A MINTANAGYSÁG ÉS A MEGHIÚSULÁSOK KAPCSOLATA REPREZENTATÍV FELVÉTELEKBEN*

ÉLTETŐ ÖDÖN – DR. MARTON ÁDÁM

Reprezentatív statisztikai felvételeket már a múlt században alkalmaztak. Ennek az volt a lényege, hogy a vizsgálni kívánt sokaságnak csak valamely célszerűen kiválasztott részéről gyűjtöttek adatokat, amelyeket összesítve az egész sokaságra vonatkozó következtetéseket vontak le.[1] A valószínüségi minták elméletét azonban csak *Jiři Neyman* lengyel matematikus dolgozta ki[2] az 1930-as években. A jelenleg használt valószínűségi minták azon az elven alapulnak, hogy

– pontosan ismerjük azokat a kritériumokat, amelyek alapján eldönthető, hogy egy elem a vizsgálni kívánt célsokasághoz tartozik-e vagy sem;

– a kiválasztási eljárás a lehetséges minták $M_1, ...,M$ véges számú halmazát eredményezi, és mindegyik M_i mintához ismert $\Pi_i > 0$ kiválasztási valószínűség tartozik (ez speciálisan azt is jelenti, hogy a célsokaság minden eleméhez is tartozik egy $P_i > 0$ adott kiválasztási valószínűség).

A valószínűségi minták legfontosabb előnye, hogy a mintából becsülhető a számított mutatók szórása, ami lehetővé teszi a mintavételi hiba becslését.

Esetenként az elemek kiválasztási valószínűsége még bonyolult felépítésű minták esetében is az egyes rétegekben azonos lehet, tehát a mintavételi terv rétegenként egyszerű véletlen kiválasztásnak tekinthető. Természetesen bármilyen más, előre meghatározott, például valamilyen nagysággal arányos valószínűségi mintavétel is elképzelhető. A továbbiakban tételezzük fel, hogy bármilyen bonyolult, rétegzett, többlépcsős mintavételi tervről legyen is szó, a kiválasztási valószínűségek a struktúra egyes részein belül azonosak.[3] A felvétel végrehajtása történhet postai úton, telefonos vagy személyes megkérdezés útján. Mint ismeretes, a terepmunka, a felvétel végrehajtása számos nehézségbe ütközik, amelyek közül a továbbiakban csak a meghiúsulásokkal foglalkozunk. Azt vizsgáljuk, milyen módon lehet a kívánt mintanagyságot, illetve azáltal a becslések elvárt megbízhatóságát elérni, illetve a meghiúsulások negatív hatását a

* Készült az Országos Kutatási Alap (OTKA) támogatásával. (385. sz. téma)

[1] Lásd: *A. N. Kiaer*: Reprezentative method of statistical surveys. Norwegian Academy of Science and Letters. II. The Historical Philosophical Section. I. 1897. No. 4. Oslo Reprint: CBS Norway, Oslo. 1976.

[2] *J. Neyman*: On the two different aspects of the representative method. *Journal of the Royal Statistical Society.* 1934. évi 97. sz. 558—625. old.

[3] Napjaink statisztikai gyakorlatában általában valószínűségi mintákat alkalmaznak. Előfordulnak azonban ún. kvóta minták vagy tudatos kiválsztáson alapuló felvételek is. Ezek tulajdonságainak vizsgálatával azonban nem kíván e tanulmány foglalkozni.

becslésekre csökkenteni. (Egyéb válaszolási, kódolási, feldolgozási stb. hibákkal most nem foglalkozunk. Azaz feltételezzük, hogy a rendelkezésünkre álló minta egyedi adatai pontosak.)

A MINTANAGYSÁG MEGHATÁROZÁSA

Reprezentatív felvételek tervezése során az egyik fontos kérdés a mintanagyság meghatározása. Ideális esetben a mintából kapott becslésekre vonatkozó pontossági követelmények alapján lehet a szükséges mintanagyságot számítani, feltéve, hogy – mint azt már említettük – rendelkezünk információval a vizsgálni kívánt ismérv(ek) alapsokaságbeli szóródásáról. A reprezentatív felvételek alapján általában nem egyetlen ismérv átlagát, értékösszegét stb. akarjuk becsülni, hanem több változóét, több ismérv szerinti eloszlást, esetleg változók kapcsolatát is. Így még ideális esetben sem lehet a szükséges mintanagyságot egyértelműen meghatározni, illetve a különböző ismérvekre különböző mintaelemszámok adódnak, hiszen azok alapsokaságbeli szórásai eltérők. A gyakorlatban persze sokszor nincs pontos információnk az alapsokaságbeli szórásokról, legfeljebb feltevéseink, továbbá a becslésekre vonatkozó pontossági követelmények sincsenek többnyire egzaktan megfogalmazva. Máskor egy-egy felvétel végrehajtásához, feldolgozásához és elemzéséhez rendelkezésre álló anyagi források szabnak határt a mintanagyságnak. Akárhogyan történik is azonban a minta elemszámának a meghatározása és a minta elosztása területi egységekre, ágazatokra (alágazatokra), vállalkozási formákra stb., fontos követelmény, hogy az adott felvételben ténylegesen közreműködő mintavételi egységek (lakások, háztartások, személyek, vállalkozások, telephelyek stb.) száma ne térjen el számottevően a kijelölt mintaelemszámtól.

A kijelölt mintanagyság és a minta struktúrájának biztosítása nem könnyű feladat. A gyakorlat azt mutatja, hogy az utóbbi években mind a lakossági, mind a gazdaságstatisztikai felvételeknél jelentős a meghiúsulási arány, olykor ez az 50–60 százalékot is eléri. Ennek következtében a tényleges mintanagyság – hacsak nem gondoskodunk előre ennek megfelelő szinten tartásáról – általában számottevően alatta marad a tervezettnek, ami természetesen negatívan befolyásolja a mintából kapott eredmények megbízhatóságát.

A nemválaszolás, ami felöleli nemcsak a megtagadásokat, hanem az egyéb okból történt meghiúsulásokat is, két módon hat az eredményekre:

a) a kisebb mintanagyság miatt a becslések mintavételi hibája a tervezettnél nagyobb lesz;

b) a rétegenként, területi egységenként eltérő meghiúsulási arány következtében a tényleges minta struktúrája nemcsak véletlen, hanem bizonyos szisztematikus hatások miatt is különbözik az alapsokaság összetételétől, ami kisebb-nagyobb mértékben torzíthatja az eredményeket.

Mielőtt a nemválaszolás hatásainak csökkentésére szolgáló módszerek tárgyalására rátérnénk, érdemes felhívni a figyelmet arra, hogy a meghiúsulások egy része esetenként nem igazán meghiúsulás, így ezeket a válaszolási arány számításánál sem kell figyelembe venni. Olyan esetben ugyanis, amikor a kiválasztás alapjául szolgáló mintavételi keret tökéletlen olyan értelemben, hogy az alapsokasághoz nem tartozó elemeket is tartalmaz – például a címjegyzékben üres lakások vagy nem lakásnak használt címek is szerepelnek, a regiszter már megszűnt vagy sohasem működött vállalkozásokat is tartalmaz stb. –, ilyen kiválasztott mintaelemeknél a nemválaszolás nem tekinthető meghiúsulásnak. A

mintanagyság ugyan csökken ebből adódóan, de a minta összetétele nem torzul szükségképpen. Sőt, ha a mintanagyság egy adott kiválasztási aránnyal lett meghatározva, ettől a tényleges alapsokaságra vonatkozó kiválasztási arány sem módosul.

Egy nem mezőgazdasági kisvállalkozókra (50 főnél kevesebb foglalkoztatott) vonatkozó 1993. évi felvétel végrehajtása során kiderült, hogy a mintavételi keret – a Központi Statisztikai Hivatal (KSH) regisztere – több mint 16 százalékban tartalmaz nem létező vagy a célsokasághoz nem tartozó vállalkozásokat, ami természetesen önmagában is csökkentette az eredetileg kétezresre tervezett mintelemszámot, hiszen a célsokaság elemszáma is körülbelül ugyanilyen arányban kisebb volt a feltételezettnél. A tényleges célsokaságra vonatkozó kiválasztási arányok – 0,1 százalék az egyéni vállalkozásokra és 1 százalék a társas vállalkozásokra – azonban ettől még nem módosultak. Más kérdés, hogy az egyébként jelentős arányú és differenciált válaszmegtagadás és egyéb meghiúsulások miatt a minta struktúrája végül eltért a tervezettől.

Utalnunk kell azonban arra, hogy a célsokaság, illetve a mintavételi keret hiányosságai eleve kihatnak a becslések megbízhatóságára. Ez bizonyos mértékig úgy korrigálható, hogy a mintavétel végrehajtása, a terepmunka során szerzett tapasztalatok alapján korrigáljuk a célsokaság számosságát, „nagyságát", ami végül is visszahat a felszorzásra, esetleg rétegenként különböző mértékben. Példa a fent említett kisvállalkozókra vonatkozó felvételből, illetve az ELAR-felvételek (Egységes Lakossági Adatfelvétel Rendszer) gyakorlatából található.

Tételezzük fel a továbbiakban, hogy a mintavételi keret önmagában már hibátlan, s így a minta sem tartalmaz hibát, nem létező vagy időközben átalakult elemeket. (Ez a feltétel a gazdaságstatisztika alapjául szolgáló regiszterek esetében nem teljesül, ami tehát az említettek szerinti korrekciót igényli. A lakossági felvételek esetében, részben a mintavételi keret – lakáscímek – folyamatos továbbvezetése eredményeként az említett feltétel jobban teljesül).

A MEGHIÚSULÁSOK HATÁSA A BECSLÉSEKRE

A nemválaszolás az esetek többségében azt eredményezi, hogy a minta és az alapsokaság struktúrája eltér egymástól, és ezért a mintából kapott eredmények torzítottak lesznek. Nézzük először is, hogyan jelentkezik ez a becsléseknél.

Jelölje valamilyen y ismérv esetén $\overline{Y}_1$ a válaszolók alapsokaságbeli átlagát, $\overline{y}_1$ a válaszolók mintabeli átlagát, $\overline{Y}_2$ a nemválaszolók alapsokaságbeli átlagát, W_1 és W_2 a két réteg súlyát az alapsokaságban.

A teljes sokaság átlaga: $\overline{Y} = \frac{Y}{N} = W_1\overline{Y}_1 + W_2\overline{Y}_2$

Az $\overline{y}_1$ mintaátlag relatív torzításának várható értéke: $RT(\overline{y}_1) = \frac{\overline{Y}_1 - \overline{Y}}{\overline{Y}} = W_2\frac{\overline{Y}_1 - \overline{Y}_2}{\overline{Y}}$

E formulából látható, hogy a mintaátlag relatív torzítása akkor jelentős, ha a nemválaszolók aránya (W_2) nagy és/vagy a két réteg átlaga ($\overline{Y}_1$ és $\overline{Y}_2$) erősen különbözik. Ha például a nemválaszolási arány 40 százalék és a válaszolók és nemválaszolók átlagának relatív eltérése 25 százalékos, akkor a mintaátlag relatív torzítása várhatóan 10

százalékos lesz, ami gazdasági idősorok esetén alapvetően megváltoztathatja a trendeket, de a személyes megkérdezésen alapuló lakossági felvételeknél is számottevően csökkentheti az eredmények megbízhatóságát.

ELJÁRÁSOK A MEGHIÚSULÁSOK KÖVETKEZMÉNYEINEK CSÖKKENTÉSÉRE

A jó, torzítatlan becslés feltételeinek a kijelölt minta felel meg. Minden lehetséges eszközzel arra kell törekedni, hogy a véletlenül kiválasztott minta célsokasághoz tartozó minden eleméről a kívánt információkat beszerezzük. Megfelelő erőfeszítések jelentős eredményeket hozhatnak, azonban valamilyen – némely esetben számottevő – meghiúsulással mindig számolni kell.

A gyakorlatban leginkább az előző pontban említett torzítás jelenti a nagyobb veszélyt az eredmények megbízhatósága szempontjából, már csak azért is, mert nehezebb megfelelő módszert találni a torzítás kiküszöbölésére vagy legalábbis csökkentésére.

A) Helyettesítés

Több országban elterjedt gyakorlat, amit néhány éve Magyarországon is alkalmaztunk például egyes ELAR-felvételeknél, hogy a minta kiválasztása során eleve pótmintaelemeket (például pótcímeket) is kiválasztunk, amelyek akkor kerülnek felhasználásra, ha az elsődleges mintaelemek valamelyikénél valamilyen ok miatt meghiúsul a felvétel. Ha rétegenként, területi egységenként stb. elegendő pótmintaelem áll rendelkezésre, ezzel a helyettesítő módszerrel általában biztosítani lehet az előírt mintaelemszámot. A pótcímek használata – ami az imputálás (lásd később) speciális esetének is tekinthető – azonban a Neyman-féle megközelítés alapján vitatható, s általában a nemzetközi szakirodalom sem ajánlja.[4] Mindazonáltal a helyettesítési eljárásnak, a pótcímek használatának lehetnek bizonyos előnyei, illetve ezek esetenként nagyobbak, mint a hátrányok,

– ha a helyettesítés megfelelő szinten (például település, körzet, alágazat, vállalkozási forma stb.) történik, megőrizhető a minta eredeti struktúrája, legalábbis a struktúra azon jellemzői, amelyeket a mintavételi terv tartalmazott;

– minimális lehet a helyettesítésből eredő torzítás, ha az eredeti és pótcímek kiválasztása megfelelően rétegzett mintavételi keret ugyanazon rétegeiből történik (például a népességregiszter ugyanazon a településen élő, azonos nemű és korcsoportú személyei közül, gazdálkodó egységek esetén ugyanazon alágazathoz, vállalkozási formához, nagyságkategóriához tartozó, ugyanolyan típusú településen bejegyzett vállalkozásai közül); ilyen típusú helyettesítést alkalmazott 1991-ig bezárólag, illetve készül újra alkalmazni 1996-tól a KSH a háztartási költségvetési felvételnél, ahol a mintába tartozó körzetek összes háztartása előzetes felmérés alapján háztartástípus és taglétszám alapján rétegekbe lett sorolva és meghiúsulás esetén a pótcím kiválasztása azonos rétegből történt (természetesen az ilyen típusú helyettesítés sem tud minden torzító hatást kiküszöbölni, mivel további tényezők – például a háztartási költségvetési felvételnél a jövedelmi szint – ilyen esetben is gyakorolhatnak torzító hatást az eredményekre);

– biztosítható a kívánt mintanagyság, valamint az összeírók egyenletes munkaterhelése, bár ez utóbbi a sikertelen felkeresések számának különbözősége miatt nem teljesen;

[4] 1994 szeptemberében, Ottawában nemzetközi szeminárium foglalkozott a meghiúsulások kezelésével lakossági felvételek esetében. A szlovén és a lengyel előadó (Lásd: *Vasja Vehovar:* The substitution procedure for the unit nonresponse, *Jan Kordos*: Nonresponse problems in the Polish household surveys) foglalkozott részletesen e problémával, feltárva annak előnyeit és hátrányait. Néhány gondolatot felhasználunk előadásaikból.

– pótcímek segítségével gyakorlatilag kiküszöbölhető, hogy egyes klaszterek (körzetek) egyáltalán ne szerepeljenek a megvalósult mintában, jóllehet a kiválaszott mintában benne voltak.

Az előbbiekben felsorolt előnyök ellenére a helyettesítési eljárás általában nem tudja megoldani a meghiúsulások miatt fellépő problémákat, egyik legnagyobb veszélye éppen az e tekintetben táplált hamis illúzió. Önmagában az a tény, hogy biztosítjuk az előírt mintanagyságot, egyáltalában nem biztosítja, hogy mintánk valóban tükrözi a vizsgálni kívánt sokaság fontos tulajdonságait.

A helyettesítéses módszer további hátrányai:

– ha az összeíró tudja, hogy használhat pótcímet, kisebb erőfeszítést tesz az eredetileg kijelölt háztartások meggyőzésére a felvételben való közreműködésre, így növekszik a nemválaszolási arány;

– még központilag megadott pótcímek esetén is az összeíró könnyebben választja a pótcímet nehezebben elérhető mintaelemek helyett (például fiatal egyedül álló helyett idősebb család vagy nyugdíjasok); ez másként úgy is fogalmazható, hogy a tényleges megfigyelések között túlsúlyba kerülnek azok, akik készségesebbek közreműködni, könnyebb otthon találni őket (például nők, nyugdíjasok);

– növeli a területi munka időszükségletét;

– esetenként előfordul, hogy a mintavételi keret tökéletlensége miatt a célsokasághoz nem tartozó mintaelemek helyett is pótmintaelemet használnak.

Bizonyos feltételezések mellett kimutatható, hogy a helyettesítési torzítás nagyobb, mint a nemválaszolás miatt fellépő torzítás (amikor is a meghiúsulásokat nem pótoljuk), bár a különbség általában nem túl jelentős. Az is bizonyítható bizonyos elfogadható feltételek fennállása esetén, hogy bár a nagyobb mintaelemszám miatt a szórás kisebb, mint ha nincs helyettesítés, a nettó helyettesítési torzítás viszont nagyobb, így a mindkettőt figyelembe vevő átlagos négyzetes eltérés (Mean Square Errors – MSE) általában helyettesítés esetén lesz nagyobb.

B) Nagyobb minta

Ha vannak bizonyos tapasztalataink a célsokaság viselkedéséről (szórás, meghiúsulási arány), akkor a tervezett mintanagyság biztosításának másik módja egy eleve nagyobb – a várható nemválaszolási arány alapján számított mértékben nagyobb – minta kiválasztása. Ilyen esetben természetesen célszerű minél differenciáltabban – területi egységenként, rétegenként, vállalkozási formánként, nagyságkategóriánként stb. differenciáltan – figyelembe venni a várható meghiúsulási arányt. A pótcímek alkalmazásával szemben ez az említett eljárás ugyan nem javít a minta nemválaszolás okozta torzítottságán, de legalább nem fokozza azt, így inkább ajánlható. A meghiúsulások helyett igénybe vett pótmintaelemeknél ugyanis épp olyan arányú nemválaszolás és ennek következtében ugyanolyan irányú torzítás várható, mint az eredeti mintaelemeknél.

1995-től két folyamatos ELAR-felvételnél, a munkaerő-felmérésnél, illetve a háztartási költségvetési felvételnél ezt a módszert alkalmazza a KSH. A számítások a különböző nagyságkategóriájú településeken 1994. I. félévben (munkaerő-felmérés), illetve 1993. II. félévben (háztartási költségvetési felvétel) az alapcímeknél tapasztalt válaszadási arányokon alapultak.

A munkaerő-felvételnél ezek az arányok a jelzett időszakban a megyék különböző nagyságú településein 85-87 százalék körül voltak, Budapesten valamivel 77 százalék

alatt. Az eredeti mintavételi terv szerint negyedévenként 8273 körzetben 3-3 címen, azaz összesen 24 819 címen kell a felmérést végrehajtani. Mivel egy körzetből három helyett négy cím kiválasztása a mintanagyság egyharmados növelését jelenti, ami Budapest kivételével jelentősen meghaladja a meghiúsulási arányt, a mintavételi terv által előírt mintaelemszámot oly módon lehetett tervezni a várható válaszadási arányok mellett, hogy a fővárosban és az 1994. I. félévben legalacsonyabb válaszadási arányú 2-20 ezer lélekszámú településeken körzetenként három helyett négy cím került kiválasztásra, a többi településen maradt az eredeti három cím. Ilyen kiválasztási eljárás mellett 24 750 közreműködő háztartásra lehetett számítani. Az 1995. I. félévi adatok azt mutatják, hogy a mintanagyság tekintetében helyesnek bizonyult a módosított kiválasztási eljárás, az I. negyedévben 24 940, a II. negyedévben 24 360 háztartásnál volt sikeres a munkaerő-felmérés, tehát ily módon lényegében el lehetett érni az eredeti mintavételi terv szerinti mintanagyságot.

A háztartási költségvetési felvételnél a lényegesen magasabb meghiúsulási arányok miatt nehezebben lehetett megoldani, hogy pótcímek használata helyett nagyobb induló minta biztosítsa az eredeti mintavételi tervben szereplő mintanagyságot. A *H* mintában szereplő 3221 számlálókörzet mindegyikéből éves szinten 3 cím, összesen 9663 cím, illetve háztartás jövedelmi és kiadási adatainak összeírását írta elő a mintavételi terv. Az előző években 100 százalékban pótcím alkalmazásával sem sikerült elérni a kívánt mintanagyságot. A meghiúsulási arány különösen a fővárosban volt magas. Itt 1993. II. félévben az elsődlegesen kiválasztott címeknek csupán 37 százaléka vállalkozott a felvételben való közreműködésre. A megyékben ennél magasabb, kisebb településeken 70 százalékos, nagyobb községekben és városokban 63 százalékos volt a közreműködési arány. Így 1995-ben az 5000-nél kisebb településeken körzetenként 4 cím került kiválasztásra a háztartási költségvetési felvétel céljára, az ennél nagyobb vidéki településeken körzetenként 5, Budapesten pedig 8 cím került a mintába. A számítások szerint, ha 1995-ben is hasonlók lesznek a közreműködési arányok, mint 1993. II. félévben, a mintából körülbelül 9640 közreműködő háztartásra lehet számítani. 1995. I. negyedévben közel 2750 háztartás működött közre a felvételben, azaz éves szintre vetítve több, mint ami várható volt az 1993. évi meghiúsulási arányok alapján. A KSH területi igazgatóságainak tapasztalatai szerint a többlet főként abból adódik, hogy a kisebb településeken a megnövelt minta (körzetenként 3 helyett négy cím) többségénél az összeírók sikeresen rá tudták beszélni a háztartást a felvételben való közreműködésre. (Hozzájárult a jobb válaszadási arányokhoz az összeírói hálózat szervezeti átalakítása is.) Az előző évek adatai viszont azt mutatják, hogy a II. és III. negyedévben – főleg a nyári hónapokban – kisebb a közreműködési készség, így lehet, hogy éves szinten nagyjából a várt mintanagyság fog adódni a közreműködő háztartások számát illetően.

C) A minta súlyozása az eredeti kiválasztási valószínűségek alapján

Olyan reprezentatív felvételeknél, ahol az egyes mintaelemek kiválasztási valószínűsége (P_i) például rétegenként különböző,

$$\sum_{i=1}^{N} P_i = n ,$$

az alapsokasági értékösszeg, Y torzítatlan becsléséhez az i-edik mintaelem megfigyelt y_i értékét a kiválasztási valószínűség reciprokával kell súlyozni, azaz

$$\hat{Y} = \sum_{i=1}^{n} y_i / P_i$$

Ez az ún. Horvitz-Thomson-becslés azonban csak akkor alkalmazható, ha minden mintaelemre ismerjük y_i értékét, azaz nincs nemválaszolás. Meghiúsulások esetén a fenti becslés az

$$\hat{Y}_r = \sum_{i=1}^{n} \frac{y_i}{P_i \cdot R_i}$$

formulára módosul, ahol R_i annak pozítiv és ismert valószínűsége, hogy az i-edik mintaelem közreműködik a felvételben. A gyakorlatban természetesen lehetnek olyan elemei a sokaságnak, akik vagy amelyek határozottan elzárkóznak a válaszadástól, akikre tehát R_i=0. De még ha el is tekintünk ettől, R_i általában nem ismert a priori. E problémát át lehet hidalni oly módon, hogy a mintát utólag olyan rétegekre – például területi egységek, ágazat, vállalkozási forma, háztartásnagyság, a háztartásfő társadalmi-aktivitási csoportja stb. szerint – bontjuk, amelyeken belül feltehető, hogy az R_i válaszadási valószínűségek közel azonosak, és ezeket a tényleges n_{hr}/n_h közreműködési arányokkal becsüljük, ahol n_h a h-adik utólag képzett réteg elemszáma az eredeti mintában, n_{hr} pedig a válaszolók elemszáma. Ha még az is fennáll, hog az eredeti P_i kiválasztási valószínűségek is azonosak a h-adik rétegben, akkor az alapsokasági értékösszegre az

$$\hat{Y}_r = \sum_{h=1}^{H} \frac{y_h}{n_h/N_h \cdot n_{hr}/n_h} = \sum_{h=1}^{H} N_h \bar{y}_{hr}$$

becslést alkalmazhatjuk, ahol:

H – a képzett rétegek száma,
N_h – a h-adik réteg elemszáma az alapsokaságban,
$\bar{y}_{hr}$ – a válaszolók átlaga a h-adik rétegben.

Érdemes felhívni a figyelmet arra, hogy a fenti becsléshez a mintavételi tervben szükségképpen szereplő P_i kiválasztási valószínűségeken kívül nincs szükség a mintán kívüli „külső" információkra.

D) Súlyozás külső információk felhasználásával

Utólagos rétegzés és átsúlyozás külső információk alapján is történhet, ha ezen információk alapján kiderül, hogy a minta struktúrája a különböző nemválaszolási arányok miatt bizonyos változó(k) szerint jelentősen eltér az alapsokaság struktúrájától. Ezzel az egyébként jelentkező torzítás határozottan csökkenthető. Ilyen külső információ lehet

például a népesség kor és nem szerinti összetétele területi egységenként, a lakott lakások teljes körű száma megyénként és település-nagyságcsoportonként, a vállalkozások száma ágazatonként, településtípusonként és vállalkozási formánként, esetleg más olyan adatok, amelyekre egyrészt teljes körű statisztikák állnak rendelkezésre, másrészt kapcsolatban lehetnek a nemválaszolási aránnyal.

A folyamatos munkaerő-felvételnél például több lépcsőben történik az utólagos rétegzés és korrekció. Első lépcsőben az eltérő válaszolási arányból adódó potenciális torzítás elkerülése érdekében a mintabeli adatok a teljes körű lakott lakások és a mintában közreműködő lakások arányával kerülnek teljeskörűsítésre részletes területi egységenként. Azonban még az ehhez alkalmazott rétegeken belül is eltérők a válaszadási arányok, például korcsoportonként, nemenként. Ezért utólagosan a személyi adatok megyénként város-község bontásban nemenként és korcsoportonként rétegezve vannak, és a továbbvezetett népességszámok alapján az első lépcsőben megállapított felszorzó faktorok korrigálásra kerülnek.[5]

Ezt az utólagos rétegzést és korrekciót egyébként más ELAR-felvételeknél is alkalmazzák, sőt olyan felvételeknél is, ahol nem a személy, hanem a háztartás a számbavételi egység, például a háztartási költségvetési felvételnél, külön program gondoskodik arról, hogy adott háztartáshoz tartozó személyek mind ugyanazt a súlyt kapják.

E) A nemválaszolók jellemzőinek megállapítása

Azt, hogy a nemválaszolás milyen mértékű torzításokat okozhat a becslésekben, annak alapján lehet behatárolni, ha releváns információink vannak arról, milyen jellemzők tekintetében térnek el a nemválaszolók a válaszolóktól. Ha a területi elhelyezkedésükön – esetleg ágazati hovatartozásukon – kívül semmit sem tudunk a nemválaszolókról, nehéz megbecsülni a nemválaszolás okozta torzítás mértékét, még nehezebb a mintát megfelelően korrigálni. Ezért minden olyan felvételnél, ahol jelentős arányú nemválaszolással kell számolni, alapvető fontosságú, hogy legyenek lényeges információink a nemválaszolók jellemzőiről. Ennek egyik módja, hogy a nemválaszolók közül veszünk egy kisebb mintát, tagjait postai felvétel esetén személyes felkereséssel, illetve kérdőbiztosok alkalmazása esetén jobb, tapasztaltabb összeírók alkalmazásával igyekszünk megnyerni a felvételben való közreműködésre, vagy legalábbis alapvető jellemzőiket próbáljuk megtudakolni. Postai úton végrehajtott gazdaságstatisztikai felvételeknél ez az eljárás információt szolgáltathat arra is, milyen arányú a regiszterben szereplő vállalkozások közül a már megszűntek vagy nem létezők száma, amit az adatok teljeskörűsítésénél figyelembe kell venni. A kisvállalkozókra vonatkozó, 1993 szeptemberében végrehajtott szociológiai felvételnél például egyértelműen kiderült, hogy a KSH regisztere, amelyből mint mintavételi keretből a kiválasztás történt, elég jelentős arányban tartalmaz nem létező, már megszűnt vagy más ágazatban, más vállalkozási formában működő cégeket. Mivel e felvételt gyakorlott ELAR-összeírók végezték, s nem postán küldték ki a kérdőíveket, a felvétel esetleges meghiúsulási okát minden mintaelemnél megbízhatóan meg lehetett állapítani, s így jól meg lehetett becsülni, hogy

[5] Részletesebben lásd: *Mihályffy László*: Meghiúsulások kompenzálása lakossági felvételekben: egy speciális lineáris inverz probléma. *Szigma*. 1994. évi 4. sz. 191-202. old.

a regiszter a különböző településtípusoknál, ágazatokban és vállalkozási formáknál milyen arányban tartalmaz az alapsokasághoz nem tartozó elemeket. Ez az információ aztán az adatok teljeskörűsítése során felhasználásra is került. Az utólagos, illetve megismételt személyes felkeresés akkor is fontos információkat szolgáltathat a nemválaszolók jellemzőire, összetételére, ha a felvételben való aktív közreműködésre ekkor sem sikerül őket rábeszélni. Ezek az információk felhasználhatók a válaszolók mintájának utólagos átsúlyozására.

Olyan reprezentatív felvételek esetén, ahol a minta utólagos átsúlyozása külső információk alapján nem tudja a nemválaszolás okozta torzítás döntő részét kiküszöbölni, mert a nemválaszolási arányt olyan tényezők is számottevően befolyásolják, amelyekre nem állnak rendelkezésre megbízható külső, teljes körű statisztikák, célszerű megkísérelni a kijelölt mintánál még a tényleges felvétel végrehajtása előtt néhány olyan információt megtudakolni, amelyek meghiúsulás esetén is alkalmasak a nemválaszolók összetételének jellemzésére. Ez történik 1995-ben a háztartási költségvetési felvételnél, ahol ismeretes, hogy például a háztartásfő kora, a háztartás gazdasági aktivitása, illetve taglétszáma is jelentősen befolyásolják a felvételben való közreműködési készséget. Így remény van rá, hogy e fontos változók szerint ismeretes lesz a teljes véletlen minta összetétele területi rétegenként, s amennyiben a közreműködő háztartások összetétele e változók szerint jelentősen eltér a teljes kijelölt minta összetételétől, utólagos rétegzéssel és átsúlyozással mód lesz e tényezők szerinti különböző válaszadási arány okozta torzítás kiküszöbölésére. Az I. negyedévi adatok szerint a megyék többségében ez az eljárás valóban alkalmas a nemválaszolók jellemzőinek megállapítására és ennek alapján a minta utólagos átsúlyozására. A fővárosban és néhány (például Pest) megyében ugyanakkor úgy tűnik, nem volt sikeres e kezdeményezés, mert a nemválaszolók többsége a jellemzésükre szolgáló néhány alapvető adatról sem volt hajlandó felvilágosítást adni.

Említést kell tenni végül még egy eljárásról, amelyet gyakran alkalmaznak a nemválaszolás kezelésére. Ez az eljárás az imputálás, azaz a nemválaszolók adatainak pótlása a minta véletlenszerűen kiválasztott elemének adataival. E módszer hatákonysága nagymértékben függ attól, mennyire tudjuk a nemválaszoló mintaelem helyett egy hozzá hasonló elem adatait imputálni. Ha csak azt tudjuk biztosítani, hogy azonos területi egységhez tartozó mintaelemek közül választjuk ki az imputáláshoz felhasználandó elemet, akkor az eljárás lényegében azzal egyenértékű, amikor eleve nagyobb mintát választunk ki a nemválaszolás okozta mintanagyság-csökkenés ellensúlyozására. Az imputálás ugyanakkor olcsóbb, hiszen nem kell több mintaelemet felkeresni (személyesen vagy levélben) és kitölteni a kérdőívet.

*

Sok szakstatisztikusban él az a tévhit, hogy a minta nagysága, pontosabban a begyűjtött válaszok száma dönti el a felvétel eredményeinek pontosságát, megbízhatóságát. Ez nincs így, mint az előzőkben elmondottakból is következik. Csak az elméletileg megfelelően kiválasztott minta egyedeinek válaszaiból kapott becslések azok, amelyek a kiszámítható keretek közötti pontossággal írják le a vizsgált paramétereket. Ha ez nem így lenne, nagyon egyszerű (s nem is nagyon költséges) megoldás kínálkozna: addig kell

címről címre, házról házra járni, amíg a kellő számú választ meg nem kaptuk. Különösen csalóka, vonzó ez az eljárás, ha azt az elvet követ(het)jük, hogy az első kimaradó, megtagadó egység helyett valami hozzá hasonlót keresünk. Ebben az esetben valójában felesleges lenne a nemválaszolás bonyolult, módszertanilag igényes, néha nehézkes, időrabló s nem is nagyon olcsó módszereivel bajlódni.

Más a helyzet a mintavételi keret (regiszter) hiányosságaival. Az ilyen irányú tapasztalatok eleve visszahatnak a célsokaság nagyságára, esetleg belső struktúrájára vonatkozó információkra. Ez a hiányosság a célsokaság terjedelmének korrekcióját igényli.

Végül az egész mintavételi terv és becslési eljárás függ természetszerűleg attól, hogy a rendelkezésre álló (feltételezetten helyes) információk, például a rétegsúlyok mennyire helytállók.

A kvótaminták azért is tűnhetnek vonzóknak, mert látszólag teljesen kiküszöbölik a meghiúsulások torzító, mintavételi hibát növelő hatását. Az más kérdés, hogy a kvótaminták megfigyelt adatai alapján a mintavételi hiba nem becsülhető! A pótcímzés, nagyon tágan értelmezve, illetve átvitt értelemben a kvótaminta egy változatának tekinthető. Arról természetesen nincs szó, hogy a kvótamintákból kapott eredmények egyáltalán ne lennének használhatók. A kvótaminták módszertani vizsgálata, elemzése azonban nem célja e tanulmánynak.

Végül még arra szeretnénk nyomatékosan felhívni a figyelmet, hogy a torzítás adott mintavételi terv, technika mellett a minta nagyságának növelésével nem csökkenthető. A torzítás becslése, kezelése, csökkentése általában más módszereket, lehetőleg további (külső) információk beszerzését igényli.

TÁRGYSZÓ: Reprezentatív mintavétel.

SUMMARY

The authors analyse in their study the correlation of failure and sample size in sample surveys.

They start out from the supposition that in a complex, stratified, multi-stage sampling design, the probability of selection and the response probability of any selected unit are the same in certain parts of the structure. Then they analyse how the required sample size and in this way the desired reliability of estimates can be achieved, that is how to reduce the negative effect of failures on estimates.

As a concluding thought they direct the attention of experts to the fact that, using specific sampling design and technique, bias can not be reduced through increasing the sample size. Estimating, handling and reducing bias usually require other methods (it may be, among other things, primarily extracting further outside information).