MINTAVÉTEL A LESLIE KISH-KULCS ALKALMAZÁSÁVAL*

NÉMETH RENÁTA – RUDAS TAMÁS

A megkérdezendő személy háztartáson belüli kiválasztásának problémája gyakran merül fel a többlépcsős mintavételi eljárás utolsó lépcsőjében; például telefonos felmérésekben, miután a telefonszámok véletlen kiválasztásával háztartási mintához jutottunk. A Kish-kulcs a háztartáson belüli választásra ad eljárást. Eredményeink szerint, bár a kulcs a háztartáson belül véletlen mintát biztosít, az általánosan elterjedt nézettel szemben nem alkalmas a sokaságban jellemző nemek és korcsoportok szerinti arány beállítására. A tévedés feltehetőleg abból származik, hogy amikor Kish az 1950-es évek Amerikájában az eljárást kidolgozta, az ország aktuális háztartásszerkezete miatt ez a két cél egyszerre volt megvalósítható. Számításaink szerint a mai magyarországi helyzetben ez nincs így. A Kish-kulcs olyan módosítását javasoljuk, mely alkalmasabb reprezentatív minták kiválasztására.

TÁRGYSZÓ: Felvételtervezés. Mintavétel. Leslie Kish-kulcs.

A Leslie Kish-kulcsot vagy módosított változatait gyakran alkalmazzák hazai és külföldi felmérésekben.[1] A kulcs ugyanúgy a háztartáson belüli véletlen kiválasztásra szolgál, mint például a „legutóbbi születésnap" módszere. Alkalmazásakor a mintavétel során először a háztartásokat választjuk ki, majd a mintába került háztartásokon belül jelöljük ki a kérdezendő személyt. Az *1.* részben tárgyaljuk azokat a tényezőket, amelyek ennek a kétlépcsős mintavételi módszernek az alkalmazását indokolhatják.

* A tanulmány alapjául *Németh Renáta* egyetemi szakdolgozata szolgált, melyet az ELTE Szociológiai Intézetének szociológus szakán készített, *Rudas Tamás* témavezetésével.

[1] Néhány példa: *1. A közigazgatás megítélése* (1999). MONITOR Társadalomkutató Intézet és Módszertani Központ. http://www.monitor-tki.hu. *2. Csaba TV nézettségi mutatók* (1998). Jelenkutató Intézet. http://www.bekes.hungary.net/csabatv/nezettseg.htm. *3. Felmérés Veresegyházán és környékén* (1998). (Táblaképek az egészségről – A veresegyházi példa) (2001). *4. ISSP* (1994). Család II., TÁRKI. http://www.tarki.hu. *5.* Az Egyesült Államokban: telefonos politikai közvéleménykutatások 1998-ban és 1992-ben, a felméréseket vezette a CBS és a New York Times. (*Voss–Gelman–King*; 1995). *6.* National fertility and family survey Európa több országában, 1990-es évek. Kivitelező: Population Activities Unit of the United Nations Economic Commission for Europe. http://www.cpc.unc.edu/pubs/paa_papers/1995/klijzing.html. 7. International Social Justice Project, 1991, 1996. 12 európai ország, köztük Magyarország bevonásával. http://www.isjp.de/. *8. SCSU Survey*, Minnesota State Lottery (1997). http://www.lottery.state.mn.us/gambling/stcloud.html. *9. Los Medios y Mercados de Latinoamérica* (1997), több latin-amerikai országban. http://www.zonalatina.com/Zldata25.htm. *10. Northern Ireland social attitudes survey* (1989–1996). Social and Community Planning Research http://cain.ulst.ac.uk/othelem/research/nisas/robinson.htm. *11. Scottish health survey* (1995). Social and Community Planning Research http://www.official-documents.co.uk/document/scottish/shealth/shch1.htm\#1.2.1. *12. Oregon health behavior surveys* (1998). State of Oregon, Health Division http://pub.das.state.or.us/purchasing/contracts/8042.html. *13. Australian national social science survey*, 1984. The Australian National University http://ssda.anu.edu.au/SSDA/CODEBOOKS/FR84/description.html.

A tanulmány középpontjában a Kish-kulcs használatával kapott minták nem és kor szerinti reprezentativitása áll. A Kish-kulcs szokásos használatát a *2.* részben ismertetjük. Tapasztalatok szerint a minták nem megfelelően illeszkednek a sokasághoz, és az eltérések szisztematikus hasonlóságot mutatnak.[2] A *3.1.* részben kifejtett hipotézisünk szerint ezek az eltérések levezethetők magából a mintavételi eljárásból, s magyarázatukkor nincs szükség a megvalósítás során adódó problémák (például szisztematikus válaszmegtagadás) feltételezésére. A hipotézis alátámasztásaként egy magyarországi felmérésből származó mintát veszünk példának a *3.2.* részben. A példa esetlegességén túlmutatva a *3.3.* részben a hipotézis bizonyítására is alkalmas számítást közlünk. A bizonyítás során az elméletileg várható minta összetételét határozzuk meg. Ugyanitt rámutatunk arra, hogy nem a kulcs okozza a hibát; ugyanez a probléma lép fel minden, háztartásmintára épülő módszer esetén. A kulcs esete azért is szerencsés, mert használatakor ellenőrizhető a minta nem és kor szerinti összetétele. Ezt kihasználva hajtjuk végre a kulcs módosítását a *4.* részben. Eredményünk szerint a tökéletes reprezentativitás még elméletileg sem érhető el, de a minta illeszkedése javítható.

1. A HÁZTARTÁSMINTA HASZNÁLATÁNAK OKAI

A felvétel kivitelezésekor a kutatási terv létrehozása az első lépés. A kutatási tervben a felvétel tárgyát képező egyedek összességét, a célsokaságot, és a megismerni kívánt célsokasági jellemzőket határozzuk meg. Ezen kívül a terv rögzíti a célsokaság gyakorlati elérésére alkalmas keretet. A keret köztes állomás a célsokaság és a minta között, azaz a mintát a keretből választjuk.

A mintavétel módjának meghatározása ezen paramétereknek megfelelően történik, figyelembe véve a felvétel kivitelezésének költségvonzatát. A megfelelő eljárás kiválasztása minden felvétel tervezésekor egyedi döntést igényel.

A háztartáson belüli kiválasztás problémája olyan kétlépcsős mintavételi eljárás alkalmazásakor merül fel, amikor az első lépcsőben a mintába kerülő háztartásokat választjuk ki a keretből, majd a háztartás (felnőtt) tagjai közül jelöljük ki a kérdezendő személyt a második lépcsőben.[3]

Miért alkalmazunk két lépcsőt, miért nem választjuk ki a keretből egy lépésben, közvetlenül a kérdezetteket? Ennek gyakori oka, hogy nem áll rendelkezésre megfelelő, keretként használható lista a célsokaságról. A választói névjegyzékek adatai például több országban sok hibás adatot tartalmaznak a populáció mobilitása vagy a nyilvántartásba nem vétel miatt. Ugyanakkor, bár nem megfelelő a minőségük a személyek kiválasztásához, de eredményesen használhatók címlistaként, mert hibáik a személyek és lakcímek

[2] A magyar statisztikai irodalom a sokaság mellett a populáció kifejezést is használja.

[3] A háztartás mint mintavételi egység gyakran fordul elő a felmérésekben. Leggyakrabban az együtt élők olyan közösségét értik alatta, akik funkcionálisan is együttműködnek. Egyik ilyen megfogalmazás szerint a részben közös jövedelemből élő, közös konyhát használók tartoznak egy háztartásba. Hasonló mintavételi egység lehet a család. A család általában a rokoni kötelékkel is összekapcsolt együtt élőket jelentheti, de például a Központi Statisztikai Hivatal meghatározása szerint a család házastársi vagy élettársi kapcsolatban álló, gyermekes vagy gyermek nélküli személyeket jelöl, illetve egyedül álló szülőt nőtlen/hajadon gyermekkel. A család és a háztartás itt tehát nem teljesen átfedő kategóriák, az egy fedél alatt élő háztartás több generációja külön családokat alkot. Ugyanígy létezik a szakirodalomban a „nem családháztartás" kifejezés, ami egymással családi kapcsolatban nem álló együtt élőket jelent. Egy harmadik, a háztartás és a család fogalmát megint csaknem átfedő fogalom az együtt, egy lakásban élő személyek összessége. Problémafelvetésünk szempontjából nem okoz nehézséget a három fogalom megkülönböztetése, hisz bárhogy is határozza meg a felméréstervező a mintavételi egységet, ugyanaz a probléma jelentkezik mindhárom esetben.

összekapcsolásában jelentkeznek, de magukat a lakcímeket megfelelő minőségben tartják nyilván.[4] Egyes felmérésekben ezért a nyilvántartási listát a lakások vagy háztartások kiválasztásához használják, míg a kérdezett kiválasztását a háztartásból újabb véletlen mintavétellel oldják meg.[5]

A megfelelő nyilvántartások hiányának kezelésére fejlesztettek ki egy másik többlépcsős módszert, a területi mintavételt (area sampling) (*Kish*; 1965). Alkalmazásának feltétele, hogy a célsokaság egy adott földrajzi egység lakosságaként legyen meghatározható. A területet blokkokra osztják (például kerületek), ezeket kisebb egységekre (utcák). A területi módszer lényege, hogy földrajzilag definiált mintavételi egységeken működik, így jó területi diszperzitást biztosít. Végső lépcsőben (például véletlen sétával) itt is háztartási mintához jutunk, és a kérdezendőt újabb mintavétellel választjuk.

A telefonos felvételek során is felmerül a háztartások véletlen kiválasztása után a kérdezett kiválasztásának problémája. Ilyenkor a háztartások kiválasztásának lépcsőfoka a mintába kerülő telefonszámok kiválasztásának feleltethető meg.

Tehát a második lépcsőben a háztartás tagjai közül kell kiválasztanunk a kérdezendő személy(eke)t. Nem merülne fel a kérdezendő kiválasztásának problémája, ha a felvétel kutatási céljának megfelelően ő a háztartás meghatározott tagja, például a családfő lenne. Ugyanígy nem okozna nehézséget a választás, ha a mintavételi egység maga a háztartás lenne, ekkor bármely háztartástag egyaránt adekvát információkat szolgáltathatna.

Általában a háztartások több tagja tartozik a kerethez. A kiválasztás ekkor sem okoz nehézséget, ha az alábbi feltételek bármelyike teljesül.

a) Bármelyik személytől beszerezhetők az összes többi háztartásbelire vonatkozó szükséges információk.

b) A háztartás minden tagját bevonva a felmérésbe sem csökken jelentősen a becslések megbízhatósága az azonos mintanagyság mellett, de egyszerű véletlen mintavétellel kapott mintából számolt becslések megbízhatóságához képest.[6] Ilyenkor a felmérés kérdéseire adott válaszoknak elhanyagolható a háztartásokon belüli korrelációja. Ellenkező esetben, vagyis amikor egyfajta homogenitás tapasztalható a háztartásokon belül, megbízhatóbb becsléseket kapunk, ha (azonos mintanagyság mellett) háztartásonként csak egyetlen személyt választunk.[7]

[4] Hazánkban a Központi Nyilvántartó és Választási Hivatal listája alkalmazható a személyi minta kiválasztására. A TÁRKI tapasztalatai szerint (http://www.tarki.hu/kozvelemeny-h/adatfelv/metodol.html) a Hivatal listájának alkalmazásakor a tervezett minta körülbelül 55 százalékának sikeres lekérdezésére lehet számítani. Ebben a 45 százalékos kiesési arányban a címlista hibái mellett persze más okok, így a válaszmegtagadás is közrejátszik. Úgy találták, hogy a kiesés szisztematikus torzítást okoz, a megmaradó minta demográfiai összetétele eltér a kiinduló mintáétól. Másik tapasztalatra is hivatkozhatunk: az Országos Lakossági Egészségfelmérés 2000 szintén a Hivatal címlistájára támaszkodott. (OLEF2000. A kutatást az Egészségfejlesztési Kutatóintézet vezette, a terepmunka kivitelezője a Magyar Gallup Intézet volt. Az adatfelvétel időpontja 2000. november-december. A Központi Népességnyilvántartó és Választási Hivatal címlistájából készült országos minta. Módszere: kérdőív kérdezőbiztosokkal. Esetszám: 5503.) A sikertelen interjúk aránya 21 százalékos, ebből 7 százalék adódott a címlista valamely hibája miatt (például a kérdezendő elköltözött, vagy meghalt). A kiinduló minta ezen 7 százaléka szignifikánsan eltér a teljes mintától a személyek életkorát tekintve (a kiesők főként fiatalok), illetve a személyek lakhelye szerint (több köztük a budapesti lakcímmel szereplő).

[5] Például Nagy-Britanniában: 1. lábjegyzet 11-es felmérése, illetve lásd a Market research methods forrást.

[6] A becslés érvényessége (validity) arra utal, vajon a többszöri mintavétel után kapott becslássorozat mennyire mozog a valódi sokasági paraméter körül. Ha az ingadozás centruma egy másik érték, az eljárás torzított. Az érvényesség mértéke a torzítás segítségével ragadható meg: a torzítás csökkenésével nő a becslés érvényessége. A becslés megítélésének másik szempontja a megbízhatóság (reliability). A becslés megbízhatósága attól függ, mennyire hasonló becsléseket kapnánk a mintavétel többszöri megismétlésével. A megbízhatóság a becslés varianciájának segítségével mérhető: minél kisebb a varianciája, annál megbízhatóbb a becslés. Fontos megjegyezni, hogy mindkét párhuzam (torzítás-érvényesség, illetve variancia-megbízhatóság) csak akkor érvényes, ha feltesszük, hogy nem lép fel nem-mintavételi hiba (*Levy–Lemeshow*; 1999).

[7] Ezen homogenitás mértéke az osztályon belüli korreláció (intraclass correlation) együtthatója, a ρ. A csoportos mintavétellel kapott mintából számolt becslés varianciája növekszik, ha a ρ nő, illetve ha nagyobb a csoportok átlagos nagysága. A variancia növekedése kevésbé megbízható becsléseket eredményez (*Kish*; 1965).

Ezek a feltételek kevés felmérés esetén teljesülnek. Tehát egy személyt kell háztartásonként kiválasztani. Rögzítsük továbbá, hogy feleslegesen nem keresünk fel háztartásokat: minden mintaháztartásból kiválasztunk egy kérdezendőt. Ennek az előfeltételnek a költségkímélés az ésszerű indoka. Vegyük még fel feltevéseink közé, hogy a háztartások egyenlő valószínűséggel kerültek a mintába.

A kérdezés gyakorlatában a kontrollálatlan kiválasztás lenne a legegyszerűbb módszer, amikor is azt a személyt választanánk, aki kinyitja az ajtót, vagy – telefonos felmérésben – aki felveszi a telefont. Ám ez a kiválasztás erős torzítást eredményezne a mintában, mivel azokat kérdeznénk, akik könnyen elérhetők, illetve akik inkább hajlandók együttműködni.

2. A KISH-KULCS

Leslie Kish[8] szándéka szerint a kulcs egy világos és előre rögzített eljárást ad a kérdezendő kiválasztására (*Kish*; 1949, 1965). Lényege, hogy segítségével a véletlen kiválasztást jól imitáló módon választhatjuk ki a háztartás egyik tagját. Másik előnye, hogy használatának megfelelő dokumentációja birtokában a választás utólag is ellenőrizhető.

A Kish-módszert alkalmazó felmérésben minden kérdőívhez egy fedőlapot csatolnak. A fedőlapon a háztartástagok felsorolására alkalmas tábla (példa az 1. tábla) és egy kiválasztási tábla (példa a 2. tábla) van. Az 1. tábla egy, a kérdező által a terepmunka során már kitöltött táblát illusztrál.

1. tábla

Kiválasztási adminisztráció a terepmunka során

Rokonsági viszony a családfőhöz	Nem	Kor	Sorszám	Kiválasztás
Családfő	F		2	
Feleség	N	40	5	
Családfő apja	F		1	
fia	F		3	
lánya	N		6	
A feleség nagynénje	N	44	4	✔

Forrás. Itt és a 2–5. táblánál: *Kish*; 1965.

A háztartás tagjai közül elsőként a férfiakat állítjuk sorba, életkor szerint (az idősektől a fiatalabbak felé), majd utánuk a nőket soroljuk fel, szintén életkoruk alapján. Az interjú első néhány kérdése az ehhez szükséges információkat gyűjti össze. Célszerű minél kevesebb kérdésből megállapítani a sorrendet, ezt segíti a rokonsági viszony megállapítása. Az 1. táblabeli példában csupán a feleség és a feleség nagynénje életkorának ismeretében

[8] *Leslie Kish* (1910–2000) magyar származású amerikai statisztikus volt. A világ elsőszámú mintavételi szakértőjeként jegyezték. 1965-ös könyve, a „Survey Sampling", ma is használt forrás szerte a világon. Véletlen mintavételi technikájának megalapozottsága először az 1948-as amerikai elnökválasztáskor nyert bizonyítást. (Kevesebb, mint 1000 háztartást tartalmazó országos mintája *Dewey* és *Truman* szoros eredményét jósolta, kis fölénnyel Truman javára, míg a sajtó és a szokott felmérések Dewey megsemmisítő győzelmét várták.) Számos eredményéből a legfontosabbak közé tartozik a válaszmegtagadási arány figyelembevételének szükségessége a valódi véletlen minta eléréséhez; vagy a tervezéshatás-együttható megalkotása, amely alkalmas a felmérések hatékonyságának mérésére. Kish az elsők között javasolta a guruló minták évenkénti kivitelezését.

végrehajtható volt a sorrendezés. A felsorolt háztartástagok közül a kiválasztandó személyt a Kish-tábla jelöli ki. Példánk esetén 6 felsorolt van, a 2. tábla szerint ekkor a sorban a 4. személy a kérdezendő. (Ha a kiválasztott nem található otthon, a kérdező újra felkeresi a háztartást.)

2. tábla

A 8 Kish-tábla egyike – D kiválasztási tábla

Felnőttek száma a háztartásban	A kiválasztott sorszáma
1	1
2	2
3	2
4	3
5	4
6 vagy több	4

A D kiválasztási tábla csak egyike a nyolc hasonló táblának. (Lásd a 2. táblát.) Ezeket a táblákat a kérdőívekhez csatolják, véletlenszerűen párosítva őket. A nyolc tábla nem egyenlő arányban oszlik meg a kérdőívek között, például a C jelzetű minden 6. kérdőíven található meg, míg az E csak minden 12.-en. Az elsődleges cél a háztartásokon belül egyenlő kiválasztási valószínűséget rendelni minden személyhez (anélkül hogy túl sok táblát kellene létrehozni). A 3. táblát elemezve kitűnik, hogy ez csaknem minden háztartás esetén megvalósul. A 3. tábla alapján határozhatók meg a 4. táblában található háztartáson belüli kiválasztási valószínűségek. Például a 3 felnőtt tagot számláló háztartások sorban 1. tagját 1/6+1/12+1/12=1/3 valószínűséggel, a 2. tagját 1/6+1/6=1/3 valószínűséggel, a 3. tagját 1/12+1/12+1/6=1/3 valószínűséggel választjuk ki. Az egyenlő esélyű kiválasztás két esetben nem valósul meg. Az egyik a 6-nál több felnőtt tagot számláló háztartások esete. A 7. vagy nagyobb sorszámú tag 0 valószínűséggel kerül a mintába. A másik eltérés az 5 felnőtt tagú háztartásoknál figyelhető meg. Ott az 1. 2. és 4. tag 1/6 valószínűséggel kerül kiválasztásra, a 3. és az 5. 1/4 valószínűséggel. A várhatóan fiatal, nőnemű 5. személy felülreprezentálása Kish szerint ellensúlyozza a 6 tagú háztartásokból 0 valószínűséggel bekerülő, várhatóan szintén fiatal lányok kiesését.

3. tábla

A Kish-kulcs kiválasztási szabálya

A tábla aránya	A tábla jelzete	A kiválasztott sorszáma, ha a felnőttek száma					
		1	2	3	4	5	6 vagy több
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3	5
1/12	E2	1	2	3	4	5	5
1/6	F	1	2	3	4	5	6

4. tábla

Háztartáson belüli kiválasztási valószínűségek a Kish-kulcs alkalmazása esetén

A tag sorszáma a sorban	A háztartástagok száma					
	1	2	3	4	5	6 vagy több
1	1	½	1/3	1/4	1/6	1/6
2		½	1/3	1/4	1/6	1/6
3			1/3	1/4	1/4	1/6
4				1/4	1/6	1/6
5					1/4	1/6
6						1/6
7 vagy nagyobb						0

A Kish-féle módszer elsősorban a véletlen kiválasztást imitálja. A leggyakrabban a telefonos felmérésekben is ezzel az eljárással választják ki a kérdezettet, bár adott a technika a véletlenszám-generálásra, és ezért a kiválasztásra elegendő információ lenne a háztartástagok azonosítása, például keresztnév alapján. Vélhetően azért ragaszkodnak ma, a számítógépes háttértechnika birtokában is, a nem/kor alapján történő felsoroláshoz, mert ez a módszer bizonyos mértékig biztosítja – a szisztematikus mintavételhez hasonlóan – a minta nem- és korösszetételének kiegyenlítettségét (Kish maga nem említi ezt az előnyt).

3. A MINTA REPREZENTATIVITÁSA

A mintavételi módszer megítélésének egyik szempontja a mintából levezethető becslések megbízhatósága. A becslés megbízhatósága (reliability) attól függ, mennyire hasonló becsléseket kapnánk a mintavétel többszöri megismétlésével. A megbízhatóság szempontjából a minta kívánatos tulajdonsága a gyakran emlegetett reprezentativitás. A reprezentativitás szó jelentésének nagyszámú változatával találkozhatunk. A legtágabb értelemében az adatok jóságát, a minta tipikus voltát értik rajta. Szűkebben a sokaság valamely jellemzője szerinti megoszlásának mintabeli reprodukálását jelenti. Ez azért kívánatos, mert ha a kutatás kérdései összefüggésben állnak az adott jellemzővel, akkor a becslés várhatóan megbízhatóbb lesz a jellemző szerinti reprezentativitás esetén.

A Kish-kulcsot vagy módosított változatait használó hazai és külföldi felmérések során a minták jellemzésekor azonos problémák tapasztalhatók (*Kennedy*; 1993, *Táblaképek az egészségről...*; 2001, az 1. lábjegyzet 4. és 8. felmérése). Gyakori a nők felülreprezentáltsága és az idősek, főként az idős nők vártnál nagyobb mintabeli aránya. A hozzáférhető források az esetek legnagyobb hányadában az eltéréseket megvalósíthatósági problémákkal magyarázzák, azaz bizonyos demográfiai jellemzők szerinti szisztematikus válaszhiánnyal. Hasonlóan, amikor Leslie Kish az általa létrehozott eljárással kapott minták reprezentativitását értékelte, a férfiak alulreprezentáltságát ritka otthon tartózkodásukkal és nagyobb arányú válaszmegtagadásukkal magyarázta (*Kish*; 1949).

3.1. A reprezentativitás sérülésének feltételezhető fő oka

A nehezen ellenőrizhető gyakorlati okokon kívül más forrása is lehet a problémának. A kulcs a háztartáson belüli egyenlő esélyű kiválasztást hivatott teljesíteni, és ez nem je-

lenti a mintába kerülési valószínűségek egyezését. A nagyobb létszámú háztartásokból nyilván nehezebb a mintába kerülni, míg az egyszemélyes háztartások lakóját mindenképpen megkérdezzük, ha háztartása már bekerült a háztartásmintába.

Az általunk felkutatott források jó része a minta reprezentativitásának értékelésekor nem veszi figyelembe az eltérő bekerülési esélyből fakadó elvi szintű problémákat. Ha a háztartások nagyságától függ a bekerülési esély, a háztartások nagysága viszont karakteres jellegzetességeket hordoz bizonyos demográfiai jellemzők szerint, akkor maga a mintavétel okozza a szisztematikus demográfiai eltéréseket. Vagyis ekkor akár tökéletesen véletlen háztartásminta, és nem szisztematikus válaszmegtagadás mellett sem kapunk reprezentatív mintát. Szélsőséges példával: akkor is „rossz" mintát kapunk, ha minden célsokaságbeli háztartást bevonunk a mintába (kiküszöbölve a háztartásválasztás mintavételi hibáit), és teljes együttműködést tételezünk fel a kérdezettek részéről.

A kiválasztási valószínűségek kiegyenlítését Kish súlyozással javasolja megoldani. A kulccsal kapott minták súlyozásáról Kish megállapította, hogy nem okoz jelentős változást a becslésekben, ezért a gyakorlatban elhagyták az elemzésből (*Kish*; 1949, 1965). Kish tapasztalata a súlyozás elhagyhatóságáról a korabeli egyesült államokbeli háztartásszerkezetből eredeztethető. Más körülmények között – más országban, más korban – a minta elégtelen reprezentativitása esetleg már magából a mintavételi módszerből levezethető, és nem csak a szisztematikus válaszmegtagadás az előidézője a torzulásoknak.

A háztartásmintán alapuló felmérések mintája és a sokaság közötti nem és kor szerinti eltérés visszavezethető a háztartásstruktúrára, ha a háztartások nagysága összefügg nemek szerinti és korösszetételükkel. Ekkor a háztartáslétszámmal fordítottan arányos kiválasztási valószínűség torzítja a minta e két jellemző szerinti összetételét. Az alábbiakban ezt a feltevést igazoljuk.

A Központi Statisztikai Hivatal 1996-os mikrocenzusa szerint 1996-ban a háztartások több mint negyede (26%) egyetlen lakóval rendelkezett. A népesség öregedési folyamata megváltoztatta a háztartások korösszetételét is. 1980 és 1996 között 12-ről 8,5 százalékra esett vissza a csak 30 évesnél fiatalabb személyekből álló háztartások aránya. A csak 60 évesekből vagy idősebbekből álló háztartások aránya viszont 18-ról 24 százalékra emelkedett. A háromgenerációs háztartások hányada 8-ról 5 százalékra esett vissza. Jellegzetes az egyszemélyes háztartások nem és kor szerinti összetétele: legalább 60 éves az egyszemélyes háztartások lakóinak 58 százaléka. Az egyszemélyes háztartások kétharmadában nő él egyedül, méghozzá az összes egyszemélyes háztartás 45 százalékában legalább 60 éves nő.

Az adatok szerint tehát a háztartások nagysága erősen összefügg nem- és korösszetételükkel. A háztartások azonos mintába kerülési esélyei mellett ez már önmagában előidézhet reprezentativitási problémákat. Például vélhetően a nők a sokasághoz viszonyítva nagyobb számban kerülnek be a mintába, mivel nagy az egyszemélyes háztartások aránya, és ebben a háztartástípusban zömmel nők élnek. A fiatalok alulreprezentáltsága és az idősek nagyobb száma szintén háztartásszerkezeti jellemzőkre vezethető vissza, hiszen, mint láttuk, igen alacsony a csak fiatalokból álló háztartások aránya, míg minden negyedik háztartásban csak idősek élnek.

Fontos hangsúlyozni, hogy ez a hiba nem a Kish-kulcs sajátja, hasonló probléma jelentkezik minden más, háztartásmintán alapuló kétlépcsős mintavétel esetén. A probléma forrása ugyanis, mint láttuk, az egyenlőtlen kiválasztási valószínűségek megléte. Ez a hi-

ba jelentkezik, ha azonos a háztartások bekerülési valószínűsége, és a háztartáson belül véletlen kiválasztást alkalmazunk. Ezért találtak egy 1990-es amerikai felmérésben azonos reprezentativitási problémákat mind a Kish-kulcs, mind a legutóbbi és a legközelebbi születésnapmódszer használata esetén (*Kennedy*; 1993). Ezért is szerepelhetnek a nők és főként az idős nők a tervezettnél nagyobb arányban a KSH legutóbbi születésnapmódszerrel készült mintájában (*Egészségi Állapot Felvétel...*; 1994).

Érdemes rövid összehasonlítást végezni a mikrocenzus szerinti magyar és a Kish-kulcs létrehozáskor fennálló amerikai háztartásszerkezet között. Mint említettük, Kish a kiválasztási valószínűségek eltérését kiegyenlítő súlyozás elhagyhatósága mellett döntött (*Kish*; 1965). Állásfoglalását a súlyozott és a súlyozatlan becslések kis különbségére alapozta, a kis különbségek okaként pedig a korabeli háztartásszerkezet jellemzőit jelölte meg. Az általa hivatkozott háztartási adatok a következők.

5. tábla

Háztartások aránya felnőtt tagjaik szerint, 1957, Egyesült Államok

A felnőttek száma a háztartásban	1	2	3	4	5	6 vagy több
A háztartások százalékaránya	14,6	73,0	9,0	2,8	0,4	0,2

Kish értékelése szerint a kiválasztási valószínűségek szórása nagyon kicsi, mivel a háztartások zöme két felnőtt taggal rendelkezik, a többiek közül pedig gyakorlatilag mind 1, 3, vagy 4 felnőtt lakót számlál.

Ezzel szemben a hazai, 1996-os mikrocenzus szerint a magyar háztartások 26,1 százaléka egytagú. Ez csaknem kétszerese a Kish által tapasztalt aránynak.[9] Bár a nagy létszámú háztartások nálunk sem képviseltetik magukat jelentős számban, ez az eltérés önmagában is olyan mértékű, hogy óvatosnak kell lenni a Kish-kulcs súlyozás nélküli honosításával.

3.2. A kulcs mint a reprezentativitási problémák forrása – a feltételezést alátámasztó példa

A fentiekben a reprezentativitási problémák egyik feltételezhető forrásaként jelöltük meg a kulcsot. Általában azonban nem választható el egymástól számszerűen a felmérésben alkalmazott két mintavételi lépcsőben, a háztartások, illetve a személyek kiválasztásakor fellépő hiba. Ezért szerencsés a következőkben leírt példa, ahol a rendelkezésre álló információk alapján jól elkülöníthető a két képcső hatása a minta létrejöttében.

Az ISSP Család II. felmérést a TÁRKI végezte 1994. május–júniusban. Ennek során a KSH 1990-es népszámlálásának lakásmintájából készült országos reprezentatív mintát használták, és a Kish-kulcsot alkalmazták a kérdezendő kiválasztására. A minta elem-

[9] Az Egyesült Államok háztartásszerkezete is jelentősen megváltozott az utóbbi évtizedekben. Növekvő háztartásszám és csökkenő átlagos háztartásnagyság volt megfigyelhető: 2000-ben az egyszemélyes háztartások aránya már 26 százalékos, ebből 9 százalékot tesznek ki a legalább 65 éves egyedül élők (www.census.gov). Az egyedül élők és az egyszülős háztartások arányának növekedése, illetve az együtt élő párok arányának csökkenése általános jelenség a mai modern társadalmakban (*Bonvalet–Lelievre*; 1997, *de Jong*; 1996).

száma 1500 fő. A háztartásokat véletlen (vagy legalábbis véletlent elérni kívánó) módon választották, majd minden háztartásból pontosan egy tagot jelöltek ki. Jól jellemezi a háztartásmintát használó kétlépcsős módszer kikerülhetetlen mintatorzító hatását már a minta háztartási struktúrája is. A mikrocenzus szerint a népesség 10,1 százaléka él egytagú háztartásban, ezzel szemben a Család II. mintájának 19,7 százaléka él egyedül.

Az alábbiakban a minta nem- és korösszetételét vizsgáljuk. A 6. tábla a minta korcsoport és nem szerinti százalékos megoszlását mutatja.

6. tábla

A Család II. mintájának nem- és korösszetétele (n=1500)

Korcsoport (éves)	Nem (százalék)		
	Férfi	Nő	Együtt
18–39	16,47	18,27	34,74
40–59	14,93	19,40	34,33
60 és idősebb	10,93	20,00	30,93
Együtt	*42,33*	*57,67*	*100,00*

A 7. táblában az 1994. év eleji sokasági arányokat láthatjuk.

7. tábla

A magyarországi populáció 1994 elején (n=7 855 559)

Korcsoport (éves)	Nem (százalék)		
	Férfi	Nő	Együtt
18–39	20,64	20,11	40,75
40–59	16,24	17,73	33,97
60 és idősebb	10,02	15,26	25,28
Együtt	*46,90*	*53,10*	*100,00*

Forrás: Magyar statisztikai évkönyv, 1993.

Összevetve az arányokat, szembeszökő a férfiak vártnál alacsonyabb számú előfordulása. A korcsoportokat nézve a középkorúak sokasági arányuknak megfelelően találhatók a mintában, míg a fiatalok alul-, az idősebbek felülreprezentáltak. A korcsoportok megoszlásának eltérése szignifikáns (khi-négyzet próbával vizsgálva, 0,000 szinten). A minta és a sokaság különbözősége a nemeket tekintve is szignifikáns (0,000 szinten). Ha nemcsak a marginálisok eltérését vizsgáljuk, hanem egyszerre a táblázat hat cellájának illeszkedését is,[10] megint csak szignifikáns eltérést kapunk (0,000 szinten).

A kulcs torzító hatását akkor lehetne érdemben elemezni, ha a kétlépcsős mintavétel lépcsőnként lenne jellemezhető. Ugyanis a fenti eltérések csak akkor tulajdoníthatók a végső lépcsőbeli kiválasztásnak, ha feltételezhetjük, hogy a háztartások egyforma valószínűséggel kerültek a mintába.

[10] A két nem és három korcsoport kategória kereszt-kombinációikból képzett hatértékű változó eloszlására végeztünk khi-négyzet tesztet.

Viszont megkísérelhető a kulcs okozta hatás vizsgálata, ha ismert a minta háztartásainak összes tagjából összeállított pszeudominta összetétele. Ekkor ugyanis ennek a pszeudomintának a nem- és korösszetételéhez viszonyíthatjuk a tényleges mintabeli megoszlásokat. Továbbá a pszeudominta és a sokasági adatok összevetésével a háztartásminta kiválasztási lépcsője is értékelhető.

A Család II. felvétel esetében a rendelkezésünkre állt nemcsak a kérdezettek, hanem a háztartások minden tagjának nemére és korára vonatkozó információ is. 47 esetben nem ismert valamelyik lakó neme vagy kora. Ezekkel a hiányokkal az eredeti 1500-as mintából 4397 használható személyi rekordhoz jutunk; a kérdezettek közül 3398-an töltötték be a 18-adik életévüket. A pszeudominta arányainak százalékos megoszlását a 8. tábla mutatja.

8. tábla

A Család II pszeudomintájának nem- és korösszetétele (n=3389)

Korcsoport (éves)	Nem (százalék)		
	Férfi	Nő	Együtt
18–39	19,50	19,92	39,42
40–59	16,50	19,21	35,71
60 és idősebb	10,00	14,87	24,87
Együtt	*46,00*	*54,00*	*100,00*

A pszeudominta és az eredeti személyi minta összehasonlításában a tendenciák egyeznek az előbbi összehasonlításkor kapottal: a személyi mintában felülreprezentáltak az idősek, míg a fiatal korosztály alulreprezentált. Az eltérés szignifikáns (0,000 szinten). Az eltérés a nemeket tekintve is szignifikáns (0,004 szinten). A marginálisokon kívül a hat kategória illeszkedése is szignifikáns eltérést mutat (0,000 szinten).

A két összehasonlítás tapasztalatai egyeznek, mert a pszeudominta igen jól illeszkedik a sokasághoz. A pszeudominta és a sokaság összehasonlításában most is alulreprezentáltak a férfiak, de az eltérés nem szignifikáns (p=0,31). A korcsoportok eltérése sem szignifikáns (p=0,09), és a cellákat vizsgálva a hat kategória illeszkedése is elfogadható (p=0,222).

Eredményeink lehetővé teszik a mintaválasztás egymástól szétválasztott lépcsőfokainak értékelését. A pszeudominta jól illeszkedik a várható arányokhoz. Eszerint a háztartásválasztás véletlen volta nem cáfolható, a háztartásminta még reprezentatívnak minősíthető. A személyválasztási eljárást, azaz a kulcsot a hiba okai között önálló tényezőként különítettük el.

3.3. A kulcs mint a reprezentativitási problémák forrása – a feltételezés bizonyítása

A fenti eredményt megkérdőjelezheti az, hogy csak egy mintát vizsgáltunk, és annak összetétele akár a véletlen egyszeri játékával is magyarázható. Tegyük most ezt az értékelést jobban általánosíthatóvá.

Elemzésünkben a TÁRKI adatbázisát, a Magyar Háztartási Panel IV. hullámfelmérést használtuk. A kutatást vezető intézmények a BKE Szociológiai Tanszék és a TÁRKI. Az

adatfelvétel időpontja 1995. Módszere: kérdőíves adatfelvétel. Esetszám: 8043. Ez országos háztartási minta, 6306 személy adatait tartalmazza a bázis. Kizártuk az elemzésből azokat a személyeket, akiknek hiányzott az információ a koráról vagy a neméről. A többiek közül 4188-an töltötték be 1994-re a 18-adik életévüket. Így 4188 személy – 1986 háztartás – adatait használtuk fel. 813 esetben hiányzott az információ valamely háztartástag neméről vagy koráról, de elsősorban nem válaszmegtagadás miatt, nem vettek fel adatokat a 16 éven aluliakról, a háztartásból kivált, de még nyilvántartott személyekről stb.

Az elemzés során mintavételi szimulációt végzünk, azaz úgy kezeljük a mintát, mintha egy virtuális sokaság lenne. A továbbiakban a minta megnevezést is pszeudosokaságra váltjuk. Ezen a pszeudosokaságon teszteljük a Kish-kulcs használatának hatását a minták nem- és korösszetételére.

9. tábla

A Háztartás Panelből kapott sokaság összetétele (n=4188)

Korcsoport (éves)	Nem (százalék)		
	Férfi	Nő	Együtt
18–39	19,79	20,51	40,30
40–59	14,66	17,60	32,26
60 és idősebb	10,94	16,50	27,44
Együtt	*45,39*	*54,61*	*100,00*

Kiszűrendő a mintaösszetételnek a háztartásválasztás által okozott ingadozását, a kulcs teljesítményének értékelését nem egyetlen minta kiválasztásával értékeltük. Ehelyett a várható minta összetételét számítottuk ki. A várható minta összetételét a következőképpen kaphatjuk: egy adott nem/korcsoport kategória mintabeli arányának mint valószínűségi változónak a várható értéke adja a kategória várható mintabeli arányát (részletesebben lásd alább az /1/ egyenletet).

10. tábla

Várható mintaösszetétel Kish-kulcs alkalmazásakor

Korcsoport (éves)	Nem (százalék)		
	Férfi	Nő	Együtt
18–39	17,27	19,47	36,74
40–59	12,92	17,04	29,96
60 és idősebb	11,87	21,43	33,30
Együtt	*42,06*	*57,94*	*100,00*

Megfigyelhetjük az idős korosztály felülreprezentáltságát, leginkább az idős nőkét. A másik két korosztály a kellőnél kisebb számban szerepel, főként a fiatal és a középkorú férfiak. A férfiak a vártnál kisebb számarányban képviseltetik magukat. Az eltérések megegyeznek a valós felmérések tapasztalataival.

Eredményeink szerint tehát szisztematikus torzítás lép fel a Kish-kulcsot alkalmazó mintavételben. Ezt a torzítást maga a mintavétel módja okozza: az, hogy minden háztartásból a véletlent imitáló módon, azonos valószínűségekkel választunk. Nem a kulcs okozza a hibát, ugyanez a probléma lép fel például a legutóbbi születésnap módszerének alkalmazásánál is. Ugyanakkor a kulcs esete azért szerencsés, mert használatakor ellenőrizhető a minta nem és kor szerinti összetétele, azaz esetleg javítható is.

4. A KISH-KULCS MÓDOSÍTÁSA

Az alábbiakban a kulcs módosítására teszünk kísérletet, olyan céllal, hogy a minta a nem/korcsoport jellemzők szerint minél közelebb legyen a sokasági paraméterekhez.

A Kish-kulcs módosításakor a mintavételi eljárás alapvető vonásait állandónak tételezzük fel. Így nem változtatunk azon, hogy a háztartások azonos valószínűséggel kerülnek a mintába, s hogy minden háztartásból pontosan egy személyt választunk ki. A kulcsból azt a jellemzőt is megtartjuk, hogy a háztartás tagjait nemük és koruk szerint sorba tesszük, és a kérdőívhez csatolt táblából keresheti ki a kérdező, hogy a sor hányadik tagját kérdezze. A módosítás csak ezekre a kiválasztási táblákra terjed ki.

Felhasználva a sokaság háztartásszerkezetének ismeretét, a kiválasztási táblákból egyértelműen meghatározhatjuk a minta várható nem- és korcsoportszerkezetét. A feladat tehát olyan táblák elkészítése, amelyek használata esetén a minta nem és korcsoport szerinti összetétele megegyezik, vagy a „lehető legközelebb van" a sokasági megoszláshoz. A feladat még egy paramétert tartalmaz, ez a kérdőívekhez csatolt táblák típusainak száma. Minél több táblafajtát használhatunk, annál finomabb valószínűség-eloszlás jön létre, s vélhetően annál közelebb hozható a minta a sokasághoz. A Kish-kulcs eredetije 8 táblát használ fel, és ezeket különböző arányban osztja ki a kérdőívek között, 1/12-ekben adva meg a kiosztási arányt. Az alábbi módosításban 12-ben rögzítjük a táblatípusok számát, míg az egyes táblatípusok kiosztási arányát továbbra is 1/12-ekben adjuk meg. A meghatározásra váró ismeretlenek tehát a 3. táblához hasonló tábla elemei, 12×6 érték.

Az ismeretlenek segítségével egyértelműen meghatározható, hogy egy i létszámú háztartásnak a nem/kor szerinti sorban j-edik tagja mekkora valószínűséggel kerül kiválasztásra, feltéve, hogy háztartása már bekerült a mintába. Ha $j > 6$, ez a valószínűség 0. Más esetben ez a valószínűség 12-ed része annak, ahányszor a keresett táblázat i-edik oszlopában a j szerepel. Tehát a kulcs módosításának feladata egyszerűsíthető: nem a táblákat kell meghatározni, csak azokat a valószínűséget (12-edekben), amekkora eséllyel a táblák használatával az i tagú háztartások j-edik tagja kiválasztásra kerül. Jelöljük a meghatározandó feltételes valószínűségeket p_{ij}-vel (i tagú háztartás sorban j-edik tagjához tartozó kiválasztási valószínűség, miután a háztartás már bekerült a mintába). A 12-edekben adott p_{ij} valószínűségek segítségével a 12 tábla már egyértelműen meghatározható; például p_{ij}=5/12 esetén a keresett táblázat i-edik oszlopában a j-t 5-ször kell szerepeltetni (az oszlop tetszőleges celláiban). Az ismeretlenek tehát: p_{21}, p_{22}, p_{31}, p_{32}, p_{33},…, p_{61}, p_{62}, p_{63}, p_{64}, p_{65}, p_{66}. (Értelmezhető p_{11} is, ennek értéke nyilván 1.)

Az i tagú háztartásban lakó j-edik személy kiválasztási valószínűsége a háztartás kiválasztási valószínűségének és az egyén háztartáson belüli feltételes kiválasztási valószínűségének szorzata. Az utóbbi érték a p_{ij}-vel jelölt ismeretlen paraméter, az előbbi érték viszont a sokaság háztartásszerkezetének függvénye. Mivel a háztartások azonos valószí-

nűséggel kerülnek a mintába, az i létszámú háztartás kiválasztási valószínűsége az i tagú háztartások sokasági arányával egyezik meg. Jelölje ezt az értéket H_i, a feladat input paramétere tehát: $H_1,\ldots,H_6$.

A feladat megoldásához szükség van annak valószínűségére, hogy egy i tagú háztartássorban j. tagja fiatal férfi, fiatal nő, középkorú férfi, középkorú nő, idős férfi, illetve idős nő. Ez 3×2 érték, amely egy 3×2 dimenziós mátrixszal adható meg. A fentieken kívül a feladat input paramétere tehát összesen még 21 darab mátrix, jelölve: $\mathbf{a}_{11}$, $\mathbf{a}_{21},\ldots,\mathbf{a}_{66}$. Például $a_{21}[11]$ jelöli annak a valószínűségét, hogy egy 2 tagú háztartás sorban első tagja fiatal férfi, $a_{21}[21]$ annak a valószínűségét, hogy ugyanő középkorú férfi, $a_{21}[12]$ pedig annak valószínűségét, hogy fiatal nő.

Most már megadható a mintában várható adott nemű, adott korcsoporthoz tartozó személyek aránya, és ez a hat érték az előbbi módon ismét egy 3×2 dimenziós mátrixba rendezhető. Jelölje a mátrixot **a**. $a[11]$ a fiatal férfiak arányát jelöli, $a[21]$ a középkorú férfiak arányát stb. Ezek az értékek kifejezhetők a fenti paraméterek függvényében:

$$a[ij] = \sum_{k=1..6} H_k \left(\sum_{l=1..k} p_{kl} a_{kl}[ij] \right) \quad i = 1, 2, 3 \; j = 1, 2. \qquad /1/$$

A célként jelölt sokasági nem/korcsoport megoszlás szintén egy 3×2 dimenziós mátrixszal adható meg. A mátrixot jelölje **A**, cellái a mintabeli megoszlást jelölő $a[ij]$-khez hasonlóan értelmezhetők. Így például $A[11]$ a fiatal férfiak sokasági arányával egyezik meg. A feladat szerint keressük p_{ij}-k függvényében azon $a[ij]$ értékeket, amelyeknek értéke a megfelelő $A[ij]$ értékkel egyezik meg, ekkor a minta nem/korcsoport megoszlása pontosan reprodukálja a sokaságét.

A feladat megfogalmazása tehát: adott az **A** mátrix, adottak az $\mathbf{a}_{21},\ldots,\mathbf{a}_{66}$ mátrixok. Keressük azon $p_{21},\ldots,p_{66}$ skalárokat, amelyek kielégítik az alábbi egyenletet:

$$\sum_{i=1,2,3\, j=1,2} \left| a[ij] - A[ij] \right| = 0, \qquad /2/$$

az alábbi feltételekkel:

$$\sum_{j=1\ldots i} p_{ij} = 1 \quad \forall i\text{-re},$$
$$p_{ij} > 0 \quad \forall i, j\text{-re},$$
$$p_{ij} = k_{ij}/12 \quad \forall i, j\text{-re, ahol } k_{ij} \text{ egész.}$$

0-nál nagyobb kiválasztási valószínűségeket kívánunk meg, ugyanis a mintaválasztási eljárásokkal szembeni alapvető követelmények közé tartozik a pozitív kiválasztási valószínűségek megléte. A sokaság megváltoztatását jelentené a 0 értékű kiválasztási esély megengedése, hiszen ezzel szűkítenénk a kiválaszthatók körét.

A Microsoft Excel Solver bővítményprogramját használjuk a feladat megoldásának keresésére.[11] A feladat egy többváltozós nemlineáris egyenlet egészértékű megoldását keresi, korlátozó feltételek mellett.

[11] A Microsoft Excel Solver az általános redukált gradiens (GRG2) nemlineáris optimalizálási eljárást használja. A lineáris és az egész értékű problémákra a változókat korlátozó szimplex módszert, valamint az elágazás és korlátozás (branch-and-bound) módszert alkalmazza.

Az input paraméterek megadásakor a korábban is használt Magyar Háztartás Panel mintáját szerepeltetjük sokaságként. A sokasági nem/korcsoport arányok, az $A[ij]$-k értéke a 9. táblában szerepel.

Az egyenletnek nincs olyan megoldása, amely a feltételeket kielégítené.

Felvethető az a kérdés, hogy megoldható-e a probléma, ha nem köt minket a táblatípusok számának korlátja. A táblák számának növelése nyilván költségnövekedést eredményez, nagyobb adminisztrációt, nagyobb nyomdaköltséget igényelhet. Azonkívül a minta nagysága maga is felső korlátja a kioszthatő táblák számának. Az elvi probléma azonban végiggondolásra érdemes. A probléma megfogalmazásakor ekkor a kiválasztási valószínűségekre vonatkozó feltételek közül elhagytuk az egészértékűségi feltételt, és a pozitivitás feltétel helyett 1/100-os alsó határt adtunk meg.

A feladatnak azonban ilyenkor sincs megoldása.

Tehát nem reprodukálható pontosan a sokaság nem/korcsoport megoszlása. Ehelyett kereshetjük azokat a táblákat, amelyek használata esetén a sokasághoz „legközelebb levő" mintát kapunk. A legközelebb levő fogalmának operacionalizálásához a távolság fogalmát is meg kell határozni. A távolságot kétféleképpen határozhatjuk meg. Elsőként a sokaság és a minta távolságát a khi-négyzet statisztikához hasonlóan határozzuk meg: a mintából várt és a sokasági arányok négyzetes eltérésének és a várt aránynak a hányadosát összegeztük a hat cellára:

$$f(a) := \sum_{i=1,2,3\ j=1,2} (a[ij] - A[ij])^2 / A[ij]. \qquad /3/$$

Az optimális táblák esetén ez a távolság minimális, azaz most a feladat megfogalmazása:

$$f(a) \to \min,$$

a korábbi feltételekkel:

$$\sum_{j=1\ldots i} p_{ij} = 1 \quad \forall i\text{-re},$$
$$p_{ij} > 0 \quad \forall i, j\text{-re},$$
$$p_{ij} = k_{ij}/12 \quad \forall i, j\text{-re, ahol } k_{ij} \text{ egész.}$$

Ez egy nemlineáris egészértékű optimalizálási probléma, megszorító feltételekkel.

Ugyanakkor más megközelítésben is definiálhatjuk a távolság fogalmát. A másik távolságfüggvény alkalmazása a felmérések elemzésekor gyakran alkalmazott súlyozási eljárással kapcsolatban merül fel. A súlyozás célja a torzítások csökkentése. A súlyok alkalmazásakor viszont gyakran megnő a becslések varianciája. (Lásd a 6. lábjegyzetet.) A Kish-kulcs módosításakor a nem és kor szerinti reprezentativitást tartottuk szem előtt. A minta megoszlásának eltérése a kívánatostól súlyozással korrigálható, ez az utólagos rétegzés a minta nem és korcsoport arányait illeszti a sokaságéhoz. A variancia növekedése ebben az esetben a súlyok változékonyságával hozható összefüggésbe. Pontosabban, a variancianövekedés a súlyok négyzetösszegének monoton függvénye. Így adódik a kulcs újabb szempont szerinti módosítása: keressük azt az eljárást, amely a minimális

négyzetösszegű súlyokat eredményezi (a súlyok összege itt állandó: 1). A korábbi jelölésekkel a súlyok négyzetösszege:

$$\begin{aligned}\sum_{k=1..n} W_k^2 &= \sum_{i=1,2,3\ j=1,2} \sum_{k:korcsoport=i\ nem=j} W_k^2 \\ &= \sum_{i=1,2,3\ j=1,2} \sum_{k:korcsoport=i\ nem=j} A[ij]^2 / a[ij]^2 \\ &= \sum_{i=1,2,3\ j=1,2} n_{ij} A[ij]^2 / a[ij]^2 = \sum_{i=1,2,3\ j=1,2} n_{ij} A[ij]^2 / (n_{ij}/n)^2 \\ &= n\sum_{i=1,2,3\ j=1,2} A[ij]^2 / (n_{ij}/n) = n\sum_{i=1,2,3\ j=1,2} A[ij]^2 / a[ij],\end{aligned}$$

ahol n a mintanagyság, n_{ij} az i-edik korcsoporthoz és a j-edik nemhez tartozók száma a mintában.

Mint látható, a súlyok négyzetösszege függ a mintanagyságtól. Ez a függés kiküszöbölhető, ha a mintanagyságot rögzítettnek tekintjük. Ekkor az optimumkeresés során a négyzetösszeg helyett a

$$g(a) := \sum_{i=1,2,3\ j=1,2} A[ij]^2 / a[ij] = (1/n)\sum_{k=1..n} W_k^2 \qquad /4/$$

függvényt optimalizáljuk. A korlátozó feltételek ugyanazok, mint fent, az f függvény használata esetén.

A 10. táblában látható, hogy a Kish-kulcs alkalmazása esetén milyen $a[ij]$-ket kapunk, azaz milyen a mintabeli várható aránya az adott nemű, adott korcsoporthoz tartozó személyeknek Ezeket az értékeket f-be helyettesítve a függvény értéke 0,021573301. Ha találunk olyan paraméterértékeket, ahol az f ennél kisebb értéket vesz fel, akkor a Kish-kulcsnál jobb nem/korcsoport illeszkedést produkáló kiválasztási eljárást határozhatunk meg.

Hasonlóan, a g-be helyettesítve a Kish-kulcs alkalmazásakor kapott $a[ij]$-ket 1,018901199-et kapunk. Ennél kisebb értéket adó paraméterértékek megtalálása esetén olyan kiválasztási eljárást határozhatunk meg, amely az utólagos súlyozás szempontjából kedvezőbb mintát produkál.

A függvényeknek létezik elvi optimuma a megadott tartományon. f és g ugyanis folytonos függvény, a megadott tartomány zárt és korlátos. *Weierstrass* tétele szerint a folytonos függvénynek van legkisebb és legnagyobb értéke korlátos és zárt tartományon. Az elvi optimum megtalálása azonban nem egyszerű matematikai probléma. Speciális esetektől eltekintve a nemlineáris optimalizálási feladatok olyan algoritmussal oldhatók meg, mellyel általában nem garantálható, hogy a talált optimum globális és nem csupán lokális szélsőérték. A megoldás keresésére ismét az Excel Solver programját használjuk.

Mindkét feladatot megoldjuk az egészértékűségi feltétel mellett (tehát 12 tábla használatával) és annak elhagyásával is. A 11. tábla foglalja össze az eredményeket.

Az **a** mátrixokat összehasonlítva megállapíthatjuk, hogy a négy variáns gyakorlatilag egyező eredményt ad: az értékek első két tizedes jegyre kerekítve megegyeznek a négy esetben. Eszerint nem ad jobb eredményt az egészértékűségi feltétel elhagyása, tehát nem érdemes 12-nél több táblával dolgozni.

Összevetve a 9. táblabeli sokasági nem/korcsoport megoszlást az eredményül kapott $a[ij]$-kel, láthatóan javul a Kish-kulcshoz képest a fiatal korosztályok illeszkedése,

ugyanígy javul az idős nők aránya. Viszont a középkorú nők és az idős férfiak részaránya kissé távolabb kerül a sokaságtól, mint a Kish-kulcs esetén.

Az eredményül kapott p_{ij} értékeket értelmezve elmondható, hogy a várt tendenciák tapasztalhatók. A kétszemélyes háztartásokból a Kish-kulcsban adottnál nagyobb valószínűséggel választjuk az első, várhatóan inkább férfi tagot, a nők Kish-kulcsos felülreprezentáltságát korrigálva. Ugyanígy a három- és négytagú háztartásokból sorban utolsó, várhatóan fiatal lány tagot választjuk nagyobb valószínűséggel, korábbi alulreprezentáltságukat javítva.

11. tábla

Optimalizálási eredmények

Eredeti Kish-kulcs										
	Függvényértékek a Kish-kulcs behelyettesítésekor	Várható nem/kor eloszlás (a mátrix)		p_{ij}-k						
	az *f* értéke: 0,021573301 a *g* értéke: 1,018901199	0,1727	0,1947	p_{21}	1/2	p_{31}	1/3	p_{41}	1/4	
		0,1292	0,1704	p_{22}	1/2	p_{32}	1/3	p_{42}	1/4	
		0,1187	0,2143			p_{33}	1/3	p_{43}	1/4	
								p_{44}	1/4	
				p_{51}	1/6	p_{61}	1/6			
				p_{52}	1/6	p_{62}	1/6			
				p_{53}	1/4	p_{63}	1/6			
				p_{54}	1/6	p_{64}	1/6			
				p_{55}	1/4	p_{65}	1/6			
						p_{66}	1/6			
f optimalizálása										
Korlátozó feltételek:	optimum	Az optimummal kapott várható nem/kor eloszlás (a mátrix)		Optimumhely (p_{ij}-k)						
$\Sigma_{j=1..i}p_{ij}=1$ $\forall i$-re $p_{ij}>0$ $\forall i, j$-re $p_{ij}=k_{ij}/12$ $\forall i, j$-re, ahol k_{ij} egész szám.	0,013938589	0,1856	0,1994	p_{21}	2/3	p_{31}	2/12	p_{41}	1/12	
		0,1281	0,1616	p_{22}	1/3	p_{32}	1/12	p_{42}	3/12	
		0,1311	0,1942			p_{33}	9/12	p_{43}	1/12	
								p_{44}	7/12	
				p_{51}	1/12	p_{61}	1/12			
				p_{52}	8/12	p_{62}	7/12			
				p_{53}	1/12	p_{63}	1/12			
				p_{54}	1/12	p_{64}	1/12			
				p_{55}	1/12	p_{65}	1/12			
						p_{66}	1/12			
$\Sigma_{j=1..i}p_{ij}=1$ $\forall i$-re $p_{ij}>0$ $\forall i, j$-re	0,013216638	0,1875	0,2015	p_{21}	0,6539	p_{31}	0,2351	p_{41}	0,0100	
		0,1293	0,1584	p_{22}	0,3461	p_{32}	0,0100	p_{42}	0,3728	
		0,1307	0,1926			p_{33}	0,7549	p_{43}	0,0100	
								p_{44}	0,6072	
				p_{51}	0,0100	p_{61}	0,0100			
				p_{52}	0,8877	p_{62}	0,9500			
				p_{53}	0,0100	p_{63}	0,0100			
				p_{54}	0,0823	p_{64}	0,0100			
				p_{55}	0,0100	p_{65}	0,0100			
						p_{66}	0,0100			

(A tábla folytatása a következő oldalon.)

(Folytatás.)

g optimalizálása

Korlátozó feltételek:	Optimum	Az optimummal kapott várható nem/kor eloszlás (a mátrix)		Optimumhely (p_{ij}-k)					
$\Sigma_{j=1..6} p_{ij}=1$ $\forall i$-re	1,012869186	0,1854	0,1960	p_{21}	2/3	p_{31}	2/12	p_{41}	2/12
$p_{ij}>0$, $\forall i, j$-re		0,1314	0,1607	p_{22}	1/3	p_{32}	1/12	p_{42}	3/12
$p_{ij}=k_{ij}/12$ $\forall i, j$-re,		0,1323	0,1942			p_{33}	9/12	p_{43}	1/12
ahol k_{ij} egész szám.								p_{44}	6/12
				p_{51}	1/12	p_{61}	1/12		
				p_{52}	7/12	p_{62}	7/12		
				p_{53}	1/12	p_{63}	1/12		
				p_{54}	2/12	p_{64}	1/12		
				p_{55}	1/12	p_{65}	1/12		
						p_{66}	1/12		
$\Sigma_{j=1..6} p_{ij}=1$ $\forall i$-re	1,012269154	0,1863	0,1993	p_{21}	0,6574	p_{31}	0,2543	p_{41}	0,0270
$p_{ij}>0$, $\forall i, j$-re		0,1311	0,1591	p_{22}	0,3426	p_{32}	0,0100	p_{42}	0,3886
		0,1323	0,1919			p_{33}	0,7357	p_{43}	0,0100
								p_{44}	0,5744
				p_{51}	0,0100	p_{61}	0,0100		
				p_{52}	0,5594	p_{62}	0,7638		
				p_{53}	0,0100	p_{63}	0,1962		
				p_{54}	0,4106	p_{64}	0,0100		
				p_{55}	0,0100	p_{65}	0,0100		
						p_{66}	0,0100		

A két függvény által az egészértékűségi feltétel mellett, illetve annak elhagyásával kapott eljárások nem különböznek lényegesen a p_{ij}-ket tekintve, és gyakorlatilag azonos eredményt adnak a várható mintaösszetétel szerint is. Továbbá, ha *f* optimumát *g*-be helyettesítjük, akkor *g* optimumához igen közeli értéket kapunk, és fordítva. Ez azt jelenti, hogy az optimális megoldások mindkét szempont szerint megfelelők. Mivel nem különböznek lényegesen a megoldások, válasszuk az *f* által az egészértékűségi feltétel mellett adott optimumot. A kapott p_{ij} értékhez a következő Kish-tábla illeszthető.

12. tábla

Módosított Kish-tábla

A tábla aránya	A tábla jelzete	A kiválasztott sorszáma, ha a felnőttek száma					
		1	2	3	4	5	6 vagy több
1/12	1.	1	1	1	1	1	1
1/12	2.	1	1	1	2	2	2
1/12	3.	1	1	2	2	2	2
1/12	4.	1	1	3	2	2	2
1/12	5.	1	1	3	3	2	2
1/12	6.	1	1	3	4	2	2
1/12	7.	1	1	3	4	2	2
1/12	8.	1	1	3	4	2	2
1/12	9.	1	2	3	4	2	3
1/12	10.	1	2	3	4	3	4
1/12	11.	1	2	3	4	4	5
1/12	12.	1	2	3	4	5	6

A 12. tábla közvetlenül használható a felmérésekben.

*

Legfontosabb megállapításaink a következők.

– A kulccsal készített magyarországi minták nem és kor szerinti jellemzők alapján lényegesen eltérnek a sokaságtól. Ennek elvi oka az, hogy kétlépcsős eljárásban, háztartási mintát alkalmazva a végső, személyi minta az aktuális háztartásszerkezet függvénye.

– A kulcs eredményesen módosítható, ha célunk a minta adott jellemzők szerinti reprezentativitásának javítása.

Sikerült bizonyítanunk a kulccsal kapott minta és az aktuális sokasági háztartásszerkezet kapcsolatát. Az általunk felkutatott források nem bizonyították ezt a kapcsolatot, sőt, általában nem is említették ezt a tényezőt. Másik eredményünk, a kulcs módosítása önmagában is értékelhető. De fontos szempont, hogy a megoldásra alkalmazott algoritmus sokkal szélesebb körben is alkalmazható. Használatával meghatározhatunk más, nem a kor és nem jellemzők szerinti reprezentativitást javító optimális kulcsvariánsokat is.

Természetesen felvetődnek további, még nyitva álló kérdések. Elemzésünket a nem és kor szerinti reprezentativitásra korlátoztuk. Vizsgálandó lenne a Kish-kulccsal kapott minták más, lényeges jellemzők szerinti illeszkedése. A módosítás eredményeként kapott eljárás tesztelése is szükséges lenne, ugyanis a nem és kor szerinti illeszkedés javulása nem feltétlenül jelenti a minta más jellemzők szerinti javulását. A módosított kulcs használatával megváltozott mintavételi valószínűségek is további vizsgálatokra szorulnak. Továbbá a kulcs módosítását csupán a Magyar Háztartás Panelből vezettük le, a megnyugtató érvényességhez szükséges lenne ugyanezt az optimalizáló feladatot a teljes magyar háztartásstruktúra jellemzőiből kiindulva, népszámlálási adatokon végrehajtani.

FORRÁS- ÉS IRODALOMJEGYZÉK

BINSON, D. – CANCHOLA, J. A. – CATANIA, J. A. (2000): Random selection in a telephone survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods. *Journal of Official Statistics*, 16. évf. 1. sz. 53–59. old.

BONVALET, C. – LELIEVRE, E. (1997): The transformation of housing and household structures in France and Great Britain. *International Journal of Population Geography*, 3. évf. 3. sz. 183–201.

Egészségi Állapot Felvétel (1994) – Életmód, kockázati tényezők (1996). Központi Statisztikai Hivatal, Budapest.

GROVES, R. M. (1989): *Survey errors and survey costs*. John Wiley and Sons, Inc., New York.

GROVES, R.M. – BIEMER, P. P. – LYBERG, L. E. – MASSEY, J. T. – NICHOLLS, W. L. – WAKSBERG, J. (szerk.) (1988): *Telephone survey methodology*. John Wiley and Sons, Inc., New York.

DE JONG, A. H. (1996): National household forecasts 1996: fewer and fewer couples are married. *Maandstatistiek van de Bevolking*, 45. évf. 5. sz. 18–27. old.

KENNEDY, J. M. (1993): *A comparison of telephone survey respondent selection procedures*. http://www.indiana.edu/~csr/aapor93.html

KISH, L. (1949): A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44. évf. 380–387. old.

KISH, L. (1965): *Survey sampling*. John Wiley and Sons, Inc., New York.

LAVRAKAS, P. J. (1993): *Telephone survey methods: Sampling, selection and supervision*. Applied Social Research Methods. 7. köt.

LEVY, P. S. – LEMESHOW, S. (1999): *Sampling of populations*. John Wiley and Sons, Inc. New York.

Magyar Statisztikai Évkönyv 1993 (1994). Központi Statisztikai Hivatal, Budapest.

Magyar Statisztikai Évkönyv 1999 (2000). Központi Statisztikai Hivatal, Budapest.

Market research methods. University of Southampton. http://lispstat.alcd.soton.ac.uk/am306/quant5.txt

Mikrocenzus, 1996. Főbb eredmények (1996). Központi Statisztikai Hivatal, Budapest.

Mikrocenzus, 1996. A népesség és a lakások jellemzői (1996). Központi Statisztikai Hivatal, Budapest.

OLDENDICK, R. W. – BISHOP, G. G. – SORENSON, S. B. – TUCHFARBER, A. J. (1988): A comparison of the Kish and Last Birthday Methods of respondent selection in telephone surveys. *Journal of Official Statistics*, 4. évf. 4. sz. 307–318. old.

PIAZZA, T.: *Respondent selection for CATI/CAPI (Equivalent to the Kish method)*. http://srcweb.berkeley.edu:4229/res/rsel.html
SCSU Survey, 1997. Minnesota State Lottery. http://www.lottery.state.mn.us/gambling/stcloud.html
Táblaképek az egészségről – A veresegyházi példa (2001). MTA Szociológiai Kutatóintézet – Fekete Sas Kiadó, Budapest.
VOSS, D. S. – GELMAN, A. – KING, G.: The Polls – A review. Preelection survey methodology: Details from eight polling organizations, 1998 and 1992. *Public Opinion Quarterly,* 59. évf. 98–132. old.
TARJÁNYI J.: *Módszertani problémák a telefonos közvélemény-kutatásokban*. http://www.c3.hu/scripta/scripta0/replika/1920/14vita.htm

SUMMARY

The problem of drawing a person from a household often occurs at the final stage of a survey design e. g. in telephone surveys, after contacting the households using random digit dialing. The Kish grid offers an algorithm for this random selection. The authors found that in contrast with the widely held opinion, the grid is not capable of presenting the population by gender and age. This misconception stems from the fact that when the Kish grid was developed in the 1950's, both randomness and representativity could be achieved through the method due to the household structure in the USA. On the basis of their calculation, this does not hold for the recent Hungary. Finally, the authors suggest a modification to the Kish grid that is more appropriate for selecting a representative sample.