Vállalkozáselmélet és –gyakorlat Doktori Iskola
MISKOLCI EGYETEM
Gazdaságtudományi Kar

Szilágyi Roland

Mintavételen alapuló becslések hibáinak kezelése különös tekintettel a nemválaszolás okozta problémákra

Ph.D. értekezés

A Doktori Iskola vezetője:

Prof. Dr. Szintay István
egyetemi tanár

Témavezető:

Prof. Dr. Besenyei Lajos
egyetemi tanár

Miskolc, 2011.

Köszönöm a szakmai útmutatást és segítséget
témavezetőmnek, Prof. Dr. Besenyei Lajosnak,
közvetlen munkahelyi vezetőmnek, Dr. Varga Beatrixnak,

előopponenseimnek, Dr. Rappai Gábornak és Dr. Telegdi Lászlónak;

tanszéki kollégáimnak, különösképpen Kassai Editnek;

emellett
Dr. Szép Katalinnak, Mihályffy Lászlónak, Horváth Beátának és Horváth Gergelynek;

valamint a támogatást és megértést Családomnak!

TARTALOMJEGYZÉK

ÖSSZEFOGLALÓ

A mintára épülő vizsgálatok és következtetések egyre nagyobb szerepet kapnak a gazdasági döntések meghozatalában és az információ képzésben egyaránt. A mintavételek terjedését elsősorban a költségek és a vizsgálathoz szükséges idő csökkentése indukálja. A mintán alapuló felmérések nem csak mikro szinten, hanem makrogazdasági vizsgálatoknál is egyre népszerűbbek, de a mintavételek terjedésének azonban nagy veszélye is van, pontosan a minta minősége miatt.

Kutatási munkám elsődleges céljának a mintavételen alapuló kutatások lehetséges hibáinak, és azok eredményekre gyakorolt negatív hatásainak feltárását jelöltem meg. Ezt követően megoldási változatokat kerestem a hibák, majd elsődlegesen a nem véletlen jellegű hibák kezelésére. A hibák kezelésének leghatékonyabb módja, ha megelőzzük a keletkezésüket. Azonban ha a hiba már bekövetkezett, akkor a kezelés első fázisaként fel kell térképezni a hiba okát.

Munkám során 53 különböző mintavételi eljárás alapján generáltam mintákat a sokaságból – a Háztartási Költségvetési Felvétel adatbázisából –, annak érdekében, hogy minél részletesebben vizsgálhassam a mintavételi tervek eredményre gyakorolt hatását. Következtetéseim, becslési eredményeim ellenőrzésére lehetőséget biztosított az, – a gyakorlatban nem teljesülő feltétel – hogy a vizsgált jelenség sokasági információi a birtokomban voltak.

Ezt követően olyan szempontok kidolgozására vállalkoztam, amelyek alapján – ha nem is a minőség összes értelmezhető kritériuma tekintetében, de néhány fontosabb jellemző alapján – minősíteni, rangsorolni lehet a különböző mintákból nyert adatokat, becslési eredményeket.

A hazai és nemzetközi kutatások tapasztalatai alapján egyaránt elmondható, hogy a válaszadás hiányossága talán az egyik legnagyobb probléma, ami a felmérések készítésénél felmerül.

Kutatásom utolsó fázisában a modell alapú eljárások szerepét teszteltem a nemválaszolás kezelésében.

A nemválaszolás okozta torzítás kiküszöbölésére tett lépések között fontos szerepe van a tendenciák azonosításának. Érdemes megvizsgálni a válaszadók és nemválaszolók vizsgált ismérvbeli tendenciái közötti különbséget. Ezek felhasználásával megalkottam a súlyozott tendenciák becslési modelljét, mely a torzítás mértékét megfelelő keretek közé szorítja.

SUMMARY

Research based on samples and their conclusions play an increasing role in making business decisions and also in creating information. The spread of sampling is mainly due to the lower expenses and shorter time needed for an investigation. Surveys based on samples are becoming more popular not only on a micro-level, but also in case of macroeconomic investigations. However, the spread of sampling has a great risk because of the quality of the samples.

As the priority of my work I chose the exploration of the possible faults of surveys based on samples and the negative effect they have on the result. Consecutively,I searched for solution variants for handling of mistakes, primarily for non-random errors. The most effective way of handling mistakes is to prevent them from happening. If the error has already occurred, then as the first phase of the treatment we have to map the reasons for the mistake.

During the process I generated samples for the population from the Household Budget Survey (HBS) database - using 53 different sampling methods in order to study in more detail the effect of sampling plans on the results. This was to check my conclusions and estimates, as I had the population information (which is not given in real practice).

Furthermore, I aimed at working out points of view based on which results and estimates gained from several samples can be ranked. Of course I tried this in the case of several hypothetical conditions.

Based on domestic and international research findings, we can state that the biggest problem when carrying out surveys is when answers are not given.

The last phase of my research was testing the role of model-based approaches in handling non-response. Identifying tendencies plays a significant role in eliminating the bias caused by non-response. It is worth examining the differences between response and non-response criterion tendencies. In order to achieve success the sample should be grouped based on different criteria. Using these methods I created a model for estimating weighing tendencies, which limits the degree of bias to within an acceptable range.

„Gondoljunk arra, hogy tévedni emberi, megbocsátani Isteni dolog – a tévedést beépíteni a tervembe, ez viszont statisztikusi dolog”

(Leslie Kish)

1. BEVEZETÉS

1.1. A KUTATÁSI PROBLÉMA MEGFOGALMAZÁSA

A piaci szerkezet módosulása, a gyors környezeti változások, az információs technikák fejlődése eredményeképpen a döntések előkészítéséhez szükséges időtáv jelentősen lerövidült. Az üzleti gazdasági élet döntéshozói csak úgy vehetik fel hatékonyan a versenyt – az idővel és a versenytársakkal egyaránt – ha folyamatosan fejlesztik a döntés-előkészítés során alkalmazott technikákat. A tömegesen előforduló jelenségek jellemzésére irányuló kvantitatív elemzési módszerek széles tárháza nyújt lehetőséget az egyedek jellemzőinek vizsgálatára. Egy sikeres kutatásnak azonban csak egyik kulcspontja a megfelelő kvantitatív módszer megválasztása és tudományos alkalmazása. Másik kulcspontja mindenképp a felhasznált adatok információtartalmának megbízhatósága kell, hogy legyen.

A mintára épülő vizsgálatok és következtetések egyre nagyobb szerepet kapnak a gazdasági döntések meghozatalában és az információ képzésben egyaránt. A mintavételek terjedését elsősorban a költségek és a vizsgálathoz szükséges idő csökkentése indukálja. A mintán alapuló felmérések nem csak mikro szinten, hanem makrogazdasági vizsgálatoknál is egyre népszerűbbek, jó példa erre Franciaországban a klasszikusan teljes körű népszámlálás adatgyűjtésének mintavétellel történő helyettesítése. Mikro szinten pedig egyértelműen az időmegtakarítás, a gyorsabban nyerhető információk alapján a döntések előkészítéséhez szükséges idő csökkentése fokozza a mintavétel alkalmazását.

A mintavételek terjedésének azonban nagy veszélye is van, pontosan a minta minősége miatt. A mintán alapuló adatokból nyert információk nagyon sok és sokféle hibát hordozhatnak magukban. Ezeknek a hibáknak a feltárása és matematikai-statisztikai módszerekkel történő értékelése, valamint az eredményre gyakorolt negatív hatásuk csökkentése képezi kutatásom tárgyát. Célom olyan elemzési módszer, szempontrendszer kidolgozása,

mely alkalmazkodik a modern információs technikák és technológiák nyújtotta lehetőségekhez, és jól kiegészíti az eddig alkalmazott módszertant.

A vállalatok, a mindenkori kormány, és a gazdasági-, társadalmi élet egyéb szereplői növekvő számban készítenek, igényelnek statisztikai információkat, melyek nem minden esetben megalapozottak. Kutatásom fontos céljának tartom mindenekelőtt a döntéseket sikeresen támogató elemzések megfelelő alapjának megteremtését.

Milyen okokra vezethetők vissza a társadalomtudományi kutatások hibái? Melyek lehetnek azok az eredő tényezők, amelyek alapján egy kutatás, vagy annak eredménye „kontár" jelzővel illethető? Ezekre a kérdésekre a szakirodalom a következő válaszokat adja: megbízhatatlan, megalapozatlan, kis mintára épül, magas a relatív hiba, nem szignifikáns az eredő problémára irányuló hatása, stb. További kritikai tényezők merülhetnek fel abban az esetben, amikor a kutatás egy mintavétel során nyert primer adathalmazt dolgoz fel, elemez, és von le különböző következtetéseket. Ennek oka, hogy a mintavétel során a hibaforrások száma fokozódik. Ez a tény korántsem jelenti, hogy a dolgozatban említett hibákat a kutatók feltétlenül elkövetik, illetve, hogy e hibák következményeinek, a kutatás eredményére gyakorolt negatív hatása ne lenne mérsékelhető.

A mintavétel során elkövethető hibák már az előállított forrásadatok mennyiségében és minőségében is torzulást eredményezhetnek, ami azért veszélyes, mert a rossz alapadatokból a legjobb módszertan alapján végzett alapos számítások is téves következtetésekhez vezetnek.

1.2. HIPOTÉZISEK

A kutatással és munkával töltött évek alatt sok negatív tapasztalatot szereztem az alkalmazott mintavétellel, a következtetésekkel, illetve a módszertani indoklásokkal kapcsolatban. Ezek egy része abból fakad, hogy a kutatók talán éppen óvatosságból – annak érdekében, hogy a módszertani hiányosságok okozta hibákat minimalizálják – nem alkalmazzák az összetettebb mintavételi eljárásokat, megmaradva ezáltal az egyszerű mintavétel nyújtotta lehetőségeknél. Másik részük a mintából nyert adatokra épülő következtetések hiányosságaira vezethető vissza. Olyan információkra, melyek rejtve maradnak az elemzések eredményeit felhasználók előtt, (például a nemválaszolás okozta torzítás mértéke).

A disszertációban a következő, a mintavételes felmérések alapjait érintő hipotéziseket vizsgáltam. Ezek ellenőrzése közben állításaimat a KSH (Központi Statisztikai Hivatal) háztartásokra vonatkozó adatbázisából képzett minták adatainak elemzésével támasztottam alá. A hipotézisek vizsgálata közben kapott eredményeket, megállapításokat tézisek formájában összegeztem, melyek az empirikus vizsgálatokat tartalmazó fejezetekben kerültek elhelyezésre.

- Mind hatásosság, mind pontosság szempontjából relevánsabb eredményeket biztosítanak azok a részletesebb információk alapján képzett minták, melyeknél a rétegzéshez több, a vizsgált változóval sztochasztikus viszonyban álló változó kerül bevonásra a mintavételi terv kidolgozásában,
- A mintavételi tervnek a becslési eredményekre gyakorolt hatását megtestesítő Deff mutató az egyszerű véletlen mintához képest kevésbé hatékony mintavételi tervek hatását méri, azonban azonos méretű, azonos változóra vonatkozóan azonos becslőfüggvény esetében nem alkalmas egyértelműen a minták közül eldönteni melyik a hatásosabb.
- A vizsgált mintaelemek (háztartások) hasonlóságot mutatnak különböző változók, leginkább demográfiai és a fogyasztással összefüggésbe hozható változók mentén. A hasonlóságokra épülő klasszifikációs módszerek alkalmazása segít a nemválaszoló egyedek okozta információ veszteség csökkentésében.
- A mintavételi terv minősége hatással van a becslés pontosságára, hatásosságára, megbízhatóságára. Ezért feltételezhető, hogy az alaposan, előrelátóan megtervezett mintavétel csökkenti a becslési eredmények nemválaszolás okozta torzításának mértékét.
- A kiegészítő információk alkalmazása segítséget nyújt a hibák mértékének csökkentésében. A kutatóknak azonban nem mindig van lehetősége külső információk beépítésére, ezért a belső (mintabeli) információkat kell a lehető legnagyobb mértékben kiaknázni. A mintaegyedek megfelelő részletességű csoportosításával a válaszadói csoportokban megfigyelhető tendencia kivetíthető a teljes mintára, ezáltal a nemválaszoló egyedekre. A tendenciák modellezésével a nemválaszolás torzító hatása csökkenthető.

1.3. ALKALMAZOTT MÓDSZERTAN

Dolgozatomban nemcsak a mintavételen alapuló felmérések potenciális hibáinak meghatározásával, az egyes hibatípusok egyszerű bemutatásával foglalkozom, hanem olyan eljárásokat dolgozok ki, melyek a hiba mértékének minimálisra csökkentését, a torzítás mérséklését teszik lehetővé. Mindezeket olyan általánosított formában teszem, hogy hasznos segítségül szolgáljanak a társadalomtudomány területén tevékenykedő egyéni kutatók és kutató szervezetek számára egyaránt.

A mintabeli adatok alapján készített elemzések hibáinak vizsgálatára nagyméretű adatbázisok a legmegfelelőbbek. Anyagi források hiányában az elemzést megbízható szekunder adatokon végeztem. Azonban szükségképpen foglalkoztam az adatgyűjtés, megfigyelés és mérés különböző módszereivel, hiszen azok más-más típusú hibákat generálhatnak.

Az adatelemzés során a HKF (Háztartási Költségvetési Felvétel) adatbázisán végeztem kutatásokat. Ennek keretében a meglévő eredményeket egy újszerű logikai struktúra alapján csoportosítottam. Olyan módszereket alkalmaztam, melyek nemcsak makroszinten, hanem akár kis vállalkozások szintjén is hasznosíthatók. Hiszen a vállalkozások általában nincsenek olyan hibaszámítási szoftverek és algoritmusok birtokában, mint a hivatalos statisztika képviselői. Megfelelő módszerek hiányában pedig nem képesek jó minőségű információk előállítására. Ennek az űrnek a kitöltését célozzák az empirikus kutatásom módszertani eredményei.

A feldolgozás során az alkalmazott bonyolult és időigényes módszerek, matematikai számítások elvégzése, illetve a grafikus ábrák, táblázatok szemléletesebb formában történő megjelenítés érdekében a Windows alapú SPSS 17.0 statisztikai szoftvert, valamint a Microsoft Excel táblázatkezelő szoftvert használtam.

A kutatásmódszertan elméleti követelményeinek megfelelően praktikusan itt kellene elhelyezni a disszertációban használatos fogalmak felsorolását, szintetizálását, de mivel a dolgozat témája szorosan matematikai-, statisztikai módszertanhoz kapcsolódik, így az alkalmazott definíciók egyértelműen determináltak. Az esetleges specializációk részletes ismertetésére a dolgozat megfelelő fejezeteiben kerül sor.

1.4. A DOLGOZAT FELÉPÍTÉSE

Jelen dolgozat a mintavételen alapuló felmérések hibáinak feltárásával és azok kezelésével foglalkozik.

A bevezetés, a kutatási célok és módszerek felvázolása után a második fejezetben áttekintem a statisztikai következtetések elméleti hátterét. Valamint ismertetem azokat a mintavételi eljárásokat, hibaszámítási módszereket, amelyek alkalmazásra kerülnek a dolgozatban. Ezek egy része ugyan nem haladja meg jelentősen az egyetemi képzés standard tananyagát, de mindenképpen fontosnak tartom megjeleníteni, mivel a dolgozat empirikus részeinek alappilléreit képezik és strukturálisan nélkülözhetetlenek a mintavételes vizsgálatokra vonatkozó megállapítások kidolgozásában.

A harmadik fejezetben a gyakran elkövetett potenciális hibaforrásokat ismertetem, emellett jelen dolgozat témájával kapcsolatos általam fontosnak ítélt, valamint a szakterület jeles képviselői szerint irányadónak tekintett korábbi kutatások eredményeinek összefoglalására kerül sor. Számos könyv és tanulmány készült – ezen meglehetősen összetett probléma egyes részleteinek feltárására –, amelyek nemcsak elméleti, hanem gyakorlati oldalról, nemzetközi tapasztalatokra építve mutatják be a kérdéskör alapjait. A hazai és külföldi szakirodalom áttekintése során többnyire különféle folyóiratokban, illetve konferenciákon megjelent publikációkra támaszkodtam. A hazánkban megjelent vonatkozó irodalomban az elméleti megközelítések bemutatása meglehetősen specializált, kevésbé foglalkozik általánosságban a témával, túlnyomórészt parciális problémákat érintő vizsgálatok láttak napvilágot. Szintén ebben a fejezetben kapott helyet a – hibák gyakori forrását jelentő – mintaméret meghatározását segítő informatikai alkalmazás bemutatása.

A negyedik fejezetben a vizsgált adatbázis és a kialakított minták rövid bemutatása után a megfelelő mintavételi terv megválasztásának hiányából fakadó hibát szemléltetem, valamint annak relatív mértékének meghatározásával foglalkozom, ezek alapján rangsorolva a különböző mintavételi terveket.

Az ötödik fejezetben a meghiúsulás okozta problémát, és ennek a nem mérhető, de igen jelentős hibának a hatását és mérséklési lehetőségeit vizsgálom különböző a (felsőoktatásban is oktatott) módszerek segítségével.

A hatodik fejezetben a nemválaszolók mintán belüli adatokból történő azonosításának lehetőségeit tárom fel. A nemválaszolás különböző szintjei mellett érvényesülő tendenciák felhasználásával alkotott modell segítségével a nemválaszolók okozta torzítás csökkentésének lehetőségét mutatom be.

A dolgozatban gyakorlati, praktikai szempontok érvényesítésére törekszem, a végeredmények minél szélesebb körű felhasználási lehetőségét tartva szem előtt. Ezért a hipotézisek vizsgálatát, az állítások helytállóságát empirikus úton bizonyítom.

2. KÖVETKEZTETÉSELMÉLETI ALAPVETÉSEK

A statisztikai elemzések módszereinek két nagy csoportja különböztethető meg:

- leíró statisztika eszközei,
- statisztikai következtetések módszerei.

A leíró statisztika eszköztára olyan adathalmazok, sokaságok jellemzésére használatos, melyeknek minden egyes eleme ismert. Így az elemzések eredményeiből levonható következtetések kizárólag a vizsgált egyedek összességére vonatkoznak, és nem érvényesek azon túl más egyedekre.

A következtetéses statisztikai módszerek lehetőséget nyújtanak a megfigyelt adatok alapján következtetések levonására arra a populációra vonatkozóan, ahonnan az adatok származnak. A következtetés elméleten belül két típus különböztethető meg: a klasszikus és a bayesi következtetéselmélet.

A bayesi következtetéselméletben nemcsak a mintából nyert információk kerülnek felhasználásra, hanem azokat kiegészítik előzetes (prior) információkkal. A dolgozat empirikus kutatásokat tartalmazó fejezeteiben kizárólag a mintából nyert információk kerülnek felhasználásra. Ezért a klasszikus következtetéselmélet módszereire összpontosítok.

„A következtetéses statisztikai módszerek célja, hogy az alapsokaságból megfelelően kiválasztott részsokaság (minta) alapján

- közelítő értéket nyerjünk az alapsokaság valamely jellemző paraméterére (pl. várható értékére, szórására, egy kitüntetett ismérvváltozat előfordulásának valószínűségére, két ismérv várható értékének különbségére vagy hányadosára); vagy
- döntsünk az alapsokaságra megfogalmazott valamilyen állítás (előfeltevés, prekoncepció) igazságtartalmáról; vagy
- meghatározzuk az alapsokaság jellemző összefüggéseit leíró oksági kauzalitási) modellek formáját.” Pintér – Rappai (2007. p.272.)

A klasszikus következtetéselmélet Maddala (2004.) szerint két feltevésre épül:

- a mintabeli adatok minden lényeges információt tartalmaznak,
- a különböző következtetési eljárások szerkesztése és értékelése azon alapul, hogy lényegében azonos körülmények között végbemenő hosszú távú viselkedést vizsgálunk.

Emellett azonban nem szabad figyelmen kívül hagyni, hogy a következtetéses statisztika eredményei mindig tartalmaznak valamilyen bizonytalanságot. Ez azonban a gyakorlati üzleti élet szereplői számára nem újdonság, hiszen döntéseiket nap, mint nap bizonytalan körülmények között hozzák. A bizonytalanság tényének és mértékének beépítése a statisztikai következtetésekbe és ezen keresztül a döntésekbe elengedhetetlen. A bizonytalanság kezelése azonban sokszor nehéz feladat a kutató számára.
A következtetéses statisztika módszereinek sikeres alkalmazása megköveteli a mintavételi ismeretekben és a valószínűségszámítási ismeretekben való jártasságot.

2.1. VALÓSZÍNŰSÉGSZÁMÍTÁSI ALAPOK

A következtetéselmélet feltételezése szerint a rendelkezésre álló adatokat egy ismeretlen folyamat állítja elő, és ez a folyamat leírható egy valószínűségeloszlással, amely bizonyos ismeretlen paraméterekkel jellemezhető. Például normális eloszlás esetében ilyen paraméterek a várható érték és a variancia. Mint ismeretes, a valószínűségszámítás alkalmazási területeinek egyike, amikor ismert egy valószínűségi változó eloszlása és ennek segítségével meghatározható a vizsgált jelenséggel kapcsolatos egyéb események bekövetkezésének valószínűsége. A valószínűség erre irányuló elméletei és összefüggései tehát plasztikusan alkalmazhatók a következtetéses statisztika céljának megvalósításához.
A következtetés elmélet alkalmazásához rendelkezésre álló adatokat valamilyen eljárással kiválasztott minta adatai, illetve a minta egyedeinek adatai jelentik. Mint azt a későbbiekben látni fogjuk, a mintaegyedek kiválasztása adott valószínűség mellett történik, ezáltal a minta egyedeinek valamely vizsgált ismérvre vonatkozó adatai valószínűségi változóként értelmezhetők. A statisztikai változók mintabeli értékei megfeleltethetők egy olyan véletlen változó reprezentációjának, amely leírható:

- értékeinek felsorolásával, és az egyes értékek bekövetkezési valószínűségének megadásával; vagy
- eloszlás-, illetve sűrűségfüggvényével.

A következtetéselmélet alkalmazásához tehát ismerni kell a nevezetesebb statisztikai eloszlásokat, melyeket itt részletesen nem mutatok be. Szükségesnek tartom azonban a dolgozat empirikus részeiben alkalmazott mintákra vonatkozó következő alapvetések rögzítését.

„Feltételezve, hogy az n független megfigyelésből álló (x_1, x_2, …, x_n) minta μ várható értékű és σ^2 varianciájú normális eloszlású sokaságból származik, illetve a minta átlaga:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ és}$$

korrigált tapasztalati varinaciája

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

a következő megállapítások rögzíthetők:

- A mintaátlag mintavételi eloszlása szintén normális, μ várható értékkel és σ^2/n varianciával
- Az $(n-1)s^2/\sigma^2$ kifejezés (n-1) szabadságfokú χ^2 eloszlást követ. Ezen felül $\bar{x}$ és s^2 eloszlásai függetlenek.
- Mivel $\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0,1)$ és $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, és ezek az eloszlások függetlenek, a $\frac{\sqrt{n}(\bar{x}-\mu)}{s}$ kifejezés (n-1) szabadságfokú t-eloszlást követ.
- Továbbá $E(\bar{x}) = \mu$ és $E(s^2) = \sigma^2$ ezért $\bar{x}$ és s^2 torzítatlan becslőfüggvényei μ-nek és σ^2-nek.” Maddala (2004. pp.57-58.)

Ezek a megállapítások és összefüggések független azonos eloszlású minta esetén érvényesek. Bár empirikus kutatásaim során nem FAE mintákkal dolgozom, azonban az általam alkalmazott képletek és összefüggések minden esetben ezekből vezethetők le követve a mintavétel elméletének szakirodalmi előírásait.

2.2. MINTAVÉTELI ISMERETEK

A reprezentatív megfigyelés sikere elsősorban azon múlik, hogy a megfigyelendő egységek kiválasztása milyen módon történik. A reprezentatív megfigyelés szervezésének első és alapvető lépése, hogy alkalmazásának előfeltételei biztosítva legyenek.

A reprezentatív megfigyelés során populációnak, vagy alapsokaságnak nevezik azt a sokaságot, melyre vonatkozóan következtetések levonására kerül sor, a kiválasztott elemek összességét pedig mintasokaságnak nevezik.

A mintasokaság akkor a legjobb, azaz akkor tükrözi legpontosabban az alapsokaság tulajdonságait, ha a kiválasztás során semmiféle tudatos szubjektív befolyásolás nem érvényesül, vagyis az alapsokaság minden egyes elemének egyenlő esélye van arra, hogy a mintasokaságba belekerüljön. Ha a kiválasztásnál a szándékosság vagy a részrehajlás alkalmazása megengedett, akkor a tudatosan alkalmazott kiválasztás előre befolyásolja a vizsgálat eredményeit, így a véletlen tömegjelenségekre vonatkozó valószínűségszámítási összefüggések nem alkalmazhatók. Éppen ezért alapvető szabály az, hogy a reprezentatív módszer alkalmazásánál a minták kiválasztását olyan módszerrel kell végezni, amely bármiféle szubjektív befolyásolás érvényesülését eleve kizárja. Attól függően, hogy ez a követelmény hogyan érvényesül a mintavételnél, több módszer különböztethető meg. Ezen módszerek közül csak az általam alkalmazottakat mutatom be, melyekre Köves Pál, Párniczky Gábor, Hunyadi László, Vita László, Mundruczó György, Éltető Ödön munkáiban megfogalmazott tulajdonságokat, feltételeket és definíciókat tartottam irányadónak, az alábbiak szerint.

Annak érdekében, hogy a minta adatainak elemzése során alkalmazni lehessen a valószínűségszámítás összefüggéseit, a mintát, és ezáltal a minta egyedeit véletlenszerűen kell megválasztani. A matematikai statisztika elmélete pontosan definiálja a véletlen mintára vonatkozó tulajdonságokat:

- A mintavételi eljárás alkalmazása egy adott véges sokaságra a különböző M_1, M_2, …, M_L minták olyan véges (L) számú halmazát eredményezi, hogy pontosan meg lehet mondani, az alapsokaság mely elemei tartoznak M_1-hez, M_2-höz, stb.
- Mindegyik lehetséges M_i mintához ismert P_i kiválasztási valószínűség tartozik.
- A mintavétel során a lehetséges M_i minták közül kerül kiválasztásra egy úgy, hogy M_i kiválasztási valószínűsége P_i legyen.
- Előre adva van, milyen statisztika alapján készül becslés a mintából és a módszernek olyannak kell lennie, hogy egy adott mintából egyetlen becslést eredményezzen a szóban forgó statisztika alapján. Éltető (1970. pp. 8-9.)

A véletlen mintavétel tehát olyan eljárás kell legyen, amelyik rendelkezik a fent megfogalmazott tulajdonságokkal, máskülönben a kapott minta adataira nem alkalmazhatóak a valószínűségszámítás törvényei, a matematikai statisztika módszerei.

2.2.1. EGYSZERŰ VÉLETLEN MINTA

A gyakorlatban is legtöbbször használt, illetve a további véletlen mintavételi eljárások alapjaként említhető az egyszerű véletlen mintavétel. A módszer alkalmazása során egy N elemből álló sokaság minden elemének egyenlő esélyt kell biztosítani a mintába való bekerülésre. Ez azonban sokféleképpen érhető el, pl. bonyolult mintavételi módokkal. Az ún. egyszerű véletlen mintavétel esetén még azt is biztosítani kell, hogy minden lehetséges n elemű mintának azonos legyen a kiválasztási valószínűsége.

„Az egyszerű véletlen minta elkészítéséhez komplett lista, ún. mintavételi keret szükséges; ennek összeállítása az első feladat. A következő lépés a mintanagyság meghatározása, amelyet a pontossági követelmények, a sokaság szóródása, valamint a rendelkezésre álló költségkeret determinálnak. A kiválasztás leggyakrabban tervezett véletlen módon történik, mely során egyenlő esélyt biztosítunk minden egyednek a mintába jutáshoz. Az egyenlő esély biztosítása természetesen ellentmond minden önkényes vagy tudatos válogatásnak.” Hunyadi – Vita (2002. p.280)

2.2.2. RÉTEGZETT MINTA

A mintavétel alapján kapott becslések megbízhatóságának növelésére szolgáló módszerek közül a gyakorlatban legáltalánosabban a rétegzést alkalmazzák. A rétegzés azt jelenti, hogy a szóban forgó heterogén alapsokaságot „L” számú csoportra – rétegre – kell bontani, úgy, hogy a rétegek elemidegenek legyenek és együttesen kiadják az egész alapsokaságot. Ezután az egyes rétegekben egymástól függetlenül kell végrehajtani a mintavételt és a rétegekre vonatkozó becslések egyesítése útján nyerhető becslés az egész alapsokaságra vonatkozóan. Ha a rétegeken belül a mintavétel egyszerű véletlen kiválasztással történik, akkor ez a mintavételi eljárás rétegzett véletlen kiválasztás. Éltető (1970. pp.31-32.)

A rétegzett kiválasztási eljárás alkalmazásakor valamilyen ismérv alapján csoportokra (rétegekre) kell felosztani a sokaságot. Hogy milyen ismérv alapján, az megfontolás tárgyát kell, képezze. Ugyanis akkor hatékony a rétegképző ismérv kiválasztása, ha az sztochasztikus kapcsolatban van a vizsgálat célját képező ismérvvel. A rétegek számának meghatározása viszonylag egyértelmű abban az esetben, ha a rétegképző ismérv nominális skálázású és véges számú ismérvváltozattal rendelkezik. Amennyiben a rétegképző ismérv arányskálán

mérhető, úgy a rétegek számáról objektív szakmai szempontok alapján kell dönteni, figyelembe véve, hogy a túl kevés számú réteg nem növeli a mintavétel hatékonyságát, a túlzottan nagy rétegszám pedig esetenként indokolatlanul sok többletfeladatot jelent az elemzés során.

A rétegképzés (csoportosítás) során arra kell törekedni, hogy a vizsgált ismérv szempontjából egynemű elemek azonos csoportba kerüljenek. Egy egyed csakis egy csoportba kerülhet, tehát a csoportok átfedés mentességét biztosítani kell. Ez után az egyes csoportokon belül egyszerű kiválasztást kell végrehajtani oly módon, hogy végeredményben az egyes csoportokból (rétegekből) kiválasztott elemek összessége a kívánt nagyságú minta legyen.

Amennyiben a rétegképző ismérve megválasztása megfelelő, úgy a kapott rétegek homogénebbek a teljes sokaság egyedeihez képest. Ami egyben azt is jelenti, hogy az egyes rétegekből pontosabb becslések készíthetők, melyeket egyesítve a sokasági paraméter is pontosabban becsülhető egy egyszerű véletlen mintához képeset.

A rétegzés pozitív tulajdonságai között kell említeni, hogy helyesen alkalmazva nem csupán a becslés pontosságán javít, hanem az egyes rétegekre külön-külön is végezhetők elemzések, ami újabb következtetések levonására ad lehetőséget.

Az empirikus kutatás kezdetén az elméletet a gyakorlatba átültetve a feladat számomra az volt, hogy az ismert teljes minta elemszáma hogyan kerüljön szétosztásra az egyes rétegek között, azaz mennyi legyen $n_1, n_2, ..., n_j, ..., n_L$? A probléma megoldására több, a későbbiekben ismertetésre kerülő elosztási tervet alkalmaztam.

2.2.2.1. Egyenletes elosztás

Az egyenletes elosztás során arra kell törekedni, hogy minden egyes rétegbe azonos számú mintaelem kerüljön, azaz $n_j = \frac{n}{L} = \bar{n}$. Az egyenletes elosztás előnyös tulajdonságai között említhető, hogy egyszerű, nem igényel bonyolult szervezési-tervezési előkészítést, végrehajtása kényelmes. Továbbá ha végeredményként az egyes rétegek jellemző paramétereire külön-külön is következtetéseket kell levonni (nem csak a rétegek összességére, pl. főátlag), akkor az egyenletes elosztás a későbbiekben tárgyalt elosztásnál kedvezőbb eredményekhez vezet. Ha a rétegek egyforma nagyságúak, akkor az egyenletes elosztás egyben arányos is lesz, így annak kedvező tulajdonságaival is rendelkezik. Hunyadi – Mundruczó – Vita (2000. p.304)

2.2.2.2. Arányos elosztás

Amikor a kiválasztás minden egyes rétegre nézve azonos kiválasztási arányszám mellett történik, úgy arányos rétegzésnek tekinthető. Az arányos elosztás lényege az, hogy a mintába a sokasági arányoknak megfelelően kell megválasztani az elemszámot.

Az arányos mintaelosztás sok előnyös tulajdonsággal rendelkezik. Végrehajtása ugyanis szintén egyszerű, másrészt a mintában ugyanazok a súlyarányok érvényesülnek, mint a sokaságban, azaz a minta összetétele megegyezik a sokaság összetételével, ami a későbbi számítások szempontjából kedvező tulajdonság. „Kedvező tulajdonsága az arányos elosztással kapott mintának az is, hogy ha a rétegenkénti sokasági szórásokat nem ismerjük, illetőleg azonosnak tekintjük, akkor az alapvető mutatók esetében optimálisnak tekinthető, azaz az ebből számított mutatók mintavételi hibája minimális. Ezért ez az elosztás a gyakorlat számára kivételesen fontos.” Hunyadi – Vita (2002. p.286.)

3. POTENCIÁLIS HIBAFORRÁSOK

3.1. A MINTÁN ALAPULÓ KUTATÁSOK HIBÁIRÓL ÁLTALÁNOSSÁGBAN

Milyen okokra vezethetők vissza a társadalomtudományi kutatások hibái? Melyek lehetnek azok az eredő tényezők, amelyek alapján egy kutatás, vagy annak eredménye megalapozatlannak minősül? Ezekre a kérdésekre a szakirodalom a következő válaszokat adja: megbízhatatlan, megalapozatlan, kis mintára épül, magas a relatív hiba, nem szignifikáns az eredő problémára irányuló hatása, stb.

További kritikai tényezők merülhetnek fel abban az esetben, amikor a kutatás egy mintavétel során nyert primer adathalmazt dolgoz fel, elemez, és von le különböző következtetéseket. Hiszen a mintavétel során a hibaforrások száma fokozódik. Ez a tény korántsem jelenti, hogy a következőkben említett hibákat a kutatók feltétlenül elkövetik, illetve, hogy e hibák következményeinek, a kutatás eredményére gyakorolt negatív hatása nem mérsékelhető.

A mintavétel során elkövethető hibák már az előállított forrásadatok mennyiségében és minőségében is torzulást eredményezhetnek, ami azért veszélyes, mert a rossz alapadatokból a legjobb módszertan alapján végzett alapos számítások is téves következtetésekhez vezethetnek. Mindemellett kijelenthető, hogy a mintán alapuló társadalomtudományi kutatások során végzett statisztikai adatfelvételek mindig tartalmaznak hibákat.

A teljes hiba mintavételi hibából és nem mintavételi hibából tevődik össze. A kétféle hiba között a leglényegesebb különbség – a statisztika, mint tudomány szemszögéből – az, hogy míg a mintavételi hiba matematikai-statisztikai eszközökkel becsülhető, addig a nem mintavételi hiba nehezen mérhető, számszerűsítésére korábbi tapasztalatok, analógiák, illetve szakértői becslések állnak csak rendelkezésre.

Mind a mintavételi, mind a nem mintavételi hiba mérsékelhető, amennyiben kellő körültekintéssel történik a mintavétel megtervezése, pontosabban közvetlenül a tervezést követő lépések elvégzése. A gazdasági-, szociológiai kutatások általános hibájaként említhető, hogy nélkülözik a jó mintavételi terv elkészítését és a mintavételi mód megalapozott kiválasztását. Ez a probléma azonban korántsem újkeltű, ezt bizonyítja, hogy közel hatvan évvel ezelőtt a következő kritikával illették a nem megfelelően előkészített felméréseket:

„A kontár statisztikák készítőit többek között az jellemzi, hogy a statisztikai megfigyelés megszervezésének egyes részleteit nem ismerik. Ebből adódik, hogy a kontár statisztikusok megalapozatlanul és kellő előkészítés nélkül küldik ki a kérdőíveket, amik minden lehetséges és lehetetlen adatra kiterjednek, nem csak azt kérdezik amire szükségük van, hanem azt is amire véleményük szerint szükségük lehet. A kontár statisztikusoknak legtöbbször fogalmuk sincs arról, hogy egy-egy adat megválaszolása, a válaszok hitelességének biztosítása milyen nagy munkát jelent a kérdőív kitöltői számára. A kontár statisztikát jellemzi a feltett kérdések nagy bősége, kérdések meg nem alapozottsága, és az, hogy a kérdésekre adható válaszok jelentős része nem összesíthető, vagy ha összesítik is, abból elemzések nem készíthetők.” Péter (1955. pp.207-208.)

Az alapos tervezést követően a kutatók általában belevetik magukat a megfigyelésbe, ezáltal egy fontos lépést ugorva át. A mintavételi egységek meghatározása legalább olyan fontos mozzanata a tervezésnek, minta a mintavételi mód kiválasztása. A mintavételi egységek definiálása során ugyanis a későbbiekben jelentkező nem mintavételi hibák több típusa is mérsékelhető.

Azonban kicsit általánosabban tekintve a gazdasági-, szociológiai kutatásokat, – felülemelkedve a mintavételi hiányosságokon – újabb ok fedezhető fel, ami kontárrá minősíti az elemzést. Ez az ok pedig nem más, mint a kutatás indoka. Éles különbséget kell tenni célszerű és kényszerű elemzések között. A társadalomtudományi kutatásokban alkalmazott statisztikai elemzések alapvető feltétele a célszerűség, amely napjainkban igen gyakran csorbát szenved. A statisztikai elemzések „újszerű” alkalmazása a célszerűség többoldalú értelmezéséből fakad. A statisztikai elemzések indukción alapulnak, az alkalmazásuk során feltárt összefüggések, tulajdonságok bizonyos – nem kevés számú – feltétel megléte esetén általánosíthatók. A kutatók egy része azonban inverz módon arra használja az elemzéseket, hogy az általa felfedezni vélt összefüggéseket valamilyen módon alátámassza. Ennek érdekében áthág matematikai törvényszerűségeken, mellőzi az elemzések feltételrendszerének teljesítését, bizonytalanabb elemzési módszereket választ. Mindezt azért, hogy a kutatási munkálatai, eredményei nehogy homályos képpé formálódjanak a mutatószámok tükrében. Az elemzési módszereket, mint eszközrendszert nem arra használja, hogy eddig nem ismert, rejtett összefüggéseket tárjon fel, hanem az általa vélt, deduktív, általánosságban elfogadtatható eredményeket megerősítse.

A másik eset, amikor egy kutató megalapozatlanul hajlandó elemzéseket végezni, a kényszerűségből végzett kutatás. Ezt az eshetőséget nyomatékosan említi Earl Babbie: „A napjainkban végzett társadalomtudományi kutatások jelentős részét indokolt kényszer szülte kutatásnak nevezzük. Amennyiben külső nyomás hatására vállalkoznak rájuk a kutatók. Ennek a jelenségnek két fő kategóriája van: (1) a ranglétrán alul álló egyetemi, főiskolai oktatók, akiknek a szakmai biztonsága és előmenetele részben a tudományos publikációkon múlhat és (2) olyan egyetemi, főiskolai hallgatók, akik csak akkor kapnak jegyet kutatásmódszertanból, ha elvégeznek egy kutatást." Babbie (1999. p.65.) A társadalomtudományi kutatások között azonban nem csupán ez a két indok fordulhat elő, amely kényszerű kutatást eredményez.

Abban a helyzetben, amikor a társadalomtudományi kutatás egy gazdasági döntés előkészítése, megalapozása céljából jön létre szintén születhetnek kényszerű elemzések. Ezeket éppen egy általánosított döntési modell hívja életre. A gazdaság különböző területein működő vállalkozások vezetői gyakorta szembesülnek a „make or buy" döntési problémával a hatékonysági és gazdaságossági célkitűzésekre való törekvéseik során. Napjainkban, amikor egyre több képzett közgazdász tevékenykedik a munkaerőpiacon, jogosan merül fel az említett döntési probléma. Megvásárolni valamilyen külső, megfelelő szakmai tapasztalattal, tudományos háttérrel és infrastruktúrával rendelkező szakértő cég által készített kutatást, vagy megbízni egy a vállalkozás által alkalmazott, elméleti képesítéssel rendelkező szakembert a feladat elvégzésével. Ebben a szituációban automatikus egyoldalú kényszer érvényesül a kutatással kapcsolatban. Ugyanis a kutatást nem a kutató munkája során felmerülő probléma indukálja, hanem a vezetők, döntéshozók elvárása. A kutató az elemzés végrehajtásával nem a probléma megoldására törekszik, hanem egy vezetői utasítás teljesítésére. A hasonló jellegű kényszer szülte kutatások során elkövethető hibák mérséklésének a kutató anyagi motiválásán kívül több lehetősége van. A döntéshozókkal történő kapcsolattartás rendkívül fontos. A döntéshozóknak ismerniük kell a kutatás lehetőségeit, illetve korlátait. A kutatások vezetői döntéseket segítő információkat nyújtanak, ám nem biztosítanak megoldást, hiszen az menedzseri megítélést igényel. Fordítva is igaz ez: a kutatók számára világosnak kell lennie, milyen döntéssel állnak szemben a döntéshozók − mi a vezetői probléma − és milyen eredményekre számítanak a kutatásból. A vezetői probléma meghatározása érdekében a kutatóknak különleges képességekkel kell rendelkezniük a döntéshozók megértéséhez. Kettejük viszonyát több tényező is megnehezítheti. Nehézkes lehet a kapcsolatfelvétel; a kutató vagy a kutatási részleg szervezeten belüli elhelyezkedé-

se is megnehezítheti a megfelelő emberek elérését. A hasonló kutatási helyzetekben a döntéshozók helytelen szemléletéből is fakadhatnak kontár statisztikák. Az ilyen vezetők mindenre kiterjedő kutatások, hatalmas mennyiségű statisztikák révén próbálják irányítani a szervezetet. Ebből a célból végzett kutatások nagy hibaforrást jelentenek azért is, mert megterhelik a kutatót és elvonják a figyelmét a kutatás operatív tevékenységeiről, a valódi értékelő, elemző, ellenőrző munkáról.

Még mélyebben keresve az elhivatott, megalapozott kutatások hiányának okait a célszerűség és kényszerűség ellentéte helyett azok kapcsolatára is érdemes figyelmet fordítani. A tudományos – Babbie által kényszerűnek tartott – kutatások során a statisztikai elemzések a célszerűség másfajta értelmezését adják. Annak érdekében, hogy a különböző kényszerű okok indukálta kutatások tudományosabbnak, megalapozottabbnak tűnjenek, előszeretettel alkalmaznak a kutatók statisztikai elemzéseket anélkül, hogy azok alkalmazási feltételeit ellenőrizték volna. Ez gyakorlatilag egy marketing fogásként értékelhető, mert ezáltal a kutatás „eladhatóbbá" válik. Követve a gondolatmenetet, a jelenséget Naresh K. Malhotra munkája nyomán Kutatásmarketingnek neveztem el. Nem találkozunk tudományos kutatói munkával mátrixok, diagramok, koordináta rendszerek, statisztikai elemzések nélkül. Nagyrészük szakmailag igényes, értékes elemzés, de nem kis hányaduk híján van ezeknek az erényeknek. Azonban törekedni kell arra, hogy az említett eszközök, módszerek ne szükséges formai kellékévé váljanak a kutatásoknak, hanem tartalmi alapköveivé.

A fent említett hibákat, melléfogásokat kikerülendően, a tudomány és technika fejlődését követő, a fejlődés eredményeit tudatosan alkalmazó kutató a modern elemző szoftverek alkalmazását előtérbe helyezi. A tudományos kutatók, szakmai elemzők fegyvertárában fellelhető informatikai eszközök, adatbázis kezelő, adatelemző, statisztikai szoftverek jelentős magabiztosságot nyújthatnak a kutatás során. Meg kell említeni azonban, hogy ezek a programok feltételezik olyan alapvető ismeretek meglétét, melyek nélkül ugyan tökéletesen alkalmazhatjuk őket, de az általuk generált eredmények felhasználhatósága ismételten korlátozott lesz. Hasonlóan a manuális elemzésekhez. Ily módon ezek a fegyverek könnyen magunk ellen fordulhatnak, és visszavethetik a kutatás eredményeit. Vita László szavait idézve: "A legfőbb hátrány az, hogy a statisztikai programcsomagok bármilyen adathalmazra „rászabadíthatók", és minden fajta elemzés elvégezhető velük, akár van értelme az adott elemzésnek, akár nincs, illetve akár fennállnak az elemzés alkalmazásának feltételei, akár nem." Balogh –Vita (2005. p.556.)

Az érett kutató számára nyilvánvaló kell legyen, hogy a megbízható társadalomtudományi kutatás során végzett statisztikai elemzés nem végezhető el egy-egy módszertani jegyzet, vagy könyv néhány fejezetének átlapozásával és tartalmának nagyvonalú alkalmazásával. Jól definiált módszertani feltételekhez történő következetes és teljes körű igazodás alapján sikeres kutatási eredmények publikálhatók.

3.2. MINTAVÉTELI HIBA

A mintavételi hiba tulajdonképpen a felmérések legismertebb hibaforrásaként említhető. Ez a hibaforrás az adatok változékonyságára vezethető vissza. Egészen pontosan arra, hogy az elméletileg kiválasztható minták közül csak egyetlen egy kerül megvalósításra, így a változékonyságnak köszönhetően akár szélsőséges eredményeket is produkálhat. A mintavételi hiba tehát abból fakad, hogy egy sokasági jellemző becslésekor a sokaságnak csak egy része ismert, nem pedig az egész sokaság. Mintavételi hibán ezért a mintából kapott becslés és a „valós" érték (a sokaság összes egyedének ismeretében kiszámítható érték) közötti különbséget kell érteni.

A véletlen mintavételi hiba alapvető tulajdonságai a következőkben foglalhatók össze:

- csökken a mintanagyság növekedésével (de nem egyenesen arányosan),
- függ a vizsgált sokaság nagyságától,
- függ a megismerni kívánt jellemző szóródásától,
- mérhető és kontrollálható véletlen mintavétel esetén,
- csökkenthető egy megfelelő mintavételi terv elkészítésével, megfelelő mintavételi mód kiválasztásával.[1]

Felmerülhet a kérdés, mit kell megfelelő mintavételi terv alatt érteni, hogyan mérhető a megfelelőség és hogyan lehet összehasonlítani a különböző mintavételi terveket. Többek között ezekre a kérdésekre keresem a választ a dolgozat következő fejezeteiben.

[1] www.statcan.ca: Power from Data (2010.07.06.)

3.2.1. AZ EGYSZERŰ VÉLETLEN MEGFIGYELÉS HIBÁJA

Az egyszerű véletlen megfigyelés – amint erre elnevezése is utal – a reprezentatív megfigyelési módszer legegyszerűbb formája. Egyszerű véletlen megfigyelés esetén a minta elemeit – már ismertetett módon – egyszerű véletlen módszerrel kell kiválogatni.

A mintavételi hiba meghatározásának módszereit, képleteit nem kívánom teljeskörűen bemutatni. Mivel a különböző paraméterek becslésénél alkalmazható képleteket és analógiákat a szakirodalom részletesen ismerteti, így én csupán egyetlen alapesetet jelenítek meg, a sokasági várható érték becslés hibájának számítását. A várható érték becslésénél a végső célkitűzés az alapsokaság átlagának meghatározott valószínűséggel és becslési hibahatárok közötti megadása.

Bevezetve a következő jelöléseket:

- N az alapsokaság elemeinek száma,
- n elemek száma a mintában; a minta nagysága,
- X_i adott változó értéke az alapsokaság i-edik eleménél *(i= 1,2, ..., N),*
- x_i adott változó értéke a minta i-edik eleménél *(i= 1,2, …, n).*

Meg kell említeni, hogy X_i értéke nem feltétlenül egyezik meg x_i-vel azonos *„i”* értékeknél, hanem általában különbözik attól.

$\overline{X}$ Adott változó értékeinek számtani átlaga az alapsokaságban:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$\overline{x}$ Adott változó értékeinek számtani átlaga a mintában:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

σ^2 Az adott változó szórásnégyzete az alapsokaságban:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}$$

s^2 Az adott változó mintából számított korrigált tapasztalati szórásnégyzete:

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

A gyakorlati statisztikai munkában éppen a standard hiba megállapításánál szükség van az alapsokaság szórásának a σ-nak az ismeretére is. Az esetek nagy többségében azonban az alapsokaság szórása nem ismert. A számításoknál tehát csak a minta korrigált tapasztalati szórásának ismeretére lehet támaszkodni. A minta alapján az alapsokaság szórásának torzítatlan becslésére nincs mód, a gyakorlati munkában azonban jól bevált a fenti összefüggés, amely becslés a szórásnégyzetre torzítatlan.

STANDARD HIBA

A standard hiba képletének felírása előtt egy fontos körülményre kell tekintettel lenni. A kiválogatás ugyanis kétféleképpen történhet:

- a kiválogatott elemek az értékeik feljegyzését követően valamilyen módon visszahelyezésre kerülnek az alapsokaságba, vagy
- a kiválasztott elemek a további kiválasztások során nem kerülnek figyelembe vételre.

Az első eset végtelen számú elemből álló alapsokaságot eredményez, az utóbbi eset ezzel szemben csak véges számú elemből állót. Kétségtelen, hogy az utóbbi esetben az egyszerű véletlen reprezentatív megfigyelésnek az a feltétele, hogy minden egyes elemnek egyenlő lehetősége legyen arra, hogy a mintába belekerüljön, nem tud maradéktalanul érvényesülni, (lásd Antal – Tillé (2011.)). Az ebből eredő esetleges hibák ellensúlyozására használatos egy korrekciós tényező[2], mellyel a standard hiba képletét megszorozva az ismétlés nélküli kiválasztás esetén a standard hiba képlete a következő:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

ahol n/N tört, kiválasztási arány jelölésére szolgál.

[2] A korrekciós tényező értéke 1-nél mindig kisebb. Ha n=N, azaz ha teljes körű felvételről van szó, úgy az $s_{\bar{x}} = 0$, mert a gyökjel alatti kifejezés értéke zérussal egyenlő. Ha ismétléses mintavételről van szó, úgy nyilvánvalóan N→∞ és így a gyökjel alatti kifejezés =1.

Az alapsokaság szórását, σ-t, a mintából számított értékével helyettesítve a képlet a következőképpen módosul:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

3.2.2. A RÉTEGZETT MEGFIGYELÉS HIBÁJA

Az alapsokaság átlagának a mintára támaszkodó becslése a következő tényezőktől függ:

- a minta nagyságától,
- a sokaság méretétől,
- a vizsgált jellemző szóródásától,
- az alapsokaság homogén vagy heterogén jellegétől, mely a standard hiba nagyságában jut kifejezésre.

Eltekintve attól a lehetőségtől, hogy a becslés pontosságát és megbízhatóságát a mintaelemek számának növelésével javítani lehet, a becslés pontosságának növekedése csak az alapsokaság heterogenitásának csökkentésétől remélhető. A reprezentatív megfigyelésnek egyik ilyen módja, melynek célja éppen az alapsokaságon belül homogén csoportok létrehozása, a rétegzett megfigyelés.

Az egyszerű véletlen mintavétellel szembeállítva fő jellemzője, hogy az alapsokaság valamennyi rétegének mintabeli reprezentációját megköveteli. Ennek következtében az alapsokaság keresztmetszetéről pontosabb képet ad, mint a rétegezés nélküli eljárás. A jellemzők becslésének pontossága fokozódik. Alkalmazását azonban megnehezíti, hogy a rétegek nagyságát ismerni kell, valamint a mintaelemek számát minden egyes rétegben előre kell tudni.

A továbbiakban azokat a hiba-meghatározási eljárásokat ismertetem, melyeket ennél a módszernél alkalmaznak.

Felhívva a figyelmet arra, hogy azokat a korrekciókat, melyek az egyszerű véletlen reprezentatív megfigyelésnél használatosak, itt is figyelembe kell venni.

Mindezek előrebocsátása után a következő jelöléseket használom:

- N_j az elemek száma az alapsokaság j-edik rétegében; a j-edik réteg nagysága,
- L a rétegek száma,
- n_j a j-edik rétegből kiválasztott minta elemeinek száma,
- $\sum_{j=1}^{L} N_j = N$ az alapsokaság összes elemeinek száma,
- $\sum_{j=1}^{L} n_j = n$ a mintasokaság összes elemeinek száma,
- X_{ij} = a változó értéke az alapsokaság i-edik eleménél a j-edik rétegben,
- x_{ij} = az i-edik mintaelemhez tartozó érték a j-edik rétegben.

A j-edik rétegben a elemek értékeinek átlaga,

az alapsokaságban:

$$\overline{X}_j = \frac{\sum_{i=1}^{N_j} X_{ij}}{N_j},$$

a mintában:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}.$$

Az alapsokaság átlaga:

$$\overline{X} = \frac{1}{N} \sum_{j=1}^{L} N_j \overline{X}_j \ .$$

Az minta átlaga sokasági arányokkal súlyozva:

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{L} N_j \bar{x}_j \ .$$

Az alapsokaság értékeinek szórásnégyzete a j-edik rétegben:

$$\sigma_j^2 = \frac{\sum_{i=1}^{N_j} \left(X_{ij} - \overline{X}_j \right)^2}{Nj}.$$

A minta korrigált tapasztalati szórásnégyzete a j-edik rétegben:

$$s_j^2 = \frac{\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j \right)^2}{nj}.$$

A rétegzett mintából történő becslés standard hibájának meghatározásához az átlag rétegen belüli becsléséből kell kiindulni. Az átlagbecslés standard hibájának négyzete az ismert képlet alapján: (a rétegeken belül egyszerű véletlen mintavételt feltételezve)

$$s_{\bar{x}_j}^2 = \frac{s_j^2}{n_j} k_j$$

ahol:

$$k_j = 1 - \frac{n_j}{N_j}$$

a korábban említett korrekciós tényező.

A standard hiba nagysága - a mintaátlagok szórása – a következő képletek alkalmazásával számítható:

$$s_{\bar{x}}^2 = \frac{1}{N^2} \sum_{j=1}^{L} N_j^2 \frac{s_j^2}{n_j} k_j = \sum_{j=1}^{L} \left(\frac{N_j}{N} \right)^2 s_{\bar{x}_j}^2 k_j$$

A közölt képletek értékelésénél megállapítható, hogy a standard hiba nagysága az alap sokaság elemeinek rétegen belüli szóródásától függ. Ebből következik, hogy minél több és minél homogénebb réteg kerül kialakításra, annál nagyobb lesz a becslés pontossága.

A rétegzett megfigyelés pontossága tehát szorosan összefügg az egyes rétegekből kiválasztott minta elemszám *„n_j”* nagyságával. Ez a körülmény szükségessé teszi a rétegen belüli kiválasztás kérdésének behatóbb vizsgálatát. Indokolja annak részletesebb elemzését, hogy miként lehet egyrészt megkövetelt becslési pontosságot minimális mintaterjedelemmel biztosítani, másrészt adott mintaterjedelem esetén lehető legkisebb standard hibát elérni. A fentebb leírtak alapján, azokat az eljárásokat melyek a rétegeken belüli kiválasztást ezeknek az alapelveknek a figyelembevételével írják elő, optimális elosztásnak nevezik. Ha a minta *„n”* terjedelmének meghatározása és ezen belül a kiválasztás az egyes rétegekben oly módon történik, hogy a standard hiba nagysága a minta nagyságához képest a legkisebb legyen, úgy *„n_j”* megválasztásánál az alábbi követelményeknek kell eleget tenni:

1. Minél nagyobb a réteg, annál nagyobb mintát kell belőle kiválogatni.
2. Minél nagyobb az alapsokaság elemeinek szóródása a rétegen belül, annál nagyobbra kell méretezni a minta terjedelmét is.

Ezeknek az alapelveknek megfelelően a rétegen belül az egyes minták optimális nagysága a Neyman-féle optimális allokáció már ismert képlete alapján adható meg:

$$n_j = n \frac{N_j \sigma_j}{\sum_{j=1}^{L} N_j \sigma_j}$$

illetve amennyiben s_j^2 torzítatlan becslése σ_j^2-nek :

$$n_j = n \frac{N_j s_j}{\sum_{j=1}^{L} N_j s_j}$$

Összehasonlítva a rétegezett megfigyelés hibájának általános képletét az egyszerű véletlen megfigyelés hibájának formulájával, a következő megállapítások tehetők:

Az egyszerű véletlen megfigyelés hibája a mintaelemek szórásnégyzetének felhasználásával került kiszámításra. A rétegezett megfigyelésnél ezzel szemben a mintaelemeknek csak a rétegen belüli szóródása került figyelembevételre. A képletben csak a rétegen belüli szórásnégyzetek összege szerepelt. A mintaelemek szóródása ennek megfelelően rétegezés esetén két összetevőre bontható $\sigma^2 = \sigma_K^2 + \sigma_B^2$, ahol σ_K a rétegek közötti, σ_B pedig a rétegen belüli szórást jelenti. A rétegezett megfigyelésnél a mintaelemek szórásnégyzete a rétegek közötti szórásnégyzettel kisebb, mint az egyszerű véletlen mintavételnél.

Az alapsokaság rétegekre bontásának tehát – amint az a fentiekből kitűnik – az az értelme, hogy ezzel a módszerrel a reprezentatív megfigyelés hatékonysága növelhető. Másként fogalmazva, rétegezett megfigyelést alkalmazva a becsléseknek kisebb lesz a standard hibája, mint egyszerű véletlen mintavétel esetén.

3.3. NEM MINTAVÉTELI HIBA

A mintavételen alapuló felméréseknek az előbbiekben bemutatott mintavételi hibán kívül számos másfajta hibaforrása is létezik, melyeket gyűjtőnéven nem mintavételi hibáknak neveznek.

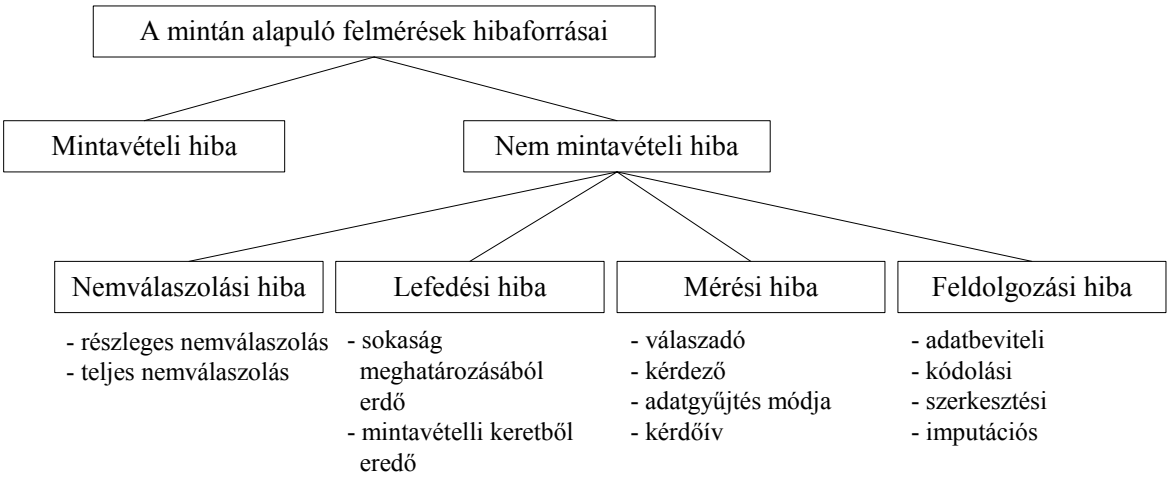
A nem mintavételi hibák a mintavételen kívüli, egyéb forrásokból erednek, és véletlen vagy nem véletlen jellegűek lehetnek. Mivel a nem mintavételi hibák a felmérés, lekérdezés, kódolás feldolgozás folyamatában egyaránt keletkezhetnek, okozhatja őket a probléma helytelen meghatározása vagy megközelítése, a skálák vagy a kérdőívek felépítése, az

interjú módszere, valamint az adatfeldolgozás és az elemzés is. Emellett előfordulhat az is, hogy a kérdezőbiztos pontatlanul kérdez, a válaszadó akarattal vagy akaratlanul rosszul válaszol, stb.

A nem mintavételi hiba tulajdonságai közé tartozik, hogy:

- mind teljes körű, mind pedig részleges adatfelvétel során létezik,
- nehezen mérhető,
- a megfigyelés különböző aspektusaiban jelentkezhet,
- nem csökkenthetők a mintanagyság növelésével (mint az megtehető a mintavételi hiba esetén).[3]

A mintavételen alapuló felmérések hibáinak, főbb forrásai a Statistical Policy Working Paper 31 (2001.) alapján a következőkben foglalhatók össze. (1. ábra)



1. ábra: A mintavételen alapuló felmérések hibaforrásai

3.3.1. NEMVÁLASZOLÁSOK A MINTÁBAN

A nemválaszolási hiba általánosan ismert és könnyen azonosítható hibaforrás. A nemválaszolási hiba akkor merül fel, ha néhány, a mintában szereplő válaszadó nem válaszol, így az adott egyedektől a kívánt kérdésre nem kapható információ. A visszautasítás és az elérés hiánya az elsődleges okai a nemválaszolásnak. A nemválaszolás következménye, hogy a realizált minta méretében vagy összetételében eltér az eredeti mintától. A válasz

[3] www.statcan.ca: Power from Data (2010.07.06.)

megtagadás csökkenti a valós minta méretét, ezzel potenciálisan növeli a varianciát és torzítást okoz a becslési eredményekben. A magas válaszadási arány csökkenti annak az esélyét, hogy a nemválaszolásból eredő torzítás túl nagy legyen. A nemválaszolási hiba úgy definiálható, mint a vizsgált változó eredeti mintában szereplő átlagértéke és a megvalósult mintában szereplő átlagérték közötti eltérés. Ezek alapján a becslési eredmények megfelelő értékeléséhez ismerni kell a mintavétel során bekövetkező nemválaszolási arányt, ami a felmérés minőségének egyfajta mérőszámaként is értelmezhető. A nemválaszolási arányt mindig fel kell tüntetni a mintavételes felmérésekben, és ha lehetséges a nemválaszolás hatásait becsülni kell. Csak így lehet az adatelemzés és az értelmezés számára megfelelő alapot biztosítani.

A nemválaszolás elsődleges okai a visszautasítás és az elérés hiánya, azonban emellett további tényezők is generálhatnak nemválaszolás okozta adathiányt. Ezeket Hunyadi – Vita (2002.) a következőképpen csoportosítja:

a) Egyszerű, de nem ritka oka lehet a nemválaszolásnak az, hogy a keresett személy (átmenetileg) nem található meg. Leggyakrabban azért, mert nem tartózkodik otthon.
b) Az is fontos eset, amikor a kérdezett nem képes válaszolni a kérdésekre. Ez leginkább olyan esetekben fordul elő, amikor személyes megkérdezéskor a megkérdezett nincs a kellő információk birtokában, azoknak utána kell néznie, régi, kéznél nem levő dokumentumokat kell beszereznie stb.
c) A leglényegesebb azonban mindezeken túl az a helyzet, amikor a megkérdezett szándékosan megtagadja a válaszadást. Ez történhet egyszerűen a kérdőívtől vagy az interjútól való általános viszolygásából, bizalmatlanságból, abból, hogy sajnálja rá az időt és a fáradságot, de leggyakrabban abból, hogy bizonyos kényes kérdésekre jól vagy rosszul felfogott érdekét szem előtt tartva nem kíván válaszolni. Ilyen kényes kérdések lehetnek a faji, vallási hovatartozásra, egészségi állapotra, vagy manapság kiváltképp jellemző módon a jövedelmi illetőleg vagyoni helyzetre vonatkozó kérdések.

Emellett az okozott adathiány mértéke szempontjából is megkülönböztetik a nemválaszolásokat. Létezik teljes nemválaszolás (unit-nonresponse) amikor a kijelölt mintaelem egyáltalán nem tud, vagy nem akar részt venni a megfigyelésben, azaz egyetlen kérdésre sem ad választ. A másik eset az, amikor a kijelölt mintaelem csak bizonyos kérdésekre nem ad választ; ekkor részleges nemválaszolásról (item non-response) van szó.

„Bár lényegét tekintve a két eset hasonló, kezelésük azonban részben más megoldásokat követel. Utalunk itt előzetesen arra, hogy a felvétel során beszerzett adatokat általában adatmátrixokba rendezik, ahol a mintaelemek a sorokat, az egyes kérdésekre adott válaszok az oszlopokat jelölik. Teljes nemválaszolás esetén a mátrixból egész sorok esnek ki, míg részleges nemválaszolás esetén csak egyes elemek.

A nemválaszolások téves következtetésre vezetnek, hiszen könnyű belátni, hogy ha a nemválaszolók és a válaszolók viselkedése, egy kérdésre adott válasza tendenciaszerűen eltérő, vagy másként fogalmazva a válaszadás tartalma és a válaszolás/nemválaszolás egymással sztochasztikus kapcsolatban lévő ismérvek, akkor önmagában az a tény, hogy létezik nemválaszolás, hibás következtetésekre vezet.” Hunyadi – Vita (2002. p.300.)

Bármilyen fajtája is forduljon elő a nemválaszolásnak egy mintavétel során, az így nyert adatokon végzett elemzésekből levont következtetések, becslések torzítottak lesznek. A nemválaszolás okozta torzítás mértéke (nonresponse bias) nem számszerűsíthető olyan „közvetlen” módon, mint a mintavételi hiba mértéke. Ennek oka a nemválaszolás fenti sokféleségében kereshető. A torzítás mértékének csökkentésére viszont számos módszer létezik. Mielőtt ezen módszerek átfogó ismertetésére rátérnék, meg kell említeni, hogy a nemválaszolás nem okoz minden esetben torzítást. Amennyiben a nemválaszolás véletlenszerűen következik be és nincs összefüggésben a vizsgálni kívánt jelenséggel, akkor a becslési eredményeket, következtetéseket nem fogja torzítani az adathiány. Abban az esetben, ha valamely megfigyelési egység szándékosan nem-, vagy rosszul válaszol, akkor a torzítás egyértelműen érzékelhető.

Az adathiány mértéke és a vizsgált jelenség (változó) közötti kapcsolat alapján Oravecz (2008.) Little és Rubin (1987.) munkája nyomán háromfajta adathiányt különböztet meg:

- Teljesen véletlenszerű adathiány: a válaszoló és a hiányzó adatokat tartalmazó egységek teljesen egyformák (ez a gyakorlati kutatásokban elég valószerűtlen)
- Részben véletlenszerű adathiány: esetében a hiányzó adatokat tartalmazó egységek eltérnek a hiánytalan adatokkal bíró egységektől, de a hiány jellegzetességei nyomon követhetők, előre jelezhetők az adatbázis más változói segítségével. Az adathiány tehát más változókkal kapcsolatban van, de azzal a változóval, amelyikben a hiányzás felmerül nincs közvetlen kapcsolatban

– Nem véletlenszerű adathiány: esetében az adathiány nem véletlenszerű, és más változókkal sem becsülhető, mert közvetlenül az adathiányt tartalmazó változóval van kapcsolatban. Ez az adathiány legveszélyesebb, legnehezebben kezelhető formája. Ez a mechanizmus fordul elő például, ha a magasabb jövedelemmel rendelkezők nagyobb valószínűséggel tagadják meg a jövedelemre vonatkozó kérdések megválaszolását, és a jövedelemre nem lehet következtetni a felmérés más változóiból.

3.3.2. EGYÉB NEM MINTAVÉTELI HIBAFORRÁSOK

LEFEDÉSI HIBA

A lefedési hiba általában az alapsokaság meghatározásából fakad, ami a felmérés szempontjából releváns, tényleges sokaság és a kutató által meghatározott sokaság közötti eltérésként definiálható. A sokaság helyes meghatározásának problémája egyáltalán nem triviális. Sokszor egy adott jelenséget régóta vizsgáló kutató sem tudja egzakt módon lehatárolni a vizsgált sokaság kereteit.

A lefedési hiba abból is eredhet, hogy a kutató által meghatározott alapsokaság és a mintavételi lista által érintett alapsokaság között eltérések tapasztalhatók. Ezt a mintavételi keretből eredő hibának nevezik. Ilyen eset gyakran előfordul helytelen, hibás regiszterek esetén, amikor a kijelölt cím nem létezik, vagy ott nem található meg a keresett megfigyelési egység. A mai magyar gyakorlatban – akár a gazdaságstatisztika, vagy társadalomstatisztika területén – nem ritka ez az eset. Például, egy magyarországi felmérésnél, ahol személyes elérést biztosító lekérdezést kell megvalósítani, a lakcímnyilvántartó rendszer adatbázisához lehetne segítségül fordulni. Azonban sajnos ez sem szolgáltat teljes mértékben pontos ismereteket, (például megyei szinten) hiszen sokan elmulasztják lakcímbejelentési kötelezettségüket, különösen ideiglenes lakcím esetén.

MÉRÉSI HIBA

A mérési hiba úgy definiálható, mint a változó megfigyelt és a valós értéke közötti különbség. A mérési hibák forrása négy alapvető tényező köré csoportosítható:

– a kérdőív, mint a keresendő információk bemutatásának eszköze;
– az adatgyűjtési módszer, mint a keresendő információ megszerzésének módszere;

- a kérdező, mint a kérdések feltevője, az információ megszerzője (kivéve az önkitöltős kérdőívek esetében);
- a válaszadó, mint a kérdések befogadója, az információ szolgáltatója.

A kérdőív olyan hibákat hordozhat magában, mint a kérdés megfogalmazása, a kérdések hossza, magának a kérdőívnek a hossza, a kérdések sorrendje, válaszkategóriák, nyitott és korlátozott válaszlehetőségek megadása.

A felmérések lebonyolítása eltérő technikákat követel meg különböző adatgyűjtési módszerek esetén. A postai-, az elektronikus-, a személyes adatgyűjtések, illetve a naplózáson alapuló felmérések mind-mind sajátos hibaforrásokat rejtenek magukban.

A kérdezőbiztos által okozott válaszadási hibák a válaszadó kiválasztásából, a kérdezés módjából fakadó, a rögzítésből származó és a csalási hibákat jelentik. Emellett a kérdezési hiba azokat az eseteket jelenti, amikor a kérdésfeltevés során követnek el hibákat vagy amikor nem kérdeznek rá tovább valamire, holott több információra lenne szükség. Például, a kérdezőbiztos a kérdéseket nem a kérdőív szóhasználatával teszi fel.

A válaszadásból akkor származik hiba, ha a válaszadók pontatlan válaszokat adnak, vagy válaszaikat rosszul rögzítik, illetve azokat félreértelmezik. A válaszadás hibáját "elkövethetik" a kutatók, a kérdezőbiztosok, vagy a válaszadók.

A válaszadó által elkövetett hibák a képtelenségből és a válasz megtagadásából eredhetnek.A képtelenségből származó hibákat az jelenti, hogy a válaszadó nem tud pontos válaszokat adni. Ismeretlen a téma, fáradt, unatkozik, rosszul emlékszik, nem érti a kérdést, vagy más miatt ad pontatlan feleletet. A válaszmegtagadási hibák abból erednek, hogy a válaszadó nem hajlandó pontos információt adni. A válaszadó szándékosan "félreválaszolhatja" a kérdéseket, mert társadalmilag elfogadható válaszokat akar adni, vagy el akarja kerülni, hogy megütközzenek a válaszán, zavarba jöjjön, vagy egyszerűen a kérdezőbiztosnak akar imponálni.

FELDOLGOZÁSI HIBA

Adatbeviteli hiba esetén a válaszadó válaszainak a meghallásában, értelmezésében és rögzítésében vannak hibák. A válaszadó például semleges választ ad (határozatlan), de a kérdezőbiztos ezt úgy értelmezi, mintha pozitív lett volna.

Előfordulhat, hogy a papír alapú kérdőívek digitalizálásakor keletkezik hiba, ami egyaránt elkövethető elektronikus scannelés és manuális adatrögzítés esetén.

A különböző irányított kérdésekre adott válaszok kódolása szintén nagy rizikófaktor, a kódszámok összekeverhetők, emellett a helytelen skálázás alkalmazása és a hiányzó értékek megfelelő kezelése is problémát okozhat.

Az adatstruktúra kialakítása szintén nagy jelentősséggel bír. A nem megfelelően szerkesztett adatok, a helytelen csoportosítás, aggregálás hibáit sokszor egyáltalán nem, vagy csak rendkívül nagy és fölösleges energia befektetésével lehet helyrehozni.

A hiányzó adatok pótlására alkalmazott módszerek szintén kellő körültekintést igényelnek, hiszen a végeredményekre gyakorolt hatásuk jelentős lehet.

Az adatelemzési hiba olyan hibákat ölel fel, amelyek akkor keletkeznek, amikor a kérdőívekből származó nyers adatokat kutatási eredményekké alakítják át. Például, egy helytelen statisztikai eljárás helytelen értelmezést és következtetéseket okoz.

„Fontos megjegyezni azt, hogy sokféle hibaforrás létezik. A kutatási terv kialakításakor a kutatónak a teljes hiba, és nem csak egy-egy hibalehetőség minimalizálására kell törekednie. Ezt a figyelmeztetést a diákok és a nem eléggé képzett kutatók körében tapasztalható azon szokás indokolja, ami a minta növelésével próbálja ellensúlyozni a mintavételből származó hibákat. A mintanagyság növelése csökkenti a mintavételből fakadó esetleges hibákat, de ez együtt járhat a nem mintavételből eredő hibák növekedésével, mivel így megnőhet a kérdezőbiztosok által elkövethető hibák száma.” Malhotra (2005. p.146.)

„Valószínűleg a nem mintavételből eredő hibák több problémát okoznak, mint a mintavételből származó hibák. A mintavételből származó hibák kiszámíthatók, míg a nem mintavételi hibák sokféle formája lehetetlenné teszi becslésüket. Ráadásul, a teljes hiba legnagyobb része a nem mintavételi hibákból ered, míg a véletlen mintavételből származó hibák viszonylag kismértékűek.” Corlett (1996. p. 312.) A különböző vizsgálatoknál azonban egyöntetűen a teljes hiba mértéke a meghatározó. Egy-egy hibatípus annál lényegesebb, minél nagyobb hatást gyakorol a teljes hiba növekedésére.

3.4. A NEM MINTAVÉTELI HIBA KIKÜSZÖBÖLÉSE

A legtöbb probléma kezelésében az egyik leghatékonyabb módszer a megelőzés. Azt a célt kell kitűzni, hogy a lehető legtöbb megfigyelési egység az összes feltett kérdésre korrekt választ adjon. A megfigyelési egységek ösztönzését a válaszadásra számos marketing eszköz támogatja. Ezen kívül fontos szerepe van a felmérés, lekérdezés tervezésének. A megfelelő módon feltett kérdések, vagy a kérdezés módja megváltoztathatja az emberek hozzáállását a felméréshez, még akkor is ha a felmérésben való részvétel nem szolgálja az ő személyes érdeküket.

A preventív eszközök és technikák a kutatók által elkövethető hibák kiküszöbölésére is szolgáltatnak megoldást, bár ezek többségében nem matematikai- statisztikai eszközök.

A kutatási probléma kellően pontos definiálása és a folyamatos konzultáció a kutatást végző és az eredményeket felhasználó személy, vagy szervezet között (amennyiben ezek különböznek egymástól) nagymértékben elősegíti a mérési hiba csökkentését. Ezt a típusú problémát egyaránt hordozhatják a szekunder információkat felhasználó kutatások és az alacsony költségvetési kerethez kötött kutatások. Ezekben az esetekben ugyanis a megszerzett információk halmaza gyakran nem fedi pontosan a keresett információk körét. A keresett információ ugyanis többnyire nem lelhető fel pontosan szekunder információként, vagy ha mégis, akkor olyan strukturált-, vagy aggregált formában, amely további elemzések elvégzésére nem nyújt lehetőséget. Primer információként való beszerzése viszont hatalmas költségeket emésztene fel, akár olyan mértékben, amely még a kutatás eredményeinek legsikeresebb felhasználásával sem ellentételezhető.

Az ilyen hibák megelőzésére a kutatási probléma, a kutatási cél módosítása jelenthet megoldást. Jelen kutatás esetében ezek a hibák nem merülhetnek fel, mivel a kutatási cél nem az adatokban rejlő információk feltérképezése, hanem a mintavételi és hibaszámítási módszerek fejlesztése.

Az alapsokaság meghatározásából eredő hiba gyakran előforduló jelenség, mely sok esetben feltáratlan marad, hiszen a kutató maga sem sejti ennek a hibának a megjelenését, így nem is számol ennek torzító hatásával a becslési eljárások alkalmazásakor. A gyakorlati kutatások azonban a legritkább esetben vonatkoznak jól behatárolt, egyértelműen kijelölhető sokaságokra. Gondoljunk a népszámlálásra, mint egy igen kézenfekvőnek tűnő kuta-

tásra, melynek sokasága Magyarország népessége. Mégis megnehezíti a helyzetet a népességre vonatkozó definíciók különbözősége (pl.: de facto népesség, de jure népesség, stb.), vagy éppen a nemzetközi összehasonlítási igények. Emellett a sokaság rejtett egyedei is problémát okozhatnak, mely már a nemválaszolási hibák közé sorolható. Ez okozza azt a tényt, hogy a népszámlálásokból olykor több százezer ember kimarad (pl.: hajléktalanok).

A rejtett sokaságok felmérésénél pedig esély sincs a sokaság nagyságának behatárolására, így az egyedek megtalálása még nehezebb feladat. Ezért okoz nem kevés gondot a rejtett fekete, vagy szürke gazdaság feltérképezése, vagy a látens bűnözés mértékének becslése. Ilyen esetekben alternatív mintavételi módszerek alkalmazása lehet célravezető, melyekről részletesen beszámol Kapitány (2010.).

Az adatelemzési hiba látszólag a legelhanyagolhatóbb méretet ölti a kutatási eredmények torzítottságának növelésében. Azt gondolnánk, hogy a tapasztalt kutatók a fejlett technikának köszönhetően minimálisra csökkenthetik, vagy meg is szüntethetik ennek a hibának a jelenlétét. Azonban a modern technikai eszközök, elemző szoftverek, tanuló algoritmusok nem kellő szakértelemmel történő alkalmazása legalább akkora problémát okoz az adatelemzési hiba kialakulásában, mint amekkora segítséget nyújt. Az ellenőrző tesztek és próbák sokfélesége lehetőséget biztosít az elemzési eredmények igazolására, amennyiben megtaláljuk a legenyhébb feltételekkel rendelkező teszteket. Ezt a problémát tekintve azonban kénytelenek vagyunk élni azzal a hipotézissel, hogy ilyen nem fordul elő, még az üzleti célú kutatások esetében sem!

A kérdező biztosok által elkövethető hibák kezelése a legegyszerűbb, mégis olykor a legnagyobb arányban előforduló és emellett sokszor látens hibaforrásról van szó. A kérdező biztosok által elkövethető hibák közül a legjellemzőbb a válaszadó helytelen kiválasztása. Sajnos személyes gyakorlati tapasztalataim, valamint a közelmúltban kipattant botrányos híresztelések nagy, magyarországi kutatócégekkel kapcsolatban nemcsak alátámasztják feltételezésemet, hanem hatványozottan megerősítik azt.

Ezek a tapasztalatok sajnos megkérdőjelezik, hogy bármilyen erőfeszítés eredményes lehet-e abban az esetben, ha a kérdezést végrehajtó személy nem kellően elhivatott a kutatás sikeressége iránt.

A kérdezőbiztosok megfelelően alapos képzésével kivédhető a kérdezés módjából, a kérdezőbiztosokkal szembeni bizalmatlanságból és a rögzítésből eredő hibák nagy része. Az említett hibatípusok csökkentésére alkalmazzák továbbá a kérdőívbe épített, és annak struktúrájában megfelelően elhelyezett kontroll és inverz kérdéseket is.

A csalások elkerülése érdekében pedig jól követhető és ellenőrizhető mintavételi tervet kell készíteni minden esetben, nem engedve meg a kérdezőbiztosoknak a szubjektív kiválasztást[4] a lekérdezés során. Anyagias világunkban sajnos azt kell mondanunk, hogy a kérdezőbiztos munkája nem minden esetben van kellően megfizetve, az elhivatottságot pedig a szakmai érdeklődés és lojalitás híján csak jól megfizetett munkával lehet elérni.

A válaszadó által adott válaszok képtelenségéből eredő hibák jelentős hányadát meg lehet előzni a megfelelően szerkesztett kérdőívvel. Így például a szükségtelen kérdések elhagyásával, rövid egyértelmű, könnyen megfogalmazott kérdőívek megírásával.

A válaszmegtagadás megelőzésére sajnos csak olyan módszerek alkalmazhatók, melyek a kérdezőbiztos szubjektív beavatkozását igénylik, ami a fent említett okokra hivatkozva legalább akkora károkat okozhat, mint amekkora annak potenciális jótékony hatása. Ezért a válaszmegtagadás jobbára csak utólag orvosolható.

3.4.1. A NEMVÁLASZOLÁSI HIBA KIKÜSZÖBÖLÉSÉNEK LEHETŐSÉGEI

A nemválaszolás okozta hiba az egyik legnagyobb probléma a mintavételen alapuló kutatások terén. Egyes szerzők szerint ennek mértéke bizonyos esetekben meghaladhatja a mintavételi hiba mértékét, vagy akár annak többszörösét is. Ezért a hazai és nemzetközi szakirodalom egyértelműen ennek a hibaforrásnak a vizsgálatával foglalkozik a legtöbbet. A probléma súlyosságát jól érzékeltetik Az – Vita (1998) kísérleti jövedelem felvételének tapasztalatai, melyek szerint az előzetes felkérések során közel 90%-os nemválaszolással szembesültek. A nemválaszolás mértékének ismerete és figyelembe vétele azonban nem csupán a statisztikai értelmezhetőséget javítja, hanem költségcsökkenést is eredményezhet,

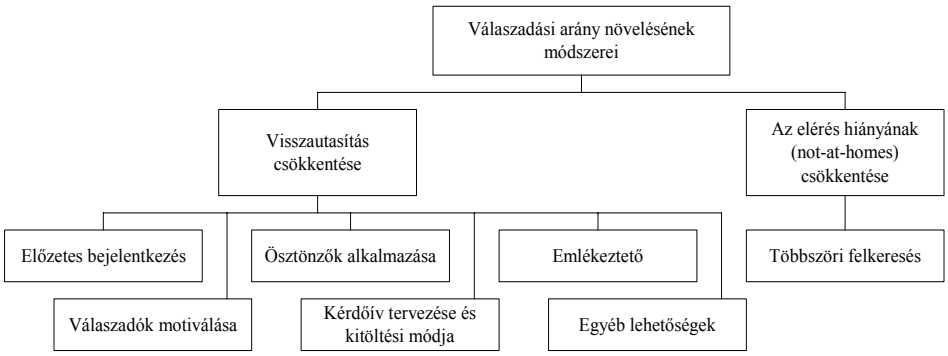
[4] Fentiekben szubjektív kiválasztás alatt az ellenőrzés nélküli önkényes kiválasztást értem. A félreértések elkerülése végett meg kell jegyezni, hogy a szubjektív kiválasztás egyébiránt egy jól bevált mintavételezési eljárás abban az értelemben, ha a kiválasztást végző személy szakmai tapasztalataira támaszkodva annak érdekében befolyásolja a kiválasztást, hogy igazoltan javítsa a minta tulajdonságait, esetleg éppen a reprezentativitását.

amint azt Szép Katalin a következőképpen említi. „A lakossági felvételeknél egyre nagyobb gond az adatszolgáltatási hajlandóság romlása, amely egyes rétegeknél már-már az adatok megbízhatóságát veszélyezteti. A Háztartási Költségvetési Felvétel az első kísérlet arra, hogy az egyes rétegek válaszolási hajlandóságában megmutatkozó különbségeket már a mintavételi eljárásnál figyelembe vegyük, és így kisebb költséggel érjük el a kívánt pontosságú eredményeket." Szép K. (2004. p.646.)

A megelőzés során elsődleges cél nyilvánvalóan a válaszadási arány minél magasabb szintre történő növelése, lásd Peytchev – Conrad – Couper – Tourangeau (2010.) .A preventív eszközök a nemválaszolás jelentős mértékű csökkentésében is hathatós segítséget nyújtanak.

- Ilyen például az, ha a minta tagjait a személyes felkeresés előtt levélben tájékoztatják arról, hogy mikor és milyen céllal fogja őket egy kérdezőbiztos felkeresni.
- Lehetőség van előzetes telefonos felkérésre és időpont egyeztetésre is.
- Ha a kérdezőbiztos nem találja otthon a megkérdezendőt, hagyhat neki üzenetet, hogy milyen céllal kereste és mikorra várható, hogy ismét megkeresi.
- Befolyásolja a válaszadási hajlandóságot az is, hogy milyen a kérdezőbiztos megjelenése, beszédmódja, milyen a bevezető, amellyel megkéri a válaszadásra a megkérdezendő személyt.
- A válaszadók motiválása, melynek során fel kell kelteni érdeklődésüket, növelni érintettségüket, esetleg anyagi ösztönzést alkalmazni.

A mintavétel során a nemválaszolás két fő kérdésköre a válaszadási arány növelése és a nemválaszolásból eredő hiba korrigálása. A nemválaszolók eltérnek a válaszadóktól demográfiai, személyes, magatartási, stb. tulajdonságokban. Ha a kutatott jellemzőben eltérnek a válaszadók a nemválaszolóktól, akkor a becslés komoly mértékben torzul. Az alacsonyabb válaszadási arány növeli a nemválaszolásból eredő hiba valószínűségét, ezért kísérletet kell tenni a válaszadási arány növelésére. Malhotra (2005.)



Forrás: Malhotra (2005. p.443.)

2. ábra: Válaszadási arány növelésének módszerei

A válaszadási arány növelése azonban nem garantálja a nemválaszolás okozta torzítás csökkentését, ez derül ki Billiet et. al. (2007.) Európai szociális felmérés során szerzett tapasztalataiból is.

A NEMVÁLASZOLÁS KORRIGÁLÁSÁNÁL KÖVETHETŐ STRATÉGIÁK A KÖVETKEZŐK:

„Több eljárás közül talán Hansen és Hurwitz módszere a legismertebb. Ennek lényege az, hogy amennyiben a nemválaszolók köréből véletlen almintát választanak ki, és intenzív erőfeszítéseket tesznek annak érdekében, hogy ebben a körben a válaszadás javuljon, úgy a becslés torzítása megszüntethető vagy legalábbis csökkenthető. A módszer jellegzetesen olyan esetekben alkalmazható, amikor első lépésben postai kérdőívet használunk, majd a nemválaszolók köréből választott véletlen mintaelemeket személyes interjúk segítségével próbáljuk meg válaszadásra bírni. Ez az eljárás a kétfázisú mintavételre épülő becslések egy alkalmazása.” Hunyadi – Vita (2002. p.301.)

„Pótlás esetén a kutató kicseréli a nemválaszolókat a mintavételi keret olyan elemeivel, amelyekről feltételezik, hogy válaszolni fognak, és amelyek egyéb jellemzőikben hasonlítanak a nemválaszolókra. Az eljárás nem csökkenti a torzítás mértékét, ha a helyettesítők hasonlítanak a már mintában szereplő válaszadókhoz.” Malhotra (2005. p.447)

„A nemválaszolások okozta torzítás nagyon gyakran abból adódik, hogy kényes kérdésekre az emberek nem szívesen válaszolnak, illetve válaszaik a valóságot szépítik, meghamisítják. Az ilyen kényes kérdések ugyanakkor gyakran igen lényeges részét képezik a felvételeknek. Ezért dolgoztak ki a statisztikusok olyan eljárásokat, amelyekkel a kérdezett őszinte választ adhat és ugyanakkor nem kell felfednie anonimitását. Erre szolgálnak a manapság egyre népszerűbbé váló randomizálási eljárások, amelyek lényege az, hogy a

válaszoknak csak az összessége, statisztikai értékelése lehetséges, az egyediség az alkalmazott eljárás során elvész. Tekintettel a személyes adatok egyre fokozottabb védelmére, ezeknek az eljárásoknak az alkalmazása ennél jóval szélesebb körben is várható.” Hunyadi – Vita (2002. p.301.)

3.4.2. A HIÁNYZÓ ADATOK KEZELÉSE

A hiányzó adatok kezelésére Oravecz (2008) a következő széles körben is alkalmazott módszereket említi Little-Rubin (2002.) alapján:

- teljesen megfigyelt, vagy elérhető egységek elemzésén alapuló eljárások,
- átsúlyozás,
- imputáció,
- modell alapú eljárások.

ADATHIÁNYOS ESETEK MELLŐZÉSE

A módszer lényege, hogy azokra a megfigyelési egységekre vonatkozó adatokat, amelyek teljes, vagy részleges nemválaszolók egyszerűen nem veszik figyelembe az elemzés során. Ez a módszer akkor alkalmazható, ha a nemválaszolás mértéke alacsony és teljesen véletlenszerű nemválaszolásról van szó. Ezt persze a gyakorlatban igen nehéz megállapítani előzetes kutatások és kiegészítő információk híján. A módszer sikeresen alkalmazható olyan kutatások esetében, melyek nem a különböző lekérdezett változók között keresnek összefüggéseket, azaz nem feltáró jellegű kutatások, csupán leíró jelleggel a meglevő teljes körűen megválaszolt változók alapján történő jellemzéssel foglalkoznak. A longitudinális, vagy feltáró kutatások esetében azonban vizsgálni kell a hiányos és a rendelkezésre álló változók közötti összefüggést is a hatékony eljárás megválasztásához.

ÁTSÚLYOZÁS

Ez a módszer abból indul ki, hogy a válaszoló egyedek és a nemválaszolók között minden bizonnyal vannak hasonlóságok, legalábbis az egyedek rétegezhetők valamilyen, a vizsgált változóval sztochasztikus kapcsolatban levő ismérv alapján. Ebben az esetben az egyes rétegekbe tartozó válaszolók értékeinek nagyobb súlyt adnak, mint amekkora arányuk valójában van a tervezett mintában, mégpedig annak érdekében, hogy képviseljék a hasonló

tulajdonságokkal rendelkező, de nemválaszoló egyedeket. Amennyiben a j-edik alcsoportban a válaszadók aránya p_j, akkor az itt szereplő elemek $1/p_j$ súlyt kapnak, azaz itt mindegyik elem ennyiszer több sokasági elemet képvisel.

IMPUTÁCIÓ

Az imputáció során a hiányzó adatot pótolják egy ahhoz hasonlónak feltételezett értékkel. A hasonló értékek imputálására alkalmazható számos módszer közül az adathiány természete-, a becslés célja-, illetve a kutatói tapasztalatok függvényében kell választani. Vannak esetek amikor logikailag kikövetkeztethető a hiányzó adat a többi változóból, vagy korábbi felmérésekből.

„Ha csak egy-egy válasz (nem pedig egy egész elem) hiányzik a mintából, akkor a következő módszerek jöhetnek szóba:

A deduktív imputáció ritkán használatos módszer, mert lényege az, hogy az adott kérdőív más kérdései alapján lehet következtetni a hiányzó válaszra. Elvben az is előfordulhat, hogy a többi kérdés alapján a hiányzó választ pontosan rekonstruálni lehet, de ez valóban csak ritka és kivételes esetekben képzelhető el.

A főátlag-imputálás annyit jelent, hogy a hiányzó választ a megvalósult minta főátlagával pótoljuk. Ez az eljárás az átlag becslése szempontjából nem torzít, de az elemek változékonyságát feltehetően csökkenti, ezért a szórást (és ebből adódóan a becslések hibáit is) általában alulbecsli.

A részátlag imputálás ehhez hasonló, de némileg kifinomultabb módszer. Ekkor külső információk alapján imputációs csoportokat hozunk létre, képezzük ezen csoportok átlagait, és a hiányzó megfigyelést a neki megfelelő csoport átlagával becsüljük. Ez az eljárás pontosabb ugyan, mint az előző, de a szórás alulbecslése itt is jellemző, bár az előzőnél kisebb mértékben.” Hunyadi – Vita (2002. pp.302-303.)

A teljes válaszhiány esetén gyakran alkalmazott eljárás a regresszión alapuló imputálás. Ebben az esetben a hiányzó adatokat tartalmazó változó lesz a regresszió eredményváltozója, a többi változó pedig magyarázóváltozóként funkcionál. A regresszión alapuló imputációs módszereknél a logit, probit modellek mellett maximum likelihood becsléseket is alkalmaznak.

Az adathiányos esethez leginkább hasonló − később a nemválaszolás pótlásául szolgáló − egyedek megtalálásában klaszter-elemzés is alkalmazható, amennyiben az adathiánytól mentes változók nem korrelálnak egymással.

3.5. A SZÜKSÉGES MINTAELEMSZÁM PARAMÉTERES BECSLÉSE

A mintavételre épülő felmérések végrehajtásának kezdetén általában felmerül az a kérdés, hogy mekkora mintára van szükség. Részletesebben vizsgálva a problémát, megállapítható, hogy a gyakorlati életben a döntéshozók gondolatai elsősorban nem a kiválasztandó minta mérete körül forognak. A döntési mechanizmusban meghatározzák, hogy mekkora bizonytalanságot tudnak tolerálni. Másképpen fogalmazva a következtetések eredményének mekkora az elvárt megbízhatósága és pontossága. Ezért tartom fontosnak bővebben szólni az adott megbízhatósági követelmények mellett alkalmazható minta elemszámának meghatározásáról.

A minta méretét, azaz a sokaságból kiválasztott elemek számát több külső dolog determinálja, úgymint a rendelkezésre álló anyagi források, a sokaság nagysága, összetétele, változékonysága, stb. A mintanagyság meghatározásának kvalitatív szempontjait Malhotra (2005.) a következőkben foglalja össze:

- a döntés súlya,
- a kutatás természete,
- a változók száma,
- az elemzés módja,
- hasonló tanulmányokban használt mintanagyság,
- az előfordulási arány,
- a megvalósulási arány,
- a rendelkezésre álló források.

A kvalitatív tényezők mellett azonban kvantitatív szempontok is érvényesülnek a mintaelemek számának meghatározásakor. Statisztikai szempontból a minta méretét – az intervallumbecslés elméletéből kiindulva – az elvárt megbízhatóság és pontosság is befolyásolja. A standard hiba képletét szemlélve egyértelmű, hogy a hiba és a minta elemeinek száma fordított négyzetes viszonyban áll egymással. Tehát a pontosság növeléséhez növelni kell a mintalemszámot. A megalapozott következtetésekhez megbízható információkra

van szükség, melyeket kellően nagy minta vizsgálatával lehet előállítani. Egy gyakran emlegetett, már-már közhellyé vált megfogalmazás szerint „a kis minta nem minta”, viszont tekintettel kell lenni arra a tényre, hogy a minta méretének növelése jelentős költségnövekedéssel párosul. A minta elemszámának emelése azonban egy bizonyos méreten túl már nem eredményez jelentős javulást a pontosságban (különösen igaz ez homogén sokaságok esetében). Az elméleti közgazdaságtan oldaláról szemlélve a minta növelése által felemésztett költségnövekmény és a következtetések pontosságának viszonyát, a csökkenő határhaszon elve írja le. Ezért az indokolatlanul nagy mintaelemszám fölösleges költségtöbbletet eredményezhet, az amúgy is szűkös kutatási keret terhére. Ezért a minta méretének kvantitatív szempontok szerinti meghatározása rendkívül hasznos és fontos folyamat.

3.5.1. STATISZTIKAI MINTAILLESZTÉS PROGRAM

A Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központjának munkájába bekapcsolódva a Miskolci Egyetem Üzleti Statisztika és Előrejelzési Tanszékének kollektívájával karöltve kifejlesztettünk egy, a minta szükséges elemszámának meghatározásával kapcsolatos döntéseket támogató programot.

3. ábra: A Statisztikai Mintaillesztés szoftver induló oldala

A megismert mintavételi eljárások elméleti módszertanának keretrendszere segítségével végrehajtható, a matematikai statisztikai elemezhetőséget is feltétlenül biztosító minta kiválasztása. Az egyes eljárások megvalósítási menetét követve, a kiválasztott egyedek által közvetített információkból megalapozott döntések hozhatók. Erre akkor kerülhet sor, ha az alkalmazott mintavételi eljárások közül a kutató képes kiválasztani a vizsgálat szempontjából optimálisnak tekinthetőt. A program alkalmazása megkönnyíti a mintavétel körülményeihez leginkább igazodó mintavételi technika kiválasztását.

3.5.2. A MINTAVÉTELEZŐ PROGRAM TERVEZÉSÉNEK, MEGVALÓSÍTÁSÁNAK SZEMPONTJAI

A program segítségével meghatározható az adott kutatáshoz leginkább illeszkedő mintavételi eljárás, melynek eredményeként a partnerek indukálta kutatásokban lehetőség nyílik az összehasonlítható, megismételhető elemzésekre.

Ezen statisztikai mintaillesztő program alkalmazásával a vállalkozások, és egyéb szervezetek az innovációs tevékenységüket megalapozó piackutatást idő- és költséghatékonyan tudják megvalósítani. A program használatával lehetőségük nyílik arra, hogy önállóan is megtervezzenek és elindítsanak egy felmérést, mely által megteremthetik az innováció hasznosulásának alapját, s ezáltal a vállalat, szervezet versenyképessége, foglalkoztatási potenciálja és stabilitása növekedhet.

A szoftver tervezésének első lépéseként modelleztük a rendszer várható működését, amely alapján elvben teljes és ellentmondásmentes struktúra kialakítására került sor.

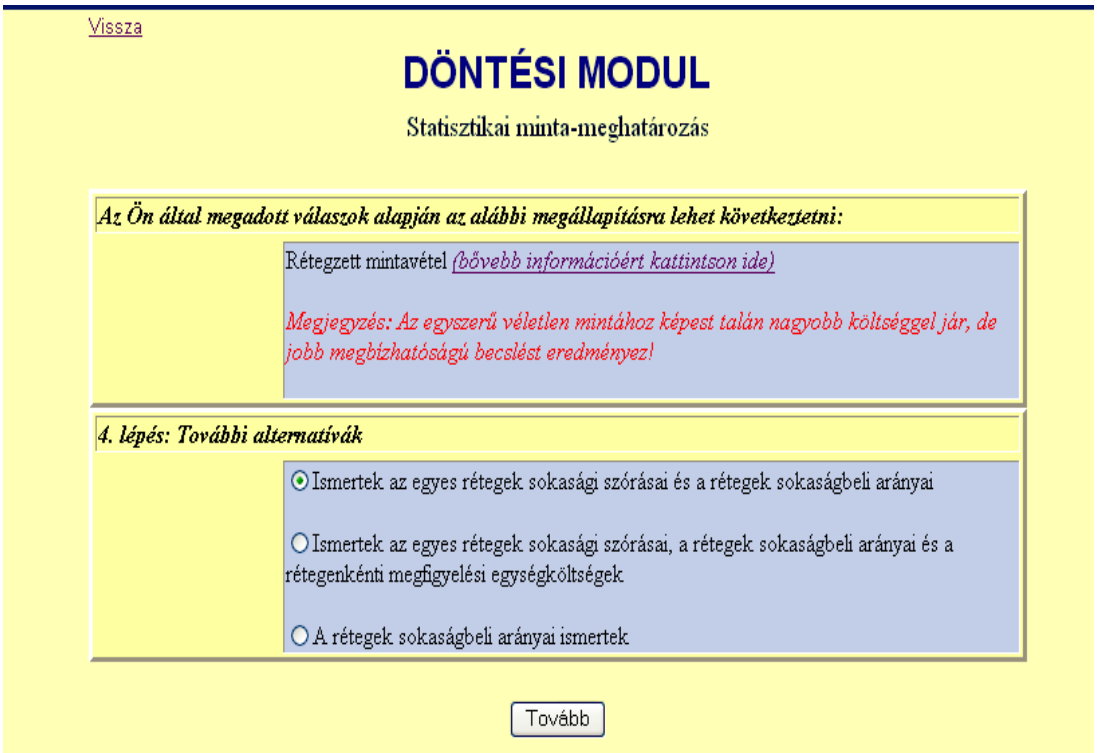
A program a következő tématerületeken végzett kutatások tervezéséhez alkalmazható:

- termékekkel kapcsolatos kutatás,
- vevőkkel, szállítókkal, meglevő, vagy potenciális partnerekkel, megelégedettséggel kapcsolatos kutatás,
- szociológiai, humán kutatás,
- pénzügyi műveletek ellenőrzése.

A program alkalmazásához elengedhetetlen a mintavételezési tevékenységhez szükséges és potenciálisan a felhasználó rendelkezésére álló inputok, valamint az outputok meghatározása.

Feltétlenül szükséges, hogy a felhasználó behatóan ismerje a kutatni szándékozott területet. A vizsgálni kívánt sokaság egységeinek jellemző tulajdonságait, összetételét, elérhetőségét, fizikai alkalmasságát a vizsgálatok elvégzésére.

Ezen információk és tudásbázis birtokában a program által feltett kérdésekre felkínált válaszokból a megfelelőt kiválasztva juthat el a felhasználó az első szintű eredményekig, miszerint meghatározta az alkalmazandó optimális eljárást.



4. ábra: A program által indukált alkalmazható optimális döntés

A továbbiakban szót kell ejteni néhány a sokaságot, és a mintavétel körülményeit érintő információról, melyek ismerete feltétlenül nélkülözhetetlen a megfelelő minta kialakításához.

Az előbbiekben már említett első szintű eredmények után a program a következő imputok és képletek felhasználásával határozza meg a kiválasztott mintavételi eljárás keretein belül vizsgálni szükséges mintaelemszámot.[5]

[5] A program alkalmazásáról, működéséről bővebben lásd (Besenyei et. al. 2006.)

1. táblázat: A felhasznált információk és képletek

Független azonos eloszlású minta	
Szórás (s), Megbízhatósági szint (π), Maximális hiba értéke (Δ)	$n = \left(\frac{z_\pi \cdot s}{\Delta} \right)^2$
Egyszerű véletlen minta: Mennyiségi ismérv várható értékének becslésére	
Szórás (s), Megbízhatósági szint: (π), Maximális hiba értéke (Δ), Sokaság elemeinek száma (N)	$n = \frac{s^2 \cdot N}{\left(\frac{\Delta}{z_\pi} \right)^2 \cdot N + s^2}$
Egyszerű véletlen minta: Arány becslésére	
Megbízhatósági szint: (π), Maximális hiba értéke (Δ), Sokaság elemeinek száma (N), Adott tulajdonsággal rendelkező egyedek aránya (p)	$n = \frac{z_\pi^2 \cdot p \cdot q \cdot N}{\Delta^2 \cdot N + z_\pi^2 \cdot p \cdot q}$
Rétegzett mintavétel: Arányos rétegzés	
A minta elemszáma (n), A sokaság elemszáma (N), A rétegek sokasági elemszáma (N_j)	$n_j = n \cdot \frac{N_j}{N}$
Rétegzett mintavétel: Egyenletes rétegzés	
A rétegek száma: (L), A minta elemszáma (n)	$n_j = \frac{n}{L}$
Rétegzett mintavétel: Neyman-féle optimális rétegzés	
A minta elemszáma (n), A rétegek szórása (σ_j), A rétegek sokasági elemszáma (N_j)	$n_j = n \cdot \frac{N_j \cdot \sigma_j}{\sum_{j=1}^{L} N_j \cdot \sigma_j}$
Költség-optimális rétegzés: A rétegek aránya	
A rétegek szórása (σ_j), A rétegek sokasági elemszáma (N_j), Az egyes rétegekhez tartozó megfigyelések egyedi költsége: (c_j)	$\frac{n_j}{n} = \frac{\left(N_j \cdot \sigma_j \right) / \sqrt{c_j}}{\sum_{j=1}^{L} \left(N_j \cdot \sigma_j / \sqrt{c_j} \right)}$
Költség-optimális rétegzés: Adott költségkeret mellett optimális bizonytalansággal	
A mintavétel költségkerete: (C), A rétegek szórása (σ_j), A rétegek sokasági elemszáma (N_j), Az egyes rétegekhez tartozó megfigyelések egyedi költsége: (c_j)	$n = \frac{C \cdot \sum_{j=1}^{L} \left(N_j \cdot \sigma_j / \sqrt{c_j} \right)}{\sum_{j=1}^{L} \left(N_j \cdot \sigma_j \cdot \sqrt{c_j} \right)}$
Költség-optimális rétegzés: Adott megbízhatóság mellett a költségek optimalizálásával	
A sokaság elemszáma (N), Megbízhatósági szint: (π), Maximális hiba értéke (Δ), A rétegek szórása (σ_j), A rétegek sokasági elemszáma (N_j), Az egyes rétegekhez tartozó megfigyelések egyedi költsége: (c_j)	$n = \frac{\sum_{j=1}^{L} \left(\frac{N_j}{N} \cdot \sigma_j \cdot \sqrt{c_j} \right) \cdot \sum_{j=1}^{L} \left(\frac{N_j}{N} \cdot \sigma_j / \sqrt{c_j} \right)}{\left(\frac{\Delta}{z_\pi} \right)^2 + \frac{1}{N} \cdot \sum_{j=1}^{L} \frac{N_j}{N} \cdot \sigma_j^2}$

A független azonos eloszlású mintától eltekintve, a többi módszer esetében véges sokaságból történő mintavételt alkalmazva adódnak a mintaelemszámok. (Ezen túlmenően a végtelen sokaságok, illetve a visszatevéses kiválasztás esetén szükséges összefüggéseket nem mutatom be, hiszen a tapasztalt kutatók a becslő függvényekben eleve alkalmazzák a szükséges korrekciós tényezőket.)

A program által meghatározott mintanagyság adja azt a mintát, mellyel biztosító, hogy a paramétereket egy megkívánt pontossági fokon és egy adott megbízhatósági szinten becsülni lehessen. Fel kell hívni a figyelmet azonban arra, hogy a program eredményei általános estet és normális eloszlást feltételeznek. A kalkuláció során nem kerülnek beépítésre olyan speciális tényezők, mint az extrém eloszlások, vagy a lehetséges ismérvváltozatok determinálta hatás. Például Kehl – Rappai (2006.) tanulmányában bizonyította, hogy Likert-skála alkalmazása esetén lényegesen nagyobb mintára van szükség, mint ami az általános feltételek alapján becsülhető.

A mintavétel tervezésénél mindemellett figyelembe kell venni a meghiúsulások mértékét is. A végső mintanagyság kialakítása érdekében, sokkal nagyobb számú lehetséges válaszadóval kell kapcsolatba lépni (lásd nemválaszolási hiba).

4. A MINTAVÉTELI TERVEK MINŐSÍTÉSE

A fejezet a különböző mintavételi eljárások, mintavételi tervek alapján kialakított minták adatai, valamint az ezeken alapuló becslések minőségének, minősítésének néhány kérdésével foglalkozik.

Természetesen különbséget kell tenni az egyedi statisztikai adatok minősége és a hivatalos statisztikai szolgálat által publikált adatok minősége között, ahogy arról beszámol Fellegi (2001.). Noha a hivatalos statisztikák esetében jogosan várható el a minőség iránti fokozottabb igény, azt azonban el kell fogadni, hogy az egyedi statisztikák minőségének javításához követni kell a hivatalos statisztikák számára megfogalmazott követelményeket. A statisztikai minőség értelmezésére megfogalmazott főbb irányelveket részletesen mutatja be Szép-Vígh (2004.) elsősorban az Encyclopedia of Statistical Sciences, valamint az ISI, az ESR, az IMF, az OECD, és további minőségdefiníciók alapján. Ezek szerint a legfontosabb kritériumok a pontosság, tartalom, az időszerűség, a koherencia és összehasonlíthatóság, valamint a hozzáférhetőség és átláthatóság. Jól érzékelhető tehát, hogy a teljes körű minőségrendszerek fejlődése a statisztikai minőség megfogalmazását a szűken értelmezett statisztikai pontosság jelentésén túlra is kiterjesztette.

A következőkben olyan szempontok kidolgozására vállalkoztam, amelyek alapján – ha nem is a minőség összes értelmezhető kritériuma tekintetében, de néhány fontosabb jellemző alapján – minősíteni, rangsorolni lehet a különböző mintákból nyert adatokat, becslési eredményeket. Természetesen mindezt számos elméleti feltétel fennállása mellett kíséreltem meg, mely feltételek az alkalmazott gyakorlatban meglehetősen ritkák, olykor egyáltalán nem teljesülnek.

Mindenekelőtt feltételeztem a minta teljes realizációját, miszerint minden megkérdezett releváns választ ad minden feltett kérdésre. Ez a feltétel egyértelművé teszi, hogy nem törekedtem a teljes hiba mértékének becslésére. A 100%-os válaszadási arány mellett feltételeztem, hogy további a kérdezőbiztos által elkövetett, és az adatrögzítés során elkövetett hibák, valamint semmilyen egyéb nem mintavételi hiba nem torzítja az eredményt.

Nem vettem figyelembe továbbá a mintavételi tervek azon törekvését, hogy bármilyen értelemben vett költséghatékonyságot biztosítsanak, sem a tervezési, sem a lekérdezési, sem pedig az adatrögzítési, illetve kiértékelési munkaszakaszokban.

Nem célom biztosítani az egyes munkafázisokban tevékenykedő kutatók egyenletes terheltségét és semmilyen egyéb – akár gazdasági szemléletű, akár időmegtakarítási – kritériumot. Ezen feltételek szellemében a mintavételnek a becslési eredmények szempontjából értelmezhető hatékonyságának vizsgálatára törekedtem.

4.1. A VIZSGÁLT ADATBÁZIS BEMUTATÁSA

Az empirikus vizsgálatok elvégzése előtt szükségesnek tartom bemutatni az elemzések alapját nyújtó adatbázis tartalmát és forrását. Egyrészt bizonyítandó, hogy a doktori munka hiteles adatokat tartalmaz. Másrészt a kutatás eredményeinek megfelelő felhasználása érdekében. Hiszen kutatási eredményeim, mint általában az elemzések outputjai csak az adatok forrásának és hátterének megfelelő ismeretében értékelhetők.

A háztartási költségvetési felvétel (HKF) olyan adatgyűjtés, mely alkalmas állomány jellegű (stock) és folyamatot tükröző (flow) információk szolgáltatására – elsősorban a lakosság fogyasztási kiadásairól, valamint jövedelmi, vagyoni és egyéb gazdasági viszonyairól, folyamatairól. A KSH több évtizedes múltra visszatekintő adatfelvétele, melynek módszertani alapjai lényegében a ma alkalmazott módszerek bázisát jelentik, megfelelő alapot biztosít a mintán alapuló kísérleti számítások, becslések, következtetések vizsgálatára. A következőkben röviden bemutatom a HKF célját, módszereit, [6] jellemzőit.

4.1.1. A HÁZTARTÁSI KÖLTSÉGVETÉSI FELVÉTEL NÉHÁNY JELLEMZŐJE

A HKF olyan önkéntes adatgyűjtés, amely információt szolgáltat a háztartások és személyek fogyasztási színvonaláról és szerkezetéről, a jövedelmekről, a háztartástagok demográfiai, gazdasági aktivitási, iskolázottsági jellemzőiről. Emellett tájékoztatást nyújt a különböző társadalmi rétegek lakáskörülményeiről, a vásárolt tartós fogyasztási cikkek állományáról.

[6] A HKF bemutatásakor a KSH szakembereinek, valamint az MTA KTI kutatóinak (elsősorban Kapitány – Molnár 2001.) munkáira támaszkodom.

A felvétel országosan reprezentatív többlépcsős rétegzett mintavételi eljárást alkalmaz a magánháztartások körében, ami egész évben folyamatosan zajlik. Az éves felmérés havi szintű adatgyűjtésekre oszlik, amit egy az egész éves időszakra visszatekintő (ellenőrzési funkciót is betöltő) lekérdezés követ. A minta reprezentativitási nehézségei, valamint bizonyos rétegek (pl.: leggazdagabb és legszegényebb rétegek) teljes hiánya miatt számos kritika érte a felvételt a társadalomtudományi kutatók részéről. Lásd: Molnár-Kapitány (2006.), Havasi (2002.). A reprezentativitás legfőbb problémáit Kapitány – Molnár (2001.) abban látja, hogy a nyugdíjasok és a munkanélküliek túlreprezentáltak, az aktív keresők, valamint a felsőfokú végzettségűek pedig alulreprezentáltak a mintában. Emellett problematikusnak találják a vállalkozók csekély számát is.

A felvételből származó adatok reprezentativitásának biztosítása érdekében a szakstatisztika készítői több eljárást alkalmaznak, melyek között demográfiai és gazdasági aktivitás szerinti kalibrációt végeznek.

A hiányzó adatok kezelését illetően a HKF-ben a kiadási tételek esetében a hasonlósági elven történő imputálás és az adatbázisból történő arányos pótlás (hot deck) alkalmazására kerül sor. „A célvizsgálatok azt jelzik, hogy a nemválaszoló háztartások általában a magasabb jövedelműek közül kerülnek ki. Ez hiányt okoz a nagy értékű cikkek vásárlásánál, illetve a háztartás életvitelében fontos szerepet játszó kiadásoknál.”[7]

A HKF (mint mikroszemléletű lakossági felvétel) mellett makroszemléletű adatforrás is informálja a kutatókat a lakossági fogyasztási kiadásokról, mégpedig a nemzeti számlarendszer háztartási szektor számlái alapján. A HKF azonban önkéntes lakossági adatgyűjtésből származik, ezért csak a háztartások fogyasztásában jelenlévő termékekre és szolgáltatásokra terjed ki. Ennek alapján a két adatforrás a következő okokból kifolyólag eltérő információkat nyújt.

- „A háztartási költségvetési felvételek megfigyelési körébe csak a magánháztartásokban élő népesség tartozik, az intézeti háztartásban élőket nem veszi figyelembe.
- Az adatbázis csak a kiadásokat tartalmazza, a folyó termelő felhasználást pedig nem.
- A háztartási felméréseknél a kiadások számbavétele a háztartások oldaláról történik és nem pedig a kibocsátókéról, ahogy a makroszemléletű nemzeti számlarendszerben.

[7] http://portal.ksh.hu/pls/ksh/ksh_web.meta.objektum?p_lang=HU&p_menu_id=110&p_almenu_id=104&p_ot_id=100&p_obj_id=ZHC&p_session_id=66509802 (letöltve: 2010. 09. 21.)

– A háztartások adatai pénzügyi szemléletűek szemben az eredmény szemléletű makrogazdasági adatokkal.„ KSH (2007. p.19.)

A HKF ezek alapján alacsonyabb fogyasztási színvonalat és más szerkezetet mutat, mint a háztartási szektor számlái, ami a kutatási eredményeim makro szintű összehasonlítását nem teszi lehetővé, de mintavételi módszertani következtetések levonására tökéletesen alkalmas.

4.2. AZ ALKALMAZOTT MINTAVÉTEL

Egyes minták kiválasztása során a 3.6. alfejezetben bemutatott Statisztikai Mintaillesztés szoftver volt segítségemre. Továbbá olyan minták is kiválasztásra kerültek, melyek elemszámát előre meghatároztam annak érdekében, hogy az eltérő mintaméretek ne befolyásolják az összehasonlítás eredményeit.

4.2.1. EGYSZERŰ VÉLETLEN MINTÁK (EV)

Az egyszerű véletlen mintavételt három különböző méretű mintán alkalmaztam. Elsőként egy standardnak számító mintaméretet választottam 150 háztartás kiválasztásával, ami az alapsokaság 1,66%-át jelentette. A minták elnevezése a következő:

– **EV1150**
– **EV2150**
– **EV9150**

Emellett az elméleti síkon jobbnak tekintett 10%-os kiválasztási arányt produkáló kb. 900 elemű mintákat képeztem.

– **EV1900**
– **EV2900**
– **EV9900**

Valamint a kismintás tulajdonságok tesztelése érdekében 30 elemű véletlen kiválasztással további három mintát generáltam:

- **EV130**
- **EV230**
- **EV930**

Mindhárom mintaméret esetében, ahogy az fent látható, a minták kiválasztása három típusban történt (EV1…, EV2…, EV9…) az SPSS program véletlenszám generátorait felhasználva. A véletlen számok generálása a különböző statisztikai és adatelemzési, adatkezelési szoftverekben eltérő algoritmusok alapján végezhető, melyek között esetenként jelentős eltérések tapasztalhatók. Annak érdekében, hogy az említett esetleges eltérésekre rávilágítsak, három típusú véletlen minta készült minden mintanagyságra. A mintaegyedeket az adatbázisban szereplő háztartások/háztartástípusok sorszáma alapján választottam ki.

4.2.2. RÉTEGZETT MINTÁK

A rétegzett minták kiválasztásakor három mintacsoportot különítettem el, a rétegzés módjának függvényében.

Az első csoport az egyszeresen rétegzett mintákat tartalmazza. Ezeknél a mintáknál mindig egyetlen rétegképző ismérv alapján rétegeztem a sokaságot. A kiválasztott rétegképző ismérvek általában vagy ordinális skálán mérhető tulajdonságok, vagy diszkrét mennyiségi ismérvek voltak (természetüknél fogva viszonylag kevés ismérvváltozattal). Minden kiválasztásnál arányos rétegzést alkalmaztam egy kivétellel, ami egyenletes rétegzést eredményezett. – Meg kell említeni, hogy voltak olyan rétegképző ismérvek, melyek szintén közel egyenletes rétegzést eredményeztek. – Minden kiválasztott ismérv szerinti rétegzést alkalmaztam egyaránt egy 1,66%-os és egy 10%-os kiválasztási arányt érvényesítő mintára. Ez utóbbi minták neve mindig 900-ra végződik.

A kialakított minták rendre a következők:

REG_HÁZT_HKF/ REG_HÁZT_HKF900: A minta az adatbázisban szereplő adatok régiókra történő csoportosítása után lett kiválasztva, régiónként egyszerű véletlen minta alkalmazásával, az SPSS 17.0 komplex mintatervező modulja segítségével a régió szerinti arányoknak megfelelően.

SŰRŰSÉG/ SŰRŰSÉG900: A minta az adatbázisban szereplő adatok területi csoportosítása után lett kiválasztva, ahol a csoportképző ismérv változatait az adott terület népsűrűségi kategóriái alkották az adatbázisban HA09 (Population density domain). A csoportokon belül egyszerű véletlen minta alkalmazásával, az SPSS komplex mintatervező modulja segítségével történt a kiválasztás, melynek eredményeként arányosan rétegzett mintát kaptam. Meg kell jegyezni, hogy a három kialakított csoport adatbázisbeli (így egyúttal mintabeli) megoszlása közelíti az egyenletes eloszlást.

AKTIV/ AKTIV900: A háztartásfő aktivitási státusza alapján történt a rétegek képzése, ami 7 különböző rétegbe sorolta a háztartásokat. A rétegek a HC12 (Current activity status of the reference person) változó ismérvváltozatai alapján különülnek el, melyek részletes leírása a változókat bemutató 2. számú mellékletben található.

CSALÁD/ CSALÁD900: Ezekben a mintákban a háztartás tagjainak száma alapján a HB05 (Household size) változó mentén 10 különböző réteget hoztam létre, melyekből arányos rétegzett mintákat választottam.

HÁZT TIP/ HÁZT TIP900: A háztartások, demográfiai jellemzőik alapján négy különböző csoportba kerültek besorolásra a HB07_3 (Household type) változó tartalma alapján.

NEM/ NEM900: A rétegzést természetesen nemek szerint is elvégeztem, ami a családfő (a felmérésben a legmagasabb jövedelemmel rendelkező személy) nemére vonatkozik a HC03 (Sex of reference person) változó alapján.

AUTO/ AUTO900: Feltételezve, hogy a fogyasztási kiadásokkal sztochasztikus összefüggésben van, a háztartásban fenntartott autók száma alapján is rétegeztem a sokaságot, amihez a HD14_02 (Number of cars) változót alkalmaztam. Mivel maximum négy autó létezik egy-egy háztartásban, így öt különböző réteget különítettem el.

TV/TV900: Az előbbihez hasonló megfontolás alapján a háztartásban levő televíziókészülékek száma szerint is rétegeztem. A HD14_14 (Number of televisions) változó segítségével öt réteget különítettem el.

REG_EGY/ REG_EGY900: Egyenletesen rétegzett mintát képeztem, melyben a statisztikai régiók csoportjaiból ugyanakkora mintát választottam egyszerű véletlen módszerrel. Annak érdekében, hogy az egyenletes elosztás könnyen megvalósítható legyen, az eddigiektől kissé eltérő elemszámú mintákat kaptam a hét statisztikai régióból.

A rétegzett minták következő csoportját a többszintű rétegzések alkotják. Az egy ismérv szerinti reprezentativitás ugyanis nem elégséges feltétel a gyakorlatban megvalósított mintavételes kutatások alkalmazásakor. A következő mintákban 2-3-szoros rétegzést alkalmaztam. Figyelembe véve annak kockázatát, hogy a többszörös rétegzés jelentősen megnöveli a rétegek számát, ezáltal csökkentve az egy rétegre jutó megfigyelési egységeket, az 1,66%-os kiválasztási arány mellett itt is alkalmaztam a 10%-os kiválasztást. Emellett, mivel nem szándékozom következtetéseket levonni az egyes rétegekre vonatkozóan, ezért nem jelent problémát, ha néhány rétegben rendkívül alacsony elemszámot produkálnak a minták.

REG_DENS/ REG_DENS900: Ebben a mintában először a földrajzi régiók szerint, majd azon belül a település népsűrűsége alapján rétegeztem a háztartásokat, arányos kiválasztást alkalmazva. (Ennél a rétegzésnél észrevettem, hogy az észak-alföldi régióban nincs reprezentálva a sűrűn lakott települések rétege a sokaságban. Ennek hatása a kutatás későbbi fázisaiban érzékelhető, ezért a problémára ott térek ki részletesebben.)

REG_AUTO_AKTIV/ REG_AUTO_AKTIV900: Ezekben a mintákban háromszintű rétegzést alkalmaztam: először regionálisan, ezután az autók száma, majd az aktivitási státusz alapján történt meg a rétegzés.

REG_AUTO_CSALÁD/ REG_AUTO_CSALÁD900: A regionális és az autók száma szerinti rétegzést a háztartásban élők száma szerinti rétegzés követte.

REG_AUTO_SZOBA/ REG_AUTO_SZOBA900: A regionális és az autók száma szerinti rétegzést az állandó lakásukban található szobák száma szerinti rétegzés követte.

REG_TV_AKTIV/ REG_TV_AKTIV900: Ezekben a mintákban szintén három szintű rétegzést alkalmaztam, először regionálisan, ezután az televíziók száma, majd az aktivitási státusz alapján történt meg a rétegzés.

REG_ TV _CSALÁD/ REG_ TV _CSALÁD900: A regionális és a televíziók száma szerinti rétegzést a háztartásban élők száma szerinti rétegzés követte.

REG_ TV _SZOBA/ REG_ TV _SZOBA900: A regionális és a televíziók száma szerinti rétegzést az állandó lakásukban található szobák száma szerinti rétegzés követte.

A rétegzett minták harmadik csoportjába az ún. mesterségesen rétegzett minták kerültek, ahol a rétegképző ismérv egy folytonos, vagy egy sok változattal rendelkező diszkrét

mennyiségi ismérv volt. Ennek eredményeként a rétegek számát nem határozta meg természetszerűleg a rétegképző ismérv változatainak száma, hanem azokat mesterségesen a decilisek segítségével különítettem el. Így a következő rétegzések 10 egyenlő nagyságú réteget eredményeztek a háztartások sokaságában. A mesterséges minták kiválasztásakor az egyszerű véletlen mintákhoz hasonlóan három különböző mintaméretet alakítottam ki minden rétegképző ismérv segítségével, melyek rendre 30, 150, 900 háztartásból álltak.

Az egyik rétegképző ismérv a vizsgálatok középpontjában levő háztartási fogyasztási kiadás HE00C (Total consumption expenditure), mellyel a következő mintákat alakítottam ki:

- **MR_FOGY_900**
- **MR_FOGY_150**
- **MR_FOGY_30**

Ezt követően a jövedelem alapján is képeztem a háztartások tizedeit a HA09_05 (Monetary net income) változó segítségével.

- **MR_JÖV_900**
- **MR_JÖV_150**
- **MR_JÖV_30**

A kor szerinti reprezentativitás sokszor jelentkezik követelményként a demográfiai vonatkozású felmérésekben, ezért a HC04 (Age (in completed years) of reference person) változó alapján a családfő életkora alapján is 10 réteget képeztem.

- **MR_KOR_900**
- **MR_KOR_150**
- **MR_KOR_30**

Végül pedig az állandó lakhelyül szolgáló lakás alapterülete segítségével végeztem el a rétegzést a HD07 (Useful living area in m^2 (principal residence)) változó segítségével.

- **MR_NM_900**
- **MR_NM_150**
- **MR_NM_30**

A rétegzett minták kiválasztásakor az adatbázis adatai mellett a KSH által publikált 2005. évi mikrocenzus adatait használtam a rétegek meghatározásához.

REG_HÁZT_KSH: Ebben a mintában arányos rétegzést alkalmaztam, ahol a rétegképző ismérv a NUTS I. szintű statisztikai régiókat jelentette. A mikrocenzus eredményeként

megjelenített háztartásszámmal arányos méretű mintát választottam minden régióból egyszerű véletlen módszerrel, az SPSS véletlenszám generátora segítségével.

REG_LAK_KSH: Ebben a mintában arányos rétegzést alkalmaztam, ahol a rétegképző ismérv a NUTS I. szintű statisztikai régiókat jelentette. A mikrocenzus eredményeként megjelenített lakosságszámmal arányos méretű mintát választottam minden régióból egyszerű véletlen módszerrel, az SPSS véletlenszám generátora segítségével.

4.3. A MINTAJELLEMZŐK ÖSSZEHASONLÍTÁSA

A különböző mintavételi tervek összehasonlítása céljából két változóra végeztem statisztikai becsléseket. Az egyik egy folytonos, arányskálán mérhető ismérv: a háztartások teljes fogyasztása, a másik változó pedig egy diszkrét ismérv szintén arányskálázással: a háztartások főben kifejezett nagysága. A változók részletesebb tartalmi bemutatása az 2. számú mellékletben található. A becsült paraméter mindkét esetben a várható érték, illetve az értékösszeg, melyből a későbbiekben meghatározható a két változó hányadosából számított – és a KSH által is publikált – háztartások átlagos egy főre jutó kiadási összege.

4.3.1. A MINTÁK RANGSOROLÁSA

A vizsgálatok során az összes generált minta esetében kiszámítottam a becslő függvényeket és azok szórását a 3.2.1. és 3.2.2. fejezetben megfogalmazottak alapján. A pontbecslés, – jelen esetben az átlag – szórása, másképpen a becslés standard hibája köztudottan jelentős szerepet játszik a becslés pontosságának, megbízhatóságának megítélésében, valamint elengedhetetlen az intervallumbecslések készítéséhez. Önmagában azonban alkalmatlan arra, hogy több egymástól független minta összehasonlítását lehessen elvégezni a segítségével annak eldöntésére, hogy melyik minta alkalmasabb a vizsgált paraméter becslésére. Ezért az elemzéshez további – a rangsorolásnál szóba jöhető – mintajellemzőket kellett megvizsgálnom.

Mivel speciális kísérletjellegű vizsgálatról van szó, ezért a minta kiválasztásának és megvalósításának jelen esetben nincsenek költségei, így az egységköltségek sem határozhatók meg és nem is szerepeltethetők a vizsgálatban.

Mivel annak ellenére, hogy egymástól független mintákat alkalmaztam, de a vizsgált változó minden esetben azonos, és a sokaság is változatlan, így a sokasági szórás alapján, – ami a gyakorlatban egyébként rendszerint stabilnak mutatkozik – szintén nem lehet hatást gyakorolni a rendezés eredményeire.

Kézenfekvően adódik a minták leginkább rugalmas jellemzőjének a minta nagyságának a vizsgálata, mely a fentebb leírtak alapján több méretben is realizálódik. A különböző mintavételi eljárások során kapott eredmények minden bizonnyal rávilágítanak arra a közismert feltevésre, mely a nagyobb méretű minták előnyös tulajdonságait hangoztatja.

Figyelembe véve azt, hogy számos mintavételi eljárást, mintavételi tervet alkalmaztam, feltétlenül ki kellett térnem a mintavételi terv becslésre gyakorolt hatásának vizsgálatára. A mintavételi terv hatásának számszerűsítését kiválóan leírja Marton – Mihályffy (1988.), Marton (1991.), Kish (1989.), és még számos további tanulmány. A mintavételi terv hatásának mérése a becslőfüggvény szórásának, illetve varianciájának felbontásán alapul, és a következő összefüggés alapján határozható meg:

$$DEFF = \frac{SE_A^2}{SE_{EV}^2}$$

ahol a számlálóban az aktuális mintavételi tervnek megfelelően számított paraméter varianciája található, a nevező pedig egy ugyanolyan méretű, de egyszerű véletlen mintavétel esetén számított variancia. Ez a mutató gyakorlatilag a klasszikus szórásnégyzet felbontás alapelvén működik, hiszen mint az köztudott, egy átlagbecslés standard hibájának négyzete – például rétegzett mintavétel esetén – megegyezik a belső szórásnégyzettel, míg az egyszerű véletlen kiválasztásnak megfelelő becslés esetén a rétegek csoportok varinaciája hozzáadódik az egyedi varianciákhoz.

A mutató értékelése egyszerűnek mondható, hiszen az egynél nagyobb értékek azt mutatják, hogy az adott mintavételi terv alapján készített becslés kevésbé hatásos (nagyobb szórással rendelkezik), mint egy egyszerű véletlen mintavételi terv becslése. Természetesen egynél kisebb eredmény esetében hatásosabb becslést jelez.

Ezek alapján feltételezhető, hogy mind hatásosság, mind pontosság szempontjából relevánsabb eredményeket biztosítanak azok a részletesebb információk alapján képzett minták, melyeknél a rétegzéshez több, a vizsgált változóval sztochasztikus viszonyban álló változó kerül bevonásra a mintavételi terv kidolgozásában.

Nem csak rétegzett minták esetében megfogalmazott elvárás a reprezentativitás. A gyakorlatban számos jellemző alapján várnak reprezentativitást a mintán alapuló felvételektől. Ezzel kapcsolatban két megállapítást tehető.

Az elvárás valós előnyöket realizál a becslési végeredményekben, abban az esetben, ha megfelelő változók mentén valósul meg a reprezentativitás. Vannak ugyan olyan esetek is, amikor bizonyos változók alapján indukált reprezentativitás kifejezetten a költségek csökkentésére, az adatgyűjtési munka megkönnyítésére, ésszerűbbé tételére irányul. Ezekben az esetekben a becslési eredményekre a reprezentativitás nincs bizonyítottan közvetlen hatással.

A több szempont szerinti reprezentativitásnak az a hátránya, hogy több ismérv alapján rengeteg réteg, illetve keresztosztály képződik, melyek mérete indokolatlanul kicsivé válhat. Bár elméleti síkon két elem elégséges lehet a variancia kiszámításához, gyakorlati szinten azonban köztudott, hogy ezek a számítások könnyen téves irányba terelhetnek. Mindezek ellenére a lehető legtöbb információ beépítése a mintavételi tervbe célravezetőnek tűnik. Annak ellenére, hogy a többszörös rétegzés eredményeként a variancia összetevőinek a száma rendkívül magas lehet, így a hibaszámítás nehézkessé válhat, a több ismérv szerinti reprezentativitás előnyei érzékelhetők a végeredményekben. A túl kis méretű részosztályok esetében pedig lehetőség van a rétegek összevonására, ami további hibákat generálhat, de a paraméter „durva" közelítésére alkalmas lehet. Leslie Kish Professzort idézve: „Megengedhetjük, sőt néha meg kell engednünk a paraméterek durva közelítő értékeinek használatát, mivel ezek hibája nem befolyásolja eredményeink érvényességét, bár csökkentheti a kutatás tervének hatékonyságát." Kish (1989. p.208.)

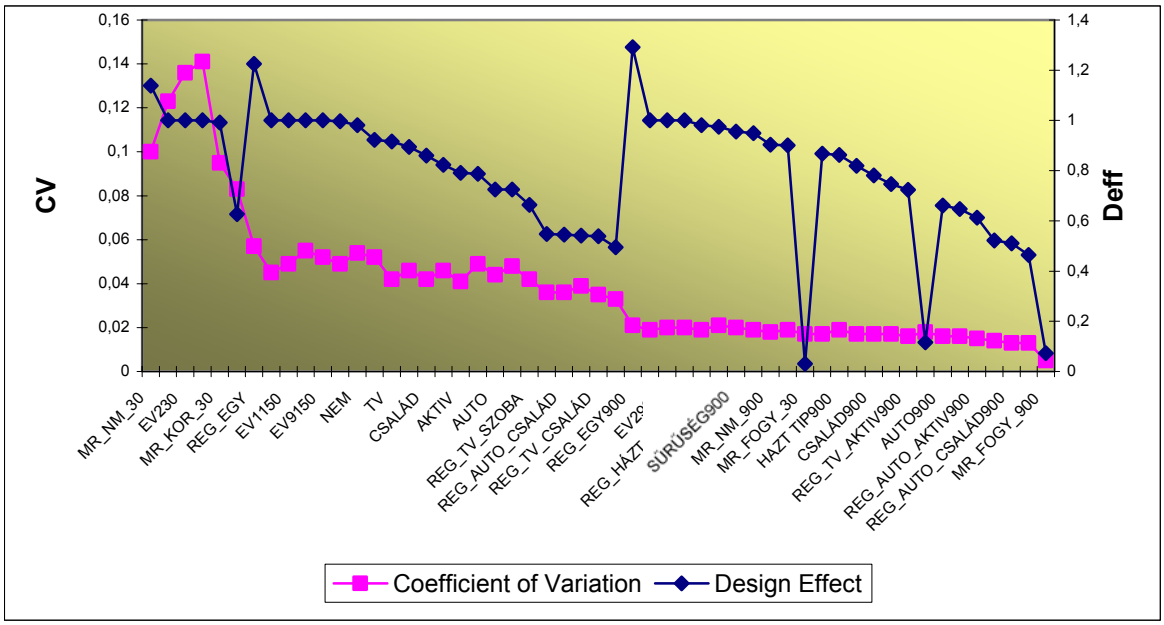
A vizsgált mintákat a következő jellemzők alapján rangsoroltam:

- mintavételi terv hatása/ *Design Effect* – Deff,
- paraméter relatív szórása / *Coefficient of Variation – CV* (variációs együttható), $\frac{SE\hat{\Theta}}{\hat{\Theta}}$ átlag esetében: $\frac{s_{\bar{x}}}{\bar{x}}$
- effektív mintanagyság: n/Deff.

Az effektív mintanagyságon itt azt értem, hogy a Deff értékét ismerve, kimutatható, hogy az eredeti mintához képest mekkora mintával lehetne ugyanolyan becslési eredményeket kapni. Tehát egynél nagyobb Deff érték esetében az effektív mintanagyság megmutatja,

hogy egy jobb mintavételi terv felhasználásával mekkora (vagy mennyivel kisebb) mintát kell vennünk azonos becslési eredményekhez.

Azt tapasztaltam, hogy az effektív mintanagyság tekinthető vezérelvnek a rangsorolás során, ami biztosítja a másik két szempont szerinti rangsor alakulását is.



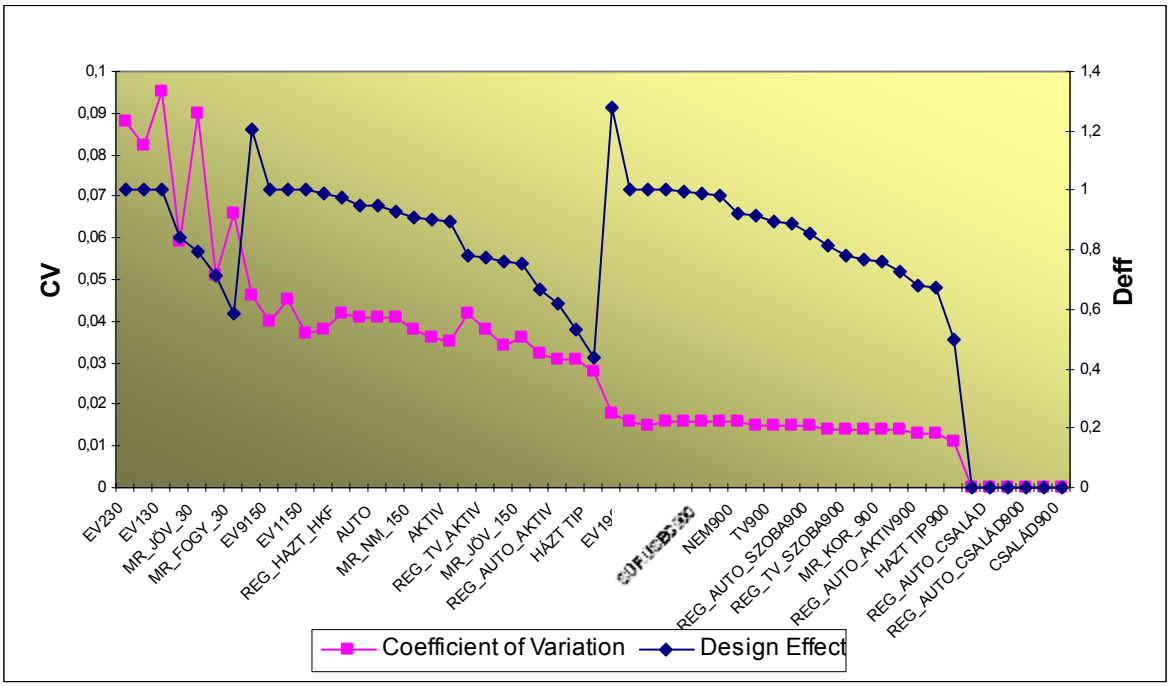
5. ábra: A háztartások teljes fogyasztása becslésének jellemzői az effektív mintanagyság szerinti rangsorolásban

A ábrából jól látszik, hogy a mintanagyságoknak megfelelően a minták három egyértelműen elkülöníthető csoportba tartoznak. Ez alól három mesterségesen rétegzett minta képez kivételt, melyek a fogyasztás tényleges értékei alapján lettek rétegezve, a sokaság decilisei szerint. Ezeknél a mintáknál a Deff mutató értéke rendkívül alacsony 0-hoz közeli értéket mutat. A többi minta Deff-jétől szintén alacsonyabb értéket mutatnak a fogyasztással jelentősen korreláló jövedelemváltozó alapján képzett mesterséges minták is.

A relatív standard hiba vizsgálatakor megállapítható, hogy a kisebb méretű mintáknak egyértelműen rosszabb eredményei vannak, mint a nagyobb méretűeknek. A mintavételi terv hatása pedig a minták összetettségével egyre javuló tendenciát mutat mindegyik mintaméret esetében.

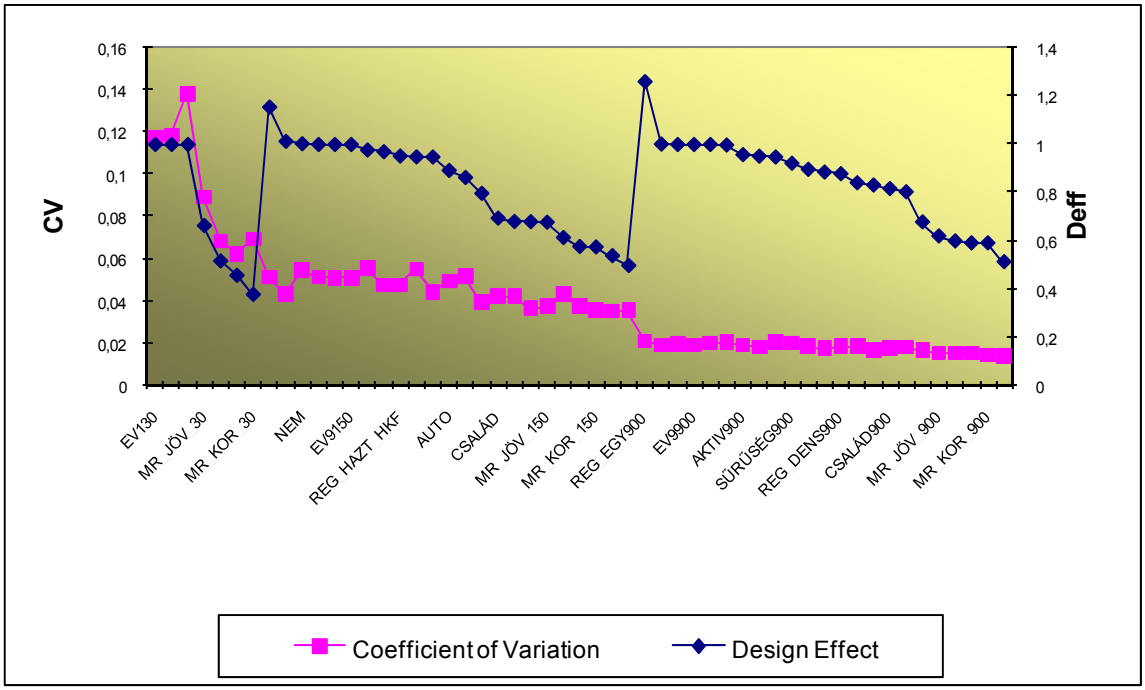
A több ismérv szerinti rétegzés tehát minden szempontból jobb eredményekkel kecsegtet, feltéve, hogy a rétegek nem túlzottan elaprózottak, mert akkor egy egyszerű véletlen kiválasztás eredménye felé halad, nagy mennyiségű, fölösleges többletmunka árán. Az ábrából

leolvasható, hogy azok a jó rétegzések, melyekben lehetőség szerint több olyan rétegképző ismérv szerepel, melyek a fogyasztással sztochasztikus viszonyban vannak, mint például az autók száma, a szobák száma, vagy a háztartás mérete. Az ábrák készítésének alapját képező számítások a 3. számú mellékletben találhatók.



6. ábra: A háztartások átlagos mérete (fő) becslésének jellemzői az effektív mintanagyság szerinti rangsorolásban

A teljes fogyasztás becslésénél tett megállapítások itt is helytállónak bizonyulnak. A legsikeresebb becslést a háztartás típusa szerinti rétegzés biztosította mind a 150, mind pedig a 900 elemű minták esetében. A grafikon végén található minták esetében azért kaptam 0 értékeket, mert kísérleti jelleggel olyan rétegképző ismérvet választottam, amely determinisztikus kapcsolatban van a háztartás méretével, így a vizsgált jellemzők értékelhetetlenné váltak. Ennek a problémának a részletes tárgyalására a következő alfejezetben térek ki. Látható még az ábrán a Deff mutató jelentős csökkenése néhány minta esetében. Ezek a minták a háztartás méretével jelentősen korreláló háztartástípus alapján lettek rétegezve, ami magyarázza a kiemelkedően jónak mondható mintavételi terv hatását.



7. ábra: A háztartások átlagos egy főre jutó fogyasztásának becslési jellemzői az effektív mintanagyság szerinti rangsorolásban

Az egy főre eső fogyasztás becslésénél a több ismérv szerint rétegzett minták mellett a mesterséges rétegzések is jól minősíthető becslési eredményeket adtak.

Mindhárom fenti ábrából látszik, hogy a régiónként egyenletes rétegzéssel tervezett minta eredményei az összes alkalmazott mintaméret esetében a legrosszabbak. A fenti jellemzők alapján rangsorolni lehet tehát az egyes mintavételi terveket, mely rangsorolás minden vizsgált változóra releváns eredményeket mutat. Ezt igazolja a felállított rangsorokra számított rangkorrelációs együtthatók értéke és szignifikanciája is.

2. táblázat: Rangkorrelációs együtthatók mátrixa

			R_fogyperfő	***R_fogy***	***R_házt***
Spearman's rho	R_fogyperfő	Correlation Coefficient	1,000	,896**	,863**
		Sig. (2-tailed)	–	,000	,000
		N	53	53	53
	R_fogy	Correlation Coefficient	,896**	1,000	,815**
		Sig. (2-tailed)	,000	–	,000
		N	53	53	53
	R_házt	Correlation Coefficient	,863**	,815**	1,000
		Sig. (2-tailed)	,000	,000	–
		N	53	53	53

**. Correlation is significant at the 0.01 level (2-tailed).

A korrelációs mátrixban minden együttható szignifikáns értéket mutat a korrelációs együttható abszolút értékének felső korlátját jelentő 1-hez igen közeli eredményekkel.

Az a feltételezés tehát, miszerint relevánsabb eredményeket biztosítanak azok a részletesebb információk alapján képzett minták, melyeknél a rétegzéshez több, a vizsgált változóval sztochasztikus viszonyban álló változó kerül bevonásra igaznak bizonyult, mivel mindhárom vizsgált változó esetében kiemelkedő eredményeket hoztak az összetett minták.

(Hangsúlyozom, hogy összetett mintán jelen esetben bonyolult felépítésű, de nem többlépcsős kiválasztást, hanem többszörös rétegzést értek).

1. tézis

A szakirodalomból ismeretes, hogy a mintavételen alapuló kutatásokból származó megalapozott eredmények, valamint relatíve alacsony hibák előállítása és publikálása csak megfelelő adatbázisok alapján garantálható. Kutatásaim során megállapítottam, hogy Magyarországon a lakossági bázisú felmérések mintavételen alapuló kutatási eredményeinek javításához szükséges adatok (nem minden esetben hozzáférhetően) diszpergálva megtalálhatók a különböző hivatalok és szervezetek adatgyűjtéseinek köszönhetően. A kutatási eredmények hozadékának javítása érdekében azonban ezek integrációjára van szükség.

Egy nagy mennyiségű, széles spektrumú információkat és megbízható adatokat tartalmazó információs bázis fenntartása egyáltalán nem utópisztikus elvárás. Különösen nem a mai világunkban, amikor a különböző közösségi oldalak, hűségkártyák adta lehetőségek révén az életünk egy részét a nyilvánosság előtt éljük, nem beszélve az adatvédelmi fegyelem jelentős lazulásáról. A kutatónak olykor az az érzése támad, hogy a hivatalos statisztika világa lemarad az informatika és információáramlás korszakában. Figyelemre és dicséretre méltó, hogy megpróbáljuk matematikai módszerekkel, közelítő eljárásokkal pótolni azokat az ismereteket, melyeket egy komplex adatbázis tartalmazhatna, illetve tartalmaz is, csak a hivatalos statisztikának ahhoz nincs megfelelő intézményesített hozzáférése.

A mintavételen alapuló kutatásokban nem szokatlan jelenség a kiegészítő információk használata, melyeket elsősorban a mintavételi hiba meghatározására, a válaszadási arány növelésére, a torzítások feltárására alkalmaznak több-kevesebb sikerrel. Lásd Estevao –

Särndal (2002.), Roy – Safiquzzaman (2006.) kétfázisú minták alkalmazására vonatkozó tanulmányaiban.

A nemválaszolás okozta torzítás kezelésében elengedhetetlen szerepet tulajdonít Särndal és Lundström (2008.) a kiegészítő információknak, kiemelve, hogy egyáltalán nem mindegy, milyen minőségű kiegészítő információkat alkalmazunk.

Valamint a részosztályok, csoportok becslésére alkalmazott módszerekben is nagy segítséget nyújtottak a kiegészítő információk, Estevao – Särndal (2004.) alapján.

A komplex adatbázis létrehozásának számos hátránya vonultatható fel, de azok korántsem akkorák, mint a szerteágazó adatszolgáltatási kötelezettségek révén létrehozott adatbázisoké. Értem ezalatt a lakcímnyilvántartó rendszer hiányosságait, az egészségbiztosítási rendszerben kallódó egyedek problémáit, és még sorolhatnám.

Problémát okozhat a különböző nyilvántartások informatikai háttérrendszerének eltérése. Azonban az informatikusok számára nem jelent megoldhatatlan feladatot a rendszerek kompatibilitását lehetővé tevő interface-ek kialakítása. A fenti tanulmányok kiemelik, hogy a Skandináv országokban működő átfogó információkat tartalmazó nyilvántartások jelentős előnyt jelentenek a statisztikai hivatalok és a statisztikai kutatások számára.

A KSH anonimizálási gyakorlatának alkalmazásával az összegzett információkkal való visszaélés elkerülhető, ezáltal az információbázis működtetése nem ütközne adatvédelmi előírásokba. Meggyőződésem, hogy számos gazdasági kérdésben megoldást jelentene, de legalábbis közelebb vinne egy valós kép kialakításához, akár különböző gazdasági teljesítmények mérésében – mikro szinten –, akár a fekete vagy szürke gazdaság hozzávetőleges méretének meghatározásában.

Természetesen egy komplex információbázis működtetéséhez szükségeltetik némi személyzetfejlesztés a hivatalokban, de ennek mértéke, valamint költségtöbblete megtérülne a információk és azok összefüggéseinek pozitív hozadékaiban.

4.3.2. A MINTAVÉTELI TERV HATÁSÁT MÉRŐ DEFF KRITIKÁJÁNAK CÁFOLATA

A mintavételi elmélettel kapcsolatos szakirodalom a Deff muatató értelmezésének tekintetében elég szűkszavúan fogalmaz. Nyilvánvalóan indikálja a kevésbé hatékony mintavételi terveket, de csupán közvetetten esik szó az alacsonyabb varianciát biztosító mintatervek-

ről. Ezek alapján az feltételezhető, hogy a mintavételi tervnek a becslési eredményekre gyakorolt hatását megtestesítő Deff mutató az egyszerű véletlen mintához képest kevésbé hatékony mintavételi tervek hatását méri, azonban azonos méretű, azonos változóra vonatkozóan azonos becslőfüggvény esetében nem alkalmas egyértelműen a minták közül eldönteni melyik a hatásosabb.

Az előző fejezetben végzett rangsorolás összehasonlítási szempontjainak mintegy ellenőrzéseképpen megkíséreltem a mintákat különböző csoportokba rendezni klaszteranalízis segítségével. Ebben az esetben nem volt célom, hogy a következtetéseket általánosítsam, pusztán azt vizsgáltam, hogy a kapott eredmények alapján a csoportosítási, rangsorolási tényezők szerepe egymáshoz képest milyen.

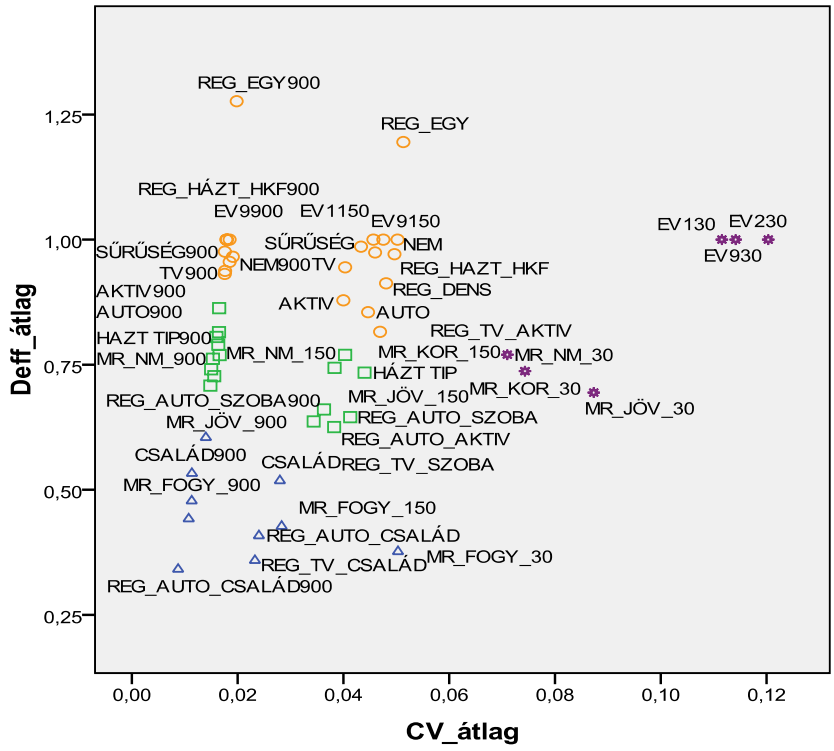
Annak érdekében, hogy az eredményeket megfelelően és világosan szemléltethessem, pusztán két változó, a Deff és a CV alapján végeztem az osztályozást. A mintaméretet a már említett kardinális szempontok miatt, miszerint nagyobb mintákból megbízhatóbb becslések adhatók, bátorkodtam kihagyni a vizsgálatból. Tettem ezt persze azért is, mert a mintaméretet mint változót – mivel 3 különböző méretű mintákat tartalmazó mintacsoport áll rendelkezésemre – ordinális skálázási szemléletben tudnám szerepeltetni, ez pedig jelen vizsgálat esetében rontaná az eredményeink értelmezhetőségét.

Az osztályozás során hierarchikus klaszterezési módszert alkalmaztam, annak itt most részletesen nem ismertetett előnyei miatt, azokat lásd: Besenyei et. al. (2010.). A csoportképzési módszerek közül a Ward módszert alkalmaztam a variancianövekmény minimalizálásának érdekében, négyzetes euklidészi távolságokra építve. Bár a vizsgált változók értékei nagyságrendileg kis mértékű eltérést mutatnak, alakulásuk teljesen más tendenciát és értelmezést hordoz. Ezért a pontosabb csoporteloszlás kirajzolásához standardizáltam a változók értékeit.

Az elemzést elvégeztem mindhárom eddig vizsgált paraméter dimenziójában (egy főre jutó fogyasztás, átlagos fogyasztás, átlagos háztartásméret). Tapasztalataim alapján az eredmények nem mutattak jelentős eltéréseket valamennyi mintacsoportban. Erre alapozva, valamint figyelembe véve Marton Ádám megállapítását, – miszerint: „Egy találomra kiragadott mutatóhoz tartozó Deft érték semmiképp sem alkalmas arra, hogy egy minta, kivált egy többcélú minta hatékonyságáról felvilágosítást adjon; ha egy mintát a Deft segítségével akarunk a vele azonos nagyságú egyszerű véletlen mintához hasonlítani, akkor lehetőség szerint több, célszerűen megválasztott mutató Deft-értékének az átlagával kell

dolgoznunk.” Marton (1991. p.34.) – a Deff és CV érékek átlagolásával folytattam a vizsgálatot.

Az elemzés összevonási táblázatának koefficienseiből készített vonaldiagram alapján 4 klaszter létrehozását láttam indokoltnak, melyet a következő ábrán szemléltetek.



8. ábra: A mintákon végzett klaszteranalízis eredményei az egy főre jutó fogyasztás becslésének eredményi alapján

Meg kell jegyeznem, hogy az ábra áttekinthetőségét rontja az első ránézésre zavaró zsúfoltság, de szükségesnek tartottam a minták nevének megjelenítését a könnyebb azonosíthatóság érdekében. Az ábrából jól követhető, hogy a kis méretű EV minták mindkét változó alapján rossz minősítésűek, hiszen rendkívül magas Deff értéket és hasonlóan magas CV értéket képviselnek. Ugyanebbe (az ábrán lilával jelölt), kevésbé hatékony csoportba tartoznak a szintén kis méretű mesterségesen rétegzett minták, ahol a minta mérete valószínűleg korlátozta a rétegzés kedvező hatásának kialakulását.

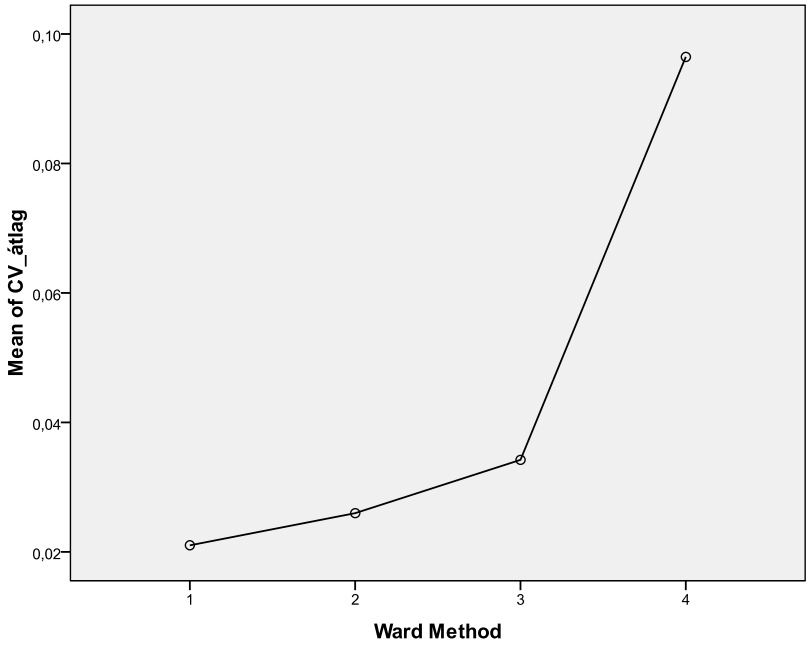
A legjobb csoport a kék színű, többszörösen rétegzett, illetve nagyobb méretű mesterséges mintákat tartalmazó csoport.

Érdemes megfigyelni az ábra felső részén megjelenő egyenletes rétegzést tartalmazó mintákat, melyek Deff értéke 1-nél magasabb, ezeket a klaszterezés sikeresen kiemelte a többi minta sokaságából.

Azt kell tehát állítanom, hogy a Deff mutató képes meghatározni, hogy melyik mintavételi terv tűnik rosszabbnak egy egyszerű véletlen mintavételi tervnél, de az EV mintáktól jobbnak tekintett rétegzett minták esetében bizonytalan eredményeket ad.

Kétségtelen, hogy a szakirodalmakban, valamint az alkalmazott statisztikákban a Deff mutató gyakorlati haszna, leginkább csoportos mintavételen alapuló becslések értékelésében jelentkezik. Az azonban elismerhető, hogy a mai számítógépes infrastruktúra mellett a csoportos mintáknak nagyon kevés előnyös tulajdonsága van. A szakirodalmak egy része szerint kifejezetten egy előnyös – és itt ki kell emelnem, hogy gyakorlati szempontból az országos lefedettségű mintavételeknél rendkívül hasznos – tulajdonsága ismert, az pedig az adatgyűjtés fázisában jelentkező költségmegtakarítás. Mivel jelen dolgozat a költségtényezőn kívüli mintavételi jellemzőkkel foglalkozik, így az említett előnyös jellemzőt nem vettem figyelembe.

A klaszteranalízis eredményeiből viszont kiderül, hogy a Deff mutató jelentős szerepet játszik a minták hatékonyság szerinti csoportokba való rendezésében. Annak ellenére, hogy csak néhány minta esetében haladja meg az egységnyi értéket, amikor is értelmezése közismert és felhasználása determinált. A 8. ábra alapján viszont azt kell állítanom, hogy a függőleges tengely mentén sokkal határozottabban különülnek el a mintacsoportok, mint a vízszintes tengely mentén, vagyis a CV alapján. Megvizsgáltam, hogy a két jellemző alapján képzett csoportokban hogyan alakulnak a jellemzők átlagai.

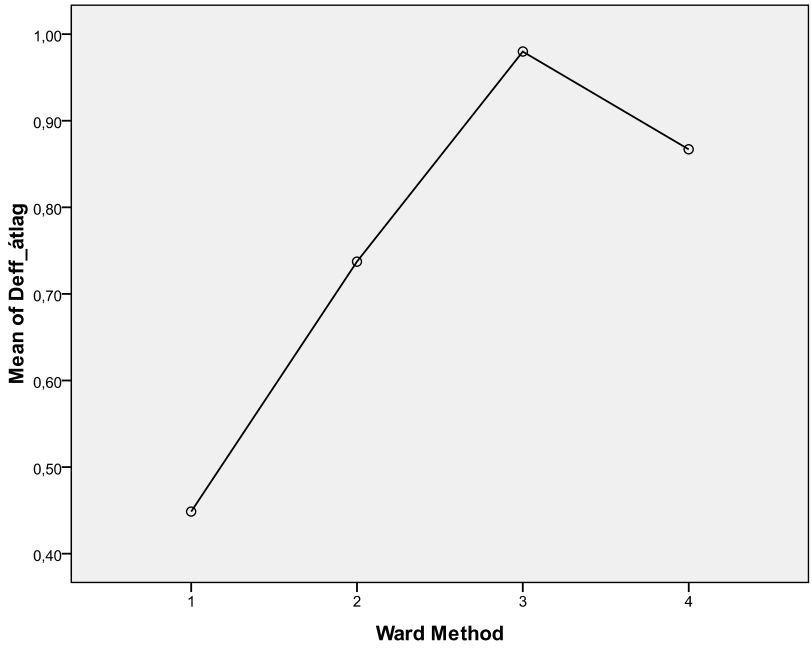


9.a. ábra: A CV átlagának alakulása az egyes mintacsoportokban

A 9.a. ábra alapján megállapítható, hogy a relatív standard hiba átlagai közel exponenciális képet mutatnak, vagyis a negyedik csoportba tartozó minták rendkívül magas relatív hibával becsülik a paramétert. Az első 3 csoport esetében viszont nincsenek szignifikáns különbségek, vagyis ezeknél a mintáknál a relatív hiba önmagában nem biztosít megfelelő támpontot az érékeléshez.

A 9.b. ábra szerint, ami a Deff mutató csoport átlagait jeleníti meg, az egyes csoportok között jelentős különbségek tapasztalhatók. Az egyes számú klaszterbe került minták esetében az átlagos Deff 0,5 alatti értékkel szerepel, ami a mutató klasszikus értelmezése alapján kétszer olyan jónak számít, mint egy azonos méretű egyszerű véletlen minta.

A 3. és 4. csoport átlagai alapján nem vonhatunk le egyértelmű következtetéseket, mivel az ábra szerint a legkevésbé hatékonynak tartott negyedik mintacsoport átlagos Deff értéke alacsonyabb, mint a hármas számú csoport átlaga. Erre a minimális zavaró tényezőre azonban a klaszteranalízis szigorú feltételei között kereshetjük a választ. Mindkét említett klaszterben vannak ugyanis olyan – akár outliernek is tekinthető – értékek, melyek a klaszter centroidokat eltolhatják a függőleges tengely mentén. Ilyenek a kis méretű EV minták, melyek becslési eredménye korlátozottan megbízható, illetve a már említett egyenletes rétegzéssel megvalósított minták, melyek tartalmára későbbiekben ki szeretnék térni.

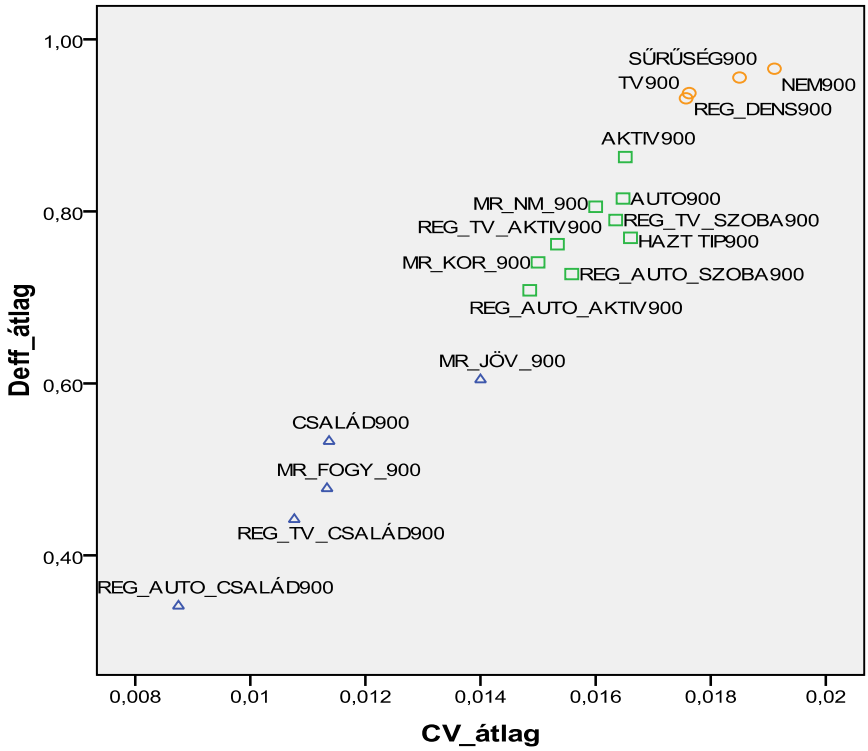


9.b. ábra: A Deff átlagának alakulása az egyes mintacsoportokban

2. tézis

A szakirodalom alapján a matematikai, statisztikai módszerek demonstrálják a Deff mutató azon tulajdonságát, hogy képes megmutatni, mennyivel rosszabb, vagy jobb az adott mintavételi terv, egy ugyanolyan méretű egyszerű véletlen mintánál. Empirikus kutatásaim során ezen túlmenően azt is feltártam, hogy a Deff mutató más mutatókkal együtt képes az egyszerű véletlen mintánál jobbnak bizonyuló mintavételi tervek esetében hatékonysági rangsor elkészítésére.

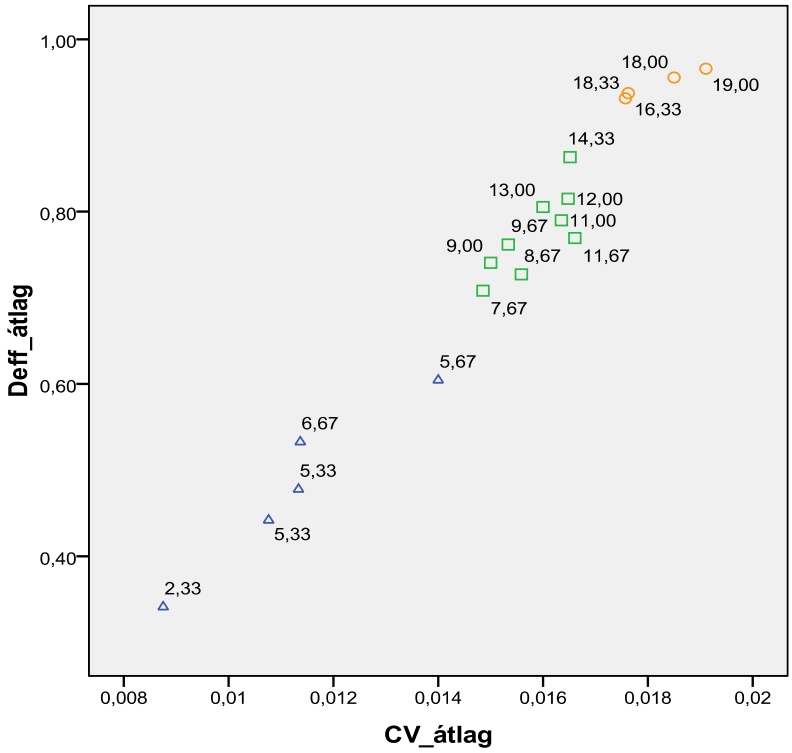
Az előző alfejezetben előtérbe helyeztem azt a feltételezést, miszerint a nagyobb mintákból jobb becslési eredmények nyerhetők. Ezt továbbra is fenntartom, de megvizsgálom, milyen következtetések vonhatók le kizárólag a nagyobb méretű mintákra koncentrálva. Ennek eredményeképpen csak a 900 elemű mintákat jelenítettem meg az átlagos Deff és az átlagos CV dimenzióiban. Fenntartva azt, hogy a klaszterelemzés sikeres volt, vagyis a minták különböző homogén csoportokba sorolhatók (a Deff és CV mutatók segítségével mért) hatékonyságuk alapján. Vagyis, kiküszöbölve a szélsőségesen kicsi mintaméret, illetve kevésbé hasznos rétegzési eljárás eredményezte minták hatásait, a klaszterek körvonalai is másképp rajzolódnak ki.



10.a. ábra: A 900 elemű minták klaszterei

A 10.a. ábrán látható, hogy amennyiben megszűnik a különböző méretű minták okozta szélsőséges ingadozás, a minták a két dimenzió között húzódó átló mentén helyezkednek el. Ami azt jelenti, hogy mindkét dimenzió arányosan jelentős szerepet játszik a klaszterek kialakításában. A Deff és a CV azért is tűnik jó párosításnak, mert a relatív hiba a standard hibának és a minta elemszámának a sajátos fordított négyzetes viszonya okán összefüggésben vannak egymással, a Deff viszont – éppen ellenkezőleg – a minta méretétől független eredményeket nyújt.

A fenti klasztereket a 10.b. ábrán is megjelenítettem, viszont ott nem a minták neve szerepel a feliratokon, hanem a minták különböző változók alapján képzett rangsorszámainak átlaga.



10.b. ábra: A 900 elemű minták klaszterei és rangsorszámai

Egyértelműen kivehető, hogy az origótól távolodva, ahogy mindkét dimenzió értéke romlik, a rangsorszámok értéke is egyre nagyobb lesz. A Deff mutató, a CV relatív hibával együtt alkalmas arra, hogy minősítse az azonos méretű, de egyszerű véletlen mintavételnél hatékonyabb mintavételi terveket. Az állítás igazolását elvégeztem a vizsgált változók mindegyikére külön-külön is, ezeket az eredményeket a 4. számú melléklet tartalmazza.

A változónként végzett elemzések arra is lehetőséget adtak, hogy megvizsgáljam, hogy a mintavételi tervben szereplő rétegképző ismérvnek vagy ismérveknek a vizsgált változóhoz fűződő sztochasztikus viszonya mutat-e valamilyen összefüggést a Deff és a CV által állított rangsorral. Megfogalmazható, hogy a rétegképző ismérv vagy ismérvek korrelációs együtthatója – többszörös rétegzés esetén többszörös korrelációs együtthatója – determinisztikus viszonyban van a Deff és CV által kialakított rangsorral. Elmondható, hogy minden esetben a kevésbé hatékony minták klaszterébe kerültek a 0,4-nél kisebb korrelációs együtthatójú rétegképző ismérvvel rendelkező minták.

5. A MEGHIÚSULÁSOK HATÁSA ÉS KEZELÉSE

A válaszadás hiányossága talán az egyik legnagyobb probléma, ami a felmérések készítésénél felmerül. Manapság nem ritkák az 50 %-on aluli válaszadási aránnyal rendelkező kérdőívek. Nyilvánvaló, hogy a szelektív válaszadás nemcsak a mintanagyságot csökkenti, hanem növeli a becslések varianciáját is, valamint a torzítás mértékét. A fejezet matematikai-statisztikai módszerek alkalmazását javasolja a mintavételi hiba és a nemválaszolás okozta torzítás csökkentésére. Különböző elemzési módszerek segítségével arra kerestem a választ, vajon a megtagadások pótlása mekkora hatást gyakorol a leíró modellek paramétereire, eredményeire.

5.1. A FOGYASZTÁSI KIADÁSOK BECSLÉSE

A vizsgálat alapvető célja, a költségvetési felvétel adataiból képzett minták alapján megbecsülni az egy háztartásra eső átlagos teljes fogyasztási kiadás értékét.

A becsléseket több különböző mintavételi eljárás és mintavételi terv alapján végeztem el, melyek részletes bemutatása az 4.2. alfejezetben található. A számos minta közül egy mesterséges információk alapján rétegzett minta bizonyult a leghatásosabbnak, melyben a teljes sokaság közel 10%-a került kiválasztásra a fogyasztási kiadások deciliseinek megfelelő arányos rétegekben.[8] A minták közötti választásnál a hatásosság és pontosság szempontjainak érvényesítése volt az elsődleges, ennek megfelelően a Deff és CV mutatók értékei alapján történt a szelekció. A mintavételi terv hatásossága 0,27, a relatív standard hiba pedig 0,5% értéket mutattak.

A továbbiakban a fenti minta adatai tekinthetők az elemzések alapjának. A mintából kísérleti jelleggel eltérő szisztémák alapján bizonyos elemeknél töröltem a fogyasztásra vonatkozó értékeket, így mesterségesen generálva a nemválaszolást.[9] Ennek a szimulációnak több hasznos tulajdonsága is volt. Először is az eredeti adatbázis ismeretében rendelkezé-

[8] A választás előzményét és bővebb indoklását lásd 4.3.2. fejezet, illetve 3. számú melléklet.
[9] Meg kell jegyezni, hogy jelen fejezet kizárólag a részleges nemválaszolás hatásaival és elemzésével foglalkozik, és nem tér ki arra a jelentős problémára, amikor teljes megtagadás/meghiúsulás erodálja az adatbázist!

semre állt a sokasági várható érték, melynek minél hatásosabb és pontosabb becslése az alapvető cél. Másodszor, jól láthatóvá váltak a teljes minta és a meghiúsulások által csonkított minta, illetve az ezekből származó becslési eredmények közötti eltérések, torzítások.

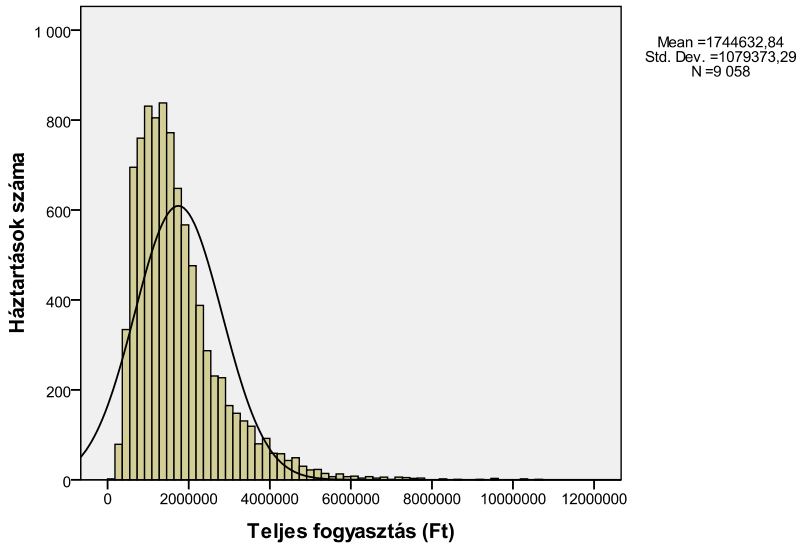
Az adathiány korrigálását azokkal a módszerekkel végeztem, melyek a piaci statisztikai szoftverek módszertárában fellelhetők, így olyan kutatók számára is elérhetők és használhatók, akik nem feltétlenül mintavétellel, illetve hibaszámítással foglalkozó szakemberek.

Ennek megfelelően a fent említett módszerek közül elsődlegesen az imputációt alkalmaztam. Az imputációs technikák közül – túllépve az egyszerű átlaggal való pótláson – a regressziós összefüggéseken alapuló, valamint a mintaegyedek hasonlóságára építő eljárásokat alkalmaztam. Természetesen az összehasonlíthatóság érdekében a későbbiekben bemutatom azokat az eredményeket is, melyek a nemválaszolás figyelmen kívül hagyásával nyerhetők.

5.1.1. NEMVÁLASZOLÁS GENERÁLÁSA

A nemválaszolások generálásakor figyelembe vettem azt a feltételezést, (melyet már több tapasztalati kutatás is igazolt pl.: Keszthelyiné (2006.), Havasi (1997.), Havasi-Schnell (1996.), miszerint a jövedelemmel, fogyasztással kapcsolatos kérdésekre a magasabb jövedelemmel rendelkezők válaszolnak inkább vonakodva, belőlük kerül ki a nemválaszolók jelentős hányada. Így a nemválaszolók figyelmen kívül hagyásával számított becslések jelentősen alulbecslik az átlagos fogyasztási, jövedelmi viszonyokat.

Ezt támasztja alá az említett jelenségek eloszlásában a 11. ábra alapján tapasztalható jelentős baloldali aszimmetria és koncentráció is.



11. ábra: A háztartások teljes fogyasztásának hisztogramja az alapsokaságban

A szimulációk során tehát egyelőre azokat az eseteket vizsgáltam, amikor a háztartások „gazdagabb" rétegeiből kerülnek ki a nemválaszolók. (Természetesen mellőzve azt a lehetőséget, hogy a nemválaszolás véletlenszerű, hiszen ekkor a becslés végeredményeire bizonyítottan nem lenne hatással.)

Kísérleteim során vizsgáltam a válaszadási arány többféle mértékét. Egyértelműen bizonyítást nyert, hogy a nemválaszolás mértékének növekedése rontja a becslési eredményeket, itt azonban nem kívánom összehasonlítani a különböző mértékű nemválaszolások esetén levonható következtetéseket. Csupán egyetlen esetet emelek ki – egy, a gyakorlatban manapság igen kedvezőnek számító 10% nemválaszolást tartalmazó mintát –, melyen bemutatom az imputálás „jótékony" hatását a különböző módszerek alapján.

5.2. A HIÁNYZÓ ADATOK KEZELÉSÉNEK EMPIRIKUS VIZSGÁLATA

A következő alfejezetekben tapasztalati úton vizsgáltam a hiányzó adatok kezelésére alkalmazott módszereket, a különböző elveken alapuló imputációs eljárásokat, valamint a kalibráció gyakorlati hasznosítását. Ezek becslési eredményekre gyakorolt hatását később összehasonlítottam minden bemutatott módszerre.

5.2.1. A MINTAELEMEK HASONLÓSÁGÁN ALAPULÓ ELJÁRÁS

A Háztartási költségvetési felvétel kiadási tételei esetében a hasonlósági elven történő (hot deck) imputálást alkalmazzák az adatok pótlására. A hasonlóságon alapuló eljárások akkor alkalmazhatók, ha a vizsgált mintaelemek (jelen esetben háztartások) hasonlóságot mutatnak különböző változók, leginkább demográfiai és a fogyasztással összefüggésbe hozható változók mentén. Ilyen esetben joggal feltételezhető, hogy a hasonlóságokra épülő klasszifikációs módszerek alkalmazása mindig segít a nemválaszoló egyedek okozta információ veszteség csökkentésében.

A vizsgálatot klaszter-analízisre építettem, melyben a legközelebbi szomszéd módszerét alkalmaztam a csoportképzésre. A módszer bővebb leírását lásd: Falus – Ollé (2000.), (2008.), Sajtos – Mitev (2007.), Ketskeméty – Izsó (2005.). Első feladat volt tehát olyan változókat találni, amelyek a hasonlóságot eredményezik, valamint sztochasztikus összefüggést mutatnak a háztartások fogyasztásával, és nem korrelálnak egymással. Minthogy a változók korrelálatlansága (megfelelően alacsony korrelációja) a klaszterelemzésnek is elengedhetetlen feltétele, így ez lett a változók kiszűrésének fő szempontja, melynek eredményeit a következő táblázat tartalmazza:

3. táblázat: A vizsgált változók korrelláció mátrixa

Variables		***Current activity status of the reference person***	***Number of cars***	***Number of televisions***	***Useful living area in m^2 (principal residence)***	***Household size***
Current activity status of the reference person	r	1	-,300**	-,190**	-,044	-,344**
	Sig.		,000	,000	,188	,000
	N	900	900	900	900	900
Number of cars	r	-,300**	1	,287**	,260**	,291**
	Sig.	,000		,000	,000	,000
	N	900	900	900	900	900
Number of televisions	r	-,190**	,287**	1	,289**	,348**
	Sig.	,000	,000		,000	,000
	N	900	900	900	900	900
Useful living area in m^2 (principal residence)	r	-,044	,260**	,289**	1	,279**
	Sig.	,188	,000	,000		,000
	N	900	900	900	900	900
Household size	r	-,344**	,291**	,348**	,279**	1
	Sig.	,000	,000	,000	,000	
	N	900	900	900	900	900

**. Correlation is significant at the 0.01 level (2-tailed).

Látható, hogy a kritikusnak tekintett r=0,3 értéket szinte egyik változó sem haladja meg, és a legtöbb eredmény szignifikánsan különbözik a nullától.

A klaszterelemzés további feltételeinek vizsgálatakor outliernek minősülő kiugró értékeket nem tapasztaltam az adatok között. Az elemzés két megoldásváltozatot eredményezett, melyek homogén csoportokat hoztak létre, egy három- és egy kétklaszteres változatot. Mivel célom a nemválaszolók elkülönítése a válaszolóktól, ezért a kétklaszteres megoldást tartottam elfogadhatónak. Az egyik klaszter az átlagos, illetve átlagon felüli életszínvonalat tükröző csoportnak bizonyult, míg a második csoport tagjai érezhetően alacsonyabb életkörülményeket tudhatnak magukénak.[10] Meggyőződésem, hogy a klaszterelemzés eredményei tovább finomíthatók, arra alkalmas, de az adatbázis korlátai miatt jelen dolgozat számára nem hozzáférhető információk bevezetésével, lásd például Bukodi – Altorjai – Tallér (2006.) a foglalkoztatási rétegek és anyagi életkörülmények összefüggésére vonatkozó megállapításait.

A klaszterezés menetét és mellékszámításait terjedelmi korlátok miatt itt most nem mutatom be (lásd 5. melléklet). Ehelyett az eredmények ellenőrzése céljából egy kontingencia táblázatban szerepeltetem a klaszterhez tartozás és a nemválaszolás összefüggéseit.

4. táblázat: Kontingencia tábla a klaszterhez tartozás és a nemválaszolás csoportosításához

Klasztertagság	***Mértékegységek***	***Válaszolók***	***Nemválaszolók***	***Összesen***
1. klaszter (jobb körülmények)	(fő)	531	84	615
	Megoszlás a klaszterek között (%)	65,6	93,3	68,3
2. klaszter (rosszabb körülmények)	(fő)	279	6	285
	Megoszlás a klaszterek között (%)	34,4	6,7	31,7
Összesen	(fő)	810	90	900
	Megoszlás a klaszterek között (%)	100,0	100,0	100,0

A fenti táblázatból jól látható, hogy sikeresen azonosítottam a nemválaszolók táborát, hiszen 93,3%-uk abba a klaszterbe sorolható, amelyik jobb életkörülményeket mutat.

[10] Ebben a megfogalmazásban életszínvonal alatt nem feltétlenül a korrekt, klasszikus közgazdasági, statisztikai megfogalmazást értem, hanem az adatbázisban rendelkezésre álló változók alapján az életkörülmények bizonyos fokú leírását lehetővé tevő jellemzőt.

3. tézis

Abban az esetben, ha a nemválaszolók a vizsgált anyagi jellegű változók tekintetében (fogyasztás, jövedelem, stb.) hasonlóságot mutatnak, akkor egyéb, a vizsgált változóhoz kapcsolódó jellemzők, valamint további demográfiai, társadalmi, gazdasági szempontok szerinti hasonlóságok is detektálhatók a mintaegyedekben, ami lehetővé teszi a klasszifikációs módszerek alkalmazását a nem mintavételi hibák csökkentésére.

Ezt a megállapítást erősíti, hogy a különböző társadalmi jelenségek is, mint például mobilitás, foglalkoztatás, eltérő fogyasztási mintázatokat eredményeznek. Grusky – Weeden (2001.).

A 4. táblázat eredményei alapján az imputálást is az 1. számú klaszter adataiból érdemes elvégezni. Az imputálás során több eljárást követtem. Fontosnak tartom ezek közül kiemelni a talán szélsőségesnek mondható következő eseteket, amikor a klaszter elemeiből véletlenszerűen választottam az imputálás alapjául szolgáló egyedeket, illetve egy másik eljárásban, a klaszter felső 10%-a képezte az imputálandó adatokat. A becslési eredményeket az 5.4. alfejezetben mutatom be.

5.2.2. MAHALANOBIS TÁVOLSÁG ALKALMAZÁSA A DONOR KIVÁLASZTÁSBAN

A hot deck imputáció lényege, megtalálni az adathiányt tartalmazó egyedhez leginkább hasonló olyan egyedet, amelyre a szükséges adat rendelkezésre áll és ezzel pótolni a hiányzót. A hasonlóság mértékének és szempontjainak meghatározására számos lehetőséget kínál a tudomány (Mahalanobis távolság, négyzetes euklidészi távolság, stb.) A következőkben egy klasszifikációs eljárás a diszkriminancia-analízis segítségével állítok elő olyan diszkriminancia függvényt, ami segít elkülöníteni a nemválaszolókat a válaszolóktól több predictor változó alkalmazása mellett. A módszer bővebb leírását lásd: Székelyi – Barna (2002.), Szűcs (2002.), Sajtos – Mitev (2007.), Ketskeméty – Izsó (2005.). A diszkrimináló hatás mérésére a Mahalanobis távolságot alkalmaztam, ezzel választva ki az egymáshoz leginkább hasonló háztartásokat. Az elemzés független változói a következők :

- családfő neme,
- családfő életkora,
- családfő családi állapota,

- családfő aktivitási státusza,
- személygépjárművek száma,
- televíziókészülékek száma,
- háztartás mérete,
- jövedelem kategória.

Ezen változók közül stepwise módszerrel választottam ki azokat, melyek szignifikáns kapcsolatot mutattak a diszkriminancia függvénnyel. Ezt a következő három változó teljesítette: jövedelem kategória, személygépjárművek száma, a családfő életkora. Bár az eredmények a Wilk's lambda statisztika alapján szignifikánsnak tekinthetők, a kanonikus korrelációs együttható értéke mégis 0,5 alatt maradt (0,447). A elemzés további végeredményei a 5. mellékletben találhatók.

5. táblázat: A klasszifikációs eredmények

			Válaszadó	***Nem válaszoló***	***Összesen***
Eredeti	N	Válaszadó	798	12	810
		Nem válaszoló	73	17	90
	%	Válaszadó	98,5	1,5	100,0
		Nem válaszoló	81,1	18,9	100,0
Validált	N	Válaszadó	797	13	810
		Nem válaszoló	73	17	90
	%	Válaszadó	98,4	1,6	100,0
		Nem válaszoló	81,1	18,9	100,0

A klasszifikációs eredmények táblázatából kiszámítható, hogy kb. 90%-ban jól azonosítja az egyedeket a diszkriminancia függvény, viszont az eredmények korlátozottan felhasználhatók, mivel a válaszadókat azonosítja nagyobb arányban helyesen, míg összesen 29 háztartást jelöl nemválaszolónak a tényleges 90 helyett. Mivel azonban a program elmenti a becsült függvényértékeket és Mahalanobis távolságokat, így lehetőség nyílik az imputációra. Azoknál az eseteknél, ahol adott nemválaszolóhoz több azonos távolságú egyed tartozik, ott ezen egyedek átlagos értékét imputáltam. Az imputált adatokkal történő becslés eredményeit mutatja az alábbi táblázat.

6. táblázat: Az imputált adatokkal történő becslés eredményei

	Átlag (Ft)	***Standard hiba (Ft)***	***Relatív standard hiba (%)***
TC a felső 10% imputálva mah diszk alapján	1620586	10633,907	0,7

Az adatok további eredményekkel való összehasonlítása az 5.4. fejezetben található.

5.3. REGRESSZIÓS ÖSSZEFÜGGÉSEKEN ALAPULÓ IMPUTÁCIÓ

A következőkben azt a kísérleti eljárást szemléltetem, amikor az SPSS 17.0 statisztikai szoftver segítségével komplex többváltozós imputációt hajtottam végre.

Az imputáció során fontos szempont volt, hogy csak a becsülni kívánt teljes fogyasztásváltozóban szerepelt adathiány, a többi változó tekintetében nem. A regressziós imputáció független változói a következők voltak:

- családfő neme,
- családfő életkora,
- családfő családi állapota,
- családfő aktivitási státusza,
- személygépjárművek száma,
- televíziókészülékek száma,
- háztartás mérete,
- jövedelem kategória.

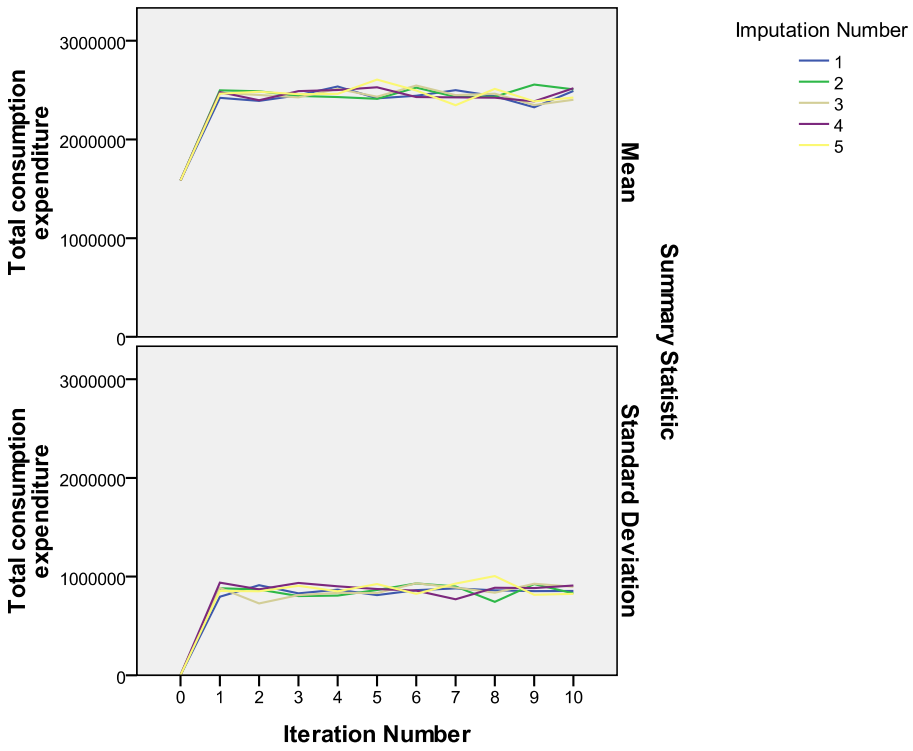
A regresszión alapuló imputáció egy iteratív eljárás, melynek eredményeként a 7. táblázatban szereplő megoldásverziókat kaptam. Közülük természetszerűleg az utolsó, 5. számú imputációra végeztem összehasonlító elemzéseket. Az imputálás utáni becslési eredményeket a következő alfejezetben mutatom be a többi becslési részeredménnyel együtt.

7. táblázat: Az összes fogyasztási kiadás hiányzó értékeinek regresszión alapuló imputációja

Data	***Imputation***	***N***	***Mean***	***Std. Deviation***	***Minimum***	***Maximum***
Original Data		810	1584876,04	928135,319	177292,00	7695496,00
Imputed Values	1	90	2492164,96	854677,547	681064,95	5384039,08
	2	90	2507749,93	830403,517	967829,96	4801665,53
	3	90	2402891,17	896478,581	769979,83	5070877,09
	4	90	2519055,67	909808,909	698372,48	5756997,38
	5	90	2425085,83	824802,684	1177892,99	5984624,70
Complete Data After Imputation	1	900	1675604,93	960041,621	177292,00	7695496,00
	2	900	1677163,43	959270,775	177292,00	7695496,00
	3	900	1666677,55	956581,950	177292,00	7695496,00
	4	900	1678294,00	967353,086	177292,00	7695496,00
	5	900	1668897,02	951919,703	177292,00	7695496,00

Az iterációk közötti választásban a program adta sorrend mellett az is nagy mértékben előremozdította a választást, hogy az 5. iterációban az imputált adatok minimális értéke jelentősen (némely iterációhoz képest 70%-kal) meghaladja a többi iteráció minimumát, miközben a maximális értékekben nincs jelentősebb különbség.

Az imputációk sikerességének ellenőrzésére egy konvergenciatesztelő ábrát készítettem, mely az átlag és a szórás konvergenciájának alakulását mutatja az egyes iterációk, imputációk során. A 12. ábrából jól látható, hogy semmilyen különleges mintázat, szignifikáns eltérés nem tapasztalható sem az átlag, sem pedig a szórás értékeiben az iterációk során a különböző imputációk tekintetében.



12. ábra: A teljes fogyasztás (Total consumption expenditure) átlagának és a szórásának konvergencia ábrája

5.4. AZ IMPUTÁLT ADATOKKAL KAPOTT BECSLÉSI EREDMÉNYEK ÖSSZEHASONLÍTÁSA

Az előzőekben bemutatott módszerek becslési eredményeinek összehasonlítása céljából a háztartások fogyasztási kiadásainak átlagbecslését végeztem a teljes mintán, az imputálás nélküli adatokon, valamint a különböző imputált adatbázisokon. Mivel a minta rétegzett eljárással lett kiválasztva, a becslési eredményeket (átlagot, standard hibát) egyszerű rétegzett becslés esetére számítottam.

Az imputált adatok megbízhatósága azonban sohasem olyan, mint az eredeti adatoké, és mint az köztudott, az egyszerű véletlen minta standard hibája mindig nagyobb, mint egy azonos nagyságú heterogén rétegzett mintáé.[11] Ezért, hogy nagyobb konfidencia intervallumot kapjak, megkíséreltem az imputált adatokból az egyszerű véletlen mintának megfelelő átlagbecslés és hibaszámítás elvégzését. Az 8. táblázat a becslési végeredmények összehasonlítását tartalmazza.

[11] Természetesen helyesen megválasztott rétegzési módot és rétegképző ismérvet feltételezve.

8. táblázat: A háztartások átlagos fogyasztási kiadásainak becslése

Módszer	***Becslés***	***Standard hiba***	***Coefficient of Variation***	***Relatív eltérés a sokasági paramétertől***	***Az eltérés mértéke (%)***	***Konf: (0:igen), (1:nem)***	***Relatív standard hiba***
Teljes minta	1728151,47	8721,276	0,005	99,055%	0,945%	0	0,4999%
Nemválaszolt 10% a felső 30%-ból	1584815,61	7964,054	0,005	90,839%	9,161%	1	0,4565%
Imputált 5. iteráció	1668815,81	12595,015	0,008	95,654%	4,346%	1	0,7219%
Klaszterből imputált R becsléssel	1665379,74	10121,469	0,006	95,457%	4,543%	1	0,5801%
Imputált EV becsléssel	1668897,02	31730,657	0,019	95,659%	4,341%	1	1,8188%
Klaszterből imputált EV becsléssel	1665488,15	31896,867	0,019	95,464%	4,536%	1	1,8283%
Klaszterből a felső 10% imputlásával EV becsléssel	1716799,67	33390,132	0,019	98,405%	1,595%	0	1,9139%
Klaszterből a felső 10% imputlásával R becsléssel	1716685,51	8618,436	0,005	98,398%	1,602%	1	0,4940%
Mahalanobis távolság alapján imputált	1620586,09	10633,907	0,007	92,890%	7,11%	1	0,6095%

Az 8. táblázat adatainak elemzéséhez tudnunk kell, hogy a sokasági paraméter, melynek becslésére vállalkoztam, jelen esetben ismert: 1.744.633,- Forint. Jól látható, hogy a teljes mintából megfelelően és viszonylag alacsony standard hibával becsülhető a sokasági paraméter. Abban az esetben viszont, amikor adathiány lépett fel, a becslés értéke jelentős mértékben több mint 9%-kal alulmúlta a sokasági értéket.

Az imputációk eredményeként kapott mintákból származó becslések javultak az adathiányos mintához képest, hiszen többségében a sokasági paramétertől kevesebb, mint 4,5%-kal kisebb értéket becsülnek.

A torzítás azonban továbbra is jelen van a becslésekben, hiszen a 95%-os megbízhatósági szint mellett számított konfidencia intervallumok nem fedik a sokasági paraméter értékét. Ezt mutatja a táblázat utolsó előtti oszlopa, melyben „1” érték szerepel abban az esetben, ha a sokasági paraméter kívül esik a konfidencia intervallumon, és „0” abban az esetben, ha az intervallum tartalmazza az országos átlagot. Ilyen értelemben tehát nem tekinthetők

sikeresnek az imputálási módszerek, hiszen jelentősen alulbecsülik a paramétert. Itt emlékeztetek arra, hogy mindössze 10%-os nemválaszolásról van szó, ami kedvezőnek számító körülmény.

A becslések konfidencia intervallumának nagyságának növelése céljából a rétegzett mintára vonatkozó hibaszámítás helyett az egyszerű véletlen (EV) minta hibáját alkalmazva a relatív standard hiba romlásán kívül más eredményt nem tudtam elérni.

Abban az esetben viszont, amikor a klaszterelemzésen alapuló imputáció alkalmazásakor önkényesen a feltérképezett klaszter legnagyobb értékeit imputáltam, jelentősen megközelítette a becslés a teljes mintán alapuló becslési eredményeket. Az egyszerű véletlen hibaszámítás alkalmazásával pedig a megbízhatóságon is sikerült „javítani".

4. tézis

Nemválaszolás esetén az egyszerű[12] imputáció nem képes kellő pontossággal visszaadni az eredeti, teljes minta tulajdonságait, így a megbízható becslésre igen kicsi az esély, különösen, ha a vizsgált sokaság eloszlása nem szimmetrikus. Megállapítható, hogy egy baloldali aszimmetriát mutató sokaság lineáris statisztikáit becsülve az egyszerű imputált becslések torzítottan alulbecslik a sokasági paramétert.

5.5. KALIBRÁCIÓ: KOMBINÁLT MÓDSZER A HIBÁK KEZELÉSÉRE

A kutatási eredmények potenciális torzításainak csökkentésére számos módszer áll rendelkezésre. Ezek egy része azonban korlátozottan alkalmazható, mert a hazánkban lefolytatott felmérések adatbázisai, az adminisztratív nyilvántartási rendszerek adatai nem összekapcsolhatók, megakadályozva ezzel számos kiegészítő információ releváns hasznosítását. Lundström és Särndal (1999.) szerint a kiegészítő információk alkalmazása jól hasznosítható a nemválaszolás okozta torzítás csökkentésében is, ennek igazolására egy logit és egy exponenciális növekedési modellt alkalmaztak. Eredményeik szerint viszont a torzítás csökkentésében rendkívül jelentős szerepe van a kiegészítő információkat tartalmazó változók megválasztásának.

[12] Egyszerű alatt itt azt értem, amikor egy imputációs módszert alkalmazunk egy változó mentén történő pótlásra nélkülözve a különböző módszerek kombinációját és egyéb összetett eljárásokat.

A kalibráció adott kiegészítő információkon alapuló kalibrációs egyenletekkel végzett súlyszámítás, annak érdekében, hogy a súlyokat felhasználva lehetőséget biztosítson a sokasági paraméterek lineárisan súlyozott becslésére. Célja, hogy közel torzításmentes becslést adjon mind a nemválaszolás, mind a mintavételi hiba kiküszöbölése mellett. A módszer meglehetősen bonyolult eljárást takar, melyet a hazánkban használatos statisztikai programok közül nagyon kevés támogat.

Mint köztudott, a statisztikai becsléseknek számos követelményt kell teljesíteniük. Valamennyi követelmény egyidejű teljesítése hívta életre a kalibráció alkalmazását. A kalibráció, vagy kalibrálás meghatározása lényegesen bonyolultabb annál, minthogy egy, vagy akár néhány definíció segítségével megoldható legyen. Särndal (2007.) az elmúlt évtizedekben megjelent, számos szerző által publikált tudományos munkára hivatkozva a kalibráció meghatározását a következőkben foglalja össze.

„Definíció: A kalibrálás egy véges sokaságból vett minta alapján történő becslés esetében a következőkből áll:

a) adott kiegészítő információk alapján, olyan mintasúlyok kialakítása, melyek különböző kalibrálási egyenleteket teljesítenek,
b) ezeknek a súlyoknak a használata annak érdekében, hogy kiszámítsuk a sokasági értékösszeg lineárisan súlyozott becslését,
c) a módszer célja, a sokasági értékösszeg közel torzításmentes becslése, feltételezve, hogy nincs válaszmegtagadás, és egyéb nem mintavételi hibák mértéke nulla.

A módszer legfőbb előnyeként Harms, Duchesne (2006.) kiemeli, hogy, a kalibráció eredményeit könnyű interpretálni, indokolni, valamint a súlyozások tervezése és a kalibrációs egyenletek egyszerűen áttekinthetők.

5.5.1. LINEÁRIS SÚLYOZÁSI MÓDSZER

A kalibráció új fogalom a mintavételes eljárás terminológiájában – mintegy 15-20 éve - de alapvetően nem a súlyozás technikájaként vált ismertté. Az elmúlt 15 évben kibővült az alkalmazási terület és a technika használata iránti hajlandóság. A kalibrációval rokon súlyozást régóta alkalmazzák a magán közvélemény kutató intézetek pl. kvótás vagy a jelen dolgozat tárgyát nem képező nem-valószínűségi mintavételes eljárások esetében. A

kalibráció ma már rendkívül elterjedt eljárás a mintavételes felmérések gyakorlatában, szerepe és jelentősége általánosan elfogadott. A módszer alkalmazása mellet szól, hogy a nemzeti statisztikai hivatalok munkájában a különböző súlyozási eljárások jelenleg is hatékony segítséget nyújtanak az adatok feldolgozásban és az eredmények publikálásában. A súlyozás alapgondolata az átlagszámítás ismeretét feltételezve minden kutató számára triviális dolog. A súlyozást a statisztikai hivataloknál, a különféle alapsokasági paraméterek: értékösszegek, átlagok, stb. becsléséhez használják. A súlyozás előnye, hogy könnyen értelmezhető eredmények produkálhatók mind az adatokat előállító, mind pedig a felhasználó kutatók számára.

A mintavétel területén az egyedek kiválasztási valószínűségének inverze alapján történő súlyozás gondolata Horwitz – Thompson (1952.) munkája nyomán széles körben elfogadottá vált. A kiválasztási valószínűség inverze alapján történő súlyozás eredményeként a megfigyelt egyed önmagát reprezentálja, valamint a többi meg nem figyelt egyedet.

A kalibrált súlyok a kiválasztási valószínűség inverze alapján képzett mintasúlyokból vezethetők le, azoknak lehetőleg csekély mértékű módosításával. Így a becsült értékösszegek a felhasználók számára is könnyen értelmezhetők, és megfelelően általánosíthatók.

5.5.2. A KIEGÉSZÍTŐ INFORMÁCIÓK SZISZTEMATIKUS FELHASZNÁLÁSA

A kalibrálás szisztematikus eljárást biztosít a kiegészítő információk számításba vételéhez. Ahogy arra Rueda – Martinez – Martinez – Arcos (2007.) rámutat, számos standard beállítás esetében a kalibrálás egyszerű és praktikus lehetőséget nyújt a kiegészítő információk becsléshez történő hatékony felhasználására. A módszer alkalmazása sikeresen megoldott a kiegészítő információk különböző szintjei mellett (Pl.: sokasági aggregát értékek, mintabeli értékek, megválaszolt értékek).

A kiegészítő információkat már jóval a kalibráció népszerűvé válása előtt is alkalmazták a becslési eljárások pontosságának javítása érdekében. Számos cikk erősítette ezt a felfogást, többé-kevésbé speciális esetekre vonatkozóan. Például a nemválaszolás jelenségének feltérképezéséhez György (2004.) a kérdezőbiztosok tulajdonságait is sikeresen alkalmazza kiegészítő információként.

A kiegészítő információk hasznosításában a kalibráció mellett az általános regressziós becslésnek (GREG) is jelentős szerepe van. A két módszer között gyakran vonnak párhuzamot, sőt bizonyos beállítások esetében megegyező eredményeket is produkálhatnak, mégpedig két fontos okból:

- a GREG becslés szintén képes szisztematikus megoldást adni a becslésekhez szükséges kiegészítő információk felhasználására;
- néhány (de nem az összes) GREG becslő függvény kalibrációs becslőfüggvény, ezért határozott részei a kalibrált lineáris súlyozás gyakorlatának.

A GREG becslőfüggvény koncepciója a kalibrációhoz képest kicsit régebbre nyúlik vissza, hiszen fokozatosan alakult ki a hetvenes évek közepén. A módszer alkalmazásának és becslésben betöltött szerepének részletesebb leírása megtalálható többek között Särndal – Lundström (2005.) és Fuller (2002.) munkáiban.

5.5.3. A KALIBRÁCIÓ ALAPJAI

A kalibráció alapgondolatát amellett a két alapfeltevés mellett mutatom be, hogy a mintaelemek egyfázisú véletlen mintavétel alapján lettek kiválasztva és teljes a válaszadás. A gyakorlatban a módszer feltételei nem ennyire egyszerűek és tökéletesek, de a műveletek lényegi megértése ezekkel az egyszerűsítésekkel megkönnyíthető.

Adott $U = \{1, 2, ..., ,..., N\}$ véges sokaságból véletlen módszerrel kiválasztásra kerül az $s = \{1, 2,..., ,..., n\}$ minta. A véletlen mintavételi terv alapján a „*k*” elem ismert kiválasztási valószínűsége, $\pi_k > 0$, és a vele összefüggő mintavételi terv súly $d_k = 1/\pi$. A keresett „*x*” változó x_k értéke rögzített minden $k \in s$ esetén.

A feladat megbecsülni a sokasági értékösszeget

$$X = \sum_{k=1}^{N} x_k$$

a kiegészítő információk használatával.

A keresett „*x*" változó lehet folyamatos, vagy − mint számos vizsgálatban − kategorikus. Például, ha „x" bináris: x_k=0, vagy x_k=1 értékkel attól függően, hogy a „k"-adik személy munkában áll vagy munkanélküli, akkor a paraméter megbecsülhető a munkanélküliek számával a sokaságban. *X* torzítatlan becslőfüggvénye:

$$\hat{X}_{HT} = \sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{d_k}\ \mathrm{x_k}\ ,$$

a Horwitz−Thompson becslőfüggvény szerint.

A kiegészítő vektor általános jelölése $\mathbf{y}_k$. A kalibrációs feltételek megadásakor létfontosságú pontosan rögzíteni a kiegészítő információ tartalmát. Így két esetet célszerű megkülönböztetni $\mathbf{y}_k$- val kapcsolatban:

a) $\mathbf{y}_k$ ismert minden $k \in U$ esetben (komplett kiegészítő információknál),

b) $\sum_{k=1}^{N} \mathrm{y_k}$ ismert (külső forrásból), és $\mathbf{y}_k$ ismert (megfigyelt) minden $k \in s$ esetben.

Komplett kiegészítő információk alkalmazása valósulhat meg az első esetben, amikor $\mathbf{y}_k$ adott a mintavétel keretei között minden $k \in U$ (és így minden $k \in s$) esetben is. Tipikusan ez eset fordul elő a személyek és háztartások vonatkozásában végzett vizsgálatoknál Skandináviában és más Észak-Európai országokban ahol jó minőségű adminisztrációs regiszterekkel rendelkeznek, ezáltal nagy számú potenciális kiegészítő változót biztosítanak.

Összefoglalva a kalibráció segítségével az alábbi feltételek teljesülésével nyílik lehetőség a sokasági értékösszeg becslésére:

1. ismertek a keresett x_k változó értékei, ha $k \in s$,
2. az ismert terv súlyok d_k=1/π_k ha $k \in U$, és
3. az ismertek $\mathbf{y}_k$ vektor értékei ha $k \in U$ (vagy a $\sum_{k=1}^{N} \mathrm{y_k}$ értékösszeg külső forrásból).

A fenti feltételek mellett kalibráljuk a d_k súlyokat w_k kalibrált súlyokká úgy, hogy eleget tegyenek a következő elvárásoknak Horváth – Mihályffy (2008.):

a) a kiegészítő információkat tartalmazó változóknak a kalibrált súlyokkal becsült értékösszege legyen egyenlő a megfelelő sokaságbeli értékösszeggel

$$\hat{Y}_{kal} = \sum_{k=1}^{N} \mathrm{y_k} \ ,$$

b) a kalibrált súlyok minél jobban közelítsék a mintavételi terv súlyokat.

Az utóbbi feltétel teljesülése egy megfelelő távolságfüggvény alkalmazásával oldható meg, vagyis minimalizálni kell az $f_k(w,d)$ távolságfüggvényt. Méghozzá úgy, hogy a következő feltételek teljesüljenek:

$$\begin{array}{lcl} Y_1 & = & y_{11}w_1 + y_{12}w_2 + \ldots + y_{1n}w_n \\ Y_2 & = & y_{21}w_1 + y_{22}w_2 + \ldots + y_{2n}w_n \\ \vdots & & \qquad\qquad + \ldots + \\ Y_p & = & y_{p1}w_1 + y_{p2}w_2 + \ldots + y_{pn}w_n \end{array}$$

Az így kapott w_k kalibrálási súlyokkal becsülhető a

$$\hat{X}_{kal} = \sum_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{w_k\ x_k}$$

értékösszeg ami az

$$X = \sum_{k=1}^{N} x_k$$

sokasági értékösszeg konzisztens becslésének tekinthető.

A kalibrálás alapvető módszerei olyan egyszerű mintát tételeznek fel, amelyben nincs nemválaszolás. Ezért ez pusztán csak elméleti alapokat biztosít a gyakorlati kutatások számára, hiszen teljes válaszadás csak nagyon ritkán, vagy sohasem fordul elő. A válaszhiány bár nem kívánatos, de teljesen természetes jelenség a mintavételes kutatásokban, ezért a kalibrációs elméletnek ezt tudnia kell kezelni. A fentebb megadott jelöléseket használva, nemválaszolást tartalmazó minták esetén a kalibráció módszere valamelyest módosul. Az „*s*” véletlen minta továbbra is az U= {1, 2, ..., k,…., N} sokaságból kerül kiválasztásra és a „*k*” elem kiválasztási valószínűség alapján számított mintasúlya továbbra is $d_k=1/\pi_k$. Válaszhiányos esetekben, jelölje „*r”* a válaszadást, ekkor az „*s*” mintában az x_k változó értéke csak $k \in r$ esetben figyelhető meg. Az ismeretlen „*k”* elem kiválasztási valószínűsége ekkor: $\Pr(k \in r \,|s)=\theta_k$ feltételes valószínűséggel módosul, vagyis $\pi'_k=\pi_k\theta_k$ lesz. A θ_k értéke ekkor nem ismert de a kalibráció tradicionális alkalmazásánál egy $\hat{\theta}_k$ heu-

risztikus becsléssel közelítik. Ezt felhasználva állítják elő a $d'_k = 1/\hat{\theta}_k \pi_k$ mintasúlyokat, amikre végül elvégezhető a kalibráció a már megadott módon. A módszer hátránya, hogy túlzottan magas nemválaszolás esetén már nem szolgáltat megfelelő eredményeket.

5.5.4. A KALIBRÁCIÓ GYAKORLATI ALKALMAZÁSA

A kalibráció gyakorlati bemutatásához továbbra is az előbbiekben használt adatállomány képezi a vizsgálatok bázisát. A kalibráció alapjául szolgáló mintasúlyok a mintavételi terv alapján adottak (illetve viszonylag egyszerűen számíthatók, jelen esetben pedig az SPSS program automatikusan generálja a súlyokat). A kalibrációhoz a következő kiegészítő információkat, valamint értékösszegeket alkalmaztam:

9. táblázat: A kalibráláshoz használt kiegészítő információkat tartalmazó változók paraméterei

Változók	***N***	***Értékösszeg***	***Átlag***	***Szórás***
Aktivitási státusz	3.837.087	8.708.654	2,27	1,50
Lakás alapterülete	3.837.087	301.059.105	78,46	32,77
Autók száma	3.837.087	2.039.584	,53	,61
Televíziók száma	3.837.087	5.535.524	1,44	,71
Háztartás mérete	3.837.087	9.932.097	2,59	1,38
Jövedelem	3.837.087	20.519.224	5,35	2,93
Kor	3.837.087	23.015.595	6,00	3,08

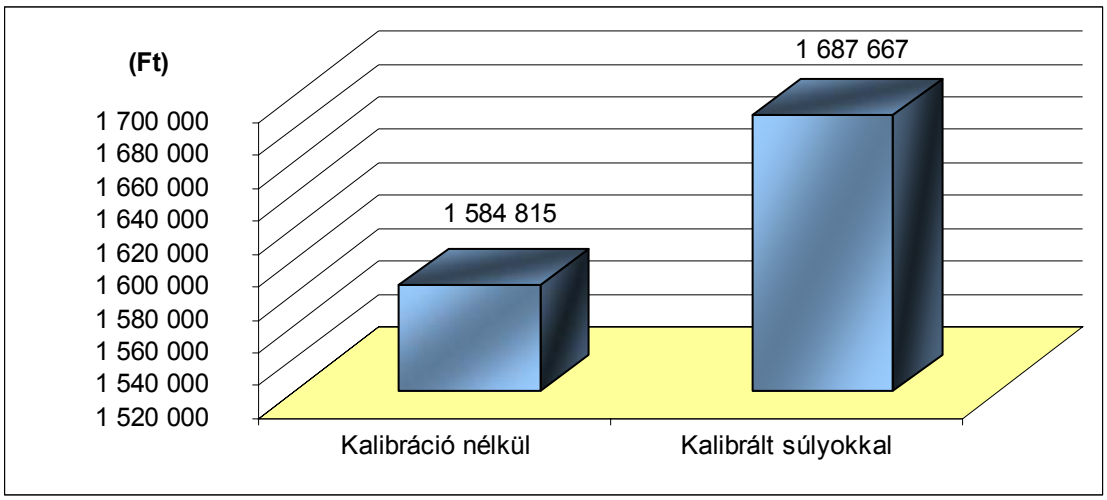
Tehát a kalibráció során a súlyokat úgy kell átalakítani, hogy a 9. táblázat értékösszegeit megfelelően közelítsék. Ehhez az iteratív művelethez természetesen megfelelő számítástechnikai háttér szükséges.[13] A kiegészítő információkat tartalmazó változók értékei és az eredeti súlyok megadása után a program futtatható. Az alapbeállítások természetesen módosíthatók. Jelen esetben a kiegészítő változók értékösszegeit ±1%-os pontossággal közelítettem, maximum 1000 iterációt végrehajtva. A program futása után a kalibrált új súlyok alkalmazásával a kiegészítő változók paramétereire az alábbi eredmények adódnak:

[13] Számomra Horváth Gergely a KSH Módszertani Főosztályának munkatársa biztosította a kalibrációt végző program SPSS-ben megírt változatát, melyért ezúton is szeretnék köszönetet mondani. Emellett köszönettel tartozom Mihályffy Lászlónak és Horváth Gergelynek a kalibráció gyakorlati alkalmazásához nyújtott hasznos tanácsaikért.

10. táblázat: A kalibrált súlyokkal számított paraméterek

Változók	*N*	*Értékösszeg*	*Átlag*	*Szórás*
Aktivitási státusz	3.859.519	8.699.149	2,25	1,49
Lakás alapterülete	3.859.519	301.708.003	78,17	33,04
Autók száma	3.859.519	2.041.701	0,53	0,60
Televíziók száma	3.859.519	5.499.196	1,42	0,69
Háztartás mérete	3.859.519	9.855.706	2,55	1,40
Jövedelem	3.859.519	20.367.979	5,28	2,85
Kor	3.859.519	23.048.050	5,97	3,13

A 9. és 10. táblázat összehasonlításából látható, hogy a kalibrált súlyokkal számított értékösszegek milyen mértékben közelítik az eredeti súlyokkal számított értékösszegeket. Amennyiben az eltérések nem megfelelően alacsonyak, úgy a módszer tovább folytatható újabb iterációkkal, esetleg további kiegészítő információk felhasználásával. A kapott eredmények most mind ±1% határon belül vannak, így a kalibráció sikeresnek tekinthető. A kalibrált súlyokkal becsülve a háztartások teljes fogyasztását a következő eredmények adódtak, melyek a torzítás jelentős csökkenését mutatják.



13. ábra: A kalibrálás hatása a teljes fogyasztás pontbecslésére

Ez az eredmény azonban nem tekinthető kevésbé torzítottnak az előző alfejezetekben bemutatott imputáláson alapuló pontbecslésekhez képest.

6. A NEMVÁLASZOLÁS OKOZTA TORZÍTÁS CSÖKKENTÉSE

A nemválaszolás okozta torzítás csökkentésére alkalmas eszközöket és módszereket felhasználva a háztartások teljes fogyasztására vonatkozó becslés eredményében bekövetkező torzítás mértékének redukálására törekedtem. Azzal az általános feltételezéssel élve, hogy az anyagi jellegű információkra vonakodva válaszolnak a mintaalanyok, különösen a vizsgált ismérv magasabb változatait reprezentáló egyedek. Ennek okán a fogyasztási kiadások becslések valószínűleg alulbecsüli a sokasági paramétert. Ezt a feltételezést támasztja alá az a tény is, miszerint a vizsgálat idejének hosszúsága determinálja, hogy a felmérésben részt vevők elfelejtenek rögzíteni több-kevesebb (feltehetően kisebb összegű) kiadást. Ezért a következőkben az alulbecslés mérséklésére törekedtem.

6.1. A MINTAVÉTELI TERV HATÁSA A NEMVÁLASZOLÁS OKOZTA TORZÍTÁSRA

Ahogy azt a 4.3. fejezetben bemutattam, a jó mintavételi terv nagymértékben hozzájárul a pontos, megbízható becslési eredmények publikálásához. A továbbiakban azt vizsgáltam, van-e pozitív szerepe a torzító hatás kiküszöbölésében olyan esetekben, amikor különböző mértékű válaszadási arányok valósulnak meg.

A mintavételi terv minősége hatással van a becslés pontosságára, hatásosságára, megbízhatóságára. Ezért feltételezhető, hogy az alaposan, előrelátóan megtervezett mintavétel csökkenti a becslési eredmények nemválaszolás okozta torzításának mértékét.

Ennek a hipotézisnek az ellenőrzésénél a fogyasztási kiadások átlagos mértékének (egy háztartás átlagos fogyasztási kiadásának) becslése biztosította az alapot. A paraméter becsléséhez legmegfelelőbb mintavételi terv kiválasztása a 4.3. fejezet eredményei alapján történt. A legjobb mintavételiterv-hatást és a legkisebb relatív hibát mutató minta az MR_FOGY_900 elnevezésű, ahol a háztartások kiválasztása rétegzett mintavétel alapján történt, a fogyasztási kiadások decilisei alapján, mesterségesen kialakított tízrétegű összetételben. A minta közel 10%-os kiválasztási arányt képviselve 900 háztartásból áll. (A minta jellemzőinek részletesebb leírását lásd: 4.2.2. fejezet.)

Alapvető célom volt, hogy a rétegképző ismérv minél szorosabb sztochasztikus kapcsolatot mutasson a becslés tárgyát képező ismérvvel. Esetemben ez az elvárás természetszerűen teljesült, mivel ($r=0,903^{**}$), a két változó szignifikáns kapcsolatát jelzi.

A gyakorlatban problémát jelenthet a fentihez hasonló jellemzőkkel rendelkező rétegképző változó megtalálása, különösen akkor, ha a vizsgált sokaságról semmilyen információ nem áll rendelkezésre. Ellenkező esetben külső információk sikeresen felhasználhatók a rétegzéshez –ezzel a megállapítással utalni kívánok az első alapozó tézis fontosságára –. A külső információk természetesen származhatnak egy teljes körű információkat nélkülöző, ugyancsak mintavételen alapuló előzetes felmérés eredményeiből is. Ebben az esetben viszont a következőkben alkalmazott módszerek nem képesek ignorálni az előzetes felmérésben rejlő torzító hatásokat.

A megfelelő mintavételi terv kiválasztása és a mintavétel elvégzése után a nemválaszolások generálása történt. Élve a fentebb, valamint az előző fejezetben megfogalmazott, gyakorlati tapasztalatokon nyugvó feltételezéssel, a nemválaszolókat a fogyasztási kiadás szerint csökkenő sorrendbe rendezett mintából választottam ki, kezdve egy igen kedvezőnek számító 10%-os nemválaszolási aránnyal. A következőkben nemválaszolás alatt részleges nemválaszolást fogok érteni, vagyis azt, amikor a megkérdezett háztartások csak az általam vizsgált változó tekintetében nem válaszoltak a többi feltett kérdésre választ adtak. Ennek megfelelően jelen esetben a 10%-os nemválaszolási arány azt jelenti, hogy a legnagyobb fogyasztói tizedbe eső háztartások a fogyasztási kiadásukat firtató kérdésre nem válaszoltak, míg az összes többi kérdésre válaszoltak.

Ezt követően, lépésenként 5-5%-kal növeltem a nemválaszolók arányát (vagyis a teljes minta adatbázisából lépésenként újabb 5-5% esetében töröltem a fogyasztási kiadásra vonatkozó értékeket), egészen az 50%-os mértékű nemválaszolásig. Ezen túlmenő nemválaszolás mértékét nem vizsgáltam.

Úgy vélem, a korábban felsorolt, valamint az említett szakirodalmakban bemutatott válaszadási arány növelésére kidolgozott módszerek és ösztönzési eszközök hatékony alkalmazása (ha kell többszöri próbálkozásra) segítséget nyújt a válaszadási arány legalább 50%-os biztosításában.

Annak ellenőrzéséhez, hogy az alaposan, előrelátóan megtervezett mintavétel csökkenti-e a becslési eredmények nemválaszolás okozta torzításának mértékét további, hasonló mére-

tű mintákban is – a fentebb megfogalmazott elvek alapján – generáltam adathiányokat. Ezeket, mint kontrollmintákat kezeltem. Közülük két esetet mutatok be, egy egyszerű véletlen kiválasztást és egy többszörös rétegzést. A különböző nemválaszolási szinteken kapott becslési eredmények az egyes mintavételi tervek esetében a következő táblázatban olvashatók.

11. táblázat: A fogyasztási kiadás becslési részeredményei különböző válaszadási arányok mellett, eltérő mintavételek esetén

adatok: Ft-ban

Nemválaszolás mértéke	***MR_FOGY_900***		***EV9900***		***REG_AUTO_CSALÁD900***	
	Mean	Standard Error	Mean	Standard Error	Mean	Standard Error
Total consumption expenditure	1.728.151	8.721	1.764.318	36.872	1.729.946	22.933
TC a felső 10% nem válaszolt	1.475.398	3.530	1.488.033	22.894	1.485.821	15.772
TC a felső 15% nem válaszolt	1.389.274	2.747	1.406.879	20.660	1.407.971	14.677
TC a felső 20% nem válaszolt	1.316.725	2.807	1.337.702	19.178	1.337.700	14.022
TC a felső 25% nem válaszolt	1.253.037	2.627	1.274.724	18.004	1.275.130	13.294
TC a felső 30% nem válaszolt	1.193.976	2.758	1.215.273	16.955	1.216.864	12.758
TC a felső 35% nem válaszolt	1.138.382	2.735	1.159.687	16.094	1.162.519	12.364
TC a felső 40% nem válaszolt	1.084.923	2.933	1.108.284	15.495	1.111.130	11.899
TC a felső 45% nem válaszolt	1.032.088	3.007	1.057.295	14.915	1.059.337	11.491
TC a felső 50% nem válaszolt	979.840	3.266	1.007.642	14.446	1.007.776	11.425

A 11. táblázat adatai meglepő eredményeket tartalmaznak, mivel a különböző válaszadási szinteken, a mesterségesen rétegzett minta adta a legkisebb becsléseket, másként fogalmazva ez tartalmazza a legnagyobb torzítást. (Hiszen eleve alulbecsüljük a fogyasztási kiadásokat, mivel a sokasági érték jelen esetben 1.744.632,- Ft.) Tisztán látszik az az egyértelmű tény is, hogy a nemválaszolás mértékének növekedésével az alulbecslés egyre drasztikusabb méreteket ölt, 50% válaszadás mellett akár 46%-kal is alábecsülhetjük a sokasági paramétert. A számítási eredmények alapján a feltételezést el kell utasítanom.

5/a. tézis

Abban az esetben, ha a nemválaszolás vélt vagy valós oka sztochasztikus összefüggést mutat a rétegképző ismérvvel, a rétegzés növeli a nemválaszolás okozta torzítás mértékét.

Az általam várttól eltérő eredmények magyarázatát a szórásnégyzet felbontása során találtam meg. Az MR_FOGY_900 mintában a rétegképző ismérv változatait a fogyasztás deciliseihez való tartozás határozza meg. Ez az ismérv, mint említettem, sztochasztikusan kapcsolódik a tényleges fogyasztáshoz. A nemválaszolások mesterséges generálása minden vizsgált mintavételi terv esetében a fogyasztási decilisek mentén történt, így ez alapján kerestem az összefüggéseket.

A varianciaanalízis során megállapítottam, hogy a különböző mintavételi tervek esetében a legnagyobb különbségek a külső eltérés négyzetösszegekben mutatkoztak. A belső eltérés négyzetösszegek viszonylagos stabilitása azzal magyarázható, hogy a fogyasztási decilisekben a háztartások fogyasztási kiadásai eleve jelentős különbségeket mutatnak. Mivel az MR_FOGY_900 minta esetében a rétegképző és a nemválaszolást előidéző ismérv azonos, így a csoportok közötti eltérés négyzetösszeg (és annak aránya a teljes eltérés négyzetösszeghez viszonyítva) itt a legmagasabb – lásd 12. táblázat –. Ez eredményezi az erőteljesebb alulbecslést, hiszen előfordul, hogy teljes csoportok tartoznak a nemválaszolók táborába. Abban az esetben, ha teljes csoportok adatai hiányoznak, a többi csoport nagyobb szóródása mérsékli a negatív torzító hatást. A kontrollminták egy részében nincs rétegzés, a másik részük pedig kisebb sztochasztikus összefüggést mutat a fogyasztással (egyúttal a nemválaszolással), mint az MR_FOGY_900 minta.

12. táblázat: Varianciaanalízis eredményeinek összehasonlítása

	Eltérés négyzetösszeg megoszlása (%)		
	MR_FOGY_900	***EV9900***	***REG_AUTO_CSALÁD900***
Külső	92,79	86,23	87,36
Belső	7,21	13,77	12,64
Teljes	100,00	100,00	100,00

A fentiek alapján megállapítható, hogy a mintavétel tervezésekor, a rétegképző ismérv megválasztása tekintetében figyelemmel kell lenni a nemválaszolás várható mértékére. A feladatot tovább nehezíti, hogy a nemválaszolás mértéke nem jelezhető előre. De még ha valamilyen külső információ alapján tervezhető is a nemválaszolás mértéke, a rétegek ki-

alakításakor további megfontolásokat kell szem előtt tartani, a megfelelő reprezentativitás megtartásához.

A legtöbb kutató a mintavétel megtervezésekor azt az optimális esetet veszi alapul, miszerint minden kiválasztott mintaegyed lelkes válaszadó is egyben. Ezt a feltételezést megtartva teljesen logikus az az elvárás, hogy a rétegképző ismérv a vizsgálat tárgyát képező ismérvvel szoros kapcsolatban legyen, hiszen így csökkenteni lehet a standard hibát, ezáltal a becslés pontossága javul. Viszont ha a vizsgált ismérv tekintetében nemválaszolások tapasztalhatók, akkor a sztochasztikus rétegzés hatása rontja a becslési eredményeket.

5/b. tézis

A kutatónak a mintavétel megtervezésekor figyelembe kell vennie a vizsgált ismérvvel kapcsolatos megtagadási várakozásokat, és ha azok potenciálisan teljes rétegeket érintenek, akkor érdemes lazítani a vizsgált ismérvvel sztochasztikus kapcsolatban levő rétegképző ismérvhez fűződő reprezentativitási követelményeken.

6.2. A NEMVÁLASZOLÁS TÉNYÉNEK BECSLÉSE

A nemválaszoló egyedek becslésekor fenntartom az előzőekben már megfogalmazott azon feltételezéseket, miszerint a nemválaszolás összefüggésben van a vizsgált jelenséggel, a fogyasztási kiadások pedig alulbecsültek. A becslési módszerek alkalmazását a 3. tézisre alapoztam, miszerint a nemválaszolók egyéb jellemzők alapján is mutatnak hasonlóságokat. Ezt a tulajdonságot kihasználva azonosítottam a nemválaszoló háztartásokat, majd figyelembe véve azok jellemzőit, a választ megtagadó háztartások nemválaszolási valószínűségeit térképeztem fel.

6.2.1. A NEMVÁLASZOLÁS BECSLÉSÉRE ALKALMAZHATÓ MODELLEK

A rendelkezésre álló módszerek közül olyat szerettem volna választani, amely a lehető legkevesebb külső információ bevonását igényli, hiszen a kutatók számára ezek igen szűkösen állnak rendelkezésre. Többféle módszer alkalmazására tett kísérletről számol be Foster (1996), melyekben a legsikeresebb eredményeket egyértelműen akkor érték el, ha a

jelenséghez fűződő olyan változókat alkalmaztak, melyek egy cenzusból vagy mikrocenzusból származnak.

Jelen dollgozatban azokat a lehetőségeket kívántam feltárni, melyek elsősorban a vállalkozások, illetve a vállalkozások által alkalmazott kutatók számára biztosítanak megfelelő alapot a helyes becslések készítéshez. Így a mintában fellelhető információkat felhasználva három klasszifikációs algoritmussal kísérleteztem:

- diszkriminancia-analízis,
- CHAID döntési fa,
- logisztikus regresszió.

Minden algoritmusnál először az optimális változó összetételt kerestem, az SPSS alapbeállításait használva, amelyek már önmagukban is alkalmasak modellek kiértékelésére.

Minden létrehozott modell kiértékelését ugyanazon mutatók mentén végeztem, így az eredmények könnyen összehasonlíthatóak. Az alkalmazott mutatók a következők:

- Pontosság: a helyes osztályozás arányát mutatja meg, vagyis, hogy az elemek hány % - t jelzi előre helyesen a modell.
- Elsőfajú hiba: valójában „True" értéket „False"-nak jelez előre a modell, azaz nemválaszoló megkérdezettet válaszolóként jelez. Ennek a hibának a mértékét a „True" értékek számához, és nem a teljes mintanagysághoz viszonyítottam, mert így jobban érzéklehetővé válik a hiba jellege, mértéke.
- Másodfajú hiba: a valójában „False" értéket „True"-nak jelez előre a modell, azaz válaszadó megkérdezettet nemválaszolóként azonosít. Ennek a hibának a mértékét pedig a „False" értékek számához viszonyítottam.

A modellek kiértékelésénél nemcsak a modell pontosságát, hanem a túltanulás lehetőségét is meg kell vizsgálni. Erre kiváló lehetőséget nyújtanak a keresztérvényesség-vizsgálatok. Túltanulásról akkor van szó, ha a modell túlzottan illeszkedik a tanulómintához. Ennek hátránya, hogy bár az eredeti mintára nagyon pontos eredményt lehet kapni, egy új mintánál a modell már sokkal pontatlanabb eredményt ad. Túltanulás esetében a tesztmintán a modell pontossága jóval alacsonyabb, mint az eredeti mintán.

A modellek lefuttatását és kiértékelését követően mindegyik algoritmusból kiválasztottam a legjobb modellt. Itt figyelembe vettem a pontosságukat, az első- és másodfajú hiba nagyságát, valamint a túltanulást is.

13. táblázat: A klasszifikációs módszerek eredményei

Algoritmus	*Eredeti minta pontossága*	*Elsőfajú hiba*	*Másodfajú hiba*	*Teszt minta pontossága*
Diszkriminancia analízis	90,6%	20,0%	19,3%	90,4%
CHAID döntési fa	90,2%	12,2%	23,7%	82,6%
Logisztikus regresszió	91,1%	22,6%	15,5%	90,7%

Az eredmények első megítélésre hasonlónak tűntek, hiszen a nemválaszolók becslésének pontossága közel azonos volt a három modellben. (Az elemzési eredmények az 6. mellékletben találhatók.)

Az algoritmusok outputjainak részletesebb vizsgálatakor azonban a döntési fa bizonyult kevésbé megbízhatónak, valamint a keresztérvényesség-vizsgálat eredményei is a CHAID módszer esetében voltak a legrosszabbak.

A diszkriminancia-analízis során a kovariancia mátrixok azonosságára vonatkozó hipotézist tesztelő Box's M teszt eredménye nem bizonyult szignifikánsnak. Így a logisztikus regresszió alkalmazása mellett döntöttem. Természetesen döntésemet befolyásolta az a tény is, hogy a logisztikus regresszióval a nemválaszoló egyedek azonosításán túl a nemválaszolás valószínűségét is meg lehet határozni.

6.3. A LOGISZTIKUS REGRESSZIÓ

A logisztikus függvény paramétereinek becslését és tesztelését az SPSS program Complex Samples parancsának Logistic Regression alparancsa segítségével végeztem, ezáltal a mintavételi terv hatását is figyelembe vettem az elemzés során.

6.3.1. A VÁLTOZÓK KÖRÉNEK LEHATÁROLÁSA

A megfelelő magyarázó változók kiválasztásakor a következő szempontokat érvényesítettem:

- a modellben szereplő változók számának korlátozása a létrehozandó keresztosztályok számának mérséklése érdekében,
- a modellben szereplő változók skálázásának vizsgálata a létrehozandó keresztosztályok számának mérséklése érdekében,

- a modellben a lehető legtöbb változó szerepeljen, melyek befolyással vannak a nemválaszolásra,
- a modellben szereplő változók szignifikánsan befolyásolják a nemválaszolást,
- a modellben szereplő magyarázó változók függetlenek legyenek egymástól,
- a modell jól illeszkedjen az adatokra.

A fenti feltételek között található néhány egymásnak teljesen ellentmondó, mint ahogy azt a regressziószámítás során is általános jelenség. Természetesen a kompromisszumos megoldásra törekedtem, és azt a változó összetételt preferáltam, ahol a legtöbb feltétel teljesült. A folytonos változók szerepeltetése érdekében újraskálázást végeztem, kategorizálva a változók értékeit.

Különböző változókombinációkkal dolgozva, a magyarázó változók optimális összetétele a következő volt:

- HD14_02: autók száma a háztartásban,
- HA09: lakóhely népsűrűsége,
- HC08: iskolai végzettség,
- Jöv: jövedelemkategória.

Várakozásaimmal ellentétben semmilyen szinten nem bizonyult szignifikánsnak a háztartások mérete, illetve a lakás mérete (alapterülete) sem. Holott ezek a változók a fogyasztással jelentős kapcsolatban vannak.

6.3.2. A LOGISZTIKUS REGRESSZIÓFÜGGVÉNY MEGHATÁROZÁSÁNAK MÓDSZERTANA

A logisztikus regressziószámítást a klasszifikációs módszerek között tartják számon. Ezáltal alkalmas valamely adatbázison az egyedek diszjunkt csoportokba sorolására a csoporthoz tartozás jellemzőjének ismerete nélkül. A besorolás tárgyát képező csoportok száma diszkrét. Esetemben két csoport létezik: egyik csoportba a válaszoló egyedek, míg a másikba a megtagadó egyedek kerülhetnek. A csoportba tartozás előrejelzését magyarázó változók és azok szintjeinek rögzített kombinációja alapján lehet elvégezni. Esetemben ez a kombináció (ún.: kovariáns) az előző pontban meghatározott.

Az elemzés során azonban nem csupán a csoportba történő besorolás végezhető el, hanem meghatározható annak a valószínűsége, hogy egy egyed adott csoportba esik, méghozzá a következők alapján.

„A logisztikus regresszió két egymást kölcsönösen kizáró kategória bekövetkezési esélyeinek az egymáshoz való arányát, vagyis az *odds* mértékét modellezi magyarázó változók értékeinek az ismeretében. Az adott kovariáns mellett kalkulálva az *odds* mértékét, azt a kategóriák bekövetkezési valószínűségévé konvertáljuk, majd e feltételes valószínűségek mérlegelésével a vizsgált egyedet a kategóriák valamelyikéhez rendeljük."[14]

Mivel a nemválaszolás detektálása az alapvető feladatom, ezért az eredményváltozóban az „1" érték jelzi a nemválaszolást, és a „0" érték jelzi a válaszadást. Vagyis az odds a nemválaszolás feltételes valószínűsége és annak 1-től való különbségének hányadosaként írható fel. Tehát x_i (i=1, 2, …,p) magyarázó változók adott kovariánsa mellett a nemválaszolásra vonatkozó *odds*:

$$odds_x = \frac{P_x}{1 - P_x}$$

„A logisztikus regresszió feltételezése szerint az odds logaritmusa – másképpen a siker valószínűségének logitja – a magyarázó változók lineáris függvénye." Hajdu (2003.) Ahonnan:

$$odds_x = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}$$

Mivel P_x bizonyíthatóan:

$$P_x = \frac{odds_x}{1 + odds_x}$$

Ezért:

$$P_x = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

[14] Az idézet, valamint a további képletek forrása: Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003. 291. o.

Vagyis a β paraméterek ismeretében a nemválaszolás valószínűsége előre jelezhető. Mivel az eredményváltozó nominális skálán mért bináris változó, ezért a β paramétereket célszerű maximum likelihood módszerrel becsülni. Vagyis a nemválaszolás valószínűségét előrejelző függvény paramétereinek becsült értékei adott x magyarázó változók mellett a Likelihood függvény maximumában találhatók, vagyis ahol:

$$L = \prod_{i=1}^{n} \frac{e^{y_i b x_i}}{1 + e^{y_i b x_i}} \rightarrow \max$$

6.3.3. PARAMÉTERBECSLÉS, MODELLTESZTELÉS

A nemválaszolás előrejelzését különböző nemválaszolási szinteken végeztem 10%-50% terjedelemben, 5%-onként növelve a nemválaszolás szintjét. Minden esetben meghatároztam a logisztikus regresszió által becsült paramétereket, valamint a feltételes valószínűségeket, melyeket külön változóként elmentettem az adatbázisban. A következőkben az első (10%-os nemválaszolási szinthez tartozó) előrejelző modell részleteit mutatom be a megfelelő SPSS outputok segítségével.

Regressziós modellek értékelésekor rendkívül nagy hangsúlyt fektettem a modell magyarázóerejének vizsgálatára. A lineáris összefüggésekhez képest az R^2 ebben az esetben nem értelmezhető megfelelően, ezért annak korrigált változatait (ún.: pszeudo R^2) számítja a program. Ezek eredményinek értelmezése a lineáris változathoz hasonlóan történik.

14. táblázat: A logisztikus modellek Pszeudo R^2 együtthatói

Nemválaszolás mértéke	***Cox and Snell***	***Nagelkerke***	***McFadden***
TC a felső 10% nem válaszolt	,214	,447	,370
TC a felső 15% nem válaszolt	,272	,477	,376
TC a felső 20% nem válaszolt	,346	,547	,425
TC a felső 25% nem válaszolt	,392	,581	,443
TC a felső 30% nem válaszolt	,428	,607	,457
TC a felső 35% nem válaszolt	,455	,626	,468
TC a felső 40% nem válaszolt	,458	,619	,454
TC a felső 45% nem válaszolt	,449	,601	,433
TC a felső 50% nem válaszolt	,472	,630	,461

A *Cox and Snell* R^2 a modell log likelihoodjának értékét egy alapmodell log likelihodd értékéhez viszonyítja. A mutató elméleti maximális értéke (ami egy tökéletes modellt feltételez) kisebb, mint egy.

A *Nagelkerke* R^2 az előző mutató skálázási problémáinak korrigálásával határozható meg, míg a *McFadden* R^2 a becslő modell log likelihoodját egy olyan alapmodelléhez viszonyítja, ahol csak a tengelymetszet került meghatározásra. A kezdeti modellben, ami 10%-os nemválaszolással számol, mindhárom mutató értéke alacsonyabb a vártnál, de eltérő kovariánsokat alkalmazva a mutatók értéke ehhez képest romlott. A nemválaszolás mértékének növekedése viszont egyre javította az illeszkedést. Ennek oka, hogy a nemválaszolás mértékének növekedésével, nőtt a nemválaszoló egyedek száma, ami az azonosítás lehetőségét könnyebbé tette a vizsgált magyarázó változók mentén.

A modell eredményeinek gyakorlati szempontból hasznos bemutatását tartalmazza a 15. táblázat.

15. táblázat: A klasszifikáció eredményei

Observed	***Predicted***		
	0	***1***	***Percent Correct***
0	7971,800	181,200	97,8%
1	623,444	281,556	31,1%
Overall Percent	94,9%	5,1%	91,1%

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

A táblázat diagonálisában a teljes sokaságban helyesen azonosított egyedek találhatók, míg a mellékátlóban a rosszul előrejelzett egyedek száma olvasható. Fontos információtartalma van a sarokszámnak, miszerint a modell a sokaság 91,1%-át helyesen azonosította válaszadás szempontjából.

16. táblázat: A modell és a változók szignifikanciájának vizsgálata

Source	***df1***	***df2***	***Wald F***	***Sig.***
(Corrected Model)	4,000	887,000	33,692	,000
(Intercept)	1,000	890,000	65,048	,000
HC08	1,000	890,000	1,448	,229
HD14_02	1,000	890,000	35,442	,000
jöv	1,000	890,000	44,290	,000
HA09	1,000	890,000	3,250	,072

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

A Wald statisztika azt a null hipotézist teszteli, miszerint az egyes változókhoz tartozó paraméterek nullával egyenlők, vagyis nincs hatásuk a klasszifikációra. Ezért a megfelelően alacsony szignifikancia szintű változók minősülnek jelentős hatást gyakorló változónak. A program a változónkénti vizsgálat mellett teszteli a teljes modell szignifikanciáját is, mely a 16. táblázat első sorában található.

Megállapítható, hogy a HC08: iskolai végzettség változó hatása nem minősül jelentősnek, ami egyértelműen annak tudható be, hogy az iskolai végzettség sem a fogyasztással sem a jövedelemmel nem mutat determinisztikus kapcsolatot. Emellett a népsűrűség szignifikancia szintje is meghaladja a társadalomtudományokban általánosan alkalmazott 5%-ot. Ennek ellenére szerepeltetem a modellben, mert hatását számos nemválaszolást vizsgáló elemzésben kimutatták. Varga (1999.), György (2004.), Johansson – Klevmarken (2008.)

A regresszió függvények esetében általános feltétel, hogy a független változókat ne lehessen felírni egymás kombinációjaként, vagyis, hogy egymástól is függetlenek legyenek. Ennek a feltételnek az ellenőrzésére a VIF mutatót alkalmaztam, melynek eredményei alapján megállapítható, hogy nincs zavaró mértékű multikollinearitás a változók között.

17. táblázat: A változók függetlenségének vizsgálata

Model	***Collinearity Statistics***	
	Tolerance	***VIF***
HC08	,760	1,315
HD14_02	,725	1,379
jöv	,659	1,518
HA09	,876	1,141

A logisztikus regressziófüggvény paramétereire vonatkozó információkat a 18. táblázat tartalmazza.

18. táblázat: A logisztikus regresszió paraméterbecslésének eredményei

Parameter	***B***	***Std. Error***	***B/Std. Error***	***Design Effect***	***Exp(B)***
(Intercept)	-7,816	,969	-8,06526	,990	,000
HC08	,207	,172	1,203288	1,007	1,230
HD14_02	1,195	,201	5,95332	,993	3,303
jöv	,640	,096	6,655089	,998	1,896
HA09	-,344	,191	-1,80266	1,006	,709

A nemválaszolás/válaszadás odds-arányt becslő logisztikus regresszió egyenlete:

$$\ln(odds)=-7{,}816+0{,}207x_1+1{,}195x_2+0{,}64x_3-0{,}344x_4$$

(ahol x_1, x_2, x_3, x_4 a magyarázó változókat jelöli a 18. táblázatban található sorrendjük alapján). A B/Std. Error arány nagyságából szintén lehet következtetni a változók szignifikáns voltára, hiszen a nagyobb értékek szignifikánsabb hatást mutatnak. A paraméterek pozitív értéke azt jelzi, hogy az adott változó növeli a nemválaszolás feltételes valószínűségét, a negatív paraméter pedig csökkenti. (Ez azonban természetesen függ a skálabeosztástól.)

A magyarázó változóknak az oddsra gyakorolt parciális hatását az Exp(B) értékek mutatják. Eszerint az iskolai végzettség egy szinttel való növekedése 23%-kal növeli a nemválaszolásnak a válaszadáshoz viszonyított esélyét, minden egyéb tényező változatlansága mellet. A háztartás tulajdonában levő autók számának növekedése ceteris paribus 230,3%-kal növeli a nemválaszolás odds arányát. A jövedelemkategória egységnyi növekedése – jelen adatbázisban adott jövedelmi tizedből eggyel magasabb tizedbe kerülve – 89,6%-kal növeli a nemválaszolásnak a válaszadáshoz viszonyított arányát, ha az egyéb feltételeket változatlannak tekintem. A lakóhely népsűrűségét megtestesítő változó B paraméterének előjele arra utal, hogy negatív hatással van a nemválaszolás odds arányára, azonban figyelembe véve, hogy az adatbázisban a „1" érték jelenti a sűrűn lakott és „3" érték a ritkán lakott településeket, kiderül, hogy a hatás nem negatív. Ezáltal azt mondhatom, hogy a sűrűbben lakott települések lakói 29,1%-kal nagyobb eséllyel lesznek nemválaszolók, mint a kisebb településeken élők.

Mindezek alapján például egy egyetemi végzettségű háztartásfő által vezetett, egy autóval rendelkező 8. jövedelmi tizedbe eső, sűrűn lakott településen élő háztartás esetében a becsült logit:

$$\ln(odds)=-7{,}816+0{,}207*3+1{,}195*1+0{,}64*8-0{,}344*1=-1{,}22566$$

Melyből a becsült feltételes valószínűség:

$$P_{3,1,8,1}=\frac{e^{-1,22566}}{1+e^{-1,22566}}=0{,}773$$

Vagyis a fenti tulajdonságokkal rendelkező háztartás 77,3% valószínűséggel nem fog választ adni a fogyasztási kiadását felmérő kérdésre.

6.4. A MINTAEGYEDEK ÁTSÚLYOZÁSA

Az előző fejezetben meghatározott logisztikus regresszió függvény paramétereinek ismeretében a becsült nemválaszolási valószínűségek meghatározása egyszerű feladat. Az SPSS programban opcionálisan beállítható, hogy külön változóként elmentse a becsült feltételes valószínűségeket. A program biztosítja, hogy bármelyik kimenet kiválasztható referenciaértékként, így az eredmények minden esetben meghatározhatók. Ezáltal akár a válaszadási valószínűségek is becsülhetők lennének.

A becsült feltételes valószínűségeket Varga (1999.) munkája alapján mintasúlyokká alakítottam annak érdekében, hogy a potenciálisan nemválaszoló háztartások nagyobb súlyt kapjanak a fogyasztási kiadások becslése során. Ugyanis hasonlóságot feltételezek a potenciálisan nemválaszoló és ténylegesen nemválaszoló háztartások között. Mivel erre építettem a regressziós modellt is. A súlyok meghatározása a következőképpen történik:

$$súly = \frac{1}{1 - P_x}$$

például a 6.3.3. alfejezetben meghatározott tulajdonságokkal rendelkező háztartás esetében a számított súly:

$$súly = \frac{1}{1 - 0{,}773} = 4{,}405$$

Vagyis egy felsőfokú végzettségű háztartásfő által vezetett, egy autóval rendelkező 8. jövedelmi tizedbe eső, sűrűn lakott településen élő háztartás több, mint négyszeres súllyal kerül figyelembe vételre a számítások során. Ezzel a súlyozási rendszerrel csökkenthető a fogyasztási kiadások alulbecslésének mértéke különböző nemválaszolási szinteken, ezt mutatja a következő táblázat.

19. táblázat: A fogyasztási kiadások átlagának becsült értéke (Ft)
különböző nemválaszolási szinteken súlyozott és súlyozatlan adatokkal számolva

Nemválaszolás mértéke	***Súlyozatlan átlagos fogyasztási kiadás***		***Súlyozott átlagos fogyasztási kiadás***		***Az átlagok relatív eltérése súlyozatlan=100%***
	Ft	***A várható érték %-ában***	***Ft***	***A várható érték %-ában***	
TC a felső 10% nem válaszolt	1.475.398	84,57%	1.554.136	89,08%	105,3%
TC a felső 15% nem válaszolt	1.389.274	79,63%	1.494.852	85,68%	107,6%
TC a felső 20% nem válaszolt	1.316.725	75,47%	1.428.680	81,89%	108,5%
TC a felső 25% nem válaszolt	1.253.037	71,82%	1.371.626	78,62%	109,5%
TC a felső 30% nem válaszolt	1.193.976	68,44%	1.316.734	75,47%	110,3%
TC a felső 35% nem válaszolt	1.138.382	65,25%	1.266.237	72,58%	111,2%
TC a felső 40% nem válaszolt	1.084.923	62,19%	1.214.319	69,60%	111,9%
TC a felső 45% nem válaszolt	1.032.088	59,16%	1.172.525	67,21%	113,6%
TC a felső 50% nem válaszolt	979.840	56,16%	1.105.332	63,36%	112,8%

A táblázat utolsó oszlopából látható, hogy a súlyozás hatására legalább 5%-kal sikerült javítani a nemválaszolás torzító hatásán. A modell és a súlyozás helyességét mutatja, hogy a nemválaszolási szintek növekedésével a súlyozási módszer torzítást csökkentő hatása egyre javul. Azonban meg kell jegyezni, hogy a gyakorlati kutatásoknak nem csupán a torzítás negatív hatásának enyhítése a célja, hanem a sokasági paraméter minél pontosabb és megbízhatóbb becslése. Ezt a célt azonban láthatóan csak részben sikerült teljesíteni. A sokasági érték ismeretében ugyanis látható, hogy a nemválaszolás mértékének szisztematikus növelésével a torzítás a súlyozás ellenére is drasztikus méreteket ölt.

6. tézis

A nemválaszolások valószínűségének becslésén alapuló átsúlyozás képes csökkenteni a nemválaszolás okozta torzítást, de a nemválaszolás szisztematikus növekedése esetében a torzítást csökkentő hatás lényegesen elmarad a tényleges torzítás mértékéhez képest.

6.5. A NEMVÁLASZOLÁSI TENDENCIA VIZSGÁLATA

A nemválaszolás okozta torzítás kiküszöbölésére tett lépések között fontos szerepe van a tendenciák azonosításának. Érdemes megvizsgálni a válaszadók és nemválaszolók vizsgált

ismérvbeli tendenciái közötti különbséget. Persze mindezek előtt fel kell térképezni, hogy léteznek-e egyáltalán valamiféle tendenciák.

A következőkben az MR_JÖV_900 minta adatai alapján tettem kísérletet a torzítás csökkentésére a nemválaszolási tendenciák feltárásának segítségével. A kiválasztott mintában a rétegzés a jövedelmi adatok mentén történt, ahol a rétegeket a háztartások jövedelmi tizedei képezik. Célom továbbra is a fogyasztási kiadás becslési torzításának csökkentése. A rétegképző ismérv sztochasztikus kapcsolatban van a fogyasztási adatokkal, ezt jelzi a lineáris korrelációs együttható értéke (r=0,719**). Így feltételezhető, hogy amennyiben a nemválaszolás oka a jövedelmi adatok eltitkolásában keresendő, úgy az összefüggésbe hozható a fogyasztási kiadások elfedésével is.

A kiegészítő információk alkalmazása segítséget nyújt a hibák mértékének csökkentésében. A kutatóknak azonban nem mindig van lehetősége külső információk beépítésére, ezért a belső (mintabeli) információkat kell a lehető legnagyobb mértékben kiaknázni. A mintaegyedek megfelelő részletességű csoportosításával a válaszadói csoportokban megfigyelhető tendencia kivetíthető a teljes mintára, ezáltal a nemválaszoló egyedekre. A tendenciák modellezésével a nemválaszolás torzító hatása pedig csökkenthető.

6.5.1. CSOPORTKÉPZÉS

Az eljárás sikere érdekében megfelelő csoportokat kell kialakítani a mintában a vizsgált ismérvvel sztochasztikus kapcsolatban álló és a nemválaszolást generáló ismérv(ek) alapján. Ilyen ismérv megtalálása nehézségeket okozhat, de az előző fejezetekben bemutatott módszerek segítenek a nemválaszolók azonosításában. Nem feltétlenül szükséges, hogy a csoportok száma a rétegképző ismérv változatainak számával egybe essen. A csoportok kialakítása – amellett, hogy olyan változó mentén történik, ami sztochasztikusan összefügg a potenciális, illetve megvalósult nemválaszolással – tetszőleges, lehetnek tizedek, századok, de természetesen maguk a rétegek is. A következő példában a csoportok kialakítása nem okoz gondot, hiszen a nemválaszolást egyértelműen a jövedelem függvényének tekintem.

A csoportok képzésének alapvető célja tehát a szélsőséges értékek hatásának csökkentése, valamint a nemválaszolók lehetőség szerint minél pontosabb azonosítása.

A csoportképzés során szem előtt kell tartani azokat az általános statisztikai ismereteket, miszerint túl kevés csoport nem segíti a hatékony elemzést, túl sok csoport létrehozása pedig a csoportosítás nélküli adatokon végzett elemzések eredményei felé vezet. Ezért amennyiben a csoportok megegyeznek a rétegekkel, vigyázni kell, nehogy túl kevés réteg kerüljön kialakításra, mert két-három csoport elemzése meglehetősen félrevezető lehet. A továbbiakban a jövedelmi adatok decilisei segítségével osztottam 10 jövedelemkategóriába a háztartásokat. Természetesen a sokasági paraméter megfelelő becsléséhez – mint ahogy az a rétegzett mintavétel elméletéből megismerhető – célszerű a sokaság jellemzőit használni a csoportok kialakításakor. Ehhez mindenképpen külső információkra van szükség. Amennyiben ezek nem állnak rendelkezésre, úgy a mintabeli jellemzők is használhatók.

6.5.2. TENDENCIÁK FELTÉRKÉPEZÉSE

A csoportok kialakítása után meghatároztam az egyes jövedelemkategóriákban a fogyasztási kiadások átlagos értékét, ezt követően pedig a csoportátlagok tendenciáit vizsgáltam. A csoportok a jövedelmi tizedeknek felelnek meg, így a szerényebb jövedelmű háztarásoktól tartanak a legmagasabb jövedelműek felé. Ez alapján feltételezhető, hogy a magasabb jövedelmű csoportokba tartozó háztartások fogyasztási kiadásai is magasabbak lesznek.

Tehát ha a csoportátlagok valamilyen irányú tendenciát mutatnak, akkor az leírható adott matematikai függvény segítségével. A kiválasztott minta esetében az egyes háztartási csoportok átlagos kiadásainak alakulása exponenciális függvénnyel jellemezhető, melynek magyarázó ereje 96,5%. Amennyiben reprezentatív a mintavétel, úgy a sokaság megfelelő csoportjainak átlagai is hasonló (jelen esetben exponenciális) görbét rajzolnak.

Kiinduló feltételezésem szerint a nemválaszolók a magasabb jövedelemmel rendelkező háztartások. Ezt az egyszerűsítést a következő fejezetben bemutatásra kerülő modell könnyebb megértése és világosabb interpretációja érdekében teszem. A későbbiekben a fenti szűkítést elhagyva életszerűbb, eltérő helyzetű nemválaszolásokat is vizsgálok. Figyelembe véve, hogy a KSH tapasztalatai szerint nem feltétlenül a vagyonosabb rétegekből kerülnek ki a nemválaszolók. A rosszabb életkörülményekkel rendelkezők, munkanélküliek, alacsonyan képzettek is könnyen válasz megtagadóvá válnak sokszor pusztán az általános közönyösség, érdektelenség okán, vagy esetleg abból fakadóan, hogy a felmérés során vizsgált jelenséget nem érzik saját életüket érintő körülménynek.

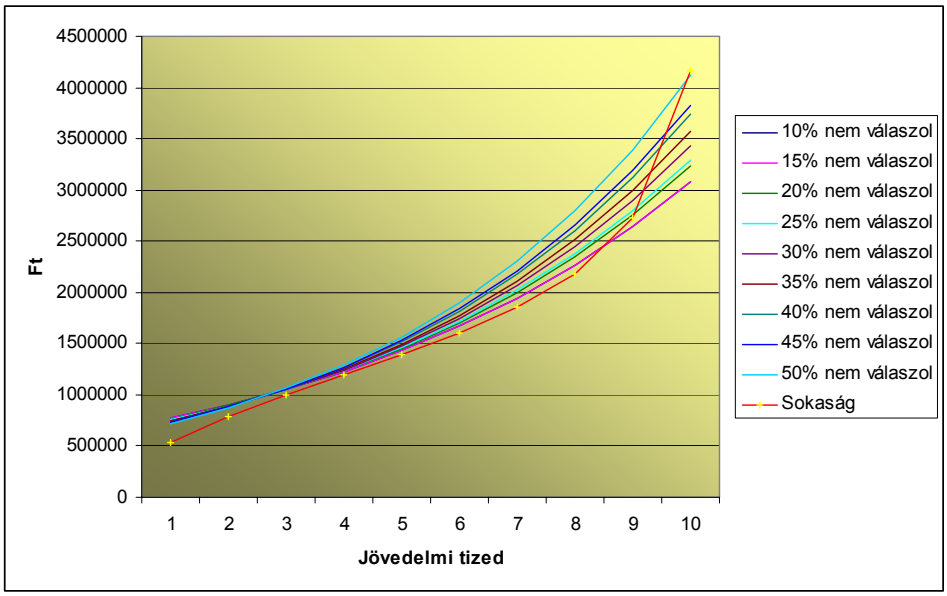
Vizsgálatom során ezúttal is különböző szintű nemválaszolásokkal dolgoztam úgy, hogy mindig szisztematikusan a felsőbb tizedekből kerültek ki a nemválaszolók. Ebben az esetben ugyanis az alacsonyabb jövedelmi rétegek kvázi teljesen válaszadónak számítanak, ami azt jelenti, hogy az adataikban feltárt tendencia kevesebb torzítást tartalmaz, mint egy imputált, vagy súlyozott minta belső tendenciái. Ezt felhasználva a válaszadók adataiban megtalálható tendenciát extrapolálva becsültem a felsőbb csoportok átlagait. A válaszadók adataira épített exponenciális függvények főbb adatai a következők:

20. táblázat: A válaszadók adataira épített exponenciális függvények paraméterei és magyarázó ereje

Nemválaszolás mértéke	***b***	***a***	***R^2***
10% nemválaszolás	1,16594538	664262,24	0,952
15% nemválaszolás	1,16596981	664215,84	0,9521
20% nemválaszolás	1,17427231	648690,46	0,9441
25% nemválaszolás	1,17744969	643453,1	0,9483
30% nemválaszolás	1,18393551	632936,05	0,9317
35% nemválaszolás	1,19079143	623265,03	0,9408
40% nemválaszolás	1,19810866	613166,01	0,9167
45% nemválaszolás	1,20202532	608514,3	0,9218
50% nemválaszolás	1,21322396	595488,81	0,8856

Mivel a nemválaszolás szisztematikus, így a válaszadók tendenciáit leíró függvények magyarázó ereje minden nemválaszolási szinten rendkívül jónak mondható.

A különböző nemválaszolási szintek tendenciái természetesen eltérnek egymástól, hol alul-, hol pedig felülbecsülve a sokasági paramétert. A 14. ábra a különböző nemválaszolási szinteken a becsült adatokat mutatja.



14. ábra: Az egyes jövedelmi tizedekbe eső háztartások átlagos fogyasztási kiadásának becsült értékei különböző nemválaszolási szinteken

Az exponenciális függvények a 9. és 10. tizedben láthatóan jelentősen alulbecsülik a fogyasztási kiadás átlagát, míg a többi tizedben kisebb mértékű felülbecslés tapasztalható.

Az első 5 csoport felülbecslési torzításával a végeredmények meghatározásakor nem kell számolni, mert ezek a csoportok nagy valószínűséggel maximális válaszadási arányt képviselnek, ezért ezeknek a csoportoknak a becslésére nincs szükség, így valós adataikat fel lehet használni.

6.5.3. SÚLYOZOTT TENDENCIÁK BECSLÉSI MODELLJE

A 14. ábrából látható, hogy az 50%-os nemválaszolás esetén generált függvény még a 10. tizedben is felülbecsli a sokasági értéket, azonban figyelembe kell venni, hogy ennek a függvénynek a magyarázó ereje a legkisebb. A nemválaszolás alacsonyabb szintjein a függvények magyarázó ereje jobb, viszont a magasabb tizedekben jelentősen alulbecsülnek.

Mindezek alapulvételével a végső modellemben az egyes jövedelmi tizedek várható értékének becslését a fenti függvények becsült értékeinek a függvények korrigált magyarázóerejének arányával súlyozott átlagos becsült értékeként határoztam meg. Így az a függvény, amelyik alacsonyabb magyarázó erővel rendelkezik, relatíve alacsonyabb súlyt kapott a becsült érték meghatározásában. Ez azt jelenti, hogy az alacsonyabb nemválaszolási

szinteken generált függvények (melyek egyre több réteg figyelembevételével lettek meghatározva, ezáltal pontosabbak is) nagyobb súllyal szerepelnek a végeredmények kialakításában.

Következő lépésben az átlagos becsült értékek változását határoztam meg az egyes csoportok között, vagyis az átlagos becsült értékek növekedésének ütemét a jövedelmi kategória növekedése mellett. Ezek ismeretében adott felső réteg nemválaszolása esetén a válaszadó adatait a növekedés ütemével kiegészítve az átlagos fogyasztási kiadás közelítő, bár nem torzítatlan becslése adható.

A modell különböző nemválaszolási szintek mellett alkalmazható. Ezért a megvalósult válaszadási arány mellett a kutatónak mesterségesen kell további nemválaszolásokat generálni a csoportokban. A nemválaszolás generált mértékei közötti lépték természetesen változtatható azzal a kitétellel, hogy a lépték mértéke a nemválaszoló csoportok méretével arányos legyen. Például 70%-os válaszadási arány mellett generálhatók további 35, 40, 45, 50 százalékos nemválaszolási mértékeket, ezáltal biztosítva, hogy öt különböző függvény súlyozásából származzanak a becslési eredmények. A 21. táblázatban szemléltetem 30%-os nemválaszolás esetében a becslés menetét, a modell alkalmazását.

A táblázat 3-6. oszlopaiban találhatók az exponenciális függvények által becsült átlagos fogyasztási kiadások a különböző mesterséges nemválaszolási mértékek mellett. (A függvények paraméterei a 18. táblázatban találhatók.) A táblázat sárgával jelölt sorában a függvények korrigált R^2 értékei láthatók, melyek alapján az utolsó sor tartalmazza a kiszámított függvénysúlyokat. Minden jövedelmi tizedben a 3-6. oszlopban szereplő becsült értékek és a megfelelő függvénysúlyok felhasználásával határozhatók meg a 7. oszlop adatai, az átlagos becsült fogyasztási értékek. Ezekből a tizedek közötti növekedés relatív mértéke egyszerűen számítható. (Meg kell jegyezzem, hogy a növekedés mértéke, az exponenciális függvények súlyozásának köszönhetően egy tökéletesen illeszkedő exponenciális függvényt eredményez.)

Mivel a megkérdezetteknek 30%-a nem válaszolt, és azt feltételeztem, hogy a nemválaszolás a jövedelmi viszonyok függvénye, így belátható, hogy az utolsó 3 tized adatainak becslésére kell koncentrálni. Ezért a végső becsült értékeket tartalmazó utolsó oszlop első hét sora (legalsó 7 jövedelmi tized) megegyezik a második oszlopban található teljes válaszadás melletti értékekkel. Az alsó hét tizednél a becsült értékek helyett a tényleges adatokkal számolva, jelentős mértékben csökkentők a becslési módszer hibái. A

tényleges becslés ebben az esetben tehát a 7. tizedtől kezdődik, annak értékét rendre megszorozva a növekedés ütemével. Így 30%-os nemválaszolás mellett, figyelembe véve a nemválaszolók tendenciáit, a fogyasztási kiadások átlagos értéke 1.767.559,- Ft-ra becsülhető. Ez az érték a sokaságban 1.744.633,- Ft.

21. táblázat: A súlyozott tendenciák becslési modellje
30%-os ténylegesen megvalósult nemválaszolás esetében

Jövedelmi tizedek	***Fogyasztás a teljes válaszadás mellett***	***30%NV***	***35%NV***	***40%NV***	***45%NV***	***50%NV***	***Átlagos becsült függvény-érték***	***A tizedek becsült átlagai***
1	*650.298*	749.355	742.179	734.640	731.450	722.461	736.179	*650.298*
2	*916.414*	887.189	883.780	880.178	879.221	876.507	881.437	*916.414*
3	*1.170.972*	1.050.374	1.052.398	1.054.549	1.056.846	1.063.400	1.055.428	*1.170.972*
4	*1.418.208*	1.243.575	1.253.186	1.263.464	1.270.355	1.290.142	1.263.851	*1.418.208*
5	*1.374.019*	1.472.313	1.492.283	1.513.767	1.526.999	1.565.231	1.513.536	*1.374.019*
6	*1.739.427*	1.743.123	1.776.998	1.813.658	1.835.492	1.898.976	1.812.674	*1.739.427*
7	*1.944.533*	2.063.746	2.116.034	2.172.959	2.206.308	2.303.883	2.171.083	*1.944.533*
8	*2.214.489*	2.443.342	2.519.755	2.603.441	2.652.038	2.795.126	2.600.538	2.329.175
9	*2.475.128*	2.892.759	3.000.503	3.119.205	3.187.817	3.391.114	3.115.157	2.790.094
10	*3.291.167*	3.424.840	3.572.973	3.737.147	3.831.836	4.114.181	3.731.872	3.342.456
átlag	1.719.465	1.797.062	1.841.009	1.889.301	1.917.836	2.002.102	1.888.176	1.767.559
$\overline{R^2}$		0,9317	0,9408	0,9167	0,9218	0,8856		
függvény súlyok	1	0,20269	0,20467	0,19943	0,20054	0,19266		

7. tézis

Mesterséges nemválaszolási szintek generálásával, amennyiben a válaszadókra nézve jól kitapintható tendencia érvényesül, a súlyozott tendenciák becslési modellje segítségével hatásosabb becslés adható a sokasági paraméterre.

Megállapítható, hogy a modell a megvalósult nemválaszolási szint növekedésével egyre nagyobb mértékben torzít. Kis arányú nemválaszolás esetében azonban alulbecsli a sokasági paramétert. A súlyozott tendenciák becslési modellje az alábbi feltételezésekkel alkalmazható:

- léteznek olyan – a vizsgált tulajdonsággal összefüggő ismérv vagy ismérvek, melyek a nemválaszolást determinálják,
- ezen ismérv/ismérvek mentén a sokaság (lehetőleg egyforma méretű) csoportokba rendezhető,

– létezik a csoportok tendenciáit szignifikánsan leíró, megbízható matematikai függvény,
– a válaszadási arány nagyobb, mint 50%.

A feltételek teljesülésével a modell a háztartások átlagos fogyasztási kiadásainak relatíve jó közelítő érékét adja. Rendkívül előnyös tulajdonsága, hogy a jelentős mértékű alulbecslést, melyet az imputálási, illetve átsúlyozási módszerek esetében tapasztalhattunk képes ellensúlyozni. Az átsúlyozáson, illetve imputáláson alapuló cold deck módszereknek nagy hátránya, hogy nem alkalmasak a becslésre olyan esetekben, amikor a mintából teljes rétegek maradnak ki a nemválaszolás miatt. Ilyen esetekben ugyanis – különösen aszimmetrikus eloszlású ismérvek becslésekor – a hiányzó réteg (esetleg rétegek) információi teljesen elvesznek.

A modell sajnos a nemválaszolás magas szintjénél látszólag zavaró mértékű felülbecslést eredményezhet. A modell nemválaszolás okozta torzításra gyakorolt csökkentő hatását mutatja be, illetve viszonyítja a súlyozásos eredményekhez a 22. táblázat.

22. táblázat: A nemválaszolást kezelő módszerek eredményeinek viszonyítása a sokasági paraméterhez

Nemválaszolás mértéke	***Súlyozatlan átlagos fogyasztási kiadás***		***Súlyozott átlagos fogyasztási kiadás***		***Súlyozott tendenciák becslési modellje***	
	Ft	***A várható érték %-ában***	***Ft***	***A várható érték %-ában***	***Ft***	***A várható érték %-ában***
TC a felső 10% nem válaszolt	1.475.398	84,57%	1.554.136	89,08%	1.684.073	96,53%
TC a felső 15% nem válaszolt	1.389.274	79,63%	1.494.852	85,68%	1.719.168	98,54%
TC a felső 20% nem válaszolt	1.316.725	75,47%	1.428.680	81,89%	1.721.414	98,67%
TC a felső 25% nem válaszolt	1.253.037	71,82%	1.371.626	78,62%	1.762.698	101,04%
TC a felső 30% nem válaszolt	1.193.976	68,44%	1.316.734	75,47%	1.767.559	101,31%
TC a felső 35% nem válaszolt	1.138.382	65,25%	1.266.237	72,58%	1.850.261	106,05%
TC a felső 40% nem válaszolt	1.084.923	62,19%	1.214.319	69,60%	1.858.925	106,55%
TC a felső 45% nem válaszolt	1.032.088	59,16%	1.172.525	67,21%	1.806.398	103,54%
TC a felső 50% nem válaszolt	979.840	56,16%	1.105.332	63,36%	1.826.144	104,67%

Míg a súlyozásos, illetve a nemválaszolások elhagyásával számított átlagos fogyasztási kiadások a magasabb nemválaszolási szinteken 40%-körüli torzítást is eredményezhetnek, addig a súlyozott tendenciák becslési modellje csupán 5%-körüli torzítást mutat. Mindemellett a különböző módszerek együttes alkalmazása ajánlott, hiszen nem felejthető el, hogy adott minta csupán egy lehetséges realizációja a mintavételi tervnek, a vizsgált tulajdonság pedig valószínűségi változó, melyet a véletlenen kívül számos más tényező befolyásolhat, melyekre a fenti modellek külön-külön nem képesek megoldást nyújtani.

6.5.4. ELTÉRŐ HELYZETŰ NEMVÁLASZOLÁSOK VIZSGÁLATA

A következőkben a modell működésének ellenőrzésére olyan eseteket vizsgálok, amikor a nemválaszolás nem kizárólag egyoldalú. Eddigi leegyszerűsített feltételezéseim szerint kizárólag a magasabb jövedelemmel rendelkezőket tekintettem nemválaszolónak. Ebből eredően olyan megvalósulásokat vizsgáltam, amikor a minta terjedelmének csak az egyik oldalánál (a maximális értékek felől) következik be csökkenés. Mint említettem ez azonban a mintavételen alapuló kutatások tapasztalatai szerint nem teljesen életszerű. Ezért azokat a lehetőségeket is elemeznem kell, ahol a nemválaszolás két oldalról csonkolja a mintát. Az eddigi problémával analóg módon a nemválaszolás itt is tetszőleges méreteket ölthet, azonban továbbra is feltételezem, hogy nem lépi túl együttesen az 50%-os mértéket. Annak érdekében, hogy számításaimat bemutathassam, 10%-os léptékkel növelem a nemválaszolás mértékét a minta valamely oldalán. Még ebben az esetben is elég jelentős számú kombináció merülhet fel. Ezért, hogy korlátozzam a lehetséges összetételek számát és mégis szemléletes eredményeket produkáljak a nemválaszolási lehetőségeket a magasabb jövedelemmel rendelkezők esetében maximum 30%-ra, az alacsony jövedelemmel rendelkezők esetében pedig maximum 20%-ra teszem. Így végeredményben 50%-os nemválaszolás is megvalósulhat, ezenkívül kétféleképpen érhető el 40%-os, valamint három 30%-os, három 20%-os és két 10%-os nemválaszolás is vizsgálható.

23. táblázat: A súlyozott tendenciák becslési modelljének eredményei a teljes fogyasztás becslésében, különböző nemválaszolási realizációkban

Nemválaszolás mértéke	***Becslés nemválaszolások mellett***		***Súlyozott tendenciák becslési modellje***	
	Becsült teljes fogyasztás (Ft)	***A várható érték %-ában***	***Becsült teljes fogyasztás (Ft)***	***A várható érték %-ában***
nv_f10	1.544.832	88,55%	1.675.538	96,04%
nv_f20	1.428.545	81,88%	1.693.421	97,06%
nv_f30	1.316.267	75,45%	1.706.211	97,80%
nv_a10	1.838.262	105,37%	1.734.666	99,43%
nv_a20	1.953.493	111,97%	1.756.714	100,69%
nv_f10a10	1.656.649	94,96%	1.687.891	96,75%
nv_f10a20	1.762.396	101,02%	1.708.690	97,94%
nv_f20a10	1.539.723	88,25%	1.698.880	97,38%
nv_f20a20	1.643.608	94,21%	1.715.116	98,31%
nv_f30a10	1.427.262	81,81%	1.699.990	97,44%
nv_f30a20	1.529.432	87,66%	1.708.190	97,91%

A táblázat rövidített jelölései némi magyarázatra szorulnak, miszerint „f” jelzi a felső jövedelmi tizedek nemválaszolását, „a” pedig az alsó jövedelmi tizedek nemválaszolását. Ez alapján *nv_f20a10* azt jelenti, hogy a minta felső 20%-a és az alsó 10%-a nem válaszolt, tehát összességében 30%-os nemválaszolással realizálódott.

A táblázat középső oszlopa mutatja a torzítás mértékét az eltérő nemválaszolások esetében. Mivel alsó és felső szintekről is számoltam nemválaszolással, ezért bizonyos arányok eltalálása (vagy éppen el nem találása) folytán a várható értékhez egészen közeli eredmények is elérhetők. Amennyiben a nemválaszolás hatását nem veszem figyelembe, úgy 75%-tól 112%-ig változó mértékű alul és felülbecslés között ingadoznak a végeredmények. De az elemzés nem bízható a szerencsére, azt remélve, hogy összességében kiegyenlítik egymást a hiányzó magas és alacsony értékek.

Az utolsó oszlopból látható, hogy a súlyozott tendenciák becslési modelljének alkalmazásával a becsült értékek többnyire csupán néhány százalékos alulbecsléssel, stabilan közelítik a sokasági várható értéket. Azzal, hogy az alsó és felső értékekből is realizálódik nemválaszolás a modell kiegyensúlyozottabb lesz, aminek köszönhetően eltűnik az a szimmetrikus torzítás, ami a 22. táblázat utolsó oszlopában látható a nemválaszolási szint növekedésével. Ez azt igazolja, hogy a modell plasztikusan alkalmazható különböző nemválaszolási szinteken, és mértékek mellett.

7. ÖSSZEGZÉS

Kutatási munkám elsődleges céljának a mintavételen alapuló felmérések lehetséges hibáinak, és azok eredményekre gyakorolt negatív hatásainak feltárását jelöltem meg. Ezt követően megoldási változatokat kerestem a hibák, majd elsődlegesen a nem véletlen jellegű hibák kezelésére. A hibák kezelésének leghatékonyabb módja, ha megelőzzük a keletkezésüket. Azonban ha a hiba már bekövetkezett, akkor a kezelés első fázisaként fel kell térképezni a hiba okát. Milyen okokra vezethetők vissza a társadalomtudományi kutatások hibái? Melyek lehetnek azok az eredő tényezők, amelyek alapján egy kutatás, vagy annak eredménye megalapozatlannak minősül? Ezekre a kérdésekre a szakirodalom a következő válaszokat adja: megbízhatatlan, megalapozatlan, kis mintára épül, magas a relatív hiba, nem szignifikáns az eredő problémára irányuló hatása, stb.

További kritikai tényezők merülhetnek fel abban az esetben, amikor a kutatás egy mintavétel során nyert primer adathalmazt dolgoz fel, elemez, és von le különböző következtetéseket. A mintavétellel ugyanis a hibaforrások száma fokozódik. A mintavétel során elkövethető hibák már az előállított forrásadatok mennyiségében és minőségében is torzulást eredményezhetnek, ami azért veszélyes, mert a rossz alapadatokból a legjobb módszertan alapján, professzionális elemzési technikák és eszközök segítségével végzett alapos számítások mellett is téves következtetésre juthatnak a kutatók.

A fenti problémákra és kérdésekre a megoldásokat és a válaszokat a KSH által rendelkezésemre bocsátott 2005. évi Háztartási Költségvetési Felvétel adatbázisának vizsgálatával kerestem. Az adatok hitelességének alátámasztásához bemutattam a HKF adatainak forrását, a felvétel módszertanát. Különböző minták elemzésén keresztül törekedtem a hibák hatásának bemutatására, és a negatív hatások csökkentésének tudományos magyarázatára.

Doktori munkám első részében a következtetések alapját képező minta kialakításának körülményeit vizsgáltam. Első lépésben áttekintettem a mintavételi eljárásokra vonatkozó elméleti alapvetéseket, majd a mintavételi- és hibaszámítási módszerek hazai és nemzetközi szakirodalmát tanulmányoztam át, ami segített az eddigi kutatási eredmények felhasználásában.

Tudományos munkám további részében a potenciális hibaforrások azonosításával és rendszerezésével foglalkoztam, melyhez a statisztikai-matematikai megközelítésen kívül szá-

mos adalékkal szolgáltak más diszciplínák, úgymint marketing, szociológia területén végzett kutatások eredményei is. A kérdőíves felmérések, közvélemény-kutatások tapasztalati segítettek megismerni a hibák kezelése érdekében tett lépéseket, és a hibák méretének csökkentésére alkalmazott eszközöket, módszertani elgondolásokat.

Munkám során 53 különböző mintavételi eljárás alapján generáltam mintákat a sokaság (a HKF adatbázisa) adataiból, annak érdekében, hogy minél részletesebben vizsgálhassam a mintavételi tervek eredményre gyakorolt hatását. Következtetéseim, becslési eredményeim ellenőrzésére lehetőséget biztosított az, – a gyakorlatban nem teljesülő feltétel – hogy a vizsgált jelenség sokasági információi a birtokomban voltak. A minták egyszerű véletlen, rétegzett, illetve több ismérv szerint rétegzett eljárással készültek, mivel tapasztalataim szerint a vállalati kutatásokban jellemzően nem használnak bonyolultabb mintavételi eljárásokat.

Ezt követően olyan szempontok kidolgozására vállalkoztam, amelyek alapján – ha nem is a minőség összes értelmezhető kritériuma tekintetében, de néhány fontosabb jellemző alapján – minősíteni, rangsorolni lehet a különböző mintákból nyert adatokat, becslési eredményeket. Természetesen mindezt számos elméleti feltétel fennállása mellett kíséreltem meg, mely feltételek az alkalmazott gyakorlatban meglehetősen ritkák, olykor egyáltalán nem teljesülnek. A dolgozat empirikus kutatásokat tartalmazó részeiben pedig kizárólag olyan eljárásokat alkalmaztam, amelyek nemcsak a hivatalos statisztikák készítésénél, hanem a vállalkozások gyakorlatában is „egyszerűen” alkalmazhatók.

A mintavételi tervek rangsorolásánál átlag- és hányadosbecslés eredményeinek javítása szempontjából rangsoroltam a mintavételi terveket. A minősítéshez két indikátort használtam: a relatív standard hibát (CV) és a Design Effekt (Deff) mutatót. A Deff mutató, a CV relatív hibával együtt alkalmas arra, hogy minősítse az azonos méretű, de egyszerű véletlen mintavételnél hatékonyabb mintavételi terveket.

A változónként végzett elemzések arra is lehetőséget adtak, hogy megvizsgáljam, vajon a mintavételi tervben szereplő rétegképző ismérvnek vagy ismérveknek a vizsgált változóhoz fűződő sztochasztikus viszonya mutat-e valamilyen összefüggést a Deff és a CV által állított rangsorral. Ennek eredményeként kimutattam, hogy a rétegképző ismérv vagy ismérvek korrelációs együtthatója – többszörös rétegzés esetén többszörös korrelációs együtthatója – determinisztikus viszonyban van a Deff és CV által kialakított rangsorral. Elmondható, hogy minden esetben a kevésbé hatékony minták klaszterébe kerültek a 0,4-nél kisebb korrelációs együtthatójú rétegképző ismérvvel rendelkező minták.

A hazai és nemzetközi kutatások tapasztalatai alapján egyaránt elmondható, hogy a válaszadás hiányossága talán az egyik legnagyobb probléma, ami a felmérések készítésénél felmerül. Manapság nem ritkák az 50 %-on aluli válaszadási aránnyal rendelkező kérdőívek. Nyilvánvaló, hogy a szelektív válaszadás nemcsak a mintanagyságot csökkenti, hanem növeli a becslések varianciáját, valamint a torzítás mértékét is.

Éppen ezért a dolgozat további részeiben a nem mintavételi hibák egyik legfontosabb típusának, a nemválaszolási hibának a vizsgálatával foglalkoztam. Ezen belül is a részleges, vagy item szintű nemválaszolással. Különböző elemzési módszerek segítségével azt kutattam, vajon a megtagadások pótlása mekkora hatást gyakorol a leíró modellek eredményeire. A kérdés megválaszolásához a háztartások összes fogyasztási kiadásának becslését végezve különböző mértékű nemválaszolásokat generáltam, azoknak a kutatási tapasztalatoknak a figyelembe vételével, melyek szerint a nemválaszolók elsősorban a magasabb jövedelmi rétegekből kerülnek ki. A nemválaszolók közötti demográfiai-, társadalmi-, gazdasági hasonlóságok feltételezésével sikerült olyan klasztereket előállítani, melyekből az egyik – a jobb életkörülményekkel rendelkezők klasztere – tartalmazta a nemválaszolók 93,3%-át. Így lehetőség nyílt a klaszter tagságon alapuló imputáció alkalmazására. Párhuzamosan, hasonló változók mentén regresszión alapuló imputációt is végeztem. Ennek az iteratív eljárásnak a megoldásai közül azt választottam, melyben az imputált adatok minimális értéke jelentősen (némely iterációhoz képest 70%-kal) meghaladja a többi iteráció minimumát, remélve, hogy ezáltal jelentősebben csökkenthetem a nagy mértékű alulbecslést. A különböző módszerek közül nagyobb részletességgel mutattam be a kalibrációs eljárás alkalmazásának elméleti hátterét. Tettem ezt egyrészt azért, mert az alkalmazott statisztikai kutatások egyre nagyobb jelentőséget tulajdonítanak a módszer sikeres alkalmazásának. Másrészt a kalibráció olyan módszer, melynek algoritmusait nem tartalmazzák a széles körben alkalmazott statisztikai elemző szoftverek. A módszer alkalmazására speciális számítógépes program szükséges, amely nehezen hozzáférhető a vállalkozások számára. Emellett a legközelebbi szomszéd módszere alapján is végeztem inputációt, ahol a hasonlóságot a Mahalanovis távolság alapján vizsgáltam.

A kapott eredményeket a teljes minta, a hiányzó adatokat tartalmazó minta és természetesen a sokaság megfelelő paramétereinek viszonylatában értékeltem.

Azt tapasztaltam, hogy a teljes mintából megfelelően és viszonylag alacsony standard hibával becsülhető a sokasági paraméter. Abban az esetben viszont, amikor adathiány lépett

fel, a becslés értéke jelentős mértékben, több mint 9%-kal alulmúlta a sokasági értéket. Az imputációk eredményeként kapott, valamint a kalibrált mintákból származó becslések javultak az adathiányos mintához képest, hiszen többségében a sokasági paramétertől kevesebb, mint 4,5%-kal kisebb értéket becsülnek. A torzítás azonban továbbra is jelen volt a becslésekben, hiszen a 95%-os megbízhatósági szint mellett számított konfidencia intervallumok nem fedték a sokasági paraméter értékét. Megállapítható, hogy abban az esetben, ha egy baloldali aszimmetriát mutató sokaság lineáris statisztikáit becsüljük, akkor az egyszerű módszerekkel imputált becslések továbbra is torzítottan alulbecslik a sokasági paramétert.

Az utolsó fejezetben azt vizsgáltam, hogy a mintavételi terv hatásának van-e szerepe a nemválaszolás okozta torzítás kialakulásában. Különböző mintavételi tervek alapján a fogyasztási kiadás átlagos értékének eltérő válaszadási arányok mellett történő becsléseiből azt tapasztaltam, hogy abban az esetben, ha a nemválaszolás vélt vagy valós oka sztochasztikus összefüggést mutat a rétegképző ismérvvel, a rétegzés növeli a nemválaszolás okozta torzítás mértékét.

A kutatónak a mintavétel megtervezésekor figyelembe kell vennie a vizsgált ismérvvel kapcsolatos megtagadási várakozásokat, és ha azok potenciálisan teljes rétegeket érintenek, akkor érdemes lazítani a vizsgált ismérvvel sztochasztikus kapcsolatban levő rétegképző ismérvhez fűződő reprezentativitási követelményeken.

Kutatásom utolsó fázisában a modell alapú eljárások szerepét teszteltem a nemválaszolás kezelésében. A mintában fellelhető információk alapján három klasszifikációs algoritmus: diszkriminancia-analízis, döntési fa és logisztikus regresszió alapján azonosítottam a nemválaszoló háztartásokat, majd figyelembe véve azok jellemzőit, a választ megtagadó háztartások nemválaszolási valószínűségeit térképeztem fel. A pontosság, elsőfajú hiba és másodfajú hiba alapján a logisztikus regresszió bizonyult a legmegfelelőbb eljárásnak.

A logisztikus regressziófüggvény becsült válaszadási valószínűsége alapján képzett súlyokat használtam a fogyasztási kiadás becslésének javítására. A súlyozás eredményeként legalább 5%-kal sikerült javítani a nemválaszolás torzító hatásán. A modell és a súlyozás helyességét mutatja, hogy a nemválaszolási szintek növekedésével a súlyozási módszer torzítást csökkentő hatása egyre javul. Azonban meg kell jegyezni, hogy a gyakorlati kutatásoknak nem csupán az a célja, hogy enyhítsék a torzítás negatív hatásait, hanem a sokasági paraméter minél pontosabb és megbízhatóbb becslése. Ezt a célt azonban csak rész-

ben sikerült teljesíteni. A sokasági érték ismeretében ugyanis elmondható, hogy a nemválaszolás mértékének szisztematikus növelésével a torzítás a súlyozás ellenére is drasztikus méreteket ölt.

A nemválaszolás okozta torzítás kiküszöbölésére tett lépések között fontos szerepe van a tendenciák azonosításának. Érdemes megvizsgálni a válaszadók és nemválaszolók vizsgált ismérvbeli tendenciái közötti különbséget. Persze mindezek előtt fel kell térképezni, hogy léteznek-e egyáltalán valamiféle tendenciák. Az eljárás sikere érdekében megfelelő csoportokat kell kialakítani a mintában a vizsgált ismérvvel sztochasztikus kapcsolatban álló és a nemválaszolást generáló ismérv(ek) alapján. A tendenciákat ezen csoportok mentén kell vizsgálni és modellezni. Dolgozatomban a fogyasztási kiadás becsléséhez a háztartások jövedelme alapján képeztem csoportokat és a különböző jövedelmi tizedekbe eső háztartások fogyasztási kiadásaiban tapasztalható exponenciális tendenciákat azonosítottam, eltérő válaszadási arányok mellett. Tapasztalataim szerint a nemválaszolás alacsonyabb szintjein a függvények magyarázó ereje jobb volt, viszont a magasabb tizedekben jelentősen alulbecsülték a fogyasztási kiadások átlagát. Ezért a *súlyozott tendenciák becslési modelljében* a fenti tendenciák becsült értékeit a függvények magyarázóerejének súlyozásával számított *átlagos becsült értékként* határoztam meg. Ezáltal az alacsonyabb nemválaszolási szinteken generált függvények (melyek egyre több réteg figyelembevételével lettek meghatározva, ezáltal pontosabbak is) nagyobb súllyal szerepelnek a végeredmények kialakításában.

Meghatározva az *átlagos becsült értékek* növekedésének mértékét a jövedelmi kategória növekedése mellett, adott felső réteg nemválaszolása esetén a válaszadó adatait a növekedés mértékével kiegészítve az átlagos fogyasztási kiadás minimálisan torzított becslése adható.

A modell természetesen számos elméleti feltétel mellett működik, ezeket figyelembe véve, a gyakorlatban is elfogadható 30%-os nemválaszolás esetében a súlyozott tendenciák becslési modellje csupán 4%-körüli torzítást mutat, 11%-os relatív hiba mellett.

Végezetül megállapítható, hogy a különböző módszerek együttes alkalmazása ajánlott, hiszen nem felejthető el, hogy adott minta csupán egy lehetséges realizációja a mintavételi tervnek, a vizsgált tulajdonság pedig valószínűségi változó, melyet a véletlenen kívül számos más tényező befolyásolhat, melyekre a fenti modellek külön-külön nem képesek megoldást nyújtani.

IRODALOMJEGYZÉK

Antal E. – Tillé Y.: Simple random sampling with over-replacement In.: Journal of Statistical Planning and Inference, Volume 141, Issue 1, January 2011, pp. 597-601.

Ay János − Vita László: Egy kísérleti jövedelmi felvétel főbb tapasztalatai; Statisztikai Szemle, 1998. 76. évf. 6. szám pp. 515-532.

Benedetti R. – Bee M. – Espa G.: A Framework for Cut-off sampling in business survey design; Journal of Official Statistics Vol.26, No.4, 2010 pp. 651–671.

Besenyei Lajos – Varga Beatrix – Domán Csaba – Szilágyi Roland: Az elemez-hetőséget biztosító mintaillesztés megvalósítása; In: Innovációmenedzsment, Tudásteremtés – Tudástranszfer Konferencia Kiadvány, Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2006, ISBN-13: 978-963-661-729-5.

Besenyei Lajos – Varga Beatrix – Domán Csaba – Szilágyi Roland: Kvantitatív információképzési technikák; Miskolci Egyetem Elektronikus tananyag 2010.

Biemer P. P.: The twelfth Morris Hansen lecture simple response variance: then and now; Journal of Official Statistics, Vol. 20, No. 3, 2004. pp. 417-439.

Biggs, D. – B. de Ville – E. Suen.: A method of choosing multiway partitions for classification and decision trees; Journal of Applied Statistics, 18, pp. 49-62. 1991.

Billiet J. – Philippines M. – Fitzgerald R. – Stoop I.: Estimation of nonresponse bias in the European Social Survey: Using Information from reluctant respondents; Journal of Official Statistics, Vol. 23,No. 2, 2007. pp. 135-162.

Bolla Marianna – Krámli András: Statisztikai következtetések elmélete; Typotex Kiadó, Budapest, 2005.

Bukodi E. – Altorjai Sz. – Tallér A.: A magyar foglalkoztatási rétegszerkezet az ezredforduló után; Statisztikai Szemle 2006. 84. évf. 8. szám, pp. 733-763.

Corlett T.: Sampling errors in practice; Journal of Market Research Society October, 1996. pp. 307-318.

Éltető Ödön – Marton Ádám: A mintanagyság és a meghiúsulások kapcsolata a reprezentatív felvételekben; Statisztikai Szemle, 1995. 10. sz. pp. 789-798.

Éltető Ödön: Mintavétel véges alapsokaságból; Budapest, 1970.

Estevao V. M. – Särndal C. E.: Borrowing strength is not the best technique within a wide class of design-consistent domain estimators; Journal of Official Statistics, Vol. 20, No. 4, 2004, pp. 645–669.

Estevao V. M. – Särndal C. E.: The ten cases of auxiliary information for calibration in two-phase sampling; Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 233–255.

Estevao V. M. – Särndal C. E.: A functional form approach to calibration; Journal of Official Statistics, Vol. 16, No. 4, 2000, pp. 379–399.

Falus Iván – Ollé János: Statisztikai módszerek pedagógusok számára; Okker kiadó, 2000.

Falus Iván – Ollé János: Az empirikus kutatások gyakorlata; Nemzeti Tankönyvkiadó, Budapest, 2008.

Fellegi, I. P.: Comment; Journal of Official Statistics. 2001. 17. évf. 1. sz. pp. 151–155.

Folsom R. E.−Singh A. C.:The generalized exponential model for disegn weight calibration for extreme values, nonresponse and poststratification, proceedings, section on survey research method; American Statistical Association. 2000. pp. 598-603.

Foster, K.: Weighting the Family Expenditure Survey in Great Britain to compensate for non-response: an investigation using census-linked data. Helsinki. 1996.

Kröpfl B. – Peschek W. – Schneider E. – Schönlieb.: Alkalmazott statisztika; Műszaki Könyvkiadó, Budapest, 2000.

Fuller W. A. [2002]: Regression estimation for survey samples; Survey Methodology, 28, pp. 5-23.

Gambino J. G. – Pedro Luis do Nascimento Silva: Sampling and estimation in household surveys; In.: Handbook of Statistics, Volume 29, Part 1, 2009, Chapter 16, pp. 407-439.

Ghosh-Dastidar B. – Schafer J. L.: Outlier detection and editing procedures for continuous multivariate data; Journal of Official Statistics, Vol. 22, No. 3, 2006. pp. 487-506.

Grusky, D. B. – Weeden, K. A. [2001]: Decomposition without death: a research agenda for the new class analysis; Acta Sociologica. 44. Vol. pp. 203–218.

György Erika: A nemválaszolás elemzése a munkaerő-felvételben; Statisztikai Szemle, 82. évf. 2004. 8. sz. pp. 747-772.

Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Hajdu – Pintér – Rappay – Rédey: Statisztika; Pécs, 1994.

Harms T. − Duchesne P. [2006]: On calibration estimation for quantiles; Survey Methodology, 32, pp. 37-52.

Havasi Éva – Schnell Lászlóné: Az 1996-os jövedelmi felvételre nem válaszoló háztartások – A megtagadások természete, a megtagadók sajátosságai; Központi Statisztikai Hivatal. Budapest. 1996.

Havasi Éva: Szegénység és társadalmi kirekesztettség a mai Magyarországon; Szociológiai Szemle 2002/4. pp. 51–71.

Havasi Éva: Válaszmegtagadó háztartások; Statisztikai Szemle 1997. 10 sz. pp. 831-843.

Hidiroglou M. A. – Lavallée P.: Sampling and estimation in business surveys; In.: Handbook of Statistics, Volume 29, Part 1, 2009, Chapter 17, pp. 441-470.

Horváth Beáta – Mihályffy László: Hibaszámítás Jakknife módszerrel bonyolult felépítésű, kalibrált minták esetén; Statisztikai Szemle 2008. 86. évf. 6. szám pp. 591-613.

Horwitz D. G. − Thompson D. J.: A generalization of sampling without replacement from a finite universe; Journal of the American Statistical Association, 47, pp- 663-685. 1952.

Hunyadi L. – Mundruczó Gy. – Vita L.: Statisztika; Aula Kiadó, 2000. Budapesti Közgazdaságtudományi és Államigazgatási Egyetem.

Hunyadi László – Vita László: Statisztika közgazdászoknak; Budapest, 2002.

Hunyadi László: A mintavétel alapjai; Számalk Kiadó , 2001.

Hunyadi László: Statisztikai következtetéselmélet közgazdászoknak; Budapest, 2005.

Johansson F. – Klevmarken A.: Explaining the size and nature of response in a survey on health status and ecinimic standard; Journal of Official Statistics, Vol. 24, No. 3, 2008. pp. 431-449.

Kapitány Balázs: Mintavételi módszerek ritka populációk esetén; Statisztikai Szemle, 2010. 88. évf. 7-8 szám, pp. 739-754.

Kapitány Zs. – Molnár Gy.: A magyar háztartások jövedelmi-kiadási egyenlőtlenségei és mobilitása 1993–1998 in: KTK/IE Műhelytanulmányok 2001/15, MTA Budapest 2001.

Kehl D. – Rappai G.: Mintaelemszám tervezése Likert-skálát alkalmazó lekérdezésekben; Statisztikai Szemle, 2006. 84. évf. 9. szám pp. 848-875.

Kemény Sándor – Papp László – Deák András: Statisztikai minőség- (megfelelőség-) szabályozás; Műszaki Könyvkiadó, Budapest, 1998.

Ketskeméty László – Izsó Lajos: Bevezetés az SPSS programrendszerbe; ELTE Eötvös Kiadó, Budapest, 2005.

Keszthelyiné Rédei Mária: A lakossági jövedelmek mérésének megbízhatóbb módszere; Statisztikai Szemle, 2006. 84. évf., 5-6. szám pp. 518-551.

Kim, J. O. – Curry, J.: The treatment of missing data in multivariate analysis; Sociological Methode Researche. 1977. 6. Vol. 2., pp. 215–240.

Kish, L: Kutatások statisztikai tervezése; Budapest, 1989.

Knottnerus P. – Duin C.V.: Variance in repeated weighting with an application to the dutch labour force survey; Journal of Official Statistics, Vol. 22 ,No. 3, 2006. pp. 565-584.

Köves P. − Párniczky G.: Általános statisztika; Tankönyvkiadó, Budapest. 1989.

Köves P. − Párniczky G.: Általános statisztika; Tankönyvkiadó, Budapest, 1980.

KSH (1997): A háztartási költségvetési felvétel módszertana; Módszertani Füzetek 37. sz. KSH, Budapest.

KSH: A háztartások fogyasztása 2006, KSH, 2007.

Lee D. – Mathiowetz N. A. – Tourangeau R.: Measuring disability in surveys: consideny over time and across respondents; Journal of Official Statistics, Vol. 23, No. 2, 2007. pp. 163-184

Little, R. J. A. – Rubin, D. B.: Statistical analysis with missing data. 2. szerk. John Wiley & Sons. New York 2002.

Lundström S. – Särndal C. E.: Calibration as a standard method for treatment of nonresponse; Journal of Official Statistics, Vol. 15, No. 2. 1999. pp. 305-327.

Maddala G. S.: Bevezetés az ökonometriába; Nemzeti Tankönyvkiadó, Budapest, 2004.

Marton Ádám – Mihályffy László: A mintavételi hiba kiszámításának néhány kérdése; Statisztikai Szemle, 1988. 4. sz. pp. 350-366.

Marton Ádám: A mintavételi hiba kiszámítása és felhasználása a hivatalos statisztikában; Statisztikai Szemle, 83. évf. 2005.

Marton Ádám: A reprezentatív felvételek megbízhatósága 1991.

Mihályffy László: Meghiúsulások kompenzálása lakossági felvételekben: egy speciális lineáris inverz probléma; (1994) Szigma XXV. évf. 4. sz. pp. 191-202.

Molnár György – Kapitány Zsuzsa: Mobilitás, bizonytalanság és szubjektív jóllét Magyarországon; Közgazdasági Szemle, LIII. évf., 2006. október pp. 845–872.

Mundruczó György: Útmutatás a statisztikai modellezés gyakorlatához; Budapest, 1998.

Murray R. Spiegel: Statisztika elmélet és gyakorlat; Panem-McGraw-Hill, Budapest, 1995.

Naresh K. Malhotra – Marketingkutatás; Budapest, 2002. KJK-KERSZÖV Jogi és Üzleti Kiadó Kft.

Oravecz Beatrix: Hiányzó adatok és kezelésük a statisztikai elemzésekben; Statisztikai Szemle 2008. 86. évf. 4. szám pp. 365-384.

Péter György: Általános statisztika; Tankönyvkiadó Budapest, 1955.

Peytchev A. – Conrad F. G. – Couper M. P. – Tourangeau R.: Increasing respondents' use of definitions in web surveys; Journal of Official Statistics Vol.26, No.4, 2010 pp. 633–650.

Pintér J. – Rappai G.: A mintavételi tervek készítésének néhány gyakorlati megfontolása; In: Marketing &Menedzsment 2001. 35. évf. 4. sz. pp. 4-10.

Pintér J. – Rappai G.: Statisztika; Pécsi Tudományegyetem Közgazdaságtudományi Kar, Pécs, 2007.

Quian J.: Sampling; In.: International Encyclopedia of Education, 2010, Pages 390-395.

Roy D. – Safiquzzaman Md.:Variance estimation by Jackknife method under two-phase complex survey design; Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 35–51.

Rudas Tamás: Hogyan olvassunk közvélemény-kutatásokat? Új Mandátum Könyvkiadó, Budapest 1998.

Rueda M.−Martinez S.−Martinez H.−Acros A.: Estimation of the distribution function with calibration methodes, Journal of Statistical Planning and Inference, 137, 2007. pp. 435-448.

Sajtos László – Mitev Ariel: SPSS kutatási és adatelemzési kézikönyv; Alinea Kiadó, Budapest, 2007.

Särndal C.E. – Lundström S.: Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator; Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 167–191.

Särndal C. E. [2007]: The calibration approach in survey theory and practice; Survey Methodology 2007, Vol. 33, No. 2, pp. 99-119.

Särndal C. E.−Lundstöm S.: Estimation in Surveys with Nonresponse, New York: John Wiley & Sons, Inc. 2005.

Singh H. P. – Kumar S.: Improved estimation of population mean under two-phase sampling with subsampling the non-respondents In.: Journal of Statistical Planning and Inference, Volume 140, Issue 9, September 2010, pp. 2536-2550.

Statistical Policy Working Paper 31.: Measuring and reporting sources of error in surveys; Statistical Policy Office, July 2001.

Székelyi Mária – Barna Ildikó: Túlélőkészlet az SPSS-hez; Typotex Kiadó, Budapest, 2002.

Szép Katalin: A Mintavételi és Módszertani Osztályon folyó műhelymunka; Statisztikai Szemle, 82. évf. 2004. 8. sz. pp. 646.

Szép Katalin – Vígh Judit: A minőség a hivatalos statisztikában; Statisztikai szemle, 82. évf. 8. sz. 2004.

Szűcs István: Alkalmazott statisztika; Budapest, 2002.

Varga Sára: A jövedelemfelvétel hiányzó adatainak pótlása; Statisztikai Szemle 1999. 77. évf. 2-3. sz. pp. 112-130.

PUBLIKÁCIÓS JEGYZÉK

A minta jellemzői; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 26-33. ISBN 978-963-06-7100-2

Besenyei L. – Domán Cs. – Szilágyi R. – Varga B.: „Statisztikai mintaillesztés" program tervezése és megvalósítása; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 8-16, ISBN: 978-963-661-798-1

Besenyei L. – Domán Cs. – Szilágyi R. – Varga B.: Faktoranalízis alkalmazásának lehetősége az innovációs potenciál mérése során; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 45-52, ISBN: 978-963-661-798-1

Besenyei L. – Domán Cs. – Szilágyi R. – Varga B.: Klaszteranalízis alkalmazásának lehetősége az innovációs potenciál mérése során; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 53-64, ISBN: 978-963-661-798-1

Besenyei L. – Varga B. – Domán Cs. – Szilágyi R.: Az elemezhetőséget biztosító mintaillesztés megvalósítása, Innovációmenedzsment, Tudásteremtés – Tudástranszfer Konferencia, Miskolc, 2006. november 15-16.

Grafikus ábrázolás; In.: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 58-73. ISBN 978-963-06-3135-8

Hipotézisvizsgálat; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 53-80. ISBN 978-963-06-7100-2

Ismérvek közötti sztochasztikus kapcsolatok elemzése; In.: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 140-153. ISBN 978-963-06-3135-8

Mintavételi eljárások; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 9-25. ISBN 978-963-06-7100-2

Statisztikai becslés; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 33-52. ISBN 978-963-06-7100-2

Szilágyi R. – Domán Cs.: Az adathiány kezelése mintavételes felmérésekben; Erdei Ferenc V. Tudományos konferencia – „Globális kihívások, lokális megoldások„ Kecskeméti Főiskola Kertészeti Főiskolai Kar Kecskemét, 2009. pp. 75-80. ISBN978-963-7294-74-7

Szilágyi R. – Varga B.: Faktoranalízis In: Kvantitatív információképzési technikák Miskolci Egyetem, Elektronikus tananyag, 2011. . (megjelenés alatt)

Szilágyi R. – Varga B.: Klaszteranalízis In: Kvantitatív információképzési technikák Miskolci Egyetem, Elektronikus tananyag, 2011. (megjelenés alatt)

Szilágyi R. –Domán Cs.: Kalibráció a statisztikai becslésekben; „Gazdaság és társadalom" Nemzetközi tudományos konferencia Nyugat-magyarországi Egyetem Közgazdaságtudományi Kar Sopron, 2009. november 3. ISBN 978-963-9871-30-4

Szilágyi R.: A nemválaszolás torzításának becslése a mintavételes felmérésekben; „HITEL, VILÁG, STÁDIUM" Tudományos konferencia, Sopron 2010. november 3.

Szilágyi R.: Analysis of nonresponse; International Conference "Economic & Social Challenges and Problems, at The time of Crisis 2009" Faculty of Economy, University of Tirana, Albania, 2009.

Szilágyi R.: Kontár statisztikák; In: Doktoranduszok Fóruma Gazdaságtudományi Kar Szekciókiadványa, Miskolc, Miskolci Egyetem Innovációs és Technológia Transzfer Centrum, 2006, pp. 168-172.

Szilágyi R.: Minőségügyi statisztika; Oktatási segédlet Miskolci Egyetem, 2006.

Szilágyi R.: Mintavételes eljárások; Oktatási segédlet Miskolci Egyetem, 2007.

Szilágyi R.: Pénzbeli ellátások beilleszkedési kölcsönhatásai; In:„Globális és hazai problémák tegnaptól holnapig", VI. Magyar (Jubileumi) Jövőkutatási Konferencia, 30 éves az MTA IX. Osztály Jövőkutatási Bizottsága, Konferenciakötet 2., Budapest, Aristotelész Stúdium Bt., 2007, pp. 91-97, ISBN: 978-963-86670-8-3

Szilágyi R.: Statisztika az üzleti életben In: Informatikai statisztikus és gazdasági tervező felsőfokú képzés II. kötet 6. fejezet HEFOP-3.2.2-P.-2004-10-0011-/1.0 sz. projekt, Miskolc, 2007.

Szilágyi R.: The infiltration of the unfounded statistical information in the forming mechanism of competitiveness In.: XXII. microCAD International Scientific Conference 2009. Miskolc pp. 233-238. ISBN 978-963-661-881-0

Viszonyszámok; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 42-57. ISBN 978-963-06-3135-8

ÁBRAJEGYZÉK

TÁBLÁZATOK JEGYZÉKE

MELLÉKLETEK

1. melléklet

A dolgozatban előforduló legfontosabb fogalmakat és mutatószámokat a következő definíciók alapján használom:

STATISZTIKAI SOKASÁG

A statisztikai megfigyelés tárgyát képező egyedek összessége, halmaza.

STATISZTIKAI ADAT

Valamely statisztikai sokaság tagjainak a száma vagy a sokaság valamilyen másféle számszerű jellemzője.

STATISZTIKAI ISMÉRV

A statisztikai ismérv a statisztikai sokaság egységeit jellemző tulajdonság.

NOMINÁLIS SKÁLA

A névleges (nominális) skála a számok kötetlen hozzárendelését jelenti. Nominális skálát alkalmazunk a területi és minőségi ismérvek szerinti megfigyeléseknél.

A számok között a nagyobb, kisebb relációk, illetve különböző matematikai műveletek nem értelmezhetőek.

ORDINÁLISI SKÁLA

A sorrendi (ordinális) mérési skála a sokaság egyedeinek egy közös tulajdonsága alapján való sorba rendezése. A skálán az egyes egyedek nem feltétlenül egyenlő távolságra helyezkednek el egymástól.

ARÁNYSKÁLA

Az arányskála a legmagasabb mérési szint. Ez nyújtja a legtöbb információt. Zérus pontja természetesen adódik. E skála esetében bármely két érték aránya független a mértékegységtől, és az értékek összege is értelmezhető. E skálán mért számokkal a statisztikai elemzésekhez szükséges összes művelet elvégezhető.

MINTA

Mintának nevezzük a sokaságnak azt a részét, amelyre az adatgyűjtés kiterjed. Véges sokaság esetén a mintát a sokaság elemeiből választjuk ki; ezt az eljárást mintavételnek nevezzük. Végtelen sokaság esetén a mintát a rendelkezésre álló megfigyelési egységek,

illetve kísérleti eredmények alkotják, amelyek ugyancsak a sokaságból származnak. A minta tehát mindig véges számú elemből áll.

NEM MINTAVÉTELI HIBÁK

A nem mintavételi hibák a mintavételen kívüli, egyéb forrásokból erednek, és véletlen vagy nem véletlen jellegűek lehetnek. Okozhatja a probléma helytelen meghatározása vagy megközelítése, a skálák vagy a kérdőívek felépítése, az interjú módszere, valamint az adatfeldolgozás és az elemzés is. A nem mintavételi hibák nemválaszolási és válaszadási hibákból állhatnak.

VÉLETLEN MINTAVÉTELI HIBA

A véletlen mintavételi hiba azért lép fel, mert adott minta nem tükrözi tökéletesen a vizsgált sokaságot. A véletlen mintavételi hiba úgy határozható meg, mint a minta valódi átlagértéke és az alapsokaság valódi átlagértéke közötti eltérés. A mintavételi hiba abból fakad, hogy egy sokasági jellemzőt úgy próbálunk meg becsülni, hogy a sokaságnak csak egy részét tekintjük, nem pedig az egész sokaságot. Tulajdonképpen a mintából kapott becslés és a „valós" érték közötti különbséget értjük alatta, melyet akkor kaptunk volna, ha a sokaság összes egyedét megvizsgáltuk volna.

PARAMÉTER

Paraméternek nevezzük és Θ szimbólummal jelöljük a sokasági eloszlás valamely jellemzőjét. Ilyen jellemző lehet pl. a várható érték, szórás, stb. Adott sokaságra nézve a paraméter állandó. A becslés célja a paraméter értékének közelítő meghatározása a mintából származó megfigyelések alapján.

KONFIDENCIA INTERVALLUM

Konfidencia intervallumnak (megbízhatósági tartománynak) nevezzük a $(\hat{\Theta}_{\hat{a}};\hat{\Theta}_{f})$ értékközt, ha teljesül az $\Pr(\hat{\Theta}_{\hat{a}} < \Theta < \hat{\Theta}_{f}) = \pi$ egyenlőség. $\hat{\Theta}_{\hat{a}}$ az intervallum alsó, $\hat{\Theta}_{f}$ pedig a felső korlátja. Az intervallumbecslés célja valamely adott π megbízhatósági szinthez tartozó konfidencia intervallum számítása.

STANDARD HIBA

A mintából származó megfigyelések bizonyos függvényeinek (pl átlagainak) szórását nevezzük standard hibának. Jelentése: az összes lehetséges módon választható minták átlagai átlagosan mennyivel térnek el a sokasági átlagtól (várható értéktől).

2. melléklet

A változók ismertetése

VÁLTOZÓ KÓDJA	VÁLTOZÓ NEVE	LEHETSÉGES KÓDOLÁSI ÉRTÉKEK + MAGYARÁZAT	FORMÁTUM
HA08	Régió	NUTS1 használata 00 – NUTS-ban nem szereplő országok 1 99 – nem meghatározott	INT2
A NUTS1 besorolási osztály egy számjegyet és betűket használ abban az esetben, ha a kód nagyobb, mint 9. A HKF-nél ezt a rendszert két számjegyű kódolás váltja fel.			
HA09	Népsűrűség	1-sűrűn lakott terület (legalább 500 lakos/km²) 2-közepesen sűrűn lakott terület (100 és 499 lakos/km²) 3- gyéren lakott terület (kevesebb, mint 100 lakos/km²) 9- nem meghatározott	INT 1
Magyarázat: 1: sűrűn lakott ter.→ egymással összefüggő helyi területek, amelyek népsűrűsége egyenként túllépi az 500 lakos/km² -t és ahol a terület összlakossága legalább 50.000 fő. 2: közepesen sűrűn lakott ter. → olyan egymással összefüggő területek halmaza, amely nem tartozik sűrűn lakott területhez, és amelyben az egyes területek népsűrűsége meghaladja a 100 lakos/km²-t de az egész halmaz népessége nem haladja meg az 50.000 lakos/km²-t és nem határos sűrűn lakott területtel. 3: gyéren lakott ter.→ olyan területek halmaza, amelyek nem tartoznak a sűrűn vagy a közepesen lakott területek közé.			
HA10	Minta súly	Az EUROSTAT által alkalmazott minta súlyozása megegyezik a tagállmok által számított súlyozással a felmérések eredményeit bemutató nemzeti kiadványokban.	DEC 6.2
HB05	Háztartás mérete	01-xx	INT 2
Magyarázat: HB05 = háztartás tagjainak száma A SILC projektben megfogalmazódik, hogy melyek azok a személyek akik a háztartás tagjainak tekintem. Egy háztartás tagja az a személy, aki: 1. többnyire ottlakó, a háztartás rokona 2. többnyire ottlakó, nem rokona a háztartásnak 3. bennlakó diák, albérlő, bérlő 4. vendég 5. bennlakó háztartási alkalmazott, au-pair 6. ottlakó, rövid időre távollévő (nyaralás, munka, tanulás miatt) 7. háztartás gyermekei tanulás miatt távol a háztartástól 8. hosszabb ideig távollévő, háztartási kötelékkel rendelkező (pl. távmunka) 9. átmenetileg távollévő, háztartási kötelékkel rendelkező pl. kórházban, szanatóriumban, vagy más intézményben tartózkodó személy, aki a, hozzájárul a háztartás kiadásaihoz. 3., 4., 5. kategóriák esetén b, akinek jelenleg nincs máshol bejelentett lakcíme vagy a tervezett vagy aktuális tartózkodási ideje 6 vagy több hónap. 6. kategória esetén b, akinek jelenleg nincs máshol bejelentett lakcíme és a tervezett vagy aktuális tartózkodási ideje kevesebb, mint 6 hónap. 7. és 8. kategóriák esetén b, jelenleg nincs máshol bejelentett lakcíme, a háztartás valamely tagjának gyermeke, élettársa, és aki folyamatos kapcsolatban van háztartással, és az adott lakcímet fő tartózkodási helyének tekinti. 9. kategória esetén b, akinek pénzügyi kapcsolata van a háztartással, és az aktuális vagy tervezett távolléte a háztartástól kevesebb, mint 6 hónap. A háztartási *kiadások megosztása* magábafoglalja a kiadásokhoz való hozzájárulást és a kiadásból való haszonszerzést is (pl. gyermekek). Ha a kiadások nincsenek megosztva, akkor az adott személy egy másik háztartást testesít meg, ugyanazon a lakcímen.			

Többnyire ottlakó a háztartásnak az a tagja, aki a napja nagyrészét ott töltötte az elmúlt 6 hónapot tekintve. Azok a személyek, akik új háztartást kezdenek vagy már meglévőhöz csatlakoznak általában új tartózkodási helyük háztartásába tartoznak. Azok, akik máshol kezdenek élni, elköltöznek nem tartoznak az adott háztartáshoz.

Az a személy, aki *határozatlan időre* vagy 6 hónapra vagy azt meghaladó időtartamra költözött az adott háztartásba a háztartás tagjának számít, még abban az esetben is, ha még nincs 6 hónapja, hogy odaköltözött és az eltelt idő alatt ideje nagyrészét máshol töltötte. Azok, akik elköltöztek a háztartásból 6 vagy annál több időre nem tekinthetők a háztartás tagjainak.

Ha azok a személyek, akik *ideiglenesen* távol vannak a háztartástól és privát szálláson tarózkodnak, akkor az, hogy ők az adott háztartás tagjai attól függ, hogy mennyi ideig vannak távol. Kivéve azokat a személyeket, akik szorosan kötődnek a háztartáshoz. Ők a távollévő időtől függetlenül a háztartás tagjainak tekinthetők, abban az esetben, ha nem tagjai más háztartásnak.

HB07.3	Háztartás típusa (egyszerűsítve)	1 → egy 65 évesnél idősebb korú személy vagy pár (HB07.1=01,05) 2 → más háztartás, egy személy vagy pár gyermek nélkül (HB07.1=02,03,06) 3 → 18 évnél idősebb gyermekkel rendelkező pár vagy egyedülálló szülő (HB07.1= 04,07,08,09) 4 → egyéb (HB07.1=10,11,12)	INT1

Az új háztartás típus csupán két kritériumot használ: felnőttek száma és eltartott gyermekek száma. Eltartott gyermek:

- ✓ 18 évesnél fiatalabb,
- ✓ 18 éves vagy idősebb illetve 24 éves vagy fiatalabb, nem dolgozó és nem munkanélküli.

Számítás szabálya:
Ha (HB07.1=01 vagy 05) → HB07.3=1
Ha (HB07.1=03 vagy 03 vagy 06) → HB07.3=2
Ha (HB07.1=04 vagy 07 vagy 08 vagy 09) → HB07.3=3
Ha (HB07.1=10 vagy 11 vagy 12) → HB07.3=4

HC03	Családfő neme	1 – férfi 2 – nő 9 – nem meghatározott	INT1
HC04	Családfő életkora (betöltött évek száma)	00 ⇓ 98 (98 vagy több éves) 99 – nem meghatározott	INT2
HC12	Családfő* aktivitási státusza	→ gazdaságilag aktív 1 foglalkoztatott 2 foglalkoztatott, de átmenetileg kieső 3 munkanélküli → gazdaságilag inaktív 4 nyugdíjas 5 tanuló vagy katonai szolgálatot töltő 6 gazdaságilag nem aktív vagy háztartásbeli 7 munkaképtelen 8 nem foglalkoztatható (nem munkaképes korú) 9 nem meghatározott	INT 1

* Családfő (reference person) általában a legmagasabb jövedelemmel rendelkező személy a háztartásban.
Magyarázat:
A **foglalkoztatotti** stádiumban lévő személyek kódolási értéke 1 vagy 2. Az átmenetileg nem dolgozó személyt dolgozónak tekintjük, ha hivatalosan kapcsolatban áll alkalmazójával.
A **munkanélküli** kategóriát (kód 3) nehéz megállapítani. A munkaerő felmérés szerinti fogalommeghatározás által – amely megfelel az ILO által javasolt fogalomnak - munkanélküli az a személy, aki bizonyos kort meghalad a kérdezési periódusban és:

- ✓ éppen munka nélkül van, pl. nem áll sem fizetett, sem fizetés nélküli alkalmazásban.
- ✓ munkaképes.

✓ munkát keres, pl. lépéseket tett annak érdekében, hogy fiezetett vagy fizetés nélküli alkalmazást szerezzen. Mindazokat a személyeket, akik sem az „alkalmazotti" sem a „munkanélküli" kategóriába nem tartoznak inaktívnak tekintjük. Az inaktívak az alábbi kategóriákba sorolhatók: ✓ nyugdíjasok → azok, akik nyugdíjas korukat betöltötték ✓ tanulók; katonai szolgálatukat töltők; otthonhoz kötött személyek, akik gazdasági aktivitást nem végeznek 7-es kód → azok a munkaképtelen személyek, akik fizikai vagy más akaratlan okból kifolyólag nem képesek munkát végezni.			
HC18	Családfő foglalkozása (ISCO 1988(COM))	01 törvényhozó, idősebb tisztviselő, menedzser 02 diplomás 03 műszaki szakember 04 irodai dolgozó 05 szolgáltatásban dolgozó, bolti vagy kereskedelmi dolgozó 06 szakképzett mezőgazdasági vagy halászati dolgozó 07 kézműves vagy mesterember 08 gépész, műszerész 09 alapfokú végzettséggel rendelkező 00 katona 88 nem alkalmazható (nem munkaképes korú) 99 nem meghatározott	INT 2
HC23	Családfő szocio-gazdasági helyzete	→ Magánszektor 01 fizikai dolgozó kivéve mezőgazdaságban dolgozó 02 nem fizikai dolgozó kivéve mezőgazdaságban dolgozó → Közszektor 03 fizikai dolgozó kivéve mezőgazdaságban dolgozó 04 nem fizikai dolgozó kivéve mezőgazdaságban dolgozó → Egyéb 05 önfoglalkoztató kivéve mezőgazdasági vállalkozó 06 mezőgazdasági dolgozó 07 munkanélküli 08 nyugdíjas 09 tanuló vagy kötelező katonai szolgálatot töltő 10 háztartásbeli vagy gazdasági aktivitást nem végző 11 munkaképtelen 88 nem foglalkoztatható (nem munkaképes korú) 99 nem meghatározott	INT 2
HD06	Szobák száma (a fő lakó/tartózkodási helyen)	1-7 8+ (8 vagy több szoba) 9 → nem meghatározott	INT1
Az ENSZ javaslata alapján „normál hálószobák, étkezők, nappalik, lakható pincehelyiségek és padlások, a személyzet szobái, konyhák vagy más belakásra alkalmas elkülönülő helyiségek számítanak szobának. A konyhasarok (4m²-nél kisebb konyha), folyosók, verandák, előszobák, különböző hasznossági szobák (pl. kazánhelyiség, mosószoba), előtér nem számítanak szobának, még a fürdőszobák és a WC helyiségek sem (még ha 4m²-t meg is haladják).			
HD07	Hasznos lakóterület m²-ben (fő lakóhely)	001 ⇓ 998 999 → nem meghatározott	INT3

A hasznos lakóterület fogalma szintén ENSZ javaslata által értelmezhető. „Hasznos alapterület a lakóhelynek a külső falak által határolt belső részének alapterülete, kivéve a nem-lakható pincehelyiségeket, padlásokat és a társasházak közös helyiségeit. Általánosságban a HD06 és a HD07 változók erős összefüggésben vannak egymással és a regressziós modellekben nem használhatóak együtt a bérelt lakások élvezeti árának megállapításakor. Viszont vannak országok, amelyeknek el kell hagyniuk ezt az összefüggést a két változó egyikének a HD07/HD06 hányados (szobák átlagos nagysága) helyettesítésével.			
HD14.02	Személygépjárművek száma	0 ⇓ 3 4+ (4 vagy több személygépjármű) 9 → nem meghatározott	INT1
Beleértve a szolgálati személygépkocsikat is.			
HD14.04	Televízió készülékek száma	0 ⇓ 3 4+ (4 vagy több készülék) 9 → nem meghatározott	INT1
HD14.14	Használatban lévő mobiltelefon készülékek száma (hálózati hozzáféréssel)	0 ⇓ 3 4+ (4 vagy több készülék) 9 → nem meghatározott	INT1
HE00	Összes fogyasztási kiadás		INT14
Annak érdekében, hogy teljesebb képet kapjunk a fogyasztási szerkezetről, különös tekintettel az élelmiszer szektorra és a juttatások területére, a fogyasztási kiadást három kategóriába soroljuk. A kategóriákat betű jelöli, amelyet az eredeti változó kódja/száma mellé írunk. „A”: pézbeli formában megvalósiló kiadás „B”: nem pénzbeli formában megvalósuló kiadás „C”: teljes összeg			
HH09.5	Pénzbeli nettó jövedelem (összes forrásból származó összes pézben kifizetett jövedelem mínusz jövedelemadó)		INT 14
HH09.9	Nettó jövedelem (összes forrásból származó összes jövedelem, beleértve a nem pézbeni elemeket is mínusz jövedelemadó)	HH09.9=HH09.5+HH01.2+HH02.3+HH03.2	INT 14
Magyarázat: HH01.2 →foglalkoztatásból származó jövedelem (fizetés, munkabér). Fizetett foglalkoztatás keretében nyújtott juttatások. (kivéve felszámított bérleti díj (HH02.3) HH02.3 →nem fizetett tevékenységből származó jövedelem. Saját termelés magánfogyasztásra (kert, vállalkozás). (kivéve felszámított bérleti díj (HH02.3) HH03.2 →felszámított bérleti díj: a tulajdonos által felszámított bérleti díj és az ingyenes bérlet. Megegyezés szerint a bérleti díjak számos típusa beletartozik ebbe a változóba – akár a tulajdonost, akár az ingyenesen bérlő személyt illeti.			

A háztartás tagjaihoz tartozó magyarázat forrása:
Household Budget Survey in the EU – Methodology and recommandations for harmonisation 2003, 17.-20.o.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-BF-03-003/DE/KS-BF-03-003-DE.PDF (letöltve 2010. 09. 22.)

3. melléklet

Fogyasztási kiadások becslési jellemzői

Minták	Estimate	Standard Error	Coefficient of Variation	Design Effect	Square Root Design Effect
MR_FOGY_30	1664195	27708,782	0,017	0,03	0,173
MR_FOGY_900	1728151,5	8721,276	0,005	0,073	0,27
MR_FOGY_150	1788553,2	32470,877	0,018	0,116	0,341
MR_JÖV_900	1719409,9	22060,666	0,013	0,465	0,682
REG_AUTO_AKTIV	1858192,6	61634,431	0,033	0,495	0,703
REG_AUTO_CSALÁD900	1729945,8	22933,482	0,013	0,51	0,714
REG_AUTO_SZOBA900	1722865,9	24077,846	0,014	0,522	0,723
REG_TV_CSALÁD	1691313,9	58894,337	0,035	0,538	0,734
REG_AUTO_SZOBA	1849614	71925,612	0,039	0,541	0,735
REG_AUTO_CSALÁD	1692150,5	61070,172	0,036	0,545	0,738
MR_JÖV_150	1780605,9	64679,676	0,036	0,548	0,741
REG_AUTO_AKTIV900	1751275	26607,137	0,015	0,613	0,783
MR_JÖV_30	1941405,9	161483,44	0,083	0,627	0,792
REG_TV_CSALÁD900	1756807,8	28620,671	0,016	0,647	0,804
AUTO900	1754322,9	28659,533	0,016	0,661	0,813
REG_TV_SZOBA	1664096,3	70157,395	0,042	0,663	0,814
REG_TV_AKTIV900	1765400,5	29023,234	0,016	0,724	0,851
AUTO	1814498,2	80549,058	0,044	0,725	0,851
REG_TV_AKTIV	1723966,4	82568,965	0,048	0,725	0,851
REG_TV_SZOBA900	1684680,6	28790,592	0,017	0,746	0,864
CSALÁD900	1684855,1	28819,246	0,017	0,781	0,883
HÁZT TIP	1831834	89015,865	0,049	0,788	0,888
AKTIV	1646391,9	66705,907	0,041	0,791	0,889
AKTIV900	1741873,6	30399,619	0,017	0,819	0,905
MR_NM_150	1846676,5	85702,302	0,046	0,823	0,907
CSALÁD	1753421,3	74031,385	0,042	0,86	0,927
HAZT TIP900	1737711,9	32798,864	0,019	0,863	0,929
MR_KOR_900	1785689,2	31197,607	0,017	0,868	0,931
MR_KOR_150	1678424,9	76789,423	0,046	0,894	0,946
TV900	1701398,4	31900,535	0,019	0,901	0,949
MR_NM_900	1751526	32022,982	0,018	0,904	0,951
TV	1759092,2	73949,76	0,042	0,916	0,957
REG_DENS	1745440,1	90463,136	0,052	0,923	0,961
REG_DENS900	1725342,3	33430,437	0,019	0,949	0,974
SŰRŰSÉG900	1741054,2	34156,232	0,02	0,956	0,978
NEM900	1769166,2	36774,062	0,021	0,975	0,987
NEM	1748126,4	94921,834	0,054	0,981	0,991
REG_HÁZT_HKF900	1723271,5	31981,133	0,019	0,981	0,99
MR_KOR_30	1791503,3	170578,48	0,095	0,991	0,995
REG_HAZT_HKF	1905964,9	92847,015	0,049	0,997	0,999
EV1150	1831324,8	89189,577	0,049	1	1
EV2150	1793904,9	98938,941	0,055	1	1
EV9150	1736354,6	90229,132	0,052	1	1
EV1900	1711361,9	32769,49	0,019	1	1
EV2900	1758497,9	34573,749	0,02	1	1
EV9900	1764317,8	34991,858	0,02	1	1
EV130	1825972	225116,76	0,123	1	1
EV230	1779368,9	241839,35	0,136	1	1
EV930	1701792,4	240515,77	0,141	1	1
SŰRŰSÉG	1777688,5	79938,146	0,045	1,001	1,001
MR_NM_30	1906348,8	189818,88	0,1	1,139	1,067
REG_EGY	1806059,3	103066,66	0,057	1,225	1,107
REG_EGY900	1686210,7	35056,391	0,021	1,292	1,136

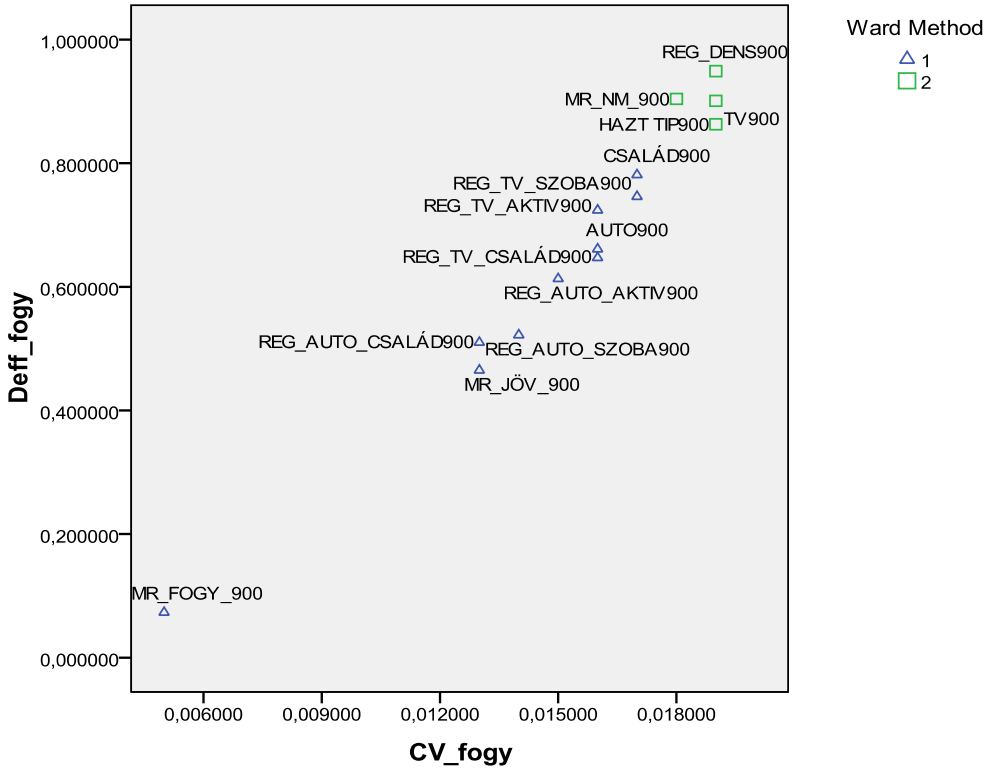
Háztartások átlagos létszáma

Minták	Estimate	Standard Error	Coefficient of Variation	Design Effect	Square Root Design Effect
AKTIV	2,64	0,093	0,035	0,897	0,947
AKTIV900	2,7	0,038	0,014	0,813	0,902
AUTO	2,61	0,107	0,041	0,948	0,973
AUTO900	2,74	0,042	0,015	0,887	0,942
CSALÁD	2,69	0	0	0	0
CSALÁD900	2,71	0	0	0	0
EV1150	2,91	0,109	0,037	1	1
EV130	2,93	0,279	0,095	1	1
EV1900	2,67	0,043	0,016	1	1
EV2150	2,87	0,13	0,045	1	1
EV230	2,63	0,232	0,088	1	1
EV2900	2,72	0,044	0,016	1	1
EV9150	2,7	0,108	0,04	1	1
EV930	2,73	0,224	0,082	1	1
EV9900	2,69	0,042	0,015	1	1
HÁZT TIP	2,83	0,08	0,028	0,437	0,661
HAZT TIP900	2,75	0,031	0,011	0,497	0,705
MR_FOGY_150	2,83	0,089	0,032	0,668	0,817
MR_FOGY_30	2,53	0,166	0,066	0,583	0,763
MR_FOGY_900	2,71	0,039	0,014	0,769	0,877
MR_JŐV_150	2,79	0,1	0,036	0,757	0,87
MR_JÖV_30	2,4	0,216	0,09	0,793	0,891
MR_JÖV_900	2,69	0,037	0,014	0,728	0,853
MR_KOR_150	2,64	0,089	0,034	0,763	0,874
MR_KOR_30	2,83	0,168	0,059	0,843	0,918
MR_KOR_900	2,73	0,037	0,014	0,763	0,873
MR_NM_150	2,65	0,101	0,038	0,91	0,954
MR_NM_30	2,92	0,148	0,051	0,715	0,845
MR_NM_900	2,73	0,041	0,015	0,913	0,955
NEM	2,78	0,115	0,041	0,929	0,964
NEM900	2,66	0,042	0,016	0,925	0,962
REG_AUTO_AKTIV	2,64	0,082	0,031	0,618	0,786
REG_AUTO_AKTIV900	2,75	0,036	0,013	0,68	0,825
REG_AUTO_CSALÁD	2,55	0	0	0	0
REG_AUTO_CSALÁD900	2,67	0	0	0	0
REG_AUTO_SZOBA	2,84	0,118	0,042	0,781	0,884
REG_AUTO_SZOBA900	2,71	0,04	0,015	0,856	0,925
REG_DENS	2,64	0,109	0,041	0,952	0,976
REG_DENS900	2,65	0,043	0,016	0,985	0,992
REG_EGY	2,61	0,119	0,046	1,205	1,098
REG_EGY900	2,7	0,049	0,018	1,277	1,13
REG_HAZT_HKF	2,86	0,12	0,042	0,974	0,987
REG_HÁZT_HKF900	2,63	0,043	0,016	0,996	0,998
REG_TV_AKTIV	2,74	0,104	0,038	0,774	0,88
REG_TV_AKTIV900	2,7	0,035	0,013	0,676	0,822
REG_TV_CSALÁD	2,54	0	0	0	0
REG_TV_CSALÁD900	2,66	0	0	0	0
REG_TV_SZOBA	2,57	0,08	0,031	0,533	0,73
REG_TV_SZOBA900	2,69	0,038	0,014	0,783	0,885
SŰRŰSÉG	2,66	0,102	0,038	0,987	0,994
SŰRŰSÉG900	2,66	0,042	0,016	0,989	0,994
TV	2,74	0,099	0,036	0,905	0,951
TV900	2,67	0,04	0,015	0,892	0,944

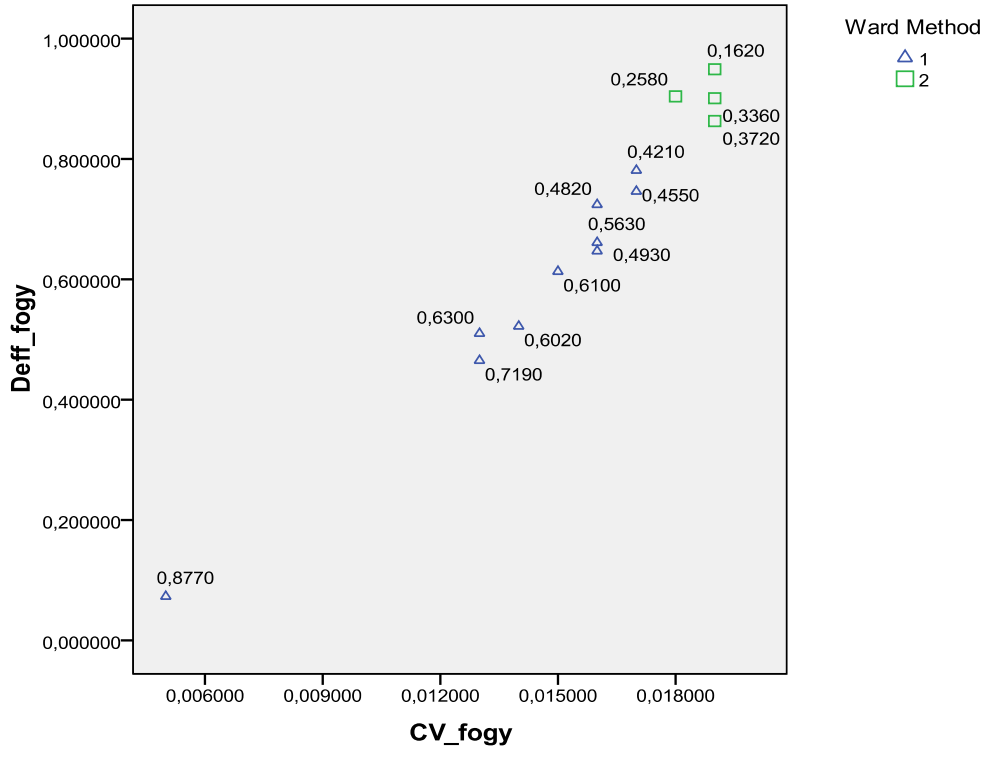
Háztartások egy főre jutó kiadása

Minták	Ratio Estimate	Standard Error	Coefficient of Variation	Design Effect	Square Root Design Effect
AKTIV	622541,333	27499,537	0,044	0,948	0,974
AKTIV900	644472,0132	11952,38	0,018546	0,9575309	0,978535057
AUTO	694415,937	33891,075	0,049	0,892	0,944
AUTO900	639979,2059	11795,695	0,0184314	0,896761	0,946974663
CSALÁD	651695,59	27515,308	0,042	0,694	0,833
CSALÁD900	621671,6364	10633,62	0,0171049	0,8170417	0,903903607
EV1150	630042,9335	32160,315	0,0510446	1	1
EV130	622490,4659	72616,954	0,1166555	1	1
EV1900	641493,4061	12172,942	0,0189759	1	1
EV2150	625780,786	31747,273	0,0507323	1	1
EV230	675709,7089	80077,329	0,1185085	1	1
EV2900	647299,8294	12616,338	0,0194907	1	1
EV9150	643094,3086	32583,233	0,0506663	1	1
EV930	622606,9878	85909,669	0,1379838	1	1
EV9900	656693,9541	12190,7	0,0185637	1	1
HÁZT TIP	647267,989	35276,361	0,055	0,977	0,988
HAZT TIP900	633031,4118	12545,378	0,0198179	0,9478606	0,973581348
MR_FOGY_150	632728,386	21991,362	0,035	0,497	0,705
MR_FOGY_30	656888,547	44741,402	0,068	0,516	0,718
MR_FOGY_900	637939,195	9550,349	0,015	0,591	0,769
MR_JÖV_150	638967,328	23621,623	0,037	0,677	0,823
MR_JÖV_30	808876,967	71853,541	0,089	0,663	0,814
MR_JÖV_900	640232,407	9683,85	0,015	0,62	0,788
MR_KOR_150	635888,972	22504,178	0,035	0,574	0,758
MR_KOR_30	633175,231	43888,449	0,069	0,377	0,614
MR_KOR_900	653302,67	9211,275	0,014	0,591	0,769
MR_NM_150	696241,24	25634,933	0,037	0,576	0,759
MR_NM_30	653749,649	40251,591	0,062	0,456	0,675
MR_NM_900	640654,122	9380,425	0,015	0,599	0,774
NEM	628305,755	33668,342	0,054	1,003	1,002
NEM900	664280,792	13499,048	0,0203213	0,9976496	0,998824112
REG_AUTO_AKTIV	704024,4751	27565,111	0,0391536	0,7967781	0,892624276
REG_AUTO_AKTIV900	637277,3886	10557,078	0,0165659	0,8316101	0,911926602
REG_AUTO_CSALÁD	663835,7365	23958,012	0,0360903	0,6790517	0,82404594
REG_AUTO_CSALÁD900	648023,6519	8590,6962	0,0132568	0,5131	0,71631
REG_AUTO_SZOBA	651274,9258	27914,419	0,0428612	0,6131878	0,783063066
REG_AUTO_SZOBA900	635458,5248	11283,545	0,0177565	0,8033367	0,896290527
REG_DENS	661783,2965	33897,947	0,0512221	0,8629539	0,928953119
REG_DENS900	651266,2885	11644,08	0,0178791	0,8784919	0,937279004
REG_EGY	691989,21	35455,347	0,051	1,155	1,075
REG_EGY900	624451,9745	12783,353	0,0204713	1,2610817	1,122978923
REG_HAZT_HKF	665293,903	31442,118	0,047	0,952	0,976
REG_HÁZT_HKF900	655997,3233	11647,778	0,0177558	0,9512634	0,975327354
REG_TV_AKTIV	628492,1043	34436,455	0,0547922	0,9482055	0,973758429
REG_TV_AKTIV900	652871,373	11106,617	0,017012	0,8855213	0,941021403
REG_TV_CSALÁD	665272,3498	23165,88	0,0348216	0,5378063	0,733352773
REG_TV_CSALÁD900	661318,077	10773,727	0,0162913	0,6783547	0,823622908
REG_TV_SZOBA	648531,6408	27088,184	0,0417685	0,6800618	0,824658612
REG_TV_SZOBA900	625668,7837	11291,499	0,0180471	0,8407939	0,91694816
SŰRŰSÉG	667902,126	31458,862	0,047	0,97	0,985
SŰRŰSÉG900	653702,7607	12749,163	0,019503	0,9218983	0,960155342
TV	642252,469	27694,998	0,043	1,013	1,007
TV900	637557,965	11933,139	0,0187169	1,0013826	1,00069105

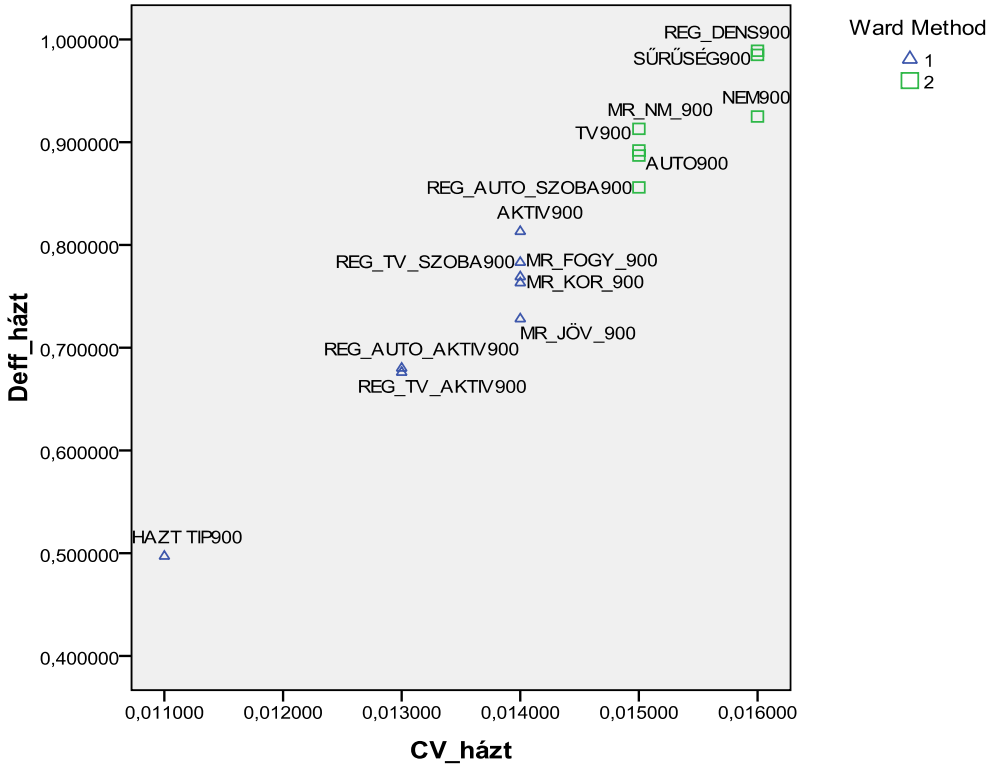
A 900 elemű minták rangsorolása a háztartások teljes fogyasztásának becslésére



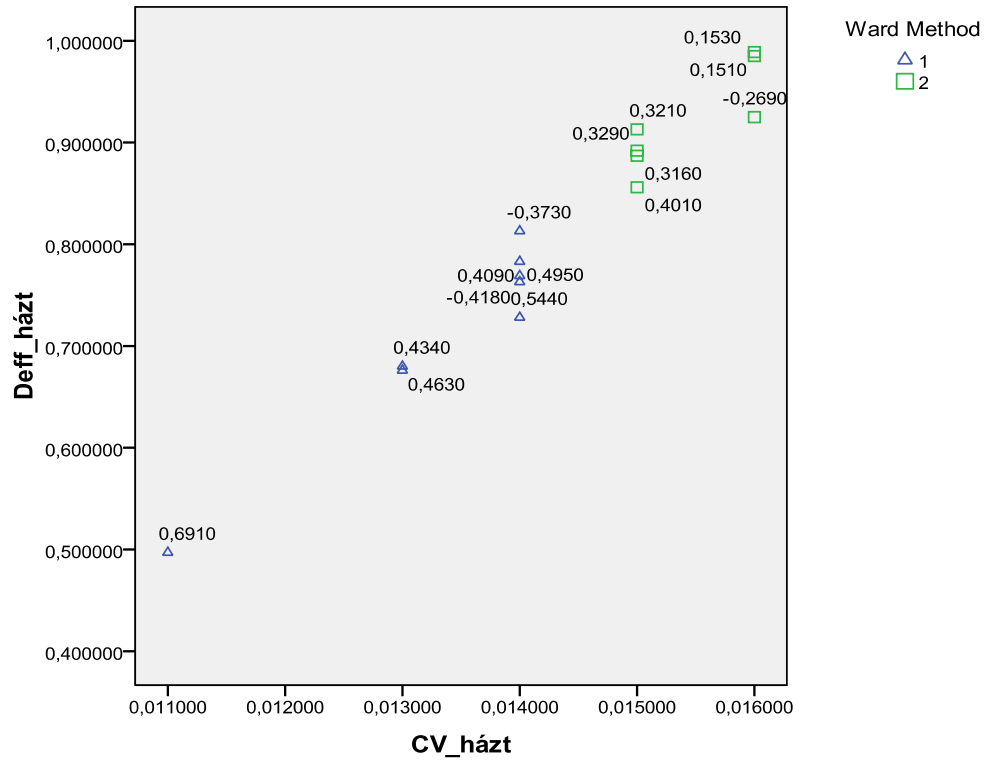
A 900 elemű minták rangsorolása a háztartások teljes fogyasztásának becslésére, valamint a rétegképző ismérv és a teljes fogyasztás kapcsolatát jellemző korrelációs (illetve többször korrelációs) együtthatók értékei



A 900 elemű minták rangsorolása a háztartások átlagos nagyságának becslésére



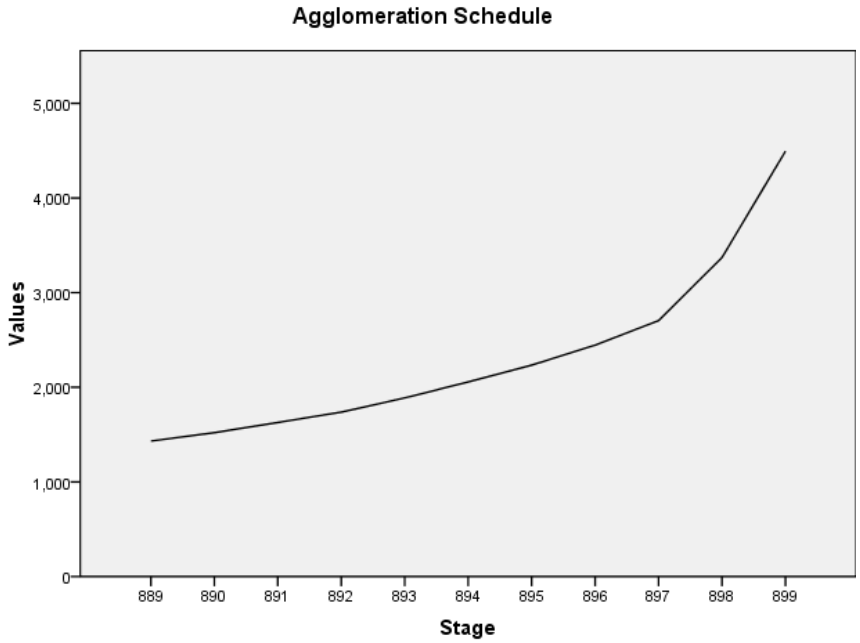
A 900 elemű minták rangsorolása a háztartások átlagos méretének becslésére, valamint a rétegképző ismérv és a háztartás méret kapcsolatát jellemző korrelációs (illetve többször korrelációs) együtthatók értékei



Az 5. fejezetmellékszámításai

5.2.1. alfejezet

Cluster
Nearest Neighbor



CROSSTABS /TABLES=CLU2_1 BY felső30 /FORMAT=AVALUE TABLES /CELLS=COUNT COLUMN /COUNT ROUND CELL.

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Nearest Neighbor * 90 from the first 270 cases (csökkenő)	900	100,0%	0	,0%	900	100,0%

Nearest Neighbor * 90 from the first 270 cases (csökkenő) Crosstabulation

			90 from the first 270 cases (csökkenő)		
			0	1	Total
Nearest Neighbor	1	Count	531	84	615
		% within 90 from the first 270 cases (csökkenő)	65,6%	93,3%	68,3%
	2	Count	279	6	285
		% within 90 from the first 270 cases (csökkenő)	34,4%	6,7%	31,7%
	Total	Count	810	90	900
		% within 90 from the first 270 cases (csökkenő)	100,0%	100,0%	100,0%

Correlations

Correlations

		Current activity status of the reference person	Number of cars	Number of televisions
Current activity status of the reference person	Pearson Correlation	1	-,300**	-,190**
	Sig. (2-tailed)		,000	,000
	N	900	900	900
Number of cars	Pearson Correlation	-,300**	1	,287**
	Sig. (2-tailed)	,000		,000
	N	900	900	900
Number of televisions	Pearson Correlation	-,190**	,287**	1
	Sig. (2-tailed)	,000	,000	
	N	900	900	900
Useful living area in m? (principal residence)	Pearson Correlation	-,044	,260**	,289**
	Sig. (2-tailed)	,188	,000	,000
	N	900	900	900
Household size	Pearson Correlation	-,344**	,291**	,348**
	Sig. (2-tailed)	,000	,000	,000
	N	900	900	900

**. Correlation is significant at the 0.01 level (2-tailed).

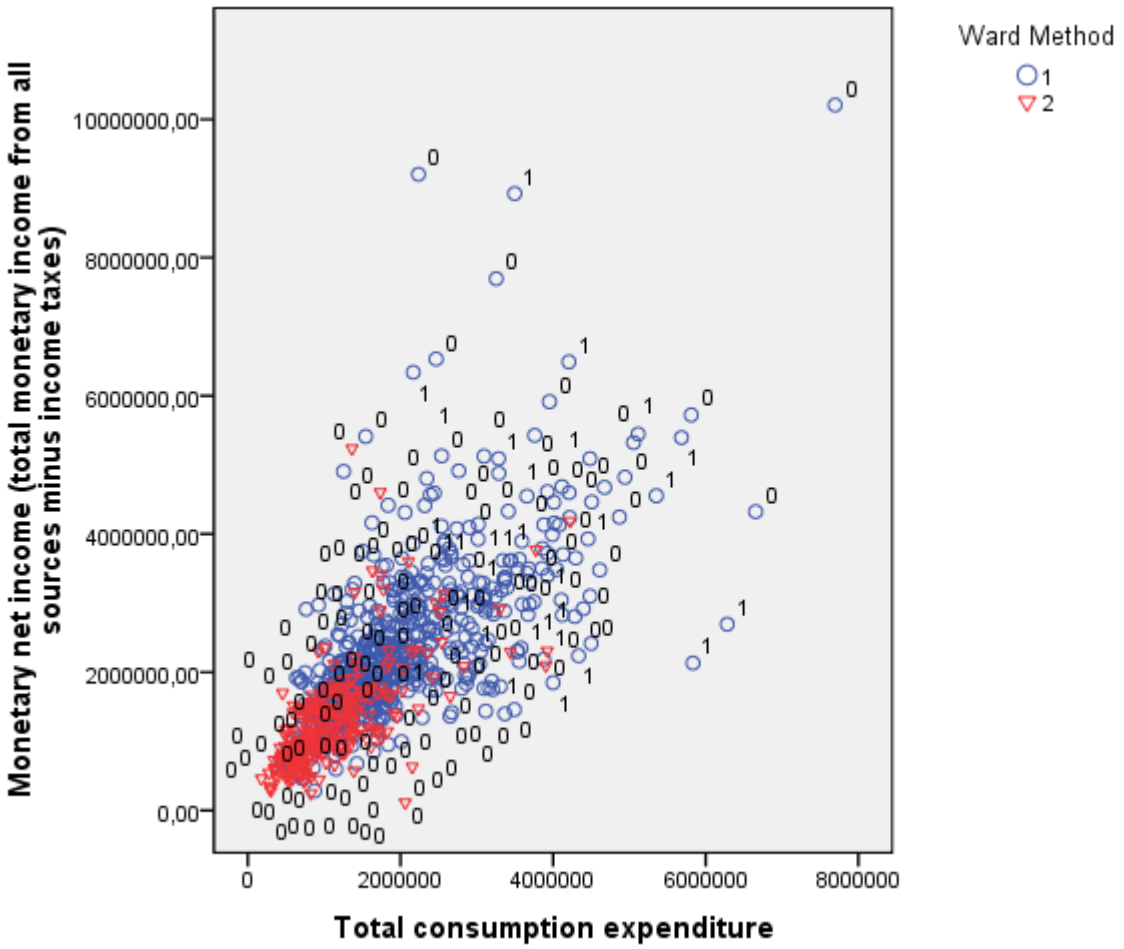
Correlations

		Useful living area in m? (principal residence)	Household size
Current activity status of the reference person	Pearson Correlation	-,044	-,344**
	Sig. (2-tailed)	,188	,000
	N	900	900
Number of cars	Pearson Correlation	,260**	,291**
	Sig. (2-tailed)	,000	,000
	N	900	900
Number of televisions	Pearson Correlation	,289**	,348**
	Sig. (2-tailed)	,000	,000
	N	900	900

Useful living area in m? (principal residence)	Pearson Correlation	1	,279**
	Sig. (2-tailed)		,000
	N	900	900
Household size	Pearson Correlation	,279**	1
	Sig. (2-tailed)	,000	
	N	900	900

**. Correlation is significant at the 0.01 level (2-tailed).

Graph



Complex Samples: Descriptives

Univariate Statistics

		Estimate	Standard Error	Coefficient of Variation	Design Effect
Mean	Total consumption expenditure	1584815,61	7964,054	,005	,066

Univariate Statistics

		Square Root Design Effect	Population Size
Mean	Total consumption expenditure	,257	8152,367

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	
	Statistic	Statistic	Statistic	Statistic	Std. Error
Total consumption expenditure	810	177292	7695496	1584876,04	32611,351
Valid N (listwise)	810				

Descriptive Statistics

	Std. Deviation
	Statistic
Total consumption expenditure	928135,319

5.2.2. alfejezet

Group Statistics

NR_10pc		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
0	Household size	2,6259	1,39647	810	810,000
	jöv	5,1284	2,69657	810	810,000
	Number of cars	,4753	,56254	810	810,000
	Current activity status of the reference person	2,1025	1,48103	810	810,000
	Useful living area in m? (principal residence)	77,5679	32,03985	810	810,000
	Number of televisions	1,4037	,65680	810	810,000
	Marital status of the reference person	1,7025	1,40338	810	810,000
	Age (in completed years) of reference person	49,5864	15,65064	810	810,000
	Sex of reference person	1,4432	,49707	810	810,000
1	Household size	3,4556	1,15302	90	90,000
	jöv	8,9333	1,45957	90	90,000
	Number of cars	1,2000	,56489	90	90,000
	Current activity status of the reference person	1,2889	,87723	90	90,000

	Useful living area in m? (principal residence)	93,2667	40,77171	90	90,000
	Number of televisions	1,9111	,89499	90	90,000
	Marital status of the reference person	1,2222	,96893	90	90,000
	Age (in completed years) of reference person	44,7333	10,07812	90	90,000
	Sex of reference person	1,3222	,46995	90	90,000
Total	Household size	2,7089	1,39590	900	900,000
	jöv	5,5089	2,83882	900	900,000
	Number of cars	,5478	,60306	900	900,000
	Current activity status of the reference person	2,0211	1,45247	900	900,000
	Useful living area in m? (principal residence)	79,1378	33,32502	900	900,000
	Number of televisions	1,4544	,70049	900	900,000
	Marital status of the reference person	1,6544	1,37332	900	900,000
	Age (in completed years) of reference person	49,1011	15,25117	900	900,000
	Sex of reference person	1,4311	,49551	900	900,000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Household size	,968	29,520	1	898	,000
jöv	,838	173,423	1	898	,000
Number of cars	,870	134,316	1	898	,000
Current activity status of the reference person	,972	26,124	1	898	,000
Useful living area in m? (principal residence)	,980	18,322	1	898	,000
Number of televisions	,953	44,559	1	898	,000
Marital status of the reference person	,989	10,005	1	898	,002
Age (in completed years) of reference person	,991	8,268	1	898	,004
Sex of reference person	,995	4,850	1	898	,028

Pooled Within-Groups Matrices

		Household size	jöv	Number of cars	Current activity status of the reference person	Useful living area in m? (principal residence)	Number of televisions	Marital status of the reference person	Age (in completed years) of reference person	Sex of reference person
Correlation	Household size	1,000	,499	,247	-,324	,260	,322	-,291	-,430	-,282
	jöv	,499	1,000	,421	-,449	,203	,326	-,345	-,375	-,251
	Number of cars	,247	,421	1,000	-,261	,226	,229	-,231	-,180	-,158
	Current activity status of the reference person	-,324	-,449	-,261	1,000	-,021	-,159	,242	,632	,151
	Useful living area in m? (principal residence)	,260	,203	,226	-,021	1,000	,268	-,085	,064	-,092
	Number of televisions	,322	,326	,229	-,159	,268	1,000	-,071	-,102	-,064
	Marital status of the reference person	-,291	-,345	-,231	,242	-,085	-,071	1,000	,351	,391
	Age (in completed years) of reference person	-,430	-,375	-,180	,632	,064	-,102	,351	1,000	,172
	Sex of reference person	-,282	-,251	-,158	,151	-,092	-,064	,391	,172	1,000

Log Determinants

NR_10pc	Rank	Log Determinant
0	3	5,955
1	3	4,164
Pooled within-groups	3	5,855

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		70,203
F	Approx.	11,556
	df1	6
	df2	134503,896
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Variables Entered/Removed[a,b,c,d]

		Min. D Squared					
				Exact F			
Step	Entered	Statistic	Between Groups	Statistic	df1	df2	Sig.
1	jöv	2,141	0 and 1	173,423	1	898,000	2,436E-36
2	Number of cars	2,690	0 and 1	108,827	2	897,000	4,842E-43
3	Age (in completed years) of reference person	2,762	0 and 1	74,400	3	896,000	5,673E-43

At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

a. Maximum number of steps is 18.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Variables in the Analysis

Step		Tolerance	F to Remove	Min. D Squared	Between Groups
1	jöv	1,000	173,423		
2	jöv	,823	72,625	1,658	0 and 1
	Number of cars	,823	37,234	2,141	0 and 1
3	jöv	,730	76,972	1,666	0 and 1
	Number of cars	,823	37,700	2,203	0 and 1
	Age (in completed years) of reference person	,858	4,657	2,690	0 and 1

Wilks' Lambda

Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	,838	1	1	898	173,423	1	898,000	,000
2	2	,805	2	1	898	108,827	2	897,000	,000
3	3	,801	3	1	898	74,400	3	896,000	,000

Pairwise Group Comparisons[a,b,c]

Step	NR_10pc		0	1
1	0	F		173,423
		Sig.		,000
	1	F	173,423	
		Sig.	,000	
2	0	F		108,827
		Sig.		,000
	1	F	108,827	
		Sig.	,000	
3	0	F		74,400
		Sig.		,000
	1	F	74,400	
		Sig.	,000	

a. 1, 898 degrees of freedom for step 1.

b. 2, 897 degrees of freedom for step 2.

c. 3, 896 degrees of freedom for step 3.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,249[a]	100,0	100,0	,447

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,801	199,407	3	,000

Standardized Canonical Discriminant Function Coefficients

	Function
	1
jöv	,737
Number of cars	,496
Age (in completed years) of reference person	,174

Canonical Discriminant Function Coefficients

	Function
	1
jöv	,283
Number of cars	,882
Age (in completed years) of reference person	,011
(Constant)	-2,606

Unstandardized coefficients

Functions at Group Centroids

NR_10pc	Function
	1
0	-,166
1	1,496

Unstandardized canonical discriminant functions evaluated at group means

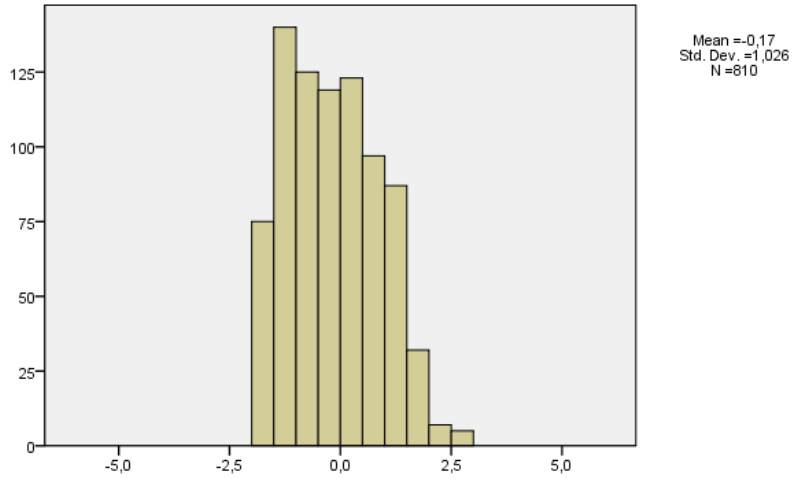
Classification Function Coefficients

	NR_10pc	
	0	1
jöv	1,409	1,880
Number of cars	,255	1,720
Age (in completed years) of reference person	,307	,326
(Constant)	-11,394	-19,027

Fisher's linear discriminant functions

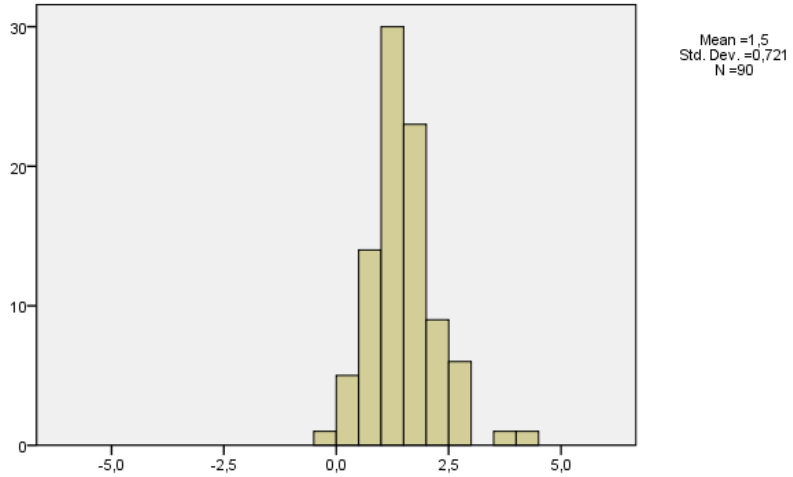
Canonical Discriminant Function 1

NR_10pc = 0



Canonical Discriminant Function 1

NR_10pc = 1



Classification Results[b,c]

		NR_10pc	Predicted Group Membership		Total
			0	1	
Original	Count	0	798	12	810
		1	73	17	90
	%	0	98,5	1,5	100,0
		1	81,1	18,9	100,0
Cross-validated[a]	Count	0	797	13	810
		1	73	17	90
	%	0	98,4	1,6	100,0
		1	81,1	18,9	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 90,6% of original grouped cases correctly classified.

c. 90,4% of cross-validated grouped cases correctly classified.

Complex Samples: Descriptives

Univariate Statistics

		Estimate	Standard Error	Coefficient of Variation
Mean	TC a felső 10% imputálva mah diszk alapján	1620586,09	10633,907	,007

5.3. alfejezet

Multiple Imputation

Imputed Values

Imputation Results

	Imputation Method	Fully Conditional Specification
	Fully Conditional Specification Method Iterations	10
Dependent Variables	Imputed	HE00C
	Not Imputed(Too Many Missing Values)	
	Not Imputed(No Missing Values)	HB05,HC03,HC04,HC05,HC12,HD14_04,HD14_14,HH09_5
	Imputation Sequence	HE00C,HB05,HC03,HC04,HC05,HC12,HD14_04,HD14_14,HH09_5

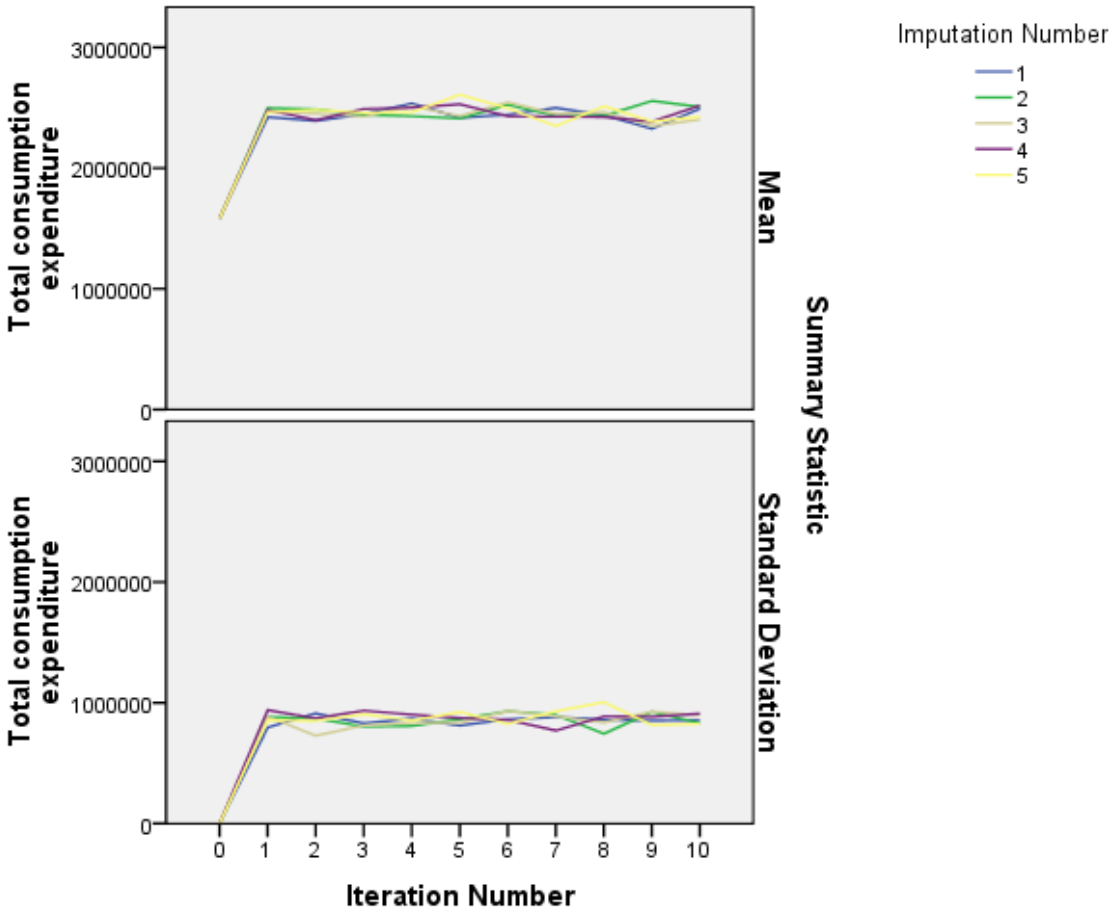
Imputation Models

	Model			
	Type	Effects	Missing Values	Imputed Values
Total consumption expenditure	Linear Regression	HC03,HC05,HC12,HB05,HC04,HD14_04,HD14_14,HH09_5	90	450

Descriptive Statistics

Data	Imputation	N	Mean	Std. Deviation	Minimum	Maximum
Original Data		810	1584876,04	928135,319	177292,00	7695496,00
Imputed Values	1	90	2492164,96	854677,547	681064,95	5384039,08
	2	90	2507749,93	830403,517	967829,96	4801665,53
	3	90	2402891,17	896478,581	769979,83	5070877,09
	4	90	2519055,67	909808,909	698372,48	5756997,38
	5	90	2425085,83	824802,684	1177892,99	5984624,70
Complete Data After Imputation	1	900	1675604,93	960041,621	177292,00	7695496,00
	2	900	1677163,43	959270,775	177292,00	7695496,00
	3	900	1666677,55	956581,950	177292,00	7695496,00
	4	900	1678294,00	967353,086	177292,00	7695496,00
	5	900	1668897,02	951919,703	177292,00	7695496,00

GGraph



Complex Samples: Descriptives

Univariate Statistics

		Estimate	Standard Error	Coefficient of Variation	Design Effect
Mean	Total consumption expenditure	1668815,81	12595,015	,008	,175

Univariate Statistics

		Square Root Design Effect	Population Size
Mean	Total consumption expenditure	,418	9058,000

Descriptives

Descriptive Statistics

Imputation Number		N	Mean		Std. Deviation
		Statistic	Statistic	Std. Error	Statistic
5	Total consumption expenditure	900	1668897,02	31730,657	951919,703
	Valid N (listwise)	900			
Pooled	Total consumption expenditure	900	1668897,02	31730,657	
	Valid N (listwise)	900			

Descriptive Statistics

Imputation Number		Fraction Missing Info.	Relative Increase Variance	Relative Efficiency
		Statistic	Statistic	Statistic
Pooled	Total consumption expenditure	.	.	.

Descriptives

Descriptive Statistics

	N	Mean		Std. Deviation
	Statistic	Statistic	Std. Error	Statistic
Total consumption expenditure	900	1716799,67	33390,132	1001703,975
Valid N (listwise)	900			

Descriptive Statistics

	N	Mean		Std. Deviation
	Statistic	Statistic	Std. Error	Statistic
Total consumption expenditure	900	1665488,15	31896,867	956906,012
Valid N (listwise)	900			

Complex Samples: Descriptives

Univariate Statistics

		Estimate	Standard Error	Coefficient of Variation	Design Effect
Mean	Total consumption expenditure	1716685,51	8618,436	,005	,074

Univariate Statistics

		Square Root Design Effect	Population Size
Mean	Total consumption expenditure	,272	9058,000

A 6.1. fejezet mellékszámításai

Descriptives

[DataSet2] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\elemzések\EV\EV9900.sav

Descriptive Statistics

	N	Mean		Std. Deviation
	Statistic	Statistic	Std. Error	Statistic
Total consumption expenditure	809	1488033,18	22893,522	651158,728
Total consumption expenditure	765	1406879,05	20660,211	571433,188
Total consumption expenditure	720	1337702,27	19177,599	514588,986
Total consumption expenditure	675	1274723,55	18003,814	467752,806
Total consumption expenditure	630	1215273,15	16954,909	425564,828
Total consumption expenditure	585	1159686,72	16094,454	389272,911
Total consumption expenditure	540	1108283,77	15494,554	360060,895
Total consumption expenditure	495	1057294,56	14915,371	331846,065
Total consumption expenditure	450	1007642,37	14445,521	306435,780
Valid N (listwise)	450			

Oneway

[DataSet3] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\elemzések\Rétegzett\többszintű rétegzett 900\REG_AUTO_CSALÁD900.sav

ANOVA

Total consumption expenditure

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8,548E14	9	9,498E13	685,254	,000
Within Groups	1,236E14	892	1,386E11		
Total	9,785E14	901			

Oneway

[DataSet4] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav

ANOVA

Total consumption expenditure

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8,719E14	9	9,688E13	1272,541	,000
Within Groups	6,775E13	890	7,613E10		
Total	9,396E14	899			

Oneway

[DataSet5] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\elemzések\EV\EV9900.sav

ANOVA

Total consumption expenditure

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9,485E14	9	1,054E14	619,170	,000
Within Groups	1,515E14	890	1,702E11		
Total	1,100E15	899			

Complex Samples: Descriptives

[DataSet3] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\elemzések\Rétegzett\többszintű rétegzett 900\REG_AUTO_CSALÁD900.sav

Univariate Statistics

		Estimate	Standard Error
Mean	Total consumption expenditure	1729945,81	22933,482
	Total consumption expenditure	1485820,61	15771,759
	Total consumption expenditure	1407971,07	14676,638
	Total consumption expenditure	1337700,17	14021,838
	Total consumption expenditure	1275129,64	13293,710
	Total consumption expenditure	1216864,30	12757,580
	Total consumption expenditure	1162518,88	12364,157
	Total consumption expenditure	1111130,39	11899,241
	Total consumption expenditure	1059336,87	11490,738
	Total consumption expenditure	1007776,35	11424,595

Correlations

[DataSet4] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav

Correlations

		Total consumption expenditure	fogy
Total consumption expenditure	Pearson Correlation	1	,903**
	Sig. (2-tailed)		,000
	N	900	900
fogy	Pearson Correlation	,903**	1
	Sig. (2-tailed)	,000	
	N	900	900

**. Correlation is significant at the 0.01 level (2-tailed).

A 6.2. fejezet mellékszámításai

Discriminant

[DataSet1] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav

Group Statistics

NR_10pc		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
0	jöv	5,1284	2,69657	810	810,000
	Number of cars	,4753	,56254	810	810,000
	Population density domain	2,0099	,79858	810	810,000
	Level of studies completed by the reference person	1,6765	,77448	810	810,000
1	jöv	8,9333	1,45957	90	90,000
	Number of cars	1,2000	,56489	90	90,000
	Population density domain	1,5556	,75120	90	90,000
	Level of studies completed by the reference person	2,3222	,74695	90	90,000
Total	jöv	5,5089	2,83882	900	900,000
	Number of cars	,5478	,60306	900	900,000
	Population density domain	1,9644	,80520	900	900,000
	Level of studies completed by the reference person	1,7411	,79534	900	900,000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
jöv	,838	173,423	1	898	,000
Number of cars	,870	134,316	1	898	,000
Population density domain	,971	26,519	1	898	,000
Level of studies completed by the reference person	,941	56,691	1	898	,000

Pooled Within-Groups Matrices

		jöv	Number of cars	Population density domain	Level of studies completed by the reference person
Correlation	jöv	1,000	,421	-,169	,351
	Number of cars	,421	1,000	,018	,235
	Population density domain	-,169	,018	1,000	-,292
	Level of studies completed by the reference person	,351	,235	-,292	1,000

Analysis 1

Box's Test of Equality of Covariance Matrices

Log Determinants

NR_10pc	Rank	Log Determinant
0	4	-,618
1	4	-1,714
Pooled within-groups	4	-,666

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

	Box's M	54,257
F	Approx.	5,337
	df1	10
	df2	104757,642
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,260[a]	100,0	100,0	,454

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,794	207,097	4	,000

Standardized Canonical Discriminant Function Coefficients

	Function
	1
jöv	,579
Number of cars	,493
Population density domain	-,216
Level of studies completed by the reference person	,111

Structure Matrix

	Function
	1
jöv	,862
Number of cars	,758
Level of studies completed by the reference person	,493
Population density domain	-,337

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

Functions at Group Centroids

NR_10pc	Function
	1
0	-,170
1	1,528

Unstandardized canonical discriminant functions evaluated at group means

Classification Statistics

Classification Processing Summary

	Processed	900
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
	Used in Output	900

Prior Probabilities for Groups

NR_10pc		Cases Used in Analysis	
	Prior	Unweighted	Weighted
0	,900	810	810,000
1	,100	90	90,000
Total	1,000	900	900,000

Classification Results[b,c]

		NR_10pc	Predicted Group Membership		
			0	1	Total
Original	Count	0	798	12	810
		1	73	17	90
	%	0	98,5	1,5	100,0
		1	81,1	18,9	100,0
Cross-validated[a]	Count	0	797	13	810
		1	73	17	90
	%	0	98,4	1,6	100,0
		1	81,1	18,9	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 90,6% of original grouped cases correctly classified.

c. 90,4% of cross-validated grouped cases correctly classified.

Complex Samples: Logistic Regression

[DataSet1] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_10pc[a]	0	8153,000	90,0%
	1[b]	905,000	10,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
jöv	5,5089
Number of cars	,55
Population density domain	1,96
Level of studies completed by the reference person	1,74

Pseudo R Squares

Cox and Snell	,214
Nagelkerke	,447
McFadden	,370

Dependent Variable: NR_10pc (reference category = 1)
Model: (Intercept), jöv, HD14_02, HA09, HC08

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	33,692	,000
(Intercept)	1,000	890,000	65,048	,000
jöv	1,000	890,000	44,290	,000
HD14_02	1,000	890,000	35,442	,000
HA09	1,000	890,000	3,250	,072
HC08	1,000	890,000	1,448	,229

Dependent Variable: NR_10pc (reference category = 1)
Model: (Intercept), jöv, HD14_02, HA09, HC08

Parameter Estimates

NR_10pc	Parameter	B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper
0	(Intercept)	7,816	,969	5,914	9,717
	jöv	-,640	,096	-,828	-,451
	HD14_02	-1,195	,201	-1,589	-,801
	HA09	,344	,191	-,031	,718
	HC08	-,207	,172	-,545	,131

Dependent Variable: NR_10pc (reference category = 1)
Model: (Intercept), jöv, HD14_02, HA09, HC08

Parameter Estimates

NR_10pc	Parameter	Design Effect	Exp(B)	95% Confidence Interval for Exp(B) Lower	95% Confidence Interval for Exp(B) Upper
0	(Intercept)	,990	2478,816	370,063	16604,026
	jöv	,998	,527	,437	,637
	HD14_02	,993	,303	,204	,449
	HA09	1,006	1,410	,970	2,051
	HC08	1,007	,813	,580	1,140

Dependent Variable: NR_10pc (reference category = 1)
Model: (Intercept), jöv, HD14_02, HA09, HC08

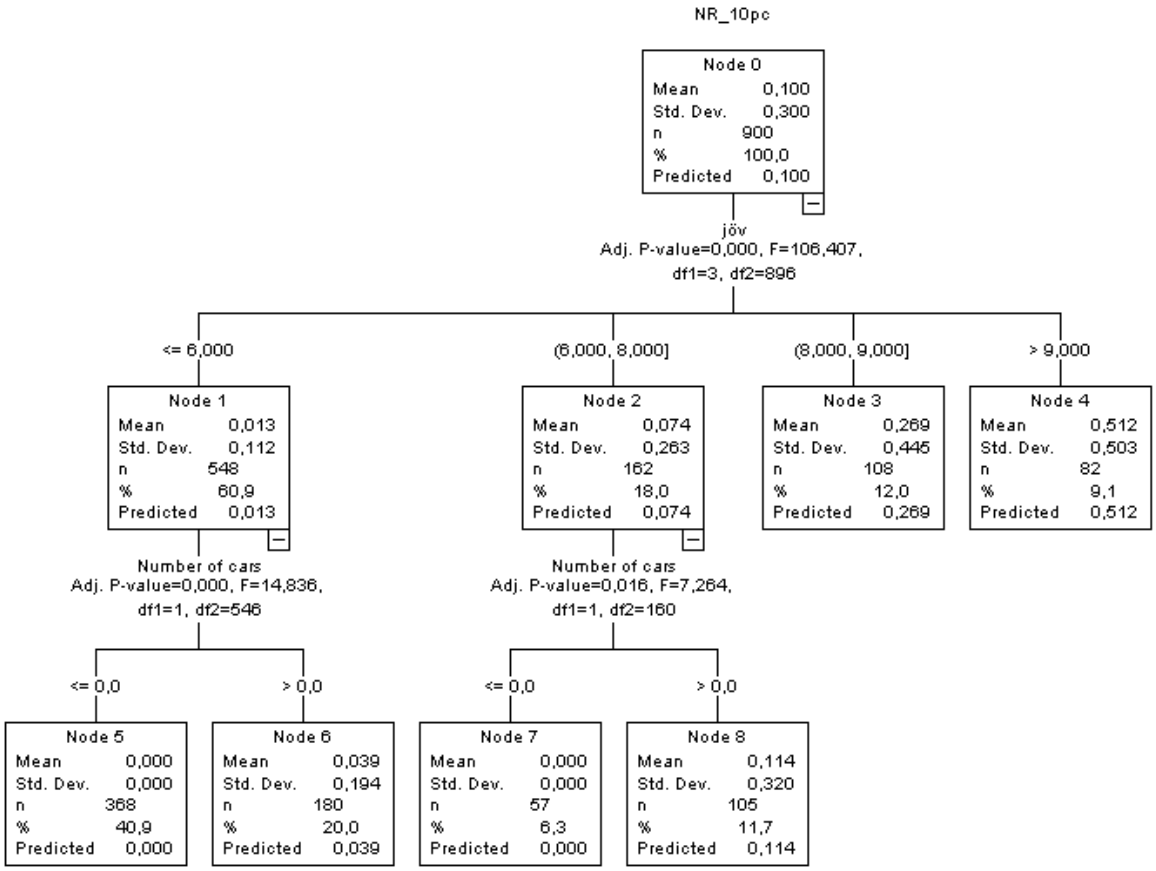
Classification

	Predicted		
Observed	0	1	Percent Correct
0	7971,800	181,200	97,8%
1	623,444	281,556	31,1%
Overall Percent	94,9%	5,1%	91,1%

Dependent Variable: NR_10pc (reference category = 1)
Model: (Intercept), jöv, HD14_02, HA09, HC08

Classification Tree

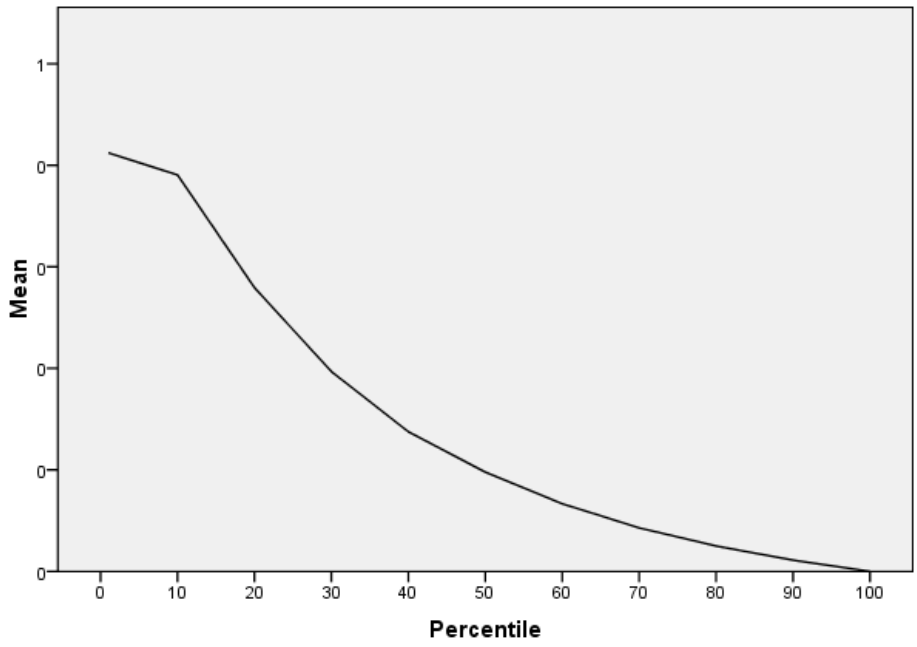
[DataSet1] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav



Gain Summary for Nodes

	Node-by-Node			Cumulative		
Node	N	Percent	Mean	N	Percent	Mean
4	82	9,1%	,51	82	9,1%	,51
3	108	12,0%	,27	190	21,1%	,37
8	105	11,7%	,11	295	32,8%	,28
6	180	20,0%	,04	475	52,8%	,19
5	368	40,9%	,00	843	93,7%	,11
7	57	6,3%	,00	900	100,0%	,10

Growing Method: CHAID
Dependent Variable: NR_10pc



Growing Method:CHAID

Dependent Variable:NR_10pc

Risk

Method	Estimate	Std. Error
Resubstitution	,066	,005
Cross-Validation	,068	,006

Growing Method: CHAID
Dependent Variable: NR_10pc

```
CROSSTABS  /TABLES=NR_10pc BY PredictedValue_1  /FORMAT=AVALUE TABLES
/CELLS=COUNT TOTAL  /COUNT ROUND CELL.
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
NR_10pc * Predicted Value	900	100,0%	0	,0%	900	100,0%

NR_10pc * Predicted Value Crosstabulation

			Predicted Value					
			0	0	0	0	1	Total
NR_10pc	0	Count	425	173	93	79	40	810
		% of Total	47,2%	19,2%	10,3%	8,8%	4,4%	90,0%
	1	Count	0	7	12	29	42	90
		% of Total	,0%	,8%	1,3%	3,2%	4,7%	10,0%
	Total	Count	425	180	105	108	82	900
		% of Total	47,2%	20,0%	11,7%	12,0%	9,1%	100,0%

A 6.3. fejezet mellékszámításai

CSLOGISTIC NR_10pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_10pc[a]	0[b]	8153,000	90,0%
	1	905,000	10,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,214
Nagelkerke	,447
McFadden	,370

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	33,692	,000
(Intercept)	1,000	890,000	65,048	,000
HC08	1,000	890,000	1,448	,229
HD14_02	1,000	890,000	35,442	,000
jöv	1,000	890,000	44,290	,000
HA09	1,000	890,000	3,250	,072

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_10pc	Parameter	B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Design Effect	Exp(B)
1	(Intercept)	-7,816	,969	-9,717	-5,914	,990	,000
	HC08	,207	,172	-,131	,545	1,007	1,230
	HD14_02	1,195	,201	,801	1,589	,993	3,303
	jöv	,640	,096	,451	,828	,998	1,896
	HA09	-,344	,191	-,718	,031	1,006	,709

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_10pc	Parameter	95% Confidence Interval for Exp(B) Lower	95% Confidence Interval for Exp(B) Upper
1	(Intercept)	6,023E-5	,003
	HC08	,877	1,725
	HD14_02	2,227	4,897
	jöv	1,570	2,290
	HA09	,488	1,031

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted 0	Predicted 1	Percent Correct
0	7971,800	181,200	97,8%
1	623,444	281,556	31,1%
Overall Percent	94,9%	5,1%	91,1%

Dependent Variable: NR_10pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

WEIGHT BY súly. DESCRIPTIVES VARIABLES=TC_10pcNR /STATISTICS=MEAN STDDEV MAX SEMEAN.

Descriptives

[DataSet1] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 10% nem válaszolt	898	3149694	1554136,44	23295,622	698205,746
Valid N (listwise)	898				

CSLOGISTIC NR_15pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_15pc[a]	0[b]	7700,000	85,0%
	1	1358,000	15,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,272
Nagelkerke	,477
McFadden	,376

Dependent Variable: NR_15pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	50,527	,000
(Intercept)	1,000	890,000	102,342	,000
HC08	1,000	890,000	2,908	,089
HD14_02	1,000	890,000	39,681	,000
jöv	1,000	890,000	80,860	,000
HA09	1,000	890,000	3,159	,076

Dependent Variable: NR_15pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_15pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-6,857	,678	-8,187	-5,527	,969	,001
	HC08	,239	,140	-,036	,514	1,004	1,270
	HD14_02	1,249	,198	,860	1,639	,988	3,488
	jöv	,580	,064	,453	,706	,966	1,785
	HA09	-,276	,155	-,581	,029	1,002	,759

Dependent Variable: NR_15pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_15pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,000	,004
	HC08	,965	1,671
	HD14_02	2,363	5,148
	jöv	1,573	2,026
	HA09	,559	1,029

Dependent Variable: NR_15pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted		
	0	1	Percent Correct
0	7327,533	372,467	95,2%
1	744,478	613,522	45,2%
Overall Percent	89,1%	10,9%	87,7%

Dependent Variable: NR_15pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

WEIGHT OFF. COMPUTE súly_15pc=1/(1-PredPr_15pc2). VARIABLE LABELS súly_15pc 'ln Total Consumption'. EXECUTE. WEIGHT BY súly_15pc. DESCRIPTIVES VARIABLES=TC_15pcNR /STATISTICS=MEAN STDDEV MAX SEMEAN.

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 15% nem válaszolt	902	2691488	1494852,39	20533,476	616737,691
Valid N (listwise)	902				

WEIGHT OFF.
CSLOGISTIC NR_20pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_20pc[a]	0[b]	7247,000	80,0%
	1	1811,000	20,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,346
Nagelkerke	,547
McFadden	,425

Dependent Variable: NR_20pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	53,529	,000
(Intercept)	1,000	890,000	123,886	,000
HC08	1,000	890,000	,988	,320
HD14_02	1,000	890,000	47,096	,000
jöv	1,000	890,000	110,147	,000
HA09	1,000	890,000	3,988	,046

Dependent Variable: NR_20pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_20pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-6,668	,599	-7,844	-5,492	,963	,001
	HC08	,131	,132	-,128	,390	,998	1,140
	HD14_02	1,499	,218	1,070	1,928	,996	4,478
	jöv	,629	,060	,511	,746	,979	1,875
	HA09	-,283	,142	-,560	-,005	1,004	,754

Dependent Variable: NR_20pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_20pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,000	,004
	HC08	,880	1,476
	HD14_02	2,917	6,875
	jöv	1,667	2,109
	HA09	,571	,995

Dependent Variable: NR_20pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted		
	0	1	Percent Correct
0	6753,733	493,267	93,2%
1	704,411	1106,589	61,1%
Overall Percent	82,3%	17,7%	86,8%

Dependent Variable: NR_20pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

COMPUTE súly_20pc=1/(1-PredPr_20pc2). VARIABLE LABELS súly_20pc 'ln Total Consumption'. EXECUTE. WEIGHT BY súly_20pc. DESCRIPTIVES VARIABLES=TC_20pcNR /STATISTICS=MEAN STDDEV MAX SEMEAN.

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 20% nem válaszolt	894	2391272	1428680,30	18224,586	544983,316
Valid N (listwise)	894				

CSLOGISTIC NR_25pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_25pc[a]	0[b]	6794,000	75,0%
	1	2264,000	25,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,392
Nagelkerke	,581
McFadden	,443

Dependent Variable: NR_25pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	67,186	,000
(Intercept)	1,000	890,000	148,514	,000
HC08	1,000	890,000	3,189	,074
HD14_02	1,000	890,000	42,266	,000
jöv	1,000	890,000	149,720	,000
HA09	1,000	890,000	3,344	,068

Dependent Variable: NR_25pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_25pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-6,442	,529	-7,480	-5,405	,956	,002
	HC08	,234	,131	-,023	,491	1,003	1,263
	HD14_02	1,365	,210	,953	1,777	,982	3,917
	jöv	,646	,053	,542	,750	,979	1,908
	HA09	-,242	,133	-,502	,018	1,007	,785

Dependent Variable: NR_25pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_25pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,001	,004
	HC08	,977	1,634
	HD14_02	2,594	5,914
	jöv	1,720	2,116
	HA09	,605	1,018

Dependent Variable: NR_25pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted		
	0	1	Percent Correct
0	6230,267	563,733	91,7%
1	724,622	1539,378	68,0%
Overall Percent	76,8%	23,2%	85,8%

Dependent Variable: NR_25pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

COMPUTE súly_25pc=1/(1-PredPr_25pc2). VARIABLE LABELS súly_25pc 'ln Total Consumption'. EXECUTE. WEIGHT BY súly_25pc. DESCRIPTIVES VARIABLES=TC_25pcNR /STATISTICS=MEAN STDDEV MAX SEMEAN.

Descriptives

[DataSet1] C:\Documents and Settings\HP\Asztal\PhD\dolgozat\6 fejezet\MR_FOGY_900.sav

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 25% nem válaszolt	889	2168619	1371625,71	16408,198	489103,828
Valid N (listwise)	889				

CSLOGISTIC NR_30pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_30pc[a]	0[b]	6341,000	70,0%
	1	2717,000	30,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable
b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,428
Nagelkerke	,607
McFadden	,457

Dependent Variable: NR_30pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	66,553	,000
(Intercept)	1,000	890,000	155,391	,000
HC08	1,000	890,000	9,769	,002
HD14_02	1,000	890,000	36,867	,000
jöv	1,000	890,000	150,461	,000
HA09	1,000	890,000	,969	,325

Dependent Variable: NR_30pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_30pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-6,480	,520	-7,500	-5,460	,971	,002
	HC08	,401	,128	,149	,653	1,004	1,494
	HD14_02	1,237	,204	,837	1,637	,973	3,446
	jöv	,657	,054	,552	,763	,986	1,930
	HA09	-,126	,128	-,376	,125	1,006	,882

Dependent Variable: NR_30pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_30pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,001	,004
	HC08	1,161	1,922
	HD14_02	2,310	5,141
	jöv	1,737	2,144
	HA09	,687	1,133

Dependent Variable: NR_30pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted		
	0	1	Percent Correct
0	5656,467	684,533	89,2%
1	744,789	1972,211	72,6%
Overall Percent	70,7%	29,3%	84,2%

Dependent Variable: NR_30pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

COMPUTE súly_30pc=1/(1-PredPr_30pc2). VARIABLE LABELS súly_30pc 'ln Total Consumption'. EXECUTE. WEIGHT BY súly_30pc. DESCRIPTIVES VARIABLES=TC_30pcNR /STATISTICS=MEAN STDDEV MAX SEMEAN.

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 30% nem válaszolt	881	2001130	1316734,39	14881,940	441816,740
Valid N (listwise)	881				

CSLOGISTIC NR_35pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_35pc[a]	0[b]	5888,000	65,0%
	1	3170,000	35,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable
b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,455
Nagelkerke	,626
McFadden	,468

Dependent Variable: NR_35pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	75,810	,000
(Intercept)	1,000	890,000	171,412	,000
HC08	1,000	890,000	13,196	,000
HD14_02	1,000	890,000	49,040	,000
jöv	1,000	890,000	170,744	,000
HA09	1,000	890,000	,371	,542

Dependent Variable: NR_35pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_35pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-6,217	,475	-7,149	-5,285	,972	,002
	HC08	,462	,127	,212	,712	1,005	1,587
	HD14_02	1,336	,191	,961	1,710	,976	3,802
	jöv	,647	,050	,550	,744	,973	1,910
	HA09	-,076	,124	-,319	,168	1,008	,927

Dependent Variable: NR_35pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_35pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,001	,005
	HC08	1,237	2,038
	HD14_02	2,615	5,529
	jöv	1,733	2,105
	HA09	,727	1,183

Dependent Variable: NR_35pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

	Predicted		
Observed	0	1	Percent Correct
0	5284,000	604,000	89,7%
1	775,033	2394,967	75,6%
Overall Percent	66,9%	33,1%	84,8%

Dependent Variable: NR_35pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 35% nem válaszolt	886	1839433	1266237,08	13328,578	396683,477
Valid N (listwise)	886				

CSLOGISTIC NR_40pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_40pc[a]	0[b]	5435,000	60,0%
	1	3623,000	40,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,458
Nagelkerke	,619
McFadden	,454

Dependent Variable: NR_40pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	81,069	,000
(Intercept)	1,000	890,000	160,966	,000
HC08	1,000	890,000	8,982	,003
HD14_02	1,000	890,000	40,811	,000
jöv	1,000	890,000	180,985	,000
HA09	1,000	890,000	,077	,781

Dependent Variable: NR_40pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_40pc	Parameter	B	Std. Error	95% Confidence Interval Lower	95% Confidence Interval Upper	Design Effect	Exp(B)
1	(Intercept)	-5,596	,441	-6,462	-4,730	,958	,004
	HC08	,377	,126	,130	,623	1,005	1,457
	HD14_02	1,181	,185	,818	1,544	,973	3,257
	jöv	,648	,048	,553	,742	,971	1,911
	HA09	-,034	,124	-,277	,208	1,008	,966

Dependent Variable: NR_40pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_40pc	Parameter	95% Confidence Interval for Exp(B) Lower	95% Confidence Interval for Exp(B) Upper
1	(Intercept)	,002	,009
	HC08	1,139	1,865
	HD14_02	2,266	4,681
	jöv	1,739	2,101
	HA09	,758	1,231

Dependent Variable: NR_40pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

	Predicted		
Observed	0	1	Percent Correct
0	4750,467	684,533	87,4%
1	825,411	2797,589	77,2%
Overall Percent	61,6%	38,4%	83,3%

Dependent Variable: NR_40pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 40% nem válaszolt	879	1720898	1214319,01	12162,043	360568,893
Valid N (listwise)	879				

CSLOGISTIC NR_45pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_45pc[a]	0[b]	4982,000	55,0%
	1	4076,000	45,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,449
Nagelkerke	,601
McFadden	,433

Dependent Variable: NR_45pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	86,106	,000
(Intercept)	1,000	890,000	149,239	,000
HC08	1,000	890,000	6,613	,010
HD14_02	1,000	890,000	37,047	,000
jöv	1,000	890,000	192,794	,000
HA09	1,000	890,000	,030	,863

Dependent Variable: NR_45pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_45pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-5,005	,410	-5,809	-4,201	,949	,007
	HC08	,316	,123	,075	,557	1,001	1,372
	HD14_02	1,069	,176	,725	1,414	,963	2,914
	jöv	,630	,045	,541	,719	,954	1,877
	HA09	,021	,119	-,214	,255	1,008	1,021

Dependent Variable: NR_45pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_45pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,003	,015
	HC08	1,078	1,746
	HD14_02	2,064	4,113
	jöv	1,717	2,052
	HA09	,808	1,291

Dependent Variable: NR_45pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted		
	0	1	Percent Correct
0	4247,133	734,867	85,2%
1	885,833	3190,167	78,3%
Overall Percent	56,7%	43,3%	82,1%

Dependent Variable: NR_45pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 45% nem válaszolt	886	1611028	1172524,66	11180,322	332780,580
Valid N (listwise)	886				

CSLOGISTIC NR_50pc(LOW) WITH HC08 HD14_02 jöv HA09

Complex Samples: Logistic Regression

Categorical Variable Information

		Weighted Count	Weighted Percent
NR_50pc[a]	0[b]	4529,000	50,0%
	1	4529,000	50,0%
[a]	Population Size	9058,000	100,0%

a. Dependent Variable

b. Reference Category

Covariate Information

	Mean
Level of studies completed by the reference person	1,74
Number of cars	,55
jöv	5,5089
Population density domain	1,96

Pseudo R Squares

Cox and Snell	,472
Nagelkerke	,630
McFadden	,461

Dependent Variable: NR_50pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Tests of Model Effects

Source	df1	df2	Wald F	Sig.
(Corrected Model)	4,000	887,000	84,899	,000
(Intercept)	1,000	890,000	135,809	,000
HC08	1,000	890,000	8,096	,005
HD14_02	1,000	890,000	48,797	,000
jöv	1,000	890,000	196,797	,000
HA09	1,000	890,000	,030	,861

Dependent Variable: NR_50pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_50pc	Parameter			95% Confidence Interval			
		B	Std. Error	Lower	Upper	Design Effect	Exp(B)
1	(Intercept)	-4,925	,423	-5,755	-4,096	,959	,007
	HC08	,357	,126	,111	,604	1,002	1,430
	HD14_02	1,285	,184	,924	1,646	,979	3,616
	jöv	,656	,047	,564	,747	,960	1,927
	HA09	,022	,126	-,226	,270	1,007	1,022

Dependent Variable: NR_50pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Parameter Estimates

NR_50pc	Parameter	95% Confidence Interval for Exp(B)	
		Lower	Upper
1	(Intercept)	,003	,017
	HC08	1,117	1,829
	HD14_02	2,520	5,189
	jöv	1,758	2,112
	HA09	,798	1,310

Dependent Variable: NR_50pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Classification

Observed	Predicted		
	0	1	Percent Correct
0	3693,467	835,533	81,6%
1	664,378	3864,622	85,3%
Overall Percent	48,1%	51,9%	83,4%

Dependent Variable: NR_50pc (reference category = 0)
Model: (Intercept), HC08, HD14_02, jöv, HA09

Descriptives

Descriptive Statistics

	N	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Std. Error	Statistic
TC a felső 50% nem válaszolt	857	1495152	1105331,58	10018,668	293268,216
Valid N (listwise)	857				