

Futó Péter

SZÉKELYI MÁRIA–BARNA ILDIKÓ:

TÚLÉLŐKÉSZLET AZ SPSS-HEZ.

Többváltozós elemzési technikákról társadalomkutatók számára (Budapest: Typotex, 2002)

BKÁE Szociológia és Szociálpolitika Tanszék
H-1093 Budapest, Fővám tér 8.; e-mail: futo@freemail.hu

A többváltozós statisztikai elemzés viszonylag fiatal tudományág, elméleti apparátusának java részét mintegy két nemzedékkal ezelőtt dolgozták ki. Napjainkra kiderült, hogy nincs értelme elkülönült módon beszélni a szociometria, a pszichometria, a kvantitatív piackutatás, a biometria és a technometria módszereiről, mert amelyik modell bevált az egyik területen, az a másikon is használható, ami viszont divathóbortnak minősült az egyik alkalmazásnál, az valószínűleg máshol sem állja meg a helyét. Kialakult egy univerzális empirikus segédtudomány, melynek terjedéséhez nagyban hozzájárult a számítástechnika fejlődése. Míg 15-20 évvel ezelőtt még vezető kutatók is csak nehezen férhettek hozzá a főleg nagy számítógépeken futó elemző szoftverekhez, addig ma már Magyarországon is egyetemek, kutatóhelyek és diákok számítógépeinek sokaságán lehet akár tízezrekre kiterjedő kérdőíves felméréseket elemezni.

Az SPSS programcsomag az utóbbi másfél évtizedben Magyarországon lekörözte vetélytársait, és a többváltozós statisztikai elemzés leggyakrabban alkalmazott eszköze lett a társadalomkutatás, a kísérleti lélektan, a piac- média- és közvélemény-kutatás, sőt a gyógyászati elemzések területén is. SPSS fájlokban nemcsak egyének vagy háztartások válaszait rögzítik. A gazdaságtudomány is sokszor támaszkodik a cégek ezreire kiterjedő megkérdezésekre, a politológia pedig az önkormányzatok százaira kiterjedő felmérésekre, és ma ehhez a legtöbben ezt a szoftvert használják. Bár az empirikus felmérések és a kísérleti adatok feldolgozásának műhelyeiben általában sajátos belső szakmai tolvajnyelvek alakultak ki, és ez gyakran egymáshoz közel álló szakterületeket is elválaszt egymástól. Ilyen körülmények között az SPSS által szabványosított módszerek és az egymásnak küldözgetett SPSS fájlok mintegy tolmácsként működnek az egyes oktató- kutató- és tanácsadó intézmények között. A sikerhez hozzájárult, hogy a szoftver gyártói idejében felismerték: már a korábbi verziók is sokkal szélesebb elemzési eszköztartalmat tartalmaztak, sokkal több fajta statisztikai szolgáltatást nyújtottak, mint amire az átlagos empirikus kutatóknak szüksége lenne. Ezért az utóbbi évtizedben az egymást követő verziók elsősorban - bár nem kizárólag - olyan irányban léptek előre, hogy használatuk egyre inkább felhasználó-barát, kezelésük egyre inkább Windows-kompatibilis legyen.

A felsőfokú társadalomtudományi oktatásba mára mélyen beépült az SPSS használatára való felkészítés. Ugyanakkor a szoftver használata sokkal mélyebb statisztikai tudást feltételez, mint amilyen a tipikus szociológushallgató rendelkezik, amikor megismerkedik ezzel a szoftverrel. Mivel a program diákverziójának elterjedésével

egyre több PC-n fut SPSS, ezért ma még nagyobb szükség van olyan kézikönyvekre, amelyek pusztán elemi statisztikai tudásra építve mutatják meg: hogyan kell felmérést tervezni, és ennek megvalósítása után elkészíttetni, majd értelmezni, megszólaltatni az SPSS output táblázatait, külsők számára is érthetővé tenni, hogy mit is üzennek a felmérésekből származó adathalmazok. De az SPSS-t tanuló diákokat idáig nem kényeztették el magyar nyelvű szakirodalommal. A cég prospektusain túlmenően az egyetlen használható kiadvány a Dr. Ketskeméty László és Dr. Izsó Lajos által írt *Az SPSS for Windows programrendszer alapjai* című felhasználói útmutató és oktatási segédlet volt.

A Túlélőkészlet az SPSS-hez hézagpótló tanulási és oktatási segédlet, elemzési esettanulmányok rendszerezett gyűjteménye, amely többnyire valós, de néha didaktikai célból kreált, az Internetről letölthető adathalmazokon mutatja be a kiválasztott eljárásokat. Az ismertetés három úgynevezett adatredukciós módszerre: a *főkomponens- a faktor- és a klaszterelemzésre*, valamint hat úgynevezett magyarázó modellre: *a variancia-analízisre, a lineáris regresszió-számításra, az útmodellekre, a diszkriminancia-analízisre, a többdimenziós skálázásra és a logisztikus regresszióra* terjed ki.

A könyv sokkal több és egyben sokkal kevesebb, mint amit a címe sugall. Azért több, mert az itt tárgyalt statisztikai elemző módszerek valójában nincsenek semmilyen konkrét szoftverhez kötve. Az input adatbázisok, a számítógépes parancsok és az outputok formailag az SPSS szabványait követik, de a piacon kapható más szoftverek is alkalmasak lehetnek ugyanezeknek a modelleknek a felépítésére, illesztésére, ugyanezeknek a szociológiai kérdésfeltevéseknek a megválaszolására. Olyan konkurens szoftverek, mint a SAS, a STATA és a MINISTAT más szintakszissal ugyan, de a legtöbb többváltozós módszerre kiterjednek. Sőt, a leggyakrabban alkalmazott modellek – a lineáris regresszió-számítás és a variancia-elemzés – a kedvelt irodai szoftverrel, az Excellel is megvalósíthatók, bár ott ez a legritkábban alkalmazott szolgáltatások közé tartozik. Másfelől a könyv azért kevesebb annál, hogy túlélőkészletként szolgáljon a többváltozós elemzés dzsungelében, mert a leggyakrabban alkalmazott SPSS elemzési eljárásokat ismertnek tételezi fel: gyakorlatilag mindazt, ami egy kezdő SPSS tanfolyamon téma lehet.

Példák segítségével. A szerzők jól ismerik az átlagos szociológia szakos hallgatót, aki türelmesen meghallgatja az elméleti okfejtéseket is, de végül mindig konkrét szám-
példákból, színes hasonlatokból és jól megkonstruált ábrákból érti meg a többváltozós statisztikai módszerek megválasztásának és értelmezésének kereteit, lehetőségeit és veszélyeit. Ennek megfelelően a könyv szemléletes, friss stílusban, régi jó ismerőseiként mutatja be a kiválasztott módszereket, melyekről, amíg lehet, közvetlen, ismeretterjesztő hangon szól. Ha tehetik, a szerzők kerülnek a megcélzott olvasóközönség számára hamar érthetlenné váló matematikai részleteket.

Máshol elmagyarázatlan statisztikai mutatók. *A Túlélőkészlet* további erénye, hogy magyarázatai számos olyan statisztikai mutató jelentésébe avatnak be, amiről az SPSS outputok és a programhoz csatolt magyarázatok (a Help és a Tutorial) csak szűkszavúan vagy egyáltalán nem szólnak. Sőt, a szerzők „kerülő úton” maguk is kreálnak olyan érzékletes mutatókat, amik az SPSS hiányosságait pótolják (például a logisztikus regressziós fejezetben).

Hibák. Célszerű lett volna a kéziratot valamivel alaposabb szerkesztői kontroll alá vetni, mert nem hiányoznak belőle a kisebb hibák, kifejejtések sem. Például az egyik fejezet címében ott szerepel a „Lazarsfeld-paradigma” fogalom, a könyv azonban nem magyarázza meg, hogy ez mi és azt sem, hogy a fejezet anyagának mi köze van a fent nevezett paradigmához. Egy másik ilyen figyelmetlenség, hogy – bár a bevezetőben ígéretet kapunk arra, hogy tárgymutató is lesz a könyvhöz – hiába keressük, még fogalomjegyzéket sem találunk a kiadvány végén.

Szerkesztési problémák. A bevezetőben a szerzők utalnak arra, hogy a könyvet nem folyamatosan kell olvasni. Valóban, a bemutatott modellek szinte tetszőleges sorrendben elővehetők és felépítésük-lebontásuk-átépítésük önállóan is gyakorolható az Internetről letölthető demonstratív SPSS adatbázisokkal. Ugyanakkor nehéz belátni, hogy a kiadvány fejezeteinek, a tárgyalt modelleknek miért éppen ez a sorrendje. Lehet, hogy célszerűbb lett volna először a gyakrabban alkalmazott regresszió-számítást és a variancia-analízist tárgyalni, és a könyv végére hagyni a ritkábban használt, valamint a nehezebben értelmezhető modelleket. Ez egyben egyfajta nehézségi sorrend érvényesítése is lett volna. További szerkesztési hiányosságra utal a könyv első fejezete: A hasznos véletlen hiba. Az átlagos olvasó számára érthetetlen – de mindenképpen megmagyarázatlanul marad —, hogy ez az okfejtés hogyan került oda és mi köze van a többi fejezethez. Ráadásul itt annyi technikai apróságon kell átrágnia magát az olvasóknak, hogy sokuknak bizonyára az lesz az érzésük: kissé nehézkesen vezetik el őket a túlélőpróba helyszínére. Pedig igazán kár lenne, ha emiatt túl sokan morzsolódnának le, mert ezután mindjárt következnek a fontos és érthető fejezetek.

Az olvasói célcsoport leszűkítése. Az SPSS leggyakrabban alkalmazott statisztikai szolgáltatása a gyakoriságszámítás, a keresztábra-készítés és a lineáris korreláció-számítás. A szerzők ezeket a modelleket ismertnek tételezik fel, támaszkodnak rájuk, de nem mélyednek el ezek finomságaiban. Ebből és más, ismertnek feltételezett elemi statisztikai fogalmakból is látszik, hogy a könyv által megcélzott olvasóközönség már nem kezdő a többváltozós elemzés területén. A teljesen kezdő olvasó számára amúgy is megnehezítené a gondolatmenet követését, hogy az egyes modellek építéskor figyelni kell a sajátos szintaktikájú SPSS nyelven kiadott számítási parancsokra is. Igaz, a függelékben a könyv elmagyarázza, hogyan kell ezeket az utasításokat egy menürendszer segítségével generálni, ez azonban nem változtat azon a tényen, hogy a modellek építése a könyvben végig SPSS programnyelven történik. Eközben az SPSS-nek éppen az egyik legnagyobb versenyelőnye a többi statisztikai szoftverrel szemben, hogy szinte valamennyi szolgáltatása menüből is elérhető és éppen ennek köszönhető, hogy ha nem akarunk, akár nem is kell tudnunk e sajátos parancsnyelv létezéséről. Ez az oka annak, hogy a Windows alkalmazásokhoz szokott diákok egyre szélesebb köre nyúl magabiztosan a statisztikai elemzés ezen eszközehez. A fentiek miatt a könyv olvasói célcsoportját azok alkotják, akik a többváltozós technikákat már középfeladói szinten ismerik és alkalmazzák.

A kétely operacionalizálása. A könyv nem abszolutizálja témáját: egy pillanatra sem rejti véka alá, hogy a bemutatott módszerekkel elért eredmények sohasem adnak megfellebbezhetetlen és végső választ az elemzett társadalomkutatási kérdésfeltevésekre. Sőt: a könyv gyakran él azzal a gyanúval, hogy a kapott számítási eredmények érvényessége korlátozott és éppen ezért számos módszert ajánl az érvényesség határait mintegy próbálgató, fészegető kísérletek elvégzésére. A címben ajánlott túlélőkészlet-

nek talán ezek a legértékesebb darabjai, melyekkel mintegy “körbekísérletezgethjük”, a sokdimenziós térben körbetapogathatjuk a vizsgált jelenségeket, tendenciákat. A könyv minden fejezetében vannak receptek arra, hogyan kell az eredményekben kételkedni: valamennyi ismertetett módszer esetében megtudjuk, miként vessük latba a kísérletezés számos – az SPSS-ben rendelkezésre álló, vagy csak trükkökkel, kerülő úton létrehozható – eszközét az eredmények javítására, megerősítésére, árnyalására vagy azok elvetésére. Megtanulunk résen állni, hogy ha lehet, kivédjük a semmitmondás vagy – épp ellenkezőleg – a belemagyarázás mindig ott leselkedő veszélyeit.

A kísérleti eszköztár azonban korántsem teljes: a könyv adósunk marad ezeknek az érvényesség-vizsgálati eljárásoknak egyfajta áttekintésével, rendszerezésével, enumerációjával. Például, a bemutatott eljárások mellett, további érzékenységi vizsgálat tárgya lehetett volna az is, hogy bizonyos magyarázó jellegű modellek, módszerek eredményei mennyire függenek a magyarázatként szolgáló alacsony mérési szintű, kategóriális változók aggregátsági szintjétől. Így például a variancia-analízisnél és a logisztikus regressziónál mindennapos tapasztalat, hogy ha a válaszolók foglalkozását tekintjük az egyik magyarázó tényezőnek, akkor az eredmény – a magyarázat – attól függően lesz szignifikáns, vagy sem, hogy a válaszolókat foglalkozásuk szerint 5 vagy esetleg 10 kategóriába soroljuk-e be.

Hol vannak az ingoványos területek? Azokkal a modellekkel kapcsolatban, ahol a legnagyobb az elemző szabadságfoka, amelyek a legtöbbféleképpen paraméterezhetők és, amelyek így a legszélesebbre tárják az ajtót az önkényes értelmezések előtt – így különösen a faktor- és klaszterelemzésnél –, a szerzők számos praktikus óvatossági rendszabályt megfogalmaznak. Az olvasó mégsem kap kézhez egy olyan áttekinthető térképet, amelyen fel lenne tüntetve, hogy merre vannak azok az ingoványos területek, ahová csak nagyon indokolt esetben szabad bemenészkedni. Pedig ez az egyik legfontosabb szabálya annak, hogy túléljük a túl merész – vagy éppen ellenkezőleg, túl közhelyes – elemzésért kapható lesújtó kritikát. Nem a szerzők tehetnek arról, hogy egyes módszerek abba a hírbe keveredtek, hogy eredményeik gyakorlatilag reprodukálhatatlanok, tehát, hogy két különböző kutató csak ritkán hozza ki ugyanazt az eredményt egyazon adatbázis elemzésekor, ugyanazt a kérdésfeltevést vizsgálva. Ugyanakkor elvárható lett volna, hogy egyfajta egységes *érzékenység-vizsgálat* segítségével a felkínált modelleket egybevetve értékeljék azokat abból a szempontból, hogy melyiknél mekkora és a belemagyarázás vagy a semmitmondás veszélye.

Az elemzés: algoritmus vagy művészet? Talán a könyvnek valamivel határozottabban kellett volna állást foglalnia amellelt, hogy egyik sem. Természetesen nem lehet olyan determinisztikus, zárt, minden lehetőséget figyelembe vevő elemzési módszert alkotni, ami valamiféle beprogramozható “szakértői rendszer” módjára írná elő, hogy adott adathalmaz és kérdésfeltevés esetében miként vizsgálódjunk, bármilyen nagy igény is lenne erre a szociológushallgatók körében. A másik véglet sem igaz, mely szerint a többváltozós elemzés nem más, mint egy nehezen áttekinthető terepen végzett túlélőpróba, ahol csak az lehet sikeres, aki rugalmasan tudja alkalmazni a bennfentes mesterek által engedélyezett ravasz trükköket. Az igazság a két véglet között van: valamennyi többváltozós módszer alkalmas arra, hogy segítségükkel sejtéseket, intuíciónkat szerezzünk, ezen belül a spekulációknak túl tág teret engedő modellek kizárólag erre alkalmasak. Biztosan azonban csak akkor állíthatjuk egy átfogó összefüggés fenn-

állítását, ha előzetesen megszerzett sejtéseinket sokféleképpen alátámasztjuk magyarázó modellekkel, elsősorban variancia-analízissel és regresszió-számítással, valamint robusztus egy- és kétváltozós módszerek alkalmazásával, és ha szükséges, az elemi adatokhoz való minél gyakoribb visszatéréssel.

A többváltozós elemzési módszerek akkor alkalmazhatók a leghatékonyabban, ha azokat megfelelőképpen párosítják a kvalitatív analízis eszközeivel. Azt, hogy milyen kérdésfeltevést akarunk megfogalmazni, már akkor tudnunk kell, mielőtt empirikus adatainkhoz hozzányúlunk. A feltárt összefüggések jellegét, azt, hogy oksági kapcsolatról vagy pusztán “együtt mozgásról” van-e szó, sohasem tudhatjuk meg a számszaki eredményeinkből, az SPSS outputból. A szerzők egy következő könyvének éppen ezek a módszerillesztési kérdések lehetnek a tárgyai.

A Túlélőkészlet pedig minden kétséget kizáróan sok szociológus diáknak és kutatóknak fog segíteni abban, hogy túléljék a többváltozós elemzés viszontagságait, a tanszéki könyvtárak pedig helyesen teszik, ha egyszerre sokat rendelnek belőle, mert kapós lesz.