Szeitl Blanka[1] – Fellner Zita[2]

Kis hiányból nagy hiba[3]

Internetes kérdőíves adatgyűjtésekből származó becslések torzítása a magyar lakosság jellemzői alapján

https://doi.org/10.18030/socio.hu.2023.3.1

Absztrakt

A társadalomtudományi kutatások során egyre elterjedtebbek az online kérdőíves adatgyűjtések *(survey-k).* Ez a folyamat leginkább a lakosság emelkedő internet-ellátottságának köszönhető, de fontos szerepet játszik az is, hogy egyre nehezebben és drágábban kivitelezhetőek a személyes és a telefonos adatgyűjtésen alapuló kutatások. Online kérdőíves adatgyűjtés esetén szinte megvalósíthatatlan a valószínűségi minta, emiatt az eredmények általánosíthatósága sérülhet, a becslések pedig torzítottak lehetnek. A tanulmány azt vizsgálja, hogy milyen mértékben és milyen szempontok szerint torzítanak az online adatgyűjtésekből származó becslések. Szimuláció segítségével modellezzük az online kérdőíves adatgyűjtés során felmerülő elméleti korlátokat. Valós surveyadatok segítségével pedig bemutatjuk, hogy milyen problémákra kell számítania a kutatónak online kérdőíves adatgyűjtések használatakor. A vizsgálat a magyar lakosság demográfiai és internetpenetrációs jellemzőin alapul, így eredményeink kizárólag a magyar lakosságra vonatkoztatott kutatásokra érvényesek. A szimuláció peremszámait a Központi Statisztikai Hivatal adminisztratív adatai alapján állítottuk össze, a valós surveyadatok pedig a *European Social Survey* 9. magyarországi adatgyűjtési hullámából származnak. A tanulmány általános megállapítása az, hogy habár a magas internetes ellátottság látszólag kézenfekvő adatgyűjtési lehetőséget nyújt a társadalomkutatás számára, az eredmények pontossága és megbízhatósága szempontjából egyelőre nem alkalmas eszköz a teljes lakosság vizsgálatára. Szimulációs eredményeink alapján azt találtuk, hogy az online kérdőíves felmérések adatai az esetek 67%-ában téves becslésekhez vezethetnek akkor, amikor a kapott eredményeket a teljes magyar populációra vonatkoztatjuk. A valós adatokon végzett becslés online kérdőíves adatgyűjtés esetén az olyan gyakran vizsgált kérdések mentén is jelentősen alul- vagy felülbecsüljük a vélhetően valós populációs arányokat, mint a pártokhoz való kötődés, vallásosság, egészségi és családi állapot.

Kulcsszavak: empirikus társadalomkutatás, internetes kérdőíves adatgyűjtés, becslés

1 ELTE Társadalomtudományi Kar, Statisztika Tanszék, Survey Methods Room Budapest; HUN-REN Társadalomtudományi Kutatóközpont.

2 ELTE Társadalomtudományi Kar, Statisztika Tanszék, Survey Methods Room Budapest; Magyar Nemzeti Bank.

3 Készült az RRF-2.3.1-21-2022-00013 azonosítószámú „Társadalmi Innovációs Nemzeti Laboratórium” elnevezésű projektben, Magyarország Helyreállítási és Ellenállóképességi Tervének keretében, az Európai Unió Helyreállítási és Ellenállóképességi Eszközének támogatásával.

From minor data gaps to major errors

Simulation study to demonstrate potential bias of online surveys

Abstract

Online data collection is one of the new directions for surveys. With online-only data collection, a probability sample is almost infeasible, which may compromise the generalisability of the results and bias the estimates. In this paper the authors present a case study of Hungary, where internet access is far from reaching full penetration (80 percent) but considered average by European standards. The study focuses on two main points: (1) the extent to which estimates from online surveys are biased in general, and (2) the socio-demographic and attitudinal aspects relevant to the magnitude of the bias. The main method of analysis is simulation, which is based on multiple data sources. Based on administrative data the demographic composition is modeled for both offline and online populations, while for the attitude dimensions face-to-face survey data of the European Social Survey is used. The study evaluates estimates from simulated online and face-to-face data collections involving several post-stratification processes. The overall conclusion of the study is that although online data collection seems to be a convenient data collection tool for social research, given the relatively high internet penetration, it is not yet a suitable method. The study found that even a minor data gap from the offline population leads to major error in the estimates: based on the characteristics of internet penetration in Hungary, in 67 percent of the cases erroneous estimates were obtained. For relevant research dimensions such as interest in politics, religiosity, health and marital status, the online data collection significantly under- or overestimates the likely real population proportions.

Keywords: empirical social research, online survey, bias

Szeitl Blanka – Fellner Zita

Kis hiányból nagy hiba

Internetes kérdőíves adatgyűjtésekből származó becslések torzítása a magyar lakosság jellemzői alapján

Bevezetés

Az önálló, mintavételen alapuló társadalomtudományi adatgyűjtések Magyarországon az 1970-es években kezdődtek, és ekkor kezdett el a Központi Statisztikai Hivatal is foglalkozni a társadalmi jelzőszámokkal, illetve a nemzetközi adatgyűjtési trendek hazai környezetbe történő átültetésével (Andorka et al. 1990). Az 1990-es években már kifejezetten nagy igény volt a társadalmat a tények, azaz az adatok nyelvén leírni. Ebben az időszakban javarészt személyes adatgyűjtési technikákat használtak, azaz az adatgyűjtő vállalkozás országos kérdezői hálózatot üzemeltetett, a kutatás során pedig a kérdezők személyesen keresték fel a potenciális válaszadókat. Habár a technikai apparátus a maihoz képest jóval fejletlenebb volt, az adatgyűjtések jellemzően rövid idő alatt megvalósíthatóak voltak az átlagos 1000–3000 fő közötti mintanagysággal. Ennek legfőbb oka a magas válaszadási arány volt: egészen a 2000-es évek közepéig a felkeresett személyek közel 70%-a adott választ a különböző vizsgálatok kérdőíveire. A válaszadási arány ezután folyamatosan és drasztikusan zuhant (Havasi 1997, Szeitl–Tóth 2020), 2020-ban pedig átlagosan a kijelölt személyek legfeljebb 40%-ára lehetett számítani a válaszadói bázisban. A személyes kutatások helyettesítésére már a 2000-es évek elején is gyakoriak voltak a telefonos kutatások, melyek egy ideig valóban magasabb válaszadási arányt biztosítottak, de a lefedettségi és mérési hibák miatt nem tekinthetőek a személyes felmérések valódi alternatíváinak (Kmetty 2012).

A válaszadási arány csökkenése a világ legnagyobb részén megfigyelhető, és magyarázatára számos elmélet létezik: egyesek szerint a lakosság „túlkérdezésének" eredménye a magas visszautasítás (Meyer et al. 2015), mások természetes attitűdváltozást látnak a folyamat mögött, amely a bizalmatlanság és a bezárkózás eredménye lehet (Davern 2013). Ezzel párhuzamosan technológiai oldalról is változás történt, hiszen az internetpenetráció emelkedése új lehetőségeket nyújt az adatgyűjtési piacnak: míg 2006-ban a magyar lakosság alig 30%-a rendelkezett internettel, 2014-ben már 65%, 2021-ben pedig közel 90% (KSH 2021). Ez a survey típusú adatgyűjtések szempontjából azt jelenti, hogy online sokkal gyorsabban, olcsóbban és helyenként nagyobb hatékonysággal összegyűjthetőek a kitöltött kérdőívek, mint személyes megkeresés által.

Habár a válaszadási arányok csökkenése és az internetpenetráció növekedése két különböző folyamatnak tűnhet, egy dolog közös a két jelenségben: nem általánosan és nem ugyanolyan mértékben figyelhetőek meg a különböző társadalmi csoportokban. Annak a valószínűsége, hogy egy kiválasztott személy válaszol-e egy személyes felmérésre, a válaszadó neme és korcsoportja szempontjából is jelentős különbséget mutat, és további releváns mintázat figyelhető meg például a lakóhely szerinti bontásban is (Szeitl–Tóth 2021). Ugyanez a helyzet az internetpenetráció esetében megjelenő mintázattal is: bizonyos társadalmi csoportok esetében az internetes ellátottság szinte teljes körű, míg más csoportok esetében alig éri el az 50%-ot (KSH

2021). A válaszadási jellemzők a véletlen mintás, személyes felmérések esetében is módszertani problémát jelentenek, de míg ezek komoly matematikai-statisztikai háttérrel rendelkeznek, és ellenőrzött módszerekkel képesek korrigálni a terepmunka során generált eltéréseket, az online felmérések esetében a kiinduló populáció pontos definiálása is kérdéses (internettel rendelkezők/bizonyos gyakorisággal internetet használók), valamint nem áll rendelkezésre lista a populáció elemeiről, így a véletlen mintavétel sem valósítható meg az esetek többségében. Így az online felmérések szerkezete nem teljesíti az olyan alapvető követelményeket sem, melyek alapján a becslések megbízhatósága egyáltalán közelítően meghatározható lenne. A személyes és az online adatgyűjtési módok nem kizárólag módszertani, technikai szempontok szerint, hanem tartalmi vonatkozásban is jelentős különbségeket mutatnak: az online adatgyűjtések nélkülözik a személyes interjúszituáció minőségi elemeit, illetve a lehetséges mintalefedés és elvi elérés kritériumait, melyek a közvélemény teljeskörű megismerését alapvetően lehetővé tehetik (Angelusz–Tardos 2009). A gyorsabb, olcsóbb és egyszerűbb megvalósíthatóság miatt viszont egyre gyakrabban találkozhatunk társadalomtudományi vizsgálatok esetében is online adatgyűjtésekkel.

Tanulmányunk azt vizsgálja, hogy az egyre elterjedtebb internetes kérdőíves adatgyűjtés milyen mértékben eredményez torzított becsléseket, és ezáltal hogyan vezethet hibás következtetésekhez. Online adatgyűjtést kétféleképpen alkalmazhatunk: vagy válaszadási módként, azaz a mintába került személyeknek lehetőséget adunk az online válaszadásra, vagy úgy, hogy csak online módon van lehetőség válaszadásra (például online paneleken keresztül vagy online hirdetett kérdőív használatával). Az előbbi típusú adatfelvételek jellemzően megvalósíthatóak valószínűségi mintavételként, míg az utóbbiak általában nem valószínűségi mintavételnek számítanak (Baker et al. 2013). Tanulmányunkban az online kérdőíves adatgyűjtések alatt a teljes lakosságra nézve nem valószínűségi mintavételen alapuló módszert értjük.

A következőkben a magyar lakosság ismérvei alapján végezzük el az online minták kiértékelését elméleti és gyakorlati szinten. A tanulmány szerkezete a következő: a 2. részben ismertetjük a kapcsolódó matematikai-statisztikai alapokat; bemutatjuk, hogy az online adatgyűjtések hogyan jellemezhetőek a klasszikus minőségi kritériumok és értékelési keretrendszerek mentén. Ezután a 3. részben ismertetjük a hazai internetpenetráció trendjeit és aktuális jellemzőit a KSH és az Eurostat statisztikái alapján. Ezek az adatok nemcsak az internetes kérdőíves adatgyűjtések lehetőségeit mutatják be, hanem kiindulópontjai a 4. részben bemutatott szimulációs munkának is. A jelenlegi internetes ellátottság mellett kapott becslések potenciális torzításait szimulált adatok elemzésével modellezzük, és bemutatjuk, hogy a teljes népesség és az internethasználó népesség közötti kismértékű eltérés is elég ahhoz, hogy szignifikánsan különböző becsléseket kapjunk. Ezután valós adatokból származó becsléseket mutatunk be az 5. részben. A becsléseket a *European Social Survey* (ESS) 9. hullámának magyarországi adatai alapján számoljuk a teljes és az internethasználó népességre, amelyek különbségei alapján látható, hogy a vizsgált változók mindegyike esetében jelentősen eltérő eredményt kapunk az online adatgyűjtés esetében ahhoz képest, mintha a teljes népességből vennénk mintát.

Klasszikus minőségi kritériumok az online surveykutatások esetében

A surveykutatások két alapvető minőségi kritériuma az érvényesség és a megbízhatóság. Az érvényesség a kutatási koncepció megfelelőségére vonatkozik: az adatfelvétel alapsokasága megegyezik azzal az alapsokasággal, amelyre a kutatási kérdés vonatkozik. Ezt úgy definiáljuk, hogy a kutatási kérdés alanyainak van-e esélye (azaz nullától különböző valószínűsége) bekerülni a mintába (Marshden et al. 2010). Amikor az adatfelvételi és a kutatási populáció nem azonos, és ennek ellenére az adatfelvétel alapján a kutatott sokaságra vonatkozó megállapításokat teszünk, akkor kutatásmódszertani szempontból a kutatás nem érvényes. Ez a helyzet áll elő akkor, amikor a teljes népességre vonatkozó eredményként közöljük az online elérhető, internethasználó népesség megfigyeléséből származó statisztikai következtetéseket. Statisztikai szempontból ez azt eredményezi, hogy a becslés (becslőfüggvény) torzított lesz, vagyis a becslés várható értéke nem egyezik a becsülendő sokasági paraméterrel. A statisztikai becslésekben megjelenő torzítás nem „végzetes hiba", a torzítás mértékének ismeretében ezt figyelembe lehet venni az eredmények interpretációjában. Az érvényesség akkor válik igazán problematikussá, ha erre a torzításra nem fektetünk hangsúlyt.

A torzítás mértékét egzakt módon leírhatjuk. Amennyiben online adatgyűjtéssel tervezzük a kutatásunkat, a megfigyelést értelemszerűen nem tudjuk végrehajtani a teljes népességből választott mintán, hiszen lesznek olyanok, akik internethozzáférés, megfelelő eszköz vagy a válaszadáshoz szükséges technikai tudás hiányában nem fognak tudni részt venni a felmérésben. Így az adatfelvétel végrehajtása után kapott minta nem lesz teljes.

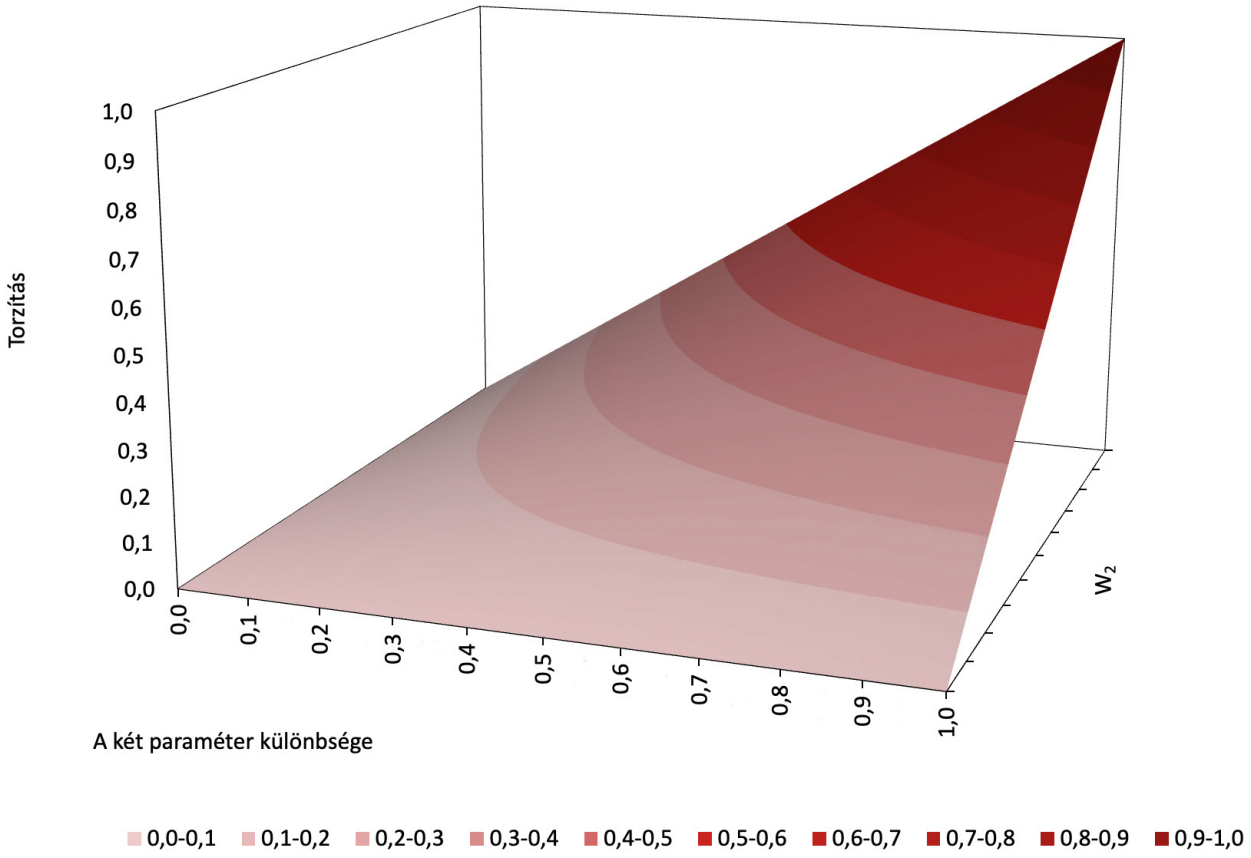
Ahhoz, hogy lássuk ennek következményeit, osszuk fel elméletben a sokaságot két rétegre: egy „online népességre", azaz olyanokra, akik, ha kiválasztásra kerülnek, szerepelnének a végső mintában; és egy „offline népességre", azaz olyanokra, akik kiválasztásuk ellenére a végső mintában nem lennének benne.

Jelölje N_1 az online, N_2 az offline népesség számát az alapsokaságban, $W_1 = N_1/N$ és $W_2 = N_2/N$ a két réteg súlyát, N pedig a teljes alapsokaságot, és $N = N_1 + N_2$. Jelölje továbbá n_1 és n_2 a két réteg elemszámát az n elemű mintában $(n = n_1 + n_2)$. Jelölje továbbá $\overline{Y}$, $\overline{Y}_1$, $\overline{Y}_2$ rendre a vizsgált y ismérv átlagát (vagy arányát) az alapsokaságban, illetve a két rétegben. Az eredetileg kijelölt n mintaelem közül csak n_1-re van megfigyelésünk. Amennyiben ezen n_1 elemű minta $\overline{y}_1$ átlagával becsüljük a teljes sokaság $\overline{Y}$ paraméterét, akkor ennek a becslőfüggvénynek a torzítása a két paraméter különbségének várható értéke:

$$E(\overline{y}_1 - \overline{Y}) = \overline{Y}_1 - \overline{Y} = \overline{Y}_1 - (W_1\overline{Y}_1 + W_2\overline{Y}_2) = W_2(\overline{Y}_1 - \overline{Y}_2)$$

A torzítás tehát két tényezőtől, az offline népesség arányától és a két réteg paraméterének különbségétől függ (1. ábra). Online adatgyűjtés esetén csak az előbbiről kapunk információt, a $w_2 = n_2/n$ ugyanis torzítatlan becslés W_2-re. Ha az offline arány elenyésző, vagy a két réteg az alapsokaság egy véletlen felosztásának tekinthető (azaz nincs kapcsolat az internethasználat és a vizsgált ismérv értéke között, s így $\overline{Y}_1$ és $\overline{Y}_2$ lényegében megegyezik), akkor a torzítás elhanyagolható, az online hozzáférés helyenkénti hiánya nem befolyásolja számottevően az eredményeket. Ha viszont jelentős azok aránya, akiktől nem sikerült adatot kapnunk online, s ráadásul megalapozottan vélelmezhető, hogy $\overline{Y}_2$ számottevő mértékben eltér $\overline{Y}_1$-től, akkor a $\overline{y}_1$ erősen torzított becslése lesz $\overline{Y}$-nak, ami sok esetben kérdésessé teheti a mintából származó eredmények használhatóságát. Ezt az összefüggést szemlélteti az 1. ábra: a torzítás azokban az esetben a legnagyobb, amikor magas az offline népesség aránya és/vagy jelentős a különbség az online és az offline népesség válaszai között.

1. ábra. A torzítás az offline népesség arányának (W_2) és a két réteg paramétereinek különbségének függvényében



Forrás: Saját szerkesztés.

A becslés értékelésének másik szempontja a megbízhatóság. Ha nem tudjuk a sokaság minden elemét megfigyelni, akkor az adatfelvételből származó becslés eltér attól, amit akkor kapnánk, ha teljes körű adatfelvételt végeznénk (Rudas 2006). Az így kapott eltérést nevezzük mintavételi hibának, statisztikai értelemben pedig a becslés standard hibájának (szórása a várható érték körül), amelynek felhasználásával adott megbízhatósági szintű konfidenciaintervallumot számíthatunk, így kiegészítve a pontbecslést intervallumbecsléssé.

A mintavételi hiba számszerűsítéséből adódnak az olyan, a becslés pontosságára vonatkozó állítások, mint hogy „a becslés hibahatára +/– 2,6%". Az egyébként minden további nélkül *matematikailag* kiszámítható hibahatár/mintavételi hiba/standard hiba csak abban az esetben számítható ki *statisztikai* értelemben is, ha a felmérésben részt vevők *véletlenszerűen* kiválasztott elemei a sokaságnak. Tehát, amíg a minta kifejezést a sokaság bármely tetszőleges részhalmazára használhatjuk, a sokaságra való általánosítást csak a véletlen (valószínűségi) minta engedi meg. Csak a valószínűségszámítási alapokon nyugvó adatfelvétel biztosítja – a nagy számok törvényein keresztül – a belőle származó becsléseknek azt a tulajdonságát, hogy a becslés nagy valószínűséggel csak kis mértékben tér el a becsülni kívánt populációparamétertől.

Az érvényesség és a megbízhatóság dimenziói mellett gyakran vizsgáljuk a kérdőíves kutatások minőségét a *total survey error* (TSE) keretrendszer segítségével, ami összegzi az összes potenciális hibát, melyek egy surveykutatás során felmerülnek (Groves et al 2009). A TSE a hibákat két csoportra osztja, melyek a mintavételi és a mérési komponensek. A mintavételi hibák oldala a populációtól a válaszadói bázis kiválasztásáig történő

lépéseket foglalja magában, míg a mérési hibák oldala az egyes válaszadók válasza és az ebből származtatott becslés közötti folyamatot összegzi (Biemer–Lyberg 2003). Az érvényesség és a megbízhatóság a mintavételi oldalhoz sorolhatóak, viszont a mérési komponensek esetében is releváns különbségekre számíthatunk akkor, ha a személyes és az online kérdőíves kutatásokat vetjük össze. Egy survey esetében maga a mérés a válaszadást jelenti, és amennyiben a válaszadás helyzete eltér, a mérési hiba is eltérhet. Egy személyes kérdezés esetében a kérdéseket kérdező teszi fel és ő jelöli a válaszokat, így az összehasonlításnál egyrészt számolni kell a kérdezői hatással (a kérdező befolyásolhatja a válaszadót szándékosan vagy akaratlanul is), illetve az eltérő válaszadói viselkedéssel (érzékeny kérdésekre könnyebb őszintén válaszolni önkitöltős módon, amilyenek például online felmérések). Ezeket együttesen módhatásnak nevezzük (Dillman–Christian 2005). A kérdezési mód hatásának vizsgálata a nemzetközi szakirodalom régóta vizsgált kérdése. A módhatás rendkívül sokrétű, és vannak olyan kutatási kérdések, ahol kisebb a mérési hiba az online kitöltés során: a szavazási szándék egy példa lehet ezekre a kérdéskörökre, hiszen e kérdés esetében a kérdező hatása miatt a személyes felmérések sem feltétlenül adnak megbízható eredményt (Duffy et al. 2005). Magyarországon is zajlott speciális módszertani kutatás a *European Social Survey* keretében. Ennek eredményei alapján valószínűségi minta alkalmazása mellett is szignifikáns különbségek tapasztalhatóak az online és a személyes kitöltők válaszai között: az érzékeny társadalmi kérdésekben nyitottabb, elfogadóbb véleményeket tükröznek még a lakossági paraméterekhez korrigált adatok is (Messing–Ságvári–Szeitl 2022). A mérési hiba tendenciájára viszont nincsenek általánosítható eredmények, emiatt ez az egyik legnehezebben korrigálható torzítás.

A TSE keretrendszere alapján a mintavételi és a mérési komponensek nem tudják kiegyenlíteni egymást, azaz az online és a személyes kérdőíves adatgyűjtések összehasonlításakor ezeket külön-külön érdemes számba venni. A módhatás általános mintázatára egyelőre nincsenek validált hazai eredmények, így a szimulációs módszert ez alapján nem tudjuk specifikálni. Emiatt a következőkben az online kérdőíves adatgyűjtésekből származó becsléseket kizárólag a mintavételi hibák szempontjából értékeljük.

Az internethasználó populáció jellemzése

A következőkben röviden bemutatjuk az internethasználó populáció vizsgálatának mintavételi problémáit, valamint hogy mennyiben tér el az online népesség a teljes népességtől Magyarországon. Az online populáció körülhatárolása nem egyértelmű: a Központi Statisztikai Hivatal által gyűjtött és közzétett egyéni információs és kommunikációs (IKT) eszközhasználatot tekintve a 2021. évre az 1. táblázatban látható internethasználati gyakoriságok érhetőek el.

1. táblázat: A hazai felnőtt népesség internethasználata 2021-ben

Mutató	Arány a teljes népességen belül
3 hónapon belüli közösségi háló használatának aránya (%)	77,2
Csaknem minden napos internethasználat aránya – 3 hónapon belül (%)	82,3
Legalább heti egyszeri internethasználat aránya – 3 hónapon belül (%)	87,4
Egyéni internethasználat aránya – 3 hónapon belül (%)	88,6
Egyéni internethasználat aránya – 12 hónapon belül (%)	89,1
Egyéni internethasználat aránya – összesen (%)	90,1

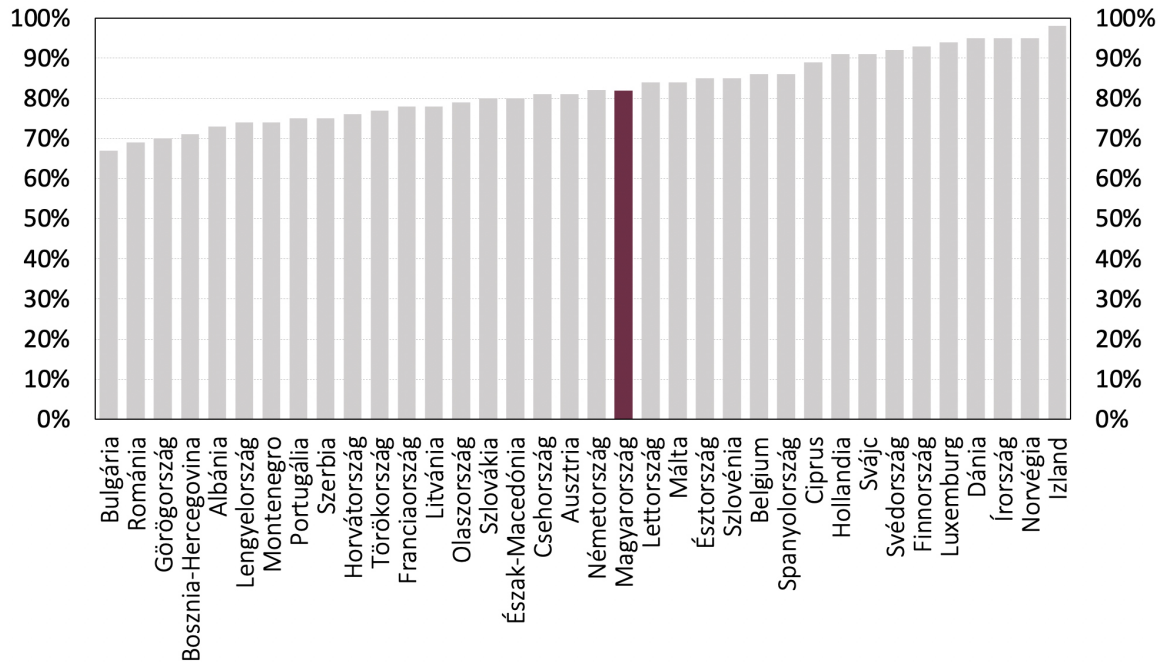
Forrás: KSH adatok alapján, saját szerkesztés.

Az adatok jól mutatják, hogy definíciótól függően széles skálán, 77–90% között szóródik az internetezőként aposztrofálható népesség. Az, hogy ezek közül melyik sokaságot tekinthetjük internethasználónak, az adatfelvétel céljától függ. Egy mintavételi keret megtervezésekor először is kiindulhatunk abból, hogy mely sokaság fedi azokat, akiknek *elméleti* esélyük van a mintába kerülésre. Ehhez választhatjuk a legbővebb – ilyen értelemben a leginkább optimista – definíciót, miszerint aki használt valaha internetet, annak volt valamekkora (nullától különböző) mintába kerülési valószínűsége. Ebben az értelemben a hazai teljes népesség 90%-a adja a mintavételi keretet.

Ugyanakkor, ha arra gondolunk, hogy a népesség melyik szegmense az, akinek *gyakorlatilag* is van potenciális bekerülési valószínűsége egy általunk tervezett online adatfelvétel mintavételi keretébe, már kevéssé életszerűnek találhatjuk, hogy elérjük a kérdőívünkkel azokat, akik használtak már valaha internetet, de az elmúlt három hónapban nem voltak aktív használók. Egy jól kialakított dizájn mellett minél gyakrabban használ az egyén internetet, átlagosan annál nagyobb eséllyel jut el hozzá a kérdőív. Így egy online adatfelvétel esetében – amely tipikusan relatíve rövid idő, néhány nap vagy hét alatt folyik – plauzibilis feltevés, hogy a KSH által mért kategóriák közül az elmúlt három hónapban csaknem mindennap internethasználó népesség lehet az a sokaság, amelyet mintavételi keretnek tekinthetünk. Így a teljes népesség 82%-a tekinthető elérhetőnek. Amennyiben a közösségi médián keresztül valósítanánk meg adatfelvételt, a népesség 77%-a válna elérhetővé.

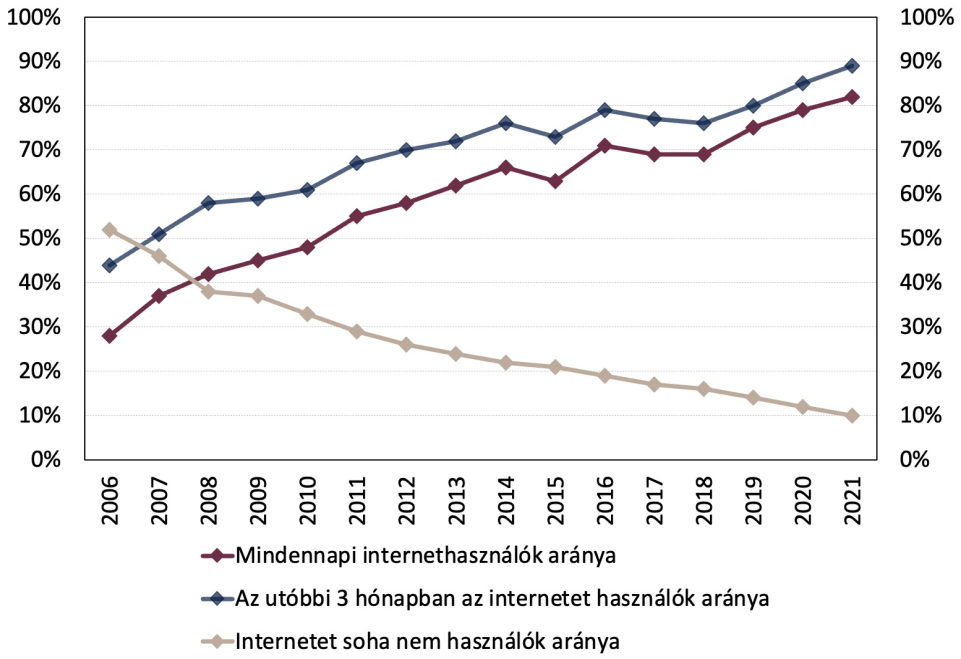
Európai összehasonlításban az látható, hogy az északi országok (Izland, Norvégia, Dánia, Írország) közelítették meg leginkább a teljes internetpenetrációt (2. ábra). A régiós országok mindegyikében kissé alacsonyabb a mindennapos internethasználók aránya, mint Magyarországon. Így középtávon nem valószínűsíthető, hogy a „teljes népesség” és az „internetező népesség” ekvivalens halmazokká váljanak.

2. ábra. A mindennapos internethasználók aránya, 2021



Forrás: Eurostat adatok alapján, saját szerkesztés.

3. ábra. A hazai népesség megoszlása az internethasználat gyakorisága szerint

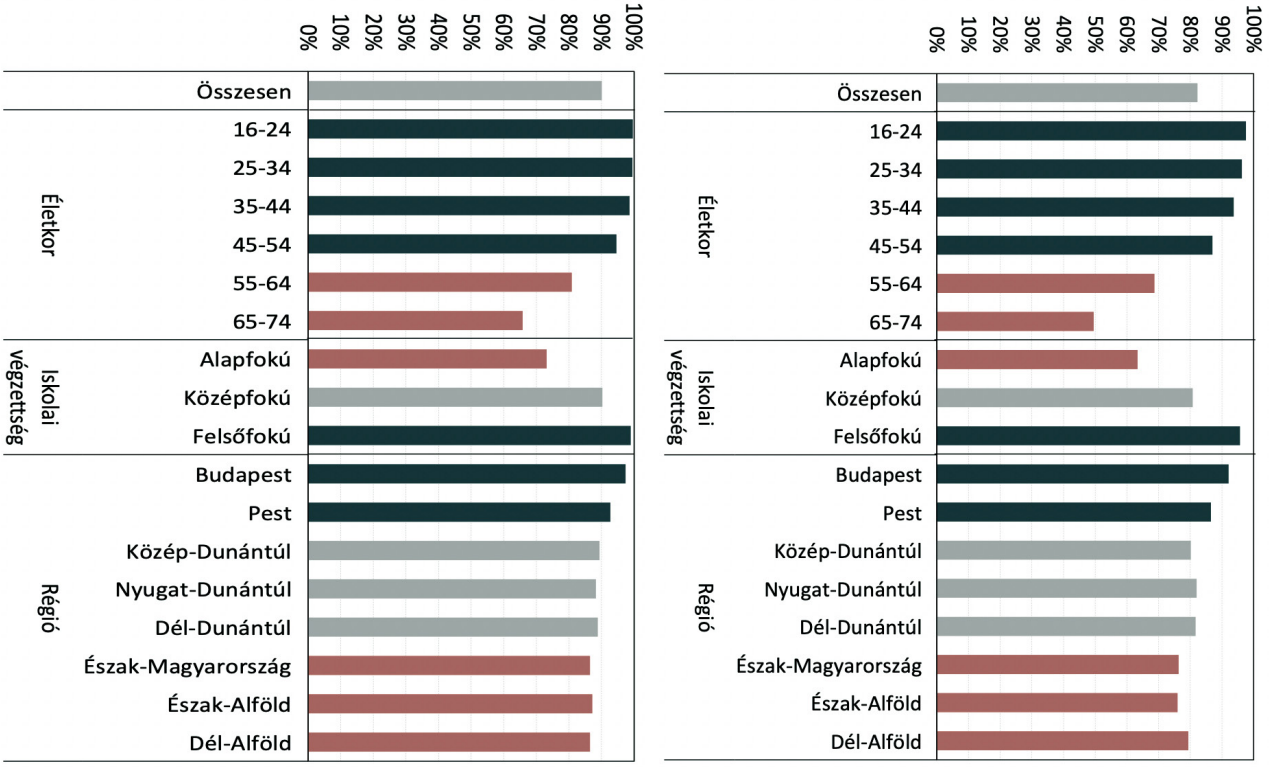


Forrás: KSH adatai alapján, saját szerkesztés.

Az internetező népesség az elmúlt két évtizedben gyökeresen megváltozott. A minden nap internetet használók aránya a 2000-es évek közepén még 30% alatt volt, ez 2014-ig relatíve gyorsan emelkedett, összefüggésben a digitalizáció fejlődésével és a fejlettebb technológiák elérhetőségének javulásával (pl. okostelefonok). A sűrűn internetezők aránya a teljes népességen belül még az utóbbi három évben is több mint 10 százalékponttal nőtt (3. ábra). Mivel az internetpenetráció emelkedésével egyre szélesebb rétegek vonódnak be és lehetnek online adatfelvételek célcsoportjai, amíg a mindennapos internethasználat aránya nem állandósul egy adott szinten, addig a különböző időszakokban készített online felmérések az alapsokaság folyamatos változása miatt egyértelműen nem hasonlíthatóak össze egymással.

A 4. ábrán és a 2. táblázatban azt szemléltettük, hogy a 2021. évi hazai népesség esetében az internettel rendelkezők és a rendszeres internethasználók mennyiben térnek el a teljes népességtől a főbb demográfiai ismérvek szerint. Az internethasználat relatív gyakorisága a KSH adatai alapján meghaladja az országos átlagot az 55 év alattiak esetében (ezen belül is az alacsonyabb életkori kategóriákban magasabb), a felsőfokú iskolai végzettségűek körében, valamint a fővárosi és Pest megyei lakosok esetében. Az idősebbek, az alapfokú végzettségűek és az ország keleti régióiban élők között alacsonyabb az internethasználat aránya (4. ábra). Az online elérhető populációkban leginkább az 55 év felettiek és az alapfokú végzettségűek alulreprezentáltak, ezek a csoportok 2–6 százalékponttal kisebb arányban képviseltetik magukat, mint a teljes népességben. Ezzel szemben a fiatalok, a felsőfokú végzettségűek és a közép-magyarországi régióban élők felülreprezentáltak (2. táblázat).

4. ábra. Az internettel rendelkezők (bal) és a csaknem minden nap internetet használók (jobb) aránya Magyarországon 2021-ben



Megjegyzés: Feketével azokat a csoportokat jelöltük, amelyekben az internethasználat nagyobb, mint a teljes népességben; bordóval azokat, ahol kisebb; szürkével a teljes népességben megfigyelt aránnyal közel azonos arányban internethasználókat. Forrás: KSH adatai alapján, saját szerkesztés.

Ezek az adatok alátámasztják, hogy az internethasználó népesség és a teljes népesség érdemben eltér egymástól.[4] Tehát egy online elérhetőség alapján vett valószínűségi véletlen minta esetében sem biztosított, hogy a felmérésben részt vevők a teljes népességet képviseljék. A bemutatott differenciák az alapvető demográfiai ismérvek szerint állnak fenn, így felmerül a kérdés, hogy a teljes népességbeli eloszlást súlyként használva feloldható-e a két sokaság eltérése. Ez az érvelés természetesen csak akkor állja meg a helyét, ha az internetező és a nem internetező népesség véleménye az egyes demográfiai csoportokban nem tér el egymástól. A következő részben azt mutatjuk be, hogyan viszonyulhatunk ehhez a feltevéshez.

4 Az internethasználat számos tevékenységet lefedhet a közösségimédia-használattól az online ügyintézésig, amelyek az online aktivitás és a digitális fejlettség eltérő fokát jelentik. A KSH magáncélú internethasználatot mérő, 2019-re vonatkozó adatai alapján a leggyakoribb felhasználás az emailek írása és fogadása (90%), a legkevésbé gyakori az internetes banki szolgáltatások igénybevétele (58%) (KSH 2021). Az efféle differenciálás az internethasználók között még nagyobb eltéréseket implikálhat a teljes népesség és az online kérdőívvel elérhető népesség között. Az egyes internethasználati módok szociodemográfiai ismérvek szerinti bontása azonban nyilvánosan nem elérhető.

2. táblázat. A teljes népesség, az internettel rendelkezők és a mindennapos internethasználók életkor, iskolai végzettség és régió szerint 2021-ben

Ismérv	Demográfiai csoport	Arány a teljes népességen belül	Arány az internettel rendelkezőkön belül	Arány a mindennapos internethasználókon belül
Életkor	16–24	13%	15%	16%
	25–34	16%	18%	19%
	35–44	19%	21%	22%
	45–54	20%	21%	21%
	55–64	16%	14%	13%
	65–74	16%	11%	9%
Iskolai végzettség	Alapfokú	19%	16%	15%
	Középfokú	56%	56%	55%
	Felsőfokú	25%	27%	29%
Régió	Budapest	17%	19%	19%
	Pest	14%	14%	14%
	Közép-Dunántúl	11%	11%	11%
	Nyugat-Dunántúl	10%	10%	10%
	Dél-Dunántúl	9%	9%	9%
	Észak-Magyarország	11%	11%	10%
	Észak-Alföld	15%	14%	14%
	Dél-Alföld	13%	12%	12%

Forrás: KSH adatok alapján, saját számítás.

Szimulációs eredmények

A következőkben azt mutatjuk be, hogy elméletben[5] az esetek mekkora arányában és milyen körülmények között vezethet egy online mintából vett becslés a valódi populációs paraméterhez viszonyítva téves eredményhez a magyar lakosság összetétele szerint.

Az 1. ábrán bemutattuk a torzítás alakulását két réteg esetén. Több *(H)* rétegre általánosítva egy adott *h* réteg esetében adódó torzítás a következő formát ölti:

$$E(\bar{y}_h - \bar{Y}) = \bar{Y}_h - \bar{Y} = \bar{Y}_h - \sum_{h=1}^{H} W_h \bar{Y}_h$$

Látható, hogy kettőnél több réteg alkalmazása esetén a torzítás mértéke az összes réteg arányától és a rétegenkénti paraméterek különbségétől is függ, ami három réteg esetében hat paraméter együttese. Ezek hatásainak elméleti bemutatását jelen tanulmányban mellőzzük annak összetettsége okán; helyette egy szimulációval illusztráljuk a torzítás alakulását.

Az internetes kérdőíves adatgyűjtésekből származó becslések pontosságának megismeréséhez online adatfelvételek eredményeit szimuláljuk. A szimuláció célja, hogy illusztrációként szolgáljon az online véletlen minták használatából adódó hiba mértékének bemutatásához. A szimuláció során a hazai felnőtt népességet iskolai végzettség szerint három rétegre bontjuk. Azért az iskolai végzettség alapján végezzük a bontást, mert

5 A szimulációban implicite két, az alapsokaságtól eltekintve minden szempontból ugyanúgy kivitelezett surveyfelmérés eltérését modellezzük. Így az nem számszerűsíti az adatfelvétel módjából, a válaszadási hajlandóságból és az egyéb nem-mintavételi hibákból származó eltéréseket, valamint azt, hogy az online népességből jellemzően nem véletlen mintavétellel kerülnek kiválasztásra a válaszadók. Tulajdonképpen tehát azt példázzuk, hogy egy survey módszertani szempontból megfelelően kivitelezett, ideális kutatási szituációban mennyiben reális, hogy az online népességből származó mintára alapuló következtetések általánosíthatóak a teljes népességre.

ez egy olyan demográfiai változó, amely egyrészt összefüggésben áll az internetezéssel, másrészt a mintavétel és az utólagos rétegzés során is gyakran használt változó. A Központi Statisztikai Hivatal adminisztratív adatait használva a három réteget az (1) alapfokú, (2) középfokú, (3) felsőfokú végzettséggel rendelkezők csoportjai alapján definiáljuk,[6] és a csoportok populáción belüli arányait is az adminisztratív adatok alapján állapítjuk meg.

Megjegyezzük, hogy az iskolai végzettség három rétege szerint végzett szimuláció szemléltetési célokat szolgál. A gyakorlatban tipikusan ennél több dimenzió szerinti rétegzés vagy utólagos súlyozás használt. Eredményeinket annyiban nem befolyásolja a használt rétegek vagy a rétegzéshez használt változók száma, hogy egy klasszikus négydimenziós (nem, életkor, iskolai végzettség, településtípus) rétegzés esetében is felmerül az online és a teljes népesség közötti szisztematikus eltérés.[7]

Vegyük tehát a három réteg (h = 1 – alapfokú; h = 2 – középfokú; h = 3 – felsőfokú végzettségűek) súlyát (a KSH 2021. évi adatai alapján) a teljes népességben *(W_h):*

$$W_1 = 0{,}195$$
$$W_2 = 0{,}559$$
$$W_3 = 0{,}246$$
$$\sum_{h=1}^{3} W_h = 1$$

és a három rétegen belül vesszük az internetező *($W_{h\ online}$)* és a nem internetező *($W_{h\ offline}$)* népesség súlyát (a KSH 2021. évi adatai alapján):

$$W_{1_{online}} = 0.634 \qquad W_{1_{offline}} = 0.366$$
$$W_{2_{online}} = 0.808 \qquad W_{2_{offline}} = 0.192$$
$$W_{3_{online}} = 0.957 \qquad W_{3_{offline}} = 0.043$$

A sokaság így definiált hat rétegéhez potenciális értékeket generáltunk a becsülendő paraméterből. Ez a paraméter az egyszerűség kedvéért egy populációs arány (például a következő választáson szavazást tervezők aránya, a napi szinten templomba járók aránya vagy a legalacsonyabb jövedelmi kategóriába tartozók aránya). A szimuláció során a rétegekben egymástól függetlenül szimuláljuk a vizsgált paraméter összes lehetséges kombinációját úgy, hogy az arány rétegenkénti értéke 0 és 1 közötti értékeket vesz fel 0,1-es osztásközzel. Az internetezőkből vett rétegzett minta esetében jelölje ezt az arányt *i* (az egyes rétegekben i_h), a nem internetezők esetében pedig *t* (az egyes rétegekben t_h). Az így létrejövő adatbázis egy részletét sematikusan a 3. táblázat tartalmazza.[8]

---

6 A rétegeket a következő csoportosítással hoztuk létre: az alapfokú végzettséghez soroltuk az általános iskolai vagy ennél alacsonyabb végzettséget, illetve az érettségi nélküli középfokú vagy szakmai végzettségeket; a középfokú végzettséghez soroltuk az érettségi bizonyítványt is tartalmazó végzettségeket, illetve az olyan érettségire épülő felsőfokú szakképzettségeket, melyek nem számítanak főiskolai/egyetemi végzettségnek; a felsőfokú végzettséghez a legalább főiskolai/egyetemi végzettségeket soroltuk.

7 A szimuláció kettő réteg használatával is a későbbiekben bemutatotthoz hasonló eredményt ad, háromnál több réteg esetében azonban számítási kapacitásbeli problémák merülnek fel. Mindazonáltal az online véletlen minta torzított jellege a teljes népességre való általánosításkor elméleti szempontból egyértelmű, ha az internethasználó népesség nem véletlen mintája a teljes népességnek. A probléma mértékének illusztrálására – ami a szimuláció célja – pedig három réteg is elégséges.

8 Összesen 6^{11}, azaz 1.771.561 szimulált eset jött létre.

3. táblázat. A szimulált arányokat tartalmazó adatbázis sematikus ábrája

i_1	i_2	i_3	t_1	t_2	t_3
0,0	0,0	0,0	0,0	0,0	0,0
0,1	0,0	0,0	0,0	0,0	0,0
0,1	0,1	0,0	0,0	0,0	0,0
…					
0,1	0,1	0,2	0,1	0,1	0,1
0,1	0,1	0,2	0,1	0,1	0,2
…					
0,5	0,6	0,7	0,5	0,6	0,7
…					
0,8	0,8	0,9	0,9	0,9	0,9
0,8	0,9	0,9	0,9	0,9	0,9
0,9	0,9	0,9	0,9	0,9	0,9

Megjegyzés: Minden sor különböző szimulált eseteket jelent. Forrás: Saját szerkesztés.

Ezt a paramétert szeretnénk becsülni a teljes népességből és az internetezőkből vett rétegzett, a rétegeken belül egyszerű véletlen mintával. Jelölje a kérdéses arányt a teljes népességben *p* (az egyes rétegekben p_h, mely mindhárom réteg esetében az online és az offline népességhez szimulált értékek súlyozott átlaga. Tegyük fel, hogy *p* valódi értéke a populációban létezik és mérhető.

Mindegyik szimulált kombináció[9] esetében elvégezzük a vizsgált arány torzítatlan becslését a rétegen belül szimulált arányok súlyozott átlagával:

$$\hat{p} = \sum_{h=1}^{3} W_h \cdot p_h$$

Így tehát a kérdéses paraméter becsült értéke, ha a rétegek becsült értékei a teljes sokaságon alapulnak, és ugyanez akkor, ha az online népességén – mindkét esetben a teljes népességbeli arányok szerint súlyozva. Már ebből is látható, hogy amennyiben a p_h és i_h paraméterek nem egyeznek meg rétegenként, akkor az online népességből vett minta még akkor is torzított becslést eredményez, ha maga a kutatás szerkezete minden más szempontból meg is felel egy, a teljes népességre vonatkozó kutatásnak, így például van lista a populáció elemeiről, és megvalósítható a véletlen mintavétel.

Felmerülhet azonban, hogy ez a torzítás – mértéket tekintve – releváns eltérést jelent-e az eredményekben. A szimuláció keretében statisztikai szempontból minősítjük ezt a relevanciát. Aránybecslés esetén a becslés 95%-os konfidenciaintervalluma a következő:

$$Int(\hat{p})_{0,95} = \hat{p} \pm 1{,}96 \cdot SH_{\hat{p}}$$

ahol a standard hiba (SH)

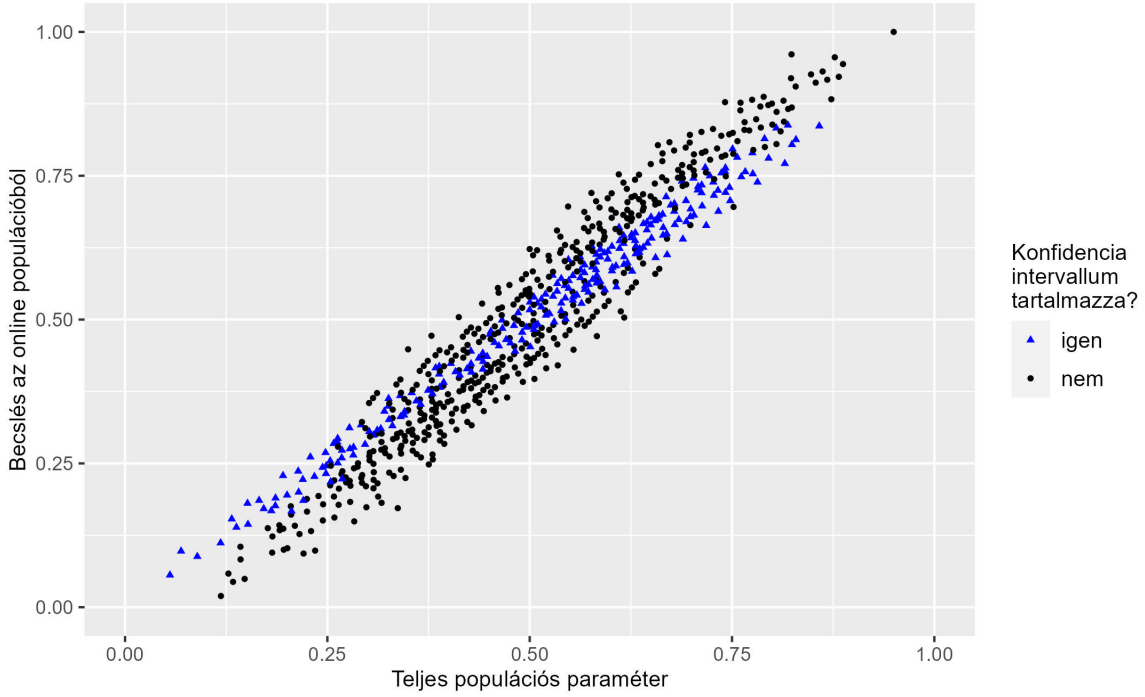
$$SH_{\hat{p}} = \sqrt{\sum_h W_h^2 \frac{p_h \cdot (1 - p_h)}{n_h}}$$

9 A rétegképző ismérv (végzettség) és a becsült paraméter közti összefüggésben a szimuláció során minden lehetséges esetet megengedtünk; azaz szerepel az az eset is, amelyben minden rétegen belül azonosak a paraméterek. Ezzel az volt a célunk, hogy az összes lehetséges szcenáriót bemutassuk, amelyek között az aktuális társadalomkutatási kérdés szempontjából vannak szerencsésebb és kevésbé szerencsés esetek (az online populációból származó becslés általánosíthatósága szempontjából).

Ezt az intervallumbecslést minden paramétervariáció mellett kiszámítottuk annak érdekében, hogy láthatóvá váljon, hogy az internetezőkön alapuló becslés hibahatáron belülre esik-e.

Az 5. ábrán az látható, hogy az összes lehetséges kombináció alapján számolt becslések közül melyek lesznek azok, amik „elég pontosak", azaz beleesnek a valós arány 95%-os megbízhatósági szintű konfidencia-intervallumába. Az összes eset kisebb része, 33%-a felel meg ennek a kritériumnak: a becslések döntő többsége szignifikánsan alul- vagy felülbecsüli a vizsgált paramétert.

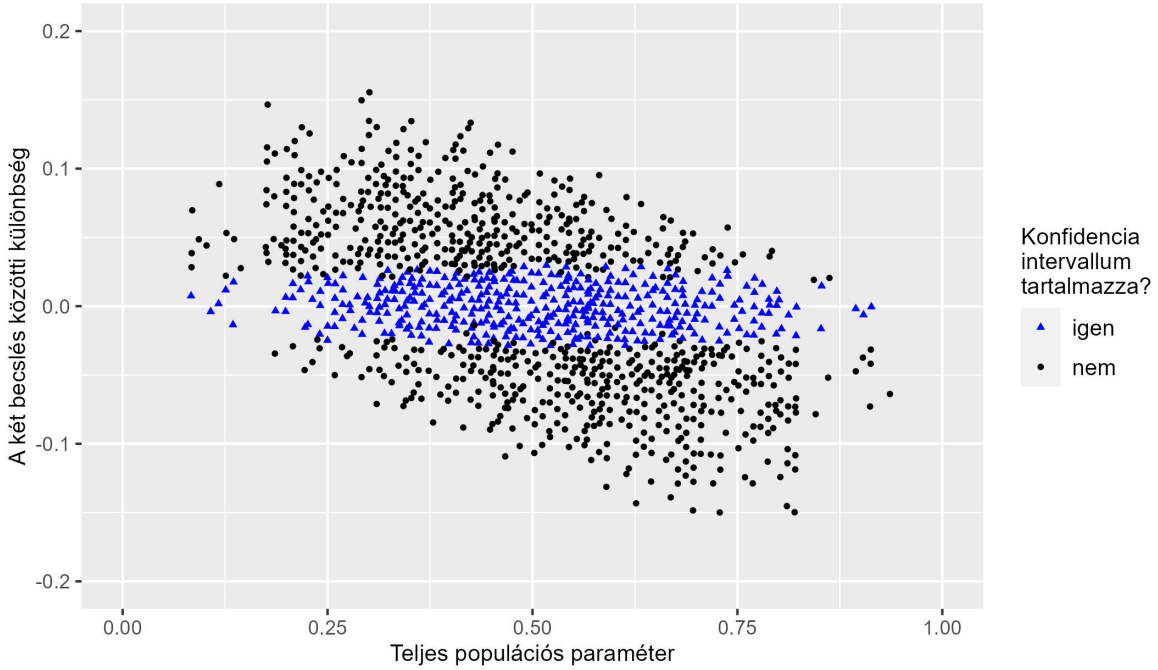
5.ábra. Az internetezők alapján a teljes populációra becsült és a tényleges populációs arány



Megjegyzés: A konfidencia-intervallum a populációs paraméter köré szerkesztett 95%-os megbízhatósági szintű konfidencia intervallumot jelenti. Az ábrán az összes szimulációnak egy 0,0006%-os (nagyjából 1000 elemű) véletlen mintáját jelenítettük meg. Forrás: saját számítás.

Az internetező népesség alapján a teljes népességre adott becslés távolsága a tényleges populációs paramétertől abszolút értékben átlagosan 4,7 százalékpont, a medián eltérés 4,1 százalékpont, és az átlagszámításból adódóan az internetezőkből számított becslés alacsony populációs paraméter mellett alulbecsli azt, a paraméter nagyobb értékeinél pedig felülbecsli (6. ábra).

6. ábra. Az internetezők alapján a teljes populációra becsült és a tényleges populációs arány különbsége



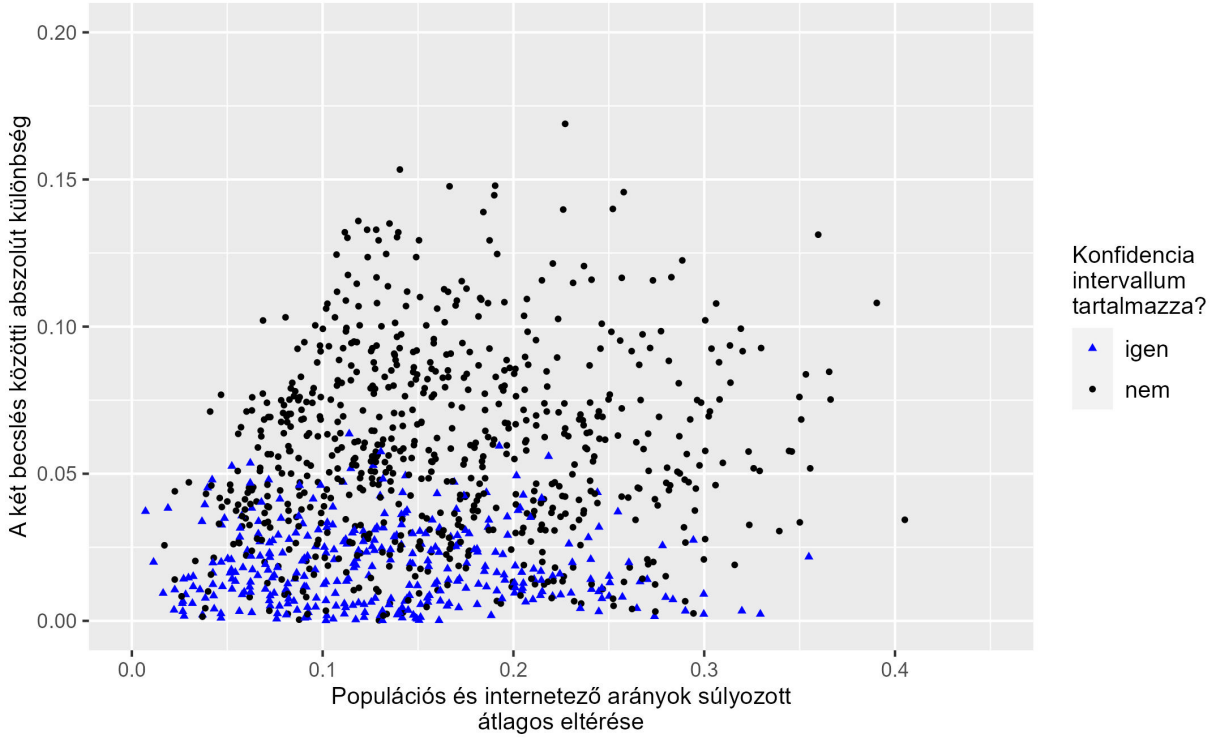
Megjegyzés: A konfidencia intervallum a populációs paraméter köré szerkesztett 95%-os megbízhatósági szintű konfidencia intervallumot jelenti. Az ábrán az összes szimulációnak egy 0,0006%-os (nagyjából 1000 elemű) véletlen mintáját jelenítettük meg. Forrás: saját számítás.

Azt, hogy mennyiben tér el az internetezők véleménye a populációs véleményektől az egyes rétegekben, egy összesített mutatóval mérjük, mely a teljes népesség és az internethasználó népesség között megfigyelhető rétegenkénti eltérések (réteg relatív gyakoriságával) súlyozott átlaga:

$$\sum_{h=1}^{3} N_h \cdot |p_h - i_h|$$

Az eltérés hatása jól látható a 7. ábrán: minél nagyobb az internetező és a teljes népesség véleménykülönbsége, annál nagyobb a különbség a két becslés között. Fontos, hogy az eltérés már viszonylag kis eltérés esetén is jelentős: az ábra alapján jól látható, hogy már 5 százalékpontos összesített eltérés esetén is (vízszintes tengely) kiugróan nagy eltérést tapasztalhatunk egy online mintából származó becslés esetében a populációs paraméterhez viszonyítva.

7. ábra. Az online minta alapján becsült és a populációs arány különbsége a rétegenkénti online mintán belüli, populációs aránytól vett különbsége alapján



Megjegyzés: A konfidencia intervallum a populációs paraméter köré szerkesztett 95%-os megbízhatósági szintű konfidencia intervallumot jelenti. Az ábrán az összes szimulációnak egy 0,0006%-os (nagyjából 1000 elemű) véletlen mintáját jelenítettük meg. Forrás: saját számítás.

Becslés valós surveyadatokon

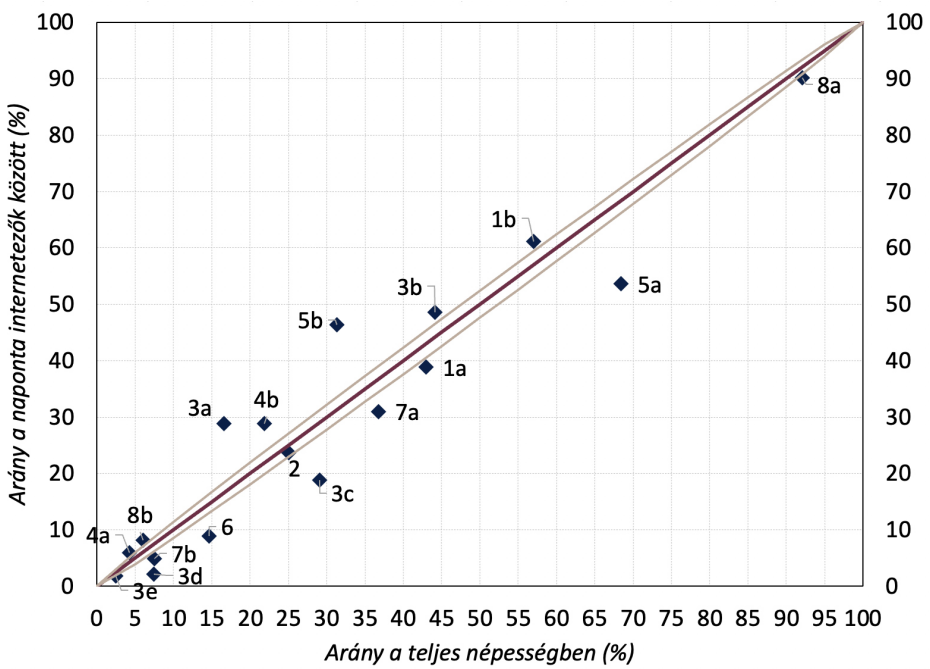
Az előző eredmény szerint az összes arány kombinációinak harmadában kaptunk olyan becslést, ami a valódi paraméter 95%-os konfidencia intervallumába esik, azaz „jó" becslés; illetve azt láttuk, hogy viszonylag kicsi, 5 százalékpont körüli átlagos eltérés is elég az internetező és a teljes populáció viszonylatában ahhoz, hogy az internetezők alapján készített becslés lényegesen eltérjen a teljes populációra jellemző paramétertől. Jogos kérdés viszont, hogy mennyire tekinthető az ilyen mértékű különbség reálisnak, azaz, hogy egy adott vélemény-/attitűdkérdés esetében valóban számolnunk kell-e azzal, hogy a teljes lakosság és az internethasználó lakosság megegyező demográfiai csoportjai a vizsgált kérdés esetében különbözőképpen vélekednek. Erre a kérdésre azért nehéz választ találni, mert a valódi populációs arányokat nem ismerjük (hiszen éppen ezért végzünk adatfelvételeket). Az internetező és a nem internetező lakosság vélemény-/attitűdkülönbségeit adminisztratív adatok hiányában surveyadatokon keresztül vizsgálhatjuk. A továbbiakban az ESS[10] 9. hullámának magyarországi adatait használjuk arra, hogy bizonyos arányok becslését elvégezzük a teljes minta esetében[11] és az internetező minta esetében. Mindegyik arány esetében két becslést végzünk: (1) a teljes népességre vonatkoztatva (azaz függetlenül az internetezési arányoktól a teljes mintán); (2) az

10 Az ESS-ről részletesebben itt található több információ: https://www.europeansocialsurvey.org/.

11 Az ESS esetében klasszikus címlistás (csökkenőmintás) kutatásról van szó, ahol a kérdezők személyesen látogatnak el előre megadott címekre, azaz a rekrutálás offline módon történik. A 9. hullám részletes mintavételi leírása itt érhető el: https://www.europeansocialsurvey.org/docs/round9/methods/ESS9_sampling_guidelines.pdf

internetező almintán belül (az internetező alminta azokat jelenti, akik legalább napi szinten használják az internetet). A becslés a három réteg súlyozott átlagát jelenti úgy, hogy eredményekben a becslés a három réteg teljes populációs mérete alapján kerül súlyozásra.

8. ábra. A European Social Survey egyes változóinak becslése a teljes népességre és az internetező népességre vonatkozóan



Megjegyzés: A pontok az egyes változók becsléseit jelölik (iskolai végzettség szerinti rétegzésből a teljes népességre vonatkozó arányokkal súlyozva). Pirossal a 45 fokos egyenes, azaz amikor a teljes népességbeli és a naponta internetezők között mért arány egybeesik. Sárgával az aránybecslés 95%-os konfidenciaintervalluma. 1a: Közel érzi magát valamelyik párthoz – igen; 1b: Közel érzi magát valamelyik párthoz – nem; 2: Érdekli a politika; 3a: Nagyon jó egészségi állapotú; 3b: Jó egészségi állapotú; 3c: Közepes egészségi állapotú; 3d: Rossz egészségi állapotú; 3e: Nagyon rossz egészségi állapotú; 4a: Nagyon vallásos; 4b: Nem vallásos; 5a: Házasságban élő; 5b: Nem házasságban élő; 6: Nagyon elégedetlen az életével; 7a: Nagyon kötődik Magyarországhoz; 7b: Nagyon nem kötődik Magyarországhoz; 8a: Alkalmazotti munkaviszonyban áll; 8b: Vállalkozó. Forrás: ESS alapján saját számítás.

A legtöbb vizsgált változó esetében jelentősen eltérnek a 3 rétegen belül megfigyelt arányok. A 8. ábrán pedig az látható, hogy mindegyik változó esetében eltérő becslést kapunk a teljes népességre és az internetes populációra vonatkozóan – miközben az iskolai végzettség szerinti arányok szempontjából reprezentatív a kapott minta. Például a politikai kérdés esetében az internetezőket vizsgálva jelentősen felülbecsüljük azok arányát, akik nem érzik magukat közel egyik párthoz sem (61%-os arányt becsülünk a teljes minta alapján becsült 55%-os arányhoz képest). Az egészségi állapot szempontjából egy internetes minta alapján jelentősen jobb egészségi állapotot találnánk az internetes minta alapján (30% a nagyon jó egészségi állapotot jelentők aránya a teljes mintán belüli 17%-os arányhoz képest).

KÖVETKEZTETÉSEK

A tanulmányban az empirikus társadalomkutatás szempontjából is releváns jelenséggel, az online adatgyűjtések értékelésével foglalkoztunk. Bemutattuk, hogy melyek azok a matematikai-statisztikai dimenziók, amik a nem valószínűségi mintavételek során sérülnek, és azt, hogy miért következik mindebből, hogy potenciálisan megbízhatatlan eredményeket kapunk.

Az online adatfelvételekben kulcskérdés a populáció meghatározása és az internetellátottság monitorozása. Tanulmányunkból kiderült, hogy az internettel rendelkezők és a mindennapos internethasználók sokasága az alapvető demográfiai szempontok szerint eltér a teljes felnőtt népességtől. Ezen populációkban leginkább az 55 év felettiek és az alapfokú végzettségűek alulreprezentáltak, míg a fiatalok, a felsőfokú végzettségűek és a Közép-Magyarország régióban élők felülreprezentáltak. Azt, hogy elméletben mekkorát tévedhetünk egy online mintából származó becsléssel, szimulációs elemzéssel vizsgáltuk: az esetek több mint kétharmadában a valódi értéket alul- vagy éppen felülmérjük, és egészen kis összesített különbség az internethasználó és a teljes népesség között is elegendő ahhoz, hogy téves következtetésekre jussunk. A szimulációs eredmények mellett valós adatokat is vizsgáltunk: az ESS mintája alapján a vizsgált összes változó esetében a teljes mintától jelentősen eltérő becslést kaptunk akkor, ha csak az internethasználókat kérdeztük, még akkor is, ha az iskolai végzettség szerinti arányokat a populációs arányokhoz igazítottuk. Összességében tehát fontos szem előtt tartani, hogy egy online adatfelvétel még a látszólag magas internetellátottság mellett sem tekinthető véletlen mintavételnek, és az ilyen adatokból számolt becslések még akkor is félrevezetőek lehetnek, ha bizonyos demográfiai szempontok szerint jól reprezentálják a teljes népességet.

Az új technológiai lehetőségek használata természetesen jó, de fontos, hogy azokat leginkább az adatgyűjtési módokhoz és nem a mintavételi módokhoz kell kapcsolni: valószínűségi mintavételt követően érdemes online módon is lehetőséget biztosítani a válaszadásra, ami (a mérési hibák figyelembevétele mellett) a válaszadási arányon is javíthat. Erre már több nemzetközi felmérés is lehetőséget biztosít hibrid (azaz kevert) módszertanú kutatás formájában, illetve *„Push-to-web”* (webre terelés) felmérés keretein belül már Magyarországon is készült kísérleti felmérés (Messing–Ságvári–Szeitl 2022). Online mintán végzett felmérésekkel továbbra is megfelelően tudunk becsülni olyan jellemzőket, melyek a kutatás céljai szerint megfelelő pontossággal körülhatárolható alpopulációra vonatkoznak. Az adatok lehetséges komplexebb korrekciós eljárásairól itt most nem esett szó. A nem valószínűségi online mintán alapuló kutatások esetében statisztikai szempontból javasolt az elemzés megkezdése előtt részletesen megvizsgálni a válaszadói bázis összetételét, a mintaszelekciós hatásokat, illetve azt, hogy az internetes rekrutáció miatt milyen társadalmi csoportok kerülhettek a válaszadói bázisba kisebb valószínűséggel, vagy melyek azok a csoportok, melyeknek egyáltalán nem is volt esélyük választ adni. E szempontok vizsgálata alapján valamelyest korrigálhatóak a különbségek, habár a tanulmányban ismertetett alapvető problémák nem kezelhetőek Elképzelhető, hogy a mérési hibák bevonása módosítaná az eredményeinket, viszont a mintavételi hibákkal, az érvényesség és a megbízhatóság dimenziói szerint nem változtatna a következtetéseken.

Hivatkozások

Andorka R. – Kolosi T. – Vukovich Gy. (1990) Előszó. In Andorka R. – Kolosi T. – Vukovich Gy. (szerk.) *Társadalmi Riport, 8–9.* Budapest: TÁRKI.

Angelusz R. – Tardos R. (2009) Demoszkópiai reprezentativitás, és demokratikus reprezentáció. Módszertani problémák és tartalmi dilemmák. In Enyedi Zs. (szerk.) *A népakarat dilemma.* Budapest: Demokrácia Kutatások Magyar Központja Közhasznú Alapítvány, 293–327.

Baker, R. – Brick, M. – Bates, N. – Battaglia, M. – Couper, M. – Dever, J. – Gile, K. – Tourangeau, R. (2013) Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology,* 1(2), 90–143. https://doi.org/10.1093/jssam/smt008

Biemer, P. – Lyberg, L. (2003) Coverage and Nonresponse Error. In Biemer, P. – Lyberg, L. (szerk.) *Introduction to Survey Quality.* Hoboken, New Jersey: John Wiley & Sons, 63–115.

Dillman, D. – Christian, L. (2005) Survey mode as a source of instability in responses across surveys. *Field Methods,* 17(1), 30–52. https://doi.org/10.1177/1525822X04269550

Davern, M. (2013) Nonresponse Rates are a Problematic Indicator of Nonresponse Bias in Survey Research. *Health Services Research,* 48(3), 905–912. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3681235/

Duffy, B. – Smith, K. – Terhanian, G. – Bremer, J. (2005) Comparing Data from Online and Face-to-face Surveys. *International Journal of Market Research,* 47(6), 615–639. https://doi.org/10.1177/147078530504700602

Groves, R. – Fowler, F. – Couper, M. – Lepkowski, J. – Singer, E. – Tourangeau, R. (2009) *Survey Methodology.* Hoboken, New Jersey: John Wiley & Sons. 347–360.

Havasi É. (1997) Válaszmegtagadó háztartások. *Statisztikai Szemle,* 75(10), 831–843.

Kmetty Z. (2012) A telefonos kutatások speciális problémái. *Statisztikai Szemle,* 90(1), 41–63.

KSH (2021) Az információs és kommunikációs eszközhasználat főbb jellemzői a háztartásokban. Elérhető: https://www.ksh.hu/docs/hun/xftp/idoszaki/ikt/2019/02/ikt_2019_02.pdf [Letöltve: 2023-10-11].

Loosveldt, G. – Sonck, N. (2008) An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods,* 2(2), 93–105. https://doi.org/10.18148/srm/2008.v2i2.82

Marsden, P. – Wright, J. (2010) *Handbook of survey research.* Bingley, UK: Emerald.

Messing V. – Ságvári B. – Szeitl B. (2022) Webre terelés a személyes lekérdezés alternatívája?: Egy „push-to-web" hibrid survey tapasztalatai. *Statisztikai Szemle,* 100(3), 213–233. https://doi.org/10.20311/stat2022.3.hu0213

Meyer, B. – Wallace M. – Sullivan, J. (2015) Household Surveys in Crisis. *Journal of Economic Perspectives,* 29(4), 199–226. https://doi.org/10.1257/jep.29.4.199

Meyer, B. – Mok, W. – Sullivan, J. (2015) Household Surveys in Crisis. *NBER Working Paper No. 21399.* Elérhető: https://www.nber.org/papers/w21399 [Letöltve: 2023-10-11].

Peyre, H. – Leplège, A. – Coste, J. (2011) Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research,* 20, 287–300. https://doi.org/10.1007/s11136-010-9740-3

Pew Research Center (2016) Evaluating Online Nonprobability Surveys. Online Report. Elérhető: https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2016/05/Nonprobability-report-May-2016-FINAL.pdf [Letöltve: 2023-10-11]

Rässler, S. – Riphahn, R. (2006) Survey item nonresponse and its treatment. *Allgemeines Statistisches Archiv,* 90, 217–232. https://doi.org/10.1007/s10182-006-0231-3

Rudas T. (2006) *Közvélemény-kutatás. Értelmezés és kritika.* Budapest: Corvina.

Szeitl B. – Tóth I. Gy. (2020) Hova tovább a nemválaszolókkal? A European Social Survey alapján végzett módszertani kísérlet eredményei. *Szociológiai Szemle,* 30(3), 96–114. https://doi.org/10.51624/szocszemle.2020.3.5