Országos Korányi Tbc és Pulmonológiai Intézet

Az SPSS statisztikai programcsomag használatának tapasztalatai

Angyal István, Hofhauser Béla, Kiss Péter

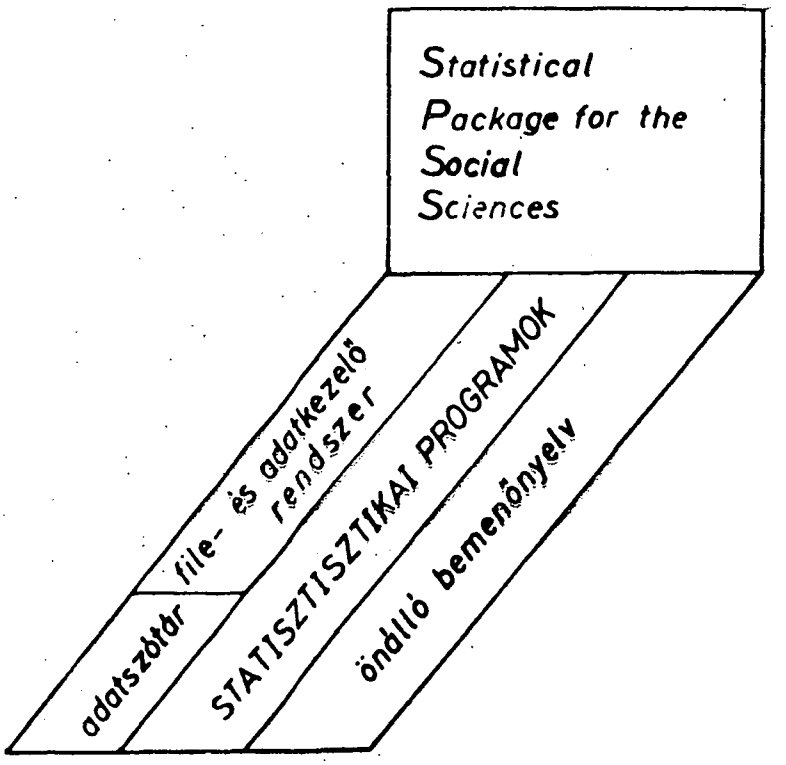
Az Intézetben folyó számitástechnikai tevékenység legfőbb célja az Országos Intézeti Modell kidolgozása. A Modell és a létrehozandó rendszerek feltételezik egy olyan szoftver eszköz mindennapos használatát, mely alkalmas nagytömegü adatok megbizható statisztikai leirására és elemzésére.

Egy ilyen szoftvér eszköz saját erőből való létrehozása a még ma is fennálló létszámhiány miatt kizártnak tekinthető. A hazai programcsomagok között nem találtunk olyan szavatolt és karbantartott terméket, amely statisztikai elemzéseket és egyben tág adat- és file-kezelési lehetőségeket is megenged. A külföldi programcsomagok közül az SSP /Scientific Subrutine Package-et/ és a BMD-t adatkezelési illetve esetszám korlátozási nehézségeik miatt vetettük el.

A külföldről beszerzett ajánlatok közül az SPSS Inc. ajánlata tünt a legkedvezőbbnek. 30.000 Ft-nak megfelelő devizáért hozzájutottunk az SPSS programcsomag örökös bér-

letéhez. E programcsomagot a világon kb. 1500 helyen installálták már és igen széles felhasználói körben alkalmazzák. Az alapváltozatának kidolgozása kb. 100 emberévi programozói munkát jelentett. Az ujabb változatok mind több statisztikai eljárást tartalmaznak, melyek minőségét szavatolják és a teljes programcsomag karbantartását vállalják.

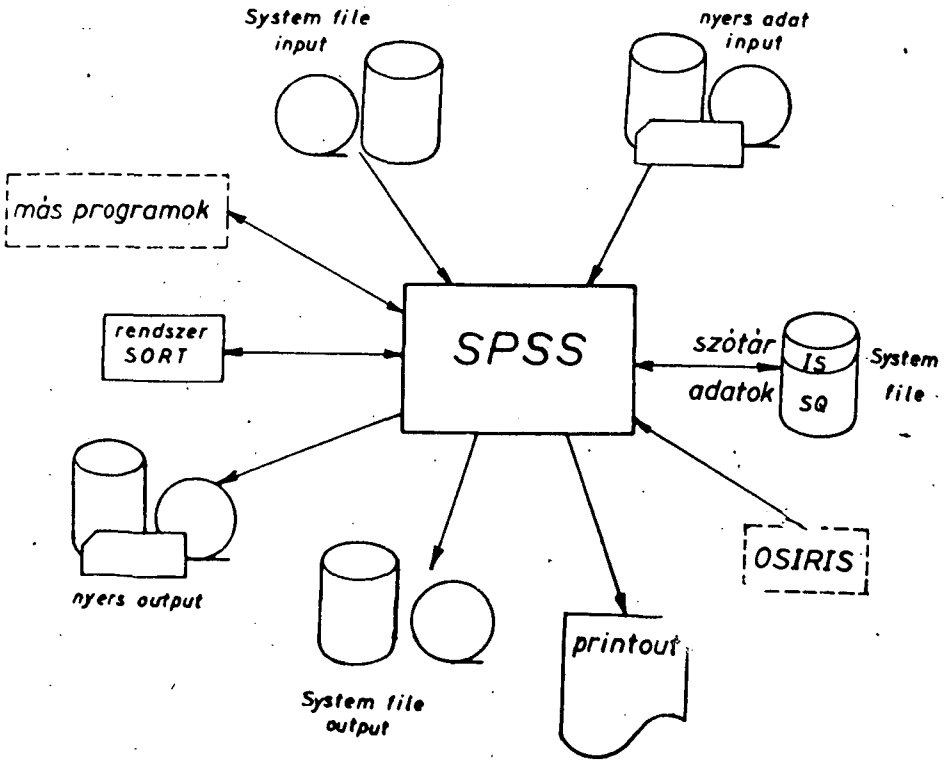
A Statistical Package for the Social Sciences programcsomag /1. ábra/ önálló bemenőnyelven keresztül fix szerkezetü adatállományok statisztikai leirását és elemzését végzi el és egyben adatszótára segitségével bonyolult file és adatkezelési müveleteket is lehetővé tesz.



1. ábra

Az önálló bemenőnyelv minden lehetőséget számitva 80 kulcsszóból áll, szintaxisa és logikája egyszerü, számitástechnikában nem járatos emberek számára is viszonylag könnyen elsajátitható. Részletes hibaüzenetei a felhasználót gyorsan rávezeti a tévesztésekre.

A file és adatkezelés a programrendszer saját adatszótárán és adatállományán /2. ábra/ az u.n. system fileon keresztül történik. Az adatszótár indexszekvenciális szervezésü és a vizsgált adatállomány adatainak fizikai, logikai és egyéb releváns információit tartalmazza. A system file adatrésze az input adatállománynak a rendszer szempontjából standardizált szekvenciális leképezése. A system file magasfoku integráltsága a változók kezelését egyszerüvé, áttekinthetővé teszi és ez többek között a tévesztési és hibalehetőségeket jelentősen csökkenti.



2. ábra

A programrendszer fontos tulajdonsága az adatokkal és más szoftverekkel szembeni nyitottság. Tetszőleges formátumú FORTRAN, PL/1 vagy COBOL programok által létrehozott output adatállományokat tud fogadni, illetve ezek számára input-ot tud generálni. A meglévő statisztikai subrutinokhoz továbbiak is beilleszthetőek és az OSIRIS programcsomaghoz kész interface-sel rendelkezik. Az adatállományok rendezése során a rendszer-SORT programot aktivizálja.

Az SPSS szempontjából nyers input adatállományból a VARIABLE LIST és INPUT FORMAT kártyák paraméterei szerint az első statisztikai alprogram hivásakor a rendszer létrehozza az ideiglenes /temporary/ system file-t, melyet a további tetszőleges számban meghivható alprogramok mindegyike végigolvas. A létrehozott temporary file-t a SAVE FILE utasitás hatására a rendszer teljes egészében kimenti és további futtatások során a GET FILE utasitással behivható, nincs szükség további adatleirásra.

A nyers output-ra tetszőleges illetve szelektált adatok vagy a statisztikai programok közbenső eredményei mint pl. korrelációs mátrix, residuumok, decriptiv információk stb. vihetők ki. Ezeket a rendszer standard módon inputállományként is tudja kezelni.

A beolvasott system file-hoz lehetőség van nyers in-

putról esetek, subfile-ok vagy változók hozzáadására. A system file kimentésekor pedig mód van feleslegessé vált változók vagy esetek törlésére. A file-kezelési müveletek tág lehetőséget biztositanak a feldolgozási szempontoknak megfelelő esetek kialakitásához. Erre azért is van szükség, mert az egészségügyi szervezés feldolgozásai gyakran merőben más eset-definiálást igényelnek. Mások a vizsgálandó eset kritériumai, ha egy egészségügyi egység terhelését vizsgáljuk és mások, ha egy bizonyos betegség ápolási, gondozási szükségletét elemezzük.

A tág file-kezelési lehetőségeket sokrétü adatkezelési müveletek egészitik ki. A COMPUTE utasitással meglévő vagy uj változóknak értéket adhatunk FORTRAN-szerü értékadó utasitással /3./a. ábra/. A BROCA nevü változó ezen utasitással az u.n. Broca-indexet fogja tartalmazni, melyet a testsúlyból és a testmagasságból lehet kiszámitani. Az IF feltételes értékadó utasitás tetszőleges logikai feltétel teljesülése esetén változtatja meg a hivatkozott változó értékét /3./b. ábra/. A systolés VERNYS és diastoles VERNYD vérnyomás értékeitől függően a vérnyomásmérés minőségét kifejező változó VERNY 1, 2 vagy 3-as értéket vesz fel, mely kifejezi a hypo-, a normo- és a hypertoniát.

A RECODE-utasitás a változók értékkészletének átkódolását teszi lehetővé /3./c. ábra/. Az életkort kifeje-

ző KOR változó értékeit korcsoportnak megfelelően átkódoljuk és ezek után a KOR változó 1-es értéke a tizenéveseket, a 2-es a fiatalokat stb. fogja jelenteni. A COUNT utasitás végigszámolja, hogy egy esetnél hányszor fordulnak elő bizonyos változók megadott értékei /3./d. ábra/. Pl. egy személy betegség-kódjait a BNO1, BNO2, BNO3 változók tartalmazzák. Amennyiben az egyik BNO-kód 250, azaz a diabetes kódja, akkor a DIABET változó értéke 1 lesz.

A programrendszer tetszőleges szelektálást, az esetek súlyozását és mintavételezését is megengedi.

A SELECT IF utasitása alapján csak a megadott logikai feltételnek eleget tevő rekordok vesznek részt a feldolgozásban /3./e. ábra/. Tehát csak a 110-nél kisebb és 90-nél nagyobb Broca-index-el rendelkező u.n. atletikus termetü személyeken végzi el a megadott feldolgozásokat. A WEIGHT utasitásban megadott változó értékének megfelelően súlyozza az egyes eseteket. /3./f. ábra/ Példánkban a SULYOZ változót úgy határoztuk meg, hogy a diabetes diagnózissal nem rendelkező személyek csak fele súllyal vegyenek részt a feldolgozásokban. Mindezek az adattranszformációk és szelektálási lehetőségek megadhatók permanensen, tehát egy egész futtatásra, vagy pedig időlegesen tehát egy statisztikai program végrehajtásának idejére. Nagy és viszonylag homogén adatállományok elemzését segiti elő a SAMPLE utasitás. Példánk /3./g. ábra/ 10 %-os

mintavételezést ir elő, mely egy beépitett és újra inditható véletlenszámgenerátorral történik.

```
a./ COMPUTE     BROCA=TESTS/(TESTM-1ØØ)*1ØØ

b./ COMPUTE     VERNY=2
    IF          (VERNYS GT 16Ø OR VERNYD GT 1ØØ) VERNY=3
    IF          (VERNYS LT 1ØØ OR VERNYD LT  7Ø) VERNY=1

c./ RECODE      KOR (1Ø THRU 19=1)(2Ø THRU 39=2)
                (4Ø THRU 59=3)(6Ø THRU HIGHEST=4)

d./ COMPUTE     DIABET=0
    COUNT       DIABET=BNO1,BNO2,BNO3 (25Ø)

e./ SELECT IF   (BROCA LT 11Ø AND BROCA GT 9Ø)

f./ COMPUTE     SULYOZ=Ø.5+Ø,5*DIABET
    WEIGHT      SULYOZ

g./ SAMPLE      Ø.1
```

3. ábra

Az SPSS 6 leiró és 14 elemző statisztikai eljárást tartalmaz /4. ábra/. Az egyes programok jellemzőinek ismertetése, algoritmusainak, módszereinek leirása, más programcsomagokkal való összehasonlitása túlnő e 15 perces előadás keretein. Az elemző eljárások közül jelenleg még hiányzik a clusteranalysis, a többváltozós varianciaanalysis és mindenek előtt az idősorok elemzésére alkalmas eljárások. Ujabb kibocsájtások előrejelzések szerint a többváltozós varianciaanalysist már biztosan tartalmazni fogják és a további programok beépitését is tervezik.

AGGREGATE	ANOVA
BREAKDOWN	CANCORR
CONDESCRIPTIV	DISCRIMINANT
CROSSTABS	FACTOR
FREQUENCIES	GUTTMAN SCALE
SCATTERGRAM	MULTI RESPONSE
	NONPAR CORR
	NPAR TESTS
	ONEWAY
	PARTIAL CORR
	REGRESSION
	RELIABILITY
	T-TEST

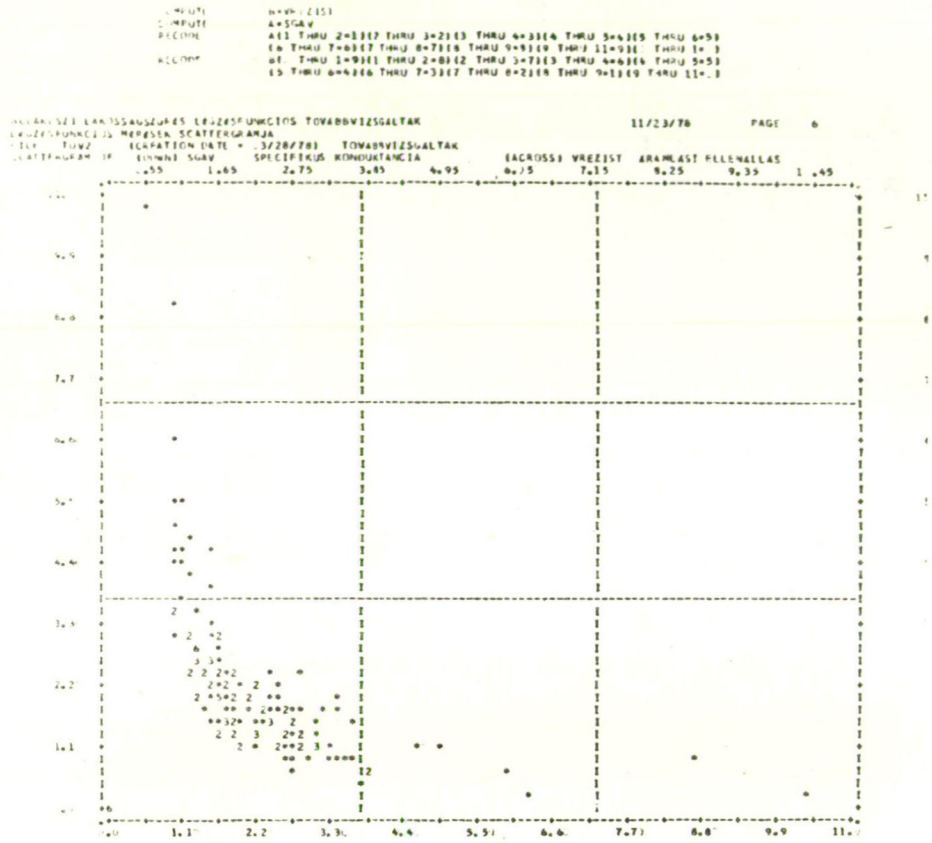
4. ábra

A programcsomag lehetővé teszi a hiányzó értékek egységes ellenőrizhető kezelését az adat és file kezelés során és valamennyi statisztikai programban. Minden változóra maximálisan három hiányzó érték /missing value/ adható meg. Igy lehetőség van arra, hogy különbséget tegyünk pl. azok között, akik nem akartak, nem tudtak, illetve helytelenül válaszoltak a feltett kérdésekre.

Az 5. ábrán légzésfunkciós értékek SCATTERGRAM-ját láthatjuk. Felette ugyanezen értékek VREZIST illetve SGAV transzformációját és átkódolását az A és B változókba. Az átkódolás érdekessége még, hogy B értékeinek átkódolása során a skálát megforditottuk és igy az A és B kereszttábláján /6. ábra/, az előző összefüggés tükörképe figyelhető meg /CORSSTABS/. A kért statisztikák a kontingencia táblában számitható különböző együtthatókat szemléltetik.

Az SPSS installálása óta eltelt másfél év alatt több kisebb, nagyobb feldolgozást végeztünk /7. ábra/. A változók és az esetek száma kétszer logaritmikus koordinátarendszerében minden egyes pont egy adatállományt illetve feldolgozást jelöl.

- GY : alkalmi kisebb volumenü klinikai-hatástani gyógyszer tesztek, melyek az Intézetünkben folytak és az SPSS-el értékeltünk ki.
- B76 és B78 : az 1976-os Budakeszi lakosságszürés és ugyanezen populáció légzési panaszokkal rendelkező részének 1978-ban újraszürése.
- TOV : az 1976-os Budakeszi lakosságszürés légzésfunkciós továbbvizsgáltjairól készült adatállomány. Az 5. és 6. ábrán erről a populációról láthattunk egy kiragadott adatösszefüggési példát.


COMPUTE
COMPUTE
RECODE
RECODE
11/23/78 PAGE 6
EGRESS) VREZIST ARAMLAST ELLENALLAS
9.25 9.35 1 .45
11/23/78 PAGE 7
STATISTICS..
CORRELATION (R)- -.64596 R SQUARED - .41688 SIGNIFICANCE - 0.00001
STD ERR OF EST - 1.34691 INTERCEPT (A) - 3.27457 SLOPE (B) - -.65628
PLOTTED VALUES - 155 EXCLUDED VALUES- 0 MISSING VALUES - 0

5. ábra

BUDAKESZI LAKOSSAGSZORES LEGZESFUNKCIOS TOVABBVIZSGALTAK 11/23/78 PAGE 9
KECODOLT ERTEKEK KERESZTTABLAT
FILE TUV2 (CREATION DATE = 03/28/78) TOVABBVIZSGALTAK

* CROSSTABULATION OF *
A AZ SAV ERTEKEINEK ATKODOLASA BY B A VREZIST ERTEKEINEK ATKODOLASA
* PAGE 1 OF 2

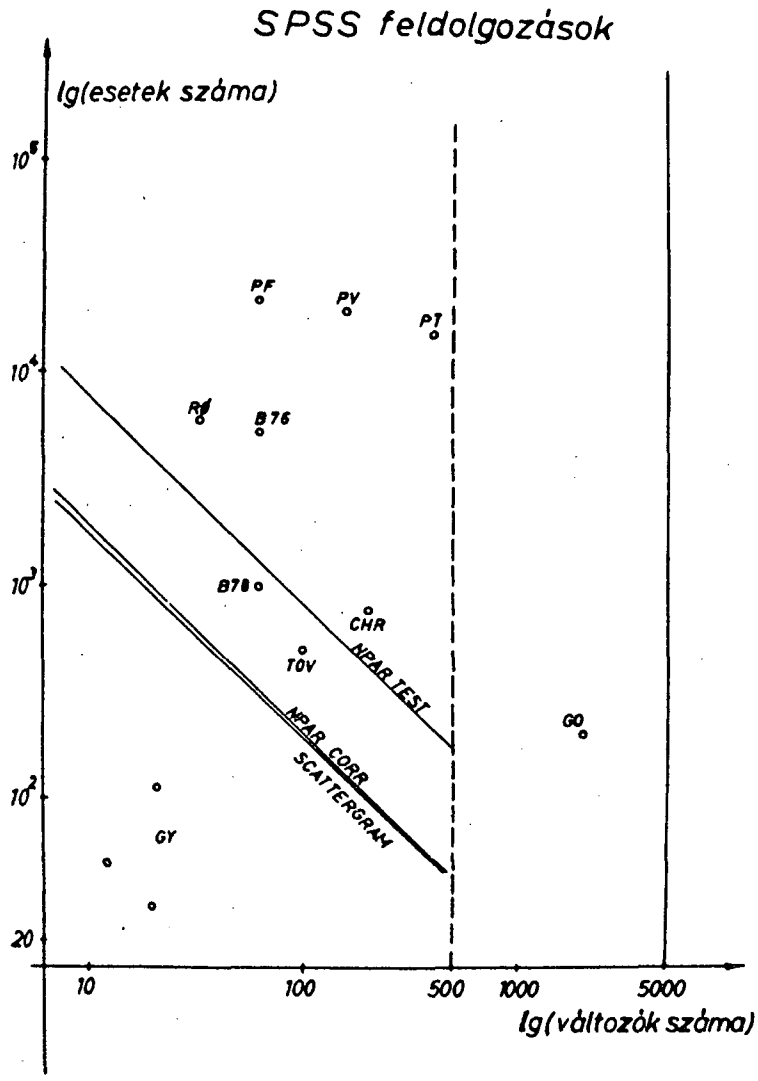
A	B: COUNT / ROW PCT / COL PCT / TOT PCT	10-11	8-9	6-7	5-6	4-5	3-4	2-3	1-2	0-1	ROW TOTAL
	0	1	1	1	1	0	7	4	0	6	21
0-1		4.8	4.8	4.8	4.8	0.0	33.3	19.0	0.0	28.6	13.5
		100.0	33.3	33.3	33.3	0.0	41.2	9.1	0.0	37.5	
		0.6	0.6	0.6	0.6	0.0	4.5	2.6	0.0	3.9	
	1	0	0	0	0	2	10	34	22	0	68
1-2		0.0	0.0	0.0	0.0	2.9	14.7	5.0	32.4	0.0	43.9
		0.0	0.0	0.0	0.0	100.0	58.8	77.3	31.6	0.0	
		0.0	0.0	0.0	0.0	1.3	6.5	21.9	14.2	0.0	
	2	0	0	0	0	0	0	0	6	36	42
2-3		0.0	0.0	0.0	0.0	0.0	0.0	14.3	85.7	0.0	27.1
		0.0	0.0	0.0	0.0	0.0	0.0	13.6	57.1	0.0	
		0.0	0.0	0.0	0.0	0.0	0.0	3.9	23.2	0.0	
	3	0	0	0	0	0	0	0	8	3	11
3-4		0.0	0.0	0.0	0.0	0.0	0.0	0.0	72.7	27.3	7.1
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.1	10.8	
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.2	1.9	
	4	0	0	0	0	0	0	0	5	3	8
4-5		0.0	0.0	0.0	0.0	0.0	0.0	0.0	62.5	37.5	5.2
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.9	10.8	
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	1.9	
	5	0	0	0	0	0	0	0	1	1	2
5-6		0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	50.0	1.3
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	3.3	
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.6	
	6	0	0	0	0	0	0	0	0	1	1
6-7		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.6
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	
	9	0	0	0	0	0	0	0	0	2	2
9-11		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1.3
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.5	
		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	
	COLUMN TOTAL	1 0.6	1 0.6	1 0.6	1 0.6	2 1.3	17 11.0	44 28.4	72 46.5	16 10.3	155 100.0

RAW CHI SQUARE = 150.77316 WITH 56 DEGREES OF FREEDOM. SIGNIFICANCE = 0.0000
CRAMER'S V = 0.37278
CONTINGENCY COEFFICIENT = 0.7222
LAMBDA (ASYMMETRIC) = 0.27586 WITH A DEPENDENT. = 0.26526 WITH B DEPENDENT.
LAMBDA (SYMMETRIC) = 0.27059
UNCERTAINTY COEFFICIENT (ASYMMETRIC) = 0.32622 WITH A DEPENDENT. = 0.34871 WITH B DEPENDENT.
UNCERTAINTY COEFFICIENT (SYMMETRIC) = 0.33379
KENDALL'S TAU B = .52705. SIGNIFICANCE = 0.0000
KENDALL'S TAU C = 0.41814. SIGNIFICANCE = 0.0000
GAMMA = 0.69613
SOMERS'S D (ASYMMETRIC) = 0.53743 WITH A DEPENDENT. = 0.51688 WITH B DEPENDENT.
SOMERS'S D (SYMMETRIC) = 0.52695
ETA = 0.54296 WITH A DEPENDENT. = 0.51998 WITH B DEPENDENT.
PEARSON'S R = 0.64651 SIGNIFICANCE = 0.0000

6. ábra

- CHR : a chronicus fertőző tbc-s betegek országos felmérésének adatállománya, mely lényegében még ma Magyarországon található súlyos tbc-s betegeket és egyben a legveszélyesebb fertőzési potenciált jelentő személyeket tartalmazza.
- RØ : a negativ tüdőröntgen-előzménnyel rendelkező friss tbc-s betegekről készült országos felmérés során a negativ rtg előzmények és a friss tbc-s megbetegedések összefüggését vizsgáltuk.
- GO : a tüdőgondozó intézetek évi statisztikai jelentései /melyek manuális módszerrel készültek/ fontos támpontot jelentenek terveink, megvalósitandó rendszereink kialakitásában.
- PF, PV és PT : a pécsi Komplex Szürőállomás 1977. évi tevékenységének retrospektiv vizsgálata, mely több mint 6 millió adatával az eddigi legnagyobb feldolgozásunk.

Az ábrán szemléltetni igyekeztünk az SPSS korlátait. 500 változóig standard kezelést biztosit, speciális kezeléssel 5000 változót lehet a system file-ba bevinni. Egy futtatás során viszont maximálisan 500 változóra lehet hivatkozni. A programok jelentős részénél nincsen esetszámkorlátozás. A hivatkozott változók számától függően csak meghatározott számú eset vehet részt az NPAR TEST, az NPAR CORR és a SCATTERGRAM programokban. Amennyiben több



7. ábra

esetünk van mint a megengedett, akkor az SPSS véletlenszerü mintát vesz a megengedhető maximális esetszámnak megfeleleően az összes esetből.

Az SPSS adat és file-kezelésre igen alkalmas, a változók paramétereit tartalmazó adatszótár létrehozása, kezelése, karbantartása egyszerü. Az egyes statisztikai subru-

tinok gyorsaságának, effektivitásának megitélésére még nem teljesen kialakult a véleményünk. Más hasonló szoftvereket használó csoportokkal való megbeszéléseink, esetleg összehasonlitó futtatások elemzése során erről reálisabb képet kaphatunk. Minden esetre elmondható, hogy kis esetszámoknál /néhány 100 eset/ minden szempontból jónak találjuk a gyorsasági tényezőket, olyannyira, hogy az SPSS akár ilyen adatállományok validálására is alkalmas. Közepes feldolgozásoknál /néhány ezer esetnél/ a file és adatkezelés még kielégitő, de a statisztikai programoknál már körültekintőnek kell lenni. Ebben a tartományban az SPSS-t már nem lehet validálásra ajánlani. Nagy feldolgozásoknál /több tiz- esetleg százezer esetnél/ a végrehajtási idő optimalizálása érdekében körültekintően kell eljárni, pontosan meg kell tervezni a feldolgozásokat és az SPSS kezelésében gyakorlattal kell rendelkezni. Nagy esetszámok statisztikai elemzésénél pedig mindenképpen célszerü a rendszer mintavételezési lehetőségeit kihasználni.

A program közel teljes dokumentálása megtalálható az SPSS Manual-ben /1975 McGraw-Hill, Inc. New York/. E kézikönyv szinte minden szempontból jó, talán egyedüli hiányosságának az adható meg, hogy a matematikai statisztikai eljárások módszereinek, feltételezéseinek, kikötéseinek leirása nem mindig teljes és azok csak a megadott irodalmak gondos áttanulmányozása során derithetőek ki. Ez bi-

zonyos mértékig akadályozza az adequát statisztikai elemzések gyors alkalmazását.

Összefoglalva elmondhatjuk, hogy az SPSS széles alkalmazási területe és magasfokú integráltsága alapján a számitástechnika mai fejlettségi szintjén a világszinvonalú, professzionista alkalmazási programcsomagok közé tartozik. A statisztikai programcsomagok közül az SPSS-nek elsőnek sikerült teljesiteni azokat a nagyon nehezen elérhető kritériumokat, melyek alapján 1977-ben megkapta az alkalmazási szoftverek részére adható egyik legnagyobb elismerést a Datapro Software Honor Roll-t.

Közép vagy nagy számitógéppel rendelkező intézményeknek csak javasolni tudjuk e programcsomag beszerzését. Ezen gépek üzemeltetőinek és más érdeklődőknek is szivesen adunk tájékoztatást további részletekről.