KSH Államigazgatási Számitógépes Szolgálat

Országos felmérések multifaktoriális vizsgálatára készült programrendszer tapasztalatai

Srajber Benedek, Dabóczi Ákos és munkatársaik

1. Bevezetés

Egy számitástechnikai szakemberekből és orvosokból álló munkacsoport többéves munkájának eredményeként megszületett, okozati összefüggések feltárására alkalmas, moduláris felépitésü tipikus feldolgozó programrendszer számitástechnikai és nagyméretü feladatokra történő alkalmazási tapasztalatairól számolunk be.

Az első alkalmazásokat a magyarországi ujszülött populáció egészét /120.000 eset/ reprezentáló mintaanyagon végeztük. A programrendszer felhasználhatósági köre kiterjed többek között demográfiai, szociológiai és klinikai adatok értékelésére.

2. A programrendszer készitésének és alkalmazásának tapasztalatai

Induljunk ki egy adott S_0 mintából, amely

$$X_j = (\rho_j, \xi_{j1}, \xi_{j2}, \ldots, \xi_{jm}) \quad (j = 1,2,\ldots,n)$$

alaku vektorokból áll, ahol $\rho_j = 1,2,\ldots,k$ aszerint, hogy a j-edik vektor /eset, beteg, egyed/ melyik kategóriába tartozik, ξ_{ji} pedig a j-edik beteg i-edik jellemző

tulajdonságát /tényezőjét/ jelöli. Bontsuk fel az S_0 halmazt $S_1,S_2,\ldots,S_r$ diszjunkt részhalmazokra, továbbá határozzuk meg a vizsgálandó $C_1,C_2,\ldots,C_k$ kategóriákat az S_i $(i = 1,2,\ldots,r)$ halmazok megfelelő kombinációinak egyesitéseként. A programrendszer a kategóriák páronkénti összehasonlitását, a közöttük levő okozati összefüggések feltárását végzi. A moduláris felépitést az 1. ábra mutatja.

2.1 *A feldolgozás módszere*

Az 1. és 3. modul előkészitő jellegü rutinokból áll; az adatok beolvasását, átkódolását, rendezését, közvetett ellenőrzését és más modulok input file-jainak elkészitését végzi.

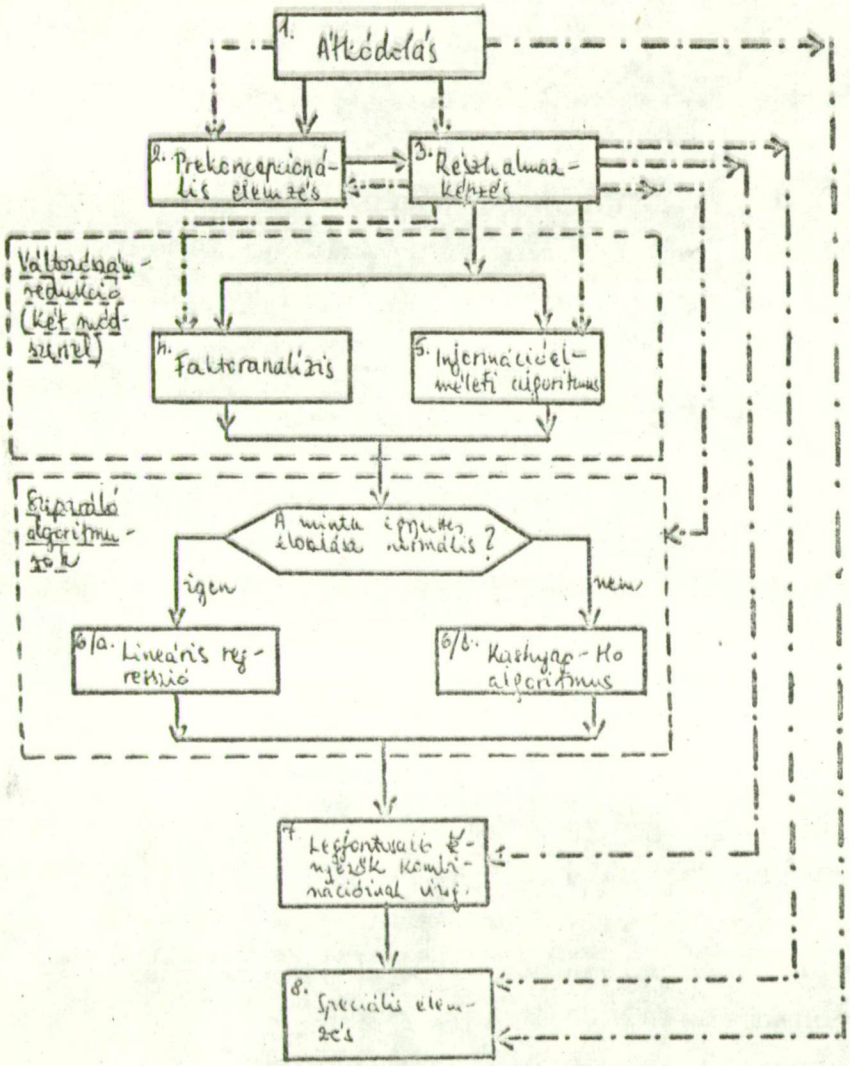
A prekoncepcionális vizsgálat /2. modul/ lényege, hogy minden kategóriára tényezők szerinti gyakoriságokat és feltételes relativ gyakoriságokat határoz meg, továbbá kategória-páronként hipotézisvizsgálatot /u-próbát/ végez és szignifikanciákat számit, hogy "első közelitésben" megadja a kategóriákra meghatározó jelleggel szóba jöhető tényezőket. Ezek száma sok tulajdonság /nagy m érték/ esetén matematikailag még nehezen kezelhető. Ezért végezzük el a tényezők /változók/ számának redukcióját a 4. és 5. modulok együttes alkalmazásával.

A faktoranalizis /4. modul/ főfaktor módszere a matematikai statisztikából ismert, programját a BMD-P-ből vettük, csupán nagyméretre használhatóvá dolgoztuk át a Honeywell gépen.

Az információelméleti módszeren alapuló lényegkiemelési eljárást /5. modul/, amelyet a klasszikus statisztikai módszerek korlátai nem zavarnak, magunk dolgoztuk ki. A módszer lényege, hogy kiválasztja az $X_1,X_2,\ldots,X_n$ változók közül azt az r számu változót $(r<<n)$, amely kö-

1. sz. ábra.

A kódolt kérdőíveket elemző programrendszer moduljai



——— elemzés menete —·—·— adatáramlás

1. ábra

zel annyi információt tartalmaz valamely vizsgált Y változóról /Y reprezentálhat rendellenes kategóriákat, betegségosztályokat, tulélést stb./, mint az összes X_j változó. Formálisan az algoritmus befejezésével az

$$i(Y_i X_{s_1},\ldots,X_{s_r}) \geq (1-\varepsilon)\cdot i(Y_i X_1,\ldots,X_n)$$

egyenlőtlenség teljesül, ahol ε megfelelően kis pozitiv szám /pl. $\varepsilon = 0{,}01$/. A módszer részletes ismertetése a 8. Kollokviumon hangzott el Szegeden és megtalálható (2)-ben.

A változók redukciója után most már könnyebb kezelnünk a megmaradt változókat, amelyek az egyes kategóriák elhatárolásánál szóba jöhetnek. Ha a vizsgált változókról megállapitható /esetleg feltételezhető/, hogy együttes eloszlásuk normális, akkor a jól ismert lineáris regresszióval /6/a modul/ egyszerü és rövid uton /kevés gépidő!/ juthatunk el a kategóriák sulyozott tényezők /változók/ szerinti szétválasztásához. Ha az együttes eloszlásról biztosat nem tudunk vagy feltételezhető, hogy nem normális eloszlásról van szó, akkor a Kahyap-Ho algoritmust /6/b modul/ használjuk a kategóriák szeparálására, s egyben a tényezők sulyainak meghatározására.

A kategóriák szeparálására szolgáló módszerek kérdésével behatóbban foglalkozik a 7. szegedi kollokviumon elhangzott előadásunk és megtalálható (15)-ben.

Miután a kategóriák szeparálásában szerepet játszó tényezőket meghatároztuk, és jelentőségüknek megfelelő sulyokat rendeltünk hozzájuk, hátra van még annak a fontos kérdésnek eldöntése, hogy a legnagyobb sulyt kapott tényezők kombinációinak /"cluster"-einek/ milyen befolyása van az egyes kategóriákra. Ezt a funkciót tölti be a 7-es modul, amely a helyes szakmai következtetések alátámasztására megbizhatósági intervallumokat is szolgáltat.

Minden komplex feldolgozásnál számitani lehet /és kell is!/ speciális igények felmerülésére. Ezek kielégitésére épithető be rugalmasan a 8. modul, amely a hazai ujszülött populáció vizsgálatában nálunk az intrauterin fejlődési görbék legisztikus illesztéssel történő előállitását jelentette. /Erről beszámoltunk az 1978-as kollokviumon./

2.2 *A feldolgozó rendszer alkalmazása*

Feladat: A hazai ujszülött populáció vizsgálata, különös tekintettel a koraszülés okainak és okozati összefüggéseinek feltárására.

A feladat méretei:

a vizsgált ujszülött populáció:	110157
a reprezentativ minta nagysága:	29745
az esetenként figyelembe vett tényezők száma:	129

A minta diszjunk részhalmazait az 1. sz. táblázat, a vizsgált kategóriákat pedig a 2. sz. táblázat mutatja.

A prekoncepcionális elemzés eredményének egy tényező kimeneteleire vonatkozó részlete: 3. sz. táblázat. Az adatredukciós algoritmusok összevont eredménye: 4. sz. táblázat. A szeparáló algoritmus /Kashyap-Ho/ szolgáltatta eredmény: 5. sz. táblázat. Azon tényezők listája, amelyeknek összes kombinációját megvizsgáltuk: 6. sz. táblázat.

A teljes hazai populációra jellemző intrauterin fejlődési görbék megrajzolására vizsgálataink alapján került sor először. Az intrauterin suly- és hossz-fejlődés görbéi közül illusztrativ példaként hármat mutatunk be /2., 3. és 4. ábra/.

1. sz. táblázat

A minta diszjunkt részhalmazai és elemszámai

A halmaz megnevezése	Azonosítása	Esetek	Mintán belüli arány
A 6 napot túlélő koraszülöttek	S_1	6751	0.193
A 0-6 napon elhalt koraszülöttek	S_2	2540	0.085
Normálisak /a 6 napot túlélő, nem koraszülött csecsemők/	S_3	20103	0.678
Halvaszülöttek	S_4	1067	0.036
A 0-6 napon elhalt, nem koraszülöttek	S_5	284	0.010

1.sz. táblázat

2. sz. táblázat

A vizsgált kategóriák alapadatai

A halmaz megnevezése	Azonosítás	Kapcsolat	Esetek	Teljes populáción belüli arány
A teljes populáció	S		110157	
A teljes minta	S_o	$S_o = \bigcup_{i=1}^{5} S_i$	n=29745	q_o=0,270
a./ A 6 napot túlélő koraszülöttek	KOR	KOR=S_1	n_1=5751	q_1=0,052
b./ A normálisak ötödrésze	NOR	NOR=S_3	n_2=20103*	q_2=0,912
c./ Halvaszülöttek	HSZ	HSZ=S_4	n_3=1067	q_3=0,010
d./ 0-6 napon elhaltak	MHA	MHA=$S_5 U S_2$	n_4=2824	q_4=0,026
e./ Összes élveszületett koraszülöttek	TKOR	$S_1 U S_2$	n_5=8291	q_5=0,075

* A q_2 kiszámításánál az összes normális esetet /az n_2 ötszörösét/ vettük figyelembe.

2. sz. táblázat

3. sz. táblázat

A kategóriapárok első közelítésű összehasonlításának szemléltetése

/Részlet a 3-as modul eredmény listájából./

Egy főre eső jövedelem	Gyakoriságok				SZIGNIFIKANCIA			Feltételes rel.gyak /q'_k/		
	KOR	NOR	HSZ	MHA	ELŐ HAL	KOR NOR	HAL NOR	0,965	0,054	0,037
≤ 600	818	9805	168	362	0,000	0,000	0,000	0,952	0,077	0,051
601 - 900	922	13315	170	407	0,016	0,000	0,008	0,961	0,065	0,042
901-1300	1244	22570	198	584	0,003	0,200	0,002	0,968	0,052	0,033
1301-1800	1456	28300	260	736	0,005	0,000	0,003	0,968	0,049	0,034
1801-2500	966	20565	182	492	0,000	0,000	0,000	0,969	0,045	0,032
> 2500	345	5960	89	244	0,000	0,833	0,000	0,949	0,056	0,053

3. sz. táblázat

4. táblázat

A koraszülött kategória adatredukciós eljárásokkal /4. és 5. modul/ nyert fontos befolyásoló tényezői*

Sorszám*	Azonosító	Befolyásoló tényező	Bevonás alapja: faktor. anal.	Bevonás alapja: inf. elm.
1.	68	Az anya iskola-éveinek száma	+	+
2.	115	Az apa iskolaéveinek száma	+	+
3.	59	Az anya életkora	+	+
4.	112	Az apa életkora	+	
5.	81	Kereső-e az anya?	+	+
6.	85	Az anya munkájának jellegzetességei	+	
7.	72	Az egy szobára jutó személyek száma	+	
8.	8o	Az egy főre eső jövedelem	+	+
9.	1o1	Az anya dohányzott a terhesség első felében	+	
1o.	1o2	Az anya dohányzott a terhesség második -"-	+	
11.	24	Ikerszülések száma		+
12.	27	Spontán abortuszok száma	+	
13.	26	A művi abortuszok száma		+
14.	17	Szövődmények a terhesség alatt I.		+
15.	2o	Szövődmények a terhesség alatt IV.		+
16.	22	Szövődmények a terhesség alatt VI.		+
17.	21	Szövődmények a terhesség alatt V.		+
18.	96	Terhességi tünetek		+
19.	88	Az anya betegségei a terhesség előtt	+	
2o.	89	Az anya betegségei a terhesség alatt	+	
21.	61	Az anya magassága		+
22.	23	Előző terhesség volt-e?	+	
23.	62	Az anya törvényes családi állapota	+	+

* A tényezőket nem fontossági sorrendben tüntettük fel.

4.sz. táblázat

5. sz. táblázat

A Kashyap-Ho algoritmus által szolgáltatott súlytényezők a koraszülés okainak vizsgálatánál

Sorrend	Azonosító	Befolyásoló tényező	Súlytényező
1.	20	Terhesség és szülés alatti szövődmények IV.	0.47034
2.	24	Élveszületések száma	0.33968
3.	26	Müvi vetélések száma	0.33317
4.	27	Spontán vetélések száma	0.30703
5.	62	Az anya törvényes családi állapota	0.23545
6.	101,102	Dohányzás a terhesség első, III. második felében /összevont változó/	0.20746
7.	22	Terhesség és szülés alatti szövődmények VI.	0.20413
8.	96	Terhességi tünetek /nincs, kóros hányás, fehérje vizelés, magas vérnyomás, rángógörcs, egyéb/	0.12246
9.	88	Terhesség előtti betegség /450 WHO betegségkód/	0.10485
10.	17	Terhesség és szülés alatti szövődmények I.	0.10475
11.	59,112	Az anya III. az apa születési ideje /összevont változó/	0.09105
12.	61	Az anya magassága	0.08802
13.	81,85	Kereső-e az anya? Milyen volt a munkája? /összevont változó/	0.03443
14.	72,80	Az egy szobára jutó személyek száma és az egy főre jutó jövedelem /összevont változó/	0.03332
15.	68,115	Az anya és az apa iskolában töltött éveinek a száma /összevont változó/	-0.00243
16.	89	Terhesség alatti betegség /450 WHO betegségkód/	-0.02741
17.	23	Előző terhesség volt-e?	-0.05086

A helyes osztályba-sorolás (diagnózis) valószinüsége: 0.759
A vizsgált esetek száma: 22798
Az AC vektor nem negativ komponenseinek száma: 17217

6. táblázat

A koraszülés szempontjából legfontosabbnak bizonyult 9 tényező összevonása a kombinációk vizsgálatához

1./ Művi vetélések száma igen: ha ≥ 1,
különben: nem.

2./ Terhesség és szülés szövődményei I.
/toxaemia/ + terhességi tünetek
igen: volt,
különben: nem.

3./ Terhesség és szülés szövődményei IV.
/placenta anomáliák/ bármelyik: igen,
különben: nem.

4./ Terhesség és szülés szövődményei VI.
/Idő előtti burokrepedés, cervix insufficientia, stb./
bármelyik: igen,
különben: nem.

5./ Az anya betegségei a terhesség előtt
igen: volt,
különben: nem.

6./ A család egy főre eső jövedelme ≤ 900 Ft +
Az anyával egy szobában lakók száma ≥ 3 +
Az anya nem kereső.
Ha bármelyik kettő fennáll: igen,
különben: nem.

7./ Az anya családi állapota (házas, nem házas).

8./ Az anya életkora igen: ha > 35 év,
különben: nem.

9./ Spontán vetélések száma igen: ha ≥ 1,
különben: nem.

Jelmagyarázat: A VI-2., VI-3., VI-4. táblázatokhoz:

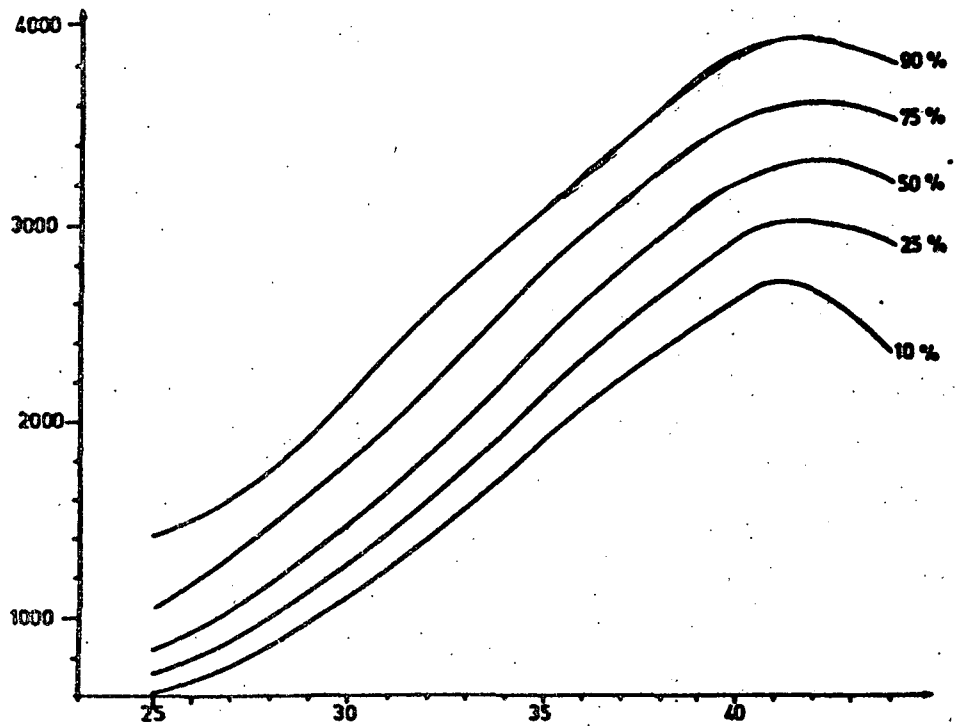
- : A tényezőt nem vettük tekintetbe
0 : A tényező nem áll fenn
1 : A tényező fennáll

7. sz. táblázat

A legszükebb konfidencia intervallummal rendelkező kombinációk

sor-szám	Lehetséges kimenetelek azonositása	Gyakoriság az alap halmazban	Gyakoriság a teljes halmazban	Alap halmaz feltételes rel. gyakorisága	Konfidencia intervallum
1.	1 0 1 — — —	167	260	0.64231	0.115750
2.	1 1 1 — — —	209	305	0.68525	0.103695
3.	1 - 1 1 — —	267	373	0.71582	0.091183
4.	- 0 1 1 — —	213	346	0.61561	0.101982
5.	- 1 1 1 — —	346	548	0.63139	0.080524
6.	1 - 1 - 0 — —	251	374	0.67112	0.094807
7.	1 - 1 - 1 — —	125	191	0.65445	0.133687
8.	- 1 1 - 1 — —	161	257	0.62646	0.117471
9.	- - 1 1 0 — —	366	607	0.60297	0.077614
10.	- - 1 1 1 — —	193	287	0.67247	0.107971
11.	1 - 1 - - 0 —	274	423	0.64775	0.090671
12.	1 - 1 - - 1 —	102	142	0.71831	0.146464
13.	- 1 1 - - 1 —	150	226	0.66372	0.122279
14.	- - 1 1 0 1 —	376	620	0.60645	0.076685
15.	- - 1 1 - 1 —	183	274	0.66788	0.110857
16.	- - 1 - 1 1 —	82	123	0.66667	0.164388
17.	1 - 1 - - - —	376	565	0.66549	0.077580
18.	- - 1 - - - 1 —	249	387	0.64341	0.095018
19.	- - 1 - - 1 —	241	371	0.64960	0.096647
20.	- - 1 - - - —	559	894	0.62528	0.063334

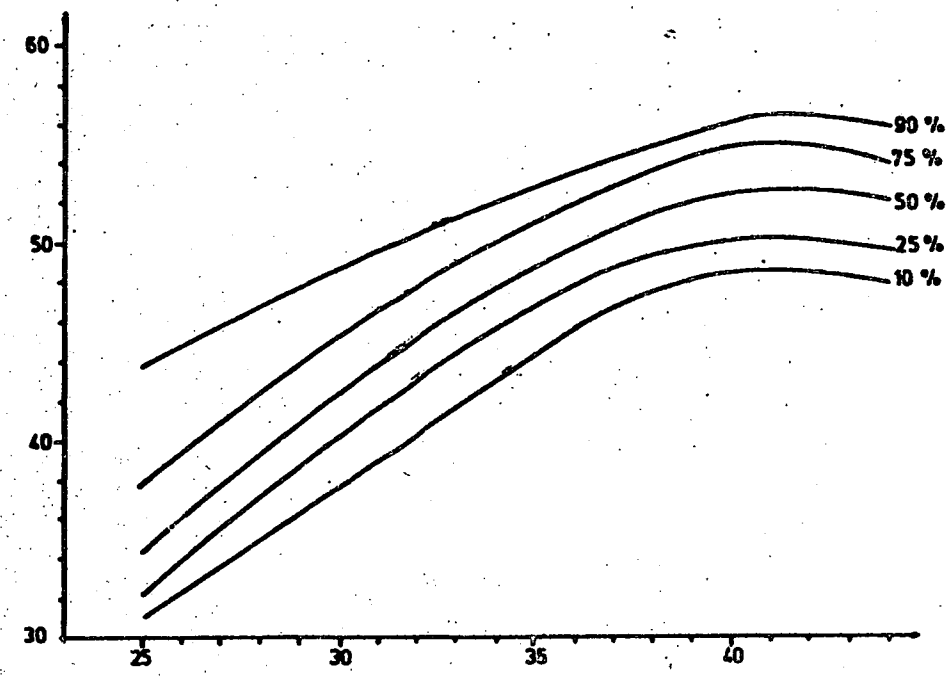
7. sz. táblázat



2. ábra

Az intrauterin súly-fejlődés.
Egész anyag.
Abscissa: gesztációs idő-hét.
Ordinata: g.

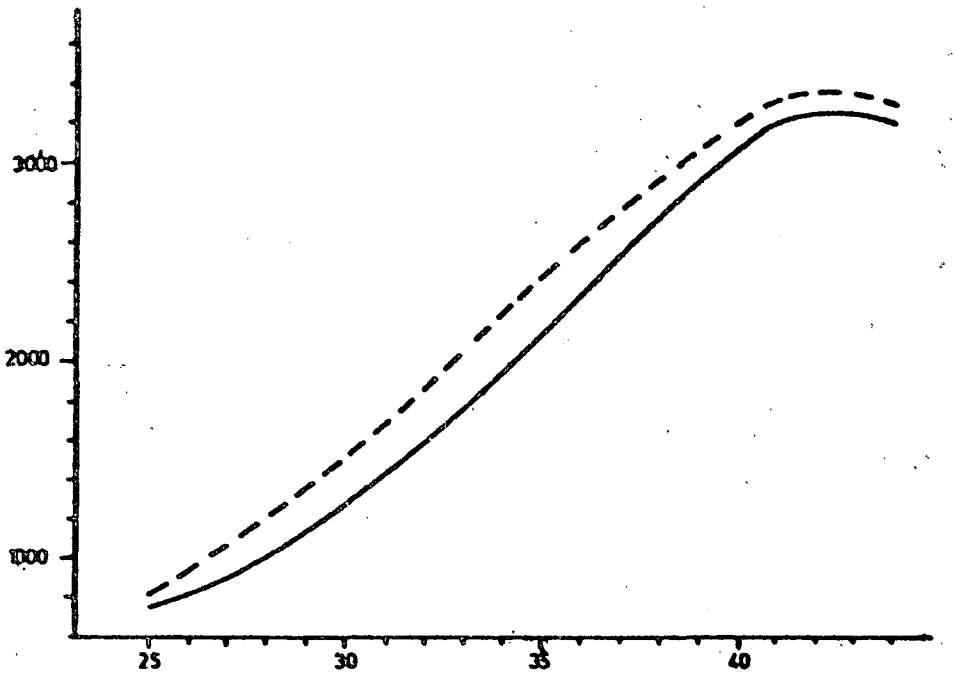
2. ábra



3. ábra

Intrauterin hossz-fejlődés.
Egész anyag.
Abscissa: gesztációs idő-hét.
Ordinata: cm.

3. ábra



4. ábra

Hyperemesis gravidarum szerepe az intrauterin súly-növekedésre
Hyperemesis -, egész minta ---
Abscissa: gesztációs idő-hét.
Ordinata: g.

4. ábra

2.3 *A fejlesztés és alkalmazás során felmerült főbb nehézségek*

Technikaiak:

a./ Alkalmas file-struktura kialakitása /2.modul/,

b./ az algoritmusok nagy méretre tervezése, ill. módositása /valamennyi modul/,

c./ memória- és gépidő megtakaritás /5., 6/b. és 7. modul két bites kódolás, GMAP rutinok/,

d./ programok interaktiv irányithatósága /5. modul/,

e./ közelitő algoritmusok konvergenciájának biztositása.

Informatikaiak:

a./ A közhiedelemmel nem egyező eredmények magyarázata,

b./ szociális okok jelentőségének elismertetése a szakterületen,

c./ a kombinációs vizsgálat eredményeinek interpretálása.

Irodalomjegyzék

(1) Baird, D., Thomson, A.M.: General factors underlying perinatal mortality rates in Perinatal problems. Livingstone. Edinburgh, 1969.

(2) Balogh, G., Götl, Gy., Srajber, B.: Klinikai adatbázis redukciója információ-elméleti módszerrel. Orvos és Technika, 1979.

(3) Char-Tung Lee, R.: Application of Information Theory to Relevant Variables. Math. Biosci. 11: 153, 1971.

(4) Csiszár, I., Fritz, J.: Információelmélet, Tankönyvkiadó, 1976.

(5) Fritz, J.: Az alakfelismerés statisztikus módszerei, MTA MKI, 1974.

(6) Hoel, P.G.: Introduction to Mathematical Statistics, John Wiley and Sons. Inc. New York, 1971.

(7) Kiszel, J.: Perinatalis mortalitás és morbiditás. Magy.Nőorv. L. 42: 5. /1979/

(8) Kryspin, J., Norwich, A.M.: Use of Information Theory in Analysis of Medical Data. IEEE Conference on Inf. Theory, Asilomar, California, 1972.

(9) Lubchenco, L.O és mtsai: Neonatale mortality rate: Relationship to birth weight and gestational age. J. Pediat. 81: 814, 1972.

(10) Meisel, W.S.: Computer-oriented Approaches to Pattern Recognation Acad. Press. New York, 1972.

(11) Mendel, J.M., Fu, K.S.: /szerk./ Adaptive, Learoing and Pattern Recognition System. Acad. Press, 1970.

(12) Prékopa, A.: Valószinüségelmélet. Müszaki Könyvkiadó, 1962.

(13) Sárkány, J.: Csecsemőhalálozás és az anya iskolai végzettsége. Népegészségügy, 53: 109, 1973.

(14) Simonovits, I. és mtsai.: Secular trend in birth length and birth weight of newborne in Hungary 1920-1972. Acta Paed Acad. Sci. Hung. 16: 97, 1975.

(15) Srajber, B. és mtsai: Két betegség-osztály megkülönböztetésére szolgáló matematikai módszerek alkalmazása. Alkalmazott Matematikai Lapok, 6: 219, 1976.

(16) Thomson, A.M.: Foetalis növekedés. Orvosképzés, 48: 7, 1973.

(17) Vincze, I.: Matematikai statisztika ipari alkalmazásokkal. Müszaki Könyvkiadó, Budapest, 1968.

(18) Paksy, A., Srajber, B., Sebők, J., Dabóczi, Á., Kiszel, J.: Multifaktoriális elemző programrendszer a koraszülés okainak és az intrauterin fejlődés tényezőinek a vizsgálatára. Akadémiai pályázat, 1979.