BEVEZETÉS A SZÁMÍTÓGÉPES STATISZTIKÁBA

Csendes Tibor

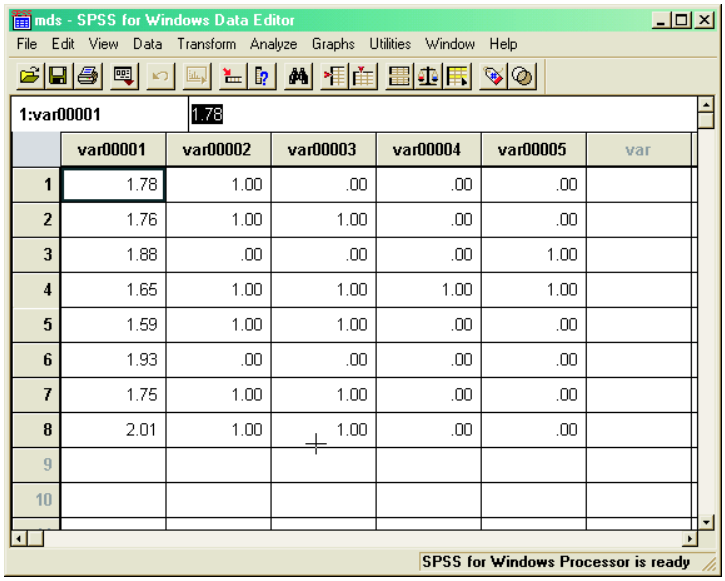
Szeged, 2001.

Lektorálta:

Dr. Boda Krisztina tudományos főmunkatárs, Szegedi Tudományegyetem, Orvosi Informatikai Intézet,

Dr. Petres Tibor egyetemi docens, Szegedi Tudományegyetem, Gazdaságtudományi Kar és

Dr. Viharos László egyetemi docens, Szegedi Tudományegyetem, Bolyai Intézet.



	var00001	var00002	var00003	var00004	var00005	var
1	1.78	1.00	.00	.00	.00	
2	1.76	1.00	1.00	.00	.00	
3	1.88	.00	.00	.00	1.00	
4	1.65	1.00	1.00	1.00	1.00	
5	1.59	1.00	1.00	.00	.00	
6	1.93	.00	.00	.00	.00	
7	1.75	1.00	1.00	.00	.00	
8	2.01	1.00	1.00	.00	.00	
9						
10						

1. ábra: Az SPSS statisztikai program beolvasott adattal.

Előszó

A jelen jegyzet[1] a Szegedi Tudományegyetemen több tárgy: a TTK-s hallgatók Bevezetés az Informatikába II (Statisztikai programcsomagok) speciálkollégiuma, a közgazdász hallgatók Bevezetés az Informatikába II. féléves és a biológus hallgatók A Számítástechnika alkalmazásai című tárgyának közös 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. Aktuális változata egy része, a hozzá kapcsolódó feladatok, gyakorlatok és adataik elérhetők a

`http://www.inf.u-szeged.hu/~csendes/stat.ps.gz`

címen, illetve anonim ftp-vel az

`ftp.jate.u-szeged.hu`

címen a `/pub/math/optimization` könyvtárban.

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. Részben emiatt nem is csak egy statisztikai programot ismertet, hanem az SPSS (1. ábra) részletesebb tárgyalása után más könnyen elérhető statisztikai programokat is bemutat röviden. A kapcsolódó statisztikai ismeretek persze mindegyikre érvényesek.

[1] Minden megjegyzést szívesen látok és előre is köszönök, különösen, ha hibákra hívják fel a figyelmem. Az email címem: `csendes@inf.u-szeged.hu`

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 ismertetem. A teljesen önálló statisztikai munkához ez persze nem elegendő. Ennek ellenére bízom 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.

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 mondok köszönetet korábbi hallgatóimnak és munkatársaimnak a jegyzet létrejöttéhez, illetve a javításához nyújtott segítségükért. Külön köszönet illeti a lektorokat, akik alapos és gyors munkát végeztek, és számos hasznos tanáccsal segítették munkámat.
Várom a további véleményeket és javaslatokat is.

Szeged, 2001. október

a szerző

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
$F(x) = P\{\xi < x\}$	eloszlásfüggvény
$f(x)$	sűrűségfüggvény
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).
$\bar{x}$	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, vagy általános numerikus programcsomagokban is találunk ilyeneket. Jegyzetünkben ilyen statisztikai eljárásokat a StarOffice és az Excel programokkal kapcsolatban ismertetünk röviden (lásd a 4. Fejezetet). Bár ezek a programok kevés statisztikai tesztet, algoritmust bocsátanak rendelkezésre, mégis elterjedtségük miatt és viszonylag könnyű kezelhetőségük révén fontos eszközök. Ide sorolható még a hálózatról letölthető shareware As-Easy-As táblázatkezelő program is (bármely keresőprogrammal megtalálható, csak egy hónapig használható).

Az egyszerűbb statisztikai programok, mint a SigmaStat is, csak egyváltozós statisztikákat képesek kiszámolni, cserében 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

StatGraphics, 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.

Linux operációs rendszerhez is számos programot lehet találni. Egy bő lista van ezekről a `http://chps06.ch.unito.it/linux/A/3` internetes címen (további linkekkel és rövid ismertetéssel minden programról).

Az általános célú numerikus programok közül a NAG programcsomagját, a Maple és a Mathematica szimbolikus algebrarendszereket kell megemlíteni, de ide tartozik a Mathlab is. Az ilyen 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[2]. 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.

[2] 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.

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

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

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. Márészt 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 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).

1.1.4 A minta jellemzői

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

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). *Cenzorált minta* az, amikor az eredeti minta elemeinek csak egy részét használjuk fel a következtetések levonásához.

Az alábbiakban röviden áttekintjük a minta legfontosabb jellemzőit, egyszerű definíciókkal, illetve ahol kell, rövid magyarázattal. A minta egyszerű 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 szám*át 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 mondjuk 0 értékkel beleszámítódik a hiányzó érték 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.

Az *átlag*, vagy *számtani átlag* egy adott mennyiségi, metrikus változó értékei összege osztva az elemszámmal.

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 $\pm\ 2 * \mathrm{SD}$ 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.

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_{25} 25%-os percentil pl. az a szám, aminél a mintaelemek 25%-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. 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 baloldali (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: $F(x) = P(\xi < x)$, ahol P a $\xi < x$ esemény valószínűsége. Az $F(x)$ abszolut folytonos eloszlásfüggvény deriváltja $f(x)$, az ún. *sűrűségfüggvény*.

A sűrűségfüggvénnyel adott változók *várható érték*ét az

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

képlettel (ahol $f(x)$ a megfelelő sűrűségfüggvény), a diszkrét változókét a súlyozott középpel definiáljuk: $\sum_{i=1}^{n} x_i P(\xi = x_i)$. 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.

Poisson-eloszlás

A ξ diszkrét valószínűségi változót λ $(0 < \lambda < \infty)$ paraméterű *Poisson-eloszlás*ának 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 (pl. a vérsejtek száma a mikroszkóp látómezejében, radioaktív anyag adott idő alatt elbomlott atomjainak a száma).

Egyenletes eloszlás

Az egyik leggyakrabban használt 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 $x_1, x_2, \ldots, x_n$.

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. 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 megfigyelés *szabadságfok*a a magyarázó rendszeren belül önkényesen megválasztható értékek száma, speciális esetben az egymástól független összeadandók száma.

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 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). 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. Ez a gyakorlat abból az időből származik, amikor csak táblázatok álltak rendelkezésre. Jelenleg 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 elfogadjuk azt a nullhipotézist, hogy nincs eltérés, de az elkövetett hibáról csak annyit tudunk, hogy nagy mintaelemszám esetén elég kicsi. ha a nullhipotézis nem igaz (másodfajú hiba) .

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 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 végrahajtásának 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.

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 ad. Ezután jön a véletlen minta előállítása, és a próbastatisztika kiszámítása. (Az érintett változó a nullhipotézis fennállása esetén valamely ismert eloszlást követ.)

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 *korrelációs* eljárások két valószínűségi változó közötti öszefüggés szorosságát mérik, ami aztán a predikció minősége mértékeként is használható. A regressziótól eltérően itt nem szükséges az egyik változó, mint eredményváltozó kijelölése. Az r korrelációs együttható egy -1 és 1 köztött 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. Ha a kapott korreláció abszolútértéke nincs közel 1-hez vagy nullához, akkor nem állapíthatunk meg korrelációt.

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.

1.2.4 Egy dolgozat feladatai

1. Adjon példát mindegyik skálatípusra úgy, hogy épp az legyen a legjobb skálatípus, amely az adott mintára érvényes!

2. Határozza meg az átlag, a medián és a módusz értékét a következő adatsorra: 1, 2, 2, 3, 4, 5!

3. Indokolja a hiányzóadat kód fontosságát!

4. Határozza meg, milyen mérési skála felel meg a színeknek, a hőmérsékleti fokoknak, a fizetés összegének és a {nem felelt meg, megfelelt, kiválóan megfelelt} értékelésnek!

5. Mutassunk egy rövid mérési adatsort, amelyre az átlag, a medián és a módusz három különböző érték!

6. Milyen hiányzóadat kódot használna a cipőméretek megadásakor? Jelölje meg az alábbi öt lehetőség egyikét:

   semmilyent 99-et 42-est tetszőlegest -

2. Fejezet

Az SPSS programcsomag

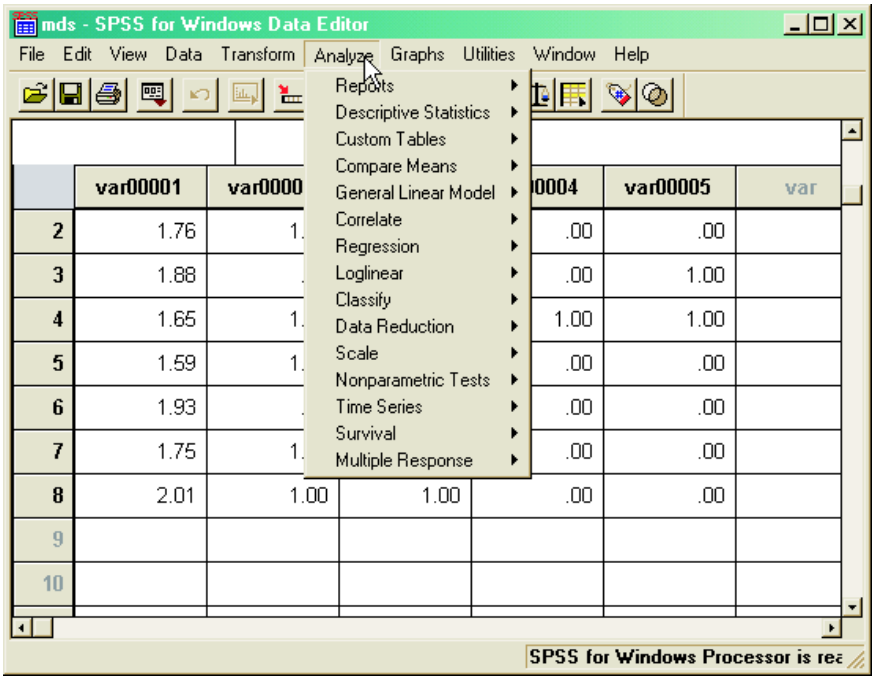
Ahogy a bevezetőben is olvasható volt, az SPSS teljes körű statisztikai eljárásokat kínáló program. A későbbiek megértéséhez sokat segít, ha tudjuk, hogy egyrészt ezt a programot nagy mennyiségű adat kezelésére tervezték, másrészt pedig azt, hogy eredetileg nagy számítógépen, kötegelt (batch) módban 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 nagygépes múlt ellenére a jelenlegi, 9-es változat mind külsőre, mind a legfontosabb szolgáltatások szempontjából teljes értékű grafikus felülettel segíti a munkánkat. Ennek

ellenére a táblázatkezelő programokat ismerők néhány kényelmi funkciót hiányolni fognak majd. Ezeket azonban egyértelműen feledtetik a sokszínű, sokféle módon használható statisztikai eljárások (amelyeknek csak töredéke található meg a táblázatkezelő programokban). Az SPSS 9.0.0 változat minden esetre évekkel ezelőtt készült (1998 decemberi), ami azon is látszik, hogy a görgős egerek használatát nem támogatja.



2. ábra: Az SPSS statisztikai programcsomag Analyze menüsora.

Az SPSS programcsomag (lásd a 2. ábrát) elérhető oktatási célra a teljes felsőoktatás számára (korábbi változatai is). A programot tartalmazó CD-t a megfelelő kulcsszóval együtt a Szegedi Tudományegyetemen a Számítóközpontban lehet megkapni. Az installálás néhány kérdés után automatikusan (és általában problémamentesen) végbemegy. A későbbi használatot megkönnyítendő, érdemes az SPSS ikont az "asztalra" vagy az ún. tálcára tenni, így két (illetve egy) kattintással (a bal egérgombbal) indíthatjuk a programot, nem kell keresnünk a Start / Programok menűsorban. A program hozzáférésének megújítása során előfordulhat, hogy a program korábbi engedélye már lejár. Ilyenkor egy figyelmeztető üzenetet kapunk, de külön-

ben teljesértékűen használható még 1-2 hónapig. A program korábbi változatai is lényegében ugyanazokat az eljárásokat tartalmazzák, inkább a kényelmi és tanácsadó szolgáltatásokkal bővült az utóbbi években.

A jegyzet ezt a programot ismerteti a legrészletesebben, mert ezzel a statisztikai programcsomagok legjellegzetesebb tulajdonságai jól bemutathatók. A tanulmányozás 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 felhasználói ismertetőt [9] (bár az a korábbi, 7.5-ös változatra vonatkozik).

2.1 Az adatok bevitele

A programot az installálás után létrejött SPSS 9.0 for Windows nevű ikonnal indíthatjuk el. A program maga angol nyelvű, de magyar nyelvű operációs rendszerrel, 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 azt kérdezi, hogy mivel szeretnénk kezdeni:

- a programleírás (tutorial) olvasásával,
- adatok begépelésével,
- egy meglevő adatbázis kérdőívvel (run an existing query),
- az adatoknak egy adatbázisból való betöltésével (database capture), vagy
- egy korábbi SPSS adatállomány betöltésével (ilyeneket mutat is a program nekünk egy kis ablakban).

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 [9] 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 arra kell klikkelni az egérrel.

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.

2.1.1 A változók beállításai

A változóink (a táblázatbeli oszlopok) típusát, tulajdonságait a Define variable utasítással lehet megadni, illetve módosítani. Ezt két módon lehet elérni: vagy az érintett változó nevére kell a jobb egérgombbal kattintani, és a menüből ezt az utasítást választani, vagy a Data menüsorból indítani.

Változónév, cimke

A kapott párbeszédes ablakban először is 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 gomb megnyomása után adhatjuk meg. Ez az ún. *cimke* a táblázatkezelő programokhoz hasonlóan magyarázatként

megjelenik egy kis ablakban, ha a kurzort a változó neve fölé visszük. Ez a cimke lehet hosszabb is, tartalmazhat szóközt is, és a későbbi statisztikai eljárások által írt jelentések is használják majd. A cimkék használata különösen akkor fontos, ha több hasonló jelentésű változónk is van, és ezek eltérését nem, vagy csak nehezen tudjuk jelezni a változó nevében.

A mindenkori érvényes beállítások az első keretes részben olvashatók (Variable description): pl. Type, Variable Label stb. Mint más helyen is, itt is elég a jobb egérgombbal egy ismeretlen fogalomra kattintani, és egy rövid magyarázatot kapunk (világossárga hátterű ún. buborékban). Ezeket a beállításokat a következő keretben lehet megváltoztatni, itt van az előbb említett Labels gomb is.

A változók típusa

A *változók típusa* a következők egyike lehet: numeric, comma, dot, scientific notation, date, dollar custom currency és string. 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[3], pl. 123.45 $\rightarrow$ 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.

[3] A tapasztalataim szerint a tizedesjegyek számát itt hibásan alkalmazza.

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

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

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).

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 lehetőség az előző kettő kombinációja: egy tartomány és egy különálló érték.

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 Column Format gombbal tudjuk beállítani. Ennek hatására egy párbeszédes ablakban megadhatjuk az oszlop szélességét, és 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.

2.1.2 Gyakorlat

Az adatbevitel legegyszerűbb formájának gyakorlásához indítsuk el az SPSS programot, és válasszuk az adatok begépelése lehetőséget. Ezek után vigyünk be egy olyan kis táblázatot, amely a következő változókat

tartalmazza az adott tanulócsoportra, vagy néhány ismerősünkre: a név kezdőbetűi, cipőméret, és tanulmányi átlag. Azt fogjuk majd később megvizsgálni, hogy milyen összefüggés van, illetve mutatkozik a cipőméret és a tanulmányi átlag között.

Bár ez egy ártalmatlan kis felmérés, de didaktikai célból érdemes tudatosítani, hogy ebben az esetben sem ragaszkodunk éles adatokhoz abban az értelemben, hogy lehetőleg ne legyenek azonosíthatók az eseteknek megfelelő személyek. Ez most annyit jelent, hogy bár a vizsgálat értelmét az adja, hogy lényegében valós adatokat adunk meg, és használunk, de ezekhez nem kell, hogy a kapcsolódó név, vagy annak kezdőbetűi igaziak legyenek. Ha az adatokat a tanulmányi csoport adja össze, akkor mindenki adhat érvényes adatot az ismeretségi köréből a sajátja helyett.

A kezdőbetűk begépelése előtt állítsuk át az illető változó típusát szövegesre (enélkül be se tudjuk vinni az adatokat). A korábban említett változókra vonatkozó beállításokat mind adjuk meg a tartalomnak megfelelően: tehát a szöveges változónk szélessége legyen arányos a tartalommal, a változónevek rövid formában és ékezet nélkül tükrözzék a tartalmat, a cimkék adjanak bővebb magyarázatot a változókra vonatkozóan, és töltsük ki megfelelően a hiányzóadat kód és skálatípus mezőket is. Az oszlopok formáját hagyhatjuk alapértelmezésben, de azok szélességét az összes adat begépelése után igazítsuk ki azok hosszának megfelelően. Amikor mindezzel elkészültünk, mentsük el a táblázatunkat.

2.2 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. Mindegyik esetben a szokásos kérdéseket teszi fel a párbeszédes ablak a létrejövő adatállomány nevéről, az érintett könyvtárról és a mentett file típusáról (egyben kiterjesztéséről).

Alaphelyzetben érdemes az SPSS saját fileformátumát használni (*.sav) és a saját floppynkra menteni (az egyetemi gépeken nem biztos, hogy a legközelebbi órán is meg lesz még az adatállomány). A saját hajlékony lemez használata a statisztikai programok tanulásához azért is hasznos, mert így a gyakorlatok közti időben, más gépeken is tudunk ismerkedni a programmal korábbi adatainkon.

Az SPSS a saját adatformátumán kívül a következő file-típusokban tud adatot menteni: a 7.0 változatú SPSS formátumban (.sav), kétféle, fix formátumú és tabulátorokkal elválasztott ASCII szöveges file-ként (.dat) és többféle adatbázis adatállományként (a fontosabbak több verziójával kompatibilis módon): pl. Excel, dBASE, 1-2-3. Adatbázis formátumban való mentés esetén a párbeszédes ablak felajánlja azt a lehetőséget, hogy a változóinkkal együtt azok nevét is mentsük. Ha szövegszerkesztővel szeretnénk keresni az adatunkban, vagy annak az átírását így kell megcsinálnunk, akkor föltétlen szöveges file-ként érdemes mentenünk, mert ekkor nem kell vezérlőjelekkel bajlódnunk.

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.

2.3 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 Megnyitás lehetőség választása után[4] egy ún. varázsló (Text Import Wizard) segítségével adhatjuk meg részletesen azt a formátumot, amelynek megfelelően a szöveges állományból az adatainkat be kell olvasnunk. Ez a lehetőség a program korábbi változataiból hiányzott (más statisztikai programok ilyen segédprogramot korábban is tartalmaztak, mint pl. a SigmaStat).

[4]Figyelem, a magyar nyelvű operációs rendszer egyes alapszavakat magyarul ad az SPSS számára, bár maga a program angol nyelvű. Ez a lehetőség persze csak megfelelő beállítások mellett érvényes, különben a megadott más nyelven kapjuk ezeket a kulcsszavakat. Mivel általában a magyar nyelv beállítása gyakori, ezért a továbbiakban külön magyarázat nélkül hagyjuk az alapszavak használatát.

A varázsló 6 párbeszédes lapon kérdez ki bennünket, e lapok között a Tovább vagy a Vissza gombokkal tudunk mozogni, az utolsót (miután minden lényeges adatot megadtunk) a Befejezés 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) acélból, 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 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: 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 file-t, hogy helyes-e a megadott elválasztás. Új elválasztó pontot az egérrel a megfelelő helyre való klikkelésel

lehet adni, a fölöslegest pedig a bal egérgombbal megfogva ki lehet vinni az adott ablakból.

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 Befejezés gomb megnyomása után az adatfile-ból a változók tartalma beolvasódik, és egy output file 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 file-ok pl. standard kimenetekre íródtak. Az output ablakok tartalmát szerkeszthetjük, elmenthetjük, vagy ezeket az ablakokat a teljes feldolgozás végéig egyszerűen a háttérbe tehetjük. A Befejezés 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.

2.3.1 Gyakorlat

Írjunk rövid, táblázatos formájú adathalmazt valamely szövegszerkesztővel, és mentsük el egyszerű karakteres formában, bináris vezérlőjelek nélkül (tehát nem pl. a Word saját .doc formátumában). Erre a célra bármely egyszerű szövegszerkesztő is megfelel. Fontos, hogy az eredmény állományban legfeljebb csak sorvége jelek legyenek, más formázási utasítások ne.

Mindkét beolvasási formára, a fix szélességű adatok és az elválasztó jelek esetére is generáljunk input állományt, és olvassuk be azokat az előző szakasznak megfelelően. A beolvasás során kisebb problémák merülhetnek fel bonyolultabb esetekben, ezért érdemes több ilyen beolvasást is végrehajtani. A beolvasott adatokat az előző szakaszban írtaknak megfelelően lássuk el minden hozzájuk tartozó beállítással a korábbi ajánlások szerint. Tanulságos egy szövegszerkesztőbeli (Word, LaTex) táblázatból kiindulva annak adatait beolvasni.

Külön vizsgáljuk meg a beolvasandó esetek lehetséges beállításait:

- olvassunk be adatot úgy, hogy nem az első sorban van az első eset, hanem pl. a másodiktól,

- amikor nem minden sor egy eset, hanem egy megadott változóértékkel specifikáljuk azt,

- generáljunk egy olyan adatbevitelt is, amikor csak az első, mondjuk 10 esetet olvassuk be,

- és olyan példát is mutassunk, amikor az esetek valamely véletlenül választott 20 %-át kell beolvasni az adatállományunkból.

Ellenőrizzük, hogy minden adat helyesen került-e az új állományba, azt hogy minden változó tartalmát a megfelelő formátumban, szélességgel jelenítjük-e meg, és hogy a hiányzóadat kód, skálatípus megfelel-e a tartalomnak.

2.4 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 Database capture 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 Tovább, illetve a Vissza gombokkal lehet mozogni, és a Befejezés 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, FoxPro és Access. Ugyanezen a lapon lehet ezek körét bővíteni is.

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 jobboldali ablakba. 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 kimarad.

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. A varázsló mindenképp ad (alapértelmezés) neveket a változóknak. Ilyen névek lehetnek pl. az Excel táblázat első sorának elemei (az első 8 karakter). 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 Befejezés 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).

2.4.1 Gyakorlat

Vigyünk be az Excel táblázatkezelő programba egy kisméretű táblázatot, amelynek sorai eseteket, oszlopai pedig változókat reprezentálnak. Az oszlopoknak legyen címe. Erre vonatkozóan mindkét lehetőséget próbáljuk ki: olvassuk be a címeket mint változóneveket, illetve ugorjuk át ezt a sort az adatbevitel során. A Database capture parancsnak igyekezzünk minden beállítását kipróbálni.

A sikeresen használt beolvasási beállításokat mentsük ki, és használjuk fel egy kicsit megváltoztatott táblázat beolvasására. A beolvasott adatokat

ki tudjuk menteni adatbázis formátumban is. Tegyük ezt meg, majd olvassuk be ismét az SPSS-be! Vessük össze a táblázatkezelő programból való adatbevitel CPU időszükségletét a szöveges file-ból való beolvasáéval.

2.5 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 a jelen jegyzet keretében nem is térhetünk ki minden előforduló utasításra, illetve beállításra. A File menüsor kapcsán ezen az alapon nem tárgyaljuk a Display Data Info, Apply Data Dictionary parancsokat.

A Print utasításról se sokat kell 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 file 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.

2.5.1 Új file

A New 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.

A New utasítás nem csak új munkalapot tud megnyitni, hanem beolvasási szintaxis-fájlt, output állományokat és egy teendőket leíró ún. script állományt is. Ez utóbbiak használatával itt nem foglalkozunk, de a program súgója, leírása segítségével ezek használata nem nehéz.

2.5.2 Korábban létrehozott állomány megnyitása

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

Ez az utasítás nem olyan sokoldalú, mint a Database capture varázsló: újabb programváltozatok állományait nem tudja kezelni. Így például az Excel 5.0 és későbbi változatú táblázatkezelő által írt file már nem olvasható be ilyen úton. A hibajelzés is az adatbázis varázslót javasolja.

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 file mentése pedig a Ctrl-S gombokkal (a Ctrl gombot lenyomva és nyomvatartva egyszer kell megnyomni a kötőjel után álló betűvel jelzett gombot). Ezt az utasításkiadási módot a jelen kurzus során nemigen fogjuk használni, és az SPSS-el való munka első fázisában sem ajánlom: a gyorsító billentyűket akkor érdemes bevetni, ha a hatásuk minden részletével tisztában vagyuk, és rutinszerűen alkalmazzuk standard helyzetekben, sok, egymáshoz hasonló adaton.

2.5.3 Adataink mentése

A beolvasott vagy begépelt, és kijavított adatunkat a későbbi feldolgozás céljából érdemes az SPSS saját formátumában elmenteni. Erre két utasítás szolgál: a Save és a Save as. 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 file 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 (esetleg nem is változtatott) 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).

2.6 A szerkesztési és a nézet menüsor

2.6.1 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 (a program ugyan nagy betűket ad itt meg, de csak kicsit használ).

Ezeket két további parancs egészíti ki. A Clear csak kitörli a megadott cellákat a vágólapra másolás nélkül.[5] 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) keresési utasítás a szokásos módon jelsorozatot (számokat vagy szöveget) keres a kijelölt tartományban. Érdekes módon a keresési tartomány kijelölése nélkül nem működik. Azt ikonok között (a szokásoknak megint megfelelően) egy távcső jelzi.

2.6.2 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 teljesértékűen használható a beállítások módosítása nélkül is.

[5] Érdekes módon a törlés (Clear) utasítás ugyan működik, a kijelölt cella tartalmát kitörli, de a megadott gyorsító billentyű, a Del önmagában nem hatásos: utána meg kell nyomni az Enter-t is, vagy a kurzor mozgató nyilakkal egy másik cellára kell ugrani.

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.

Az általános beállítások közül kiemelendő a munkaterület mérete (Special Workspace Memory Limit) és az ideiglenes állományok könyvtára (a Temporary Directory). Említésre méltó az Interactive munkalapról a Data saved with chart és a Print resolution. Az előbbivel a grafikonokkal együtt elmentett adatok mennyiségét, az utóbbival a nyomtatás feloldását adhatjuk meg. A Data munkalapról fontos a Transformation and Merge Options és a Set Century Range for 2-Digit Years. Az előbbivel azt mondhatjuk meg, hogy a kért átalakításokat azonnal hajtsa-e végre az SPSS, vagy csak a használatuk előtt. Az utóbbival pedig a rövid évszámok értelmezési tartományát rögzíthetjük. A 2000. év elmúltával világos, hogy a kapcsolódó probléma már nem jelent veszélyt.

A többi beállítással először akkor érdemes foglalkozni, amikor valamelyik formátummal vagy működésmóddal elégedetlenek vagyunk.

2.6.3 A nézet menüsor

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

- az állapot sor (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 a leggyakrabban használt utasításokat teszi könnyebben elérhetővé a táblázatunk feletti ikonokkal.

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

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

- A Value Labels lehetőséget nem tárgyaljuk.

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 jobb lesz.

2.6.4 Gyakorlat

Egy minta adattáblázaton gyakoroljuk a szerkesztési menüsor utasításainak használatát, különös tekintettel a Clear és az Undo parancsokra. Keressünk meg egy az adatunkban előforduló számot a Find utasítással (ne felejtsük el a keresési tartomány kijelölését).

Nézzük át az Options... beállítási lehetőségeit, és a Help / Topics menüpont segítségével tisztázzuk az érdekesebbek jelentését, hatását. A beállítások megváltoztatása után vizsgáljuk meg a hatást. A gyakorlat végén állítsuk vissza a végrehajtott változtatásokat.

A View menüsorral kapcsolatban nézzük meg, hogy az eszköztár és az állapotsor kihagyása után mennyivel több sornyi adat látszik. A betűtípusok változatait a változóneveken láthatjuk. Szintén próbáljuk ki a táblázatunkat az elválasztó vonalak nélkül is. A gyakorlat végén most is állítsuk vissza a végrehajtott változtatásokat.

2.7 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 Az egyes változók tulajdonságait lehet beállítani ezzel a paranccsal. Korábban az illető változóra jobb egérgombbal való kattintással tudtuk kiváltani, illetve a beolvasás során kellett ezeket az adatokat megadnunk. Mivel ezt a parancsot korábban részletesen tárgyaltuk, itt most nem térünk ki rá. Az utasítás gyorsító billentyű kombinációja: Alt-d d, tehát az Alt gombot nyomjuk le és tartsuk nyomva, majd nyomjuk meg egyszer a d gombot. Erre megjelenik a

data menüsor, amiből egy (Alt nélküli) d betű hatására végrehajtódik az utasítás.

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: Alt-d e.

Több furcsaságot is láthatunk: az egyik mindjárt, hogy a magyar nyelvi környezetben talán leggyakoribb év – hónap – nap dátumforma 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. Végül az ígéretes Custom beállítás (amely a neve alapján tetszőleges forma megadását tenné lehetővé) hatástalan.

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).

Templates... Ez az utasítás a változóink definiálását könnyíti meg, amennyiben tipikus beállításokat rögzíthetünk itt, ezeknek neveket adhatunk, és elmentésük után új változókra alkalmazhatjuk őket. Vannak előre beállított minták is. Ezekben példát láthatunk az értékcimkék használatára (value label). A parancs jelentősége inkább nagy adathalmaz esetén nyilvánvaló. De már néhány változó esetén is gyorsítja a munkánkat, ha nem kell mindent újra begépelnünk (hiányzóadat kódok, cimke, típus stb.). Az utasítás gyorsító billentyű kombinációja: Alt-d t.

Insert variable Egy új változót szúr be a meglévők közé, az aktuális cella változója elé. Az új változó az alapértelmezésnek megfelelő beállításokkal rendelkezik majd. Ezeken nyilván módosítani kell ezután a változó tartalmának megfelelően. A változóban természetesen nem lesznek adatok. Az utasítás gyorsító billentyű kombinációja: Alt-d v.

Insert Case 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. Az utasítás gyorsító billentyű kombinációja: Alt-d i.

Go to Case Egy megadott esetre ugrik a kurzor a parancs végrehajtása után. Az eset sorszámát egy párbeszédes ablakban kell megadni. Az utasítás gyorsító billentyű kombinációja: Alt-d s.

A menüsor parancsainak következő csoportja az adatállományunk alakításával foglalkozik.

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

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 baloldali kis ablakban felsorolt változók közül az érintetteket a bal egérgombbal egyenként megjelöljük, majd a két kis ablak közti nyíllal átvisszük a jobboldali ablakba. Ezt az átvitelt vontatással (drag and drop) nem lehet megoldani. Az átmozgatás előtt be kell állítani az illető változóra vonatkozó rendezés irányát (növekvő vagy csökkenő: ascending – descending).

A 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. A Sort parancsot nem lehet visszavonni.

Transpose A Transpose parancs, ahogy a neve is mondja, a táblázatunkat transponálja, 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 ekkor 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 transponálása nélkül is meg tudják oldani a feladatot — paramétereik megfelelő beállításával.

A kapott párbeszédes ablakban megadhatunk egy olyan változót is (Name Variable), amelyből 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-d n.

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

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 utolsó lépés gyorsító billentyűje a v). A másik file 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 baloldali 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 jobboldali 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 file 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. Az utolsó lépés gyorsító billentyűje a c.

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

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-d a.

Ilyen átalakításra van szükségünk például akkor, 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 cimkéjét 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 eredmény állomány beállítás szerint pótolhatja a jelen aktuális munkalap tartalmát, vagy kerülhet egy új fájlba is.

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

Orthogonal Design Ezt az utasítást nem tárgyaljuk. A parancs gyorsító billentyű kombinációja: Alt-d h.

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.

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

A kapott párbeszédes ablak alapértelmezésben minden esetet megtart (valószínűleg azért, nehogy véletlenül az OK gomb megnyomásával kitöröljünk eseteket — bár legtöbbször az állományunk megvan elmentve is). 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, csak 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 a szűrés után megmaradtak (Unselected Cases Are Filtered). A másik lehetőség az, amikor a ki nem választott esetek törlését kérjük (Unselected Cases Are Deleted).

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 baloldalról 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 abszolutérték függvényt jelöli, aminek egy numerikus kifejezés lehet az argumentuma[6].

A függvény beszúrása után egy kérdőjelet látunk az argumentum helyén. Ha ekkor egy változót hozunk be a baloldali 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 rádiógombja 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. Azért csak "kb.", mert az eljárás véletlen jellege csak közelítőleg biztosítja a megadott arányt. Az ilyen esetkiválasztás olyankor hasznos, ha egy nagyon nagy mintából kell reprezentatív részmintát megadnunk.
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.

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

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 nyillal) 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 paranccsal tetszőleges ilyen változó létrehozható. Ilyen szűrőváltozót a többi esetkiválasztási mód is létrehoz.

Esetek súlyozása 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-d w 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.

A jobboldali gombok közül a Paste is említést érdemel: ezzel tudjuk itt az aktuális beállításra vonatkozó szintaxis állományt (syntax file, belső programozható leírás, ami az SPSS program működését meghatározza) megnézni, illetve a meglevőhöz hozzáadni.

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.

2.7.1 Gyakorlat

A menüsor bőven nyújt gyakorolnivalót. A feladatokat a tárgyalt utasítások sorrendjében adjuk meg. A Define Variable paranccsal nem foglalkozunk ismét, ennek használatát gyakoroltuk már korábban. Ugyanígy kihagyjuk a nem tárgyalt utolsó utasítást sem.

A változó (Insert Variable), illetve eset beszúrásra (Insert Case) és egy adott esetre való ugrásra (Go to case) külön gyakorlatot nem írunk elő. Ezek nyilvánvaló módon könnyen kipróbálhatók, és használatuk semmi különös meglepetést, veszélyt nem jelent.

Dátum definiálása

Egy meglevő adatállományunkhoz generáljunk dátumot. Először a talán legegyszerűbb módon, hogy csak napokat engedünk meg. Ekkor két új változót kapunk (day_ és date_), mindkettő neve aláhúzás jelre végződik. Az ilyen változókat általában a program generálja, és ezek típusának megállapításakor a Define Variable parancs figyelmeztet is bennünket,

hogy a beállítások megváltoztatása veszélyezteti a helyes működést. A két változó tartalma különben megegyezik, attól eltekintve, hogy az első numerikus, a második pedig szöveges változó. Vigyázat, a date_ változó típusa *nem* dátum formátumú!

Hasonló módon generáljunk dátumot a hét, munkanap formának és a munkanap, munkaóra formának megfelelően. Figyeljük meg, hogy a kapott változóértékek periódikusak, és a periódus megfelel a beállításban megadott időszakoknak.

Végezetül adjunk meg egy olyan példát, amelyben a dátumok előállításának éppen ez a módja a kényelmes és hatékony. Vigyünk be a megfelelő adatállományból néhány sornyit, és illusztráljuk az eredményt!

Templátok

Először használjuk a beépített templátokat, és vizsgáljuk meg, hogy korábbi adataink megjelenésén mit változtatnak. Bár a beépített templátokat könnyű (talán túlságosan is) törölni, ezt ne próbáljuk ki. Ezután írjunk önállóan új templátokat, mentsük el őket (nyilván új néven), és alkalmazzuk ezeket. Adjunk meg egy olyan példát, amelyben templátokkal legalább tízszer gyorsabban tudjuk az új adatunk változóinak leíró paramétereit megadni, mint azok nélkül.

Az esetek rendezése

Az esetek rendezésének gyakorlásához hozzunk létre egy olyan adatállományt, amelyben az abc-sorrendet nem követő rendben nevek szerepelnek egy változóban, egy másikban pedig cipőméretek. Ezek után egy új változóban sorszámozzuk meg az eseteinket (erre alkalmas az előbb említett dátum generálás is). Nem kell sok esetet bevinni, a lényeg, hogy legyen olyan cipőméret, amely többször is előfordul.

Első lépésben rendezzük az állományunkat a nevek alapján szokásos abc-sorrendbe (ez tehát növekvő sorrend lesz). Figyeljük meg, hogy az eredeti esetsorszámok hogyan változnak meg. (Ez utóbbi változó nélkül az eredeti sorrendet nem lehet visszaállítani ezzel a paranccsal). Ezután rendezzük az eseteket a cipőméret szerint csökkenő sorrendbe — és mint másodlagos rendezési szempont: a nevek szerint abc-sorrendbe. Mutassuk olyan esetpárt, amely illusztrálja, hogy mindkét rendezési szempont hatásos volt.

Keressünk olyan gyakorlati példákat, alkalmazási lehetőségeket, amelyekben a két vagy több rendezési változó elengedhetetlen.

Transzponálás

Egy kicsi és egyszerű adatállományt gépeljünk be, és ezen gyakoroljuk a transzponálást. Az első gyakorlat legyen olyan, amelyben minden adat a táblázatban marad, csak (többségük) új helyre kerül. Ezután mutassunk egy olyan feladatot, amelyben egy korábbi változó tartalma az új táblázatban alkalmas az új változók neveinek.

Figyeljük meg a kapott új változót az esetcimkékkel. Keressünk olyan gyakorlati feladatot, ahol valamely statisztikai eljárást mind az esetekre, mind a változókra el kell végezni.

Adatállományok összeillesztése

Hozzunk létre két teljesen eltérő adatállományt (hogy a különbségeket az eredmény táblázatban felismerjük). Ezek után olvassuk be az elsőt az SPSS programba, és kérjük a Data / Merge Files / Add Variables parancsot, majd töltsük be a második állomány első változóját az eredetiek mögé.

Ezek után kérjük a két állomány egyesítését az esetek hozzáadása alapján. A legegyszerűbb az az eset, amikor a két állomány azonos változókat tartalmaz, csak az esetek térnek el. Tipikusan ez a helyzet, ha egy felmérés űrlapja és adatszerkezete is jól előkészített, és a felmérésben résztvevő biztosok adatait kell összevonni.

Végül alakítsuk addig az alapul vett két állományt, hogy a kulcsváltozó szerinti esetegyeztetés működjön. Nem egyszerű feladat: a program a beállítások kései fázisában igényel olyan dolgokat, amiket korábban kellett volna az adatállományokban kialakítani (azonos, illetve eltérő változónevek, a kulcsváltozó értékeinek eltérőek kell lenniük, a kulcsváltozónak mindkét állományban benne kell lennie, ...).

Aggregálás

Hozzunk létre egy olyan állományt, amely egy dolgozatírás eredményeit tartalmazza: neveket, A – B csoportba tartozást (ábrázoljuk ezt számmal, pl. 0 – 1), és a kapott pontszámot. Ezek után egy új változót képezzünk, amiben a fiúk 0, a lányok 1 értéket kapnak.

Az aggregálás segíségével adjuk meg a kapott pontok átlagát egyrészt az A, illetve a B csoportra, másrészt a nemek szerint is. Egy következő aggregálási lépésben határozzuk meg a legkisebb és a legnagyobb pontszámot mindegyik említett csoportra.

Adjuk meg aggregálással azt, hogy hányan értek el egy-egy pontszámot (minden előforduló pontszámra)!

Esetek kiválasztása

Ehhez a gyakorlathoz használjuk az előző pontban már létrehozott adatállományt. Az esetkiválasztási paranccsal adjunk meg egy-egy példát mindegyik lehetőségre. Tehát válasszuk ki azokat az eseteket, amelyekre a kapott pontok száma 3 és 5 közé esik, aztán válasszuk ki a fiúkra vonatkozó eseteket.

Kérjük azokat az eseteket véletlenszerűen, amelyek az összes esetszámnak kb. 10%-át adják. Generáltassunk esetkiválasztást addig, amíg pontosan teljesül a kért kiválasztás (nem minden százalék értékre teljesülhet).

Kérjük az esetek közül az éppen adott sorrendnek megfelelően az 5. – 9. esetek kiválasztását. Adjunk meg a létező esetszámoknál tágabb korlátokat is.

Válasszuk ki előbb a lányok, majd a fiúk adatait szűrőváltozóval. Próbáljuk ki, hogy a korábban létrejött szűrőváltozó egyes értékeit módosíthatjuk közvetlenül az érintett cellába való beírással is. A feldolgozásra kijelölt esetek köre ekkor újabb Select Cases parancs nélkül is megváltozik.

Esetek súlyozása

Először is, az aggregálási parancs is tud súlyozáshoz változót generálni. Ehhez a Save Number of Cases in Variable opciót ki kell választani a párbeszédes ablak alján, és egy változónevet megadni. Ez a változó alkalmas lesz súlyozásra.

A korábbi adatállományunkban súlyozzuk az eseteinket egy új, súly nevű változóval, amit előzőleg feltöltünk pozitív számokkal. Figyeljük meg a súlyozás jelzését a státuszsorban, és próbáljuk ki a súlyozás hatását az Analyse / Descriptive Statistics / Descriptives... paranccsal.

2.8 Az adatok módosítása

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

Ebből a sorból kilóg a Random Number Seed (Alt-t s). Ez a program által használt pszeudovéletlenszám generátor számára adja meg az induló

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 Random Seed (véletlen magpont) lehetőséget.

2.8.1 Új változó kiszámítása

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

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 az alatta levő ablakban felsorolt beépített függvényeket és a bal alsó ablakban levő változóinkat. 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. Különben minden billentyű jelentését megtudhatjuk, ha a jobb egérgombbal ráklikkelünk.

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.

2.8.2 Előfordulások megszámlálása

A Count... parancs arra való, hogy adott változóban egyes értékek előfordulásainak számát meghatározzuk. Olyankor használatos, ha az így kapott értékekkel még számolni szeretnénk a továbbiakban, mert különben több olyan statisztikai eljárás létezik, amely ilyen információt szolgáltat az

output ablakban (tehát szöveges file-ban). A gyorsító billentyűkombináció az Alt-t o.

A parancs paraméterezése az előzőekhez hasonlóan történik. A megjelenő párbeszédes ablakban meg kell adnunk az új változó nevét és esetleg cimkéjét. Ezután azt a változót vagy változókat áttesszük a baloldali listából a jobboldali 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 egyes (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. A legegyszerűbb az az eset, amely egyben az alapbeállítás is: amikor a megszámolandó értékeket egyesével megadhatjuk. 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).

2.8.3 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 billentyűkombináció az Alt-t r. Két további választási lehetőségünk van,

amelyek menüpontként jelennek meg: az eredményt vagy ugyanabba a változóba kérjük (Into Same Variables) vagy egy újba (Into Different Variables). Ez utóbbi választásokat is lehet gyorsítógombbal kérni, az előző billentyűk után s, illetve d kell. Éles gyakorlati feladatok megoldása során nem szokás az eredeti adatot felülírni, inkább a változók, esetek számát szaporítják.

A kapott párbeszédes ablakban először az érintett változót vagy változókat át kell tenni a jobboldali ablakba. A lényeges beállításokat a régi és új értékek megadásával kell megtennünk (Old and New Values gomb, vagy az o gyorsítógomb). A 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. 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.

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 szóbajövő régi értékre egy-egy bejegyzést 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).

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.

2.8.4 Változók kategorizálása

A Categorize Variables... 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-t z.

A kapott párbeszédes ablak is egyszerű: csak kétféle beállítást tehetünk. Először azon változókat kell kijelölnünk, amelyekre a kategorizálást kérjük (át kell őket tenni a jobboldali ablakba, mint szokásosan). Ezután még azt kell megmondanunk, hogy hány kategória kialakítását kérjük. Az alapértelmezés 4. Ezeket a kategóriákat a program a kvantilis értékek alapján határozza meg, és körülbelül egyenlő nagyságú csoportokat alakít ki a változók értékeiből.

A program ismét szöveges változók tartalmát is tudja kezelni, ekkor az abc-sorrend alapján történik a csoportosítás. Az új változó neve a régiből képződik: egy n betű kerül a régi név elé. A különböző csoportok kialakításához szükséges az is, hogy az illető változó értékei eltérőek legyenek. Ha minden érték megegyezik, akkor a beállított katagóriaszámtól függetlenül csak egy csoport fog képződni.

2.8.5 Az esetek rangsorolása

A Rank cases... parancs, ahogy a neve is mondja, 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-t k.

A párbeszédes ablakban először a figyelembe vett változókat kell megadni. A 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. Ennek tartalmát lehet szerkeszteni, és kitörlése nem befolyásolja a 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, de mindegyik beállítási lehetőségről informálódhatunk, ha a jobb egérgombal a szövegre kattintunk.

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).

2.8.6 Automatikus újrakódolás

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

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.

2.8.7 Idősorok létrehozása

A Create Time Series... parancs idősor jellegű változóból más típusú idősort hoz létre. A kiindulási változó leírásában numerikus típusú kell hogy legyen, és a változó tartalma olyan, amely egymásra következő, ezt a sorrendet tükröző. Ilyen például a részvények árfolyama napokon át, vagy a növények magasságadatai időrendben stb. A gyorsító billentyűkombináció az Alt-t m.

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 jobboldali 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).

2.8.8 A hiányzóadat kódok pótlása

A Replacing 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-t v.

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.

2.8.9 Gyakorlat

Új változó kiszámítása

Nyissunk egy új munkalapot, és vigyünk be adatokat egy olyan változóba, amelyik cipőméreteket tartalmaz az ún. európai mérték szerint (34 – 45 pl.). Ezután keressük meg azt a képletet, és hozzuk létre azt az új változót, amely ezeknek az értékeknek az angol egységben számolt értékeit tartalmazza (pl. a 42-es méret helyett 8-as). Az új számok legyenek kerekítve úgy, hogy csak egész értékek és egész érték és egy fél alakú számok fordulhassanak elő, ahogy a cipőméreteknél is.

Egy újabb változóban terheljük meg a kiszámított új változónkat véletlen zajjal (normális eloszlás szerint, 1 szórással, de ismét kerekítve az előbb megadott formában). Mutassuk ki a zajjal terhelt és a terheletlen változók eltérését.

Keressünk egy olyan, a gyakorlatban is előforduló példát, amelynek megoldásához szükség van a Compute... parancs feltételes értékadására. Alkalmazzuk olyan feltétellel, amelyik egyáltalán nem ad értéket, és olyan feltétellel is, amely minden esetre ad értéket.

Előfordulások megszámlálása

Első feladatként határozzuk meg egy adatállományunk valamely változóinak adott értékintervallumba eső értékei számát esetenként. Ezek után figyeljük meg, hogy az egyoldali tartományok figyelembevétele esetén a hiányzóadat kódokat is beleszámítja a program. Adjunk példát arra, hogy az SPSS hogyan tudja meghatározni a program által generált hiányzóadat kódok számát. Végül dolgozzunk ki egy olyan feladatot, amelyben az egyes értékek előfordulását feltételesen kell összeszámolni.

Újrakódolás

Első gyakorlatként vigyünk be egy kis táblázatot nevekkel, és a hozzájuk tartozó személyi számok első jegyét (ami a nemet jelöli). Ennek értéke lehet 3 vagy 4 is (ha 1900 előtt született az illető, vagy 5 és 6, ha külföldi). Tegyük fel, hogy egy későbbi statisztikai feldolgozáshoz ebből a nemeket egyértelműen kódoló változót kell létrehoznunk — oldjuk ezt meg újrakódolással.

Vizsgáljuk meg, hogy mi történik, ha újrakódolással egy x értéket y-ra, és az y-t x-re változtatjuk (Ezek különböző számok)! Mi történne, ha ezt a két értéket a szövegszerkesztőkben szokásos csere utasítással két lépésben oldanánk meg? Oldjuk meg ezt a kis feladatot az új változó kiszámítása paranccsal.

Az újrakódolással cseréljünk ki szöveges információt. Figyeljük meg, hogy az új kódok felsorolásában hogyan adja meg ezek értékét az SPSS. Mutassunk példát arra, amikor a feltételes újrakódolás a hatékony megoldás!

Egy egyszerű példán próbáljuk ki mindkét lehetőséget: az azonos változóba és egy újba való kódolást is.

Változók kategorizálása

Először vigyünk be egy új változót, amelynek minden értéke megegyezik néhány kitöltött esetre, és adjunk meg néhány kategóriát (vagy hagyjuk meg az alapértelmezést). Ezek után fokozatosan változtassuk meg a változó értékeit, és minden esetben kérjünk újra kategorizálást.

Adjunk meg egy 10-20 elemű névsort, és kérjünk erre is kategorizálást. Figyeljük meg az egyes kategóriákba tartozó elemszámokat. Adjuk meg egy dolgozat szerzett pontszámait egy tanulócsoportra, és kérjük ezek kategorizálását. A kvantiliseken alapuló csoportosítás miatt a szokásosnál sokkal jobb jegyek alakulnak így ki. A korábban tanult transzformációs eljárások valamelyikével hozzunk létre egy olyan csoportbasorolást, amely az UNESCO szabványt követi: a lehetséges maximális pontszám 65%-áig egyes, ezután 75%-ig kettes, 85%-ig hármas, 95 %-ig négyes, afölött pedig ötös.

Az esetek rangsorolása

Adjunk meg két változót néhány esettel úgy, hogy ezek rangsorolása ne legyen triviális. Ezek után hozzunk létre olyan csoportképző ismérveket, amelyek együttesen egyedi (egyelemű halmazokat eredményező) csoportosítást adnak. Erre az adathalmazra vizsgáljuk meg a rangsorolás lehetőségeit: a kétféle rangsorolási sorrendet, a több változó egyidejű rangsorolását, a több csoportképző ismérv együttes hatását, valamint az azonos értékkel rendelkező esetek kezelési lehetőségeit.

Minden ilyen beállításra keressünk olyan gyakorlati feladatot, amire épp ez a megfelelő rangsorolási algoritmus.

Automatikus újrakódolás

Gépeljünk be egy rövid adatállományt, amely tartalmaz egy szöveges és egy numerikus változót néhány esetre vonatkozóan. Kérjünk automatikus újrakódolást a szöveges változóra az abc-sorrenddel ellentétes rendezésnek megfelelően. Az új változó neve legyen a régi egy r betűvel megtoldva.

A numerikus változót az alapértelmezésnek megfelelően kódoljuk az automatikus újrakódolás utasítással. Ennek kapcsán próbáljuk ki, hogy az alapbeállítást a Reset gombbal lehet visszaállítani. Ellenőrizzük a cimkék átvitelét.

Idősorok létrehozása

Keressünk az interneten idősor jellegű adatokat, pl. a `http://www.bet.hu`, vagy a `http://www.fornax.hu` vendégoldalakon valamely részvény vagy index értékeit. Gyűjtsünk össze több napra vonatkozó számsort, és ezekre hajtsuk végre a rendelkezésre álló függvényekkel az új idősorok kialakítását.

Vizsgáljuk meg az eltérések esetén a rend, mozgó átlagok esetén pedig az időszak hossza paraméterek hatását.

A hiányzóadat kódok pótlása

Adjunk meg egy olyan adatállományt, amelyben lehetőleg sok hiányzóadat kód van, vagy hozzunk létre ilyeneket. Ezek után hajtsuk végre a hiányzóadat kód pótlása parancsot minden rendelkezésre álló függvénnyel, és hasonlítsuk össze a kapott új értékeket. Adjunk meg mindegyik új értéket számító eljáráshoz olyan gyakorlati helyzeteket, ahol az illető módszer célravezető lehet. Indokoljuk is meg ezt.

2.9 A statisztikai eljárások

Az Analyse menüsor tartalmazza a statisztikai eljárásokat. Ide van sorolva az összes, a statisztika körébe tartozó algoritmus, ezen kívül csak a Graphs menüsorban találhatunk statisztikai jellegű megjelenítési parancsokat. A még nem érintett további menük kiegészítő jellegűek, segédprogramokat tartalmaznak (Utilities), az ablakkezelést segítik (Window), és sokféle súgó, leírás van a Help menüben. A gyorsító billentyűkombináció az Alt-a a menüsorhoz.

Az előzőek után kevésnek tűnhet hogy csak egy menüsorra való eljárás van beépítve, de az Analyse menüsor valójában nem közvetlenül parancsokat tartalmaz, hanem almenüket, a feladatkörök szerint csoportosítva. Didaktikai szempontból nagyon dícséretes, 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 magasabbrendű 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. Nem foglalkozunk az általánosított lineáris modellekkel, a loglineáris analízissel, a nemparaméteres tesztekkel, az idősorokkal, a túlélési statisztikákkal és a többértékű statisztikák menükben foglalt eljárásokkal. Ezen túl 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 kevesebbnek a kicsit mélyebb megismerése tűnt hasznosabbnak. A kiválasztás során szempont volt az eredmények látványossága is, hogy egy statisztikai eljárás hasznossága mennyire nyilvánvaló egy kezdő számára.

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

2.10 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 SigmaStat ad. Magyarázó, értelmező szöveget nem kapunk, csak a statisztikai adatok, eredmények táblázatait, esetenként a jelölések megfejtésével. A gyorsító billentyűkombináció az Alt-a p, illetve az Alt-a e.

A jelentésírás menüponthoz a következő parancsok tartoznak: 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 menüsorból is csak egy parancsot tárgyalunk részletesen, a gyakoriságok (Frequencies...) nevűt. A rendelkezésre álló utasítások a Frequencies..., a Descriptives..., az Explore.. és a Crosstabs...

2.10.1 Eset összefoglalás

Az Analyse / 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 baloldali listából át kell tenni a jobboldali, 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 jobboldali ablakban megadott sorrend az eredmények megjelenési sorrendje is lesz.

2.10.2 A Gyakoriságok (Frequencies) parancs

Az Analyse / Descriptive Statistics / Frequencies menüpont kiválasztása után egy párbeszédes ablakot kapunk. 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 ablak alján.

Itt négyféle statisztikai mutatót kérhetünk: percentilis értékeket (Percentile Values), középértékeket (Central Tendencies), a szóródásra (Dispersion) és 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-el 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 (With normal curve).

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).

2.10.3 Példa

Az elsőtől-harmadik 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 vátozó típusát előzőleg

szövegesre (string) kell 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 a többi is megegyezik ezzel lényegében):

```
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 .
```

Frequencies

Statistics

		VAR00001	VAR00002	VAR00003
N	Valid	4	4	4
	Missing	0	0	0
Mean		2,5000	-,3250	
Std. Error of Mean		,6455	,4328	
Median		2,5000	-,2500	
Mode		1,00	-1,30	
Std. Deviation		1,2910	,8655	
Variance		1,6667	,7492	
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        4,0000        ,5000
            90        4,0000        ,5000
a   Multiple modes exist. The smallest value is shown

Frequency Table

VAR00001
                                Valid  Cumulative
          Frequency Percent Percent  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.

2.10.4 Gyakorlat

Az eset összefoglaláshoz tekintsünk egy olyan adatállományt, amely többféle változót tartalmaz számos különböző változóértékkel. Legyen közöttük diszkrét (csoportosításra alkalmas) változó is. Válasszunk ki néhány változót, amelyekre az összefoglaló adatokat kérjük, és keressünk egy csoportképző ismérvet is. Ezek alapján kérjük az összefoglalást minden rendelkezésre álló mutatóra. Próbáljuk ki, hogy a statisztikai mutatók táblázatbeli sorrendjét hogyan lehet szabályozni a beállító ablakbeli sorrenddel. A párbeszédes ablak alján levő opciókat is használjuk ki: korlátozzuk a megjelenítendő esetek számát, csak az érvényes eseteket vizsgáljuk, és kérjük az esetek sorszámozását is.

Az Options gomb lenyomásával tudjuk a táblázataink feliratát átírni, magyarázó szöveget adni a táblázatokhoz: használjunk magyar ékezetes szöveget. Kérjük a hiányzó adatot tartalmazó esetek kihagyását az összefoglaló statisztikákból. Végül írjuk át a hiányzóadat kódokat pl. a "hiány" szövegre a táblázatban.

Indokoljuk, hogy miért hasznos a leíró statisztikák használata a statisztikai feldolgozás korai fázisában, amikor az adatok között még lehetnek elgépelt, hibásan bevitt értékek is. Mely mutatók hasznosak ebből a szempontból?

A Gyakoriságok parancs gyakorlásához vegyük elő az előző gyakorlat adatállományát. Legyenek hiányzóadat kódok az adatok között. Ezek után kérjük a Gyakoriságok parancs végrehajtását úgy, hogy beállítjuk minden lehetséges statisztikai mutató kiszámítását. Figyeljük meg, hogy ezek hogyan jelennek meg a táblázatokban. Kérjük egyenként mindegyik diagramot is, és próbáljuk végig a lehetséges opciókat. Végül a Format... gomb megnyomása után az alapbeállítás szerinti sorrendeket változtassuk meg. Hasonlóan próbáljuk ki a táblázatok kategóriaszám szerinti korlátozását is.

Mondjunk olyan gyakorlati helyzetet, amikor épp a kördiagram, illetve amikor az oszlopdiagram a megfelelő ábrázolási forma. Indokoljuk, hogy miért hasznos a több változó statisztikai mutatóinak közös táblázatban való összefoglalása. Keressünk olyan alkalmazási példát, amikor a táblázataink sorainak sorrendje épp az előfordulások száma szerint csökkenő sorrendbe rendezett kell hogy legyen! Mondjunk egy példát olyan feladatra, amelynek megoldása során hasznos, hogy a hisztogram megjelenítése mellett láthatjuk az arra illesztett normális eloszlású görbét is. Figyeljük meg, hogy az adatok változtatásával hogyan módosul az illesztett görbe.

2.11 Az output ablakok tartalmának szerkesztése

Az előző két szakaszban sokat foglalkoztunk az output ablakokkal, az eredményeket a továbbiakban is 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 azért vannak köztük. A Window menüsorban most láthatunk második ablakot is, mert egy munkalap használata közben egy másiknak a megnyitása kötelezően az előző bezárását jelenti. 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-w 1, Alt-w 2 stb. gyorsító billentyűk is használhatók.

A menüsorok kettő kivételével megegyeznek: az adatállományt tartalmazó ablak Data és Transform menüsora helyett az output ablakban a Insert és Format menük vannak. Az egyes menüsorok tartalma is eltér, például a File menüsor csak az output ablak esetén tartalmaz nyomtatási kép utasítást (Print preview). Az új menüsorokat itt tárgyaljuk, a többi használatára vonatkozó tudnivalókat pedig a munkalapokra vonatkozóan adjuk meg. Annál is inkább, mert például az Analyse menüsor mindenképpen az adatállományunkat érinti.

Az Insert menüsor különböző formázójeleket, 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 találkozhattunk. 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 ábrákat tudunk az output ablakba vinni (ezek az Interactive 2D Graph..., Interactive 3D Graph... és Old Graph). Ezeket a lehetőségeket itt nem tárgyaljuk részletesen, de a szövegszerkesztő programokba épített rajzoló programokhoz hasonlóan egy szerkesztőablakban ikonokkal és más segédeszközökkel hozhatjuk létre az új ábrát. Az utolsó utasítás pedig egy korábban létrehozott ábrát szúr be.

A Text File parancs szöveget szúr be, de az csak .LST kiterjesztésű file lehet. Ennek az az oka, hogy az SPSS korábbi verziói ilyen szöveges állományokban adták meg az eredményeket. A jelen parancs arra való, hogy az SPSS korábbi változatai által létrehozott outputot megtekinthessük. Az Object utasítás pedig más objektumot, mint például hangot, képet, vagy videót illeszt be.

A Format menüsor mindössze három utasítást tartalmaz. Ezek csak akkor aktivizálódnak, amikor az output ablakban valamely objektumot: szövegrészt, ábrát vagy táblázatot kijelöltünk. Ezek után az Align Left paranccsal

azt a lapon belül balra igazíthatjuk (a gyorsító billentyű Ctrl+[)[7], a Center középre igazítja (a gyorsító billentyű most Ctrl+e), az Align Right pedig jobbra (a gyorsító billentyű Ctrl+])[7]. 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.

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

Oldalbeállítás

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 vannak a Nyomtató... (Printer...) és az Options... lenyomása után kapott ablakokban. Az előbbiben a szokásos módon megadhatjuk, hogy az installált nyomtatók közül melyiket választjuk. Vigyázat, valamit akkor is installáljunk, 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 Options ablakban adhatjuk meg az élőfejet és az élőlábat. Aki nem ismeri, annak furcsa lehet ez a két kifejezés: ezek a minden lap tetején és alján látható rögzített szöveget jelentik. A jelen jegyzet élőfejében például az oldalszám és a fejezet címe van. Az SPSS nem támogatja a páros és páratlan

[7] Legalábbis a leírás szerint, tapasztalatom szerint nem működik.

oldalak megkülönböztetését ebből a szempontból sem, de például a kötés miatt szokásos belső margót sem, ami szintén a páros és páratlan oldalakon hol a bal, hol a jobb oldalon jelenne meg.

Az Options párbeszédes ablak Options lapján ;-) az ábráink nyomtatáskori méretét specifikálhatjuk: ez lehet akkora, amekkora az most éppen 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 alapértelmezéssé tehetjük a Make Default gomb megnyomásával.

A nyomtatási kép

A Nyomtatási kép (Print preview) parancshoz nem tartozik gyorsító billentyű kombináció, de a szokásos ikon megtalálható az eszközsorban (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 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 is nyomtató kötve. A létrejövő file kiterjesztése .prn lesz (mint más programok esetén is). Ha a beállított nyomtatónk egy postscript nyomtató, akkor ez a nyomtatási adatállomány postscript formátumú lesz (még ha a file kiterjesztése nem .ps is. Bár ezt nem szokás szerkeszteni, mégis jól használható, sok más program tudja kezelni, például a TEX, LATEX szövegszerkesztők és számos rajzolóprogram is.

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 objects, Ctrl-k 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, Ctr-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, gyorsító billentyű vagy ikon nincs hozzá.

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 (Output Document), az ábrák nélküli dokumentumot (Output Document (No Charts)), vagy csak ábrákat (Charts Only). Ezek után meg kell adnunk a létrejövő file nevét könyvtárral együtt, de kiterjesztés nélkül. Ehhez segítséget ad a jobb oldalon levő Tallóz (Browse...) gomb.

Az Export What rovatban arról kell nyilatkoznunk, hogy minden objektumot ki akarunk-e vinni, vagy csak a nem rejtetteket. A rejtetteket az output ablak bal oldalán lévő összefoglaló diagramban (outline) lehet kijelölni a + és - jelek használatával. Az adatkivitel formátumát a következő rovatban tisztázhatjuk. A lehetőségek függnek attól, hogy mit akarunk kivinni: szövegre .HTM vagy .TXT típusú lehet, ábra esetén pedig .CGM, .JPG, .PCT, .EPS, .TIF, .BMP és .WMF is. Ennyi grafikus file forma minden bizonnyal a legtöbb alkalmazás számára elegendő.

További finomabb beállításokat tehetünk az Options... vagy a Chart Size... gombok megnyomása után. Utóbbival az ábrák méretét adhatjuk meg, hogy a jelenleginek hány százaléka legyen a ténylegesen kivitt. Az előbbi beállítási lehetőségei a kiválasztott filetípustól 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.

2.12 Táblázatok

A Custom Tables menüsor[8] különböző táblázatok írásához ad segítséget. Ennek négy almenü pontja van: Basic Tables..., General Tables..., Multiple Response Tables... és Tables of Frequencies... Ezek közül most csak az elsővel foglalkozunk. Az Analyse / Custom Tables / Basic Tables parancs elérhető az Alt-a t b gyorsító billentyűkkel.

Az itt képezhető táblázatok arra szolgálnak, hogy a változóinkra vonatkozó statisztikákat az esetek alcsoportjai szerint rendezve láthassuk. Így például a nemek, csoportok azonos értékeire kaphatunk különböző leíró

[8] Ez a menüsor csak akkor áll rendelkezésre, ha a TABLES modul benne van a programcsomagban (a felsőoktatás számára jelenleg biztosítottban elérhető).

statisztikákat a szokásos elrendezésnek megfelelően. Az adatok csoportosítása a fő erősség, hiszen az itt kezelt statisztikai mutatók különben elérhetők a korábbiakban tárgyalt jelentések és leíró statisztikák menüpontok parancsaival is.

A Basic Tables... utasítás egy párbeszédes ablakkal tisztázza, hogy mi a teendője. A baloldali változólistából először válasszuk ki azokat, amelyekre a statisztikákat kérjük. Ezeket a jobboldali kisebb ablakba kell megint áttenni. Ezek után jön a korábban nem szokásos csoportok (subgroups) kialakítása.

Az alcsoportok megjelenítésére a táblázatok celláinak sorait, oszlopait és egymást követő azonos szerkezetű táblázatokat fogunk használni. Ezen a módon az azonos alcsoportba tartozó statisztikai mutatók közvetlenül egymás mellett lesznek, megkönnyítve ezzel az összehasonlítást.

Arra, hogy a program mely változó azonos értékei szerint szedje szét a táblázatot sorokra, a Subgroups / Down ablak vonatkozik, ide kell az érintett változót áttenni. Az egyes oszlopokat az Across ablakba bevitt változó megegyező értékei jelölik ki. Ilyen táblázatokból többet fog a program létrehozni aszerint, hogy a Separate Tables ablakba bevitt változónak hány különböző értéke van[9]. Ezekbe a pozicionáló ablakokba több változót is tehetünk, ilyenkor ezek kombinációja fog hatni, ha az All combinations (nested) lehetőséget választottuk. Különben, az Each separately (stacked) opció választásával olyan sorokat kapunk, amelyekben először az első, majd a további változók szerint egyenként szétszedett csoportokra vonatkozó statisztikákat kapjuk meg. Vigyázat, a kombinációs esetben nem kérhetjük később a rendezést a Statistics nyomógomb megnyomása után kapott ablakban!

2.12.1 A Statistics... nyomógomb párbeszédes ablaka

Azt, hogy mely statisztikai mutatókat kapjuk az egyes cellákban, a Statistics nyomógomb megnyomása után kapott párbeszédes ablakban állíthatjuk be. Az egyik fontos mutató az illető csoportbeli érvényes esetek száma (Count). Kérhetjük ezen kívül az előbbi, csoporton belüli érvényes esetek számának soron belüli relatív, százalékos arányát (Row %). Hasonlóan az egy oszlopra, illetve az egész táblázatra vonatkozó százalékos arányt a Col %, illetve a Table % adja. A jobb egérgombbal a megfelelő feliratra kattintva azt tudjuk meg, hogy ezek csak csoportképző ismérvvel használhatók-e. Ha

[9] Megint egy eltérés a várható eredménytől: az itt megadott változónak csak egy értékére készül táblázat.

ilyenek nincsenek, akkor is ki tudjuk választani mind a Row %, a Col % és a Table % mutatókat, de ezekre minden helyen a 100% értéket kapjuk.

A maximum, minimum, átlag, medián, módusz és a különböző kvantilis értékekkel már korábban is találkoztunk, ezek itt is elérhetők. Ide tartozik még a terjedelem, a standard hiba, a szórás, az értékek összege, a variancia és további relatív mutatók a változóértékek teljes összegére vonatkozóan (az összes esetet figyelembevéve, illetve csak az érvényes esetekre).

A táblázatbeli kvantilis értékeket vagy a leggyakoribbak közül közvetlenül választjuk, vagy megadhatjuk, hogy melyik másikra gondoltunk (egyszerűen a Percentile megadásával).

A táblázat számainak formátumát is specifikálhatjuk a Format rovatban. A label ablakban pedig az aktuális, kiválasztás előtti statisztikai mutató feliratát módosíthatjuk: a standard angol név helyett akár ékezetes magyar nevet is használhatunk. Végül a Sorting by Cell Counts rovatban azt lehet beállítani, hogy a táblázatbeli cellák mi szerint legyenek rendezve. Az alapértelmezés az, hogy nem rendezettek (a kategóriaértékek sorrendjének megfelelőek). Lehetnek ezen túl a csoporton belüli esetszám szerint csökkenő és növekvő sorrendűek. Az alkalmazásokban talán az előbbi a gyakoribb.

2.12.2 További beállítások

A továbbiakat több nyomógomb segítségével lehet elérni. Mivel a táblázatok alakja sokféle lehetőséget rejt, és egy-egy alkalmazói kör más és más formához szokott hozzá, így ezek a beállítások nagyon hasznosak ahhoz, hogy további szerkesztésre ne legyen szükség az eredmények megjelenítése, összefoglalása során.

A Layout gombhoz tartozó párbeszédes ablakban adhatjuk meg, hogy az összegző változók (tehát amikre a statisztikák vonatkoznak) cimkéi hol legyenek: a bal oldalon, a táblázat tetején, vagy külön táblázatban[10]. Hasonló módon adhatjuk meg a statisztikák nevének helyét is a következő rovatban (Statistics Labels). Ha az összegző változók cimkéit a baloldalra kértük, akkor még egy rovatot tölthetünk ki aszerint, hogy az összegző változók, vagy a csoportokat kialakító változók képezzék-e a bővebb kategóriát a cimkék sorrendjét tekintve (és persze a statisztikák megjelenítését tekintve is).

[10] Ez megint nem működik: csak egy változóra kaptam táblázatot.

Szintén abban az esetben, amikor az összegző változók cimkéit a baloldalra kértük, a Totals gomb hatására olyan beállítási lehetőségekhez jutunk, amelyekkel kérhetjük a teljes csoportra vonatkozó statisztikákat az egyes csoportképző ismérvekre, illetve a teljes táblázatra vonatkozóan (a hiányzóadat kódok figyelembevételével).

A Format gombbal arról dönthetünk, hogy az üres cellák hogyan jelenjenek meg: szó szerint üresen vagy nullával jelölve. Szintén itt adhatjuk meg azt, hogy mivel jelöljük a túl sok hiányzóadat kód miatt meghatározhatatlan értékeket.

Végül az utolsó formázási célú gomb a Title feliratú, aminek segítségével valóban a táblázatok feliratait adhatjuk meg: a címét (title), a magyarázó feliratot (caption), és a táblázat bal felső sarkába írandó szöveget. Ez utóbbit nem mindig találjuk meg itt, megjelenésének feltételét pontosabban a krémszínű buborékban adott magyarázat adja meg.

Ha nagyon sok adatot jelenítünk meg a Basic Tables paranccsal, akkor a program azokat értelmes módon felosztva több kisebb táblázatba szervezi.

2.12.3 Példa

Tekintsük a következő kis adathalmazt (ez csak az utolsó oszlopban tér el a Gyakoriságok parancs példájában használttól): 1, 3, 2, 4; 0,30, -1,30, 0,50, -0,80; 1, 0, 0, 1. Kérjük erre a három változóra a Basic Tables eljárás végrehajtását. A csoportokat a harmadik változó megegyező értékei alapján képezzük (ennek a nevét tegyük be a Subgroups Down ablakba). A többi kettőre pedig kérjük a minimum, átlag, median, módusz és maximum mutatókat.

A következő eredményt kapjuk:

```
* Basic Tables.
TABLES
  /FORMAT BLANK MISSING('.')
  /OBSERVATION var00001 var00002
  /TABLES var00003 > (var00001 + var00002)
  BY (STATISTICS)
  /STATISTICS
  maximum( )
  mean( )
  median( )
  minimum( )
  mode( ).
```

```
Tables

                          Maximum Mean Median Minimum Mode
VAR00003     ,00 VAR00001  3,00   2,50  2,50   2,00   2,00
                 VAR00002   ,50   -,40  -,40  -1,30  -1,30
            1,00 VAR00001  4,00   2,50  2,50   1,00   1,00
                 VAR00002   ,30   -,25  -,25   -,80   -,80
```

Világosan leolvasható, hogy az egyetlen létrejött táblázatban a harmadik változó értékei szerint szétbontva kapjuk a kért statisztikai mutatók értékeit. Az első sorban ennek megfelelően a var00003 = 0 feltételnek megfelelő esetekre határozta meg a program az első változó maximumát, átlagát stb. Ezt követi ugyanezen esetekre a második változó sora, majd ennek a két sorhoz hasonlóan azon értékek, amik a var00003 = 1 feltétel által kijelölt esetekre vonatkoznak.

2.13 Átlagok összehasonlítása

Ez a menüpont azokat az eljárásokat foglalja össze, amelyek az átlagok összevetése alapján való összefüggést tisztázzák. Ide tartozik az egyszerű átlagszámítás csoportonként (angolul 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). Itt most ezek közül csak a párosított t-próbát tárgyaljuk részletesen. Ez elérhető az Alt-a m p gyorsító billentyű kombinációval is.

2.13.1 Párosított t-próba

Maga az eljárás két normális eloszlású sokaság várható értékének (illetve átlagának) eltérésének ellenőrzésére való. Ez a próba kis mintaelemszám esetére alkalmas, két összetartozó (nem független) minta összehasonlítására. 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 értekek a beadás előtti és 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 hatására egy olyan párbeszédes ablakot kapunk, amely az eljáráshoz illeszkedő módon kéri a paraméterezést: a baloldali változólistából most változó párokat kell kiválasztanunk és a jobboldali ablakba vinni. Ennek megfelelően az átmozgató nyíl addig nem is aktív, amíg mindkét változót ki nem jelöltük (eltérően a korábbiaktól). A kijelölés közben egyrészt a kijelölt változó inverz módban jelenik meg a listában, másrészt a lista alatti rovatban konkrétan meg is nevezi a kijelölt változókat a program.

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 a próba konfidenciaszintjét adhatjuk meg. Az alapértelmezés a szokásos 95%. Érdekes módon itt megadhatunk 100%-ot is, vagy nagyobb értéket, és csak a Continue gomb megnyomása után panaszkodik a program, hogy az érték az elfogadható tartományon kívül van.

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 egy vagy minusz egy, akkor pedig a megadott változók között szigorúan vett függvényszerű kapcsolat mutatható ki. Az első esetben a két változó együtt nő, minusz egy esetén pedig az egyik növekedése a másik csökkenésével jár.

A szignifikancia ez esetben azt fejezi ki, hogy a korrelációs együtthatóra kapott érték mennyire megbízható, mennyire jól becsüli a minta alapján a populációbeli összefüggést. 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 táblázaból visszakereshetünk, vagy a szignifikancia 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 occured 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 következté-

ben adódott, nem elég nagy ahhoz hogy kizárjuk annak a lehetőségét, hogy az eltérés a véletlen műve (P=...). Az ilyen szöveges értelmezés jelentősége vitatott, a szerző inkább hasznosnak tartja, 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.

2.13.2 Példa

Tekintsük az előző szakasz példájában szereplő 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 var00001 var00002  WITH
  var00002 var00003 var00003 (PAIRED)
  /CRITERIA=CIN(.95)
  /MISSING=ANALYSIS.
```

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	VAR00001	2,5000	4	1,2910	,6455
	VAR00002	-,3250	4	,8655	,4328
Pair 2	VAR00001	2,5000	4	1,2910	,6455
	VAR00003	,5000	4	,5774	,2887
Pair 3	VAR00002	-,3250	4	,8655	,4328
	VAR00003	,5000	4	,5774	,2887

Paired Samples Correlations

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

```
Paired Samples Test

                  Paired Differences
        Mean    Std.      Std.     95% Conf. Int.    t  df Sig.
             Deviation Error Mean  Lower   Upper
Pair 1 2,8250  2,0288   1,0144     -,4032  6,0532  2,785 3 ,069
Pair 2 2,0000  1,4142    ,7071     -,2503  4,2503  2,828 3 ,066
Pair 3 -,8250   ,9912    ,4956    -2,4022   ,7522 -1,665 3 ,195
```

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.

2.13.3 Gyakorlat

Gyakorlásként először generáljunk egy viszonylag kis elemszámú, normális eloszlást követő mintát. Ezt másoljuk át egy másik változóba, és kíséreljük meg az ezek közötti eltérést kimutatni a párosított t-próbával. A program tiltakozni fog. Ezek után változtassuk meg egy kicsit az egyik változó értékeit, és ismételjük meg a statisztikai eljárást. Értelmezzük az eredményt. A szignifikanciaszint változtatásával vizsgáljuk meg azt, hogy milyen szinten mutatható ki szignifikáns eltérés az átlagra vonatkozóan.

A következő feladat: generáljunk egy nagyobb elemszámú normális eloszlású mintát ugyanazzal az átlaggal és szórással, mint az előbb, majd ennek eseteit megosztva ebből hozzunk létre két új változót. Vizsgáljuk meg, hogy a két új változó átlaga eltér-e szignifikánsan a párosított t-próba szerint. Az alapminta mérete változtatásával (tehát új adat generálásával) keressük meg azt a határt, amelynél a 95%-os szignifikanciaszint mellett statisztikai értelemben vett eltérés mutatható ki a megosztott változó két felének átlagára vonatkozóan.

2.14 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[11]. A menüsorból elérhető parciális korrelációt nem tárgyaljuk.

2.14.1 Páronkénti korreláció

A páronkénti korrelációt (Bivariate) az Analyse / Correlate / Bivariate utasítással kaphatjuk, vagy az Alt - a c b gyorsító billentyűkombinációval. A kapott párbeszédes ablakban először a baloldali változólistából át kell tenni azokat a jobboldaliba, 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 szignifikanciatesztet 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.

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

Példa

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

```
CORRELATIONS
  /VARIABLES=var00001 var00002 var00003
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

```
Correlations

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

Az egyes cellékban a Pearson-féle korrelációs együtthatót, a signifikancia értékét és a figyelembe vett esetek számát találjuk. A főátlóban minden változónak az önmagára vonatkozó korrelációja nyilván egy. A kapott korrelációs mátrix nyilván szimmetrikus. A három érdemi korrelációs együttható érték megegyezik a 2.13.2 szakaszban kapottakkal (a szignifikancia-értékek szintén).

Gyakorlatok

Generáljunk egy standard normális eloszlású x változót 20 esettel, és ebből számítsunk ki egy olyan új y változót, amire $y = 2x + 3 + 0.1z$ érvényes, ahol z is standard normális eloszlású. Ezek után vizsgáljuk meg x és y korrelációját a páronkénti korreláció paranccsal. Változtassuk a kezdeti 2 együtthatót úgy, hogy negatív korrelációt kapjunk, és a z szorzószámát úgy, hogy a korrelációs együttható abszolút értéke 0,05 alatt legyen!

Vizsgáljuk meg a rendelkezésre álló korrelációs együttható típusokat azonos adaton, és módosítsuk adatunkat úgy, hogy a három érték közel azo-

nos legyen. Hozzunk létre egy nagyobb adatállományt sok egyenletes eloszlású változóval, és kérjük ezek mindegyikére a páronkénti korrelációt. Próbáljunk olyan adatot megadni, amelyre a hiányzóadat kódok kezelésének megváltoztatása esetén eltérő előjelű korrelációt kapunk.

2.14.2 Távolságok

Az esetek vagy változók hasonlóságának vagy eltérésének mértékét az Analyse / Correlate / Distances... paranccsal kaphatjuk meg írott formában, az output ablakban. A gyorsító billentyű kombináció az Alt - a c d. 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 baloldali listából át kell tenni a jobboldali kisebb ablakba. Esetleg megnevezhetünk egy olyan változót lejjebb, amelyik az esetcimkéket 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. 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%     2  100,0%

Proximity Matrix

                Euclidean Distance
          VAR00001    VAR00002    VAR00003
VAR00001                1,414       2,828
VAR00002    1,414                   1,414
VAR00003    2,282       1,414

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).

Gyakorlatok

Gépeljünk be egy olyan adatállományt, amelyben körülbelül annyi változó van, mint ahány eset. Kérjünk hasonlósági mátrix számítást az Analyse / Correlate / Distances.. paranccsal mind az esetekre, mind a változókra (a

paramétereket alapértelmezésben hagyva). Vessük össze az eredményeket, és indokoljuk az eltéréseket!

Mutassuk ki az eltérést az eredményben akkor, ha az intervallum skálán mért bináris adatunkat ennek megfelelően értékeljük ki a távolság generáló utasítással, illetve ha ezeket bináris adatoknak minősítjük. Mindkét esetben használjuk az euklideszi távolságot! Magyarázzuk meg a kapottakat.

2.15 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 paraméterbecslést (Curve Estimation, Alt - a r c) és a nemlineáris regressziót (Nonlinear..., Alt - a r n) tárgyaljuk. Ez a két parancs lényegében csak abban tér el egymástól, hogy az első az 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 azon túlmutat.

2.15.1 Paraméterbecslés

A paraméterbecslési utasítást tehát 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 kell kijelölnünk. Itt többet is megadhatunk, és ebben az esetben a megadott változószámnak megfelelő darab paraméterbecslési 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. Megadhatunk esetcimkéket is.

Beállíthatjuk azt is, hogy kérjük-e az egyenletben, a magyarázófüggvényben szereplő additív konstans kiíratását. Szintén itt, a párbeszédes

ablak közepén lehet azt is megadni, hogy kérjük-e az eredmény grafikán való bemutatását.

Ezután jön a modell típusának beállítása. A leggyakoribb regresszió-függvények itt megtalálhatók, 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).

Az egyes függvények pontos leírását megtalálhatjuk képletekkel együtt, ha a jobb egérgombbal a nevükre kattintunk. Ekkor láthatjuk azt is, hogy a modellfüggvény mely paramétereit állítja majd be az algoritmus az adatpontoknak megfelelően. Ezeket a b0, b1 stb. nevekkel illetett paramétereket aztán megtalálhatjuk 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.

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 a jósolt értékre vonatkozó konfidencia intervallumokat. Ez utóbbinak a százalékos konfidenciaszintjét 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 adatunk alapján a jelentős mértékű zaj ellenére. A párbeszédes

ablakban megjelöltük z-t, mint a eredményváltozót, x-et mint a magyarázóváltozót, és a lineáris *és*[12] kvadratikus modellfüggvénnyel kérjük az illesztést. A kapott eredmény:

```
* Curve Estimation.
TSET NEWVAR=NONE . PREDICT THRU END. CURVEFIT /VARIABLES=z
WITH x
  /CONSTANT
  /MODEL=LINEAR QUADRATIC
  /PLOT NONE.

  Curve Fit

MODEL:  MOD_7.

Independent:  VAR00001

Dependent Mth   Rsq  d.f.       F  Sigf      b0      b1      b2

Y        LIN  ,976     4  163,37  ,000 -5,8772  8,8533
Y        QUA 1,000     3 3740,12  ,000  2,9005  2,2700   ,9405
```

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

1. Az eljárás eredményének illusztrálása céljából tekintsünk egy változópárt, amelyek közül az első tartalma legyen néhány közeli, egymásra következő egész szám. A második változó tartalmát pedig az elsőből számítsuk ki úgy, hogy az első változó értékének háromszorosához adjunk hozzá kettőt, és még egy standard normális eloszlásból származó véletlen számot is.

Ezek után a görbeillesztés segítségével vizsgáljuk meg, hogy a 3 és 2 értékeket milyen mértékben lehet rekonstruálni az adataink alapján, ábrázoljuk az illesztett egyenest, és tanulmányozzuk az eltéréseket. Változtassuk meg a zaj hatásának erejét úgy, hogy azt egy egyre kisebb pozitív számmal szorozzuk be.

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

2. Második gyakorlatként generáljunk az előzőekhez hasonló módon egy, a független változótól exponenciális módon függő új változót, és vizsgáljuk meg, hogy milyen mértékben lehet ezt a függvényt lineáris, kvadratikus, illetve köbös modellfüggvényekkel közelíteni. Ismételjük meg a kísérletet zajjal terhelt esetben is, és vessük össze a kapott grafikonokat.

2.15.2 Nemlineáris regresszió

A nemlineáris regresszió szabad modellfüggvény beállítást tesz lehetővé, az összefüggéseket tetszőleges nevű paraméterekkel és standard függvényekkel határozhatjuk meg.

Az induló párbeszédes ablakban először adjuk meg a baloldali 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 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 abszolut 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[13].

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 mi-

[13] 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.

nimalizálni kell. Ehelyett a felhasználó megadhat egy új célfüggvény alakot, 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: a jósolt értékeket (Predicted values), a reziduumokat (Residuals) és a deriváltakat (Derivatives). Itt jegyzem 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 pl., 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 alapulvett 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 értekek: 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 a 2,5-es induló értékkel 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:

```
All the derivatives will be calculated numerically.

 Iteration  Residual SS           A
     1      5,566800047  2,50000000
     1.1    5,123559253  2,43802949
     2      5,123559253  2,43802949
     2.1    4,407607396  2,35861569
     3      4,407607396  2,35861569
     3.1    1,909518704  2,19438392
     4      1,909518704  2,19438392
     4.1    ,0000291852  1,99925085
     5      ,0000291852  1,99925085
     5.1     5,0512E-12  2,00000031
     6       5,0512E-12  2,00000031
     6.1     1,7120E-25  2,00000000
     7       1,7120E-25  2,00000000
     7.1    ,0000000000  2,00000000

Run stopped after 14 model evaluations and 7 derivative
evaluations. Iterations have been stopped because the relative
difference between successive parameter estimates is at most
PCON = 1,000E-08

Nonlinear Regression Summary Statistics    Dependent Variable Y

  Source                  DF  Sum of Squares  Mean Square

  Regression               1        2,46759        2,46759
  Residual                 4         ,00000         ,00000
  Uncorrected Total        5        2,46759
  (Corrected Total)        4        2,40957

  R squared = 1 - Residual SS / Corrected SS =    1,00000

                                              Asymptotic 95 %
                           Asymptotic     Confidence Interval
  Parameter   Estimate     Std. Error     Lower         Upper
  A          2,000000000   ,000000000  2,000000000  2,000000000
```

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

Gyakorlásként először keressük meg az $y = a * x^2$ függvény legjobb regressziós illesztését biztosító a értéket több indulóponttal is. Alapadatként adjunk meg legalább 5 pontot az x^2 függvényértékkel együtt (tehát $a = 1$). Adjunk magyarázatot, hogy miért tudja az eljárás az eredményt lényegében egy lépésben megtalálni?

Generáljunk most egy olyan adatot, amely a $\sin(x)$ értékeit adja meg 6-8 pontra. Ezután határozzuk meg az a paramétert a $\sin(3.1415927a + x)$ képletben eltérő indulópontokkal. Indokoljuk a kapott eredményeket! Miért nem kapunk pontosan egész számokat, és mi köze van ehhez a megadott 3.1415927 számnak? A szerzett tapasztalatok alapján mely indulópontból mely megoldást kapjuk meg?

Keressünk olyan nemlineáris regressziós feladatot, amelynek több helyi minimuma is van, és ezek reziduum értékei lényegesen eltérőek. Javasoljunk ennek alapján egy követendő startégiát, amellyel ilyen jellegű feladatok legjobban illeszkedő megoldását meg tudjuk találni.

2.16 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 rajta, amelynek elemei közelebb vannak egymáshoz, mint más klaszterek pontjaihoz.

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 idetartozó eljárásokat az Analyse / 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 gyorsítóbillentyű kombinációk: Alt - ayk, illetve Alt - ayh.

2.16.1 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 klasztereket 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 jobboldali kis ablakba. Az eseteket a sorszámuk azonosítja alapértelmezésben, de kérhetjük azt is, hogy ehhez az esetcimkéket használja a program. Fontos megadni az output klaszterek kívánt számát. 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, a bonyolultabb esetben ezt a centrumok iteratív pontosítása előzi meg. Ezt a választott módszer rádiógombjával lehet megadni. A centrumokkal külön kérésre a program további műveleteket is végez: a kiindulási centrumokat betölti egy adatállományból, vagy az eredményül kapott centrumokat kimenti egy megadott file-ba. Ezeket a lehetőségeket a Centers >> gomb megnyomása után 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 (az alapértelmezés 2%).

A Using running means opció bekapcsolásával az új centrumok mindig

a régi klaszter átlagaiból adódnak. Elmenthetjük a klaszterhez tartozás mutatóját (Cluster membership) és a klasztercentrumoktól 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 hiányzó adat 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 kértük az Options ... ablakban az ANOVA nélkül.

A kapott eredmény[14]:

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

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

```
Convergence achieved due to no or small distance change. The
maximum distance by which any center has changed 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 való megjelenését.

Gyakorlat

A gyakorlat a K-közép klaszterezés két egyedi sajátosságára koncentrál: az induló centrumokra és az előre megadott számú kialakítandó csoportra.

Először tekintsük azt az esetet, hogy egy dolgozat értékelése után a kapott pontszámoknak adott egy mintája, amiket szeretnénk 5 osztályba sorolni (remélve, hogy ezeket majd meg lehet feleltetni osztályzatoknak). Vigyük be egy változóba a következő pontszámokat (25 darab, valós adatok): 7,5; 8,5; 8,5; 8,5; 8,0; 9,0; 11,0; 12,0; 8,0; 4,5; 5,5; 10,0; 10,0; 7,5; 8,5; 11,0; 11,5; 9,5; 9,0; 11,0; 8,0; 6,5; 5,5; 2,5; 6,0. Kérjünk K-közép klaszterezést úgy, hogy 5 csoport jöjjön létre az eljárás végrehajtása során, és használjunk iterációt a centrumok javítására. Értékeljük ki a kapott eredményt, hogy az milyen mértékben használható az osztályzatok meghatározására.

Ezután ismételjük meg ezt a vizsgálatot úgy, hogy a kapott centrumokat mentsük ki egy állományba. Módosítsuk ezeket a centrumokat úgy, hogy lehetőleg közel azonos számú legyen minden érdemjegyből a klaszterezés újrafuttatása után! A centrumok kimentett állománya egy szabványos SPSS file, így magával az SPSS programmal tudjuk módosítani is. Magyarázzuk meg, hogy pusztán a centrumok megváltoztatásával miért nem lehet elérni, hogy pontosan megegyező számú legyen minden érdemjegyből.

Végül az előző adaton végezzünk el egy olyan kísérletet, hogy a kezdeti centrumok megadása nélkül hogyan változik a klaszterek elhelyezkedése, miközben a klaszterek számát fokozatosan növeljük. Indokoljuk meg, hogy ez az eljárás miért nem igazán alkalmas az osztályzatok meghatározására, és említsünk olyan alkalmazási feladatot, ahol viszont az előre adott számú klaszter megkeresése az eredeti feladatot legjobban kifejező statisztikai modell.

2.16.2 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:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

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 jobboldali kis ablakba). Megadhatunk egy esetcimkéket 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. icile plot. Ez szintén karakteres formában jeleníti meg az összetartozó eseteket (vagy változókat): ezek sorszáma lesz a táblázat fejlécében megadva, és alatta klaszterezés szintjeinek megfelelő magasságú X-ekből álló oszlop mutatja, hogy az adott eset mely másokkal tartozik egy klaszterbe az adott szinten.

A legfontosabb beállítási lehetőségeket a Method gomb rejti: itt kell megadnunk, hogy melyik klaszterezési módszert választjuk (a klaszterek közti eltérés mi), 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 elfogadott 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 felmerült feladatot a hierarchikus klaszterezés alapbeállításaival, de kérjünk dendrogramot a kiértékelés megkönnyítése kedvéért (Plot / Dendrogram). A kapott eredmény dendrogramja (az eredeti ábra kifejezőbb, itt a részhalmazok nem mindig elkülönültek):

```
* H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S *

 Dendrogram using Average Linkage (Between Groups)

                    Rescaled Distance Cluster Combine

 C A S E    0         5        10        15        20        25
Label  Num  +---------+---------+---------+---------+---------+

        11   -+
        23   -|
        25   -+---+
        22   -+   |-----------------+
        10   -----+                 +-------------------------+
        24   -----------------------+                         |
         1   -+                                               |
        14   -+-+                                             |
         9   -| |                                             |
        21   -| |                                             |
         5   -+ +-------------+                               |
         4   -+ |             |                               |
        15   -| |             |                               |
         2   -| |             |                               |
         3   -+-+             |                               |
         6   -|               +-------------------------------+
        19   -|               |
        18   -+               |
        12   -+---+           |
        13   -+   |           |
        16   -+   +-----------+
        20   -|   |
         7   -|   |
        17   -+---+
         8   -+
```

A dendrogramból leolvasható, hogy a megadott sorrendben a 11-24 sorszámú esetek lehetnek az egyes érdemjegyűek, 1-5 a kettes, 4-18 a hármas, 12-13 a négyes és a 16-8 esetek az ötös osztályba tartozók. A rajzon az egyes esetek nem nagyság szerinti sorrendben vannak. A klaszterező eljárás által javasolt csoportosítás az osztályzástól elvárható tulajdonságokat

közel teljesíti, leszámítva, hogy négyes érdemjegyet csak két hallgató kap.

Érdemes megfigyelni, hogy amíg a kettes és hármas érdemjegyű csoport alacsony szinten elválik, addig az egyes jegyet kapók közötti eltérés sokkal nagyobb, és a 2,5 pontos 24-es számú eset csak nagyon magas szinten vonható össze a többi egyes osztályzatot kapottal. Természetesen az osztályozás a tudásszintet kell hogy felmérje, tehát a példánk inkább csak arra utal, hogy az adatunk 5 csoportba osztható be, de nyilván nem kell mindenképpen egy egyes osztályzatúak csoportját kialakítani.

Gyakorlat

Gyakorlásként vegyük az eredeti feladat megfordítását: adjunk meg olyan, lehetőleg egyszerű alapadatot, amelyre a hierarchikus klaszterezés pontosan egy bináris fát ad eredményül!

Rajzoljunk ezután egy olyan kétdimenziós ponthalmazt, amelyben világosan felismerhető 5 darab többelemű részhalmaz. Vizsgáljuk meg, hogy ezek mely módszerrel és milyen távolságfüggvénnyel azonosíthatók.

Végül vegyünk egy valós anonim statisztikai mintát, amely egy tanulmányi csoport, vagy néhány ismerős következő adatait tartalmazza: cipőméret, magasság, tanulmányi átlag, a kapott ösztöndíj összege. Vizsgáljuk most meg a változókra vonatkozó klaszterezéssel hogy mely változók tartoznak össze a leginkább ez alapján. Figyeljük meg a változók standardizálásának a hatását, és magyarázzuk meg a kapott eredményeket!

2.17 Az adat tömörített jellemzése

Az adattömörítés (Data reduction) menüsorba három eljárás tartozik, a faktoranalízis, a korrespondancia analízis és az optimális skálázás. Ezek közül részletesen csak az elsővel fogunk foglalkozni. Ennek a gyorsítóbillentyű kombinációja az Alt a d f.

2.17.1 Faktoranalízis

A faktoranalízis arra való, hogy egy több változóhoz kijelöljünk kevés számú egymással korrelálatlan 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 paraméterrel 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 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 figyelembeveendő változókat kell megadni úgy, hogy az érintetteket áttesszük a baloldaliból a jobboldali 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 figyelembevenni. 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 döl 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 mód-

szerrel a faktorok rotációját, illetve megadhatjuk a kívánt hiányzóadat kód kezelési eljárást.

Példa

Generáljunk egy olyan adatot, amely bár három 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 VAR00001 változóból a Transform / Compute utasítással képezzünk három újat: az x, y és z változókat a következő képletekkel rendre

$$x = VAR00001 * SQRT(3)/2 + NORMAL(0.1),$$

$$y = VAR00001/4 + NORMAL(0.1)$$

és

$$z = VAR00001 * SQRT(3)/4 + NORMAL(0.1).$$

Ezzel lényegében egy olyan háromdimenziós pontfelhőt adtunk meg, amelynek a formája egy erősen elnyújtott elipszoid, 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[15]:

```
Communalities

    Initial Extraction
X    1,000     ,994
Y    1,000     ,977
Z    1,000     ,990

Extraction Method: Principal Component Analysis.
```

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

```
                    Total Variance Explained
          Initial Eigenvalues          Extraction Sums
                                       of Squared Loadings

Component Total   % of Variance Cumul.% Total % of Var. Cumul.%
1        2,961      98,712     98,712   2,961   98,712   98,712
2        3,400E-02   1,133     99,846
3        4,634E-03    ,154    100,000

Extraction Method: Principal Component Analysis.

Component Matrix (1 components extracted)

    Component
        1
X     ,997
Z     ,995
Y     ,989

Extraction Method: Principal Component Analysis.
```

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 hármas sajátérték (2,961) tartozik, és ez a variancia közel 99%-át megmagyarázza. A másik két sajátérték lényegesen kisebb, ami a további magyarázóváltozó szükségtelenségét jelzi. A program alapbeállításában nem is engedi egy alatti saját értékekhez tartozó faktorok meghatározását.

Gyakorlat

Adjunk meg egy, a példában szereplőhöz hasonló, de két lényegi magyarázóváltozót tartalmazó ellipszoidot a háromdimenziós térben. Keressük meg a faktorokat, és diszkutáljuk a kapott eredményt.

Második gyakorlatként pedig gyűjtsünk adatokat a környezetünkből (pl. a tanulócsoportban) egyrészt a tanulmányi eredményekről, másrészt a focicsapatokat illető preferenciákról. Ezután faktoranalízissel határozzuk meg, hogy milyen lényegi összefüggések vannak az adataink mögött. Ismételjük meg a vizsgálatot úgy is, hogy sok faktort engedünk meg,

és nem ragaszkodunk az egynél nagyobb sajátértékekhez sem. Hogyan értelmezhető az eredmények eltérése?

2.18 Skálázás

A skálázás algoritmusai közül csak a többdimenzionális skálázással foglalkozunk, a megbízhatósági analízissel (Reliability analysis) nem. Az előbbi elérhető az Alt a a m 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.

2.18.1 Többdimenzionális skálázás

A többdimenzionális 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 többdimenzionális 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 jobboldali 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 dis-

tances 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[16]. Az iterációs algoritmus megállási feltételeit pedig a Criteria rovatban hangolhatjuk.

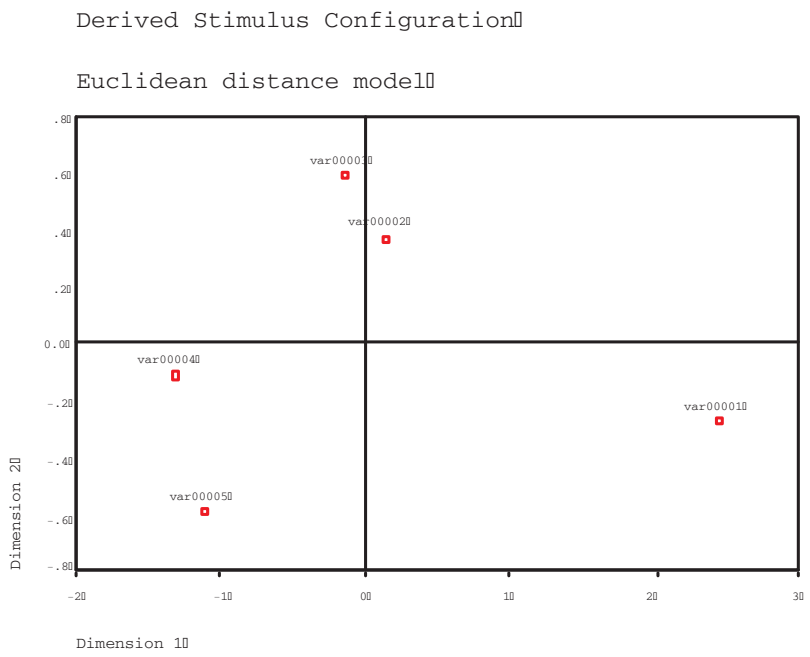
Példa

Egy egyszerű példán vizsgáljuk meg a többdimenzionális 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
```

Szemre 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 csökenteni kellene, akkor amiatt, hogy az adathalmaz eltéréseinek nagy részét három változó már hozza, elegendő lehet csak ezeknek a figyelembevétele, vagy a következő felmérésre való használata a mérési adatok sokféleségének reprezentálására. Nézzük meg, hogy mit ad ebben az esetben az SPSS többdimenzionális skálázása!

[16]Egy Windows ME operációs rendszer használata mellett következesen lehetetlen volt az ábrák megjelenítése, míg más Windows változatokkal ez problémamentesen sikerült. Az eredmény más részletei minden esetben elérhetők voltak.



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

A bevezető leírásnak megfelelő beállításokkal végrehajtva az eljárást (tehát minden változóra, távolságmátrix nélkül, kérve a Group plots lehetőséget) a várakozásokak megfelelő eredményt kapunk. Az eredmény a 3. 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 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.

Gyakorlat

Első gyakorlatként vigyük be a példában említett adatokat, és vizsgáljuk meg, hogy milyen hatása van a változókra kapott összefüggésekre a figyelembevett távolságfogalomnak, illetve annak, ha megváltoztatjuk a változóinkra érvényes mérési skálák típusát. Értelmezzük, és próbáljuk megmagyarázni a kapott eltéréseket.

Ezt követően tervezzünk egy autóvásárlást előkészítő kérdőívet, amelynek alapján egy vásárlás döntését szeretnénk előkészíteni. A változók az egyes típusokat leíró mutatók (fogyasztás, sebesség, ár, a felszereltség részletei stb.), az egyes esetek a különböző autótípusoknak felelnek meg. Az adatokat töltsük fel emlékezetből, vagy véletlenszerűen, de igyekezzünk reális számokat használni. Az egyes változók definiálása során adjuk meg a megfelelő mérési skálát is. A számos figyelembevehető szempont közül a többdimenzionális skálázás segítségével határozzuk meg azt a néhány, lényegesen kevesebb változót, amelyek értékei már lényegében leírják az ajánlatok sokféleségét, eltéréseit. Említsünk olyan hasonló alkalmazási lehetőségeket, ahol szintén a többdimenzionális skálázás segíthet a leginkább.

2.19 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 tudunk 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 (Gallery) választjuk ki azt az ábrafajtát, ami az elképzeléseinknek megfelel, és a részleteket ezután tisztázzuk egy megfelelő párbeszédes ablakban. Ennek a lehetőségnek a gyorsítóbillentyű kombinációja az Alt g g. Minden esetre ez az út inkább csak bemutatja és elmagyarázza a tennivalókat: konkrét ábra nem jön létre, csak a súgó megfelelő részéhez juthatunk el (How to).

A következő fokozat az interaktív ábrarajzolás, amikor az ábra típusá-

nak nevét ismernünk kell, de különben sok segítséget nyújtó, több lapra terjedő párbeszédes ablak támogatásával juthatunk el a grafikonunkhoz (a gyorsító billentyűkombináció Alt g a). Ez az az ábrarajzolási mód, amellyel a grafikonokat a legrészletesebben be tudjuk állítani.

Ezzel szemben a legegyszerűbb mód az, amikor a Graphs 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 g b. 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.

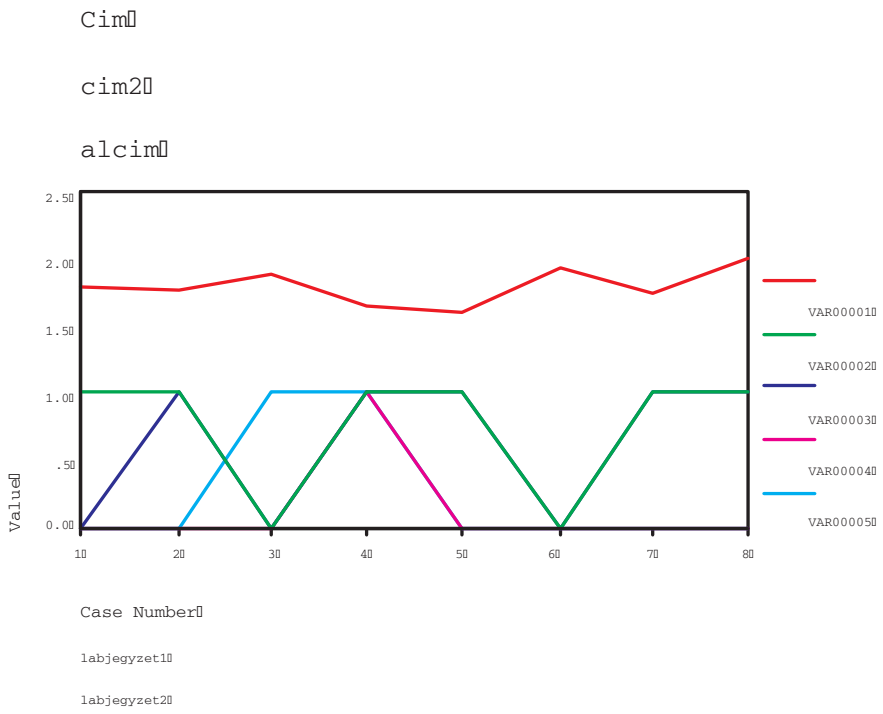
Az SPSS ábrarajzolási eljárásainak hátránya, hogy a beállítások közben nem látjuk a korábbi megadások eredményét, és arra sincs lehetőség hogy egy diagramvarázslóval a korábban létrehozott grafikonokon módosítsunk (mint egyes táblázatkezelőkben). Bár az output ablakban elérhetők a Graphs menüsor utasításai, 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, itt most csak egy példán keresztül világítjuk meg hogy hogyan kell ezeket az eljárásokat használni. Sajnos több esetben előfordul, hogy amit el lehet érni a grafikonszerkesztővel (pl. Interactive / Bar), azt egy adott szerkesztési mód (Bar) nem teszi lehetővé. Ilyen esetben a súgó segítségével közvetlenül a végrehajtandó szintaxisú utasításokat is megírhatjuk.

2.19.1 Példa

Vegyük az előző szakaszbeli adatot, és jelenítsük meg egy alkalmas grafikonon. Ehhez válasszuk a Graphs menüsorból a vonalas ábrát (Line) (több változó egyidejű megjelenítéséhez ez talán a leggyakoribb). Ahhoz, hogy több változót is együtt láthassunk, válasszuk a multiple lehetőséget: egy ábra mutatja is, hogy körülbelül hogy fog majd kinézni a diagramunk. Ez azonban egy rögzített rajz, nem az adatunkra vonatkozik, mint egyes táblázatkezelő programokban.

Ezután adjuk meg, hogy mit kívánunk megmutatni: a változóértékeket esetenként. Ez nem az alapbeállítás, hanem a “Values of individual cases”



4. ábra: A példában rajzolt ábra a feliratokkal együtt. Az egyes változókat különböző színek jelenítik meg, amik a fekete-fehér nyomtatásban különböző szürkeként látszanak.

opció, ennek a rádiógombját nyomjuk meg. A Define gomb megnyomásával tudunk továbbhaladni a paraméterezéssel. A változóinkat tegyük át a Lines represent ablakba, ezzel adjuk meg azt, hogy mely változókat szeretnénk az ábrán látni. Az x-tengely felirata lehet az esetek sorszáma (alapértelmezés), vagy ennek tartalmát megadhatjuk egy változóban is. A jelen esetben válasszuk az előbbit. Az ábra stílusát előre megírt templátokkal is kialakíthatjuk. A professzionális, gyakori használat során ezekkel hatékonyan lehet egységesíteni és gyorsítani az ábrarajzolást.

Az ábra feliratait a Titles gomb megnyomása után kapott ablakban adhatjuk meg. Itt három dolgot tisztázhatunk: a címet, az alcímet és a lábjegyzetet. Az első és a harmadik két sorra is terjedhet. Ezeket a szövegeket közvetlenül be kell írni a megfelelő rovatba: Title, Subtitle és Footnote. Itt

magyar, ékezetes betűket is használhatunk — amennyiben ezt az operációs rendszer, illetve a rendelkezésre álló betűtípusok megengedik. Az ábra kimentésénél figyeljünk arra, hogy egyes formátumok csak az ábrákra érhetők el (Chart only), az egész output ablakra (Output document) nem. Szintén említésre méltó, hogy bár az így létrejött encapsulated postscript file (.eps) használható, például a ghostview program meg tudja mutatni, és ki tudja nyomtatni, de a létrejött eps állomány illegális karakterekkel kezdődik és végződik, amire egyes programok (pl. a LaTeX szövegszerkesztő) érzékenyek. Ebben az esetben ezeket töröljük ki a file elejéről és végéről. A létrehozott grafikont a 4. ábra tartalmazza.

2.20 Feladatok

2.20.1 Egy dolgozat feladatai

1. Olvassa be a zh.txt file-ból az Elso, Masodik, Harmadik, Negyedik és a Szazalek nevű változókat (az elválasztó jel az &). Adja meg a 9, illetve a 99 értékeket mint hiányzóadat kódokat. Írja le a beolvasott utolsó sor tartalmát! 1 pont

2. Határozza meg a min, max, átlag és szórás értékeket a Szazalek változóra azon esetekre vonatkozóan, amelyeket az Elso, Masodik, ..., Negyedik változók 1 értékei jelölnek ki (4-4 érték). Írja ide a kapott értékeket! 1 pont

3. Vizsgálja meg klaszter analízis segítségével, az Elso — Negyedik változók összetartozásának szorosságát, kérjen dendrogrammot is! Foglalja össze itt az eredményt szavakban! 1 pont

4. Vizsgálja meg, hogy a Szazalek változó értékét az Elso — Negyedik változók milyen lineáris függvénye írja le (lineáris regresszió)! Értelmezze az eredményt!

   1 pont

5. Generáljon egy új változót Ujszazalek néven, amelynek tartalmát úgy kapjuk, hogy a Szazalek változóhoz a standard normális eloszlással generált véletlen számok ötszörösét adjuk, majd ezt egészre kerekítjük. Adja meg a kapott új értékeket! 1 pont

6. Ismételje meg a 4. feladat vizsgálatát az 5. feladatban kapott változóra, és értelmezze az eredményt! 1 pont

Az első feladatban említett zh.txt nevű adatállomány tartalma:

```
0 &  1 &  0 &  1 &  20
1 &  0 &  0 &0 &   63
0 &  1 &  1 & 0 & 37
1  & 1 &  0 &0 &   33
0 &  0 &  0& 0 & 67
1 &  0 &  0 &  0 &    66
0 &  1 &1&0 & 33
0  & 0&0 &  1  &    50
1& 1& 0& 1&   45
0  & 0& 1& 0& 55
```

2.20.2 Egy másik dolgozat feladatai

1. Olvassa be a `http://www.inf.u-szeged.hu/~csendes/statzh` file-t. Töltse be belőle a BUX, MOL, MATAV, OTP és BP-Növekedési nevű változókat (az elválasztó jel a szóköz) az SPSS programba. Adja meg a megfelelő beállításokat (cimke, szélesség, tizedes jegyek száma). Adja meg a 6789 értéket mint hiányzó adatkódot minden változóra. Írja ide a beolvasott utolsó sor tartalmát! 1 pont

2. Határozza meg a min, max, átlag és szórás értékeket a MOL, MATAV és OTP változókra azon esetekre vonatkozóan, amelyekre a BUX változó 5300-nál nagyobb értéket vesz fel (4-4 érték). Írja ide a kapott értékeket! 1 pont

3. Vizsgálja meg klaszter analízis segítségével, a BUX, MOL, MATAV, OTP és BP-Növekedési változók összetartozásának szorosságát, kérjen dendrogrammot is! Foglalja össze az eredményt szavakban! 1 pont

4. Vizsgálja meg, hogy a BP-Növekedési változó értékét a BUX, MOL, MATAV és OTP változók milyen lineáris függvénye írja le (többváltozós lineáris regresszió)! Értelmezze az eredményt! 1 pont

5. Generáljon 5 új változót, amelyek tartalmát úgy kapjuk, hogy a BUX, MOL, MATAV, OTP és BP-Növekedési változók értékét elosztjuk az első napi értékkel és beszorozzuk 100-al (százalékos relatív érték). Ábrázolja az új változókat egy megfelelő grafikonon! Ismételje meg

a 3. – 4. feladat vizsgálatát a kapott új változókra, és értelmezze az eredményt! 3 pont

Az első feladatban említett statzh nevű adatállomány tartalma:

Nap	BUX	MOL	MATAV	OTP	BP-Novekedesi
XI.1.	5489	4890	1279	10955	1.8789
XI.2.	6789	5120	1343	11034	1.8894
XI.3.	5834	5235	1355	11870	1.9292
XI.4.	5589	6789	1312	11345	1.9111
XI.5.	5309	5102	1298	6789	1.9034
XI.6.	5103	5023	1176	10766	1.8943
XI.7.	5302	5301	1212	11034	1.9102
XI.9.	5332	5395	1235	11203	1.9302
XI.10.	5401	5378	1320	11301	1.9422
XI.11.	5578	5567	1349	11650	1.9388
XI.12.	5632	6789	1377	11703	1.9401
XI.13.	5276	5423	1336	11329	1.9213
XI.14.	5303	5546	6789	11543	1.9324

3. Fejezet

SigmaStat eljárások

A SigmaStat a SIGMA programcsomag egyik tagja, ezekkel, mint például a SigmaPlot programmal lehet együtt használni. A jelenlegi legújabb változata, a SigmaStat 2.0 jól használható demo módban is, és elérhető a

```
http://www.spssscience.com/sigmastat/
```

internetes címen. Ez a program főleg megjelenésben sokat fejlődött a korábbi DOS változathoz képest. A program fő erőssége az érthető, világos fogalmazás, a statisztikai eljárások közti eligazodást segítő statisztikai tanácsadó, az ún. Advisor. További kellemes tulajdonsága a SigmaStat programnak, hogy teljes értékű jelentést ír a végrehajtott statisztikai eljárások eredményeiről (angolul), és ennek értelmező megállapításai különösen a kezdők számára hasznosak. Ugyanakkor épp a kezdők körültekintéssel kell hogy használják.

A demo mód több korlátozással is jár, pl. nem tudunk beolvasni vagy eredmény jelentést szerkeszteni. Ennek ellenére statisztikai szempontból a program így is teljes értékűnek tekinthető. Az első benyomás szerint ez a korszerű statisztikai programok számos sajátosságával rendelkezik, így a szokásos ablakkezeléssel, táblázatkezelő szerű adatmegjelenítéssel és ügyes menürendszerrel.

Egy minta jelentés egy egyszerű korrelációszámításról:

```
Pearson Product Moment Correlation Wednesday, November 29, 2000

Data source: Data 1 in Notebook

Cell Contents:
Correlation Coefficient
```

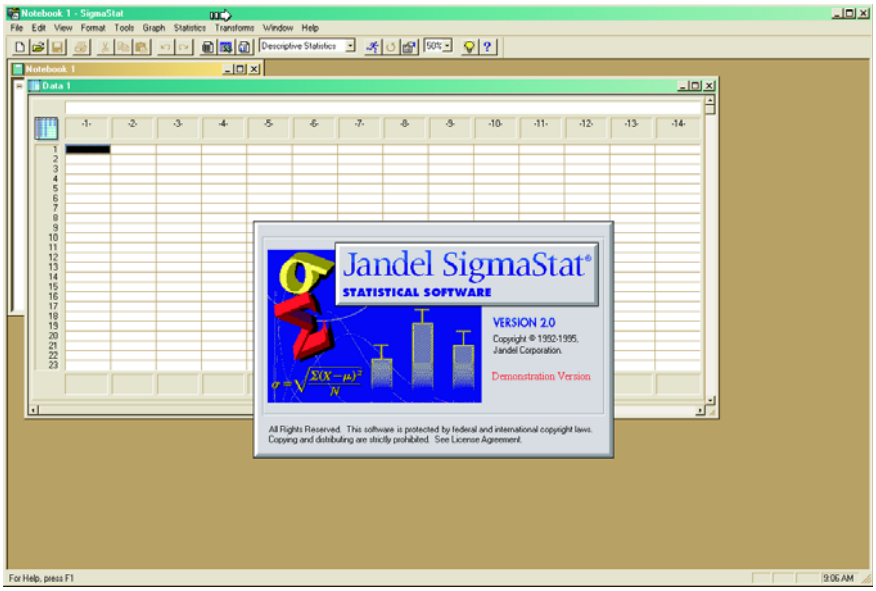
```
P Value
Number of Samples

                    Col 2    Col 3
Col 1               0.978    -0.973
                    0.000729 0.00110
                    6        6

Col 2                        -0.997
                             0.0000163
                             6

Col 3

The pair(s) of variables with positive correlation coefficients
and P values below 0.05 tend to increase together. For the
pairs with negative correlation coefficients and P values below
0.05, one variable tends to decrease while the other increases.
For pairs with P values greater than 0.05, there is no
significant relationship between the two variables.
```

A SigmaStat kezelését nem tárgyaljuk olyan részletességgel, mint az SPSS programét. Ennek ellenére az ott tanultak alapján, de magában a rendelkezésre álló súgó, Help parancsok segítségével, valamint a nagyon barátságos tanácsadóra (Advisor) támaszkodva elég gyorsan kitanulhatjuk a használatát. Ezért itt inkább csak az eltéréseket, a különlegességeket ismertetjük, a teljesség igénye nélkül.

A program elindítása után panaszkodhat, hogy egy magyar nyelvű modul nem érhető el, de ezután (OK után) egy angol nyelvűre támaszkodva helyesen működik. A program elindulása után az 5. ábrán látható kép fogad bennünket. A görgős egerek sajnos ezzel a programmal se használhatók.

3.1 Az adatállományok kezelése

A File menüsor lényegében megfelel a korszerű statisztikai programokénak. Így megnyithatunk egy új adatállományt (New, vagy Ctrl-N), egy korábban már létrehozottat (Open, Ctrl-O), vagy bezárhatjuk a jelenlegit (Close, F11). A különböző mentési funkciók nem működnek a demo változatban. Ez utóbbi komoly korlátozás, mert így minden esetben előről kell kezdenünk a feldolgozást.



5. ábra: A SigmaStat program nyitóablaka.

Az Import utasítás (Ctrl-I) viszont elérhető, és nagyon hasznos. Ez a SigmaStat program egy újítása volt korábban, ma már a statisztikai programok rendszerint rendelkeznek ilyen adatbeviteli lehetőséggel. Az elvárható SIGMA állományok (DOS és Windows változatok, különböző verziók, SigmaStat és SigmaPlot is) beolvasásán túl számos más program formátumát is támogatja. Ide tartoznak a táblázatkezelők egyes verziói (Excel, Lotus, Quattro, dBase), de egyszerű szöveges állományok is.

A szöveges állományok beolvasása ügyesen támogatott, öröm vele dolgozni. A valószínű adatszerkezetet a program felismeri, és felajánlja a felhasználónak — amin természetesen lehet módosítani. Ennek során megadhatjuk, hogy az adatokat rögzített oszloppozíciókban kell-e keresni (Fixed columns), vagy valamilyen elhatárolójelek határozzák meg az egyes változókba kerülő adatokat (White spaced, Tab delimited, Delimiter). A beolvasandó határokat fix oszlopok esetén egy egyszerű jelölés mutatja, pl. `[--]` jelenti, hogy az adott négy pozíció tartalmazza az érintett változó tartalmát. A teljes beolvasandó adatállományt átnézhetjük a párbeszédes ablakban, hogy az esetleges eltéréseket tisztázzuk. Az Import gomb megnyomása után az új adatot a kurzor által kijelölt cellától kezdve találjuk meg.

Az Export menüpont ismét ki van kapcsolva a demo változatban. A

Summary info az aktuális feldolgozás jellemzőit mutatja meg: szerző, cím, létrehozás dátuma stb. A Preferences menüpontban lényegében csak az adatformátumot lehet beállítani. A nyomtatás ismét működik, és a legutóbb használt adatállományok listája is kényelmesebbé teszi a munkát.

3.2 Szerkesztési utasítások

A szerkesztési utasítások egy része ki van kapcsolva a demo módban, így egyes cellákat ki tudunk ugyan jelölni, de ezekkel másolási, beillesztési stb. műveleteket nem tudunk végezni. Ezek különben természetesen elérhetők az Edit menüsorban, sok más szerkesztő utasítással együtt mint amilyen a törlés, beszúrási mód, cellák beszúrása és törlése és egy adott cellára való ugrás. Csaknem teljes értékűen szerkeszthető viszont a jelentés ablak, bele is írhatunk, formázhatjuk.

A View menüsor különböző megjelenítési parancsokat foglal össze. Ezek nagy része demo módban is hatásos. Így a különböző létrehozott ablakokra való ugrás (pl. Data 1 (F4), vagy Statistics-Data 1 (F6)). Az utóbbi nem csak megjelenít, hanem a változóinkra meg is határoza az alapvető leíró statisztikákat. A zoomolás ezzel szemben nem működik (az eszközsorból se, ahol különben jobbra egy kiugró menüben adhatnánk meg a zoomolás mértékét).

A Format menüsor utasításai a képernyőn való megjelenés formázására adnak lehetőséget (betűtípus, kiigazítás, sortáv, tabulálás). Ezek a jelentéseket tartalmazó ablakokban hatásosak. Az egyes oszlopok (változók) nevét megadhatjuk a Format / Column Titles paranccsal (F9) az adatablakban. Demo módban ki van kapcsolva a teljes Tools menüsor.

Szám szerint ezekből a demo módban el nem érhető utasításokból sok van, de ezek inkább csak a professzionális alkalmazásnál elengedhetetlenek. Ezzel szemben a ritkán, kisebb adaton végrehajtandó statisztikai eljárások végrehajtása lehetséges.

3.3 Az adatok átalakítása

Bár az adatok alakítása, transzformálása meg kell hogy előzze a tényleges statisztikai feldolgozást, a SigmaStat az SPSS-el szemben a Statistics menüsor után szerepelteti a Transforms menüt. Itt meglehetősen sok parancs érhető el, és ezek lényegében le is fedik az adatelőkészítéshez leginkább szükségeseket. Ahogy korábban már említettem, az adatok átalakítása,

módosítása a teljes statisztikai feldolgozás meglepően nagy részét jelenti tényleges gyakorlati feladatok esetén.

A SigmaStat által rendelkezésre bocsátott átalakító utaasítások a következők:

gyors átalakítások (Quick transforms, a gyorsító billentyűkombináció Alt r q),

a kiválasztott cellák rendezése (Sort selection, Alt r s),

indexelés (Indexing, Alt r i), ezen belül egy- és kétutas indexelés (One way, Two way),

az indexelés megszüntetése (Unindexing, Alt r u), ezen belül ismét egy- és kétutas indexelés (One way, Two way),

áttöltés (Stack, Alt r t),

a nulla átlagú eloszlás kialakítása (Center, Alt r c),

standardizálás (Standardize, Alt r a),

rangsorolás (Rank, Alt r r),

szorzat (Interaction, Alt r i),

segédváltozók (Dummy Variables, Alt r d),

késleltetett változók idősorok vizsgálatához (Lagged Variables, Alt r l),

szűrés (Filter, Alt r f),

véletlen számok (Random Numbers, Alt r n), egyenletes és normális eloszlással,

hiányzóadat kódok (Missing Values, Alt r m) és végül a

felhasználó által definiált (User-Defined, Alt r d).

Ezeknek az utasításoknak a használata egyszerű, bár kissé eltér más statisztikai programok szokásos eljárásától. Nézzünk egy példát: generáljunk normális eloszlású pszeudovéletlen számokat egy új változóba (oszlopba). A Random Numbers parancs kiadása és a Normal menüelem kiválasztása után egy párbeszédes ablakot kapunk, amely a bal oldalon megadja, hogy mi a teendőnk: válasszuk ki azt az oszlopot, amelybe az eredmény

kerül majd. Ezt az egyik táblázatbeli oszlop fejlécére való klikkeléssel tudjuk megtenni. Ha az ablak jobboldalán megadott beállításokkal elégedettek vagyunk, akkor nyomjuk meg a Finish gombot.

A következő ablakban a részleteket tisztázhatjuk: hány ilyen számot kérünk (Quantity), mi legyen a generált számok átlaga (Mean), mennyi legyen a szórás (Standard Deviation), és hogy mi legyen az indulóérték (Seed). Az utóbbinak az alapértelmezése az, hogy ez is egy véletlenszerű érték legyen, de itt megadhatunk valamely konkrét számot is. Ha az utóbbi utat választjuk, akkor a statisztikai kísérletünk megismételhető lesz, és többszöri végrehajtás esetén mindig ugyanazokat a véletlen számokat kapjuk.

A másik fontos, gyakran használatos parancs a felhasználó által definiált transzformáció. Ennek szerkesztési ablakába bonyolult képletet írhatunk, aminek a részletes ismertetésére itt most nincs hely. Minden esetre erről olvashatunk a program súgójában is, de a legegyszerűbb összefüggés a következő alakú kell hogy legyen: `col(3) = col(1)+col(2)`. Itt használhatunk számos matematikai függvényt is.

3.4 Statisztikai eljárások

Az adathalmazok vizsgálatára SigmaStat eljárásokat lehet meghívni. A megfelelő eljárás kiválasztását segítheti a tanácsadó (Advisor), amit a Help / Advisor menüponttal lehet elindítani, vagy az Alt h a billentyűkkel. Alapjában véve a legtöbb egyváltozós statisztikai eljárás rendelkezésre áll, de a bonyolultabb többváltozósak, mint pl. a klaszter analízis, vagy a többdimenzionális skálázás, hiányoznak.

A kiválasztott eljárás kéri, hogy adjuk meg azokat az oszlopokat (column), amelyekre a vizsgálat vonatkozik. Ezt az adatok transzformálásánál már bemutatott módon lehet megtenni: amikor az input vagy output változók megadását kéri az aktuális párbeszédes ablak, akkor a megfelelők fejlécére kattintva kell ezeket kijelölni.

Az eljáráshoz tartozó alapértelmezés tesztek automatikusan futnak (már amiket a felhasználó engedélyezett, illetve nem tiltott). Ha valamelyik feltevés nem teljesült, akkor alkalmas eljárásra kapunk javaslatot.

Az eljárás eredményéről jelentés (report) készül, amit a jelentésszerkesztővel átírhatunk, vagy a kimentés után (ASCII szöveget kapunk), bármely más szövegszerkesztővel feldolgozhatunk[17]. Az utoljára végrehajtott Sig-

[17] Már ha nem demo változatú a program, amivel dolgozunk.

maStat eljárást a Statistics / Run Current Test..., illetve a Statistics / Rerun Current Test parancsokkal lehet ismét futtatni.

A statisztikai eljárások opcióinak beállítására a Statistics / Current Test Options menüpont ad lehetőséget. Ahogy a parancs nevéből is kitűnik, mindig az aktuális, éppen végrehajtott teszt paramétereit tudjuk itt megváltoztatni. Az aktuálisnak tekintett teszt az eszközsorban is látható.

A rendelkezésre álló eljárások fő kategóriái, csoportjai:

- **Describe Data:** leíró statisztikák,
- **Compare Two Groups:** kétmintás próbák,
- **Compare Many Groups:** ANOVA tesztek, többmintás próbák,
- **Before and After:** önkontrolos kísérletek adatainak összehasonlítása,
- **Repeated Measures:** ismételt mérésekre vonatkozó ANOVA próbák,
- **Rates & Proportions:** csoportok eloszlásainak összehasonlítása (pl. férfi – nő esetek aránya, ...),
- **Regression:** az adatpontok adott függvénnyel való közelítése,
- **Correlation:** a változók közti lineáris összefüggés kimutatása,
- **Normality:** normalitásvizsgálat,
- **Power & Sample Size:** méréstervezés, a kívánt érzékenységhez szükséges mintaszám meghatározása.

Az itt említett almenükben további eljárások érhetők el. A legfontosabbak közül mintaként néhányat tárgyalunk röviden a következő szakaszokban.

3.4.1 Leíró statisztikák

Az opciók lehetnek: valamely mutató kiválasztása vagy kihagyása, a kvantilisek beállítása az alapértelmezés szerinti 25% és 75% helyett, a normalitási teszt szignifikancia-szintjének megadása (0,05 helyett) és a konfidencia-intervallumok specifikálása.

Az érintett változókat tartalmazó oszlopokat a szokásos módon kell kijelölni. A korábbi SigmaStat program változathoz képest itt tetszőleges oszlopokat jelölhetünk ki, nem kell azoknak egymás mellett lenniük. A kijelzett számjegyek számát a Statistics / Report Options menüpontban lehet beállítani. Normalitás: itt a Kolmogorov-Szmirnov teszt eredményét kapjuk. A magyarázó szövegeket is itt lehet ki-, vagy bekapcsolni.

3.4.2 Korreláció

A predikció regressziót és korrelációt használ két vagy több változó közti kapcsolat leírására. Kétféle korrelációs együttható van a SigmaStat rendszerben:

- a Pearson-féle 'product-movement' korreláció, amely a reziduumok normális eloszlását és konstans varianciáját tételezi fel a szignifikancia számításánál és a lineáris összefüggés szorosságát méri,
- a Spearman-féle rangkorreláció, ami egy nemparaméteres összefüggési teszt, amely nem igényli a fenti feltevéseket.

Ezeknek eljárásparaméterei nincsenek, a Statistics / Correlation almenüből hívhatók. Ezen túl természetesen több más statisztikai eljárás is ad korrelációs együtthatókat.

3.4.3 Regresszió

A regresszió egy vagy több magyarázóváltozó alapján jósolja meg az eredményváltozó értékét. A magyarázóváltozók az ismert, vagy prediktor változók. Ezek változása a függő-, vagy válasz változó megfelelő értékét eredményezi.

A regressziót a SigmaStat-ban hat modell szerint lehet végrehajtani (arra az esetre, ha ezek valamelyikének a használatát igényli a feladatunk):

1. egyszerű lineáris,
2. többszörös logisztikus,
3. többváltozós lineáris,
4. polinomiális,
5. lépésenkénti, mind előre, mind visszafelé,
6. legjobb részhalmaz és
7. nemlineáris.

Az adatokat egyszerűen változónként külön oszlopba kell csoportosítani. A regresszió figyelmen kívül hagyja az olyan sorokat, amelyekben hiányzó adat van. A regresszióhoz tartozó beállítási lehetőségeket a Statistics / Current Test Options menüpontnál lehet megtalálni.

Kérhetjük például a jósolt értékek megmutatását, az eltéréseket, a standard reziduumokat (ez a reziduum osztva a becslés standard hibájával, láthatóvá teszi a kilógó adatpontokat), a Student reziduumokat (ez figyelembe veszi azt is, hogy az adatok közepén jobb illeszkedés várható) és a Student törölt reziduumokat (ez kihagyja az aktuális adatpontot a regressziós egyenlet meghatározásából, a kiugró, outlier pontokat alapértelmezésben a regressziós populáció 95%-os konfidencia intervalluma határozza meg).

3.4.4 Nemlineáris regresszió

Akkor kell nemlineáris regressziót használni, ha az adatsor egy nemlineáris függvény görbéjét követi. A háttérben a Levenberg-Marquardt algoritmus dolgozik, amely valamely indulópontból megkeresi a legjobb illeszkedést biztosító paramétereket, amelyek legkisebb négyzetösszeg értelemben a legjobban megfelelnek. Az eljárás iteratív, és a konvergencia-feltétel teljesüléséig tart. Az algoritmus csak helyi minimumok megkeresésére alkalmas, tehát az abszolút vagy globális minimum becsléséhez több, különböző indulóértékkel való futás szükséges.

A feladat leírására használatos fogalmak:

Paraméterek: ezzel a névvel jelöljük az optimalizálandó paramétereket. De nem szabad ezeket a (más szóhasználatnak megfelelő) paraméteres feladatokra vonatkoztatni. Ezekre kezdőértéket kell majd megadnunk, és az eljárás ezeket javítja majd meg.

Változók: Itt kell megadni az eredmény- és a magyarázóváltozókat, amelyek a mért adatokat tartalmazzák, illetve az összegzés valamilyen indexelését adják meg. Legalább két változót meg kell adni — de szerepelhet itt súlyozó faktor is. Az alkalmazandó szintaxis: variable = col(szám). A változók neve tetszőleges, de ajánlott a rövid név. Legfeljebb 10 magyarázóváltozónk lehet.

Modellfüggvény: Itt határozzuk meg a magyarázóváltozók és a paraméterek összefüggését. Az összes magyarázóváltozót tartalmaznia kell. Több egyenlet is megadhatja az összefüggést, pl.:

f=if(x < 1,constant(x),line(x))
constant(x)=c
line(x)=a+b*x.

Illesztési utasítás: megadja, hogy mit mihez illesztünk. Megadható egy súlyváltozó is: fit f to y with weight w.

Optimalizálási feltételek: Megadhatunk olyan feltételeket, amelyek teljesülését előírjuk a megoldásra. Itt szerepelhetnek korlátok a paraméterekre, egyenlőségek vagy pl. valamely paraméter nemnegativitása is. Csak lineáris feltételt szabad megadni. A feltételek ellentmondásmentességét az eljárás ellenőrzi.

Opciók: a regresszió végrehajtását lehet velük szabályozni, a rendelkezésre állók:
iterations: az iterációk számának felső korlátja (alapértelmezésben ez 100),
stepsize: korlát a lépésközre (alaphelyzetben 100) és
tolerance: a megállási feltétel. Két megállási feltétel szokásos: megállás, ha két egymást követő iterációs lépésben a NORM változása kisebb, mint a megadott érték, illetve ha a paraméterek értéke nem változik a szignifikáns decimális helyeken (15 jegy).

A nemlineáris regressziós feladat megadásánál ';' után tetszőleges megjegyzéseket fűzhetünk az illető sorhoz. Az elkészült feladatleírást a Save gombbal elmenthetjük (de nem a demo módban).

Egy jellemző feladat megadása (ne felejtsük el adatainkat az első két oszlopba írni):

```
[Parameters]
a=120 ; range for mean arterial pressure
b=0.1 ; slope coefficient
c=90 ; sinus pressure at maximum rate of change
d=40 ; minimum mean arterial pressure
e=2

[Variables]
x=col(1)
y=col(2)

[Equations]
f=a/(1+exp(b*(x-c)))^e+d
fit f to y ; fit statement
```

[Constraints]
e > 0

[Options]
iterations=5

3.5 Grafikonok

A korábbi változatú (DOS alapú) SigmaStat program önállóan nem támogatta a grafikonok rajzolását, ezt a SigmaPlot programra bízta. A jelen változat a Graph menüsorbeli utasításokkal támogatja az eredmények megjelenítést. Ezek közül csak a közvetlen ábrarajzolás érhető el a demo változatban (az F3 gombra is indul), a grafikon részleteit beállító Graph Properties, és a Run SigmaPlot nem aktív ebben az esetben.

Ennek ellenére demo módban is a grafikonok összes változatát létrehozhatjuk — de nem lehet se kimenteni, se kinyomtatni. A grafikonok sokfélesége (16-féle) a táblázatkezelő programok által nyújtott ajánlatra emlékeztet, de azok kifinomult megoldásait itt nem tudjuk elérni. Mégis, a tényleges statisztikai megjelenítéshez elegendő támogatást kapunk.

A párbeszédes ablakok kezelése az előzőekhez hasonló, semmi különösre nem kell számítanunk. A grafikonokat külön ablakban fogjuk kapni.

3.6 Súgó

A SigmaStat súgója a program egyik erőssége, ahogy már korábban írtuk is. A ma már szokásos szerkezetben nyújt segítséget a program használatához, így van benne egy hipertext jellegű felhasználói leírás (SigmaStat Help), amely teljesértékű bevezetést ad angol nyelven. Ennek tartalma kereshető index vagy tartalomjegyzék alapján is. Az index egy egyszerű szószedet, nem egy megadott betűnként továbbugró kereső. Van egy magyar nyelvű segítő a súgó használatához (Help on Help).

További információkat találunk itt a technikai támogatásról (Technical support), a Microsoft Office-ra vonatkozó kompatibilitásról (Office compatible...) és a SigmaStat programról általában (About SigmaStat).

A leghasznosabb azonban a már többször említett statisztikai tanácsadó (Advisor). Ennek előnye, hogy kontextustól függő tanácsot ad, kiindulva a rendelkezésre álló adatból és az elvégzett tesztekből. A feltett kérdéseire rádiógombokkal lehet felelni, és a felhasználó céljának tisztázása után a

tanácsadó olyan statisztikai eljárásokból álló sort javasol, amely szakmailag korrekt kiértékelést tesz lehetővé. Ennek követése különösen a kezdő alkalmazó dolgát könnyíti meg. Ha már tudjuk, hogy mely tesztekre van szükség, akkor természetesen ezek közvetlenül is végrehajthatók. Érdemes a jelentések angol nyelvű szövegét is gondosan tanulmányozni, kiértékelni, ezen az úton indokolatlan következtetések levonását lehet elkerülni.

A tanácsadó közvetlenül, a menü nélkül is elérhető az eszközsorban lévő izzólámpát ábrázoló ikonnal. A súgó ikonja e mellett található, egy kérdőjelet tartalmaz.

3.7 Feladatok SigmaStat-tal való megoldásra

Az alábbi feladatokhoz készítsünk olyan file-t, amely a tesztekhez tartozó jelentések lényegét tartalmazza!

1. Olvassunk be egy adatfile-t az Import utasítással úgy, hogy a benne lévő adatok a táblázat megfelelő oszlopába kerüljenek! Javítsuk ki az adódó eltéréseket!

2. Generáljunk egy táblázatot a Transforms eljárások segítségével úgy, hogy két oszlop normális eloszlású, a köztük lévő különbség pedig szignifikáns legyen!

3. Kérjünk leíró statisztikákat egy adathalmazról, majd az elkészült jelentést szerkesszük át magyar nyelvre!

4. Generáljunk olyan adathalmazt, amelyben a 2. oszlop az elsőnek regresszióval jól leírható transzformáltja — de az összefüggés nem lineáris! Értelmezzük a jelentésben szereplő korrelációs együttható és hibanégyzet értékeket!

5. Generáljunk két normális eloszlást követő oszlopot úgy, hogy a köztük levő korreláció (kb.) -0,5 legyen!

6. Olvassuk be a MINTA.TXT állományt, hozzuk létre belőle a megfelelő táblázatot, és végezzük el rajta a Describe Data, Before and After nevű eljárásokat! (Mind a négy, az állományban lévő táblázat 3., 4., 6. és 8. oszlopai azonos jellegű mennyiségeket tartalmaznak kezelés nélkül, illetve különböző kezelések után.)

7. Generáljunk egy lineáris, polinomiális, illetve nemlineáris összefüggést jelentő adathalmazt kis véletlen, normális eloszlású zajjal terhelve, majd keressük meg az összefüggést regressziós módszerrel!

8. Adjunk meg egy olyan nemlineáris regressziós feladatot, amelynek több lényegesen különböző helyi optimuma van, majd illusztráljuk a SigmaStat lehetőségeit a feladat teljes megoldására!

A 6. feladatbeli MINTA.TXT file tartalma:

```
S5    4    0.17    0.16   94   0.17  100      0.17    100 \\
S7    4    0.25    0.24   96   0.24   96      0.25    100 \\
S10   4    0.35    0.35  100   0.35  100      0.36    103 \\
H3    3    0.48    0.22   46   0.21   44      1.07    223 \\
H6    6    5.84    3.66   63   3.09   53     16.30    279 \\
GP    2  558.57  449.94   81 474.15   85 81 089.76 14 517 \\
SHCB  2    0.07    0.07  100   0.06   86      0.09    129 \\
RCOS  2    0.04    0.04  100   0.03   75      0.04    100 \\
RB    2    0.01    0.01  100   0.01  100      0.01    100 \\
```

3.7.1 Egy dolgozat feladatai

1. Olvassa be (az import utasítással) a statzh file 5, számokat tartalmazó oszlopát (cím: `http://www.inf.u-szeged.hu/~csendes/statzh`)! Ügyeljen arra, hogy a számokon belül ne maradjon szóköz! Törölje az üres sorokat! 1 pont

2. Az adatokban talált 6789-as számokat nyilvánítsa hiányzóadat kódnak! 1 pont

3. Generáljon egy új (a 6.) oszlopba annyi standard normális eloszlású véletlen számot, ahány sor van a táblázatban! 1 pont

4. Képezze az 5. oszlop és a 6. oszlop 0,1-szeresének összegét, és tegye az eredményt a 7. oszlopba! Mentse el az eredmény táblázatot ZHA néven! 1 pont

5. Vizsgálja meg a 4. és a 7. oszlop lineáris függését az 1. oszlop adatától lineáris regresszióval! Az eredményekhez fűzzön magyar nyelvű rövid értelmezést, és mentse el őket ZHA1.RPT és ZHA2.RPT néven! 2 pont

Az első feladatban említett statzh nevű adatállomány tartalma:

Nap	BUX	MOL	MATAV	OTP	BP-Novekedesi
XI.1.	5489	4890	1279	10955	1.8789
XI.2.	6789	5120	1343	11034	1.8894
XI.3.	5834	5235	1355	11870	1.9292
XI.4.	5589	6789	1312	11345	1.9111
XI.5.	5309	5102	1298	6789	1.9034
XI.6.	5103	5023	1176	10766	1.8943
XI.7.	5302	5301	1212	11034	1.9102
XI.9.	5332	5395	1235	11203	1.9302
XI.10.	5401	5378	1320	11301	1.9422
XI.11.	5578	5567	1349	11650	1.9388
XI.12.	5632	6789	1377	11703	1.9401
XI.13.	5276	5423	1336	11329	1.9213
XI.14.	5303	5546	6789	11543	1.9324

4. Fejezet

StarOffice, Excel és As-Easy-As

A táblázatkezelő programok nem elsősorban statisztikai adatfeldolgozásra valók, de nagyon alkalmasak nagy mennyiségű numerikus és szöveges adat egységes kezelésére. Elterjedtségük miatt sok esetben mégis jól használhatók egyszerűbb statisztikai mutatók meghatározására, illetve néhány statisztikai teszt, eljárás végrehajtására. Az adatok grafikonon való megjelenítésében pedig gyakran lényegesen rugalmasabbak, több szolgáltatást kínálnak, mint a kimondott statisztikai programok.

4.1 StarOffice

A StarOffice a SUN terméke, ingyenesen letölthető az internetről (az ide vonatkozó cím: `http://www.sun.com/staroffice`). A program nem csak Windows operációs rendszerre érhető itt el, hanem Linux-ra és PC-s valamint munkaállomásra való Solaris operációs rendszerre is.

A következő megállapítások a StarOffice német nyelvű, PC-s 5.2 változatára vonatkoznak. Sajnos épp a függvény nevek az aktuális verzió függvényében eltérnek. Ennek ellenére ezek köre állandó, és az egyes egymást követő verziókban is csak kisebb módosítások várhatók. A nyelvenként változó függvényneveknek az az előnyük, hogy alkalmas nyelven könnyebben megjegyezhetők, illetve nem is kell őket megtanulni. A magyar nyelv sajnos nem támogatott, de számos nagy nyelv (összesen 12, közte az angol, német, spanyol, orosz és a francia) igen, így könnyen választhatunk egy alkalmasat.

A StarOffice táblázatkezelő programját elindítva a képlet bevitelét segítő szerkesztőablak mellett találjuk a a függvényvarázslót (Funktionsautopilot, a szumma jeltől balra). A statisztikai jellegű parancsok nagy része itt található, leszámítva a statisztikai megjelenítést szolgáló diagramvarázslót

(itt: AutoFormat Diagram), ami az objektum beszúrás ikonnal érhető el.

4.1.1 Statisztikai függvények

A függvények képletben való beszúrását támogató függvényvarázsló szokás szerint egy párbeszédes ablakban, interaktív módon segíti a statisztikai eljárások, képletek megfelelő használatát.

A függvények kategóriákba sorolva, csoportosítva is elérhetők, ezeket az első kis ablakban (Kategorie) láthatjuk. Az Alle minden elérhető függvényt megad, a Zuletzt verwendet az utolsó néhány használtat sorolja fel (ez gyakran sokat segít, ha egy konkrét alkalmazáson hosszabb ideje dolgozunk), a Mathematik a matematikai jellegűeket stb. Bennünket most a Statistik feliratú csoport érdekel.

A statisztikai függvények közé 77 tartozik. Ezeket az egyes cellák értékének meghatározásakor, a szerkesztés során használjuk. Ekkor tehát egy bizonyos munkalap egy cellájára mutatunk az egér bal gombjával, majd a szerkesztőablakban megadjuk azt a képletet, amely a cella tartalmát meghatározza. Ennek a képletnek lehet a része egy statisztikai függvény.

A rendelkezésre álló statisztikai függvények tehát a Funktion felirat alatti listában találhatók. Minden statisztikai függvényről egy rövid leírást kapunk az ablak közepén (a StarOffice aktuális nyelvén). Ez megadja az érintett függvény szintaxisát és egy pár soros magyarázatot, hogy mire való az.

Egy ilyen statisztikai függvényt a listában szereplő nevére való kétszeri kattintással jelölhetünk ki. Ekkor további részleteket kell tisztáznunk, így például a függvények argumentumát megadni. Eközben a kialakuló képlet látható az alsó kis ablakban, és a képlet értékét is láthatjuk az Ergebnis ablakban (vagy egy hibakódot, ha a képletet még nem lehet valami miatt kiértékelni).

Az elérhető statisztikai függvényeket az alábbi listában mind felsoroljuk, hogy ez alapján el lehessen dönteni, hogy érdemes-e a munkát a StarOffice-al kezdeni. A statisztikai függvények nagy száma miatt nincs mindegyikre részletes magyarázat.

Achsenabschnitt (tengelymetszet),
Anzahl (az argumentumlistában lévő számok száma),
Anzahl2 (az argumentumlista értékeinek száma),

B (a binomiális eloszlás szerinti valószínűsége egy kísérlet eredményének),

Betainv (a bétaeloszlás inverzének értékei),
Betavert (a bétaeloszlás értékei),
Binomvert (a binomiális eloszlás értékei),

Chinv (a khi-négyzet eloszlás inverzének értékei),
Chitest (a khi-négyzet függetlenségi próba),
Chivert (a khi-négyzet eloszlás értékei),

Exponent (az exponenciális eloszlás értékei),

Finv (az F-inverz eloszlás értékei),
Fisher (a Fisher-transzformáció végrehajtása),
Fisherinv (a Fisher-transzformáció inverzének végrehajtása),
Ftest (az F-próba),
Fvert (az F-eloszlás értékei),

Gammainv (a gamma-eloszlás inverzének értékei),
Gammaln (a gamma-függvény logaritmikus értékei),
Gammavert (a gamma-eloszlás értékei),
Gauss (a standard normális eloszlás értékei),
Geomittel (egy minta mértani átlaga),
Gestutztmittel (egy minta peremértékei nélküli középértéke),
Gtest (Gauss-próba, a kétoldali normalitásvizsgálat),

Harmittel (a minta harmonikus átlaga),
Hypgeomvert (a hipergeometriai eloszlás értékei),

Kgrösste (a k. legnagyobb érték a mintában),
Kkleinste (a k. legkisebb érték a mintában),
Konfidenz ($(1-\alpha)$-konfidencia intervallum a normális eloszlásra),
Korrel (a korrelációs együttható),
Kovar (a kovariancia),
Kritbiom (a binomiális eloszlás határértékei),
Kurt (egy eloszlás csúcsossága),

Loginv (az inverz lognormális eloszlás értékei),
Lognormvert (a lognormális eloszlás értékei),

Max (az argumentumlistában lévő legnagyobb szám),

Maxa (az argumentumlista legnagyobb eleme, a szöveget nullának tekinti),
Median (a minta mediánja),
Min (az argumentumlistában lévő legkisebb szám),
Mina (az argumentumlista legkisebb eleme, a szöveget nullának tekinti),
Mittelabw (a minta átlagos abszolút eltérése az átlagtól),
Mittelwert (a minta átlaga),
Mittelwerta (a minta átlaga, a szöveget nullának tekinti),
Modalwert (a minta módusza, leggyakoribb értéke),

Negbinomvert (a negatív binomiális eloszlás értékei),
Norminv (az inverz normális eloszlás értékei),
Normvert (a normális eloszlás értékei),

Pearson (a Pearson-féle korreláció),
Phi (a standard normális eloszlás értékei),
Poisson (a Poisson eloszlás értékéi),

Quantil (egy minta egy megadott kvantilise),
Quantilsrang (egy minta egy értékének százalékos rangja),
Quartil (egy minta kvartilise),

Rang (egy minta egy értékének a rangja),

Schätzer (a regressziós egyenes egy értéke),
Schiefe (egy eloszlás ferdesége),
Stabw (egy minta korrigált standard hibája),
Stabwa (egy minta korrigált standard hibája, a szöveget nullának tekinti),
Stabv (a minta standard hibája),
Stabva (a minta standard hibája, a szöveget nullának tekinti),
Standardisierung (egy véletlenszám átalakítása standard értékké),
Standnorminv (az inverz standard normális eloszlás értékei),
Standnormvert (a standard normális eloszlás értékei),
Steigung (a regressziós egyenes meredeksége),
Stfehleryx (a lineáris regresszió standard hibája),
Sumquadabw (a részminta átlagtól való teljes négyzetösszege),

Tinv (az inverz Student-eloszlás értékei),
Ttest (a t-próba),
Tvert (a Student-eloszlás értékei),

Varianz (egy minta korrigált szórásnégyzete),
Varianza (egy minta korrigált szórásnégyzete, a szöveget nullának tekinti),
Varianzen (a minta szórásnégyzete),
Varianzena (a minta szórásnégyzete, a szöveget nullának tekinti),
Variationen (az argumentumok ismétlés nélküli variációinak száma),
Variationen2 (az argumentumok ismétléses variációinak száma),

Wahrschbereich (egy intervallum diszkrét valószínűségei),
Weibull (a Weibull-eloszlás értékei).

Ezeken a kimondottan statisztikai függvényeken túl természetesen számos más, pl. matematikai függvény is felhasználható a statisztikai feldolgozás során. Ezekre itt most nem térünk ki, de ide tartozik azért a Zufallszahl, amely a [0, 1] intervallumban egyenletes eloszlással ad egy pszeudo-véletlen számot.

Példa

Írjuk be az első (A1) cellába, hogy az A2 cella tartalmának, és a 4, 23 számoknak a varianciáját kérjük: a szerkesztő sorba ekkor VARIANZ(A2, 4, 23) kerül. Mivel az A2 cella ekkor még üres, így az A1 tartalma 180,5 lesz (figyelem, itt tizedesvessző van, nem tizedespont!).

Ha most az A1 cella képletét kiterjesztjük az A2 cellára is (a bal egérgombbal megfogjuk az A1 cella jobb alsó sarkában található ún. kiterjesztő négyzetet, és azt lehúzzuk az A2 cellába), akkor az A1 tartalma 9386,58-ra változik, A2 pedig az előbbi 180,5 értéket kapja. Az utóbbira akkor kapunk magyarázatot, amikor a bal egérgombbal ráklikkelünk az A2 cellára. Ekkor a szerkesztőablakban megjelenik az erre érvényes képlet: VARIANZ(A3, 4, 23). Márpedig az A3 cella megint csak üres, ezért kaptuk az A1 korábbi tartalmát. Ha ezt az eljárást tovább ismételjük (kiterjesztve az előző cella képletét az A3 cellára), akkor az A1 cellában három darab kettős kereszt jelzi majd, hogy az illető számot a megadott helyen nem lehet megjeleníteni. Ezen az adott oszlop szélességének megnövelésével lehet segíteni: az oszlop jobb szélét a bal egérgombbal jobbra kell húzni.

4.1.2 Célértékkeresés

További fontos segítőeszköz a célértékkeresés (itt, németül Zielwertsuche). Ennek az a lényege, hogy egy a cellatartalmak összefüggését is tartalmazó

táblázatban megjelölhetünk egy cellát, amelynek a tartalmát egy általunk előírt értékre szeretnénk megváltoztatni úgy, hogy közben a meglevő összefüggések érvényben maradjanak.

Ennek eléréséhez jelöljük ki a megváltoztatandó cellát a bal egérgombbal, és kérjük az Extras / Zielwertsuche utasítást. A kapott párbeszédes ablak első sorában már be is lesz állítva a módosítandó cella koordinátája. Adjuk meg ezután a második sorban a kívánt értéket, a harmadikban pedig annak a kijelölttel összefüggő cellának a koordinátáját, amelyiknek a változását megengedjük (aminek a terhére a tervezett módosulás létrejöhet az összefüggések fenntartása mellett). A koordináták megadásához minden esetben elegendő az érintett cellára klikkelni. Ezután az OK gomb megnyomását követően a táblázatkezelő program a megfelelő átalakításokat elvégzi.

Összefüggőnek itt akkor nevezünk két cellát, ha az egyik képletében vagy közvetlenül szerepel a másik, vagy a definiáló képleteken keresztül el lehet jutni a független celláig. Célértékkeresésnek nyilván csak függő cella esetén van értelme. A függő cellákra a közvetlen összefüggést megadja pl. a függvényvarázsló is a Struktur nevű lapon, de egyszerűen az egyes cellákra klikkelve is láthatjuk a szerkesztőablakban, hogy ezek milyen képlettel, mely más celláktól függenek.

A célértékkeresés természetesen statisztikai függvényekkel megadott összefüggésekre is érvényes, ilyenekkel együtt is használható.

Példa

Adjunk meg egy oszlopban néhány érdemjegyet, majd klikkeljünk az ezután következő cellába, és nyomjuk meg a szerkesztőablak melletti $\sum$ feliratú gombot. Ekkor kellemes meglepetésként épp a kívánt adathalmaz jelölődik ki, és a szerkesztőablakban megjelenik előkészítve az a képlet, amely a jegyeket összeadja. Fogadjuk ezt el az ENTER gomb megnyomásával. Az érdemjegyek összege ekkor megjelenik az új cellában. Klikkeljünk most erre, és a szerkesztőablakban a képlet mögé írjunk oda egy / jelet és a begépelt jegyek számát. Ezzel megkapjuk a megadott jegyek átlagát.

Jelöljük ki most a legutóbb meghatározott cellát, és kérjük az Extras / Zielwertsuche utasítást. A keresett értékként írjunk be egy hármast, és módosuló cellának adjuk meg az elsőt. Az utasítás végrehajtása után csak az első és az utolsó cella tartalma változik meg, éspedig úgy, hogy az utolsó a megadott hármas számot kapja, az első pedig egy alkalmas számot (nem feltétlen tanulmányi jegynek megfelelőt, esetleg nem egész, vagy nem is

pozitív — a többi jegy függvényében).

4.1.3 Grafikonok

A grafikonok rajzolása a táblázatkezelő programok szokásos erőssége. A StarOffice is kényelmesen paraméterezhető, egy diagramvarázsló segíti a legkifejezőbb megjelenítés megtalálását. Az adatainkat a StarOffice rengeteg formátum felismerésével támogatja, így a fontosabb adatbázisprogramokét is és más táblázatkezelő programok saját adatformátumát is elfogadja.

Az adatok előkészítéséről érdemes tudni, hogy kész, feliratozott táblázatot könnyű ábrázolni. Ezért aztán ha csak lehet, ilyeneket alakítsunk ki először. Ezután jelöljük ki az adatokat tartalmazó táblázatot a feliratokkal együtt, majd kattintsunk a táblázat mellett balra lévő Objekt Einfügen feliratú, kördiagramot ábrázoló ikonra. Ezután egy oszlopdiagramot ábrázoló kurzor jelzi, hogy a program arra vár, hogy megmutassuk, hol helyezze majd el az ábrát.

A kapott párbeszédes ablakokkal a diagramvarázsló segít a grafikonok kialakításában. Az első az adatokat adja meg: ez a legtöbb esetben az előzetes kijelölésnek köszönhetően már helyesen ki van töltve. Ennek ellenére érdemes ellenőrizni. Az első helyen az adat tartománya szerepel (Bereich), ez alatt bejelölve, hogy az első sor, illetve az első oszlop feliratot tartalmaz-e.

A következő lapon (amelyet a Weiter gomb megnyomásával kaphatunk), a grafikon alaptípusát lehet kiválasztani. A jobboldali kínálatból kell klikkeléssel választani, és a kijelölt forma az adatainkra vonatkoztatva meg is jelenik előzetesen a baloldali ablakban. Azt, hogy ezen a mintán lássuk-e az ábra feliratait is, az alatta levő boxban lehet bejelölni (Textobjekte in Vorschau darstellen). Itt változtathatjuk meg azt is, hogy az adataink sorokban (Zeilen) vagy oszlopokban (Spalten) tartoznak-e össze. A következő ablakban alvariánsokból lehet egyre rámutatni, és a rácsvonalakat kérni a tengelyek mentén (Gitternetzlinien).

Az utolsó oldalon aztán a feliratokat véglegesíthetjük: a grafikon főcímét (Haupttitel) és az egyes tengelyek elnevezését (Achsentitel). Ha az utóbbiakat kérjük, akkor ne felejtsük el a felirat előtti kis négyzetbe a pipát kitenni. A különböző színnel jelölt változóink magyarázófeliratát is kérhetjük (Legende hinzufügen).

Ha az elkészült ábra mutatott mintájával elégedettek vagyunk, akkor a Fertig stellen gombot nyomjuk meg, különben pedig a Zurück gombbal

lehet korábbi ablakok beállításain változtatni.

Példa

Jelenítsük meg a következő táblázatot egy alkalmas grafikonon!

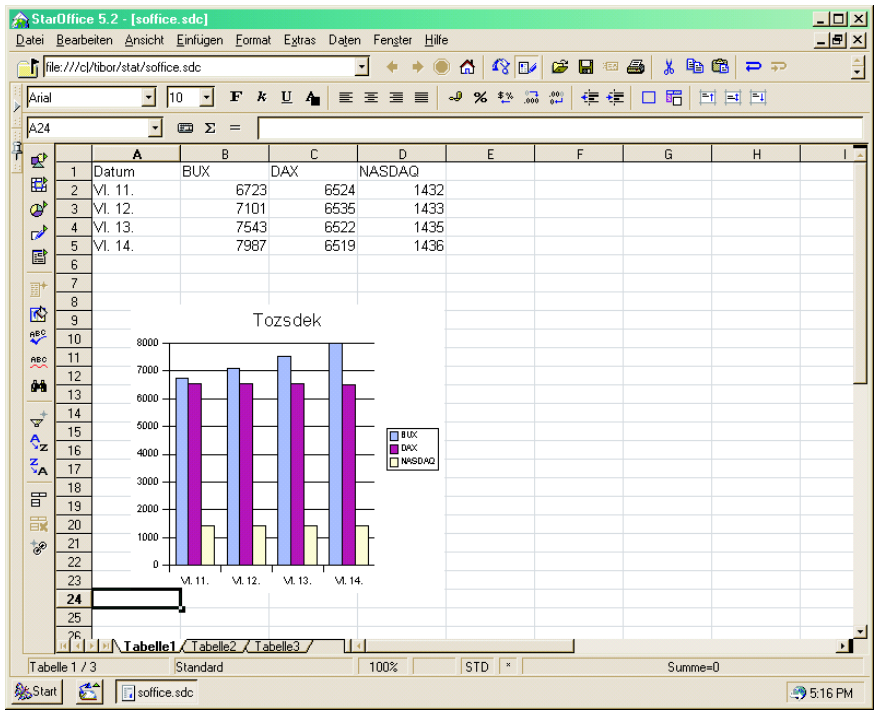
Datum	BUX	DAX	NASDAQ
VI. 11.	6723	6524	1432
VI. 12.	7101	6535	1433
VI. 13.	7543	6522	1435
VI. 14.	7987	6519	1436

Ezt azzal kezdjük (ami más hasonló esetben is nagyon hasznos), hogy a bevitt táblázatot először is kiemeljük úgy, hogy a bal egérgombbal rákattintunk a bal felső cellára, és lenyomott gombbal elvisszük a kurzort a jobb alsó celláig.

Ezután kattintsunk a bal egérgombbal a táblázat bal oldalán található Objekt Einfügen feliratú, kördiagramot ábrázoló ikonra. Ekkor a kurzor alakja megváltozik (egy grafikont mutat), és a program arra vár, hogy megmutassuk, hogy hova kerüljön a készülő ábra bal felső sarka. Miután egy cellába klikkeltünk, megjelenik egy párbeszédes ablak.

Vegyük észre, hogy a program kitalálta, hogy honnan vegye az adatokat (Bereich, tartomány), és azt is, hogy mind az első sort, mind az első oszlopot feliratot tartalmazóként kell kezelnie. A Weiter (tovább) gomb megnyomása után kapott ablakban kiválaszthatjuk az ábra típusát, legyen ez a harmadik. A következő ablakban ne változtassunk semmit, és az utolsóban megadhatjuk, hogy mi legyen az ábra címe (Diagramtitel), valamint a tengelyek felirata (X-Achsentitel, Y-Achsentitel). Ezeket a feliratokat az előttük álló kis kockában lehet ki-bekapcsolni. Hasonló módon kérhetjük a magyarázó felirat megadását vagy elhagyását (Legende hinzufügen).

Az eredmény a 6. ábrán látható.



6. ábra: A StarOffice-al kapott ábra, amely a példában megadott tőzsde-indexek alakulását mutatja.

Gyakorlat

Gyakorlásként először a példában megadott egyszerű kis adathalmaznak hozzuk létre más, kifejező grafikus megjelenítéseit. Ügyeljünk arra, hogy a táblázat minden szöveges információja kerüljön fel az ábrára, és hogy az összehasonlítható mennyiségek jól összevethetők legyenek.

Második feladatként keressünk ki egy gazdasági lapban megjelent profi grafikont, és igyekezzünk azt minél jobban megközelíteni a StarOffice által rendelkezésre bocsátott eszközökkel.

4.2 Excel

Az Excel nevű táblázatkezelő program is statisztikai függvényeket kínál adataink kiértékeléséhez. Az előző szakaszban tárgyalt StarOffice-hoz

hasonlóan ez is rendelkezik célérték kereső eljárással. Mind a statisztikai függvények, mind a célérték keresés használata lényegében megegyezik. Itt ezért inkább csak az eltérésekre hívjuk fel a figyelmet. Az elérhető statisztikai függvények köre és használatuk módja lényegében megegyezik a StarOffice-nál ismertetettel. Itt az ilyen függvényeket egy f_x feliratú ikonnal tudjuk elérni. A diagramvarázslót itt egy oszlopdiagramot ábrázoló ikonnal lehet hívni. Az ábrarajzolás is hasonlóan megy, mint az előző esetben, itt talán kicsit több választási, beállítási lehetőségünk van, és egy kész ábrára klikkelve a diagramvarázsló jelentkezik, hogy a kért módosításokat kényelmesen megtehessük.

Az Excel nem ingyenes program, de előnye, hogy nagyon elterjedt, sok helyen elérhetjük, és gyakran már az új számítógéppel együtt szállítják. Van magyar nyelvű változata is, és különösen magyar nyelvű operációs rendszerrel együtt ez idegennyelvismeret nélkül is könnyen használható.

Részben az elterjedtségéből adódik, hogy számos kiegészítő modult találhatunk hozzá, amelyek az alkalmazási körét lényegesen bővítik. Egy alaphelyzetben is elérhető bővítménye az ún. Solver eljárás. Ez a célértékkeresés kiterjesztésén túl (több megváltozó cellát enged meg és a cellatartalmakra feltételeket is megadhatunk) minimalizálni és maximalizálni is tud, akár nemlineáris összefüggéseket is — persze az induló értékek függvényében a megoldások nem mindig meg bízhatók. A Solver eljárás nem mindig található meg az Extrák menüsorban. Ha ez lenne a helyzet, akkor kérjük a bővítménykezelő utasítással ennek a beillesztését. Az esetek többségében ez problémamentesen megtörténik.

Gyakorlat

Ismételjük meg a StarOffice-nál leírt gyakorlatot azzal, hogy figyeljük meg az Excel eltéréseit.

A Solver lehetőségeivel való ismerkedést kezdjük azzal, hogy egy olyan célértékkeresést kérünk, ahol több cella megváltozását is megengedjük, miközben a feltételek között olyanokat adunk meg, amelyek csak egyetlen megoldást engednek meg: $z = x + y$, $x \leq 100$, $y \leq 200$, és z értékét kérjük, hogy a Solver állítsa 300-ra. Ügyeljünk rá, hogy az x és y cellák kezdeti tartalma kisebb legyen a határértéknél!

A minimalizálási és maximalizálási lehetőségek illusztrálására tekintsünk egy (a posteriori) befektetési döntéselőkészítést. Adjuk meg három részvény árfolyamértékeit néhány napra. Ezután három cellában adjuk meg azt, hogy kiindulásként hány darabot veszünk az egyes részvényekből (ez

lesz a portfólió összetétele). Ezután határozzuk meg a portfóliónk értékét minden egyes napra, amelyre a részvények árfolyama ismert.

Amikor ezzel kész vagyunk, paraméterezzük úgy a Solvert, hogy maximális legyen a portfóliónk értéke az utolsó napon, miközben előírjuk a feltételekben, hogy a kiindulási befektetett tőke pontosan 1 000 000 Forint legyen, és hogy minden részvényből nemnegatív mennyiséget lehet csak venni.

Magyarázzuk meg, hogy a legtöbb esetben miért csak egy részvény vásárlását javasolja a program. Figyeljük meg, hogy a nemnegativitási feltétel elhagyása milyen következményekkel jár.

4.3 As-Easy-As

A Bevezetésben említett As-Easy-As táblázatkezelő programban a következő néhány statisztikai függvény érhető el:

@AVG: átlag,

@AVEDEV: az átlagtól való átlagos eltérés,

@COMB: a kombinációk száma,

@COUNT: az elemek száma,

@COV: relatív szórás,

@COVAR: kovariancia,

@FACT: faktoriális,

@GAUSS: a (Gauss-, illetve) normális eloszlás eloszlásfüggvényének értéke,

@KURT: a csúcsosság értéke,

@MAX: a maximális érték,

@MEDIAN: a medián értéke,

@MIN: a legkisebb érték,

@PERM: a permutációk száma,

@POISSON: a Poisson eloszlás eloszlásfüggvényének az értéke adott x helyen,

@SKEW: a ferdeségi (aszimmetriai) együttható,

@PSUM: két adatsor szorzatainak összege,

@RAND: 0 és 1 közötti egyenletes eloszlású véletlen számot ad,

@STD: a szórás (N-el súlyozva),

@STDU: a korrigált szórás ($N-1$-el súlyozva),

@SUM: összeg,

@VAR: szórásnégyzet és

@VARU: a korrigált szórásnégyzet ($N-1$-el súlyozva).

A statisztikai eljárások közül rendelkezésre áll a többváltozós lineáris és nemlineáris regresszió, a célértékkeresés (itt GoalSeek) és számos pénzügyi függvény. Az említett statisztikai függvények és eljárások használatát angol nyelvű súgó segíti.

Irodalomjegyzék

[1] A Klinikai Biostatisztikai Társaság Jubileumi Évkönyve, 2001. `http://kbt.zona.hu`

[2] Boda Krisztina, Eller József, Nyári Tibor: Biostatisztika oktatási segédlet. Szeged, 1998-99.
`http://www.szote.u-szeged.hu/dmi/educat/stat/index.htm`

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

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

[5] 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.

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

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

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

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

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

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

Az irodalomjegyzék persze nem teljes, csak néhány fontosabb, illetve a tananyaghoz közvetlenül kapcsolódó könyvet, jegyzetet adunk meg. Magyar nyelven is elég kiterjedt szakirodalom érhető el, a legtöbb könyvtárban kulcsszó alapján ezek megtalálása minden nehézség nélkül megoldható.

Az itt megadott a könyvek, jegyzetek kiegészítésül ajánlhatók jegyzetünkhöz, de a félév anyagához nem feltétlenül szükségesek. A matematikai alapok alapos megismeréséhez Tandori Károly két jegyzetét [10, 11] javaslom. Az SPSS program korábbi változatait kicsit más megközelítésben tárgyalja Ketskeméty László és Izsó Lajos oktatási segédkönyve [6]. Tudomásom szerint jelenleg ez az egyetlen magyar nyelvű ismertetés az SPSS-ről. Ennek a programnak a teljes dokumentációja számos kötetből áll, az átlagos felhasználó legtöbb problémájára megoldást tud találni a felhasználói leírásban [9]. Ez a jelenlegi szerződések alapján minden hazai egyetemen legalább egy példányban elérhető, a Szegedi Tudományegyetemen például a Számítóközpontban. Az általános statisztikát és alkalmazásait tárgyalják például az [3, 4, 7, 8] könyvek, illetve jegyzetek. Az irodalomjegyzékben speciális szerepet tölt be az [5] ismerettár: ez szótárszerűen össszefoglalja a legfontosabb statisztikai fogalmakat képletekkel, rövid magyarázattal és az Excel valamint SPSS használatát segítő útmutatóval.

Tárgymutató

adatbevitel SPSS-be 23
    szöveges állományból 29
    adatbázis állományból 32
adatok mentése SPSS-ben 35
adattípusok 10, 25
alternatív hipotézis 18
arányskála 11
As-Easy-As 137

általános statisztika 8
átlag 13
    számtani átlag 13

bináris változó 10
binomiális eloszlás 15

cellaformátumok 27
centrum 92
cenzorált minta 12
célértékkeresés 132
célfüggvény 87
cimke 24
comma adattípus 25
custom currency adattípus 26

csúcsosság 14

date adattípus 25
dichotom változó 10
diszkrét valószínűségi változó 9

egyenletes eloszlás 15
egyoldali próba 18
elemszám 12
eloszlásfüggvény 14
elsőfajú hiba 18
eltérésfüggvény 91
eset 8
érvényes esetek száma 12
Excel 135

faktoranalízis 99
ferdeségi együttható 14
folytonos valószínűségi változó 9
függetlenség 12

globális optmalizálás 87
görbeillesztés 85

hasonlósági mérték 80
hiányzóadat kód 12
    SPSS-ben 26
hierarchikus klaszterezés 95
hisztogram 12

intervallum skála 11

kapcsolt rang 53
kétoldali próba 18
khi-négyzet eloszlás 16
K-közép klaszterezés 92
klaszter 91
konfidencia intervallum 17
korreláció 19
középérték 13
különbségi skála 11

kvalitatív adattípus 10
kvantilis 14
kvantitatív adattípus 10

lapultság 14

matematikai statisztika 8
másodfajú hiba 18
medián 13
megbízhatósági intervallum 17
megbízhatósági szint 17
megbízhatósági tartomány 17
mérési skálák 10
    SPSS-ben 27
minta 8
a minta eloszlása 12
módusz 13

nem szignifikáns 17
nemlineáris regresszió 87
nemparaméteres eljárás 16
névleges skála 10
nominális skála 10
normális eloszlás 12, 16
nullhipotézis 17
numeric adattípus 25

ordinális skála 11

paraméter 16
    paraméteres módszer 16
    paraméteres próba 16
paraméterbecslés 84
percentilis 14
Poisson-eloszlás 15
populáció 12

rangskála 11
regresszió 20
relatív szórás 14
reprezentativitás 12

scientific notation adattípus 25
SigmaStat 113
sokaság 12
Solver 136
sorrendi skála 11
SPSS 21
    File menüsor 34
    Window menüsor 58
standard hiba 14
standard normális eloszlás 16
StarOffice 127
    célértékkeresés 132
    grafikonok 133
    statisztikai függvények 128
statisztika 8
    matematikai statisztika 8
    statisztikai becslés 17
    statisztikai függvény 8
    statisztikai minta 8, 12
    statisztikai próba 17
string adattípus 26
sűrűségfüggvény 14

szabadságfok 17
szignifikancia 17
    szignifikancia-szint 17
    szignifikáns eltérés 17
szórás 13
szórásnégyzet 14
szöveg beolvasása az SPSS-be 29

távolságfüggvény 91
terjedelem 14
többdimenzionális skálázás 103

valószínűség 8
valószínűségi változó 8
változók típusa 25
várható érték 14

Magyar-angol szószedet

Itt a leggyakoribb szakkifejezéseket gyűjtöttem össze azok angol nyelvű változatával. Ez remélhetőleg segít majd az inkább az angol kifejezéseket ismerőknek, és fordítva, megkönnyíti majd az angol szakszöveg, illetve programutasítások olvasását, felismerését azoknak, akik nem ismerik az angol szakirodalmat.

Az nehezíti a dolgot, hogy valószínűleg marketing okokból a statisztikai programcsomagok írói mintha tudatosan igyekeznének egyéni szóhasználatot kialakítani, néhány alapvető kifejezés kivételével a programok gyakran más-más szót használnak ugyanarra a fogalomra (talán úgy gondolják, hogy így az egyszer megszerzett felhasználójuk nehezebben tér át más programra).

adatbázis	database
adathiányos eset	missing case
arány skála	continuous scale, ratio scale
aszimmetria	skewness
ábra	graph
átlag	mean, average
centrum	center
célérték keresés	GoalSeek
célfüggvény	loss function, objective function
csoport	group, class
csökkenő	descending
csúcsosság	kurtosis
diagram	chart, plot, graph

egyenesillesztés	linear regression
együttható	coefficient
eloszlás	distribution
eltérés	dissimilarity, deviation
eltérésfüggvény	dissimilarity measure
eset	case
exponenciális	exponential
érték	value
értékcimke	value label
érvényes esetek	valid cases
fadiagram	dendrogram
faktoranalízis	factor analysis
ferdeség	skewness
független változó	independent variable
függő változó	dependent or response variable
görbeillesztés	curve fitting
gyakoriság	frequency
hasonlóság	similarity
hiba	error
hisztogram	histogram
illeszkedés	fit
illesztési utasítás	fit statement
intervallum	interval
jelentés	report
jósolt értékek	predicted values
khi-négyzet próba	chi-square test
kiugró értékek	extremes, outliers
klaszter	cluster
konfidencia intervallum	confidence interval
korlátozás nélküli feladat	unconstrained problem
korlátozó feltétel	constraint

korrelációs együttható	correlation coefficient
köbös	cubic
kördiagram	pie chart
lapultság	kurtosis
legjobb részhalmaz regresszió	best subset regression
leíró statisztikák	description statistics
lépés	step
lépésenkénti módszer előre	stepwise forward method
lépésenkénti módszer visszafelé	stepwise backward method
lineáris regresszió	linear regression
magyarázó változó	predictor, explanatory variable
maximum	maximum (plural: maxima)
medián	median
mérési skála	measurement scale
minimum	minimum (plural: minima)
minta	sample
mintaelemszám	sample size
modell függvény	model function, equations
módusz	mode
nemlineáris	nonlinear
négyzetes	quadratic
négyzetösszeg	sum of squares (SS)
névleges vagy nominális skála	nominal scale or observations in categories
növekvő	ascending
optimalizálási feltételek	constraints
ordinális skála	ordinal scale, rank scale
oszlopdiagram	bar chart
osztályközép	midpoint
osztott kördiagram	pie chart
összeg	sum, total
paraméter	parameter
percentilis	percentile

p-érték	p-value
polinomiális regresszió	polynomial regression
pontdiagram	scatter plot
predikció	prediction
próba	test
rangskála	ordinal scale, rank scale
regresszió	regresion
regressziós egyenlet	regression equation
reziduum	residual
sorrendi skála	ordinal scale, rank scale
standard hiba	standard error or standard error of the mean
statisztika	statistics
Student reziduum	studentized residual
Student törölt reziduum	studentized deleted residuals
súly	weight
szabadságfok	degree of freedom
szórás	standard deviation
szöveges adat	text data, string
távolság	distance
távolságfüggvény	distance function
terjedelem	range
többváltozós lineáris	multiple linear
t-statisztika	t-ratio
valószínűségi változó	variable
variancia-analízis	analysis of variance (ANOVA)
variancia növelési faktor	variance inflation factor
variációs koefficiens	coefficient of variation
válasz változó	dependent or response variable
változó	variable
változókra vonatkozó korlát	bound constraint
vonaldiagram	line chart

Tartalomjegyzék

Előszó **3**

Jelölések **5**

1. Bevezetés **7**
1.1 Statisztikai alapfogalmak 8
1.1.1 Változótípusok 9
1.1.2 Adattípusok 10
1.1.3 Mérési skálák 10
1.1.4 A minta jellemzői 12
1.1.5 Eloszlások 14
1.1.6 Az eloszlásokkal kapcsolatos alapfogalmak 16
1.2 Statisztikai próbák 17
1.2.1 A statisztikai próbákkal kapcsolatos további alapfogalmak 18
1.2.2 Statisztikai próba végrehajtása 18
1.2.3 Változók összefüggése 19
1.2.4 Egy dolgozat feladatai 20

2. Az SPSS programcsomag **21**
2.1 Az adatok bevitele 23
2.1.1 A változók beállításai 24
2.1.2 Gyakorlat 27
2.2 Az adatok kimentése 28
2.3 Adat beolvasása szöveges állományból 29
2.3.1 Gyakorlat 31
2.4 Adat beolvasása adatbázis állományból 32
2.4.1 Gyakorlat 33
2.5 A File menüsor további utasításai 34
2.5.1 Új file 34

2.5.2 Korábban létrehozott állomány megnyitása 35
2.5.3 Adataink mentése 35
2.6 A szerkesztési és a nézet menüsor 36
2.6.1 A szerkesztési menüsor 36
2.6.2 Beállítások 36
2.6.3 A nézet menüsor 37
2.6.4 Gyakorlat 38
2.7 Az adatok beállításai, rendezése és mozgatása 38
2.7.1 Gyakorlat 46
2.8 Az adatok módosítása 49
2.8.1 Új változó kiszámítása 50
2.8.2 Előfordulások megszámlálása 50
2.8.3 Újrakódolás 51
2.8.4 Változók kategorizálása 53
2.8.5 Az esetek rangsorolása 53
2.8.6 Automatikus újrakódolás 54
2.8.7 Idősorok létrehozása 55
2.8.8 A hiányzóadat kódok pótlása 55
2.8.9 Gyakorlat 56
2.9 A statisztikai eljárások 58
2.10 Jelentések és Leíró statisztikák 59
2.10.1 Eset összefoglalás 60
2.10.2 A Gyakoriságok (Frequencies) parancs 61
2.10.3 Példa 63
2.10.4 Gyakorlat 65
2.11 Az output ablakok tartalmának szerkesztése 66
2.11.1 Az output ablak tartalmának nyomtatása 68
2.12 Táblázatok 71
2.12.1 A Statistics... nyomógomb párbeszédes ablaka 72
2.12.2 További beállítások 73
2.12.3 Példa 74
2.13 Átlagok összehasonlítása 75
2.13.1 Párosított t-próba 75
2.13.2 Példa 78
2.13.3 Gyakorlat 79
2.14 Korreláció 80
2.14.1 Páronkénti korreláció 80
2.14.2 Távolságok 82
2.15 Regresszió 84

2.15.1 Paraméterbecslés 84
2.15.2 Nemlineáris regresszió 87
2.16 Osztályozás 91
2.16.1 A K-közép klaszterezés 92
2.16.2 Hierarchikus klaszterezés 95
2.17 Az adat tömörített jellemzése 99
2.17.1 Faktoranalízis 99
2.18 Skálázás 103
2.18.1 Többdimenzionális skálázás 103
2.19 Grafikonok 106
2.19.1 Példa 107
2.20 Feladatok 109
2.20.1 Egy dolgozat feladatai 109
2.20.2 Egy másik dolgozat feladatai 110

3. SigmaStat eljárások 113
3.1 Az adatállományok kezelése 114
3.2 Szerkesztési utasítások 116
3.3 Az adatok átalakítása 116
3.4 Statisztikai eljárások 118
3.4.1 Leíró statisztikák 119
3.4.2 Korreláció 120
3.4.3 Regresszió 120
3.4.4 Nemlineáris regresszió 121
3.5 Grafikonok 123
3.6 Súgó 123
3.7 Feladatok SigmaStat-tal való megoldásra 124
3.7.1 Egy dolgozat feladatai 125

4. StarOffice, Excel és As-Easy-As 127
4.1 StarOffice 127
4.1.1 Statisztikai függvények 128
4.1.2 Célértékkeresés 131
4.1.3 Grafikonok 133
4.2 Excel 135
4.3 As-Easy-As 137

Irodalomjegyzék 139

Tárgymutató 140

Magyar-angol szószedet **143**

Tartalomjegyzék **147**