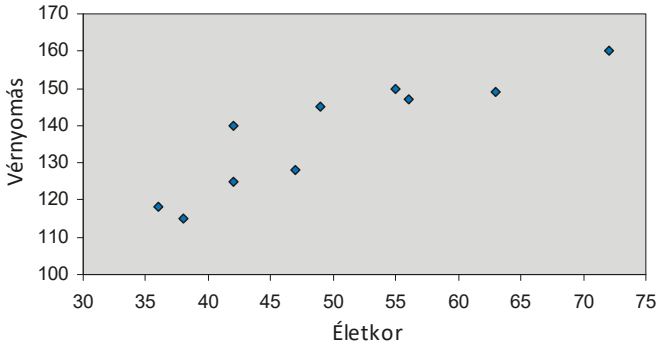
▶ *Korrelációszámítás*

A 28. táblázat (fiktív adatok) 10 véletlenszerűen kiválasztott nő életkorát és vérnyomását mutatja.

28. táblázat. *Két változóra felvett értékek és részszámítások (43. példa)*

Életkor (X)	Vérnyomás (Y)	$d_x = X_i - \overline{X}$	$d_y = Y_i - \overline{Y}$	$d_x \cdot d_y$	d_x^2	d_y^2
36	118	36-50 = -14	118-137,7 = -19,7	275,8	196	388,09
38	115	38-50 = -12	115-137,7 = -22,7	272,4	144	515,29
42	125	-8	-12,7	101,6	64	161,29
42	140	-8	2,3	-18,4	64	5,29
47	128	-3	-9,7	29,1	9	94,09
49	145	-1	7,3	-7,3	1	53,29
55	150	5	12,3	61,5	25	151,29
56	147	6	9,3	55,8	36	86,49
63	149	13	11,3	146,9	169	127,69
72	160	22	22,3	490,6	484	497,29
			Σ	**1408**	**1192**	**2080,1**

Rajzoljuk fel a pontdiagramot, hogy lássuk, van-e értelme lineáris összefüggést keresni (61. ábra). A pontdiagramunk azt jelzi, hogy joggal feltételezhetjük egy pozitív lineáris kapcsolat létét.



61. ábra. *A pontdiagram (43. példa)*

Számítsuk ki és értelmezzük a korrelációs és determinációs együtthatókat.

Első lépésben kiszámoljuk a két változó számtani átlagát.

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{36 + 38 + \ldots + 72}{10} = 50$$

$$\overline{Y} = \frac{Y_1 + Y_2 + \ldots + Y_N}{N} = \frac{118 + 115 \ldots + 160}{10} = 137{,}7$$

Második lépésben egy-egy új oszlopba kiszámoljuk a d_x és d_y különbségeket.
Harmadik lépésben összeszorozzuk a d_x- és d_y-értékeket, majd összeadjuk őket (Σ).
Negyedik lépésben négyzetre emeljük a d_x értékeket és összeadjuk (Σ), majd ugyanezt elvégezzük d_y-ra is (az eredmények a 28. táblázatban szerepelnek).
Ötödik lépésben kiszámítjuk a Pearson-féle korrelációs együtthatót:

$$r = \frac{\sum d_x \cdot d_y}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{1408}{\sqrt{1192 \cdot 2080{,}1}} = \frac{1408}{1574{,}636} = 0{,}894$$

Négyzetre emeléssel kiszámoljuk a determinációs együtthatót:

$$r^2 = 0{,}7995$$

Értelmezés szerint a korrelációs együttható értéke egy erős, pozitív kapcsolatot mutat. Tehát minél idősebb egy nő, annál nagyobb a vérnyomása. A determinációs együttható azt jelzi, hogy az életkor ismerete 80%-kal csökkenti a vérnyomás ismeretével kapcsolatos bizonytalanságot.

Végül pedig számoljuk ki a t értékét, hogy alapsokaságunkra is tudjunk következtetni.

$$t = r_{xy} \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^{\ 2}}} = 0{,}89 \cdot \frac{\sqrt{10-2}}{\sqrt{1-0{,}80}} = 0{,}89 \cdot 6{,}32 = 5{,}6248$$

Mivel elemszámunk 10 ($n < 120$), a t-táblázatot használjuk ($df = n-2 = 10-2 = 8$). A t-táblázatból idevágó értékek a 29. táblázatban szerepelnek.

29. táblázat. *A szabadságfoknak megfelelő t értékek (43. példa)*

Szabadságfok	Szignifikanciaszint		
	p = 0,05	**p = 0,01**	**p = 0,001**
8	2,306	3,355	5,041

Mivel a számított t-értékünk (5,62) nagyobb, mint a 29. táblázatban szereplő bármelyik *k*-küszöbérték ($k_{max} = 5{,}041$), a két változó közötti összefüggés 99,9%-os valószínűségi szint mellett szignifikáns (99,9%-os biztonsággal állíthatjuk, hogy alapsokaságunkban is a két változó összefügg egymással): minél idősebb egy nő, annál nagyobb a vérnyomása, és az életkorból mintegy 80%-ban ($r^2 = 0{,}8$) lehet következtetni a vérnyomás nagyságára.

Korreláció kiszámítása az SPSS-sel

Az SPSS segítségével kétféleképpen számolhatunk korrelációt.

1. A már ismert módon, az *ANALYZE* főmenü *Descriptive Statistics* almenüjének *Crosstabs* parancsával, a *Statistics* mezőnél a Pearson-féle korrelációs együttható (*Correlations*) lekérésével (a jobb felső sarokban található). Ha a főablakban a változók alatt beklikkeljük a *Suppress tables*-t, a kereszttábla nem fog megjelenni (erre, akárcsak a gamma esetében, nincs szükség).
2. Az *ANALYZE* főmenü *Correlate* almenüjénél a *Bivariate* opcióra klikkelve.

Az SPSS-program mindkét esetben szignifikanciaszintet is számol, így csak arra kell figyelnünk, hogy releváns adatokkal dolgozzunk, vagyis tisztítsuk meg adatainkat az érvénytelen válaszoktól, és skálamérési szintű változókkal dolgozzunk.

Végül pedig a szórásdiagram és a regressziós egyenes ábrázolására válasszuk a *GRAPHS* főmenüt, majd azon belül a *Legacy Dialogs* almenüt, itt a *Scatter/Dot* menüpontban a *Simple Scattert*, majd *Define*. A függő változót az Y tengelyre, a magyarázó változót pedig az X tengelyre ábrázoltatjuk. Kétszer az ábrára kattintva az *Elements, Fit Line at Total* segítségével a ponthalmazra illesszük a regressziós egyenest.

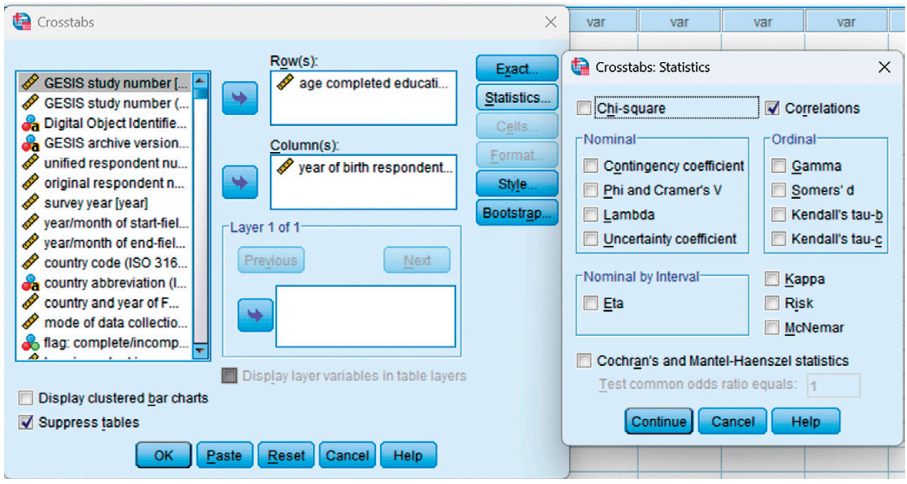
44. példa ▼

▶ *Korreláció az SPSS-ben*

Az adatbázisban, mint a legtöbb társadalomtudományi kutatás esetében, nagyon kevés magas mérési szintű (intervallum vagy arányskála) változónk van, összesen kettő. Az eredetileg arányskálaként kérdezett háztartásban élők száma esetében (K61-es kérdés, *v240*-es változó) a 6-nál több háztartástag adatai az adatbázisban egyetlen kategóriában szerepelnek, a jövedelmek decilisekként, illetve további EVS-ben mért skálaváltozók az erdélyi adatbázisban nem szerepelnek (pl. *v239_r – a háztartásban élő gyerekek száma*, *v241 – a legfiatalabb háztartástag életkora* stb.).

Vizsgáljuk meg a két mennyiségi változó, a születési év (*v226, year of birth respondent (Q64)*, az adatbázisban a 266-os sorszámú változó, a K58-as kérdés) és a nappali iskolai tanulmányok befejezésének életkora (*v242*, *age completed education respondent (Q80)*, az adatbázisban a 291-es sorszámú változó, a K62-es kérdés) közötti összefüggést. Azt feltételezzük, hogy a születési év növekedésével nő a formális képzettség megszerzésének életkora is (egy szignifikáns, pozitív korrelációs együtthatót várunk).

Először az első módszerrel lefuttatunk egy korrelációt (62. ábra).



62. ábra. *A korrelációs együttható lekérése a Descriptive Statistics, Crosstabs almenüből (44. példa)*

Az *Output* ablakban megjelenik a kért statisztikánk (63. ábra).

Symmetric Measures

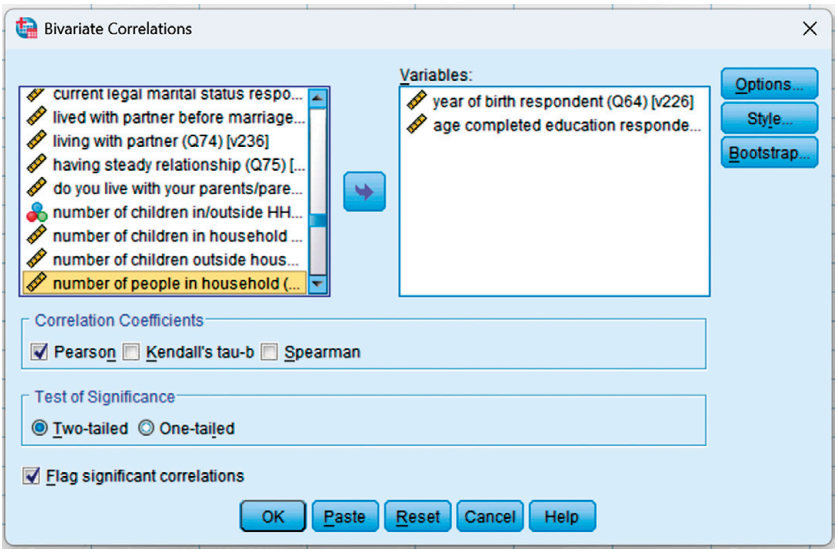
		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.164	.034	5.243	.000[c]
Ordinal by Ordinal	Spearman Correlation	.256	.032	8.368	.000[c]
N of Valid Cases		1002			

a) Not assuming the null hypothesis.
b) Using the asymptotic standard error assuming the null hypothesis.
c) Based on normal approximation.

63. ábra. *A korrelációs együttható (44. példa)*

A Pearson-féle korrelációs együttható szignifikáns (p = 0,000), értéke egy nagyon gyenge pozitív kapcsolatot jelez a két változó között (r = 0,164). 99,9%-os valószínűségi szint mellett kijelenthetjük, hogy a születési év növekedésével együtt enyhén nő a nappali iskolai tanulmányok befejezésének életkora is. Másképpen fogalmazva azt mondhatnánk, hogy a fiatalabbak egy kicsit későbbi életkorban fejezik be a nappali tagozatos tanulmányaikat, azaz enyhén nőtt az oktatási intézményekben eltöltött idő. Tehát a hipotézisünk beigazolódott.

Természetesen, ha a *Correlate* almenüből kérjük le a korrelációs együtthatót (64. ábra), akkor is ugyanezt az értéket kapjuk (65. ábra). Ebben az esetben az értelmezést megkönnyíti a szignifikáns összefüggések csillagokkal való kiemelése.



64. ábra. *A korrelációs együttható lekérése a Correlate almenüből (44. példa)*

Correlations

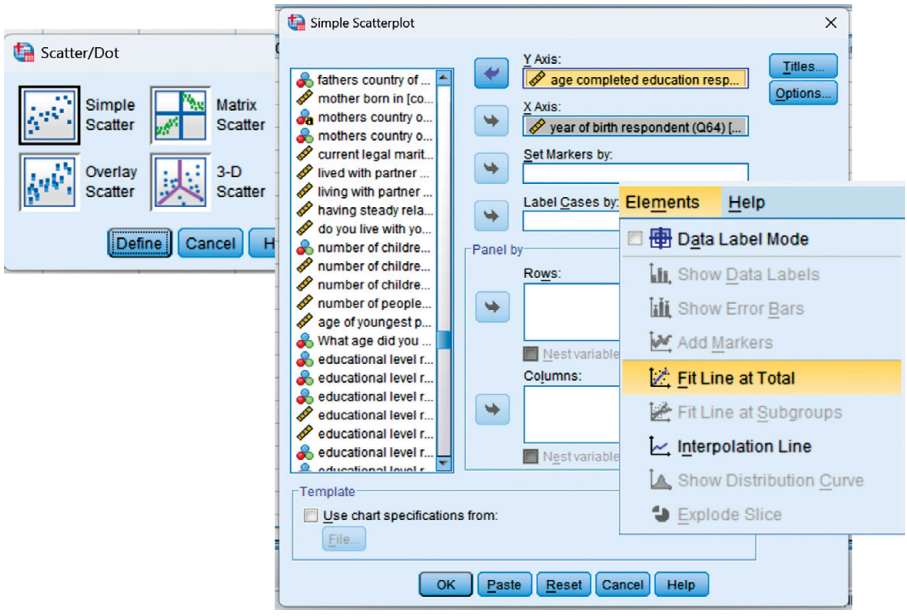
		year of birth respondent (Q64)	age completed education respondent (Q80)
year of birth respondent (Q64)	Pearson Correlation	1	.164**
	Sig. (2-tailed)		.000
	N	1106	1002
age completed education respondent (Q80)	Pearson Correlation	.164**	1
	Sig. (2-tailed)	.000	
	N	1002	1002

**. Correlation is significant at the 0.01 level (2-tailed).

65. ábra. *A korrelációs együttható a Correlate almenüből (44. példa)*

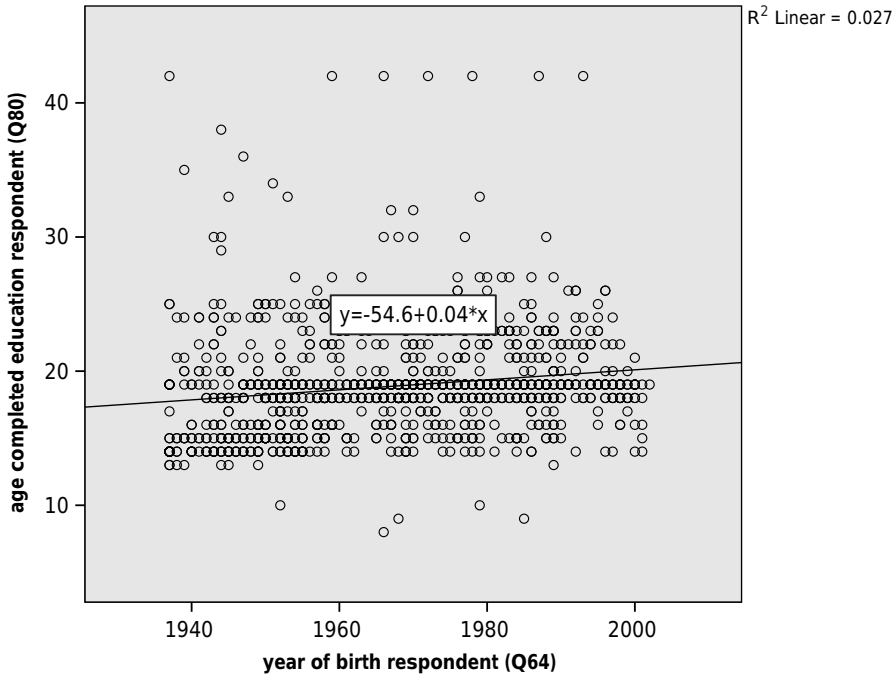
Ahogyan az SPSS is két csillaggal jelzi, az összefüggés p = 0,01 (99%-os) valószínűségi szint mellett is szignifikáns.

Végül pedig kérjük le a két változó közötti kapcsolatot egy szórásdiagramon, a korábbiakban leírt módon (66. ábra).



66. ábra. *A szórásdiagram lekérése (44. példa)*

A kapott diagram a 67. ábrán látható. A ponthalmazra illesztett lineáris regressziós egyenes felett látható az egyenes egyenlete is, illetve a jobb felső sarokban a determinációs együttható értéke ($R^2 = 0{,}027$). Ez egy nagyon alacsony érték, tehát a vizsgált változók között egy szignifikáns, de nagyon gyenge lineáris kapcsolat van. A lineáris regressziós modell (bővebben az *5.2. A többváltozós lineáris regresszió* alfejezetben) a vizsgált jelenség varianciájának mindössze 2,7%-át magyarázza, tehát a magyarázóerő rendkívül gyenge. Ez arra utal, hogy egy kétváltozós lineáris regressziós modell nem igazán alkalmas jó előrejelzésre, mert alig ragadja meg a változásokat.



67. ábra. *A születési év és a nappali képzés befejezésének életkora közötti kapcsolat (44. példa)*

5. FEJEZET

TÖBBVÁLTOZÓS ELEMZÉSEK

Ebben a fejezetben a legismertebb többváltozós elemzések: regresszió, útelemzés, idősorok elemzése, faktorelemzés, klaszterelemzés, diszkriminanciaanalízis és logisztikus regresszió rövid, lényegi összegzésére törekszem, majd rátérek a társadalomtudományokban leggyakrabban alkalmazott három többváltozós módszer, a lineáris regresszió, a faktorelemzés és a klaszterelemzés részletes bemutatására.

5.1. A többváltozós elemzések fajtái

1. Regresszióelemzés

Két mennyiségi változó közötti sztochasztikus kapcsolat leírása az $Y = f(X)$ függvénnyel történik. A konkrét függvény paramétereinek meghatározása a regresszióelemzés módszerével történik. A regresszióelemzés arra a kérdésre keres választ, hogy melyik az a függvény (lineáris vagy nem lineáris), amelynek segítségével az egyik változó (X) értékét megismerve előrejelzést tehetünk egy másik változó (Y) értékére. Ahogyan a korrelációszámításnál már láttuk, két mennyiségi változó pontdiagramjából leolvashatjuk, hogy van-e, és ha van, milyen jellegű a kapcsolat. A regresszióelemzés fajtáit megkülönböztethetjük az elemzésbe bevont független változók száma szerint (egyváltozós, illetve két- és többváltozós), a függvény típusa szerint (lineáris és nem lineáris) stb.

A regresszióelemzés alapvető fajtái:

1. lineáris regresszió,
2. többváltozós lineáris regresszió,
3. parciális regresszió,
4. nem lineáris regresszió.

Lineáris regresszió

Két mennyiségi változó közötti kapcsolat legegyszerűbb formája a lineáris kapcsolati típus, amikor az összefüggést egy függvény írja le (grafikus képe egy egyenes, amilyent a 60. ábrán láthattunk a 44. példában). A lineáris regresszióanalízis az a statisztikai eljárás, amellyel megtalálhatjuk a két változóra együttesen felvett értékekhez (a pontdiagram pontjaira) legjobban illeszkedő egyenest. Tehát a lineáris regresszióban a *regressziós egyenes* alkalmas a két vál-

tozó kapcsolatának grafikus ábrázolására, a *regressziós egyenlet* pedig a kapcsolat összegzésére használható.

A regressziós egyenlet leíró és következtetési szempontból is hasznos: megkapjuk a két változó közti kapcsolat matematikai leírását, valamint lehetőségünk van arra, hogy *X* ismeretében következtessünk *Y* értékére. Mivel a pontokra legjobban illeszkedő egyenest arra akarjuk használni, hogy *X* értékeiből az *Y* értékeire következtessünk, a legjobb egyenes az lesz, amellyel az előrejelzés hibája a legkisebb.

Ha a lineáris függvény alakja $Y = a + bX$, akkor az X_i-értékhez becsült Y-érték:

$$\hat{Y} = a + bX_i$$

Az *a* és *b* értékeit úgy számítják ki, hogy a tényleges *Y*-értékek és a becsült értékek (*X* alapján adott becslések) közötti eltérés minimális legyen. A regressziós becslés jóságának mérésére a becsült $\hat{Y}$- és a valódi *Y*-érték varianciájának hányadosa használható, amely nem más, mint a korreláció kapcsán számolt *determinációs együttható* (r^2).

Többváltozós lineáris regresszió (részletesebben lásd az 5.2-es alfejezetet)

A valóságban előforduló jelenségek olyan bonyolultak, hogy legtöbbször az egyszerű lineáris regresszió nem elég jó a leírásukra. Sokszor előfordul, hogy egy adott függő változóra egyszerre több független változó is hatással van (pl. a havi alkoholfogyasztás mennyiségét befolyásolhatja az életkor, a különleges események száma, a hőmérséklet, a szabadidő mennyisége stb.). Ilyen esetek kezelésére nyújt megoldást a többváltozós regresszió. Ilyenkor a regressziós egyenletben több *X* változó kerül az egyetlen *X* helyébe, és a *b* paraméterek száma is megváltozik, de a logika ugyanaz: minden egyes *b*-érték megadja az egyes független változók szerepét a végső érték meghatározásában. A többváltozós lineáris regressziót a *többszörös korrelációs együttható* értékével mérik (több független változó együttes hatását méri).

Parciális regresszió

A parciális regresszió arra az esetre vonatkozik, amikor azt szeretnénk vizsgálni, hogy milyen kapcsolat van két változó között akkor, ha egy vagy több másik változót állandó szinten tartunk (az előző példánknál maradva, ha megegyezik az életkor, a szabadidő mennyisége és az alkoholfogyasztás között megmarad-e az összefüggés). A változók közötti összefüggést leíró egyenletet úgy számoljuk ki, hogy állandó szinten tartjuk a kontrollváltozókat, és az így kapott eredményt összevetjük a két változó közötti eredeti kapcsolattal. A parciális regressziót a *parciális korrelációs együtthatóval* mérjük.

Nem lineáris regresszió

Empirikus vizsgálatok esetén nem feltételezhetjük, hogy minden változócsoportban lineáris összefüggések volnának. Sokszor előfordul, hogy egy görbe

vonalú regresszióval jobban magyarázhatóak az adatok, mint bármilyen lineáris modellel, ugyanakkor a regressziós modellek kettős funkciójából következik az is, hogy bár egy bonyolult egyenlettel a kapcsolat tökéletesen leírhatóvá válik, de nem használható szinte semmiféle következtetésre. Általában a regresszióelemzés extrapolációra való felhasználása nem igazán megbízható.

2. Útelemzés

Az útelemzés oksági modell a változók közötti kapcsolatok megértéséhez. A regressziószámításon alapul, de szemléletesebb képet ad több változó kapcsolatáról. Abból indul ki, hogy egyik változó értékeit más változók értékei okozzák, tehát elengedhetetlen a függő és független változók megkülönböztetése. Útelemzés révén grafikusan megjeleníthető a változók közötti összefüggések hálózata a kapcsolat erősségének feltüntetésével. A kapcsolaterősségeket *parciális regresszióelemzés* alapján számítják ki. Az útegyütthatók (*path coefficients*) két változó kapcsolatát mutatják úgy, hogy a modellben szereplő összes többi változót konstans szinten tartjuk. Az útelemzés kiváló módja a változók közötti komplex oksági láncok és hálózatok kezelésének, de *az okság rendjét nem az útelemzés, hanem a kutató mondja meg.* A kutató határozza meg a változók közötti lehetséges kapcsolatok szerkezetét, a számítógép csak az útegyütthatókat számolja ki.

3. Idősorok elemzése

Gyakran használunk regressziószámítást idősoros adatok elemzésére, amikor az egyes változók időbeli alakulását, változását kívánjuk vizsgálni. Az idősorelemzés *hosszú távú trendek kifejezésére*, egy trend magyarázatára adott hipotézisek tesztelésére, valamint a jövőben várható változások előrejelzésére is alkalmas. Szintén *parciális regresszión* alapszik, amikor az idő (év, hónap, perc stb.) változó az elemzési egység. Az idősoros összefüggések sokszor nagyon bonyolultak, ilyenkor használatos az *időeltolásos regresszióelemzés*, amikor az idő változó egy korábbi értékét (pl. előző év) tekintjük alapnak, és ez alapján becsüljük valamely változó alakulását. A társadalomban előforduló számos oksági viszonyt ilyen időeltolás jellemez. A különböző előforduló esetekben sokféle regressziós egyenlet képzelhető el, de az idősorok elemzésénél a lényeg mindig az, hogy a kutatónak mennyire sikerült megmagyaráznia a függő változó megfigyelt értékeit.

4. Faktorelemzés *(részletesebben lásd az 5.3-as alfejezetet)*

A faktoranalízis lényegesen eltér a regresszióelemzéstől. Statisztikai alapjai elég bonyolultak, és különböznek az eddig tárgyaltakétól. A faktorelemzés arra szolgál, hogy mintázatokat fedezzünk fel egy nagyobb változórendszerben. A faktoranalízis tulajdonképpen úgy történik, hogy olyan mesterséges dimenziókat, faktorokat hozunk létre, amelyek erősen korrelálnak egy sor megfigyelt változóval, és amelyek egymástól függetlenek. Minden faktorhoz hozzátartoznak a megfelelő faktorsúlyok, amelyek az egyes változók és az egyes faktorok közötti

korrelációk. A faktorelemzés a gyakorlatban úgy történik, hogy számos változóból kapunk néhány faktort a megfelelő faktorsúlyokkal, majd a kutatónak kell meghatároznia az egyes faktorok jelentését aszerint, hogy az illető faktornál mely változók szerepelnek nagy súllyal. A faktorok kialakításánál a számítógép csak két szempontot vesz figyelembe: 1. a faktor magyarázza meg a vizsgált változók összes varianciájának viszonylag nagy hányadát, 2. minden faktor legyen teljesen korrelálatlan a többi faktorral.

A módszer előnyei:
– a faktorelemzés hatékony módszer nagyszámú változó fő összefüggéseinek vizsgálatára,
– számos többszörös, egyszerű és parciális korreláció egybevetése helyett a számítógép végzi el a faktorelemzést,
– a faktorelemzés eredményei könnyen értelmezhetőek: az alapján, hogy egy adott faktornál mely változók szerepelnek nagy súllyal, megállapítható, hogy miként csoportosulnak a változók,
– az is könnyen megállapítható, hogy egy adott változó mely faktorokkal korrelál jelentős mértékben, és melyekkel nem.

A módszer hátrányai:
– az elemzés a tényleges jelentésre való tekintet nélkül állítja elő a faktorokat,
– faktorokat mindig létre lehet hozni, de ezek létezése egyáltalán nem garancia arra, hogy értelmük is van.

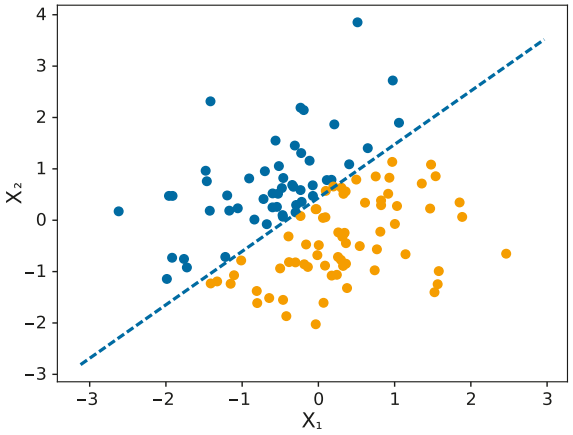
5. Klaszterelemzés *(részletesebben lásd az 5.4-es alfejezetet)*

A társadalomtudományokban az egyének, intézmények, települések vagy országok hasonlósága általában nem egyetlen, hanem számos ismérv/változó alapján állapítható meg (pl. országok esetén hasonló nagyságú az egy főre jutó GDP, a gazdaság növekedése, a munkanélküliségi ráta, a születéskor várható átlagos élettartam, az iskolázottság stb.). A klaszterelemzés (klaszter = csoport, angolul: *cluster*) célja előre nem ismert csoportok képzése, keresése, a keresés eredménye pedig a különböző homogén csoportok létrehozása. A klaszteranalízis tehát egy vizsgált sokaság egyedeinek csoportokba való sorolását jelenti, figyelembe véve az egyes egyedeknek egy bizonyos ismérvrendszerben felvett értékeit. Az elemzés nem tesz különbséget függő és független változó között, és a változókon belüli kölcsönös összefüggést vizsgálja. A klaszterbe helyezés legelterjedtebben a megfigyelési egységek páronkénti *távolságának* használatával történik. Az egy csoportba került egységek értelmezése ennél az eljárásnál is a kutató feladata.

6. Diszkriminanciaanalízis és logisztikus regresszió

A diszkriminanciaanalízis olyan adatelemzési módszer, amelyet kategóriába tartozás előrejelzésére lehet használni, és amelynél 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 fogyaszt alkoholt, vagy nem, mekkora mértékben magyarázza az életkor, jövedelem stb.). Az előbb ismertetett lineáris regresszióhoz hasonlóan a diszkriminanciaanalízisben is egyenest illesztünk: olyan egyenest keresünk, amely a legjobban szétválasztja az elemzendő csoportokat (68. ábra).



68. ábra. *A diszkriminanciaanalízis tipikus modellje*

A diszkriminanciaelemzés altervatívája a *logisztikus regresszió*, amelynek alkalmazási előfeltételei sokkal kevésbé szigorúak. Logisztikus regressziót akkor használunk, ha a megmagyarázni kívánt függő változónk kétértékű (bináris, dichotóm vagy dumy változó), a magyarázó, független változóink pedig mennyiségi vagy kategoriális változók.

5.2. A többváltozós lineáris regresszió

A lineáris regresszióelemzés tehát (ahogyan az *5.1-es alfejezet*ben ismertettük) egy magyarázó modell. Ebben egy skálamérési szintű függő változó viselkedését olyan független változókkal magyarázzuk, amelyek szintén magas mérési szintűek. Legegyszerűbb formája, amikor csak egy magyarázó változónk van, ezért a kétváltozós elemzéseknél ismertetett *korreláció* (*4.3. alfejezet*) kiegészítéseként is lehet használni.

A lineáris regressziós modellben arra a kérdésre keressük a választ, hogy melyik az a lineáris függvény, amelynek segítségével egy (X) vagy több (X_j) magyarázó változó értékét megismerve becslést tehetünk egy másik változó (Y) értékére.

A lineáris regresszió az a statisztikai eljárás, amellyel megtalálhatjuk a két vagy több változóra együttesen felvett értékekhez (a pontdiagram pontjaira) leg-

jobban illeszkedő egyenest. A legjobb egyenes az lesz, amellyel az előrejelzés hibája a legkisebb: erre a legalkalmasabb a legkisebb négyzetek módszere (*Ordinary Least Squares* vagy *OLS*). Az OLS az egyenes paramétereit úgy becsüli meg, hogy a tényleges megfigyeléseket jelentő pontok és az egyenes közötti távolságok négyzetösszege minimális legyen (mindegyik értékhez a lehető legközelebb legyen).

A regressziós egyenlet leíró és következtetési szempontból is hasznos:

1. megkapjuk a két vagy több változó közti kapcsolat matematikai leírását,
2. lehetőségünk van arra, hogy az X_i ismeretében következtessünk *Y* értékére úgy, hogy kiszűrjük belőle a többi magyarázó változó hatását (ha több magyarázó változós modellünk van).

A regressziós egyenes egyenlete:

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + ... + B_iX_i + e_i$$

ahol:

$\hat{Y}$ – *a függő változónak az egyenes alapján becsült értéke*,

$X_1, X_2, \ldots, X_i$ – a *magyarázó változók*,

B_0 – *az egyenes magassága, konstans*: ahol a regressziós egyenes metszi az Y tengelyt (a függő változó átlagos szintje, ahol a magyarázó változó értéke 0 – nem mindig van értelme),

$B_1, B_2, \ldots, B_i$ – az egyenes meredeksége vagy *regressziós együttható*: azt mutatja meg, hogy a többi változót állandó szinten tartva mennyi az adott változó hatása (ha az adott magyarázó változó egy egységgel nő, mennyivel változik átlagosan a függő változó értéke úgy, hogy közben a többi változó értékét nem változtatjuk): ha pozitív, növekvő függvény, ha negatív, csökkenő függvény,

e_i – *reziduum* vagy hibatényező (*reziduális*) a becsült és megfigyelt *Y*-érték közötti különbség.

A lineáris regresszióelemzésben lehetőségünk van arra is, hogy kategoriális (alacsony mérési szintű változókat) vonjunk be az elemzésbe. Ennek legegyszerűbb módja a kétértékű (bináris, dummy, alternatív) változók alkalmazása. Ez egy mesterséges (0–1 értékű) változó, amivel egy kétértékű (bináris) kategóriát kódolunk regresszióban: 1 akkor, ha az adott megfigyelés a kategoriális változó egyik kategóriájába esik, és 0 akkor, ha a másikba. A B_0 együttható ilyenkor azt mutatja meg, hogy mekkora a függő változó átlaga a 0-val kódolt kategóriában, a B_1 együttható azt mutatja meg, hogy mekkora a különbség a két kategória függő változóra vonatkozó átlaga között. Egy dummy magyarázó és egy skálafüggő változó esetén a kétváltozós elemzéseknél ismertetett *t-teszt* (*4.4. alfejezet*) eredményeit és annak kiegészítéseit kapjuk.

A többváltozós lineáris regresszióelemzés lépései:

1. az elemzés céljának megfogalmazása, a vizsgálatba bevont változók,
2. az adatok előkészítése,
3. a regresszió lefuttatása és a becslési (változó-szelektálási) módszer meghatározása,
4. a lineáris regresszióelemzés alkalmazhatósági feltételeinek vizsgálata,
5. a regressziós modell értelmezése.

Többváltozós lineáris regresszió az SPSS-sel

1. Az elemzés céljának megfogalmazása, a vizsgálatba bevont változók

A többváltozós lineáris regresszió első lépése annak tisztázása, hogy mit kívánunk magyarázni vagy előre jelezni. Meg kell határozni a függő változót (Y), amelynek értékeit a modell becsülni fogja. Meg kell indokolni, miért választottuk ezt a változót függő változónak, és milyen gyakorlati vagy elméleti probléma megoldására törekszünk vele (pl. alkoholfogyasztás magyarázata).

A következő lépés a magyarázó (független) változók (X_i) kiválasztása. Ezek olyan tényezők, amelyek várhatóan hatással vannak a függő változóra. A változók lehetnek: skálaváltozók (pl. életkor), kategóriás változók, amelyek dummy változóként kerülnek be (pl. nem, iskolai végzettség). A változók kiválasztásánál fontos szempont:

1. elméleti indokoltság (szakirodalmi előzmények: korábbi kutatások, hipotézisek),
2. statisztikai szempontok: nem lehet túlzott multikollinearitás (két vagy több magyarázó változó közötti erős lineáris kapcsolat, ami megnehezíti az egyes változók egyedi hatásának megbízható becslését), és a minta méretéhez illeszkedjen a változók száma (általános szabály szerint 15–20 eset magyarázó változónként, vagyis pl. 5 magyarázó változó esetén legalább 100 esetes minta).

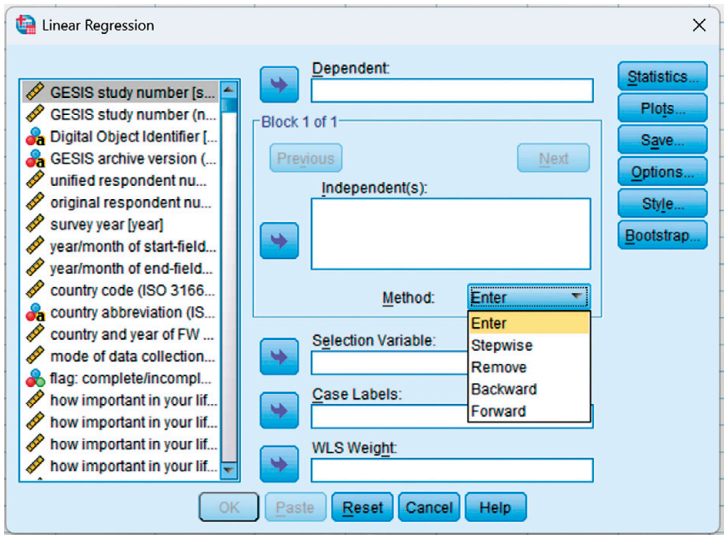
2. Az adatok előkészítése

Itt a függő és független változók adatbázisban való azonosítása történik. Kategóriás magyarázó változók esetén létre kell hozni a dummy változókat, ha szükséges. Ekkor történik a hiányzó értékek és a szélsőséges értékek ellenőrzése. Erre két egyszerű módszer is van az SPSS-ben:

1. *ANALYZE* főmenü, *Descriptive Statistics, Explore, Statistics, Outliers* úton, itt a *Boxplot* és *Stem-and-leaf* ábrák mutatják a kiugró értékeket.
2. Standardizált érték vagy z-érték, Z-score számítás segítségével: *ANALYZE* főmenü, *Descriptive Statistics, Descriptives, Save standardized values as variables.* $|Z| > 3{,}29$ esetén szélsőséges értékünk van. A z-érték azt mutatja meg, hogy egy adott érték hány szórásnyi távolságra van az átlagtól – pozitív szám esetén az átlag felett, negatív esetén az átlag alatt helyezkedik el. Számítása: (érték - átlag) / szórás.

3. A lineáris regresszió lefuttatása és a becslési (változó-szelektálási) módszer meghatározása

A lineáris regresszió lekérése és ezen belül a módszer kiválasztása az *ANALYZE* főmenü, *Regression, Linear…* almenüben történik. A *Dependent* mezőbe kerül a függő változó, az *Independent(s)* mezőbe a magyarázó változók. A magyarázó változók alatt a *Method*-nál megjelenő legördülő menüben (69. ábra) választjuk ki a módszert.



69. ábra. *A többváltozós lineáris regresszió módszerének kiválasztása*

Az SPSS-ben öt módszer is van a többváltozós lineáris regresszió lefuttatására, ezek közül a leggyakrabban (ez a program alapbeállítása is) az *Enter* módszert szokás használni. A módszerek:

1. *Enter* (bevitel): minden előre kiválasztott változót egyszerre épít be a modellbe, függetlenül azok statisztikai szignifikanciájától. Ez a legegyszerűbb módszer, ahol a kutató elméletileg indokoltan határozza meg a változókat, és nem alkalmaz automatikus szelekciót.
2. *Stepwise* (lépcsőzetes): a *Forward* és *Backward* módszerek kombinációja – minden lépésben megvizsgálja, hogy kell-e változót hozzáadni vagy eltávolítani a modellből. Ez a legflexibilisebb módszer, mivel egy korábban bevont változót később eltávolíthatunk, ha más változók jelenléte miatt már nem szignifikáns.
3. *Remove* (eltávolítás): kutatói kontroll alatt álló módszer, ahol előre meghatározott változókat távolít el a modellből, hogy vizsgálni lehessen azok hatását a modell teljesítményére és a többi változó magyarázóerejére.

4. *Backward* (hátramutató): a teljes modellből indul (minden változóval), majd lépésenként eltávolítja a legkevésbé szignifikáns változókat. A folyamat akkor fejeződik be, amikor minden modellben maradó változó szignifikáns ($p < 0,05$), vagy elérte a meghatározott eltávolítási kritériumot.
5. *Forward* (előremutató): üres modellből (csak konstans) indul, majd lépésenként hozzáadja azt a változót, amely statisztikailag a legszignifikánsabb javulást eredményezi a modellben. A folyamat akkor áll meg, amikor egyetlen változó hozzáadása sem javítja szignifikánsan ($p < 0,05$) a modell illeszkedését.

4. A lineáris regresszió alkalmazhatósági feltételeinek vizsgálata

A többváltozós lineáris regressziónak két további főbb alkalmazhatósági feltételét szükséges megvizsgálnunk az SPSS-ben.

1. A magyarázó változókra vonatkozó feltétel: ***multikollinearitás***

A magyarázó változókra vonatkozó legfontosabb feltétel, hogy egymástól lineárisan függetlenek kell legyenek, vagyis egyik magyarázó változót se lehessen a többi magyarázó változó lineáris kombinációjaként előállítani (multikollinearitás). Ha káros multikollinearitás lép fel, megkeressük azokat a magyarázó változókat, amelyek a zavart okozzák és elhagyjuk őket a modellből, vagy ha több ilyen van, akkor inkább faktorelemzést érdemes használni.

A multikollinearitás tesztelése:

a) korrelációs együtthatóval: két magyarázó változó közötti páronkénti Pearson-féle korrelációs együttható nem szabad nagyobb legyen 0,7-nél: *ANALYZE, Correlate, Bivariate*,
b) VIF mutatóval: a VIF értéke ne legyen 2-5-nél nagyobb: *ANALYZE, Regression, Linear, Statistics, Collinearity diagnostics*.

A VIF (*Variance Inflation Factor* – variancianövelési tényező) megmutatja, hogy egy adott magyarázó változó varianciája mennyivel nagyobb a multikollinearitás miatt, mint amennyire egyedül lenne. Értékelése:

VIF = 1: nincs multikollinearitás,
1 < VIF ≤ 2: gyenge multikollinearitás (elfogadható),
2 < VIF ≤ 5: erős zavaró multikollinearitás,
VIF > 5: nagyon erős, káros multikollinearitás.

2. A hibatagra (e_i, reziduális) vonatkozó feltétel: ***homoszkedaszticitás***

A lineáris regresszióban a hibatagra (reziduálisok) vonatkozó alapfeltétel, hogy varianciája konstans, állandó minden *X*-érték mellett, a becsült értékek teljes tartományában. Ez azt jelenti, hogy a függő változó körüli szóródás egyforma kell legyen minden független változóra felvett érték esetén, vagyis a reziduálisok „sávja” egyenletesen széles kell legyen a regressziós egyenes körül. Lényegét tekintve azt mutatja, hogy a modell egyformán jól illeszkedik az adatok minden tartományában. Ha a homoszkedaszticitás feltétele nem teljesül és heteroszke-

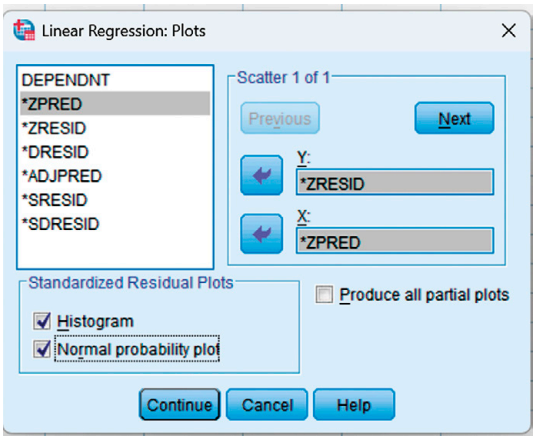
daszticitás lép fel, a lineáris regresszióval mért becslésünk nem lesz pontos, és más becslési módszert (más típusú regressziót) kell használni, vagy pl. logaritmizálni kell (*Compute* almenüvel) az adatokat.

A homoszkedaszticitás tesztelése az SPSS-ben grafikusan történik (70. ábra):

a) a hibatagok normális eloszlásának ellenőrzése *hisztogram*mal: a standardizált reziduálisoknak szimmetrikusan és sűrűn kell elhelyezkedniük a ráillesztett Gauss-görbe (a standard normális eloszlás középértéke 0, szórása 1) körül, követve annak alakját: *ANALYZE, Regression, Linear, Plots: Histogram*,

b) a hibatagok normális eloszlásának ellenőrzése *Normal P-P Plot*-tal: a standardizált reziduálisoknak szorosan a 45 fokos átlós egyenes mentén kell elhelyezkedniük, minimális eltérésekkel (a megfigyelt és az elméleti vagy normális kumulált eloszlás megegyezik): *ANALYZE, Regression, Linear, Plots: Normal probability plot*,

c) a becsült hibatagokat (*ZRESID*) a becsült értékek ($\hat{y}$) függvényében (*ZPRED*) ábrázoljuk standardizált formában: a hibatagoknak véletlenszerűen, egyenletesen kell szóródniuk a nullavonal körül, anélkül, hogy bármilyen rendszert (tölcsér, ív vagy más mintázat) mutatnának: *ANALYZE, Regression, Linear, Plots, Scatter 1 of 1*-nél az X tengelyre a *ZPRED*, az Y tengelyre a *ZRESID*.



70. ábra. *A homoszkedaszticitás grafikus tesztelése*

A homoszkedaszticitást más statisztikai tesztekkel is lehet ellenőrizni, mint pl. a Goldfeld–Quandt-féle teszt, de erre az SPSS-ben nincsen beépített funkció, ezért manuálisan vagy más statisztikai programban, pl. R-ben kell elvégezni.

5. A regressziós modell értelmezése

Alapbeállításban a többváltozós lineáris regresszió eredményeit négy táblázatban látjuk.

Az 1. táblázat a modellbe bevitt és kivett változókat (*Variables Entered/Removed*) és a választott módszert (pl. *Enter*) mutatja.

A 2. táblázat (*Model Summary*) a modell magyarázóerejét mutatja. Itt láthatjuk az alábbi mutatókat:

R – korrelációs együttható,

R square – determinációs együttható, vagyis hogy a függő változó hány százalékát magyarázza a regressziós modell,

Adjusted R square – az alapsokaság-beli megmagyarázott hányad torzítatlan becslése, ezt vesszük figyelembe a modell illeszkedésének, magyarázóerejének vizsgálatakor: egy 0 és 1 közötti értékeket felvevő mutató, pl. egy 0,25-ös érték azt jelenti, hogy a független változók együtt 25%-ot magyaráznak a függő változó varianciájából,

Std. Error of the Estimate – a reziduálisok szórása, szintén illeszkedésmérő (minél nagyobb, annál több a függő változóra felvett olyan érték, ami távol esik a regressziós egyenlettel becsült értéktől).

A 3. táblázat (*ANOVA*) a modell szignifikanciáját mutatja, vagyis azt, hogy sikerült-e akkora részt megragadni a függő változó varianciájából, hogy a független változó hatását szignifikánsnak tekinthessük. Ebben a táblázatban az F-érték szignifikanciaszintje (*Sig.)* kisebb kell legyen, mint 0,05 ahhoz, hogy szignifikáns legyen a modellünk.

Végül az utolsó, 4. táblázatunk tartalmazza a regressziós együtthatókat (*Coefficients*). Ebben a mutatók értelme:

(Constant) – a B_0 vagy konstans, a regressziós egyenes magassága: ahol a regressziós egyenes metszi az Y tengelyt (a függő változó átlagos szintje, ahol a magyarázó változó értéke 0),

B – standardizálatlan regressziós együttható, megmutatja, hogy a függő változó hány egységgel változik (pozitív *B*-érték estén nő, negatív *B*-érték esetén csökken), ha az illető független változó értéke 1 egységgel nő, miközben az összes többi változó értéke állandó marad és az eredeti mértékegységekben van kifejezve (pl. ha a függő változó lej, akkor *B* is lejben van megadva),

Beta – béta-súly vagy standardizált regressziós együttható (kétváltozós regressziónál a korrelációs együttható): azt mutatja meg, hogy melyik független változó hatása erősebb,

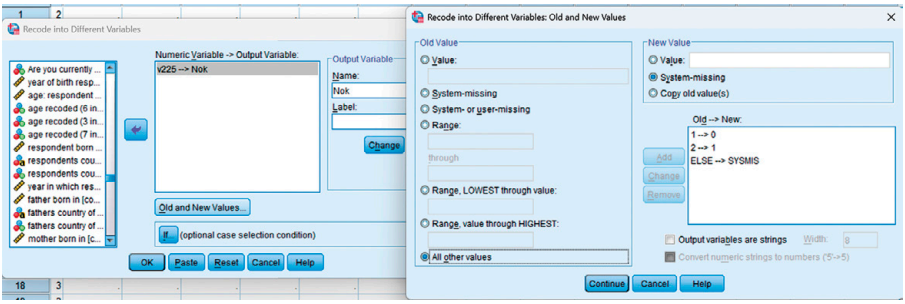
t-érték és a neki megfelelő szignifikancia *(Sig.)* – azt mutatja, hogy az illető független változó hatása szignifikáns-e a függő változóra.

45. példa ▼

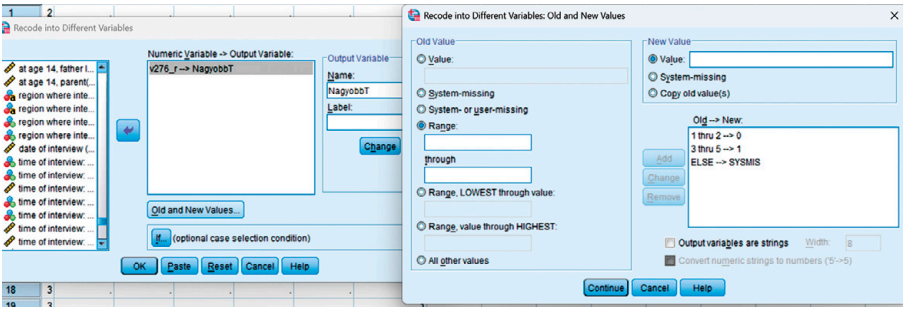
▶ *Többváltozós lineáris regresszió az SPSS-ben*

Ahogyan már a korreláció számításakor (44. példa) is jeleztük, adatbázisunkban csak két magas mérési szintű változónk van, ezért a lineáris regressziós modellünkben a nappali iskolai tanulmányok befejezésének életkora (*v242*, *age completed education respondent (Q80)*, az adatbázisban a 291-es sorszámú változó, a K62-es kérdés) lesz a függő változónk. A másik skála változónk, a születési év (*v226*, az adatbázisban a 266-os sorszámú változó, a K58-as kérdés), illetve további két kategoriális változó, a nem (*v225*, az adatbázisban a 265-ös sorszámú változó, K57-es kérdés) és a település lélekszáma (*v276_r*, K79-es kérdés, a 377-es sorszámú változó az adatbázisban) lesz a három magyarázó változónk a regressziós modellben. A regressziós modellel azt vizsgáljuk, hogy a három független változóval (születési év, nem és település nagysága) mennyiben magyarázható a formális oktatás befejezésének életkora. A függő változónk az iskolai végzettség egy helyettesítő (*proxy*) változója, amit skála változóként lehet mérni: a nagyobb formális oktatás befejezési életkor magasabb iskolai végzettségi szinttel társul ($r = 0{,}731$, $p = 0{,}000$). Ezért azt feltételezzük (az a hipotézisünk), hogy a fiatalabbak, a nők és a nagyobb településeken élők magasabb életkorban fejezik be tanulmányaikat (több időt szánnak tanulásra), mint az idősebbek, férfiak és kisebb településeken élők.

Első lépésben hozzuk létre a két kategoriális változóból a két dummy változót (*TRANSFORM, Recode into Different Variables*). A nem változó esetében (*v225*) az új változóban (*Nok*) a nő legyen az 1-es, a férfi a 0-ás kategória, minden más érték (*All other values*) legyen *System-missing* (71. ábra). A település nagyságánál (*v276_r*) az új változóban (*NagyobbT*) a 20 000 fő feletti település legyen az 1-es (régi 3-5 kategóriák) és az ez alatti településnagyság a 0-ás kategória (régi 1-2 kategóriák), minden más érték pedig hiányzó adat (72. ábra). Definiáljuk is a változókat az átkódolásnak megfelelően (tizedesek, változó- és értékcímkék).



71. ábra. *Dummy változó létrehozása: Nők (45. példa)*



72. ábra. *Dummy változó létrehozása: Nagyobb településen élők (45. példa)*

Átkódolás után ellenőrizzük adatainkat egy gyakorisággal (577 nő és 441 nagyobb településen élő kérdezett) (73. ábra).

Nők

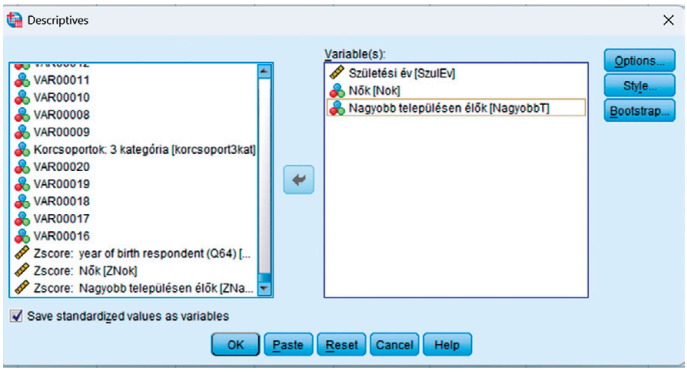
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	férfi	529	47.8	47.8	47.8
	nő	577	52.2	52.2	100.0
	Total	1106	100.0	100.0	

Nagyobb településen élők

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	kisebb településen él	665	60.1	60.1	60.1
	nagyobb településen él	441	39.9	39.9	100.0
	Total	1106	100.0	100.0	

73. ábra. *Az új dummy változók gyakoriságai (45. példa)*

Mielőtt továbblépnénk, készítsünk egy másolatot a *születési év* változóról (*v226*), és nevezzük el *SzulEv*-nek (*TRANSFORM, Recode into Different Variables: All other values – Copy old Value(s)-el*). A szélsőséges értékek tesztelésére a Z-score módszert használjuk, először létrehozva a standardizált változókat a korábban leírtak szerint (74. ábra).



74. ábra. *A magyarázó változók standardizálása (45. példa)*

Az új standardizált változókra leíró statisztikákat kérünk (*ANALYZE, Descriptive Statistics, Descriptives*), és azt látjuk, hogy nincsenek kiugró adataink, $|Z| < 3{,}29$ (75. ábra).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Zscore: Születési év	1106	-1.74173	1.78371	.0000000	1.00000000
Zscore(Nok) Nők	1106	-1.04391	.95707	.0000000	1.00000000
Zscore(NagyobbT) Nagyobb településen élők	1106	-.81398	1.22743	.0000000	1.00000000
Valid N (listwise)	1106				

75. ábra. *A standardizált változók leíró statisztikái (45. példa)*

Az *Enter* módszer mellett döntünk, és rátérünk a regressziós modell alkalmazhatósági feltételeinek vizsgálatára. Elsőként a multikollinearitást teszteljük: megnézzük a páronkénti korrelációs együtthatókat (76. ábra), majd a VIF mutatót (77. ábra), a korábbiakban leírtak szerint lekérve (a regressziós együtthatók és a modell illeszkedése nélkül). Azt látjuk, hogy egyetlen szignifikáns korreláció sincs a magyarázó változók között, illetve a VIF értéke is az 1-es érték körül mozog, tehát a multikollinearitást kizárhatjuk.

Correlations

		Születési év	Nők	Nagyobb településen élők
Születési év	Pearson Correlation	1	-.013	-.026
	Sig. (2-tailed)		.669	.385
	N	1106	1106	1106
Nők	Pearson Correlation	-.013	1	.000
	Sig. (2-tailed)	.669		.993
	N	1106	1106	1106
Nagyobb településen élők	Pearson Correlation	-.026	.000	1
	Sig. (2-tailed)	.385	.993	
	N	1106	1106	1106

76. ábra. *A magyarázó változók közötti korrelációk (45. példa)*

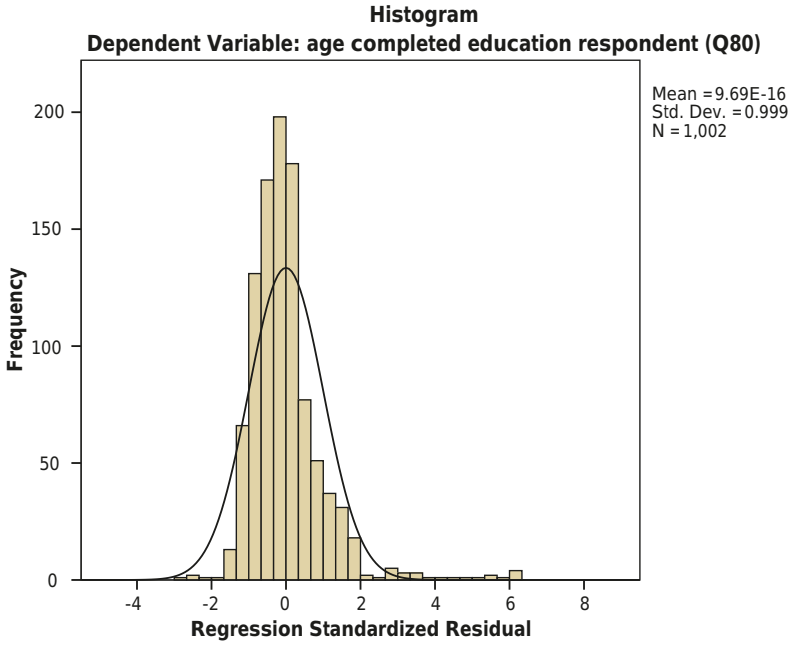
Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Születési év	.994	1.006
	Nők	1.000	1.000
	Nagyobb településen élők	.995	1.005

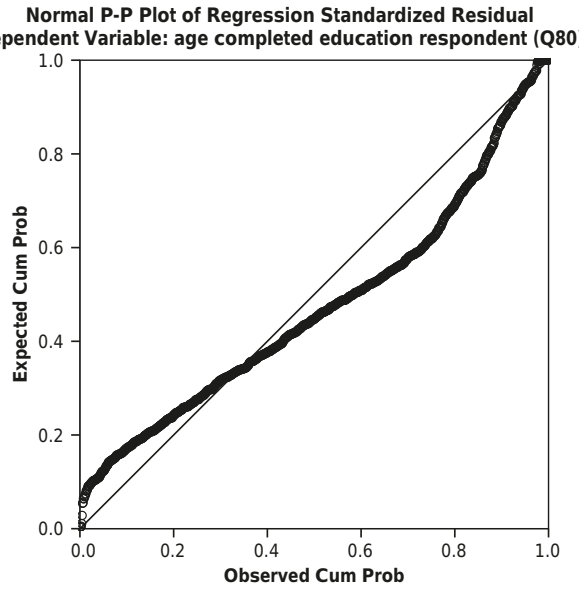
a) Dependent Variable: age completed education respondent (Q80)

77. ábra. *A VIF mutatók (45. példa)*

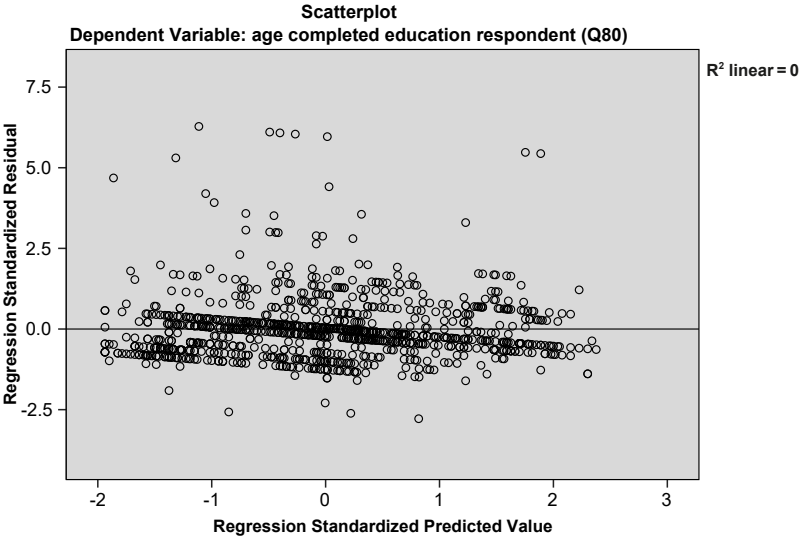
A homoszkedaszticitás tesztelésére az előzőekben leírtak szerint lekérjük mindhárom ábrát: a hisztogramot (78. ábra) a *Normal P-P Plot*-ot (79. ábra), és a standardizált becsült hibatagokat (*ZRESID*) a becsült értékek függvényében (*ZPRED*) (80. ábra). Mindhárom ábra azt jelzi, hogy a hibatagok közelítőleg normális eloszlást követnek, vagyis a homoszkedaszticitás feltétele teljesül.



78. ábra. *A standardizált reziduálisok szóródása: hisztogram (45. példa)*



79. ábra. *A standardizált reziduálisok szóródása: Normal P-P Plot (45. példa)*



80. ábra. *A standardizált becsült hibatagok és értékek pontdiagramja (45. példa)*

Végül pedig nézzük a többváltozós lineáris regressziós modell eredményét. Az első táblázatból annyit látunk, hogy a modellben 3 magyarázó változó

(*Születési év, Nők* és *Nagyobb település*) szerepel, amelyeket Enter módszerrel (egyszerre) vittünk be a modellbe. A függő változó a formális tanulmányok befejezésének életkora (*age completed education respondent*). A modell magyarázóereje (*Adjusted R Square*) 0,073, tehát a modell a függő változó varianciájának körülbelül 7,3%-át magyarázza meg, vagyis a három bevont magyarázó változó csak kismértékben járul hozzá a függő változó alakulásának magyarázatához, a variancia döntő része (92,7%) továbbra is a modell által nem magyarázott tényezőknek vagy véletlen hibának tudható be (81. ábra).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.270[a]	.073	.070	3.873

a) Predictors: (Constant), Nagyobb településen élők, Nők, Születési év
b) Dependent Variable: age completed education respondent (Q80)

81. ábra. *A regressziós modell magyarázóereje (45. példa)*

Az alacsony magyarázóerő ellenére a regressziós modell szignifikáns (F = 26,222, p = 0,000) (82. ábra).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1180.286	3	393.429	26.222	.000[b]
	Residual	14973.818	998	15.004		
	Total	16154.104	1001			

a) Dependent Variable: age completed education respondent (Q80)
b) Predictors: (Constant), Nagyobb településen élők, Nők, Születési év

82. ábra. *A regressziós modell szignifikanciája (45. példa)*

Végül az együtthatókat bemutató táblázat (83. ábra) alapján megállapítható, hogy a *Nők* változó szignifikanciaszintje meghaladja a 0,05-ös küszöbértéket (p = 0,111), ezért a válaszadó neme nem befolyásolja szignifikánsan a formális oktatás befejezésének életkorát. A másik két magyarázó változó viszont szignifikáns hatással bír: mindkét regressziós együttható pozitív, ami azt jelzi, hogy növelik a formális képzés befejezésének életkorát úgy is, ha a többi magyarázó változó értékét állandó szinten tartjuk. Konkrétan: az egy évvel későbbi születés átlagosan 0,41 évvel (B_1), míg a nagyobb településen való lakhely 1,7 évvel (B_3) hosszabbítja meg a formális oktatásban eltöltött időt. A település nagyságának modellben való hangsúlyosabb szerepét a béta együtthatók is jelzik (a születési év esetében 0,178, míg a nagyobb településnél 0,210). A modellben a konstans (B_0) nem értelmezhető (nincs 0-ás születési év).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-61.649	13.739		-4.487	.000
	Születési év	.041	.007	.178	5.831	.000
	Nők	-.390	.245	-.049	-1.594	.111
	Nagyobb településen élők	1.729	.251	.210	6.882	.000

a) Dependent Variable: age completed education respondent (Q80)

83. ábra. *A regressziós modell együtthatói (45. példa)*

Összességében tehát a többváltozós lineáris regressziós modell magyarázóereje alacsony, vagyis a bevont tényezők a függő változó varianciájának mindössze 7,3%-át magyarázzák, a fennmaradó hányad más, a modellben nem szereplő tényezőknek vagy véletlen hatásoknak tulajdonítható. Eredeti hipotézisünk, hogy a fiatalabbak, a nők és a nagyobb településeken élők magasabb életkorban fejezik be tanulmányaikat (több időt szánnak tanulásra), mint az idősebbek, férfiak és kisebb településeken élők, csak részben igazolódott be, mivel a nem változó nem bizonyult szignifikáns magyarázó tényezőnek. Továbblépésként érdemes lenne további, releváns változókat bevonni a modellbe (pl. szülők iskolai végzettsége, családi háttér, munkaerőpiaci helyzet), hogy pontosabb képet kapjunk a formális oktatás befejezésének életkorát befolyásoló tényezőkről.

5.3. A faktorelemzés

A faktorelemzés egy gyűjtőfogalom, amely a többváltozós elemzések egy csoportjára vonatkozik. A faktorelemzést arra használjuk, hogy adatainkat tömörítsük, vagy hogy nagyszámú változó mintázatát, belső struktúráját feltárjuk. A faktorelemzés célja, hogy sok, általunk mért változót úgynevezett faktorváltozókba (háttérváltozókba) vonjon össze, amelyek közvetlenül nem figyelhetők meg. A vizsgálatba bevont változók legalább ordinális mérési szintűek kell legyenek, és egymással korrelálniuk kell (ha nincs közöttük összefüggés, nem érdemes tömöríteni őket).

A faktoranalízis tehát olyan adatredukciós eljárás, amellyel az egymással lineáris összefüggésben lévő változók közös lényegét kifejező faktorok tárhatók fel. Az elemzés azt feltételezi, hogy a változók hátterében olyan nem mérhető, látens struktúrák állnak, melyeket e módszerrel kiragadva kis információveszteséggel leírható az adathalmaz. Az analízis során kapott faktorok száma lénye-

gesen kevesebb, mint az eredeti változóké, és ha ezekkel szeretnénk dolgozni, tudnunk kell, hogy milyen következményekkel jár az adatredukciónk. A két csoport illeszkedését két korrelációs mátrix összehasonlításával mérjük, melyek egyformaságának megítélésére kiválóan alkalmas a λ^2 próba. A faktoranalízisnek ez a variációja *exploratív* (feltáró) jellegű, hiszen sok mért változóból kevés ismeretlen aggregált változót hoz létre, míg a *konfirmatív* (megerősítő) elemzés egy előzetes hipotézis (korábban talált faktorok) tesztelésére alkalmas. A konfirmatív faktorelemzés sokkal komplexebb, ezért a továbbiakban ezzel nem foglalkozunk.

Az exploratív faktorelemzés folyamata:

1. az elemzés céljának megfogalmazása, a vizsgálatba bevont változók,
2. az adatok előkészítése,
3. a faktorelemzés lefuttatása és a faktorelemzés módszerének meghatározása,
4. a faktorelemzés alkalmazhatóságának vizsgálata,
5. a faktorok/főkomponensek számának meghatározása,
6. a faktorok értelmezése,
7. értelmezés rotálással,
8. a faktorváltozók elmentése,
9. további felhasználás.

A faktorelemzés folyamatát az SPSS-programcsomag használatával mutatjuk be.

Faktorelemzés főkomponensmódszerrel az SPSS-ben

1. Az elemzés céljának megfogalmazása, a vizsgálatba bevont változók

A faktorelemzés első lépése a kutatás céljának pontos meghatározása. Ez azt jelenti, hogy tisztáznunk kell, miért szeretnénk faktorelemzést végezni: például a változók mögött meghúzódó rejtett dimenziók, faktorok azonosítása, a változók számának csökkentése vagy a mintázatok feltárása érdekében. Ezt követően ki kell választani és röviden bemutatni azokat a változókat, amelyeket a vizsgálatba bevonunk. A változók jellemzően azonos mérési szintűek: *ordinális (Likert-skálák), intervallum- vagy arányskálák,* és olyan tulajdonságokat, attitűdöket, véleményeket mérnek, amelyek feltételezhetően összefüggenek egymással.

A változók kiválasztásánál, ahogyan a többváltozós lineáris regressziónál is láttuk, két fontos szempontot kell figyelembe venni:

1. elméleti indokoltság (szakirodalmi előzmények: korábbi kutatások, hipotézisek),

2. statisztikai szempontok: a változóknak összefüggést kell mutatniuk egymással, különben nem lesz értelme közös faktorokat keresni, és a minta méretéhez illeszkedjen a változók száma (általános szabály szerint 5-10 eset változónként, vagyis pl. 10 változó esetén 100 esetes minta javasolt).

Vegyünk két konkrét fiktív példát faktorelemzésre, a munkahelyi elégedettséget (a) és a vevői (fogyasztói) elégedettséget (b).

a) A faktorelemzés célja a munkahelyi elégedettséget befolyásoló mögöttes tényezők azonosítása. A vizsgálatba bevont változók a munkavállalók tapasztalatait és attitűdjeit mérik különböző szempontok szerint. Idetartozik a *fizetés és juttatások megítélése, a munkakör érdekes és motiváló volta, a vezetők támogató hozzáállása, a munkahelyi légkör és a kollégákkal való kapcsolatok*, valamint a *szakmai fejlődési lehetőségek megléte*. Feltételezhető, hogy ezek a változók bizonyos közös dimenziók (pl. *anyagi megbecsülés, emberi kapcsolatok, szakmai előrelépés*) mentén rendeződnek. A faktorelemzés így segít a komplex elégedettségi struktúra egyszerűsítésében és átláthatóbbá tételében.

b) A faktorelemzés célja a fogyasztói elégedettséget meghatározó tényezők feltárása, amelyek segítségével jobban megérthető a vásárlói döntések háttere. A vizsgálatban olyan változók szerepelnek, mint *a termék minőségének megítélése, az ár-érték arány, a vásárlás egyszerűsége, az ügyfélszolgálat színvonala, a márkába vetett bizalom*, valamint *a termék elérhetősége és dizájnja*. Ezek a szempontok várhatóan néhány átfogó dimenzióba sűríthetők, például a *termékjellemzők, szolgáltatási élmény* és *bizalom* faktorokba. A faktorelemzés lehetővé teszi, hogy az egyéni változókból közös mintázatok rajzolódjanak ki, és jobban átláthassuk, mi áll a fogyasztói elégedettség mögött.

2. Az adatok előkészítése

Ebben a lépésben történik a faktorelemzésbe bevonni kívánt változók adatbázisban való azonosítása. Ekkor történik a hiányzó értékek és a szélsőséges értékek ellenőrzése – az SPSS-ben erre vonatkozó módszereket a többváltozós lineáris regressziónál (*5.2. alfejezet*) már ismertettük.

3. A faktorelemzés lefuttatása és a faktorelemzés módszerének meghatározása

Faktorelemzést az *ANALYZE* főmenü *Dimension Reduction*, *Factor* menüpontnál kérhetünk. Ahogyan minden elemzésnél, a bal oldalról átvisszük a jobb oldalra a vizsgálatba bevont változókat.

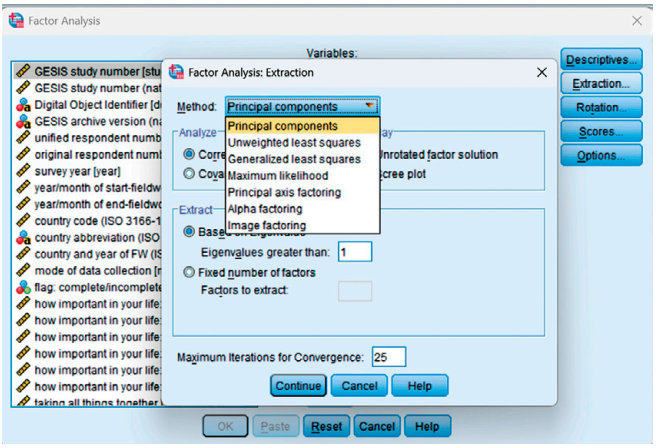
A faktorelemzés menüben az *Extraction* parancskötegnél adhatjuk meg a tömörítés módszerét. A faktorelemzés extrakciós módszerei (7 db):

1. **Főkomponens-elemzés** (*Principal components*), alapbeállítás: ez a módszer (más néven Hotelling-módszer) első faktorként egy olyan standardizált (0-ás átlagú, 1-es szórású) változót állít elő, amelyik a legjobban korrelál az összes modellbe vitt változóval, második faktorként egy olyat, amelyik korrelálatlan a már előállított faktorral, és legjobban korrelál az összes modellbe vitt változóval.

2. **Súlyozatlan legkisebb négyzetek módszere** (*Unweighted least squares*): minimalizálja a megfigyelt és az újonnan létrehozott korrelációs mátrixok közötti különbségek négyzeteinek összegét, előnye, hogy a változók eloszlása lényegtelen, viszont skálatranszformációt hajt végre, ezért standardizált változókkal érdemes végezni.
3. **Általánosított legkisebb négyzetek módszere** (*Generalized least squares*): minimalizálja a megfigyelt és az újonnan létrehozott korrelációs mátrixok közötti különbségeket, de a korrelációk súlyozásra kerülnek.
4. **Maximum-likelihood-módszer** (*Maximum liklehood*): a megfigyelt korrelációs mátrixból indul ki és olyan becsléseket ad, amelyek ezt a korrelációs mátrixot a legnagyobb valószínűség mellett létrehozhatták, feltételezve a változók normális eloszlását.
5. **Főtengelyelemzés** (*Principal axis factoring*): hasonlít a főkomponens-elemzéshez, viszont a kezdeti kommunalitásokként az eredeti korrelációs mátrix átlójában a többszörös korrelációs együtthatók négyzeteit használja.
6. **Alfa-eljárás** (*Alpha factoring*): feltételezi, hogy az elemzésbe bevont az összes lehetséges változónak csak egy mintáját képezik, a faktorok alfaértékét maximalizálja.
7. **Image-eljárás** (*Image factoring*): a változókat egy lineáris regresszió részeként kezeli, nem egy mesterséges változó (faktor) függvényeként.

A főkomponens-, a főtengely-, az alfa- és a maximum-liklehood elemzés nagyon sok esetben ugyanahhoz az eredményhez vezet. Ha nagyon sok változóval dolgozunk, a maximum-liklehood-, az image- és az alfa-elemzés használata javasolt.

Mivel a főkomponens-elemzés a faktorelemzési eljárások közül a leggyakrabban használt és legkönnyebben alkalmazható módszer, a továbbiakban ezzel a faktorelemzési módszerrel fogunk dolgozni (84. ábra).



84. ábra. *A faktorelemzés módszerének kiválasztása*

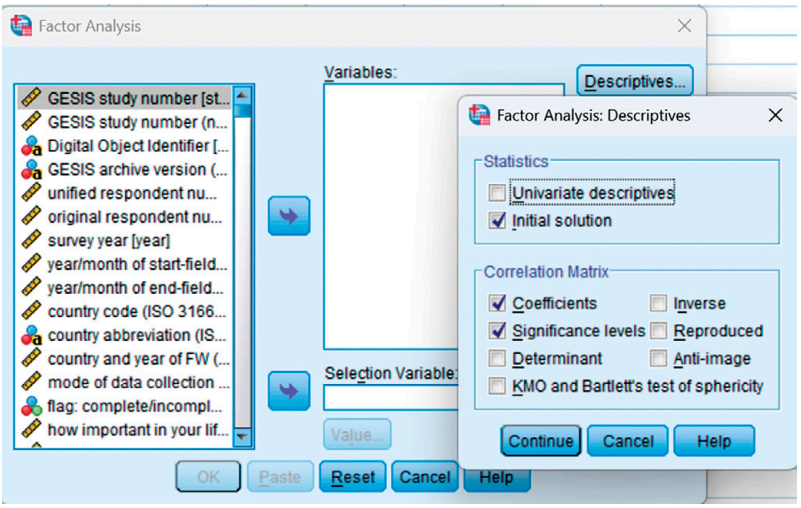
A főkomponens módszere tulajdonképpen a vizsgálatba bevont változók közti korrelációs együtthatók mátrixából úgynevezett *sajátértéket* és *sajátvektort* számít közelítő (iterációs) módszerrel. A sajátérték (*Eigenvalue*) azt mutatja meg, hogy az adott faktor(ok) az eredeti változók teljes varianciáját mennyiben magyarázzák meg (a faktorok számának behatárolására használatos). A faktorelemzésben a sajátvektor (*Eigenvector*) az a vektor, amely megmutatja, hogy az adott faktor mekkora súllyal kapcsolódik az egyes változókhoz, és így a faktor irányát, értelmezését határozza meg. Alapértelmezésben az SPSS maximum 25 iterálást végez (*Maximum Iteration for Convergence* mező, 77. ábra), amíg megkapja a sajátértékeket és faktorsúlyokat (a pontosabb értékek kiszámíttatása céljából a 25-ös szám átállítható egy nagyobb értékre). A sajátvektor komponensei a *faktorsúlyok*, amelyek valójában egy, a sajátértékhez tartozó faktornak a mért változókkal való korrelációs együtthatói (–1 és 1 közötti érték), a *sajátérték* pedig ezen faktorsúlyok négyzetösszege. A *faktorsúlyok* tehát azt mutatják meg, hogy egy változó milyen erősen és milyen irányban kapcsolódik az adott faktorhoz, így alapvető szerepük van a faktorok értelmezésében és elnevezésében.

4. A faktorelemzés alkalmazhatóságának vizsgálata

Az alkalmazhatóság vizsgálatára négy lehetőségünk van: a korrelációs mátrix elemzése (1), a KMO és a Bartlett-teszt (2), a kommunalitások vizsgálata (3) és az anti-image mátrix (4). Az anti-image mátrix a faktorelemzésben a korrelációs mátrix inverzéből származtatott, részkorrelációkon alapuló mátrix, amelyben az átlóelemek az MSA-értékeket (*Measure of Sampling Adequacy*, vagyis a változó faktorelemzésre való alkalmasságának mutatója), a többi elem pedig a (negatív) részkorrelációkat jelzi, így mutatva, mennyire illeszkednek a változók a faktorstruktúrába. Ebből a négy alkalmazhatósági feltételből az anti-image mátrix vizsgálatától (4) eltekintünk, hiszen ez akkor hasznos, amikor a KMO azt mutatja, hogy a változók rendszere alkalmatlan faktorelemzésre. Tulajdonképpen az anti-image mátrix világít rá az ok helyére: a változók mindegyikében van valami, ami miatt nem alkalmasak vagy csak egyik-másik nem alkalmas a faktorelemzésre (ez utóbbi esetben kihagyva az oda nem illő változót, már elemzésre alkalmas változórendszert kapunk).

1. A korrelációs mátrix elemzése

A szignifikáns korrelációk arra utalnak, hogy a változóink alkalmasak a faktorelemzésre, ugyanakkor a túlságosan magas korrelációs együtthatók ($r > 0{,}9$) nem mindig jók, mert akkor minden változónk egy faktorba tömörülne (ugyanakkor ez is lehet a faktorelemzés célja). A faktorelemzés tehát akkor lesz jó, ha a változók között van, de nem túlzottan erős az összefüggés. A korrelációs mátrix a faktoranalízis menüben, az *ANALYZE, Dimension Reduction, Factor, Descriptives* úton kérhető le (85. ábra).



85. ábra. *A korrelációs mátrix lekérése a faktorelemzés menüből*

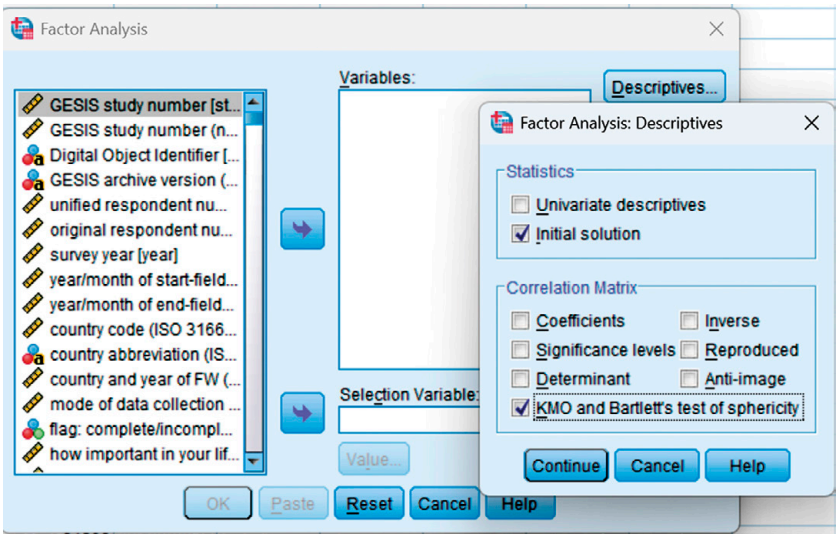
2. A KMO mutató és a Bartlett-teszt

A *KMO* (Kaiser–Meyer–Olkin) mutató az egyik legfontosabb mérőszám annak megítélésére, hogy a változók mennyire alkalmasak a faktorelemzésre. A KMO mutatót a már említett anti-image mátrix alapján számolják ki. Értelmezése:

KMO ≥ 0,9: adataink kiválóak a faktorelemzésre,
KMO ≥ 0,7: adataink megfelelőek,
KMO ≥ 0,5: adataink még elfogadhatóak a faktorelemzésre,
KMO < 0,5: a faktorelemzés elfogadhatatlan.

A *Bartlett-teszt* a korrelációkkal kapcsolatos teszt, amely azt vizsgálja, hogy a változók az alapsokaságban korrelálnak-e. Ha a szignifikanciaszint kisebb, mint 0,05 ($p < 0,05$), akkor 95%-os valószínűséggel állíthatjuk, hogy a változók közötti korreláció nem a véletlen műve, tehát a változók között van összefüggés, így alkalmasak a faktorelemzésre.

A Bartlett-teszt és a KMO mutató szintén a faktoranalízis menüben, a *Descriptives* mezőnél kérhető le (86. ábra).



86. ábra. *A KMO mutató és a Bartlett-teszt lekérése*

3. A kommunalitások vizsgálata

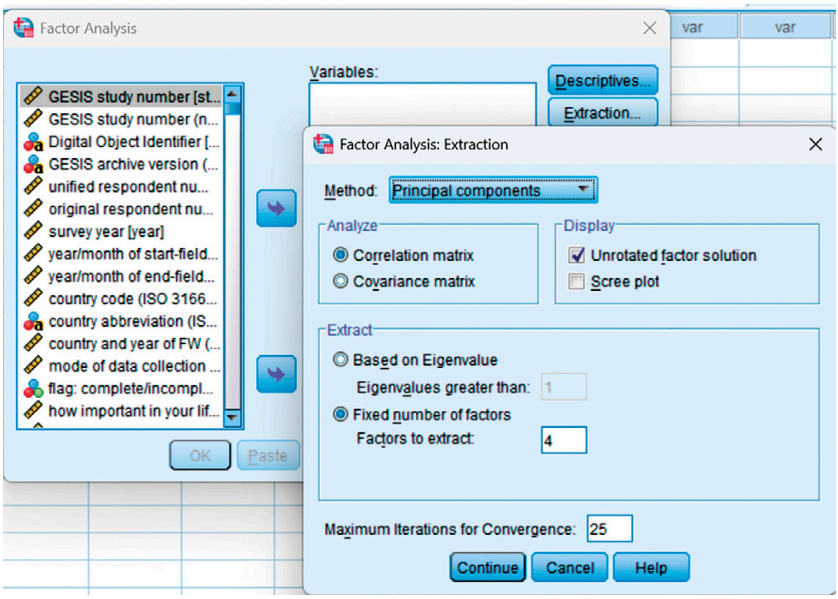
A kommunalitások vizsgálata azt a célt szolgálja, hogy megállapítsuk, minden változó hozzájárul-e a faktorstruktúra kialakításához. *A kommunalitások* a többszörös korrelációs együtthatók négyzetei, és azt mutatják meg, hogy a faktorok együtt milyen mértékben magyarázzák az adott változó szóródását. Azt a változót tekintjük a főkomponens alkotóelemének, amelynek a *kommunalitása* $\geq 0{,}25$, vagyis a főkomponens és az eredeti változó közötti kapcsolat szorossága legalább 0,5 értékű korrelációval írható le. Amennyiben ez a feltétel nem teljesül, az illető változó nem járul hozzá a faktorstruktúra kialakításához, és ki kell vennünk a modellből.

A kommunalitásokat az SPSS alapértelmezésben kiszámolja a faktorelemzés lefuttatásakor.

5. A főkomponensek számának meghatározása

A létrehozni kívánt faktorok számának megállapítására három lehetőségünk is van.

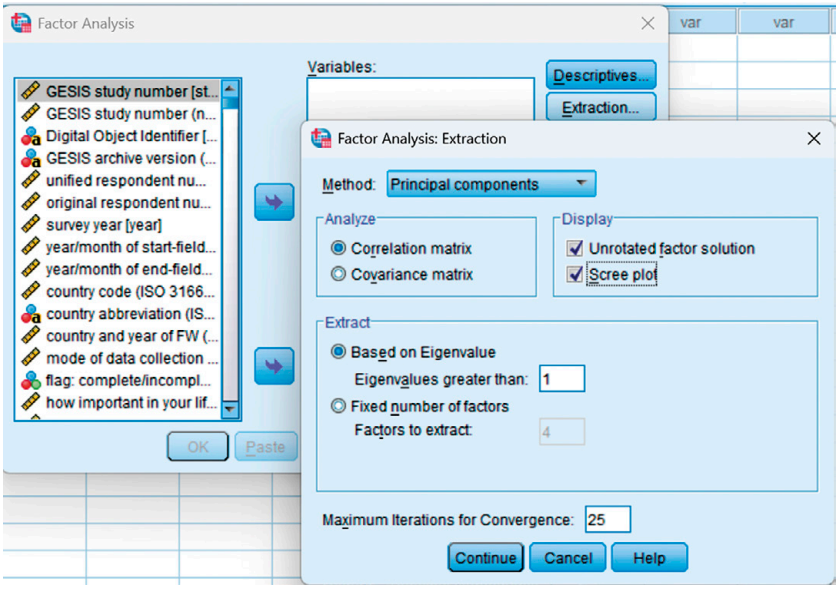
1. Az SPSS-program lehetőséget ad arra, hogy *mi határozzuk meg a faktorok számát*. A legkézenfekvőbb az, amikor a faktorok számát egy elméleti modell vagy korábbi vizsgálatok alapján határozzuk meg, ilyenkor a faktorelemzés főablakában, az *Extraction* parancskötegnél az alapértelmezett *Eigenvalues over 1* helyett a *Number of factors* mezőnél beírjuk a kívánt faktorok számát (a 87. ábra szerint, pl. 4 faktor).



87. ábra. *A főkomponensek számának megadása*

2. Feltételezzük, hogy a vizsgált változóinkkal kapcsolatosan nem rendelkezünk előzetes feltételezésekkel a látens dimenziók számáról. Ilyenkor a legegyszerűbben a *Kaiser-kritérium* alapján határozhatjuk meg a faktorok számát (az SPSS alapértelmezésben ezt használja). A Kaiser-kritérium azt mondja, hogy csak az 1 sajátérték feletti faktorokat vegyük figyelembe. A *sajátértéket* (*Eigenvalue*) viszonyítva a változók számához azt kapjuk, hogy a sajátértékhez tartozó faktor mennyit képes magyarázni a mért változók varianciájából. A sajátértékek pozitívak, számuk egyenlő a bemenő változók számával, és összegük is ugyanennyi. Tehát a sajátértékek átlaga 1, ezért lesznek közöttük 1-nél nagyobbak is és 1-nél kisebbek is (amikor minden sajátérték 1, akkor a bemenő változók egymással teljesen korrelálatlanok, tehát már faktorváltozók). Abból, hogy a sajátértékek pozitívak és átlaguk 1, az is következik, hogy általában több 0 és 1 közötti lesz köztük, mint 1-nél nagyobb (ha van egy 4-nél is nagyobb sajátérték, akkor ehhez négy 1-nél kisebb sajátérték is kell, hogy átlagban 1-et hozzanak ki). Amikor a változókban sok a közös információ, akkor igen nagy sajátérték(ek) is előfordul(nak), és sok lesz a nagyon kicsi, tehát sok faktor fog kevés magyarázóerővel bírni. Ha egy faktor sajátértéke kisebb, mint 1, akkor ez azt jelenti, hogy kevesebb információt hordoz, mint akármelyik változó, és azt a faktort nem feltétlenül érdemes használni.

A Kaiser-kritérium alkalmazását könnyíti az *Extraction* menüpontnál, a *Display* ablakrészben található *Scree Plot* elnevezésű ábra lekérése, amely a faktorok által megtestesített sajátérték nagyságát szemlélteti (88. ábra).



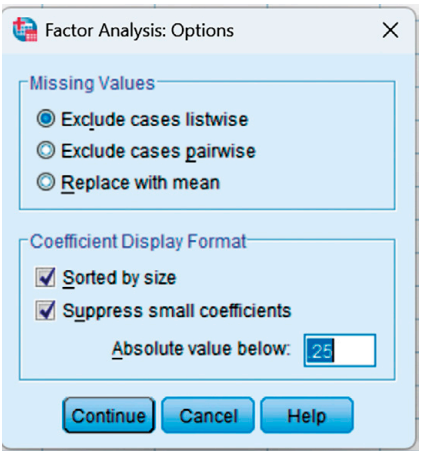
88. ábra. *A Kaiser-kritérium alkalmazásának lekérése*

3. Egy másik alapvető módszer a faktorszám meghatározására a varianciahányad-módszer. A faktorok számát meghatározhatjuk a variancia kumulált százaléka alapján is. Társadalomtudományokban elfogadott szabály, hogy főkomponens-elemzés esetén a faktorok által hordozott információérték ne legyen kevesebb, mint 50% (más faktorelemzési eljárásoknál 33%). A faktorok által magyarázott varianciát az SPSS alapértelmezésben megadja, a *Total Variance Explained* táblázatban.

5. A faktorok értelmezése

A *faktorsúly* (*Factor Loading*) tehát nem más, mint az eredeti változó és az adott faktor közötti korrelációs együttható (értéke –1 és 1 közötti érték). A faktorok értelmezésére a faktorsúlyokat tartalmazó faktorsúlymátrixot használjuk. Általános szabály, hogy *a faktorsúly értéke legalább a 0,25 értéket el kell érje* (abszolút értékben). Kisebb mint 100 fős mintákon a faktorsúly értéke legalább 0,5 kell legyen. Minél magasabb egy faktorsúly értéke (abszolút értékben), annál nagyobb szerepet játszik az illető változó a faktor értelmezésében. Továbbá egy változó akkor tartozik egyértelműen egyik faktorhoz, ha *faktorsúlya csak egy faktoron nagyobb, mint 0,25*, vagy ha *faktorsúlya az egyik faktoron nagyobb, mint bármelyik más faktoron lévő faktorsúlya értékének kétszerese.*

A könnyebb értelmezés kedvéért a faktorelemzés főablakban, az *Options* menünél állítsuk be, hogy adatainkat csökkenő sorrendben jelenítse meg az SPSS, és csak a 0,25-nél nagyobb faktorsúlyokat lássuk majd az *Output*-ban megjelenő táblázatban (89. ábra).



89. ábra. *A faktorsúlyok értékeinek megjelenítése*

A rotálatlan faktorsúlymátrixot az SPSS a kommunalitásokhoz hasonlóan automatikusan megjeleníti az eredménykijelző ablakban. Azonban a legtöbb esetben a rotálatlan faktorsúlymátrix alapján nem tudjuk értelmezni a faktorainkat.

6. Értelmezés rotálással

A faktoranalízis alapegyenletének végtelen sok matematikailag helyes megoldása van, a főkomponens-módszer (de a többi is) valamilyen közelítő módszerrel (számítógépen csak ilyen módszerekkel dolgoznak a programok) meghatároz egyet, majd ebből kiindulva, újra csak közelítő módszerrel olyan megoldást szolgáltat, amelyik bizonyos szempontból optimálisabb, mint a többi megoldás. A faktoregyenlet egyik megoldásából a többi megoldást úgynevezett mátrixtranszformációval lehet megkapni, és a geometriában ennek a transzformációnak a neve: *forgatás (rotáció)*.

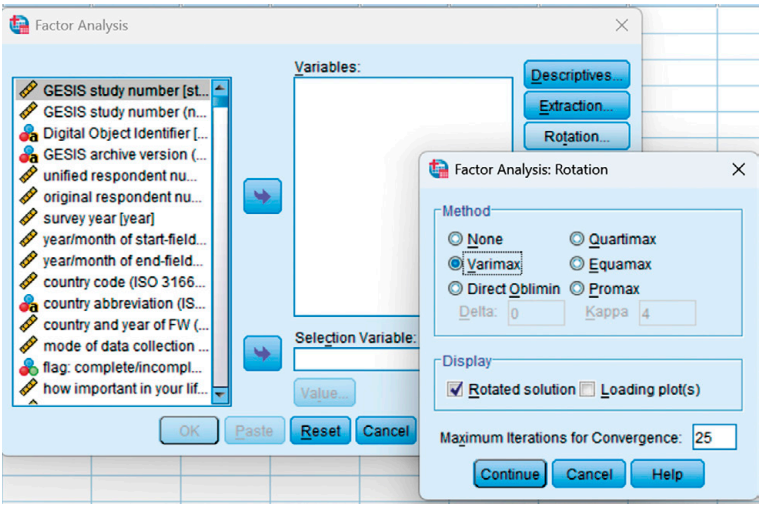
A társadalomkutató számára fő optimalizációs szempont az, hogy a különböző faktorok a mért változók csak egy jól elkülönülő részével korreláljanak nagyon jól, a többiekkel pedig a legkorrelálatlanabbak legyenek. A faktorelemzés során azonban nagyon gyakran előfordul, hogy olyan változók korrelálnak ugyanazzal a faktorral (tartoznak ugyanahhoz a faktorhoz vagy ülnek ugyanazon a faktoron), amelyeknek semmi közük egymáshoz, vagy egyszerre két faktorral is korrelálnak (keresztkötődés), és így nem tudjuk őket értelmezni. Ebben segít a forgatás vagy rotálás, ami a gyakorlatban azt jelenti, hogy a faktorok tengelyeit elforgatjuk úgy, hogy egyszerűbb és főként értelmezhetőbb faktorokat nyerjünk. A rotálás nem

változtatja meg sem a kommunalitásokat, sem pedig az összes magyarázott varianciát, csak a faktorok magyarázott varianciáit módosítjuk.

Kétféle rotálási típust szokás megkülönböztetni: derékszögű vagy orthogonális, valamint hegyesszögű rotálást. A hegyesszögű rotálás eredményeképpen a faktorok korrelálni fognak egymással (a tengelyek tetszőleges szöget zárnak be), a derékszögű forgatás eredményeként pedig a faktorok korrelálatlanok maradnak egymással (a tengelyek derékszöget zárnak be). Ha a faktorelemzés eredményeit további elemzésekbe kívánjuk bevonni, akkor az orthogonális, ha pedig csak értelmezni akarjuk a faktorokat, akkor a hegyesszögű forgatás ajánlott.

Az SPSS által használt derékszögű forgatási módszerek a *Varimax* (csökkenti az egy faktorra eső magas faktorsúlyú változók számát), *Quartimax* (az egy változó megmagyarázásához szükséges faktorok számát csökkenti) és *Equimax* (az első kettő kombinálása). Hegyesszögű forgatási módszerek a *Direct Oblimin* és a *Promax*.

A rotálás a faktoranalízis menüben a *Rotation* menüpontnál kérhető le, a választott forgatási módszer, pl. *Varimax* bejelölésével (90. ábra).



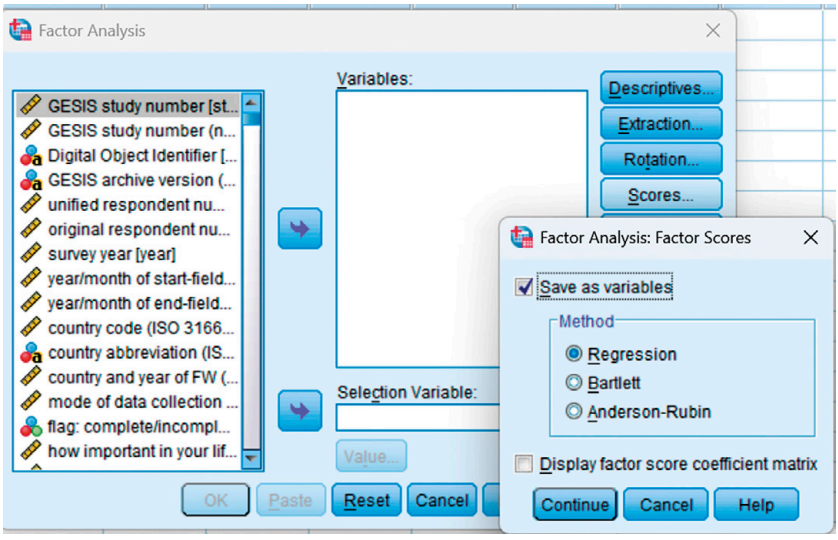
90. ábra. *Varimax forgatás kérése*

A rotálással, egyik forgatási módszerrel sem értelmezhető változók kezelésére három lehetőségünk van:

1. megvizsgáljuk, hogy *több vagy kevesebb faktorszám* esetén ezek a változók hogyan viselkednek,
2. *kizárhatjuk* az elemzésből ezeket a változókat és újrafuttatjuk a faktorelemzést, vállalva, hogy lényeges információkat veszítettünk,
3. a változókat benne hagyjuk az elemzésben, de az *értelmezésnél nem vesszük figyelembe* őket (1-2 ilyen lehet).

7. A faktorváltozók elmentése

Az SPSS a létrehozott új faktorváltozókhoz úgynevezett faktorszkórokat rendel, ami azt jelenti, hogy minden esethez rendel egy értéket a faktorok jellemzésére. Tehát a *faktorértékek* (*Factor scores*) az egyes vizsgált esetek „eredményei" a létrehozott háttérváltozóban, faktorban az eredetileg mért változók alapján. A faktorszkórokat tartalmazó faktorokat három módszerrel menthetjük el: regressziós módszerrel (1), Bartlett-módszerrel (2) és Anderson–Rubin-módszerrel (3). A három módszer közötti különbséget csak nagyon bonyolult matematikai apparátus segítségével lehet megmagyarázni. Elég, ha azt tudjuk, hogy a három módszerrel elmentett faktorszkórok között nincs lényeges különbség. Azonban ha a faktorokat további elemzésre kívánjuk felhasználni, a regressziós módszer használata ajánlott. A mentés a *Scores* mezőnél történik (91. ábra).



91. ábra. *A faktorok mentése regressziós módszerrel*

Az adatbázisunk végén ilyen módon megjelennek az új faktorváltozók (FAC1_1, REGR factor score 1 for analysis 1 stb. név alatt), amelyek standardizált skálaváltozók lesznek, és az értelmük szerint definiálni kell őket (változónév stb.).

8. További felhasználás

A létrehozott faktoraink a továbbiakban kétváltozós elemzésekre és klaszterelemzésre is jól használhatóak. A faktorszkórok értelmezéséhez legcélszerűbb először leíró statisztikákat kérni (*ANALYZE, Descriptive Statistics, Descriptives*). A faktorok tehát egységnyi szórású, 0 körüli átlagú, standardizált mennyiségi változók. A maximális és minimális értékek a leíró statisztikák táblázatában sze-

repelnek, tehát az adatok értelmezésekor ehhez kell viszonyítsunk. Általában a pozitív értékek a magasabb, a negatív értékek az alacsonyabb értéket jelölik (az eredetileg mért változók szerint).

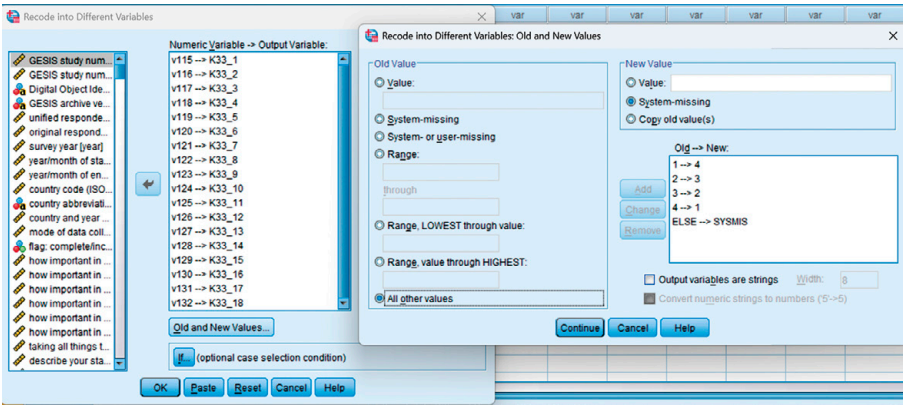
46. példa ▼

▶ *Faktorelemzés az SPSS-ben*

Az adatbázisban a K33-as kérdés változói 18 különböző intézménybe vetett bizalmi szintet mérnek egy 1-től 4-ig terjedő Likert-skálán. Tehát az elemzésünkben 18 ordinális mérési szintű változót (*v115–v132*, az adatbázis 148–165-ös sorszámú változói) elemzünk faktorelemzéssel. Arra vagyunk kíváncsiak, hogy ezen bizalmi attitűdök mögött milyen általánosabb dimenziók vagy közös faktorok húzódnak meg.

Azt feltételezzük, hogy az emberek bizalma bizonyos intézményekben együtt mozog, így kialakulhatnak nagyobb kategóriák, mint pl. politikai intézményekbe vetett bizalom (parlament, kormány, pártok) vagy jogi-igazságszolgáltatási bizalom (pl. igazságszolgáltatási rendszer, rendőrség, fegyveres erők), társadalmi-közösségi intézmények (pl. oktatási intézmények, társadalombiztosítási rendszer, egyház, civil szervezetek) stb. Így a faktorelemzés fő célja: a 18 különálló változó mögött meghúzódó kevesebb számú, átfogóbb bizalmi dimenzió azonosítása, amelyek egyszerűbben értelmezhetővé teszik az emberek intézményekbe vetett bizalmának szerkezetét.

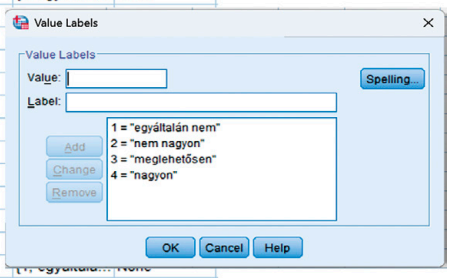
Első lépésben kódoljuk át az eredeti változóinkat új változókba úgy, hogy az eredeti változóink sértetlenek maradjanak (*TRANSFORME, Recode into Different Variables*). Az átkódoláskor az új változókat a kérdőív szerint nevezzük el *K33_1-K33_18*, az értékek sorrendjét pedig fordítsuk meg úgy, hogy az új változókban az értékek növekedése a bizalmi szint növekedését mutassa (pl. az eredeti változókban az 1-es érték a legmagasabb szintű bizalmat jelölte, az új változókban ez legyen 4-es stb.) (92. ábra).



92. ábra. *A faktorelemzésbe bevont változók átkódolása (46. példa)*

A *Variable View*-ban címkézzük fel az új változókat a kérdőív szerint (a kérdőívben hiányzik a *v125*-ös változónak megfelelő sor, ami az ENSZ-et jelöli), megadva az új értékcímkéket is (93. ábra).

Name	Type	W...	De...	Label	Values	Missing
K33_1	Numeric	8	0	Egyház	{1, egyáltalá...	None
K33_2	Numeric	8	0	Fegyveres erők	{1, egyáltalá...	None
K33_3	Numeric	8	0	Oktatási rendszer	{1, egyáltalá...	None
K33_4	Numeric	8	0	Sajtó		
K33_5	Numeric	8	0	Szakszervezetek		
K33_6	Numeric	8	0	Rendőrség		
K33_7	Numeric	8	0	Parlament		
K33_8	Numeric	8	0	Közigazgatás		
K33_9	Numeric	8	0	Társadalombiztosítási rendszer		
K33_10	Numeric	8	0	Európai Unió		
K33_11	Numeric	8	0	ENSZ		
K33_12	Numeric	8	0	Egészségügyi rendszer		
K33_13	Numeric	8	0	Igazságszolgáltatási rendszer		
K33_14	Numeric	8	0	Nagyvállalatok		
K33_15	Numeric	8	0	Környezetvédelmi szervezetek		
K33_16	Numeric	8	0	Politikai pártok	{1, egyáltalá...	None
K33_17	Numeric	8	0	Kormány	{1, egyáltalá...	None
K33_18	Numeric	8	0	Közösségi média	{1, egyáltalá...	None



93. ábra. *A faktorelemzésbe bevont átkódolt változók az adatbázisban (46. példa)*

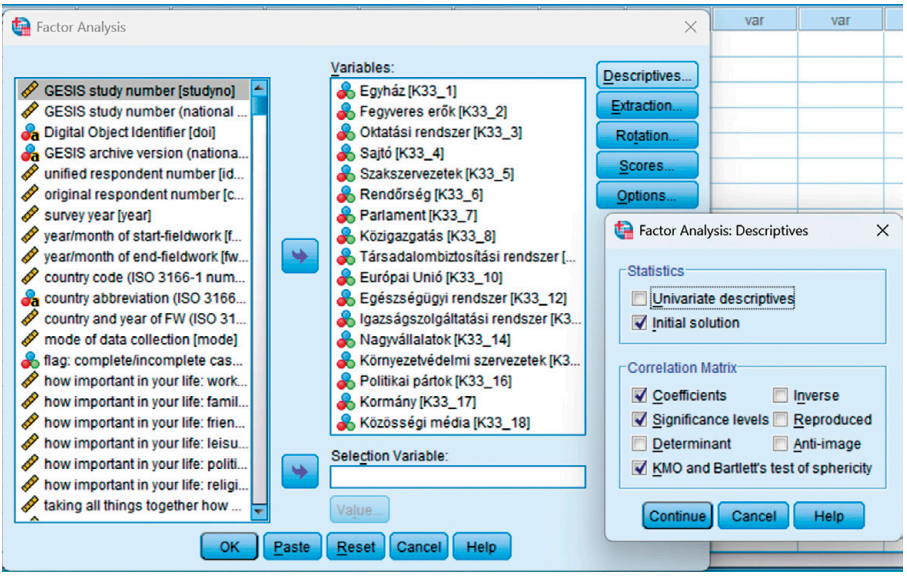
A faktorelemezni kívánt K33_1-K33_18 változók leíró statisztikáit (*ANALYZE, Descriptive Statistics, Descriptives*) a 94. ábra mutatja. Itt láthatjuk, hogy az ENSZ-be vetett bizalomra vonatkozó változó (*k33_11*) egyetlen adatot sem tartalmaz, ezért a továbbiakban csak a többi 17 változóval dolgozunk.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Egyház	1099	1	4	3.08	.852
Fegyveres erők	1061	1	4	2.32	.880
Oktatási rendszer	1084	1	4	2.53	.864
Sajtó	1085	1	4	2.13	.798
Szakszervezetek	1017	1	4	2.11	.792
Rendőrség	1094	1	4	2.39	.875
Parlament	1078	1	4	1.76	.790
Közigazgatás	1082	1	4	2.27	.795
Társadalombiztosítási rendszer	1065	1	4	2.21	.815
Európai Unió	1060	1	4	2.23	.886
ENSZ	0				
Egészségügyi rendszer	1084	1	4	2.41	.861
Igazságszolgáltatási rendszer	1074	1	4	2.22	.867
Nagyvállalatok	1021	1	4	2.08	.808
Környezetvédelmi szervezetek	1058	1	4	2.36	.863
Politikai pártok	1063	1	4	1.66	.722
Kormány	1068	1	4	1.68	.743
Közösségi média	1038	1	4	1.94	.811
Valid N (listwise)	0				

94. ábra. *A faktorelemzésbe bevont változók leíró statisztikái (46. példa)*

Következő lépésben ellenőrizzük az alkalmazhatósági feltételeket. Ehhez futtassuk le a korábbiakban leírtak szerint a faktorelemzést főkomponens (*Principal Components*) módszerrel (*ANALYZE, Dimension Reduction, Factor*), vigyük át a 17 elemezni kívánt változót, majd a *Descriptives…*-nél kérjük le a korrelációs mátrixot, valamint a KMO mutatót és a Bartlett-tesztet (95. ábra).



95. ábra. *A faktorelemzés feltételeinek vizsgálata (46. példa)*

A korrelációs mátrix számos szignifikáns összefüggést mutat, a legmagasabb korrelációs együttható 0,730 (a 0,9 küszöbérték alatti), így e szerint a változóink alkalmasak a faktorelemzésre. A KMO mutatónk faktorelemzésre kiváló változókat jelez (KMO = 0,915), a Bartlett-teszt is szignifikáns (p = 0,000) összefüggést mutat a változók között (96. ábra).

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.915
Bartlett's Test of Sphericity	Approx. Chi-Square	6132.609
	df	136
	Sig.	.000

96. ábra. *A KMO-mutató és a Bartlett-teszt értéke (46. példa)*

A kezdeti kommunalitások értéke főkomponens-elemzésnél mindig 1. A faktorelemzés utáni kommunalitások értékei 0,4 felettiek (97. ábra), így a 0,25-ös küszöbérték fölött vannak, tehát minden alkalmazhatósági feltétel teljesülése után nagy reményekkel foghatunk neki a faktorelemzésnek.

Communalities

	Initial	Extraction
Egyház	1.000	.419
Fegyveres erők	1.000	.532
Oktatási rendszer	1.000	.631
Sajtó	1.000	.730
Szakszervezetek	1.000	.538
Rendőrség	1.000	.602
Parlament	1.000	.656
Közigazgatás	1.000	.530
Társadalombiztosítási rendszer	1.000	.569
Európai Unió	1.000	.520
Egészségügyi rendszer	1.000	.673
Igazságszolgáltatási rendszer	1.000	.663
Nagyvállalatok	1.000	.473
Környezetvédelmi szervezetek	1.000	.574
Politikai pártok	1.000	.722
Kormány	1.000	.771
Közösségi média	1.000	.642

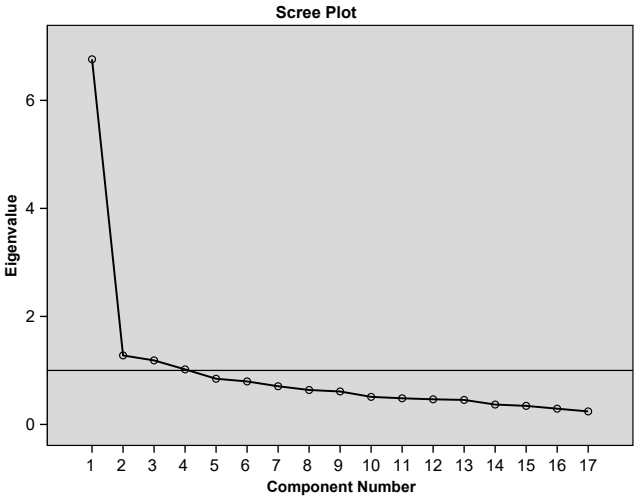
Extraction Method: Principal Component Analysis.

97. ábra. *A kommunalitások (46. példa)*

Mivel a vizsgált változóinkkal kapcsolatosan nem rendelkezünk előzetes feltételezésekkel a látens dimenziók számáról, a *Kaiser-kritérium* alapján határozzuk meg a faktorok számát és ábrázoltatjuk is (*Extraction, Display, Scree Plot)*. Az ábrára az 1-es sajátértéknél egy referenciavonalat is kérünk (98. ábra).

A *Scree Plot* (98. ábra) azt mutatja, hogy 4 sajátérték feletti faktorunk van, és ezeket érdemes megtartani (a függőleges tengelyen a sajátérték nagysága, a vízszintes tengelyen pedig a faktorok száma található).

A 99. ábrán szereplő táblázatban (*Total Variance Explained*) is azt látjuk, hogy 4 db egynél nagyobb sajátértékű faktorunk van. Ahogyan ez a főkomponens-elemzéstől elvárható, az első faktornak van a legnagyobb magyarázóereje, és a négy faktor által hordozott információmennyiség az eredeti 17 változó által megtestesített információ 60,3%-a. Ez az érték az 50%-os küszöbérték felett van.



98. ábra. *A sajátértékek grafikus megjelenítése (46. példa)*

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.761	39.768	39.768	6.761	39.768	39.768
2	1.278	7.515	47.284	1.278	7.515	47.284
3	1.186	6.975	54.259	1.186	6.975	54.259
4	1.019	5.996	60.255	1.019	5.996	60.255
5	.846	4.977	65.232			
6	.796	4.685	69.917			
7	.707	4.159	74.077			
8	.638	3.751	77.828			
9	.610	3.590	81.418			
10	.511	3.004	84.422			
11	.484	2.848	87.270			
12	.465	2.737	90.007			
13	.454	2.671	92.677			
14	.368	2.167	94.844			
15	.344	2.022	96.866			
16	.292	1.715	98.582			
17	.241	1.418	100.000			

Extraction Method: Principal Component Analysis.

99. ábra. *A faktorok által magyarázott összvariancia (46. példa)*

Tehát azáltal, hogy 17 változó helyett 4 faktorváltozóval dolgozunk, háromötödére csökkent a rendelkezésünkre álló információmennyiség (a tömörítés nyomán 40%-os az információveszteség). Ha értelmezni tudjuk a faktorainkat, ez jó cserének tűnik.

A faktorok értelmezéséhez a korábban jelzett módon az *Options*-nál állítsuk be, hogy a faktorsúlyok nagyság szerint legyenek rendezve, és csak a 0,25-ös értéknél nagyobbak legyenek feltüntetve. Ugyanakkor kérjük a *Varimax* rotálást is (*Rotation*-nél), mivel a rotálatlan faktorsúlymátrix ritkán értelmezhető. Ezért a továbbiakban csak a rotált faktorsúlymátrixo(ka)t nézzük (100. ábra).

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Környezetvédelmi szervezetek	.689			.266
Társadalombiztosítási rendszer	.673			
Európai Unió	.651	.269		
Egészségügyi rendszer	**.650**		**.463**	
Igazságszolgáltatási r.	**.599**	**.385**	**.394**	
Közigazgatás	**.579**			**.378**
Nagyvállalatok	**.564**	**.352**		
Kormány	.339	.782		
Politikai pártok	.251	.769		
Parlament	.281	.668	.293	
Oktatási rendszer	.260		.705	.254
Fegyveres erők			.677	
Rendőrség	**.351**	.279	**.632**	
Sajtó				.818
Közösségi média	**.352**	**.436**		**.562**
Szakszervezetek		.271	**.342**	**.562**
Egyház	**.369**	**-.280**		**.377**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a) Rotation converged in 8 iterations.

100. ábra. *A rotált faktorsúlymátrix a négyfaktoros kiinduló modellre (46. példa)*

A rotált faktorsúlymátrix nem túlságosan biztató: a 17 változóból 8 keresztkötődésű (a félkövérrel jelöltek), vagyis nagy faktorsúllyal kötődik legalább két faktorhoz is. Mivel ezeknek a változóknak a faktorsúlya egyik faktoron sem legalább kétszer akkora, mint a többin, nem lehet eldönteni, hogy melyik faktorhoz tartoznak. Például az *Egészségügyi rendszerbe vetett bizalom* változó az 1. és a 3. faktorhoz is tartozik, és ahhoz, hogy egyértelműen az 1. faktorhoz tartozzon, a faktorsúlya legalább 2×0,463 = 0,926 kellene legyen, de ennél kisebb (0,650). Ugyanez a helyzet az *Igazságszolgáltatási rendszer, Közigazgatás, Nagyvállalatok, Rendőrség, Közösségi média, Szakszervezetek* és *Egyház* változókkal. A másik két derékszögű rotálási módszerrel sem születtek könnyebben értelmezhető eredmények.

Mivel az elsődleges cél az értelmes faktorváltozóink megragadása, ezért megnézzük a faktorok tartalmát. Azt látjuk, hogy az első, a harmadik és a negyedik faktorunk elég vegyes intézményeket tartalmaz, nehezen értelmezhető. A második faktor jól értelmezhető (politikai intézményekbe vetett bizalom: *Kormány, Politikai pártok, Parlament*). Két választásunk van: 1. vagy lépésenként, egyenként kiveszünk 1-1 keresztkötődésű és kevésbé fontos változót, vagy 2. próbáljuk növelni a faktorok számát, hogy több dimenzióban ragadjuk meg a sokrétűnek látszó intézményi bizalmat.

Először a 2. lehetőséget nézzük. Az ötfaktoros modell esetén sokkal biztatóbb a kép: marad 6 keresztkötődésű változónk és az *Egyház* változó külön faktorba kerül, a magyarázóerő (összvariancia) 65,2%-ra nő, és minden más alkalmazhatósági mutató (korrelációs mátrix, KMO és Bartlett-teszt, kommunalitások) rendben van. A hatfaktoros modellünk sokkal nehezebben értelmezhető, ezért az ötfaktoros modellt folytatjuk (101. ábra) az 1. lehetőséggel.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
Környezetvédelmi szervezetek	.711			.274	
Társadalombiztosítási rendszer	.695				
Egészségügyi rendszer	**.654**		**.455**		
Európai Unió	.637	.287			
Igazságszolgáltatási r.	**.618**	**.359**	**.394**		
Nagyvállalatok	.603	.294			
Közigazgatás	**.555**			**.328**	
Kormány	.317	.828			
Politikai pártok		.822			
Parlament	.277	.686	.290		
Oktatási rendszer			.688		
Fegyveres erők			.677		
Rendőrség	**.376**		**.634**		
Sajtó				.836	
Szakszervezetek	.250		**.359**	**.632**	
Közösségi média	**.339**	**.462**		**.531**	
Egyház					.939

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

101. ábra. *A rotált faktorsúlymátrix az ötfaktoros kiinduló modellre (46. példa)*

A továbbiakban a több faktorhoz is nagy súllyal tartozó változókat sorra (többféle kombinációt és sorrendet kipróbálva) kivesszük az ötfaktoros modellből, amíg megkapunk egy értelmezhető, statisztikailag is érvényes faktormodellt. Ezt pl. az *Oktatási rendszer, Közigazgatás, Egészségügyi rendszer, Igazságszolgáltatási rendszer* és a *Közösségi média* változók modellből való elhagyásával érhetjük el (102. ábra), de más jó megoldások is lehetnek.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
Környezetvédelmi szervezetek	.762				
Nagyvállalatok	.701				
Európai Unió	.658	.315			
Társadalombiztosítási rendszer	.650		.301		
Kormány	.292	.846			
Politikai pártok		.833			
Parlament		.731	.266		
Fegyveres erők			.843		
Rendőrség	.297	.292	.664		
Sajtó				.882	
Szakszervezetek	.313		.260	.684	
Egyház					.974

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a) Rotation converged in 6 iterations.

102. ábra. *A rotált faktorsúlymátrix az ötfaktoros végső modellre (46. példa)*

A faktorokat most már tudjuk értelmezni és el tudjuk őket nevezni, mivel minden modellben maradt változó egyértelműen csak egyetlen faktorhoz tartozik.

1. faktor: a *Környezetvédelmi szervezetek, Nagyvállalatok, Európai Unió és Társadalombiztosítási rendszer* változók tartoznak ide, ezért ez a faktor a *Nemzetközi és gazdasági szervezetek iránti bizalom* elnevezést kapta,

2. faktor: a már korábban is jelzett, világosan kirajzolódó *Politikai intézményekbe vetett bizalom* (*Kormány, Politikai pártok, Parlament* változók alkotják),

3. faktor: *Állami biztonsági intézményekbe vetett bizalom* (*Fegyveres erők, Rendőrség* változók alkotják),

4. faktor: *Érdekképviseleti és véleményformáló intézményekbe vetett bizalom* (*Sajtó és Szakszervezetek* változók alkotják).

5. faktor: *Vallási intézményekbe vetett bizalom* (*Egyház* változó alkotja).

A végső, 12 változóból létrehozott 5 faktoros modell (102. ábra) KMO mutatója 0,887, a Bartlett-teszt szignifikáns ($p = 0{,}000$, $\chi^2 = 3533{,}534$), a kommunalitások 0,562-nél nagyobbak, és a modell magyarázóereje 71,1%. Tehát a

modell statisztikai mutatói kiválónak tekinthetőek, és a 12 intézményi bizalmat mérő változóból egy jól értelmezhető, 5 faktoros modellt hoztunk létre.

Utolsó lépésként a végső faktormodellt regressziós módszerrel (*Scores, Save as variables, Regression*) elmentjük, majd az adatbázis végén megjelenő faktorváltozókat felcímkézzük a dimenziók nevével (1. faktor: *Nemzetközi és gazdasági szervezetek iránti bizalom* stb.). A faktorok leíró statisztikáit a 103. ábra mutatja.

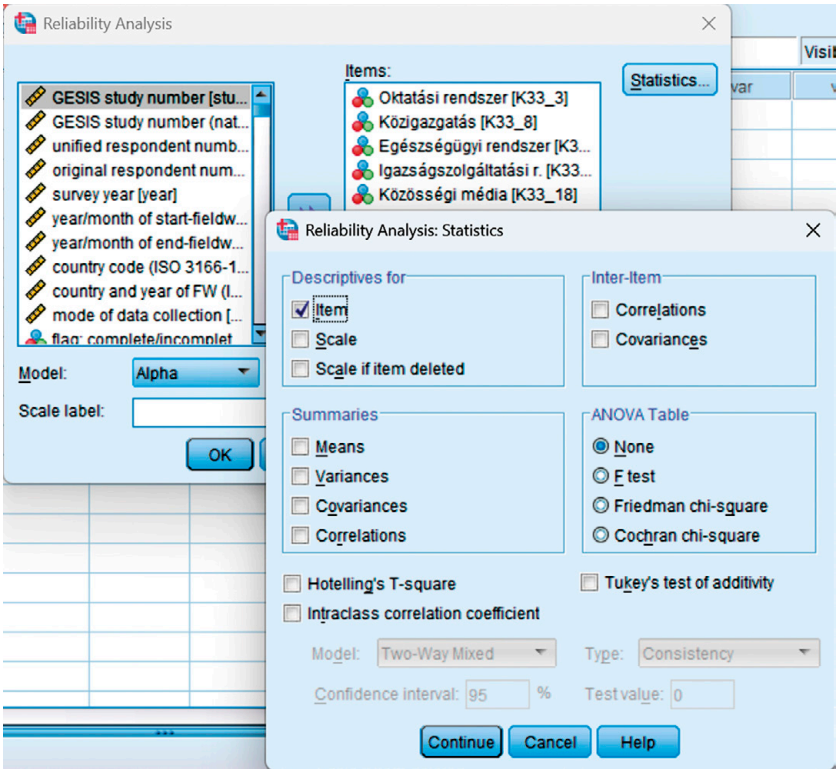
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Skewness
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Nemzetközi és gazdasági szervezetek iránti bizalom	909	-2.70587	4.10640	.0000000	1.00000000	.104	.081
Politikai intézményekbe vetett bizalom	909	-2.24525	3.76730	.0000000	1.00000000	.498	.081
Állami biztonsági intézményekbe vetett bizalom	909	-2.85822	3.53111	.0000000	1.00000000	-.105	.081
Érdekképviseleti és véleményformáló intézményekbe vetett bizalom	909	-2.52058	3.51270	.0000000	1.00000000	.214	.081
Vallási intézményekbe vetett bizalom	909	-2.92113	1.70914	.0000000	1.00000000	-.560	.081
Valid N (listwise)	909						

103. ábra. *A faktorváltozók leíró statisztikái (46. példa)*

A modellből kimaradt 5 változót (*Oktatási rendszer, Közigazgatás, Egészségügyi rendszer, Igazságszolgáltatási rendszer* és *Közösségi média*) egy újabb főkomponens-elemzéssel (a korábbiakban leírtak szerint) újra tömörítjük. Mind az öt változó egy faktorba tartozik, ezért nincs szükség rotálásra. A modell mutatói jók: a KMO = 0,751, a Bartlett-teszt szignifikáns (p = 0,000), a kommunalitások 0,3 felettiek, a magyarázott összvariancia 50,3%, így létrehozunk egy újabb faktorváltozót (6), a *Közintézményi bizalom* háttérváltozót (el is mentjük regressziós módszerrel). Összességében a 17 intézményi bizalmat mérő változót tehát hat háttérváltozóba sűrítettük faktorelemzés segítségével.

Végül érdemes ellenőrizni a faktorváltozók belső konzisztenciáját. A *Cronbach's alpha* mutató ennek a mérésére szolgál, és azt mutatja meg, hogy a faktort alkotó változók mennyire mérik konzisztensen ugyanazt a látens konstruktumot. Minél magasabb az érték (ideálisan 0,7 felett), annál megbízhatóbb a faktor. Az SPSS-ben ez az *ANALYZE* menü *Scale* almenüjében található *Reliability Analysis* opcióval kérhető le úgy, hogy a faktorokat alkotó változókat az *Items* mezőbe helyezzük, a *Statistics* opciónál bejelöljük az *Item* lehetőséget, majd az Ok paranccsal lefuttatjuk az elemzést (104. ábra). Az *Output*-ban a Cronbach's Alpha érték mutatja a belső konzisztencia mértékét, és ha 0,7 alatti, érdemes megfontolni egyes változók eltávolítását a faktorból.



104. ábra. *A Cronbach's Alpha lekérése (46. példa)*

A *Közintézményi bizalom* (6) faktorváltozónk esetében (97. ábra) a Cronbach's Alpha értéke 0,748, tehát a faktorunk belső konzisztenciája jó. Ugyanígy hasonlóan jó eredményeket kapunk a *Nemzetközi és gazdasági szervezetek iránti bizalom* (1) faktor komponenseire (Cronbach's Alpha = 0,739) és a (2) *Politikai intézményekbe vetett bizalom* faktor (0,843) mért változóira is. Az (3) *Állami biztonsági intézményekbe vetett bizalom* (0,544) és a (4) *Érdekképviseleti és véleményformáló intézményekbe vetett bizalom* (0,611) faktorváltozók esetén a Cronbach's Alpha értékek alacsonyabbak, csak közepes szintű belső konzisztencia van a konstruktumok között.

△ *Gyakorlófeladatok főkomponens-elemzésre*

1. Végezzünk főkomponens-elemzést az SPSS-ben a K1-es kérdés (különböző *értékek*) változóin (*v1-v6*) a bemutatott 46. példa alapján!
2. Végezzünk főkomponens-elemzést az SPSS-ben a K22-es kérdés (*a sikeres házasság komponensei*) változóin (*v65–v70*) a bemutatott 46. példa alapján!

5.4. A klaszterelemzés

Miként a többváltozós elemzések rövid összefoglalásánál láttuk (*5.1. alfejezet*), a klaszterelemzés előre nem ismert csoportok képzésére használatos eljárás. Tehát a klaszterelemzést arra használjuk, hogy a vizsgálatba bevont minden egyes ismérv szerint a hasonló egységek (egyének) azonos, a különbözők pedig eltérő csoportokba (klaszterekbe) kerüljenek. Akárcsak a faktorelemzésnél, ennél az eljárásnál sem kell megkülönböztetni a függő és a független változókat.

A módszer alapvetően feltáró jellegű, vagyis nem vonható le belőle következtetés az alapsokaságra nézve. Akárcsak a faktorelemzés esetében, a klasztereket létre lehet hozni, de a kutatónak kell eldöntenie, hogy tudja-e értelmezni őket. A klaszterelemzésbe bevont változóknak magas mérési szintűeknek (skála) kell lenniük, az alacsony mérési szintű változókat (nominális és ordinális) dummy változókként (0 és 1 kódú) lehet bevinni a klaszterelemzésbe.

A klaszterelemzésnek két alapvető típusa van: hierarchikus és nem hierarchikus klaszterelemzés. Mivel a hierarchikus klaszterelemzés nagy adatfájlokon (amelyekkel a társadalomtudományi adatfelvételek nyomán dolgozunk) nem javasolt, ezért csak a nem hierarchikus klaszterelemzéssel (ha n >30) foglalkozunk a jegyzetben.

A nem hierarchikus klaszterelemzés folyamata:

1. az elemzés célja, a vizsgálatba bevont változók,
2. az adatok előkészítése,
3. a klaszterelemzés lekérése,
4. a klaszterelemzés alkalmazhatósági feltételeinek vizsgálata,
5. a klaszterelemzés folyamata,
6. a klaszterek értelmezése és jellemzése,
7. a megbízhatóság és érvényesség vizsgálata, mentés.

A nem hierarchikus klaszterelemzés folyamatát az SPSS-programcsomag használatával mutatjuk be.

Nem hierarchikus (K-közép módszerrel) végzett klaszterelemzés az SPSS-ben

1. Az elemzés céljának megfogalmazása, a vizsgálatba bevont változók

A klaszterelemzés célja általában az, hogy homogén csoportokat azonosítson a vizsgált sokaságon belül, vagyis hasonló tulajdonságokkal rendelkező eseteket/egységeket csoportosítson.

A klaszterelemzés során az SPSS minden esetben létrehoz klasztereket, függetlenül attól, hogy azok ténylegesen léteznek-e. Mivel a klasztermegoldások teljesen az elemzésbe bevont változóktól függenek, nagyon kell vigyáznunk, hogy milyen változókat választunk ki az elemzésre. Továbbá a gyakorlati tapasztalat

azt mutatja, hogy amikor előzetes elgondolás nélkül vonjuk be a változókat, nem igazán reménykedhetünk sikeres értelmezésben.

A változók kiválasztásánál fontos szempont:

1. az elméleti indokoltság (szakirodalmi előzmények: korábbi kutatások, hipotézisek),

2. a statisztikai szempontok: nem lehet túlzott multikollinearitás, és a minta méretéhez illeszkedjen a bevont változók száma (általános szabály szerint legalább 2-3-szor több eset kell legyen, mint változó, és a legkisebb mintanagyság 100 eset).

Vegyünk két fiktív példát klaszterelemzésre, a munkahelyi elégedettséget (a) és a szabadidős tevékenységeket (b).

a) A klaszterelemzés célja a munkavállalók elégedettségi és elköteleződési csoportosulásainak feltárása. Az elemzésbe bevont változók: munkahelyi elégedettség, szervezeti elkötelezettség, stresszszint, munka–magánélet egyensúly, fejlődési lehetőségek, bérezés és juttatások megítélése stb. Az eredmények alapján homogén dolgozói csoportok azonosíthatók, amelyekhez differenciált HR-stratégiák rendelhetők.

b) A klaszterelemzés célja, hogy a lakosságot szabadidős tevékenységeik, társas kapcsolataik és kulturális fogyasztási szokásaik alapján csoportosítsuk. A vizsgálatba bevont változók: sportolási gyakoriság, kulturális programokon való részvétel, baráti találkozások gyakorisága, önkéntes aktivitás, médiafogyasztási szokások stb. Az elemzés eredményeként különböző életstíluscsoportok (pl. *aktív közösségi életet élők*, *kulturális fogyasztók*, *passzív szabadidő-eltöltők*) azonosíthatók, amelyek hozzájárulnak a társadalmi rétegződés és közösségi mintázatok jobb megértéséhez.

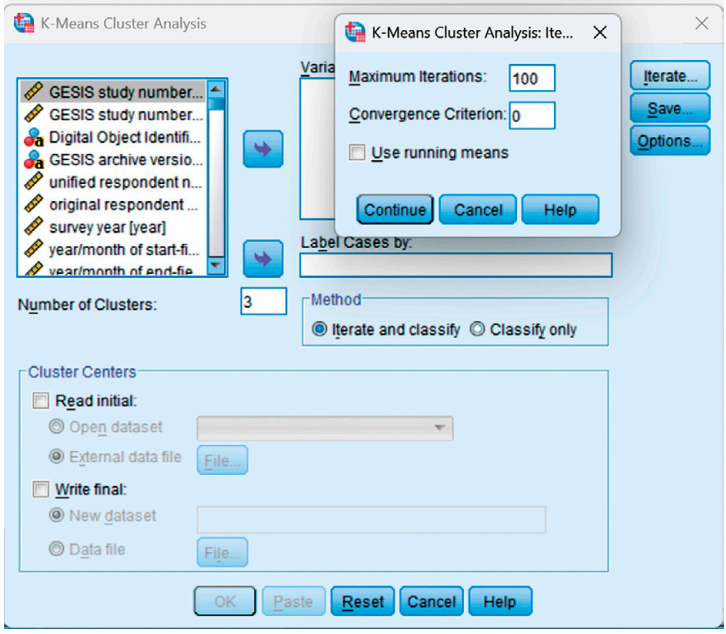
2. Az adatok előkészítése

Ebben a lépésben történik a klaszterelemzésbe bevonni kívánt változók adatbázisban való azonosítása. Miként már korábban is említésre került, klaszterelemzést *csak mennyiségi változók* vagy *dummy* változók bevonásával végezhetünk. Továbbá a klaszterelemzés szempontjából rendkívül fontos, hogy ne legyenek *túlságosan kiugró adataink* (*Outliers)*, ezért nagyon figyeljünk az adattisztításra – az SPSS-ben erre vonatkozó módszereket a többváltozós lineáris regressziónál (*5.2. alfejezet*) már ismertettük.

Mivel a klaszterelemzés a távolságra alapszik, nem mindegy, hogy milyen nagyságrendű adataink vannak. Ha a változóink nem egyforma skálán lettek mérve, akkor nagyon torz adatokat kapunk, ezért a változókat standardizált formában kell bevinnünk a klaszterelemzésbe (a *Zscore* változókat). A *standardizálás,* ahogyan ezt már korábban is jeleztük, tulajdonképpen azt jelenti, hogy az átlagot kivonjuk az egyes értékekből, és a különbséget elosztjuk a szórással. Az SPSS-ben ezt a többváltozós lineáris regressziónál (Az adatok előkészítése lépésnél) leírtak szerint hozzuk létre (*ANALYZE* főmenü, *Descriptive Statistics, Descriptives, Save standardized values as variables).*

3. A klaszterelemzés lekérése

Nem hierarchikus (K-közép) klaszterelemzést az ANALYZE főmenü *Classify, K-Means Cluster* menüpont alatt kérhetünk. A megszokott módon balról jobbra átvisszük az elemzésbe bevont változókat. Mielőtt azonban lefuttatnák a klaszterelemzést, meg kell adnunk a klaszterek számát a változók alatt szereplő *Number of Clusters* mezőnél. Ez azt jelenti, hogy vagy előzetes elvárásokra támaszkodva, vagy „vakon" kell eldöntsük, hogy hány klaszterbe kívánjuk besorolni eseteinket (pl. 3). Az *Iterate* mezőben átállítjuk az ismétlések számát 10-ről 100-ra, mivel feltételezzük, hogy 10 ismétlés nem vezet végleges klaszterstruktúrához, és lefuttatjuk a klaszterelemzést (105. ábra).



105. ábra. *A klaszterelemzés lefuttatása*

Az *Output* ablakban megtekinthetjük a kezdeti klaszterközéppontokat tartalmazó táblázatot, az *Iteration History* tábla azt jelzi, hogy a program hány iterálás (közelítés, ismétlés) után jutott el a végső klaszterstruktúrához.

4. A klaszterelemzés alkalmazhatósági feltételeinek vizsgálata

A klaszterelemzés lefuttatásának két főbb alkalmazhatósági feltételét szükséges megvizsgálnunk az SPSS-ben.

1. A bevont változókra vonatkozó feltétel: a *multikollinearitás*

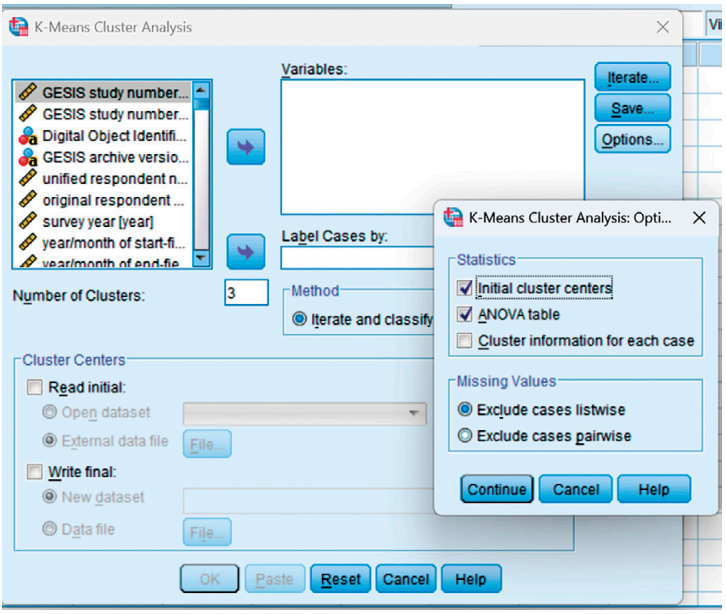
A modellben elemezni kívánt változókra vonatkozó legfontosabb feltétel (a többváltozós lineáris regresszióhoz hasonlóan), hogy egymástól lineárisan füg-

getlenek kell legyenek, vagyis egyik változót se lehessen a többi változó lineáris kombinációjaként előállítani (multikollinearitás). Ha káros multikollinearitás lép fel, vagyis az eljárásban szereplő változók között erős korreláció van ($r > 0{,}7$), ezek a változók nagyobb szerepet fognak kapni az elemzésben és így az eredményekben is, ezért egyiküket ki kell zárni az elemzésből. Másképp fogalmazva, a redundáns információk torzításhoz vezetnek, ezért a klaszterelemzésben minden változónak azonos kell legyen a súlya.

A multikollinearitás tesztelését Pearson-féle korrelációs együtthatók lekérésével végezhetjük: *ANALYZE, Correlate, Bivariate.*

2. A klaszterek szignifikáns elkülönülésére vonatkozó feltétel ellenőrzése az elemzésbe bevont változók mentén: *ANOVA.*

Az ANOVA a klaszterelemzésben nem a klasszikus varianciaanalízis feltételeit (normalitás, szóráshomogenitás) ellenőrzi, hanem arra szolgál, hogy megmutassa, a klaszterek szignifikánsan különböznek-e az adott változó mentén. Bár az SPSS arra törekszik, hogy olyan csoportokat hozzon létre, amelyek egymástól jól elkülönülnek, mégis érdemes megvizsgálni a K-közép klaszterelemzés menü *Options* almenüjében lekérhető (nem klasszikus) ANOVA-tesztet (106. ábra).



106. ábra. *Az ANOVA-teszt lekérése*

Tehát itt az ANOVA feltétele inkább az, hogy a klaszterek ténylegesen különböző átlagú csoportokat alkossanak. Ez biztosítja, hogy az adott változó ténylegesen hozzájárul a klaszterek elkülönítéséhez. A táblázatban a p-érték jelzi, hogy

melyik változók járultak hozzá leginkább a klaszterek elkülönítéséhez és melyek nem. Ha egy változónál szereplő F-érték nem szignifikáns (p>0,05), akkor az azt jelenti, hogy az illető változó mentén nem sikerült homogén csoportokat kialakítani, és el kell hagyni a modellből vagy nagyobb klaszterszámmal kell próbálkozni. Minél nagyobb az F értéke, annál fontosabb szerepet játszik az illető változó a klaszterstruktúra kialakításában.

5. A klaszterelemzés folyamata

A nem hierarchikus vagy dinamikus klaszterelemzést tehát a K-közép (*K-Means*) módszerrel végezzük az SPSS-ben. A K-közép klaszterezés algoritmusa az euklideszi távolságszámításon (az egyes változók közötti különbségek négyzetösszegének a négyzetgyöke) alapszik. A K-közép eljárás a kiinduláskor megadott klaszterszám alapján választ ki kezdeti klaszterközéppontokat (*Initial cluster centers*), vagyis minden klaszterhez egy középpontot rendel. A *kezdeti klaszterközéppontok* tulajdonképpen az adatfájl első *k* (*k* a kért klaszterek száma) elemének adatait jelentik (ezek a kezdőpontok nem láthatóak, mivel a valódi kezdeti középpontokat egy algoritmussal alakítja ki a program), és ezek után kerül behelyezésre a többi elem. Tehát a klaszterelemzésben fontos lehet az esetek sorrendje. A program akkor cserél ki egy már kiválasztott klaszterközéppontot, ha az új eset távolsága (euklideszi) a hozzá legközelebb eső klaszterközépponthoz képest nagyobb, mint a két egymáshoz legközelebb eső klaszterközéppont távolsága. A klaszterbe sorolás kritériuma pedig az, hogy egy elem abba a klaszterbe kerül, amelynek a középpontjához a legközelebb van. Amennyiben új klaszterközéppontot talál a program, a klaszterképző változók átlagai alapján újra kiszámítja az új klaszterközéppontokat, és minden esetet újra behelyez. Mindez a folyamat több iterálás (ismétlés, közelítés) révén addig folytatódik, míg kialakul egy stabil klaszterstruktúra, vagyis a klaszterközéppontok tovább nem változnak. A klaszterek értelmezése a végső klaszterközéppontok (*Final cluster centers*) alapján történik.

Az SPSS által automatikusan generált és az *Output*-ban megjelenő *Final Cluster Centers* táblázat a végleges klaszterközéppontokat tartalmazza, a *Number of Cases in each Cluster* táblázatban pedig az egyes klaszterekhez tartozó esetszámok vannak feltüntetve.

6. A klaszterek értelmezése és jellemzése

A klasztereket tehát az euklideszi távolság alapján számolt végleges klaszterközéppontok alapján jellemezzük (*Final Cluster Centers*).

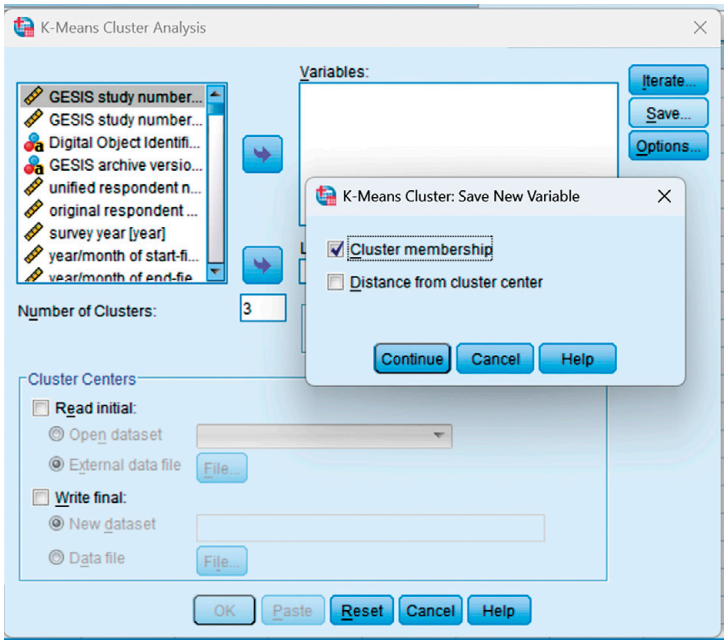
A Final Cluster Centers táblázat mutatja meg, hogy az egyes változók átlagértékei hogyan alakulnak a kialakított klaszterekben. Ezek a középpontok a klaszterek tipikus profilját írják le. Az értelmezés során az a feladat, hogy összehasonlítsuk a klasztereket az egyes változók mentén, és megfogalmazzuk, hogy miben különböznek egymástól. Fontos, hogy nem önmagukban, hanem relatív különbségeikben értelmezzük a klaszterközéppontokat.

Ha az alfejezet elején felhozott b) fiktív alkalmazási példát nézzük, erre a *Final Cluster Centers* segítségével három életstíluscsoportot különíthetünk el (fiktív, szemléltető adatok). Az első klaszterben az emberek gyakran sportolnak (átlag: 4,2), rendszeresen járnak kulturális programokra (4,0), és aktív közösségi életet élnek (baráti találkozások: 4,5). A második klaszter tagjai ritkán sportolnak (2,1), de sokat járnak színházba, koncertre (4,3), miközben kevésbé társaságiak (2,5). A harmadik klaszterben lévők szinte minden mutatóban alacsony értékeket mutatnak (sport: 1,5, kultúra: 1,8, barátok: 2,0). Ezek alapján az 1. klaszter lehetne pl. az *Aktív közösségépítők* (magas sportolási, kulturális és társas aktivitás), a 2. klaszter a *Kulturális egyéniségek* (erős kulturális érdeklődés, de gyenge sport- és társas aktivitás) és a 3. klaszter a *Passzív visszahúzódók* (alacsony sport, kevés kulturális részvétel és ritka társas kapcsolatok). Az összehasonlításból jól látszik, hogy a középpontok segítenek a klaszterek tipikus profiljának értelmezésében.

7. A megbízhatóság és érvényesség vizsgálata, mentés

Mielőtt elmentenénk klaszterváltozóinkat, még egyszer ellenőrizzük le a kapott klaszterstruktúrát. Miként már korábban említésre került, a klaszterstruktúra kialakítását befolyásolja az adatbázisban szereplő esetek sorrendje, mivel a használt klaszterezési eljárás az adatfile első *k* darab elemének adataiból kiindulva határozta meg az iniciális klaszterközéppontokat. Ezért ellenőrizni kell, hogy az elemzési egységek más sorba rendezése után (más iniciális klaszterközéppontok) is ugyanezt a végső klaszterstruktúrát adják-e. Egy olyan változó szerinti új sorba rendezés javasolt, amivel a klaszterstruktúra változói gyengén korrelálnak (*DATA, Sort Cases)*. A sorba rendezés után (akár 2-3 változó szerint is) futtassuk le újra a klaszterelemzést, és ellenőrizzük le, hogy lényegét tekintve változatlanok maradtak-e a végső klaszterközéppontok alapján kirajzolódó csoportok.

Amennyiben a klaszterváltozóinkat további elemzésekbe kívánjuk bevonni, akárcsak a faktorok esetén, a klaszterkódokat tartalmazó változó is elmenthető. Ezt a K-közép klaszterelemzés *Save* menüpontja segítségével tehetjük meg (107. ábra). A *Save New Variable, Cluster membership* révén egy kategoriális változót kapunk, amelyben az 1-es érték az első klaszterhez, a 2-es a második stb. klaszterhez való tartozást jelzi. A *Save New Variable, Distance from cluster center* utasítással a klaszterváltozó egy mennyiségi ismérv lesz, amely a klaszterközépponttól való távolságot (euklideszi) jelzi.



107. ábra. *A klaszterek mentése*

A gyakorlatban a könnyebb értelmezhetősége miatt a klaszterbe tartozás szerint szokás menteni a kapott klaszterváltozót.

47. példa ▼

▶ *K-közép klaszterelemzés az SPSS-ben*

Adatbázisunkban a *v103–v107* (az adatbázisban a 133–137. sorszámú) változók *különböző témákkal kapcsolatos ellentétes vélemények*nek egy 1–10 fokú skálán való elhelyezését tartalmazzák a kérdőív K30-as kérdése szerint. Ahogyan korábban a vegyes kapcsolat esetében is jeleztük, bár ezek a változók sem klasszikus értelemben vett skálaváltozók, de elég széles intervallumban lettek mérve ahhoz, hogy a társadalomtudományokban klaszterelemzésre használhassuk őket.

A K30-as kérdés elemezni kívánt 5 változója a társadalmi-gazdasági ideológiát és az állami szerepvállaláshoz kapcsolódó attitűdöket méri, konkrétabban az individualizmus vs. kollektivizmus vagy egyéni vs. állami felelősség kérdéskörét. Ezek a skálák a politikai-ideológiai orientáció fontos indikátorai, amelyek mentén a társadalom tagjai elhelyezhetők a szociáldemokrata/baloldali értékektől (magasabb állami szerepvállalás) a liberális/jobboldali értékekig (egyéni felelősség hangsúlyozása). Klaszterelemzéssel ezekből az ideológiai skálákból különböző társadalmi csoportok rajzolódhatnak ki, amelyek

eltérő jólétiállam-felfogással és munkaetikával rendelkeznek. Tipikusan 3-5 klaszter alakulhat ki, például egy *szociáldemokrata csoport* (erős állami szerepvállalás, bőkezű munkanélküli támogatás), egy *neoliberális csoport* (egyéni felelősség hangsúlyozása, szigorú munkamorál), egy *szocioliberális csoport* (állami gondoskodás, erős munkamorál) és más különböző ambivalens csoportok, amelyek vegyes attitűdökkel rendelkeznek. Ez lehetőséget ad arra, hogy a társadalmat ne egy egydimenziós bal–jobb tengelyen, hanem összetett ideológiai profilok alapján kategorizáljuk, és megértsük, hogy különböző társadalmi rétegek hogyan viszonyulnak a munka, az állam és az egyéni felelősségvállalás kérdéseihez.

Első lépésben vizsgáljuk meg az elemezni kívánt öt változót. Vegyük észre, hogy a K30-as kérdés D kérdése (*v106*) esetében az értékelés fordított irányú: az alacsonyabb értékek a szociáldemokrata/baloldali irányultságot jelzik, míg a többi 4 változó esetében a liberális/jobboldali irányultságot. Ezért a *v106*-os változót fordítva kell kódolni az elemzéshez (az 1-esből 10-es, a 2-esből 9-es stb. értékeket kell kódolni). A faktorelemzéshez hasonlóan (itt csak az 1-es baloldali és 10-es jobboldali értékcímkéket és a magyar változóneveket kell megadni) hozzuk létre az 5 új változót úgy, hogy a *v103–v105* és *v107*-es változókat az eredeti értékekkel (a nem releváns válaszokat *Missing System*-ként kódoljuk), a *v106*-ost pedig a megfordított értékekkel (*K30_4*) reprodukáljuk (*TRANSFORM, Recode into Different Variables*). Az új változókat átkódoláskor itt is a kérdőív szerint nevezzük el *K30_1-K30_5*, majd a *Variable View*-ban felcímkézzük a kérdőív szerint (108. ábra). Tehát így minden változónkban az alacsony értékek a liberális/jobboldali irányultságot (1. jobboldali), a magasabb értékek a szociáldemokrata/baloldali irányultságot (10. baloldali) mutatják.

Name	Type	Width	D...	Label	Values
K30_1	Numeric	8	0	Az egyén vs. az állam felelőssége a szolgáltatások biztosítása	{1, jobb...
K30_2	Numeric	8	0	Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására munkanélküliség esetén	{1, jobb}...
K30_3	Numeric	8	0	A verseny jó vs. káros	{1, jobb}...
K30_4	Numeric	8	0	Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése	{1, jobb}...
K30_5	Numeric	8	0	Magántulajdonú vs. állami tulajdonú vállalkozás	{1, jobb}...

108. ábra. *A klaszterelemzésbe bevont átkódolt változók az adatbázisban (47. példa)*

A klaszterelemezni kívánt *K30_1-K30_5* változók leíró statisztikáit (*ANALYZE, Descriptive Statistics, Descriptives*) a 109. ábra mutatja.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Az egyén vs. az állam felelőssége a szolgáltatások biztosítása	1071	1	10	5.10	3.287
Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására munkanélküliség esetén	1069	1	10	3.92	2.904
A verseny jó vs. káros	1071	1	10	3.71	2.833
Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése	1073	1	10	5.35	3.313
Magántulajdonú vs. állami tulajdonú vállalkozás	1026	1	10	4.93	2.924
Valid N (listwise)	996				

109. ábra. *A klaszterelemzéssel elemzendő változók leíró statisztikái (47. példa)*

Mivel nincsenek kiugró adataink, és a változók ugyanolyan skálán lettek mérve (nem szükséges standardizált változókkal dolgoznunk), a multikollinearitást teszteljük Pearson-féle korrelációs együtthatók lekérésével (*ANALYZE, Correlate, Bivariate*). A korrelációs mátrixból azt látjuk, hogy a multikollinearitást ki lehet zárni, hiszen a legnagyobb korrelációs együttható értéke 0,328 (jóval a 0,7-es küszöbérték alatt van) (110. ábra).

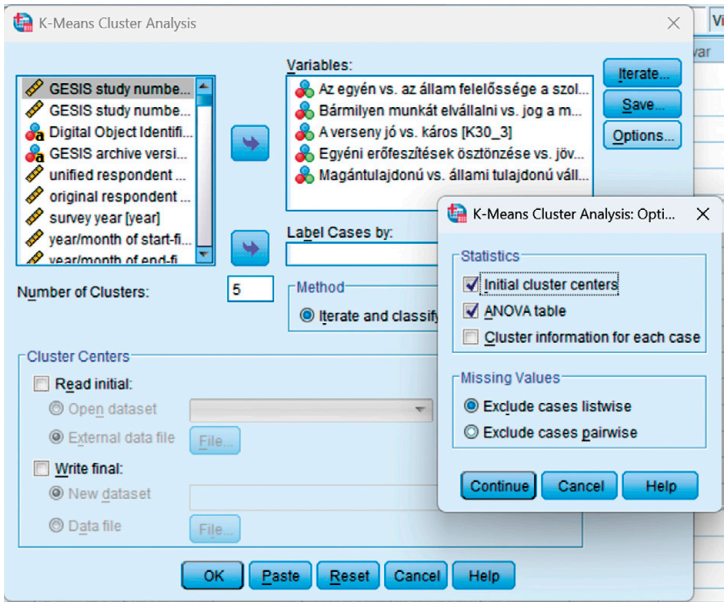
Correlations

		Az egyén vs. az állam felelőssége a szolgáltatások biztosítása	Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására	A verseny jó vs. káros	Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése	Magántulajdonú vs. állami tulajdonú vállalkozás
Az egyén vs. az állam felelőssége a szolgáltatások biztosítása	Pearson Correlation	1	.176**	.149**	-.146**	.194**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	1071	1060	1061	1060	1016
Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására	Pearson Correlation	.176**	1	.234**	-.113**	.174**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1060	1069	1057	1058	1014
A verseny jó vs. káros	Pearson Correlation	.149**	.234**	1	-.121**	.328**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	1061	1057	1071	1062	1019
Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése	Pearson Correlation	-.146**	-.113**	-.121**	1	-.084**
	Sig. (2-tailed)	.000	.000	.000		.007
	N	1060	1058	1062	1073	1019
Magántulajdonú vs. állami tulajdonú vállalkozás	Pearson Correlation	.194**	.174**	.328**	-.084**	1
	Sig. (2-tailed)	.000	.000	.000	.007	
	N	1016	1014	1019	1019	1026

**. Correlation is significant at the 0.01 level (2-tailed).

110. ábra. *A korrelációs mátrix*
(47. példa)

Lefuttatjuk a klaszterelemzést a korábbiakban leírtak szerint, az elméleti feltevésekben megfogalmazottaknak megfelelően egy árnyaltabb, 5 klaszteres modellt kérve, az iterálások számát 100-ra emelve, és lekérve az ANOVA-tesztet is (111. ábra).



111. ábra. *Az ötklaszteres modell lekérése (47. példa)*

Az ötklaszteres modellt 19 ismétlési folyamat (iterálás) után kaptuk meg. Mind az öt változó szignifikánsan járul hozzá az öt klasztercsoport kialakításához ($p = 0{,}000$). Az F-értékek azt jelzik (112. ábra), hogy a modellben az *Az egyén vs. az állam felelőssége a szolgáltatások biztosítása*, az *Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése* és a *Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására munkanélküliség esetén* változók mentén sikerült a leghomogénebb csoportokat kialakítani. A modell az esetek 90,1 százalékát ($N = 996$) tartalmazza.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Az egyén vs. az állam felelössége a szolgáltatások biztosítása	1556.447	4	4.392	991	354.363	.000
Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására munkanélküliség esetén	1078.052	4	4.105	991	262.591	.000
A verseny jó vs. káros	543.309	4	5.851	991	92.863	.000
Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése	1398.076	4	5.275	991	265.051	.000
Magántulajdonú vs. állami tulajdonú vállalkozás	467.057	4	6.643	991	70.311	.000

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.

112. ábra. *Az ANOVA-teszt (47. példa)*

A klasztereket tehát a *Final Cluster Centers* táblázat alapján értelmezzük (113. ábra).

Final Cluster Centers

	Cluster				
	1	2	3	4	5
Az egyén vs. az állam felelőssége a szolgáltatások biztosítása	4	8	2	8	6
Bármilyen munkát elvállalni vs. jog a munkavállalás megtagadására munkanélküliség esetén	4	4	2	2	8
A verseny jó vs. káros	6	3	2	3	5
Egyéni erőfeszítések ösztönzése vs. jövedelmek kiegyenlítése	6	9	6	2	2
Magántulajdonú vs. állami tulajdonú vállalkozás	6	5	3	5	6

113. ábra. *Az ötklaszteres modell végső klaszterközéppontjai (47. példa)*

A végső klaszterközéppontok értelmezését és a klaszterek kiolvasását, elnevezését megkönnyíti, ha az értékek helyett a nekik megfelelő értékcímkéket használjuk (30. táblázat).

30. táblázat. *Az ötklaszteres klasztermodell jelentése (47. példa)*

	Klaszterek				
	1	**2**	**3**	**4**	**5**
Az egyén *vs.* az állam felelőssége a szolgáltatások biztosítása	kicsit jobb	bal	jobb	bal	kicsit bal
Bármilyen munkát elvállalni *vs.* jog a munkavállalás megtagadására munkanélküliség esetén	kicsit jobb	inkább jobb	jobb	jobb	bal
A verseny jó *vs.* káros	kicsit bal	jobb	jobb	jobb	közép
Egyéni erőfeszítések ösztönzése *vs.* jövedelmek kiegyenlítése	kicsit bal	nagyon bal	kicsit bal	jobb	jobb
Magántulajdonú *vs.* állami tulajdonú vállalkozás	kicsit bal	közép	jobb	közép	kicsit bal

A klaszterek jellemzése:

1. klaszter: közepesnél kicsit nagyobb egyéni felelősségvállalás az önmagukról való gondoskodásban és a munkavállalásban, illetve közepesnél kicsit nagyobb egyetértés azzal, hogy a verseny káros, a jövedelmek egyenlőbbek kellene legyenek, és több állami vállalkozás kellene legyen az üzleti és ipari szférában (ötvözi a mérsékelt individualizmust és a mérsékelt közösségi, egalitárius értékeket): ***Szolidaristák csoportja***,

2. klaszter: az állam jobban kell gondoskodjon az emberekről, inkább munkát kell vállalni, a verseny jó, a jövedelmeket egyenlőbbé kell tenni, és kiegyenlített állami-magán tulajdonú vállalkozások szükségesek az üzleti és ipari szférában (az állami szerepvállalás és az egyenlőség erős támogatása a munka és a verseny pozitív jelentőségének elismerése mellett): ***Szocioliberálisok csoportja***,

3. klaszter: nagyobb egyéni felelősségvállalás, bármilyen munkát el kell vállalni, a verseny jó, kicsit ki kellene egyenlíteni a jövedelmeket, és több magántulajdonú vállalkozás kellene (egyéni felelősség hangsúlyozása, szigorú munkakényszer, kiegyenlítettebb jövedelmek) ***Mérsékelt neoliberálisok csoportja***,

4. klaszter: az állam jobban kell gondoskodjon az emberekről, inkább munkát kell vállalni, a verseny jó, az egyéni teljesítményt kell ösztönözni, és kiegyenlített állami-magán tulajdonú vállalkozások szüksége-

sek az üzleti és ipari szférában (nagyobb állami felelősségvállalás és szolidaritás, erős munkamorál és piaci dinamizmus): ***Szociáldemokraták csoportja***,

5. klaszter: az állam egy kicsit jobban kell gondoskodjon az emberekről, az embereknek joguk van a munkavállalás megtagadására munkanélküliség esetén, a verseny se nem jó, se nem káros, az egyéni erőfeszítéseket nagyobb mértékben kellene jutalmazni, és több állami vállalkozás kellene legyen az üzleti és ipari szférában (inkább államibeavatkozás-párti az egyéni erőfeszítések erőteljesebb értékelésével): ***Pragmatisták csoportja***.

Mielőtt elmentenénk klaszterváltozóinkat, még egyszer leellenőrizzük a kapott klaszterstruktúrát, pl. a 100. sorszámú *v74* változó szerint. Ez a változó egyetlen klaszterváltozóval sem korrelál, így az e szerint sorba rendezett (*DATA, Sort Cases)* adatbázison újra lefuttatjuk a klaszterelemzést. Azt látjuk, hogy bár a klaszterek sorrendje változott (a *Szolidaristák csoportja* a második modellben a 3. klaszter, a *Szocioliberálisok* az 5. klaszter stb.), de az egyes klaszterecsoportok értelme semmit és elemszámuk csak minimálisan módosult (pl. 155 helyett 158 eset). Tehát elfogadjuk ezt az ötcsoportos modellt végső klasztermodellnek.

Utolsó lépésként elmentjük a klaszereket a klaszercsoport tagsága szerint, majd a klaszernevek alapján felcímkézzük a klaszterváltozó attribútumait, és kérünk rá egy gyakorisági eloszlást (114. ábra).

Egyéni vs. állami felelősség és munkaetika szerinti klaszterek

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Szolidaristák csoportja	270	24.4	27.1	27.1
	Szocioliberálisok csoportja	155	14.0	15.6	42.7
	Mérsékelt neoliberálisok csoportja	237	21.4	23.8	66.5
	Szociáldemokraták csoportja	188	17.0	18.9	85.3
	Pragmatisták csoportja	146	13.2	14.7	100.0
	Total	996	90.1	100.0	
Missing	System	110	9.9		
Total		1106	100.0		

114. ábra. *Az öt klasztercsoport nagysága a mintában (47. példa)*

A klaszterek mérete jól mutatja, mely értékkombinációk a legelterjedtebbek. A legnagyobb csoport a *Szolidaristáké* (27%), akik mérsékelt egyéni felelősséget és mérsékelt egalitarianizmust ötvöznek. Őket szorosan követik a

Mérsékelt neoliberálisok (24%), ahol az egyéni felelősség és a verseny előnyei dominálnak. A középmezőt a *Szociáldemokraták* (19%) adják, akik az állami gondoskodást erős munkamorállal kapcsolják össze. Kisebb, de markáns csoport a *Szocioliberálisoké* (16%), akik az állami szerepvállalást és az egyenlőséget támogatják, miközben pozitívan értékelik a versenyt. A legkisebb klaszter a *Pragmatistáké* (15%), akik enyhén állampártiak, de az egyéni erőfeszítések jutalmazását is kiemelik.

MELLÉKLETEK

A χ^2-eloszlás táblázata ($p = 0,05$, $p = 0,01$ és $p = 0,001$)

Szabadságfok	Szignifikanciaszint		
	p = 0,05	p = 0,01	p = 0,001
1	3,841	6,635	10,827
2	5,991	9,210	13,815
3	7,815	11,345	16,268
4	9,488	13,277	18,465
5	11,070	15,086	20,517
6	12,592	16,812	22,457
7	14,067	18,475	24,322
8	15,507	20,090	26,125
9	16,919	21,666	27,877
10	18,307	23,209	29,588
11	19,675	24,725	31,264
12	21,026	26,217	32,909
13	22,362	27,688	34,528
14	23,685	29,141	36,123
15	24,996	30,578	37,697
16	26,296	32,000	39,252
17	27,587	33,409	40,790
18	28,869	34,805	42,312
19	30,144	36,191	43,820
20	31,410	37,566	45,315
21	32,671	38,932	46,797
22	33,924	40,289	48,268
23	35,172	41,638	49,728
24	36,415	42,980	51,179
25	37,652	44,314	52,620
26	38,885	45,642	54,052
27	40,113	46,963	55,476
28	41,337	48,278	56,793
29	42,557	49,588	58,302
30	43,773	50,892	59,703

A t-eloszlás táblázata (p = 0,05, p = 0,01 és p = 0,001)

Szabadságfok	Szignifikanciaszint		
	p = 0,05	p = 0,01	p = 0,001
1	12,706	63,657	636,619
2	4,303	9,925	31,598
3	3,182	5,841	12,941
4	2,776	4,604	8,610
5	2,571	4,032	6,859
6	2,447	3,707	5,959
7	2,365	3,499	5,405
8	2,306	3,355	5,041
9	2,262	3,250	4,781
10	2,228	3,169	4,587
11	2,201	3,106	4,437
12	2,179	3,055	4,318
13	2,160	3,012	4,221
14	2,145	2,977	4,140
15	2,131	2,947	4,073
16	2,120	2,921	4,015
17	2,110	2,898	3,965
18	2,101	2,878	3,922
19	2,093	2,861	3,883
20	2,086	2,845	3,850
21	2,080	2,831	3,819
22	2,074	2,819	3,792
23	2,069	2,807	3,767
24	2,064	2,797	3,745
25	2,060	2,787	3,725
26	2,056	2,779	3,707
27	2,052	2,771	3,690
28	2,048	2,763	3,674
29	2,045	2,756	3,659
30	2,042	2,750	3,646
40	2,021	2,704	3,551
60	2,000	2,660	3,460
120	1,980	2,617	3,373
∞	1,960	2,576	3,291

Az IBM SPSS Statistics 22.0 program menüsor parancsainak rövid leírása

Áttekintés a program menürendszeréről

Menüpont	Fő funkciók
File (Fájl)	Új adatfájl létrehozása, meglévő fájlok megnyitása, mentés, exportálás, legutóbbi fájlok kezelése, nyomtatás beállítása.
Edit (Szerkesztés)	Adatok, változók és szintaxisfájlok szerkesztése; keresés és csere; adattranszformációs beállítások módosítása.
View (Nézet)	Munkaterület és eszköztárak testreszabása; eredmények megjelenítési beállításainak módosítása.
Data (Adat)	Adatrendezés, fájlok összefűzése vagy szétválasztása, transzponálás, hiányzó értékek definiálása.
Transform (Transzformálás)	Új változók létrehozása, értékek számítása, újrakódolás, feltételes transzformációk alkalmazása.
Analyze (Elemzés)	Leíró statisztikák, átlagok összehasonlítása, általános lineáris modellek, regresszió, nemparaméteres próbák, faktor- és klaszterelemzés, fejlett statisztikai eljárások.
Direct Marketing (Közvetlen marketing)	Eszközök célcsoport-kiválasztásra, ügyfélprofilozásra és kampányhatékonyság elemzésére.
Graphs (Grafikonok)	Oszlopdiagram, hisztogram, pontdiagram, dobozdiagram és további vizualizációk létrehozása.
Utilities (Eszközök)	Változótulajdonságok, fájlinformációk és egyéni attribútumok kezelése.
Add-ons (Bővítmények)	Telepített modulok, egyéni párbeszédpanelek, R és Python integrációk elérése.
Window (Ablak)	Adat-, szintaxis- és eredményablakok közötti navigáció és elrendezés.
Help (Súgó)	Felhasználói kézikönyvek, oktatóanyagok, szintaxisreferencia, online források.

A FILE menü

New (Új)

⊕ **Data**
Üres adatlap létrehozása új adatok beviteléhez (új adatfájl létrehozása).

⊕ **Syntax**
Új szintaxisfájl nyitása SPSS-parancssorok kézi írásához.

⊕ **Output**
Új eredményfájl létrehozása, amelybe a futtatott elemzések eredményei kerülnek.

⊕ **Script**
Egy új parancssoregyüttes, „script" lehívása (a szkript bizonyos helyzetekhez vagy feltételekhez kapcsolódó programrészlet, amely a helyzet vagy a feltétel változásakor lefut).

Open (Megnyitás)

⊕ **Data**
SPSS (.sav) formátumú adatfájlok.*

⊕ **Syntax**
Szintaxisfájlok (.sps).*

⊕ **Output**
Eredményfájlok (.spv, *.htm, *.mht).*

⊕ **Script**
Parancssoregyüttesek, „script"-ek (.sbs).*

Open Database (Adatbázisok megnyitása)

*Meglévő külső ODBC (Open Database Connectivity) adatbázisok importálása (pl. Excel, MS Access) és lehetőség a különböző változók szelektív beolvasására (Database Query, *.spq formátum).*

Read Text Data (Szöveges adatfájl beolvasása)

*Szövegfájlokból (pl. *.txt, *.csv, *.dat) történő adatimportálás, megadhatjuk az oszlopokat elválasztó karaktert (pl. vessző), a változónevek helyét stb.*

Read Cognos Data (Cognos adatok beolvasása)

Cognos-adatok (üzleti intelligencia, BI-rendszerből származó vállalati adatok) közvetlen importálása, a releváns változók kiválasztásával és a kódolások automatikus kezelésével.

Close (Bezárás)

A megnyitott adatfájl bezárása, a módosítások mentésének lehetőségével.

Save (Mentés)

Az aktív munkafájl mentése az aktuális néven és a régebbi változat felülírása SPSS formátumban (.sav).

Save As (Mentés / Mentés másként)

*Más névvel való mentés, másik mappába való mentés, más formátumba történő mentés (pl. *.por, *.xls, *.csv).*

Save All Data (Minden adat mentése)

Az összes megnyitott adatfájl mentése egyszerre.

Export to Databas (Mentés adatbázisba)

Az aktuális adatok külső adatbázisba történő mentése (pl. Excel, MS Access).

Mark File Read Only (Fájlt csak olvashatóként jelöl)

A fájl csak olvashatóként való megjelölése, így nem módosítható véletlenül.

Rename Dataset (Adatfájl átnevezése)

Az adatfájl átnevezése.

Display Data File Information (Adatfájl-információk megjelenítése)

Az SPSS formátumú (.sav kiterjesztésű) adatfájlokról és annak változóiról ad információt, a legfontosabb attribútumoknak az output ablakban való kiírásával (pl. változók nevei, típusai, értékcímkék).*

Cache Data (Az adatok gyorsítótárba helyezése)

Az adatok ideiglenes tárolása a gyorsabb elérés érdekében, így a műveletek gyorsabban hajthatók végre.

Repository (Rendszerezett adattár)

Egy központi hely, ahol mentett adatforrások, jelentések, elemzések és modellfájlok tárolhatók és újra felhasználhatók.

Print Preview (Nyomtatás előnézete)

Nyomtatás előtt lehetőség van a formázásra és a tartalom kiválasztására.

Print (Nyomtatás)

Fájlok vagy adatlapok nyomtatása.

Recently Used Data (Legutóbbi adatok)

Gyors elérés az utoljára megnyitott adatfájlokhoz (csak adatfájlok).

Recently Used Files (Legutóbbi fájlok)

Gyors elérés az utoljára megnyitott különböző típusú fájlokhoz (minden SPSS fájl).

Exit (Kilépés)

A program bezárása, mentési lehetőséggel.

Az EDIT menü

Undo/Redo (Visszavonás/Újra)

Az utolsó művelet visszavonása vagy újbóli végrehajtása (a syntax, az output és a script fájlokban nem aktív).

Cut/Copy/Paste (Kivágás/Másolás/Beillesztés)

Kijelölt adatok vagy szöveg kivágása, másolása és beillesztése.

Paste Variables (Változó beillesztése)

Változó beillesztése a kijelölt helyre.

Clear (Törlés)

A kijelölt cellák vagy objektumok tartalmának törlése, és a törlés következtében nem keletkeznek üres sorok vagy oszlopok.

Insert Variable / Insert Cases (Változó beszúrása / Eset beszúrása)

Új változó vagy új sor (eset) beszúrása az adatfájlba.

Select All (Mindet kijelöl)

Az összes adat vagy szöveg kijelölése.

Find/Find Next/Replace (Keresés/Következő keresés/Csere)

Adatok vagy szövegek keresése, illetve cseréje.

Go To Case (Ugrás esetre)

Adott sor (eset) gyors elérése az adatfájlban.

Go To Variable (Ugrás változóra)

Adott változó gyors elérése a változólistában.

Go To Imputation (Ugrás imputáláshoz)

Az adatfájlban az adott imputált (pótlással előállított) értékhez való gyors navigálást teszi lehetővé.

Options (Lehetőségek)

Az SPSS működését előzetesen szabályozó parancsok találhatók meg itt, beállítható a munkaterület a háttértárolón, a műveletek végrehajtásának a módjai, a grafikus megjelenítés módjai, a nyelv (a magyar és román nem elérhető) stb.

A VIEW menü

Status Bar (Állapotsor)

Az SPSS ablak alján található sáv, amely információt jelenít meg az aktuális műveletekről, fájlállapotról és az adatok feldolgozási folyamatáról, kijelzi a számításoknál figyelembe vett esetek számát, jelzi, ha csak bizonyos

esetekkel dolgozunk, ha az adatfájlt több csoportra osztottuk vagy esetleg súlyozott adatbázissal dolgozunk.

Toolbars (Eszköztárak)

Az SPSS-ablak tetején található ikon- és gombsorok, amelyek gyors hozzáférést biztosítanak a leggyakrabban használt parancsokhoz és funkciókhoz.

Fonts (Betűk)

A betűtípus és betűméret beállítása.

Grid Lines (Rácsvonalak)

Az adatbázis oszlopait és sorait elválasztó vonalak megjelenítése.

Value Labels (Értékcímkék)

Az ismérvértékek szöveges vagy numerikus (számokkal kódolt) formában való megjelenítése.

Mark Imputed Data (Imputált adatok megjelölése)

Az imputálással (adatpótlással) létrehozott értékek kiemelése vagy jelölése az adatfájlban.

Customize Variable View (Változónézet testreszabása)

Lehetővé teszi a változónézet oszlopainak megjelenítését, elrejtését és sorrendjük módosítását az SPSS-ben.

Data/Variables (Adatok/Változók)

Az adattábla nézet (Data View) vagy változó nézet (Variable View) közötti váltást teszi lehetővé.

A DATA menü

Define Variable Properties (Változótulajdonságok meghatározása)

Lehetővé teszi a változók jellemzőinek, például típusának, címkéinek és értékcímkéinek beállítását vagy módosítását.

Set Measurement Level for Unknown
(Mérési szint beállítása ismeretlen értékekhez)

Automatikusan vagy manuálisan meghatározza a változók mérési szintjét (nominális, ordinális, arányskála), ha az még nincs megadva.

Copy Data Properties (Adattulajdonságok másolása)

Lehetővé teszi változók tulajdonságainak, pl. típus, címkék, értékcímkék, mérési szint átmásolását egyik adatfájlból vagy változóból a másikba.

New Custom Attribute (Új egyéni attribútum)

Lehetővé teszi felhasználó által meghatározott, egyedi jellemzők vagy metaadatok létrehozását és hozzárendelését a változókhoz.

Define Dates (Időbeállítás)

Az időbeállítás formátumát lehet megadni, olyan időváltozók generálására alkalmas, amelyekkel megadható az idősorok periodicitása.

Define Multiple Response Sets (Többválaszos halmazok definiálása)

Olyan változók csoportosítására szolgál, amelyeknél a válaszadók több választ is megadhatnak egyetlen kérdésre. Az almenü lehetővé teszi Multiple Dichotomy Sets (dichotóm változók halmaza) és Multiple Category Sets (kategoriális változók halmaza) létrehozását. A definiált halmazok speciális elemzési eljárásokban használhatók, amelyek figyelembe veszik a többszörös válaszok sajátosságait, és megfelelő statisztikai számításokat végeznek a gyakorisági és kereszttáblás elemzések során.

Validation (Érvényesség-ellenőrzés)

Az adatok minőségének és konzisztenciájának ellenőrzése, lehetővé teszi szabályok definiálását az adatmezők értékeire vonatkozóan, pl. hiányzó vagy hibás azonosítása.

Identify Duplicate Cases (Duplikált esetek azonosítása)

Az adatbázisban található ismétlődő rekordok automatikus felderítésére és megjelölésére szolgál.

Identify Unusual Cases (Szokatlan esetek azonosítása)

Olyan rekordok felderítésére szolgál, amelyek statisztikai szempontból kiugrónak vagy atipikusnak minősülnek az adathalmazban.

Compare Datasets (Adathalmazok összehasonlítása)

Két adathalmaz közötti különbségek és eltérések azonosítására szolgál.

Sort Cases (Esetek rendezése)

Adatsorok rendezése egy vagy több változó szerint.

Sort Variables (Változók rendezése)

A változók sorrendjének átrendezésére szolgál az adathalmaz változólistájában.

Transpose (Transzponálás)

Sorok és oszlopok felcserélése az adatfájlban.

Merge Files (Fájlok egyesítése)

*Adatfájlok összekapcsolása, egyesítése sorok (**Add Cases**) vagy változók (**Add Variables**) szerint.*

Restructure (Adatszerkezet átalakítása)

Az adattábla sor- és oszlopstruktúrájának módosítására szolgál, lehetővé téve a változók és esetek elrendezésének átalakítását.

Aggregate (Aggregálás)

Az adatok csoportosítására és összesítő statisztikák számítására (átlag, legkisebb érték, összeg stb.) szolgál meghatározott változók szerint.

Orthogonal Design (Ortogonális terv)

Kísérleti elrendezések tervezésére szolgál, ahol a faktorok hatásai egymástól függetlenül vizsgálhatók: új adatbázist hoz létre, amely néhány változó vagy változóegyüttes statisztikai tesztelését teszi lehetővé (független leképzésen alapuló minta).

Copy Dataset (Adathalmaz másolása)

Az aktív adathalmaz teljes másolatának létrehozására szolgál a memóriában.

Split File (Fájl felosztása)

Az adatmátrixot egy megadott változó értékei szerint részekre lehet bontani, hogy a részeken külön-külön statisztikai analízist vagy grafikus megjelenítést lehessen végezni.

Select Cases (Esetek kiválasztása)

Az adathalmaz meghatározott feltételeknek megfelelő rekordjainak szűrésére és kiválasztására szolgál a leszűkített elemzéshez.

Weight Cases (Esetek súlyozása)

A sorok, esetek (minta) súlyozása az elemzéshez anélkül, hogy ténylegesen megsokszoroznánk őket az adatmátrixban (az alulreprezentált eseteket nagyobb, a túlreprezentált eseteket kisebb értékkel súlyozzuk).

A TRANSFORM menü

Compute Variable (Változó számítása)

Új változók létrehozására szolgál matematikai műveletek, függvények és logikai kifejezések segítségével. A funkció lehetővé teszi összetett számítások végrehajtását meglévő változók alapján, feltételes értékadást (ilyenkor csak azoknál az eseteknél képződik számított érték, amelyekhez a beállított logikai kifejezés igaz, a többi helyre system missing value kerül) és különféle átalakítások alkalmazását az adatokon.

Count Values within Cases (Értékek számlálása az eseteken belül)

Egy új változó létrehozására szolgál, amely megszámolja, hogy egy meghatározott érték vagy értéktartomány hányszor fordul elő az egyes esetek kiválasztott változóiban.

Shift Values (Értékek eltolása)

A változók értékeinek időbeli vagy logikai eltolására szolgál az adathalmazban, lehetővé téve az előző vagy következő időpontok értékeinek elérését Lag vagy Lead változók létrehozásával.

Recode Into Same Variable (Átkódolás azonos változóba)

A meglévő változók értékeinek módosítására és újrakódolására szolgál, ahol az eredeti változó értékei helyettesítődnek az új kódolással.

Recode Into Different Variable (Átkódolás eltérő változóba)

A meglévő változók értékeinek újrakódolására szolgál egy új változó létrehozásával, miközben az eredeti változó változatlan marad.

Automatic Recode (Automatikus átkódolás)

Kategorikus (szöveges) változók automatikus numerikus kódolására szolgál, ahol a program egymást követő egész számokat rendel az egyedi értékekhez.

Visual Binning (Vizuális csoportosítás)

Folytonos változók kategóriákba sorolására szolgál grafikus felület segítségével, ahol interaktívan meghatározhatók a határértékek és a csoportok száma.

Optimal Binning (Optimális csoportosítás)

Folytonos változók statisztikailag optimális kategóriákba sorolására szolgál, ahol az algoritmus automatikusan meghatározza a legjobb határértékeket a célváltozóval való kapcsolat maximalizálása érdekében.

Prepare Data for Modeling (Adatok előkészítése modellezéshez)

Az adatok automatikus előfeldolgozására szolgál machine learning és statisztikai modellek számára, beleértve a hiányzó értékek kezelését, változók átalakítását és skálázását.

Rank Cases (Esetek rangsorolása)

Az adatok rangsor szerinti rendezésére és rangszámok hozzárendelésére szolgál a kiválasztott változók értékei alapján.

Date and Time Wizard (Dátum és idő varázsló)

Dátum- és időadatok létrehozására, átalakítására és manipulálására szolgál, beleértve a különböző dátumformátumok közötti konverziót és időszámítási műveleteket.

Create Time Series (Idősor létrehozása)

Új idősor változók generálására szolgál meglévő adatok alapján, beleértve a Lag változók, differenciák, mozgóátlagok és egyéb idősor-specifikus transzformációk létrehozását.

Replace Missing Values (Hiányzó értékek pótlása)

A hiányzó adatok helyettesítésére szolgál különböző módszerekkel, mint például átlag, medián, trend alapú becslés vagy más statisztikai eljárások alkalmazásával.

Random Number Generator (Véletlenszám-generátor)

Véletlenszámok vagy véletlenszerű mintavételi sémák létrehozására szolgál különböző eloszlások és paraméterek alapján az adatelemzési és szimulációs célokhoz.

Run Pending Transforms (Függő átalakítások futtatása)

Az előzőleg definiált, de még nem végrehajtott adatátalakítási műveletek alkalmazására szolgál az aktuális adathalmazon.

Az ANALYZE menü

Reports (Jelentések)

⊕ **Codebook (Kódkönyv)**
Az adathalmaz változóinak részletes dokumentációjának létrehozására szolgál, beleértve a változónevek, címkék, értékkészletek és alapstatisztikák összefoglalását.

⊕ **OLAP Cubes (OLAP, Online Analytical Processing kockák)**
Többdimenziós adatelemzésre és összegző táblázatok létrehozására szolgál hierarchikus adatstruktúrákkal. Lehetővé teszi az adatok összegzését és aggregálását különböző kategóriák mentén, az adatokat többdimenziós struktúrába (kocka formájába) szervezi, ahol különböző részletességi szinteken lehet elmélyülni az elemzés során.

⊕ **Case Summaries (Eset-összefoglalók)**
Az egyes esetek részletes adatainak (leíró statisztikák) táblázatos megjelenítésére szolgál kiválasztott változók alapján.

⊕ **Report Summaries in Rows (Összefoglalók sorokban)**
Összesítő statisztikák megjelenítésére szolgál soronkénti elrendezésben csoportosított adatok esetén.

⊕ **Report Summaries in Columns (Összefoglalók oszlopokban)**
Összesítő statisztikák megjelenítésére szolgál oszloponkénti elrendezésben különböző változócsoportok számára.

Descriptive Statistics (Leíró statisztikák)

⊕ **Frequencies (Gyakoriságok)**
Egy vagy több változóhoz gyakorisági táblázatokat és leíró statisztikákat, valamint az eloszlást szemléltető ábrákat készít.

⊕ **Descriptives (Leíró mutatók)**
Az egyváltozós statisztikákat számolja (átlag, szórás, ferdeség, csúcsosság stb.), és ezek standard hibáit (az elméleti értékektől való eltérések becslései). A statisztikákat a változók átlagértékei szerinti csökkenő vagy növekvő sorrendben írathatjuk ki. Lehetőség van egy-egy változó standardizáltjának új változóként való előállítására is.

⊕ **Explore (Feltárás)**
Az eloszlást jellemző további statisztikákat számol, illetve grafikonokat rajzol. Az adatok közepét, az esetleges adathibákat kiszűrve, úgynevezett robosztus becslésekkel (M-estimators) közelíti, megkeresi és kijelzi a tipikustól jelentősen elütő eseteket (outliers), kiszámolja a kvartiliseket és a mediánt. Gyors grafikus normalitásvizsgálat végezhető el, ha a hisztogramra kikérjük a Gauss-görbét. A változók eseteit csoportképző változók segítségével részcsoportokba oszthatjuk, és a részcsoportok statisztikáit különböző grafikonokkal együtt elkészíttethetjük.

⊕ **Crosstabs (Kereszttáblák)**
Kereszttáblák készíthetőek itt két vagy több diszkrét változó együttes előfordulásainak megjelenítésére. A táblázatból különféle, a függetlenség ellenőrzésére szolgáló statisztikák kérhetőek ki (khí-négyzet statisztikák, asszociációs mérőszámok, korrelációs együttható stb.).

⊕ **Ratio (Arányok)**
Két változó közötti arányok és kapcsolódó statisztikák számítására szolgál.

⊕ **P-P Plots (P-P, Probability-Probability diagramok)**
Egy változó empirikus eloszlásfüggvényét a normális eloszlás eloszlásfüggvényével együtt lehet kirajzoltatni.

⊕ **Q-Q Plots (Q-Q, Quantile-Quantile diagramok)**
Kvantilisek összehasonlítására szolgál az elméleti és tapasztalati eloszlások között, a normalitás és más eloszlási feltételezések vizsgálatára.

Tables (Táblázatok)

⊕ **Custom Tables (Egyéni táblázatok)**
Komplex, többdimenziós táblázatok létrehozására szolgál rugalmas formázási és csoportosítási lehetőségekkel.

⊕ **Multiple Response Sets (Többválaszos halmazok)**
Olyan kérdések elemzésére szolgál, ahol a válaszadók egyszerre több választ is megadhatnak speciális gyakorisági és kereszttáblás elemzésekkel.

Compare Means (Átlagok összehasonlítása)

⊕ **Means (Átlagok)**
Egy vagy több csoportképző változó segítségével kialakított alcsoportok leíró statisztikáit számolja.

⊕ **One-Sample T Test (Egymintás t-próba)**
Egy minta átlagának összehasonlítására szolgál egy elméleti értékkel.

⊕ **Independent-Samples T Test (Független mintás t-próba)**
Két független csoport átlagainak összehasonlítására szolgál.

⊕ **Paired-Samples T Test (Páros mintás t-próba)**
Ugyanazon alanyok két mérése közötti különbség vizsgálatára szolgál átlagok alapján.

⊕ **One-Way ANOVA (Egyszempontos varianciaanalízis)**
Egyszeres szórásanalízist hajt végre a különböző csoportok átlagai eltéréseinek ellenőrzésére.

General Linear Model (Általános lineáris modell)

⊕ **Univariate (Egyváltozós)**
Azt vizsgáljuk, hogy egyetlen függő változót hogyan befolyásol egy vagy több független változó és azok kölcsönhatásai.

⊕ **Multivariate (Többváltozós)**
Több függő változó egyidejű elemzésére szolgál ugyanazon független változók hatásának vizsgálatával.

⊕ **Repeated Measures (Ismételt mérések)**
Ugyanazon alanyokon végzett többszöri mérések elemzésére szolgál az időbeli vagy feltételbeli változások vizsgálatára.

⊕ **Variance Components (Varianciakomponensek)**
A véletlen hatásoknak a függő változó varianciájára gyakorolt hatását becsülhetjük meg.

Generalized Linear Models (Általánosított lineáris modellek)

⊕ **Generalized Linear Models (Általánosított lineáris modellek)**
Nem normális eloszlású függő változók modellezésére szolgál különböző link (linearizáló) függvények és hibaeloszlások használatával. Az általános lineáris modellek (ALM) kiterjesztései, és alkalmasak bináris, kategorikus, Poisson- vagy más nem normális eloszlású adatok elemzésére is.

⊕ **Generalized Estimating Equations (Általánosított becslő egyenletek)**
Egy olyan becslési módszer, amely képes figyelembe venni az egyedi alanyokhoz tartozó több válasz közötti korrelációt, tehát akkor használjuk, amikor ismételt mérésekkel vagy fürtözött adatokkal dolgozunk, ahol a megfigyelések nem függetlenek egymástól.

Mixed Models (Vegyes modellek)

Fix és véletlen hatások együttes modellezésére szolgál hierarchikus vagy többszintű (pl. egyéni és iskolai) adatstruktúrák elemzésében. Különösen hasznosak klaszterezett adatok (a megfigyelések klaszterekbe vannak csoportosítva) elemzéséhez, ahol a klaszterek nem függetlenek.

Correlate (Korreláció)

⊕ **Bivariate (Kétváltozós)**
Két változó közötti lineáris kapcsolat erősségét és irányát méri. Lehetőség van a Pearson-féle korrelációs együttható és a Kendall és Spearman-féle rangkorrelációs együtthatók (két változó rangsorai közötti kapcsolat) kiszámítására. A korrelációs együtthatók nagyságára vonatkozó statisztikai próba is elvégezhető.

⊕ **Partial (Parciális)**
Azt méri, hogy két változó között milyen kapcsolat van, miután egy vagy több harmadik változó hatását kontrolláljuk.

⊕ **Distances (Távolság)**
Két véletlenszerű vektor közötti függőséget méri a mintaelemek közötti euklideszi távolságok segítségével. Lineáris és nem lineáris összefüggések kimutatására egyaránt használható.

Regression (Regresszió)

⊕ **Automatic Linear Modelling (Automatikus lineáris modellezés)**
Automatizálja a lineáris regressziós modellek építésének folyamatát. Egyszerűsíti a változók kiválasztását és az adatok előkészítését (pl. kiugró értékek eltávolítása, a kategóriák egyesítése), hogy javítsa a modell illeszkedését,

így különösen hasznos nagy és összetett adatbázisok esetén. Mindezek ellenére nagyon fontos kritikus szemmel értékelni az eredményül kapott modelleket, mivel az automatikus folyamatok néha figyelmen kívül hagyhatnak fontos szempontokat.

⊕ **Linear (Lineáris)**
Egy- és többváltozós lineáris regressziót hajt végre. A célváltozót vagy függő változót egy vagy több független változó lineáris függvényeként írja le. Az együtthatókat a legkisebb négyzetek elvével határozza meg, amelyek a független változó és a függő változó parciális korrelációs együtthatóival arányosak. Az összefüggésben részt vevő változók kiválasztására különböző modellépítési stratégiák (pl. Enter, Stepwise, Forward, Backward) vehetők igénybe.

⊕ **Curve Estimation (Görbebecslés)**
Olyan görbe keresésének folyamata, amely a legjobban kifejezi a függő változó és egy vagy több független változó közötti kapcsolatot.

⊕ **Partial Least Squares (Legkisebb négyzetek módszere)**
Több független változó és több függő változó közötti kapcsolatok modellezésére használják, különösen akkor, ha a független változók (predictors) erősen korrelálnak egymással, vagy ha számuk meghaladja a megfigyelések számát. A főkomponens-elemzés és a többváltozós regresszió egyes elemeit ötvözi, célja a független és a függő változók közötti tér közötti maximális kovarianciát leíró látens változók feltárása.

⊕ **Binary Logistic Regression (Bináris logisztikus regresszió)**
Kétértékű (igen/nem) függő változó és független változók kapcsolatának vizsgálata valószínűségek becslésére.

⊕ **Multinomial Logistic Regression (Multinomiális logisztikus regresszió)**
Többkategóriás, nem sorrendfüggő függő változók elemzése logisztikus regressziós módszerrel.

⊕ **Ordinal Regression (Ordinális regresszió)**
Sorrendi skálán mért függő változók kapcsolatának vizsgálata prediktorváltozókkal.

⊕ **Probit Regression (Probit regresszió)**
Bináris kimenetek modellezése, ahol a kapcsolat feltételezése normális eloszlású függvényen alapul.

⊕ **Nonlinear Regression (Nemlineáris regresszió)**
Olyan modellillesztési eljárás, amely a függő és független változók közötti kapcsolatot nem egyenes (nem lineáris) matematikai függvénnyel írja le, például exponenciális vagy logaritmikus formában.

⊕ **Weight Estimation (Súlybecslés)**
A regressziós modell paramétereinek becslésére szolgáló eljárás, amely figyelembe veszi az egyes megfigyelésekhez rendelt súlyokat. Ez különösen akkor hasznos, ha a minta nem egyenletesen reprezentálja a populációt, és a súlyozással pontosabb, torzításmentesebb eredmények érhetők el.

⊕ **2-Stage Least Squares (Kétlépcsős legkisebb négyzetek módszere)**
Elsősorban akkor alkalmazzák, ha a magyarázó változók között endogenitás áll fenn. Az első lépésben az endogén változót instrumentális változók segítségével becslik, a második lépésben pedig ezt a becsült értéket használják a fő regresszióban, így kiküszöbölve a torzítást.

⊕ **Optimal Scaling (Optimális skálázás)**
Olyan statisztikai technika, amely a nominális vagy ordinális változókat numerikus skálára alakítja át úgy, hogy a transzformáció maximalizálja a változók közötti kapcsolat erősségét a modellben. Gyakran használják regressziós és faktoranalízishez, hogy javítsa a magyarázóerőt és a modell illeszkedését.

Loglinear (Loglineáris)

⊕ **General (Általános)**
Maximum likelihood módszerrel próbát végez el és megbecsüli az általános loglineáris modell paramétereit, ahol a független változók között nominális mérési szintűek is lehetnek.

⊕ **Logit (Logit)**
Általános eljárás a kategóriás adatok közötti kapcsolat elemzésére loglineáris módszerrel, beleértve a modell specifikálását és illesztését. A függő nominális változó és több független kategóriaváltozó közötti kapcsolat feltárására szolgáló modell.

⊕ **Model Selection (Modellválasztás)**
Többváltozós loglineáris modellek közötti kiválasztás és összehasonlítás a legjobb illeszkedés megtalálásához.

Neural Networks (Mesterséges neurális hálózatok)

⊕ **Multilayer Perceptron (Többrétegű perceptron)**
Többrétegű, előrecsatolt neurális hálózat, amely rejtett rétegekkel modellezi a bemeneti és kimeneti változók közötti nemlineáris kapcsolatokat.

⊕ **Radial Basis Function (Radiális bázisfüggvény)**
Olyan előrecsatolt neurális hálózat, amely radiális bázisfüggvényeket használ az aktivációhoz, különösen jól alkalmazható gyors és pontos mintázatfelismerésre.

Classify (Osztályozás)

⊕ **TwoStep Cluster (Kétlépcsős klaszterezés)**
Ez az eljárás lehetővé teszi nagy méretű adathalmazok automatikus klaszterezését. Az első lépésben az adatok előfeldolgozásával kisebb előklasztereket hoz létre, a második lépésben ezekből finomítja a végső klasztereket a legjobb illeszkedés érdekében. Különösen előnyös vegyes típusú változók (folytonos és kategorikus) esetén, mivel mindkettőt figyelembe veszi a csoportok kialakításánál.

⊕ **K-Means Cluster (K-közép klaszterezés)**
Megfigyelések csoportosítása K darab klaszterbe úgy, hogy a klaszteren belüli eltérések minimálisak legyenek. Nagy adatfájlokon alkalmazható, a klaszterstruktúrához nem hierarchikus úton jutunk, azaz előre megadott K számú klaszterbe csoportosítjuk az eseteket a klaszterközéppontok alapján.

⊕ **Hierarchical Cluster (Hierarchikus klaszterezés)**
Objektumok fokozatos egyesítése vagy szétválasztása hierarchikus klaszterstruktúra kialakításához. Azon az elgondoláson alapul, hogy első lépésben valamennyi klaszterezésre váró esetet külön-külön egyszemélyes klaszterekben képzelünk el, majd az egymáshoz legközelebb álló eseteket ugyanahhoz a klaszterhez soroljuk (hierarchikusan építjük ki az osztályokat).

⊕ **Tree (Döntési fa)**
A döntési fa módszer a független változók alapján iteratív módon bontja az adatokat csoportokra, hogy meghatározza a függő változó értékeinek legjobb előrejelzését. A faágak logikai feltételeken alapulnak, és vizuálisan is ábrázolhatók, így jól szemléltetik a döntési szabályokat és a változók fontosságát.

⊕ **Discriminant (Diszkriminanciaanalízis)**
Több csoport előre meghatározott besorolása független torváltozók alapján, lineáris kombinációkkal. Segítségével meghatározható, mely változók járulnak leginkább hozzá a csoportok megkülönböztetéséhez, és előre jelezhető az új esetek csoporttagsága.

⊕ **Nearest Neighbor (Legközelebbi szomszéd)**
Objektumok besorolása a legközelebb eső ismert kategóriájú esetek alapján. A döntést a minta pontjai közötti távolságok alapján hozza, így egyszerű, de hatékony algoritmus a kategóriák előrejelzésére.

Dimension Reduction (Dimenziócsökkentés)

⊕ **Factor (Faktoranalízis)**
Több változó mögötti rejtett tényezők, látens dimenziók azonosítása, amelyek magyarázzák a változók közötti korrelációt, és közvetlenül egyetlen változóval sem mérhetőek. Segítségével egyszerűsítjük az adatstruktúrát, csökkentve a változók számát.

⊕ **Correspondence Analysis (Korrespondenciaanalízis)**
Kategóriás változók közötti kapcsolatok vizualizálására szolgáló statisztikai módszer. Segítségével a táblázatos adatok két- vagy többdimenziós térben ábrázolhatók, így könnyen felismerhetők a mintázatok, asszociációk és csoportosulások a változók között.

⊕ **Optimal Scaling (Optimális skálázás)**
Olyan statisztikai technika, amely a nominális vagy ordinális változókat numerikus skálára alakítja át úgy, hogy a transzformáció maximalizálja a változók közötti kapcsolat erősségét a modellben. Gyakran használják regressziós és faktoranalízishez, hogy javítsa a magyarázóerőt és a modell illeszkedését.

Scale (Skála)

⊕ **Reliability Analysis (Megbízhatósági elemzés)**
A skálák belső konzisztenciájának és megbízhatóságának vizsgálata, például Cronbach-alfa számításával. Segítségével meghatározható, hogy a tételek mennyire mérik következetesen ugyanazt a konstrukciót. Ezzel azonosíthatóak a gyenge vagy felesleges tételek, és javítható a mérőeszköz megbízhatósága.

⊕ **Multidimensional Unfolding (Többdimenziós kibontás)**
Olyan statisztikai módszer, amely a preferenciák vagy választások elemzésére szolgál, és mind a tárgyakat, mind a válaszadókat ugyanabban a többdimenziós térben helyezi el. A módszer célja, hogy a hasonló preferenciákkal rendelkező válaszadók közelebb, a különbözők távolabb kerüljenek egymástól, így vizuálisan ábrázolható a mintázat és a kapcsolatok szerkezete.

⊕ **Multidimensional Scaling PROXSCAL**
(Többdimenziós skálázás, Proximal Scaling)
Olyan MDS-eljárás, amely a disszimilaritási vagy távolságadatokból többdimenziós térképet hoz létre, minimalizálva az eltéréseket az adatok és a pontok közötti távolságok között.

⊕ **Multidimensional Scaling ALSCAL**
(Többdimenziós skálázás, Alternating Least Squares Scaling)
Iteratív módszer, amely optimalizálja a pontok elhelyezkedését a többdimenziós térben, hogy a megfigyelt hasonlóságok vagy távolságok a lehető legjobban tükröződjenek.

Nonparametric Tests (Nemparaméteres próbák)

⊕ **One-Sample (Egymintás)**
Olyan teszt, amely egy minta eloszlását hasonlítja össze egy elméleti értékkel vagy mediánnal, amikor az adatok nem követik a normális eloszlást. Segítségével megállapítható, hogy a minta szignifikánsan eltér-e a feltételezett középértéktől, például a Wilcoxon Signed-Rank Test használatával.

⊕ **Independent Samples (Független minták)**
Két vagy több független csoport adatainak összehasonlítására szolgál, amikor a feltételezett normális eloszlás nem teljesül. Gyakori példái a Mann–Whitney U- és a Kruskal–Wallis-tesztek, amelyek a csoportok mediánjainak vagy rangsorainak szignifikáns különbségét vizsgálják.

⊕ **Related Samples (Kapcsolt minták)**
Páros vagy ismételt mérésekkel rendelkező minták összehasonlítására szolgál, amikor az adatok nem normális eloszlásúak. Tipikus példái a Wilcoxon Signed-Rank Test és a Friedman Test, amelyek a mérések közötti rangkülönbségek szignifikanciáját vizsgálják.

⊕ **Legacy Dialogs (Régi párbeszédablakok)**
Lehetőséget ad a nemparaméteres tesztek klasszikus, korábbi verziókból ismert felületén történő végrehajtására. Itt elérhetők például a One-Sample, Independent Samples és Related Samples nemparaméteres próbák, egyszerű párbeszédablakokban, amelyek megkönnyítik a tesztek gyors beállítását és futtatását.

Forecasting (Előrejelzés)

A jövőbeli értékek előrejelzésére szolgál meglévő időbeli adatok alapján. Lehetővé teszi különböző előrejelzési modellek, például exponenciális simítás vagy ARIMA alkalmazását a trendek és szezonális mintázatok azonosítására.

Survival (Túlélés)

⊕ **Kaplan-Meier (Kaplan-Meier)**
Lehetővé teszi a túlélési görbék becslését és összehasonlítását különböző csoportok között.

⊕ **Cox Regression (Cox-regresszió)**
Segítségével a túlélési időt befolyásoló kovariánsok hatása vizsgálható többváltozós modellben.

⊕ **Life Tables (Élettáblák)**
A túlélési valószínűségek és a halálozási arányok időbeli alakulását mutatja.

⊕ **Cox w/ Time-Dependent Covariates (Cox-regresszió időfüggő kovariánsokkal)**
Olyan túlélési modell, amely lehetővé teszi, hogy a független változók (kovariánsok) értékei az idő során változzanak, így pontosabban tükrözik a kockázati tényezők hatását.

Multiple Response (Többszörös válaszok)

⊕ **Define Multiple Response Sets (Többszörös válaszhalmaz definiálása)**
Lehetővé teszi, hogy a felhasználó kijelölje, mely változók tartoznak egy többszörös válaszhalmazba. Így a későbbi elemzések során az SPSS ezeket egy egységként kezeli.

⊕ **Frequencies (Gyakoriságok)**
A többszörös válaszhalmaz gyakorisági eloszlását és százalékos arányát számítja ki. Segít áttekinteni, mely válaszok a leggyakoribbak a mintában.

⊕ **Crosstabs (Kereszttáblák)**
Lehetőséget ad a többszörös válaszhalmazok és más változók közötti összefüggések vizsgálatára. Segítségével csoportok szerinti bontásban is látható a válaszok eloszlása.

Missing Value Analysis (Hiányzó adatok elemzése)

A hiányzó adatok kezelésére és vizsgálatára szolgáló eszközöket tartalmazza. Azonosíthatóak a hiányzó értékek mintázatai, elemezhető azok hatása az adatokra, és különböző imputációs módszereket lehet alkalmazni a hiányzó adatok pótlására.

Multiple Imputation (Többszörös imputáció)

A hiányzó adatok kezelésére szolgál, lehetővé téve azok több lehetséges pótlással történő helyettesítését. A módszer csökkenti a hiányzó adatok miatti torzítást, így megbízhatóbb statisztikai következtetésekhez vezet.

Complex Samples (Komplex minták)

A bonyolult mintavételezési tervek, például rétegzett vagy klaszteres minták elemzését teszi lehetővé. Lehetővé teszi a mintavételi tervek definiálását is, a minták súlyozását és a torzítások figyelembevételével történő statisztikai elemzést.

Simulation (Szimuláció)

Statisztikai modellek és hipotézisek szimulációjára szolgál véletlenszerű adatok generálásával. Különböző forgatókönyveket lehet megvizsgálni, a modellek érzékenységét tesztelni, és a várható eredményeket előre lehet jelezni.

Quality Control (Minőség-ellenőrzés)

A termelési és folyamatadatok minőségének nyomon követésére szolgál. Ellenőrizhető a folyamat stabilitása, azonosíthatóak a hibák, és figyelemmel lehet kísérni a szabályozó diagramokat a minőség javítása érdekében.

ROC Curve (Receiver Operating Characteristic, ROC-görbe)

A bináris osztályozási modell teljesítményének grafikus ábrázolása az összes osztályozási küszöbértékre vonatkozóan (pl. logisztikus regresszió eredményei).

A DIRECT MARKETING menü

Direct Marketing (Közvetlen marketing)

A marketingkampányok tervezését, elemzését és értékelését támogatja statisztikai módszerekkel. Lehetővé teszi a célcsoportok szegmentálását, a válaszadási mintázatok feltárását és prediktív modellek alkalmazását a kampányhatékonyság növelésére. Különböző statisztikai eszközöket kínál a prediktív modellezéshez, mint pl. regressziós elemzéseket és döntési fákat, amelyek segítségével meghatározható a célcsoport legvalószínűbb reakciója a marketingakciókra. Támogatja a kampánykimenetek összehasonlítását, a költséghatékonyság vizsgálatát és a ROI (Return on Investment) elemzését. Előrejelezhetők a vásárlói magatartásminták, azonosíthatók a legértékesebb ügyfélcsoportok, optimalizálhatók a marketingstratégiák és a kampánykimenetek.

A GRAPHS menü

Chart Builder (Diagramépítő)

Interaktív eszköz különböző típusú grafikonok készítéséhez, például oszlop-, vonal-, pont- vagy területi diagramokhoz. Az ábrákon szerepeltetni kívánt minden változóhoz előre megfelelően be kell állítani a mérési szintet és kategoriális változók esetén az értékcímkéket.

Graphboard Template Chooser (Graphboard sablonválasztó)

Előre elkészített grafikonsablonok alkalmazása a gyors és egységes vizualizációkhoz.

Legacy Dialogs (Régi párbeszédablakok)

⊕ **Bar (Oszlopdiagram)**
Kategóriák összehasonlítása oszlopok segítségével.

⊕ **3D Bar (3D oszlopdiagram)**
Kategóriák összehasonlítása oszlopok segítségével háromdimenziós nézetben, amely térhatású vizualizációt biztosít az adatoknak.

⊕ **Line (Vonaldiagram)**
Idősoros vagy folyamatos adatok trendjeinek ábrázolása vonalakkal.

⊕ **Area (Területdiagram)**
Vonaldiagramhoz hasonló, de a vonal alatti területet kiemeli a vizuális hatás növelésére.

⊕ **Pie (Kördiagram)**
Kategóriák arányának ábrázolása körszeletek formájában.

⊕ **High-Low (Magas-alacsony)**
Pénzügyi adatok, például nyitó, záró, maximum és minimum értékek vizualizálása vonaldiagramon, a trendek és ingadozások szemléltetésére.

⊕ **Boxplot (Dobozdiagram)**
Az adatok eloszlásának, mediánjának és szórásának vizualizálása.
A változók eseteinek elhelyezkedését szemlélteti oly módon, hogy az esetek túlnyomó többsége a doboz által kijelölt intervallumba esik.

⊕ **Error Bar (Hibasávos diagram)**
Átlagok és hibahatárok (pl. szórás, standard hiba) ábrázolása.

⊕ **Population Pyramid (Korfa)**
Egy adott populáció kor- és nemi megoszlásának ábrázolása.

⊕ **Scatter/Dot (Pontdiagram)**
Két vagy három változó közötti kapcsolat ábrázolása pontok segítségével.

⊕ **Histogram (Hisztogram)**
Az adatok eloszlásának vizualizálása oszlopdiagram formájában.

Az UTILITIES menü

⊕ **Variables... (Változóinformációk)**
Megjeleníti a kiválasztott változó tulajdonságait, például változónév, címke, formátum, hiányzó értékek, értékcímkék, mérési szint és elhelyezkedés az adatnézetben. Lehetőséget ad a változók szintaktikai ablakba való beillesztésére is.

⊕ **OMS Control Panel (OMS-vezérlőpanel)**
Az Output Management System (OMS) működését szabályozza, amellyel az SPSS-kimenetek kezelhetők és szűrhetők. Lehetővé teszi az eredmények külön fájlba irányítását vagy automatikus feldolgozását.

⊕ **OMS Identifier (OMS-azonosító)**
Az OMS-kimenetek egyedi azonosítóval való ellátását biztosítja.
Segíti a különböző outputok megkülönböztetését és rendszerezését.

⊕ **Scoring Wizard (Pontozási varázsló)**
Egy lépésenkénti varázsló, amely előre betanított prediktív modellek alkalmazását teszi lehetővé új adatokra. A modellek eredményeinek pontszámokká alakításában hasznos.

⊕ **Merge Model XML (Modell-XML egyesítése)**
Lehetőséget ad különálló prediktív modell XML-fájlok kombinálására. Ez elősegíti a modellek integrálását és újrahasznosítását egyetlen struktúrában.

⊕ **Data File Comments (Adatfájl-megjegyzések)**
Az adatfájlhoz fűzhető szöveges megjegyzéseket kezeli. Ezek dokumentációs célokat szolgálnak, például adatforrás vagy előfeldolgozás leírását.

⊕ **Define Variable Sets... (Változókészletek definiálása)**
Lehetővé teszi változók csoportosítását és elnevezését, ami megkönnyíti a navigációt és a munkát összetettebb adatállományokban.

⊕ **Use Variable Sets... (Változókészletek használata)**
Kiválasztható, hogy a definiált változókészletek mely tagjai jelenjenek meg az adatnézetben és a párbeszédablakokban.

⊕ **Show All Variables (Minden változó megjelenítése)**
Egy kattintással megjeleníti az összes változót, függetlenül a korábban beállított változókészletektől vagy szűrésektől.

⊕ **Spelling (Helyesírás-ellenőrzés)**
Lehetővé teszi a szöveges változók és címkék helyesírásának ellenőrzését az adatfájlban. Hibás szavak esetén javítási javaslatokat kínál, hasonlóan a szövegszerkesztő programokhoz.

⊕ **Run Script (Parancsfájl futtatása)**
Lehetővé teszi külső scriptek futtatását, például Python vagy Java kód integrálását az SPSS-be. Ez kiterjeszti az SPSS funkcionalitását automatizálással.

⊕ **Production Facility (Gyártóüzem funkció)**
Automatizált elemzési folyamatok létrehozását és futtatását biztosítja. Olyan feladatokra alkalmas, ahol ismétlődő elemzéseket kell rendszeresen végrehajtani.

⊕ **Map Conversion Utility (Térképkonverziós segédprogram)**
Régebbi SPSS-verziókból származó térképfájlok új formátumba alakítását végzi. Ez biztosítja a kompatibilitást a frissített grafikus modulokkal.

⊕ **Custom Dialog (Egyéni párbeszédablak)**
Lehetőséget nyújt saját menük és párbeszédablakok létrehozására az SPSS-en belül. Ezzel felhasználóbarát módon integrálhatók egyedi elemző parancsok.

⊕ **Extension Bundles (Bővítménycsomagok)**
Olyan kiegészítések telepítését és kezelését teszi lehetővé, amelyek új funkciókat és parancsokat adnak az SPSS-hez. A bővítmények többek között Python, R vagy Java alapú eszközöket tartalmazhatnak.

Az ADD-ONS menü

⊕ **Applications (Alkalmazások)**
További SPSS-hez kapcsolódó alkalmazások elérését és telepítését biztosítja. Ezek az eszközök kiegészítő funkciókkal bővítik az alapprogram lehetőségeit.

⊕ **Services (Szolgáltatások)**
Az SPSS-hez kapcsolódó online és szerveralapú szolgáltatások kezelésére szolgál. Idetartozhat például a licencelés, frissítések vagy kiegészítő erőforrások elérése.

⊕ **Programmability Extension (Programozhatósági kiterjesztés)**
Lehetővé teszi külső programozási nyelvek, például Python vagy R integrálását az SPSS-környezetbe. Ez bővíti az elemzési lehetőségeket és támogatja az automatizált adatfeldolgozást.

A WINDOW menü

⊕ **Split (Ablak felosztása)**
Az aktív adatfájlt vízszintes vagy függőleges részre osztja, így egyszerre több részlet is látható. Ez különösen hasznos nagy adathalmazok áttekintésénél.

⊕ **Minimize All Windows (Minden ablak minimalizálása)**
Az aktív ablakot ikonméretűre csökkenti. Az ablak a tálcáról vagy az SPSS-ablaklistából újra megnyitható.

⊕ **Reset Dialog Sizes and Position**
(Párbeszédpanelek méretének és helyzetének visszaállítása)
Visszaállítja az összes párbeszédpanel alapértelmezett méretét és elhelyezkedését. Akkor célszerű használni, ha az ablakok elrendezése nehezen áttekinthetővé vált.

⊕ **Ablaklista (Megnyitott ablakok listája)**
A menü alján megjelenik az összes megnyitott adat- és kimeneti ablak felsorolása. Kiválasztva az adott ablak azonnal aktívvá válik.

A HELP menü

Help (Súgó)
Átfogó segítséget nyújt a felhasználóknak a statisztikai szoftver használatához. A menü tartalmazza a Topics-ot (Súgótémák), amelyek részletes dokumentációt kínálnak az eljárásokról és adatkezelési funkciókról.
A Tutorial (Bemutató) opció lépésről lépésre vezeti végig a felhasználókat a gyakori statisztikai elemzéseken. A menü keresőfunkciója lehetővé teszi a konkrét témák vagy eljárások gyors megtalálását.

Az Erdélyben lekérdezett EVS-kérdőív

Forrás:
https://search.gesis.org/research_data/ZA7550?doi=10.4232/1.13562

European *Values* Study

EUROPEAN VALUES STUDY 2017

Questionnaire
Romania – Hungarian minority

(Hungarian)

European Values Study
Erdély, 2019/2020

Kérdőív
A válaszadás önkéntes és névtelen!

Kérdezőbiztos neve: .....

A kérdezőbiztos azonosító száma: **.....**

Kérdezés dátuma

Év	Hónap	Nap

Kérdezés időtartama (óra, perc):

Kezdés óra	Kezdés perc		Befejezés óra	Befejezés perc

Megye: .....

Település: ..... Utca: .....

Jó reggelt/napot/estét kívánok! A nevem....., a TT Research and Communications kérdezőbiztosa vagyok. Mi egy olyan Európa-szerte végzett felmérés erdélyi részét készítjük, amely azt vizsgálja, hogy mit értékelnek az emberek az életben. Ebben a felmérésben olyan emberek kiválasztott csoportjával beszélgetünk, akik az európai embereket képviselik. Az Ön nevét véletlenszerűen választottuk ki az erdélyi magyar lakosság reprezentatív mintája részeként. Meg szeretném kérdezni az Ön véleményét néhány témával kapcsolatban. Az Ön segítsége hozzájárul ahhoz, hogy jobban megértsük, miben hisznek és mit várnak az élettől az emberek egész Európában.
A kérdőív kitöltése 25 percet igényel, a válaszok névtelenek, statisztikai módszerekkel lesznek feldolgozva.

K1. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 1 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Kérem, mondja meg, hogy a következők mennyire fontosak az Ön életében?

		Nagyon fontos	Elég fontos	Nem fontos	Egyáltalán nem fontos	*NT*	*NV*
v1	Munka	1	2	3	4	*8*	*9*
v2	Család	1	2	3	4	*8*	*9*
v3	Barátok és ismerősök	1	2	3	4	*8*	*9*
v4	Szabadidő	1	2	3	4	*8*	*9*
v5	Politika	1	2	3	4	*8*	*9*
v6	Vallás	1	2	3	4	*8*	*9*

K2. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 2

Mindent összevetve, mit mondana magáról, Ön

	Nagyon boldog	Meglehetősen boldog	Nem nagyon boldog	Egyáltalán nem boldog	*NT*	*NV*
v7	1	2	3	4	*8*	*9*

K3. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 3

Mindent összevetve, hogyan tudná Ön a mostani egészségi állapotát jellemezni? Azt mondaná, hogy:

	Nagyon jó	Jó	Elég jó, elfogadható	Nem nagyon rossz	Nagyon rossz	*NT*	*NV*
v8	1	2	3	4	5	*8*	*9*

K4. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 4 - NE OLVASD FEL A LISTÁT – KÓDOLD MINDET – BIZONYOSODJ MEG, HOGY A VÁLASZADÓ A TELJES LISTÁT ELOLVASSA

Kérem, nézze meg gondosan a következő önkéntes szervezetek listáját, és mondja meg, hogy Ön ezek közül tartozik-e valamelyikhez, s ha igen, melyikhez?

			Említette	Nem említette	*NT*	*NV*
v9	A	Vallási vagy egyházi szervezetek	1	2	*8*	*9*
v10	B	Oktatási, művészeti, zenei vagy kulturális tevékenység	1	2	*8*	*9*
v11	C	Szakszervezet	1	2	*8*	*9*
v12	D	Politikai párt vagy csoport	1	2	*8*	*9*
v13	E	Környezetvédelem, az élővilág megőrzése, állatok jogai	1	2	*8*	*9*
v14	F	Szakmai szervezet	1	2	*8*	*9*
v15	G	Sport vagy aktív pihenés	1	2	*8*	*9*
v16	H	Humanitárius vagy jótékonysági szervezet	1	2	*8*	*9*
v17	I	Fogyasztói érdekvédelmi szervezet	1	2	*8*	*9*
v18	J	Önsegítő csoport, kölcsönös segítségnyújtásra irányuló csoport	1	2	*8*	*9*
v19	K	Más csoportok	1	2	*8*	*9*
v20	L	Egyik sem (spontán említés)	1	2	*8*	*9*

K5. Végzett-e önkéntes munkát az elmúlt hat hónapban?

	Igen	Nem	*NT (spontán)*	*NV (spontán)*
v21	1	2	*8*	*9*

K6. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 6 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Ezen a listán különböző típusú emberek vannak. Ki tudná választani azokat, akiket Ön nem szeretne szomszédjainak, ha vannak ilyenek?

			Említette	Nem említette	*NT*	*NV*
v22	A	Más fajhoz tartozó emberek	1	2	*8*	*9*
v23	B	Erősen iszákosak	1	2	*8*	*9*
v24	C	Bevándorlók, külföldi vendégmunkások	1	2	*8*	*9*
v25	D	Kábítószeresek	1	2	*8*	*9*
v26	E	Homoszexuálisok	1	2	*8*	*9*
v27	F	Keresztények	1	2	*8*	*9*
v28	G	Muzulmánok	1	2	*8*	*9*
v29	H	Zsidók	1	2	*8*	*9*
v30	I	Cigányok	1	2	*8*	*9*

K7. Általában véve hogyan vélekedik inkább: a legtöbb emberben meg lehet bízni, vagy az ember nem lehet elég óvatos másokkal szemben?

	A legtöbb emberben meg lehet bízni	Az ember nem lehet elég óvatos	*NT (spontán)*	*NV (spontán)*
v31	1	2	*8*	*9*

K8. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 8 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Most arról szeretném kérdezni, hogy mennyire bízik meg különböző csoportokhoz tartozó emberekben. Kérem, mindegyikre mondja meg, hogy teljesen, valamennyire, nem nagyon, vagy egyáltalán nem bízik-e meg az ezekhez a csoportokhoz tartozó emberekben.

		Teljesen megbízik	Valamennyire megbízik	Nem nagyon bízik meg	Egyáltalán nem bízik meg	*NT*	*NV*
v32	Családjában	1	2	3	4	*8*	*9*
v33	A szomszédságában élő emberekben	1	2	3	4	*8*	*9*
v34	Azokban az emberekben, akiket személyesen ismer	1	2	3	4	*8*	*9*
v35	Azokban az emberekben, akiket először lát	1	2	3	4	*8*	*9*
v36	Más vallású emberekben	1	2	3	4	*8*	*9*
v37	Más nemzetiségű emberekben	1	2	3	4	*8*	*9*

K9. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 9

Egyes emberek úgy érzik, hogy teljesen szabadon határozzák meg életüket, mások viszont úgy érzik, hogy semmi befolyásuk sincs sorsuk alakítására. Kérem, a skála segítségével mondja meg, hogy Ön szerint mekkora befolyása van arra, hogy miként alakul az élete!

	Semmi									Nagyon nagy	*NT*	*NV*
v38	1	2	3	4	5	6	7	8	9	10	*88*	*99*

K10. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 10

Mindent egybevetve, összességében mennyire elégedett jelenlegi életével? Kérem, használja ezt a kártyát a válaszadásban!

	elégedetlen									elégedett	*NT*	*NV*
v39	1	2	3	4	5	6	7	8	9	10	88	99

K11. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 11 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Ön egyetért, vagy nem ért egyet a következő kijelentésekkel?

		Teljesen egyetért	Egyet-ért	Sem-leges	Nem ért egyet	Egyáltalán nem ért egyet	*NT*	*NV*
v46	Ahhoz, hogy adottságait teljesen kibontakoztassa, kell hogy legyen valamilyen munkája	1	2	3	4	5	*8*	*9*
v47	Megalázó dolog pénzt kapni, ha az ember nem dolgozott meg érte	1	2	3	4	5	*8*	*9*
v48	Akik nem dolgoznak, ellustulnak	1	2	3	4	5	*8*	*9*
v49	A munka a társadalommal szembeni kötelesség	1	2	3	4	5	*8*	*9*
v50	Az első helyen mindig a munka áll, még akkor is, ha emiatt kevesebb szabad idő marad	1	2	3	4	5	*8*	*9*

K12. Tartozik-e Ön valamilyen vallási felekezethez?

	Igen	Nem	*NT*	*NV*
v51	1	2	*8*	*9*
		MENJ K14		

K13. *Ha igen*, MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 13

Melyikhez?

v52	1. Római katolikus	4. Görög katolikus	7. Adventista	10. Egyéb	*88. NT*
	2. Református, kálvinista	5. Evangélikus-lutheránus	8. Ortodox	_____	*99. NV*
	3. Unitárius	6. Baptista	9. Jehova Tanúi	v52a	

K14. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 14/15 - OLVASD FEL, ÉS CSAK EGY VÁLASZT KÓDOLJ

Esküvőktől, temetésektől és keresztelőktől eltekintve, mostanában milyen gyakran szokott részt venni vallási szertartáson?

	Hetente többször	Hetente	Havonta	Csak bizonyos ünnepekkor	Évente	Ritkábban	Soha	*NT*	*NV*
v54	1	2	3	4	5	6	7	*8*	*9*

K15. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 14/15 - OLVASD FEL, ÉS CSAK EGY VÁLASZT KÓDOLJ

Amikor 12 éves volt, esküvőktől, temetésektől és keresztelőktől eltekintve, milyen gyakran szokott részt venni vallási szertartáson?

	Hetente többször	Hetente	Havonta	Csak bizonyos ünnepekkor	Évente	Ritkábban	Soha	*NT*	*NV*
v55	1	2	3	4	5	6	7	*8*	*9*

K16. KÉRDEZŐI INSTRUKCIÓ: - OLVASD FEL, ÉS CSAK EGY VÁLASZT KÓDOLJ

Függetlenül attól, hogy jár-e templomba vagy sem, mit mondana magáról, Ön …

	Vallásos ember	Nem vallásos ember	Meggyőződéses ateista	*NT*	*NV*
v56	1	2	3	*8*	*9*

K17. KÉRDEZŐI INSTRUKCIÓ: - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Az alábbiak közül miben hisz Ön és miben nem hisz?

		Igen	Nem	*NT*	*NV*
v57	Isten	1	2	*8*	*9*
v58	A halál utáni élet	1	2	*8*	*9*
v59	Pokol	1	2	*8*	*9*
v60	Mennyország	1	2	*8*	*9*

K18. Hisz-e Ön a reinkarnációban, azaz abban, hogy voltak előző életeink, és itt a Földön újra meg fogunk születni?

	Igen	Nem	*NT*	*NV*
v61	1	2	8	9

K19. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 19 - OLVASD FEL, ÉS CSAK EGY VÁLASZT KÓDOLJ

A következő állítások közül melyik áll a legközelebb az Ön hitéhez?

	Létezik egy Isten, mint személy	Van valamilyen szellemi lény vagy életerő	Valójában nem is tudom, hogy mit gondoljak	Nem hiszem, hogy létezne valamilyen Isten, szellemi lény vagy életerő	*NT*	*NV*
v62	1	2	3	4	8	9

K20. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 20

És mennyire fontos Isten az Ön életében? Kérem, határozza meg ennek a kártyalapnak a segítségével - a 10-es azt jelenti, hogy nagyon fontos, az 1-es pedig hogy egyáltalán nem fontos.

	Egyáltalán nem fontos									Nagyon fontos	*NT*	*NV*
v63	1	2	3	4	5	6	7	8	9	10	88	99

K21. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 21 - OLVASD FEL, ÉS CSAK EGY VÁLASZT KÓDOLJ

Milyen gyakran imádkozik istentiszteleten kívül? Nagyjából…

	Minden nap	Hetente többször	Hetente	Legalább havonta	Évente többször	Ritkábban	Soha	*NT*	*NV*
v64	1	2	3	4	5	6	7	*8*	*9*

A KÖVETKEZŐ KÉRDÉSEK A CSALÁDI ÉLETRE ÉS A HÁZASSÁGRA VONATKOZNAK

K22. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 22 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Ezen a listán olyan dolgokat sorolunk fel, amelyekről egyesek azt tartják, hogy szükségesek a sikeres házassághoz vagy élettársi kapcsolathoz. Kérem, mondja meg mindegyikről, hogy az Ön szerint nagyon fontos, eléggé fontos, vagy nem nagyon fontos a sikeres házassághoz.

			Nagyon fontos	Inkább fontos	Nem nagyon fontos	*NT*	*NV*
v65	A	Hűség	1	2	3	*8*	*9*
v66	B	Megfelelő jövedelem	1	2	3	*8*	*9*
v67	C	Megfelelő lakhatási körülmények	1	2	3	*8*	*9*
v68	D	A háztartási teendők megosztása	1	2	3	*8*	*9*
v69	E	Gyerekek	1	2	3	*8*	*9*
v70	F	Mindenkinek legyen valamennyi ideje a saját barátaira és a személyes hobbijaira/tevékenységeire	1	2	3	*8*	*9*

K23. Egyetért Ön, vagy nem ért egyet azzal az állítással, hogy: A házasság egy idejétmúlt intézmény?

	Igen	Nem	*NT (spontán)*	*NV (spontán)*
v71	1	2	*8*	*9*

K24. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 24 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Kérem jelezze, hogy mennyire ért egyet vagy mennyire nem ért egyet a következő állításokkal. Teljesen egyetért, egyetért, nem ért egyet, vagy egyáltalán nem ért egyet?

		Teljesen egyetért	Egyet-ért	Nem ért egyet	Egyáltalán nem ért egyet	*NT*	*NV*
v72	Ha egy anya fizetett munkát vállal, azt megszenvedik a gyermekek	1	2	3	4	*8*	*9*
v73	Lehet munkát vállalni, de amit egy nő igazán akar, az az otthon és a gyermekek	1	2	3	4	*8*	*9*
v74	Mindent egybevetve, a család élete megsínyli, ha a nő teljes munkaidőben dolgozik.	1	2	3	4	*8*	*9*
v75	A férfi dolga pénzt keresni, a nő dolga a háztartásról és a családról gondoskodni	1	2	3	4	*8*	*9*
v76	Mindent egybe vetve a férfiak jobb politikai vezetők, mint a nők	1	2	3	4	*8*	*9*
v77	Az egyetemi végzettség fontosabb a fiúknak, mint a lányoknak	1	2	3	4	*8*	*9*
v78	Mindent egybe vetve a férfiak jobb gazdasági vezetők, mint a nők	1	2	3	4	*8*	*9*
v79	Egyik legfontosabb célom az életben, hogy szüleim büszkék legyenek rám	1	2	3	4	*8*	*9*

K25. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 25 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Kérem jelezze, hogy mennyire ért egyet vagy mennyire nem ért egyet a következő állításokkal. Teljesen egyetért, egyetért, nem ért egyet, vagy egyáltalán nem ért egyet?

		Teljesen egyetért	Egyetért	Nem ért egyet, de nincs ellenvéleménye sem	Nem ért egyet	Egyáltalán nem ért egyet	*NT*	*NV*
v80	Ha kevés a munkahely, a munkáltatóknak előnyben kell részesíteni a bevándorlókkal szemben azokat, akik itt születtek	1	2	3	4	5	*8*	*9*
v81	Ha kevés a munkahely, a férfiaknak több joguk van a munkához, mint a nőknek	1	2	3		4	*8*	*9*

K26. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 26 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Mi az Ön véleménye a következő kijelentésekről? Egyetért velük, vagy nem ért velük egyet?

		Teljesen egyetért	Egyet-ért	Nem ért egyet, de nincs ellenvéleménye sem	Nem ért egyet	Egyáltalán nem ért egyet	*NT*	*NV*
v82	A homoszexuális párok éppolyan jó szülők, mint más párok	1	2	3	4	5	*8*	*9*
v83	A gyermekvállalás a társadalommal szembeni kötelesség	1	2	3	4	5	*8*	*9*
v84	A felnőtt gyermekeknek kötelességük hosszútávon gondoskodni a szüleikről	1	2	3	4	5	*8*	*9*

K27. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 27 - LEGFELJEBB ÖT EMLÍTÉST KÓDOLJ

Most felsorolok néhány tulajdonságot, amelyekre a szülők a gyermekeiket otthon nevelhetik. Ezek közül melyek azok, amelyeket Ön különösen fontosnak tart? Kérem, válasszon legfeljebb ötöt.

			Említette	Nem említette	*NT*	*NV*
v85	A	Jó modor	1	2	*8*	*9*
v86	B	Önállóság	1	2	*8*	*9*
v87	C	Szorgalom	1	2	*8*	*9*
v88	D	Felelősségérzet	1	2	*8*	*9*
v89	E	Képzelőerő, fantázia	1	2	*8*	*9*
v90	F	Mások tisztelete, tolerancia mások iránt	1	2	*8*	*9*
v91	G	Takarékosság	1	2	*8*	*9*
v92	H	Elszántság, állhatatosság.	1	2	*8*	*9*
v93	I	Vallásos hit	1	2	*8*	*9*
v94	J	Önzetlenség	1	2	*8*	*9*
v95	K	Engedelmesség	1	2	*8*	*9*
v96	L	Egyik sem (spontán módon)	1	2	*8*	*9*

MOST AKTUÁLIS TÁRSADALMI KÉRDÉSEKRŐL FOGUNK KÉRDEZNI

K28. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 28

Mennyire érdekli Önt a politika?

	Nagyon érdekli	Némileg érdekli	Nem nagyon érdekli	Egyáltalán nem érdekli	*NT*	*NV*
v97	1	2	3	4	*8*	*9*

K29. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 29

Politikai kérdésekben az emberek beszélnek "bal" illetve "jobb" oldalról. Ön általában véve hová helyezné el a saját nézeteit ezen a skálán?

	Bal									Jobb	*NT*	*NV*
v102	1	2	3	4	5	6	7	8	9	10	*88*	*99*

K30. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 30 - OLVASD FEL

Ezen a kártyán különböző témákkal kapcsolatos ellentétes véleményeket lát. Hová helyezné saját nézeteit ezen a skálán?

v103	A	Az egyéneknek nagyobb mértékben kellene felelősséget vállalniuk a saját magukról való gondoskodásban	1	2	3	4	5	6	7	8	9	10	Az államnak nagyobb mértékben kellene felelősséget vállalnia az emberekről való gondoskodásban	*NT* *88*	*NV* *99*
v104	B	A munkanélkülieknek minden lehetséges munkát el kellene vállalniuk, vagy veszítsék el a munkanélküli segélyt	1	2	3	4	5	6	7	8	9	10	A munkanélkülieknek meg kellene lenni a joguknak, hogy visszautasítsák a munkát, amelyet nem akarnak elvégezni	*NT* *88*	*NV* *99*
v105	C	A verseny jó dolog	1	2	3	4	5	6	7	8	9	10	A verseny káros dolog.	*NT* *88*	*NV* *99*
v106	D	A jövedelmeket egyenlőbbé kellene tenni	1	2	3	4	5	6	7	8	9	10	Az egyéni erőfeszítéseket nagyobb mértékben kellene jutalmazni	*NT* *88*	*NV* *99*
v107	E	Az üzleti szférában és az iparban növelni kellene a magántulajdon arányát	1	2	3	4	5	6	7	8	9	10	Az üzleti szférában és az iparban növelni kellene az állami tulajdon arányát	*NT* *88*	*NV* *99*

K31. Természetesen mindannyian azt reméljük, hogy nem lesz még egy háború, de ha mégis arra kerülne a sor, Ön hajlandó lenne harcolni az országáért?

	Igen	Nem	*NT (spontán)*	*NV (spontán)*
v112	1	2	*8*	*9*

K32. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 32 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Itt van két olyan változás, amelyek talán a közeljövőben bekövetkezhetnek életvitelünkben. Kérem, mondja meg mindegyikről, hogyha bekövetkezne, akkor az jó, vagy rossz, vagy mindegy lenne az Ön számára?

		Jó	Rossz	Mindegy	*NT*	*NV*
v113	A munka életünkben betöltött fontosságának csökkenése	1	2	3	*8*	*9*
v114	A tekintélytisztelet erősödése	1	2	3	*8*	*9*

K33. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 33 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Kérem, nézze meg ezt a kártyát, és a listán szereplő mindegyik elemről mondja meg, hogy mennyire van bizalma benne! Nagyon, meglehetősen, nem nagyon, vagy egyáltalán nem?

		Nagyon	Meglehetősen	Nem nagyon	Egyáltalán nem	*NT*	*NV*
v115	Az egyház	1	2	3	4	*8*	*9*
v116	A fegyveres erők	1	2	3	4	*8*	*9*
v117	Az oktatási rendszer	1	2	3	4	*8*	*9*
v118	A sajtó	1	2	3	4	*8*	*9*
v119	A szakszervezetek	1	2	3	4	*8*	*9*
v120	A rendőrség	1	2	3	4	*8*	*9*
v121	A parlament	1	2	3	4	*8*	*9*
v122	A közigazgatás	1	2	3	4	*8*	*9*
v123	A társadalombiztosítási rendszer	1	2	3	4	*8*	*9*
v124	Az Európai Unió	1	2	3	4	*8*	*9*
v126	Az egészségügyi rendszer	1	2	3	4	*8*	*9*
v127	Az igazságszolgáltatási rendszer	1	2	3	4	*8*	*9*
v128	Nagyvállalatok	1	2	3	4	*8*	*9*
v129	Környezetvédelmi szervezetek	1	2	3	4	*8*	*9*
v130	Politikai pártok	1	2	3	4	*8*	*9*
v131	A kormány	1	2	3	4	*8*	*9*
v132	Közösségi média	1	2	3	4	*8*	*9*

K34. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 34 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Számos dolog kívánatos, azonban közül nem mindegyik nélkülözhetetlen a demokrácia működése szempontjából. Kérem, a következő dolgokról mondja meg, hogy mennyire tartja azokat a demokrácia nélkülözhetetlen jellemzőjének. Kérem, használja ezt a skálát, ahol az 1-es azt jelenti, hogy az adott tényező egyáltalán nem fontos jellemzője a demokráciának, és 10-es pedig azt, hogy nélkülözhetetlen jellemzője a demokráciának

		Egyáltalán nem fontos jellemzője									**Nagyon fontos jellemzője**	**Ez ellentétes a demokráciával (*NE OLVASD FEL*)**	*NT*	*NV*
v133	A kormányzat megadóztatja a gazdagokat, és támogatja a szegényeket	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v134	A vallási képviselőknek döntő szavuk van a törvények értelmezésénél	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v135	Az emberek szabad választásokon választják meg vezetőiket	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v136	A munkanélküliek állami segélyben részesülnek	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v137	Ha a kormány alkalmatlan ellátni a feladatát, a hadsereg veszi át a hatalmat	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v138	A polgári jogok védelmezik az embereket az állam elnyomó hatalmával szemben.	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v139	Az állam az emberek jövedelmét egyenlővé teszi.	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v140	Az emberek engedelmeskednek vezetőiknek	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*
v141	A nőknek ugyanolyan jogaik vannak mint a férfiaknak	1	2	3	4	5	6	7	8	9	10	0	*88*	*99*

K35. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 35 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Mi a véleménye a következő állításokról? Az 1 az jelenti, hogy soha nem engedhető meg, a 10, hogy mindig megengedhető. Köztes értékeket is használhat, hogy árnyalja a véleményét.

		Soha									**Mindig**	***NT***	***NV***
v149	Állami juttatásokat jogtalanul igénybe venni	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v150	Csalni az adóval, ha van rá mód	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v151	Marihuánát vagy hasist szívni	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v152	Csúszópénzt elfogadni a kötelesség teljesítése során	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v153	Homoszexualitás	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v154	Abortusz	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v155	Elválni	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v156	Eutanázia (gyógyíthatatlan beteg életét kioltani)	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v157	Öngyilkosság	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v158	Alkalmi szexuális kapcsolatot létesíteni	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v159	Jegy és bérlet nélkül utazni egy tömegközlekedési eszközön	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v160	Prostitúció	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v161	Mesterséges vagy lombikban történő megtermékenyítés	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v162	Politikai indíttatású erőszak	1	2	3	4	5	6	7	8	9	10	*88*	*99*
v163	Halálbüntetés	1	2	3	4	5	6	7	8	9	10	*88*	*99*

K36. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 36 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Az emberek különbözőképpen vélekednek önmagukról és a világhoz való viszonyukról. Meg tudná nekem mondani, hogy mennyire érzi magát közel a következőhöz?

		Nagyon közel	Közel	Nem nagyon közel	Egyáltalán nem közel	*NT*	*NV*
v164	Az Ön falujához vagy városához	1	2	3	4	*8*	*9*
v165	A régiójához	1	2	3	4	*8*	*9*
v166	Romániához	1	2	3	4	*8*	*9*
v167	Európához	1	2	3	4	*8*	*9*
v168	A világhoz	1	2	3	4	*8*	*9*

K37. KÉRDEZŐI INSTRUKCIÓ: - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Választásokkor Ön mindig szavaz, általában szavaz, vagy soha nem szavaz? Kérem, hogy egyesével válaszoljon a következő szintekre vonatkozóan!

		Mindig	Általában	Soha	*NT*	*NV*
v171	Önkormányzati választások	1	2	3	*8*	*9*
v172	Országos választások	1	2	3	*8*	*9*
v173	Európai parlamenti választások	1	2	3	*8*	*9*

K38. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 38/39

Melyik (politikai) párt a legrokonszenvesebb Önnek?

v174		
1	Partidul Social Democrat (PSD)	
2	Partidul Național Liberal (PNL)	
3	Uniunea Salvați România (USR)	
4	Romániai Magyar Demokrata Szövetség (RMDSZ)	
5	Partidul Alianța Liberalilor și Democraților (ALDE)	
6	Partidul Mișcarea Populară (PMP)	
7	Pro Romania	
8	Partidului Libertății, Unității și Solidarității (PLUS)	
9	Erdélyi Magyar Néppárt (EMNP)	
10	Magyar Polgári Párt (MPP)	
26	Egyéb, Kérem, nevezze meg. (ÍRD BE):	v174a
88	*Nem tudja (spontán módon)*	
99	*Nincs válasz (spontán módon)*	

K39. Van még másik párt is, amely rokonszenves Önnek?

v175		
1	Partidul Social Democrat (PSD)	
2	Partidul Național Liberal (PNL)	
3	Uniunea Salvați România (USR)	
4	Romániai Magyar Demokrata Szövetség (RMDSZ)	
5	Partidul Alianța Liberalilor și Democraților (ALDE)	
6	Partidul Mișcarea Populară (PMP)	
7	Pro Romania	
8	Partidului Libertății, Unității și Solidarității (PLUS)	
9	Erdélyi Magyar Néppárt (EMNP)	
10	Magyar Polgári Párt (MPP)	
31	Egyéb, Kérem, nevezze meg. (ÍRD BE):	v175a
88	*Nem tudja (spontán módon)*	
99	*Nincs válasz (spontán módon)*	

K40. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 40 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Ön szerint Romániában a választások alkalmával milyen gyakran fordulnak elő a következő dolgok?

		Nagyon gyakran	Elég gyakran	Nem gyakran	Egyáltalán nem gyakran	*NT*	*NV*
v176	A szavazatokat tisztességes módon számolják össze.	1	2	3	4	*8*	*9*
v177	Megakadályozzák, hogy ellenzéki jelöltek indulhassanak a választáson	1	2	3	4	*8*	*9*
v178	A tévé hírműsora előnyben részesíti a kormányzó pártot	1	2	3	4	*8*	*9*
v179	Szavazókat lefizetnek	1	2	3	4	*8*	*9*
v180	Az újságírók pártatlanul tudósítanak a választásokról	1	2	3	4	*8*	*9*
v181	A választásokat biztosító hivatalnokok és a választási bizottságok tisztességesek	1	2	3	4	*8*	*9*
v182	A gazdag emberek megvásárolják a választást	1	2	3	4	*8*	*9*
v183	A szavazókat erőszakkal megfélemlítik a szavazásnál	1	2	3	4	*8*	*9*

K41. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 41

Most meg szeretnénk tudni a véleményét azokról az emberekről, akik más országból jöttek Romániába lakni – a bevándorlókról. Hogy értékelné ezeknek az embereknek Románia fejlődésére gyakorolt hatását?

	Nagyon rossz	Elég rossz	Sem jó, sem rossz	Elég jó	Nagyon jó	*NT*	*NV*
v184	1	2	3	4	5	*8*	*9*

K42. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 42 - OLVASD FEL

Kérem, nézze meg a következő állításokat, és jelölje meg, hova helyezné a véleményét ezen a skálán.

v185	A	A bevándorlók munkahelyeket vesznek el az itteniektől	1	2	3	4	5	6	7	8	9	10	A bevándorlók nem vesznek el munkahelyeket az itteniektől	*NT* *88*	*NV* *99*
v186	B	A bevándorlók rontják a bűnözési helyzetet	1	2	3	4	5	6	7	8	9	10	A bevándorlók nem rontják a bűnözési helyzetet	*NT* *88*	*NV* *99*
v187	C	A bevándorlók megterhelik egy ország jóléti rendszerét	1	2	3	4	5	6	7	8	9	10	A bevándorlók nem terhelik meg egy ország jóléti rendszerét	*NT* *88*	*NV* *99*
v188	D	Jobb, ha a bevándorlók megőrzik sajátos szokásaikat és hagyományaikat	1	2	3	4	5	6	7	8	9	10	Jobb, ha a bevándorlók nem őrzik meg sajátos szokásaikat és hagyományaikat	*NT* *88*	*NV* *99*

K43. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 43 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Az emberek különböző módon vélekednek arról, hogy mit jelent európainak lenni. Az Ön véleménye szerint mennyire fontosan a következők ahhoz, hogy valaki európai legyen?

		Nagyon fontos	Elég fontos	Nem fontos	Egyáltalán nem fontos	*NT*	*NV*
v194	Európában szülessen	1	2	3	4	*8*	*9*
v195	Legyenek európai felmenői	1	2	3	4	*8*	*9*
v196	Keresztény legyen	1	2	3	4	*8*	*9*
v197	Elsajátítsa az európai kultúrát	1	2	3	4	*8*	*9*

K44. MUTASD A KÖVETKEZŐ KÁRTYALAPOT:44

Néhányan azt mondják, hogy az Európai Unió kibővítésének tovább kell folytatódnia. Mások szerint az Unió kibővítése már így is túl messzire ment. A kártya segítségével mondja meg, melyik szám fejezi ki a legjobban az Ön véleményét. Az 1-es jelenti, hogy "tovább kellene folytatódnia", a 10-es, hogy "már így is túl messzire ment"

	Tovább kellene folytatódnia									Már így is túl messzire ment	*NT*	*NV*
v198	1	2	3	4	5	6	7	8	9	10	*88*	*99*

K45. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 45 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Kérem, mondja meg a következő kijelentésekről egyenként, hogy mennyire ért velük egyet!

		Teljesen egyetért	Egyetért	Semleges	Nem ért egyet	Egyáltalán nem ért egyet	*NT*	*NV*
v199	Odaadnám a jövedelmem egy részét, ha biztos lehetnék benne, hogy a pénzt a környezetszennyezés megelőzésére fordítják	1	2	3	4	5	*8*	*9*
v200	A hozzám hasonló embereknek túl nehéz bármi lényegeset tenni a környezetért	1	2	3	4	5	*8*	*9*
v201	Vannak fontosabb teendők az életben, mint a környezetet védeni	1	2	3	4	5	*8*	*9*
v202	Nincs értelme megtenni a környezetért azt, amit tudok, amíg mások nem teszik ugyanezt	1	2	3	4	5	*8*	*9*
v203	Számos, a környezetet fenyegető veszélyről szóló állítás túlzás	1	2	3	4	5	*8*	*9*

K46. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 46 - OLVASD FEL ÉS KÓDOLJ EGY VÁLASZT

Itt van két olyan kijelentést, amit az emberek tesznek néha, amikor a környezetről és a gazdasági növekedésről beszélgetnek. Melyik áll közelebb az Ön véleményéhez?

	A környezetvédelmet akkor is előnyben kell részesíteni, ha ez lelassítja a gazdasági növekedést és néhány munkahely megszűnését okozza.	A gazdasági növekedésnek és a munkahelyteremtésnek kell a legfontosabbnak lennie, még akkor is, ha ez egy bizonyos mértékben negatív hatással van a környezetre.	Más válasz (csak ha spontánul említi)	*NT*	*NV*
v204	1	2	3	*8*	*9*

K47. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 47 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Mit gondol, az államnak jogában kellene állnia vagy nem kellene jogában állnia, hogy megtegye a következőt:

		Határozottan álljon jogában	Valószínűleg álljon jogában	Valószínűleg ne álljon jogában	Határozottan ne álljon jogában	*NT*	*NV*
v205	Közterületen biztonsági kamerákkal megfigyelni az embereket	1	2	3	4	*8*	*9*
v206	Megfigyelni az összes e-mailt és interneten cserélt egyéb információt	1	2	3	4	*8*	*9*
v207	Bárkiről, aki itt él információt gyűjteni, akár a tudta nélkül	1	2	3	4	*8*	*9*

K48. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 48 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Milyen gyakran követi figyelemmel a politikát...

		Minden nap	Hetente többször	Heti egy vagy két alkalommal	Ritkábban	Soha	*NT*	*NV*
v208	a TV-ben	1	2	3	4	5	*8*	*9*
v209	rádióban	1	2	3	4	5	*8*	*9*
v210	napilapokban	1	2	3	4	5	*8*	*9*
v211	a közösségi médiában	1	2	3	4	5	*8*	*9*

K49. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 49 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Mennyire foglalkoztatják Önt a következő emberek és csoportok életkörülményei:

		Igen nagyon	Nagyon	Bizonyos fokig	Nem olyan nagyon	Egyáltalán nem	*NT*	*NV*
v212	A szomszédságában élők	1	2	3	4	5	*8*	*9*
v213	Az Ön régiójában élők	1	2	3	4	5	*8*	*9*
v214	Az Önnel egy országban élők	1	2	3	4	5	*8*	*9*
v214	Az európaiak	1	2	3	4	5	*8*	*9*
v215	A világon élő minden ember	1	2	3	4	5	*8*	*9*

K50. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 50 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

És menyire foglalkoztatják Ön a következő emberek és csoportok életkörülményei:

		Igen nagyon	Nagyon	Bizonyos fokig	Nem olyan nagyon	Egyáltalán nem	*NT*	*NV*
v216	Idős emberek	1	2	3	4	5	*8*	*9*
v217	Munkanélküliek	1	2	3	4	5	*8*	*9*
v218	Bevándorlók	1	2	3	4	5	*8*	*9*
v219	Betegek és fogyatékosok	1	2	3	4	5	*8*	*9*

K51. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 51 - OLVASD FEL ÉS SORONKÉNT EGY VÁLASZT KÓDOLJ

Mit kellene egy társadalomnak nyújtania? Kérem, a következő kijelentések mindegyikéről mondja meg, hogy Ön szerint az nagyon fontos vagy egyáltalán nem fontos.

		Nagyon fontos	Elég fontos	Nem fontos	Egyáltalán nem fontos	***NT***	***NV***
v221	Az állampolgárok jövedelmei közötti nagy különbségek megszüntetése	1	2	3	4	*8*	*9*
v222	Az elemi szükségleteknek, az élelmezésnek, a lakhatásnak, a ruházkodásnak, az oktatásnak, az egészségügyi ellátásnak minden állampolgár számára való biztosítása	1	2	3	4	*8*	*9*
v223	Az emberek érdemeik szerinti elismerése	1	2	3	4	*8*	*9*
v224	A terrorizmus elleni védelem	1	2	3	4	*8*	*9*

ORSZÁGSPECIFIKUS KÉRDÉSEK

K52. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 52 - OLVASD FEL

Az alábbi kategóriák szerint melyikkel tudná legjobban jellemezni önmagát?

cs25	1. Vallásos vagyok az egyház tanítása szerint	3. Nem tudom eldönteni, hogy vallásos vagyok-e	5. Nem vagyok vallásos, határozottan más a meggyőződésem
	2. Vallásos vagyok a magam módján	4. Nem vagyok vallásos	*8. NT*
			9. NV

K53. Kérem, válasszon egy számot egy skáláról, ahol az 1 azt fejezi ki, hogy egyáltalán nem, az 5 pedig, hogy teljesen igaz Önre a következő állítás! Egyházak és vallásos szertartások nélkül, a magam módján teremtek kapcsolatot a Természetfelettivel.

	Egyáltalán nem				**Teljesen**	*NT*	*NV*
cs27	1	2	3	4	5	*8*	*9*

K54. Hisz Ön abban, hogy egy szerencsehozó tárgy, mint pl. egy kabala vagy egy talizmán meg tudja védeni, vagy segíteni tudja Önt? Válaszoljon egy 1 és 10 közötti skála segítségével, ahol az 1 a határozottan nem, a 10 pedig a határozottan igen.

	Határozottan nem									**Határozottan igen**	*NT*	*NV*
cs28	1	2	3	4	5	6	7	8	9	10	*88*	*99*

K55. Részt szokott-e Ön venni valamilyen egyházon belüli csoport tevékenységein (imacsoport, bibliaóra, lelkiségi csoport, ifjúsági csoport, egylet, stb.)?

	Igen	**Jelenleg nem, de korábban részt vettem**	**Nem**	*NT*	*NV*
cs29	1	2	3	8	9

K56. Betölt-e Ön jelenleg valamilyen egyházi tisztséget (egyháztanácsos, presbiter, lelkész, stb.)?

	Igen	**Jelenleg nem, de korábban igen**	**Nem**	*NT*	*NV*
cs30	1	2	3	8	9

DEMOGRÁFIA

MINDENKITŐL KÉRDEZD

K57. Neme:

	1	2	*8*	*9*
v225	Férfi	Nő	*NT (spontán)*	*NV (spontán)*

K58. Melyik évben született Ön:

V226		*8888 NT (spontán)*	*9999 NV (spontán)*

K59. MINDENKITŐL KÉRDEZD

Mi az Ön jelenlegi hivatalos családi állapota?

CSAK A HÁZASSÁGBÓL VAGY REGISZTRÁLT PARTNERKAPCSOLATBÓL MEGÖZVEGYÜLTET ÉS ELVÁLTAT KÓDOLD ÍGY, AZ EGYÜTTÉLÉSÉBŐL NEM

	Házas	**regisztrált partnerkapcsolat**	**Özvegy élnek**	**Elvált**	**különélő**	**hajadon, nőtlen (soha nem házasodott meg és nem volt regisztrált partnerkapcsolatban)**	*NT*	*NV*
v234	1	2	3	4	5	6	*8*	*9*
	MENJ K61							

K60. KÉRDEZŐI INSTRUKCIÓ: - HÁZASSÁG VAGY REGISZTRÁLT ÉLETTÁRSI KAPCSOLAT ESETÉN (A K59. KÉRDÉS 1-ES VAGY 2-ES) MENJ A K61. KÉRDÉSRE!

Van-e társa, akivel együtt él?

	1	2	7	*8*	*9*
v236	Igen	Nem	Nem vonatkozik rá	*NT (spontán)*	*NV (spontán)*

MINDENKITŐL KÉRDEZD

K61. Hányan élnek Önök egy háztartásban? Önmagát és a gyerekeket is beleszámolva hány ember él itt, mint a háztartás tagja

v240	.....személy	1. egyedül élek

K62. Hány éves korban fejezte be Ön nappali iskolai tanulmányait (főiskolát és egyetemet is beleértve, de a tanoncidőt és a szakmai gyakorlatot nem beleszámítva)?

KÉRDEZŐI INSTRUKCIÓ: - Ha a kérdezett még iskolába jár, akkor kérdezd:

Várhatóan hány éves korban fogja befejezni Ön tanulmányait?

	Írd be az életkort	0	*88*	*99*
v242	……	Nem részesült hivatalos oktatásban	*NT (spontán)*	*NV (spontán)*

K63. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 63

Mi a legmagasabb megszerzett iskolai végzettsége?

KÉRDEZŐI INSTRUKCIÓ: - A "MEGSZERZETT" AZT JELENTI, HOGY DIPLOMA/BIZONYÍTVÁNY

v243	
0. nem járt iskolába	13. művezetőképző technikum érettségi nélkül
1. befejezetlen elemi iskola	14. művezetőképző technikum érettségivel
2. befejezett elemi iskola	15. posztliceális képzés érettségi nélkül
3. befejezetlen általános iskola	16. posztliceális képzés érettségivel
4. befejezett általános iskola	17. befejezetlen egyetem
5. inasképző iskola (kiegészítő)	18. elsőfokú szakképzés, technikum vagy almérnöki
6. Művészeti és népiskola	19. befejezett egyetem - 3 év
7. Művészeti és népiskola - kiegészítő év	20. befejezett egyetem - 4 év
8. szakiskola (2 évnél kevesebb)	21. befejezett egyetem - 5 év
9. szakiskola (2-4 éves)	22. befejezett egyetem - 6 év
10. befejezetlen középiskola	23. mesteri fokozat
11. befejezett középiskola érettségivel	24. Doktori fokozat, PhD
12. befejezett szakközépiskola érettségivel	
	88 Nem tudja (spontán módon)
	99 Nincs válasz (spontán módon)

K64. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 64

Van-e Önnek jelenleg fizetett állása vagy nincs? Kérem, válassza ki a kártyáról azt a foglalkoztatási státuszt, ami Önre illik.

KÉRDEZŐI INSTRUKCIÓ: - HA TÖBB ÁLLÁSA VAN: CSAK A FŐÁLLÁSRA VONATKOZÓAN

v244			
Fizetett állás			
1	Heti 30 óra vagy több		***MENJ K66***
2	Kevesebb mint heti 30 óra		
3	Saját vállalkozásában dolgozik		
Nincs fizetett állása			
4	Katonai szolgálatot teljesít		***MENJ K65***
5	Nyugdíjas		
6	Háztartásbeli		
7	Tanuló		
8	Munkanélküli		
9	Rokkant		
KÉRDEZŐI INSTRUKCIÓ: - CSAK HA A VÁLASZADÓ ROKKANTSÁG MIATT NEM DOLGOZIK!			
10	Egyéb, kérem, nevezze meg. (ÍRD BE):.....	v244a	***MENJ K65***
88	*Nem tudja (spontán módon)*		
99	*Nincs válasz (spontán módon)*		

K65. Az UTOLSÓ állásában Ön alkalmazott volt (akár teljes munkaidőben, akár részmunkaidőben), vagy saját vállalkozásában dolgozott?

v245		
1	alkalmazott	
2	saját vállalkozásában dolgozott	
8	*Nem tudja (spontán módon)*	***MENJ K72***
9	*Nincs válasz (spontán módon)*	
6	soha nem volt fizetett állása	
7	Nem vonatkozik rá	

K66. KÉRDEZŐI INSTRUKCIÓ: - HA A KÉRDEZETTNEK JELENLEG VAN FIZETETT ÁLLÁSA (K64 KÉRDÉSRE A VÁLASZ 01, 02 VAGY 03), KÉRDEZD:

Mi a neve vagy titulusa az Ön főfoglalkozású munkájának?
KÉRDEZŐI INSTRUKCIÓ: - HA A KÉRDEZETTNEK KORÁBBAN VOLT FIZETETT ÁLLÁSA (K65 KÉRDÉSRE A VÁLASZ 1 VAGY 2), KÉRDEZD:
Mi volt a neve vagy titulusa az Ön főfoglalkozású munkájának?
KÉRDEZŐI INSTRUKCIÓ: - HA A VÁLASZADÓNAK TÖBB MUNKAHELYE VAN VAGY VOLT, A FŐÁLLÁSÚ FOGLALKOZÁSÁT KÉRDEZD!

ÍRD LE A LEHETŐ LEGRÉSZLETESEBBEN!

v246a	Írd be ……

K67. Milyen jellegű munkát végez/végzett a főállásában ideje nagy részében?

KÉRDEZŐI INSTRUKCIÓ: - ÍRD LE RÉSZLETESEN

v246b	Írd be ……

AZ INTERJÚT KÖVETŐEN KÓDOLD AZ ISCO08 SZERINT (4 SZÁMJEGY) A K66 ÉS K67 KÉRDÉSEK ALAPJÁN

	Kód	*NT (spontán)*	*NV (spontán)*	*Nem vonatkozik rá*
v246c	_ _ _ _ _	*8888*	*9999*	*7777*

K68. KÉRDEZŐI INSTRUKCIÓ: - AZOKTÓL KÉRDED, AKIK JELENLEG SAJÁT VÁLLALKOZÁSUKBAN DOLGOZNAK (K64. KÉRDÉSRE A VÁLASZ 3) VAGY A LEGUTÓBBI ÁLLÁSUKBAN SAJÁT VÁLLALKOZÁSUKBAN DOLGOZTAK (K65 KÉRDÉSRE A VÁLASZ 2)

Hány alkalmazottja van/volt?

	1	2	3	4	*8*	*9*	*7*
v247	Egy sem	1-9	10-24	25 vagy több	*NT (spontán)*	*NV (spontán)*	*Nem vonatkozik rá*
MENJ K72							

K69. KÉRDEZŐI INSTRUKCIÓ: - AZOKTÓL KÉRDEZD, AKIK JELENLEG ALKALMAZOTTAK (K64. KÉRDÉSRE A VÁLASZ 1 VAGY 2) VAGY A LEGUTÓBBI ÁLLÁSUKBAN NEM SAJÁT VÁLLALKOZÁSUKBAN DOLGOZTAK (K65 KÉRDÉSRE A VÁLASZ 1)

Felügyeli/felügyelte Ön más alkalmazottak munkáját?

	1	2	8	9	*7*
v248	Igen	Nem	*NT (spontán)*	*NV (spontán)*	*Nem vonatkozik rá*
		MENJ K71			

K70. Hány egyéb alkalmazottat felügyel/felügyelt?

	2	3	4	*8*	*9*	*7*
v248a	1-9	10-24	25 vagy több	*NT (spontán)*	*NV (spontán)*	*Nem vonatkozik rá*

K71. Hol dolgozik Ön?

	v249			
1	Állami vagy közintézmény		*8*	*Nem tudja (spontán módon)*
2	Magáncég vagy magánvállalkozás		*9*	*Nincs válasz (spontán módon)*
3	Nonprofit szervezet		*7*	*Nem vonatkozik rá*

KÉRDEZŐI INSTRUKCIÓ: - HA A VÁLASZADÓ HÁZAS VAGY REGISZTRÁLT ÉLETTÁRSI KAPCSOLATBAN ÉL (A K59 KÉRDÉS 1 VAGY 2); EGYÉBKÉNT MENJ A K73-AS KÉRDÉSRE

K72. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 72

Mi az Ön házastársa vagy partnere legmagasabb megszerzett iskolai végzettsége?

KÉRDEZŐI INSTRUKCIÓ: - A "MEGSZERZETT" AZT JELENTI, HOGY DIPLOMA/BIZONYÍTVÁNY

v252	
0. nem járt iskolába	13. művezetőképző technikum érettségi nélkül
1. befejezetlen elemi iskola	14. művezetőképző technikum érettségivel
2. befejezett elemi iskola	15. posztliceális képzés érettségi nélkül
3. befejezetlen általános iskola	16. posztliceális képzés érettségivel
4. befejezett általános iskola	17. befejezetlen egyetem
5. inasképző iskola (kiegészítő)	18. elsőfokú szakképzés, technikum vagy almérnöki
6. Művészeti és népiskola	19. befejezett egyetem - 3 év
7. Művészeti és népiskola - kiegészítő év	20. befejezett egyetem - 4 év
8. szakiskola (2 évnél kevesebb)	21. befejezett egyetem - 5 év
9. szakiskola (2-4 éves)	22. befejezett egyetem - 6 év
10. befejezetlen középiskola	23. mesteri fokozat
11. befejezett középiskola érettségivel	24. Doktori fokozat, PhD
12. befejezett szakközépiskola érettségivel	
	88 Nem tudja (spontán módon)
	99 Nincs válasz (spontán módon)

K73. MINDENKITŐL KÉRDEZD

Volt-e Ön három hónapnál hosszabb ideig folyamatosan munkanélküli az elmúlt öt év során?

	Igen	Nem	*NT*	*NV*
v259	1	2	*8*	*9*

K74. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 74

Itt egy lista a jövedelmekről, és szeretnénk megtudni, hogy az Ön háztartása melyik csoportba tartozik, beszámítva az összes bért, fizetést, nyugdíjat és egyéb befolyó jövedelmet. Csak adja meg a betűjelét annak a csoportnak, amelybe az Ön háztartása tartozik az adók és egyéb levonások után.

v261		**HETENTE körülbelül**	**HAVONTA körülbelül**	**ÉVENTE körülbelül**
1	A	Kevesebb, mint 125 lej	Kevesebb, mint 500 lej	Kevesebb, mint 6000 lej
2	B	126-250 lej	501-1000 lej	6001-12.000 lej
3	C	251-500 lej	1001-2000 lej	12.001-24.000 lej
4	D	501-750 lej	2001-3000 lej	24.000-36.000 lej
5	E	751-1000 lej	3001-4000 lej	36.001-48.000 lej
6	F	1001-1250 lej	4001-5000 lej	48.001-60.000 lej
7	G	1251-1500 lej	5001-6000 lej	60.001-72.000 lej
8	H	1501-1875 lej	6001-7500 lej	72.001-90.000 lej
9	I	1876-2250 lej	7501-9.000 LEJ	90.001-108.000 lej
10	J	Több, mint 2250 lej	Több, mint 9.000 lej	Több, mint 108.000
88		*NT (spontán)*		
99		*NV (spontán)*		

K75. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 75/76

Mi az Ön édesapja legmagasabb megszerzett iskolai végzettsége?

KÉRDEZŐI INSTRUKCIÓ: - A "MEGSZERZETT" AZT JELENTI, HOGY DIPLOMA/BIZONYÍTVÁNY

v262	
0. nem járt iskolába	13. művezetőképző technikum érettségi nélkül
1. befejezetlen elemi iskola	14. művezetőképző technikum érettségivel
2. befejezett elemi iskola	15. posztliceális képzés érettségi nélkül
3. befejezetlen általános iskola	16. posztliceális képzés érettségivel
4. befejezett általános iskola	17. befejezetlen egyetem
5. inasképző iskola (kiegészítő)	18. elsőfokú szakképzés, technikum vagy almérnöki
6. Művészeti és népiskola	19. befejezett egyetem - 3 év
7. Művészeti és népiskola - kiegészítő év	20. befejezett egyetem - 4 év
8. szakiskola (2 évnél kevesebb)	21. befejezett egyetem - 5 év
9. szakiskola (2-4 éves)	22. befejezett egyetem - 6 év
10. befejezetlen középiskola	23. mesteri fokozat
11. befejezett középiskola érettségivel	24. Doktori fokozat, PhD
12. befejezett szakközépiskola érettségivel	
	88 Nem tudja (spontán módon)
	99 Nincs válasz (spontán módon)

K76. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 75/76

Mi az Ön édesanyja legmagasabb megszerzett iskolai végzettsége?

KÉRDEZŐI INSTRUKCIÓ: - A "MEGSZERZETT" AZT JELENTI, HOGY DIPLOMA/BIZONYÍTVÁNY

v263	
0. nem járt iskolába	13. művezetőképző technikum érettségi nélkül
1. befejezetlen elemi iskola	14. művezetőképző technikum érettségivel
2. befejezett elemi iskola	15. posztliceális képzés érettségi nélkül
3. befejezetlen általános iskola	16. posztliceális képzés érettségivel
4. befejezett általános iskola	17. befejezetlen egyetem
5. inasképző iskola (kiegészítő)	18. elsőfokú szakképzés, technikum vagy almérnöki
6. Művészeti és népiskola	19. befejezett egyetem - 3 év
7. Művészeti és népiskola - kiegészítő év	20. befejezett egyetem - 4 év
8. szakiskola (2 évnél kevesebb)	21. befejezett egyetem - 5 év
9. szakiskola (2-4 éves)	22. befejezett egyetem - 6 év
10. befejezetlen középiskola	23. mesteri fokozat
11. befejezett középiskola érettségivel	24. Doktori fokozat, PhD
12. befejezett szakközépiskola érettségivel	
	88 Nem tudja (spontán módon)
	99 Nincs válasz (spontán módon)

K77. MUTASD A KÖVETKEZŐ KÁRTYALAPOT: 77

Gondoljon a szüleire, amikor Ön 14 év körüli volt! Mit mondana ezekről az állításokról, helyesen írják le őket?

		Igen	**Bizonyos mértékig**	**Egy kicsit**	**Nem**	***NT***	***NV***
v267	Az édesanyám szeretett könyvet olvasni	1	2	3	4	*8*	*9*
v268	Édesanyámmal otthon megvitattuk a politikai kérdéseket	1	2	3	4	*8*	*9*
v269	Édesanyám szerette figyelemmel kísérni a híreket	1	2	3	4	*8*	*9*
v270	A szüleimnek gondot okozott kijönni a jövedelmükből	1	2	3	4	*8*	*9*
v271	Az édesapám szeretett könyvet olvasni	1	2	3	4	*8*	*9*
v272	Édesapámmal otthon megtárgyaltuk a politikai kérdéseket	1	2	3	4	*8*	*9*
v273	Édesapám szerette figyelemmel kísérni a híreket	1	2	3	4	*8*	*9*
v274	A szüleimnek gondot okozott pótolni a tönkrement tárgyakat	1	2	3	4	*8*	*9*

KÖSZÖNÖM A VÁLASZADÁST!

A KÉRDEZŐ TÖLTI KI

K78. Régió:

Írja be:	v275	Kód _ _ _ _ _ _	v275a

KÓDOLÁSI INSTRUKCIÓ: KÓDOLD A RÉGIÓT A NUTS 3 SZERINT

K79. A település lélekszáma

v276			
1	kevesebb, mint 2000 fő	6	50 - 100.000
2	2-5.000	7	100 - 500.000
3	5-10.000	8	500.000 fő vagy több
4	10 - 20.000	*88*	*NT*
5	20 - 50.000	*99*	*NV*

K80. A kérdezés alatt a válaszadó...

V280	
1	nagyon érdeklődő volt
2	valamelyest érdeklődő volt
3	nem volt nagyon érdeklődő

SZAKIRODALOM

ANDERSEN, Erling B. – JENSEN, Niels Erik – KOUSGAARD, Nils
1987 *Statistics for Economics. Business Administration and the Social Sciencies. [Gazdaságstatisztika. A vállalatvezetés és a társadalomtudományok]*. New York, Springer–Verlag, LLC.

ANGHELACHE, Constantin
1999 *Statistică generală*. Bucureşti, Editura Economică

ANGHELACHE, Constantin – NICULESCU, Emanuela
2001 *Statistică. Indicatori, formule de calcul şi sinteze.* Bucureşti, Editura Economică

ARGYROUS, George
2011 *Statistics for research: With a guide to SPSS. [Statisztika a kutatáshoz: Útmutató az SPSS-hez]*. Third Edition. SAGE Publications Ltd.

BABBIE, Earl
1996 *A társadalomtudományi kutatás gyakorlata*. Budapest, Balassi Kiadó

BUIGA, Anuţa
2001 *Metodologii de sondaj şi analiza datelor în studiile de piaţă.* Cluj-Napoca, Presa Universitară Clujeană

CSALLNER András Erik
2015 *Bevezetés az SPSS statisztikai programcsomag használatába. Jegyzet.* Szeged, Szegedi Tudományegyetem, Juhász Gyula Pedagógusképző Kar. http://www.inf.u-szeged.hu/~banhelyi/okt/SPSS_2021tavasz/csallner-spss-javitott.pdf

EVS
2020 *European Values Study 2017: Romania - Hungarian minority (EVS 2017 Country data file). [Európai Értékek Tanulmány 2017: Románia – magyar kisebbség (EVS 2017 országadat-fájl)]*. GESIS Data Archive, Cologne. ZA7550 Data file Version 1.0.0, https://doi.org/10.4232/1.13562. https://search.gesis.org/research_data/ZA7550?doi=10.4232/1.13562

EVS, GESIS
2022 *European Values Study (EVS) 2017 Method Report. [Európai Értékek Tanulmány (EVS) 2017 Módszertani Jelentés]*. Köln, GESIS Papers 2022|07, GESIS – Leibniz-Institut für Sozialwissenschaften, GESIS Data Archive, Cologne. ZA7500, ZA7501 and ZA7502. https://www.ssoar.info/ssoar/handle/document/79215

FALUS Iván – OLLÉ János
2000 *Statisztikai módszerek pedagógusok számára.* Budapest, Okker Kiadó Zrt.

FÜSTÖS László
1988 *Az exploratív faktorelemzés módszerei.* Budapest, MTA Szociológiai Kutató Intézet, Értékszociológiai- és Társadalomtudományi Elemzések Műhelye

GUPTA, Vijay
1999 *SPSS for Beginners.* [Spss kezdőknek]. VJBooks Inc.

HAJDU Ottó
2003 *Többváltozós matematikai számítások. Statisztikai módszerek a társadalmi és gazdasági elemzésekben.* Budapest, Központi Statisztikai Hivatal

HOWITT, Dennis – CRAMER, Duncan
2006 *Introducere în SPSS pentru psihologie: Versiunile SPSS 10, 11, 12 şi 13.* Iaşi, Editura Polirom

HUNYADI László – MUNDRUCZÓ György – VITA László
2000 *Statisztika*. Budapest, Aula Kiadó

HUZSVAI László
2004 *Biometriai módszerek az SPSS-ben. SPSS alkalmazások.* Debreceni Egyetem, Mezőgazdaságtudományi Kar

HUZSVAI László – VINCZE Szilvia
2012 *SPSS-könyv.* Seneca Books. http://seneca-books.hu/doc/spsskonyv.pdf

KEMÉNY Ildikó – SIMON Judit – BEREZVAI Zombor – KUN Zsuzsanna
2021 *Marketingkutatás kvantitatív módszerei – segédanyag SPSS program használatához.* Budapest, Budapesti Corvinus Egyetem. http://unipub.lib.uni-corvinus.hu/6387/1/Kemeny_et_al_marketinkutatas_2021.pdf

KETSKEMÉTY László – IZSÓ Lajos, Dr
1996 *Az SPSS for Windows programrendszer alapjai.* Budapest, SPSS Partner Bt.

KORPÁS Attiláné (szerk.)
1996 *Általános statisztika I.* Budapest, Nemzeti Tankönyvkiadó

1997 *Általános statisztika II.* Budapest, Nemzeti Tankönyvkiadó.

KÖVESI János – ERDEI János – TÓTH Zsuzsanna Eszter – NAGY Jenő Bence
2007 *Gazdaságstatisztika*. Budapest, Budapesti Műszaki és Gazdaságtudományi Egyetem, Üzleti Tudományok Intézet, Menedzsment és Vállalatgazdaságtan Tanszék.

LANDAU, Sabine – EVERITT, Brian
2004 *A Handbook of Statistical Analyses using SPSS. [SPSS segítségével végzett statisztikai elemzések kézikönyve].* Chapman & Hall/CRC Press LLC. https://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/SPSS_Statistical_Analyses_using_SPSS.pdf

LÁZÁR Ede
2022 *Közgazdasági kutatásmódszertan. Egyetemi jegyzet.* Kolozsvár, Risoprint Kiadó. https://risoprint.ro/ebooks/lazar_ede_bt_e-book.pdf

LUKÁCS Ottó
2002 *Matematikai statisztika.* Budapest, Műszaki Könyvkiadó

MEZEI Elemér – VERES Valér
2001 *Társadalomstatisztika.* Kolozsvár, Egyetemi Kiadó

MOKSONY Ferenc
1999 *Gondolatok és adatok. Társadalomtudományi elméletek empirikus ellenőrzése.* Budapest, Osiris Kiadó

NÉMETH Renáta – SIMON Dávid
2010 *Társadalomstatisztika. Jegyzet.* Budapest, ELTE, elektronikus tananyag. https://dtk.tankonyvtar.hu/bitstream/handle/123456789/7422/0010_2A_21_Nemeth_Renata-Simon_David_Tarsadalomstatisztika_magyar_es_angol_nyelven.pdf?sequence=1&isAllowed=y

OPARIUC-DAN, Cristian
2009 *Statistică aplicată în ştiinţele socio-umane: noţiuni de bază: statistici univariate.* Iaşi, Cristian Opariuc-Dan

2012 Analiza componentelor principale pentru date categoriale (CATPCA). *Psihologia Resurselor Umane* 10. 2. 103–117.

2023 *Introducere în analiza datelor: Tomul I. Măsurarea, colectarea datelor şi modele statistice. Aplicaţii în IBM SPSS Statistics şi R.* Bucureşti, Editura Universităţii din Bucureşti – Bucharest University Press

2025 *Introducere în analiza datelor: Tomul II. 1 Analize descriptive univariate. Probabilităţi şi distribuţii. Aplicaţii în IBM SPSS Statistics şi R.* Bucureşti, Editura Universităţii din Bucureşti-Bucharest University Press

PAH, Iulian
2004 *Tehnici de analiză a datelor cu SPSS.* Cluj-Napoca, Presa Universitară Clujeană

PECK, Roxy – OLSEN, Chris – DEVORE, Jay
2008 *Introduction to Statistics and Data Analysis. [Bevezetés a statisztikába és az adatelemzésbe].* Third Edition. USA, Thomson Brooks/Cole, Belmont

ROTARIU, Traian – BĂDESCU, Gabriel – CULIC, Irina – MEZEI Elemér – MUREŞAN, Cornelia
1999 *Metode statistice aplicate în ştiinţele sociale.* Iaşi, Editura Polirom

SAJTOS László – MITEV Ariel
2007 *SPSS kutatási és adatelemzési kézikönyv.* Budapest, Alinea Kiadó

SANDU, Dumitru
1992 *Statistica în ştiinţele sociale*. Bucureşti, Universitatea din Bucureşti

SINCICH, Terry
1989 *Business Statistics by Example. [Gazdaságstatisztika példákon keresztül].* Dellen Publishing Company, Collier Macmillan Publishers

SPIEGEL, Murray R.
1995 *Statisztika. Elmélet és gyakorlat.* Budapest, Panem-McGraw-Hill

SZÉKELYI Mária – BARNA Ildikó
2002 *Túlélőkészlet az SPSS-hez.* Budapest, Typotex Kiadó

SZÉKELYI Mária – ÖRKÉNY Antal
1998 *Statistical Methods in Social Research – Adv. II.* Budapest, ELTE-UNESCO Minority Studies Program

TARLING, Roger
2009 *Statistical modelling for social researchers: Principles and practice. [Statisztikai modellezés társadalomtudományi kutatók számára: elvek és gyakorlat]*. Routledge

VARGHA András
2000 *Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal.* Budapest, Pólya Kiadó

REZUMAT

Metode statistice şi analiza datelor în ştiinţele sociale cu IBM SPSS

Cartea de faţă este un curs universitar destinat studenţilor de la specializarea ştiinţe sociale. Scopul său este de a prezenta fundamentele teoretice şi aplicarea practică a celor mai frecvent utilizate metode statistice în ştiinţele sociale. Manualul sprijină utilizarea practică a cunoştinţelor teoretice prin exemple rezolvate pas cu pas, atât clasic, cât şi cu ajutorul programului IBM SPSS.

Exemplele utilizează o bază de date online gratuită (EVS România – studiu asupra minorităţii maghiare, 2020) şi, pas cu pas, explică cele mai importante concepte statistice fundamentale (populaţie, variabilă, niveluri de măsurare), precum şi operaţiunile legate de baze de date (crearea, etichetarea, importul, îmbinarea, selecţia cazurilor, transformarea variabilelor).

Capitolul 2, dedicat analizelor univariate, prezintă distribuţiile de frecvenţă, măsurile de tendinţă centrală, indicatorii de dispersie şi indicatorii de formă, împreună cu interpretarea acestora. Capitolul 3 oferă un rezumat al elementelor fundamentale ale teoriei probabilităţilor şi ale eşantionării probabilistice, precum şi metodele de calcul ale erorii standard.

Capitolul 4 se concentrează pe analizele bivariate, detaliind tipurile de relaţii între variabile: asocierea între două variabile calitative (testul chi-pătrat şi gamma), compararea mediilor între o variabilă cantitativă şi una calitativă (testul t şi ANOVA) şi corelaţia între două variabile cantitative.

Capitolul 5 oferă o privire de ansamblu asupra analizelor multivariate şi parcurge etapele regresiei liniare multivariate, analizei factoriale (metoda componentelor principale) şi analizei clusterelor (metoda K-Means) în IBM SPSS.

Cursul subliniază două mesaje principale încă din primul capitol:

1. Înţelegerea tehnicilor statistice necesită practică în aplicarea metodelor. Cunoştinţele teoretice sprijină practica, dar competenţele de analiză a datelor se dezvoltă eficient în activitatea practică, cu ajutorul programelor informatice.
2. Instrumentele statistice nu pot fi aplicate mecanic, este necesară expertiza în ştiinţele sociale. Chiar şi cea mai complexă analiză statistică nu poate corecta erorile făcute la planificarea cercetării, iar rezultatele obţinute pot fi utilizate eficient doar cu o pregătire profesională adecvată.

ABSTRACT

Statistical Methods and Data Analysis in the Social Sciences with IBM SPSS

The present book is a university course textbook designed for students in social sciences programmes. Its purpose is to present both the theoretical foundations and the practical applications of the most commonly used statistical methods in social sciences. The textbook supports the practical use of theoretical knowledge through step-by-step solved examples, both manually and using IBM SPSS.

The examples use a freely available online database (EVS Romania – Hungarian minority study, 2020) and guide students step by step through the most important basic statistical concepts (population, variable, levels of measurement), as well as database-related operations (creation, labelling, import, merging, case selection, variable transformation).

Chapter 2, focused on univariate analyses, introduces frequency distributions, measures of central tendency, dispersion indicators, and shape indicators, along with their interpretation. Chapter 3 briefly summarizes the basic elements of probability theory and probabilistic sampling, as well as methods for calculating the standard error.

Chapter 4 focuses on bivariate analyses, detailing the different types of relationships: association between two qualitative variables (chi-square test and gamma), comparison of group means between a categorical and a quantitative variable (t-test and ANOVA), and correlation between two quantitative variables.

Chapter 5 provides an overview of multivariate analyses and guides the reader through the steps of multivariate linear regression, factor analysis (Principal Component method), and cluster analysis (K-Means method) in IBM SPSS.

The textbook emphasizes two key messages from the very first chapter:

1. Understanding statistical techniques requires practice in applying the methods. Theoretical knowledge supports practice, but data analysis skills are developed most effectively through hands-on work, aided by software packages.
2. Statistical tools cannot be applied mechanically; expertise in social sciences is also required. Even the most complex statistical analysis cannot correct mistakes made in the planning of the research, and the results can only be used effectively with appropriate professional knowledge.

A SZERZŐRŐL

Bálint Gyöngyvér szociológus, 2002 óta a Sapientia EMTE Csíkszeredai Karának Társadalomtudományi Tanszékén főállású oktató. 1997-ben szerzett szociológusi diplomát a Babeş–Bolyai Tudományegyetemen, 2009-ben pedig a Budapesti Corvinus Egyetemen PhD-fokozatot. Oktatási tevékenysége kiterjed a szociológiai és humánerőforrás-menedzsment területekre, kiemelt hangsúlyt fektetve a mennyiségi adatelemzésre és az IBM SPSS gyakorlati alkalmazására.

Kutatási érdeklődése a munkaerőpiac, a környezetszociológia, a migráció, a demográfia és a szabadidős tevékenységek területére terjed ki. Elemzéseiben arra keres választ, hogyan alakítják a társadalmi struktúrák és kulturális tényezők a munkatapasztalatokat, a környezeti magatartást és a migrációs mintákat, valamint hogyan befolyásolják a szabadidős tevékenységek az emberek mindennapi életét és társas kapcsolatait.

ORCID: 0009-0004-8286-7608
E-mail cím: balintgyongyver@uni.sapientia.ro

Scientia Kiadó
400112 Kolozsvár (Cluj-Napoca)
Mátyás király (Matei Corvin) u. 4. sz.
Tel./fax: +40-364-401454
E-mail: scientia@kpi.sapientia.ro
www.scientiakiado.ro

Műszaki szerkesztés:
Ruzsa István

Borítóterv:
Tipotéka Kft.

Korrektúra:
Szenkovics Enikő

Tipográfia:
Könczey Elemér

A *Statisztikai módszerek és adatelemzés a társadalomtudományokban IBM SPSS segítségével* című egyetemi jegyzet társadalomtudományi szakos hallgatók számára készült, és a leggyakrabban alkalmazott statisztikai eljárások elméleti hátterét, valamint gyakorlati alkalmazását mutatja be. A tananyag kézi számításokra és az IBM SPSS programban végigvezetett példákra támaszkodva segíti az elméleti ismeretek alkalmazását az empirikus adatelemzésben. Áttekinti az alapvető statisztikai fogalmakat, az adatkezelés főbb lépéseit, valamint az egy-, két- és többváltozós elemzések legfontosabb típusait, különös tekintettel azok társadalomtudományi felhasználására. Célja, hogy gyakorlatorientált módon támogassa az önálló adatelemzési készségek kialakítását és a statisztikai eszközök tudatos alkalmazását.

ISBN 978-606-975-110-7