

Mintavételi módszerek ritka populációk esetén

Kapitány Balázs,

a KSH Népeségtudományi
Kutatóintézetének
tudományos titkára

E-mail: kapitany@demografia.hu

A tanulmány áttekinti azokat a legfontosabb statisztikai mintavételi eljárásokat, melyek olyan esetekben alkalmazhatók, amikor nem áll rendelkezésre az alapsokaságról közvetlen mintavételi keret.

A bemutatott eljárások: mintanagyság növelése; minta aszimmetrikus rétegzése; minta szűrése; közbenső mintavételi egység szintjén történő szűrés; többszörös/dupla mintavételi keret; kapcsolathasznosító módszerek (hálózati, illetve adaptív csoportos mintavétel) és rejtőzködő populációk esetén alkalmazható módszerek (térben és időben meghatározott, illetve válaszadó által vezérelt mintavétel).

Ezek alkalmazása költséghatékony módon teszi lehetővé adatok gyűjtését olyan speciális társadalmi csoportokról, amelyekről a hagyományos mintavételi alapokon nyugvó adatgyűjtések nem, vagy csak nagyon költségesen tudnak megbízható információt szolgáltatni.

TÁRGYSZÓ:
Mintavétel.
Statisztikai módszertan.

A mintavétel módszertani kérdései közül az egyik legérdekesebb, legtöbbet vitatott probléma a ritka populációk (rare populations) körében végzett mintavétel. Ritka populációk alatt olyan közösségeket, társadalmi csoportokat értünk, amelyekről közvetlen mintavételi keret (tehát az adott populáció tagjairól egy csaknem teljes körű lista) nem áll rendelkezésünkre (nincs, vagy létezik, de nem elérhető). Ilyenek például a magas vérnyomásban szenvedők, egy adott vallási, etnikai csoporthoz tartozók, a 100 kilónál nehezebb személyek, a munkásokat feketén foglalkoztató vállalatok, a HIV-pozitívak és így tovább.

Minél ritkább e csoport, mint részpopuláció előfordulása az azt tartalmazó nagyobb populációhoz tartozó mintavételi kereten belül, annál nagyobb problémába ütközik a belőle történő reprezentatív, valószínűségi mintavétel. Ez utóbbi jellemzőit jelen tanulmányban úgy határozzuk meg (eltekintve a téma kapcsán létező viták ismertetésétől), hogy az adott populáció (csaknem) minden tagjának 0-nál nagyobb, meghatározható mértékű esélye van a mintába kerülésre, illetve a mintából mind az adott populáció ritkaságára, mind a ritka populációt jellemző belső arányokra statisztikai becslést lehet készíteni.

Általánosan elfogadott, hogy szélsőségesen ritka populáció (például 1/1000-es előfordulási gyakoriság) esetén az ilyen jellegű mintavételről lemondunk, és más módszerekhez (például hólabdatípusú vagy szakértői mintavételhez) folyamodunk. Az elmúlt évtizedekben azonban folyamatosan megfigyelhető módszertani törekvés arra, hogy minél alacsonyabb sűrűségű populáció esetén váljon lehetővé egy, az előző kritériumoknak megfelelő reprezentatív, valószínűségi mintavétel.

A következőkben először ez utóbbi esetben alkalmazható módszereket fogjuk bemutatni. Amellett érvelünk, hogy amennyiben egy ritka populáció sűrűsége eléri a 3–5 százalékot, megoldható feladat (igaz nehézségek árán) a reprezentatív mintavétel, még ha a ritka populáció eloszlása kis mértékben el is tér az egyenletestől. Az ismertetésben az egyszerűbb módszerektől haladunk az összetettebbek felé: 1. mintanagyság növelése; 2. minta aszimmetrikus rétegzése (disproportionate stratification); 3. minta szűrése (screening); 4. közbenső mintavételi egység szintjén történő szűrés; 5. többszörös/dupla mintavételi keret (multiple/dual frame methods);¹ 6. Kapcsolathasznosító módszerek (linkage exploitation methods): 6.1. hálózati mintavétel (network sampling); 6.2. adaptív csoportos mintavétel (adaptive cluster sampling); 7. Rejtőzködő populációk (hidden populations) esetén alkalmazható módszerek: 7.1. térben és időben

¹ Az angol kifejezések egy részének nincs még bevett magyar megfelelője. Az általunk javasoltak a magyar módszertani szakirodalomhoz illeszkednek, és nem feltétlenül az angol eredeti megnevezések szolgai fordításai.

meghatározott mintavétel (time-space sampling); 7.2. válaszadó által vezérelt mintavétel (respondent driven sampling).

A különféle módszerek egymáshoz kapcsolódnak, egymást kiegészíthetik. Természetesen más jellegű csoportosítás, beosztás is lehetséges, ezen a téren a szakirodalom sem egységes. *Kalton* [2001] például 11 módszert különböztet meg, köztük azonban van nem reprezentatív jellegű is (hólabdatípusú mintavétel).

Tanulmányunk elsősorban a mintavételi eljárások ismertetésére összpontosító szakirodalmi áttekintés, tehát csak érintőlegesen foglalkozunk az egyes minták alkalmazása után, az adatbázis elemzésekor felmerülő becslési problémákkal.

Amikor arra lehetőség nyílik, a szakirodalmi példákon túl két visszatérő témán keresztül mutatjuk be e módszerek gyakorlati alkalmazásait. Ezek közül az egyik egy valóságos, az erdélyi magyar populációra készült reprezentatív kutatás („Életünk Fordulópontjai – Erdély” vizsgálat), ami a közelmúltban valósult meg (az alkalmazott mintavételi eljárásról részletesen lásd *Kiss–Kapitány* [2009]).²

A másik elméleti jellegű: egy olyan nem megvalósult vizsgálat példája, amely azt szeretné felmérni, hogy hányan, kik és milyen körülmények között végeznek ma hivatalosan, nem hivatalosan vagy félhivatalosan mezőgazdasági bérmunkát Magyarországon.

1. A mintanagyság növelése

Legegyszerűbb lehetőség a ritka populációk vizsgálatára a teljes mintanagyság növelése. Például az Egyesült Államokban, ha elemezhető (legalább 300 fős) fekete és spanyolajkú válaszadót (jelenleg hozzávetőleg 13, illetve 15 százalékos arányuk a felnőtt populációban) szeretnének a mintában a közvélemény-kutató cégek, egyszerűen 1 000 helyett 3 000 fős mintán kérdezik le a kérdőívet. Ha a mintavétel nincs is származás szerint kontrollálva, a 300 feletti elemszámokat a módszer nagy bizonyossággal garantálja.

Ritkább populáció és nagyobb esetszámigény esetén azonban ez az út nehezen járható: egy 2 500-as elemszámú mintához 5 százalékos sűrűség esetén akár 50 000 fős mintát kellene venni, ami 50 százalékos válaszadási hajlandósággal számolva 100 000 mintavételi egység (személy, cég stb.) megkeresését jelentené.

² Az adatfelvételre, melyben a KSH Népeségtudományi Kutatóintézete (Budapest), a Nemzeti Kisebbségkutató Intézet (Kolozsvár) és a Max Weber Alapítvány (Kolozsvár) vett részt, 2007-ben került sor Erdély magyarul beszélő 20–45 éves lakossága körében. A mintanagyság 2 500 fő volt. A kutatás főbb eredményei magyarul elérhetők: *Spéder* [2009].

2. Aszimmetrikus mintarétegzés

Az aszimmetrikus mintarétegzés akkor használható, ha a vizsgált ritka populáció egyes, a mintavétel során rétegzésként használható jellemzők (például nem, kor, lakóhely) szerint ismert, de nem egyenletesen oszlik el. A mintavételnél a hasznos elemszám növelése érdekében azokat a rétegeket, ahol a ritka populáció nagyobb sűrűségét feltételezzük, felülreprezentáljuk. Ugyanakkor a feltehetően kisebb sűrűséggel rendelkező rétegeknek is hagyunk egy alacsonyabb, de ismert bekerülési valószínűséget. A terepmunka után súlyozással helyreállítjuk a helyes arányt, de ez a ritka populáció válaszadóinak a valós hasznos elemszámát már nem módosítja. (Márpedig ez meghatározó a mintavételi hiba, vagyis az adatok megbízhatósága szempontjából.)

Nézzünk példát egy ilyen jellegű mintavételre az erdélyi magyarok esetén. Első lépésként veszünk egy 3 000 fős reprezentatív mintát a tágran vett erdélyi megyék körében, megyénként rétegezve. Ebbe a mintába előreláthatólag kb. 600 (20%)³ magyar anyanyelvű kerülne. Anyanyelvet nem figyelembe véve a 3 000 főből mintegy 500 válaszadó élne Hargita–Kovászna–Maros megyékben, utóbbiak közül hozzávetőleg 300 magyar anyanyelvű.

Eközben egy kiegészítő mintában még 1 500 válaszadót választunk ki e három megyéből, ezzel még háromszorosan felülreprezentálva azokat népességarányukhoz viszonyítva. Ezek közül hozzávetőleg 900 fő magyar anyanyelvű. Így összesen a 4 500 válaszadóból 1 500 lesz magyar. Ezen 1 500 ember között persze arányon felül vannak a székely megyékben élők (80 százalékos arányban a valós 50 százalék helyett), melyet utólag súlyozással helyre lehet állítani. A magyarokra vonatkozó adatok mintavételi hibája természetesen nagyobb lesz, mintha „normális” 1 500 fős mintavétel történt volna, de még mindig kezelhető és alacsonyabb, mint egy 600 fős véletlen minta esetén. Persze ilyenkor az elemzés során nagyon óvatosan kell kezelni az elemszámokat területi megoszlásokkal összefüggő tényezők esetén (a módszerről lásd például *Kalton* [2001]).

Természetesen ennek a módszernek is erőteljes korlátai vannak. Alkalmazásához egyfelől egyértelmű sűrűsödési pontokra (mint az erdélyi magyarok esetén a Székelyföld) van szükség, másfelől elengedhetetlen, hogy a vizsgálni szándékozott ritka populáció összességének minél nagyobb aránya legyen megtalálható ezeken a ritka populáció által sűrűn lakott területeken (*Waksberg–Judkins–Massey* [1997]). Ha például magyarországi romákra szeretnénk mintát venni, ahol a megyénkénti eloszlás nem ennyire aszimmetrikus, e módszerrel is legalább 10 000 fős minta kellene 1 000 roma válaszadóhoz.

³ Az áttekinthetőség miatt kerekített számok.

3. A minta szűrése („szkríning”)

A következő megoldás az előzetes szűrés (screening, szkríning, szkríningelés). Ennek lényege, hogy nagyméretű mintát választunk ki egészen a mintavételi egységek (személyek, cégek stb.) szintjéig. Az utóbbiakkal (személyes, telefonos stb.) kapcsolatfelvételre is sor kerül, de a teljes adatgyűjtés csak ott történik meg, ahol a potenciális adatszolgáltató a keresett ritka populációba tartozik. A többi esetben csak egy rövid regisztrációs adatlapot töltünk ki. Ezzel a módszerrel a terepmunka költsége jelentősen (de a tapasztalatok szerint 50 százalékot csak ritkán meghaladó mértékben) csökkenthető, hiszen egy rövid szűrőkérdőív kitöltésének költsége lényegesen alacsonyabb a teljes kérdőívénél.

Nézzünk példát egy kombinált szkríningelésre azzal számolva, hogy egy szűrőkérdőív kitöltésének költsége 50 százaléka egy teljes kérdőív lekérdezésének. Vegyünk egy 20 000 fős, a munkavállalási korúakra reprezentatív mintát, azt feltételezve, hogy lesz benne 1 000 mezőgazdasági munkavállalásban érintett személy (5%). 2 000 személy – a 20 000 fő véletlen almintája – esetén azt az utasítást adjuk a kérdezőknek, hogy a válaszadótól a kérdőív adatait mindenképpen kérdezzék le, míg a többi esetben csak akkor, ha az illető a szűrőkérdőív alapján valóban érintett a mezőgazdasági munkavállalásban. Ezzel összesen 2 900 lekérdezett (1 900 nem érintett és 1 000 érintett) kérdőívet, valamint 17 100 szűrt, de nem kérdezett címet kapunk, utóbbit fél költségen. Így az adatfelvételi költség 20 000-ról 11 450 egységre csökkenthető. Ez a 43 százalékos megtakarítás ugyanakkor bizonyos szempontból látszólagos, hiszen összesen 2 900 kitöltött kérdőívért fizettük ki 11 450 árát.

A szűrésre jól bevett módszer más célú omnibusz adatfelvételek „szűrőkérdőívként” való használata. Havi 1 000 fős omnibusz adatfelvételekkel számolva egy év alatt 12 000 fő szűrése végezhető el. Ez 10 százalékos sűrűség esetén még az újbóli megkeresés által jelentett lemorzsolódással számolva is 1 000 fős mintát eredményezhet.

Ladányi és Szelényi [2001] a kelet-európai romákról szóló vizsgálatukban szintén ezt a módszert használták. Országoként eltérő számban, de mintegy 10–20 000 fős omnibuszos kutatás keretében készült interjúból „szűrték ki” azokat a roma válaszadókat, akikkel később a személyes beszélgetést lefolytatták.

A módszer hátulütője a korábbiakban bemutatottakkal szemben az, hogy nem eredményez a teljes alapsokaságra (tehát nem a ritka populációra) vonatkozó kontrolladatokat (hiszen a „nem kiszűrtekről” nem gyűjt információt). Márpedig ezek a kontrolladatok ritka populációkra történő mintavétel esetén minőségbiztosítási, validálási okokból lennének fontosak.⁴ Ezért szokták azt a megoldást alkalmazni,

⁴ Ha a teljes populációra vonatkozó adatok jelentősen eltérnek a korábbi felvételek hasonló értékeitől, valószínű, hogy a ritka populációhoz kapcsolódó adatokkal is baj van. Ha nincs információ a teljes populációra, ezt a kontrollt nem lehet elvégezni.

hogy a nagy minta egy kis részét (önmagában is reprezentatív almintáját) mindenképpen lekérdezik, többségét azonban csak szkríningelve. Az előzőekben említett omnibuszos jellegű szűrés erre a problémára is megoldást jelent.

A szűrés, mint módszer sok esetben kombinálható az aszimmetrikus rétegzéssel. Nézzük erre a korábbi erdélyi magyar példát. Először bontsuk szét a 3 000 fős „teljes erdélyi” mintát: 1 000 főt kérdezzünk le anyanyelvtől függetlenül, a másik 2 000-nél végezzünk előzetes szűrést. (Utóbbiak közül így 1 600 nem magyar nyelvű válaszadót csak szűrünk, s 400 magyart kérdezzünk.) A székely megyék kiegészítő 1 500 fős mintája esetén ismét előzetes szűrést végzünk nyelv szerint (600 szűrés; 900 magyar teljes kérdezés). Így összességében – 2 200 fő kiszűrése után – kapunk 800 román és 1 500 magyar válaszadót. A szűréssel járó megtakarítás – 50 százalékos szűrés költséggel számolva – jelentős, hiszen egy 3 400 fős minta költségéből megtörtént a 4 500 embert lekérdező adatfelvétel.

Ha rendelkezésünkre áll olcsó szűrési módszer (például telefon), elvileg akár 1 százalékos sűrűségű populáció is elérhető ilyen kombinált mintavétellel. A 2001/2002. évi amerikai zsidó adatfelvétel (National Jewish Population Survey) esetén például zsidók által is lakott területeket felülreprezentálva több mint 170 000 háztartás telefonos szűrését végezték el a mintegy 5 000 fős minta kialakítása érdekében. (Eközben egy 4 000 fős nem zsidó kontrollmintát is felvettek.)

Komoly veszélyt jelent ugyanakkor maga a szűrési eljárás, amely újabb hibalehetőségeket hordoz magában. Kisebb baj, ha nem a keresett ritka populáció tagjai válaszolnak, hiszen az általuk szolgáltatott adatok utólag törölhetők. Nagyobb problémával jár azonban ennek a fordítottja, amikor a keresett ritka populáció tagjai valamilyen okból nem jutnak túl a szűrőkérdőíven. Az utóbbi téves besorolás tömegessége akár a teljes kutatást is ellehetetlenítheti, mivel emiatt alulbecsülhetjük a keresett ritka populáció sűrűségét.

4. Közbenső mintavételi egység szinten történő szűrés

Az egyszerű szűrésnél módszertanilag vitathatóbb, ugyanakkor költség-hatékonyabb – és így kisebb sűrűségű populáció elérését teszi lehetővé –, ha a szűrés már a többlépcsős mintavétel valamelyik közbenső szintjén (is) megtörténik („screening with area sampling”). Ez azt jelenti, hogy a többlépcsős mintavétel során olyan mintavételi egységekben, ahol az adott ritka populációnak nincsenek, vagy alig vannak tagjai, csak elméletileg kerül sor mintakijelölésre, a gyakorlatban nem történik meg a válaszadók felkeresése. Ez a módszer értelemszerűen csak abban az esetben alkalmazható, ha a kutatni szándékozott ritka populáció eloszlása a közbenső

mintavételi szinten nem egyenletes. Így feltehető, hogy az adott népesség egyáltalán nincs jelen a közbenső mintavételi egységek (jellemzően települések) jelentős részénél.

Személyi szintű előzetes szűréssel kombinálva mi is ilyen módszert alkalmaztunk az „Életünk Fordulópontjai – Erdély” vizsgálat során az erdélyi magyar nyelvű populációra történő mintavételkor.

Első lépésben az erdélyi megyékre vonatkozóan kialakítottunk egy úgynevezett elméleti kontaktmintát. Itt – más vizsgálatokhoz hasonlóan – úgy jelöltük ki az (elméleti) kutatási pontokat, hogy az önreprezentáló települések mellett régió- és településméret szerinti rétegeket is létrehoztunk. A 10 000 fő alatti, nem önreprezentáló települések esetén az egyes települési rétegeken belül a reprezentativitáshoz szükséges elemszámot egyenletesen osztottuk el arra törekedve, hogy egy településen legalább 50 címet kijelöljünk. A települések meghatározásában nem volt szerepe azok nemzetiségi (vagy anyanyelvi) megoszlásának.

Ezután az elméleti kontaktminta és a tényleges terepmunka mintája közé egy köztes, településszintű szűrést ékeltünk. Azokra a településekre, ahol népszámlálási adatok alapján minimális volt az esély sikeres magyar nyelvű interjúra, nem küldtünk kérdezőbiztosot.⁵ A felkeresendő személyek száma így mintegy 45 százalékkal csökkent az elméleti kontaktmintához képest.

A kiválasztott mintavételi pontokon belül kijelöltük a reprezentativitáshoz szükséges interjúalanyok számát, magukat a megkérdezetteket pedig a települési névjegyzékből választottuk ki. Minden elérhető válaszadót felkerestünk, és a kérdezőbiztos közülük azokkal készített interjút, akik a szűrőkérdések alapján képesek voltak magyar nyelven válaszolni.

Az ilyen közbenső szintű szűrés esetén kritikus pont, vajon honnan és mennyire megbízhatóan tudjuk megállapítani, hogy az adott közbenső mintavételi egységben valóban nincs-e tagja a keresett ritka populációnak. Hiszen sok esetben erről nem vagy csak nagyon korlátozottan állnak külső információk a rendelkezésünkre.⁶ Ennek a speciális, de nem ritka problémának a megoldását *Sudman* formalizálta először 1972. évi cikkében. Eszerint első lépésben a közbenső mintavételi szint minden tagjából kiválasztunk véletlenszerűen egy-egy elemet. Ha a kiválasztott válaszadó tagja a keresett ritka populációnak, folytatjuk a szűréssel kombinált lekérdezést az adott klaszterben. Ellenkező esetben a klaszter többi válaszadóját már nem keressük fel.

A módszer napjainkra lényegesen finomodott: a teljes minta kiválasztását követően kijelölünk egy újabb almintát oly módon, hogy az lehetőleg egyenlő számban tar-

⁵ A határt az jelentette, hogy az adott településen a magyarok aránya a 2002. évi népszámláláson elérte-e a 8 százalékot. Ha nem, úgy vettük, mintha az adott településen egyetlen sikeres kérdés sem valósult volna meg.

⁶ Az előbb említett erdélyi példa nem ilyen volt, hiszen itt a településsoros népszámlálási adatok erre a célra elégséges információt adtak.

talmazzon elemeket minden egyes közbenső mintavételi egységből. A terepmunka első lépéseként ennek az almintának minden elemét felkeressük. Ezután azonban csak azokban a közbenső mintavételi egységekben folytatjuk a teljes mintán a lekérdezést, ahol az első almintá eredményes volt, vagyis találtunk a ritka populációhoz tartozó adatszolgáltatót.

További újítási lehetőség, hogy az első almintá esetén nem alkalmazunk szűrést, hanem mindenkit lekérdezzük. Ilyenkor kiküszöböljük a módszernek azt a hátlütő-jét, hogy az nem eredményez a teljes populációra vonatkozó kontrolladatokat, hiszen a teljes kutatás eredményeképpen nemcsak a ritka populációt reprezentáló mintát kapjuk meg, hanem egy teljes populációra vonatkozó kontrollmintát is. (Ennek jelentőségéről már a szűrés kapcsán volt szó.) Ez esetben azonban az almintavételnek önmagában is reprezentatívnak kell lennie, ami sok esetben feloldhatatlan ellentmondásban van azzal az elvárással, hogy az almintának egyenlő számban kell tartalmaznia elemeket minden egyes közbenső mintavételi egységről.

5. Többszörös/dupla mintavételi keret alkalmazása

A többszörös/dupla mintavételi keret elnevezésű módszerek (multiple/dual frame methods) egy speciális mintavételi eljárási rendet képviselnek. Ennek lényege, hogy a mintavételkor több mintavételi keret együttes alkalmazására kerül sor. A már az 1950-es évek óta ismert, de tömegesen csak az 1980-as évek óta alkalmazott módszernek a gyakorlati megvalósítást tekintve rengeteg alfaja és ezekhez kapcsolódóan kiterjedt szakirodalma van (például *Skinner–Holmes–Holt* [1994], *Lohr–Rao* [2006]).

Mi a következőkben e módszernek először azzal a speciális esetével foglalkozunk, melyet jellemzően a ritka populációkból történő mintavételek során alkalmaznak. Eszerint a mintavételi keretek között van olyan, amely a teljes ritka populációt magába foglalja („A” keret), de drágán használható, és egy vagy több másik („B” keret), amely nem teljes körű, de a keresett ritka populáció tagjait nagy sűrűségben tartalmazza. (Például „B” keretnek nevezhetők az amerikai indián törzsek törzsi névjegyzékai. Értelemszerűen ezekben nem minden indián szerepel, de akik igen, azok valóban indiánok és könnyen elérhetők.) Tehát ebben az esetben elvileg lehetőség nyílna a korábban felsorolt módszerek alkalmazására – hiszen van egy, a teljes ritka populációt magába foglaló keret –, de a ritka populáció eloszlása mintavételi szempontból mégis olyan kedvezőtlen, hogy azok alkalmazása irreális költségekkel járna.

Az első módszertani kérdés az, hogy miképp lehet kombinálni ezt a két mintavételi keretet. A legegyszerűbb eljárás erre, hogy veszünk két független mintát. Először

az „A” keretből egy olyan nagyot, ami használható elemszámú (mondjuk 500 főnyi) válaszadót eredményez a keresett populációból. Ezután veszünk egy másikat (mondjuk 2 000 főt) a „B” keretből, amelyet azután hozzáülozünk a másik keretből kapott ritka populáció eloszlásához.⁷ Ezzel a kapott mintanagyságot sikerült meglehetősen alacsony költségen megötszöröznünk. Természetesen ezután kritikus pont az eredmények pontosságának becslése (pont- és szórásbecslések) (Lohr [2007]).

Az előbb ismertetettnél lényegesen komolyabb mintavételi probléma, ha nem áll rendelkezésünkre a teljes ritka populációt magában foglaló mintavételi keret, hanem az egyes mintavételi keretek egymást átfedve léteznek, összességében lefedve a keresett ritka populáció minden tagját. Például, amennyiben a Magyarországon mezőgazdasági jellegű munkát végzőket szeretnénk tanulmányozni, szembe kell néznünk azzal, hogy a vizsgálandó populáció jelentős része nem magyar állampolgár, és magyarországi lakhellyel sem rendelkezik. Tehát a Közigazgatási és Elektronikus Közszolgáltatások Központi Hivatala (KEK KH) népszámszámításából vett esetleges minta („A” keret) nem fedi le a teljes keresett populációt. Rendelkezésünkre áll egy „B” mintavételi keret is, mivel a külföldi mezőgazdasági munkavállalók (az ellenőrzésekre számítva még a feketén munkát vállalni szándékozók is) kiváltják az alkalmi munkavállalói kiskönyvet. Ez utóbbi a vizsgált populáció szempontjából lényegesen „sűrűbb”, ugyanakkor kisebb, és nem részalmazza az „A” keretnek. Az „A” és a „B” együttesen viszont teljes lefedést biztosít.

6. Kapcsolathasznosító módszerek

A kapcsolathasznosító módszerek (linkage exploitation methods) abban az esetben alkalmazhatók, ha a keresett ritka populáció tagjai (fel)ismerik egymást, és vannak köztük kapcsolatok (Kaslsbeek 2000). (Így ez nem használható például allergiások közüli mintavételnél.)

E csoportba tartozó módszerek lényege, hogy kiválasztunk egy kiinduló (reprezentatív) mintát a keresett ritka népességből, majd ehhez kapcsolódva választunk, gyűjtünk újabb mintaelemeket. Így a mintanagyság olcsón és könnyen az eredeti kétszeresére vagy akár többszörösére növelhető.

A reprezentativitást legalább elvi szinten biztosító kapcsolathasznosító eljárások több fontos ponton megkülönböztethetők az első pillanatban hasonló „hólabdamódszertől”: esetükben 1. a kiinduló mintának reprezentatívnek kell lennie; 2. az ismerő-

⁷ Vagy mások szerint inkább az „A” mintakeret azon alpopulációjához, amely tagja a „B”-nek is. Ehhez természetesen az „A”-beli adatgyűjtéskor meg kell tudni, hogy ki tagja szintén a „B” keretnek.

söktől már nem lépünk tovább az ismerősök ismerőseire, mivel minden további lépésnél hatványozódna az esetleges torzítás mértéke; 3. a megadható kapcsolatok típusa és száma pontosan szabályozott.

A következőkben a kapcsolathasznosító mintavételi típusok két formájával, a hálózati (network sampling) és az adaptív mintavétellel (adaptive cluster sampling) foglalkozunk.

6.1. Hálózati mintavétel

A hálózati mintavétel, melynek névadója, „prófétája” *Monroe G. Sirken* (*Sirken* [1970], *Shimizu–Sirken* [2006], történeti áttekintést lásd *Sirken* [1998]), a nevével ellentétben nem mintavételi módszer, hanem csupán egy olyan eljárás, amely több esemény egy megfigyelési egységhez való társítását teszi lehetővé ritka populációk esetén. Erre példa speciális genetikai rendellenességgel élők keresése egy háztartási mintában. Nyilván sűrűn előfordul, hogy egy háztartásban több személy is szenved ilyen betegségben. „Normális” háztartási adatgyűjtés esetén egy család egy esetnek számít. A hálózati mintavétel módszere azonban megengedi, hogy egyrészt egy háztartás annyszor számítson, ahányszor a rendellenesség benne előfordul, másrészt a válaszadó a háztartástagok külön élő testvéreinek genetikai rendellenességeiről is adjon információt.

Mindezekből következik az a továbblépés, hogy minden társított eseményhez külön adatgyűjtést, kérdőívet társítsunk.

A módszer kapcsán felmerülő két legkritikusabb kérdés a következő: 1. hogyan jelöljük ki az „ismerősöket” (counting rules); 2. utólag milyen módon súlyozzuk az adatokat (nyilván a kevés kapcsolatot megadókat „értékesebbek”), és becsüljük a mintavételi hibát.

Az utóbbi problémával – hisz cikkünk elsősorban a mintavételi technikákkal foglalkozik – részletesen nem foglalkozunk. Azt azonban lényeges megjegyezni, hogy itt – egyedülként az ismertett mintavételi eljárások közül – valójában előbb volt ismeretes a becslési probléma, és utóbb született a mintavételi eljárás. Sirken ugyanis eredetileg épp azzal a kérdéssel foglalkozott, hogy milyen veszélyt jelent az adatok megbízhatóságára a többszörös kiválasztás (multiplicitás), amely nehezé teszi a hálózati mintákra vonatkozó becslést. Így az alkalmazott becslési eljárások (multiplicity estimator; weighted multiplicity estimator stb.) már rögtön a módszer kialakulásakor rendelkezésre álltak, és az alkalmazási standard részévé váltak.⁸

⁸ Nem merül tehát fel senkiben, hogy egy hálózati mintavétellel kapott adatbázisra éppoly módszerrel végezhetőek becslések, mint egy hagyományos véletlen minta eredményeire. Más ismertett módszerekről ez sajnos nem mondható el.

Az ismerősök kijelölésekor szigorúan és jól meghatározott kapcsolatok alkalmazására kerül sor. A mintákat általában a testvérekre, leszármazottakra, nagybácsikra, nagynénikre, közvetlen szomszédokra bővítjük ki. Módszertani minimum, hogy a kijelölt ismerősök száma minden válaszadó esetén ismert legyen, és az eredeti megkérdezettek tudják, ismerőseik vajon szintén a keresett ritka populáció tagjai közé tartoznak-e. Természetesen csak azon ismerősök bevonására kerül sor, akik szintén tagjai a keresett populációnak.

A módszer gyakorlati haszna nem elhanyagolható, egy szórványhelyzetű etnikai közösség esetében például két-háromszorosra lehetne növelni az elemszámot a mintába véletlenszerűen bekerült populációtagnak (életben lévő) testvéreinek (a féltestvérek közül csak a kérdezettel egyneműek) felkeresésével.

6.2. Adaptív mintavétel

Az 1990-es évek közepétől egy új módszer került előtérbe, a *Steven K. Thompson* nevéhez köthető (rétegzett/inverz) adaptív mintavétel ((stratified/invers) adaptive cluster sampling) (*Thompson* [1990], *Thompson–Seber* [1996]). (Az elnevezésben szereplő adaptív jelző arra utal, hogy a mintavételi eljárás „adaptálódik”, vagyis alkalmazkodik az adatgyűjtés folyamán talált adatokhoz.) Ezen eljárás kiindulópontja más volt, de a ritka populációk mintavételi kérdései felől közelítve lényegében hasonló eredményre jutott, mint a hálózati mintavétel. Eredetileg biostatistikai célból fejlesztették ki (ritka állatfajok, például bálnák stb. vizsgálatára). (Többek között *Philippi* [2005]). A módszer – mivel az állatok és a növények nem tudnak beszélni, és nem képesek arra, hogy megjelöljék testvéreiket, legjobb barátjukat stb. – a kiinduló mintaelemek közelében automatizált algoritmus szerint keres újabb mintaelemeket, teljességgel megszüntetve az elsődleges mintaelemek ebben játszott – akár mennyire is minimális, de – szubjektív szerepét.⁹

A módszer lényege, hogy első lépésben egy hagyományos, előre meghatározott elemszámú mintavételre kerül sor. Ezután azon elemek tekintetében, amelyek tagjai a ritka populációnak, megtörténik a „körülvevő” esetek adatfelvétele is. Ez utóbbiak közül a ritka populáció tagjainál (míg van ilyen) ismét mintába vonjuk a környező eseteket. Az ilyen mintába utólag bevont, de nem a ritka populáció részét képező elemeket peremelemeknek („edge units”) nevezik.

A végső minta végül négyfajta elemet tartalmaz: a kiinduló minta ritka populációhoz tartozó és nem tartozó elemeit, az utólag bevont, ritka populációból származó mintaelemeket, illetve a peremelemeket.

⁹ A hálózati mintavétel esetén például arra gondolhatunk, hogy a válaszadó „elfelejt” beszámolni a család „fekete bárányának” szerepét betöltő testvéréről stb.

A különböző becslések és számítások elvégzésekor más-más eseteket vonunk az elemzésbe. A ritka populáció elterjedtségére vonatkozó készítésekor értelemszerűen a kiinduló minta tagjaiból indulunk ki. A ritka populáció belső arányaihoz kapcsolódó becslésnél a kiinduló minta ritka populációhoz tartozó elemei mellett az utólag bevont mintaelemeket is figyelembe vesszük. Az utóbbi esetben azonban az adatok előzetes belső átsúlyozására van szükség, hiszen a ritka populáció nagyobb sűrűsödési pontjainak mintába kerülésére nagyobb az esély, mint a szórványos elhelyezkedésűekéinek.¹⁰

A módszer egyszerű, érthető, frappáns, de sajnos csak nem életszerű mintavételi problémák esetén alkalmazható. A korrekt becslések előfeltételei ugyanis a következők: kölcsönösség (ha az A elem szomszédja B-nek, akkor a B elem is legyen szomszédja A-nak); egyenlő „szomszédszám” minden tag esetén; és a „szomszédos” elem egyértelmű definiálhatósága. A módszert így általában földrajzi alapú mintavétellel kombinálják, mivel a valós életben ritkák a szabályos, stabil mértani formákba rendezett válaszadók.

Tegyük fel, hogy egy bejelentésre nem kötelezett mezőgazdasági kultúra (például egy ritka fajtája) elterjedését és állapotát kívánjuk vizsgálni. Ekkor első lépésben szabályos (például 100×100 méteres) négyzetekre osztjuk a vizsgált területet. Majd ezek közül veszünk egy véletlen mintát előre meghatározva azt az egyedszámot vagy -sűrűséget, amelytől kezdve az adott területet „érintettnek” tekintjük. Végül e területi alapon lefolytatjuk az előzőekben leírt eljárást.

A mezőgazdasági munkavállalókra vonatkozó példa esetén a munkavégzés helye alapján történhetne a megközelítés. A véletlenszerűen kiválasztott munkavállalókat az adatgyűjtő elkíséri a megadott napon a munkavégzés helyszínére, és az ott lévő többi munkavállaló közül – egy előre megadott algoritmus alapján – bővíti a mintát.

7. Rejtőzködő populációk esetén alkalmazható módszerek

A ritka populációk egy speciális alcsoportját képezik a rejtőzködő populációk (hidden populations). Ezek tagjai nemcsak alacsony sűrűségben lelhetők fel, hanem még arra is hajlamosak, hogy a hagyományos megkeresési módok (például előzetes telefonos szűrés, kérdezőbiztosi felkeresés a kérdezett otthonában) esetén ne vállalják fel csoporttagságukat. A rejtőzködő populációkra tipikus példák: intravénás

¹⁰ Merész hasonlat, de a közkedvelt torpedójátékban is az őrnaszádokhoz képest nagyobb az esély a több tagból álló anyahajók megtalálására. (<http://www.logikaifeladatok.hu/torpedo/torpedo.html> Elérés dátuma: 2010. május 26.)

droghasználók, örömlányok, homoszexuálisok csoportjai stb. Jelen tanulmányunkban két, ilyen populációkat elérő speciális módszert mutatunk be röviden.

7.1. Térben és időben meghatározott mintavétel

A térben és időben meghatározott mintavétel (time-space sampling)¹¹ arra alapoz, hogy a rejtőzködő ritka populációk tagjai is elérhetők bizonyos helyeken (klubokban, internetes chatszobákban, speciális nyilvános tereken stb.) és helyzetekben (csoportspecifikus felvonulásokon, ünnepeken stb.), amikor könnyebben felvállalják csoporttagságukat (*Stueve et al.* [2001], *Mansergh et al.* [2006], *Parsons–Grov–Kelly* [2008]). Az eljárás lényege, hogy mintavételi keretként ezek a helyszínek, események, illetve a helyszíneken belül az idő (nap és óra) számít. Tehát a mintavétel során az előzők közül véletlenszerűen kiválasztott helyszíneken és időpontokban az ott és akkor megjelenő összes személy közül szűrjük ki a célcsoporthoz tartozókat, s veszünk közülük mintát. Az eljárás így összességében három lépcsőből áll: először a helyszínek, majd a megfigyelési időszakok, végül a személyek közül választunk. Szokásos a mintavételi eljárás közbeni rétegzés és az aszimmetrikus felülreprezentálás is. (Például a helyszínek típusai közül rétegzünk, s felülreprezentáljuk azokat, ahol alacsony sűrűséget várunk (low yield venues).)

Statisztikai értelemben nyilvánvalóan sok probléma van a módszerrel. A legkomolyabb, hogy a minta tervezésekor nem lehet feltérképezni és felkeresni minden helyszínt, hiszen vannak kevésbé vagy egyáltalán nem nyilvánosak. A keresett ritka populáció bizonyos tagjai ezeken kívül máshol szinte soha nem tűnnek fel. Ezért a módszerrel elért populáció bizonyos szempontokból különbözhet a teljes rejtőzködő populáció jellemzőitől, ami – az egyébként szükséges – utólagos súlyozással nem kiszűrhető torzításokhoz vezethet. Emiatt a kutatók inkább a vizsgált populációt szűkítik le az ily módon elérhető sokaságokra. Ez okból készül például az „utcai szexmunkásokról” lényegesen több, empirikus adatgyűjtésen alapuló tanulmány az általában vett örömlányokhoz képest.

A hagyományos keresztmetszeti kutatásokkal való összevetések (többek között *Xia et al.* [2006]) arra utalnak, hogy nem árt fenntartással kezelni e mintavételi forma eredményeit. Vannak azonban olyan esetek, amikor a keresett populáció tagjai elkerülhetetlenül megjelennek bizonyos helyszíneken, és ez a szinte egyedüliként használható módszer. Így például, ha illegálisan foglalkoztatott, nem magyar állampolgárságú mezőgazdasági kampánymunkásokat vizsgálunk, nyugodtan állíthatjuk, hogy az alapsokaság tagjainak a meghatározásnál fogva meg kell jelenniük munkavégzésük helyén. Ezért ezek a helyszínek bizonyos időszakokban (szőlőmetszés,

¹¹ Használatos még a venue-time-space sampling, illetve a time-location sampling kifejezés is.

dinnyeszüret stb.) alkalmasak mintavételi egységnek. Vannak ezek mellett olyan helyek is, ahol ugyan a populáció nem minden tagja jelenik meg, de közöttük nagy a keresett populáció sűrűsége (például a mezőgazdasági települések „emberpiacai”, a munkaügyi központok, ahol a külföldi munkavállalók szezonális alkalmi munkavállalói kiskönyveinek¹² kiváltása folyik stb.).

7.2. Válaszadó-vezérelt mintavétel

A rejtett populációk esetén az elmúlt évtizedben gyakran alkalmazott másik módszer a válaszadó-vezérelt (respondent driven) mintavétel (*Heckathorn* [1997], [2002]). Ez gyakorlatilag olyan minta, ahol a kiindulópontok a rejtőzködő populáció szakértő elemei, akik saját maguk kutatásban való részvételre felhívó „vócsereket” (részvételi jegyeket) osztanak ki a populáció általuk ismert más tagjai között. Ez utóbbiak, ha hajlandók válaszolni, szintén kapnak ilyen vócsereket, melyeket szétoszthatnak, és így tovább, amíg csak el nem jutunk a kívánt mintanagysáig.

A vócserek kiosztási rendszerét és információtartalmát olyan módon kell megoldani, hogy ezek segítségével „visszafejthetők” legyenek az ajánlási hálózatok, illetve információt gyűjthessünk arról, hogy ki-ki hány másik tagot ajánlott (s közülük hány jelentkezett). Így a kutatás eredményeként kapott adatokat elsősorban nem személyi szintű, hanem hálózati jellegű adatbázisként kell felfogni, amely a hálózat jellemzőiről bír hasznos információkkal. Ezek már elégséges alapot nyújtanak a válaszadóktól kapott személyi szintű adatok olyan átsúlyozásához, hogy azokból valódi – a valószínűségi mintavételhez hasonlítható – becsléseket kapjunk. Ez nyilván felértékeli a kis zárt, és lebecsüli a nagy hálózatok tagjait (a levezetést és a becsléseket lásd *Salganik–Heckathorn* [2004]).

Kérdés azonban, hogy mi garantálja az elvi esélyt a populáció minden tagjának a mintába kerülésre. A módszer hívei azzal érvelnek, hogy a hálózati kutatások eredményei szerint szinte minden ember negyed- vagy ötödfokú ismerőse mindenkinek, tehát ha megfelelően hosszú ajánlási láncok működnek a válaszadók összegyűjtése során, akkor az az összes személynek esélyt nyújt a bekerülésre. Ez a logika azonban egyértelműen téves, hiszen e mintavétel esetén az „ismerős ismerősét” csak akkor tudjuk elérni, ha maga az „ismerős” is eleme a keresett ritka populációnak.

Ez tehát gyakorlatilag a hólabdatípusú módszer speciális, továbbfejlesztett és utólag súlyozott alfajának tekinthető, így megítélésünk szerint a meggyőző részeredmények ellenére sem valószínűségi mintavételi eljárás. Az alkalmazott súlyozási mód-

¹² Az alkalmi munkavállalói kiskönyvek e fajtája olyan speciális „intézmény” (volt), amelyet szinte kizárólag az illegális munkát vállalni szándékozók váltottak ki tevékenységük látszólagos lefedésére.

szerek pedig – az előző állítással szemben – nem képesek kompenzálni a mintavétel alapvető hiányosságait.¹³

Irodalom

- CHRISTMAN, M. C. – LAN, F. [2001]: Inverse Adaptive Cluster Sampling. *Biometrics*. 57. évf. 4. sz. 1096–1105. old.
- HECKATHORN, D. D. [1997]: Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*. 44. évf. 2. sz. 74–99. old.
- HECKATHORN, D. D. [2002]: Respondent-Driven Sampling II.: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*. 49. évf. 1. sz. 11–34. old.
- KISS T. – KAPITÁNY B. [2009]: Magyarok Erdélyben: A minta kialakítása és az adatfelvétel. In: *Spéder Zsolt (szerk.): Párhuzamok – Anyaországi és erdélyi magyarok a századfordulón*. KSH-NKI Kutatási jelentések. 86. Budapest. 31–54. old.
- KASLSBEEK, W. D. [2000]: *Sampling Racial and Ethnic Minorities*. Summer Public Health Conference on Minority Health. Június 12–16. Chapel Hill, Észak-Karolina, Egyesült Államok. http://chsr.sph.unc.edu/Dissemination/MinHlth_2000.ppt (Elérés dátuma: 2010. május 26.)
- KALTON, G. [2001]: *Practical Methods for Sampling Rare and Mobile Populations*. Proceedings of the Annual Meeting of the American Statistical Association. Augusztus 5–9. http://chsr.sph.unc.edu/Dissemination/asa_pres_2000_35mins.ppt (Elérés dátuma: 2010. május 26.)
- LADÁNYI J. – SZELÉNYI I. [2001]: A roma etnicitás „társadalmi konstrukciója” Bulgáriában, Magyarországon és Romániában a piaci átmenet korszakában. *Szociológiai Szemle*. 11. évf. 4. sz. 85–95. old.
- LOHR, S. [2007]: *Recent Developments in Multiple Frame Surveys*. <http://www.amstat.org/sections/SRMS/proceedings/y2007/Files/JSM2007-000580.pdf> (Elérés dátuma: 2010. május 26.)
- LOHR, S. L. – RAO, J. N. K. [2006]: Estimation in Multiple Frame Surveys. *Journal of the American Statistical Association*. 101. évf. 405. sz. 1019–1030. old.
- MANSERGH, G. ET AL. [2006]: Adaptation of Venue-Day-Time Sampling in Southeast Asia to Access Men Who Have Sex with Men for HIV Assessment in Bangkok. *Field Methods*. 18. évf. 2. sz. 135–152. old.
- PARSONS, J. T. – GROV, C. – KELLY, B. C. [2008]: Comparing the Effectiveness of Two Forms of Time-Space Sampling to Identify Club Drug-Using Young Adults. *Journal of Drug Issues*. 38. évf. 4. sz. 1061–1082. old.
- PHILIPPI, T. [2005]: Adaptive Cluster Sampling for Estimation of Abundances within Local Populations of Low-Abundance Plants. *Ecology*. 86. évf. 5. sz. 1091–1100. old.
- SALGANIK, M. J. – HECKATHORN, D. D. [2004]: Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*. 34. köt. 193–239. old.

¹³ A módszer híveinek saját honlapján (<http://www.respondentdrivensampling.org/>) Elérés dátuma: 2010. május 26.) erről további – bár némiképp elfogult – információk nyerhetők.

- SHIMIZU, I. – SIRKEN, M. [2006]: *Network Sampling for Rare Trait Inference*. American Statistical Association Proceedings of the Survey Research Methods Section. 3664–3668. old. <http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000397.pdf> (Elérés dátuma: 2010. május 26.)
- SIRKEN, M. G. [1970]: Household Surveys with Multiplicity. *Journal of the American Statistical Association*. 65. évf. 329. sz. 257–266. old.
- SIRKEN, M. G. [1998]: *A Short History of Network Sampling*. American Statistical Association Proceedings of the Survey Research Methods Section. 1–6. old. http://www.amstat.org/sections/SRMS/proceedings/papers/1998_001.pdf (Elérés dátuma: 2010. május 26.)
- SKINNER, C. J. – HOLMES, D. J. – HOLT, D. [1994]: Multiple Frame Sampling for Multivariate Stratification. *International Statistical Review*. 62. évf. 3. sz. 333–347 old.
- SPÉDER Zs. (szerk.) [2009]: *Párhuzamok – Anyaországi és erdélyi magyarok a századfordulón*. KSH-NKI Kutatási jelentések. 86.
- STUEVE, A. et al. [2001]: Time–Space Sampling in Minority Communities. *American Journal of Public Health*. 91. évf. 6. sz. 922–926. old.
- SUDMAN, S. [1972]: On Sampling of Very Rare Human Populations. *Journal of the American Statistical Association*. 67. évf. 338. sz. 335–339. old.
- THOMPSON, S. K. [1990]: Adaptive Cluster Sampling. *Journal of the American Statistical Association*. 85. évf. 412. sz. 1050–1059. old.
- THOMPSON, S. K. – SEBER, G. A. F. [1996]: *Adaptive Sampling*. Wiley. New York.
- WAKSBERG, J. – JUDKINS, D. – MASSEY, J. T. [1997]: Geographic-Based Oversampling in Demographic Surveys of the United States. *Survey Methodology*. 23. évf. 1. sz. 61–71. old.
- XIA, Q ET AL. [2006]: The Effect of Venue Sampling on Estimates of HIV Prevalence and Sexual Risk Behaviors in Men Who Have Sex With Men. *Sexually Transmitted Diseases*. 33. évf. 9. sz. 545–550. old.

Summary

The present study addresses the most important sampling methods for rare populations.

The following methods are described: 1. increase of the sample size; 2. disproportionate stratification of the sample; 3. screening; 4. screening at the level of a primary sampling unit; 5. multiple/dual frame methods; 6. linkage exploitation methods: 6.1. network sampling; 6.2. adaptive cluster sampling; 7. methods applicable in cases of hidden populations: 7.1. time-space sampling; 7.2. respondent driven sampling. The author values these methods and demonstrates cases in which it is worth applying them.

These methods allow the collection of data on special social subgroups in a cost-efficient manner, on which the traditional sampling methods cannot, or only very costly can provide reliable information.