7. SÚLYOZÁS ÉS AZ ALKALMAZANDÓ SÚLYVÁLTOZÓK

KAPITÁNY BALÁZS

A „súlyozás" kifejezést a magyar társadalomtudományi szaknyelvben sok különféle módszerre, eljárásra szokták alkalmazni. Ebben a rövid ismertetőben súlyozásnak azt az adatbázis-szintű „manipulációs" folyamatot nevezzük, amelynek az a célja, hogy az egyes – mintavételen alapuló – adatbázisokat úgy módosítsa, hogy a módosítás hatására a kapott eredmények az alapsokaság ismert vagy feltételezett arányaihoz jobban igazodjanak.

Ez a gyakorlatban azt jelenti, hogy minden egyes válaszadóhoz egy konkrét 1 körüli – de azzal jellemzően nem egyező – értéket rendelünk. Ezt az értéket az adatbázisban az úgynevezett súlyváltozó tartalmazza. (A súlyváltozók neve az *Életünk fordulópontjai* adatbázisaiban mindig „s" betűvel kezdődik. A változónevek kialakítását részletesen lásd a 6. fejezetben.) Ezt követően a statisztikai számítások esetén az egyes válaszadók által adott válaszok értékét a megfelelően beállított statisztikai program már nem egyenlő mértékben, hanem a súlyozott értékkel korrigálva veszi figyelembe. Ez az eljárás képes garantálni, hogy az elemzés során az adatbázis tényleges nagysága megmaradjon, de a kapott eredmények jobban közelítsék a populáció ismert (vagy feltételezett) valós értékeit.

Fontos hangsúlyoznunk, hogy a különféle adatbázisokhoz, sőt akár ez egyes kutatási kérdésekhez is különféle súlyok készülhetnek, például annak függvényében, hogy mely kérdések kapcsán nagyon fontos elvárás a minta minél pontosabb illeszkedése. Például egy demográfiai elemzés esetén igen fontos lehet a minta családi állapot szerinti pontossága, ezzel szemben egy területi egyenlőtlenségeket vizsgáló kutatás kapcsán a településcsoportok, rétegek teljesen pontos mintareprezentációja lehet az elvárás. Emellett akár speciális – az egyes országok méretét is figyelembe vevő – súlyok készülhetnek nemzetközi összehasonlításra, vagy háztartási súlyok olyan elemzésekre, amelyek a háztartási arányra készítenek becslést, stb. Az elemzéshez tehát a megfelelő súly kiválasztása nagyon fontos döntés, amely az eredményeket is sok esetben érdemben befolyásolja.

A következőkben az *Életünk fordulópontjai* (ÉF) program leginkább széles körben használatos adatbázisai kapcsán mutatjuk be az adott adatbázisok esetén jellemzően használandó súlyok készítésének módját. E rövid ismertetés célja, hogy az adatbázisok felhasználói számára röviden bemutassa ezen súlyok képzésének elveit. A különféle súlyok kialakításának részletes leírása megtalálható az egyes hullámok súlyozását elkészítő módszertani szakemberek által

MURINKÓ LÍVIA – SPÉDER ZSOLT (SZERK.) 2016: *FELHASZNÁLÓI KÉZIKÖNYV AZ ÉLETÜNK FORDULÓPONTJAI PANELKUTATÁS 1–4. HULLÁMÁHOZ.* KUTATÁSI JELENTÉSEK 97. KSH NÉPESSÉGTUDOMÁNYI KUTATÓINTÉZET, BUDAPEST, 77–80.

írt önálló tanulmányokban (pl. Szelényi 2003, Bartus 2015). A ritkábban használt adatbázisok elemzésekor használt súlyozási eljárásokról pedig a vonatkozó szakpublikációkban (pl. az erdélyi adatbázis esetén: Kiss – Kapitány 2009; a szülő-gyermek adatbázis esetén: Kapitány 2012) talál részletesebb ismertetést az érdeklődő felhasználó.

AZ ELSŐ HULLÁM KERESZTMETSZETI SÚLYA (S1SULY)

Az adatbázis alapsúlya a 2001. évi népszámlálás adatbázisához illeszti az eredményeket, figyelembe véve a válaszadó korcsoportját (5 kategória), nemét (2 kategória), iskolai végzettségét (3 kategória), családi állapotát (4 kategória), illetve a település jogállását (3 kategória) (összesen 480 súlycella). A népszámlálási adatok mintegy 8 hónappal megelőzték az adatgyűjtésünket, így a korra, családi állapotra és iskolai végzettségre vonatkozó adatok esetén a válaszadónak a népszámlálás eszmei időpontjában érvényes állapotát vettük figyelembe. Mivel a családi állapot jelentősen összefügg az életkorral, több cella üresen maradt, illetve nagyon kis elemszámú volt. Például igen kevés fiatal özvegy férfi van. Ilyen esetekben összevonásokra került sor, ezért a peremeloszlások nem illeszkednek tökéletesen. A kapott súlyok 0,66 és 1,36 közé estek úgy, hogy 87%-uk esett 0,7 és 1,3 közé. Ezek kifejezetten kedvező – alacsony szóródású – értéknek számítanak, figyelembe véve a súlycellák igen magas számát. A súlyok kis szóródása arra utal, hogy az első hullám nyers adatbázisának az előzetesen torzított mintavétel segítségével – lásd a 2. fejezetet – meglehetősen jól sikerült közelítenie a kiinduló sokaság néhány fontos jellemzőjét.

A NEGYEDIK, KIEGÉSZÍTŐ MINTÁVAL BŐVÍTETT HULLÁM KERESZTMETSZETI SÚLYA (S4SULY)

A súlyt az első hullámhoz módszertanában nagyon hasonló módon, megegyező dimenziók által képzett súlycellák felhasználásával állítottuk elő, csak itt referenciaként a 2011. évi népszámlálás adatállománya és eredményei szolgáltak. A feladatot valamivel összetettebbé tette, és a súlyok nagyobb szóródását eredményezte, hogy a régóta követett longitudinális alminta a nem véletlenszerű lemorzsolódások miatt – lásd az 5. fejezetet – 2012-re már valamennyire torzult. A súlyok mértékének szóródását azonban az is jelentősen növelte, hogy a súlyok megállapításakor figyelembe kellett venni azt a tényt, hogy jelentős átfedés volt a régóta követett és a kiegészítő minta korcsoportjai között. Így egyes korcsoportok kétszer is bekerülhettek a mintába. A súlyozás folyamán ezt a „duplázódást" valamilyen módon helyre kellett állítani. Emiatt a „duplázódott" korcsoportok súlya már kiinduláskor alacsonyabb volt, míg az „egyszer mért" csoportok súlya magasabb lett. Így a súlyok 0,42 és 3,43 között szóródnak.

LONGITUDINÁLIS SÚLYOK (S2LONGI, S3LONGI, S4LONGI)

Teljesen más logikát követ a longitudinális adatbázis súlyozása, a longitudinális elemzésekben használatos súlyok előállítása (lásd részletesen Bartus 2015). Itt egyfelől problémát jelent, hogy a két népszámlálás között a referencia-adatok nem állnak a kutatók rendelkezésére megfelelően részletes bontásban. Például a népszámlálások között nem áll rendelkezésre

információ az iskolai végzettség módosulásáról az egyes társadalmi csoportokban vagy családi állapot szerint. Tehát nincs mihez „hozzásúlyozni" az adatbázist. Ugyanakkor a követéses vizsgálatok lehetőséget adnak egy másik súlyozási eljárásra. Ennek az a lényege, hogy az előző hullámból igen részletes információk állnak rendelkezésünkre mindazokról a válaszadókról, akik lemorzsolódtak. Sőt, a lemorzsolódás okát is meglehetősen jól ismerjük, tehát figyelembe tudjuk venni, kik léptek ki az alapsokaságból (jelen esetben a meghaltakat soroltuk ide), illetve kiknek „kellett volna" válaszolniuk (akik megtagadták a választ, „eltűntek" stb.). Ezen információkra alapozva lehetett előállítani a longitudinális elemzésre alkalmas súlyokat, a következő lépések alapján:

1. Vesszük az első hullám kiinduló keresztmetszeti súlyát.

2. Ezután többváltozós logisztikus regressziós modellek segítségével (bevont változók: születési kohorsz, nem, a település régiója és jogállása, iskolázottság, családi állapot és a háztartásban élő, még nem iskoláskorú gyermek(ek) jelenléte) modellezzük az első és második hullám közötti lemorzsolódás esélyét. A nyers longitudinális súly a következő hullámban történő sikeres válaszadás valószínűségének reciproka. Vagyis ez az eljárás azokat a mintában maradt válaszadókat „értékeli fel", akiknek a tulajdonságaik alapján egyébként nagyobb eséllyel kellett volna lemorzsolódniuk.

3. A súlyozási eljárás következő lépése a kalibrálás (vagy utólagos rétegzés). Ennek lényege, hogy a kiinduló keresztmetszeti súly és a nyers longitudinális súly szorzataként létrejött súlyokat olyan módon korrigáljuk, hogy a minta nem, születési év és a lakóhely régiója szerinti (háromdimenziós) eloszlása illeszkedjen a népszámlálásból továbbvezetett KSH-adatokhoz. Ez a kalibrálási eljárás gyakorlatilag megegyezik egy egyszerű mátrixsúlyozás módszerével, azzal az eltéréssel, hogy itt egy csoporton belül a súlyok szórása már nem nulla. Így kialakult a második hullám nyers kalibrált longitudinális súlya.

4. Ezután a 2. és a 3. lépést megismételjük a másodikról a harmadik és a harmadikról a negyedik hullámra való továbblépés esetén, így kialakítva a harmadik és a negyedik hullám nyers kombinált longitudinális súlyait is.

5. Legvégül a kapott nyers kalibrált longitudinális súlyokat cenzoráltuk, vagyis a 2,5-ös értéknél magasabb súlyokat 2,5-re csökkentettük. (A kalibrált súly a 2–4. hullámoknál rendre 41, 87 és 132 esetben haladta meg a küszöbértéket.) Mivel azonban a súlynak normalizáltnak (is) kell lennie, ezt követően normalizáltuk a súlyok értékét. Ez azonban kismértékben megnövelheti a 2,5-ös értékeket. A cenzorálás ezért iteratív eljárás: a tulajdonképpeni cenzorálás után a súlyt normalizáltuk, majd ismét cenzoráltunk. Az eljárást addig folytattuk, míg a normalizált súlyok maximuma pontosan 2,5 lett.

Ennek az eljárásnak az eredményeképpen alakultak ki a longitudinális elemzésekhez használatos longitudinális súlyok (s2longi, s3longi, s4longi), amelyek közül mindig a legutolsó, az adott elemzésbe még bevont adatfelvételi hullám súlya a használatos.

HIVATKOZÁSOK

Bartus Tamás (2015): Lemorzsolódás és súlyozás az Életünk fordulópontjai panelfelvételben. *Demográfia*, 58(4), 287–308.

Kapitány Balázs (2003): A mintavétel és a nyers adatok megbízhatósága. In Kapitány Balázs (szerk.): *Módszertan és dokumentáció az „Életünk fordulópontjai" című demográfiai követéses vizsgálat első hullámának adatfelvételének ismertetése*. Műhelytanulmányok 2. KSH Népességtudományi Kutatóintézet, Budapest, 29–39.

Kapitány Balázs (2012): A hátrányos társadalmi helyzetek generációk közötti átörökítése: Egy magyarországi követéses vizsgálat eredményei. *Esély*, 23(2), 3–37.

Kiss Tamás – Kapitány Balázs (2009): Magyarok Erdélyben: A minta kialakítása és az adatfelvétel. In Spéder Zsolt (szerk.): *Párhuzamok. Anyaországi és erdélyi magyarok a századfordulón*. KSH NKI Kutatási Jelentések 89. KSH Népességtudományi Kutatóintézet, Budapest, 55–70.

Szelényi Barbara (2003): Az adatbázis súlyozása. In Kapitány Balázs (szerk.): *Módszertan és dokumentáció az „Életünk fordulópontjai" című demográfiai követéses vizsgálat első hullámának adatfelvételének ismertetése*. Műhelytanulmányok 2. KSH Népességtudományi Kutatóintézet, Budapest, 64–81.