MTA SZTAKI

Az 1974 óta végzett kórházi morbiditásvizsgálat által felvetett statisztikai kérdések

Ratkó István - Soltész János

Bevezetés

Az MTA SZTAKI Valószinüségszámitási és Matematikai Statisztikai Osztályának munkatársai 1973 óta vesznek részt a kórházi morbiditási vizsgálatokban. Ennek kapcsán különböző matematikai statisztikai kérdések merültek fel, melyek megválaszolása - bármilyen orvosstatisztikai adatfeldolgozás alkalmával - a kapott eredmények értékelése, megbizhatósága szempontjából alapvető fontosságu. Erről részletesebben [1]-ben beszámoltunk. Most csak négy problémára szeretnénk rávilágitani. A felhasznált adatok az 1972-73 évi kórházi morbiditási vizsgálat adatai.

1. Adott százaléku minta biztositása

Mintavételnél két fontos szempontot kell figyelembe venni: a minta lehetőleg a teljes populáció egy meghatározott hányada legyen, hogy a mintából könnyen következtethessünk a teljes populációra és ugyanezért fontos az is,

hogy a minta reprezentativ legyen.

Az évenkénti kórházi morbiditási adatokból osztályonként 10 %-os mintát kell kiválasztani. A minta pontossága az elsődleges cél, még ha ez néhány többszörösen ápolt személy egyes adatainak elvesztését vonja is maga után.

Korábbi statisztikai vizsgálatok igazolták azt a természetes feltevést, hogy a morbiditási adatok függetlenek a születésnaptól. A betegek születésnap szerinti eloszlása, ha a hónapot nem vesszük figyelembe, egyenletesnek tételezhető fel [2]. Hány születésnap /havonként/ szükséges a 10 %-os minta biztositásához feltéve, hogy a kórházak a kijelölt napokon született valamennyi beteg fejlapját beküldik. A Moivre-Laplace tétel és a Bernstein egyenlőtlenség felhasználásával bizonyitható, hogy három nap 0,5 két nap 0,00008 négy nap 0,9974 valószinüséggel elegendő a 10 %-os mintához.

A kórházak négy születésnap betegeinek lapjait /4.-i, 14.-i, 24.-i és 6.-i/ küldik el. Ezekből egy már megirt program segitségével választja ki a számitógép a 10 %-os mintát egy egyszerü itt nem részletezett algoritmus alapján.

Meggondolásaink, számolásaink nem 10 %-os mintavételnél is alkalmazhatók.

2. A többszörös ápolással kapcsolatos mintavételi problémák

Azt vizsgáljuk meg először, 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.

A probléma a következő: 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 össze, hogy a mintában hány olyan beteg van, akit előbb az A osztályon, majd a B osztályon kezeltek. Az ilyen betegeket az egyszerüség kedvéért AB betegeknek fogjuk 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 so-

rán a B osztályon kezelték.

A következő modellel fogunk dolgozni:

Az ápolási esetek rendezve vannak születésnap /először vannak a 14.-én, majd a 24.-én, 4.-én, 6.-án, stb. született emberek, majd azokon belül egyéb azonositóik /születési év, hó, név, anyja neve, stb./ szerint. Igy minden kétszer ápolt beteg két esete egymás mellé kerül. Ez a feltevés nem jelent megszoritást a kórházi morbiditási adatok statisztikai viselkedésére vonatkozóan. Egy beteg $p_1, p_2, \ldots, p_{20}$ valószinüséggel kerül az 1., 2., ..., 20. osztályra. Ha kétszer kezelik, akkor a második alkalommal az elsőtől függetlenül kerül $p_1, \ldots, p_{20}$ valószinüséggel a megfelelő osztályra. Egy beteg i.-én 1/30 valószinüséggel születik /i = 1, 2, ..., 30/, függetlenül attól, hányszor és melyik osztályon kezelték.

A mintavételezést ugy végezzük, hogy elindulunk a rendezett populáció elejéről és minden az A osztályon kezelt esetet beveszünk a mintába, egészen addig, amig $M_A/10$ esetünk nem lesz. Jelöljük ξ_A-val az A osztály mintájába bekerülő utolsó eset sorszámát. Jelölje η a teljes mintába kerülő azon esetek számát, akiket mind az A, mind a B osztályon ápoltak. A

$$h = \frac{E(\eta \mid \xi_A = a,\ \xi_B = b) - E\eta}{E\eta}$$

3. A mintavétellel kapcsolatos megbizhatósági kérdések

[1]-ben részletes leirás található a felvethető kérdésekről, a felhasznált összefüggésekről. Most csak konkrét példákon keresztül szeretnénk rávilágitani a lényegre.

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 korformáju betegséggel ápolták, stb. Másszóval ez pl. a következőt jelenti 95 %-os biztonsággal állithatjuk, hogy a Pest megyei betegek száma 14200 és 14800 közé esik, stb. Másik példa 0,95 megbizhatósági szintü konfidencia intervallumot akarunk szerkeszteni annak p valószinüségére, hogy egy adott beteg Szabolcs megyei. 7600 elemü a mintánk, igy azt kapjuk, hogy $0,433 \leq p \leq 0,0461$ 0,95 megbizhatósági szintü konfidencia intervallum. Ez azt jelenti, hogy 5 %-os biztonsággal állithatjuk a Szabolcs megyei betegek száma 7311 és 7837 = 170000.0,0461 közé esik.

b/ Feltételezve, hogy kórházainkban évente kb. 1700000 beteget ápolnak, felmerül a kérdés, hogy adott megbizhatósági szint /adott hibavalószinüség és hibakorlát/ esetén hány százalékos mintára van szükségünk.

Ha pl. azt az eseményt vizsgáljuk, hogy a beteget 333-as korformával ápolták, akkor $0,0000644 \leq p \leq 0,0002298$

adódik, ami "rossznak" mondható. Élesebb konfidenciaintervallumhoz juthatunk M növelésével. Ha pl. az intervallum két végpontja közötti távolságra 0,00005 értéket kívánjuk meg - ez olyankor fordulhat elő, amikor az A esemény valószinüsége igen kicsi, mint pl. ebben a példában is. Most azt kapjuk, hogy M≧1124000000, ami természetesen semmilyen mintavétellel sem érhető el, figyelembevéve Magyarország lakosainak számát. Vegyünk egy másik példát. Az esemény legyen most az, hogy a beteget a 10. osztályon ápolják. Ekkor 0,05 intervallum esetén az következik, hogy 1382 elemü minta is elég lenne a 0,95 megbizhatóságu szintü 0,05 hosszuságu konfidencia intervallum megadásához. Látjuk tehát, hogy adott megbizhatóságu szintü adott nagyságu konfidencia intervallumhoz más-más mintanagyság kellene. Van amikor ez problémába ütközik.

c/ Szükség lehet arra, hogy eldöntsük adott H_o feltevés /pl. a szellemi dolgozók 30 %-a infarktusban hal meg, vagy a születésnapok eloszlása egyenletes, stb./, u.n. nullhipotézis adott szinten elfogadható-e. Megadandó továbbá a H_o-t elutasitó u.n. kritikus tartomány. /Ilyen kérdésekről ld. [5]./

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 konfidenciasávot 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 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.

3. Azonositó kódok vizsgálata

Ha egy populáció egyedeinek azonositása nem lehetséges sorszámozással, akkor az egyedeket valamilyen természetes adatuk alapján lehet megkülönböztetni egymástól. Ezek az adatok személyeknél lehet pl. néhány születési adat, stb. Ilyen adatok azonban több különböző egyednél is lehetnek azonosak /pl. egyazon napon született azonos nemü emberek/. Pl. belátható, hogy már 23 ember megkülön-

böztetésére sem elég jó azonositó az év 365 napja: ha véletlenszerüen kiválasztunk 23 embert, akkor az esetek 50 %-ában eközött a 23 ember között legalább kettőnek az év ugyanazon napján van a születésnapja /az év minden napját egyenlő valószinünek tekintve/.

A kórházban ápolt személyek azonositására bizonyos adatokat használunk fel. Kérdés: a/ ezek az adatok a személyek hány %-át azonositják egyértelmüen? b/ Hány ujabb adatot kell hozzávennünk az azonositóhoz, hogy az előbbi százalékszámot növeljük?

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 karakter
beteg /leánykor/ nevének kezdőbetüi	4 karakter
anyja nevének kezdőbetüi	4 karakter
születési hely megyéje	2 karakter

Bizonyitott összefüggéseinkből következik, hogy a

duplán azonositott személyek várható száma 36.

Felvetődik a kérdés, mi történik, ha valamelyik adatot kihagyjuk az azonositóból, mennyire változik meg a rosszul azonositott emberek várható száma. Erre vonatkozik a következő táblázat.

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

4. Megjegyzések

a/ A kettőnél többször ápoltak száma elhanyagolható, az ebből adódó hiba egy nagyságrenddel kisebb, mint az általunk adott becslés hibája.

b/ Az AB betegek számára vonatkozó becslés elég pontos lesz, ha A is és B is "nagy" osztály.

c/ Az elmondott példák alapján a következő megállapitásokat tehetjük. Bizonyos értékek - a 10 %-os mintát alapul véve - nem szolgáltatnak megbizható

eredményeket, ugyanakkor vannak 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 emlitett egyéb kérdések alkalmazási lehetőségeinek pontos behatárolására - hol, milyen számitások elvégzésénél kell bizonyos próbákat, stb. kivitelezni - további vizsgálatokra van szükség.

d/ Az azonositóba tul sok adatot nem célszerü belevenni, mert ez egyrészt megnövelné a különböző helyigényeket /az adathordozókon/, másrészt meglassitaná az adatmozgatást.

e/ Az azonositás egy másik problémája közvetlenül a kódolással kapcsolatos. Ha természetes adatokkal azonositunk, akkor sok esetben igen rossz hatásfoku kódokat kell használnunk. Például a "beteg neme" kétféle érték lehet, holott a felhasznált egyjegyü decimális kód tiz érték megkülönböztetését teszi le-

hetővé. Ugyanez a helyzet a születés hónapjánál és napjánál is, de még az olyan látszólag teljesen kihasznált kódnál is, mint a születés éve, 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 hétjegyü azonositó közel sem ad 10 millió féle értéket, hanem csak néhány tizezernyit.

I r o d a l o m

[1] Králmi A., Ratkó I., Ruda M., Soltész J.: A statisztikai adatfeldolgozás matematikai és számitástechnikai problémái, MTA SZTAKI, Tanulmányok, 70/1977.

[2] Garádi J., Krámli A., Ratkó I., Ruda M.: Statisztikai és számitástechnikai módszerek alkalmazása kórházi morbiditás vizsgálatokban, MTA SZTAKI, Tanulmányok, 35/1975.

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

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

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