Egészségügyi Minisztérium, MTA SZTAKI

A kórházi morbiditás vizsgálat számitógépes feldolgozásának tapasztalatai és továbbfejlesztése

Csukás Andrásné, Greff Lajos, Krámli András és Ruda Mihály

Az elmult évben ugyanezen a kollokviumon beszámoltunk az 1972. április 1-től 1973. május 31-ig tartó kórházi morbiditás vizsgálat számitógépes feldolgozásának tervéről és a kezdeti tapasztalatokról. Ebben az előadásban az akkor körvonalazott terv megvalósulásáról és további elképzeléseinkről adunk számot.

A feldolgozás eredményeképpen az egészségügyi tervezés számára nélkülözhetetlen adatok váltak könnyen és gyorsan elérhetővé. A feldolgozás orvosi tapasztalatainak ismertetése külön előadást igényel, ezért most elsősorban a gépi megvalósitás számitástechnikai és statisztikai szempontjaival kivánunk foglalkozni.

Az érthetőség kedvéért röviden összefoglaljuk a kórházi morbiditás vizsgálatnak a tavalyi előadásban részletesen kifejtett céljait, a rendelkezésre álló adatok természetét, a mintavétel módját.

Emlékeztetőül megemlitjük, hogy hazánkban 1952 óta folynak kórházi morbiditás vizsgálatok. Ezek feldolgozása eddig kézzel, vagy Hollerith gépekkel történt, igy az összefüggések vizsgálata csak korlátozott lehetett.

Az 1972-73. évi kórházi morbiditás vizsgálat célkitüzései a következők voltak:

- a kórházi ápolást igénylő betegségek előfordulás-gyakoriságának megállapitása, a kórházi ápolás időtartamának meghatározása,

- epidemiologiai jellegü adatok biztositása a feltétlen kórházi ápolást igénylő betegségeknél,

- többször ápoltak arányának és összetételének megállapitása,

- különböző diagnózisok közötti kapcsolat vizsgálata,

- meghatározott faktorok hatásának vizsgálata,

- különböző intézmények tevékenységének értékelése,

- felső és középszintü igazgatás számára adatokat szolgáltatni az irányitáshoz és a vezetéshez,

- nemzetközi összehasonlitás biztositása.

A mintavétel rétegezett volt, a nagyforgalmu szakmáknál - melyek a kórházi ápolások tömegét jelentik - 33 %-os, a többi szakmánál 50 %-os mintavételi aránnyal dolgoztunk.

Egy-egy rétegen belül a véletlen kiválasztás a születésnap alapján történt. A 33 %-os mintavételi arány esetében a 4-re, 8-ra és a 0-ra végződő napokon -, valamint a 22-én születettek kerültek a mintába, a többi szakmánál pedig a hónap valamennyi páros napján születettek. Ez a kiválasztási technika az adatfelvevők számára könnyen elsajátitható volt, megadta a differenciálás lehetőségét, valamint megkönnyitette az ellenőrzést. Ezzel a módszerrel egy 32,6 %-os és egy 49,3 %-os mintavételi arány volt várható.

A próbafelvételnek szánt első két hónapos mintába 110 000, a teljes évi mintába 623 000 kórházi eset került.

A születésnap feltételezett egyenletes eloszlását az adataink igazolták.

Szakmánként a mintavételi arány azonban mindig alacsonyabb volt a vártnál. A vártnál alacsonyabb arányt a mintavételi arányban a megyék között fennálló különbségek okozhatják, de okozhatja az eltérést az is, hogy a hibás alapbizonylatok egy része nem került vissza a mintába.

A vizsgálat céljaira a kórházi adminisztrációban rendszeresitett ugynevezett fejlapot használtuk. Ezt a megoldást a kódolási hibalehetőségek minimálisra csökkentése indokolta, bár jelenlegi formájában nem felel meg a modern gépi feldolgozás követelményeinek, nehezen áttekinthető, nem egyértelmü, több felesleges információt is tartalmaz.

Az adatok kódolását a fekvőbeteg gyógyintézetekben orvosok és egészségügyi középkáderek végezték. Az orvosok és statisztikusok egyszer ellenőrizték a teljes minta kódolását, majd a minta 20 %-át az Egészségügyi Minisztérium szakemberei ujra ellenőrizték.

A lyukkártyára rögzitett, majd az SZKI által mágnesszalagra felvitt adatokat az SZKI Siemens gépén is ellenőrizték. A gépi ellenőrzés során kiszürték az idegen karaktereket, az értelmetlen kódértékeket és a korral, nemmel összeférhetetlen diagnózisokat tartalmazó rekordokat. A hibás rekordok aránya 3,6 % volt.

A további feldolgozást az MTA SZTAKI végezte CDC-3300-as gépén. Problémát jelentett az SZKI által szolgáltatott 9 csatornás mágnesszalagok 7 csatornássá konvertálása. Ezt a munkát az Országos Tervhivatal ICL gépén végezték el. A SZTAKI-ban a szalagokat ujra ellenőrizték.

A minta jóságát mutatja, hogy 100-nál kevesebb volt a páratlan napon születettek száma, amit gépi uton nem is ellenőriztek, és csak az eloszlások vizsgálatánál derült ki.

Az azonos személyek többszörös ápolására vonatkozó kérdések felvetették az azonositó kódok statisztikai vizsgálatát.

A személyek azonositására a tervezéskor 9 karakter jött szóba. Ezek: születési dátum, nem, az anya nevének kezdőbetüje.

Az MTA SZTAKI valószinüségszámitási osztályának munkatársai a probléma elméleti tárgyalására a klasszikus cellabetöltési problémának egy nem egyenletes eloszlásra vonatkozó általánositását javasolták, feltételezve, hogy az azonositásra használt egyes faktorok - a nem és a kor kivételével - függetlenek.

A hibásan azonositott személyek várható számának meghatározására számitástechnikailag könnyen kezelhető formulákat a Matematikai Kutató Intézet munkatársai dolgozták ki.

A számitott és az elméleti értékek jó egyezést mutatnak. A számitott értékeket ugy kapták, hogy különbözőnek tekintették azokat a személyeket, akiknek állandó lakhelye és foglalkozása különböző, majd megvizsgálták, hogy hány olyan 9 jegyü azonositó kód van, mely az ilyen módon különbözőnek tekintett személyek közül 1, 2, 3, stb. személyhez tartozik.

Megjegyezzük, hogy a belső vándorlás és az egy éven belüli foglalkozás változás - pl. nyugdijazás - viszonylag nagy száma miatt ez az eljárás nem teljesen korrekt, mégis hü képet nyujt az azonositó kódok statisztikai viselkedéséről.

A Matematikai Kutató Intézet munkatársai nem egyenletes eloszlás esetén is igazolták azt - az egyenletes eloszlásra ismert tényt -, hogy az egybeesések számának szórásnégyzete kis sürüség esetén is megegyezik a várható értékkel.

A fentiek alapján a személyek azonositását 13 számjegyü azonositó kóddal végeztük.

A feldolgozásban szereplő bemenő adatok - mintegy 70 millió karakter - négy 800 byte/inch sürüséggel teleirt mágnesszalagot foglalnak el. A feladatban szereplő minden egyes táblázat elkészitésekor ilyen tömegü adat mozgatása nem lett volna gazdaságos, sem biztonságos. Ezért a teljes adatrendszert részekre bontottuk. Módszerünket a következő, további szempontok is indokolják:

A tervezett táblázatok esetenként az adatrendszernek csak egy részhalmazára vonatkoztak, igy egy-egy táblázat elkészitésekor felesleges a teljes adattömeget mozgatni. Sikerült a táblázatok jelentős részét öt olyan csoportba sorolni, melyekhez a teljes adatrendszer 20-26 karakteres részrekordjai szükségesek csak. Emellett az egyes táblák kialakitásához szükséges rendezési eljárásokat egyszerübben és gyorsabban végrehajtottuk.

Ugyancsak a feldolgozás optimalizálásának céljából bontottuk két részre magukat a táblázat készitő programokat is. Első lépésben a táblázatokban szereplő elsődleges értékeket (ápolási esetek, napok száma) gyüjtöttük össze és helyeztük el mágnesszalagon. A táblázatok kinyomtatása a járulékos értékekkel (százalékok, átlagos ápolási napok, stb.), a feliratokkal együtt egy második lépésben történt. Ilyen módon a menetközben felmerült - kisebb módositásokra vonatkozó - ujabb igények teljesitéséhez csak kisebb időigényü listázó programokat kellett ujrafuttatni.

Ilyan nagyvolumenü adatfeldolgozási munka során különböző optimalizálási problémák merülnek fel. Meg kell állapitani a részadatrendszerek optimális tartalmát és terjedelmét, valamint a mágnesszalagon tárolt táblázatok optimális tartalmát is, mert egy-egy ilyen táblázat több hasonló tipusu kérdés megválaszolására is alkalmas. A fenti optimalizálási feladatok az időközben változó igények miatt egzakt módon nem oldhatók meg.

Optimalizálni kellett a gép tömegtároló SORT program kapacitását lényegesen meghaladó adatmennyiség rendezését is. Ebben az esetben a szalagok fizikai mozgatásának minimális idejére zárt formulát nyertünk.

Könnyen belátható, hogy korlátlan számu mágnesszalag esetén - azaz, ha az egységek e száma nagyobb, mint ahányszorosa a rendezendő r adatmennyiség a SORT program rendelkezésre álló mágneslemez memória k kapacitásának - a fizikai mozgatáshoz szükséges idő nem lehet kevesebb, mint a teljes adatmennyiség mozgatásához szükséges idő $\frac{3r - k}{r}$ - szerese.

Ha $\frac{k}{r} > e$, akkor a minimális mozgatás nagyobb, mint a teljes adatmennyiség mozgatásához szükséges idő

$$\frac{(d + 2) - \frac{k(e-1)^d - 1}{e-2}}{r}$$

-szerese, ahol d az a minimális szám, melyre

$$r \leq k \left[e\,(e-1)^d + (e-1)^{d-1} + \ldots + 1 \right] .$$

Felmerül a kérdés, hogy ilyen nagyvolumenü adatfeldolgozás esetén szükséges-e előre elkészíteni a statisztikai táblaterveket. Ezeknek valamilyen egyszerü formája mindenképpen szükséges, mert kidolgozásával tisztázódnak a szakemberek előtt a felmerült kérdések, a számitástechnikai szakemberek számára pedig látható, melyek azok a faktorok, melyekre a szekemberek különös sulyt helyeznek, és a programterveket ezeknek megfelelően lehet elkésziteni.

További tapasztalataink azonban azt mutatják, hogy a feldolgozás teljes menete előre nem rögzithető le, a feldolgozást szekvenciálisan kell végrehajtani. Az első lépésben csak egy vagy két szempont szerinti összefüggéseket lehet vizsgálni, és a további bontásokat csak olyan csoportokra szabad elvégezni, ahol statisztikailag elfogadható eredmény várható. Az ilyen tipusu feldolgozást jelenleg elsősorban adminisztrativ jellegü problémák gátolják: a szerződéskötések rendszere nem eléggé rugalmas, hosszu távra kell lekötni meghatározott mennyiségü programozási és gépi kapacitást.

Az ilyen módon elkészült, mintegy 35 nagyméretü, 2-15 ezer sort tartalmazó táblázat áttekintése nehézkes, a táblák a sok felesleges adat mellett gyakran nem tartalmazzák közvetlenül azokat az adatokat, melyekre a szakemberek kiváncsiak. Ezen a helyzeten két módon is kivánunk változtatni.

Az egyik mód egy rugalmas kérdezőrendszer kidolgozása, mely szinte tetszőleges kérdésekre perceken belül választ ad, feltéve, hogy a válasz nem tartalmaz 1-2 ezernél több számadatot. Ilyen kérdezőrendszer kialakitása folyamatban van. Az egyes esetekről felvett 70 karakternyi információból 30 karaktert fog felhasználni. Ezek a karakterek az azonos személyek többszörös ápolására vonatkozó kérdések kivételével valamennyi eddig felmerült kérdés megválaszolásához elegendőek.

A kivitelezés alatt álló kérdező-rendszerben a CDC 3300-as gép ez év januárjában végrehajtott bővitése során nyert lehetősége-

ket használtuk fel. Rendelkezésünkre áll egy 32 millió karakter kapacitásu közvetlenül címezhető tömegtároló egység (óriás disc), és ezen már elhelyezhető a teljes adatrendszerből készített olyan rendezett file, amelynek alapján a leggyakrabban felmerülő, különböző eloszlásokra vonatkozó kérdések indextáblázatos módszerrel néhány percen belül megválaszolhatók.

A másik módot ugy nevezhetnénk, hogy a kimenő adatok érdektelenségének automatikus megállapitásai, igy pl. a megbetegedési esetek havi megoszlás-vizsgálatánál csak az olyan diagnózisra vonatkozó adatokat kell kiirni, melyeknek havi eloszlása az egyenletestől szignifikánsan eltér.

Több más feladattal együtt ez a probléma is felveti az ilyen nagyvolumenü adatfeldolgozással kapcsolatban alkalmazható statisztikai módszerek kérdését.

A feldolgozás kezdeti szakaszában - amikor az eloszlásról szinte semmilyen előzetes információnk nincs - nehéz kiválasztani az alkalmazható statisztikai módszereket.

Az ápolási esetek vizsgálatánál a binomiális eloszlást, illetve kis valószinüségek esetében a Poisson eloszlást tételeztük fel. Az egyes megyéknek az országos értéktől való eltérésének kimutatására az u-próbát használtuk. Az átlagos ápolási napok, megyék közötti eltérésének vizsgálatára - egyes diagnózisokon belül - variancia analizist végeztünk. A vizsgálat alapjául a 300 diagnózist tartalmazó D jegyzéket vettük.

Az analizisek során felhasználtuk még az ápolási esetekné az ugynevezett indirekt standardizálási eljárást is, mellyel a megyék között koreloszlásban fennálló különbséget küszöböltük ki. Standardként az ország lakosságát vettük. Ez az analizis szintén a 300 diagnózisra történt, a szignifikancia vizsgálatra a chi^2 - próbát használtuk.

Finomabb statisztikai eljárások más jellegü adatok felvételét is igénylik.

Ez a kórházi morbiditás vizsgálat kiinduló pontként szolgál a további, folyamatos kórházi morbiditás vizsgálatokhoz. A jelenleg kapott adataink alapján ugy látszik, hogy egy-egy tervidőszak alatt a 10 %-os mintavétel országos vezetői szinten megfelelő tájékoztatást nyujt. A távlati tervek elkészitéséhez azonban a 10 %-os minta nem megfelelő, mivel részletesebb és sokoldalu analizis szükséges. Egy-egy nagyobb volumenü kórházi morbiditás vizsgálatot a tervidőszakok második felében kell végrehajtani, ugy, hogy a vezetők az adatokat a tervezéshez még fel is tudják használni. Ez azt jelenti, hogy az adatfeldolgozásnak gyorsnak kell lennie, és a közben felmerülő ad-hoc kérdésekre is választ kell adnia. Ezt csak kérdezőrendszerrel lehet megoldani.