Számítógépes statisztika

Bánhelyi Balázs, Csendes Tibor

2014

A tananyag a TÁMOP-4.1.2.A/1-11/1-2011-0104 "A felsőfokú informatikai oktatás minőségének fejlesztése, modernizációja" c. projekt keretében a Pannon Egyetem és a Szegedi Tudományegyetem együttműködésében készült.

BEVEZETÉS A SZÁMÍTÓGÉPES STATISZTIKÁBA

Bánhelyi Balázs és Csendes Tibor

Szeged, 2014.

Lektorálta:

Galambosné Tiszberger Mónika
Pécsi Tudományegyetem
Közgazdaságtudományi Kar

Kehl Dániel
Pécsi Tudományegyetem
Közgazdaságtudományi Kar

Előszó

Jelen jegyzet[1] a Szegedi Tudományegyetemen a TTIK-s hallgatók Számítógépes statisztika című tárgya anyagát tartalmazza. A tárgy ideális esetben heti egy óra előadást és egy óra gyakorlatot jelent, de mindenképpen legalább a fele időt számítógépes teremben, aktív gyakorlással kell tölteni.

A jegyzet nem támaszkodik más tárgyakra az érintett hallgatók különböző előképzettsége miatt, de nagyon hasznos a matematikai statisztika vagy a statisztika tárgyak anyagának előzetes ismerete. Az aktuális változat egy része, a hozzá kapcsolódó feladatok, gyakorlatok és adataik elérhetők a

`http://www.inf.u-szeged.hu/~banhelyi/stat`

címen.

A tárgy olyan tudást kíván adni, amely elegendő egyszerűbb statisztikai munkák elvégzéséhez, és amelyet önálló gyakorlással továbbfejlesztve egy-egy szakterület teljes statisztikai feldolgozását is végre lehet hajtani. Mivel a programcsomagok gyakran változnak, az anyag főleg az állandó vagy kevésbé változó ismereteket tartalmazza.

A rendelkezésre álló rövid idő (kb. 14×1.5 óra) nem elég a valószínűségszámítás és a matematikai statisztika alapfogalmainak részletes tárgyalására sem, ezért a legfontosabb definíciókat, összefüggéseket az érintett statisztikai eljárások tárgyalása előtt csak a feltétlenül szükséges terjedelemben ismertetjük. A teljesen önálló statisztikai munkához ez persze nem elegendő. Ennek ellenére bízunk benne, hogy a tárgyalt anyag segít a leggyakoribb hibákat elkerülni, és a viszonylag könnyen kezelhető programok segítségével (támaszkodva a mind több esetben rendelkezésre álló kiterjedt súgó, tanácsadó varázslókra) önálló munkával is lehetséges a további szükséges eljárások megismerése. A teljes itt közreadott anyag több, mint amit egy féléves kurzusban át lehet adni, ez némi rugalmasságot követel az előadótól, illetve a gyakorlatvezetőtől.

[1] Minden megjegyzést szívesen látunk és előre is köszönünk, különösen, ha hibákra hívják fel a figyelmet. Az e-mail cím: `banhelyi@inf.u-szeged.hu`

További cél segítséget nyújtani a statisztikai feldolgozáshoz olyanok számára is, akik ezt a hagyományos képzés keretében nem tanulták. Így a jegyzet alapján az egyszerűbb feladatok esetén az olvasó elegendő útmutatást kap ahhoz, hogy a feladatát úgy fogalmazza meg, illetve írja át, hogy az a rendelkezésre álló szoftverrel hatékonyan megoldható legyen.

A jelen jegyzet a korábbi speciálkollégiumok és gyakorlatok során csiszolódott anyagot is tartalmazza. Itt mondunk köszönetet korábbi hallgatóinknak és munkatársainknak a jegyzet létrejöttéhez, illetve a javításához nyújtott segítségükért. Külön köszönet illeti a lektorokat, akik alapos és gyors munkát végeztek, és számos hasznos tanáccsal segítették munkánkat. Várjuk a további véleményeket és javaslatokat is.

Szeged, 2014. április 7.

a szerzők

Jelölések

Itt a legfontosabb, szinte mindig a megadott formában használatos jelöléseket adjuk meg, de ezektől helyenként – ahol a tárgyalás ezt megköveteli – eltérhetünk.

A, B, C	események
α	szignifikancia-szint
$D_1, \ldots, D_9$	decilisek
$E(X)$	az X valószínűségi változó várható értéke
$F(X) = P\{\xi < x\}$	az X valószínűségi változó eloszlásfüggvénye
$f(X)$	az X valószínűségi változó sűrűségfüggvénye
H_0	a nullhipotézis
H_1	az alternatív hipotézis
Me	medián
μ	(elméleti) átlag
$N(0, 1)$	standard normális eloszlás
$N(\mu, \sigma)$	normális eloszlás
p, q	valószínűségek
$P(A)$	egy A esemény valószínűsége
$P_1, \ldots, P_{99}$	percentilisek

Q_1, Q_2, Q_3	a kvartilisek
σ	(elméleti) szórás
s	korrigált szórás
SD	szórás
SE	standard hiba
X, Y, Z	valószínűségi változók. A matematikai statisztikában inkább görög betűk használatosak (mint pl. ξ, ζ)
x_i	valószínűségi változó mintaelemei
$\bar{x}$	az $x_1, x_2, \ldots, x_n$ mintaelemek számtani átlaga

1. fejezet

Bevezetés

A statisztikával kapcsolatban két gyakran idézett mondás a: „*van hazugság, nagy hazugság és statisztika*", illetve az „*elegendő számú adatból statisztikával bármit ki lehet mutatni*". Ezek mögött az az igazság rejlik, hogy a statisztikai eljárások nem elég gondos, nem elegendően körültekintő használata esetén megkérdőjelezhetetlennek tűnő hibás eredményeket kaphatunk. A statisztikai programcsomagok ismertetése során a leggyakoribb hibalehetőségeket is megtárgyaljuk az elkerülésükhöz szükséges lépésekkel.

A jegyzet címében a számítógépes jelző arra utal, hogy közvetlenül nem a statisztika fogalmaival, összefüggéseivel foglalkozunk, hanem statisztikai eljárások, próbák, mutatók konkrét adatokra való meghatározásával. Márpedig ezeket ma a legegyszerűbbek kivételével csak számítógépen hajtják végre. Néhány statisztikai eljárás más jellegű programban is elérhető, így például gyakran táblázatkezelő programban (Excel), vagy általános numerikus programcsomagokban (Matlab, Maple) is találunk ilyeneket.

Bár ezek a programok kevés statisztikai tesztet, algoritmust bocsátanak rendelkezésre, mégis elterjedtségük miatt és viszonylag könnyű kezelhetőségük révén fontos eszközök.

Az egyszerűbb statisztikai programok, mint a SigmaPlot is, csak egyváltozós statisztikákat képesek kiszámolni, cserébe viszont könnyen kezelhetők és kisebb kapacitású gépen is futtathatók, olcsóbbak.

A statisztikai eljárások közel teljes körét rendelkezésre bocsátó professzionális programokból sok van, ezeket főleg PC-n vagy munkaállomásokon használhatjuk. Ide tartozik a részletesen tárgyalt SPSS mellett például a Statistica, a BMPD és az SAS. Ezen osztály által kínált algoritmusok köre nem nagyon tér el, és bár a használatuk nagyon különböző lehet, a céljainkra elegendő ezek közül egyet ismertetni.

Macintosh és Linux operációs rendszerhez is számos – akár

ingyenes – programot lehet találni. Egy bő lista van ezekről a `http://freestatistics.altervista.org/?p=stat` internetes címen (további linkekkel és rövid ismertetéssel minden programról). Ezek közül kiemelnénk az R-t, mely manapság az egyik legelterjedtebb ingyenes statisztikai szoftver. Ez elérhető mind windows, mind linuxos, de Mac OS környezetre is.

Az általános célú numerikus programok közül a NAG programcsomagját, a Maple és a Mathematica szimbolikus algebrarendszereket kell megemlíteni, de ide tartozik a Matlab is. Ezeket a programokat nem tárgyaljuk, mert még rövid ismertetésük is aránytalanul sok időbe kerülne.

1.1. Statisztikai alapfogalmak

A *statisztika* olyan eljárásokkal foglalkozik, amelyek mérési adatok, felmérésekre kapott válaszok vagy más véletlen eseményektől függő adatok jellemzőit vagy összefüggésük mértékét és jellegét határozzák meg. Ide tartozik a kapott eredmények olyan megjelenítése is, amely az adatok értelmezését megkönnyíti. Ezt a diszciplinát szokás *általános statisztiká*nak is hívni, szemben a matematikai alapjait tisztázó *matematikai statisztikával*. De röviden statisztikának szokás nevezni a *statisztikai függvény*eket, a mintaelemekből számított értékeket is, mint amilyen például az átlag.

A *valószínűség:* egy 0 és 1 közötti szám ($0 \leq p \leq 1$), amely azt jellemzi, hogy egy esemény bekövetkezte milyen eséllyel, gyakorisággal várható. Az 1 valószínűség csaknem biztos bekövetkezést, a nulla valószínűség csaknem lehetetlen előfordulást jelent[1]. A kísérletezés során tapasztalt relatív gyakoriságok megközelítik az elméleti valószínűséget.

Az adatokat általában egy táblázatban célszerű elrendezni. Az *eset* az összetartozó statisztikai adatok olyan egysége, amelyek amiatt képeznek egységet, mert egy egyedre, vagy mérési kísérletre vonatkoznak (pl. a kísérletben résztvevő személy, állat, vegyület stb.). Az eseteket általában egy-egy számítógépes rekordban, rendszerint a táblázat soraiban adjuk meg.

A tulajdonságokat, jellemzőket az egyes egyedekre vonatkozóan a *valószínűségi változó*k (röviden változók) tartalmazzák. Az esetekre vonatkozó változóértékek alkotják a *statisztikai mintá*t, vagy röviden *mintá*t. Sok esetben jellemző az, hogy a teljes sokaságból csak kevés egyedre vonatkozó adat áll rendelkezésre.

[1] A köznyelvben itt használhatunk biztos, illetve lehetetlen előfordulást is, a „csaknem" a matematikai pontosság kedvéért áll itt.

Számos példát lehet ezekre a fogalmakra hozni, így statisztikai mintának tekinthetjük a szavazási hajlandóságot, illetve a választási preferenciákat vizsgáló közvéleménykutatás alapadatát. Az eseteknek ekkor egy-egy megkérdezettre vonatkozó adathalmaz felel meg, míg a feltett kérdésekre kapott válaszok változók értékeit adják. A válaszadók átlagéletkora például egy olyan statisztikai mutató, amit a fenti értelemben statisztikának is szoktak röviden nevezni.

Egy másik példa egy új gyógyszer hatásosságának vizsgálatára gyűjtött adatsor. Ilyenkor két csoportra szokás osztani a pácienseket, az egyik csoport kapja a vizsgálandó kezelést, a másik (az ún. kontroll csoport) hatástalan gyógyszert kap – hogy valóban csak a szer hatását mérjük, ne az egyéb, pl. pszichés következményeket. A betegenként gyűjtött adatok tartoznak egy esethez, a mért értékek pedig egy-egy változóhoz. Olyan statisztikát szokás vizsgálni, mint a megcélzott mérhető értékek átlagos eltérése a csoportok által reprezentált sokaságok között.

A táblázatkezelő programok az eseteket sorokban, a változókat oszlopokban tárolják. Ezt követik a statisztikai programok is. Másrészt több statisztikai eljárás szempontjából az esetek és a változók szerepe felcserélhető (mint pl. a klaszterezés esetén). Ekkor az illető eljárás legtöbbször meg is kérdezi, hogy esetekre vagy változókra kérjük a végrehajtását. A legtöbb statisztikai feldolgozás nyilvánvalóvá teszi, hogy mik lesznek az esetek és mik a változók.

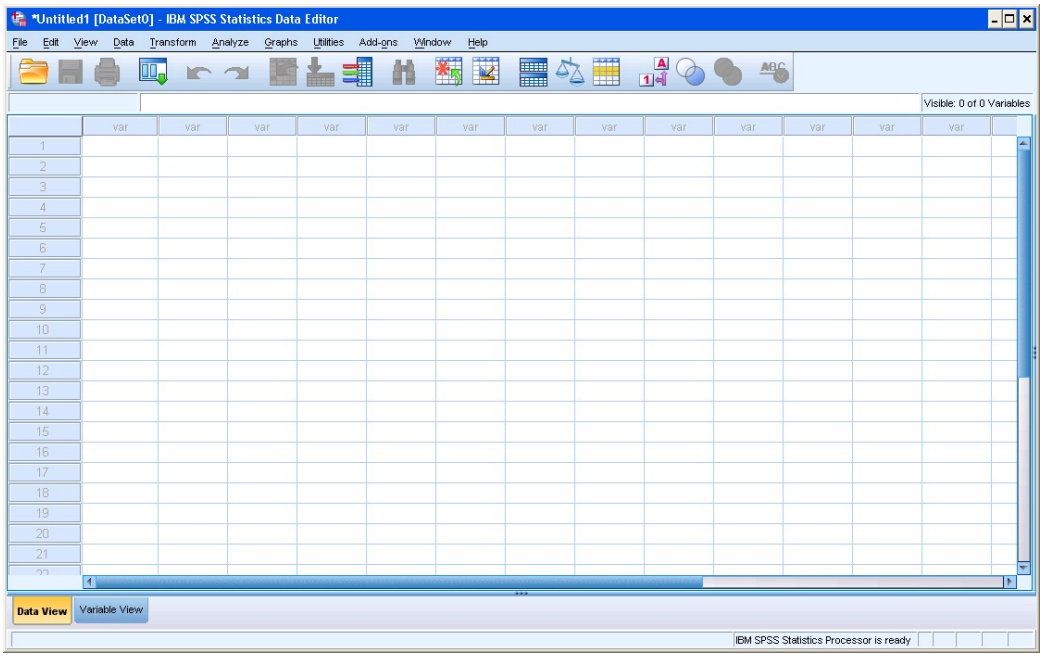
Hasonlóan az IBM SPSS Statistics program is a soros és oszlopos megvalósítást követi. Alap esetben a program felső részén a megszokott menürendszer található, míg alatta a leggyakrabban használt funkciók ikonos formában. Az SPSS program központi részét a korábban említett táblázat, illetve a változók átfogó nézete tölti ki (lásd 1.1. ábra). A mérési adatsor nézetben az oszlopokban találhatók a változók, míg a sorokban az esetek.

1.1.1. Változótípusok

A valószínűségi változók típusa fontos a végrehajtandó eljárás szempontjából, és az előzetes adatkezelést is befolyásolja. Alapvetően két típust különböztetünk meg:

1. a *diszkrét valószínűségi változó* által felvehető értékek száma véges (vagy megszámlálhatóan végtelen, mint pl. az egész számok halmaza), vagy

2. a *folytonos valószínűségi változó*: amely a valós számok halmazának egy vagy több intervallumán bármely értéket felvehet. Más szóval adott



1.1. ábra. Az SPSS program nyitóablaka.

határok között bármely valós értéket felvehet (ilyen például a valós számok halmaza 0 és 1 között).

Az előbbire példa a megfelelt – nem felelt meg – kiválóan megfelelt minősítés, illetve a kék – zöld – piros színhármas. Diszkrét valószínűségi változó például az is, ami egy kockadobás eredményét írja le. Az utóbbi csoportba tartozik a testmagasság, a termésátlag vagy az autók fogyasztása.

A diszkrét változókon belül van a *bináris* vagy *dichotom változó*k (alternatív ismérvek) csoportja: ezek csak két értéket vehetnek fel (pl. igen – nem, vagy férfi – nő).

1.1.2. Adattípusok

Az adatokat először is a jelentésük jellege alapján lehet osztályozni: eszerint az adat *kvalitatív* vagy *kvantitatív* lehet. A kvalitatív (vagy minőségi) adattípus az objektumok fajtáit adja meg (pl. neme: férfi – nő). A kvantitatív (vagy mennyiségi) adattípus a számmal kifejezhető jellemzőket mutatja (pl. életkor, jövedelem).

Az adatok ilyen osztályozása általában természetes, könnyen megadható,

mégis, ha az adatokat számokkal kódoljuk, akkor ezek a típusok csak a jellemzők eredeti jelentése alapján határozhatók meg. Így egy 100-as adatérték lehet mérési eredmény (tehát kvantitatív típusú), de például színkód is, ami pedig kvalitatív adattípusnak felel meg. Azaz csak az adatokat látva nem dönthető el egyértelműen az adatsor típusa, melyek meghatározása a statisztikai elemzések előtt fontosak és nagy figyelemmel járjunk el.

1.1.3. Mérési skálák

A mérési skálák (vagy mérési szintek) részletesebb osztályozást adnak. Ezek mondják meg, hogy az adatainkat pontosan hogyan szabad értelmezni, milyen összefüggéseket használhatnak a statisztikai eljárások. Ennek megadása döntően befolyásolhatja az eredményünket, és emiatt ez komoly hibalehetőséget is jelent.

Az alkalmazandó mérési skálát a statisztikai program nem tudja maga kiválasztani, mindenképpen a felhasználó, illetve a program kezelője segítségére lesz szükség. Ezért ennek az osztályozásnak a megfelelő ismerete elengedhetetlen a statisztikai programok megbízható használatához. Bár erről megkérdezhetjük a végső felhasználót, vagy kideríthetjük a szűkebb szakmában szokásos, elfogadott osztályozást, de ezt magunk is tisztázhatjuk. Másrészt számos később ismertetendő részletkérdésben mindenképp a szakterület elfogadott módszertanára kell támaszkodnunk, így ebben az esetben törekedni kell az önálló döntésre.

Legyen A és B két objektum, X egy változó, x_A és x_B pedig az X változó értékei A és B esetén. A következő skálatípusokat tárgyaljuk (amelyek ebben a sorrendben tartalmazzák egymást):

1. A *névleges (vagy nominális) skála* minden értéke egy önálló kategóriát jelöl, az objektumok között csak az azonosság vagy különbözőség viszonyát tételezi fel (pl. a nem, szín, születési hely). A-ról és B-ről csak annyit tudunk, hogy $x_A = x_B$ vagy $x_A \neq x_B$. Ez a legkevésbé informatív mérési skála.

   Ennek esetében tehát hiába kódoltuk az adatokat számokkal, azokkal a szokásos műveleteket nincs értelme elvégezni, hiszen az eredeti információ-tartalom azt nem engedi meg (két színnek nincs pl. sorrendje). Ennek megfelelően az adatunkra vonatkozóan mindig a legtöbb információt nyújtó érvényes mérési skálát kell megadni.

2. A *sorrendi (vagy ordinális, ill. rang-) skála* esetén az objektumok között az azonosságon kívül nagyságrendi, illetve sorrendi különbséget

is megállapíthatunk (például jó – közepes – rossz, magas – alacsony). A-ról és B-ről mondhatjuk, hogy $x_A < x_B$ vagy $x_A = x_B$ vagy $x_A > x_B$.

A statisztikai programok gyakran támogatják ezt a mérési skálát, és a rá vonatkozó eljárások természetesen eltérnek a többi mérési skálán mért változókra írtaktól.

3. Ha az adatainkat *intervallum (vagy különbségi) skálán* mérhetjük, akkor a különbségek mértékét is értelmezhetjük (például a hőmérséklet, a dátum). Ha $x_A > x_B$, akkor B az A-tól $x_A - x_B$ egységgel különbözik.

   Ez a skálatípus már a legtöbb magasszintű statisztikai eljárást megengedi, ebben az értelemben ennek megléte már nem nagyon korlátozza a végrehajtható algoritmusok körét.

4. Az *arányskálán* az előbbieken túl még értelmezhető kezdőpont is van, tehát két objektum között nemcsak a különbséget, hanem az arányt is megállapíthatjuk (pl. a sorszámok, a fizetés, az életkor). Ha $x_A > x_B$, akkor az A objektum adott szempontból x_A/x_B-szer nagyobb, mint B. Az arányskálát nevezhetjük a legmagasabb mérési szintnek.

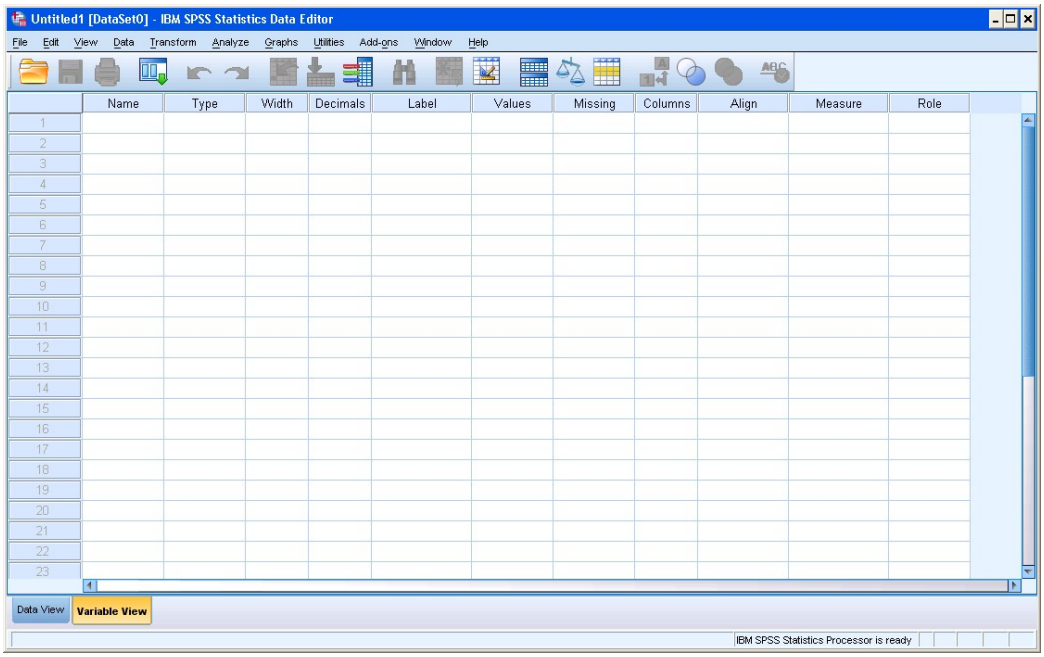
   Ismét meg kell jegyezni, hogy a számmal való kódolás miatt természetesen minden esetben van ugyan kezdőpont (hiszen kódolásra használt valós számok arányskálának felelnek meg), de a lényeges kérdés, hogy a mért mennyiségre értelmezhető-e ez, illetve hogy annak kitüntetett szerepe van-e a feldolgozás szempontjából.

Az utóbbi két mérési skálát együttesen metrikus skálának szokás nevezni. A minőségi ismérvek többnyire névleges skálán mértek (de nem mindig), a mennyiségi ismérvek pedig általában ennél erősebb mérési skálához tartoznak. A statisztikai programok nem minden mérési skála megadását teszik lehetővé, például az SPSS a scale, ordinal és nominal lehetőségeket adja meg (az első az arány- és intervallum skálát is fedi).

Az SPSS változókra vonatkozó beállításait a Variable View fülön tudjuk megtenni (lásd 1.2. ábra). A korábbiak alapján a változók az alábbi mérési skálákat vehetik fel SPSS-ben: Scale (arányskálán), Ordinal (sorrendi) és Nominal (névleges).

1.1.4. A minta jellemzői

A statisztikai feldolgozás adatait más szempontból is lehet jellemezni. A *sokaság vagy populáció* a statisztikai vizsgálat egyedeinek összessége



1.2. ábra. Az SPSS Variable View füle.

(halmaza). Ennek minden elemre kiterjedő teljeskörű vizsgálatát nem mindig lehet, vagy nem gazdaságos elvégezni. Ilyen statisztikai sokaság például a szavazati joggal rendelkezők köre, vagy egy más feldolgozásban egy gyógyszerkísérletben az emberiség.

A *statisztikai minta* ezzel szemben a vizsgált sokaságból kiválasztott egyedekhez tartozó megfigyelési adatok halmaza, részsokasága. Mintavételnél fontos szempont a *reprezentativitás* (azaz a kiválasztott mintának jól kell reprezentálnia a vizsgálni kívánt sokaságot az adott vizsgálatok szempontjából), és a *függetlenség*. Ugyanazon egyed többszöri mérése nem független adatokat eredményez – a minta elemszámát így nem szabad növelni. De más mintavételezés is okozhat nem független adatsort. Például egy család lakhelyéről, anyagi helyzetéről, stb. készülő adatsor a családtagok esetén nem független. *Cenzorált minta* az, amikor az eredeti minta elemeinek csak egy részét használjuk fel a következtetések levonásához. A szavazati joggal rendelkezők mintája lehet például 1000 ember, akit véletlenül választunk ki és kérdezünk meg a szavazással kapcsolatban. Míg a gyógyszerkísérletben minta a gyógyszerkísérletben résztvevő egyének.

Az alábbiakban röviden áttekintjük a minta legfontosabb jellemzőit, egyszerű definíciókkal, illetve ahol kell, rövid magyarázattal. A minta egy-

szerű jellemzői elsősorban a statisztikai feldolgozás első fázisában hasznosak, amikor a feldolgozandó adatok helyességét kell megállapítani. Ehhez nagy segítséget adnak a mért mennyiségek várt mutatói és a ténylegesen feldolgozott számokra adódó mutatók esetleges eltérései. Ez persze inkább nagyobb adatmennyiség esetén jelentős, kevés adatot könnyen össze lehet vetni akár teljes egészében is. Erre nagy adathalmaz esetén nincs reális lehetőség.

A minta eloszlásának (a folytonos változó értékei elhelyezkedésének) megjelenítésére általában *hisztogram*ot használunk. Utóbbi előállításához a legkisebb és a legnagyobb mintaelem közti különbséget valahány (általában 5-nél több) intervallumra osztjuk. Ezután készítünk egy ábrát, amelyben az intervallumokra olyan magas téglalapokat rajzolunk, mint ahány megfigyelés abba az intervallumba esik. Minél több a mintaelem, és minél több az intervallumok száma, a hisztogram annál jobban megközelíti az elméleti eloszlást. Ha ez az elméleti eloszlás a harang-görbe (Gauss-görbe), akkor azt mondjuk, hogy a minta *normális eloszlású* populációból származik.

Az *elemszám* a statisztikai mérések száma (az esetek száma). A *hiányzóadat kódok* olyan értékek, amik az illető változó lehetséges, értelmes értékei között nem fordulnak elő, de annak legszűkebb ábrázolásába belefférnek. Például a cipőméretek hiányzóadat kódja lehet a 99. A hiányzóadat kód figyelembevételével szokás külön megadni az *érvényes esetek számá*t is. Az adatgyűjtés során nyilván üresen maradhat a hiányzó adatok helye, de a számítógépes bevitel, tárolás során nem helyes, ha a véletlenre bízzuk, hogy milyen szám rendelődik a szóköz(ök)höz. Ha a hiányzóadat kódok elkülönített kezelését nem oldjuk meg, akkor olyan hibák adódhatnak, hogy például egy átlagba 0 értékkel beleszámítódik a hiányzó érték (0) is, és így irreális, a valós helyzetet nem tükröző eredményt kapunk.

Középértékek

Egy változó *középértéke* a gyakorisági eloszlás helyzetét tömören, egy számmal kifejező érték, azonos mértékegységű adatok olyan jellemzője, amelytől azt várjuk, hogy közepes helyzetű, könnyen meghatározható és értelmezhető legyen. A középérték mértékegysége megegyezik a jellemzett változóéval. Ide tartoznak a helyzeti középértékek: a módusz és a medián, valamint a számított középértékek vagy átlagok, mint pl. a számtani átlag.

A valószínűségi változónak is létezik átlaga, szórása, stb. A tapasztalati úton szerzett adatsorok azonos jellemzőit empírikus középérték, empirikus szórásnak, stb. hívjuk.

Az *átlag*, vagy *számtani átlag* egy adott mennyiségi, metrikus változó

értékei összege osztva az elemszámmal:

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i)}{n}.$$

A mintaelemeket nagyság szerint rendezve a középső elem (páratlan számú elem esetén), vagy a két középső elem átlaga (páros számú elem esetén) a *medián* (rövidítése Me). Ebben az értelemben ez a minta közepe. Más szóval az a szám, aminél a mintaelemek 50%-a kisebb vagy egyenlő.

A *módusz* a leggyakrabban előforduló érték(ek). A később ismertetendő normális eloszlás esetén az átlag, a módusz és a medián egybeesik.

Ha egy y változó értékei 5, 2, 3, 4, 4 és 1, akkor az ezekre vonatkozó átlag $3,1\dot{6}$ a módusz 4, a medián pedig 3,5.

Az eloszlás jellemzői

Itt a statisztikai változók, vagy más szóval ismérvek további jellemzőit ismertetjük röviden. A *szórás* (angol rövidítése SD a standard deviation-ből) a minta szórása, azaz a minta elemeinek az átlagtól való eltérésének négyzetes átlaga. Normális eloszlás esetén az átlag $\pm$ 2 $*$ SD intervallumban található a mintaelemek 95,45%-a. A szórás (elméleti) és a korrigált tapasztalati szórás:

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

ahol x_i az i-edik értéke az X valószínűségi változónak, és $\bar{x}$ a mintaelemek átlaga – inkább csak arra az esetre, ha számítógép nélkül kellene meghatározni. Bár a szórást a legtöbb kalkulátor közvetlenül is meg tudja adni. Az előbb megadott Y változó szórása $\sigma = 1,3437$, illetve $s = 1,4720$. A szórás négyzete, a *szórásnégyzet* is gyakran használt mutató, neve a variancia.

Becsléskor a becslőfüggvény szórását az illető becslés *standard hibájá*nak (SE vagy SEM) nevezzük. Átlag esetén ez az átlag szórása. Ez azt fejezi ki, hogy az adott részminta alapján kapott átlag mennyire jól közelíti a valódi populáció átlagot. Az átlag $\pm$ 2 $*$ SE jelenti azt az intervallumot, amelyben a populáció átlaga kb. 95% valószínűséggel benne van.

A *relatív szórás* = (SD/átlag)$*$100. Megadja százalékos értelemben (mértékegység nélkül), hogy a szórás hányszorosa az átlagnak. Relatív jellege miatt alkalmas a különböző nagyságrendű változók szórásának összehasonlítására.

A *percentil* vagy *percentilis* a mediánhoz hasonló mutató a minta jellemzésére. A P_{75} 75%-os percentil pl. az a szám, aminél a mintaelemek 75%-a

kisebb (vagy egyenlő). Az 5, 2, 0, 3, 1, 4, 6, 8, számokra P_{75} értéke 5, mert ennél nem nagyobb számból pont 6 van ($8 \times 0,75$). Ha az értékkészletet nem száz, hanem 4 részre osztjuk, akkor kvartilisről (Q_i), ha tízre, akkor decilisről (D_i) beszélünk. Az ilyen mutatók összefoglaló neve a *kvantilis*.

A mennyiségi jellegű minta *terjedelme* a legnagyobb és a legkisebb mintaelem közötti különbség:

$$d = max_{i=1...n}(x_i) - min_{i=1...n}(x_i)$$

Hasznos lehet a hibásan bevitt adatok kiderítéséhez. Az előző bekezdésben említett minta terjedelme 8 - 0 = 8.

A *ferdeség*, vagy *ferdeségi együttható, aszimmetria* egy mérőszám arra, hogy az eloszlás szimmetrikus-e vagy ferde. Negatív ferdeségi együttható esetén bal oldali (negatív) ferdeségről van szó, ekkor az átlagnál nagyobb értékek a gyakoribbak.

A *lapultság* (kurtózitás) is az eloszlás egy alaki tulajdonságát fejezi ki: ha ez a mutató pozitív, az azt jelenti, hogy az eloszlás a normális eloszláshoz képest csúcsosabb, negatív esetben pedig lapultabb. Ennek megfelelően szokásos a *csúcsosság* név is.

1.1.5. Eloszlások

A valószínűségi változók *eloszlásfüggvénye* azt mutatja meg, hogy ezek a változók milyen valószínűséggel vesznek fel egy adott számnál kisebb értéket: $F(x) = P(X < x)$, ahol P a $X < x$ esemény valószínűsége. Az $F(x)$ abszolút folytonos eloszlásfüggvény deriváltja $f(x)$, az ún. *sűrűségfüggvény*. Diszkrét eloszlású valószínűségi változóknak nincs sűrűségfüggvénye.

A sűrűségfüggvény ismeretében több, a valószínűségi változóval kapcsolatos esemény valószínűsége megadható. Például, annak a valószínűsége, hogy a valószínűségi változó egy adott intervallumba esik, az alábbi képlettel kapható meg:

$$P(a \leq X < b) = \int_b^a f(x)dx.$$

A sűrűségfüggvénnyel adott változók *várható értéké*t az

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

képlettel (ahol $f(x)$ a megfelelő sűrűségfüggvény), a diszkrét változókét a súlyozott középpel definiáljuk: $\sum_{i=1}^{n} x_i P(X = x_i)$. Az E betű az angol expectation szóra utal.

A következőben a leggyakrabban előforduló, illetve a statisztikai feldolgozáshoz leginkább használatos eloszlásokat mutatjuk be röviden.

Binomiális eloszlás

Tekintsünk egy olyan kísérletet, amelynek két kimenetele van, A és B, és amelyek valószínűségei p és $q = 1 - p$. Ekkor annak a valószínűsége, hogy n számú független kísérletből az A lehetőség pontosan k-szor következik be, $P_k = \binom{n}{k} p^k q^{n-k}$. A P_k valószínűségek n-edrendű p paraméterű *binomiális eloszlást* határoznak meg. A binomiális változó várható értéke np, szórásnégyzete npq.

Ilyen eloszlást mutat például az alábbi esemény: Egy 20 fős osztályról tudjuk, hogy 15-en folyamatosan készülnek. Ha a tanár a hónap során 10 főt feleltet, várhatóan hány szép feleletet fog hallani ebben a hónapban? (Természetesen egy tanuló többször is felelhet a hónap során.) A valószínűségi kísérlet a feleltetés. Az A esemény: „a felelő tudja az anyagot". Ekkor: $N = 20$, $M = 15$, $n = 10$. Az A valószínűsége $P(A) = p = 15/20 = 0,75$. A komplementer esemény – a felelő nem tudja az anyagot – valószínűsége $1 - p = 0,25$. A valószínűségi változó lehetséges értékei azt mutatják, hogy a 10 feleletből mennyi volt jó: $k = 0, 1, 2, \ldots, 10$.

Poisson-eloszlás

A ξ diszkrét valószínűségi változót λ $(0 < \lambda < \infty)$ paraméterű *Poisson-eloszlás*únak nevezzük, ha lehetséges értékei a nemnegatív egész számok, és

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-k}$$

teljesül $(k = 0, 1, 2, \ldots)$. Várható értéke és szórásnégyzete is λ.

A binomiális eloszlás határeseteként lehet megkapni a kísérletek számának (n) növelésével és a p csökkentésével úgy, hogy az $np = \lambda$ szorzat állandó maradjon. Pontok térbeli vagy időbeli véletlen elhelyezkedése akkor követ Poisson-eloszlást, ha azok egymástól függetlenül minden térrészben vagy időszakaszban egyforma valószínűséggel oszlanak meg.

Ilyen eloszlást mutatnak például az alábbi események: a vérsejtek száma a mikroszkóp látómezejében, vagy a radioaktív anyag adott idő alatt elbomlott atomjainak a száma.

Egyenletes eloszlás

Egyenletes eloszlás lényegében azt fejezi ki, hogy a szóba jöhető alternatívák egyforma valószínűségűek. Diszkrét esetben, amikor a változó csak véges számú értéket vehet fel, ezek mindegyike egyenlő valószínűségű

(mint például a kockadobás). Folytonos esetben akkor beszélünk *egyenletes eloszlásról*, ha a változónak egy adott szakaszra, tartományra esésének a valószínűsége arányos a szakasz hosszával, illetve a tartomány mértékével. Az egyenletes eloszlású ξ diszkrét változó várható értéke $\frac{1}{n}\sum_{i=1}^{n} x_i$, és szórásnégyzete $\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$, amennyiben a felvehető értékei $x_1, x_2, \ldots, x_n$.

Erre az egyik legtipikusabb példa a hagyományos hatoldalú kockával dobott érték. A valószínűségi változó lehetséges értékei azt mutatják, hogy hányast dobtunk.

Normális eloszlás

Egy valószínűségi változó *normális eloszlás*ú (jelölése $N(\mu, \sigma)$), ha az eloszlásfüggvénye

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt.$$

A binomiális eloszlás határeseteként is előáll a normális eloszlás, ha n növekedése közben p állandó marad. A képletében szereplő két paraméter a várható érték (μ) és a szórás (σ). A μ az eloszlás várható értéke, mediánja és módusza is egyben.

Független valószínűségi változók összegének az eloszlása közelítően normális eloszlású, ez biztosítja gyakori előfordulását. Hasonló okból, ha csak egyenletes eloszlású pszeudovéletlen-szám generátor áll rendelkezésre, akkor pl. n darab ($n > 10$) ilyen véletlen szám összege közelítőleg normális eloszlású véletlen számot ad.

Ezen ok miatt a természetben nagyon gyakran találkozunk normális eloszlásokkal, például fák várható magassága, vagy terméshozam nagysága.

A standard normális eloszlás a 0 várható értékű, 1 szórású normális eloszlás ($N(0, 1)$).

Khi-négyzet eloszlás

A $\xi_1, \xi_2, \ldots, \xi_n$ független, standard normális eloszlású változók négyzetei összegének eloszlása n szabadságfokú *khi-négyzet (χ^2) eloszlás*. Ennek a várható értéke n, a szórásnégyzete pedig $2n$. Az előző szakaszban elmondottak miatt nagy n szabadságfok esetén alig tér el a normális eloszlástól.

1.1.6. Az eloszlásokkal kapcsolatos alapfogalmak

Paraméter (vagy az eloszlás paramétere) az eloszlásfüggvényt meghatározó képletben szereplő valamely változó. Például a normális eloszlás paraméterei a várható érték (μ) és a szórás (σ).

Paraméteres módszer: olyan matematikai statisztikai módszerek összefoglaló neve, melyek paraméterrel vagy paraméterekkel (véges sok) leírható sokaságokra alkalmazhatók. Ebből adódóan nyilván vannak *nemparaméteres statisztikai eljárás*ok is, amelyek tehát nem a véges sok paraméterrel megadható eloszlásokon alapulnak. Hasonlóan a *paraméteres próba* a hipotézisvizsgálatnál az előírt parametrikus eloszlású sokaság valamelyik paraméterére vonatkozó próba.

Statisztikai becslés: a populáció eloszlásának valamely ismeretlen paraméterét egy alkalmas minta alapján közelítjük. A minta elemeit egy megfelelő formulába helyettesítve közelíthetjük a paraméter igazi értékét (pl. a populáció „elméleti" átlagát a mintaelemekből szokásos módon számolt átlaggal közelítjük).

Egy statisztika *szabadságfokát*, úgy definiáljuk, hogy az N mintaszámból levonjuk az adott statisztika kiszámításhoz szükséges, az adatokból már meghatározott paraméterek számát. Például az n számú minta adatból számított számtani átlag szabadságfoka n, mivel az átlag kiszámításához csak a minta adatokat használjuk fel, a képletben nincs olyan paraméter, amit az adatokból számolnánk ki.

Megbízhatósági intervallum (vagy *konfidencia intervallum, megbízhatósági tartomány):* olyan intervallum, amely (általában) nagy, előre megadott valószínűséggel tartalmazza a becsült paraméter valódi értékét.

1.2. Statisztikai próbák

Ez a szakasz a statisztikai próba felállításához és az eredmény kiértékeléséhez ad segítséget, összefoglalva a legfontosabb fogalmakat. A szokásos, gyakori hipotézisvizsgálatokat a statisztikai programok közvetlenül támogatják. A *statisztikai próba* olyan eljárás, amely valamilyen hipotézisnek (az alapsokaságra vonatkozó feltevésnek) az ellenőrzését teszi lehetővé a minta adatai és a próbafüggvény alapján.

A nullhipotézis: hipotézisvizsgálatban általában az a feltevés, hogy bizonyos különbségek vagy hatások a populációban adott értékkel egyenlők. Például, hogy két átlag különbsége 0, vagy az, hogy a korrelációs együttható nulla. De lehet az is a kiindulási feltevésünk, hogy pl. a várható érték 10.

Szignifikancia, szignifikáns eltérés: a nullhipotézistől való, adott valószínűségi szintet meghaladó eltérés. A *szignifikancia-szintet* általában valószínűséggel adjuk meg. Ez lehet pl. 5%, azaz $\alpha = 0,05$ annak a hibának a valószínűsége, hogy tévesen állapítottuk meg a különbséget, ha a nullhipotézis igaz (ez a maximális első fajú hibalehetőség, amit még hajlandóak vagyunk tolerálni). Ha tehát a próba eredménye $p < 0,05$, akkor ez azt jelenti, hogy szignifikáns különbséget vagy hatást állapítottunk meg. Ha százszor megismételnénk a kísérletet, a százból csak kb. 95 esetben kapnánk ugyanezt az eredményt, 5 esetben nem találnánk eltérést (elsőfajú hiba). A szokásos szintek: 5%, 1%, 0,1% (azaz $\alpha = 0,05$, 0,01, 0,001). A *megbízhatósági szint*ek ennek megfelelően 95%, 99% és 99,1%. A szignifikáns eredményt leggyakrabban a p-érték és a szignifikancia-szint (α) összehasonlításával szokás megállapítani. Egyre elterjedtebb, magának a p értéknek a megadása.

Nem szignifikáns: $p > 0,05$ (p nagyobb, mint 0,05). Az 5%-os szinten nem szignifikáns különbség azt jelenti, hogy nem sikerült a különbséget kimutatni. Ez nem feltétlenül jelenti azt, hogy egyáltalán nincs különbség. Ha az eredmény nem szignifikáns, akkor lényegében semmit sem tudunk mondani a vizsgált jelenségről. Ebben az értelemben végül is elfogadhatjuk a nullhipotézist, miszerint nincs eltérés, miközben a nullhipotézis nem igaz. Ekkor az elkövetett hibáról (másodfajú hiba) csak annyit tudunk, hogy nagy mintaelemszám esetén elég kicsi.

1.2.1. A statisztikai próbákkal kapcsolatos további alapfogalmak

Az *elsőfajú hiba* akkor fordul elő, amikor a nullhipotézist elvetjük, holott az igaz. Valószínűsége egyenlő a az általunk megválasztott szignifikancia-szinttel (α).

A *másodfajú hibá*t akkor követjük el, amikor a nullhipotézist elfogadjuk, bár az nem igaz. Valószínűségét (β) nem ismerjük. Ha az elsőfajú hiba valószínűségét csökkentjük, a másodfajú hibáé nő, de $\alpha + \beta \neq 1$. Nagy mintaelemszám esetén általában a másodfajú hiba valószínűsége csökken.

Egyoldali próba amikor a nullhipotézissel szemben felállított *alternatív hipotézis*ben (ellenhipotézisben) csak egyirányú változást tételezünk fel.

Kétoldali próba: ekkor a nullhipotézissel szemben felállított alternatív hipotézisben minden irányú változást figyelembe veszünk.

1.2.2. Statisztikai próba végrehajtása

A statisztikai próbák végrehajtásának a következő lépései vannak:

1. Az előzetes ismereteink alapján állítunk valamit, amit statisztikai módszerrel szeretnénk igazolni. Először a kiinduló hipotézist (H_0) kell felállítani, a nullhipotézist megfogalmazni. A nullhipotézisben sok esetben (de nem mindig) azt rögzítjük, hogy nincs változás.

2. Ezután az alternatív hipotézis (H_1) felállítása következik, amely általában a nullhipotézis tagadása, de nem feltétlenül az. A nullhipotézis és alternatív hipotézis közül csak az egyik eset állhat fenn, másként megfogalmazva a két hipotézis egyszerre nem állhat fenn

3. A következő lépés a próba szignifikancia-szintjének meghatározása ($\alpha = 0,05$, $\alpha = 0,01$, vagy $\alpha = 0,001$). Ezt az értéket az adott szakterület szokásos értékeihez kell igazítani.

4. Határozzuk meg ezután a használt véletlen minta elemszámát. Ezt idő-, illetve pénzkorlátok és előzetes ismereteink is meghatározzák, különben nyilván a nagyobb minta megbízhatóbb eredményt adhat. Ezután jön a véletlen minta előállítása, és a próbastatisztika kiszámítása.

5. Meghatározzuk a döntési szabályt, és azt a kritikus értéket vagy értékeket (ha kétoldali próbát hajtunk végre), amelynél a mintából kiszámított próbastatisztika csak kis ($< \alpha$) valószínűséggel vesz fel nagyobb értéket.

6. Ha a kiszámított próbastatisztika a kritikus értéknél nagyobb (illetve az elfogadási tartományon kívül esik), akkor elvetjük a nullhipotézist, mivel egy kis valószínűségű esemény következett be (egyúttal elfogadjuk az alternatív hipotézist). Ilyenkor azt mondjuk, hogy az eltérés szignifikáns az α szinten ($p < \alpha$), az alternatív hipotézis teljesül.

7. Ha a kiszámított próbastatisztika a kritikus értéknél kisebb (illetve az elfogadási tartományon belül van), akkor megtartjuk a nullhipotézist és azt mondjuk, hogy az eltérés nem szignifikáns α szinten. Azt is mondhatjuk, hogy nem vetjük el a nullhipotézist, ami egy óvatos megfogalmazás, és arra utal, hogy a szignifikancia-szint függvényében általában nem állíthatjuk, hogy a nullhipotézis igaz.

Az itt megadott szempontok és útmutatások új statisztikai próbák összeállításához és végrehajtásához adnak segítséget. Másrészt a leggyakoribb ilyen teszteket a tárgyalt statisztikai programok közvetlenül is támogatják, vagyis ekkor inkább csak az eredmények helyes értelmezéséhez, vagy a jó paraméterezéshez használhatjuk ezeket az ismereteket.

1.2.3. Változók összefüggése

A *korreláció*s eljárások két valószínűségi változó közötti összefüggés szorosságát mérik, ami aztán a predikció minősége mértékeként is használható. Itt nem szükséges az egyik változó kijelölése, azok egyenrangúak a statisztika szempontjából. Az r korrelációs együttható egy -1 és 1 között változó szám. Ha ennek értéke -1, akkor függvényszerű negatív lineáris összefüggés van a változók között, azaz amíg az egyik nő, addig a másik csökken. Ha a korrelációs együttható 1, akkor függvényszerű pozitív lineáris összefüggés van. A nulla korrelációs együttható pedig azt jelenti, hogy nincs lineáris összefüggés a változók között. Más érték esetén óvatos diszkusszió mellett a közelálló említett eseteknek megfelelő következtetést vonhatjuk le.

A fenti mennyiségi változók közötti mérőszám. De létezhet összefüggés minőségi-minőségi és minőségi-mennyiségi változó típusok között. Az előbbi esetében *asszociációs kapcsolat*ról, míg az utóbbi esetben *vegyes kapcsolat*ról beszélünk. Például asszociációs kapcsolat lehet nem - beosztás; nem - vezetési stílus és iskolai végzettség - beosztás között. Vegyes például a nem - kereset; beosztás - életkor, míg korrelációs kapcsolat életkor - kereset, jövedelem - fogyasztás és tanulási idő - vizsgajegy között lehet.

A *regresszió*s eljárás feltételezi, hogy olyan összefüggés van a magyarázóváltozók és az eredményváltozó között, hogy ha az adatokat térben ábrázoljuk, akkor egyenest, síkot, vagy adott típusú görbét kapunk megközelítőleg. A regresszió azt a paraméterezést keresi meg, amely a legjobb illesztést adja az aktuális adathoz. A többváltozós lineáris esetben a magyarázóváltozók (nyilván többváltozós) lineáris függvényével modellezzük az eredményváltozó értékét. A regresszió egy paraméteres statisztikai módszer, amely feltételezi, hogy a reziduumok (a becsült és a tényleges eredményváltozó értékek közti eltérések) normális eloszlásúak. Mivel a regressziós együtthatók kiszámításakor a reziduumok négyzetösszegét minimalizáljuk, ezért szokás ezt az eljárást a legkisebb négyzetek módszerének is hívni.

2. fejezet

Az SPSS programcsomag

Ahogy a bevezetőben is olvasható volt, az SPSS teljes körű statisztikai eljárásokat kínáló program. A későbbiek megértéséhez sokat segít, ha tudjuk, hogy egyrészt ezt a programot nagy mennyiségű adat kezelésére tervezték, másrészt pedig azt, hogy eredetileg nagy számítógépen, kötegelt (batch) módban futtatott programok gyűjteménye volt, ezt írták át előbb DOS, majd Windows operációs rendszer alá.

Az első szempont azért fontos, mert emiatt, ahogy látni fogjuk, a program nem a kézzel, egyedileg beállított paraméterezésre készült, hanem a tömeges, programozásszerű értékadásra. Ez megmutatkozik már ott is, hogy a különböző statisztikai eljárások számára a feldolgozandó eseteket, változókat kijelöléssel, halmazként lehet megadni, és nem például egyenként begépelni a nevüket. A nagy mennyiségű adat feldolgozására való felkészülést jellemzi az is, hogy az egyszer már kialakult statisztikai eljárássort programozáshoz hasonló módon (az ún. *.sps parancsállománnyal, vagy syntaxfile-al) lehet megismételtetni más adatokra, változókra is.

Ez utóbbi lehetőség a korábbi kötegelt futtatásra is utal. Ennek egy másik jele, hogy más, komolyabb statisztikai programcsomagoktól eltérően az SPSS csak a legfontosabb eljárásokat tudja azonnal végrehajtani, azok többségét előbb be kell töltenie (mint korábban a nagygépes rendszerekben). Ennek a moduláris szerkezetnek számos előnye van.

Az elmondottak ellenére a jelen tárgy oktatása során persze mindig kisebb adathalmazokkal dolgozunk majd, tehát kényelmi szempontból az SPSS említett két alaptulajdonsága inkább hátrányosnak tűnik majd. Ilyenkor gondoljunk mindig arra, hogy valódi gyakorlati feladatok megoldása során ezek a tulajdonságok inkább előnyösek.

A jegyzet ezt a programot ismerteti, mert ezzel a statisztikai programcsomagok legjellegzetesebb tulajdonságai jól bemutathatók. A ta-

nulmányozás nem térhet ki minden részletre a rendelkezésre álló eljárások nagy száma miatt. A bemutatás követi a tipikus statisztikai feldolgozás legfontosabb lépéseit, és kitérünk a legfontosabb, illetve legérdekesebb statisztikai eljárások használati módjára. Akit más eljárás is érdekel, illetve akinek valamilyen itt nem tárgyalt módszerre van szüksége, az tanulmányozza a irodalomjegyzékben található szakirodalmat.

2.1. Alapvető adatkezelési eljárások

2.1.1. Az adatok bevitele

Az ismertetés során a program angol nyelvű változatát vesszük. A magyar nyelvű operációs rendszernél, illetve beállításokkal egyes párbeszédes ablakok, vagy más nyelvi elemek, mint pl. a tizedesvessző használata magyarul történhet.

Az indulás után más statisztikai vagy táblázatkezelő programoktól eltérően először is egy párbeszédes ablak jelenik meg (a különben szokásos Windows-os táblázatkezelőszerű munkalap előtt), aminek a kitöltése szükséges a további munkához. Ez az ablak azt kérdezi, hogy mivel szeretnénk kezdeni:

- a programleírás olvasásával (Run the tutorial),
- adatok begépelésével (Type in data),
- egy meglevő adatbázis lekérdezéssel (Run an existing query),
- egy új adatbázis lekérdezés létrehozásával varázsló segítségével (Create new query using Database Wizard),
- egy korábbi SPSS adatállomány betöltésével, melyeket mutat is nekünk a program egy kis ablakban (Open an existing data source), vagy
- egyéb állomány betöltésével (Open another type of file).

Ha nem szeretjük ezt a fajta programindulást, akkor a párbeszédes ablak bal alsó sarkában levő kis ablakba klikkelésel kérhetjük azt, hogy a program ne ezzel a párbeszédes ablakkal induljon a továbbiakban. Tekintsük a lehetőségeinket egyenként.

A programleírás olvasása általában hasznos, mégis most ezt nem ajánljuk azoknak, akik csak a jegyzet által megadott anyagot szeretnék elsajátítani,

mert minden szükséges adat megtalálható a jegyzetben. Másrészt azok, akik a jegyzet anyagán túl, további statisztikai eljárásokat akarnak használni, tanulmányaikhoz vagy kutatómunkájukhoz további részletekre kíváncsiak, azoknak érdemes itt kezdeni a keresést, a felhasználói leírás [7] előtt.

Az adatok begépelése a leggyakoribb adatbeviteli mód lesz a számunkra, mivel a gyakorláshoz mindig elég lesz kisebb adathalmaz is. Ennek ellenére ez nem tipikus gyakorlati feladatok megoldása során, mert az utóbbi esetben a legtöbbször már valamilyen adatállományban vannak a kiinduló adataink. A számokat egyszerűen be kell gépelni, és helyes bevitelük után egy ENTER hatására kerülnek a táblázatkezelőkben szokásos szerkesztőlécből az automatikusan következő vagy kijelölt cellába. Ha egy bizonyos cellába szeretnénk adatot bevinni, akkor először azt ki kell választani.

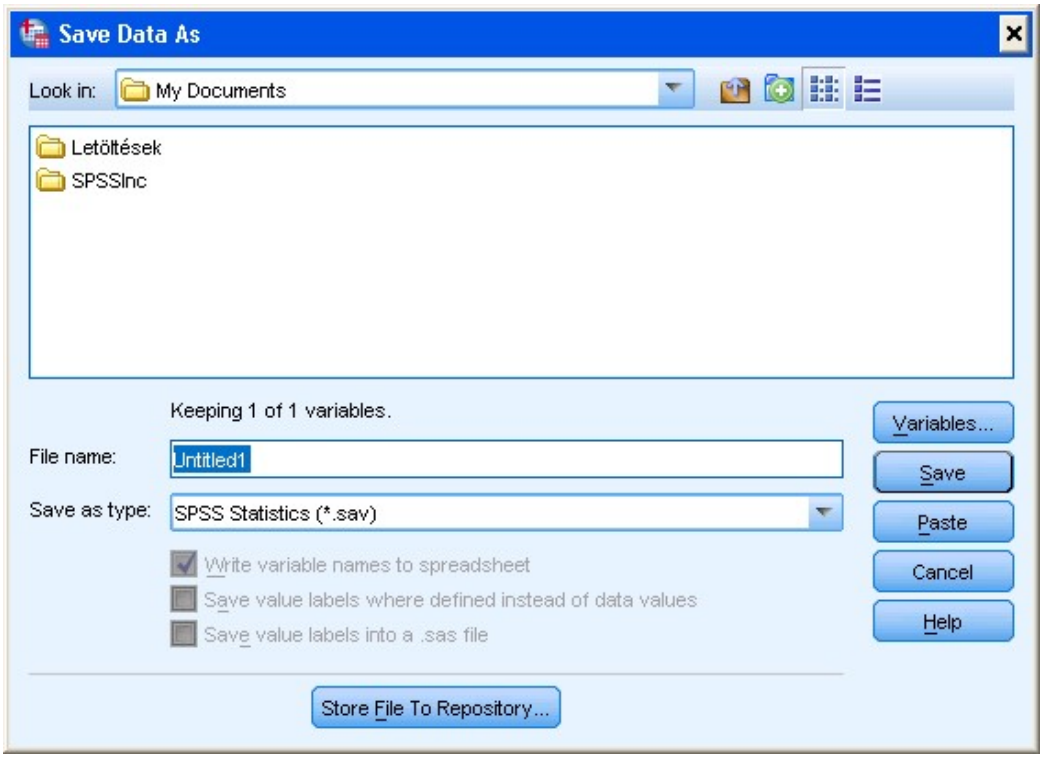
Alaphelyzetben csak számokat tudunk bevinni, és azok is 8 számjegyű, két tizedesjeggyel rendelkező számokként jelennek meg. A szám hosszába a tizedespont (az angol helyesírás szerinti) is beleértendő. Ez a formátum azonban csak a megjelenítésre vonatkozik, a belső ábrázolásban ennél többet is képes a program tárolni.

Bár az adatbevitel nem tűnik különösen fontosnak, a tapasztalat szerint az adatbevitel, az adatok javítása és átalakítása a tényleges statisztikai eljárások számára a tipikus teljes statisztikai feldolgozás idejének kb. harmadát igényli.

2.1.2. Az adatok kimentése

A programmal való ismerkedés első rövid köre az adatok kimentésével zárul. Ezután a későbbi futtatások során használhatjuk majd a korábban bevitt adatainkat. Az adatok kimentése a Windows programokban megszokott módon történik: vagy a floppyt ábrázoló ikonra kell klikkelni, vagy a File menüsorból választjuk ki a Save vagy a Save As parancsokat. Többes adat esetén a Save All Data menüpont használatos. Mindegyik esetben a szokásos kérdéseket teszi fel a párbeszédes ablak a létrejövő adatállomány nevéről, az érintett könyvtárról és a mentett fájl típusáról (egyben kiterjesztéséről). Alaphelyzetben érdemes az SPSS saját fájlformátumát használni (*.sav).

Az SPSS a saját adatformátumán kívül – többek között – a következő egyszerű fájltípusokban tud adatot menteni: fix formátumú (Fixed ASCII) és tabulátorokkal (Tab Delimited) vagy vesszővel (Comma delimited) elválasztott szöveges fájlként. Többféle adatbázis adatállományként is képes menteni, pl. Excel (hagyományos és xml alapú Excel is), Lotus 1-2-3, dBASE, 1-2-3-4, SAS, illetve Stata állományokba. Ha szövegszerkesztővel szeretnénk keresni az adatunkban, vagy annak az átírását így kell elvégeznünk, akkor föltétlen



2.1. ábra. Az SPSS mentés ablaka.

szöveges fájlként érdemes mentenünk, mert ekkor nem kell vezérlőjelekkel bajlódnunk.

Mentésnél a megjelenő ablakon (2.1. ábra) a szokásos módon beállíthatjuk a menteni kívánt állomány helyét és nevét, illetve a megadott típusok közül kiválaszthatjuk az állomány adatformátumát. A Variables gombbal megadhatjuk, hogy mely változók legyenek kiírva az állományba. Több adatformátumba való mentés esetén lehetőségünk van beállítani, hogy mentődjenek az állományba a változók nevei is és azt is, hogy a tényleges értékek helyett a beállított címkék mentődjenek. SAS esetén lehetőség van a címkék mentésére is. A Save gomb hatására mentődik az állomány.

Itt is lehetőség kínálkozik arra, hogy az állományt közvetlenül egy adatbankba helyezzük el. Ezt a Store File To Repository gomb hatására tehetjük meg. Ekkor meg kell adnunk a hely elérhetőségét és a kapcsolat paramétereit.

A jobb oldali gombok közül a Paste is említést érdemel: ezzel tudjuk itt az aktuális beállításra vonatkozó szintaxis állományt (syntax fájl, belső programozható leírás, ami az SPSS program működését meghatározza) megnézni, illetve a meglevőhöz hozzáadni. Ezzel a gombbal több ablaknál is

találkozhatunk.

Hosszabb munka, az adatokon való komolyabb változtatás esetén érdemes időnként elmenteni az adatainkat, esetleg két különböző adathordozóra is. A kimentés után győződjünk meg róla, hogy a File menüsorbeli Open utasítással adatainkat visszakapjuk-e (SPSS adatforma esetén, a többire a következő szakaszban térünk ki). A File menüsor utasításai függenek attól, hogy éppen mivel dolgozunk, tehát az adatállomány, illetve az output ablak tartalmának mentésekor, illetve betöltésekor más-más adatállomány kiterjesztés az alapértelmezés.

2.1.3. Adat beolvasása szöveges állományból

Ez a mód az egyik leggyakoribb adatbeviteli eljárás, szemben a korábban tárgyalt billentyűzetről való bevitellel. Ilyen helyzet adódik például akkor, ha az adatainkat szövegszerkesztő programmal rendeztük, ilyenbe vittük be eredetileg, vagy ha az adatainkat valamely mérési adatgyűjtő program karakteres formában, különösebb bináris vezérlőjelek nélkül írta ki.

A File menüsorból a Read text data utasítást kell kiválasztanunk. Ez egy párbeszédes ablakot ad, amelyben meg kell adnunk azt az állományt (nevet és könyvtárat), amelyből a beolvasást kérjük. A Open lehetőség választása után egy ún. varázsló (Text Import Wizard) segítségével adhatjuk meg részletesen azt a formátumot, amelynek megfelelően a szöveges állományból az adatainkat be kell olvasnunk.

A varázsló 6 párbeszédes lapon kérdez ki bennünket, e lapok között a Next vagy a Back gombokkal tudunk mozogni, az utolsót (miután minden lényeges adatot megadtunk) a Finish gombbal tudjuk elhagyni, és egyben a konkrét beolvasást megkezdeni.

Az első lapon azt kérdezi, hogy van-e már korábban létrehozott formátum a beolvasáshoz (predefined format). Ha már egyszer döntöttünk egy beolvasási formátumról, akkor annak elmentését is kérhetjük majd (az utolsó lapon) a célból, hogy más, hasonló szerkezetű adatállományokból a betöltést megkönnyítsük. Ilyen formátum tehát a legelső alkalommal még nem áll rendelkezésre. Még az első lapon kaphatunk egy mintát a beolvasandó állomány első néhány soráról. Ez a minta a későbbiekben nagyban megkönnyíti majd a paraméterezést.

A második lapon a varázsló azt a lényeges dolgot tisztázza, hogy az adataink rögzített oszlopszerkezetet követnek-e (fixed width), vagy a változó értékeink valamilyen egységes jellel pl. a szóköz, a vessző, az & jel stb. (delimited by a specific character) vannak-e elválasztva. Csak e két eset valamelyike fennállásakor tudjuk a szöveges állományból beolvasni az adatainkat. A

következő kérdés az, hogy a változónevek benne vannak-e az állományban. Ha itt igennel (Yes) felelünk, akkor az első sorból változóneveket fog a program beolvasni, és következésképp az adataink csak a második sortól kezdődhetnek majd.

A harmadik lap azt kérdezi, hogy

- melyik sortól kezdve kell az adatainkat beolvasni (az esetleges változóneveken kívül),
- az egy esethez tartozó adatok hány sorra terjednek ki (ennek minden egyes esetre azonosnak kell lennie),
- hány esetet szeretnénk beolvasni: mindet, az első n esetet, vagy az esetek x százalékát?

Ezek a beállítási lehetőségek nagy rugalmasságot biztosítanak. Ha az eredeti állományunk nem is alkalmas a közvetlen beolvasásra, legtöbbször rövid szövegszerkesztés után megfelelő alakra hozható.

A következő lapon mutatkozik meg igazán az, hogy milyen előnyös a bemutatott minta az adatállományunkból. Itt a második lapon kiválasztott fix, vagy változó szélességű adattárolástól függően különböző nézetet kapunk. Fix esetén az elválasztó pontokat (breakpoints) a program automatikusan kijelöli (az első sor alapján), és a felhasználóra bízza azt, hogy ezt esetleg módosítsa. Ehhez a beállított elválasztó pontokat figyelve átnézhetjük a teljes fájlt, hogy helyes-e a megadott elválasztás. Új elválasztó pontot az egérrel a megfelelő helyre való klikkeléssel lehet adni, a fölöslegeset pedig a bal egérgombbal megfogva ki lehet vinni az adott ablakból. Változó szélességű adattárolás esetén megadhatjuk, hogy az állományban mik választják el az adatokat egymástól. Ez lehet: tabulátor (Tab), vessző (Comma), szóköz (Space), pontosvessző (Semicolon), de megadhatunk egy egyedit is (Other). A jobb oldalt pedig megadhatjuk, hogy a szöveges mezők esetében van-e valami befoglaló jelölő karakter, például kettős idézőjel (Double quote).

Az ötödik lapon az egyes változók adataiból a legfontosabbakat lehet megadni: a változók nevét és típusát (a szélesség és a tizedesjegyek száma kivételével). Az egyes oszlopokat az egérrel való klikkeléssel tudjuk megjelölni. A változók összes jellemzőjének megadására a beolvasás után lesz lehetőség.

Az utolsó lapon a korábban már említett beolvasási formátumot lehet elmenteni (save file format), illetve egy korábbit bővíteni (paste the syntax). A Finish gomb megnyomása után az adatfájlból a változók tartalma beolvasódik, és egy output fájl nyílik meg a végrehajtott művelet hibajelentésével

és részleteivel. Ez utóbbi tipikus lesz a későbbi műveletekre is. Ez a jelenség is az SPSS nagygépes korszakára utal: akkor ezek a fájlok pl. standard kimenetekre íródtak. Az output ablakok tartalmát szerkeszthetjük, elmenthetjük, vagy ezeket az ablakokat a teljes feldolgozás végéig egyszerűen a háttérbe tehetjük. A Finish gomb már az utolsó lap előtt is aktiválható lehet, ha már minden lényeges adatot megkapott a varázsló.

A beolvasás után megmaradt hibás adatokat részben kézzel javíthatjuk, azután a hiba okát kiderítve megismételhetjük a beolvasást a javítás után.

2.1.4. Adat beolvasása adatbázis állományból

Az adatbevitel valószínűleg leggyakoribb formája az adatbázisból történő beolvasás. Ennek során egy táblázatkezelő programmal létrehozott adatsorból tudjuk a statisztikai feldolgozáshoz szükséges változókat az SPSS állományba bevinni. A beolvasást az SPSS File menüsorának Open Database parancsa végzi.

Három választási lehetőséget kapunk: új formátumú bevitel (New query), korábbi formátum szerkesztése (Edit query), vagy egy korábbi formátum szerinti beolvasás (Run query). Először természetesen az új formátumú bevitellel kell kezdenünk. A varázsló több párbeszédes lapon át tisztázza, hogy milyen formában kell az adatokat beolvasnia. Ezek a lapok között ismét a Next, illetve a Back gombokkal lehet mozogni, és a Finish gomb megnyomásával tudjuk indítani a beolvasást.

Az első párbeszédes ablakban azt adhatjuk meg, hogy milyen adatbázis program formátumát kell követni. A választási lehetőségeink dBASE, Excel és MSAccess. Ugyanezen a lapon lehet ezek körét bővíteni is (ODBC kapcsolattal rendelkezők vehetők fel).

A következő lapon meg kell adnunk az érintett adatbázis állományt névvel, könyvtárral. Ezután meg kell adnunk a megfelelő munkalapokat sorrenddel együtt. A megadás módja jellemző az SPSS-re, és később is gyakran elő fog fordulni. Két fehér színnel kiemelt (tehát hozzáférhető) ablakot kapunk, ezekből az első a választási lehetőségeinket tartalmazza, ahonnan a szükségeseket a bal egérgombbal megfogva át kell helyezni a jobb oldali ablakba, vagy áthelyezhetjük a két mező közötti nyilacskával. Erre a kijelölési módra azért volt szükség (szemben a különben szokásos egérrel való kattintással, esetleg közben nyomva tartva a Ctrl gombot), mert nagy mennyiségű adat esetén, illetve nagygépes környezetben ez volt a hatékony, másrészt ez lehetővé teszi a programozást.

A varázsló 6 lépéses információgyűjtéséből a harmadik csak akkor kell, ha több munkalapot adtunk meg. Ekkor az ezek közti összefüggéseket kell

itt megadni. Ha csak egy munkalapot adtunk meg a 2. lépésben, akkor ez a harmadik lépése a varázslónak nem jelenik meg.

A negyedik lapon a beolvasandó eseteket lehet specifikálni. Ehhez a változóinkból és a megadott függvényekből szerkeszthetünk feltételi képleteket. Ha minden esetet be szeretnénk olvasni (ez gyakori helyzet), akkor egyszerűen lépjünk a következő lapra, ne töltsünk ki semmit.

Az ötödik párbeszédes ablakban a változóneveket adhatjuk meg, vagy itt módosíthatjuk azokat. A varázsló mindenképpen ad (alapértelmezett) neveket a változóknak. Ilyen nevek lehetnek pl. az Excel táblázat első sorának elemei. Ha az SPSS olyan karaktert talál, amely nem megengedett, akkor azt itt kell kijavítani. Mivel később számos beállítást kell megadnunk a változókra, ezért a változóneveknek a varázsló segítségével történő megadása kisebb jelentőségű.

Az utolsó lapon megnézhetjük és szerkeszthetjük a varázsló által összeállított szintaxis leírást. Ez minden részletet tartalmaz, ami az adatok beolvasásához szükséges, illetve amit megadtunk. Ez ismét a korábbi nagygépes múltra utal, másrészt hasznos, ha hasonló szintaxis szerint szeretnénk ismét beolvasni. Ezek után kérhetjük, hogy a megadottak alapján olvassa be az adatunkat, mentse el a kapott szintaxist (save query to file), vagy hogy a szintaxist mentse el a vágólapra további szerkesztés céljából (paste into the syntax editor). A beolvasást a Finish gomb megnyomása indítja. A korábbiakhoz hasonló módon ez a gomb sokszor már az utolsó lap kitöltése előtt is használható (amikor a feltétlen szükséges információkat már megadtuk).

2.1.5. A File menüsor további utasításai

A legfontosabb adatállományokat kezelő utasításokat már megtárgyaltuk, tekintsük még azokat, amelyek a File menüsorban elérhetők. Ezek nagy része más Windows-os programból már ismerős, és nem is kell sok újat mondani ezekről.

Másrészt jelen jegyzet keretében nem is térhetünk ki minden előforduló utasításra, illetve beállításra. A File menüsor kapcsán ezt az elvet követve nem tárgyaljuk a Display Data File Info, Cache Data parancsokat.

Itt lehetőségünk van a szokványos mentés (Save As, Save All Data) mellett közvetlenül adatbázisba menteni az adatokat (Export to Database), a munkalapunkat átnevezni (Rename Dataset), illetve adattároló szerverekkel való kapcsolat kezelésére (Repository).

A Print utasításról sem kell sokat mondani: mint minden más program nyomtatási utasítása, ez is az operációs rendszerben beállított nyomtatóra

küldi ki a munkalap tartalmát (illetve ha más helyzetben adjuk ki, pl. egy output fájl szerepel épp az aktuális ablakban, akkor a nyomtatás nyilván arra vonatkozik).

További hasznos utasítás a Stop Processor, amely a túl sokáig futó statisztikai eljárások megállítására szolgál (ez is nagygépes eredetű). Ezt követően a legutóbb használt néhány adatállomány listája következik a könnyebb betöltés kedvéért, majd az Exit utasítás, amivel a programot tudjuk bezárni, a statisztikai feldolgozást befejezni.

2.1.6. Új fájl

A New / Data parancs egy új, üres munkalapot (vagy más SPSS állományt) nyit, amelybe az adatokat az előbb tárgyalt módok valamelyikével vihetjük be. Más programoktól eltérően az SPSS csak egy munkalapot enged egyidőben megnyitni: így egyértelmű, hogy a feldolgozást melyik adathalmazra értjük. Ez is a nagygépes múltra utal. Több megnyitása esetén új alkalmazás indul.

A New utasítás nem csak új munkalapot tud megnyitni, hanem beolvasási szintaxis-fájlt (Synatx), output állományokat (Output) és egy teendőket leíró ún. script állományt is.

2.1.7. Korábban létrehozott állomány megnyitása

A más programokból is ismert Open / Data utasítás korábban elmentett adatállományokat tud megnyitni. Ez nem csak SPSS adatállomány lehet (.sav kiterjesztéssel), hanem sok más között Excel táblázat (.xls, .xlsx), dBASE állomány (.dbf) és Lotus (.w*) fájl is. Ezeken túl különböző további SPSS állomány fajtákat is meg tudunk nyitni, így beolvasási szintaxis- és script fájlokat is.

A szöveges (.dat és .txt kiterjesztésű, esetleg tabulátorokat tartalmazó, de különben vezérlő karakterektől mentes) állományok megnyitása voltaképp a Read text data utasítást hajtja végre. Az utasítások, parancsok ilyen megtöbbszörözése a komolyabb felhasználói programok esetén tipikus: a felhasználók a számukra megszokott, vagy épp a nekik kedvesebb módon indíthatják a parancsokat.

Ide tartozik még a gyorsító billentyűk esete is. Az Open utasítás kiváltható a Ctrl-o billentyű-kombinációval, a fájl mentése pedig a Ctrl-s gombokkal (a Ctrl gombot lenyomva és nyomva tartva egyszer kell megnyomni a kötőjel után álló betűvel jelzett gombot). Ezt az utasításkiadási

módot a jelen kurzus során nemigen fogjuk használni, és az SPSS-szel való munka első fázisában sem ajánljuk: a gyorsító billentyűket akkor érdemes bevetni, ha a hatásuk minden részletével tisztában vagyuk, és rutinszerűen alkalmazzuk standard helyzetekben, sok, egymáshoz hasonló adaton.

2.1.8. Adataink mentése

A beolvasott vagy begépelt, és kijavított adatunkat a későbbi feldolgozás céljából érdemes az SPSS saját formátumában elmenteni. Erre három utasítás szolgál: a Save, a Save As és a Save All Data. Használatuk a szokásos: ha egy új munkalappal indultunk, és az adatainkat begépeltük, akkor az SPSS az untitled nevet adja az állományunknak, és ennek elmentése során mindkét gomb megnyomása esetén a Save as párbeszédes ablakot kapjuk, utalva arra, hogy valószínűleg meg fogjuk változtatni a fájl nevét. Értelemszerűen, ha egy állománynak már van neve, akkor mind a Save utasítás, mind a megfelelő, floppylemezt ábrázoló ikon, mind a Ctrl-s billentyű kombináció minden kérdés nélkül elmenti az aktuális állományunkat az eddigi névre, az őt már korábban is tartalmazó könyvtárba.

Nagyobb mennyiségű adat begépelése, vagy azokon végzett komolyabb változtatások után érdemes az adatainkat elmenteni. A programból való kilépés, illetve új adat betöltése előtt a program ezt meg is kérdezi (csak ha az aktuális állapot eltér az elmentettől).

2.2. Alapvető műveletek adatokkal

2.2.1. A változók beállításai

A változóink (a táblázatbeli oszlopok) típusát, tulajdonságait a Variable View fülön adhatjuk meg, illetve módosíthatjuk. Ezt két módon lehet elérni: vagy az érintett változó nevére kell a bal egérgombbal kétszer kattintani, vagy egyszerűen lent ezt a fület választjuk. A dupla kattintás esetén egyből az adott változó lesz aktív. Ebben a nézetben 1-1 sor a változókat jelenti, míg az oszlopok azok adott tulajdonságait.

Változónév, címke

Az első oszlopban a változó nevét adhatjuk meg (automatikusan a var0001 stb. neveket kaptuk). Ide csak rövid, egyszerű, lehetőleg ékezetes betű nélküli nevet írjunk. A hosszabb, pontosabb leírást lehetővé tevő nevet

a Labels oszlopban adhatjuk meg. Ez az ún. *címke* a táblázatkezelő programokhoz hasonlóan magyarázatként megjelenik egy kis ablakban, ha a kurzort a változó neve fölé visszük. Ez a címke lehet hosszabb is, tartalmazhat szóközt is, és a későbbi statisztikai eljárások által írt jelentések is használják majd. A címkék használata különösen akkor fontos, ha több hasonló jelentésű változónk is van, és ezek eltérését nem, vagy csak nehezen tudjuk jelezni a változó nevében.

A változóra vonatkozó beállítások a következő oszlopokban olvashatók.

A változók típusa

A *változók típusa* a következők egyike lehet: numeric, comma, dot, scientific notation, date, dollar, custom currency, string és restricted numeric. Ezek rendre a következőket jelentik:

Numeric: Numerikus adattípus, ezt használjuk a leggyakrabban. A táblázatbeli szélességét, és a tizedesjegyek számát a Width és Decimal Places rovatokban adhatjuk meg. Vigyázat, ez a formátum nem a tárolt formára vonatkozik! Ha túl hosszú számot adunk meg, akkor először a tizedesjegyekből ad meg kevesebbet, majd még hosszabb számok esetén áttér a tudományos formára.

Comma: Megegyezik a numerikus adattípussal, de az ezreseket, a milliókat angol szabály szerint vesszővel választja el.

Dot: Pont fordítva, mint a Comma típusnál: itt tizedesvessző van, és pont választja el az ezreseket stb.

Scientific notation: A szokásos tudományos forma: csak egy egészjegyet tartalmaz, és ezt a tíz megfelelő hatványával szorozva értelmezi[1], pl. 123.45 → 1.2345E2.

Date: A dátumok megadásához szükséges formátum, de lényegében csak az angol szokásokat követi (sok választási lehetőségünk van).

Dollar: Dollárokban megadott pénzösszegek számára való formátum, több szélesség és konkrét forma választható.

Custom currency: Néhány, az Options menüpontban korábban beállítandó speciális pénzformátum.

[1] A tapasztalataink szerint a tizedesjegyek számát itt hibásan alkalmazza.

String: A másik gyakori formátum: karaktersorozatot, szöveges adatot tárolhatunk ilyen formában. A megjelenítendő hosszat be lehet állítani.

Restricted Numeric: Fix szélességű egész szám, ahol a hossz kitöltésére vezető nullák vannak.

A táblázatkezelő programoktól eltérően ezek a típusok a teljes oszlopra, illetve változóra érvényesek, itt tehát nincs lehetőség szöveges fejléccel ellátott táblázatok írására (a változó nevének használatát kivéve). Ahova a program számot vár, oda nem is hajlandó szöveges adatot beolvasni (tizedesvesszőt sem), hibajelzést ad. Ez sok hibát segít korán kiszűrni. Másrészt a szöveges adatba természetesen írhatunk számot is. Ha esetleg tévesen adtuk meg a szöveges formátumot, akkor azt arról ismerhetjük fel, hogy automatikusan a szöveg balra van igazítva a cellán belül (a számok pedig jobbra).

Values

Itt megadhatunk egyszerű érték-leírás párokat. Ez akkor hasznos, ha például számokkal jelölünk különböző tulajdonságokat, és ezeket a későbbiekben szeretnénk azonosítani. Erre az oszlopra kattintva egy egyszerű párbeszédes ablak ugrik fel, mely megkérdezi az érték (Value) és leírás (Label) párokat, majd ezeket hozzáadhatjuk a listánkhoz (Add). Szükség esetén törölhetjük (Remove) és lecserélhetjük (Change) az adatokat. A fentiekre példa lehet az 1-es a férfi, míg a 2-es a nő jelölése.

Hiányzóadat kódok

A korábban már említett hiányzóadat kódok nagyon fontosak nagy mennyiségű statisztikai adat korrekt feldolgozásához. A különböző okból hiányzó adatot meg kell jelölnünk ahhoz, hogy adatunkból a lehető legtöbb információt ki tudjuk nyerni (különben minden olyan esetet ki kellene hagynunk a feldolgozásból, amelynek valamely változóértéke hiányzik). Ha egy adat hiányzik, akkor a leggyakoribb eljárás az, hogy az űrlapon a helye üresen marad. Ennek ellenére valamely kódot kell majd választanunk, és számok között a szóköz nem ajánlott. Természetesen olyan értékeket kell ilyen célra használnunk, amelyek különben érvényes adatként nem fordulhatnak elő.

A hiányzóadat kódokat a Missing values gomb megnyomásával adhatjuk meg. Az ekkor kapott párbeszédes ablakban négy lehetőség között választhatunk az ún. rádiógombok segítségével:

- nem adunk meg hiányzóadat kódot (No missing values),

- néhány egyedi kódot adunk meg (Discrete missing values), legfeljebb három különböző érték számára van hely. Csak olyan kódot adhatunk, amely különben az adott változóban érvényes, tehát például nem adhatunk szöveges hiányzóadat kódot, ha a változónk numerikus.

- Megadhatjuk számoknak egy tartományát, amelybe tartozó minden értéket hiányzóadat kódnak tekintünk. Ilyenkor nem elég csak alsó vagy csak felső korlátot megadni.

- Az utolsó lehetőségnél megadhatunk még egy egyedi kódot is. Így gyakorlatilag az előző kettő kombinációját kapjuk, egy tartományt és egy különálló értéket.

Szöveges változó esetén csak az első két lehetőséget tudjuk használni. Az SPSS a hiányzóadat kódokat minden statisztikai eljárásában jelentésüknek megfelelően kezeli, ahol szükséges, ott ezeket a műveletekből kihagyja. Ez az a szolgáltatás, amit táblázatkezelő programokkal nem tudunk elérni, vagy csak nagyon körülményesen.

Cellaformátumok

A cellaformátumot a Columns és az Align oszlopokban tudjuk beállítani. A Columns oszlopban megadhatjuk az oszlop szélességét, az Align oszlopban pedig azt, hogy a cellatartalmat balra, jobbra vagy középre igazítsa a program. Az utóbbiak kiválasztása során figyeljünk arra, hogy a számokat jobbra igazítva, míg a szöveget balra igazítva tudjuk jobban olvasni. Ezek a beállítások egyben segítenek a téves adatbevitel elkerülésében is, és egybeesnek az alapértelmezéssel az adott adattípust illetően. A cellák szélességét pedig közvetlenül a cellahatárok bal egérgombbal való mozgatásával is módosíthatjuk (fenn, a változóneveknél).

Mérési skálák

A bevezetőben már tárgyalt mérési skálák közül itt hármat lehet beállítani: az intervallum- vagy arányskálát (az SPSS-ben e kettőt nem különböztetik meg, neve scale), a rangskálát (ordinal) és a nominális skálát (nominal). Ezek megadása nagyon fontos: ezt a beosztást csak a felhasználó tudhatja, a program bizonyos esetekben nem is dönthetné el. Másrészt ezek ismerete egyes statisztikai eljárások végrehajthatóságát, eredményét, illetve azok értelmezését döntően befolyásolja.

Role

Néhány párbeszédablak képes a változókat aszerint használni, hogy milyen szabályt állítottunk be rájuk. Például egy regresszió számításnál az inputot veszi a magyarázó, míg a target változót a magyarázott változónak. Vagy például tesztek esetén a partition szabály esetén ketté szedi az adatokat teszt és tanuló halmazra. Ezen lehetőségek a legújabb verziókban kezdtek el kialakulni.

2.2.2. A szerkesztési és a nézet menüsor

A szerkesztési menüsor

A szerkesztési menüsor (Edit) alapjában véve három dolgot tartalmaz: a szokásos szerkesztési utasításokat, egy kereső eljárást és a program beállítási lehetőségeit. A szerkesztési utasítások a Word szokásait követik, tehát a Cut a kijelölt adatot kitörli és egyidejűleg a vágólapra teszi (clipboard), a Copy ezzel szemben nem törli az adatot, csak kijelöli a másolásra, és a vágólapra teszi, a Paste pedig a vágólap tartalmát az egérrel kijelölt cellától kezdve a táblázatba írja. Ezek az utasítások mind az ikonokat, mind a gyorsító billentyűket tekintve a Word-nek megfelelően is kiválthatók: Ctrl-x, Ctrl-c és Ctrl-v.

Ezeket két további parancs egészíti ki. A Clear csak kitörli a megadott cellákat a vágólapra másolás nélkül. Az Undo (Ctrl-z, illetve a balra visszamutató nyíl az ikonok között) a kiadott utasítások egyszintű visszavonására szolgál, tehát csak a legutolsó parancs vonható vissza. Az utóbbi tulajdonság miatt különösen az első fázisban érdemes körültekintően dolgozni.

A Find (Ctrl-f) és Find Next (F3) keresési utasítás a szokásos módon jelsorozatot (számokat vagy szöveget) keres a kijelölt tartományban. A Replace utasítással egyszerű cserék hajthatók végre az értékeken. Ennél az utasításnál pár alap dolgot tudunk beállítani. Például, hogy a változó ezzel kezdődjön, vagy tartalmazza a cserélendő szöveges értéket.

Ezen szokásos szerkesztési parancsok mellett megjelennek a statisztikai program mivoltából eredő utasítások is. Ilyen a változó beillesztése (Paste Variables), az új változó beszúrása (Insert Variable), vagy az eset beszúrása (Insert Cases).

Az Insert variable egy új változót szúr be a meglévők közé, az aktuális cella változója elé. Az új változó az alapértelmezésnek megfelelő beállításokkal rendelkezik majd. Ezután ezeken nyilván módosítani kell a változó tartalmának megfelelően. A változóban természetesen nem lesznek adatok.

Az előző parancshoz hasonlóan új esetet illeszthetünk be az Insert case utasítással az aktuális cella esete elé (vagy ha nem volt ilyen, pl. mert épp egy egész változó volt kijelölve), akkor első esetként. Az új esetben nem lesznek adatok.

A Go to Case parancs végrehajtása után egy megadott esetre ugrik a kurzor. Az eset sorszámát egy párbeszédes ablakban kell megadni. Hasonló a Go to Variable parancs is, mely esetében annak a változónak nevét kell kiválasztani, amire ugorni szeretnénk.

Itt találjuk még a korábbi utasításaink visszavonására szolgáló Undo és Redo parancsokat is.

Beállítások

A program beállításait az Options paranccsal érhetjük el. A lehetőségek számosak, de más, hasonló méretű programmal összevetve ezekből nincs átláthatatlanul sok. Ismét csak a legfontosabbakat említjük – részben azért is, mert az SPSS teljes értékűen használható a beállítások módosítása nélkül is.

Az Options parancs kiválasztása után egy több munkalapból álló párbeszédes ablakot kapunk. A módosítások alkalmazását az Apply billentyűvel kérhetjük. Vigyázzunk, a rádiógombokon és az egyes lehetőségek jelölőnégyzetein túl a mellettük lévő szövegre is elég kattintani a beállításhoz. Ez más programok esetén is így szokás újabban, de óvatlan használat esetén téves megadásokat kaphatunk.

A nézet menüsor

A nézet (View) menüsor a más programokban megszokott elemeket tartalmazza:

- Az állapotsor (Status Bar) az SPSS ablak alján a program működéséről tájékoztat: az egérrel éppen megjelölt (de még ki nem választott) parancs jelentéséről és az SPSS processzor állapotáról. Ez utóbbi ismét a nagygépes időkre utal.

- Az eszközsor (Toolbars) a leggyakrabban használt utasításokat teszi könnyebben elérhetővé a táblázatunk feletti ikonokkal.

- A Menu editor pontban a menü szerkezete személyre szabható.

- A Fonts menüpont az SPSS által használt betűtípusok megadását teszi lehetővé. Magától értetődően az operációs rendszerben beállított betűk közül választhatunk.

- A Grid Lines menüponttal a táblázat celláit elválasztó vonalakat lehet ki- és bekapcsolni. A kikapcsolás talán a nyomtatás előtt ajánlható, egyéb esetben az elválasztó vonalakkal jobban lehet tájékozódni a cellák között.

- A Value Labels lehetőséget nem tárgyaljuk.

- A Customize Variable View lehetőség csak akkor aktív, ha a változó fülön vagyunk. Ennél a pontnál beállíthatjuk, hogy mely változó tulajdonságok jelenjenek meg, illetve milyen sorrendben.

összefoglalóan az mondható, hogy a nézet alapbeállításai általában jók, a módosításukra ritkán van szükség. Az állapotsor és az eszközsor eltávolításának például akkor van értelme, ha ezek nélkül az adatunk áttekinthetősége lényegesen jobbá válik.

2.2.3. Az adatok beállításai, rendezése és mozgatása

Az előző, inkább általános parancsok után a Data menüsor már specifikus, a statisztikai feldolgozáshoz szorosabban kötődő utasításokat tartalmaz. Ahogy korábban már említettük, az adatok kezelése, csoportosítása és átalakítása egy komoly részt jelent a teljes feldolgozáson belül. Az adatok elsődleges kezelésére vonatkozó parancsokat tartalmazza a Data menüsor.

Az utasítások első csoportja az adatok, változók és esetek közvetlen beállításaival foglalkozik, ezek egy kis részével már korábban is találkoztunk:

Define Variable Properties: Az egyes változók tulajdonságait lehet beállítani ezzel a paranccsal. A jobb oldali listába áthelyezhetjük a feltérképezendő változókat. Továbbá beállíthatjuk, hogy a későbbiekben hány változót jelenítsen meg, és a feltérképezésbe mennyi adatot vegyen be a program. Ha ezt elvégeztük, a Continue gomb megnyomására megkapjuk a változóra vonatkozó adatokat. Bal oldalt válogathatunk a feltérképezésben részt vett változók között. Míg jobb oldalt megkapjuk a fontosabb információkat a változóról. Ilyenek például a korábban beállított változók tulajdonságai, de táblázatos formában kapunk összefoglaló információkat az adatokra vonatkozóan is.

Set Measurement Level for Unknow: Itt beállíthatjuk a mérési skálákat a még ismeretlen változókra. Gyakorlatilag a három csoport valamelyikébe helyezhetjük a kis nyilak segítségével.

Copy Data Properties: Lehetőségünk van a jelenlegi állományunkra áthozni egy korábban beállított változóhalmaz tulajdonságait. Ezt megtehetjük egy megnyitott állományból (An open dataset), vagy egy mentett SPSS állományból. Valamint itt tudjuk az egyik aktuális változónk összes tulajdonságát átmásolásolni egy másik változónkra (The active dataset).

New Custom Attribute: Ebben a menüpontban lehetőségünk van – a korábban említett alapértelmezett attribútumok mellett – újabb attribútumokat létrehozni a változóinkhoz.

Define Dates: Ezzel az utasítással dátumformákat adhatunk meg a már meglévő adatunkhoz, minden eset egy-egy új dátumot fog kapni, növekvő sorrendben. Az utasítás gyorsító billentyű kombinációja: Altde.

Több furcsaságot is láthatunk: az egyik, hogy a magyar nyelvi környezetben talán leggyakoribb év – hónap – nap dátum forma még az ellentétes, az angol nyelvnek megfelelő formájában sem áll rendelkezésre. De ugyanígy hiányzik a hónap – nap forma is (valószínűleg azért, mert az egyes hónapok napjainak számát nehezebb meghatározni). Az alapértelmezés az, hogy nincs dátum. Custom beállítás használatához előbb egy DATE parancsot kell kiadni a syntax ablakban. Itt lehet definiálni az általunk használni kívánt dátum formátumot. Például a DATE Y 1970 M paranccsal létrehoztunk egy év és hónapot tartalmazó dátumtípust, mely 1970 januárjával kezdődik. Ezen parancs után a Custom esetében a sorokhoz hozzárendeli sorban a 1970 január, 1970 február. . . dátumokat.

Ezzel együtt a parancs hasznos: az eseteinkhez könnyen tudunk dátumokat, időpontokat rendelni. Ez jól alkalmazható például tőzsdei adatok utólagos napokhoz rendelése során (és a hét munkanapjait is figyelembe vehetjük).

Multiple Response Define Variable set: Itt definiálhatjuk azon változók halmazát, amelyek több változóból állnak össze. Erre tipikus példa, amikor egy kérdőíven több lehetőség közül többet is kiválaszthatunk és ezeket a változókat egységes halmazban kell kezelni.

Identify Duplicate Cases: Ez a funkció az azonos esetek kiszűrésére szolgál. Beállítható, hogy mely változókat vizsgálja azonosság szempontjából. Az azonosoknál megadható, hogy hogyan jelölje meg őket. Például minden csoportban az első (vagy egyedi) kapjon egy 1-es értéket, míg a duplumok 0-át az új változóba (First case in each group is primary).

A menüsor parancsainak következő csoportja az adatállományunk alakításával foglalkozik.

Sort cases: Ez a parancs sok esetben nagyon hasznos lehet: az eseteinket ezzel lehet új sorrendbe rendezni valamely változó értéke alapján. Természetesen az új sorrendben minden eset mint egy egység kerül új helyre, tehát a korábban egy sorban lévő adataink ezután is egy sorban lesznek (ez így van más táblázatkezelő programok esetében is). Az utasítás gyorsító billentyű kombinációja: Alt-do.

A parancs hatására egy párbeszédes ablakot kapunk. A legfontosabb dolog kijelölni azt a változót vagy változókat, amelyek szerint a rendezést kérjük. Ezt a kijelölést az SPSS-ben szokásos módon úgy tehetjük meg, hogy a bal oldali kis ablakban felsorolt változók közül az érintetteket a bal egérgombbal egyenként megjelöljük, majd a két kis ablak közti nyíllal átvisszük a jobb oldali ablakba. Az átmozgatás előtt be kell állítani az illető változóra vonatkozó rendezés irányát (növekvő vagy csökkenő: ascending – descending).

Több változó megadása a rendezés több szintjét jelenti, tehát az elsődleges rendezési szempont az első változó szerinti lesz. Ha ezután van két vagy több olyan eset, amelyekre az első rendezési változó értéke megegyezik, akkor a második változó értéke fogja eldönteni a sorrendet.

A parancs használatával az adataink tartalma nem változik, de az eredeti, esetleg rendezetlen sorrend egy rendezés után már nem érhető el. Emiatt is érdemes minden rendezés előtt az adatállományt új néven elmenteni, vagy egy sorszám változót definiálni, mellyel visszarendezhető az adatállományunk.

Sort Variables: A parancs arra szolgál, hogy a változókat új sorrendbe helyezzük. Itt meg kell adnunk, hogy melyik tulajdonsága alapján hogyan rendezzük. A korábbi sorrend belementhető egy új változó tulajdonságába a lenti jelölőnégyzet és nevének megadásával. Típusok esetén a típus neve alapján rendez.

Transpose: A Transpose parancs, ahogy a neve is mondja, a táblázatunkat transzponálja, a sorokból oszlopokat csinál, és viszont. Az átalakítás során megadhatjuk, hogy mely változókból hozzon létre eseteket. Itt tehát nem kell minden változót megadnunk, viszont a ki nem választott változók tartalma elveszik. Ettől eltekintve az utasítás kétszeri (megfelelő) végrehajtásával visszakaphatjuk az eredeti adattáblázatot. Egyes statisztikai eljárások mind esetekre, mind változókra értelmes eredményt adhatnak. Az eljárások egy része erre fel van készülve (mint pl. a klaszterezés), és a táblázatunk transzponálása nélkül is meg tudják oldani a feladatot – paramétereik megfelelő beállításával.

A kapott párbeszédes ablakban megadhatunk egy olyan változót is (Name Variable), amelyben az új táblázatbeli változók nevei lesznek. Az alaphelyzetet az esetleges hibás bevitel után gyorsan elérhetjük a Reset gomb megnyomásával. Az utasítás gyorsító billentyű kombinációja: Alt-dn.

Merge Files: A parancs két SPSS adatállomány összeillesztését tudja megoldani. A feladat önmagában sem egyszerű, így az utasítás paraméterezése is kissé bonyolult. Az egyszerű összevonásokat minden esetre kiváltani is egyszerű. Az utasítás gyorsító billentyű kombinációja: Alt-dg.

1. Az egyik egyszerű eset, amikor a két állomány változóit szeretnénk összevonni, és ezek teljesen (páronként) különbözőek. Ekkor az egyik állományt olvassuk be a programba a szokásos módon, majd kérjük a Data / Merge Files / Add Variables lehetőséget (az utasítás gyorsító billentyű kombinációja: Alt-dmv.). A másik fájl megadása után egy olyan párbeszédes ablakot kapunk, ahol már be is van állítva, hogy a második állomány minden változóját kérjük bevonni az összeillesztésbe. Ez persze nem mindig célunk, ilyenkor a nem szükséges változókat visszatehetjük a bal oldali ablakba. A kihagyandó változók nevét meg tudjuk változtatni a Rename gombbal (hogy azután visszategyük a beillesztendők közé). Ha sok változót kell kijelölni a mozgatáshoz, akkor használhatjuk a szokásos lehetőséget: a Alt gomb nyomvatartása mellett a bal egérgombbal többet is kiválaszthatunk egy egyszeri átvitelhez.

   A jobb oldali ablakban minden változóra meg van adva, hogy melyik állományból származik. Fontos lehetőség, hogy a két állomány eseteit egyes ún. kulcsváltozók (key variables) alapján egyeztethetjük. Így az egyes állományokban hiányzó esetek nem borítják

föl az esetek egyeztetését. Például, ha a két állományunkat az eseteket egyértelműen azonosító név változó szerint illesztjük össze, akkor az esetek elcsúszását elkerülhetjük.

2. A másik egyszerű eset, amikor két állomány azonos változókat tartalmaz, és a két fájl különböző eseteit szeretnénk összevonni. Ilyenkor a Data / Merge Files / Add Cases parancs kérése után meg kell adnunk a két fájlt, ezután kapunk egy párbeszédes ablakot, amelyben a második állomány minden változója az átvivendők között van. Persze ezen megint változtathatunk a szokott módon. A parancs gyorsító billentyű kombinációja: Alt-dmc.

Az ezeknél bonyolultabb helyzetekre itt nem térünk ki, de ezek is megoldhatók rövidebb kísérletezés után. A rossz helyre került értékeket pedig az Edit menüsor parancsaival rendezhetjük.

Restructure: Ebben a pontban az adatállományunk átstruktúrálását végezhetjük el. A lehetőségek sokfélék lehetek, így inkább egy példával illusztrálnánk ezen funkció képességét. Ha például egy eseményről több mintavétel is készült és ezeket egy-egy eset különböző változóiba mentették, azonban mi olyan statisztikát szeretnénk készíteni, ahol ezen méréseket külön-külön kezeljük, akkor az egy esetet változók mentén fel kell bontani több esetre. De például ennek az ellenkezője is előfordulhat, miszerint a valójában egy eset több esetre szétszedve került az adatállományunkba. Ilyen problémák kezelésére szolgál ez a parancs, mely hatásait a színes ábrák jól mutatják. Itt is elmondható, hogy a nem kívánt átstruktúrálás visszaállítása érdekében javasolt menteni a parancs megkezdése előtt.

Aggregate: Ezzel az utasítással egy nagy esetszámú állományból egy döntési változó (break variable) alapján összevont értékeket tartalmazó új változót (aggregate variable) vagy változókat hozhatunk létre. Az utasítás gyorsító billentyű kombinációja: Alt-da.

Ilyen átalakításra például akkor van szükségünk, ha van egy kimutatásunk egy cég munkavállalóinak keresetéről, és arra vagyunk kíváncsiak, hogy az azonos beosztásúak átlagos keresete mennyi. Ekkor a beosztás kódja lesz a döntési változó, és az azonos beosztási kódú esetek keresetértékeinek átlagát kell az aggregált változóban megkapnunk. Kevés adat esetén ezt persze közvetlen átlagszámítással vagy a később tárgyalt esetkiválasztással is megoldhatnánk, de ha a döntési változónknak nagyon sok különböző értéke van, akkor az Aggregate utasítás a leghatékonyabb.

Az Aggregate utasítás kiadása után egy olyan párbeszédes ablakot kapunk, amelyikben a döntési és az aggregált változót (vagy változókat) kell megadni. Az aggregált változók nevét és címkéjét előre megadhatjuk (a többi paramétert az új állományban állíthatjuk be). A Function gomb megnyomásával az alapértelmezésbeli átlag helyett más függvénnyel lehet meghatározni az aggregálást. A döntési változó által kijelölt csoportok elemszámát egy új változóban kérhetjük (Save number of cases in break group as variable). Az eredményállomány beállítás szerint pótolhatja a jelen aktuális munkalap tartalmát, vagy kerülhet egy új fájlba is.

Vigyázat, az aggregált változó típusa megegyezik az alapul vett változóéval. Ez a számított értékeket is befolyásolhatja a célváltozó szélessége, illetve a tizedesjegyek száma (kerekítés!) miatt.

A menüsor utolsó csoport utasítása közül az elsővel, a Split File nevűvel nem foglalkozunk (az adatállományunk felosztására szolgál általában valamely olyan változó értéke alapján, amely csoportokat képez az esetek között). Olyan esetben lehet ez hasznos, amikor valamely következő statisztikai eljárás nem tud csoportokat vagy részhalmazokat kezelni. A maradék két parancs viszont fontos és gyakran használatos is.

Select Cases: A Select Cases utasítás arra szolgál, hogy egy-egy statisztikai eljárás számára kiválasszuk a teljes adatállományból azokat az eseteket, amelyekre a végrehajtást kérjük. Ezt persze megtehetnénk új állományok létrehozásával is, de az nehézkesebb lenne. A parancs gyorsító billentyű kombinációja: Alt-dc.

A kapott párbeszédes ablak alapértelmezésben minden esetet megtart. A parancs lényegét persze a többi lehetőség adja, ezek között rádiógombbal választhatunk. A kiválasztás jelenlegi érvényes beállítása az ablak bal alsó sarkában olvasható. Amíg nem sikerült elfogadtatni egy új megadást, addig a régi, illetve az alapbeállítás az érvényes.

Alaphelyzetben a ki nem választott esetek az adatállományunkban maradnak. Az SPSS a következő feldolgozási lépések során (amíg a kiválasztást meg nem változtatjuk, illetve meg nem szüntetjük) csak a szűkített esethalmazzal fog számolni. Ezt az esetkiválasztási ablak alján egy rádiógombrendszer is jelzi: a ki nem választott esetek csak azok, amik a szűrés után megmaradtak (filter out unselected Cases). A másik lehetőség az, amikor a kiválasztott elemeket új állományba helyezzük (Copy selected cases to a new dataset), a hamadik pedig a ki nem választott esetek törlését kérjük (Delete Unselected Cases).

Feltétel megadásával

Az If condition is satisfied lehetőséget akkor kell választanunk, ha a megtartandó eseteinket a változóértékek vagy azok függvényértékei közti relációk (pl. egyenlőség vagy egyenlőtlenség) határozzák meg. A feltétel megadásához nyomjuk meg az If gombot.

Egy kalkulátorszerű ablakot kapunk, és a jobb felső sarokban kell kialakítanunk a feltételt. Ehhez a bal oldalról változókat választhatunk ki, a nyomógombokkal műveleteket és relációjeleket adhatunk meg, és a jobb szélen levő listából segédfüggvényeket szúrhatunk be.

Ilyen függvényből több mint száz van, ezekre itt nem tudunk részletekbe menően kitérni. Szerencsére mindegyikről kaphatunk egy rövid magyarázatot, ha a függvény nevére klikkelünk a jobb egérgombbal. Másrészt a függvények neve is utal a jelentésükre, és az argumentum jelölése is a várt paraméterekre. Például az ABS(numexpr) az abszolút érték függvényt jelöli, aminek egy numerikus kifejezés lehet az argumentuma[2].

A függvény beszúrása után egy kérdőjelet látunk az argumentum helyén. Ha ekkor egy változót hozunk be a bal oldali listából, akkor az a kérdőjel helyére kerül. A feltétel megfogalmazása után a Continue gombbal tudunk visszatérni az esetkiválasztási ablakba, ahol a generált feltételt olvashatjuk is az IF gomb mellett.

Az OK gomb megnyomásával elfogadjuk a megadott feltételt, és a program ki is jelöli a kiválasztott eseteket. A táblázat bal szélén, a sorok azonosítására szolgáló sötét alapon írt számok közül áthúzva jelennek meg azok, amely esetek nem felelnek meg a feltételünknek. Másrészt a táblázat utolsó oszlopába kaptunk egy új, ún. szűrőváltozót is.

Az esetek egy véletlen mintája

A Random sample of cases rádiógomb kiválasztása után meg kell nyomnunk a Sample gombot a részletek tisztázásához. A véletlen minta generálásához két lehetőségből választhatunk:

1. Az első esetben azt kell megadnunk, hogy körülbelül hány százalékát tartsuk meg az eredeti mintának. A kiválasztott elemek száma, azért lesz csak körülbelül a megadott arányú, mert az eljárás véletlen jellege csak közelítőleg biztosítja ezt. Az ilyen esetkiválasztás olyankor hasznos, ha egy nagyon nagy mintából

[2]Vigyázat, az RND függvény nem a véletlen számokat generáló, hanem a kerekítésfüggvény.

kell reprezentatív részmintát megadnunk. Például ilyen eset, ha halmazunkat 2 részre szeretnénk bontani. Egy tanuló halmazra, melyen megpróbálunk megfigyeléseket eszközölni, illetve egy teszt halmazra, melyen ellenőrizhetjük a megfigyelésünket.

2. A másik lehetőség az, hogy egy pontosan megadott darabszámú részmintát generáltatunk a programmal. A beállítandó paraméterek: a kiválasztandó esetek száma és az, hogy ezeket az első hány esetből kell kiválasztani.

Az esetsorszám tartománya által kijelölt esetek

Egy alapvetően eltérő megoldás az, amikor az esetek sorszámának egy tartományát, intervallumát adjuk meg, és mindazon esetek kiválasztását kérjük, amelyekre az eset sorszáma a megadott korlátok között van. Ennek a formának akkor van értelme, ha az esetek sorszáma jellemző érték, aminek idő vagy más sorrendi jelentése is van.

A rádiógomb beállítása után meg kell nyomni a Range feliratú gombot, és a kapott párbeszédes ablakban megadhatjuk a kiválasztandó tartomány (intervallum) alsó és felső korlátját (az első és az utolsó kiválasztandó eset sorszámát).

Szűrőváltozóval kijelölt esetek

Az esetkiválasztásra használhatunk szűrőváltozókat, olyanokat, amelyeknek 1 értéke jelöli ki a megtartandó eseteket. Az érintett változót a szokásos módon a listából át kell vinni a kijelölt területre (az átmozgató nyíl csak a rádiógomb megnyomása után aktivizálódik). A szűrőváltozót úgy lehet megváltoztatni, hogy a régi kiválasztott változót (a most visszafelé mutató mozgató nyíllal) visszatesszük a többi közé, és az újjal pótoljuk.

A szűrőváltozó kialakításával itt most nem foglalkozunk, csak arra utalunk, hogy például a később tárgyalt Transform / Compute parancscsal tetszőleges ilyen változó létrehozható. Ilyen szűrőváltozót a többi esetkiválasztási mód is létrehoz.

Weight Cases: Az esetek súlyozása az adatok előkészítésének egy további fontos fázisa. Olyankor van rá szükség, ha minden változóértékhez egy további szám tartozik, amely jellemzi azt, hogy az illető változóérték a statisztikai feldolgozás szempontjából mekkora jelentőségű a többi változóértékkel összehasonlítva. A jelentőség különbözőségét okozhatja több dolog is: például az, hogy az illető értéket hány (eredeti, korábbi)

eset átlagából számítottuk, vagy hogy a mérés, amiből kaptuk, milyen pontos, vagy mennyire megbízható volt. A súlyok kialakításához fontos tudnunk, hogy a program a súlyainkat szimulált többszörözéssel fogja értelmezni.

Amennyiben ilyen szempontok szerint nem tudunk különbséget tenni az egyes esetekre vonatkozó változóértékek között, akkor vagy azonos súlyt adunk minden esetnek, vagy – ami ezzel egyenértékű – nem használunk súlyozást. A súlyozás közvetlenül nyilván nem változtatja meg a változóink értékét, de befolyásolja azok értelmezését, illetve szerepét a végrehajtott statisztikai eljárásokban.

Az utasítás végrehajtható a Data / Weight Cases menüpont kiválasztásával, vagy a Alt-dw gyorsító billentyűkombinációval. A párbeszédes ablakban látszik, hogy az alapbeállítás az, hogy nem kérjük az esetek súlyozását. Ezt a rádiógomb átállításával lehet megváltoztatni. Ez önmagában még nem elég (ezt abból is megállapíthatjuk, hogy az OK gomb nem hatásos): meg kell adnunk azt a változót, amiben az egyes esetekre vonatkozó súlyokat meg lehet találni. A mindenkori érvényes beállítást az ablak alján olvashatjuk. Az OK gombbal tudjuk a súlyozást alkalmazni.

Vigyázat, az esetek súlyozása a változók leírásában vagy az esetek sorszámán nem látszik, erre csak a munkalapunk alatt a státusz felirat utal: Weight On. Ha a későbbiekben egy statisztikai eljárás meglepő eredményt ad, akkor érdemes az esetkiválasztás mellett azt is ellenőrizni, hogy az eseteink nincsenek-e súlyozva.

2.2.4. Az adatok módosítása

A Transform menüsor is a fontosabbak közé tartozik. A benne szereplő parancsok csaknem mind a meglévő adatunk átalakítására szolgálnak, sokszor a változóinkból újakat hoznak létre.

Ebből a sorból kilóg a Random Number Generator (Alt-tg). Ez a program által használt pszeudovéletlenszám generátor számára adja meg az induló értéket, informatikai körökben Seed-nek nevezett értéket. Ez egy fontos eszköz, mert ha egy bizonyos számot itt megadunk, akkor bár véletlenszerűnek fognak tűnni a generált számok, de ezek sorozata megismételhető, ami bizonyos tudományos feladatok megoldása során elengedhetetlen. Ha teljesen (tőlünk is) véletlen sorozatot szeretnénk, akkor válasszuk a Set Starting Point-nál a Random (véletlen) lehetőséget.

Új változó kiszámítása

A Compute Variable menüponttal meglehetősen kötetlenül lehet új változókat létrehozni úgy, hogy a tartalmukat a korábbi változók függvényeként adjuk meg. A gyorsító billentyűkombináció az Alt-tc.

A párbeszédes ablak bal felső sarkában az új változó nevét adhatjuk meg. Ennek begépelése után aktiválódik a Type & Label gomb. Ennek megnyomása után megadhatjuk a kapcsolódó értékeket (de természetesen a változó kiszámítása után is megtehetjük ezt).

Ezek után a jobb felső ablakban kialakíthatjuk az új változót meghatározó képletet. Ehhez felhasználhatjuk a jobb oldalt, alul levő ablakban felsorolt beépített függvényeket és a bal ablakban levő változóinkat. A függvényeket funkciójuk szerint csoportokba rendezetten is kereshetjük (Function group). Ezeket a csoportokat jobb oldalt középen találjuk. Egy csoport kiválasztás hatására csak a kiválasztott csoportba tartozó függvények jelennek meg a jobb alsó választóban (Function and Special Variables). Ha nem tudjuk, hogy mely csoportban található a függvényünk, akkor az összes függvényt tartalmazó csoportot (All) válasszuk. Kényelmessé teszi a szerkesztést, hogy a képletek argumentumában megjelenő kérdőjel helyére egyszerűen bevihetünk egy változót, nem kell a kérdőjelet kitörölni.

A képlet összeállításához jól alkalmazhatók a képletszerkesztő ablak alatti billentyűk. Vigyázat, ha a billentyűkkel relációjeleket viszünk be, akkor a képletet a program logikai függvényként értelmezi, és így a visszaadott érték a 0 és az 1 valamelyike lesz a reláció teljesülésétől függően. A hullámos vonal a negáció jele.

érdekes lehetőséget nyújt az értékadásra a párbeszédes ablak alján levő If feliratú gomb. Ennek segítségével az értékadást korlátozhatjuk, tehát a képlet által megadott értékeket csak azon esetekre kapja meg az új változó, amelyekre a megadott feltétel teljesül. A többi esetre nincs értékadás. Ez lehetővé teszi, hogy egy változó különböző esetekre vonatkozó értékeit más és más képlettel határozzuk meg.

Előfordulások megszámlálása

A Count Values within Cases parancs arra való, hogy adott változóban egyes értékek előfordulásainak számát meghatározzuk. Olyankor használatos, ha az így kapott értékekkel a továbbiakban még számolni szeretnénk, mert különben több olyan statisztikai eljárás is létezik, amely ilyen információt szolgáltat az output ablakban (tehát szöveges fájlban). A gyorsító billentyűkombináció az Alt-to.

A parancs paraméterezése az előzőekhez hasonlóan történik. A megjelenő párbeszédes ablakban meg kell adnunk az új változó nevét és esetleg címkéjét. Ezután azt a változót vagy változókat áttesszük a bal oldali listából a jobb oldali ablakba, amelyekre az esetszámlálást kérjük. Ez még nem elég a parancs végrehajtásához: az OK gomb nem aktív. Ehhez még meg kell adnunk azt az értéket, vagy azokat az értékeket, amelyekre az összegzést kérjük.

Ha több értéket adunk meg, akkor bármelyiknek a kijelölt változóban való előfordulása esetén az új változóban megjelenik egy 1-es (különben 0 lesz az értéke). Ha több változót adtunk meg, akkor az adott esetben a változókban előforduló megadott értékek száma kerül az új változóba, tehát ennek értéke nulla és a megadott változók száma között változhat.

Ezt a parancsot is lehet feltételes formában használni, tehát az előző paramétereket még kiegészíthetjük egy feltétellel, amit az If gomb megnyomása után vihetünk be. Ekkor a számlálás csak azokra az esetekre vonatkozik majd, amelyek megfelelnek a megadott feltételnek. Tehát csak ezekre az esetekre fog az eredményváltozó új értéket kapni. A korábbi esetleges értékek megmaradnak (ami gyakran előfordul, ha nem sikerült egy lépésben minden paramétert jól megadni, és ugyanazt a változót használjuk továbbra is – ami pedig alapértelmezés)!

A megszámolandó értékek köre is sokféle lehet. Ezt a Define Values gomb megnyomása után tehetjük meg. A legegyszerűbb az az eset, amely egyben az alapbeállítás is: amikor a megszámolandó értékeket egyesével adhatjuk meg (Value). Ilyenkor az illető érték bevitele után az Add gombbal (vagy az „a” betű megnyomásával) mozgathatjuk át a figyelembe vett értékek ablakába. A következő rádiógombbal a program által értelmezett hiányzóadat kódot (System missing) számolja. A következő lehetőség a program által generált és a felhasználó által megadott hiányzóadat kódokat együttesen adja meg. Beállíthatjuk értékeknek egy intervallumát is. Ezt megtehetjük úgy is, hogy nem adunk meg konkrét határokat, hanem csak a felső korlátot (illetve fordítva: egy értéktől kezdve fölfelé).

Esetek eltolása

Itt megtehetjük, hogy egy eset egyik változója értékét valamely rákövetkező, vagy előtte lévő eset új változójának adjuk értékül. A felugró ablak bal oldalán megadhatjuk, hogy mely változó értéket szeretnénk eltolni. Ezen szabályokat láthatjuk a (Variable → New name) részen. Jobb oldalt a változóink vannak felsorolva, melyekből egy kiválasztása után a kis nyíllal felvehetjük a szabályok közé. Ekkor a szabályok között megjelenik egy új, melynél a cél változó még ???-lel van jelölve. Alul (Name and Method) a

szabály kiválasztása után megadható, hogy mi legyen az új változó neve, az, hogy lefelé, vagy felfelé tolódjanak el az értékek, és hogy ez az eltolás mekkora legyen (Number of cases to shift). Az OK gomb hatására a definiált változó eltolások sorban végrehajtódnak. A parancs gyorsító billentyűkombinációja az Alt-tf. Ez a parancs hasznos lehet, ha például egy adatsor egymásra következő adatait szeretnénk összehasonlítani.

Újrakódolás

A Recode nevű parancs az új változó kiszámítása művelethez hasonló, de ebben az esetben egy már bevitt változó minden értékére meg kell mondanunk, hogy azt milyen új értékkel kívánjuk pótolni. A gyorsító billentyűkombinációk az Alt-tr, Alt-ts és Alt-ta. Három további választási lehetőségünk van, amelyek menüpontként jelennek meg: az eredményt vagy ugyanabba a változóba kérjük (Into Same Variables), vagy egy újba (Into Different Variables) vagy automatikus újrakódolást kérünk (Automatic Recode). Az első kettő nagyon hasonló, az egyetlen különbség a felugró párbeszédes ablakban, hogy megkérdezi az új változó nevét és címkéjét, vagy nem. A harmadik egy kicsit jobban eltér, így ezt külön részletezünk.

A kapott párbeszédes ablakban először az érintett változót vagy változókat át kell tenni a jobb oldali ablakba. Új változóba mentés esetén az Output Variable rész aktívvá válik és megadhatók az új változókra vonatkozó információk. Az Old and New Values gomb után kapott újabb párbeszédes ablakban aztán minden, az értékek átalakítására vonatkozó beállítást megtehetünk.

Alapértelmezésben a program a legvalószínűbb lépésre számít: hogy több régi értékre közvetlenül megmondjuk, hogy milyen újabb érték tartozzon hozzá. A két értéket értelemszerűen a két kis ablakba kell írni (Old and new value). Az új értékek között lehet rendszer által generált hiányzóadat kód is. Ha a beállítást megtettük, akkor az Add gombbal vigyük be ezt a kész kódok ablakába (Old $\rightarrow$ New).

A régi kódok között lehet rendszer adta hiányzóadat kód (System-missing), bármiféle hiányzóadat kód (System- or user missing), értékeknek egy intervalluma (Range), egy adott értéknél nem nagyobb (Range lowest through), illetve nem kisebb értékek (Range through highest) halmaza is. Az utóbbi kettőbe beleszámítanak a hiányzóadat kódok is. Ez a párbeszédes ablak eddig a pontig megegyezik azzal, amivel már találkoztunk a korábbi előfordulások számlálása parancs esetén.

Fontos új lehetőség a minden más változóérték (All other values). Ez lehetővé teszi, hogy ne kelljen minden szóba jövő régi értékre egy-egy be-

jegyzést tennünk, ha ezek új értéke már megegyezik.

Az újrakódolás is köthető feltételhez. Ennek megadása a korábban tárgyaltak szerint történik (lásd például az előző, az előfordulások számlálása parancsnál írtakat). A feltételeknek meg nem felelő esetekre, illetve akkor, ha az aktuális régi értéket nem soroltuk fel az újrakódolási listánkban, nem változik a korábbi érték (azaz ha egy változót helyben kódoltunk át).

Az Automatic Recode parancs olyan esetben használható, amikor az eredeti változóértékek valamilyen szempontból nem alkalmasak a további feldolgozáshoz. Például szöveges változókat tudunk átkódolni numerikussá, amik egyes statisztikai eljárások használatát teszik lehetővé. Az átalakítás során a változó- és értékcímkék megmaradnak. A gyorsító billentyűkombináció az Alt-ta.

A párbeszédes ablakban először az átkódolandó változókat kell megadni a szokásos módon. Fontos, hogy megadjunk egy új nevet az átkódolt változónak (itt tehát most nem támaszkodhatunk egy ajánlott névre, valószínűleg a lehetőségek bősége miatt). A kis beviteli ablakba beírt nevet el kell fogadtatni a New Name gombbal. Ennek hatására egyrészt a jobb felső ablakban is megjelenik az új név, másrészt az OK gomb aktiválódik (ha minden kijelölt változónak van már új neve), tehát a parancs végrehajtható.

Ezek után már csak azt lehet beállítani, hogy nem értünk egyet az alapértelmezéssel: az újrakódolást mégis a legnagyobb értéktől kezdje. Az utasítás végrehajtásáról jelentést kapunk egy új output ablakban, ami tartalmazza a régi és az új kódok összefüggését.

összefoglalva, az újrakódolási parancs abban az esetben hatékony, ha csak kevés értéket kell átírnunk, illetve akkor elkerülhetetlen megoldás, ha a régi és az új értékek összefüggése nehezen adható meg képlettel.

Változók kategorizálása

A Visual Binning parancs egyszerű műveletet hajt végre: egy megadott korábbi változó értékeit kategóriákba sorolja. Ezután új változót hoz létre, amely ezen kategóriák sorszámát tartalmazza. A gyorsító billentyűkombináció az Alt-tb.

A kapott párbeszédes ablak is egyszerű: csak kétféle beállítást tehetünk meg. Először azon változókat kell kijelölnünk, amelyekre a kategorizálást kérjük (szokásos módon át kell őket tenni a jobb oldali ablakba). Alul megadható, hogy a vizualizáció során hány értéket jelenítsen meg. Ezek után a Continue gomb hatására egy újabb részletező párbeszédablakot kapunk.

Fenti részen állíthatjuk be az új változóra vonatkozó információkat, úgy mint a neve és címkéje. Alatta hisztogramszerűen jelenik meg az adatsor. Az

alsó részen táblázatos formában adható meg, hogy a kategóriák határpontjai hol legyenek. Ezt megtehetjük a táblázatban kézzel, vagy a Make Cutpoints gomb segítségével kérhetünk automatikus beállítást is. A határértékek hovatartozását a Upper Endpoints részben adhatjuk meg, ahol az Included és Excluded között választhatunk aszerint, hogy a határpontot a csoporthoz értjük vagy sem. Miután megadtunk ilyen határpontokat megjelenik 1-1 vonal a felső ábrán is, jelölve a beállított határokat. Végül a fentiek megadása után az OK gomb hatására a kategóriákba tartozó eseteket besorszámozza aszerint, hogy hányadik csoportba esett.

Az esetek rangsorolása

A Rank cases parancs, ahogy a neve is mutatja, az eseteket rangsorolja az alapul vett változó vagy változók értékei alapján. Ez a rangsorolás nagyon hasonló a sportban szokásoshoz: az is erre utal, hogy ha egy érték többször is előfordul (ún. *kapcsolt rang*ok), akkor nem pl. két első ranggal rendelkező eset lesz, hanem kettő olyan, amelynek rangszáma 1.5 (az alapértelmezésben). A gyorsító billentyűkombináció az Alt-tk.

A párbeszédes ablakban először a figyelembe vett változókat kell megadni. Jelen esetben ezeknek nem a kombinált értéke alapján képződik a rangsor, hanem mindegyik változóra kapunk egy-egy rangsoroló új változót.

A rangsorolást kérhetjük részcsoportokra is. Azt a változót, vagy változókat, amelyek szerint a csoportosítás történik, a jobb alsó ablakban (By) kell megadni. Erre a célra korábbi parancsokkal kialakított olyan változók alkalmasak, amelyekre az egyes csoportokon belüli érték azonos. Ha itt több változót adunk meg, akkor azok nem egyenként felelnek meg a rangsorolandó változóknak, hanem a megadott csoportképző ismérvek azonos érték n-esei jelölik ki a csoportokat. Tehát például az $x_1 = (1, 1, 2, 2, 3, 3)^T$ és a $x_2 = (1, 2, 1, 2, 1, 2)^T$ változók hat csoportot adnak meg.

A bal alsó sarokban lévő rádiógombokkal lehet azt megadni, hogy az egyes rang melyik értékhez tartozzon: a legkisebbhez vagy a legnagyobbhoz. Ezzel a növekvő vagy csökkenő sorrendet adjuk meg. Az összegző táblázatokat egy ettől jobbra levő pipával adhatjuk meg. Az eredmény egy külön output ablakban jelenik meg. Az output ablak tartalmát lehet szerkeszteni, melyet később tárgyalunk. Az ablak becsukása és szerkesztése nem befolyásolja az eredeti adat további feldolgozást.

A rangszámokon kívül még számos mutatót is meghatározhatunk, ezek a Rank Types gombbal, vagy a k gyorsítóbillentyűvel érhetők el, eredményük részben új változókban fog megjelenni, részben az output ablakban. Ezeknek részleteit nem tárgyaljuk.

A Ties gomb megnyomása után módosíthatjuk azt az alapértelmezést, hogy azonos értékű változókra a rájuk jutó rangszámok átlagát kérjük. A lehetséges alternatívák: ezek kapják a rájuk jutó legkisebb rangszámot (Low) vagy a legnagyobbat (High), vagy pedig az ilyen esetek után a következő rangszámot kérjük kiadni (és nem az azonos értékkel rendelkezők számához igazítottat).

Idősorok létrehozása

Két lehetőségünk van idősorok létrehozására a Transform menüben. A Date and Time wizard segítségével létrehozhatunk idősorokat már valamilyen formában megadott időkből. A varázsló minden részletét nem vizsgáljuk most. Alapjában véve itt tudjuk a korábban helytelen formában tárolt időket valós időként kezelhető idősorrá konvertálni. Ezt megtehetjük, ha például egyetlen vagy több szöveges mezőben van az időnk tárolva, vagy egy más típusú időnk van meg.

Egy szöveges mezőben tárolt idő esetén megadandó, hogy mely szöveges változók tartalmazzák az időket, illetve milyen formában. Ezen formátumok – a korábban tárgyaltakhoz hasonlóan – főleg az angolszász időformátumokat veszik alapul. Abban az esetben, ha több mező tartalmazza az esethez tartozó időket, akkor minden egyes változó egyetlen értéket tartalmazhat az alábbiak közül: év, hónap, nap, év napja, hónap napja, óra, perc, másodperc. Ezekhez egy-egy változó társítható és ebből rakja össze az esethez a megfelelő időt.

Említést érdemel még a Calculate with dates and times parancs. Ezt akkor érdemes használni, ha olyan információ áll rendelkezésünkre egy változóban, hogy az adott eset egy adott időpont után mennyivel lett mintavételezve. Ezen információból elegendő, hogy ki tudjuk kalkulálni a esemény időpontját, melyhez meg kell adjuk az ismert adott. Itt van lehetőségünk arra is, hogy egy esethez két változóban tárolt időpont között eltelt időt meghatározzuk, melyet egy új változóba menthetünk.

Az Extract a part of a date or time variables segítségével egy meglévő időből nyerhetünk ki részleteket. Például olyan változóból, ami tartalmazza a napot és azon belül az időpontot is (date and time), egy új változóba kinyerhetjük csak a napot, vagy akár csak az időt. De akár azt is megtudhatjuk, hogy az az év hányadik napja.

A fenti példák is mutatják, hogy a varázsló sokféle lehetőséget képes lekezelni. A varázsló elég segítőkész a feladatok megoldásában, ezért javasoljuk a próbálgatást az adott feladat esetén.

A másik lehetőség a Create Time Series parancs, mely egyben a varázsló utolsó lehetősége is. Itt idősor jellegű változóból más típusú idősort lehet létre

hozni. A kiindulási változó leírásában numerikus típusú kell hogy legyen. A változó tartalma pedig olyan, amely egymásra következő értékei és az eredeti sorrend azonos. Ilyen például a részvények árfolyama napokon át, vagy a növények magasságadatai időrendben stb. A gyorsító billentyűkombináció az Alt-tm.

A kapott párbeszédes ablakban a feldolgozandó változókat kell kiválasztani a szokásos módon. Minden ilyen kiválasztás során automatikusan új nevet is kapnak az új változók, amit a felhasználó megváltoztathat. A jobb oldali ablakban egy egyenlőségjel után azt is láthatjuk, hogy melyik függvény fogja az átalakítást végezni. Ezt a függvényt (Function) a középen levő listából választhatjuk ki. Vigyázat, a listából való választás után még meg kell nyomnunk a Change feliratú gombot ahhoz, hogy a változás érvényes legyen! Ha ezt elfelejtenénk, akkor a program figyelmeztet erre.

Az alapértelmezés az eltérés: két szomszédos érték különbsége lesz az új idősor egy-egy eleme. Több érdekes függvényt használhatunk még, mint például a többféle mozgó átlag (fontosak a tőzsdei technikai elemzésben), vagy a simítás. Egyes függvényekhez beállítható paraméterek vannak: az eltérésekhez annak rendje (Order), a mozgó átlagokhoz az alapul vett időszak hossza (Span).

A hiányzóadat kódok pótlása

A Replace Missing Values parancs arra alkalmas, hogy ha valamely eljáráshoz feltétlen szükség van minden eset számára érvényes értékre – amikor tehát nem felel meg a hiányzóadat kód –, akkor ez az utasítás ezekre a helyekre többé-kevésbé elfogadható értékeket generál. Ezt tehát csak akkor használjuk, ha feltétlenül szükséges. A gyorsító billentyűkombináció az Alt-tv.

A párbeszédes ablakban meg kell adnunk azokat a változókat, amelyekre ezt az átalakítást kérjük. Ezek itt is új nevet kapnak, amelyeket a program javasol, de a felhasználó szükség esetén módosíthat. A lehetséges értékek, amik a hiányzóadat kódok helyére kerülhetnek: a teljes átlag (Series mean), a szomszédos pontok átlaga (Mean of nearby points), a szomszédos pontok mediánja (Median of nearby points), lineáris interpoláció (Linear interpolation) és az adott pontra vonatkozó lineáris trend (Linear trend at point). Ahol ez szóba jöhet, ott az érintett pontok számát paraméterként meg lehet adni. A lineáris interpoláció a szomszédos pontokon alapul, a lineáris trend pedig a teljes sorozaton.

2.3. Az SPSS outputja

2.3.1. Az output ablakok tartalmának szerkesztése

Az előző szakaszokban sokat foglalkoztunk az output ablakokkal, az eredményeket sokszor így kapjuk meg. Az output ablak tartalmát sokoldalúan lehet szerkeszteni, és azt más programokba átvinni. Most ezeket a lehetőségeket tárgyaljuk.

Az output ablak nagyon hasonló külsejű, mint maga az SPSS program, de kisebb eltérések előfordulnak köztük. Hasonlóan egy másik adatállomány megnyitásához, új ablak jelenik meg. A Window menüsorban csak a megnyitott ablakok közül választhatunk, illetve kérhetjük azok mindegyikének összecsukását (a tálcán megtalálhatók maradnak). Az ablakok közötti váltásra a más hasonló programok esetén is szokásos Alt-w1, Alt-w2 stb. gyorsító billentyűk is használhatók.

A menüsorok kettő kivételével megegyeznek. Az output ablak kiegészül az Insert és Format menüpontokkal. Az egyes menüsorok tartalma is eltér, például a File menüsor csak az output ablak esetén tartalmaz nyomtatási kép utasításokat (Page Attributes és Page Setup). Az új menüsorokat itt tárgyaljuk, a többi használatára vonatkozó tudnivalókat pedig a munkalapokra vonatkozóan adtuk meg. Annál is inkább, mert például az Analyze menüsor mindenképpen az adatállományunkat érinti.

Az Insert menüsor különböző formázójeleket tartalmaz, illetve az outputhoz tartozó további szöveg, ábra beillesztését támogatja. Az első két utasítás a lapvége jel beszúrását, illetve ennek eltávolítását váltja ki (Page break, illetve Clear page break). A következő csoportban levő parancsokkal új szakaszt tudunk kezdeni, címet megadni, lapcímet beírni, és szöveget bevinni: New Heading, New Title, New Page Title, illetve New Text. Ezekkel az utasításokkal a program által létrehozott kimeneti adatot kedvezőbb alakra hozhatjuk. Minden újonnan bevitt szövegszerű objektumhoz tartozik egy olyan keret, amellyel rajzolóprogramokban találkozhatunk. Ennek segítségével adhatjuk meg a megfelelő formát, keretet a bevitt szövegnek. A beszúrási utasítások elérhetők a szerkesztési sor alján lévő ikonokkal is.

A következő csoportban levő parancsokkal szöveget (Text File) és ábrákat (Image) tartalmazó állományokat tudunk az output ablakba bevinni. Ezeket a lehetőségeket itt nem tárgyaljuk részletesen, annyit azonban érdemes tudni, hogy a fájl keresőben kiválasztott állományt szúrja be.

A Format menüsor mindössze három utasítást tartalmaz. Ezek csak akkor aktivizálódnak, amikor az output ablakban valamely objektumot (szövegrészt, ábrát vagy táblázatot) kijelöltük. Ezek után az Align Left pa-

ranccsal azt a lapon belül balra (a gyorsító billentyű Ctrl-[), a Center paranccsal középre (a gyorsító billentyű Ctrl-e), az Align Right paranccsal pedig jobbra igazíthatjuk (a gyorsító billentyű Ctrl-]). A formázás hatását közvetlenül nem láthatjuk, csak egy kis jel jelenik meg (a balra igazítás kivételével, ami alapértelmezés) a kijelölt objektum előtt. A tényleges hatást a nyomtatási kép (Print Preview) paranccsal nézhetjük meg.

2.3.2. Az output ablak tartalmának nyomtatása

Az output ablak tartalmának legegyszerűbb, talán leggyakoribb használati módja annak nyomtatása. Minden esetre a File menüsorból a Print utasítást választva az output ablak tartalma megjeleníthető a beállított alapértelmezett nyomtatón.

A programmal való gyakorlás során sokszor előfordul, hogy nem áll rendelkezésre nyomtató, például azért, mert az illető számítógépes teremben ilyen nincs is. Ekkor azt ajánljuk, hogy a nyomtatást irányítsuk fájlba (a megfelelő nyomtató operációs rendszerbeli beállítása szükséges).

A nyomtatáshoz négy utasítás tartozik, mindegyik a File menüsorból érhető el: Page Attributes, Page Setup, Print Preview és Print. Az utóbbi gyorsító billentyűje a Ctrl-p, és a szokásos ikon is megtalálható az eszközsorban.

Oldalbeállítás

A Page Attributes parancs hatására két dolgot állíthatunk be. Az első fülön a fej- és lábléc tartalmát. Erre két szöveges mező áll a rendelkezésünkre, melyeket a középen található gombokkal formázhatunk meg, illetve ezek segítségével szúrhatunk be dinamikus adatokat, olyanokat, mint például az aktuális dátum és idő, vagy az oldalszám. A második fülön az ábráink nyomtatáskori méretét specifikálhatjuk: ez lehet akkora, amekkora az ábra jelenlegi mérete az output ablakban (As is, ez az alapértelmezés), teljes-, fél- vagy negyedlap magasságú. Itt mondhatjuk meg az objektumok távolságát (pontban mint hosszegységben: 72 pont egy hüvelyk, a szokásos írógépes betű 12 pontos), és azt, hogy a nyomtatás lapszámozása melyik számmal kezdődjön (tehát nem azt, hogy melyik oldaltól kezdve történjen a nyomtatás). Az utóbbi beállításokat alapértelmezetté tehetjük a Make Default gomb megnyomásával.

A Page Setup a nyomtatási oldalak beállítását támogatja. A hozzá tartozó párbeszédes ablakban találjuk egy mintaoldal távoli képét, ez természetesen

követi a módosításokat. Itt adhatjuk meg a papír méretét és forrását, a tájolást (álló vagy fekvő) és a négy margó szélességét is.

További opciók érhetők el a Printer, Properties lenyomása után kapott ablakokban. Az előbbiben a szokásos módon megadhatjuk, hogy a telepített nyomtatók közül melyiket választjuk. Vigyázat, akkor is telepítsünk, ha konkrétan nincs elérhető nyomtató, mert bizonyos szolgáltatások, mint a nyomtatási kép enélkül esetleg nem lesznek elérhetők. Az egyes nyomtatók speciális beállításait is itt tehetjük meg, de erre van lehetőségünk a nyomtatás megkezdése előtt is.

Az SPSS nem támogatja a páros és páratlan oldalak megkülönböztetését, valamint például a kötés miatt szokásos belső margót sem, ami szintén a páros és páratlan oldalakon hol a bal, hol a jobb oldalon jelenne meg.

A nyomtatási kép

A Nyomtatási kép (Print preview) parancshoz gyorsító billentyűkombináció az Alt-fv, de a szokásos ikon megtalálható az eszközsorban is (ami egy papírlap fölötti nagyítót ábrázol). Ez az utasítás különösen betanulás során fontos, hiszen sok esetben nem is lesz lehetőség az output ablak tartalmának nyomtatására. Ráadásul néhány formázás csak jelzésszerűen jelenik meg az output ablakban, és csak a nyomtatási kép mutatja meg a tényleges hatásukat (mint például a jobbra igazítását).

A nyomtatási kép ablaka más programokból már szokásosnak mondható. A Zoom in és Zoom out gombokkal a megjelenített oldalt nagyíthatjuk és kicsinyíthetjük szűk határok között, a Close gombbal pedig bezárhatjuk a nyomtatási kép ablakot. Ez utóbbival visszakerülünk az output ablakba. A lapok között (ha több van) a Next Page, Prev Page feliratú gombokkal mozoghatunk. Arra is lehetőség van, hogy egyszerre lássunk két lapot. Az egy- és a kétlapos megjelenítés között a Two page / One page gomb vált (ugyanaz a gomb, a felirata változik értelemszerűen). A képre klikkelve is változik a nagyítás mértéke és az egy, illetve kétlapos megjelenítés.

Bár a menüsorok nem érhetők el a nyomtatási kép ablakból, de a legfontosabb nyomtatási- és oldalbeállítási parancsok itt is kiadhatók egy-egy megfelelő feliratú nyomógombbal.

Nyomtatás

Nyomtatást a Print utasítással kérhetünk. Ez a szokásos módon a menüből is elérhető (általában ez szokott lenni a legteljesebben paraméterezhető változat), de gyorsító gombbal is: Ctrl-p, és egy nyomtatót

ábrázoló ikonnal is. Az SPSS mindhárom változatban ugyanazt a párbeszédes ablakot adja (sőt, a munkalapról indítva is ezt kapjuk), tehát nem támogatja a „mindent a szokásos módon nyomtatni" lehetőséget.

érdekes, hogy az output ablakra vonatkozó nyomtatási ablak több beállítást tartalmaz, mint az, amelyik a munkalapról érhető el. Ismét felhívjuk a figyelmet a fájlba való nyomtatásra, ami akkor a legfontosabb, ha az éppen használt géphez nincs nyomtató kötve.

Ezekkel a beállítási lehetőségekkel együtt az SPSS-ből való nyomtatás nem feltétlen javasolható: más programmal jobban a kívánt alakra formálhatjuk a statisztikai program adta eredményeket.

Adatok, eredmények kivitele az output ablakból

Ha az SPSS által meghatározott statisztikai mutatókat, a statisztikai eljárások eredményeit, illetve az előállított grafikonokat más programba szeretnénk bevinni, ilyennel szeretnénk feldolgozni, akkor vagy az Export parancsot kell kérni a File menüsorból, vagy a vágólapon (clipboard) keresztül tudunk másolni.

A vágólapra az output ablak kijelölt részét többnyire a szokásos utasításokkal lehet bevinni. Ezek az Edit menüsorban találhatók: Cut (Ctrl-x), Copy (Ctrl-c) és Paste After (Ctrl-v). Ezeken felül az ábrák mozgatására a Copy Special (Ctrl-y) parancs alkalmas. Ebbe a csoportba tartozik még a vágólapra másolás nélküli törlés: Delete (Del) és a Paste Special, amivel irányítottan tudunk az output ablakba beilleszteni saját objektumként. Az output ablakban mindent ki tudunk jelölni a Select All (Ctrl-a) paranccsal.

A mindennapi munkában gyakori a vágólapon keresztüli kommunikálás az SPSS program és a további feldolgozást végző programok között. Ebben az esetben ugyanis nem kell új adatállományokat létrehozni, azok nevét meghatározni és a másik programban azt megadni vagy kikeresni. Ugyanakkor ilyen esetben nem marad másolat a statisztikai programmal létrehozott táblázatról vagy ábráról, így ha például a szövegszerkesztő program eredményét elveszítjük, akkor a statisztikai vizsgálatot meg kell ismételni.

Az Export utasítás ezzel szemben a jelentés, output adatsor kijelölt részét közvetlenül egy adatállományba írja, ahonnan aztán be lehet tölteni azt egy alkalmazói programba. Ezt a parancsot a File menüsorban találjuk, a gyorsító billentyű az Alt-ft.

A végrehajtása azzal kezdődik, hogy egy párbeszédes ablakban a szükséges adatokat meg kell adnunk. Ezek közül az első az, hogy mit szeretnénk kimenteni: az egész dokumentumot (All), az összes láthatót (All vi-

sible), vagy csak a kijelölt részt (Selected). A rejtetteket az output ablak bal oldalán lévő összefoglaló diagramban (outline) lehet kijelölni a + és - jelek használatával. Ezek után meg kell adnunk a létrejövő fájl típusát, és nevét könyvtárral együtt. Ehhez segítséget ad a jobb oldalon levő Tallóz (Browse) gomb.

Az adatkivitel formátumát a Change Options gomb hatására tisztázhatjuk. A lehetőségek függnek attól, hogy milyen típusú állományba akarjuk kivinni az adatokat.

Amennyiben az output állományunk típusa nem támogatja a képek beágyazását (például sima szöveges állomány, azaz txt), akkor lehetőségünk van a képekre vonatkozóan is beállításokat eszközölni.

További finomabb beállításokat tehetünk a Change Options gomb megnyomása után. Utóbbival például az ábrák méretét adhatjuk meg, hogy a jelenleginek hány százaléka legyen a ténylegesen kivitt. Az előbbi beállítási lehetőségei a kiválasztott fájltípustól függenek.

Az output ablakba tehát közvetlenül is beírhatunk szöveget, így abban elvileg egy teljes értékű jelentést is létrehozhatunk, mégis ez nem szokásos a kifinomult szövegszerkesztő és táblázatkezelő programok miatt.

2.3.3. Grafikonok

A statisztikai programok egyik fontos funkciója az adatok, vagy az eredmények grafikonon való megjelenítése. Sok esetben konkrét statisztikai feldolgozás nem is történik, csak az alkalmas prezentáció a fontos. Az SPSS grafikonszerkesztője meglehetősen ügyes, komoly támogatást nyújt az ábrák kialakításához, mégis más programok is számításba vehetők, ha a kívánt ábraformát nem találjuk a kínálatban, vagy ha valamely megjelenítési részletet nem tudjuk kialakítani. Ekkor az adatunkat vagy eredményünket vigyük át egy táblázatkezelő, vagy egy prezentációt támogató programba.

Az ábrák kialakítását a Graphs menüsor utasításai támogatják. Más programokhoz, így a táblázatkezelőkhöz hasonlóan itt is több úton juthatunk ugyanahhoz az ábrához. Az egyik lehetséges mód az, hogy egy ún. galériából választjuk ki azt az ábrafajtát, ami az elképzeléseinknek megfelel, és a részleteket ezután tisztázzuk egy megfelelő párbeszédes ablakban. Egyik ilyen menüpont a Chart Builder. Ekkor a megjelenő párbeszédes ablak alsó felében a Gallery fülön kiválaszthatjuk azt az ábra változatot, mely az ábrázolni kívánt adatokhoz a leginkább megfelelő. Majd ezek után lehetőségünk nyílik interaktívan a grafikonunk beállításaira. Ennek a lehetőségnek a gyorsító billentyűkombinációja az Alt-gc.

A következő fokozatot az Graphboard Template Chooses menüpontból érjük el. Itt az első fülön kiválaszthatjuk a megjelenítendő adatokat, majd ennek függvényében ajánlatokat tesz a megjelenítésre. A kiválasztott forma további beállításait a következő füleken tehetjük meg. A gyorsító billentyűkombináció Alt-gg.

Ezzel szemben a legegyszerűbb mód az, amikor a Graphs / Legacy Dialogs menüben felsorolt grafikon változatok közül kiválasztjuk a kívántat, és a kapott párbeszédes ablak sorozattal adjuk meg a szükséges paramétereket, beállításokat. Az utóbbi esetben az ábra típusától függ a gyorsítóbillentyű kombináció, az oszlopdiagramra pl. Alt-glb. Bár a közvetlen ábrarajzolási eljárások csak a legfontosabb információkat kérik, és emiatt egyes finomabb részleteket nem is tudunk befolyásolni, vannak olyan beállítások, amiket ezen a módon meg tudunk tenni, míg az interaktív módon nem. A rajzolható ábratípusok köre is eltér kissé az utóbbi két esetben.

Az SPSS ábrarajzolási eljárásainak hátránya, hogy a beállítások közben nem látjuk a korábbi megadások eredményét, és arra sincs lehetőség hogy egy diagramvarázslóval a korábban létrehozott grafikonokon módosítsunk (mint egyes táblázatkezelőkben). Bár az output ablakban elérhetők a Graphs menüsor utasításai, azok az eredeti adatokra vonatkoznak, nem az épp létrehozott ábrára. Fontos megjegyezni, hogy a létrejött grafikonok az export utasítással sok formátumban kimenthetők, és más programmal feldolgozhatók, pl. szövegbe is beilleszthetők.

A grafikonok összeállítása érthető módon nagyon összetett lehet. A későbbiekben egy-egy példán keresztül világítjuk meg hogy hogyan kell ezeket az eljárásokat használni. Sajnos több esetben előfordul, hogy amit el lehet érni a grafikonszerkesztővel, azt egy adott szerkesztési mód (Bar) nem teszi lehetővé. Ilyen esetben a súgó segítségével közvetlenül a végrehajtandó szintaxisú utasításokat is megírhatjuk.

3. fejezet

Statisztikai eljárások és grafikus megjelenítések

Az Analyze menüsor tartalmazza a statisztikai eljárásokat. Itt megtalálhatóak a leggyakoribb statisztikai algoritmusok. A még nem érintett további menük kiegészítő jellegűek. Ilyen a segédprogramokat tartalmazó Utilities menü, a kiegészítők által beillesztett menüpontok az Add-ons részben, valamint az ablakkezelést segítők szokványos menűpontot a Window alatt. Végül sokféle súgó és leírás található a Help menüben. A gyorsító billentyűkombináció az Analyze menüsorhoz az Alt-a, míg a Graphs menüsorhoz az Alt-g.

Az előzőek után kevésnek tűnhet hogy csak egy menüsorra való eljárás van beépítve, de az Analyze menüsor valójában nem közvetlenül parancsokat tartalmaz, hanem almenüket a feladatkörök szerint csoportosítva. Didaktikai szempontból nagyon dicséretes, hogy ez a beosztás a statisztikai eljárásokat megelőző lépések relatív fontosságát hangsúlyozza.

Az almenük sorrendje is fontos: ez lényegében követi a szokásos feldolgozási sorrendet. Bár a magasabb rendű statisztikai eljárások végrehajtásához nem feltétlenül szükséges a megelőzőkből akár csak egyet is végrehajtani, mégis, az alapos, szakszerű statisztikai feldolgozás rendszerint követi az itt is látható rendet.

Az itt megadott rengeteg parancsot nem tekintjük át teljes egészében, csak a legfontosabbakra kerítünk sort. Több menüsorból csak néhány eljárást tárgyalunk. Ez persze nem azt jelenti, hogy ezek a módszerek nem fontosak, de a jelen bevezető kurzus keretében nem marad elég idő erre. A sok eljárás megemlítése helyett, inkább a kevesebb eljárás kicsit mélyebb megismerése tűnt hasznosabbnak. A kiválasztás során szempont volt az eredmények látványossága is, hogy egy statisztikai eljárás hasznossága mennyire nyil-

vánvaló egy kezdő számára.

Ha önálló munka során olyan statisztikai eljárást kell használnunk, amelyet itt nem tárgyalunk részletesen, akkor támaszkodjunk a program súgójára, a beépített leírásra (Help / Tutorial), vagy használjuk a Help / Statistics Coach tanácsait.

3.1. Jelentések és leíró statisztikák

Ezekben a menüsorokban olyan parancsok vannak, amelyekkel az adatállományunk változóinak alapvető statisztikai mutatóiról lehet jelentést kérni, illetve ezeket a táblázatba bevinni. A jelentés, amit itt kapunk, nem olyan kiterjedt, mint amit a SigmaPlot ad. Magyarázó, értelmező szöveget nem kapunk, csak a statisztikai adatok, eredmények táblázatait, esetenként a jelölések megfejtésével. A gyorsító billentyűkombináció az Alt-ap, illetve az Alt-ae.

A jelentésírás (Reports) menüponthoz a következő parancsok tartoznak: Codebook (változó tulajdonságai), OLAP Cubes (összefoglaló táblázatok kategóriák szerint), Case Summaries (eset összefoglalók), Report Summaries in Rows (összefoglaló jelentés sorokban) és Report Summaries in Columns (összefoglaló jelentés oszlopokban). Ezek közül csak az eset összefoglaló utasítással foglalkozunk részletesebben, a többi inkább formai eltérést mutat – ami azért fontos lehet, ha nagy mennyiségű adat feldolgozásakor időt takaríthatunk meg a szerkesztés során. Az említett parancsok kicsit eltérhetnek az elérhető statisztikai mutatókat illetően, de mindegyik megenged egy változó értékével meghatározott csoportosítást.

Hasonló módon a Leíró statisztikák (Descriptive Statistics) menüsorból is csak egy parancsot tárgyalunk részletesen, a gyakoriságok (Frequencies) nevűt. A rendelkezésre álló utasítások a Frequencies, a Descriptives, az Explore, a Crosstabs, Ratio, P-P Plots, és a Q-Q Plots.

3.1.1. Eset összefoglalás

Az Analyze / Reports / Case Summaries menüpont kiválasztása után kapott párbeszédes ablakban először a szokásos módon, a változók áthelyezésével meg kell adnunk azokat a változókat, amelyekre az összefoglalást kérjük, illetve amelyek a csoportosítást meghatározzák. Az eljárás eredménye egy output ablakba kerül, többszöri végrehajtása ugyanabba az ablakba írja a táblázatokat.

A változókat tartalmazó ablakok alatt azt állíthatjuk be, hogy kérjük-e az egyes esetek felsorolását is (Display cases). Mivel a program nagy számú esetre van felkészülve, ezért itt korlátozni tudjuk a figyelembe vett, megjelenítendő esetek számát (Limit cases to first). Az első, a kis ablakban megadott számú eset fog így a jelentésünkben szerepelni. A beállított esetszámot az output ablakbeli táblázataink címének lábjegyzete adja meg: pl. Limited to first 100 cases. A következő sorban azt pipálhatjuk ki, hogy csak érvényes eseteket szeretnénk megjeleníteni (Show only valid cases). Ilyenkor tehát a hiányzóadat kódot tartalmazó esetek nem kerülnek be a listánkba. Végül az utolsó helyen kérhetjük az esetsorszámok kiíratását is (Show case numbers).

A további beállítási lehetőségek egy része az Options gomb megnyomásával érhető el. Itt az ablakokban megadhatjuk az eseteket összegző táblázat címét (Title, lehet magyarul is, az ékezetes betűk is elfogadottak) és a táblázat magyarázó szövegét (Caption). Ezután kérhetjük, hogy az összegsort a Totals felirattal lássa el a program (Subheadings for totals), és azt, hogy zárjon ki minden olyan esetet a felsorolásból, amelyben akár csak egy hiányzóadat kód is van (Exclude cases with missing values listwise). érdekes lehetőséget kínál az utolsó beállítási pont (Missing statistics appears as): itt azt a szöveget adhatjuk meg, amely azokba a sorokba kerül, amelyek hiányzóadat kódot tartalmaznak. Ez a szöveg ismét lehet magyarul is, és tartalmazhat ékezetes betűt is. Az utóbbi szó balra igazítva jelenik meg (mint ahogy a szövegeket általában is írni szokás) a jobbra igazított számok között.

A másik nyomógomb a Statistics feliratot viseli, és ezzel lehet az esetek összegzéséhez különböző statisztikai mutatókat kérni. A kívánt statisztikai mutatókat a bal oldali listából át kell tenni a jobb oldali, Cell Statistics feliratú ablakba. A rendelkezésre állók (ebben a menüpontban): átlag, medián, standard hiba, összeg, minimum, maximum, terjedelem, az első és az utolsó elem az előfordulás sorrendje szerint, szórás (az SPSS-ben tapasztalati korrigált), variancia, csúcsosság, a csúcsosság standard hibája, ferdeség, a ferdeség standard hibája, harmonikus átlag, mértani átlag, a teljes összeg százaléka és az összes esetek számának százaléka. Az utóbbi két mutató csoportonként értendő. Ebben a listában a párbeszédes ablakbeli sorrendet követtük. A jobb oldali ablakban megadott sorrend az eredmények megjelenési sorrendje is lesz.

A gyorsító billentyűkombináció az Alt-apm.

3.1.2. A gyakoriságok (Frequencies) parancs

Az Analyze / Descriptive Statistics / Frequencies menüpont kiválasztása után vagy az Alt-asf billentyűkombináció után egy párbeszédes ablakot ka-

punk. Először azt kell megadnunk, hogy mely változókra kérjük a gyakoriságok meghatározását. A legegyszerűbb kérésünk az lehet, hogy a program adja meg a gyakorisági táblázatokat. Ehhez a változók listája alatt ki kell pipálni a Display frequency tables opciót. Az utóbbi inkább a diszkrét, míg a Statistics gombbal megadható lehetőségek (és a felajánlott grafikonok is) inkább a folytonos adatok összegzésére valók. Ha megadtunk legalább egy változót, akkor az OK gomb aktiválódik, tehát végrehajthatjuk a parancsot.

Az eredmény az output ablakban jelenik meg. Az első táblázat a változókra megadja az érvényes értékek és a hiányzóadat kódok számát. Ezt követően minden megadott változóra kapunk egy táblázatot, ami a következő oszlopokat tartalmazza:

- A változó előforduló értékei, külön megadva az érvényeseket (Valid) és a hiányzóadat kódokat (Missing). Ezek az értékek nagyság szerint növekvő sorrendben állnak. A hiányzóadat kódok mindig az utolsó sorokba kerülnek, függetlenül az érvényes változó értékek nagyságától.
- Ezek relatív gyakorisága (Frequency, az előfordulásaik száma).
- Az előbbi relatív gyakoriságok százalékos értéke az összes esetre vetítve (beleértve a hiányzóadat kódokat is).
- Egy hasonló százalékos megoszlás, de ez esetben csak az érvényes értékekre vetítve (Valid Percent).
- Az érvényes esetekre vonatkozó kumulatív százalékos megoszlás (Cumulativ Percent): az adott értéket és az annál kisebbet felvett esetek relatív gyakorisága. A kumulált relatív gyakorisági sor is az értékek növekvő sorrendjében van megadva.

A táblázatok utolsó sora (Total) megadja az összes esetek számát, illetve a százalékok összegét (100%). Minden táblázat címe annak a változónak a neve, amire a gyakorisági kimutatás vonatkozik. Ezeket a táblázatokat további statisztikákkal tudjuk kiegészíteni. Ehhez nyomjuk meg a Statistics gombot a párbeszédes ablakon.

Itt négyféle statisztikai mutatót kérhetünk: percentilis értékeket (Percentile Values), középértékeket (Central Tendency), a szóródásra (Dispersion) és az eloszlás alakjára (Distribution) vonatkozókat.

A kvantilisekből három csoportból választhatunk: egyszerűen a 25, 50 és 75%-ra (alsó kvartilis, medián, felső kvartilis) vonatkozó elválasztó értékeket kérjük; vagy az egyes változók értékeit egy megadott számú azonos nagyságú

csoportra osztó értékeket kívánjuk látni; vagy pedig explicit megadjuk azokat a százalékos értékeket, amelyeknek megfelelő változóértékeket szeretnénk látni. Ezek a lehetőségek nem vagylagosak (ezt az is jelzi, hogy nem rádiógombbal kell őket kiválasztani): többet is megjelölhetünk közülük. Ilyenkor mindegyik kért értéket megkapjuk.

A párbeszédes ablak következő csoportjában azt állíthatjuk be, hogy melyiket kérjük kiszámítani és megjeleníteni az átlag, medián, módusz és összeg mutatókból. Ezek meghatározása során a hiányzóadat kódokat természetesen jelentésüknek megfelelően veszi figyelembe a program. Ha többször egymás után hajtjuk végre ezt a parancsot, akkor az eredményeket mindannyiszor ugyanabba az output ablakba kapjuk.

Lehetőség van arra is, hogy a kvantilis értékeket és a mediánt azon feltevés alapján határozza meg a program, hogy az illető változó értékei csoportosítás eredményeként az eredeti osztályok középpontjai (az ún. osztályközép). Ezt a helyzetet a Values are group midpoints kapcsolóval tudjuk az SPSS-sel közölni.

A változóink eloszlására vonatkozóan a következő mutatóknak a táblázatokba foglalását kérhetjük a programtól: szórás, variancia, terjedelem, minimum, maximum, standard hiba, ferdeség és csúcsosság. Ezek az értékek az első táblázatban jelennek meg a változókra vonatkozó érvényes és hiányzóadat kódot tartalmazó esetek számával együtt.

A gyakoriságok meghatározása mellett ugyanebben a párbeszédes ablakban kérhetjük grafikonok egyidejű előállítását is. A Charts gomb megnyomása után egy rádiógomb kombináció segítségével az ábra típusát adhatjuk meg. A választási lehetőségek (ezek most egymást kizárók): ne legyen grafikon, oszlopdiagramot kérünk, vagy kördiagramot, vagy pedig hisztogramot. Utóbbihoz külön kérhetjük egy illeszkedő normális eloszlási görbe ábrázolását is (Show normal curve on histogram).

Az oszlop- és a kördiagramok esetén döntenünk kell, hogy az ábrázolt értékek gyakoriságok vagy relatív gyakoriságok legyenek. Mindkét választás persze azonos arányokat eredményez, de az aktuális értékek mások lesznek. Az ábrák címe itt is az érintett változó neve lesz, az egyes oszlopok, illetve körszeletek felirata pedig az illető változóérték.

Az utolsó beállítási gomb, a Format feliratú azt teszi lehetővé, hogy a táblázataink egyes részleteit alakíthassuk. Az első rovatban azt tisztázhatjuk, hogy a táblázat soraiba milyen sorrendben kerüljenek az értékek. Az alapbeállítás – ahogy már írtuk is – az, hogy a változóértékek növekvő sorrendjében jelennek meg a gyakoriságok. Itt most ehelyett választhatunk csökkenő sorrendet, illetve az egyes értékek előfordulási gyakoriságai (counts) szerint is rendezhetjük a listát növekvő vagy csökkenő sorrendbe.

A Multiple variables rovatban azt határozhatjuk meg, hogy több változóra kért statisztikai mutatók táblázatos megadásakor több változóra vonatkozó adatokat egy táblázatba kérjük-e (Compare variables), vagy a változókra egy-egy táblázatot kérünk-e (Organize output by variables). Végül megadhatjuk azt is, hogy olyan táblázatokat nem szeretnénk megjeleníteni, amelyek több mint egy megadott számú különböző értéket tartalmaznak. Ez utóbbinak az az értelme, hogy az ilyen táblázatok áttekinthetetlenül hosszúak lehetnek. Az alapértelmezés a 10 különböző kategória legfeljebb (de az opciót még be kell kapcsolni ahhoz, hogy ennek megfelelően működjön a program).

3.1.3. Példa

Az első három változóba bevittük a következő 4-4 adatot: 1, 3, 2, 4; 0,30, -1,30, 0,50, -0,80; a, c, e, f. A harmadik változó típusát előzőleg szövegesre (string) kellett állítani. Erre az adatra kértük a Descriptive Statistics / Frequencies eljárást végrehajtani. A statisztikák közül mindent kértünk, kivéve a felhasználó által megszabott kvantiliseket. A kapott eredmények (a gyakorisági táblázatot csak az első változóra adjuk meg, mert lényegében a többi is megegyezik ezzel):

```
FREQUENCIES VARIABLES=VAR00001 VAR00002 VAR00003
  /NTILES=4
  /NTILES=10
  /STATISTICS=STDDEV VARIANCE RANGE MINIMUM MAXIMUM SEMEAN
   MEAN MEDIAN MODE SUM SKEWNESS SESKEW KURTOSIS SEKURT
  /ORDER=ANALYSIS.
```

Statistics

		VAR00001	VAR00002	VAR00003
N	Valid	4	4	4
	Missing	0	0	0
Mean		2,5000	-,3250	
Std. Error of Mean		,64550	,43277	
Median		2,5000	-,2500	
Mode		1,00a	-1,30a	
Std. Deviation		1,29099	,86554	
Variance		1,667	,749	
Skewness		,000	-,235	
Std. Error of Skewness		1,014	1,014	
Kurtosis		-1,200	-4,173	
Std. Error of Kurtosis		2,619	2,619	
Range		3,00	1,80	
Minimum		1,00	-1,30	

Maximum		4,00	,50	
Sum		10,00	-1,30	
Percentiles	10	1,0000	-1,3000	
	20	1,0000	-1,3000	
	25	1,2500	-1,1750	
	30	1,5000	-1,0500	
	40	2,0000	-,8000	
	50	2,5000	-,2500	
	60	3,0000	,3000	
	70	3,5000	,4000	
	75	3,7500	,4500	
	80	.	.	
	90	.	.	

a Multiple modes exist. The smallest value is shown

Frequency Table
VAR00001

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	1	25,0	25,0	25,0
	2,00	1	25,0	25,0	50,0
	3,00	1	25,0	25,0	75,0
	4,00	1	25,0	25,0	100,0
	Total	4	100,0	100,0	

A harmadik, szöveges változóra érthető módon csak az érvényes esetek számát lehetett megállapítani (és a gyakorisági táblázat is elkészült). Az első két változóra az egyszerűbb statisztikai mutatók (pl. átlag, terjedelem) értékét könnyen ellenőrizhetjük. A gyakoriságok táblázatában pedig az ún. felfelé menő kumulált relatív gyakoriságok oszlopát érdemes tanulmányozni: ez a példa megvilágítja a korábbi szöveges formában talán nehezebben érthető definíciót.

3.1.4. Gyakorlat

Az eset összefoglaláshoz tekintsünk a 3.1. táblázatban található 1992-es amerikai kosárlabda csapat adatait.

Vizsgáljuk meg a különböző pozíciókban lévő játékosok testmagasságát és tömegét. Ehhez készítsünk egy esetösszefoglaló táblázatot az SPSS Case Summaries parancsával. A párbeszédes ablakba adjuk meg, hogy a magasságról és a testtömegről szeretnénk adatokat. Ehhez a változók közül a

NAME	POS	HGT (in)	HGT (m)	WGT (lb)	WGT (kg)	AGE
Charles Barkley	F	78	1,98	250	113,25	29
Larry Bird	F	81	2,06	220	99,66	35
Clyde Drexler	G	79	2,01	222	100,57	30
Patrick Ewing	C	91	2,31	240	108,72	30
Earvin Johnson	G	81	2,06	220	99,66	32
Michael Jordan	G	78	1,98	198	89,69	29
Christian Laettner	F	83	2,11	235	106,46	22
Karl Malone	F	81	2,06	256	115,97	29
Chris Mullin	F	79	2,01	215	97,40	29
Scottie Pippen	G	79	2,01	210	95,13	26
David Robinson	C	85	2,16	235	106,46	27
John Stockton	G	73	1,85	175	79,28	30

3.1. táblázat. Az 1992-es amerikai olimpia csapat adatai.

megfelelő változókat (HGT és WGT) helyezzük át a Variables ablakba. Ezek után kérjük az adatok csoportosítását a csapatban betöltött pozíció szerint, melyet megtehetünk úgy, hogy a Grouping Variable(s) mezőbe behelyezzük a pozíciókat tartalmazó mezőt (POS). Majd a Statistics gomb megnyomása után állítsuk be, hogy az esetek számán (Number of cases) kívül szükségünk van az átlagra (Mean), illetve a legnagyobb (Maximum) és a legkisebb (Minimum) értékekre. A Continue gomb megnyomására visszatérhetünk az első párbeszédes ablakra, ahol az OK gombra kattintva az alábbi eredményt kapjuk:

Case Summariesa

				HGT_m	WGT_kg
POS	C	1		2.31	108.72
		2		2.16	106.46
		Total	N	2	2
			Mean	2.2350	107.5900
			Maximum	2.31	108.72
			Minimum	2.16	106.46
	F	1		1.98	113.25
		2		2.06	99.66
		3		2.11	106.46
		4		2.06	115.97
		5		2.01	97.40
		Total	N	5	5
			Mean	2.0440	106.5480
			Maximum	2.11	115.97
			Minimum	1.98	97.40

	G	1	2.01	100.57
		2	2.06	99.66
		3	1.98	89.69
		4	2.01	95.13
		5	1.85	79.28
		Total N	5	5
		Mean	1.9820	92.8660
		Maximum	2.06	100.57
		Minimum	1.85	79.28
	Total	N	12	12
		Mean	2.0500	101.0208
		Maximum	2.31	115.97
		Minimum	1.85	79.28

a Limited to first 100 cases.

Látható, hogy a C (center) pozícióban lévő emberek a legmagasabbak, míg a G (irányító) pozícióban lévő emberek a legalacsonyabbak a csapatban. A játékosok tömegeire hasonló már nem mondható el.

3.2. Átlagok összehasonlítása

Az Analyze / Compare Means menüpont azokat az eljárásokat foglalja össze, amelyek az átlagok összevetése alapján való összefüggést tisztázzák. Ide tartozik az egyszerű átlagszámítás csoportonként (Means), az egymintás és a párosított t-próba (One- és Paired-Sample T Test), a független mintás t-próba (Independent Sample T Test) és az ANOVA, vagy variancia analízis eljárás (One-Way ANOVA). Először a párosított t-próbát tárgyaljuk részletesen. Ez elérhető az Alt-amp gyorsító billentyűkombinációval is.

3.2.1. Párosított t-próba

Maga az eljárás két normális eloszlású sokaság várható értékének (illetve átlagának) eltérésének ellenőrzésére szolgál. Ez a próba két összetartozó (nem független) minta összehasonlítására alkalmas. Általában akkor kell ezt a próbát alkalmazni, ha ugyanazokra az esetekre ugyanazokat a változókat határozzuk meg két különböző szituációban (pl. egy gyógyszer hatását vizsgáljuk: a mért értékek a beadás előtti és az azt követő helyzetet

tükrözik). Az eljárás veszi a két változó azonos esethez tartozó értékeit, és azt vizsgálja, hogy ezek átlagának különbsége statisztikai értelemben eltér-e nullától.

Az utasítás kiadása után egy olyan párbeszédes ablakot kapunk, amely az eljáráshoz illeszkedő módon kéri a paraméterezést: a bal oldali változólistából most változó párokat kell kiválasztanunk és a jobb oldali ablakba vinni.A jobb oldali ablakban sorokban láthatók a kiválasztott változók. Egyszerre több párt is vizsgálhatunk, így minden esetben kapunk egy új üres sort, mely kijelölésével, újabb párokat hozhatunk létre. Sor kijelölése esetén, melyet a sor elején szereplő számra való kattintással tehetünk meg, lehetőségünk van az aktívvá váló nyilakkal az adott pár vizsgálati sorrendjét módosítani. A párok viszonyának felcserélését a kétirányú nyíllal tehetjük meg.

Az eljárással kapcsolatban megjelenő statisztikai mutatókat közvetlenül (ahogy eddig), egy gomb megnyomása után nem specifikálhatjuk, de automatikusan kiszámítódik minden változóra az átlag, a minta elemszáma, a szórás és a standard hiba. Ezen felül külön statisztikákat kapunk a megadott változópárokra, ezek: a korreláció, az átlagok átlagos eltérése, a t-próba és a konfidencia intervallum az átlagok eltérésére (ennek konfidenciaszintjét a felhasználó megadhatja). Az eltérés szórását és standard hibáját is megkapjuk.

A változópárok megadása után az Options gombot megnyomva kapunk egy újabb párbeszédes ablakot, amelyben az átlagokra vonatkozó becslések konfidenciaszintjét adhatjuk meg. Az alapértelmezés a szokásos 95%.

Egy rádiógombbal itt adhatjuk meg azt is, hogy a hiányzóadat kódokat hogyan kezelje a program: az érintett eseteket analízisről analízisre újra határozza meg (Exclude cases analysis by analysis), vagy törölje a feldolgozásból az összes olyan esetet, amelyben az érintett változókban hiányzóadat kódot talált (Exclude cases listwise).

Az eredmény értelmezése

Mivel most nem egyszerű statisztikai mutatókat kapunk eredményül, amiket valójában csak fel kell sorolnunk, amikor a statisztikai eljárás outputját összefoglaló jelentést írjuk, ezért részletesebben tárgyaljuk, hogy hogyan kell a kapott mutatókat értelmezni. Persze ez a rövid útmutató nem lehet teljes, ezért bővebb magyarázatot az irodalomjegyzékbeli általános, illetve matematikai statisztikai könyvekben találhatunk.

Az eljárás három táblázatot ad. Ezek közül az első a próba által megvizsgált változók alapvető statisztikai mutatóit tartalmazza. Ebben a páronként felsorolt változókra egyszerűen ellenőrizhetjük, hogy azok átlagai

mennyivel térnek el egymástól, valamint hogy a szórások közel azonosnak tekinthetők-e.

A második táblázat a változópárok korrelációit adja meg a figyelembe vett esetszámmal és a szignifikancia-értékekkel. A jegyzet statisztikai bevezetőjében említetteknek megfelelően a korrelációs együttható egy olyan mutató, amely az érintett két normális eloszlású változó lineáris összefüggését fejezi ki. Ha ennek értéke szélsőséges esetben nulla, akkor a változók egymástól függetlenek. Ha ez 1 vagy -1, akkor pedig a megadott változók között szigorúan vett függvényszerű kapcsolat mutatható ki. Az első esetben a két változó együtt nő, -1 esetén pedig az egyik növekedése a másik csökkenésével jár.

A szignifikancia ez esetben azt fejezi ki, hogy a korrelációs együtthatóra kapott érték mennyire valódi, és nem a véletlen műve. Ha a szignifikancia-szint nulla, vagy ahhoz közeli, akkor az eredményünk nagyon alátámasztott statisztikai szempontból. Jelen esetben ha a „Sig" alatt található szám, a p-érték nullához közeli, pl. kisebb mint 0,05 (ilyenkor általában a korreláció abszolút értéke egyhez közeli), akkor azt mondjuk, hogy a korreláció szignifikáns 5%-os szinten. Ez azt jelenti, hogy a korrelációs együttható eltér 0-tól, a két változó között lineáris összefüggés van. Ha a „Sig" alatt lévő szám nagyobb, mint a szignifikancia-szint, akkor a korreláció nem szignifikáns, a két változó között adott szinten adataink alapján nem tudunk lineáris összefüggést kimutatni.

A harmadik táblázat foglalja össze a t-próba tényleges eredményét. Az első oszlopban vannak felsorolva a kijelölt változópárok, a következő oszlopokban pedig a páronként vett eltérések statisztikai mutatói. A kiértékeléshez felhasználhatjuk a t-értéket, melyet a mellette levő szabadságfok (df) alapján táblázatból visszakereshetünk, vagy a szignifikancia értéke alapján dönthetünk. Ha ez utóbbi kicsi vagy nulla, akkor azt mondjuk, hogy az eltérés szignifikáns, a két átlag különbsége szignifikánsan eltér nullától. Ilyenkor az átlagok különbségére vonatkozó konfidencia intervallum nem tartalmazza a nullát. Ellenkező esetben, ha a „Sig (2 tailed)" alatt látható szám nagy, nagyobb, mint például 0,05, akkor azt mondjuk, hogy a két változó átlaga között az adatunk alapján nem tudunk 5% hiba mellett különbséget kimutatni.

Szöveges választ vagy értelmezést az SPSS nem ad. Hasonló esetben a SigmaStat ilyen mondatot ír például: „The change that occurred with the treatment is not great enough to exclude the possibility that the difference is due to chance (P=...).". Azaz az a változás, ami a kezelés következtében adódott, nem elég nagy ahhoz hogy kizárjuk annak a lehetőségét, hogy az eltérés a véletlen műve (P=...). Az ilyen szöveges értelmezés jelentősége vitatott, a szerzők inkább hasznosnak tartják, de a gyakorlatlan felhasználót

félre is vezetheti. Remélhetőleg az SPSS is hamarosan szolgáltat ilyen pontos értékelő szöveges összefoglalót. Ez számos félreértés elkerülését segíti, és példát is ad a pontos, óvatos fogalmazásra.

3.2.2. Példa

Tekintsük az alábbi rövid adatsort (1, 3, 2, 4; 0,30, -1,30, 0,50, -0,80; 1, 0, 0, 1), és kérjük az összes változópárra a páros t-próbát. A kapott eredmény a következő.

```
T-TEST PAIRS=VAR00001 VAR00002 VAR00001 WITH VAR00002 VAR00003 VAR00003 (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.

Paired Samples Statistics
|---------------|------|-|--------------|---------------|
|               |Mean  |N|Std. Deviation|Std. Error Mean|
|------|--------|------|-|--------------|---------------|
|Pair 1|VAR00001|2,5000|4|1,29099       |,64550         |
|      |--------|------|-|--------------|---------------|
|      |VAR00002|-,3250|4|,86554        |,43277         |
|------|--------|------|-|--------------|---------------|
|Pair 2|VAR00002|-,3250|4|,86554        |,43277         |
|      |--------|------|-|--------------|---------------|
|      |VAR00003|,5000 |4|,57735        |,28868         |
|------|--------|------|-|--------------|---------------|
|Pair 3|VAR00001|2,5000|4|1,29099       |,64550         |
|      |--------|------|-|--------------|---------------|
|      |VAR00003|,5000 |4|,57735        |,28868         |
|-------------------------------------------------------|

Paired Samples Correlations
|--------------------------|-|-----------|-----|
|                          |N|Correlation|Sig. |
|------|-------------------|-|-----------|-----|
|Pair 1|VAR00001 & VAR00002|4|-,761      |,239 |
|------|-------------------|-|-----------|-----|
|Pair 2|VAR00002 & VAR00003|4|,100       |,900 |
|------|-------------------|-|-----------|-----|
|Pair 3|VAR00001 & VAR00003|4|,000       |1,000|
|-----------------------------------------------|

Paired Samples Test
|--------------------------|-----------------------------------------|------|--|---------------|
|                          |Paired Differences                       |t     |df|Sig. (2-tailed)|
|                          |-------|----------------|----------------|      |  |               |
|                          |Mean   |Std. Dev|Std. Err|95% Confidence I|      |  |               |
|                          |       |        |        |--------|-------|      |  |               |
|                          |       |        |        |Lower   |Upper  |      |  |               |
|------|-------------------|-------|--------|--------|--------|-------|------|--|---------------|
|Pair 1|VAR00001 - VAR00002|2,82500|2,02875 |1,01438 |-,40320 |6,05320|2,785 |3 |,069           |
|------|-------------------|-------|--------|--------|--------|-------|------|--|---------------|
|Pair 2|VAR00002 - VAR00003|-,82500|,99121  |,49561  |-2,40224|,75224 |-1,665|3 |,195           |
|------|-------------------|-------|--------|--------|--------|-------|------|--|---------------|
|Pair 3|VAR00001 - VAR00003|2,00000|1,41421 |,70711  |-,25033 |4,25033|2,828 |3 |,066           |
|-------------------------------------------------------------------------------------------------|
```

Az eredeti output formáját kissé átszerkesztettük a nagyobb betűs megjelenítés kedvéért. Az első táblázat alapstatisztikákat tartalmaz, a második pedig a korreláció-értékeket. Ez utóbbiak alapján nem állapítható meg lineáris összefüggés a változóink között. Bár az első két változópárra az átlagok eltérése viszonylag nagy, a szignifikancia kapott értéke nem elég kicsi, és a 95%-os konfidencia intervallumok is tartalmazzák a nullát. Ez alapján csak

annyit állíthatunk, hogy a változók átlaga között az adatunk alapján nem tudunk 5% hiba mellett különbséget kimutatni.

Csak a példa kedvéért hajtottuk végre minden változópárra a t-próbát: az összehasonlítások számának növekedésével annak az esélye is nő, hogy egy szignifikáns különbség pusztán a véletlen műveként adódik.

3.2.3. Gyakorlat

Vigyük be az SPSS-be a 3.2 táblázat adatait. Az adatokat áttanulmányozva jól látható, hogy a kezdő és a jelenlegi fizetés jelentős mértékben eltér, de vizsgáljuk meg, hogy ezt statisztikailag is alá lehet-e támasztani. Ehhez használjuk az SPSS Paired-Samples T-test parancsát. A megjelenő párbeszédes ablakba adjuk meg az összehasonlítandó változóknak a jelenlegi és a kezdő fizetéseket tartalmazó változókat. A beállításoknál állítsunk be 99%-os konfidencia intervallumot, majd nyomjuk meg az OK gombot. Ezek után a kapott konfidencia intervallum 99% mellett \$11,595 és \$30,358, amely nem tartalmazza a 0-át, azaz nagy biztonsággal mondható, hogy a jelenlegi fizetés eltér a kezdő fizetéstől. Továbbá a mintában az átlagos eltérés nagysága \$20,977, azaz kb. 21 ezer dollárral nagyobb a jelenlegi fizetés. Ezeket az információkat az SPSS alábbi táblázata tartalmazza.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	99% Confidence Int				
					Lower	Upper			
Pair 1	V3 - V4	20,977.000	14,664.309	3,279.039	11,595.884	30,358.116	6,397	19	,000

3.3. Korreláció

A korreláció menüpontból két utasítást tárgyalunk részletesebben: a páronkénti korrelációt és a távolságok generálását. Az első a klasszikus korrelációt jelenti, amikor tehát két normális eloszlású változó közötti lineáris összefüggés meglétét vizsgáljuk. A mérési adataink változói vagy esetei közötti sokféle távolságot pedig például a később tárgyalt osztályozás, klaszterezés során használhatjuk[1]. A menüsorból elérhető parciális korrelációt nem tárgyaljuk.

[1]Bár itt a távolságmátrix az output ablakba kerül. A klaszterezés esetén ugyanezek a távolsági, illetve hasonlósági mértékek szintén elérhetők.

Id	Neme (male-female)	éves jövedelem	Kezdő fizetés
1	m	$57,000	$27,000
2	m	$40,200	$18,750
3	f	$21,450	$12,000
4	f	$21,900	$13,200
5	m	$45,000	$21,000
6	m	$32,100	$13,500
7	m	$36,000	$18,750
8	f	$21,900	$9,750
9	f	$27,900	$12,750
10	f	$24,000	$13,500
11	f	$30,300	$16,500
12	m	$28,350	$12,000
13	m	$27,750	$14,250
14	f	$35,100	$16,800
15	m	$27,300	$13,500
16	m	$40,800	$15,000
17	m	$46,000	$14,250
18	m	$103,750	$27,510
19	m	$42,300	$14,250
20	f	$26,250	$11,550

3.2. táblázat. Aktuális és kezdő fizetés alakulása egy kitalált cégnél.

3.3.1. Páronkénti korreláció

A páronkénti korrelációt (Bivariate) az Analyze / Correlate / Bivariate utasítással kaphatjuk, vagy az Alt-acb gyorsító billentyűkombinációval. A kapott párbeszédes ablakban először a bal oldali változólistából át kell tenni azokat a jobb oldaliba, amelyek közötti páronkénti korrelációt kérjük. Itt természetesen kettőnél több változót is megadhatunk, ekkor az ezekből képezhető összes párra kapunk korrelációs együtthatókat (korrelációs mátrix).

A következő rovatban a korrelációs együttható típusát határozhatjuk meg. A választási lehetőségek: Pearson, Kendall's tau-b és Spearmann koefficiens. Az utóbbi kettő rangskálán mért adatokra alkalmas, illetve csak az értékek sorrendjén alapul. Mindhárom korrelációs mutatóra érvényes viszont, hogy a kapott érték előjele a változók közötti összefüggés irányát jelzi (pozitív, ha együtt nőnek), az abszolút értéke pedig az összefüggés szorosságát jelzi.

Ezután azt adhatjuk meg, hogy a szignifikancia-tesztet egyoldali (One-tailed) vagy kétoldali (Two-tailed) vizsgálat alapján kérjük-e. Legalul pedig egy pipával jelezhetjük azt, hogy kérjük a szignifikáns korrelációk megjelölését. Ekkor a 0,05 szintnek megfelelő értékeket egy, a 0,01 szintnek megfelelőt pedig két csillag jelzi a táblázatban.

Az Options gomb megnyomása után további részleteket tisztázhatunk. Először is különböző statisztikákat kérhetünk: az átlagot és a szórást, illetve a változópárokra vonatkozó négyzet- (vagy keresztszorzat-) összeget és a

kovarianciát. Végül a hiányzóadat kódok kezelését határozhatjuk meg: hogy az eseteket akkor hagyja-e ki a feldolgozás, ha az egyes párokban előfordul ilyen adat, vagy törölje azokat az eseteket, amelyekben akár csak egy helyen is van hiányzóadat kód.

Példa

Vegyük az előző szakasz adatsorát ismét. Kérjük erre a Correlate / Bivariate eljárást. Az eredmény, amit kapunk:

```
CORRELATIONS
  /VARIABLES=VAR00001 VAR00002 VAR00003
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

		VAR00001	VAR00002	VAR00003
VAR00001	Pearson Correlation	1	-,761	,000
	Sig. (2-tailed)		,239	1,000
	N	4	4	4
VAR00002	Pearson Correlation	-,761	1	,100
	Sig. (2-tailed)	,239		,900
	N	4	4	4
VAR00003	Pearson Correlation	,000	,100	1
	Sig. (2-tailed)	1,000	,900	
	N	4	4	4

Az egyes cellákban a Pearson-féle korrelációs együtthatót, a szignifikancia-értékét és a figyelembe vett esetek számát találjuk. A főátlóban minden változónak az önmagára vonatkozó korrelációja egy és a kapott korrelációs mátrix szimmetrikus. A három érdemi korrelációs együttható érték megegyezik a 3.2.2 szakaszban kapottakkal (a szignifikancia-értékek szintén).

Gyakorlat

Első lépésként gépeljük be a 3.3. táblázat értékeit. A táblázat Magyarország népességére és élve születésére vonatkozó 1994-es évi adatokat tartalmazza hónapokra bontva. Vizsgáljuk meg, hogy melyik változó milyen viszonyban áll a többivel. Ehhez használjuk az Analyze menüpont Bivariate parancsát. A felugró párbeszédes ablakban adjuk meg, hogy az összes változó összefüggésére kíváncsiak vagyunk. Ezt megtehetjük úgy, hogy a bal oldali

Hónap	Népesség (e fő)	élveszületések száma (fő)	Halálozások száma (fő)
1	10273	10238	13888
2	10270	9285	12825
3	10267	10105	12516
4	10265	9617	11753
5	10262	9548	12328
6	10260	9717	11839
7	10258	9965	11848
8	10257	9980	11722
9	10256	9844	10968
10	10252	9021	12542
11	10249	8740	11743
12	10246	9538	12917

3.3. táblázat. Magyarország népesség alakulása 1994-ben (ezer fő).

listában az összes változót kijelöljük, majd a nyíl segítségével áthelyezzük a jobb oldalra. Ezek után az OK gomb hatására az alábbi eredményt kapjuk.

Correlations

		Honap	Nepesseg	Elveszul	Halalozas
Honap	Pearson Correlation	1	-,994**	-,496	-,430
	Sig. (2-tailed)		,000	,101	,163
	N	12	12	12	12
Nepesseg	Pearson Correlation	-,994**	1	,516	,403
	Sig. (2-tailed)	,000		,086	,194
	N	12	12	12	12
Elveszul	Pearson Correlation	-,496	,516	1	,136
	Sig. (2-tailed)	,101	,086		,674
	N	12	12	12	12
Halalozas	Pearson Correlation	-,430	,403	,136	1
	Sig. (2-tailed)	,163	,194	,674	
	N	12	12	12	12

** Correlation is significant at the 0.01 level (2-tailed).

Az eredményből látható, hogy míg a hónap és a népesség között egy elég erős lineáris összefüggés van (-0,994), addig a többi változó gyakorlatilag lineárisan nem magyarázható másik változóval. A korrelációs együttható negatív értéke azt mondja nekünk, hogy a népesség fogy. Azaz a statisztikai eredmények alapján a népesség alakulásában fellelhető negatív trendet valószínűleg nem az adott év születések és halálozások számában kell keresni.

3.3.2. Távolságok

Az esetek vagy változók hasonlóságának vagy eltérésének mértékét az Analyze / Correlate / Distances paranccsal kaphatjuk meg írott formában, az output ablakban. A gyorsító billentyűkombináció az Alt-acd. Az angol nyelvű szóhasználat ezekre a mutatókra a distance (távolság), similarity (hasonlóság) és dissimilarity (a hasonlóság ellentéte, kb. eltérés) szavakat alkalmazza.

A párbeszédes ablakban az érintett változókat a bal oldali listából át kell tenni a jobb oldali kisebb ablakba. Esetleg megnevezhetünk egy olyan változót lejjebb, amelyik az esetcímkéket tartalmazza. Az első rádiógomb párral azt kell közölnünk, hogy az esetek (alapértelmezés), vagy a változók közti hasonlóságokat, illetve eltéréseket kérjük-e. Mindegyik esetben a kijelölt változók által korlátozott adathalmazon dolgozik a program.

A következő rovatban először a mutató típusát kell megadni, hogy az eltérés vagy hasonlóság jellegű-e. Ennek függvényében változik a beállított formula, amit az alul levő felirat is megad, és amelyet a Measures gombbal megváltoztathatunk. A választott mérték függ az érintett értékek mérési skálatípusától. Nagyon sok különféle mértékből választhatunk. Azt, hogy melyikre van egy bizonyos esetben szükség, az alkalmazás dönti el. Minden, statisztikát rendszeresen használó szakterületen már kialakult az a gyakorlat, hogy melyik mérték a megfelelő a vizsgálatokhoz.

A már az előzőekben is nagyon rugalmasnak mutatkozó eljárás azt is megengedi, hogy mind az adatunkat, mind a kapott mértéket transzformáljuk (Transform Values). Az előbbit például a 0-1 intervallumra normalizálhatjuk, az utóbbinak pedig vehetjük az abszolút értékét.

Bár az eredményt az output ablakba kapjuk, és ebben az értelemben a program azt inkább egy jelentés írásához szánja illusztrációként, de a kapott táblázatot minden további nélkül bemásolhatjuk a munkalapunkba, és ott tovább számolhatunk vele.

Példa

Hogy áttekinthető eredményt kapjunk, adjunk meg most egy nagyon egyszerű adatsort: a három változónk (két-két esettel) legyen rendre 1, 1; 2, 2; és 3, 3. Kérjük a Correlate / Distances eljárást mindhárom változóra (ne felejtsük el beállítani, hogy nem az esetek közötti távolságot kérjük). A kapott eredmény:

```
PROXIMITIES VAR00001 VAR00002 VAR00003
  /VIEW=VARIABLE
  /MEASURE=EUCLID
  /STANDARDIZE=NONE.
```

```
Proximities

Case Processing Summary
|-----------------------------------------|
|Cases                                    |
|-------------|---------------|-------------|
|Valid        |Missing        |Total        |
|-----|-------|-------|-------|-----|-------|
|N    |Percent|N      |Percent|N    |Percent|
|-----|-------|-------|-------|-----|-------|
|2    |100,0% |0      |0,0%   |2    |100,0% |
|-----------------------------------------|

Proximity Matrix
|--------|-------------------------------------|
|        | Euclidean Distance                  |
|        |-------------------|--------|--------|
|        |VAR00001           |VAR00002|VAR00003|
|--------|-------------------|--------|--------|
|VAR00001|,000                |1,414   |2,828   |
|--------|-------------------|--------|--------|
|VAR00002|1,414              |,000    |1,414   |
|--------|-------------------|--------|--------|
|VAR00003|2,828              |1,414   |,000    |
|-------------------------------------------|
This is a dissimilarity matrix
```

Az első, egysoros táblázat a feldolgozott eseteket összegzi, a második pedig magukat a távolságértékeket foglalja össze. Ez utóbbi értékeit rövid fejszámolással ellenőrizhetjük (vegyük észre, hogy az alapbeállítás euklideszi távolságot jelent: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.).

Gyakorlat

A 3.4. táblázatban láthatók egyes európai országok GDP változásai az előző évekhez képest. A sorok és oszlopok között bizonyos számszerű hasonlóságok és eltérések a számokból levonhatók, de ezek kimutatása már nehézkes. Ilyenek megtalálására jó eszköz lehet a SPSS Distances parancsa. A táblázat betöltése után adjuk ki ezt a parancsot. A megjelenő párbeszédes ablakban a Variables részbe adjuk meg az éveket mint vizsgált változókat. A Compute Distances részben megadhatjuk, hogy esetekre vagy változókra kívánunk távolságokat kiszámítani. Jelen esetben mindkettő érdekes lehet. Az esetek választása esetén az eredményből az országok GDP alakulásának a hasonlóságára tudunk utalásokat tenni, míg a változók esetén az évek közti hasonlóságot tudjuk vizsgálni. Értelemszerűen a Label Cases by opció csak az esetek összehasonlítása esetén lesz engedélyezett. Ide megadhatjuk ebben az esetben az országok neveit tartalmazó változót. A változók összehasonlítása esetén a távolság mátrixban a címkék természetesen a változók nevei lesznek. Nézhetünk egyéb távolságokat is de jelen esetben elegendő lesz az alapbeállításként is szereplő euklidészi távolság (Euclidean distance).

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Austria	3,53	2,33	0,33	2,63	1,89	2,59	1,82	3,53	3,24	3,31
Belgium	1,82	1,52	-0,97	3,18	2,36	1,15	3,28	1,86	3,37	3,68
Denmark	1,29	1,96	-0,09	5,38	3,02	2,8	3,15	2,14	2,52	3,51
Finland	-6,60	-3,88	-1,25	3,86	3,39	3,72	6,01	5,50	3,79	4,94
France	1,20	1,93	-0,99	2,06	2,33	1,09	2,37	2,51	3,25	3,84
Germany	5,18	2,20	-0,81	2,62	1,87	0,99	1,79	1,81	1,99	3,16
Hungary	-12,67	-3,11	-0,58	2,9	1,48	1,31	4,47	4,72	4,03	5,08
Romania	-13,81	-9,21	1,49	3,83	6,86	3,83	-6,29	-4,92	-1,21	2,08

3.4. táblázat. Országok GPD növekedésének alakulása (előző év százaléka).

Az eset összehasonlítása esetén az alábbi távolságokat kapjuk:

Proximity Matrix

	Euclidean Distance							
	1:Austria	2:Belgium	3:Denmark	4:Finland	5:France	6:Germany	7:Hungary	8:Romania
1:Austria	,000	359,743	429,812	1315,726	342,399	333,823	1751,961	2492,712
2:Belgium	359,743	,000	317,706	1143,370	176,136	409,320	1563,390	2354,625
3:Denmark	429,812	317,706	,000	1114,110	407,783	550,329	1564,329	2298,694
4:Finland	1315,726	1143,370	1114,110	,000	1138,576	1500,207	716,392	1982,424
5:France	342,399	176,136	407,783	1138,576	,000	439,969	1519,418	2323,577
6:Germany	333,823	409,320	550,329	1500,207	439,969	,000	1925,365	2554,435
7:Hungary	1751,961	1563,390	1564,329	716,392	1519,418	1925,365	,000	1800,331
8:Romania	2492,712	2354,625	2298,694	1982,424	2323,577	2554,435	1800,331	,000

This is a dissimilarity matrix

Látható, hogy a legkisebb távolságok Ausztria, Belgium, Franciaország és Németország között található, míg nagyobb távolságokat Magyarország és Románia esetében találhatunk. Azaz elmondható, hogy ezen országok az előbbi csoporthoz képest másmilyen GDP változásokat mutattak az adott időszak alatt.

A változók összehasonlítása esetén az alábbi táblázatot kapjuk:

Proximity Matrix

	Euclidean Dista								
	@1991	@1992	@1993	@1994	@1995	@1996	@1997	@1998	@1999	@2000
@1991	,000	1146,055	2166,761	2626,189	2730,345	2518,518	2303,733	2327,340	2374,827	2683,678
@1992	1146,055	,000	1264,108	1675,723	1828,818	1582,459	1299,732	1302,973	1342,234	1687,989
@1993	2166,761	1264,108	,000	1082,584	979,844	769,864	1369,556	1253,303	1067,132	1237,952
@1994	2626,189	1675,723	1082,584	,000	434,088	411,393	1076,688	982,207	608,968	408,051
@1995	2730,345	1828,818	979,844	434,088	,000	369,232	1376,965	1254,863	870,292	679,367
@1996	2518,518	1582,459	769,864	411,393	369,232	,000	1118,815	979,138	662,569	620,150
@1997	2303,733	1299,732	1369,556	1076,688	1376,965	1118,815	,000	286,009	584,501	883,863
@1998	2327,340	1302,973	1253,303	982,207	1254,863	979,138	286,009	,000	450,032	763,363
@1999	2374,827	1342,234	1067,132	608,968	870,292	662,569	584,501	450,032	,000	400,590

```
|@2000|2683,678|1687,989|1237,952|408,051 |679,367 |620,150 |883,863 |763,363 |400,590 |,000    |
|-------------------------------------------------------------------------------------------------|
This is a dissimilarity matrix
```

Itt igazán értékes információt talán közvetlenül a főátló alatti számok hordozhatnak. Ezek megmondják, hogy az előző évhez képest nagy változás történt-e a GDP számokban. Igazán kirívó eset nincs, talán az 1996-os év tekinthető egy kicsit nagyobbnak a környezetéhez képest. Itt az eredeti GDP adatokat megvizsgálva is tapasztalhatunk változást. De hasonló éveket látunk a 90-es évek elejéből is, amikor inkább csak 1-1 ország okozza ezen nagy értékeket.

3.4. Regresszió

Míg az előző szakaszban tárgyalt korreláció a változók közötti valamely lineáris összefüggés meglétét állapította meg, illetve annak szorosságát jellemezte, addig a regresszió a változóink közötti nem feltétlenül lineáris összefüggések részleteit hivatott tisztázni, az adott szerkezetű, képletű összefüggések együtthatóit tudja statisztikai értelemben meghatározni. Az ide tartozó eljárások determinisztikus feladatok esetén az optimalizálás témakörében találhatók.

A menüsorban rendelkezésre álló nagyszámú különféle eljárásból itt csak a görbe illesztést (Curve Estimation, Alt-arc) és a nemlineáris regressziót (Nonlinear, Alt-arn) tárgyaljuk. Ez a két parancs lényegében csak abban tér el egymástól, hogy az első rögzített modellfüggvények alapján végzi a regressziót, a második pedig egy függvénykészletből való szabad építkezést enged meg. Mindkettő meglehetősen kifinomult, és ilyen értelemben a táblázatkezelő programok által nyújtott solver parancshoz hasonló, illetve túl is mutatnak azon.

3.4.1. Görbe illesztés

A görbe illesztési utasítást akkor adjuk ki, ha valamely standard modellfüggvényt használjuk az adataink leírására. Az eljárás tehát egy statisztikai minta alapján a magyarázóváltozó (független) és az eredményváltozó (függő) közötti összefüggést határozza meg.

A kapott párbeszédes ablakban először az eredményváltozókat (Dependent(s)) kell kijelölnünk. Itt többet is megadhatunk, és ebben az esetben a megadott változószámnak megfelelő darab görbe illesztési feladatot old meg az eljárás. Mindegyikről egy-egy ábrát is kapunk. A magyarázóváltozónak

vagy kiválaszthatunk egyet az adatállományunk változói közül, vagy az időt adjuk meg magyarázóváltozónak (Indepentent). Megadhatunk esetcímkéket is.

Beállíthatjuk azt is, hogy kérjük-e az egyenletben a magyarázófüggvényben szereplő additív konstans kiíratását (Include constant in equation). Szintén itt, a párbeszédes ablak közepén lehet azt is megadni, hogy kérjük-e az eredmény grafikán való bemutatását (Plot models).

Ezután következik a modell típusának beállítása. A leggyakoribb regressziófüggvények találhatók meg itt, és ezekből egyszerre többet is megadhatunk. Ebben az esetben mindegyiket (nyilván külön-külön) meghatározza a program, és azután a kapott eredményt egymás mellett, az összehasonlítást elősegítve megjeleníti. A rendelkezésre álló függvények (többek között): lineáris (linear), logaritmikus (logarithmic), hiperbolikus (itt inverse), kvadratikus (quadratic), köbös (cubic), hatványfüggvény (power) és exponenciális (exponential).

A modellfüggvény paramétereit általában a b0, b1 stb. nevekkel jelöli az output ablakban. Más esetekben a kapott paramétervektort B néven kell keresni az eredményállományban, ilyenkor ennek minden komponenséhez egy-egy sor tartozik. Ebben a kapott paraméterértéken kívül megtalálhatjuk például annak standard hibáját (SE) is.

Az eddigieken felül kérhetjük a variancia analízis (ANOVA) eredményének egyidejű nyomtatását is. Ezzel együtt a kapott output ablak tartalma alapos vizsgálatot tesz lehetővé, melyben összevethetjük az előrejelzett és az eredeti adatsorok átlagát.

Mivel a kapott eredményekkel gyakran tovább szeretnénk számolni, ezért a görbeillesztés kapott értékeit be is vihetjük az adatállományunkba a Save gomb megnyomásával. Az ekkor kapott új párbeszédes ablakban be lehet jelölni, hogy melyik mutató mentését kérjük az alábbiak közül: illesztett érték, reziduum (az adat és a becsült érték eltérése) és az előrejelzett értékre vonatkozó konfidencia intervallumok. Ez utóbbinak a százalékos konfidencia-szintjét be lehet állítani.

Példa

Az első oszlopba bevittük az 1, 2, 3, 4, 5, 6 értékeket, ezek lesznek a magyarázóváltozó, x értékei. Ezután 0 átlaggal, 0,5 szórással generáltunk normális eloszlású véletlen számokat a második oszlopba, az y változóba. Ezek segítségével definiálunk egy új változót, z-t, amelynek értéke $x^2 + 2x + 3 + y$.

A görbeillesztés segítségével az 1, 2 és 3 együtthatókat szeretnénk visszakapni az adatok alapján, a jelentős mértékű zaj ellenére. A párbeszédes ablakban megjelöltük z-t, mint az eredményváltozót, x-et mint a magyarázóváltozót, és a lineáris *és*[2] kvadratikus modellfüggvénnyel kérjük az illesztést. A kapott eredmény:

```
 Curve Estimation.
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=z WITH x
  /CONSTANT
  /MODEL=LINEAR QUADRATIC
  /PLOT FIT.


Model Summary and Parameter Estimates
Dependent Variable: z
|---------|-----------------------------------|------------------------------|
|Equation |Model Summary                      |Parameter Estimates           |
|         |-------------|--------|---|---|----|------------------|-----|----|
|         |R Square     |F       |df1|df2|Sig.|Constant          |b1   |b2  |
|---------|-------------|--------|---|---|----|------------------|-----|----|
|Linear   |,976         |163,371 |1  |4  |,000|-5,877            |8,853|    |
|---------|-------------|--------|---|---|----|------------------|-----|----|
|Quadratic|1,000        |3740,117|2  |3  |,000|2,900             |2,270|,940|
|-------------------------------------------------------------------------------|
The independent variable is x.
```

Az utolsó két sor tartalmazza az adott modell becslésének paramétereit. Látható, hogy az eredeti értéket jó közelítéssel visszakaptuk, az eltéréseket az additív zaj magyarázza.

Gyakorlat

Használjuk a 3.3. táblázat értékeit. Próbáljunk lineáris görbét illeszteni a népességi adatokra. Ezt megtehetjük a Linear parancs segítségével, amit a Regression menüpontban találunk. Ebben a feladatban arra vagyunk kíváncsiak, hogy hogyan alakul Magyarország népessége a hónapok függvényében, azaz a függő változó (Dependent) a Nepesseg, míg a magyarázóváltozóként (Independent) a Honapot szeretnénk. A default beállítások most teljesen megfelelnek a számunkra, azaz a változók áthelyezése után kérhetjük a OK gomb megnyomásával az eredményt. A korábban látott korrelációs együttható itt is látható, amely nagynak mondható (0,994). Továbbá a modell érvényességére vonatkozó vizsgálat is azt mondja, hogy a lineáris trenden felül a hiba kicsi (szórása: 0,971). A konstansra a 10 274 értéket kaptuk, míg a lineáris trend meredekségére $-2,283$. Ez azt jelenti, hogy Magyarország népessége az illesztés szerint 1994 elején 10 274 424 fő volt, és minden egyes hónapban 2283 fővel csökkent. Megjegyzendő,

[2]Ha a lineáris modellt nem adjuk meg, akkor nem kapunk becslést az együtthatókra.

hogy ez a becslés nem is olyan rossz, ugyanis így egy évben nagyjából 27 396 fővel csökken a népesség. A valóságban 1990-től 2000-ig évente átlagosan 18 700 fővel csökkent a népesség.

3.4.2. Nemlineáris regresszió

A nemlineáris regresszió szabad modellfüggvény beállítást tesz lehetővé, az összefüggéseket tetszőleges nevű paraméterekkel és standard függvényekkel határozhatjuk meg. Továbbá ebben az esetben egynél több magyarázó változó is szerepelhet a modellfüggvényben.

Az induló párbeszédes ablakban először adjuk meg a bal oldali listából azt a változót, amelyik az eredményváltozót jelenti. Ezután adjuk meg azt a képletet, amely a modellfüggvény összefüggését határozza meg (jobb középső ablak). Itt tehát egy olyan képletet kell kialakítanunk, amely a független változóinkat és a meghatározandó paramétereket is tartalmazza. Ennek megszerkesztéséhez a szokásos kalkulátorszerű billentyűzet és egy hosszú lista az alkalmazható függvényekkel áll rendelkezésre.

Új beállítási lehetőség a paraméterek megadása. Ehhez a bal oldalon lévő Parameters gombot kell megnyomni, majd a paraméterek nevét a képletben szereplőnek megfelelően megadni és indulóértéket is hozzárendelni. Ezt más regressziós vizsgálathoz hasonlóan a gyakorlati tapasztalatok alapján a várható paraméterérték közelében célszerű megválasztani. Ahogy a nemlineáris optimalizálás tanítja, az érintett regressziós algoritmusok nagy része érzékeny az indulóértékek viszonylag jó, a megoldáshoz közeli elhelyezkedésére, és csak ritkán találja meg az abszolút optimális megoldást (az ún. globális optimumot) minden indulópontból. Ennek az az oka, hogy a beépített eljárások csak függvénykiértékelésre támaszkodnak mint információ forrásra, és így a függvénynek a kiértékelési pontok közti viselkedéséről nincs megbízható tudásuk. Következésképp érdemes az indulópontokat alapos megfontolások után meghatározni, illetve több indulóponttal megismételni a regressziót[3].

Ezekkel a beállításokkal az eljárás már működik, de további finomításokat lehet végrehajtani, amikkel az algoritmus eredményét befolyásolhatjuk. A Loss gomb segítségével a célfüggvény (amit optimalizálunk) alakját lehet pontosítani. Ez alapértelmezésben a reziduumok négyzetösszege, amit minimalizálni kell. Ehelyett a felhasználó megadhat egy új célfüggvény alakot,

[3]A korábbi indulóérték megváltoztatásához van ugyan egy Change gomb, de ez csak akkor aktivizálódik, ha egy korábbi értéket, pl. az indulóértéket vagy a változó nevét már átírtuk.

amely az eredmény- és a magyarázóváltozók, a paraméterek, továbbá a reziduumok és a jósolt értékek valamely függvénye lehet. Ha ilyen célfüggvényt szeretnénk megadni, akkor a User-defined loss function feliratú rádiógombot kell megnyomni.

A görbeillesztéshez képest az is többlet ebben az eljárásban, hogy itt megadhatunk korlátozó feltételeket az optimalizálandó paraméterekre. Ehhez nyomjuk meg a Constraints gombot. Az alapértelmezés az, hogy a feladatunk korlátozás nélküli (unconstrained). Ha ettől el szeretnénk térni, akkor meg kell nyomni a Define parameter constraint rádiógombot. Ezután a paraméterek listájáról választva megadhatunk korlátozó feltételeket ezekre. Itt az ún. változókra vonatkozó korlátok (bound constraints) mellett megadhatunk összetett feltételeket is, amiket a paraméterek és az alapműveletek segítségével állíthatunk össze. Ismét, ha egy korábban megadott korláton változtatunk, akkor a hozzáadás (Add) és módosítás (Change) lehetőségek között választhatunk.

A Save gomb azt teszi lehetővé, hogy a regressziós eljárás során létrejött új változókat a későbbi feldolgozás számára elmentsük: az előrejelzett értékeket (Predicted values), a reziduumokat (Residuals) és a deriváltakat (Derivatives). Itt jegyezzük meg, hogy az optimalizálási eljárás a működéséhez szükséges deriváltakat numerikus deriválással maga állítja elő (ezt meg is írja az output ablak jelentésében).

Az Options gomb megnyomásával további eljárásparamétereket adhatunk meg. A kapott párbeszédes ablakban kérhetjük például, hogy az alapértelmezésbeli szekvenciális kvadratikus programozási módszer helyett az eljárás a Levenberg-Marquardt módszert használja. A bootstrap mintavétel esetén csak az első az elérhető. A két optimalizálási módszer más és más megállási feltétel-paraméterekkel működik, csak a maximális megengedett iterációszám a közös. Ezek a megállási feltételek döntően befolyásolhatják az eredményt, ezért részletesebben tárgyaljuk ezeket.

A szekvenciális kvadratikus programozási módszer esetén korlátozhatjuk a megtett lépések hosszát (Step limit). Ennek az a hatása, hogy elegendően kis érték esetén az induló érték vonzáskörzetében levő helyi minimumot találja meg az algoritmus. Három további tolerancia-értéket lehet még beállítani: az optimum és a célfüggvény pontosságát, valamint a túl nagynak tekintendő lépésméretet (Infinite step size). Az előbbi kettő alapértelmezésbeli értéke egy az adott számítógépes környezetre jellemző nagyon kis szám. Ha ezeknél nagyobbat adunk meg, akkor az eljárás gyorsabban megáll, de az eredmény az optimumtól távolabb lehet. Az utolsó tolerancia-paraméter hatása pedig a következő: ha az aktuális lépésköz az itt megadottnál nagyobb, akkor a módszer arra a következtetésre jut, hogy a feladat

megoldása nem korlátos.

A Levenberg-Marquardt módszer három eljárásparaméterrel dolgozik: a korábban már említett maximális iterációszám mellett két toleranciaértéket lehet megadni az ún. négyzetösszeg- és a paraméter-konvergenciára. Ha az utolsó két iterált értékre az alapul vett mennyiségek kevesebbel változnak mint a beállított szám, akkor az algoritmus leáll. Arra is van lehetőség, hogy ez utóbbi két feltételt kikapcsoljuk a Disable lehetőség kiválasztásával.

Példa

Egy egyszerű példát vegyünk: tekintsük a következő magyarázóváltozó értékeket: 1, 3, 4, 5 és 6. Ezeket beírjuk egy x változóba, majd képezünk egy újabb változót, y-t a Compute utasítással a következő képletnek megfelelően: $y = \sin(2 * x)$. A kapott értékek: 0,91, -0,28, 0,99, -0,54 és -0,54. A nemlineáris regresszióval a kettes szorzót szeretnénk visszakapni csak az adatainkra támaszkodva.

A nemlineáris regressziós utasítás által adott párbeszédes ablakban megadjuk azt, hogy az eredményváltozó az y, a modellfüggvény pedig SIN($a * x$). Vigyázat, itt a szinusz függvényt ki kell keresni a megadottak közül (Functions). Az a itt a regresszió paramétere, amit meg akarunk határozni. Ezt a bal alsó sarokban levő Parameters gomb segítségével közölhetjük. A paraméter nevét válasszuk a-nak, és induló értéknek (Starting Value) adjunk meg először egyet (1,0). Az Add gomb megnyomása után a kis ablakban meg is jelenik, hogy a(1), ami pontosan azt jelenti, hogy a rendszer tud egy a nevű paraméterről, aminek az indulóértéke egy.

Az ezek után aktívvá vált OK gomb megnyomása után az output ablakban kapjuk a meglepő eredményt: a paraméter értékére a program a 0,68318 értéket adja, és a reziduális szórásnégyzet is nagy, majdnem kettő. Ismételjük meg a vizsgálatot 2,5-es indulóértékkel az a-ra (előtte ki kell törölni a régi értéket). Ekkor pontosan megkapjuk a várt kettes paraméterértéket, és a reziduális szórásnégyzet is nulla. Ez utóbbi eredményt el is várhattuk, hiszen az adatunkat nem terhelte zaj.

A kapott végső eredmény az output ablakban:

```
Iteration Historyb
|-----------------|-----------------------|---------|
|Iteration Numbera|Residual Sum of Squares|Parameter|
|                 |                       |---------|
|                 |                       |a        |
|-----------------|-----------------------|---------|
|1.0              |5,567                  |2,500    |
|-----------------|-----------------------|---------|
|1.1              |5,124                  |2,438    |
```

2.0	5,124	2,438
2.1	4,408	2,359
3.0	4,408	2,359
3.1	1,910	2,194
4.0	1,910	2,194
4.1	,000	1,999
5.0	,000	1,999
5.1	,000	2,000
6.0	,000	2,000
6.1	,000	2,000
7.0	,000	2,000
7.1	,000	2,000

Derivatives are calculated numerically.

a Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b Run stopped after 14 model evaluations and 7 derivative evaluations because the relative reduction between successive parameter estimates is at most PCON = 1,000E-008.

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	2,000	,000	2,000	2,000

érdemes megfigyelni, hogy az iteráció szépen mutatja a várható kvadratikus konvergenciát: az iteráció végén a pontos értékes jegyek száma minden lépésben megkétszereződik. Mivel az adatunk nem volt zajjal terhelt, az eredményben az „a” paramétert pontosan meg lehetett határozni, a reziduumok négyzetösszege és a konfidencia intervallum szélessége is nulla.

Ha az indulópontnak ellentmondó korlátot adunk meg, akkor a program „The CNLR procedure has set up the problem incorrectly.” üzenetet adja, és nem kapunk megoldást.

Gyakorlat

Rendelkezésünkre áll Magyarország mobiltelefon előfizetőinek a száma 1996. IV. negyedév és 2003. II. negyedév közötti időintervallumra. Vigyük be ezen adatokat az SPSS-be, ahol az év és a Negyedév változók legyenek szöveges- (String), míg az előfizetők száma legyen szám (Numeric) típusú. A 3.5. táblázat adatait megvizsgálva látható, hogy kezdetben egy intenzív növekedés tapasztalható, majd ezen növekedés lassul. Az ilyen típusú növekedést (logisztikus) nagyon jól leírja az alábbi képlet:

$$F(t) = \frac{\beta}{1 - \alpha e^{\gamma - t}}.$$

A γ paraméter a növekedés legintenzívebb időpontját jellemzi, az α paraméter a növekedési fázis időhöz viszonyított gyorsaságát jellemzi, míg a β paraméter a várható legnagyobb érték.

Első lépésben hozzunk létre egy olyan változó oszlopot, amely az időt szimbolizálja. Ezt megtehetjük az év és Negyedév oszlopok segítségével, de mivel egyenletesek a méréseink, így a Create Time Series funkció segítségével is generálhatunk időbélyegeket az adatsorunkhoz. Javasolt további időpontokat is belerakni, hogy a „jövőbe" is kapjunk illesztett pontokat. Ezek után minden adott, hogy a nemlineáris regressziót elkezdjük a Nonlinear menüvel. A megjelenő ablakban a függő változónak állítsuk be az előfizetők száma változót. Hozzunk létre 3 új változót, melyeket nevezzünk a, b és c-nek. Induló paraméternek megadhatjuk: a=1, b=10000 és c=-20. Ezek nagyjából a várható paraméterek közelítő értékei. Az előfizetések számát vizsgálva a maximális előfizetői számot (b) körülbelül 10 millióra lehet becsülni. A legintenzívebb növekedés (c) a 20. esetnél található. Ezek után a modellbe írjuk be az alábbi képletet: b/(1+exp(-a*(Time+c))).

További beállításként kérjük az illesztett értékek mentését, melyet a Save gomb megnyomása után a Predicted values beállításával tehetünk meg. Az OK gomb megnyomása után a kapott paraméterek az alábbiak lettek: b=10892; a=0,173; c=-21,347. Azaz a várható legnagyobb előfizető szám kb. 11 millió és a legnagyobb növekedés a 21-es esetnél volt várható, azaz 2002 első félévében. Ezek értelemszerűen a legjobban illesztett görbe tulajdonságai, a valós adatok némileg eltérhetnek. Ennek vizualizálására készítsünk egy ábrát, amelyen láthatók a valós, illetve illesztett adatok. A jó címkézés érdekében készítsünk egy új címke oszlopot az év és Negyedév változóinkból, mondjuk szövegek összeillesztésével. Ezt megtehetjük a Compute Variable menüpont segítségével, ahol az új változó értéke legyen: concat(Ev, Negyedev).

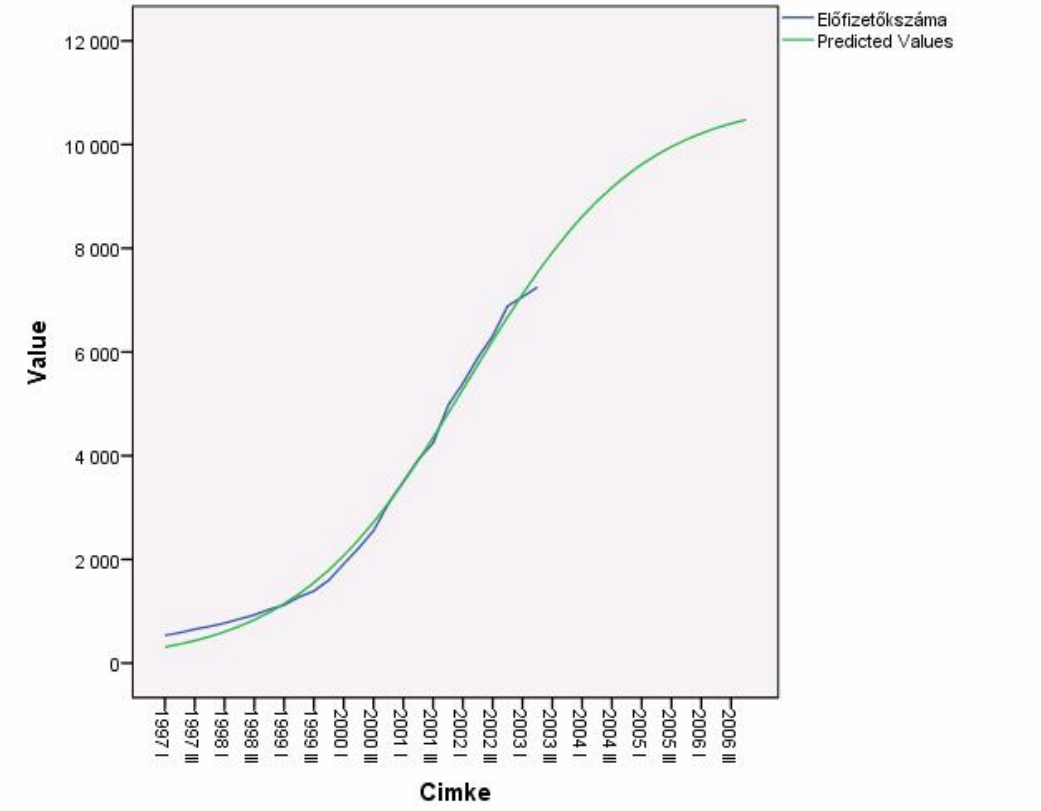
év	Negyedév	Előfizetők száma
1997	I	535
1997	II	589
1997	III	653
1997	IV	708
1998	I	771
1998	II	852
1998	III	930
1998	IV	1034
1999	I	1121
1999	II	1275
1999	III	1393
1999	IV	1601
2000	I	1912
2000	II	2216
2000	III	2553
2000	IV	3076
2001	I	3502
2001	II	3926
2001	III	4245
2001	IV	4967
2002	I	5399
2002	II	5889
2002	III	6312
2002	IV	6886
2003	I	7062
2003	II	7250

3.5. táblázat. Magyarország mobiltelefon előfizetőinek száma (ezer fő).

A fenti információk megjelenítéséhez válasszuk a Line menüpontot, ahol a Multiple és a Values of Individual Cases eseteket kell bejelölni. A Define gomb után adjuk meg, hogy a vonalak reprezentálják (Line Represent) az előfizetők számát, illetve az illesztett értékeket. A kategória címkének (Category Labels) pedig állítsuk be az új változónkat. Ezek után az OK gomb hatására a 3.1. ábrát kapjuk, melyen látható, hogy az illesztés mennyire pontos.

3.5. Osztályozás

Az osztályozás algoritmusai a számítógépes statisztika érdekes eljárásai: arra valók, hogy egy többdimenziós ponthalmazból kiválasszuk azokat, amelyek egymáshoz valamilyen szempont szerint közel esnek, illetve amelyek összetartoznak. Az összefüggés szorosságát is tudjuk majd jellemezni. A klaszter (cluster) szó szerint fürtöt vagy kupacot jelent, itt pontok egy halmazát értjük alatta, amelynek elemei közelebb vannak egymáshoz, mint más klaszterek pontjaihoz.



3.1. ábra. Magyarországi mobiltelefon előfizetők száma és az illesztett görbe.

Az osztályozás alapja egy ún. távolság-, illetve eltérésfüggvény (distance vagy dissimilarity measure), ami két pontra vagy pontok két halmazára (klaszterekre) megmondja, hogy ezek mennyivel térnek el egymástól.

Az ide tartozó eljárásokat az Analyze / Classify menüpontban találjuk. Ezek közül itt csak az ún. K-közép klaszterezést és a hierarchikus klaszterezést tárgyaljuk részletesen. Az elsőt a K-Means Cluster, a másikat a Hierarchical Cluster kulcsszónál találjuk, a megfelelő gyorsító billentyűkombinációk: Alt-afk, illetve Alt-afh.

3.5.1. A K-közép klaszterezés

Ez a klaszterezési eljárás alapvetően statisztikus jellegű szemben a következő szakaszban ismertetett determinisztikus jellegű módszerrel. A K-közép klaszterezés az eseteket sorolja előre megadott számú osztályba. Ehhez megadott számú centrumot határoz meg, és ezeket finomítja, pontosítja egy iterációs eljárással, majd az eseteket sorolja be a centrumok által meghatározott osztályokba. A centrumok nem feltétlenül felelnek meg az eseteknek, közöttük is lehetnek. Az eljárás nevében szereplő K is a centrumok számára utal. Akkor érdemes használni, ha mindenképpen adott számú klasztert szeretnénk kialakítani.

Az eljárás paramétereinek beállítása során először adjuk meg azokat a változókat, amelyek értékének eltérése alapján kell az eseteinket elkülöníteni. Ezeket át kell vinni a jobb oldali kis ablakba. Az eseteket alapértelmezésben a sorszámuk azonosítja, de kérhetjük azt is, hogy ehhez az esetcímkéket használja a program. Fontos megadni az output klaszterek kívánt számát (Number of Cluster). Ez alapértelmezésben 2, és természetesen nem lehet több mint az esetek száma.

A végrehajtott módszernek két alcsoportja van: az egyikben csak a kezdeti centrumokhoz keresi meg a legközelebbi eseteket (Classify only), a bonyolultabb esetben ezt a centrumok iteratív pontosítása előzi meg (Iterate and classify). Ezt a választott módszer rádiógombjával lehet megadni. A centrumokkal külön kérésre a program további műveleteket is végez, mint a kiindulási centrumokat betölti egy adathalmazból vagy adatállományból. Az eredményül kapott centrumokat kimenti egy megadott adathalmazba vagy fájlba. Ezeket a lehetőségeket a Cluster Centers részben érhetjük el. A fájlneveket a szokásos módon egy böngészővel kereshetjük ki.

Ezeken kívül további eljárásparamétereket adhatunk meg, illetve szolgáltatásokat kérhetünk az Iterate, a Save, illetve az Options gombok megnyomásával. Így megadhatjuk a végrehajtható maximális iterációszámot, a klaszterek centrumának azt a maximális relatív megváltozását, amelynél az

iteráció még nem áll le. A Using running means opció bekapcsolásával az új centrumok mindig a régi klaszter átlagaiból adódnak.

A Save gomb megnyomása után beállíthatjuk, hogy az algoritmus elmentse a klaszterhez tartozás mutatóját (Cluster membership) és a klaszterközpontoktól való távolságot (Distance from cluster center) is.

Kérhetünk további információt, így a kezdeti centrumokat (Initial cluster centers), egyváltozós variancia analízist (ANOVA table), vagy a végleges klaszterbe helyezést a centrumtól vett távolsággal (Cluster information for each case).

A hiányzóadat kezelésre itt is két lehetőség adott: vagy kizárjuk a feldolgozásból mindazokat az eseteket, amelyekre hiányzóadat kód van valamely változóra (Exclude cases listwise), vagy csak az esetek távolságának meghatározásából azokat a változókat zárjuk ki, amelyekben hiányzóadat kód van (Exclude cases pairwise).

Példa

Tekintsünk egy kis adathalmazt, négy változónk van: x, y, z és v. Ezek értékei rendre: $(1,5,4,3)^T$, $(4,7,2,4)^T$, $(1,1,4,1)^T$ és $(2,5,2,2)^T$. A klaszterezésbe bevonjuk mind a négy változó értékeit, és két klasztert kérünk kialakítani. A centrumok iterációs finomítását és a centrumokat mindig a régi klaszterekből képezzük. A kiegészítő statisztikát kérjük az Options ablakban az ANOVA nélkül.

A kapott eredmény[4]:

```
QUICK CLUSTER x y z v
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(2) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /PRINT INITIAL CLUSTER DISTAN.


Initial Cluster Centers
|-|------------|
| |Cluster     |
| |-------|----|
| |1      |2   |
|-|-------|----|
|x|4,00   |5,00|
|-|-------|----|
|y|2,00   |7,00|
|-|-------|----|
|z|4,00   |1,00|
|-|-------|----|
|v|2,00   |5,00|
|--------------|

Iteration Historya
```

[4] Az eredeti táblázat formázása most is szebb, de azt közvetlenül nem lehetett átvinni a LaTeX szövegszerkesztőbe.

Iteration	Change in Cluster Centers	
	1	2
1	2,749	,000
2	,000	,000

a Convergence achieved due to no or small change in cluster centers.
The maximum absolute coordinate change for any center is ,000.
The current iteration is 2. The minimum distance between initial centers is 6,633.

Cluster Membership

Case Number	Cluster	Distance
1	1	2,055
2	2	,000
3	1	2,749
4	1	1,247

Final Cluster Centers

	Cluster	
	1	2
x	2,67	5,00
y	3,33	7,00
z	2,00	1,00
v	2,00	5,00

Distances between Final Cluster Centers

Cluster	1	2
1		5,375
2	5,375	

Number of Cases in each Cluster

Cluster	1	3,000
	2	1,000
Valid		4,000
Missing		,000

Vegyük észre, hogy a második klaszter egyetlen pontból áll, és az iteráció során nem is változott (és nyilván megegyezik a centrumával). Az első klaszter centruma viszont egy olyan pont, amelyik egyik esettel sem egyezik meg (bár a kezdeti centrumok mind esetek voltak). Habár a négydimenziós térben nehéz elképzelni az eseteket mint pontokat, mégis a második eset több változóban is lényegesen eltér a többitől. Ez magyarázza a külön klaszterként

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. setosa
5.0	3.6	1.4	0.2	I. setosa
7.0	3.2	4.7	1.4	I. versicolor
6.4	3.2	4.5	1.5	I. versicolor
6.9	3.1	4.9	1.5	I. versicolor
5.5	2.3	4.0	1.3	I. versicolor
6.5	2.8	4.6	1.5	I. versicolor
6.3	3.3	6.0	2.5	I. virginica
5.8	2.7	5.1	1.9	I. virginica
7.1	3.0	5.9	2.1	I. virginica
6.3	2.9	5.6	1.8	I. virginica
6.5	3.0	5.8	2.2	I. virginica

3.6. táblázat. Nőszirom három alfajának adatait tartalmazó adatbázis (részlet).

való megjelenését.

Gyakorlat

Tekintsük az interneten sok helyen (például: http://en.wikipedia.org/ wiki/Iris_flower_data_set) fellelhető nőszirom virág adatbázist (Iris flower data set). Ez egy 150 mintából álló adathalmaz, mely 3 alfaj (I. setosa, I. versicolor, I. virginica) csésze (Sepal) és szirom (Petal) levél nagyságát tartalmazza 50-50 példányra. Ebből az adathalmazból pár sor megtalálható aa 3.6. táblázatban.

A fenti adatok beolvasása után érdekes lehet megvizsgálni, hogy a virághlevelekből milyen arányban jósolható meg, hogy melyik alfajba tartozik az adott példány. Ehhez először határozzuk meg, hogy 1-1 alfajra mik a legjellemzőbb méretek. Ezek meghatározása történhet a K-közép klaszterezés segítségével. A klaszterek középpontjai lesznek az alfajok jellemző szirom- és csésze levél méretei.

Adatok beolvasása után indítsuk el az SPSS k-Means parancsát. A párbeszédes ablakban adjuk meg, mint klaszterezési szempont, a méreteket tartalmazó 4 változót, továbbá, hogy 3 klasztert szeretnénk. Valamint érdemes a Save gombnál megadni, hogy mentse le a klaszterezés eredményét is. A klaszterek középpontjait a 3.7. táblázat tartalmazza. Azaz, ha veszünk egy nőszirom példányt és megnézzük, melyik klaszterközépponthoz van a legközelebb, az nagy valószínűséggel ahhoz az alfajhoz tartozik. Ez a jelenleg használt 150 elemű adatbázisunk 134 elemén helyes eredményt ad.

	I. virginica	I. setosa	I. versicolor
Sepallength	6.9	5.0	5.9
Sepalwidth	3.1	3.4	2.7
Petallength	5.7	1.5	4.4
Petalwidth	2.1	0.2	1.4

3.7. táblázat. Nőszirom alfajaira jellemző méretek.

3.5.2. Hierarchikus klaszterezés

A hierarchikus klaszterezés kiindulópontja eltér a K-közép klaszterezésétől: nem egy adott számú osztályt kell kialakítani, hanem azt meghatározni, hogy az adatok alapján milyen osztályozások adódnak természetes módon – különböző mértékű hasonlóság, illetve eltérés esetén.

Nyilvánvaló, hogy ha semmilyen eltérést sem engedünk meg az egy klaszterbe tartozó objektumok között, akkor csak a megegyező változóértékkel rendelkező esetek tartozhatnak egy klaszterbe. Másrészt ha tetszőlegesen nagy eltérés is megengedett, akkor az összes eset mind egyetlen klaszterbe sorolható. A két véglet között azonban az átmenet sokféle struktúrát alakíthat ki. A hierarchikus klaszterezés azt a struktúrát határozza meg, aminek értelmezése során a felhasználó az egy-egy hasonlósági szintnek megfelelő klasztereket leolvashatja.

Az eljárás a következő: tekintsünk egy ponthalmazt (ezek általában az eseteink), amely elemeinek a közelségét vagy távolságát valamely változók értékei alapján határozhatjuk meg. A feladat ezután megfeleltethető egy többdimenziós térben levő ponthalmaz osztályozásának.

A legalsó szinten minden eltérő pontot külön klaszterbe tartozónak tekintünk, tehát ekkor a megengedett eltérés nulla. Ezután a küszöbértéket megemeljük úgy, hogy a két „legközelebbi" pont egy klaszterbe kerüljön. Ezzel, mint új klaszterrel a távolságokat újra meg kell határozni. A klaszterek összevonását a küszöbérték emelésével folytatjuk mindaddig, amíg az összes eset egyetlen klasztert képez. A közben kialakult szerkezetet mindig regisztráljuk, és az eljárás végén valamely alkalmas formában megjelenítjük.

Két dolgot kell még tisztázni: az egyik, hogy a pontok eltérését hogyan értelmezzük, milyen távolságfüggvénnyel, a másik pedig az, hogy hogyan definiáljuk két klaszter, azaz két ponthalmaz távolságát.

Az eltérés mérésére használhatjuk például az euklideszi távolságot. Két ponthalmaz távolságának egyik szokásos mértéke pedig a két legközelebbi pontjuk távolsága (Single linkage, vagy itt Nearest neighbor).

Ezek a beállítások döntően befolyásolhatják a klaszterezés eredményét.

A pontok és a klaszterek távolságfüggvényét az érintett alkalmazási terület módszertana határozza meg, a kialakult gyakorlatnak megfelelően. Az SPSS segíti a megfelelő távolságfüggvény megtalálását annyiban, hogy rákérdez az adatunk típusára, és így pl. az intervallum skálán mért adatainkhoz ezekre használhatókat ajánl fel egy listában.

Az eljárás végrehajtása

Az utasítás kiadása után kapott párbeszédes ablakban be kell állítani, hogy mely változók jellemzik az esetek közti eltéréseket (ezeket át kell tenni a jobb oldali kis ablakba). Megadhatunk egy esetcímkéket tartalmazó szöveges változót, ekkor az eredményben ezeket is láthatjuk, nem csak az esetek sorszámait.

A klaszterezés nemcsak az esetekre alkalmazható, hanem az eredeti változóinkat is tudja klaszterekbe sorolni. E kettő között a Cluster Cases / Variables rádiógombokkal választhatunk. Kérhetjük az érintett statisztikáknak és ábráknak az output ablakban való megjelentetését is (Display Statistics, Plots). Ha kértük ezeket, akkor a Statistics, illetve Plots feliratú gombok aktiválódnak, és megnyomásuk után beállíthatjuk, hogy konkrétan mi is legyen az output állományban. A dendrogram megjelentetése javasolt, ha az eredmény várhatóan áttekinthető marad. Van egy további forma is itt: az ún. icicle plot. Ez más formában jeleníti meg az összetartozó eseteket (vagy változókat): ezek sorszáma lesz a táblázat fejlécében megadva, és alatta klaszterezés szintjeinek megfelelő magasságú oszlop mutatja, hogy az adott eset mely másokkal tartozik egy klaszterbe az adott szinten. Itt kérhetjük a klaszterek számára vonatkozóan csak egy intervallumra is ezt a táblázatot. Ekkor meg kell adnunk a kezdő- és vég klaszter számát, illetve a léptéket erre vonatkozóan. Továbbá beállítható, hogy vízszintesen vagy függőlegesen rajzolja ki a táblázatot.

A legfontosabb beállítási lehetőségeket a Method gomb rejti: itt kell megadnunk, hogy melyik klaszterezési módszert választjuk (mi a klaszterek közti eltérés), melyik legyen az alapul vett távolságfogalom, és hogy milyen transzformációkat kérünk még. Ezeket mind az érintett alkalmazási feladat kell hogy eldöntse, de érdemes a következőkre figyelni: ha csak szoros kapcsolat esetén szeretnénk megengedni az összevonást, akkor válasszuk a Furthest neighbor (complete link) lehetőséget, ellenkező esetben pedig a Nearest neighbor-t (single link).

A pontok közti távolságfüggvény definiálásakor vigyázni kell, mert a program maga nem ismeri fel az alapul vett változók skálatípusát, azt a felhasználónak kell itt beállítani. Ha az adott alkalmazási területen nincs el-

fogadott távolság-definíció, akkor vegyük az alapbeállítást, illetve nézzünk utána az ajánlott távolságok sajátosságainak.

Sokat segíthet a megfelelő eredmény elérésében, hogy mind az alapul vett változókat, mind az eltérési mértékeket lehet transzformálni. Az előbbieknél a standardizálás egy szokásos eljárás, az utóbbiaknál az előjeleket lehet manipulálni (amik az eltérésen túl a reláció irányát jelzik). Végül a Save gomb megnyomása után kapott ablakban határozhatjuk meg, hogy a klaszterhez tartozás mely jellemzőjét mentse ki a program egy SPSS állományba.

Példa

Oldjuk meg a K-közép klaszterezésnél ismertetett feladatot a hierarchikus klaszterezés alapbeállításaival, de kérjünk dendrogramot a kiértékelés megkönnyítése kedvéért.

```
CLUSTER   x y z v
  /METHOD BAVERAGE
  /MEASURE=SEUCLID
  /PRINT SCHEDULE
  /PLOT DENDROGRAM VICICLE.
```

Case Processing Summary[a,b]

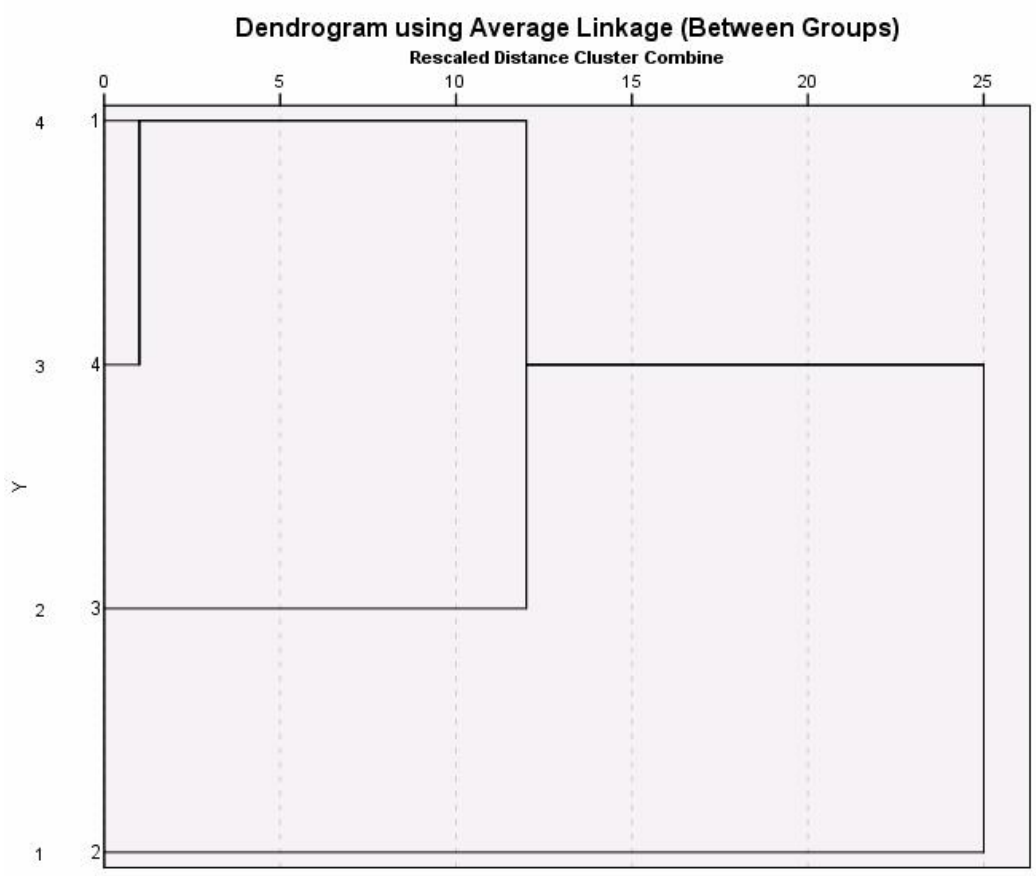
Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
4	100,0	0	,0	4	100,0

a Squared Euclidean Distance used
b Average Linkage (Between Groups)

Average Linkage (Between Groups)
Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	4,000	0	0	2
2	1	3	18,000	1	0	3
3	1	2	33,333	2	0	0

A dendrogramból (3.2. ábra) leolvasható, hogy a megadott távolság-definíció mellett az 1-es és 4-es sorszámú esetek lehetnek egymáshoz a legközelebb. Majd a 3-as sorszámú eset van hozzájuk a legközelebb, míg a 2-es eset mindegyiktől a legtávolabb. Ezeket a legkésőbb, már csak 1 klaszter esetén rakja be egy csoportba a többivel. érdemes megfigyelni, hogy amíg az 1-es és 4-es eset csak alacsony szinten válik el a 3-as esettől, addig a 2-es sorszámú eset sokkal nagyobb szinten vonható össze a többivel.



3.2. ábra. A klaszterezés dendrogramja.

Régió	2001	2002	2003
Dél-Alföld (Bács-Kiskun, Békés, Csongrád megye)	81 246	97 761	111 892
Dél-Dunántúl (Baranya, Somogy, Tolna megye)	83 857	100 729	113 947
Észak-Alföld (Hajdú-Bihar, Jász-Nagykun-Szolnok, Szabolcs-Szatmár-Bereg megye)	80 847	98 054	111 807
Észak-Magyarország (Borsod-Abaúj-Zemplén, Heves, Nógrád megye)	84 640	101 679	116 283
Közép-Dunántúl (Fejér, Komárom-Esztergom, Veszprém megye)	95 558	112 191	124 241
Közép-Magyarország (Budapest, Pest megye)	127 488	149 203	165 005
Nyugat-Dunántúl (Győr-Moson-Sopron, Vas, Zala megye)	90 926	106 405	118 928

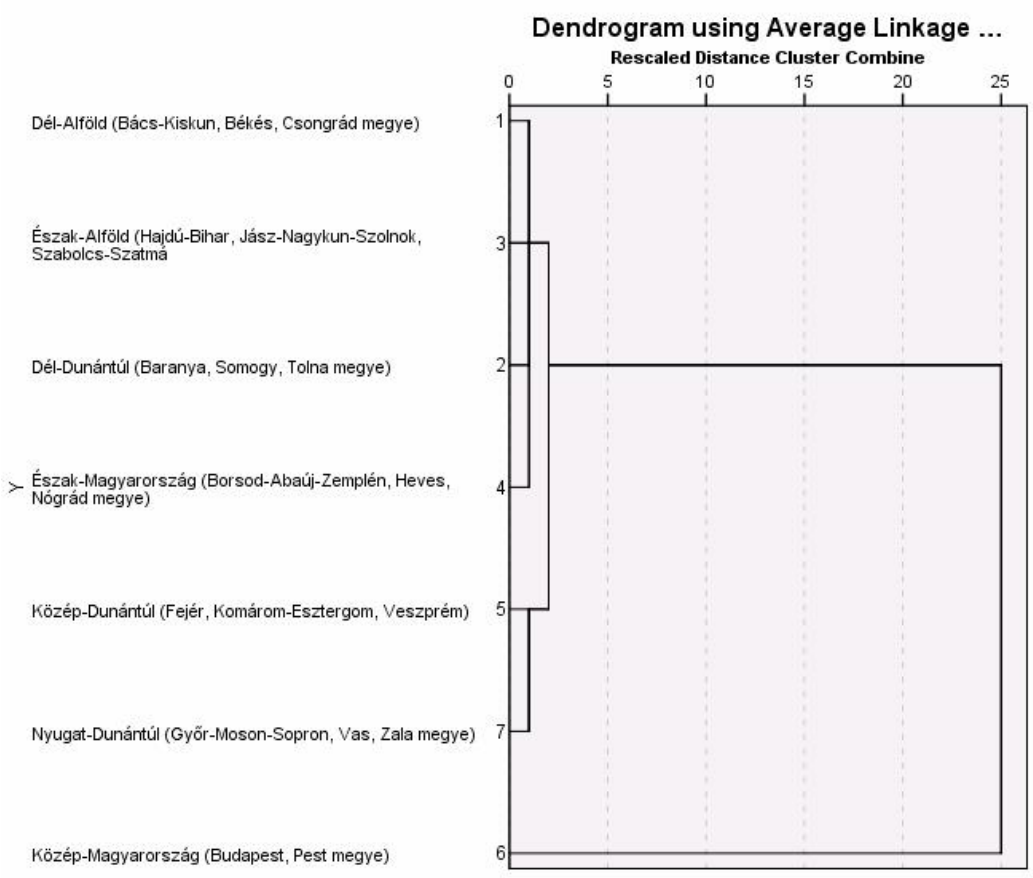
3.8. táblázat. Magyarországi havi átlagkeresetek régiónként 2001-től 2003-ig (Ft/fő).

Gyakorlat

Magyarországi jövedelem adatok állnak a rendelkezésünkre. Vizsgáljuk meg, hogy hogyan viszonyulnak egymáshoz a jövedelmek. Vizsgáljuk meg, ha csoportosítanunk kellene a régiókat jó és rossz átlagkereset alapján, akkor hogyan alakulnának ezek a halmazok. Először vigyük be a 3.8. táblázat adatait az SPSS-be.

Mivel kezdetben nem tudjuk, hogy hány csoportba volna érdemes sorolnunk a régiókat kereset alapján, így érdemes megvizsgálni a kérdést a hierarchikus klaszterezéssel. Ehhez válasszuk a Hiearchical Cluster parancsot az Analyze / Classify menüpontból. A parancs hatására megjelenő párbeszédes ablakban helyezzük át a jövedelmeket tartalmazó változókat a Variables részbe. Bármely évet bevehetünk a klaszterezésbe, de az időszakra leginkább jellemző csoportosítást minden jövedelmet tartalmazó oszlop áthelyezésével kapjuk. Ekkor minden évet figyelembe vesz a csoportosításhoz. Végül a címkézésnek beállíthatjuk a régiók neveit tartalmazó változót. Továbbá a Plots gombbal megadhatjuk, hogy kérünk dendrogramot, melynek eredménye a 3.3. ábrán látható.

Az ábrát elemezve észrevehetjük, hogy a közép-magyarországi régió nagyon magas szinten kerül egy csoportba a többi régióval. A számokat megvizsgálva látható, hogy ebben a régióban a többihez képest tényleg jelentősen nagyobb a jövedelem minden évet tekintve. Az ábra azt sugallja nekünk, hogy a leginkább kézenfekvő klaszterszám a 2. Gyakorlatilag van a nagy jövedelmű Közép-Magyarország, és vannak a kis jövedelmű egyéb régiók. Ha 3 csoportot szeretnénk létrehozni, akkor Dél-Alföld, észak-Alföld, Dél-Dunántúl és észak-Magyarország lenne egy csoport, míg Közép-Dunántúl és Nyugat-Dunántúl egy másik csoport. De ez a szétszedés már nem igazán indokolt, nagyon alacsony szinten válnak szét.



3.3. ábra. A klaszterezés dendogramja a Magyarországi havi átlagkeresetekre régiónként.

3.6. Az adat tömörített jellemzése

A dimenziócsökkentés (Dimension Reduction) menüsorban többek között a faktoranalízis található. Ennek a gyorsító billentyűkombinációja az Alt-adf.

3.6.1. Faktoranalízis

A faktoranalízis arra való, hogy több változóhoz meghatározzunk kevés számú olyan új változót (faktort), amelyekkel az alapadatunk varianciája a lehető legjobban megmagyarázható. Tehát egy statisztikai mintában segít felismerni azokat a rejtett magyarázóváltozókat, amelyek a mintát kevesebb változóval tudják leírni. A rejtett változók (faktorok) lineáris kombinációjaként előállnak az eredeti változók (kis hibától eltekintve).

A faktoranalízis különösen hasznos olyan esetekben, amikor a háttérben lévő, nem ismert magyarázóváltozókat (faktorokat) keressük, és bizonyos értelemben túl sok változó, adat írja le a rendszerünket, és valójában ehhez sokkal kevesebb is elég lenne.

Gyakran emlegetett példa ilyen esetre a politikai választások szavazási preferenciájának valódi „okainak” megkeresése. Ilyenkor sok adat áll rendelkezésre arról, hogy a közvéleménykutatásban résztvevő személy mely társadalmi rétegből származik, milyen nemű stb. A kérdés pedig az, hogy hogyan lehet azt egyszerűen, kevés adattal megmondani, hogy egy adott pártra leginkább kik szavaznak. A megtalált faktorokat aztán úgy szokás megadni, hogy az A pártra való szavazási hajlandóságot leginkább az befolyásolja, hogy az illetőre milyen mértékben igaz az pl., hogy jól kereső városlakó, aki középkorú és felsőfokú végzettsége van. Ilyenkor az utóbbi változókból képzett faktor lehetett az, ami a faktoranalízis eredményében szerepelt.

Az eljárás elindításakor kapott párbeszédes ablakban először ismét csak az analízis során figyelembe veendő változókat kell megadni úgy, hogy az érintetteket áttesszük a bal oldaliból a jobb oldali ablakba. Ha nem az összes esetet szeretnénk feldolgozni, akkor meg lehet adni egy szűrőváltozót (Selection Variable), ami megmutatja, hogy mely eseteket kell a feldolgozás során figyelembe venni. Ennyi információ birtokában az algoritmus már végrehajtható – köszönhetően az ésszerű alapbeállításoknak.

A sokféle speciális opció közül itt csak néhányat tárgyalunk. A Descriptives gomb megnyomása után például kérhetjük a korrelációs mátrix együtthatóinak, a szignifikancia-szinteknek, a determinánsnak, illetve az inverz mátrixnak a megjelenítését.

Az Extraction gomb fontos algoritmus részleteket rejt: itt lehet megadni, hogy melyik alapmódszer szerint történjen a faktorok meghatározása:

a főkomponens analízis (Principal components), a súlyozatlan- vagy az általános legkisebb négyzetek módszerével (Unweighted or Generalized least squares), a maximum likelihood elven stb. Mint más hasonló esetben, most is ezeknek a beállításoknak első sorban az érintett szakterület kialakult módszertanának kell megfelelnie. Fontos az is, hogy mi alapján dől az el, hogy hány faktort fog az eljárás megadni. Kérhetjük azt, hogy csak egy megadott számított sajátértéken felülieknek megfelelőekhez tartozókat (az alapbeállítás az 1-nél nagyobbak), vagy hogy egy megadott számú faktort kérünk. Ugyanitt adhatjuk meg a maximális megengedett iterációszámot, amivel az eljárás futásidejét, illetve a megoldás finomságát befolyásolhatjuk.

Ezeken kívül is számos beállítási lehetőség van, így kérhetjük több módszerrel a faktorok rotációját, illetve megadhatjuk a kívánt hiányzóadat kód kezelési eljárást.

Példa

Generáljunk egy olyan adatot, amely bár négy változóval van megadva, mégis lényegében egy magyarázóváltozóval leírható.

Először is vigyünk be egy változót néhány esetnyi véletlenül választott adattal: a lényeg az, hogy sok különböző érték legyen benne. Ebből a kiindulási v változóból a Transform / Compute utasítással képezzünk három újat: az x, y és z változókat rendre a következő képletekkel

$$x = v * SQRT(3)/2 + NORMAL(0.1),$$

$$y = v/4 + NORMAL(0.1)$$

és

$$z = v * SQRT(3)/4 + NORMAL(0.1).$$

Ezzel lényegében x, y és v-ben egy olyan háromdimenziós pontfelhőt adtunk meg, amelynek a formája egy erősen elnyújtott ellipszoid, és amelynek főtengelye sokkal hosszabb, mint a többi. A normális eloszlású véletlen számokra azért volt szükség a képletekben, hogy az adatunk ne pontosan egy egyenesre illeszkedjen. A kis megadott szórás azonban nem nagy eltéréseket jelent.

Az alapértelmezésbeli értékekkel végrehajtott faktoranalízis a következő eredményt adja[5]:

```
FACTOR
  /VARIABLES v x y z
```

[5] A véletlen adat miatt megismétléskor más eredményt kapunk.

```
  /MISSING LISTWISE
  /ANALYSIS v x y z
  /PRINT INITIAL EXTRACTION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /METHOD=CORRELATION.

Communalities
|-|-------|----------|
| |Initial|Extraction|
|-|-------|----------|
|v|1,000  |,995      |
|-|-------|----------|
|x|1,000  |,996      |
|-|-------|----------|
|y|1,000  |,977      |
|-|-------|----------|
|z|1,000  |,986      |
|--------------------|
Extraction Method: Principal Component Analysis.

Total Variance Explained
|---------|---------------------------------|---------------------------------|
|Component|Initial Eigenvalues              |Extraction Sums of Squared Loadin|
|         |-------|-------------|-----------|------|-------------|-----------|
|         |Total  |% of Variance|Cumulative %|Total |% of Variance|Cumulative %|
|---------|-------|-------------|-----------|------|-------------|-----------|
|1        |3,955  |98,863       |98,863     |3,955 |98,863       |98,863     |
|---------|-------|-------------|-----------|------|-------------|-----------|
|2        |,033   |,829         |99,692     |      |             |           |
|---------|-------|-------------|-----------|------|-------------|-----------|
|3        |,009   |,219         |99,911     |      |             |           |
|---------|-------|-------------|-----------|------|-------------|-----------|
|4        |,004   |,089         |100,000    |      |             |           |
|-------------------------------------------------------------------------|
Extraction Method: Principal Component Analysis.

Component Matrixa
|-|---------|
| |Component|
| |---------|
| |1        |
|-|---------|
|v|,998     |
|-|---------|
|x|,998     |
|-|---------|
|y|,989     |
|-|---------|
|z|,993     |
|-----------|
Extraction Method: Principal Component Analysis.
 a 1 components extracted.
```

A zajmentes esetben az utolsó táblázatban a komponensek egyesek lettek volna, ami azt jelzi, hogy az első komponens mindhárom kiindulási változó hatását statisztikai értelemben jól megmagyarázza.

A második táblázat szerint pedig az első komponenshez csaknem négyes sajátérték (3,955) tartozik, és ez a variancia közel 99%-át megmagyarázza. A másik három sajátérték lényegesen kisebb, ami a további magyarázóváltozó

szükségtelenségét jelzi. A program alapbeállításában nem is engedi egy alatti sajátértékekhez tartozó faktorok meghatározását.

érdemes az előző feladatot megismételni úgy, hogy egy magyarázó változó helyett több is legyen (v1, és v2). Vegyük például az alábbi képletekkel kapott új változókat:

$$x1 = v1 * SQRT(3)/2 + NORMAL(0.1),$$

$$x2 = v1/4 + NORMAL(0.1) + v2/4 + NORMAL(0.1),$$

$$x3 = v2/2 + NORMAL(0.1),$$

és

$$x4 = v1 - v2 + NORMAL(0.1).$$

Ekkor az alábbi táblázatot kapjuk:

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,338	72,292	72,292	4,338	72,292	72,292
2	1,623	27,046	99,338	1,623	27,046	99,338
3	,035	,580	99,918			
4	,004	,059	99,977			
5	,001	,018	99,995			
6	,000	,005	100,000			

Extraction Method: Principal Component Analysis.

Látható, hogy ilyenkor akár kettő 1-nél nagyobb sajátérték is megjelenhet, illetve az 1-es komponens csak 72%-ban képes magyarázni a változókat, míg az első kettő komponens már együtt több mint 99%-ban magyarázza.

Gyakorlat

Vigyük be a 3.9. táblázat adatait az SPSS-be. Vizsgáljuk meg az adatokat, és vegyük észre, hogy az adatok mindegyike jellemzően csökkenő tendenciát mutat. Első lépésben vizsgáljuk meg, hogy ezen csökkenés jól magyarázható-e valamely változó segítségével. Ehhez használjuk a SPSS Faktoranalízis parancsát.

Az eredményekből látható, hogy az első főkomponens relatíve nagy mértékben képes magyarázni az eredeti változók értékeit (76%). A komponens táblázatból látszódik, hogy leginkább az osztályok száma magyarázható,

Tanév	Tanulók/nappali	Lányok/nappali	Helyek	Pedagógusok száma	Osztályok száma
1990/91	1 166 076	566 475	3 723	96 791	52 675
1991/92	1 112 374	542 741	3 820	95 559	52 254
1992/93	1 071 727	524 686	3 901	94 980	51 932
1993/94	1 032 025	507 528	3 962	95 753	51 020
1994/95	1 001 709	496 007	4 010	96 141	50 578
1995/96	987 561	490 436	4 006	93 035	49 178
1996/97	976 423	486 347	3 965	89 792	48 184
1997/98	973 401	470 934	3 952	89 238	48 119
1998/99	973 326	470 193	3 931	89 570	48 314
1999/00	969 755	468 531	3 897	89 424	47 813
2000/01	957 850	463 064	3 875	89 750	47 845
2001/02	944 244	456 759	3 852	90 294	47 865
2002/03	930 386	449 537	3 793	89 035	46 723
2003/04	909 769	439 380	3 748	89 784	46 006
2004/05	887 785	428 173	3 690	87 116	45 057

3.9. táblázat. A magyarországi nappali képzés alapadatai 1990-től 2005-ig.

míg legkevésbé a feladat ellátási helyek száma. Tehát elmondható, hogy a hallgatók száma és a többi változó változása (kivéve a feladat ellátási helyek száma) egy egyenes mentén helyezkednek el, azaz például a tanulók számának változása jól magyarázza a többi változót. Ezen egyenes egyéb tulajdonságai további vizsgálatokat igényelnek.

3.7. Skálázás

A skálázás algoritmusai közül csak a többdimenziós skálázással foglalkozunk, a megbízhatósági analízissel (Reliability Analysis) nem. Az előbbi elérhető az Alt-aam gyorsító billentyűkombinációval is. Az utóbbi a használt skálák megbízhatóságával foglalkozik, tehát például hogy egy kérdőív kérdései mennyiben függenek össze, illetve hogy hol kell javítani ezeken.

3.7.1. Többdimenziós skálázás

A többdimenziós skálázás célja, hogy egy távolságokkal együtt megadott többdimenziós ponthalmazban meglévő struktúrát kiderítse, illetve megjelenítse. Ennek elérése kedvéért egy olyan 2- vagy 3-dimenziós térbeli modell készül, amelyben az egyes pontok távolságai amennyire csak lehetséges, megfelelnek az eredeti többdimenziós térbeli pontokénak. Ennyiben tehát az eredeti tér „átskálázásáról" van szó.

A leképezett térbeli pontokat azután meg lehet jeleníteni, és a pontok egymáshoz való viszonyát közvetlenül, vizuálisan is lehet vizsgálni. Emiatt ez a statisztikai eljárás (például a klaszterezéssel és a faktoranalízissel együtt) ismét egy olyan algoritmus, amelyhez hasonlót más számítógépes diszciplínák, mint a numerikus analízis vagy az operációkutatás nem kínálnak.

A leggyakoribb eset a többdimenziós skálázás használatára, amikor több eset által leírt változók egymáshoz való viszonyát szeretnénk vizuális módon tisztázni. Tegyük fel, hogy az adataink már rendezett módon az SPSS táblázatában vannak.

Az utasítás kiváltása után kapott párbeszédes ablakban először az érintett változókat kell megadni úgy, hogy azoknak a nevét átvisszük a jobb oldali kis ablakba. Több változó gyors kijelöléséhez elég a bal egérgombot nyomva tartva végighúzni a kurzort az érintett változók nevein. Az egyik fontos beállítás az, hogy megadjuk, az adataink távolságmátrix vagy eredeti változók formájában adottak-e. Az alapértelmezés az előbbi, habár a többdimenziós skálázás első alkalmazásai során valószínűleg nem áll rendelkezésre előre kiszámított távolságmátrix. A Distances rovatban tehát ebben az esetben a rádiógomb beállításával adjuk meg, hogy „Create distances from data". Ha az euklideszi távolsággal nem vagyunk elégedettek, akkor a Measure gomb megnyomása után kapott újabb párbeszédes ablakban választhatunk a rendelkezésre állók közül.

Fontos és nem nyilvánvaló beállítási lehetőségek vannak a Model és az Options gombokkal elérhető formában. A Model gomb megnyomása után tudjuk például a mérési skálát beállítani, az Options nyomógomb pedig az eredmények megjelenítésére vonatkozó részletek tisztázását teszi lehetővé. Így a Display rovatban a Group plots lehetőséget mindenképpen ajánlatos kiválasztani, mert csak ekkor kapunk tényleges grafikus megjelenítést az eredményünkről. Az iterációs algoritmus megállási feltételeit pedig a Criteria rovatban hangolhatjuk.

Példa

Egy egyszerű példán keresztül mutatjuk meg a többdimenziós skálázás használhatóságát. Tegyük fel, hogy egy kérdőívre kapott válaszokat tartalmaz a következő adathalmaz:

```
1,78  1,00  ,00  ,00  ,00
1,76  1,00 1,00  ,00  ,00
1,88   ,00  ,00  ,00 1,00
1,65  1,00 1,00 1,00 1,00
1,59  1,00 1,00  ,00  ,00
1,93   ,00  ,00  ,00  ,00
1,75  1,00 1,00  ,00  ,00
2,01  1,00 1,00  ,00  ,00
```

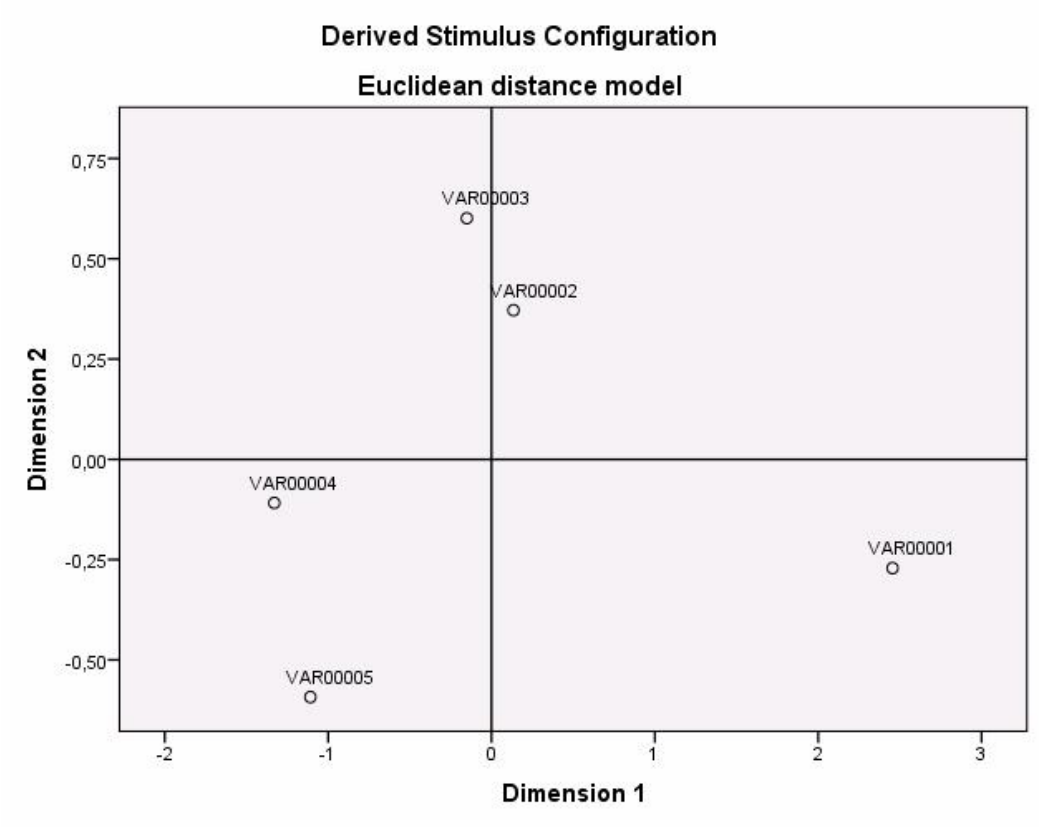
Ránézésre is látszik, hogy az első oszlopban lévő változó lényegesen eltér a többitől, és hogy a második és a harmadik, illetve a negyedik és ötödik változó mért értékei esnek közel egymáshoz. Ebben az értelemben tehát azt mondhatjuk, hogy ha a felmérésbeli változók számát csökkenteni kellene, akkor amiatt, hogy az adathalmaz eltéréseinek nagy részét három változó már hozza, elegendő lehet csak ezeknek a figyelembe vétele, vagy a következő felmérésre való használata a mérési adatok sokféleségének reprezentálására. Nézzük meg, hogy mit ad ebben az esetben az SPSS többdimenziós skálázása!

A bevezető leírásnak megfelelő beállításokkal végrehajtva az eljárást (tehát minden változóra, távolságmátrix nélkül, kérve a Group plots lehetőséget) a várakozásokak megfelelő eredményt kapjuk. Az eredmény a 3.4. ábrán látható.

Az eredmény megvitatása előtt előre kell bocsátani, hogy az első kivételével a változóink aligha intervallum skálán mértek, de az ebből adódó értelmezésbeli problémáktól most tekintsünk el: a fő cél most az eljárás szemléltetése. Az ábrán azonnal látszik, hogy az első változónk (nem adtunk meg változóneveket, ez tehát a var00001) lényegesen távolabbra került a többitől, míg a sejtett változópárok (var00002 és var00003, illetve var00004 és var00005) egyértelműen egymás közelébe kerültek. Az eredeti tér 8 dimenziós volt, abban természetesen ezeket az összefüggéseket ilyen könnyen nem tudtuk volna megállapítani.

3.7.2. Gyakorlat

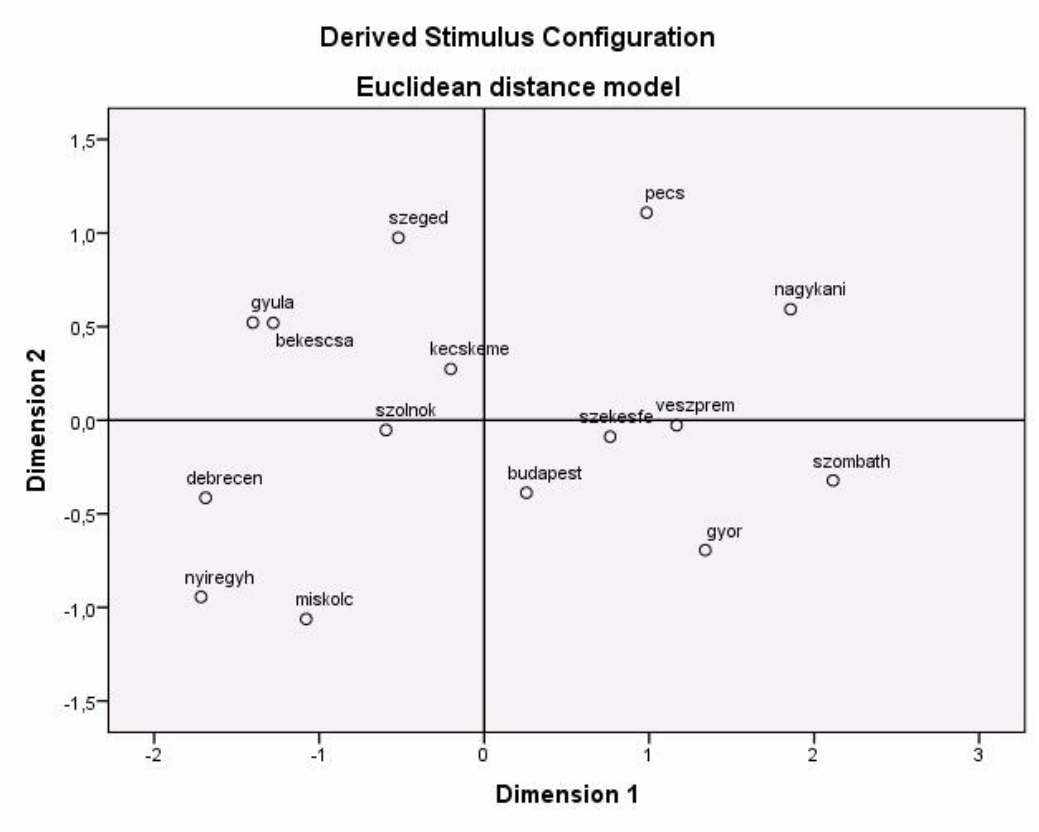
Gyakorlásként vigyük be Magyarország nagyobb városainak távolságát. Ezeket az adatokat a 3.10. táblázat tartalmazza. Ezek után hívjuk meg a Multidimensional Scaling eljárást. Mivel most a távolságmátrix adott, így a Distances résznél az első lehetőséget kell megadni. Továbbá a távolságaink szimmetrikusak, így a kezdeti beállítás is helyes (Square symmetric). A Model gombnál sem szükséges további beállításokat tenni, a kétdimenziós eset érdekel minket. Az Options lehetőségnél állítsuk be a Group plots-ot, hogy grafikusan is megjelenjen az eredmény (lásd a 3.5. ábrát). Az így megjelenő ábrán gyakorlatilag a nagyobb városok valódi egymáshoz viszonyított helyzetét látjuk. Az eljárás által adott további ábrákon látható, hogy a megadott távolságok és a megjelenített távolságok nagyon jól korrelálnak, azaz a megjelenített ábrán lévő távolságok jól illusztrálják a megadott távolságokat. Az eredmény várható volt, ugyanis a bevitt távolságok egyik lehetséges megjelenése a valóság. De mint látható a forgatással és tükrözésekkel kapott megoldások is teljesen hasonlóak.



3.4. ábra. A Multidimensional Scaling eljárással kapott ábra, amely a példában megadott változók helyzetét mutatja.

város	sz.	n.	gy.	v.	p.	sz.	b.	k.	sz.	sz.	m.	b.	gy.	d.	ny.
Szombathely	0														
Nagykanizsa	90	0													
Győr	93	144	0												
Veszprém	99	99	68	0											
Pécs	178	103	185	117	0										
Székesfehérvár	138	135	82	40	125	0									
Budapest	190	189	110	99	170	60	0								
Kecskemét	238	216	180	135	144	103	81	0							
Szeged	290	238	246	193	150	167	160	81	0						
Szolnok	270	252	202	171	193	130	94	49	103	0					
Miskolc	330	333	243	243	297	208	145	153	210	110	0				
Békéscsaba	345	310	284	248	230	212	177	112	88	85	160	0			
Gyula	360	324	297	256	243	220	189	124	97	101	166	13	0		
Debrecen	380	369	300	284	305	247	195	162	183	117	90	105	99	0	
Nyíregyháza	393	392	305	301	335	263	205	193	220	144	73	152	148	47	0

3.10. táblázat. Magyarország nagyobb városainak távolsága.



3.5. ábra. A Multidimensional Scaling eljárással kapott ábra, amely Magyarország városainak a helyzetét mutatja.

4. fejezet

További példák

4.1. Elemzések Magyarországi adatsorokon

Az alábbi részben további példákat fogunk tárgyalni. Ezek elemzése során nem feltétlenül csak egy-egy konkrét statisztikai eszközt fogunk használni, így a korábbi fejezetek egyikéhez sem lehet az alábbiakat közvetlenül társítani. A szerzők remélik, hogy ezen valóságközeli példák is segítik az olvasót a különböző statisztikai eljárások megértésében.

4.1.1. Magyarország fogyasztási szokásainak elemzése

A Magyar statisztikai évkönyvben ([12]) több érdekes elemzésre alkalmas adathalmazt találhatunk. Ezek közül most az egy főre jutó élelmiszerfogyasztást vizsgáljuk. Ebben a táblázatban (lásd a 4.1. táblázat) megtalálható Magyarország lakosságának 1 főre jutó élelmiszer fogyasztása 1960-tól 2010-ig az alábbi kategóriákban: Hús, húskészítmény, hal; Tej és tejtermékek; Tojás; Zsiradékok; Liszt és rizs; Cukor; Burgonya.

Az ilyen jellegű adatok között sok esetben összefüggés van. Az elemzés egyik módszere lehet a regresszió számítás, de ennek segítségével elég nehéz megállapítani, hogy mely változókat vegyünk is be a további vizsgálatainkba. Ehelyett az eredeti magyarázóváltozókat olyan új változókkal próbáljuk helyettesíteni, amelyek dimenziószáma kisebb, de megtartja a magyarázóváltozókban lévő információkat. Jelen példában a magyarázóváltozók 7 dimenziós terét próbáljuk csökkenteni. A transzformált tér dimenzió számára is kapunk majd javaslatot.

Első lépésként vigyük be a 4.1. táblázatban lévő adatokat (lásd a 4.1. ábra). Ezek után válasszuk az SPSS Factor Analysis menüpontját.

Itt a vizsgált változók közé vegyük fel az összes változót, kivéve az év-et tartalmazó változót. Ha a későbbiekben a kapott transzformált adatokon szeretnénk tovább dolgozni, akkor a Scores gomb megnyomása után állítsuk be a Save lehetőséget. Látni fogjuk, hogy a második főkomponens is elég nagy (bár 1-nél éppen kisebb), ezért az Extraction gomb megnyomása után állítsuk be, hogy 2 főkomponenst szeretnénk kapni a Fixed number of factors lehetőségnél. Ezek után az OK gomb megnyomása után az alábbi eredményt kapjuk:

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,121	73,158	73,158	5,121	73,158	73,158
2	,961	13,723	86,881	,961	13,723	86,881
3	,542	7,742	94,623			
4	,175	2,495	97,118			
5	,103	1,467	98,584			
6	,065	,930	99,515			
7	,034	,485	100,000			

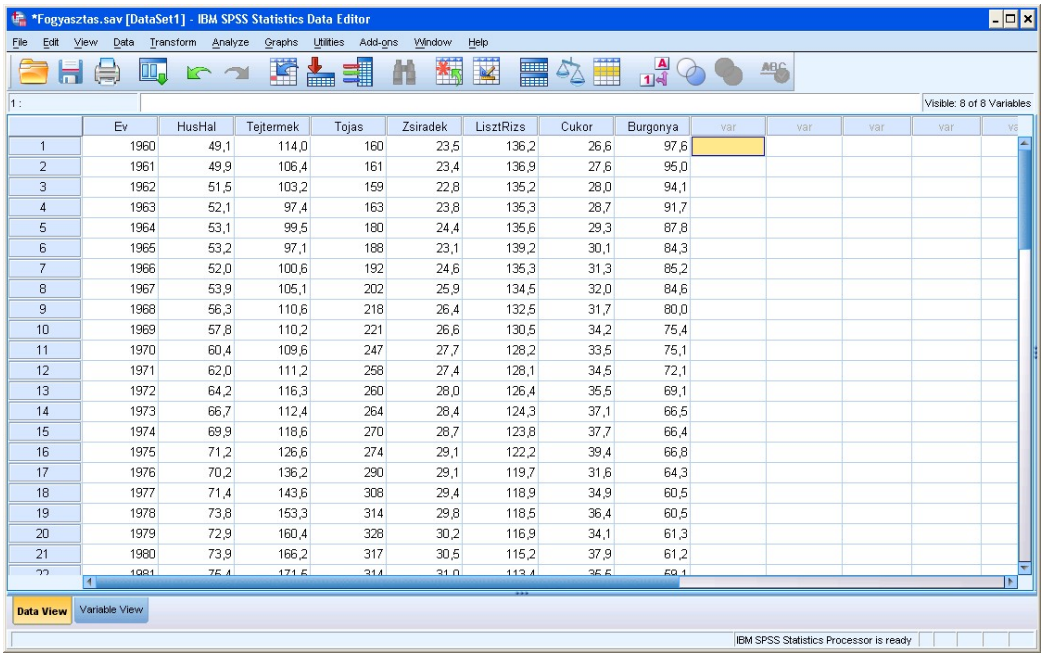
Extraction Method: Principal Component Analysis.

Component Matrixa

	Component	
	1	2
HusHal	,911	,277
Tejtermek	,890	-,014
Tojas	,929	,223
Zsiradek	,879	-,424
LisztRizs	-,705	,660
Cukor	,670	,459
Burgonya	-,959	-,088

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Az első táblázatból leolvasható, hogy az első főkomponens a magyarázóváltozók szórásnégyzetének közelítőleg 73%-át magyarázza, míg az első kettő komponens 86%-át. Az második táblázatból látható, hogy az eredeti változók és a főkomponensek között milyen szoros a kapcsolat. A legjobban a zsiradék fogyasztását magyarázza, míg legkevésbé a cukor és liszt fogyasztását. A második komponens pont ezeket a komponenseket tudja a legjobban magyarázni, bár már nem olyan erősen.



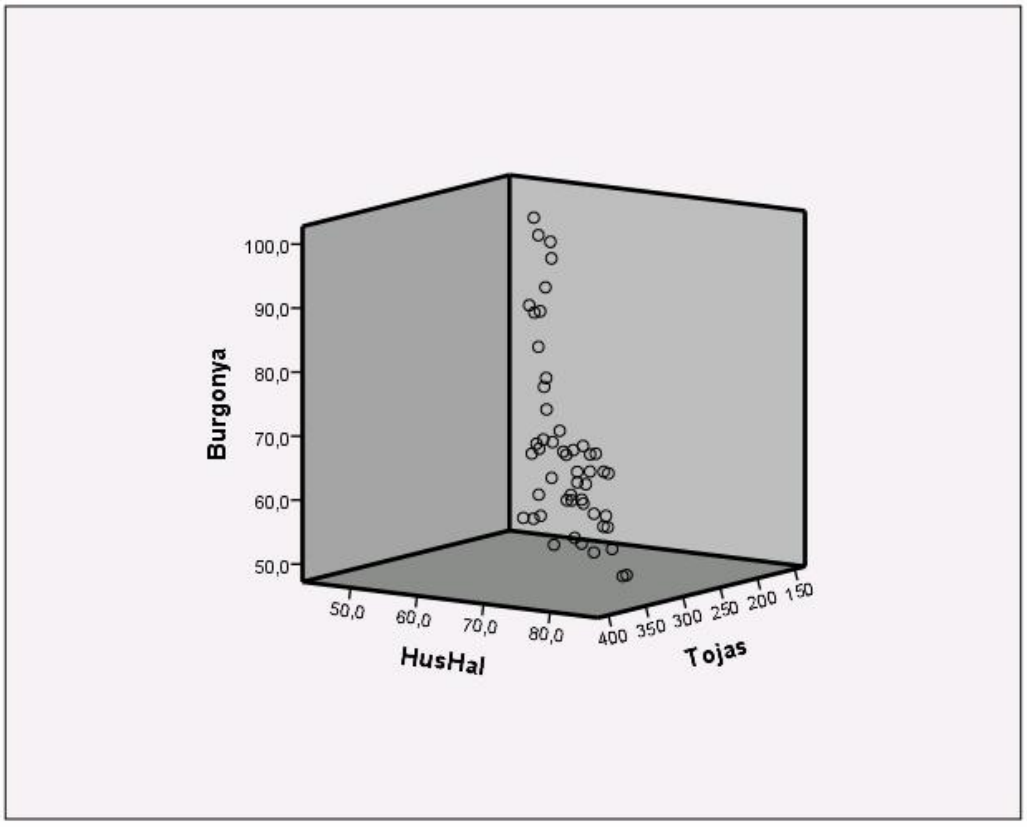
4.1. ábra. Magyarország egy főre jutó fogyasztása az SPSS-ben.

Ezután nézzük meg, hogy a három legjobban magyarázott változó hogyan helyezkedik el a térben. Ehhez használjuk a Graphs menüpont alatt található Scatter/Dot grafikus ábrázolási lehetőséget. Itt válasszuk a 3-D-s lehetőséget. Majd az X, Y és Z tengelyeknek adjuk meg a Burgonya, Hus és Tojas változókat. Látható (lásd a 4.2 ábrán), hogy megközelítőleg valóban egy egyenes mentén helyezkednek el. Ez a tulajdonság a 7 dimenziós térben is nagyjából igaz lenne. Valójában az elhelyezkedésről sokkal jobb képet mutat a főkomponensek terében lévő koordináták megjelenítése. Ehhez használjuk a Scatter/Dot grafikus ábrázolási lehetőség 2 dimenziós változatát, melynél az X és Y értékeknek állítjuk be a faktoranalízis által kapott értékeket (FAC1_1 és FAC2_1). Ezt az ábrát láthatjuk a 4.3. ábrán. Itt is látható, hogy inkább egy egyenesen helyezkednek el a pontok, de elég erős a szórás a második dimenzióban is.

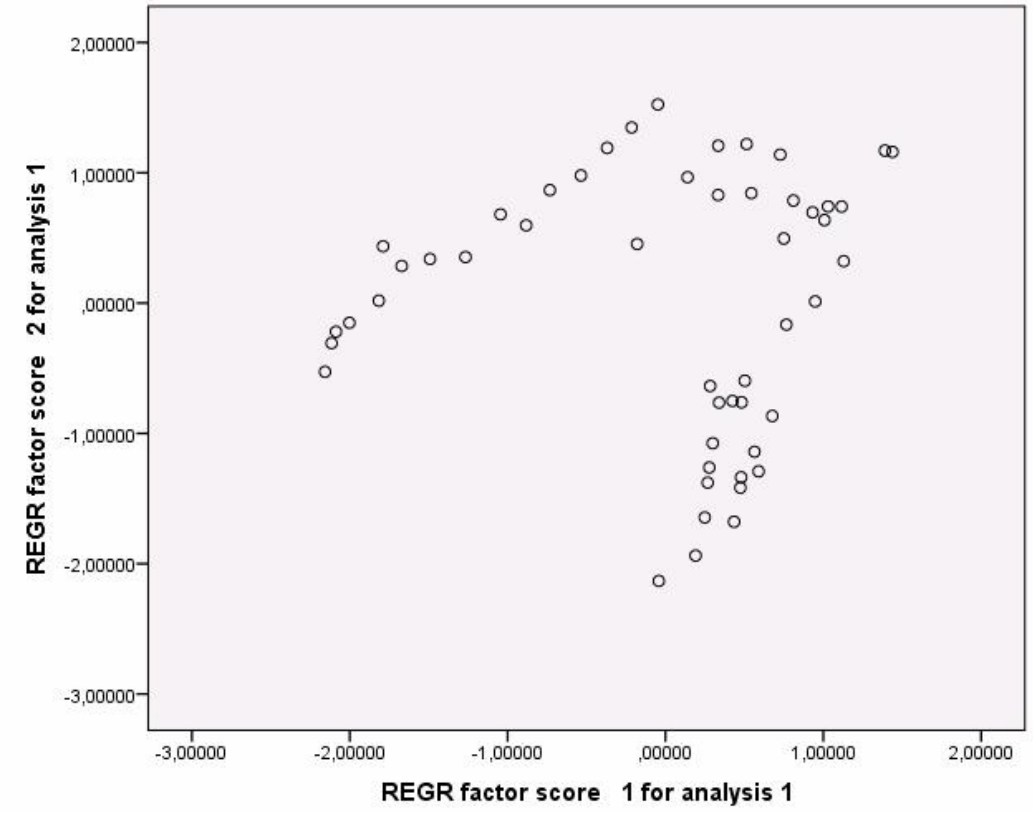
Ezen statisztikai eredményekből azt a következtetést vonhatjuk le, hogy Magyarországon az egy főre jutó fogyasztásban fellelhető egy arányos változás. Mivel ezt főként 1 komponens magyarázza, így ezen 7 élelmiszercsoportban nem lelhető fel olyan részcsoportok, melyeken belül külön-külön hatnának egymás fogyasztott mennyiségeire.

év	Húskészítmény kg/év	Tejtermék kg/év	Zsiradékok db/év	Tojás kg/év	Liszt és rizs kg/év	Cukor kg/év	Burgonya kg/év
1960	49,1	114,0	160	23,5	136,2	26,6	97,6
1961	49,9	106,4	161	23,4	136,9	27,6	95,0
1962	51,5	103,2	159	22,8	135,2	28,0	94,1
1963	52,1	97,4	163	23,8	135,3	28,7	91,7
1964	53,1	99,5	180	24,4	135,6	29,3	87,8
1965	53,2	97,1	188	23,1	139,2	30,1	84,3
1966	52,0	100,6	192	24,6	135,3	31,3	85,2
1967	53,9	105,1	202	25,9	134,5	32,0	84,6
1968	56,3	110,6	218	26,4	132,5	31,7	80,0
1969	57,8	110,2	221	26,6	130,5	34,2	75,4
1970	60,4	109,6	247	27,7	128,2	33,5	75,1
1971	62,0	111,2	258	27,4	128,1	34,5	72,1
1972	64,2	116,3	260	28,0	126,4	35,5	69,1
1973	66,7	112,4	264	28,4	124,3	37,1	66,5
1974	69,9	118,6	270	28,7	123,8	37,7	66,4
1975	71,2	126,6	274	29,1	122,2	39,4	66,8
1976	70,2	136,2	290	29,1	119,7	31,6	64,3
1977	71,4	143,6	308	29,4	118,9	34,9	60,5
1978	73,8	153,3	314	29,8	118,5	36,4	60,5
1979	72,9	160,4	328	30,2	116,9	34,1	61,3
1980	73,9	166,2	317	30,5	115,2	37,9	61,2
1981	75,4	171,5	314	31,0	113,4	35,5	59,1
1982	76,8	174,8	308	31,8	113,1	38,1	57,0
1983	78,4	181,4	328	32,9	111,4	35,7	57,9
1984	78,0	185,0	321	33,5	111,3	34,3	59,3
1985	79,6	183,2	327	34,1	110,8	35,5	54,5
1986	81,0	185,6	320	34,1	110,1	35,7	50,4
1987	81,3	199,1	328	37,6	113,0	40,1	50,5
1988	78,7	195,6	360	37,0	109,3	34,3	56,2
1989	81,0	189,6	364	39,2	112,2	40,5	55,2
1990	75,8	169,9	389	38,6	110,4	38,2	61,0
1991	74,1	167,4	356	37,0	102,6	35,0	55,3
1992	75,2	159,1	338	37,5	105,6	39,5	56,0
1993	70,5	144,2	365	36,8	97,4	35,8	59,3
1994	69,0	140,0	338	38,1	91,3	34,2	58,2
1995	65,2	132,1	297	36,7	88,2	37,3	60,3
1996	61,9	136,4	267	35,7	84,6	39,8	66,2
1997	60,8	156,4	267	36,1	88,1	39,4	65,3
1998	63,7	149,6	265	36,2	84,1	41,3	67,4
1999	63,3	151,7	274	34,2	90,4	37,7	68,0
2000	73,2	160,6	275	39,0	94,1	33,2	64,0
2001	70,4	144,2	284	37,4	95,3	32,9	68,2
2002	75,4	143,1	301	39,0	87,8	32,6	65,3
2003	71,9	138,3	288	39,2	88,3	32,8	64,5
2004	64,3	155,2	292	36,0	89,4	32,7	68,0
2005	67,1	166,8	281	36,5	97,3	31,2	66,8
2006	69,6	163,1	273	37,7	92,0	32,3	61,8
2007	67,0	163,5	270	37,4	88,3	31,2	59,7
2008	65,3	158,2	261	36,8	88,9	31,9	65,5
2009	65,4	155,9	247	36,6	88,4	29,8	60,8
2010	60,2	156,8	235	34,6	88,2	28,7	60,5

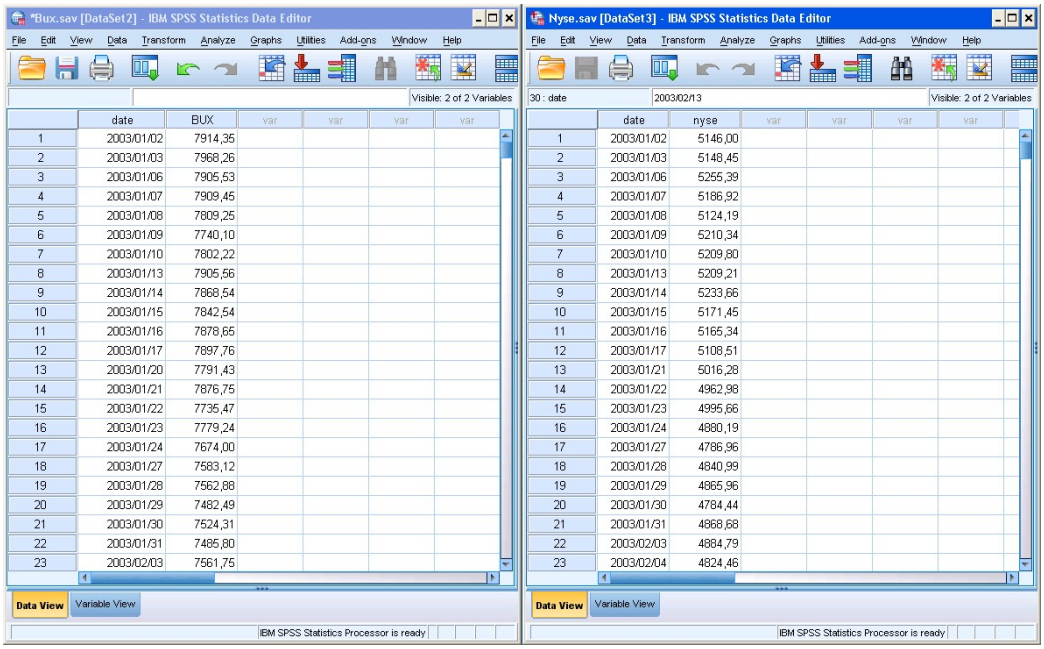
4.1. táblázat. Magyarország egy főre jutó élelmiszerfogyasztása.



4.2. ábra. Magyarország egy főre jutó fogyasztása a burgonya, hús és tojás tekintetében.



4.3. ábra. Magyarország egy főre jutó fogyasztása a főkomponensek terében.

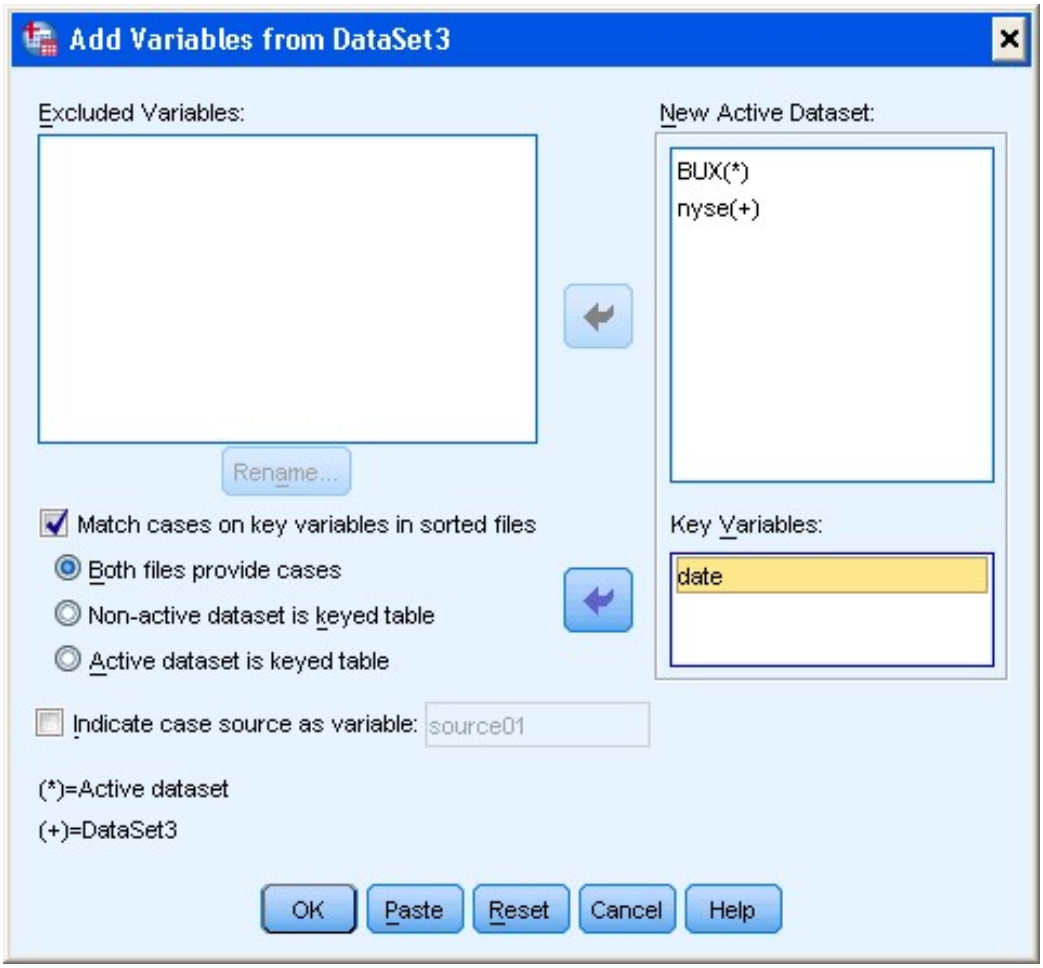


4.4. ábra. BUX és NYSE záró indexei az SPSS programban.

4.1.2. Tőzsdék elemzése

A világ tőzsdéinek a napi záró indexei általában elérhetőek az adott tőzsde weblapjain. Ebben a fejezetben a New York-i Tőzsde (New York Stock Exchange, NYSE), illetve a Budapesti Értéktőzsde (Budapest Stock Exchange, BUX) [3] indexeit fogjuk vizsgálni a 2003 és 2004-es évben. Ezen adatokat az adott értéktőzsdék weblapjairól töltöttük le [3, 13]. Mivel a weblapok általában különböző formátumban szolgáltatják az ilyen jellegű adatokat, így ezek beolvasása az SPSS-be változatos, így adatok beolvasásának gyakorlására kiválóan alkalmas. Javaslatunk, hogy ezeket külön-külön olvassuk be, mivel az azonos napi adatok fognak minket érdekelni, így ezeket a dátum alapján össze kell kapcsolni. A rendelkezésünkre álló napok a különböző országok miatt nem lesznek azonosak, az adott ország munkaszüneti napjainak változatossága miatt. A beolvasott adatokat a 4.4. ábrán láthatjuk, melyen már látható, hogy a január 20. adat a BUX esetén rendelkezésünkre áll, míg NYSE esetén nem, mert a New York-i Tőzsde ezen a napon szünetet tartott (Martin Luther King Day).

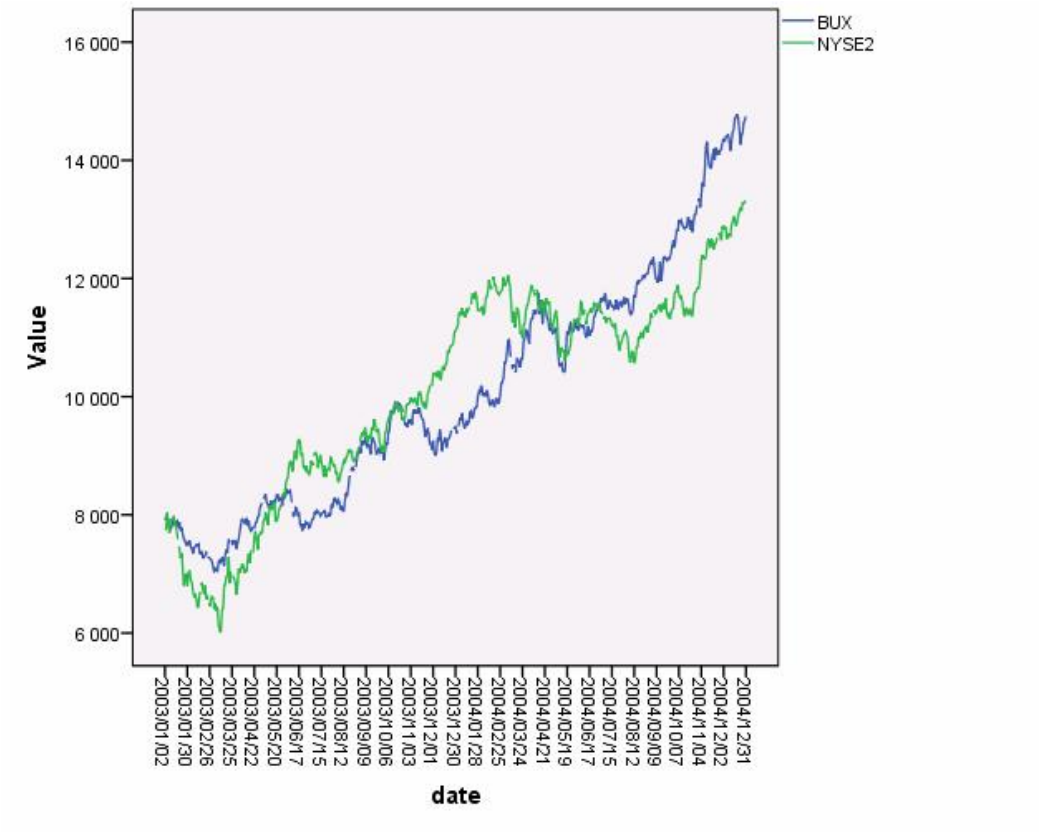
Miután ezzel megvagyunk a két SPSS állományt össze kell kapcsolni a napok mentén. Ehhez vegyük az egyik SPSS állományt, majd válasszuk a Merge Files menüpont alatt az Add Variables parancsot. A felugró ablakban



4.5. ábra. BUX és NYSE záró indexeinek összekapcsolása.

válasszuk ki a hozzácsatolandó állományt. Ezek után a megjelenő párbeszédes ablakban állítsuk be, hogy dátum alapján szeretnénk az összeillesztést. A beállításokat a 4.5. ábrán láthatjuk. Miután ezzel megvagyunk az OK gomb hatására az SPSS összekapcsolja a két állományt. Ezek után elkezdhetjük a tőzsdék egymással való összehasonlítását.

Először nézzük meg a Bivariate paranccsal, hogy tapasztalható-e lineáris összefüggés a két tőzsde záró értékei között. Az eredményül kapott 0,890-es érték nagyon erős kapcsolatra utal. Amennyiben a két tőzsde értékeit grafikonon ábrázoljuk nem igazán vehetjük észre ezt az erős kapcsolatot, mert a nagyságrendek jelentősen eltérnek. Ezért próbáljuk meg az NYSE értékét linárisan transzformálni a BUX értékeinek a közelébe. Ezen transz-



4.6. ábra. BUX és a transzformált NYSE záró indexei.

formációnál a tendenciákat tartsuk meg, de az értékek megváltozhatnak. Ezt megtehetjük egy lineáris regresszió analízissel a Linear parancs segítségével.

A párbeszédes ablakban állítsuk be, hogy a BUX értékeit szeretnék az NYSE értékeinek a függvényében meghatározni. A kapott konstans $-5806,273$, az együttható pedig 2,634, azaz az alábbi képlettel lehet közelíteni a BUX értékét: $2,634 * NYSE - 5806,273$. Ezek után e képlet segítségével számoljunk ki egy új oszlopot (NYSE2), mely valójában az NYSE értékeit tartalmazza, de nagyságrendben a BUX értékeihez hasonlít. Ezek után egy grafikus megjelenítés már jobban fogja mutatni a hasonló tendenciákat a két tőzsdén.

A grafikus Line parancs eredményét a 4.6. ábrán láthatjuk. Ezen már jól követhető, hogy a két tőzsde tényleg sokszor hasonló tendenciát mutat, továbbá láthatóak azok a napok, amikor a tendencia másképp alakult a két tőzsdén.

4.1.3. Részvények közti kapcsolatok feltárása

A Magyarországon forgalomban lévő részvények adatai elérhetők a Budapesti Értéktőzsde weblapján [3]. Innen letölthetők Excel formátumban a tőzsdei árfolyamok, amelyeket a jelenlegi elemzéseknél is használunk. Excel formátumban letölthető adatok elég megszokottak ilyen esetekben.

Ebben a fejezetben végzett elemzésekhez a részvények 2012-es záró értékeit használtuk. Egyetlen részvény vásárlása esetén nagyon ki vagyunk téve annak ingadozásainak. Azaz a nyereségünk vagy veszteségünk nagyságát ezen részvény határozza meg. Több részvény vásárlása esetén ezen kockázat várhatóan csökken. De ha a vásárolt részvények ugyanúgy reagálnak a piac eseményeire, akkor hiába van több részvényünk a kockázat teljesen hasonló lesz az egy részvényes esethez. Így egy portfolió összeállításakor fontos szempont szokott lenni, hogy a tőzsdei hatásoktól függetlenül egy bizonyos nyereségre mindig szert tegyünk. Ezt úgy tehetjük meg, ha olyan részvényeket válogatunk össze, amelyek ellentétesen reagálnak a piac eseményeire, vagy legalább is nem mindig ugyanúgy. Ezen elemzés egyik legegyszerűbb változata a korreláció analízis. Vizsgáljuk meg a letöltött adatokat ilyen szempontból. Ezen adatbázis egy részlete látható a 4.2. táblázatban).

Általában az ilyen típusú állományokat az Open menüpont alatt található egyszerűbb Data paranccsal is be tudjuk olvasni az SPSS-be. Nem szükséges a Open Database esetet használni. Miután ezt megtettük érdemes a hiányzó adatokra egy pillantást vetni és beállítani a változóknál ezen értékeket, hogy a későbbi elemzések során ne szolgáltassanak félrevezető adatokat. A beolvasott adatbázis a 4.7. ábrán látható.

Miután megvagyunk az adatok beolvasásával, illetve a kezdeti beállításokkal, nekiláthatunk az ilyen irányú vizsgálatoknak. Ezen vizsgálathoz tartozó eljárást a Correlate menüpont alatt találjuk. A Bivarite parancsot kiadása után a megjelenő párbeszédes ablakba állítsuk be az összehasonlítandó részvényeket. Ezek után az OK gomb megnyomására az alábbi eredményt kapjuk:

Correlations

		Egis	FHB	MOL	OTP	Richter	Fotex	ANY	MTELEKOM
Egis	Pearson Correlation	1	-,544**	,502**	,633**	,508**	,297*	-,010	-,354**
	Sig. (2-tailed)		,000	,000	,000	,000	,022	,876	,000
	N	245	245	245	245	245	59	236	245
FHB	Pearson Correlation	-,544**	1	,223**	-,127*	,091	,286*	,485**	,773**
	Sig. (2-tailed)	,000		,000	,047	,156	,028	,000	,000
	N	245	245	245	245	245	59	236	245

MOL	Pearson Correlation	,502**	,223**	1	,717**	,391**	,494**	,594**	,457**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,000	,000
	N	245	245	245	245	245	59	236	245
OTP	Pearson Correlation	,633**	-,127*	,717**	1	,452**	,602**	,081	-,074
	Sig. (2-tailed)	,000	,047	,000		,000	,000	,217	,248
	N	245	245	245	245	245	59	236	245
Richter	Pearson Correlation	,508**	,091	,391**	,452**	1	,284*	-,164*	-,151*
	Sig. (2-tailed)	,000	,156	,000	,000		,030	,011	,018
	N	245	245	245	245	245	59	236	245
Fotex	Pearson Correlation	,297*	,286*	,494**	,602**	,284*	1	,211	,380**
	Sig. (2-tailed)	,022	,028	,000	,000	,030		,116	,003
	N	59	59	59	59	59	59	57	59
ANY	Pearson Correlation	-,010	,485**	,594**	,081	-,164*	,211	1	,866**
	Sig. (2-tailed)	,876	,000	,000	,217	,011	,116		,000
	N	236	236	236	236	236	57	236	236
MTELEKOM	Pearson Correlation	-,354**	,773**	,457**	-,074	-,151*	,380**	,866**	1
	Sig. (2-tailed)	,000	,000	,000	,248	,018	,003	,000	
	N	245	245	245	245	245	59	236	245

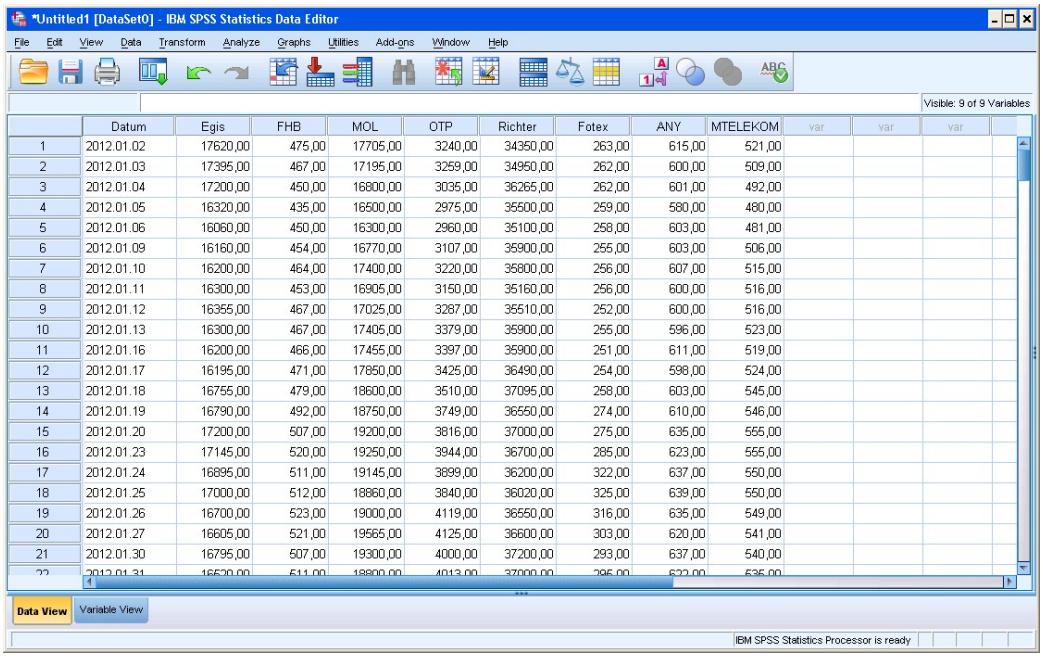
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Az eredményből látható, hogy a legnagyobb korrelációt az MTELEKOM (Magyar Telekom Távközlési Nyilvánosan Működő Részvénytársaság) és az ANY (ANY Biztonsági Nyomda Nyilvánosan Működő Részvénytársaság) részvényei mutatják. Azaz ezen részvények együttes vásárlása a múltbeli adatok alapján kerülendő kockázat csökkentés érdekében. Az árfolyamok alakulása a 4.8. ábrán látható.

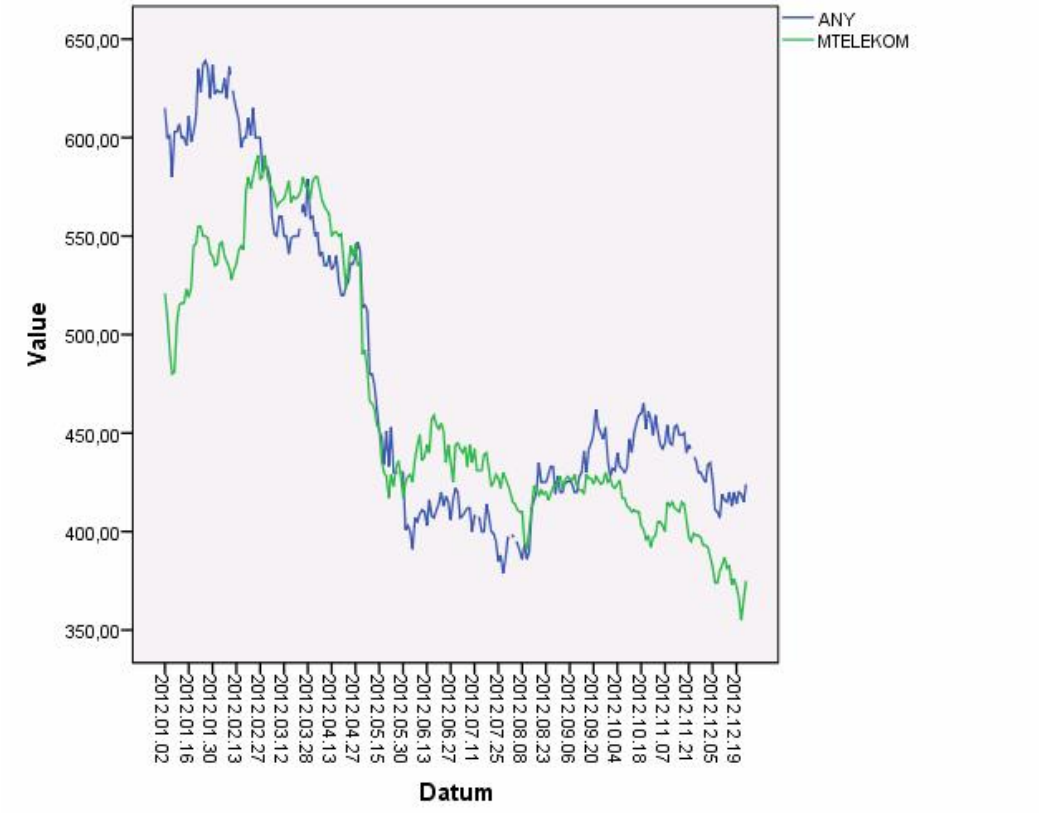
A legerősebb, de nem mérvadó ellentétes mozgást az FHB (FHB Jelzálogbank Nyilvánosan Működő Részvénytársaság) és az EGIS (EGIS Gyógyszergyár Nyilvánosan Működő Részvénytársaság) részvények mutatták. Ezen árfolyamok egymáshoz viszonyított helyzetét a 4.9. ábra mutatja. Látható, hogy az időszak alacsony és magas értékei a két részvény esetén nem feltétlenül azonos napokon voltak (a pontok nem csak jobb fent és bal lent helyezkednek el). Azaz például ezen részvények együttes vásárlása feltételezhetően a kockázat nagyságát csökkenti.

	Egis	FHB	MOL	OTP	Richter	Fotex	ANY	MTELEKOM
2012.01.02	17 620	475	17 705	3 240	34 350	263	615	521
2012.01.03	17 395	467	17 195	3 259	34 950	262	600	509
2012.01.04	17 200	450	16 800	3 035	36 265	262	601	492
2012.01.05	16 320	435	16 500	2 975	35 500	259	580	480
2012.01.06	16 060	450	16 300	2 960	35 100	258	603	481
2012.01.09	16 160	454	16 770	3 107	35 900	255	603	506
2012.01.10	16 200	464	17 400	3 220	35 800	256	607	515
2012.01.11	16 300	453	16 905	3 150	35 160	256	600	516
2012.01.12	16 355	467	17 025	3 287	35 510	252	600	516
2012.01.13	16 300	467	17 405	3 379	35 900	255	596	523

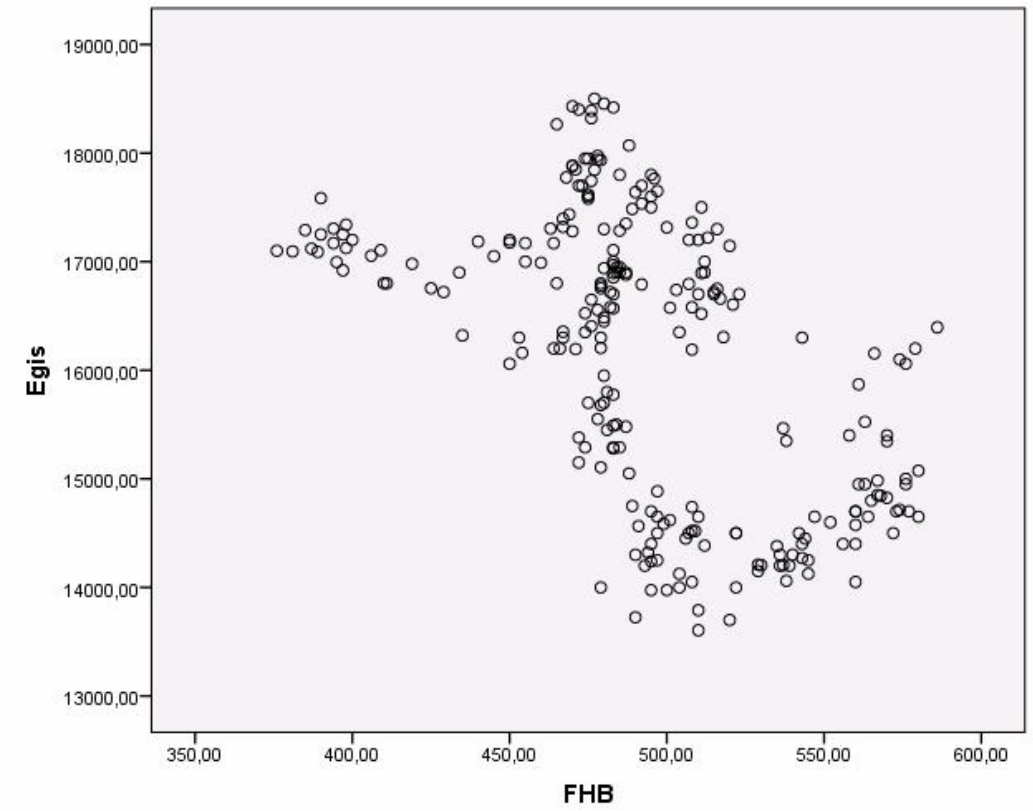
4.2. táblázat. Magyarország néhány főbb részvényeinek záró árfolyama.



4.7. ábra. Magyarország részvényadatai az SPSS-ben.



4.8. ábra. Az ANY és MTELEKOM részvény árfolyama az idő függvényében.



4.9. ábra. Az EGIS és FHB részvény árfolyamai egymáshoz viszonyított értékeinek alakulása.

Irodalomjegyzék

[1] Barna Ildikó, Székelyi Mária: Túlélőkészlet az SPSS-hez, Typotex Budapest, 2002.

[2] Boda Krisztina, Eller József, Nyári Tibor: Biostatisztika oktatási segédlet. Szeged, 1998-99.
`http://www3.szote.u-szeged.hu/dmi/downloads/biostat2011/`

[3] Budapesti Értéktőzsde weblapja, http://www.bet.hu

[4] Fegyverneki Sándor: Valószínűségszámítás és matematikai statisztika, Miskolc, 2007.

[5] Hunyadi László és Vita László: Statisztika I., Aula Kiadó, Budapest, 1991.

[6] Hunyadi László, Mundruczó György és Vita László: Statisztika, Aula Kiadó, Budapest, 2000.

[7] IBM SPSS Statistics 20 Core System User's Guide, 2011

[8] Katona Tamás és Lengyel Imre (szerk.): Statisztikai Ismerettár, szerzők: Csendes Tibor, Katona Tamás, Kovacsicsné Nagy Katalin, Lengyel Imre, Petres Tibor, Pukli Péter és Vavró István, JATEPress, Szeged, 1999.

[9] Ketskeméty László: Valószínűségszámítás és matematikai statisztika, Budapest, 1996.

[10] Ketskeméty László és Izsó Lajos: Az SPSS for Windows programrendszer alapjai, SPSS Partner Bt, Budapest, 1996.

[11] Ketskeméty-Pintér: Bevezetés a matematikai statisztikába, Budapest, 1999.

[12] Magyar statisztikai évkönyv, 2011

[13] New York-i Értéktőzsde weblapja, http://www.nyse.hu

[14] A Klinikai Biostatisztikai Társaság Jubileumi évkönyve, 2001.

[15] Obádovics J. Gyula: Valószínűségszámítás és Matematikai Statisztika, SCOLAR Kiadó, Budapest, 1998.

[16] Prékopa András: Valószínűségelmélet műszaki alkalmazásokkal, Műszaki Könyvkiadó, Budapest, 1962.

[17] Rényi A.: Valószínűségszámítás, Tankönyvkiadó, Budapest, 1966.

[18] SPSS Base 7.5 User Guide. SPSS Inc. 1997.

[19] Tandori Károly: Matematikai statisztika, JATE Szeged, 1974.

[20] Tandori Károly: Valószínűségszámítás, JATE Szeged, 1974.

[21] Tómács Tibor: Matematikai statisztika, Eger, 2012.

Az irodalomjegyzék persze nem teljes, csak néhány fontosabb, illetve a tananyaghoz közvetlenül kapcsolódó könyvet, más egyetemek statisztikai jegyzeteit adunk meg. Továbbá megadtunk pár magyar nyelvű SPSS bemutatását szolgáló szakirodalmat. Magyar nyelven is elég kiterjedt szakirodalom érhető el, a legtöbb könyvtárban kulcsszó alapján ezek megtalálása minden nehézség nélkül megoldható.

Tartalomjegyzék