tanulmányok

70/1977

MTA Számitástechnikai és Automatizálási Kutató Intézet Budapest

MAGYAR TUDOMÁNYOS AKADÉMIA
SZÁMITÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZETE

A STATISZTIKAI ADATFELDOGLOZAS MATEMATIKAI ES SZÁMITÁSTECHNIKAI PROBLÉMÁI

Hospitalizált morbiditási statisztikákkal kapcsolatos meggondolások
/Esettanulmány/

Irta:

Krámli András
Ratkó István
Ruda Mihály
Soltész János

Tanulmányok 70/1977.

A kaidásért felelős:

DR VÁMOS TIBOR

ISBN 963 311 052 1

ISSN 0324-2951

Készült az
ORSZÁGOS MŰSZAKI KÖNYVTÁR ÉS DOKUMENTÁCIÓS KÖZPONT
Budapest, VIII., Reviczky u. 6.
Sokszorosító üzemében,
F. v.: Janoch Gyula

TARTALOMJEGYZÉK

B e v e z e t é s

Ez a tanulmány elsősorban nem új statisztikai, vagy számitástechnikai eredmények publikálását tüzi ki célul, hanem olyan adatfeldolgozási kérdéseket érint, melyeknek helyes megoldása matematikai statisztikai meggondolásokat igényel. Ezzel a tanulmánnyal támogatást kivánunk nyujtani azoknak a számitógép-felhasználóknak, akik adatfeldolgozáskor olyan kérdésekkel kerülnek szembe, mint pl, a helyes mintaarány megválasztásának problémája, vagy egy jól használható azonositó kód kialakitásának feladata. A dolgozat - mint esettanulmány - konkrét feladatok megoldásán keresztül mutatja be a tárgyalt módszereket.

A tanulmány fő célja a kórházi morbiditási vizsgálatok kapcsán felmerülő matematikai statisztikai és számitástechnikai kérdések megválaszolása.
A tanulmányban a következő kérdéseket érintjük: mintavétel technikájának kialakitása /ld. 1.pont/, a mintavétellel kapcsolatos megbizhatósági szempontok vizsgálata /ld.2.pont/, azonositó kódok vizsgálata /ld.3.pont/, egyes kódok eloszlásának hatása a rendszer müködésére /ld.4.pont/, clusteranalizis alkalmazási lehetőségeinek ismertetése /ld.5.pont/ és egyéb, a rendszer müködését befolyásoló tényezőkkel kapcso-

latos megjegyzések /ld.6.pont/.
Az előforduló javaslatokat példákkal támasztjuk alá, megmutatva a javaslat célszerüségét.

1. A mintavétel technikája

Statisztikai vizsgálatoknál, a költségek csökkentésének érdekében, az esetek többségében nem készitünk a teljes populációról adatfelvételt, hanem annak csak egy kis hányadáról veszünk mintát. Mintavételnél két fontos szempontot kell figyelembe venni: a minta lehetőleg pontosan a teljes populáció egy meghatározott hányada /pl.10%/ legyen, hogy a mintából könnyen következtethessünk a teljes populációra, és ugyanezért fontos az is, hogy a minta reprezentativ legyen, azaz a vizsgált populáció egyes részei arányosan kerüljenek a mintába.

A hospitalizált morbiditás vizsgálatánál jelenleg 10%-os mintavétel van, amely kórházi szakmánként /osztályonként/ országos összesitésben reprezentativ kell hogy legyen.

1.1. A 10%-os minta kiválasztása

Az évenkénti kórházi morbiditási adatokból tehát osztályonként /országos összesitésben/ 10%-os mintát kell kiválasztani. A minta pontossága az elsődleges cél, még ha ez néhány többszörösen ápolt személy egyes adatainak elveszését vonja is maga után /erről ld.az 1.2. pontot/. Ebben a pontban megadjuk a mintavétel technikáját, továbbá megvizsgáljuk, mit jelent az a többszörösen ápoltakra.

A véletlen mintavétel technikája a következő: a mintába a hónap bizonyos napjain születettek kerülnek be. Korábbi statisztikai vizsgálatok igazolták azt a

természetes feltevést, hogy a morbiditási adatok /betegség, ápolási nap, stb./ függetlenek a születésnaptól.

A következőkben azt a kérdést vizsgáljuk, hogy hány születésnap /havonként/ szükséges a 10%-os minta biztositásához feltéve, hogy a kórházak a kijelölt napokon született valamennyi beteg fejlapját beküldik.
Jelölje M a feldolgozandó év betegeinek számát. A mintába vegyük bele azokat, akiknek a születésnapja 4.-ére, 14.-ére, vagy 24.-ére esik./Természetesen a konkrét napokat másként is ki lehet jelölni./ Az i. osztályról a mintába került betegek száma legyen m_i, $i=1,2,\ldots, 20$ /20 szakma van/.
Mivel a betegek születésnap szerinti eloszlása, ha a hónapot nem vesszük figyelembe, egyenletesnek tételezhető fel /ld.pl. [1], [3] /.

$$P\left(m = m_1+m_2+\ldots + m_{20} \geq \frac{M}{10}\right) = \sum_{k=\frac{M}{10}}^{M} \binom{M}{k} \left(\frac{1}{10}\right)^k \left(\frac{9}{10}\right)^{M-k}$$

A Moivre - Laplace tételből következik, hogy a binomiális eloszlás normálissal közelithető:

$$P\left(m \geq \frac{M}{10}\right) \approx 1 - \Phi\left(\frac{\frac{M}{10} - M \cdot \frac{1}{10}}{\sqrt{M \frac{1}{10} \cdot \frac{9}{10}}}\right) = 0{,}5.$$

Három nap tehát csak 0,5 valószinüséggel /az esetek 50%-ában/ elegendő a 10%-os mintához /a $\Phi(x)$ függvény

a standard normális eloszlásfüggvény/.
Hány napot válasszunk ki tehát?

Az előzőhöz hasonlóan kapjuk, hogy ha két, illetve négy napot választunk ki, akkor

$$P\left(m \geq \frac{M}{10}\right) = \sum_{k=\frac{M}{10}}^{M} \binom{M}{k} \left(\frac{2}{30}\right)^k \left(\frac{28}{30}\right)^{M-k}$$

illetve

$$P\left(m \geq \frac{M}{10}\right) = \sum_{k=\frac{M}{10}}^{M} \binom{M}{k} \left(\frac{4}{30}\right)^k \left(\frac{26}{30}\right)^{M-k}$$

/ha 30 napos hónapokat tételezünk fel/

Ezt a két valószínüséget a Bernstein-egyenlőtlenséggel becsülve kapjuk, hogy két születésnap kiválasztása esetén

$$P\left(m \geq \frac{M}{10}\right) \leq 2 \exp\left(-\frac{28\,M}{5041}\right) ,$$

négy nap kiválasztása esetén

$$P\left(m \geq \frac{M}{10}\right) \geq 1 - 2 \exp\left(-\frac{52M}{14161}\right)$$

A kórházi morbiditási mintavétel osztályonként történik. A legkisebb létszámú intenziv osztályon /az 1972-73. évi adatok szerint/ 1816 beteg feküdt /ld. 1. táblázat/. A Bernstein-egyenlőtlenség becsléseit kiszámitva erre az értékre /M=1816/ azt kapjuk, hogy két születésnapnál

$$P\left(m \geq \frac{M}{10}\right) \leq 0.00008 ,$$

négy születésnapnál

$$P\left(m \geq \frac{M}{10}\right) \geq 0.9974$$

Az ajánlott mintavételi eljárás tehát elég megbizható, hiszen sohasem fordulhat elő az, hogy két születésnap 10 %-nál nagyobb mintát adjon, /ennek valószinüsége kisebb mint 0,00008/ vagy hogy négy születésnap ne legyen elegendő a 10 %-os minta kiválasztásához /feltéve, hogy minden adatlapot beküldtek az egyes osztályok/.

Elegendő tehát minden hónapból négy születésnapot kiválasztani /pl. 4., 6., 14. és 24./. További vizsgálatot igényel az, hogy ha a reprezentativ mintavétel a szakmáknál kisebb egységekre történik, akkor elegendő-e négy születésnaphoz tartozó betegek adatait begyüjteni.

A tanulmányban felhasznált adatok /táblázatok/ az 1972-73 évi kórházi morbiditási vizsgálat adatai. Egy 30, illetve 50 százalékos mintából "felszorzással" nyert adatok, igy nem pontos értékek.

A felhasznált matematikai statisztikai és valószinüség-számitási módszerekkel kapcsolatban ld. pl. az [5] , vagy a [6] könyvet. Ugyanitt találhatók a nevezetes eloszlásfüggvények /pl. normális eloszlás/ táblázatai is.

Ápolási esetek száma az egyes kórházi szakmákban

Sor-szám	S z a k m a	Esetszám	Relativ gyakoriság
1	Belgyógyászat	314715	0.1845
2	R e u m a	11048	0.0065
3	S e b é s z e t	211887	0.1242
4	Traumatológia	45045	0.0264
5	O r t o p é d i a	17802	0.0104
6	U r o l ó g i a	25970	0.0152
7	S z e m é s z e t	40824	0.0239
8	Fül-, orr-, gége	86664	0.0508
9	Fog- és szájsebészet	5518	0.0032
10	Szülészet, nőgyógyászat	497268	0.2915
11	Gyermekgyógyászat	163173	0.0957
12	F e r t ő z ő	56931	0.0334
13	I d e g	45504	0.0267
14	Onkoradiológia	11012	0.0065
15	Bőr- és nemibeteg	21018	0.0123
16	I n t e n z i v	1816	0.0010
17	T B C	63514	0.0372
18	E l m e gyógyászat	40110	0.0235
19	Krónikus utókezelő	5376	0.0032
20	S z a n a t ó r i u m	40384	0.0237

1. táblázat

A mintavétel technikája a következő:

A kórházak négy születésnap betegeinek lapjait küldik el.Ezekből a számitógép állitja össze a 10%-os mintát.

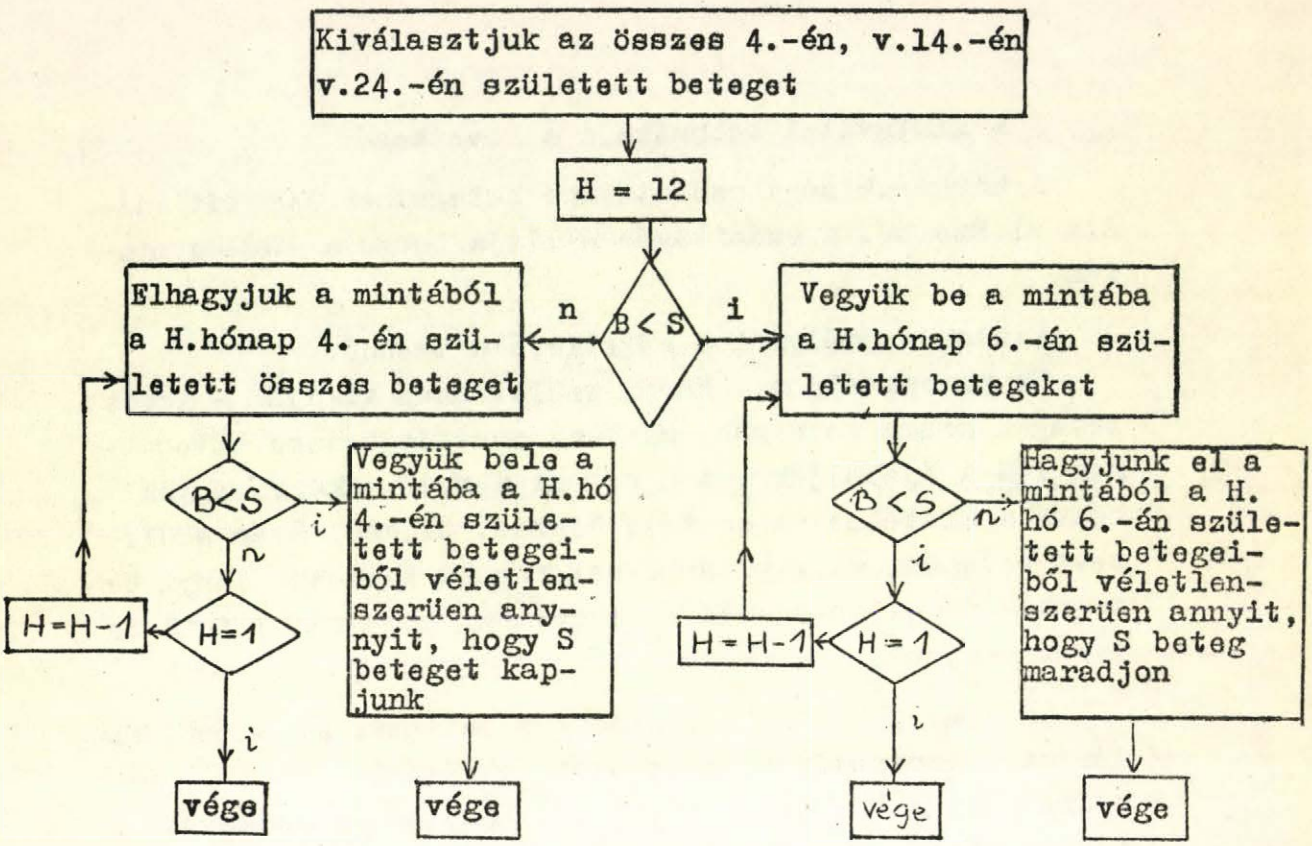
Minden osztálynál a következő a teendő:

Ha az osztályon - három születésnap alapján - kapott betegek száma kevesebb, mint az osztály összes betegeinek 10%-a /jelöljük ezt a számot S-sel/, akkor vegyük hozzá a mintához az osztály összes, pl.dec. 6-án született betegét. Ha igy már S-nél többet kapnánk, annyi beteget - véletlenszerüen - elhagyunk, hogy végül is S beteget kapjunk.

Ha még ezek után sem kapunk S beteget, az előbbi eljárást megcsináljuk a novemberi, októberi, ..., januári 6-án született betegekre /ahány hónap szükséges/. A pont elején végzett számitásokból következik, hogy ilymódon 1 valószinüséggel 10%-os mintához jutunk.

Ha az osztályon - a három születésnap alapján - kapott betegek száma nagyobb, mint S, ugyanúgy járunk el, mint az előbb, de most elhagyás helyett hozzávétel és hozzávétel helyett elhagyás értendő, s ekkor mondjuk a 4-én születettekkel kell operálni. /Ekkor is 1 valószinüséggel eljutunk a 10%-os mintához/.

A mondottakat blokkdiagrammal is leirhatjuk:



A mintavétel technikájának folyamatábrája

/adott osztálynál/

Jelölések: H : hónapszám /1 - 12/

B : a mintába belevett, utolsó utasításnak eleget tevő betegek száma

S : Az osztály összes betegei számának 1/10-e.

1. 2. A többszörös ápolással kapcsolatos mintavételi problémák

1. 2. 1. A probléma megfogalmazása

Azt fogjuk megvizsgálni, milyen hibák adódnak, amikor a mintánk alapján a többszörösen ápolt betegek számát akarjuk megbecsülni. Most csak a speciális kérdésfeltevésből és a mintavétel sajátosságaiból adódó hibákkal fogunk foglalkozni.

Csak a legegyszerűbb kérdést tárgyaljuk azt, hogy hány olyan beteg van, akit előbb egy A-val jelzett osztályon ápoltak, majd még ugyanebben az évben a B osztályon kezeltek.

Mint tudjuk, a mintavétel olyan, hogy tetszőleges C osztály esetén ha ott M_C esetet kezeltek, akkor a mintába ezekből $M_C/10$ eset kerül.

Tegyük fel, hogy egy A osztályon ápolt olyan esetek száma , amelyeknél a beteg 4-én, 14-én, vagy 24-én született, kisebb, mint 0.1 M_A. Ekkor a mintába beveszünk még néhány 6-án született, A osztályon kezelt beteget. Tegyük fel továbbá, hogy a 4-én, 14-én, vagy 24-én született B osztályon kezelt betegek eseteinek száma nagyobb mint 0,1 M_B /ekkor el kell hagynunk néhány 4-én született beteg esetét/. Számoljuk most összze, hogy a mintában hány olyan beteg van, akit előbb az A osztályon, majd a B osztályon kezeltek. /Az ilyen embereket a továbbiakban AB betegeknek fogom hivni/.

A fenti feltevések esetén a mintából az AB betegekre adódó becslés valószinüleg kisebb lesz a pontos értéknél, mivel elvesznek azok a betegek, akik 4-én születtek, de a B osztály mintájából kihagytuk őket. Akik 6-án születtek és bekerültek az A osztály mintájába, azok is elvesznek, ugyanis a minta alapján nem lehet megállapitani, hogy őket a későbbiek során a B osztályon kezelték.

A most ismertetett jelenségből adódó hibát fogjuk a továbbiakban vizsgálni.

1.2.2. A modell

A következő modellel fogunk dolgozni:

Csak egyszer és kétszer kezelt betegek vannak, /a 2-nél többször ápoltak száma elhanyagolható, az ebből adódó hiba egy nagyságrenddel kisebb, mint az általunk adott becslés hibája/.

Az ápolási esetek le vannak rendezve, elsősorban születési nap szerint /legelől vannak a 14-én, majd a 24-én, 4-én, 6-án, stb. született emberek/, majd egyéb azonositóik szerint /születési év, hó, név, anyja neve, stb./. Igy minden kétszer ápolt beteg 2 esete egymás mellé kerül. Ez a feltevés nem jelent megszoritást a kórházi morbiditási adatok statisztikai viselkedésére vonatkozóan. A további feltételek a tapasztalattal nagymértékben egyező, de idealizált esetet irnak le.

Egy beteg p_1, p_2, ..., p_{20} valószinüséggel kerül az 1., 2., ..., 20. osztályra. Ha kétszer kezelik, akkor a második alkalommal az elsőtől függetlenül kerül p_1, ..., p_{20} valószinüséggel a megfelelő osztályra.

Egy beteg i-edikén $\frac{1}{30}$ valószinüséggel születik /i=1, 2, ..., 30/, függetlenül attól hányszor és melyik osztályon kezelik.

Ezt a modellt például a következő módon épithetjük fel: először kisorsoljuk a kétszeres esetek helyét rendezett populációnkban ugy, hogy ezek párosával legyenek, és a kétszeres esetek "egyenletesen" helyezkedjenek el az egyesek közt. Ezután minden esetről kisorsol-

juk p_1, ..., p_{20} valószinüséggel, hogy a beteget melyik osztályon kezelték. Végül összeszámoljuk, hány betegünk van, /ez egy M-nél kisebb szám lesz/, és kisorsolunk annyi születésnapot. Ha s_1 beteg született elsején, ..., s_{30} 30-án, akkor azt mondjuk, hogy a rendezett populációban szereplő első s_{14} beteg 14-én, a következő s_{24} 24-én született, stb. Jól látható, hogy a modell felépitésében egyetlen pont okoz problémát, a kétszeres esetek kisorsolása. Most ezt fogjuk részletezni.

Végezzünk független kisérleteket, melyeknek eredménye p valószinüséggel egy C esemény. Ha nem következik be a C esemény /ennek 1-p a valószinüsége/, akkor azt mondjuk, hogy a rendezett populációban egy egyszeres eset következik. Ha C bekövetkezik, akkor egy kétszer ápolt beteg két esete van a populációban.

Addig végezzük a kisérleteket, amig az M hely betelik. Előfordulhat, hogy amikor az M-edik helyet akarjuk betölteni, akkor a sorsolásnál C bekövetkezik, és az M-edik helyre egy kétszeres ápolás első esete kerül, és a második esetet nem tudjuk hová tenni, mivel nincs több hely a populációban. Mivel M egy nagyon nagy szám, mindegy, hogy az M-edik helyen levő esetet egy kétszeres ápolás egyik esetének tekintjük-e vagy sem.

Jelölje ν a C esemény gyakoriságát / $0 \leq \nu \leq \frac{M}{2}$ /.

Legyen $E \cdot \nu = M \cdot \pi$ /*/

Ha van egy mintánk, akkor annak alapján π megbecsülhető. Most azt számitjuk ki, hogy ha π -t megadjuk, akkor hogyan lehet p-t ugy megválasztani, hogy /*/ fennálljon.

Legyen az 1. kétszeres ápolás 2. esetének sorszáma Y_1, a 2. kétszeres ápolás 2. esetének sorszáma Y_1+Y_2, az utolsóé pedig $Y_1+Y_2+\ldots+Y_\nu$. Ekkor az Y_i változók függetlenek és

$$P\{Y_i = k\} = /1-p/^{k-2}\, p \qquad k=2,3,4,\ldots.$$

azaz Y_i= egy elsőrendü negativ binomiális eloszlásu valószinüségi változó + 1 . Igy

$$E\, Y_i = \frac{1}{p} + 1 = \frac{p+1}{p}$$

Az u.n. elemi felujitási tétel alapján /lásd pl. [14] 116.oldal/ nagy M-re

$$E\nu \cdot EY_1 \approx M$$

Igy, ha pontos egyenlőséget veszünk

$$M\Pi = E\nu = \frac{M}{EY_1} = M\,\frac{p}{1+p}$$

$$\Pi = \frac{p}{1+p}$$

$$p = \frac{\Pi}{1-\Pi}$$

Π tulajdonképpen annak a valószinüsége, hogy egy eset egy kétszeresen ápolt beteg első esete.

1.2.3. Egy segédfeladat

Az A osztályon $M_A=M_{p_A}$ esetet kezeltek. Ezek közül a mintába $h_A=\frac{M_A}{10}$ kerül be. A mintavételt ugy végezzük, hogy elindulunk a rendezett populáció elejéről, és minden, az A osztályon kezelt esetet beveszünk a mintába, egészen addig, amig h_A esetünk nem lesz. Jelöljük r_A-val az A

osztály mintájába bekerülő utolsó eset sorszámát. Először az r_A valószinüségi változó eloszlását fogjuk pontosan, majd közelitőleg meghatározni.

Jelöljük x_i-vel két szomszédos A osztályon kezelt eset távolságát, azaz legyenek az A osztályon kezelt esetek az x_1-edik, $/x_1+x_2/$-edik, $/x_1+x_2+x_3/$-adik, stb. helyen rendezett populációnkban.

Az x_i valószinüségi változók függetlenek és elsőrendü negativ binomiális eloszlásuak, igy

$$P\left\{x_i=k\right\} = /1-p_A/^{k-1}\, p_A \qquad k=1,2,\ldots$$

$$E\, X_i = \frac{1}{p_A}$$

$$D^2\, x_i = \frac{1-p_A}{p_A^2}$$

Mivel $r_A=x_1+x_2+\ldots+x_{h_A}$

$$E\, r_A = \frac{h_A}{p_A} = \frac{M_A}{10\, p_A} = \frac{M\, p_A}{10 p_A} = \frac{M}{10}$$

$$D^2 r_A = h_A \;\; D^2\, x_i = \frac{M\, p_A}{10}\;\; \frac{1-p_A}{p_A^2} = \frac{M/1-p_A/}{10\, p_A}$$

és r_A h_A-ad rendü negativ binomiális eloszlásu változó A centrális határeloszlás-tétel alapján /lásd [13] 372. oldal/.

$$\lim_{M\to\infty} P\left\{\frac{r_A-Er_A}{Dr_A} < x\right\} = \Phi/x/$$

ahol $\Phi(x)$ a 0 várható értékü 1 szórásu normális eloszlásu változó eloszlásfüggvénye.
Igy azt mondhatjuk, hogy r_A eloszlása közelitőleg $\frac{M}{10}$ várható értékü $\sqrt{\frac{M(1-p_A)}{10p_A}}$ szórásu normális eloszlás.

Ez a közelités elég pontos, hiszen $h_A = \frac{M_A}{10} \geqslant 180$ változót adtunk össze /ld..l. táblázat/.

Legyen $\mu_A = \frac{2Dr_A}{Er_A}$.

Mivel $\Phi(2)=0.9772$, azt állithatjuk, hogy r_A az $(Er_A-2Dr_A,\ Er_A+2Dr_A) = (\frac{M}{10} - \mu_A\frac{M}{10},\ \frac{M}{10} + \mu_A\frac{M}{10})$ intervallumban lesz $2\Phi(2)-1 = 0.9544$ valószinüséggel.
A μ_i $(i=1,2,\ldots,20)$ számokat az alábbi táblázat tartalmazza /2.táblázat/.

Osztálykód /i/		μ_i	$100\,\mu_i$ /%/
Belgyógyászat	1	0.01018	1,0
Reuma	2	0.05998	6,0
Sebészet	3	0.01286	1,3
Traumatológia	4	0.02940	2,9
Ortopédia	5	0.04715	4,7
Urológia	6	0.03895	3,9
Szemészet	7	0.03093	3,1
Fül-orr-gége	8	0.02093	2,1
Fog és szájseb.	9	0.08500	8,5
Szülészet, nőgyógy.	10	0.00755	0,8
Gyermekgy.	11	0.01489	1,5
Fertőző	12	0.02606	2,6
Ideg	13	0.02925	2,9
Onkoradiológia	14	0.06007	6,0
Bőr és nemibeteg.	15	0.04336	4,3
Intenziv	16	0.14833	14,8
TBC	17	0.02462	2,5
Elme	18	0.03121	3,1
Krónikus	19	0.08612	8,6
Szanatórium	20	0.03110	3,1

2. táblázat

A második oszlop azt mutatja, hogy a $\mu_i \frac{M}{10}$ hibahatár az $\frac{M}{10}$ várható értéknek hány százaléka.

Nagy esetszám esetén /i=10,1,3,11/ a 100 μ_i számok 0,75% és 1,5% között vannak, mig kis esetszám esetén /i=2,14,9,19,16/ 5,9% és 14,9% között találhatók.

Mivel $\Phi/1/ = 0,8413$, azt mondhatjuk, hogy

r_A az $/Er_A - Dr_A , Er_A + Dr_A/ =$

$= /\frac{M}{10} - \frac{\mu_A}{2}\frac{M}{10} , \frac{M}{10} + \frac{\mu_A}{2}\frac{M}{10}/$ intervallumon kivül van elég nagy, $2/1-\Phi/1// = 0,3174$ valószinüséggel.

A $100\frac{\mu_A}{2}$ számok "kis" osztályok esetén elég nagyok, 2,9% és 7,5% közé esnek.

Várható, hogy ha mind A, mind B "nagy" osztály, akkor $r_A \approx r_B \approx \frac{M}{10}$ lesz, és igy ha egy AB beteg A esete /azaz az A osztályon való kezelése/ bekerül az A osztály mintájába /azaz ennek az A esetnek a sorszáma a rendezett populációban kisebb, mint r_A, ami körülbelül $\frac{M}{10}$/, akkor ennek az AB betegnek a B esete is majdnem mindig bekerül a B osztály mintájába. A fenti állitás megforditottja is igaz: ha egy AB beteg B esete bekerül a B osztály mintájába, akkor ennek a betegnek az A esete is majdnem mindig benne van az A osztály mintájában. Ezek szerint kevés AB beteg fog elveszni. /lásd az 1.2.1. pontot/. Igy várható, hogy az AB betegek számára vonatkozó becslés elég pontos lesz, ha A is és B is "nagy" osztály.

Ha viszont legalább az egyik osztály kis esetszámu, akkor a becslés hibája már nagyobb lehet viszonylag nagy valószinüséggel.

1.2.4. A feltételes hiba

Most azt fogjuk megvizsgálni, hogy várhatólag mekkora lesz a becslési hiba, ha ismerjük az r_A és az r_B változók értékét. Becslésünk nyilván az lesz, hogy összeszeszámoljuk, hogy a mintában hány AB eset van, és ezt a számot megszorozzuk tizzel.

Mekkora a pontos érték? A mintában $M \cdot \Pi$ kétszeres eset van. Ha találtunk egy kétszeres kezelést, akkor annak a valószinüsége, hogy az első A eset, p_A, annak hogy a második B eset, p_B. Mivel modellünkben minden független, az AB esetek számának várható értéke $M \Pi p_A p_B$.

Mekkora lesz a becslés? Rendkivül kicsi valószinüséggel előfordulhat például az, hogy $r_A = h_A$, azaz a rendezett populációban az első h_A eset A eset. Ha B=A, akkor a mintában 0 vagy 1 AB eset lesz, azaz a becslési hiba nagyon nagy lesz. Az r_A eloszlására kapott közelités szerint azonban az ilyen extrém esetek valószinütlenek.

Ha B=A, akkor a mintában körülbelül $\frac{M}{10} \Pi$ kétszeres eset van. A korábbihoz hasonló gondolatmenettel adódik,

hogy a mintában szereplő AA esetek számának várható értéke $\frac{M}{10}\pi p_A^2$, igy a várható hiba $10/\frac{M}{10}\pi p_A^2/-M\pi p_A^2=0$ lesz.

A továbbiakban csak a B≠A esettel foglalkozunk.

Legyen tehát r_A és r_B adott és tegyük fel, hogy $r_A < r_B$.

A mintában pontosan $h_A = \frac{Mp_A}{10}$ A eset van. Ezek közül körülbelül $\frac{Mp_A}{10}\pi$ lesz egy kétszeres kezelés első esete. Tegyük fel, hogy a B esetek sürüsége a $/0,r_A/$ szakaszon ugyanannyi, mint a $/0,r_B/$ szakaszon, azaz

$\frac{h_B}{r_B} = \frac{M}{10} p_B \frac{1}{r_B}$. Ez a feltevés nem jogos például a korábban emlitett extrém esetben, de elég jó közelités akkor, ha r_A az $I_A=/\frac{M}{10} - \mu_A \frac{M}{10}$, $\frac{M}{10} + \mu_A \frac{M}{10}/$ intervallumba esik. Ugyanis ha az A osztály "kicsi", akkor az I_A intervallum ugyan elég nagy, de mivel p_A nagyon kicsi, a $/0,r_A/$ szakaszon nagyon kevés hely lesz lefoglalva A esettel, igy r_A értéke szinte semmivel sem befolyásolja a B esetek sürüségét. Ha viszont A egy "nagy" osztály, akkor az I_A intervallum olyan kicsi, hogy r_A pontos értékével szinte semmit sem nyerünk. Feltéve tehát, hogy a B esetek sürüsége a $/0,r_A/$ szakaszon $\frac{M}{10} p_B \frac{1}{r_B}$, az adódik, hogy a mintában levő AB esetek számának várható értéke

$$/\frac{M}{10} p_A \pi / \frac{M}{10} p_B \frac{1}{r_B} = /\frac{M}{10} \pi p_A p_B / /\frac{M}{10} \frac{1}{r_B}/$$

Tehát a várható relativ hiba $r_A < r_B$ esetén

$$\frac{10 /\frac{M}{10} \pi p_A p_B / /\frac{M}{10} \frac{1}{r_B}/ - M \pi p_A p_B}{M \pi p_A p_B} = \frac{\frac{M}{10}}{r_B} - 1$$

Legyen most $r_B < r_A$.

A $/0, r_A/$ intervallumban $h_A = \frac{M}{10} p_A$ A eset van. Tegyük fel, hogy a $/0, r_B/$ intervallumba ezeknek $\frac{r_B}{r_A}$ -ad része esik, azaz ott $\frac{M}{10} p_A \frac{r_B}{r_A}$ A eset van. /Ezen közelités helyessége az $r_A < r_B$ esethez hasonló módon indokolható/.

Mivel a $/0, r_B/$ szakaszon körülbelül $\frac{M}{10} p_A \frac{r_B}{r_A} \pi$ A esettel kezdődő kétszeres kezelés van, és a B esetek sürüsége $\frac{h_B}{r_B} = \frac{M}{10} p_B \frac{1}{r_B}$ a $/0, r_B/$ szakaszon, a mintában levő AB esetek számának várható értéke

$$\frac{M}{10} p_A \frac{r_B}{r_A} \pi \frac{M}{10} p_B \frac{1}{r_B} = /\frac{M}{10} \pi p_A p_B / /\frac{M}{10} \frac{1}{r_A}/$$

Tehát a várható relativ hiba $r_B < r_A$ esetén

$$\frac{10 /\frac{M}{10} \pi p_A p_B / /\frac{M}{10} \frac{1}{r_A}/ - M \pi p_A p_B}{M \pi p_A p_B} = \frac{\frac{M}{10}}{r_A} - 1$$

Összefoglalva adott r_A, r_B esetén a feltételes várható

relativ hiba közelitőleg

$$\widehat{VRH} = \frac{\frac{M}{10}}{\max/r_A, r_B/} - 1$$

1.2.5. Az eredmény értékelése

Az r_A és az r_B változók közelitőleg függetlenek. Ahol a későbbiekben az $\stackrel{f}{=}$ jel látható, ott használjuk ki r_A és r_B /feltételezett/ függetlenségét. Azonban mindig adunk olyan becsléseket is, amelyek levezetése során nem tesszük fel r_A és r_B függetlenségét.

Látható, hogy amennyiben $\max/r_A, r_B/$ kisebb, mint 0.1 M, akkor $\widehat{VRH}$ pozitiv lesz. Ennek valószinüsége

$$P/\max/r_A, r_B/ < 0.1\ M/ = P/r_A < 0.1\ M\ ,\ r_B < 0.1\ M/ \stackrel{f}{=}$$

$$\stackrel{f}{=} P/r_A < 0.1\ M/\ P/r_B < 0.1\ M/ = \Phi/0/\Phi/0/ = \frac{1}{4}$$

A továbbiakban megvizsgáljuk, hogyan lehet $\widehat{VRH}$-ra konfidenciaintervallumot adni az r_A és r_B változók eloszlására nyert eredmények segitségével.

Legyen $\varepsilon > 0$. Ekkor

$$\frac{0.1\ M}{0.1\ M + \varepsilon} - 1 < 0$$

Legyen továbbá

$$P_{alsó} = P/\widehat{VRH} < \frac{0.1\ M}{0.1\ M + \varepsilon} - 1/ = P/\max/r_A, r_B/ > 0.1\ M + \varepsilon/ =$$

$$= P\left\{r_A > \frac{M}{10} + \varepsilon \text{ vagy } r_B > \frac{M}{10} + \varepsilon\right\} \leq P/r_A > \frac{M}{10} + \varepsilon / + P/r_B > \frac{M}{10} + \varepsilon /$$

Mivel $\widehat{VRH}$ r_A-ban és r_B-ben szimmetrikus, az általánosság megszoritása nélkül feltehetjük, hogy $Dr_B \leq Dr_A$. Ha $\varepsilon > 0$, akkor

$$\frac{0.1\ M}{0.1\ M \cdot \varepsilon} - 1 > 0 .$$

Legyen

$$P_{felső} = P/\widehat{VRH} > \frac{0.1\ M}{0.1\ M - \varepsilon} - 1/ = P\left\{\max/r_A, r_B/ < 0.1\ M - \varepsilon\right\} =$$

$$= P\left\{r_A < 0.1\ M - \varepsilon,\ r_B < 0.1\ M - \varepsilon\right\} \leq P\left\{r_B < 0.1\ M - \varepsilon\right\}$$

Ha feltesszük, hogy r_A és r_B független, akkor

$$P_{felső} = P\left\{r_A < 0.1\ M - \varepsilon,\ r_B < 0.1\ M - \varepsilon\right\} \overset{f}{=}$$

$$\overset{f}{=} P\left\{r_A < 0.1\ M - \varepsilon\right\}\ P/r_B < 0.1\ M - \varepsilon / \leq$$

$$\leq P/r_A < 0.1M/\ P/r_B < 0.1\ M - \varepsilon / = \frac{1}{2}\ P/r_B < 0.1\ M - \varepsilon /$$

Az $\overset{f}{=}$ utáni szám természetesen kedvezőbb, mint $\frac{1}{2}\ P/r_B < 0.1\ M - \varepsilon /$, az utóbbi azonban gyorsabban számolható. A két szám nem nagyon sokkal tér el egymástól, ha Dr_B jóval kisebb, mint Dr_A.

Lássunk most egy számpéldát. Legyen A a tizenhatos osztály /azaz a "legkisebb" osztály/, B pedig az egyes /B a második "legnagyobb" osztály/. Legyen először

$$\varepsilon = \mu_A \frac{M}{10} = 2\ Dr_A \qquad \text{Ekkor}$$

$$P\,/r_A > 0.1\,M + \varepsilon\,/ = P\,/r_A > Er_A + 2\,Dr_A/ =$$

$$= P\,/\frac{r_A - Er_A}{Dr_A} > 2/ = 1 - \Phi/2/ = 0.0228$$

$$P\,/r_B > 0.1\,M + \mu_A \frac{M}{10}/ = P\,/r_B > Er_B + \frac{\mu_A}{\mu_B}\mu_B \frac{M}{10}/ =$$

$$= P\,/r_B > Er_B + \frac{\mu_A}{\mu_B}\,/2\,Dr_B// \leq P\,/r_B > Er_B + 20\,Dr_B/ =$$

$$= 1 - \Phi/20/ \leq 10^{-6}$$

A levezetés során kihasználtuk, hogy $\frac{\mu_A}{\mu_B} = \frac{0.14833}{0.01018} > 10$.

Tehát

$$P_{alsó} \leq P\,/r_A > 0.1\,M + \varepsilon\,/ + P\,/r_B > 0.1\,M + \varepsilon\,/ = 0.0228$$

Mivel

$$\frac{0.1\,M}{0.1\,M + \varepsilon} - 1 = \frac{0.1\,M}{0.1\,M + \mu_A\,0.1\,M} - 1 = \frac{1}{1 + \mu_A} - 1 =$$

$$= -\,0.1292\quad ,$$

azt nyertük, hogy

$$P_{alsó} = P\,/\hat{VRH} < \frac{0.1\,M}{0.1\,M + \varepsilon} - 1/ = P\,/\hat{VRH} < -0.1292/ \leq 0.0228$$

Legyen most $\varepsilon = \mu_B \frac{M}{10}$. Ekkor

$$\frac{0.1\,M}{0.1\,M - \varepsilon} - 1 = \frac{1}{1 - \mu_B} - 1 = 0.0103$$

$$P_{felső} = P\,/\hat{VRH} > \frac{0.1\,M}{0.1\,M - \varepsilon} - 1/ = P\,/\hat{VRH} > 0.0103/ \leq$$

$$\leq P\,/r_B < 0.1\,M - \varepsilon\,/ = P\left\{ r_B < Er_B - 2\,Dr_B \right\} = \Phi/-2/ =$$

$= 1 - \Phi/2/ = 0.0228$

Ha még r_A és r_B függetlenségét is feltesszük

$$P_{felső} \overset{f}{=} P\ /r_A < 0.1\ M - \varepsilon\ /\ P\ /r_B < 0.1\ M - \varepsilon\ / =$$

$$= P\ /r_A < Er_A - \frac{2\mu_B}{\mu_A}\frac{\mu_A}{2}\frac{M}{10}/\ /1 - \Phi/2// =$$

$$= \Phi\ / - \frac{2\mu_B}{\mu_A}/\ /1 - \Phi/2// = /1 - \Phi\ /\frac{2\mu_B}{\mu_A}//\ /1 - \Phi\ /2// =$$

$$= /1 - \Phi/0.1373//\ /1 - \Phi/2// = 0.4454 \cdot 0.0228 = 0.0101$$

Ha a másik eredményünket használjuk

$$P_{felső} \overset{f}{=} P\ /r_A < 0.1\ M - \varepsilon\ /\ P\ /r_B < 0.1\ M - \varepsilon\ / \leq$$

$$\leq \frac{1}{2}\ P\ /r_B < 0.1\ M - \varepsilon\ / = \frac{1}{2}\ /1 - \Phi/2// = 0.0114$$

Mivel Dr_B jóval kisebb, mint Dr_A, az utóbbi két becslés /0.0101 és 0.0114/ nem sokkal tér el. Összeolvaszthatjuk egy képletbe azt a két becslést, melyeknél nem tettük fel r_A és r_B függetlenségét:

$$P\ /-0.1292 < \hat{VRH} < 0.0103/ = 1 - P\ /\hat{VRH} < -0.1292/ -$$

$$- P\ /\hat{VRH} > 0.0103/ \geq 1 - 0.0228 - 0.0228 = 0.9444$$

Könnyü belátni, hogy $\hat{VRH}$ elég nagy valószinüséggel vesz fel viszonylag nagy értékeket:

$$P\ /\hat{VRH} < \frac{0.1\ M}{0.1\ M + Dr_A} - 1/ = P\ /\frac{0.1\ M}{max/r_A, r_B/} - 1 < \frac{0.1\ M}{0.1\ M + Dr_A} - 1/ =$$

$$= P\ /\max/r_A, r_B/ > 0.1\ M + Dr_A/ \geqslant P\ /r_A > Er_A + Dr_A/ =$$

$$= P\ / \frac{r_A - Er_A}{Dr_A} > 1/ = 1 - \Phi\ /1/ = 0.1587$$

Mivel $Dr_A = \frac{\mu_A}{2} \frac{M}{10}$,

$$\frac{0.1\ M}{0.1\ M + Dr_A} - 1 = \frac{1}{1 + \frac{\mu_A}{2}} - 1 = 0.0690$$

Igy

$$P\ /\hat{VRH} < -0.0690/ \geqslant 0.1587$$
$$P\ /|\hat{VRH}| > 0.0690/ \geqslant 0.1587$$

Tehát a várható relativ hiba abszolut értéke 6,9%-nál nagyobb legalább 0.1587 valószinüséggel, ha A a tizenhatos, B pedig az egyes osztály.

2. A mintavétellel kapcsolatos megbizhatósági kérdések

Nyilvánvalóan igen fontos tudni azt, hogy a mintavétel alapján nyert táblázatok, értékek milyen megbizható eredményeket szolgáltatnak. Minden lehetséges esetre ennek megadása lehetetlen, igy csupán arra törekszünk, hogy leirjuk a felvethető kérdéseket /ld. 2.1. pont/, a kérdések megválaszolását szolgáló módszereket /ld. 2.2. pont/, továbbá konkrét példákat adjunk /ld. 2.3. pont/.

2.1. A felvethető kérdések

/a/ Leggyakrabban az a feladatunk, hogy becslést adjunk annak p valószinüségére, hogy egy beteg valamely előre adott tulajdonsággal rendelkezik, pl. a beteg Pest megyei v. adott kórformáju betegséggel ápolták, stb. Másszóval ez pl. a következőt jelenti: 95%-os biztonsággal állithatjuk, hogy a pestmegyei betegek száma 14200 és 14800 közé esik, stb.

/b/ Feltételezve, hogy kórházainkban évente kb. 1.700.000 beteget ápolnak, felmerül a kérdés, hogy adott megbizhatósági szint /adott hibavalószinüség és hibakorlát/ esetén hány %-os mintára van szükségünk.

/c/ Szükség lehet arra, hogy eldöntsük adott H_0 feltevés /pl. a szellemi dolgozók 30%-a infarktusban hal meg, vagy a születésnapok eloszlása egyenletes/,

u.n. nullhipotézis adott szinten elfogadható-e. Megadandó továbbá a H_0-t elutasitó u.n. kritikus tartomány. /Ilyen kérdésekről ld. pl. [7]/

/d/ Ha az /a/ kérdést szeretnénk megválaszolni azokban az esetekben, amikor a "tulajdonság" rendre az, hogy: a beteg négyjegyü BNO kódja 0001, 0002, ... , 9998, 9999, és az ott követendő eljárást alkalmaznánk most is, sok és felesleges számolást végeznénk. Ehelyett a Kolmogorov eloszlás alapján konfidencia sávot adunk az eloszlásfüggvényre. Hangsulyozni szeretnénk, hogy tulajdonképpen az egyes valószinüségekre adunk becslést, csak más módon, mint az /a/ pontban. Ugyancsak ezt az eloszlást használhatjuk annak eldöntésére, hogy kapott eredményeink mennyire egyeznek meg régebbi eredményeinkkel vagy külföldi eredményekkel.

/e/ Homogenitás vizsgálat alkalmazása is felmerülhet: állandó lakóhely megyéje, születési hely megyéje azonos eloszlásunak tekinthető-e.

/f/ Két tényező, amelyek egymásrahatása feltételezhető, függetlennek vehető-e; pl. beteg és beteg édesanyja keresztnevének kezdőbetüje, nem v. kor és bizonyos betegségek, keresztnév kezdőbetüje és nem, stb.

2.2. Az alkalmazott módszerek

a/-ban, b/-ben, c/-ben, ... stb. rendre a 2.1. pont a, b, c, ... stb.-ben felvetett kérdésre alkalmazható módszereket ismertetjük.

a/ A feladat nem más, mint egy rögzitett A esemény p=P/A/ valószinüségére adott 1- megbizhatósági szintü konfidenciaintervallum megadása.
Ha M nagy, a következőképpen járhatunk el: jelölje r_M az A esemény relativ gyakoriságát, ekkor

$$\frac{r_M + \frac{U_\varepsilon^2}{2M} - \frac{U_\varepsilon}{\sqrt{M}}\sqrt{r_M/1-r_M/ + \frac{U_\varepsilon^2}{4M}}}{1 + \frac{U_\varepsilon^2}{M}} < p \leqslant \qquad /1/$$

$$\leqslant \frac{r_M + \frac{U_\varepsilon^2}{2M} + \frac{U_\varepsilon}{\sqrt{M}}\sqrt{r_M/1-r_M/ + \frac{U_\varepsilon^2}{4M}}}{1 + \frac{U_\varepsilon^2}{M}}$$

intervallum megbizhatósági szintje $1-\varepsilon$, ahol U_ε-t az $1-\varepsilon = 2\Phi/U_\varepsilon/-1$ (Φ itt is a standard normális eloszlásfüggvény) összefüggésből állapithatjuk meg.

b/ Most tehát $/1-\varepsilon/$ megbizhatósági szintü δ nagyságu konfidenciaintervallum megszerkesztéséhez kell meghatároznunk az M értékét.

/1/ felhasználásával bizonyitható a következő:

Ha M elég nagy és teljesül a következő egyenlőtlenség:

/2/ $M \geqslant \Phi^{-1}/1 - \frac{\varepsilon}{2}/^2 \frac{1-2\delta}{2}$, ahol Φ^{-1} Φ inverzét jelöli, akkor az a/ pont alapján szerkesztett

/1/ konfidenciaintervallum hossza δ .

c/ Vizsgáljuk az alábbi nullhipotézist:

H_o: $P/A_i/ = /i = 1,2,\ldots,k$; $p_1+p_2+\ldots+p_k=1/$, ahol $A_1, A_2, \ldots, A_k$ teljes eseményrendszert alkot. N számu megfigyelést végezve, tegyük fel, hogy az A_i esemény ν_i-szer következik be.

Nyilván $\sum_{i=1}^{k} \nu_i = N$ és a ν_i valószinüségi változók binomiális eloszlásuak.

Belátható, hogy a

$$\sum_{i=1}^{k} \frac{/\nu_i - Np_i/^2}{Np_i}$$

kifejezés nagy N értékek esetén közelitőleg k-1 szabadságfoku χ^2-eloszlás. Ezért nullhipotézisünk vizsgálatára adott $/1-\varepsilon/$ szinthez a következő X_k kritikus tartományt konstruálhatjuk:

$$X_k = \left\{ \chi^2 \geq \chi^2_{k-1}/\varepsilon/ \right\} .$$

d/ Legyen a nullhipotézis az, hogy a ξ valószinüségi változó eloszlásfüggvénye $F/x/$, $\xi_1, \xi_2, \ldots, \xi_n$ pedig egy n-elemü minta. Rögzitett x-re jelölje K_n azt a valószinüségi változót, amely megadja az x-nél kisebb elemek számát a mintában. Ekkor a tapasztalati elosztásfüggvény: $F_n/x/ = \frac{K_n}{n}$. Adott

ε-hoz határozzuk meg azt az y_ε értéket, amelyre

$$\sum_{-\infty}^{\infty} /-1/^i \exp/-2i^2 y_\varepsilon^2/ = 1-\varepsilon .$$

Ekkor F/x/ számára a következő 1- ε megbizhatósági szintü konfidenciasávot nyerhetjük:

$$F_n/x/ - \frac{y_\varepsilon}{\sqrt{n}} < F/x/ < F_n/x/ + \frac{y_\varepsilon}{\sqrt{n}}$$

A Kolmogorov-Szmirnov-féle kétmintás próbával azt vizsgáljuk, hogy a ξ és η valószinüségi változók azonos eloszlásuak-e. Ha az eloszlásfüggvények F/x/ és G/x/, akkor a nullhipotézis:

$$H_o: \; G/x/ \equiv F/x/ .$$

Legyen a ξ -re vonatkozó n elemü minta $\xi_1, \xi_2, \ldots,$ ξ_n, az η-ra vonatkozó m-elemü minta $\eta_1, \eta_2, \ldots,$ η_m. Határozzuk meg az ezekhez tartozó $F_n/x/$ és $G_m/x/$ empirikus eloszlásfüggvényeket.

Az ellenhipotézis

$H_1: G/x/ \not\equiv F/x/$, akkor a

$D_{n,m} = \max \left| F_n/x/ - G_M/x/ \right|$ statisztikával konstruáljuk a következő 1- ε szintü kritikus tartományt:

$$X_k = \left\{ D_{n,m} \geq D'_\varepsilon \right\} , \text{ ahol } \quad D'_\varepsilon \text{ -re}$$

$$P\,/D_{n,m} < D'_\varepsilon \,|\, H_o/ = 1- \varepsilon .$$

/e/ A homogenitásvizsgálat arra a kérdésre keresi a választ, hogy két valószinüségi változó azonos eloszlásunak tekinthető-e. Jelölje a két változót ξ és η. Legyenek a két változóra vett minták $\xi_1,\dots,$ ξ_N és $\eta_1,\dots,\eta_M$. A fellépő értékkészletet osszuk fel r részre: $-\infty = z_0 < z_1 < \dots < z_r = \infty$.

Jelölje ν_i ill. μ_i a $/z_{i-1}, z_i/$ intervallumba eső ξ-k ill. η-k számát /i=1,2,...,r/. Nyilván $\sum_{i=1}^{r} \nu_i = N$, $\sum_{i=1}^{r} \mu_i = M$. Bizonyitható, hogy ha $N\to\infty$ és $M\to\infty$, akkor

/3/ $$\chi^2 = NM \sum_{i=1}^{r} \frac{/\frac{\nu_i}{N} - \frac{\mu_i}{M}/^2}{\nu_i + \mu_i}$$ /r-1/ - paraméterü

χ^2 eloszlást követ. Ilymódon nagy M és N esetén alkalmazhatjuk a χ^2 próbát.

/f/ Az a kérdés, hogy a ξ és η valószinüségi változók függetlennek tekinthetők-e? A ξ ill. η változók értékkészletét r ill. s csoportba osztjuk a

$$-\infty = x_0 < x_1 < \dots < x_r = \infty$$

$$-\infty = y_0 < y_1 < \dots < y_s = \infty$$

osztópontokkal. Tekintsük az alábbi eseményeket:

$$A_k = \{ x_{k-1} \leq \xi < x_k \} \qquad k=1,2,\dots,r$$

$$B_l = \{ y_{l-1} \leq \eta < y_l \} \qquad l=1,2,\dots,s$$

Végezzünk n független megfigyelést és jelöljük $\nu_{k\ell}$-lel az $A_k B_\ell$ esemény gyakoriságát a mintában. Vezessük be még a következő jelöléseket:

$$\nu_{k\cdot} = \sum_{i=1}^{s} \nu_{ki} \quad \text{és} \quad \nu_{\cdot\ell} = \sum_{i=1}^{r} \nu_{i\ell}$$

A függetlenségi hipotézis ellenőrzését a

$$\chi^2 = n \sum_{k=1}^{r} \sum_{\ell=1}^{s} \frac{/\nu_{k\ell} - \frac{\nu_{k\cdot}\,\nu_{\cdot\ell}}{n}/^2}{\nu_{k\cdot}\,\nu_{\cdot\ell}}$$

függvényre alapozzuk, amely a hipotézis fennállása esetén nagy n-re közelitőleg /r-1//s-1/ - paraméterü χ^2 eloszlásu.

2.3. Példák

A példák megkonstruálásánál az 1972-73 évi vizsgálat eredményeit használjuk fel: annak alapján egy "elképzelt" 10%-os mintát /betegszám: 170 000/ tételezünk fel és adjuk meg a számitásokat. Más minta alapján hasonló számitásokat lehet majd végezni.

/a/ 0,95 megbizhatósági szintü konfidencia intervallumot akarunk szerkeszteni annak p valószinüségére, hogy egy adott beteg Szabolcs megyei. M=7600 elemü a mintánk, igy /1/-et alkalmazhatjuk. Az $1-\varepsilon = 2\Phi/U_\varepsilon/-1$ összefüggésből következik, hogy

$U_\varepsilon = 2{,}81$.

$r_M = \frac{7600}{170.000} = 0,0447.$ Ezeket az értékeket /1/-be helyettesitve a $0,0433 \leqslant p \leqslant 0,0461$

0,95 megbizhatósági szintü konfidencia intervallumhoz jutunk. Ez azt jelenti, hogy 95%-os biztonsággal állithatjuk: a Szabolcs megyei betegek száma 7311=170.000 0,0433 és 7837=170.000 0,0461 közé esik.

/b/ Nézzük, mi a helyzet akkor, ha pl. az A esemény az, hogy a beteget a 333-as kórformával ápolták. Ekkor

$r_M = \frac{25}{170.000} = 0,0001471$, s igy /1/-ből

$0,0000644 < p \leqslant 0,0002298$ adódik 0,95 megbizhatósági szintü konfidenciaintervallumnak, ami "rossz"-nak mondható. Élesebb konfidenciaintervallumhoz juthatunk M növelésével.

Ha pl. az intervallum két végpontja közötti δ távolságra $\delta = 0,00005$ értéket kivánjuk meg, - ez olyankor fordulhat elő, amikor az A esemény valószinüsége igen kicsi, mint pl. az emlitett példában is - /2/ alapján, $\varepsilon = 0,0005$-tel számolva

$M \geqslant 1124 \cdot 10^6$ kellene, hogy legyen, ami természetesen semmilyen mintavétellel sem érhető el, figyelembevéve Magyarország lakosainak számát.

Vegyünk egy másik példát. Az A esemény legyen most az, hogy a beteget a 10. osztályon ápolják. Ekkor

$r_M = \frac{49.720}{170.000} = 0,2924706$, $\delta = 0,05$ esetén /2/-ből következik, hogy 1382 elemü minta is elég lenne

a 0,95 megbizhatóságu szintü 0,05 hosszuságu konfidenciaintervallum megadásához. Látjuk tehát, hogy adott megbizhatósági szintü adott nagyságu konfidenciaintervallum eléréséhez más-más mintanagyság kellene. Van, amikor ez problémába ütközik.

/c/ Itt csak néhány példát sorolunk fel, milyen esetekben merülhet fel hipotézisvizsgálat szükségessége. Annak eldöntésénél, hogy:

1. születésnapok eloszlása egyenletes-e,
2. a 8.táblázatban szereplő eloszlások azonossága milyen szinten fogadható el,
3. adott kódok eloszlása milyen szinten egyezik meg egy feltételezett eloszlással.

/d/ A konfidenciasáv meghatározásának realizálását feldolgozás közben egy külön programnak kellene végeznie. Ha a 2.1. /d/ példájában felvetett kérdésre keressük a választ 2.2. /d/ szerint kell eljárnunk.

/e/ Nézzük meg pl., hogy a születési hely és az állandó lakóhely megyéje azonos eloszlásunak tekinthető-e? A vizsgálatnál 2.2. /e/ pont /3/ formuláját kell használni.

/f/ Ilyen kérdés merülhet fel pl. az azonositó kódokkal kapcsolatban /ld. 3. rész/, de a feldolgozás után, a táblázatok ismerete is felvethet ilyen sejtést az orvosokban, s ennek ellenőrzésére használható a függetlenségvizsgálat.

Az elmondott példák alapján a következő megállapitásokat tehetjük. Bizonyos értékek - a 10%-os mintát alapul véve - nem szolgáltatnak megbizható eredményeket, ugyanakkor vannak olyan esetek, amikor kisebb mintából is megbizhatóan következtethetünk. Felmerülhet annak igénye, hogy a kapott táblázatokban valamilyen formában jelöljük, mely eredmények nem megbizhatóak - adott szinten-. Ez azonban két problémát vet fel: megnöveli a számolási időt, csökkenti a rendszer hatékonyságát, általánosságát. Mindezek ellenére nyilvánvaló, hogy bizonyos esetekben feltétlenül szükség van erre.

Ennek és az itt tárgyalt egyéb kérdések alkalmazási lehetőségeinek pontos behatárolására - hol, milyen számitások elvégzésénél kell bizonyos próbákat, stb. kivitelezni - további vizsgálatokra van szükség.

3. Azonosító kódok vizsgálata

3.1. A személyazonositás problémái

Mielőtt javaslatot tennénk a hospitalizált morbiditási vizsgálatnál használatra kerülő személyazonositóra /amely az ÁNH azonositó megjelenéséig lenne használatban/, röviden bemutatjuk, hogy milyen jellegü problémák lépnek fel "véletlen" adatokból felépitett azonositók kialakitásánál.

Ha egy populáció egyedeinek azonositása nem lehetséges sorszámozással, akkor az egyedeket valamilyen természetes adatuk alapján lehet megkülönböztetni egymástól. Ezek az adatok személyeknél lehetnek pl. a születési adatok, stb. Ilyen adatok azonban több különböző egyednél is lehetnek azonosak /pl. egyazon napon született azonos nemü emberek/. Az egybeesés véletlenszerü, de bármikor felléphet, még akkor is ha az azonositók lehetséges értékkombinációinak száma több, mint ahány azonositandó egyed van. Jó példaként szolgál erre az u.n. "születésnap paradoxon". Eszerint, ha véletlenszerüen kiválasztunk 23 embert, akkor az esetek több mint 50%-ában eközött a 23 ember között legalább kettőnek az év ugyanazon napján van a születésnapja /az év minden napját egyenlő valószinünek tekintve/. Ez egy igen érdekes, és első pillanatra meglepő jelenség, hiszen egy évben lényegesen több mint 23 nap van. Mégis, már 23 ember megkülönböztetésére sem elég jó azonositó az év 365 napja.

Ennek a jelenségnek a valószinüségszámitási hátterét a következő /3.2./ szakaszban tárgyaljuk. Most egy könnyen áttekinthető kisérletet irunk le a probléma szemléltetésére, amelyet az olvasó maga is elvégezhet /természetesen a kisérlet konkrét kimenetele bizonyára más lesz mint az itt leirtaké, statisztikai viselkedése azonban hasonló lesz/.

Végezzünk pénzdobási kisérletet! Egy dobás eredménye lehet fej vagy irás - jelölje ezeket a következőkben f és i. Ha mondjuk öt dobásból álló dobássorozatokat végzünk, akkor egy kisérletünk /dobássorozat/ eredménye pl. a következő sorozat lehet:

f f i f i .

Könnyen belátható, hogy összesen $2^5 = 32$ féle különböző eredménye lehet egy öt dobásból álló kisérletnek.

Végezzünk tehát öt dobásból álló kisérleteket, és figyeljük, hogy hányadik sorozat után lesz először két azonos dobáskombináció /legfeljebb 32 különböző sorozat lehetséges/! Bemutatunk egy ilyen kisérletsorozatot /az egyforma dobáskombinációkat * jelöli/.

a kisérlet sorszáma	a dobás-kombinációk	az első ismétlés helye
1.	i i f i i	
	f f f f i	
	f f f i f	
	i f i f f	
	i i i i f	
	i f f i f	
	* f f i f i	
	* f f i f i	8

a kísérlet sorszáma	a dobás-kombinációk	az első ismétlés helye
2.	f i i f i	
	i f f i f	
	i f f f f	
	* f i f i i	
	* f i f i i	5
3.	i f i f f	
	i i f i f	
	f i i i i	
	* i i f i i	
	i f f i f	
	i i i f i	
	* i i f i i	7
4.	* f f i i f	
	i f f i f	
	f i i f i	
	i i i f i	
	f f i i i	
	f i f i f	
	* f f i i f	7
5.	i f i i f	
	i f f i i	
	* f i i f i	
	f f f f i	
	* f i i f i	5
6.	* f f i f i	
	f f f i f	
	i f f i f	
	i i i i f	

a kisérlet sorszáma	a dobás-kombinációk	az első ismétlés helye
	f i f i f	
	i f f f f	
	i i i f i	
	* f f i f i	8
7.	* f i f f f	
	* f i f f f	2
8.	f f f i i	
	f f i i f	
	* f f i f i	
	* f f i f i	4
9.	f f f f i	
	* f i f i f	
	f i i f f	
	i f i i f	
	* f i f i f	5
10.	i f i i i	
	f i f i f	
	i i i f i	
	f i f f i	
	* f f f i f	
	i f i f i	
	* f f f i f	7

Egy ismétlés tehát rendre 8, 5, 7, 7, 5, 8, 2, 4, 5, 7 tagu kisérletsornál jött létre. Az itt látható 10 kisérletnél az átlagos sorozatszám 5.8 , tehát átlagosan minden 5-ik, 6-ik esetben azonos kombináció-

val találkozunk. Ha meghatározzuk a fenti kisérletben az első ismétlés sorszámának /mint valószinüségi változónak/ a várható értékét és szórását, akkor a 7.774... várható értéket és a 3.367... szórást kapjuk. Ezek az értékek jól illeszkednek a kisérleti eredményhez.

A bemutatott kisérlet eredményéből látható, hogy egy 32 féle értékü azonositó, már 5 vagy 6 tagu csoport egyedeinek azonositására sem alkalmas.

Hasonló a helyzet nagyobb populációk esetén is, igy pl. az évenként kórházban ápolt több mint 1 millió személy azonositására egy közel ugyanennyi értéket felvevő /pl. 7-8 jegyü/ "véletlen" azonositó kód semmiképpen sem elegendő.

Az azonositás egy másik problémája közvetlenül a kódolással áll kapcsolatban. Ha természetes adatokkal azonositunk, akkor sok esetben igen rossz hatásfokú kódokat kell használnunk. Például a "beteg neme" kétféle érték lehet, holott a felhasznált egyjegyü decimális kód tiz érték megkülönböztetését teszi lehetővé. Ugyanez a helyzet a születés hónapjánál és napjánál is, de még az olyan látszólag teljesen kihasznált kódnál mint a születés éve is, hiszen pl. a kórházi ápoltak között bizonyos viszonylag szük korosztályba tartozó betegek nagy számban fordulhatnak elő /pl. szülő nők/. Igy pl. a születési dátumból és nemből álló 7-jegyü azonositó közel sem ad 10 millió-féle értéket, hanem csak néhányszor tizezernyit.

3.2. A hospitalizált morbiditási vizsgálathoz javasolt személyazonositó

A kórházban ápolt személyek azonositására bizonyos adatokat használunk fel. /A probléma megértéséhez egyenlőre tekintsünk el ezen adatok konkretizálásától./

Kérdés: a/ ezek az adatok a személyek hány százalékát azonositják egyértelmüen? b/ hány ujabb adatot kell hozzávennünk az azonositóhoz, hogy az előbbi százalékszámot növeljük?

Nyilvánvaló az a cél, hogy ez a százalékszám minél nagyobb legyen. Az azonositóba azonban tul sok adatot nem célszerü belevenni, mert ez egyrészt megnövelné a különböző helyigényeket /az adathordozókon/, másrészt meglassitaná az adatmozgatást.

Modellként egy u.n. cellabetöltési problémát használunk /ld. [1] és [2]/: adott n cella, melyekbe egymástól függetlenül elhelyezünk N golyót ugy, hogy bármelyik golyó /a többitől függetlenül/ az i-edik cellába p_i /i=1,2,...,n/ valószinüséggel esik; $p_1+p_2+\ldots+p_n=1$.

Jelölje ν_k /k=1,2,.../ azon cellák számát, amelyekbe pontosan k golyó esik. A ν_k valószinüségi változó várható értékére és szórásnégyzetére a következő formulák adódnak: /a bizonyitás [1]-ben megtalálható/

$$E\nu_k \approx \sum_j \frac{/Np_j/^k}{k!} e^{-Np_j} \qquad /1/$$

$$D^2\nu_k \approx E\nu_k - \sum_j \frac{/Np_j/^k}{/k!/^2} e^{-2Np_j} \qquad /2/$$

Esetünkben a celláknak az azonositó kód egy-egy konkrét értéke, a golyóknak pedig az ápolt személyek felelnek meg.

[1]-ben és [2]-ben az 1972-73. évi kórházi morbiditás vizsgálat közben használt azonositó kódok elemzésének leirása található. Az akkor kapott eredményeinket mostani leirásunkban felhasználjuk, azonban, mint látni fogjuk, uj értékek számitására is szükségünk lesz.

Az azonositás hatásfokának növelése érdekében nyilván az azonositásra csak olyan adatokat célszerü használni, melyek nem változnak meg az ember élete során. Ilyen adat pl. a születési év, hó, nap, stb., de nem ilyen adat pl. az állandó lakóhely megyéje, annak "település-jellege", stb. Ennek megfelelően vizsgálatunk az alábbi adatokra terjed ki:

születési dátum	6 karakter
nem	1 "
beteg /leánykori/ nevének kezdőbetüi	4 "
anyja nevének kezdőbetüi	4 "
születési hely megyéje	2 "

Az /1/ és /2/ formulákból látható, hogy a számitások elvégzéséhez a p_j valószinüségek ismerete szükséges. Ehhez viszont felhasználjuk a 3-8. táblázatokat, melyek az ott jelzett eloszlásokat tartalmazzák. Ezek közül

néhány [1] -ben is megtalálható, a 6-at és a 7 -et a mostani vizsgálatokhoz számitottuk ki. A táblázatok a 10%-os mintára vonatkozó adatokat tartalmazzák. A születési év és a nem nem függetlenek egymástól; a többi változót, valamint ezt az együttes eloszlást függetleneknek tekintjük. A születésnapok egyenletes eloszlását tételezzük fel.

Nézzük a számitások eredményeit: /1/-ből következik, hogy

$$E\,\nu_2 \approx \sum_j \frac{/Np_j/^2}{2} \quad , \text{ felhasználva a 3.-8.}$$

táblázatokat

$E\,\nu_2 \approx 18$ adódik. Vagyis azt

kapjuk, hogy a duplán azonositott személyek várható száma 36. /A születési hely megyéjét a lakóhely megye szerinti eloszlással helyettesitettük - ld. 3.tábla./

$E\,\nu_k$ meghatározásához /1/-ben e^{-Np_j} -t hatványsorba fejtve a következőt kapjuk:

$$E\,\nu_k = \sum_{\ell=0} \sum_j \frac{/-1/^\ell}{\ell!k!} /Np_j/^{k+\ell}$$

Ennek az összefüggésnek előnye az, hogy a

$$\sum_{i,\ell,m} q_i^k r_\ell^k s_m^k = /\sum_i q_i^k/ \; / \sum_\ell r_\ell^k/ \; / \sum_m s_m^k/$$

disztributivitási törvényt alkalmazva többszáz millió

müveletet megtakaritva juthatunk eredményhez.

Felvetődik az a kérdés, mi történik, ha valamelyik adatot kihagyjuk az azonositóból: mennyire változik meg a rosszul azonositott emberek várható száma. Az, hogy ez a szám megnő, a képletekből azonnal következik. A pontos értékeket a következő táblázat mutatja /az adatok itt is a 10%-os mintára vonatkoznak/:

Kihagyott adat	Duplán azonositottak várható száma
Beteg vezetékneve	292
Születési megye	216
Beteg keresztnevének kezdőbetüje	294

Az 1972-73. évi adatok között a beteg keresztneve nem szerepelt, s az értékek szimulálása most nem adhat megfelelő eredményt. Ezért meggondolásainkban feltételeztük, hogy a beteg keresztnevének kezdőbetüje - mint valószinüségi változó - független a nemtől és a beteg születési évétől. Ez - érezhetően nincs igy -, s a függetlenség vizsgálat ezt igazolja is.

Ezen feltételezés mellett kapott számszerü eredményeink mégis használhatóak a következő értelemben. Mivel

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}^2 \geqq \sum_{j=1}^{m} /\alpha_{1j}+\alpha_{2j}+\ldots+\alpha_{nj}/^2 .$$

$$\cdot \sum_{i=1}^{n} /\alpha_{i1}+\alpha_{i2}+\ldots+\alpha_{im}/^2$$

$$/\alpha_{ij} \geqq 0 \quad , \quad i=1,\ldots n \quad , \quad j=1,\ldots,m/$$

következik, hogy a függetlenség feltételezésével kapott várható érték kisebb, mint egyébként. A 36, 292, ... stb. eredmények tehát a duplán azonositott személyekre alsó korlátokat szolgáltatnak.

Lakóhely megyéje szerinti elosztás

Budapest	36620
Bács-Kiskun	7244
Baranya	8860
B é k é s	7287
B o r s o d	12527
C s o n g r á d	6921
F e j é r	5764
Győr-Sopron	5739
Hajdú-Bihar	8131
H e v e s	5477
K o m á r o m	5190
N ó g r á d	4421
P e s t	14595
S o m o g y	5632
S z a b o l c s	8756
S z o l n o k	7208
T o l n a	4048
V a s	4137
V e s z p r é m	6903
Z a l a	4084
Szocialista külföld	388
Kapitalista külföld	68

3. táblázat

Január	13167
Február	14492
Március	15207
Április	14478
Május	14532
Június	14215
Július	14312
Augusztus	14830
Szeptember	15903
Október	13982
November	12628
December	12254

Születési hónap

4. táblázat

A	2209	LY	1
Á	769	M	11206
B	18066	N	7187
C	1368	NY	575
Cs	4464	O	2676
D	4540	Ö	291
Dz	7	P	9424
Dzs	14	Q	7
E	1731	R	4346
É	262	S	9320
F	7438	Sz	12032
G	5966	T	9801
Gy	1501	Ty	32
H	10908	U	866
I	1239	Ü	64
J	3351	V	8469
K	21298	W	912
L	5319	X	136
		Y	36
		Z	1311
		Zs	858

Vezetéknév kezdőbetüje

5. táblázat

A	10094	12453
Á	6362	3932
B	3342	3512
C	126	63
Cs	2834	4504
D	581	716
E	27434	13 915
É	10776	5058
F	148	4032
G	5085	6614
Gy	3752	5306
H	2145	1064
I	19545	18191
J	9969	16256
K	14983	8951
L,Ly	953	10744
M	24617	14459
N	684	638
Ny	0	0
O	749	708
Ö	0	25
P	1821	3914
R	3096	2911
S	451	4392
Sz	1240	938
T,Ty	6508	9521
U,Ü	0	1
V	2857	1765
W	0	0
X,Y,Q	5	2
Z	739	7341
Zs	9104	8074

A beteg keresztnevének kezdőbetüje nőknél

6.táblázat

A beteg keresztnevének kezdőbetüje férfiaknál

7.táblázat

Év	Férfi	Nő	Év	Férfi	Nő	Év	Férfi	Nő
1874	4	7	1875	0	0	1876	0	0
77	0	0	78	0	0	79	4	18
80	14	11	81	7	29	82	14	90
83	25	29	84	43	65	85	43	93
86	83	129	87	101	151	88	97	194
89	108	205	90	104	259	91	241	305
92	255	323	93	273	417	94	316	463
95	374	603	96	481	553	97	546	625
98	618	690	99	693	758	1900	751	840
1901	650	722	1902	740	736	3	715	779
4	837	797	5	805	708	6	812	801
7	815	858	8	959	998	9	988	887
10	905	887	11	884	866	12	916	923
13	1073	952	14	1048	855	15	708	625
16	506	467	17	467	524	18	503	514
19	898	927	20	1081	1279	21	884	1117
22	977	1286	23	1106	1207	24	948	1034
25	937	1336	26	884	1200	27	866	1117
28	858	1311	29	747	1185	30	880	1110
31	833	1268	32	772	1376	33	801	1476
34	743	1476	35	629	1584	36	510	1577
37	528	1609	38	542	1677	39	575	1874
40	629	1936	41	603	2133	42	611	2226
43	567	2481	44	567	2506	45	499	2664
46	546	2680	47	632	3239	48	682	3430
49	593	3455	50	557	3630	51	636	3691
52	718	3164	53	880	3296	54	908	2790
55	675	1802	56	521	1386	57	409	722
58	585	474	59	460	467	60	445	424
61	524	435	62	567	402	63	506	546
64	618	532	65	722	542	66	746	654
67	966	808	68	1357	1042	69	1342	1027
70	1249	866	71	1687	1156	72	2334	1809
73	564	424						

Születési év-, nem
8.táblázat

4. Az adattartalom szerepe a feldolgozási módszerek kiválasztásában

Számitástechnikai feladatoknál az adott cél elérésére legmegfelelőbb módszer kiválasztása nem csak a feladatban megfogalmazott logikai kapcsolatok, célok, stb. milyenségétől függ. A helyes módszer kiválasztásánál feltétlenül figyelembe kell venni a feladatokban szereplő adatok tulajdonságait /elsősorban statisztikai tulajdonságait/ is. Ez a kérdéskör a számitástechnika szinte valamennyi területén központi helyet foglal el - operációs rendszerek tervezésétől kezdve az adatbázis kezelő rendszerek előállitásáig. A felmerülő problémák sokasága és bonyolultsága miatt ebben a kérdéskörben még rengeteg megoldatlan, sőt megfogalmazatlan probléma van. A következőkben két idevágó témát érintünk, amelyek a kórházi morbiditási feldolgozáson belül is fontos szerepet játszanak.

4.1. Egyes kódok eloszlásának hatása

A nagyméretü táblák összeállitásánál problémát jelent egyes kódok nagy értékkészlete. /Például a BNO 4-jegyü diagnózis listája, mely elvben 10000 kódértéket tartalmaz/. Az eddigi tapasztalatok azt mutatják, hogy a minta nagy százalékát jóval kevesebb kód értékhez tartozó esetek teszik ki. /Pl. a

4-jegyü diagnózisnál a minta 80%-át kb. 300 kódértékhez tartozó eset adja meg/. Ezért az ilyen kódokra vonatkozó kérdések megválaszolását a minta szétválasztásával célszerü megoldani; a gyakran előforduló kódértékekre olyan sokdimenziós táblázatot állitunk elő, amelyből a kivánt táblázat összevonással nyerhető. A minta fennmaradt kisebb részét más eljárással dolgozzuk fel /ld. [4]/.
A kódok kumulativ eloszlásának ismeretében könnyen meghatározható a minta optimális szétválasztása.

Legyen például a diagnózis mellett a kérdéstipusban szereplő kódok terjedelmének szorzata n, a teljes minta elemszáma M, s az x koordináta a BNO kódok olyan permutációja, amely szerint az empirikus eloszlásfüggvény monoton csökkenő. Ekkor az

$$x \cdot n + M/1-F/x//$$ kifejezést kell x-ben minimalizálni.

Ez az eljárás természetesen csak akkor optimális, ha sok hasonló tipusu kérdést kell megválaszolni, mert az adatelőkészités költségeit nem veszi figyelembe. A várható kérdésszám ismeretében az adatelőkészités költségeinek figyelembevételével hasonló tipusu feladathoz jutunk.

4.2. Adatkeresési technikák

Az adatkeresési /adatbeillesztés, törlés/ eljárások a számitástechnika egyik központi témakörét al-

kotják. Nemcsak adatfeldolgozásnál /adatbázisok/ játszanak fontos szerepet ezek az eljárások, hiszen bármely más területen is szükség van keresési /beillesztési, törlési/ eljárásokra. Ezekkel a kérdésekkel részletesen foglalkozik pl. a [15] könyv.

Most a következő feladatot vizsgáljuk meg: Adott n-féle rögzitett egész érték, amelyek az [1,N] intervallumon helyezkednek el. Természetesen $n \leqq N$. A kérdés az, hogyan tároljuk ezeket az értékeket, hogy az érték ismeretében annak tárolási helyét a lehető leggyorsabban megtaláljuk /természetesen az adott n-féle érték mind különböző/.

Ha $N \approx n$, akkor nyilvánvaló, hogy a k értéket a legjobb a tároló k cimére helyezni, és itt közvetlen hivatkozással elérhető.

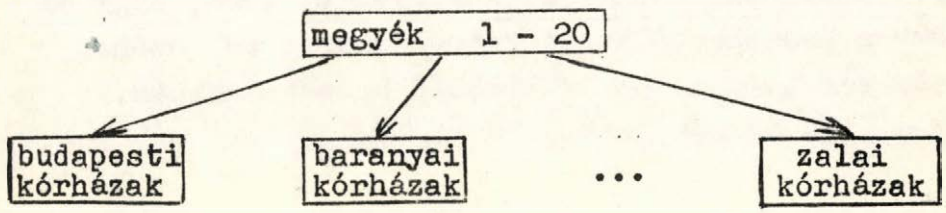
Ha az n "elég kicsi", akkor célszerü a jó ismert "bináris" keresési eljárást alkalmazni. Ilyenkor elegendő n tárolóhelyet biztositani. A bináris keresés hátránya az, hogy nagy n értékekre időigényes - a szükséges lépések száma n logaritmusával arányos. Figyelembe kell venni azt is, hogy egy lépés is több részből tevődik össze: a felezőpont kijelölése, a "kisebb, nagyobb" viszony vizsgálata és az uj intervallum kijelölése.

Ha a feldolgozandó n-féle értéken belül valamilyen kapcsolat van, akkor hatásosabb keresési eljárások is adhatók. Egy ilyen eljárást mutatunk be a kórházi morbiditási vizsgálatban alkalmazott formájában.

A jelenlegi kórházi morbiditási adatfelvételeknél a kórházakat egy négyjegyü kód azonositja. Az első két jegy a "kórház megyéje", a második két jegy egy megyén belüli sorszám. A megye kódja 1 és 20 közé esik, a sorszám egy gyakorlatilag tetszőleges kétjegyü pozitiv szám. Igy a kórházkód egy közel 2000 hosszuságu intervallumon változhat. A Magyarországon lévő fekvőbetegintézetek száma viszont a 250-et sem éri el. Célszerü tehát az érték szerinti tárolás helyett /amikor a k kód a k cimen van/ egy tömörebb tárolást alkalmazni.

A következő tárolásmódot használtuk:

Megyénként, érték szerinti cimeken tároltuk azokat a sorszámokat amelyek az adott megyén belül előfordulhatnak. Igy megyénként változó hosszuságu táblázatokat kapunk. Egy külön táblázat tartalmazza a megyénkénti sorszámokat tartalmazó tömbökre vonatkozó mutatókat /ld. a 2. ábrát/.



2. ábra

Mint ahogy a 2. ábra is mutatja a keresési eljárás egy irányitott gráffal reprezentálható hierarchikus rendszerben történik. A keresés ebben a rendszerben két egyszerü lépésből áll - mig a bináris kere-

sésnél a kb. 250 kórház esetén hét lépésből álló keresési folyamat is létrejöhet. A hierarchikus rendszerben szükséges két keresési lépés a következő: 1. a megye kiválasztása, 2. a megyén belüli sorszám kiválasztása.
Egy lépés csak egy indexezésből és egy értékadásból áll, mig a bináris keresést lényegesen bonyolultabb lépések alkotják.

Meg kell azonban jegyezni, hogy bináris keresésnél elegendő a kórházak számával /kb. 250/ egyenlő számu tárolóhely, mig a hierarchikus rendszerben majdnem 1000 tárolóhelyre van szükség. Nem ilyen nagy a különbség akkor, ha figyelembe vesszük azt, hogy bináris keresésnél a 250 pozició csupán egy cimet határoz meg. Ha pl. a kórház-kódhoz egy uj értéket akarunk rendelni /uj szorszám, tipuskód, stb./ akkor ez ujabb 250 helyet vesz igénybe. A hierarchikus rendszerben ujabb tárolóhelyekre nincs szükség.

Befejezésként felhivjuk a figyelmet arra, hogy az előzőkben összehasonlitott két eljáráson tul számos más módszer is van, pl. különböző hash-technikák, láncolási módszerek, stb.

5. A clusteranalizis alkalmazási lehetőségei

Ebben a részben a matematikai statisztika egy fiatal ágának, a clusteranalizisnek rövid ismertetését, alkalmazási lehetőségeit irjuk le. Ismertetésünkkel szeretnénk meginditani annak vizsgálatát, hogy az orvosi, egészségügyi adatfeldolgozásban - és speciálisan a kórházi morbiditási vizsgálatnál - milyen módon alkalmazható legeredményesebben a matematikai statisztikának ez a modern és igen hatékony módszere. Javaslatunk alapja az, hogy a SZTAKI Valószinüségszámitási Osztályán már évek óta sikeresen alkalmazzuk a clusteranalizist, számos alkalmazási területen /ld. pl. [11]/.

A clusteranalizis orvosi alkalmazására példát ad még [9] és [10] is. Alapvető tanulmányként [8] és [12] javasolható.

A clusteranalizist hazánkban a hetvenes évek elejétől alkalmazzák orvosi, gazdaságföldrajzi, szociológiai, kriminalisztikai és számitógép rendszerek matematikai leirásával foglalkozó kérdések leirásánál.

A cluster szó mindennapi jelentése: csoport, nyaláb, rakás, csomó, Kendall-Buckland: A Dictivnary of Statistical Terms szerint: egy statisztikai sokaság összefüggő elemeinek halmaza. A cluster-analizis csoportképző eljárás. Nem azt teszi, hogy megadott ismérvek alapján elemeket sorol be, "oszt szét" osztályokba, hanem maga alakitja ki az osztályokat. Green,

Frank és Robinson definiciója szerint a cluster-analizis olyan eljárások osztályára vonatkozó név, amelyek célja a dolgok birtokában lévő /feltétlen vagy mért/ jellemzőkből a hasonló dolgokat felismerni, azonositani. A diszkriminancia-analizistől eltérően nem ismert előre, hogy mely dolgok tartoznak egy osztályba. Az eljárás clustereket alakit ki, amelyek egyrészt differenciálják a statisztikai-demográfiai osztályokat, másrészt uj elrendezést hoznak létre a kutatás számára. A clusterezési feladat megoldásához definiálni kell a hasonlóság fogalmát mennyiségi módon, majd egy megfelelő algoritmust választani, amely a megfigyeléseket osztályokba sorolja.

Jelöljük $I=\{I_1, I_2, \ldots, I_n\}$ -vel n egyén /beteg/ halmazát. Tegyük fel, hogy megfigyelhető egy $c=\{c_1, c_2, \ldots, c_p\}$ tulajdonság vagy jellemző vektor, amely minden I-beli egyénnek birtokában van. Legyen m egy n-nél kisebb szám. A feladat: meghatározandó az I-beli egyének m clustere ugy, hogy I_i egy és csak egy részhalmazba tartozzon és azok az egyének, amelyek ugyanazon clusterbe tartoznak, hasonlók, mig a különböző clusterekhez tartozók nem hasonlók. A hasonlósági mérték definiálása függ a tulajdonságtól /változótól/.

A clusteranalizisban központi szerepet játszik a metrika.

A különböző változók esetén az irodalomban igen sok hasonlósági mértéket találhatunk. A feladat természetének legjobban megfelelő metrikát a szakembe-

rek alapos meggondolása és a felmerült mérőszámok kipróbálása után lehet megtalálni. Ez nem feltétlenül egyezik meg valamilyen ismert mértékkel.

6. Egyéb megjegyzések

a/. Ha egy programban bizonyos utasitás vagy utasitáscsoport többszázezerszer fordul elő, nem mindegy - az időtakarékosság miatt - milyen az utasitás "felépitése", ill. utasitáscsoport esetén milyen sorrendiségről van szó. Pontosabban, itt a következőt vizsgáljuk:

többezerszeres ciklusban szereplő összetett logikai kifejezést hogyan épitsünk be a programba? Például az IF/K1.GE.10.OR.K2.GE.120.OR.K3.LT.600/GO TO 1 tipusu utasitás szerepel az alábbi ciklusban:

```
      DO  1  I = 1, 100000
      IF/K1.GE.10.OR. stb./ GO TO 1
      OSSZEG = OSSZEG + 1
    1 CONTINUE
```

Ugyanezt a logikai vizsgálatot az alábbi programrészletben is elvégezzük:

```
      DO  1  I = 1, 100000
      IF/K1.GE.10/ GO TO 1
      IF/K2.GE.120/ GO TO 1
      IF/K3.LT.600/ GO TO 1
      OSSZEG = OSSZEG + 1
    1 CONTINUE
```

Ha történetesen K1 $\geq$ 10 a 2. programrészletben azonnal megtörténik az 1-es cimkére ugrás, mig az 1. programrészletben háromtagu logikai kifejezés kiértékelése tovább tart.
Könnyen látható tehát, hogy a második tipusu megoldás az idő megtakaritása miatt előnyösebb.

Az is észrevehető azonban, hogy nem mindegy az: milyen sorrendben követik egymást a 2., 3. és 4. sorok /ld. 2. programrészlet/. Nyilván azt az elemi feltételt kell a 2. sorba irni, amelyik a leggyakrabban teljesül; és igy tovább a többi sorba. Ezt a sorrendiséget néha meg lehet érezni, általában pedig a megfelelő eloszlások ismeretében meghatározhatók. Szükség szerint még program is irható, mely ezt a sorrendiséget eldönti.

A fenti probléma tipikusan egy nagyméretü adatfeldolgozási probléma, ahol nagyszámu egyednél kell bonyolult logikai feltételek vizsgálatát elvégezni.

b/. A következő megjegyzésünk az adatfelvétellel és az adatellenőrzéssel kapcsolatos.

Reprezentativ adatfelvételnél lényeges a mintaelemszámnak az egyes részpopulációkon belüli pontos beállitása /pl. a kórházi morbiditási vizsgálatnál a szakmánként reprezentativ mintavétel/.

A mintavétellel párhuzamosan adatellenőrzésre is szükség van. Ellenőrzéskor esetenként éppen a mintát meghatározó adatok /pl. a kórházi vizsgálatnál a születésnap és az osztálykód - ld. az 1. pontot/ is hi-

básak lehetnek. Ez torzitja az eredetileg pontos mintaarányt. Célszerü ezért a pontos mintaarány beállitása előtt végezni az ellenőrzést. Ha ugy járunk el, mint a kórházi morbiditási vizsgálatnál, hogy a pl. 10%-os mintát egy 13-14 százalékos mintából választjuk ki, akkor még az is előfordulhat, hogy az ellenőrzéskor kideritett és javitott hibák miatt a mintaarányok ugy módosulnak, hogy egyes csoportokon belül /pl. esetünkben a kórházi szakmákon belül/ 10% alá csökken ez az előzetes /13-14%-os/ mintanagyság - pl. szisztematikusan hibás osztály és születésnapkódok jönnek be. Ilyenkor természetesen nem lehet a 10%-os mintanagyságot biztositani.

Ezeket a szempontokat az 1. pontban leirt vizsgálatainknál nem vehettük figyelembe, hiszen az adatfelvételi hibák eloszlása függ az adatfelvétel szervezésétől, a felvitelben kialakult módszerektől és az esetleges hibaforrásoktól /hiányos tájékoztatás, utasitások hibás értelmezése, stb./.

Az adatfelvétel általános statisztikai elemzése, értékelése azonban egy külön tanulmányt igényel.

Irodalomjegyzék

[1] Az 1972-73. évi kórházi morbiditási vizsgálat számológépes feldolgozása, MTA SzTAKI dokumentáció - I-II.kötet, 1974.

[2] Garádi János - Krámli András - Ratkó István - Ruda Mihály: Statisztikai és számitástechnikai módszerek alkalmazása kórházi morbiditás vizsgálatokban, MTA SzTAKI, Tanulmányok, 35/1975.

[3] M.Csukás - L.Greff - A.Krámli - M.Ruda: An approach to the hospital morbidity data system development is Hungary, Symposium on medical data processing, Toulouse, 1975.

[4] Csukás A-né, Greff Z., Krámli A. és Ruda M.: Lekérdező rendszer a kórházi morbiditás vizsgálat adataira, Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 6. Kollokvium. Szeged, 1975.

[5] Vincze I.: Matematikai statisztika ipari alkalmazásokkal, Müszaki Könyvkiadó, Bp., 1968.

[6] Prékopa A.: Valószinüségelmélet müszaki alkalmazásokkal, Müszaki Könyvkiadó, Bp., 1962.

[7] Arató M.: Fejezetek a matematikai statisztikából számitógépes alkalmazásokkal I., MTA SzTAKI Tanulmányok, 42/1975.

[8] M.R.Anderberg: Cluster Analysis for Applications, Academic press, New York - London, 1973.

[9] Felsővályi Á., Hajtman B., Juhász P., Kopp M., Veér A.: Faktor- és clusteranalizis alkalmazása a szociálpszichiátriai kutatásban, Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 6. Kollokvium, Szeged, 1975.

[10] Fenyő I., Bánóczy J., Sima D., Siminszky M.: A clusteranalizis diagnosztikai alkalmazása leukoplákiás betegek cardinoma veszélyeztetettségének megállapitására, Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 6. Kollokvium, Szeged, 1975.

[11] Csukás A-né, Mándi A., Galgóczy G., H.Gaudi I.: A légzésfunkciós elváltozások vizsgálata faktor- és clusteranalizis segitségével, Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 6. Kollokvium, Szeged, 1975.

[12] B.S.Duran, P.L.Odell: Cluster Analysis, A Survey, Springer Verlag, Berlin-Heidelberg, New York, 1974.

[13] Rényi A.: Valószinüségszámitás, Tankönyvkiadó, Bp., 1966.

[14] Tomkó J.: A Markov-folyamatok elemei és néhány operációkutatási vonatkozása, Bolyai János Matematikai Társulat kiadványa, Bp., 1968.

[15] J.D.E.Knuth: The Art of Computer Programing, Sorting and Searching /3.kötet/, Addison-Wesley, London - California, 1973.