

A válaszadó-vezérelt mintavétel megbízhatóságának vizsgálata szimulációs módszerekkel¹

Kmetty Zoltán – Simon Dávid

zkmetty@yahoo.com; dr.david.simon@gmail.com

Beérkezés: 2013. 02. 21.

Végző átdolgozott változat beérkezése: 2013. 05. 23.

Elfogadás: 2013. 05. 25.

ÖSSZEFOGLALÓ: Cikkünkben az Egyesült Államokban és Nyugat-Európában már széles körben alkalmazott válaszadó-vezérelt mintavétellel (RDS) kapcsolatban végzett szimulációs kutatás eredményeit foglaljuk össze. A szimuláció célja a módszer megbízhatóságával kapcsolatban a kurrens kutatások kapcsán felmerült problémák vizsgálata. Elemzésünk eredményeként kimutatjuk, hogy a becslések megbízhatóságát elsősorban az alacsonyabb heterofília, másodsorban a magasabb fokszám csökkentik. Cikkünkben eredményként bemutatjuk, hogy a várhatóérték-becslések szórásának mekkora hányada magyarázható a fokszámbecslés hibájával. A szimuláció mellett röviden ismertetjük a módszerrel kapcsolatos kurrens módszertani fejleményeket és a felhasználás lehetőségeit.

KULCSSZAVAK: szimuláció, válaszadó-vezérelt mintavétel (RDS)

Az empirikus szociológiai kutatások rendszeresen foglalkoznak olyan célcsoportokkal, amelyek esetén nem áll rendelkezésre a csoportra vonatkozó teljes körű lista, a csoport mérete a teljes népességhez viszonyítva relatíve kicsiny (ilyen csoportok lehetnek egyes életmódcsoportok, szubkultúracsoportok, nemzeti és etnikai kisebbségek, az elit). Az ilyen csoportok vizsgálata során a véletlen mintavételen alapuló módszerek (többlépcsős, arányos vagy aránytalan rétegzett mintavételek) nem lehetnek költséghatékonyak. E csoportok esetén különféle alternatív mintavételi eszközök állnak rendelkezésre, amelyek közül egy lehetőség a hálózati mintavételek közé tartozó, a hólabda-mintavétel logikáját követő (Goodman 1961) válaszadó-vezérelt mintavétel. Cikkünkben – a *Statisztikai Szemlé*ben megjelent irodalmi összefoglalót (Simon 2012) követően – e Magyarországon is terjedőben lévő módszer kapcsán végzünk szimulációs vizsgálatot, amelynek célja a módszer megbízhatóságának és érvényességének tesztelése.

1 Ezúton mondunk köszönetet Dr. Bozsonyi Károlynak, aki a közös együttgondolkodásunk során több ötlettel segítette munkánkat.

A válaszadó-vezérelt mintavételt nemzetközi elismertség mellett Magyarországon is növekvő érdeklődés övezi. A válaszadó-vezérelt mintavétel a nem hagyományos mintavételi módszerek hálózati mintavételnek nevezett csoportjába tartozik, alapvetően a hólabda-mintavételnek egy továbbfejlesztett változata. A lekérdezés során a kérdezett ajánl további olyan személyeket, akik tagjai a vizsgált célcsoportnak. A módszer megalkotói és a felhasználók elsősorban a ritka és rejtett populációk (például intravénás droghasználók, homoszexuális férfiak, etnikai kisebbségek stb.) vizsgálatára ajánlják, illetve használták fel. Bár elsősorban Amerikában és Nyugat-Európában használják ezt a technikát, egyre több magyar piac- és társadalomkutatás esetében is alkalmaznak válaszadó-vezérelt mintavételt. Ilyen volt többek között a Bozsonyi Károly által irányított banki ügynököket vizsgáló kutatás (Bozsonyi – Jelenfi – Kmetty 2008), az MTA kutatása a külföldön dolgozó magyar diplomásokról (Csanády et al. 2008), a Kopint Alapítvány ausztriai munkavállalókat célzó kutatása (Hárs 2009; Hárs – Simon 2009), illetve a Táarki roma migrációs potenciál kutatása (Simon 2005; Bernát 2006).

Bár jelenleg a módszert elsősorban nehezen elérhető, rejtett, illetve kis célcsoportok vizsgálatára használják, a hagyományos survey felvételekben egyre inkább növekvő válaszmegtagadási arányok miatt, létjogosultságot nyerhetnek azok a kutatási megközelítések, melyek növelni tudják a válaszadási arányokat. Az ajánlásokon keresztül a kutatásba meghívott szereplők nagyobb valószínűséggel vesznek részt egy felvételen, mint egy előzmény nélküli megkeresés esetében. A módszer szintén fontos eszköze lehet az olyan online közösségek vizsgálatának (pl. Facebook), amelyek esetében rendelkezünk valamiféle ismerettel a kapcsolatok megoszlásáról és sűrűségéről.

A módszer sajátossága, hogy ismeretlen mintavételi keret mellett – bizonyos feltételek teljesülése esetén – a minta alapján lehetővé teszi a torzítatlan becslést viszonylag kis költségek mellett. A módszerrel kapcsolatban magyar nyelven is megjelentek szakirodalmi összefoglaló írások (Kapitány 2010; Simon 2012), melyek közül az utóbbi részletesen foglalkozik a módszer kritikájával, illetve a korábbi kutatások tapasztalataival is. A módszerrel kapcsolatban legfrissebben felmerült problémák közül kiemelkedik, hogy bár a mintavételi eljárás alapján képzett mintákból adható becslések torzítatlanságát továbbra sem kritizálják, a becslések szórása meghaladja a korábbi – részben szimulációs eredményeken, részben módosított szórásbecslésen alapuló – becsléseket (elsősorban Goela – Salganik [2010]). Ezek az újabb eredmények különösen kis mintaméret esetén megkérdőjelezzik a módszer használhatóságát, ugyanakkor növelik a becslés tulajdonságainak jobb megértését célzó kutatások jelentőségét.

Kutatásunkban különböző paraméterekkel rendelkező szimulált hálózatok, majd a hálózatokon a válaszadó-vezérelt mintavételi eljárás szimulálásán keresztül azt igyekszünk vizsgálni, milyen pontossággal lehet becsülni bizonyos sokasági paramétereket különböző elméleti populáció esetében. Ezenfelül azt is vizsgáljuk, hogy a becslés egyes elemei milyen módon járulnak hozzá a becslések szórásához. Cik-

künk célja, hogy pontosítsuk a módszer felhasználási feltételeit, és alapot teremtsünk a további, itt bemutatott problémák kiküszöböléséhez

A válaszadó-vezérelt mintavétel alapjai

A válaszadó-vezérelt mintavétel (*respondent driven sampling*, RDS) első leírása Douglas D. Heckathorn cikkében jelent meg 1997-ben (Heckathorn 1997). A szerző a mintavételi eljárást a következő lépésekben összegzi (Simon [2012] alapján rövidítve). Első lépésben kiválasztják a minta kezdőpontjait úgy, hogy már az első válaszadóknak is jutalmat ígérnek a részvételért. Az első lépcsőbe bekerült személyekkel lefolytatják az interjút, és megkérik őket a kutatás végén, hogy segítsenek további alanyokat beszervezni a vizsgált célcsoportból. A beszervezésért a kérdezettek szintén jutalmat kapnak, de csak abban az esetben, ha az általuk ajánlott személlyel megvalósul a lekérdezés. Minden kutatásban szereplő személy jutalmat kap a részvételért, és azért is, ha sikeresen beszervezett valakit. Eredetileg ezt a kutatási felépítést még további két ponttal egészítették ki. Egyrészt a bevonható további személyek száma válaszadóként korlátozott számú – általában két fő –, másrészt az interjút egy olyan helyen készítették, ahova a válaszadónak kellett elmennie.

A szerzők által leírt lépések természetesen csak irányadók, az elmúlt évtized során az RDS módszert alkalmazó kutatások a fenti leírástól sok esetben eltértek. Az RDS módszer „nagy ígérete” az egyszerűbb hólabda-adatfelvétellel szemben, hogy az egyes populációs paraméterekre torzítatlan becslést lehet adni. Ez praktikus azt jelenti, hogy egy ilyen jellegű adatfelvételtől is lehet érvényes megállapításokat tenni a sokaságra. A torzítatlan becsléshez azonban több feltételnek is meg kell hogy feleljen a minta (Salganik – Heckathorn 2004, idézi Simon 2012). Az egyik alapfeltétel az, hogy a válaszadóknak és ajánlottaknak kölcsönös (reciprok) kapcsolatban kell állniuk egymással. Tehát ha „A” személy ismeri „B” személyt, akkor „B” személynek is ismernie kell „A” személyt. Ez az előfeltétel a mintavétel technikai lebonyolítása miatt gyakorlatilag mindig fennáll. A második fontos előfeltevés az, hogy a hálózatnak egy komponensből kell állnia. Tehát a hálózatban bármely személytől el lehet jutni bármely másik személyig. Ez az előfeltétel nehezen tesztelhető, bár kidolgozhatók olyan indikátorok, melyek részben jelzik ennek teljesülését. További előfeltevés az, hogy a mintavétel visszatevéssel történjen. Tehát elviekben egy személy többször is bekerülhet a mintába. Ez az előfeltevés a gyakorlatban nem realizisztikus, mivel egy személyt csak egyszer kérdezzük le a kutatásban, ebből következően tehát valamekkora torzítás jelentkezeni fog. A torzítás annál kisebb, minél nagyobb a célcsoportunk és minél kevésbé klaszterezett. Továbbá a válaszadónak teljes pontossággal meg kell tudnia határoznia azt is, hogy hány kapcsolata van a célcsoportban (ezt nevezzük fokszámnak). A fokszámnak szerepe van a becslésekben, ezért van rá kifejezetten szükség. A Salganikék által kidolgozott módszer diszkrét változókra (pl. nem, iskolai végzettség, településtípus) lett kidolgozva, ami egyértelműen szű-

kíti a módszer hatókörét, bár ennek kapcsán mindenképp megjegyezhetjük, hogy bármely folytonos változó átalakítható diszkrétte (például a jövedelemből jövedelmi kategóriákat képezhetünk).

A fenti feltételek teljesülése esetén a hálózat egyes paramétereinek becslése alapján megbecsülhető a vizsgált diszkrét változó populációs eloszlása (dichotóm, kétértékű változóra Salganik – Heckathorn [2004] alapján). Az alábbiakban egy kétértékű „A” és „B” értékeket felvevő változó esetében írjuk le, hogyan lehet „A” sokaság megoszlását (populációs valószínűségét) megbecsülni. A könnyebb érthetőség kedvéért a „neme” változót használjuk fel példaként a képletekben:

$$\hat{P}_A = \frac{\hat{D}_B \hat{C}_{B,A}}{\hat{D}_A \hat{C}_{A,B} + \hat{D}_B \hat{C}_{B,A}} \quad (1)$$

$$\hat{D}_A = \frac{n_A}{\sum_{j=1}^{n_A} \frac{1}{d_j^A}}; \quad \hat{D}_B = \frac{n_B}{\sum_{j=1}^{n_B} \frac{1}{d_j^B}} \quad (2)$$

$$\hat{C}_{A,B} = \frac{f_{A,B}}{n_A}; \quad \hat{C}_{B,A} = \frac{f_{B,A}}{n_B} \quad (3)$$

ahol:

$n_A; n_B$

esetszámok: a változó A, illetve B értéket választó személyek száma a mintában (pl. A: férfi, B: nő)

$D_A; D_B$

fokszámok (*degree*): a változó A, illetve B értéket választó személyek átlagos fokszám becslése a mintából (a férfiaknak és a nőknek átlagosan hány kapcsolata van a minta alapján)

$C_{A,B}; C_{B,A}$

összekötöttség (*connection*): az A és B, illetve B és A értéket választó személyek közötti kapcsolat valószínűségének becslése a mintából (egy férfinak milyen valószínűséggel van női ismerőse, és egy nőnek milyen valószínűséggel van férfi ismerőse)

P_A

valószínűség (*probability*): az A értéket választó személyek arányára adott becslés (a férfiak arányára adott becslés)

$d_j^A; d_j^B$

az A, illetve B értéket választók közül a j. válaszadó fokszáma (a j-ig válaszadó férfi-, illetve nőismerőseinek száma)

$f_{A,B}; f_{B,A}$

azon válaszadók száma, akik az A, illetve B értéket választották, és az általuk ajánlott személy a B, illetve A értéket választotta (azon férfiak száma, akik nőt választottak, és azon nők száma, akik férfit választottak).

A számítás logikája alapvetően az, hogy első lépésben két-két hálózati paramétert becslünk meg, majd ez alapján teszünk becslést a vizsgált változó populációs eloszlá-

sára. Vizsgált változóként a nemet feltételezve a minta alapján először megbecsüljük a férfiak és nők átlagos kapcsolatszámát, illetve a férfiak nőismerőseinek és a nők férfiismerőseinek átlagos számát, majd ezek alapján teszünk becslést a férfiak és nők arányára a vizsgált népességben.

Ezt a becslést a továbbiakban RDS I becslésnek nevezzük. Volz és Heckathorn (2008) egy másik elterjedtté vált becslésre adtak javaslatot a Hansen–Hurwitz-becslés módosítása nyomán (a továbbiakban RDS II becslés):

$$(5) \quad \hat{p}_A = \frac{n_A}{n} \frac{\hat{D}}{\hat{D}_A}$$

$$(6) \quad \hat{D} = \frac{n}{\sum_{i=1}^n \frac{1}{d_i}}$$

ahol:

D az átlagos fokszám becslése a mintából (a mintából becsült kapcsolatszám) minden egyéb jelölés a korábban megadottal megegyezik.

Utóbbi becslés előnye elsősorban abban áll, hogy esetében a becslés szórására, azaz a megbízhatóságra vonatkozóan is rendelkezésre áll – igaz, csak közelítő módon – zárt képlettel kiszámítható becslés, míg az RDS I becslés szórására csupán idő- és számításigényes nem paraméteres iteráción alapuló becslést adhatunk.

A válaszadó-vezérelt mintavételhez kapcsolódó korábbi módszertani kutatások

A válaszadó-vezérelt mintavétel megbízhatóságának és érvényességének vizsgálatával számos korábbi kutatás foglalkozott, melyeket a korábban már idézett tanulmányunk (lásd Simon 2012) részletesen is taglal, itt azokat az – elsősorban szimulációs – kutatásokat idézzük, amelyek jelen kutatásunk szorosabb előzményét jelentik.

Salganik és Heckathorn (2004) szimulációs vizsgálatának részletes bemutatását azért tartjuk fontosnak, mivel saját szimulációnk is ezen alapul, de röviden kitérünk más – kutatási kérdésünkkel foglalkozó – elemzésekre is, melyek ismertetése során elsősorban az eredményekre koncentrálunk, amelyeket cikkünk utolsó részében öszszefoglalunk saját eredményeinkkel.

Salganik és Heckathorn (2004) a fent már idézett becslés mellett az általuk javasolt (RDS I) becslés tulajdonságait – elsősorban torzítatlanságát – szimuláció segítségével is vizsgálták. Vizsgálatukban első lépésként létrehozták a szimulációban a célpopuláció szerepét betöltő alappopulációt. Az alappopulációt két részre osztották: A, illetve B tulajdonsággal rendelkező alappopulációra (ez megfeleltethető egy

változó két értékének). Definiálták egy meghatározott eloszlás alapján, hogy mekkora legyen az egyes populációtagok kapcsolatszáma. Definiálták, mekkora legyen a kapcsolatok aránya a két csoport tagjai között a kisebb összkapcsolatszámúmal rendelkező csoport összes kapcsolatához képest, ezt nevezzük a továbbiakban heterofiliának. Tehát a heterofília az A és B tulajdonságú csoportok összekötöttségét méri. Értéke 0, ha nincs kapcsolat a két csoport között, és 1, ha a kisebb összkapcsolatszámúmal rendelkező csoport összes kapcsolata a nagyobb csoport felé irányul. Az alappopuláció mérete Salganikék szimulációjában 10 000 volt. Mind a két csoport esetén a kapcsolatszám exponenciális eloszlású volt, az A csoport esetén 10, a B csoport esetén 20-as fokszámtálgal. A heterofiliát pedig 0,6-ra állították be, tehát a kisebb kapcsolatszámú csoport kapcsolatainak 60 százaléka irányult a nagyobb csoport irányába.

A hálózat kialakítása után 5 kezdőpontból kiindulva szimulációs módszerrel vizsgálták az RDS mintavételt. Minden lépésnél a már mintába kerültekhez a velük kapcsolatban állók közül két-két további populációtagot választottak ki véletlenszerűen, akikkel bővítették a mintát. Az eljárást addig folytatták, amíg el nem érték a kívánt mintaméretet. Az eljárás során visszatevéses mintavételt alkalmaztak, azaz egy-egy populációtag többször is a mintába kerülhetett. A mintavételt 1000-szer ismételték meg minden beállítás mellett. A szimulációk során elsősorban a torzítás mértékét (azaz a populációs és a mintából becsült arány különbségét) vizsgálták egyrészt a mintaméret, másrészt a populációs arányok, harmadrészt a kezdőpont-kiválasztás szempontjai alapján. A szerzők azt találták, hogy a torzítás mértéke már 500 elemű minta esetén bármely populációs arányra 0,1%-nál kisebb. A kezdőpontok kiválasztása esetén pedig azt, hogy 500 elemű mintára egyszerű véletlen kezdőpont-kiválasztás, kapcsolatszám-arányos valószínűséggel történt kezdőpont-kiválasztás, illetve a kapcsolatszám négyzetével arányos valószínűségű kezdőpont-kiválasztás esetén is a becslések torzítása 0,1% alatt marad. A cikk ábrái a becslések szórását is tartalmazzák, azonban ezeket sajnálatos módon a szerzők nem számszerűsítik.

A továbbiakban még négy tanulmányt ismertetünk, amelyek eltérő adatokon vizsgálják a válaszadó-vezérelt mintavétel tulajdonságait. A tanulmányok ismertetése során három tényező szerepére koncentrálnak a becslések szórására és a torzításra vonatkozóan: a heterofília hatására, a hálózati tulajdonságokra és a mintavétel sajátosságaira.

Gile és Handcock (2009) 1000 modellezett hálózatot vizsgáltak, amelyek egyenként 1000 fős populációt szimuláltak (a nem vizsgált hálózati paramétereket egy korábbi kutatás valós hálózati adatai alapján állították be, a szimulált hálózatok a fent leírthoz hasonló módon kerültek előállításra exponenciális foksámeloszlással és változó heterofília-mértékkel). Az 1000 darab 1000 elemű hálózatból 500 elemű mintákat vettek. Szimulációjuk során azt találták, hogy nagyon alacsony heterofília esetén (tehát ha a kapcsolatok nagy része csoporton belülre mutat) az RDS II becslések szórása jelentősen megnő. Ugyanakkor azt is kimutatták, hogy torz (a populáci-

ős eloszlástól jelentősen eltérő) kiinduló minta esetén a hosszabb láncok (kevesebb kezdőpont) kisebb torzításhoz vezetnek, összességében a torz kiinduló minta nincs jelentős hatással az RDS II becslés szórására. Emellett a szerzők azt is vizsgálták, hogy milyen hatása van annak, ha a vizsgált változó egyes értékei által kijelölt csoportok átlagos kapcsolatszáma eltér egymástól (például ha nemek szerinti megoszlásra vagyunk kíváncsiak, és a nők átlagos kapcsolatszáma jelentősen eltér a férfiakétól). A szerzők azt találták, hogy az átlagos kapcsolatszámokban mutatkozó különbségek a nagyobb átlagos kapcsolatszámokkal jellemezhető csoport értéke gyakoriságának alulbecsléséhez vezetnek. Ez a hatás még jelentősebb, ha a gyakoriságok nagymértékben eltérnek az egyenletestől. Azt tapasztalták továbbá, hogy amennyiben a minta mérete közelít a populáció méretéhez, a vizsgált változó egyes értékeihez tartozó eltérő átlagos kapcsolatszám torzító hatása jelentős mértékben megnő. A szerzők végül azt is megállapították, hogy a visszatevés nélküli mintavétel csökkenti a becslések (mind az RDS I, mind az RDS II) szórását (valamint a kiinduló minta torzító hatását).

Goela és Salganik (2010) egy valós nagymintás kutatás hálózatának 4430 fős rész-hálóját használták (ezen belül 85 különálló komponens), amelyen az RDS I becslés feltételeinek megfelelő 500 fős mintákat szimuláltak, a mintavételt 10 ezer alkalommal ismételték meg. Eredményeik azt mutatták, hogy a becslések szórása számottevően meghaladta a várt értéket, mértéke egy egyszerű véletlen mintával összevetve 5–10-szeres is lehet. Ez azt jelenti, hogy egy 1000 fős RDS mintához számolt becslési hiba akár olyan nagy is lehet, mint egy 100 fős véletlen minta esetében.

Tomas és Gile (2010) 1000 fős szimulált hálózattal modelleztek egy populációt (a hálózat tulajdonságait úgy paraméterezték, hogy az egy korábbi kutatás valós paramétereinek, például átlagos fokszámának feleljen meg). Vizsgálatuk szerint a célcsoporthoz viszonyítva kis minta és alacsony heterofília esetén az RDS I becslés kisebb torzítást eredményez a becslések nagyobb szórása mellett, míg nagyobb heterofília mellett az RDS II becslések szórása kisebb.

Verdery és Mouw (2011) a Facebook 100, egyetemalapú ismeretségi hálóján modellezett mintavétellel vizsgálták a válaszadó-vezérelt mintavétel tulajdonságait. A kutatás során 100 ezer mintavételt modelleztek. Elemzésük során kimutatták, hogy azon változók esetén, amelyek jelentős klaszterezettséghez vezetnek (azaz a hálózaton belül olyan alcsoportok vannak, amelyekre jellemző a nagyszámú alcsoporton belüli és kevés az azon kívüli kapcsolat, valamint ezen alcsoportok esetén a vizsgált változó szóródása kisebb), a becslések szórása jelentősen megnövekedett. Ez gyakorlatilag megegyezik azzal a korábbi eredménnyel, miszerint az alacsony heterofília megnöveli a becsléseinkhez tartozó konfidencia-intervallumot.

Összefoglalva a leírtak fontosabb megállapításait a korábbi kutatások azt állapították meg, hogy a heterofília mértéke fordított összefüggést mutat a becslések szórásával (azaz a megbízhatósággal), az egyes csoportok átlagos kapcsolatszáma közötti különbség ugyancsak növeli a becslések szórását, mindezek mellett a kiin-

duló minta nagysága (amely adott mintanagyság mellett a láncok hosszát is megszabja) a becslések szórását nem befolyásolta jelentősen.

Kutatási kérdés

A vonatkozó kutatásokat áttekintve nem találtunk olyat, amely szisztematikusan vizsgálta volna a fontosabb tényezők együttes hatását. Ezért úgy döntöttünk, hogy kutatásunk során a válaszadó-vezérelt mintavétel két szakirodalomban használt becslésének (RDS I és RDS II) szóródását és torzítását vizsgáljuk különféle hálózati és populációs paraméterek kombinációja esetén. Emellett egy a korábbi irodalomban egyáltalán nem szereplő tényezőt is vizsgálunk, azt, hogy a változók populációs eloszlására adott becslések szórásának mekkora hányadát adják a becslés egyes tényezői, remélve, hogy így jobban megérthetjük, mekkora mértékben hatnak az egyes paraméterek a becslések megbízhatóságára. A hálózati és mintavételi tényezők közül vizsgáltuk az átlagos fokszámot, a heterofília nagyságának hatását, valamint a kiindulópontok számát. Mind a három paraméter esetében további három-három paraméterbeállítást teszteltünk. Az eredmények kiértékelésénél elsősorban az RDS I és RDS II becslések megbízhatóságára koncentráltunk, de vizsgáltuk az A és B tulajdonságú populációs tagok fokszámbecslését, a teljes sokaságra adott fokszámbecslést, valamint az A és B, illetve B és A csoportok közötti összekötöttség (heterofília) becslését is.

Mindezek alapján kutatási kérdéseinket a következő módon fogalmaztuk meg: *hogyan befolyásolják a hálózati és mintavételi tényezők a becslések torzítását és szórását, milyen mértékben határozzák meg a becslések szórását a becslés egyes tényezői, valamint hatással vannak-e egyéb hálózati tényezők a becslések szórására, illetve torzítására?*

Módszertan

A kutatási kérdések megválaszolásához a szakirodalmi előzményekre támaszkodva egy hálózati szimulációs módszert választottunk. A hálózatok szimulálásához Salganik és Heckathorn (2004) módszerének módosított változatát alkalmaztuk. A szimulációk során fix paraméterként kezeltük a vizsgált változó eloszlását (dichotóm, 0,3–0,7 eloszlású, a továbbiakban A értékű és B értékű), a populáció méretét (10 000), valamint a minta méretét (1000). Ezek a paraméterek az összes futtatásban állandók voltak. A szimulációk során a két hálózati és egy mintavételi paraméter esetén (átlagos foksám, heterofília, illetve kiindulópontok száma) három-három értéket határoztunk meg. A különböző paramétertényezők esetében több mintavételi szituációt lefuttattunk, és megvizsgáltuk minden esetben a kapott becslés érvényességét és megbízhatóságát, valamint azt, hogy az egyes tényezőknek mekkora a hatása a becslések torzítására és szórására. A foksám esetén a kisebbik csoport nagyobb átlagos foksámát feltételeztük (ezzel Salganik és Heckathorn modelljének feleltünk meg). A kisebbik csoport esetén (A értékkel jellemezhető)

a foks számok elméleti várható értéke: 7,66; 12,11 és 21.² A nagyobb csoport esetén (B értékkel jellemezhető) a foks számok elméleti várható értéke: 4,33; 6,55 és 11. A heterofília tekintetében azt határoztuk meg, hogy a kisebbik csoport (A értékkel jellemezhető) tagjainak összes kapcsolata milyen arányban irányul a nagyobbik csoport (B értékkel jellemezhető) felé. Az értékek a következők voltak: 15%; 30%; 60%.³ Végül a minta kiindulópontjainak száma tekintetében 20, 40 és 80 értékek mellett döntöttünk. Ez utóbbi egyúttal befolyásolta a mintavétel során a mintamérethez szükséges egymást követő lépések számát is (lánc hossz): 20 kiindulópont esetén a várható átlagos lánc hossz 5,6, 40 kiindulópont esetén 4,6, míg 80 kiindulópont esetén 3,6. Mivel a mintavétel szimulálása során – az idézett szerzőktől eltérően – a gyakorlatban alkalmazott visszatevés nélküli modellt alkalmaztuk, az itt megadott átlagos lánc hosszok kismértékben eltérnek a tapasztalttól. A szimuláció során minden paraméter-beállítással 10 hálózatot generáltunk és minden hálózat esetén 1000 mintavételt szimuláltunk, összesen 270 hálózaton 270 000, 1000 elemű mintavétel történt.

Az eredményeket paraméterkombinációként hasonlítottuk össze egymással, illetve egy azonos paraméterű egyszerű véletlen minta (EVM) várható adataival. Minden minta esetén kiszámítottuk az RDS I és RDS II becslést, a becslések hibájának szórását, valamint kiszámítottuk az RDS I becslést a valós heterofília-értékekkel és a becsült foks számokkal, illetve a becsült heterofília-értékekkel és a valós foks számokkal. Az így kapott becslések segítségével paraméterkombinációként meghatároztuk, hogy milyen mértékben járul hozzá a becslések szórásához a foks szám, illetve a heterofília. Az eltérő hálózatszimulációk hatását azonos paraméterek mellett a becslések populációs értéktől való eltérésére vonatkozóan ANOVA elemzéssel vizsgáltuk.

Eredmények

Foglaljuk össze a szimulációk eredményeit⁴ a kutatási kérdéseket követve:

Elsőként az RDS I és RDS II becslések torzítását vizsgáljuk a különböző paraméterek mentén. Az egyes paraméterek független (kontrollált) hatásának vizsgálata érdekében egy általános lineáris modellt (GLM) illesztettünk a szimulált adatokra.⁵ Modelünkben a paraméterváltozók alacsony mérési szintűként szerepeltek, így az elemzés tulajdonképpen több utas, többszintű ANOVA elemzésnek feleltethető meg.

A modell alapján elmondható, hogy mindhárom paraméter és a hálózatszimuláció is szignifikáns hatással volt mindkét becslés esetén. A hatáserőségek értékelése érdekében megvizsgáltuk a modell alapján becsülhető marginálisokat az egyes para-

2 A foks számokat exponenciális eloszlásúnak feltételeztük, technikai okokból – a hálózat összefüggését garantálandó – minden populációtag esetén eggyel megnöveltük, így nulla foks szám nem fordulhatott elő.

3 Tekintettel arra, hogy a nagyobb csoport a populáció 70%-át teszi ki, átlagos foks számuk viszont a kisebb csoport valamivel több mint fele (52–56%-a), a 15%-os arány viszonylag erőteljes homofiliát, a 30%-os arány véletlenhez közeli kapcsolatválasztást, míg a 60%-os arány kifejezett heterofiliát jelent.

4 A hálózatok mérete a szimuláció sajátosságai miatt esetenként némileg kisebb az eredeti méretnél. A foks számátlagok jól követik az előre rögzített értékeket.

5 Randomfaktorként figyelembe vettük azt is, hogy az egyes paraméterbeállítások mentén 10–10 hálózatot generáltunk, tehát összesen 270 szimulált hálózatból vettük a hálózatonkénti 1000 darab mintát.

méterek esetén (1. táblázat), amely megmutatja, hogy az egyes paraméterek, tehát a heterofília, a fokszám és a kiinduló minta mérete külön-külön milyen hatást gyakorolnak a becslések eltérésére a tényleges populációs értékektől.

1. táblázat: A heterofília, a fokszám és a kiinduló mintanagyság becsült hatása (a modell alapján kiszámított marginálisátlagok és standard hibájuk)

Heterofília	RDS I		RDS II	
	Átlag	Std. hiba	Átlag	Std. hiba
15%	0,018	<0,001	-0,007	<0,001
30%	0,014	<0,001	0,003	<0,001
60%	0,006	<0,001	0,007	<0,001

Fokszám	RDS I		RDS II	
	Átlag	Std. hiba	Átlag	Std. hiba
21	0,014	<0,001	0,004	<0,001
12,11	0,013	<0,001	0,002	<0,001
7,66	0,012	<0,001	-0,003	<0,001

Kiinduló minta	RDS I		RDS II	
	Átlag	Std. hiba	Átlag	Std. hiba
20	0,013	<0,001	<0,001	<0,001
40	0,013	<0,001	0,001	<0,001
80	0,013	<0,001	0,002	<0,001

A tábla adatai alapján látható, hogy mindkét becslés elsősorban a csoportok összekötöttségére, a heterofíliára érzékeny. Míg azonban az RDS I becslés a heterofília növekedésével szinte lineárisan egyre kisebb torzítású becslést ad, az RDS II becslés a közepes (véletlenhez közel álló) heterofília esetén adja a legpontosabb becslést. Kisebb mértékben, de mindkét becslésre hatással van a fokszám várható értéke (a táblázatban a kisebb csoport fokszámának beállított várható értékeit közöljük). Ebben az esetben is hasonló jelenséget tapasztalhatunk, mint a heterofília esetén: az RDS I becslés az egyre kisebb fokszám-várhatóértékű hálózatok esetén egyre pontosabb becslést ad, az RDS II viszont valahol a közepes tartományban adja a legpontosabb becslést. Végül a kiinduló minta nagyságának hatása – bár statisztikai értelemben szignifikáns – gyakorlati szempontból nem tekinthető jelentősnek. A hálózathatás hasonló módon nem értékelhető, hiszen itt random hatásról van szó. Az 1. tábla eltérés-négyzetösszegeket tartalmazó oszlopa alapján kiszámítható parciális magyarázóerőt mutató parciális éta-négyzet értéke RDS I becslés esetén 0,002, míg RDS II becslés esetén 0,008, ami a hálózatok struktúrájának szignifikáns, de nagyon kis mértékű hatását mutatja. Végül nem hagyható figyelmen kívül az a tény (lásd 1. táblázat), hogy az RDS I becslés torzítása kedvezőtlen paraméterezés esetén közel egy nagyságrenddel nagyobb, mint az RDS II becslése. Összefoglalva: az RDS I becslés különösen kedvezőtlen paraméterek esetén jelentősebb torzítással bír, egyértelműen csökken a torzítása a heterofília

növekedésével és kisebb mértékben a foksám csökkenésével, ezzel szemben az RDS II becslés általában kisebb torzítással bír, de mind a foksám, mind a heterofília tekintetében az általunk közepesnek tekintett értékektől való eltérés a torzítást növeli.

A következő lépésben a becslések megbízhatóságát vizsgáljuk meg. Sajnos ebben az esetben nincs lehetőségünk komplex modell alkalmazására, így egyszerű szórásegyenlőség-vizsgálatokra, illetve a paraméterkombinációk alkalmazása mellett a becslések szórásadataira hagyatkozhatunk. A három vizsgált paraméter hatását a becslések szórására varianciahomogenitás-vizsgálattal teszteltük (2. táblázat).

2. táblázat: A heterofília (H), a foksám-várhatóérték [E(D)] és a kiinduló minta méretének (NK) hatása az RDS I és RDS II becslés szórására (a Levene-teszt eredményei alapján mindhárom tényező mindkét becslési módszer esetén szignifikáns eltérést hozott létre a becslések szórásában)

H	E(D)	NK	A becslések szórása	
			RDS I	RDS II
15%	21	20	0,0352	0,0442
		40	0,0353	0,0391
		80	0,0367	0,0344
	12,11	20	0,0349	0,0436
		40	0,0350	0,0385
		80	0,0359	0,0337
	7,66	20	0,0345	0,0433
		40	0,0354	0,0381
		80	0,0362	0,0334
30%	21	20	0,0256	0,0269
		40	0,0259	0,0266
		80	0,0263	0,0257
	12,11	20	0,0246	0,0261
		40	0,0250	0,0253
		80	0,0257	0,0248
	7,66	20	0,0241	0,0255
		40	0,0243	0,0247
		80	0,0251	0,0238
60%	21	20	0,0201	0,0200
		40	0,0201	0,0200
		80	0,0201	0,0199
	12,11	20	0,0184	0,0183
		40	0,0185	0,0184
		80	0,0187	0,0184
	7,66	20	0,0171	0,0170
		40	0,0175	0,0173
		80	0,0176	0,0173

A Levene-tesztek adatai alapján mindhárom vizsgált paraméter szignifikánsan befolyásolta mindkét becslés szórását. Az RDS I becslés szórására a legjelentősebb hatást a heterofília gyakorolta, a nagyobb heterofília erősen csökkentette a becslések szórását (15% esetén 0,0345–0,0367 között, míg 60% esetén 0,0171 és 0,0201 között). A foksám és a kiinduló minta méretének hatása ennél jóval gyengébb volt: az alacsonyabb foksám-várhatóérték és az alacsonyabb kiinduló mintaméret egy nagyságrenddel kisebb mértékben csökkentette az RDS I becslés szórását. Az RDS II becslés esetén is a heterofília hatása volt a legerősebb a becslés megbízhatóságára: a nagyobb heterofília kisebb szórással járt (15% esetén 0,0334–0,0442 között, 60% esetén 0,0173–0,0200 között). Ebben az esetben azonban a kiinduló minta mérete valamivel nagyobb hatást gyakorolt a szórásra: nagyobb kiinduló minta kisebb szórást eredményezett. Emellett csak gyenge hatása volt a foksám várhatóértékének: azonos körülmények között a kisebb foksám-várhatóérték az RDS II becslés valamivel kisebb szórását eredményezte. A két becslés szórását összevetve azt láthatjuk, hogy az optimálistól eltérő körülmények esetén az RDS I becslés szórása alacsonyabb. Érdemes a tapasztalt szórásértékeket összevetni egy egyszerű véletlen minta esetén hasonló populációs eloszlású változóra adott becslés szórásával. Ez az érték 1000 elemű egyszerű véletlen minta és 0,3–0,7 eloszlású változó esetén 0,0145, amely értéket a legkedvezőbb paraméterértékek esetén (nagy összekötöttség, alacsony várható foksám) mindkét becslés szórása megközelíti. Ugyanakkor a legkedvezőtlenebb vizsgált paraméterbeállítás esetén az RDS I becslés szórása 2,5-szerese, míg az RDS II becslés szórása valamivel több mint 3-szorosa az azonos méretű EVM mintából származó becslésének. Tehát egy ugyanakkora EVM és RDS minta esetében utóbbinál akár háromszor szélesebb konfidenciaintervallum tartozhat egy változó várhatóértéknek a becsléséhez.

A hálózati szimuláció hatásának tesztelése érdekében a szóráshomogenitás-vizsgálatot minden egyes paraméterbeállítás mellett elvégeztük a 10-10 hálózatra, az eredmények azt mutatták, hogy az eltérő hálózatok általában nem okoztak szignifikáns eltérést a becslések szórásában.⁶

Végül a több hálózati paraméter (foksám és heterofília) alapján becslő RDS I esetén megvizsgáltuk, hogy a becslés eltérései milyen mértékben függnek össze az egyes hálózati paraméterbecslések eltéréseivel. Ennek érdekében az RDS I becslést úgy is kiszámítottuk, hogy ahhoz az egyik esetben a populációs heterofíliaadatokat és a foksámbecsléseket (a megmaradt eltérések a foksámbecslések hatását tükrözik), a másik esetben a populációs foksámadatokat és a heterofíliára vonatkozó becsléseket használtuk (a megmaradó hiba az összekötöttség becsléséből származik). A becslések hibájának átlagát tekintve megállapítható, hogy a heterofíliára vonatkozó becslések hibája okozza az RDS I becslések konzisztens pozitív torzítását: a populációs foksámátlag használatával a becslések átlagos eltérése minden paraméterbeállítás

6 A 27-féle paraméterkombinációból az RDS I becslés szórása 4 esetben, míg az RDS II becslés szórása 5 esetben volt szignifikánsan különböző (5%-os elsőfajú hiba mellett) hálózatonként.

mellett növekedett, míg a populációs heterofíliaadatok használatával negatívvá vált, abszolút értéke pedig általában lényegesen csökkent. A becslések szórása esetén az egyes paraméterbeállításokra kiszámítottuk a populációs fokszámtálggal, illetve populációs heterofíliaadattal adott becslések szórásnégyzeteinek összegét, amelyek jól közelítették az eredeti becslések szórásnégyzetét. Ezt az összefüggést használva a továbbiakban az egyes hálózati paraméterbecslések által hordozott hibahányadot a szórásnégyzet hányadával írjuk le. A szórásnégyzethányadok és az egyes vizsgált paraméterek (fokszám, heterofília, kiinduló minta mérete) hatását vizsgálva azt találtuk, hogy a kiinduló minta mérete nem gyakorolt jelentős hatást, így e paraméter mentén átlagoltuk a szórásnégyzethányadokat. A 3. táblázat a heterofília és a fokszám-várhatóérték mentén mutatja meg a fokszámbecslés részesedését az RDS I becslések hibájából.

3. táblázat: A heterofília és a fokszám-várhatóérték hatása a fokszámbecslés-hiba részesedésére az RDS I becslés varianciája esetén (a kezdőpontok számának hatását átlagolva)

Heterofília	Fokszám-várhatóérték		
	21	12,11	7,66
15%	0,120	0,167	0,224
30%	0,260	0,331	0,424
60%	0,515	0,604	0,678

A táblázat alapján elmondható, hogy mind a heterofília, mind a fokszám-várhatóérték befolyásolta a fokszámbecslés hibája által magyarázható hiba arányát az RDS I becslés varianciáján belül. Látható, hogy a heterofília hatása ebben az esetben is nagyobb a vizsgált paraméter-összeállítás mentén. Nagyobb heterofília és kisebb fokszám-várhatóérték esetén a fokszámbecslés hibája kevesebbet magyaráz az RDS I becslés varianciájából, ennek megfelelően ezekben az esetekben a heterofíliára vonatkozó becslés hibája magyaráz jóval nagyobb arányt a becslések szórásnégyzetéből. Ha ehhez az eredményhez figyelembe vesszük az RDS I becslés szórásának paraméterfüggésénél leírtakat is, akkor elmondható, hogy azokban az esetekben, amikor az RDS I becslés szórása nagy (elsősorban alacsony heterofília, másodsorban, de kisebb mértékben magas fokszám esetén), a becslés szórásáért jelentős részben a heterofília becslésének a hibája felelős.

Konklúzió és további kutatási lehetőségek

A következőkben igyekszünk összefoglalni szimulációs kutatásunk eredményeit, azokat összevetni az ismertetett irodalom eredményeivel, valamint mindezek alapján egyrészt praktikus javaslatokat megfogalmazni a módszert alkalmazni kívánó kutatók számára, illetve eredményeink alapján javaslatot adunk a további kutatás lehetséges irányára.

Vizsgálatunk első része az RDS I és RDS II becslés összehasonlítására vonatkozó különféle paraméterek mellett. A 4. táblázatban összefoglaltuk a korábban részletesen leírt eredményeket.

4. táblázat: Az összekötöttség, a foksám, kiinduló minta és egyéb hálózati hatások befolyása az RDS I és RDS II becslés torzítására és szórására

	RDS I becslés		RDS II becslés	
	Torzítás	Szórás	Torzítás	Szórás
Az RDS I és RDS II egymáshoz viszonyítva	Nagyobb (főleg kedvezőtlen esetben)	Legtöbbször kisebb	Legtöbbször kisebb	Nagyobb (főleg kedvezőtlen esetben)
Nagyobb heterofília	Jelentősen csökkenti	Jelentősen csökkenti	A heterofília közepes értékénél megszűnik	Jelentősen csökkenti
Nagyobb átlagos foksám	Kismértékben növeli	Kismértékben növeli	A heterofília közepes értékénél megszűnik	Kismértékben növeli
Nagyobb kiinduló minta	Lényegében nem befolyásolja	Kismértékben növeli	Lényegében nem befolyásolja	Csökkenti
Egyéb hálózati hatás	Szignifikáns, de kicsi	Nem szignifikáns	Szignifikáns, de kicsi	Nem szignifikáns

Összevetve a két becslésre vonatkozó szimulációs eredményeket, azt mondhatjuk, hogy mindkét becslés erősen érzékeny a heterofília hatásra, azaz arra, hogy a vizsgált változó értékei szerint létrehozható csoportok között milyen gyakoriságúak a kapcsolatok. Amennyiben a változó értékei mentén a kapcsolathálózat nagyon szegmentált, mindkét becslés szórása magas lesz, azonban míg az RDS I becslés torzítása is megnő, a megbízhatósága valamivel alacsonyabb marad, mint a kisebb torzítású, de nagyobb szórású RDS II becslésé. A vizsgált célcsoport átlagos foksáma mindkét becslést csak kismértékben befolyásolja (mind torzítás, mind szórás tekintetében). A nagyobb kiinduló minta közepes mértékben képes csökkenteni az RDS II minta szórását, míg az RDS I becslésre alig van hatással – meg kell azonban jegyezni, hogy szimulációnk során foksámarányos véletlen kiindulópontokat használtunk, amit a gyakorlatban nem lehet pontosan reprodukálni (bár az idézett szerzők, elsősorban Heckathorn érvelnek amellett, hogy a kezdőpontkeresésnél a nagyobb foksámú célcsoporttagok könnyebben fellelhetők, így kvázi megvalósulhat a foksámarányos kezdőpont-kiválasztás). Végül úgy tűnik, hogy legalábbis a jelen hálózati szimulációs modellben a nem kontrollált hálózati paraméterek nem játszottak jelentős szerepet a becslésekben – erre utalt az azonos paraméterekkel szimulált hálózatok közötti különbség alacsony (bár szignifikáns) magyarázóereje a becslésekre nézve.

Az első kutatási kérdésünk arra irányult, hogyan befolyásolják a hálózati és mintavételi tényezők a becslések torzítását és szórását, illetve mely paraméterkombinációk mellett melyik becslés (RDS I, RDS II) tekinthető hatékonyabbnak.

A heterofília, azaz a vizsgált változó mentén eltérő értéket választók közötti kapcsolatok aránya gyakorolja a legjelentősebb hatást mind a becslések szórására, mind torzítására, oly módon, hogy a kisebb heterofília nagyobb torzítással és a becslés nagyobb szórásával jár együtt. Ez az eredményünk megfelel az idézett korábbi kutatásoknak, azzal a pontosítással, hogy elemzéseink során a foksám-várhatóértéket és a kiinduló minta elemszámát is kontrolláltuk (azaz az összefüggések a két vizsgált paraméter minden értéke mellett igazak).

A heterofília mellett a foksám is szignifikánsan befolyásolta mind a becslések torzítását, mind szórásukat, azonban a hatás erőssége a heterofília hatásánál gyengébb volt, általában az alacsonyabb foksám-várhatóérték (kisebb sűrűségű hálózat) esetén a becslések valamivel kisebb torzításúak és szórásúak voltak (ez különösen az RDS II becslés esetén volt igaz). E paraméter hatására vonatkozó korábbi kutatási adatot nem találtunk.

A kiinduló minta nagysága és ezzel párhuzamosan az átlagos lánchossz nem gyakorolt jelentős és konzisztens befolyást sem a torzításra, sem a becslések szórására. Ez az eredményünk ismét megfelel a korábbi kutatások eredményeinek azzal a kiegészítéssel, hogy ebben az esetben is kontrolláltan történt a vizsgálat a többi paraméterre nézve.

A második kutatási kérdésünk a becslések megbízhatóságára vonatkozott, illetve ezzel összefüggésben arra, hogy az egyes paraméterek ezt mennyiben befolyásolják.

Kutatásunk igazi újdonsága, hogy az RDS I becslés esetén – amely az átlagos foksám és a heterofília becslésén alapul – kimutattuk, hogy a két becslés hibája különféle paraméterbeállítások mellett milyen mértékben járul hozzá a teljes becslés hibájához. Ezzel kapcsolatban azt állapítottuk meg, hogy elsősorban a heterofília változása határozza meg, melyik hálózati paraméter becslésének hibája gyakorolt nagyobb befolyást a becslés teljes hibájára. A nagyobb heterofília a heterofília-becslés hibájának jelentősen kisebb hatásával, míg a kisebb heterofília a heterofília-becslés jelentősen nagyobb hatásával jár együtt. Emellett a foksám-várhatóérték eltéréseinek hatása jóval csekélyebb: a magasabb foksám-várhatóérték a foksám átlagbecslésének a teljes becslés szórására gyakorolt valamivel kisebb hatásával jár együtt.

Végül, de nem utolsósorban az, hogy egy-egy paraméterbeállítással több hálózatot szimuláltunk és egy-egy hálózaton több mintavételt végeztünk, lehetővé tette azt is, hogy megvizsgáljuk az azonos paraméterű, de eltérő hálózatokból származó minták különbözőségét is. Ez alapján azt mondhatjuk, hogy a hálózat egyéb, nem kontrollált tulajdonságai bár szignifikáns hatást gyakoroltak modellünkben, de a hatás mértéke csekély volt, így lényegében az egyéb hálózati tényezők hatását elhanyagolhatjuk.

A szimulációs eredményeink több gyakorlati javaslatot is implikálnak. Amennyiben azt feltételezhetjük, hogy az általunk leginkább fontosnak tartott változók mentén klasztereződik a sokaság (azaz kicsi a heterofília), akkor érdemesebb nagyobb mintaméretet tervezni már a kutatásunk elején. Összességében sem az

RDS I, sem az RDS II becslés nem preferálható a másikkal szemben, ha azonban a mintavételi design több kiindulópontot tartalmaz és rövidebb láncokat, valamint a kiindulópontok esetén megközelíthető a fokszámarányos bekerülési valószínűség, akkor az RDS II becslés valószínűleg pontosabb. További fontos eredmény, hogy nem kell figyelembe venni a becslés pontossága szempontjából a vizsgált célcsoport kapcsolathálózatának sűrűségét. A fent leírtakon túl nincs nagy jelentősége (legalábbis a vizsgált tartományon belül és a használt mintavétel mellett) a kiindulópontok számának. Gyakorlati szempontból további fontos eredmény, hogy érdemes minden vizsgált változó esetén az elemzést megelőzően becslést tenni a változó értékei szerint képzett csoportok összekötöttségére, és amennyiben az összekötöttség alacsony, akkor szélesebb konfidenciaintervallumokkal érdemes számolni.

A gyakorlati javaslatokon túl az elemzés második részében leírt és az RDS I becslés szórásának összetételére vonatkozó eredmények alapján további kutatási lehetőséget látunk. Azt találtuk, hogy azokban az esetekben, amikor az RDS I becslés szórása magasabb, az ok jelentős részben a heterofília becslésének a hibája. Azt gondoljuk, hogy további kutatások irányulhatnak arra, hogy a heterofília becslését pontosítsuk és ezáltal az RDS I becslést hatékonyabbá tegyük.

ABSTRACT: In our paper we summarize the results of our simulation research connected to the method of respondent driven sampling (RDS). The goal of our simulation was to investigate problems related to the reliability of the method that were mentioned in recent literature. As a result we conclude that the main factors that decrease the reliability of RDS estimators are lower levels of heterophily and to a lesser extent the higher average degree of the network. We also discuss what is the share of the variance of expected value estimates that can be explained by the error of degree estimation. Besides the results of our simulation we summarise literature on the RDS methodology and the fields of possible usage of the method.

Irodalom

- Bernát A. (2006): A kapcsolathálózat jelentősége a migrációban – etnikai metszetben. In Némedi D. – Somlai P. – Szabari V. – Szikra D. (szerk.): *Kötő-jelek 2005*. ELTE Társadalomtudományi Kar Szociológia Doktori Iskola. Budapest. 123–146.
- Bozsonyi K. – Jelenfi G. – Kmetty Z. (2008): A magyarországi aktív banki ügynökök. V. HUNNET kapcsolathálózati konferencia, Budapest.
- Csanády M. – Kmetty Z. – Kucsera T. – Személyi L. – Tarján G. (2008): A magyar képzett migráció a rendszerváltás óta. *Magyar Tudomány*, 5: 603–616.
- Gile, K. J. – M. S. Handcock (2009): Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology*, 40(1): 285–327.
- Goela, S. – M. J. Salganik (2010): Assessing Respondent-Driven Sampling. *Proceedings of the National Academy of Sciences*, 107(15): 6743–6747.

- Goodman, L. A. (1961): Snowball Sampling. *Annals of Mathematical Statistics*, 32: 223–258.
- Hárs Á. (2009): *Magyarok az osztrák munkaerőpiacon: Ingázók, bevándorlók, munkaerő-migránsok?* KOPINT Konjunktúra Kutatási Alapítvány. Internet: <https://skydrive.live.com/?cid=20226f10b70b2c25&id=20226F10B70B2C25%21311>
- Hárs Á. – Simon D. (2009): Ingázók, bevándorlók, munkaerő-migránsok? Munkaerő mobilitása konferencia – Országos Foglalkoztatási Alapítvány, Budapest.
- Heckathorn, D. (1997): Respondent-Driven-Sampling. *Social Problems*, 44(2).
- Kapitány B. (2010): Mintavételi módszerek ritka populációk esetén. *Statisztikai Szemle*, 88(7–8): 739–754.
- Salganik, M. J. – D. D. Heckathorn (2004): Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34: 193–239.
- Simon D. (2005): Kapcsolathálózati mintavételi módszer a romák migrációs potenciáljának vizsgálatára: módszertani megfontolások. II. HUNNET kapcsolathálózati konferencia, Budapest.
- Simon D. (2012): Válaszadó-vezérelt mintavétel: ritka és rejtett csoportok kvantitatív vizsgálata. *Statisztikai Szemle*, 90(4): 249–275.
- Tomas, A. – K. J. Gile (2010): The Effect of Differential Recruitment. Non-response and Non-recruitment on Estimators for Respondent-Driven Sampling. <http://arxiv.org/abs/1012.4122v1>
- Verdery, A. M. – T. Mouw (2011): *Estimated Sampling Variance in Respondent Driven Sampling: Do You Know If You Have an Imprecise Sample?* Working Paper.
- Volz, E. – D. D. Heckathorn (2008): Probability Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics*, 24(1): 79–97.