Dusek Tamás

Beszámoló a „Bűn-e a reprezentativitás hiánya mintavétel esetén?" című rendezvényről

Report on the event titled "Is the lack of representativity in sampling a sin?"

DUSEK TAMÁS, a Széchenyi István Egyetem tanszékvezető egyetemi tanára
E-mail: dusekt@sze.hu
a *Statisztikai Szemle* főszerkesztője
E-mail: Tamas.Dusek@ksh.hu

Az MTA Statisztikai és Jövőkutatási Tudományos Bizottsága Statisztikai Tudományos Albizottsága „Bűn-e a reprezentativitás hiánya mintavétel esetén?" címmel tudományos ülést tartott 2019. szeptember 23-án. A rendezvény helyszíne a KSH Keleti Károly-terme volt, formája pedig moderált kerekasztal-beszélgetés, amelyben az első kört követően bármely jelenlévő aktívan részt vehetett kérdéssel, hozzászólással. A rendezvény szervezői *Kovács Péter*, *Szép Katalin* és *Vereczkei Zoltán* voltak, moderátora Szép Katalin, a kerekasztal felkért résztvevői pedig *Bolla Marianna* (Budapesti Műszaki és Gazdaságtudományi Egyetem)*, Dusek Tamás* (*Statisztikai Szemle*), *Fraller Gergely* (Központi Statisztikai Hivatal), *Kerékgyártó Györgyné* (Budapesti Corvinus Egyetem), *Rudas Tamás* (Eötvös Loránd Tudományegyetem), *Simon Judit* (Budapesti Corvinus Egyetem) és *Szilágyi Roland* (Miskolci Egyetem).

Az esemény beszámolóját a következőkben három részre tagolom. Először az előzményekről írok, majd a rendezvényen elhangzottakból emelek ki témákat, végül néhány további gondolatot fogalmazok meg a kérdéskörről és annak kapcsolódásairól, csupán röviden felvetve néhány kérdést, amelyek ugyanúgy részletesebben és szisztematikusabban tárgyalhatók lennének, mint az addigiak. Ez remélhetőleg megtörténik a közeljövőben, amelyhez a *Statisztikai Szemle* is egy lehetséges fórumot nyújt, akár hosszabb módszertani, történelmi vagy egyéb tanulmányok, akár rövidebb hozzászólások, állásfoglalások formáját öltő írások révén. Az MTA Statisztikai és Jövőkutatási Tudományos Bizottsága honlapján (http://www.ksh.hu/mta_sjtb) olvashatók lesznek majd az üléssel kapcsolatos anyagok (meghívó, emlékeztető).

Előzmények

A rendkívül tág témakörrel kapcsolatos problémák ott kezdődnek, hogy a reprezentativitásnak – széles körű használata ellenére – nincs egységes és pontos definíciója. A kerekasztal résztvevői előzetesen témakörön belüli fókuszálást segítő bevezető gondolatokat tartalmazó Power Point prezentációt és kérdéseket kaptak Szép Katalintól. A reprezentatív minta kifejezés angol használatáról (representative sampling) *William Kruskal* és *Frederick Mosteller* az 1970-es évek végén négyrészes tanulmánysorozatot írt. A kérdés komplexitásának bemutatása mellett céljuk volt annak dokumentálása is, hogy még a statisztikai szakirodalom is gyakran lazán, gondatlanul, nem kellően precízen használ egy fontos statisztikai koncepciót. Az *International Statistical Review*-ban megjelent cikksorozat első része a nem tudományos irodalomban történő használatot, második része a tudományos, de nem statisztikai szakirodalmi használatot, harmadik része az aktuális statisztikai szakirodalmi használatot (ami alatt a szerzők az 1940-es évektől a cikkük megírásáig terjedő időszakot értik) mutatta be. A negyedik rész a reprezentatív minta 1895 és 1939 közötti történetével foglalkozott, de a szerzők történeti utalásai egészen *Ciceró*ig nyúlnak vissza (*Kruskal–Mosteller* [1979a], [1979b], [1979c], [1980]). Ebből a leírásból nem egyértelmű, de a szerzők elsősorban a kérdőíves lakossági felmérésekre gondoltak, általánosságban nem említve például a szervezeti, intézményi felmérések, a környezeti felmérések és a kísérleti kutatások sajátos kérdéseit, egyes eseteknél viszont mégis környezeti felmérésekre és kísérleti elrendezésekre is hivatkoztak.

A bevezető prezentáció második diája, emlékeztetve a sokszínű értelmezésre, Kruskal és Mosteller tanulmánysorozatának harmadik részében található összegzést tartalmazta, amely a reprezentatív minta használatának kilenc (egymást részben kizáró, de olykor átfedő) jelentését különböztette meg. A szerzők a reprezentatív mintához tartozó eszme történetét dolgozták fel, és nem pusztán a szó etimológiájával foglalkoztak. A következőkben ezt a listát rövid kiegészítő magyarázatokkal ellátva közlöm:

> *1*. Az adatok jóságának általános, de meg nem erősített elismerése. Ekkor a jól hangzó reprezentatív jelzőt retorikai fogásként használják, az olvasók bizonyítás nélkül higgyék el a szerzőnek, hogy a minta reprezentatív.
>
> *2*. Speciális kiválasztási szempont hiánya. A kiválasztási szempont jelenléte torzítja a mintát, és nem teszi általánosíthatóvá az eredményeket.
>
> *3*. A sokaság tükre vagy miniatűr képmása: a minta egyes jellemzők (például kor, nem, családi állapot, végzettség, lakóhely, etnikum) szerinti megoszlása hasonló az alapsokaságéhoz. Ahogyan a figyelem-

be vett jellemzők száma és részletezettsége növekszik, úgy ennek elérése a gyakorlatban egyre kevésbé lesz lehetséges. Például 2 nem, 3 korosztály, 4 családi állapot és 4 végzettség figyelembevétele összesen 2 * 3 * 4 * 4 = 96 részsokaságot eredményez, amelyek többsége nem üres, és nem homogén, vagyis a részsokaságokon belüli eltérésekkel is számolni kell. A végeredményt befolyásoló lehetséges változók és kategóriák száma pedig az egyszerűbb társadalmi felmérésekben is jóval nagyobb ennél.

4. Tipikus vagy ideális eset. A mintába kerülő elemeket nem véletlenül választják, hanem tudatosan próbálnak olyan elemeket választani, amelyek tipikusak az egész sokaságra nézve. Például politikai közvéleménykutatásnál olyan területi egységet választanak, amelynél a választás végeredménye korábban megegyezett az országos átlaggal. Az egyik probléma ezzel kapcsolatban megegyezik az előző esetnél tárgyalttal: egyik vagy néhány szempontból tipikus egyedek nagyon sok más fontos szempontból egyáltalán nem tipikusak lehetnek.

5. A sokaság változatosságának (rétegeinek) lefedése. Ezt az igényt egyes szerzők sokféleképpen részletezhetik, de az előző használatnak, a reprezentativitás mint tipikus esetnek mindenképpen ellentmondóan, mivel a mintaelemeket a lehető legtöbb részsokaságból kívánják meríteni.

6. Általános és kezdetben bizonytalan jelentésű kifejezés, amelyet majd később pontosítanak.

Ezt a hat használatot a szerzők már az első két tanulmányukban is tárgyalták. Ehhez vesznek hozzá még három további használatot a statisztikai szóhasználatot bemutató tanulmányban:

7. Reprezentatív mintavétel mint speciális mintavételi módszer. A legtöbb szerző a reprezentatív mintavételt véletlen vagy valószínűségi mintavételként határozza meg, különböző precizitású definíciókkal.

8. Reprezentatív mintavétel, amely jó becslést tesz lehetővé. Ez az előző meghatározással majdnem megegyezik, csak inkább a becslési eljárás következményét emeli ki és nem a módszert.

9. Reprezentatív mintavétel, amely elég jó egy adott célra. Ez lehet csupán az, hogy a minta ne legyen nagyon félrevezető. Vagy például annak kiderítésére, hogy léteznek-e egy sokaságon belül bizonyos tu-

> lajdonságú egyedek, vagy sem, megfelelő egy olyan mintavétel, amelyben a keresett tulajdonságú elemek találhatók. Ekkor lehet, hogy az adott elemek arányára nem lehet következtetni, de a létezésükre igen. A reprezentatív minta ilyen értelmű használatával inkább csak alkalmazott tanulmányokban lehet találkozni, elméleti, módszertani tanulmányokban nem.

Ez a jelentésbeli sokféleség, amelyet a szerzők minden esetben számos példával illusztrálnak, a statisztikai és általában a tudományos szakterminológiában nem egyedi jelenség. Épp ellenkezőleg, tipikusnak mondható, hogy a gyakran használt fogalmak sok jelentésárnyalatot vesznek fel, amelyek egymástól nagyon eltérők lehetnek. A túlzottan terhelt terminusokat a tudományos munkákban célszerű kerülni, vagy első használatukkor egyértelműen definiálni és a szövegen belül azonos jelentésben, következetesen használni. A reprezentatív kifejezés annyira meggyökeresedett a statisztikában, olyan széles körben használt (így például a *Statisztikai Szemle* archívuma alapján a folyóiratban 2019. szeptemberig összesen 2491 cikkben fordult elő a reprezentatív kifejezés), hogy száműzése helyett inkább az egyértelműsítése, fogalmi tisztítása a járható út, amit az ismertetett rendezvény is szolgál.

Szép Katalin arra kérte a résztvevőket, hogy a reprezentatívnak szánt mintákkal foglalkozzanak, tehát azokkal az esetekkel, amikor a kutató azzal a céllal vesz mintát, hogy az alapsokaságra következtessen, és egyéb részleges megfigyelésekről, kísérletekről ne beszéljenek. A bevezető prezentáció harmadik diája egyes mintavételi eljárásokat sorolt fel. A következő dia arra hívta fel a figyelmet, hogy a mintaválasztás módja még nem jelenthet garanciát a reprezentativitásra. A tényleges vizsgálatokat a vizsgálatra előkészített, megvalósult mintán végezzük. Ez a minta torzulhat a keret lefedettségi hibái és a nemválaszolás miatt.

Az ötödik dia a résztvevőknek feltett kérdéseket sorolta fel, amelyek a következők voltak:

- „Milyen értelemben használják a reprezentatív minta kifejezést?
    - Hogyan biztosítható?
    - Hogyan ellenőrizhető?
    - Mi szerepel ebből az oktatásban ill. milyen mintákkal dolgoznak a gyakorlatban?
- Mintavételből származó adatállomány minőségét (a vizsgálat célja, módszerek alkalmazhatósága szempontjából) hogyan jellemzik?
    - Mi a javasolt eljárás?
        - Keresnek olyan módszert, ami az állomány sajátosságait is kezeli, vagy azokra nem érzékeny (adathiány, függetlenség hiánya, stb.)
        - Bár a feltételek nem teljesülnek, de adhat hasznos információt
        - Közvetlenül csak a mintára fogalmaznak meg állításokat.
- Bűn-e a reprezentativitás hiánya?” (*Szép* [2019])

A rendezvény

A rendezvényt megnyitó Kovács Péter, az MTA Statisztikai és Jövőkutatási Tudományos Bizottsága Statisztikai Albizottsága elnöke, a Szegedi Tudományegyetem Gazdaságtudományi Karának dékánja a témaválasztás indokaként a reprezentatív minta gyakori és nem egyértelmű használatát nevezte meg. Disszertációkban, doktori értekezésekben, tanulmányokban, előadásokban gyakran használják a kifejezést nem pontosan meghatározott értelemben. Az adatgyűjtés körülményei sokszor nem jól dokumentáltak, így a reprezentativitást nem lehet megítélni. Ezért az elméleti, gyakorlati és oktatói nézőpontok több képviselőjét kérték fel a beszélgetésre.

Szép Katalin a már említett prezentációnak megfelelően vázolta a főbb kérdéseket. A statisztikai adatfelvétel a tudomány és művészet határterületén található olyan értelemben, hogy az elméleti előírásokat és követelményeket a gyakorlati nehézségek mellett kell a lehető legjobban megvalósítani, úgy, hogy az adatok minőségére vonatkozó szempontok különbözők lehetnek. A művészi elemre később több hozzászóló is utalt, miként azzal is többen foglalkoztak, hogy a reprezentatív minta kifejezés használata nem egységes, még szakterületeken belül sem. Az elméleti meghatározásokban leggyakrabban a véletlen minta, a valószínűségi minta, a megfigyelési egységek ismert valószínűségű mintába kerülése kifejezések szerepelnek, ami következményként maga után vonja, hogy a megvalósuló minta néhány szempontból az alapsokasághoz hasonló arányban tartalmaz mintavételi elemeket. Ez azonban csak akkor érvényes, ha megfelelő a mintavételi keret és a válaszadási hajlandóság, amely kiegészítés sokszor elmarad.

A reprezentativitás fogalmának népszerűvé válása kapcsán Rudas Tamás egyetemi tanár, az MTA doktora az ismert közvélemény-kutatási kudarcot, az 1936-ban rosszul előre jelzett amerikai elnökválasztási eredményt említette. Ennek és az ehhez hasonló kudarcoknak az oka az, hogy a minta nem reprezentatív; tehát a sikerhez a mintának reprezentatívnak kell lennie. De amíg a negatív meghatározás egyszerűbb (mi nem reprezentatív), addig a reprezentativitás pozitív meghatározása már jóval bonyolultabb, sokszínűbb és ellentmondásosabb. A minta önmagában való reprezentativitása jelentés nélküli, nem ellenőrizhető, ugyanakkor bizonyos megnevezett szempontokból lehet beszélni a minta reprezentativitásáról. A lehetséges szempontok száma kicsi, és azokra a jellemzőkre korlátozódnak, amelyek alapsokasági eloszlása ismert. Rudas Tamás foglalkozott azzal is, hogyha a mintabeli arányok eltérnek a főbb szociodemográfiai jellemzők népszámlálásokból ismert arányaitól, akkor az alapsokasági arányoknak megfelelő súlyozással korrigált eredményeket számítanak. Alapvető demográfiai szempontokból tehát reprezentatívvá lehet tenni a mintát, de ez nem garancia arra, hogy a minta a kutatás szempontjából fontos jellemzők, például a pártpreferenciák szerint is reprezentatív lesz. A választási előrejelzések

világtörténete azt mutatja, hogy 15 évvel ezelőttig a demográfiai szempontból reprezentatívvá tett minták alapján jó előrejelzéseket lehetett tenni a választási eredményekre, az elmúlt 15 évben azonban sok látványos előrejelzési kudarc történt.

Simon Judit egyetemi tanár marketingkutatási kurzusok oktatójaként és marketingkutatói gyakorlati alkalmazóként beszélt arról, hogy a piackutatások során nincs ismert alapsokaság és mintavételi keret (nincs nyilvántartás arról, hogy kik bizonyos termékek vásárlói, használói), és így nagyon ritka a termékek elterjedtségére vonatkozó véletlen mintavételes penetrációkutatás. A piackutatások során az adatfelvételnél a főbb demográfiai ismérvek szerint reprezentatívitásra lehet törekedni, de magára a vizsgált jellemzőre vonatkozóan, az alapsokasági eloszlás ismeretének hiányában ez nem lehetséges.

Fraller Gergely rámutatott arra, hogy a lakossági felvételek során a gyakorlatban nem létezik matematikai értelemben valószínűségi minta. A lakcímnyilvántartás is minden bizonnyal tartalmaz hibákat, nincs teljes lefedettséget biztosító mintavételi keret. A nagyobb problémák pedig a megvalósult minták kapcsán lépnek fel, mert a kiválasztott mintához képest egyre nagyobb arányú meghiúsulás jellemző. Fraller Gergely a mintavételi kerettel nem rendelkező felvételek közül a turizmussal kapcsolatos adatfelvételeket említette, amikor csak nem valószínűségi minta kiválasztására van mód.

Ami a diákok által végzett adatgyűjtéseken alapuló szakdolgozati elemzéseket illeti, nem volt teljes egyetértés a résztvevők között abban, hogy ezek esetén elvárható-e a valószínűségi mintavétel értelemben reprezentatív mintavétel végzése és azon alapuló elemzés. A többségi álláspont szerint (amivel magam is egyetértek) a diákok nincsenek abban a helyzetben, hogy végre tudjanak hajtani véletlen mintavételt, nemcsak országos, de lakóhelyi szintű elemzések esetén sem. Fraller Gergely szerint többször kértek már segítséget a Központi Statisztikai Hivatal Módszertani főosztályától országos vagy helyi véletlen mintavétel lebonyolítására szakdolgozatot írók, de az sohasem valósult meg, mert a szakdolgozat írói nincsenek abban a helyzetben, hogy azt kivitelezni tudják. Esetükben nem reális elvárás, hogy reprezentatív mintával dolgozzanak, megengedhető az önkényes mintavétel, de az nem, hogy megfogalmazásuk az eredményeik általánosítását sugallaná. Ami elvárható tőlük, az az alkalmazott statisztikai elemzési módszerek korrekt használata.

Ez még kiegészíthető két további szemponttal. Egyrészt, a diákok sokszor nem általános, a teljes lakosság esetében értelmezhető jelenséget elemeznek, hanem egy nehezen meghatározható, lehatárolható, beazonosítható, mintavételi kerettel nem megfogható részsokaságban értelmezhető jelenséget, amelynél valószínűségi mintavétel eleve nem lehetséges. Másrészt, nemcsak diákok, hanem professzionális kutatók sem tudnak egyedül kivitelezni sokmilliós vagy tízmilliós költségvetésű nagy lekérdezéseket.

Arra a kérdésre nem született válasz, hogy a reprezentativitás miként biztosítható a lefedettségi, adathiánybeli és nemválaszolási problémák mellett. A reprezentativitás kérdéseinek körbejárása több ülés témája lesz. Első körben a fókusz a jelen helyzet, fogalomhasználat és gyakorlat áttekintése volt, mely alapján az Albizottság egy ajánlást kíván megfogalmazni a kutatók számára. Ezután kerül sor a jövő kihívásainak taglalására, így *Hunyadi László* által feltett azon kérdésre is egy következő alkalomra halasztódott el a válaszadás, hogy a big data típusú adatforrásokkal kapcsolatban hogyan értelmezhető a reprezentativitás. Simon Judit szerint az internetes adatfelvétel melletti fő érvként az internet növekvő elterjedtségét szokták felhozni, különösen a piackutatók számára elsődlegesen fontos korcsoportban; így a piackutatási céloknak elfogadható minőségű eredményeket szolgáltatnak az ilyen felmérések.

Mindezek mellett számos érdekes további részletkérdés merült fel a kerekasztal résztvevői és a hozzászólók részéről, amelyek mind a témakör gazdagságáról tanúskodtak. *Katona Tamás* egyetemi tanár azt a gyakorlatot bírálta, amely szerint a részminták hibahatárát nem közlik, illetve a részminták kis elemszáma miatt azokra már nem is lehet megbízható becslést adni akkor sem, ha maga a teljes minta reprezentatív. *Telegdi László* arra hívta fel a figyelmet, hogy az angol sampling survey kifejezést évtizedekig reprezentatív megfigyelésnek fordították. Ebben az esetben nem valószínűségi mintáról volt szó, hanem akármilyen mintáról, amelyek lehetnek valószínűségiek is.

További gondolatok

A tárgyalt kérdések az adatok minőségét, a statisztikai módszerek alkalmazását és az eredmények értelmezését érintik, nem a statisztikai módszertan matematikai szerkezetére vonatkoznak. A rendelkezésre álló adatok keletkezési körülménye és minősége befolyásolja azok felhasználhatóságát. Ehhez hasonló gyakorlati, adatfelvételi, minőségi, értelmezésbeli kérdésekkel szinte bármely módszer kapcsán találkozhatunk, a valószínűségszámítás alkalmazása pedig különösen sok ilyen példát eredményezett. A társadalmi-gazdasági életben többnyire nem véletlen jelenségeket vizsgálnak, hanem egymással komplex módon összefüggő jelenségeket; egyik ember véleménye, jövedelme, viselkedése részben függ a másik ember és az összes többi ember véleményétől, jövedelmétől, viselkedésétől. Így itt nem magának a jelenségnek a véletlen volta indokolja a valószínűségszámítás alkalmazását, hanem az, hogy a véletlen a minta kiválasztási módja révén (amennyiben lehetséges mesterségesen véletlen mechanizmussal a mintaelemek kiválasztása) megjelenik a rendszerben, és így az egyes jellemzők, mutatók véletlen mintavételi ingadozása a valószínűségszámítás segítségével meghatározható. Cenzus típusú (a sokaság összes elemét megfigyelni szán-

dékozó) felméréseknél ez a kérdés nem vetődik fel, ott egyéb típusú adatfelvételi hibák jelentkeznek. A statisztikai szokásjog által elterjedt gyakorlat alapján ugyanakkor bármely sokaságot kezelhetnek véletlen mintavétel eredményeként létrejöttnek, így például térben, időben, minőségben és viselkedésben heterogén elemek aggregálásával keletkező makroökonómiai mutatószámokat is, amelyekkel végzett műveletekre vonatkozóan a statisztikai szoftverek automatikusan megadják a véletlen mintavételt feltételező, különböző következtető statisztikai eljárások eredményeit.

A véletlen és nem véletlen minta fogalmi szétválasztása egyszerű, tartalmilag viszont inkább egy kontinuumról van szó, amelynek egyik végpontjában az idealizált matematikai modellek véletlen mintája található, majd a gyakorlatilag véletlen mintaként kezelhető valós eseteken át az egyre kevésbé véletlen, egyre növekvő és eltérő jellegű tökéletlenségekkel terheltek következnek, a főbb jellemzők alapján a teljesen hibátlan mintavételi kerettől a sok hibát tartalmazó mintavételi keretig, a 100 százalékos válaszadási hajlandóságtól a nagymértékű válaszmeghiúsulásig. Ezek a tökéletlenségek egy precíz és korrekt eljárás során jól dokumentáltak, így a minta minősége (elvileg) bárki számára megítélhető.

A reprezentatív minta kifejezést a valószínűségi minta szinonimájaként használva a terminológia felesleges megkettőződéséről van szó, arról, hogy a valószínűségi mintára egy másik terminust is alkalmaznak. Ez önmagában nem zavaró, csak szükségtelen. Ezenkívül azonban sok más esetben és értelemben is használatos a reprezentatív minta kifejezés, amelyről külön empirikus tanulmányt lehetne írni, a magyar szóhasználatot akár az angollal is összehasonlítva.

Irodalom

KRUSKAL, W. – MOSTELLER, F. [1979a]: Representative sampling, I: non-scientific literature. *International Statistical Review*. Vol. 47. No. 1. pp. 13–24. https://doi.org/10.2307/1403202

KRUSKAL, W. – MOSTELLER, F. [1979b]: Representative sampling, II: scientific literature, excluding statistics. *International Statistical Review*. Vol. 47. No. 2. pp. 111–127. https://doi.org/10.2307/1402564

KRUSKAL, W. – MOSTELLER, F. [1979c]: Representative sampling, III: the current statistical literature. *International Statistical Review*. Vol. 47. No. 3. pp. 245–265. https://doi.org/10.2307/1402647

KRUSKAL, W. – MOSTELLER, F. [1980]: Representative sampling, IV: the history of the concept in statistics, 1895–1939. *International Statistical Review*. Vol. 48. No. 2. pp. 169–195. https://doi.org/10.2307/1403151

SZÉP K. [2019]: „Bűn-e a reprezentativitás hiánya mintavétel esetén?" című kerekasztal-beszélgetés (Budapest, 2019. szeptember 23.) diái. http://www.ksh.hu/stab_dokumentumok