Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar
Szociológiai, Szociálpolitikai Intézet és Továbbképző Központ

HOGYAN VÁLASSZUK KI A KÉRDEZENDŐT A HÁZTARTÁS TAGJAI KÖZÜL?

Konzulens:
Rudas Tamás

Készítette:
Németh Renáta
nappali tagozat, szociológia szak
survey statisztikus szakirány

2001. április

KIVONAT

Az egyelemű véletlen minta kiválasztásának problémája gyakran merül fel többlépcsős mintavételi eljárás utolsó lépcsőjében. A Kish-kulcs alkalmazásával egyelemű véletlen mintát vehetünk olyan csoportból, ami elég kicsi ahhoz, hogy a terepmunka során egyszerűen sorba rendezhető legyen. Erre egyéb bevett eljárások is léteznek, így a legutolsó vagy a legközelebbi születésnap módszere. Ezek összehasonlításakor több dimenzió mentén kell értékelnünk a módszerek használhatóságát, a kivitelezhetőség és az elvi statisztikai hatékonyság szempontja egyaránt lényeges. A Kish-kulcs módszerének kivitelezhetőségét nemzetközi tapasztalatokra támaszkodva értékeltem. Eszerint a kulcs használata magasabb válaszmegtagadást eredményezhet a terepmunka során adódó hosszadalmas adminisztráció miatt. Újabb tapasztalatok szerint vélhetően a kérdezők hatása is jelentős szerepet tölt be a válaszmegtagadás létrejöttében. A válaszarányt tekintve nem bizonyítható a Kish-kulcs hátránya más, hasonló célú mintavételi eljárásokkal szemben. A Kish-módszert hatékonysági szempontból a minta nem és kor szerinti reprezentativitását tekintve értékeltem. Mivel empirikus adatok vizsgálatával a mintaválasztás hatása nem különíthető el más hatásoktól, a dolgozatban virtuális populáción, szimulációval teszteltem a kulcs használatát. A szimulációval kapott minták nem és korcsoport szerinti reprezentativitását értékeltem, vizsgálatom szerint jelentős eltérés tapasztalható a Kish-módszerrel kapott minta és a populáció között. Eredményeim szerint ez az eltérés csökkenthető a Kish-kulcs módosításával. A mintából kapott becsléseket javítani hivatott súlyozás több módját is elemeztem, a Kish-kulccsal kapott szimulációs mintán tesztelve azokat.

Tartalomjegyzék

1 Bevezetés **4**

2 Felmérés, felméréstervezés **6**
2.1 A felmérés tervezése a kutatási célokkal összhangban 6
2.2 Hibák és költségek 7
2.3 Az elemzési szakasz 7
2.3.1 A becslések hibája 8

3 Mintavétel, a mintavételhez kapcsolódó alapfogalmak **9**

4 Kérdezendők kiválasztása a háztartásmintából **9**

5 A Kish-kulcs **11**

6 A Kish-kulcs változatai és egyéb eljárások a kérdezendő kiválasztására **13**
6.1 Az eljárások összevetése telefonos felmérésekben 14

7 A minta reprezentativitása **16**
7.1 A magyarországi háztartásszerkezet jellemzői 18
7.2 Kish-kulccsal vett minták reprezentativitása hazai felmérésekben 19
7.3 A kiválasztás szimulációja - Kish-módszer 22
7.4 A kiválasztás szimulációja - random választás 26
7.5 A populációt változtatva 26

8 Súlyozás **28**
8.1 A súlyozásról általában 28
8.1.1 Az alapsúly 28
8.1.2 Rétegzés, utólagos rétegezés 29
8.1.3 A mintavételi módszer hatása a becslés varinciájára 31
8.1.4 A súlyozás hatása a becslés varianciájára 31
8.2 Súlyozás a Kish-kulccsal vett minta elemzésekor 32
8.2.1 Az alapsúly 32
8.2.2 Utólagos rétegezés 33
8.2.3 Utólagos rétegezés, alapsúly nélkül 33
8.3 Becslések a Kish-kulccsal vett mintából 34

9 A Kish-kulcs megváltoztatása **36**

10 Összefoglaló **41**

11 Függelék **42**

12 Hivatkozások **45**

13 Felhasznált adatbázisok **45**

1. Bevezetés

A felmérésstatisztika tudománya az utóbbi néhány évtized terméke. Az 1930-ban megjelent *Encyclopedia of the Social Sciences* figyelmeztetése szerint az emberi populációból történő mintavétel olyan nehézségekbe ütközik, hogy helyette a teljes cenzus végrehajtása javasolt. Ennek ellenére ma a világ legtöbb országában széles körben alkalmazott módszer a mintán történő felmérés. Az enciklopédia megjelenése óta eltelt hetven évben a felmérések statisztikai tervezésének tudománya gyors ütemű fejlődésen ment keresztül, az elért eredmények alkalmazásával biztosítható a mintából gyűjtött információk megfelelő általánosíthatósága.

Dolgozatom témája a kérdezendő háztartáson belüli kiválasztásában alkalmazott Leslie Kish-kulcs. Témaválasztásom indoka, hogy bár a kulcsot széles körben használják hazai és külföldi felmérésekben, mégis gyakran utalnak a kutatók a kulcs hátrányos tulajdonságaira. Célom volt, hogy a Kish-módszert lehetőleg teljességel, a felmérésstatisztika minden területe felől megközelítve értékeljem. Terveim között szerepelt a kulcs módosításának megkísérlése is, a minta kor és nem szerinti reprezentativitását javítandó. Olyan szemlélettel szerettem volna megközelíteni a problémát, mely figyelembe veszi a felmérések kivitelezésének költségvonzatát is. A "jobb" mintát produkáló eljárások nagyobb forrásokat igényelnek, míg a kevésbé hatékony módszerek általában könnyebben kivitelezhetők. Szándékaim szerint a dolgozat teljes gondolatmenetét végigkíséri ez a szemlélet, explicitté téve azt a felismerést, hogy a megfelelő eljárás kiválasztása minden felmérés tervezésekor egyedi döntést igényel.

A kulcs vizsgálatakor a legfontosabb szakirodalmi forrás a módszert elsőként tárgyaló cikk Leslie Kish tollából, a Journal of the American Statistical Association hasábjain (Kish, 1949). Ugyanő később, 1965-ben megjelent Survey Sampling c. művében önálló fejezetet szentelt a kérdezendő háztartásból történő kiválasztásnak, itt ismét visszatért a kulcsra, részletezve annak elvi, statisztikai előnyeit (Kish, 1965). Az azóta megjelent tanulmányok inkább a kivitelezhetőség oldaláról közelítették a módszert.

Hipotézisem szerint a kulcs teljesítményének megítélésekor figyelembe kell venni a felmérés célcsoportjának egyedi jellemzőit. A módszer kivitelezhetősége, a válaszmegtagadás aránya annak a társadalomnak a szociokulturális jellegétől függ, ahol a felmérést végrehajtani kívánjuk. A statisztikai szempontú jellemzésnél, a minta reprezentativitásának és a becslések pontosságának kérdésénél pedig mindenképpen figyelembe kell venni a kulcs alkotójának eredeti szándékát. Kish célja a háztartáson belüli egyenlő kiválasztási esélyek biztosítása volt. Nem volt, nem is lehetett célja reprezentatív mintát produkáló eljárás kidolgozása. Hogy mégis jó mintákat kapott akkor és ott, a 60-as évek Amerikájában, az nagyrészt a korabeli háztartásszerkezetre vezethető vissza. Előfeltevésem az, hogy a kulcs mai, hazai alkalmazása kevésbé jó mintákat ad, és ennek oka a magyar háztartásstruktúrában keresendő.

A téma feldolgozásában a kivitelezhetőségi problémákat hazai és külföldi felmérések eredményeire támaszkodva jártam körül. A statisztikai szempontú jellemzésnél mintavételi szimulációkat végeztem, így a véletlen hatásának kiszűrésével állandó tendenciák felismerésére számíthattam. Kísérletet tettem a kulcs módosítására, a minta várható nem és kor szerinti illeszkedését kívántam javítani. Ezt a problémát egy matematikai optimalizálási modellben fogalmaztam meg, és a feladat megoldásában a nemlineáris programozás eszköztárára támaszkodtam.

A dolgozat első részében a felmérések tervezésének lényeges lépéseit foglaltam össze. A tervezés a kutatási célokkal összhangban történik, a rendelkezésre álló források figyelembe vételével kell dönteni a várhatóan legkevesebb hibát eredményező eljárás kiválasztásáról. Ezután a mintavételi eljárásokat és a minták "jó" tulajdonságait tipizáltam. Adott mintavételi eljárás jellemzését a belőle származó minta tulajdonságainak általánosíthatóságával kapcsoltam össze. A negyedik fejezetben ismertetem magát a problémát: hogyan válasszunk ki pontosan egy kérdezendőt a háztartásokból. A Kish-kulcs eredeti változatának bemutatása található az ötödik fejezetben, a hatodikban pedig a kulcs módosított változatait sorolom fel, és ugyanitt írtam le más eljárásokat, amik ugyanezen probléma - a háztartásból történő kiválasztás - megoldási kísérletei. A különböző eljárások kivitelezhetőség szerinti összevetése a nemzetközi tapasztalatok alapján történt, mivel nem állt rendelkezésemre megfelelő hazai anyag. A hetedik fejezet foglalkozik a minta "jóságának" egyik dimenziójának, a reprezentativitásnak a kérdésével. Magyarországi felmérésekből származó mintákat vettem példának, illetve mintavételek szimulációs sorozatából kapott mintákat jellemeztem a reprezentativitás szempontjából. Megszorítva a problémát, csak a nem és kor szerinti reprezentativitással foglalkoztam. A felmérések elemzési szakaszában gyakran alkalmazott súlyozás

a témája a nyolcadik fejezetnek. A súlyozás több céllal történhet, megkíséreltem szétválasztani ezeket a célokat, és önállóan jellemezni hatásukat. Gyakorlati tesztelésüket a Kish-kulccsal történt szimulációból kapott mintán végeztem. Vizsgáltam őket önállóan, eloszlásuk segítségével, illetve az alkalmazásukkal kapott becslések összehasonlításával. Az utolsó fejezetben a kulcs módosítását hajtottam végre, a minta nem és korcsoport szeritni reprezentativitását javító eljárást hozva létre.

2. Felmérés, felméréstervezés

A modern társadalmakban folyamatos szükséglet mutatkozik az országok vagy egyéb populációk bizonyos jellemzőinek közigazgatási, gazdasági, politikai vagy más célból történő megismerésére. Ezek az információk a populáció tagjainak felkeresésével szerezhetők be, ha nem lelhetők fel az aggregált, tehát nem individulás szintű adatokat gyűjtő tárakban (Warwick, Lininger 1975). Aggregált infomáció például a politikai választások pártonkénti szavazataránya, de nem található ilyen adat a két választás közötti pártpreferenciáról. Ezenkívül az aggregált adatokból csak szűk körben nyerhető információ bizonyos jellemzők kapcsolatáról, mivel az egyes tulajdonságok általában csak önmagukban ismertek. Ezzel szemben a felkeresés lehetőséget ad az egyén több, összekapcsolt jellemzőjének megismerésére. Így például az aggregált adatok között a választás eredményén kívül a populáció néhány demográfiai jegye is megtalálható, de ezek összekapcsolására, így a pártpreferenciák motivációjának esetleges magyarázatára a felkeresés ad lehetőséget. A felkeresés történhet a teljes populáció cenzusával, vagy felméréssel. A felmérés viszonylag alacsony költségei és gyors kivitelezhetősége miatt preferált a teljeskörű adatfelvétellel, a cenzussal szemben. A felmérések előnyei abból származnak, hogy a tanulmányozni kívánt populációnak csak egy kiválasztott részét, a mintát vizsgálják. Az így kapott információk azután a populáció egészére kivetíthetők. Ennek az általánosíthatóságnak a megbízhatósága nagyban függ attól, hogy hogyan választották a mintát. Dolgozatomban a mintaválasztás egy speciális problémájával foglalkozom.

2.1. A felmérés tervezése a kutatási célokkal összhangban

Felmérés útján kielégíthető információs igény felmerülésekor a kutatási terv létrehozása az első lépés. Megfogalmazása két fő paramétert kell, hogy rögzítsen. Az egyik a megismerni kívánt egyedek összessége, az úgynevezett **célpopuláció**. A másik a megismerni kívánt populációs jellemzők számbavétele.

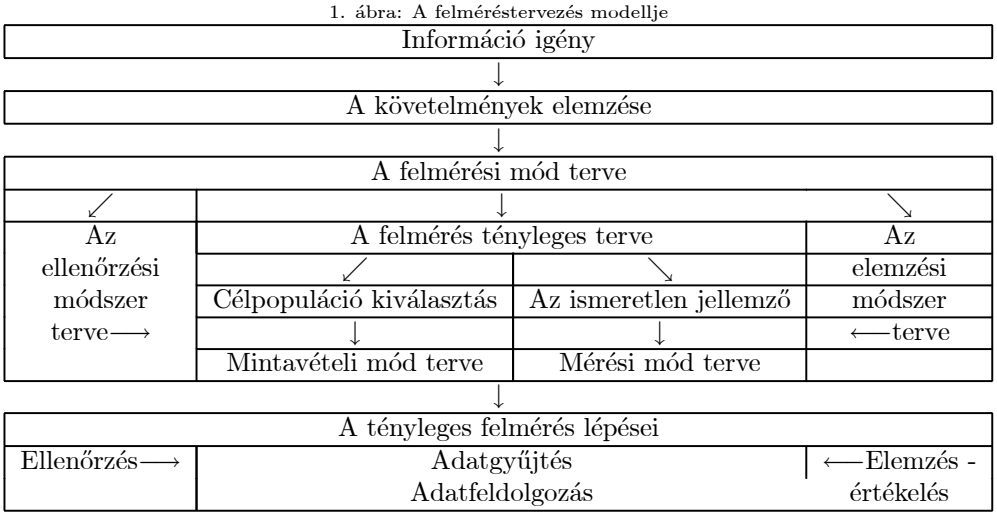
A kutatás tervezésekor ezután a célpopuláció gyakorlati elérésére alkalmas **keret**et adjuk meg. A keret megválasztása tulajdonképpen a célpopuláció konceptualizálása. Gyakori eset, hogy a megfelelő keret megválasztásának kérdése a költségek és várható hibák közötti explicit egyensúlyozást idéz elő. Ilyen döntési helyzet állt elő Groves (1989) valós példája szerint, amikor egy felmérés célpopulációja az amerikai zsidó háztartások lakóinak összessége volt. A keret meghatározására két megoldás kínálkozott. Az egyik szerint a zsidó szervezetek taglistái alapján állíthatták volna össze a keretet. A másik megoldás szerint területi valószínűségi mintát jelöltek volna ki. A területi minta lényege, hogy egy adott földrajzi egység minél egyenletesebb lefedését célozva választunk háztartásokat. Az utóbbi megoldás lényegesen költségigényesebb, hisz sok "felesleges" felkeresést kíván. Mégis, bármely zsidó háztartásnak pozitív mintába kerülési esélyt biztosít, míg a szervezetek taglistájából készített keret szűkebb a célpopulációnál. Ráadásul szisztematikus le nem fedettségi hibát okozhat az utóbbi, ha a zsidó szervezeti tagság valamilyen pszichikai jellemzővel hozható összefüggésbe (a tagok fontosnak tartják a vallási-kulturális közösséghez tartozást).

Másik példa populáció és keret kapcsolatára: ha a kutatás célpopulációja a magyar állampolgárságú, 18. életévét betöltött személyek csoportja, a keretet rögzíthetjük például a Központi Választási és Népességnyilvántartó Hivatal aktuális választói címjegyzékeként. Ez utóbbi nem teljesen fedi le a teljes felnőtt lakosságot, adatgyűjtési hibák adódhattak összeállításakor, ezenkívül a tökéletes aktualizálás kivitelezhetetlensége miatt téves lakcímadatokat vagy időközben elhalálozott személyek adatait is tartalmazhatja.

A keret köztes állomás a célpopuláció és a minta között: a mintát a keretből választjuk. A mintavétel módját a kutatási tervben rögzített vizsgálni kívánt jellemzőknek és az elérni kívánt becslési pontosságnak megfelelően választjuk. Részletesebben: a minta nagyságával arányos a mintából a célpopulációra történő általánosításnak - a **becslés**nek - a megbízhatósága. Ezenkívül a vizsgálni kívánt populációs tulajdonságok is befolyásolják a megfelelő minta tulajdonságait, a populációról meglévő előzetes ismereteket felhasználva. Ha például a munkanélküliség országos arányát kívánjuk megbecsülni, a mintánknak földrajzilag egyenletesen szórtnak kell lennie, hiszen ismerjük a munkaerőpiaci regionális külöbségeket. Továbbá nagyobb mintára van szükség, ha a vizsgált jellemzőnek vélhetően alacsony a populációs gyakorisága.

Az 1. ábrán a teljes felmérési tervezés modelljét ábrázoltam (Särndal et al. 1997).

1. ábra: A felméréstervezés modellje



2.2. Hibák és költségek

A rendelkezésre álló források felső határa korlátozza a mintából származó becslések várható pontosságát (lásd 8.1.3. fejezet). Ugyanis a becslési pontosság elsősorban nagyobb minta választásával növelhető, de a mintanagyság arányos a költségekkel. Hasonlóan, a minta földrajzi kiterjedtsége további logisztikai és emberi erőforrás igényeket támaszt. Látni fogjuk, hogy a matematikailag jobb eredményeket ígérő kiválasztási módszerek gyakran bonyolultabbak, hosszasabb adminisztrációt igényelnek a terepmunka során, ezzel a válaszmegtagadási arányt növelik, így tényleges pluszköltségeket okoznak.Ugyanígy az ellenőrzési mechanizmusok beépítése a terepmunka vagy az adatrögzítés folyamatába bár növeli a felmérés hatékonyságát, de jelentős a költségvonzata. A tervezőnek tehát a megbízó igényeinek, a kutatás céljának figyelembe vételével kell egyensúlyozni a költségek és a várható hibák között. Tulajdonképpen valódi költséghatékonysági elemzést igényel a döntéshozatal (Groves 1989).

2.3. Az elemzési szakasz

Az elemzés módszerének meghatározása is a felmérési terv része. Statisztikai eszközeinek beállításai a mintavétel módjának megfelelően választandók. Például a mintavétel egyedi jellemzőire illeszkedik az elemzésben használt súlyozás (lásd 8. fejezet). A szórásbecslés is más, ha csoportos mintát választottunk vagy ha több lépcsős eljárást alkalmaztunk. Ugyanígy rétegzett mintánál is figyelembe kell venni a rétegzéskor definiált rétegváltozókat a variancia számításához (8.1.2. fejezet). A szórásbecslés mintavételre illesztésére utal az angol elnevezés: *design-based estimate of variance*.

A becslés tehát a mintából számolt közelítése a populáció valamely jellemzőjének, a populációs paraméternek. Adott becslési eljárás több dimenzió mentén jellemezhető, a becslések elsősorban átlaguk és szórásuk szerint értékelhetők (Levy, P.S., Lemeshow, S., 1999).

A becslés átlaga a populációból az adott módszerrel vett összes lehetséges mintához tartozó becslések átlagaként definiálható:

$E\left(\widehat{d}\right) = \sum_{i=1..C} \widehat{d}_i \pi_i$

ahol

$\widehat{d}$ a d paraméter becslése

$\widehat{d}_i$ a d paraméter becslése az i. mintából

π_i az i. minta kiválasztásának valószínűsége

C az összes lehetséges minta száma

Ha a becslés átlaga nem egyezik meg a valódi populációs paraméterrel, a becslés torzít. A **torzítás** olyan eltérés, ami szisztematikus, tehát nem csupán a véletlennek tulajdonítható. A torzítás mértéke az eltérés nagysága:

$B\left(\widehat{d}\right) = E\left(\widehat{d}\right) - d$

A másik fontos tulajdonság a becslés varianciája:
$Var\left(\widehat{d}\right) = \sum_{i=1..C}\left(\widehat{d}_i - E\left(\widehat{d}\right)\right)^2 \pi_i$
A becslés varianciájának négyzetgyöke a **sztenderd hiba**.
Az **átlagos négyzetes hiba** a valódi populációs paraméter és a becslés átlagos négyzetes eltérése:
$MSE\left(\widehat{d}\right) = \sum_{i=1..C}\left(\widehat{d}_i - d\right)^2 \pi_i$
Az átlagos négyzetes hiba, a becslés varianciája és a torzítás között fennáll az alábbi összefüggés:
$MSE\left(\widehat{d}\right) = Var\left(\widehat{d}\right) + B^2\left(\widehat{d}\right)$
Ezen fogalmak segítségével jellemezhetők a becslési eljárások. A becslés **megbízhatósága** attól függ, mennyire hasonló becsléseket kapunk a mintavétel többszöri megismétlésével. A megbízhatóság a sztenderd hiba fogalmával írható le: minél kisebb a sztenderd hiba, annál megbízhatóbb a becslés[1].

A becslés **érvényessége** arra utal, vajon a többszöri mintavétel után kapott becsléssorozat menynyire mozog a valódi populációs paraméter körül. A torzítás terminusa segítségével ragadható meg ez a dimenzió: a torzítás csökkenésével nő a becslés érvényessége[2].

Ha a konkrét becslések távolsága a paramétertől átlagosan nem nagy, az a becslésünk pontosságára utal. A **pontosság** az átlagos négyzetes hiba segítségével jellemezhető. Minél kisebb az átlagos négyzetes hiba, annál pontosabb a becslési eljárás.

A becslés sztenderd hibája segítségével konstruálható meg az adott megbízhatósági szinthez tartozó **megbízhatósági intervallum**. A megbízhatósági intervallum a becslés értékét magában foglaló szakasz. A szakasz végpontjainak megadásához a mintából kapott értékeket használjuk fel, ezért ez a szakasz maga is a véletlentől függ. A szokásos 95%-os szint esetén ha torzítatlan a becslésünk, akkor annak valószínűsége, hogy ez a véletlen helyzetű intervallum tartalmazza a populációs paramétert, megközelítőleg 0,95. Ez úgy is interpretálható, hogy 100 mintát véve a kapott 100 megbízhatósági intervallum közül várhatóan 95 lefedi a populációs paramétert.

A pontbecslésen kívül a becslés sztenderd hibájának és megbízhatósági intervallumának megadása további információkkal szolgál. Megtudható belőle, mennyire megbízható az adott becslés, illetve segítségével tesztelhető két becslés eltérésének statisztikai szignifikanciája.

2.3.1. A becslések hibája

A felmérésből kapott becslés eltérését a valódi populációs értéktől több hibaforrás okozhatja (Warwick, Lininger 1975, Särndal et al. 1997)

- Azonos módon kivitelezett felmérésekből számolt becslések is különböznek egymástól. Ez a hiba az úgynevezett **mintavételi hiba**. Oka maga a módszer, azaz a mintavétel alkalmazása a célpopuláció teljes cenzusa helyett. Ez a típusú hiba csökkenthető a mintanagyság növelésével, illetve megfelelő mintavételi eljárással és becléstechnikával.

- A **nem mintavételi hibák** akkor is fellépnek, ha a teljes populációt bevonjuk a felmérésbe. Ide tartozik a mérési hiba; például ez a hiba lép fel, ha a kérdezett nem valós információt szolgáltat. Nem mintavételi hiba adódhat szisztematikus válaszmegtagadásból, vagy az adatrögzítés hiányosságaiból is. Maga az adatelemzés- és publikálás is okozhat hibát. Ilyen eset az, ha a felmérés szponzora az érdekeinek megfelelő eredményt várja el, vagy ha a kutató önnön szakmai presztízsének érdekében "szignifikáns eltérések kihozására" törekszik.

A becslések előbbi jellemzésekor - a megbízhatóság, az érvényesség, a pontosság dimenziói mentén - tulajdonképpen csak a mintavételi hiba fellépését vettük számításba, egy hipotetikus szituációt tételezve fel. Nem mintavételi hiba megléte esetén ez a hibatényező is befolyásolja a becslés eredményét, a felmérésből kapott adatok pontatlanságán keresztül. Például mérési hiba megléte esetén a becslés eltérését a paramétertől a véletlenen kívül a rosszul mért adatok pontatlansága is növeli. Ekkor a becslés érvényessége nem teljesen feleltehető meg a becslés torzításának. Ugyanígy, ha a felmérés eszköze megbízhatatlan

[1] Feltéve, hogy nem lép fel úgynevezett nem mintavételi hiba hiba. Lásd a következő fejezetet.
[2] Megint csak akkor, ha nem lép fel nem mintavételi hiba.

adatokat szolgáltat, ugyanazon a mintán kétszer egymás után elvégezve a felmérést, különböző eredményeket kaphatunk. Ekkor a becslés szórásán kívül a mérési hiba is befolyásolja a becslés megbízhatóságát.

3. Mintavétel, a mintavételhez kapcsolódó alapfogalmak

Ha a kiválasztási eljárás véletlen mechanizmussal történt, valószínűségi mintavételről beszélünk. A valószínűségi mintavételi módszernek két alapvető követelményt kell teljesítenie:

- A keret minden eleméhez ismert, pozitív kiválasztási valószínűséget rendeljen.
- Gyakorlatban megvalósítható eljárást vezethessünk le belőle.

Az első feltétel teljesüléséről. Nyilván ha a keret nem minden tagjának van esélye bekerülni a mintába -azaz ha nem minden elemhez rendelünk pozitív kiválasztási valószínűséget-, tulajdonképpen megváltoztatjuk, szűkítjük felmérésünk keretét. Ez nem valós felmérési eredményekre vezet, torzítást okoz. A konkrét kiválasztási valószínűség ismerete az eredmények kontrollálhatósága céljából szükséges. Ugyanis azonos kiválasztási valószínűségek esetén kapunk torzítatlan becsléseket a mintából. A nem azonos, de ismert kiválasztási valószínűségek viszont megfelelően korrigálhatók az elemzési szakaszban, súlyozás alkalmazásával, így biztosítva becsléseink torzítatlanságát.

Nem elhanyagolható a második feltétel jelentősége a különböző mintavételi eljárások megvalósíthatóságának összehasonlításakor. Mint később látni fogjuk, a felméréstervező dolgozóasztalán megfelelően precíz módszernek talált eljárás a gyakorlatban esetleg nehezen kivitelezhetőnek bizonyulhat a tényleges terepmunkában.

A két kritérium teljesülése csak egyik előfeltétele a mintavételi módszer használhatóságának. A feltételeket teljesítő, de más-más kiválasztásra épülő módszerek összehasonlítása új dimenzió mentén történik. Ez a dimenzió a mintából levezethető becslések megbízhatósága. A megbízhatóság szempontjából a minta egyik jó tulajdonsága az azonos vagy közel azonos kiválasztási valószínűségek megléte. Mint később a súlyozásról szóló fejezetben (8.1.4. fejezet) látni fogjuk, ha a kiválasztási valószínűségek nem azonosak, az ezt korrigáló súlyozás csökkentheti a becslések megbízhatóságát.

További jó tulajdonság a megbízhatóság szempontjából a gyakran emlegetett **reprezentativitás**. A reprezentativitás szó jelentésének nagyszámú változatával találkozhatunk a szakirodalomban és a médiában. A legtágabb értelemében az adatok jóságát, a minta tipikus voltát értik alatta. Szűkebben a populáció valamely jellemzője szerinti megoszlás mintabeli reprodukálását jelenti. Az utólagos rétegzésről szóló 8.1.2. fejezetben részletesen kitérek arra, miért kívánatos a minta ilyen értelemben vett reprezentativitása. Röviden: ha a kutatási terv kérdései előismereteink szerint pozitív összefüggésben állnak az adott jellemzővel, akkor a becslés várhatóan megbízhatóbb lesz a jellemző szerinti reprezentativitás esetén. A reprezentativitást célzott módon éri el a kvótás mintavétel vagy a rétegzett mintavétel, a populációs arányok direkt betartásával.

A dolgozatban összehasonlításra kerülő módszerek elemzésekor a gyakorlati megvalósíthatóságra is kitértem. A megvalósíthatósági kérdésekben nemzetközi tapasztalatokra támaszkodtam, illetve hazai felmérések eredményét is vizsgáltam. A módszerek megbízhatósági dimenzióit vizsgálva önálló elemzéseket végeztem.

4. Kérdezendők kiválasztása a háztartásmintából

A választási nyilvántartások adatai nemzetközi tapasztalatok szerint eredményesen használhatók címlistaként, de általában nem megfelelő a minőségük a végső mintavételi lépésben, a személyek kiválasztásakor. Ennek oka, hogy a nyilvántartások sok hiányos adatot tartalmaznak a populációs mobilitás vagy a nyilvántartásba nem vétel miatt. A gyakorlatban ezért a nyilvántartási listát a lakások vagy háztartások kiválasztásához használják, a kérdezett kiválasztását a háztartásból az utolsó lépcsőben pedig véletlen mintavétellel oldják meg. Másik, ugyanerre a végső lépcsőre vezető módszer az úgynevezett területi mintavétel (*area sampling*). Ekkor a célpopuláció egy adott földrajzi egységen lokalizálható. A területet blokkokra osztják (pl. kerületek), ezeket kisebb egységekre (utcák). Ezek a lépcsők változhatnak.

A területi módszer lényege, hogy földrajzilag definiált mintavételi egységeken működik, így jó területi diszperzitást biztosít. Végső lépcsőben (pl. véletlen sétával) a háztartási mintához jutunk.

A **háztartás** mint mintavételi egység gyakran fordul elő felmérések kivitelezésében. Definíciója problematikus, nincs konszenzus a meghatározásában. Leggyakrabban együttélők olyan közösségét értik alatta, akik funkcionálisan is együttműködnek. Egyik ilyen megfogalmazás szerint a részben közös jövedelemből élő, közös konyhát használók tartoznak egy háztartásba. Hasonló mintavételi egység lehet a család. A család a rokoni kötelékkel is összekapcsolt együtt élőket jelentheti, de például a Központi Statisztikai Hivatal meghatározása szerint a család házas- vagy élettársi kapcsolatban álló, gyermekes vagy gyermek nélküli személyeket jelöl, illetve egyedül álló szülőt nőtlen/hajadon gyermekkel[3]. A család és a háztartás itt tehát nem teljesen átfedő kategóriák, az egy fedél alatt élő háztartás több generációja külön családokat alkot. Ugyanígy létezik a szakirodalomban a "nem családháztartás"[4] kifejezés, ami egymással családi kapcsolatban nem álló együttélőket jelent. Egy harmadik, a háztartás és a család fogalmát megint csaknem átfedő fogalom az együtt, egy lakásban élő személyek összessége. A dolgozat problémafelvetése szempontjából nem okoz problémát a három fogalom megkülönböztetése, hisz bárhogy is definiálja a felméréstervező a mintavételi egységet, ugyanaz a probléma jelentkezik mindhárom esetben.

Tegyük mostantól fel, hogy tervezett felmérésünk célpopulációja egy adott földrajzi egység, például Magyarország felnőtt (18 év feletti), nem intézményben élő lakossága. A keretet definiáljuk a Magyarország területén található háztartásokban élő felnőttek összességeként. A keret megválasztása nagyon gyakorlatias, de hibája például, hogy nem érjük el a hajléktalanokat. A definíció földrajzi elhelyezkedés alapján határozza meg a mintába bekerülhető háztartásokat, ebből további hibát eredhetnek. A mintavételi módszer legyen kétlépcsős mintavétel. Az első lépcsőben választjuk ki a mintába kerülő háztartásokat, adott kiválasztási valószínűségeket rendelve a mintaháztartásokhoz. A háztartás (felnőtt) tagjai közül választjuk ki a kérdezendő személyt a második lépcsőben, egyszerű véletlen mintavétellel.

Nem merülne fel a kérdezendő kiválasztásának problémája, ha a felmérés kutatási céljának megfelelően ő a háztartás meghatározott tagja, például a családfő lenne. Ugyanígy nem okozna nehézséget a választás, ha a mintavételi egység maga a háztartás lenne, ekkor bármely háztartástag egyaránt adekvát információkat szolgáltathatna.

A mi esetünkben a háztartások több tagja tartozik a kerethez. A kiválasztás itt sem okozna nehézséget, ha az alábbi feltételek bármelyike teljesülne:

- Bármelyik személytől beszerezhetők az összes többi háztartásbelire vonatkozó szükséges információk.
- Ha nagyon kicsi az esélye, hogy a háztartásnak több tagja tartozik a kerethez. Ez elmondható például, ha háztartásonként átlagosan 1,2 keretbeli személyt találunk.
- Ha a felmérésünk kérdéseire adott válaszoknak elhanyagolható a háztartásokon belüli korrelációja. Ellenkező esetben, vagyis amikor egyfajta homogenitás tapasztalható a háztartásokon belül, statisztikailag megbízhatóbb becsléseket kapnánk az eredmények elemzésekor, ha háztartásonként csak egyetlen személyt választunk ki.[5]

Tegyük most fel, hogy a fentiek egyike sem teljesül - mint ahogy a valóságban is kevés felmérés esetén állnak. Tehát egy személyt kívánunk háztartásonként kiválasztani. A háztartásonkénti egynél nem több kérdezett kiválasztásának igénye az egyébként valószínűsíthetően kényelmetlen interjúszituációk miatt is felmerül. Rögzítsük továbbá, hogy feleslegesen nem keresünk fel háztartásokat- tehát minden mintaháztartásból kiválasztunk egy kérdezendőt. Ennek az előfeltételének a költségkímélés az ésszerű indoka. Vegyük még fel feltevéseink közé, hogy a háztartások egyenlő valószínűséggel kerültek a mintába.

[3]A KSH "gyermek" definíciója sem egyezik a hétköznapi fogalommal. Eszerint gyermek a szüleivel élő nőtlen/hajadon családi állapotú egyén, tekintet nélkül a korára.

[4]A Központi Statisztikai Hivatal háztartáskategóriái között.

[5]Ezen homogenitás mértéke az úgynevezett osztályon belüli korreláció (intraclass correlation) együtthatója , a ρ. A csoportos mintavételből kapott adatokból számolt variancia $[1+\rho(B-1)]$-szerese annak a varianciának, amit a sztenderd módon számolva, a csoportokat figyelembe nem véve kapunk. (A képletben a B a csoportok átlagos nagyságát jelöli.) A $[1+\rho(B-1)]$ szorzótényező jelölése $deff$ (*design effect,* tervezés-hatás együttható), ami a mintavételi mód varianciaváltozásra gyakorolt hatására utal. Ha tehát a háztartásoknak átlagosan B tagját bevonva végzünk csoportos mintavételt, a $deff$ értékével szorzódik a varianciabecslésünk, a sztenderd hiba pedig a $deff$ négyzetgyökével. Minél homogénebbek a háztartások, annál nagyobb a ρ értéke, s annál kevesebb új információ nyerhető több háztartástag bevonásával (Kish, 1965).

Tehát az előzőleg egyenlő eséllyel kiválasztott háztartásokból kell egy interjúalanyt kiválasztani. A kérdezés gyakorlatában a kontrollálatlan kiválasztás lenne a legegyszerűbb módszer, amikor is azt a személyt választanánk, aki kinyitja az ajtót, vagy -telefonos felméréskor- aki felveszi a telefont. Ám ez a kiválasztás erős torzítást eredményezne a mintában, mivel azokat kérdeznénk, akik könnyen elérhetőek, illetve akik inkább hajlandóak együttműködni. Ők általában az idősebb korosztályhoz tartoznak és a nők, az iskolázottak is felülreprezentáltak köztük (Lavrakas, 1993).

5. A Kish-kulcs

Leslie Kish[6] szándéka szerint a kulcs egy világos és előre rögzített eljárást ad a kérdezendő kiválasztására (Kish, 1949, 1965). Lényege, hogy sztenderd eljárást kapunk a háztartástagok sorbarendezésére, majd véletlen módon választhatjuk ki a sor egyik elemét. Másik előnye, hogy használatának megfelelő dokumentációja birtokában a választás utólag is ellenőrizhető. Nem biztosítaná ugyanezt a kontrollálási lehetőséget például a pénzfeldobás vagy más gyakorlati véletlen módszer.

A Kish-módszert alkalmazó felmérésben minden kérdőívhez egy fedőlapot csatolnak. A fedőlapon a háztartástagok felsorolására alkalmas táblázat (példa az 1. táblázat) és egy úgynevezett Kish-tábla (példa a 2. táblázat) van.

1. táblázat: Kiválasztási adminisztráció a terepmunka során

Rokonsági viszony a családfőhöz	Nem	Kor	Sorszám	Kiválasztás
Családfő	F		2	
Feleség	N	40	5	
Családfő apja	F		1	
Fia	F	22	3	
Lánya	N	20	6	
A feleség nagynénje	N	44	4	✓

A háztartásbeliek sorba tételekor elsőként a férfiakat állítjuk sorba, életkor szerint (az idősektől a fiatalabbak felé), majd utánuk a nőket soroljuk fel - szintén életkoruk alapján. Az interjú első néhány kérdése az ehhez szükséges információkat gyűjti össze. Célszerű minél kevesebb kérdésből megállapítani a sorrendet, ezt segíti a rokonsági viszony megállapítása. Az 1. táblázatbeli példában a családfő és fia életkorának ismerete nélkül is végrehajtható volt a sorrendezés. A felsorolt háztartástagok közül a kiválasztandó személyt a Kish-tábla jelöli ki. A példánk esetén 5 felsorolt van, a 2. táblázat szerint ekkor a sorban a 4. személy a kérdezendő. (Ha a kiválasztott nem található otthon, a kérdező újra felkeresi a háztartást.)

[6]Leslie Kish (1910.-2000. október 11.) magyar származású statisztikus volt. A világ elsőszámú statisztikai mintavételi szakértőjeként jegyezték. 1965-ös könyve, a "Survey Sampling," ma is használt forrás szerte a világon. Véletlen mintavételi technikájának megalapozottsága először az 1948-as amerikai elnökválasztáskor nyert bizonyítást. Kevesebb mint 1000 házartást tartalmazó országos mintája Dewey és Truman szoros eredményét jósolta, kis fölénnyel Truman javára, míg a sajtó és a szokott felmérések Dewey megsemmisítő győzelmét várták. Számos megfontolásából a legfontosabbak közé tartozik a válaszmegtagadási arány figyelembe vételének szükségessége a valódi véletlen minta eléréséhez; továbbá a tervezés-hatás együttható összetett matematikai kiszámítása, amellyel a felmérés eredményeinek megbízhatóságát kívánta mérni. Kish az elsők között javasolta a guruló minták évenkénti kivitelezését.

Kish 1910-ben született, a mai Szlovákia területén fekvő Poprádon. Családjával 1925-ben emigrált az Egyesült Államokba, Bronxban telepedtek le. Apja egy év után váratlanul meghalt, ekkor anyja tésztaboltot nyitott Manhattan-ben, amit idővel Eleanor Roosevelt is szívesen látogatott. Kish nappal dolgozott, a középiskolát esti tagozaton végezte. 1937-ben, látva a fasizmus előretörését szerte Európában, önkénteseként csatlakozott a spanyolországi Nemzetközi Brigádokhoz, egy magyar csapat tagjaként harcolt. Huesca-nál megsebesült, 1939-ben tért vissza Amerikába. Hazatérte után a City College of New York-on szerzett matematikusi diplomát. Először Washington-ban dolgozott a U.S. Census Bureau-nál, majd a mezőgazdasági minisztériumban, ahol társadalomkutatók egy csoportjához csatlakozott, a minisztérium felméréskutató egységét alkotva. A háború után a michigan-i egyetemen Likert-tel és másokkal megalapították az Institute for Social Research intézetet, közben matematikai statisztikából és szociológiából szerzett tudományos fokozatot. Számos tudományos cím és kitüntetés birtokosává vált. Az Amerikai Statisztikai Szövetség elnökévé, a Nemzetközi Statisztikai Intézet és a Magyar Tudományos Akadémia tiszteletbeli tagjává választották, a Bologna-i Egyetem díszdoktorává avatták az egyetem 900-adik jubileumán.

2. táblázat: A 8 Kish-tábla egyike

D kiválasztási tábla	
Felnőttek száma a háztartásban	A kiválasztott sorszáma
1	1
2	2
3	2
4	3
5	4
6 vagy több	4

A D kiválasztási tábla csak egyike a nyolc hasonló táblázatnak (lásd a 3. táblázatot). Ezeket a táblázatokat a kérdőívekhez csatolják, véletlenszerűen párosítva őket. A nyolc táblázat nem egyenlő arányban oszlik meg a kérdőívek között, például a C jelzetű minden 6. kérdőíven található meg, míg az É csak minden 12.-en. Az elsődleges cél a háztartásokon belül egyenlő kiválasztási valószínűséget rendelni minden személyhez. A 4. táblázatot elemezve kitűnik, hogy ez csaknem minden háztartás esetén megvalósul. Például a 3 felnőtt tagot számláló háztartások sorban 1. tagját $1/6+1/12+1/12 = 1/3$ valószínűséggel, a 2. tagját $1/6+1/6 = 1/3$ valószínűséggel, 3.tagját $1/12+1/12+1/6 = 1/3$ valószínűséggel választjuk ki. Az egyenlő esélyű kiválasztás két esetben nem valósul meg. Az egyik a 6-nál több felnőtt tagot számláló háztartások esete. A 7. vagy nagyobb sorszámú tag 0 valószínűséggel kerül a mintába. A másik eltérés az 5 felnőtt tagú háztartásoknál figyelhető meg. Ott az 1. 2. és 4. tag 1/6 valószínűséggel kerül kiválasztásra, a 3. és az 5. 1/4 valószínűséggel. A várhatóan fiatal, nőnemű 5. személy felülreprezentálása Kish szerint kompenzálja a 6 tagú háztartásokból 0 valószínűséggel bekerülő, valószínűleg szintén fiatal lányok kiesését.

3. táblázat: A Kish-kulcs kiválasztási szabálya

A tábla aránya	A tábla jelzete	Ha a felnőttek száma: 1	2	3	4	5	6 vagy több
		A kiválasztott sorszáma:					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3*	5
1/12	E2	1	2	3	4	5*	5
1/6	F	1	2	3	4	5	6

Fontos megjegyezni, hogy Leslie Kish saját indoklása szerint a nem és kor változókat csupán a háztartástagok egyszerű sorbatételére használta, nem törekedett kimondva a populációs nem- és korösszetételt reprodukáló, reprezentatív minta létrehozására. Cikkében 4 felmérés súlyozatlan mintájának nem és kor összetételét vetette össze a cenzus adataival. Egyetlen szignifikáns eltérésnek a férfiak alulreprezentáltságát értékelte, ezt a férfiak ritka otthon tartózkodásával és nagyobb arányú válaszmegtagadásával magyarázta.

4. táblázat: Háztartáson belüli kiválasztási valószínűségek a Kish kulcs alkalmazása esetén

Háztartástagok száma						
Tag sorszáma a sorban	1	2	3	4	5	6 vagy több
1	1	1/2	1/3	1/4	1/6	1/6
2		1/2	1/3	1/4	1/6	1/6
3			1/3	1/4	1/4	1/6
4				1/4	1/6	1/6
5					1/4	1/6
6						1/6
7 vagy több						0

6. A Kish-kulcs változatai és egyéb eljárások a kérdezendő kiválasztására

Hazai és külföldi nagymintás felmérésekben is gyakran alkalmazzák a kulcsot[7]. A fenti kulcs 8 táblája (Kish, 1965) csak egy példa, több különböző mutánsa létezik a módszernek. Kish önmaga ajánlja a kulcs speciális céloknak megfelelő módosítását, így például vázolja a populáció adott százalékát kiválasztani képes variáns megvalósítását. Közöl egy másik módosítást is, mellyel a kérdezőbiztos munkáját könnyítendő, az A-C jelzetű táblákat úgy konstruálja, hogy azokkal várhatóan férfi kérdezettet válasszunk. A kérdező az A-C előlappal borított kérdőíveket a nap végére sorolhatja, így nagyobb eséllyel találhatja otthonában a jelöltet. Kish célja a módosítások felsorolásával az, hogy megvilágítsa, mennyire rugalmas és saját célra könnyen átalakítható eljárással állunk szemben.

A sorbarendezést követően alkalmazott Kish-féle kiválasztási módszer elsősorban a - véletlenszám-generálásra épülő - véletlen kiválasztást helyettesíti, a gyakorlatban is megvalósítható formában. A terepmunka során a tollal-papírral dolgozó kérdezőbiztosok ebben a formájában használják a Kish-kulcsot. A CATI[8] és CAPI[9] felmérésekben a kérdőívekhez a számítógép random módon rendeli a 8 tábla valamelyikét, a kívánt arányban. Az ilyen felmérésekben néha a kulcs módosított változata szerepel. A módosítás lényege, hogy ilyenkor adott a technika a véletlenszám-generálásra, ezért csak a háztartástagok azonosítása lenne szükséges. Így a keresztnevük alapján azonosítják a személyeket, majd a keresztnevekből választanak véletlenszerűen (Piazza, Respondent Selection for CATI/CAPI). Mégis, a leggyakrabban a telefonos felmérésekben is a Kish módszerrel sorolják fel a háztartástagokat. Vélhetően azért ragaszkodnak a kutatók számítógépes háttértechnika birtokában is a nem/kor alapján történő felsoroláshoz, mert ez a módszer valamilyen mértékben biztosítja a minta nem- és korösszetételének kiegyenlítettségét, bár Kish explicit módon nem hangsúlyozza ezt a szándékát. Hasonló a módszer elve a szisztematikus mintavételhez. A szisztematikus mintavétel során valamely fontosnak tartott jellemző alapján, például nemenként és azon belül kor szerint soroljuk fel a keretet, majd állandó lépéshosszal végiglépegetve a soron választjuk ki a mintát. Így csökkentjük a lehetséges minták nem és kor szerinti változékonyságát. A Kish módszernél a háztartás tagjait soroljuk így fel, szintén a minták variábilitását csökkentve. Ha keresztnevük alapján azonosítanánk a háztartástagokat, és így tennénk őket random módon sorba, pozitív valószínűséggel előállhat az az eset, hogy minden, női taggal is rendelkező háztartásból éppen nőt választunk ki. Ugyanez, a Kish kulccsal végezve a kiválasztást, csak nagyon speciális háztartás-minta esetén fordulhat elő. Például ahhoz, hogy a női taggal is rendelkező kéttagú háztartásokból a kulccsal mindenképpen nőt válasszunk ki, a háztartások sorban első tagja nő kell, hogy legyen, ha az első négy tábla volt a kérdőívhez rendelve (lásd a 3. táblázatot). Mivel a férfiakkal kezdjük a felsorolást, ez csak úgy jöhet létre, ha ezen háztartások mindkét tagja nő. Egy ilyen mintában a 3. táblázat szerint minden

[7] néhány példa:

A közigazgatás megítélése, 1999. MONITOR Társadalomkutató Intézet és Módszertani Központ. http://www.monitor-tki.hu

ISSP 1994. Család II., TÁRKI. http://www.tarki.hu

Csaba TV nézettségi mutatók, 1998. Jelenkutató Intézet. http://www.bekes.hungary.net/csabatv/nezettseg.htm

National Fertility and Family Survey, Európa több országában, 1990-es évek. kivitelező: Population Activities Unit of the United Nations Economic Commission for Europe. http://www.cpc.unc.edu/pubs/paa_papers/1995/klijzing.html

International Social Justice Project, 1991, 1996. 12 európai ország, köztük Magyarország bevonásával. http://www.isjp.-de/

SCSU Survey, Minnesota State Lottery, 1997. http://www.lottery.state.mn.us/gambling/stcloud.html

Los Medios y Mercados de Latinoamérica, 1997, több Latin-Amerikai országban. http://www.zonalatina.com/Zldata25.htm

Northern Ireland Social Attitudes Survey, 1989-1996. Social and Community Planning Research http://cain.ulst.ac.uk/othelem/research/nisas/robinson.htm

Scottish Health Survey, 1995. Social and Community Planning Research http://www.official-documents.co.uk/document/scottish/shealth/shch1.htm#1.2.1

Oregon Health Behavior Surveys, 1998, telefonos felmérés. State of Oregon, Health Division http://pub.das.state.or.us/-purchasing/contracts/8042.html

Australian National Social Science Survey, 1984. The Australian National University http://ssda.anu.edu.au/SSDA/CODEBOOKS/FR84/description.html

[8] *computer-assisted telephone interview*, számítógéppel támogatott telefonos kérdezés

[9] *computer-assisted personal interview*, számítógéppel támogatott személyes kérdezés. Közvetlenül a gépbe viszik az adatokat, a kérdőívnek nincs papír-verziója.

egytagú háztartás lakója nő, a kéttagú háztartások legalább fele két nőt foglal magába, és a többi kéttagú egyik lakója is nő stb.

Az úgynevezett legutóbbi születésnap módszer is a háztartástagok egyikének véletlen kiválasztására szolgál, a 80-as évek közepe óta nemzetközi szinten is nagyon elterjedt eljárás. Alkalmazásakor azt a személyt választjuk ki, akinek a felkeresés időpontjához időben visszafelé legközelebb esik a születésnapja. A legközelebbi születésnap módszere szintén használatos, ekkor időben előre haladva keressük az első születésnapot. A legutóbbi vagy legközelebbi születésnaposok nem különböznek szisztematikusan a teljes populácitótól (az asztrológia ezzel ellenkező csillagjegy-jellemezéseit most figyelmen kívül hagyva). Elméletileg a metódus valódi véletlen mintát eredményez, de a gyakorlatban nem ez a tapasztalat. A kérdezettek nehezen értik meg a legutóbbi születésnap fogalmát, sokszor a legkésőbb született, legfiatalabb személyt választják ki (Binson et al. 2000). Egy célzott vizsgálatban a legutóbbi születésnap módszerrel kiválasztott kérdezettet az interjú során a háztartás minden tagjának születési adatainak felsorolására kérték (Lavrakas 1993). Az összehasonlítás szerint a háztartások 20-25%-ában nem egyezett meg a teljes felsorolás alapján a kiértékelők által megjelölt személy a ténylegesen kiválasztottal. Egy másik hibára mutatnak rá a 4. táblázatban bemutatott adatok, melyek két, Egyesült Államokbeli felmérésből származnak (Lavrakas 1993). A vizsgálatban a legutóbbi születésnap módszerét alkalmazták. A kérdezettek által megjelölt hónapok gyakorisága meglepő módon szisztematikusan eltér az egyenletes eloszlás esetén várt gyakoriságoktól. Mindkét esetben a legnagyobb arányban megjelölt hónapok éppen a felmérést megelőző hónapok voltak. Ezt a megfigyelést támasztották alá korábban mások tapasztalatai is (Oldendick et al. 1988).

5. táblázat: A legutóbbi születésnap módszerrel kapott legutóbbi születésnapok megoszlása

	Az adott hónapot megjelölők aránya	
A születés hónapja	1990 Chicago április végén - május elején	1992 Egyesült Államok március végén - április elején
Január	10,6	10,0
Február	7,9	10,1
Március	12,7	11,1
Április	12,7	7,8
Május	7,9	6,0
Június	4,5	6,4
Július	5,3	7,8
Augusztus	6,4	8,0
Szeptember	8,7	8,0
Október	8,5	7,0
November	6,7	7,8
December	8,0	10,0

A megfigyelt torzítások oka egyfajta konrollálhatalan ön-kiválasztás lehet. Vélhetően az a személy került kiválasztásra, aki ajtót nyitott, vagy felvette a telefont. Önmagát jelölte ki legutóbbi születésnaposként, mivel hajlandóságot érezhetett az együttműködésre. Ezért saját születésnapjaként a felmérést megelőző hónapot jelölte meg.

A legutóbbi születésnap módszere tehát a háztartások le nem fedésének hibáját növeli, nyilvánvaló előnye, az egyszerű alkalmazhatóság mellett. Bár a hatás mértéke nem pontosan becsülhető meg, a felmérés hibáinak növelésében jelentkező költsége talán nagyobb, mint a haszna.

A kérdezett kiválasztásának egyik bevett módja a kvótás mintavétel, ami a háztartáson belüli szelekcióban is alkalmazható. A nem és kor szerinti kvóták megadása azért okoz itt nehézséget, mert feltettük, hogy minden felkeresett háztartásból választunk valakit. A feltételt teljesítő kvótás kérdezés egyik megvalósítása a következő fejezetben leírt Hagen-Collier módszer.

6.1. Az eljárások összevetése telefonos felmérésekben

A telefonon végzett felmérések dinamikus módon nyernek teret a nemzetközi gyakorlatban. Alkalmazásuk korlátai nyilvánvalóak, a telefonnal való ellátottság nem teljes, ezenkívül hazai eredmények szerint a telefon tulajdonlása és az iskolázottság, a vagyoni helyzet között erős összefüggés van (Tarjányi Jó-

zsef, Módszertani problémák a telefonos közvélemény-kutatásokban). Mégis, a telefonon végzett interjú költsége jóval alacsonyabb, és az adatrögzítés is hatékonyabb. A telefonos felmérések során is felmerül a háztartások véletlen kiválasztása után a kérdezett kiválasztásának problémája. A háztartások kiválasztásának egyszerű módja a telefonszámok random kiválasztása. A háztartáson belül az interjúalany kiválasztására gyors, egyszerű és a felhívott fél számára kevéssé zavaró eljárást kell találni, hiszen telefonon sokkal könnyebb megtagadni az együttműködést.

Több kutató vélekedik úgy, hogy a Kish-kulcs alkalmazásához szükséges nem / kor információk beszerzése kényelmetlen szituációt teremt a kérdezettek számára, amiből gyakran válaszmegtagadással szabadulnak meg. A háztartás felsorolása kellemetlen lehet az egyedül élő személyeknek (főként a nőknek), illetve azoknak, akik szeretnék elkerülni a velük együtt élő azonosítását (Binson et al. 2000).

A visszautasítás arányának csökkentésére tettek kísérletet több vizsgálatban az interjú forgatókönyvének variálásával (Groves et al. 1988, Groves 1989). A Kish-kulcs alkalmazásakor kedvezőbbnek találták azt a változatot, amikor a többi háztartástag nem/kor adminisztrációja előtt a telefont felvevő személy adatait vették fel, így teremtve vele kapcsolatot. Ez az eredmény alátámasztja azok vélekedését, akik szerint nagyon nehéz az interakció legkorábbi pillanataiban a háztartástagok sorrendezéséhez szükséges információk megszerzése.

A Kish-kulcs mutánsaként létrejött, a 70-es évektől széles körben használt módszer Verling Trodahl és Roy Carter Jr. munkája (1964). Két kérdést tesz fel, a háztartásbeli felnőttek majd köztük a nők számára vonatkozóan. Ezután a kérdezendőt nem és relatív kora alapján azonosítja, pl. "a legfiatalabb férfi", a "második legidősebb nő". Tesztelése során viszonylag nagy számban produkált hibát (Groves et al. 1988). Amikor a Trodahl-Carter (T-C) módszerrel kiválasztott kérdezetől az interjú során a háztartás teljes leírását is beszerezték, az esetek 9%-ában különbséget találtak ezek között az adatok és a kiválasztásra felhasznált információ között.

Másik Kish-kulcs utód a Hagen-Collier módszer (Lavrakas 1993). Alkalmazásakor a kérdezők az alábbi négy személy egyikét kérik a háztartás tagjai közül a kérdőívhez csatolt kiválasztási algoritmus alapján: a legfiatalabb vagy a legidősebb nőt, illetve a legfiatalabb vagy a legidősebb férfit. Ha egy nőt tartalmaz a háztartás, ő egyben a legfiatalabbnak és a legidősebbnek is minősül. Ha a kívánt nem nem fordul elő a háztartásban, a további lépéseket a kérdezőknek mellékelt instrukciók rögzítik. Amikor egyenlő arányban elosztva használják a négy verziót, a Lavrakas hivatkozta források szerint a középkorúak illetve a férfiak alulreprezentáltak lesznek a mintában. Ugyanő azt ajánlja, hogy gyakorlatban a korábban létrejött minták nem és kor szerinti eltolódásainak megfelelően állítsuk be a 4 verzió megfelelő arányát. Egy ilyen beállítással egyszerre korrigáljuk a kiválasztási módszer torzító hatását és minden más potenciális torzító hatást, mint például a háztartások egyenlő kiválasztásának esetleges sérülését, a válaszmegtagadás torzítását stb. A kiválasztási eljárás önálló hatása egyébként sem azonosítható csupán a megvalósult mintából . A Hagen-Collier módszer a tapasztalatok szerint megőrizte a Kish-kulcs nehéz kivitelezhetőségének problémáját. Az algoritmus csak a módszerben járatos kérdezői gárdával alkalmazható hatékonyan.

Oldendick és társai a Kish-kulcs és a legutóbbi születésnap módszerének összehasonlítása céljából célzott vizsgálatot terveztek (Oldendick et al. 1988). Az Egyesült Államok három államában végeztek telefonos felmérést, államonként mintegy 1000 darabos volt a megvalósult interjúk mennyisége. A felmérésekben véletlen módon rendelték a háztartásokhoz a két kiválasztási módszer egyikét. A kérdezők gyakorlott munkatársak voltak, akik a Kish-módszert jól ismerték és zökkenőmentesen voltak képesek azt alkalmazni. Eredményeik szerint bár a személykiválasztást követő válaszmegtagadás aránya mindhárom felmérés esetén a születésnap–módszerrel volt kedvezőbb, a különbség nem volt nagy, sem szignifikáns. Sőt, a Kish módszerrel adminisztrált háztartások tagjai mindhárom felmérésben nagyobb arányban voltak hajlandók a kezdeti visszautasítás utáni rábeszélést követően együttműködni. Így a végleges visszautasítást nézve a születésnap módszer előnye még kevésbé jelentős. A minták reprezentativitását vizsgálva egyetlen demográfiai jellemző mentén találtak szignifikáns eltérést. A feketék aránya a születésnap-módszer alkalmazásakor volt közelebb a cenzus becsléséhez, a Kish-módszerrel ez az arány inkább alatta volt a valódinak. A kutatók hipotézise szerint a feketék talán tolakodóbbnak tartják a Kish-kulcshoz szükséges adminisztrációt. Két másik jellemző is eltérést mutatot, de nem szignifikáns módon. Az egyetlen felnőtt taggal bíró háztartások aránya a születésnap-módszernél magasabb, a visszautasítás oka talán a Kish módszerhez szükséges háztartásnagyság kérdés zavaró volta. A másik különbség a férfiak száma, itt a Kish módszerrel kapott minta közelebb volt a populációs arányhoz, bár mindkét esetben alatta volt a

valódi értéknek a mintabeli arány. Nem találtak lényeges eltérést más kédésekre adott válaszokban sem. Figyelmet érdemel viszont a születésnap módszerrel felvett kérdezettek születési hónapjának eloszlása. A felmérést közvetlenül megelőző hónapok jóval gyakrabban fordulnak elő. Az elemzésben nem interpretálták ezt a jelenséget, mondván, kevéssé valószínű, hogy a születési hónap bármiféle kapcsolatban lenne más, lényeges jellemzőkkel. Fentebb, a fejezet elején már sor került ennek a megfigyelésnek a tárgyalására.

Az Oldendick és társai által végzett vizsgálat tehát nem talált lényeges eltérést a két különböző módszerrel produkált minta között. Fontos nyomatékosítani azt a háttértényezőt, hogy a felmérést a Kish módszerben járatos kérdezőkkel vették fel. Más elemzések azt mutatják, hogy a Kish-kulcsban gyakorlatlan kérdezők nem csupán magas visszautasítási arányt produkálnak, de magát a módszert sem megfelelően végzik (Lavrakas 1993).

A kérdezői hatás fontossága Binson és társai (2000) által végzett kísérletben is bizonyítást nyert. A kutatók rendkívül érdekes és újszerű eredményeket kaptak. Három eljárást hasonlítottak össze: a Kish--kulcsot, a legutóbbi és a legközelebbi születésnap módszerét. Céljuk elsősorban a visszautasítási arány vizsgálata volt. A felmérésben az Egyesült Államok 48 államát szerepeltették. A szokásos módon véletlen telefomszámokkal érték el a háztartásmintát, mintegy 3500 interjút vettek fel. A 6. táblázatban látható, hogy valóban eltérést találtak a három eljárás között, a legnagyobb kiesési arány a Kish-kulcsnál mutatkozott. A táblázatban azoknak az aránya látható, ahol valaki felvette a telefont, de az interjú folyamán valahol megszakadt a beszélgetés. Meglepő, hogy a két születésnap-módszer is eltérő eredményt produkált.

6. táblázat: O2sszehasonlítás a félbeszakadt interjűk aránya alapján

Interjű félbeszakadása	Legutóbbi születésnap	Legközelebbi születésnap	Kish módszer
Igen	34,1%	32,0%	37,0%
Nem	65,9%	68,0%	63,0%
(n)	(1.116)	(1.143)	(1.166)

A Kish és a legközelebbi születésnap módszer között találtak szignifikáns eltérést, de az eltérés nem volt szignifikáns a Kish és a legutóbbi születésnap módszer között, sem a két születésnap módszer között. Az interjú félbeszakadásának helyét is rögzítették. Eszerint már a beszélgetés első szakaszában, még a módszerek különbségét okozó adminisztratív kérdések előtt megfigyelhető a legutóbbi születésnap módszer előnye a másik két eljárással szemben. A különbség szignifikáns, a kérdezett már a beszélgetés elején nagyobb eséllyel szakítja meg a vonalat, még mielőtt kiderülne a számára, milyen kérdéseket tesznek majd fel neki. Sőt, az interjúk későbbi részében már nem mutatható ki szignifikáns eltérés a három módszer visszautasítási aránya között. A felmérést vezetők hipotézise szerint a kérdezők hozzáállása okozza a magasabb visszautasítási arányt. A kérdezők a Kish és a legutóbbi születésnappal szemben negatív attitűddel viseltetnek, mert ezekkel az eljárásokkal végezve a kiválasztást, gyakrabban kerülnek kényelmetlen helyzetbe. (Például a Kish módszernél a kérdezettek gyakran visszakérdeznek: mi is a célja ezeknek a részletes információknak.) A legutóbbi módszer pedig a már ismertetett félreértéseket eredményezi (nem annak van a legutóbbi születésnapja aki a legutóbb született). Az eltérések fő forrása tehát vélhetően a kérdezők magatartása. Annak eldöntésére, hogy ez általános kérdezői probléma-e, vagy csupán néhány kérdező okozza, khi-négyzet tesztet végeztek. Eszerint a 17-ből csak egyetlen kérdező különbözik szignifikánsan a többitől, vagyis általános kérdezi magatartásról van szó. A kiválasztási eljárások értékelésekor tehát figyelembe kell venni a kérdezők hatását is. Képzésük során érdemes ezekre a problémákra koncentrálni.

A kérdezőbiztosok eddig figyelembe nem vett tényezője okozhatja a módszerek értékeléséről hozzáférhető eltérő szakirodalmi eredményeket is. Mivel a módszerek alkalmazhatóságának vélhetően erős a kulturális meghatározottsága, csak korlátozottan általánosíthatók az előbbi tapasztalatok a magyarországi viszonyokra. Hazai, célzott összehasonlításról nem áll rendelkezésemre hozzáférhető információt. A Kish-kulcsot az alábbiakban ezért nem jellemzem más módszerrel összevetve.

7. A minta reprezentativitása

A fentiekben a Kish-módszert gyakorlati megvalósíthatósága felől közelítettem meg. A következőkben megkísérlem az elméleti, statisztikai értelemben vett jóságát körüljárni.

A mintavételt tárgyaló 3. fejezetben említettem, miért javít a becslés hatékonyságán a minta különböző demográfiai jellemzők szerinti reprezentativitása. Általános kutatói gyakorlat a minta ilyen jellemzők szerinti jellemzése, mintegy az eredmények általánosíthatóságát alátámasztva. A hozzáférhető források áttekintésével megpróbáltam a Kish-kulcs alkalmazásával vett minták ilyen szempontú jellemzését rendszerezni. Az esetek legnagyobb hányadában a populációs eloszlásoktól való eltéréseket a fenti fejezetben tárgyalt megvalósíthatósági problémákkal magyarázzák, azaz a demográfiai jellemzők szerinti szisztematikus válaszmegtagadással.

A nem kontrollálható, gyakorlati okokon kívül más forrása is van a problémának. Kish hangsúlyozza, hogy a kulcs a háztartáson belüli egyenlő esélyű kiválasztást hivatott teljesíteni, és ez nem jelenti a valódi bekerülési valószínűségek egyezését. Nyilván, hiába azonos a háztartáson belüli kiválasztási esély, a nagyobb létszámú háztartásokból nehezebb a mintába kerülni, míg az egyszemélyes háztartások lakóját mindenképpen megkérdezzük, ha a háztartása már bekerült a háztartásmintába. A kiválasztási valószínűségek kiegyenlítését Kish súlyozással javasolja megoldani. Ezt a módszert a 8. fejezetben részletesebben elemzem. A kulccsal kapott minták súlyozásáról Kish megállapította, hogy nem okoz jelentős változást a becslésekben, ezért a gyakorlatban elhagyják az elemzésből (Kish, 1949, 1965).

Az általam felkutatott források jó része a minta reprezentativitásának jellemzésekor nem veszi figyelembe az eltérő bekerülési esélyből fakadó elvi szintű problémákat. Ezen elvi megfontolások szerint ugyanis ha a háztartások nagyságától függ a bekerülési esély, a háztartások nagysága viszont karakteres jellegzetességeket hordoz bizonyos demográfiai jellemzők szerint, akkor a mintavétel utolsó lépcsője okozza a szisztematikus demográfiai eltéréseket. Vagyis ekkor akár tökéletes véletlen háztartásminta, és nem szisztematikus válaszmegtagadás mellett sem kapunk reprezentatív mintát. Szélsőséges példával: akkor is "rossz" mintát kapunk, ha minden célpopulációbeli háztartást bevonunk a mintába (kiküszöbölve a háztartásválasztás mintavételi hibáit), és tökéletes együttműködést tételezünk fel a kérdezettek részéről.

Néhány példát választottam olyan kutatók elemzéséből, akik Kish-kulccsal vagy Kish-kulcs variánssal vett mintájuk nem megfelelő reprezentativitását magyarázzák. Magam kommentárként minden esethez olyan háztartásszerkezet-modellt rendeltem, melyben várhatóan a kapott minta valósul meg (azaz nincs szükség gyakorlati megvalósítási problémák feltételezésére). Természetesen tényleges értékelés csak az aktuális háztartásszerekezeti viszonyok figyelembe vételével történhet.

Az 1997-es SCSU Survey (Minnesota State Lottery. http://www.lottery.state.mn.us/gambling/stcloud.html) Egyesült Államok-beli telefonos felmérés volt. Véletlenszám módszerrel választották a felhívott telefonszámokat, földrajzilag rétegzett kiválasztással. A Hagen-Collier módszert használták. Erre a módszerre a Kish-kulcs kapcsán fent részletezett, a háztartások demográfiai jellenzőire vonatkozó megfontolások ugyanúgy érvényesek. Azt találták, hogy a nők kissé felülreprezetáltak a mintában, hozzátéve, hogy ugyanez a jelenség fordul elő minden telefonos kutatásban. Az eredmény azonban anélkül is magyarázható, hogy a nők nagyobb válaszolási hajlandósága mellett érvelnénk. Olyan háztartásmodell mellett várható ez a minta, ahol a háztartásnagyság befolyásolja a nemi összetételt. Így például ha gyakoriak az egyszemélyes háztartások, és az egyszemélyes háztartások nagy részét nők lakják, az eredmény nem meglepő. Ekkor ugyanis nagyobb valószínűséggel választunk nőt, mint férfit.

A közigazgatás megítélése című felmérést 1999-ben végezte a MONITOR Társadalomkutató Intézet. A mintát véletlen sétás módszerrel és Kish-kulccsal választották. Az elemzés szerint "a véletlen kiválasztás miatt a minta összetétele a nemek és az életkor szerint kismértékben eltér a népszámlálási adatok alapján várhatótól. Ezeket az eltéréseket az 1996-os mikrocenzus felhasználásával (...) korrigáltuk." (http://www.monitor-tki.hu/) A véletlen kiválasztás miatt, azaz a mintavételi hibából eredően valóban történhet eltérés. De a talált demográfiai eltérés szisztematikus is lehet, a háztartásstruktúrára visszavezethetően. Mint a következő fejezetben látni fogjuk, a magyar háztartások létszáma erősen összefügg nem- és korösszetételükkel, így a háztartáslétszámmal fordítottan arányos kiválasztási valószínűség tozítja a minta e két változó szerinti összetételét.

Kennedy viszont nem feledkezik meg a háztartásstruktúra hatásáról az 1993-ban az American Association for Public Opinion Research (AAPOR) éves találkozóján elhangzott előadásában (http://www.indiana.edu/ csr/aapor93.html). Expliciten, a háztartásszerkezetből vezeti le a minta demográfiai jellemzőit egy 1990-es Indiana állambeli felmérést elemezve. A kutatásban a Kish-kulcs, a Hagen-Collier, az legutóbbi és a legközelebbi születésnap módszert hasonlította össze. Mind a négy eljárás olyan mintát eredményezett, melyben a 65 éven felüliek a populációs arányukat felülmúlva voltak reprezentálva - a

magyarázat szerint azért, mert az idősebbek kisebb háztartásokban élnek. Hasonlóan a fiatal felnőttek alacsony arányának indokát a szerző abban keresi, hogy a fiatalok inkább sokfelnőttes háztartások tagjai - szüleikkel, vagy más fiatalokkal osztják meg otthonukat. A Kennedy által összehasonlított négy kiválasztási eljárás közös vonása, hogy pontosan egy felnőttet választanak ki, elméletileg háztartáson belül egyenlő eséllyel. Emiatt mindegyikük ugyanazt a szisztematikus torzítást produkálja, és ez a torzítás az amerikai háztartásnagyság és a korcsoportok kapcsolatából ered.

Kish tapasztalata a súlyozás elhagyhatóságáról a korabeli Egyesült Államok-beli háztartásszerekzetből eredeztethető. Más körülmények között -más országban, más korban- a minta elégtelen reprezentativitása esetleg már magából a mintavételi módszerből levezethető, és nem feltétlenül a szisztematikus válaszmegtagadás az előidézője a torzulásoknak.

Az aktuális populációs háztartásszerkezet tehát nagyban befolyásolja a Kish-kulccsal vett minta demográfiai jellemzőit. A továbbiakban áttekintem a mai magyarországi háztartások tulajdonságait, és ezzel vetem össze egy Kish-kulccsal készült, Magyarországon felvett felmérés nem és kor szerinti reprezentativitását.

7.1. A magyarországi háztartásszerkezet jellemzői

A magyar háztartások lényeges változáson mentek keresztül a 80-as évek folyamán. Nagy háztartásszámnövekedés volt megfigyelhető, ezt a generációk különélését előidéző tendenciákkal és a megnövekedett válások miatti családszétaprózódással magyarázhatjuk. Megnövekedett azoknak a háztartásoknak a száma, ahol a szűk családmag él együtt, illetve ahol egy személy él egyedül. A 90-es évekre ezek a nagy változást előidéző mozgások lényegében elcsitultak, de az egyszemélyes háztartások aránya számottevően tovább nőtt. 1996-ban 100 háztartásra 258 személy jut. A 7. táblázatban az 1980-as és az 1990-es népszámlálás, illetve az 1996-os Központi Statisztikai Hivatal mikrocenzusának eredményei találhatók[10]: a háztartások taglétszám szerinti százalékos megoszlása aszerint, hogy hány fős háztartásban élnek.

7. táblázat: A magyarországi háztartások megoszlása taglétszámuk szerint

	Háztartások		
taglétszám	1980	1990	1996
1	19,6	24,3	26,1
2	28,1	29,0	28,2
3	22,3	20,6	19,8
4	19,5	18,0	17,5
5	6,9	5,6	5,7
6	2,4	1,6	1,8
7	0,8	0,5	0,6
8+	0,5	0,3	0,4
együtt	100	100	100
	(n=3,719M)	(n=3,890M)	(n=3,869M)

1996-ban a háztartások több mint negyede egyetlen lakóval rendelkezik, számuk meghaladja az egymilliót. A háztartások 92%-a legfeljebb négy lakóval rendelkezik. A háztartásban élők megoszlása háztartásuk taglétszáma szerint az Y. táblázatban látható. Háztartásban él 1996-ban a népesség 98%-a, 2%-uk, mintegy 203 ezer fő intézeti háztartásokban él (diákszálló, kollégium, idősek otthona). Minden tizedik ember egyszemélyes háztartásban él, a népesség 82%-a él négy vagy kevesebb személyt magában foglaló háztartásban.

[10] A KSH háztartás-definíciója szerint háztartást alkotnak az együtt élő személyek, ha megélhetési költségeiket legalább részben együtt viselik, s ha a hét egy vagy több napján együtt étkeznek.

8. táblázat: Magyarországi háztartásban élők megoszlása háztartásuk létszáma szerint

	Háztartásban élők		
taglétszám	1980	1990	1996
1	7,0	9,3	10,1
2	20,1	22,2	21,8
3	24,0	23,8	23,0
4	27,9	27,7	27,0
5	12,3	10,9	11,1
6	5,1	3,8	4,2
7	1,9	1,3	1,6
8+	1,6	1,0	1,2
együtt	100	100	100
	(n=10,377M)	(n=10,124M)	(n=9,995M)

A népesség öregedési folyamata megváltoztatta a háztartások korösszetételét is. 1980 és 1996 között 12%-ról 8,5%-ra esett vissza a csak 30 évesnél fiatalabb személyekből álló háztartások aránya. A csak 60 évesekből vagy idősebbekből álló háztartások aránya viszont 6 százalékpontos növekedéssel 24%-ra emelkedett. A háromgenerációs háztartások hányada 5%-ra esett vissza, 3 százalékpontos csökkenéssel.

9. táblázat: Az egyszemélyes háztartások nem- és korösszetétele

	-39	40-59	60-	együtt
férfi	9,71	12,59	12,48	34,79
nő	6,03	13,81	45,38	65,21
együtt	15,74	26,40	57,86	100 (n=1010147)

A 9. táblázat az egyszemélyes háztartások nem és kor összetételét jellemezi. A táblázat szerint legalább 60 éves az egyszemélyes háztartások lakóinak 58%-a. Az egyszemélyes háztartások kétharmadában nő él egyedül, méghozzá az összes egyszemélyes háztartás 45%-ában legalább 60 éves nő. A 60 év feletti nők 38%-a egyedül él, nagyrészt özvegyen, a magyarországi várható élettartam tendenciák következményeként.

Mint láttuk, a magyar háztartások struktúrája a 80-as években lezajlott nagy változások óta viszonylag stabil. A 2001-es népszámlálás eredménye még várat magára, de ha ezt a stabilitást tartósnak tételezzük fel, az 1996-os mikrocenzus adataira ma is támaszkodhatunk. A következő fejezetekben több, az utóbbi időben végzett felmérés mintáját elemezzük, és ehhez az 1996-os mikrocenzust használjuk fel.

Érdemes rövid összehasonlítást végezni a mikrocenzus szerinti magyar háztartásszerkezet és a Kish--kulcs létrehozáskor fennálló amerikai háztartásstruktúra között. Mint említettem, Kish a minta elemzésekor alkalmazandó súlyozás elhagyhatósága mellett döntött (Kish. 1965). Állásfoglalását a súlyozott és a súlyozatlan becslések kis különbségére alapozta, a kis különbségek okaként pedig a korabeli háztartásszerkezet jellemzőit jelölte meg. Az általa hivatkozott háztartási adatok a következők:

10. táblázat: Háztartások aránya felnőtt tagjaik szerint (1957, Egyesült Államok)

Felnőttek száma a háztartásban	1	2	3	4	5	6 vagy több
A háztartások aránya	0,146	0,730	0,090	0,028	0,004	0,002

Kish értékelése szerint a kiválasztási valószínűségek szórása nagyon kicsi, mivel a háztartások zöme két felnőtt taggal rendelkezik, a többiek közül pedig gyakorlatilag mind 1, 3, vagy négy felnőtt lakót számlál.

Ezzel szemben a hazai, 1996-os mikrocenzus szerint a magyar háztartások 26,1%-a egytagú. Ez csaknem kétszerese a Kish által tapasztalt aránynak. Bár a nagylétszámú háztartások nálunk sem képviseltetik magukat jelentős számban, ez az eltérés önmagában is olyan mértékű, hogy óvatosnak kell lenni a Kish-kulcs súlyozás nélküli honosításával. A következő fejezetekben a hazai alkalmazhatóságot vizsgálom.

7.2. Kish-kulccsal vett minták reprezentativitása hazai felmérésekben

A Társadalomkutatási Informatikai Egyesülés 1994-ben módszertani célú közvélemény-kutatást végzett a 17 évesnél idősebb budapesti lakosok körében egy 500 fős telefonos és egy 500 fős személyes megkeresésen alapuló mintán. A személyes megkérdezés mintaválasztásánál többlépcsős valószínűségi mintavételt alkalmaztak, a harmadik lépcsőjében Kish-kulccsal választották ki a kérdezendőt. Tarjányi József (Módszertani problémák a telefonos közvélemény-kutatásokban) elemzése szerint: "A mintát az 1990-es

népszámlálás főbb demográfiai adatai (nem, iskolai végzettség, korcsoport) alapján teszteltük. Nem szerint -mint a fenti módon választott minták általában- a minta összetétele a nőknek a népszámlálásnál 2,8%-kal nagyobb arányát mutatja. Vizsgálataink szerint ebben szerepet játszhat az is, hogy a lakásminta megnöveli az egyszemélyes háztartások mintába kerülési esélyét, ebben a háztartástípusban pedig a nők vannak túlsúlyban. A hat korcsoportra osztott kor szerinti illeszkedés 95%-os szinten elfogadható, és különösen jó a közép korosztályok csoportjaiban. 60 év felett a várható értéknél némileg többen, 30 év alatt valamivel kevesebben kerültek a mintába. Az iskolai végzettség szerint a minta nem illeszkedik a népszámlálás adataihoz. Különösen felülreprezentáltak a szakmunkásképzőt, ill. a főiskolát végzettek, és különösen kevesen vannak a várható számhoz viszonyítva a nyolc általánosnál kevesebbet és a nyolc általánost végzettek. ”

Tarjányi a populációs arányokhoz képest a nemek és a korcsoportok szerint is eltéréseket talált. A nők felülreprezentáltságát az egyszemélyes háztartások nagyobb mintába kerülési esélyével magyarázza. Vélhetően a háztartások azonos mintába kerülési esélyei mellett is a nők kerülnek be nagyobb számban a mintába, mivel a háztartások 26%-a egyszemélyes (az 1996-os adatokat használva), és valóban, ebben a háztartástípusban 65%-os arányban inkább a nők képviseltetik magukat. A fiatalok alulreprezentáltsága és az idősek nagyobb száma szintén háztartásszerkezeti jellemzőkre vezethető vissza, hiszen, mint láttuk, például 8,5% a csak 30 évesnél fiatalabb személyekből álló háztartások aránya, míg a csak 60 évesekből vagy idősebbekből álló háztartások aránya 24%. Ezek a számok csak kiragadott jellemzők, szisztematikus vizsgálat csak úgy lenne kivitelezhető, ha a Kish-kulccsal vett átlagos mintát a teljes országos háztartásstruktúra függvényében írnánk le. Erre teszek kísérletet a következőkben.

Az ISSP Család II. felmérést a Tárki végezte 1994-ben. Az 1990-es KSH Népszámlálás lakásmintájából készült országos reprezentatív mintát használták, és a Kish-kulcsot alkalmazták a kérdezendő kiválasztására. A mintaelemszám 1500 fő. A háztartásokat véletlen (vagy legalábbis véletlent elérni akaró) módon választották, majd minden háztartásból pontosan egy tagot jelöltek ki. Jól jellemezi ennek a módszernek a kikerülhetetlen mintatorzító hatását már a minta háztartási struktúrája is. A mikrocenzus szerint a népesség 10,1%-a él egytagú háztartásban, ezzel szemben a Család II. mintájának 19.7%-a él egyedül. Viszont ha felhasználjuk a kérdezettek háztartáslétszámáról rendelkezésre álló információkat, kiderül, hogy az összes háztartástagnak csupán 8,46%-a tartozik egytagú háztartáshoz. A háztartásminta ilyen szempontból még nem tért el jelentősen a populációtól, de a háztartástagok kiválasztása után már jelentős eltérés tapasztalható. Fontos szem előtt tartani, hogy ez a differencia elkerülhetetlen, ha a mintavételi eljárásunk szerint véletlen módon választunk háztartást, majd a háztartásból pontosan egy embert veszünk ki. A Kish-kulcs azonban tovább specifikálja ezt a módszert, a nem- és kor szerinti kiválasztás folytán vélhetően a minta nem- és kormegoszlására is befolyással van.

Az alábbiakban a minta nem- és korösszetételét vizsgáljuk. A 11. táblázat a minta korcsoport- és nem szerinti százalékos megoszlását mutatja, míg a 12. táblában az 1994. év eleji populációs arányokat[11] láthatjuk.

11. táblázat: A Család II. mintájának nem- és korösszetétele

	nem		
korcsoport	férfi	nő	együtt
18-39 éves	16,47	18,27	34,73
40-59 éves	14,93	19,40	34,33
60- éves	10,93	20,00	30,93
együtt	42,33	57,67	100 (n=1500)

12. táblázat: A magyarországi populáció nem- és korösszetétele 1994 elején

	nem		
korcsoport	férfi	nő	együtt
18-39 éves	20,64	20,11	40,75
40-59 éves	16,24	17,73	33,97
60- éves	10,02	15,26	25,28
együtt	46,90	53,10	100 (n=7854559)

[11] A Magyar Statisztikai Évkönyv 1993. alapján

Összevetve az arányokat, szembeszökő a férfiak a vártnál alacsonyabb számú előfordulása. A korcsoporokat nézve a középkorúak a populációs arányuknak megfelelően találhatók a mintában, míg a fiatalok alul-, az idősebbek felülreprezentáltak. Megfigyeléseink megegyeznek a 7.2. fejezetben hivatkozott Tárki módszertani felmérés Tarjányi József közölte tapasztalataival. Az eltérés a hármas korcsoport változót tekintve szignifikáns (khi-négyzet teszttel vizsgálva, 0,000 szinten). A minta és a populáció különbözősége a nemeket tekintve is szignifikáns (khi-négyzet teszttel, 0,000 szinten). Ha nem csak a marginálisok eltérését vizsgáljuk, hanem egyszerre a táblázat hat cellájának illeszkedését[12], megint csak szignifikáns eltérést kapunk (0,000 szinten).

A kulcs torzító hatását akkor lehetne érdemben elemezni, ha a többlépcsős mintavétel lépcsőnként lenne jellemezhető. Ugyanis a fenti eltérések csak akkor tulajdoníthatók a végső lépcsőbeli kiválasztásnak, ha garantálni tudjuk, hogy a háztartások egyforma valószínűséggel kerültek a mintába. Ez a gyakorlatban is kevéssé biztosítható, a kérdezőbiztosok ellenőrzésének költségigénye miatt.

Viszont megkísérelhető a kulcs okozta hatás vizsgálata, ha ismert a minta háztartásainak összes tagjából összeállított pszeudominta összetétele. Ekkor ugyanis ennek a pszeudomintának a nem- és korösszetételéhez viszonyíthatjuk a tényleges mintabeli megoszlásokat. Ráadásul a pszeudominta és a populációs adatok összevetésével a háztartásminta kiválasztási lépcsője is tesztelhető.

A Család felmérés esetében a rendelkezésemre állt nem csak a kérdezettek, hanem a háztartások minden tagjának nemére és korára vonatkozó információ is. A pszeudominta kialakításának akadálya lehetne a magas válaszmegtagadás ezekkel a kérdésekkel kapcsolatban. Ám arra a kérdésre, hogy hány felnőtt lakója van a lakásnak, mindössze 3 esetben nem kaptunk választ. 47 esetben nem ismert valamelyik lakó neme vagy kora. Ezekkel az elhanyagolható hiányokkal, az eredeti 1500-as mintából 4397 darab használható személyi rekordhoz jutunk, a kérdezettek közül közülük 3398-an töltötték be a 18-adik életévüket.

A Családból kapott "pszeudominta" arányainak százalékos megoszlása:

13 táblázat: A Család II "pszeudomintájának" nem- és korösszetétele

	nem		
korcsoport	férfi	nő	együtt
18-39 éves	19,47	19,92	39,39
40-59 éves	16,52	19,21	35,73
60- éves	10,00	14,87	25,23
együtt	46,00	54,00	100 (n=3389)

A pszeudominta és az eredeti személyi minta összehasonlításában a tendenciák egyeznek az előbbi összehasonlításkor kapottal: a középkorúak illeszkedése a legjobb, a személyi mintában a másikhoz képest felülreprezentáltak az idősek, míg a fiatal korosztály alulreprezentált. Az eltérés szignifikáns (khi-négyzet teszttel mérve, 0,000 szinten). A nemi összetételt vizsgálva, a nők itt is felülreprezentáltak a személyi mintában a háztartásmintához képest. Az eltérés a nemeket tekintve is szignifikáns (khi-négyzet teszttel, 4 ezrelékes szinten). A marginálisokon kívül a hat kategória illeszkedése is szignifikáns eltérést mutat (0,000 szinten).

A két összehasonlítás tapasztalatai egyeznek, mert a pszeudominta igen jól illeszkedik a populációhoz. Közelebb van a populációs adatokhoz, mint az eredeti minta. Kismértékben most is alulreprezentáltak a férfiak, de a nem változó eltérése nem szignifikáns (p=0,31). A marginálisokon kívül a cellákat vizsgálva kitűnik, hogy a fiatal férfiak alacsony képviselete okozza az eltérést. Kevesebb fiatalt és több középkorút találunk a pszeudomintában, mint amit az országos megoszlás alapján várnánk. A középkorú nők nagyobb arányú képviselete okozza az eltérést. Viszont a korcsoportok eltérése sem szignifikáns (p=0,09). A cellákat vizsgálva, a hatkategóriás változó illeszkedése is elfogadható (0,222 szinten).

Eredményeink interpretációja lehetővé teszi a mintaválasztás egymástól szétválasztott lépcsőfokainak értékelését. A pszeudominta jól illeszkedik a várható arányokhoz. Eszerint a háztartásválasztás véletlen volta nem cáfolható, azaz a háztartásminta még reprezentatívnak minősíthető. A személyválasztási eljárást, azaz a kulcsot a hiba okai között önálló tényezőként különítettük el.

Vélhetően a minta (illetve a populáció) háztartásszerkezete idézi elő az eltérést. Ha ugyanis pl. az

[12] A két nem és három korcsoport kategória keresztkombinációibók képzett hatértékű változó eloszlására végeztem khi--négyzet tesztet.

idős egyedüdálló nők aránytalanul nagy hányadát alkotják a lakosságnak, akkor bármely, a háztartástagok között egyenlő valószínűséggel választó mintavétel előidézi azokat a szisztematikus hibákat, amiket fent találtunk. Hipotézisünk szerint tehát ha minden háztartásból pontosan egy személyt választunk - azonos bekerülési esélyeket biztosítva minden háztartástagnak - ugyanezeket a rendszerszerű hibákat találjuk.

Teszteljük a hipotézisünket. Válasszunk tökéletesen véletlen személyi mintát a Család háztartásmintájából a számítógépes véletlenszámgenerálás lehetőségét kihasználva. Teljesüljön a két feltétel: minden háztartásból pontosan egy személyt választunk, és a háztartáson belül egyforma eséllyel választjuk ki bármelyik személyt. A kapott minta összetétele:

14. táblázat: A Család II. háztartásaiból random módon választva

	nem		
korcsoport	férfi	nő	együtt
18-39 éves	18,60	18,80	37,40
40-59 éves	13,53	18,80	32,33
60- éves	10,27	20,00	30,27
együtt	42,40	57,60	100 (1500)

Összehasonlítva a pszeudominta nem/kor arányaival, az eltérések tendenciája most is ugyanaz mint a valódi minta eltérésekor. A nők felülreprezentáltak, az idősek is, míg a fiatal korosztály a kellőnél kisebb arányban képviseltetei magát. A korcsoportok gyakoriságának eltérése a háztartásmintáétól most is szignifikáns (és most is 0,000 szinten). Ugyanígy a nemek arányának eltérése is jelentős (0,006 szinten szignifikáns, a Kish-kulccsal választott minta eltérése 4 ezrelékes szinten volt az). A marginálisok után a hat cella illeszkedését egyszerre vizsgálva is szignifikáns eltérést találtam (0,000 szinten).

7.3. A kiválasztás szimulációja - Kish-módszer

A Kish kulccsal vett és a számítógéppel generált véletlen minta reprezentativitását hasonlítottuk össze az előbb. Az eredményt megkérdőjelezheti az, hogy csak 1-1 mintát vizsgáltunk, így a minták összetétele akár a véletlen egyszeri játékával is magyarázható. Tegyük most ezt az összehasonlítást jobban általánosíthatóvá. Végezzünk szimulációsorozatot, egy nagyobb háztartásmintán. Alkalmazzuk egymás után többször a Kish-kulcsot is, a számítógépes véletlen mintát is, és hasonlítsuk össze a minták tipikus viselkedését. Igy az előbb talált hibák változékonysága is tesztelhető.

A vizsgálat során a háztartások kiválasztása is random módon történik. Ezért a kapott minta összetételét csak maga a kiválasztási eljárás befolyásolhatja torzító módon, nem lép fel kérdezői hatás, sem mérési hiba stb. A háztartások választásának lépcsője önállóan nem lép fel torzító hatásként, a háztartásminta lakóinak összetétele várhatóan egyezik a populációval. Bár előfordulhat például, hogy az egyik mintába a véletlen játékaként csak egyszemélyes háztartások kerülnek be, de átlagosan - az összes mintát figyelembe véve - ez a hatás nem okoz eltérést a populációhoz képest. Tehát ha a szimulációk eredményeként kapott minták nagy részének összetétele mégis eltér a populációétól, az a háztartáson belüli kiválasztás módjának tudható be.

Elemzésemben mintaként a Tárki adatbázisát, a Magyar Háztartási Panel IV. hullám felmérést használtam. Ez országosan reprezentatív háztartási minta, 6306 személy adatait tartalmazza a bázis. Ezek közül 813 esetben hiányzott valamelyik háztartástag koráról vagy neméről az információ[13], ezeket a könnyebb elemzés érdekében kizártam az elemzésből. Közülük 4188-an töltötték be 1994 folyamán a 18-adik életévüket. Így 4188 személy - 1986 háztartás - adatait használtam fel.

Az elemzés során mintavételi szimulációkat végzek, azaz úgy kezelem a mintát, mintha egy virtuális populáció lenne. A továbbiakban a "minta" megnevezést is "populáció"-ra váltom. Ezen a populáción tesztelem a Kish-kulcs használatának hatását a minták nem- és korösszetételére. A Kish-módszer után, mint azt fent is tettem, az egyszerű random kiválasztást tesztelem. A két módszer eredménye ugyanis elméletileg sem egyezik meg feltétlenül. Az 5. fejezetben, a Kish-táblák elemzésekor említettem, hogy több ponton sérül a használatukkal elérni kívánt háztartásokon belüli egyenlő esélyű kiválasztás. Az első

[13]Nem csak válaszmegtagadás miatt: nem vettek fel adatokat a 16 éven aluliakról; a háztartásból kivált, de még nyilvántartott személyekről stb.

probléma, hogy háztartásonként csak az első 6 személy kerülhet be a mintába. A másik "hiba" az 5 tagú háztartások tagjainak egyenlőtlen kiválasztási valószínűségeiből ered (lásd 4. táblázat).

Most a (virtuális) populáció nem- és korösszetétele:

15. táblázat: A Háztartás Panelből kapott populáció összetétele

	nem		
korcsoport	férfi	nő	együtt
18-39 éves	19,79	20,51	40,31
40-59 éves	14,66	17,60	32,26
60- éves	10,94	16,50	27,44
együtt	45,39	54,61	100 (n=4188)

A minta elkészítésekor először a háztartások közül választottam véletlen módon tíz százalékos mintát, ez 199 háztartás kiválasztását jelenti. A háztartásmintából ezután háztartásonként választottam ki a Kish-kulccsal egy személyt. A kulcsnak a 3. táblázatban leírt változatát használtam.

Mivel tendeciózus hatásokat szerettem volna vizsgálni, több mintavétel szimulációját hajtottam végre egymás után: száz mintát állítottam elő. A minták nemi összetételét vizsgáltam elsőként. A férfiak arányának eloszlása a száz mintában a következő volt:

16. táblázat: A férfiak arányának eloszlása - háztartáson belül Kish-kulcsos választás

	Percentilis	Szélsőértékek		
1%	0,3668	Legkisebbek		
5%	0,3894	0,3518		
10%	0,3920	0,3819	Minta elemszáma	100
25%	0,4070	0,3819		
		0,3869	Átlag	0,4267
50%	0,4271		Szórás	0,0256
		Legnagyobbak	Variancia	0,0007
75%	0,4422	0,4774	Ferdeség	0,1370
90%	0,4573	0,4824	Csűcsosság	3,2951
95%	0,4749	0,4874		
99%	0,4925	0,4975		

A populációban a férfiak aránya 45,39%. A kapott minták férfiarányának átlaga 42,67%, vagyis torzító hatás lép fel a Kish-kulcsos kiválasztáskor. A férfiak inkább alulreprezentáltak a mintákban.

A legszélsőségesebb mintákban előfordult 35,18%-os férfiarány, illetve 49,75%-os férfiarány is. Ezenkívül a kapott minták negyedének férfiaránya 40,7% alatt van, míg a felső szélsőséges kvartilis alsó határa 44,22%. A populációs arány (45,39%) a negyedik kvartilis középső régiójában van. Tehát várhatóan a minták több mint háromnegyedében alulreprezentáltak lesznek a férfiak. A minták felében az arány 42,71% alatt, a másik felében ezen érték felett van a férfiak aránya. Az eloszlás kissé ferde a normális eloszlás esetén várthoz képest, de gyakorlatilag mediánja és átlaga megegyezik. Továbbá az eloszlásnál nyújtottab kissé a normális eloszlás (csúcsossági mutató). Összefoglalva elmondhatjuk, hogy viszonylag nagy az értékek variábilitása, az átlagostól eltérő értékek előfordulási esélye számottevő. Ez a mintavételre visszatérve azt jelenti, hogy a választási módszer torzítja a férfiak arányát a mintában, és az arányváltozás mértéke is változó.

A korcsoport szerinti összetételt korcsoportonként külön vizsgálom. A fiatal korosztály (18-39) évesek mintabeli arányainak eloszlása:

17. táblázat: A fiatalok arányának eloszlása - háztartáson belül Kish-kulcsos választás

	Percentilis	Szélsőértékek		
1%	0,2965	Legkisebbek		
5%	0,3116	0,2965		
10%	0,3216	0,2965	Minta elemszáma	100
25%	0,3467	0,3065		
		0,3116	Átlag	0,3640
50%	0,3618		Szórás	0,0301
		Legnagyobbak	Variancia	0,0009
75%	0,3819	0,4271	Ferdeség	0,08570
90%	0,4045	0,4322	Csűcsosság	2,9711
95%	0,4121	0,4372		
99%	0,4372	0,4372		

A populációban az első korcsoport aránya 40,31%. A kapott mintákban a fiatalok arányának átlaga 36,40%, azaz torzító hatás lép fel a véletlen választáskor is: ez a korcsoport alulreprezentált.

A legszélsőségesebb mintákban előfordult 29.56%-os arány, illetve 43,72%-os arány is. A kapott minták alsó kvartilise 34,67%, míg a minták egynegyedében 38,19% felett van a fiatalkorúak hányada. A populációs arány a negyedik kvartilis középső régiójában található, tehát hozzávetőlegesen öt mintából négy esetben a fiatalok alulreprezentáltak. A medián 36,18%, a minták felében ez alatt van, a másik felében ezen érték felett van az első korcsoport aránya. Az eloszlás majdnem szimmetrikus. Továbbá az eloszlás kissé nyújtottab a normális eloszlásnál, de az eltérés nem jelentős. Itt is azt mondhatjuk, hogy viszonylag nagy az értékek változékonysága.

A középső korosztály (40-59 évesek) arányának eloszlása a mintákban:

18. táblázat: A középkorűak arányának eloszlása - háztartáson belül Kish-kulcsos választás

	Percentilis	Szélsőértékek		
1%	0,2236	Legkisebbek		
5%	0,2563	0,2161		
10%	0,2613	0,2312	Minta elemszáma	100
25%	0,2764	0,2362		
		0,2462	Átlag	0,3005
50%	0,2965		Szórás	0,0326
		Legnagyobbak	Variancia	0,0011
75%	0,3241	0,3618	Ferdeség	0,0859
90%	0,3467	0,3668	Csűcsosság	2,6214
95%	0,3568	0,3668		
99%	0,3693	0,3719		

A populációban a középső korosztály aránya 32,26%. A kapott minták arányának átlaga 30,05%, vagyis torzító hatás lép fel a véletlen választáskor is. A középkorúak alulreprezentáltak a mintákban, kisebb mértékben ugyan, mint az első korosztály - itt két százalékponttal alacsonyabb az átlagos képviselet, ott átlagosan több mint négy százalékponttal volt kisebb az arány.

A legszélsőségesebb mintákban előfordult 21,61%-os, illetve 37,19%-os arány is. Ezenkívül a kapott minták negyedének középkorú-aránya 27,64% alatt van, míg a felső kvartilis alsó határa 32,41%. A populációs arány a harmadik kvartilisben található. Az eloszlás kissé ferde a normális eloszlás esetén várthoz képest, a medián jobb odalán laposabb a sűrűségfüggvény. Az eloszlás csúcsossági mutatója kisebb, mint a normális eloszlásé. Azt mondhatjuk, hogy viszonylag nagy az értékek variábilitása, az átlagostól eltérő értékek előfordulási esélye számottevő.

Mivel az alsó két korosztály átlagosan alulreprezentált a mintákban, ennek következményeként az idősebbek a populációs hányaduknál várhatóan nagyobb számarányt képviselnek:

19. táblázat: Az idősebbek arányának eloszlása - háztartáson belül Kish-kulcsos választás

	Percentilis	Szélsőértékek		
1%	0,2739	Legkisebbek		
5%	0,2814	0,2714		
10%	0,2915	0,2764	Minta elemszáma	100
25%	0,3116	0,2764		
		0,2814	Átlag	0,3355
50%	0,3317		Szórás	0,0335
		Legnagyobbak	Variancia	0,0011
75%	0,3618	0,3969	Ferdeség	0,1991
90%	0,3769	0,4070	Csűcsosság	2,4752
95%	0,3945	0,4070		
99%	0,4146	0,4221		

A populációban a harmadik korcsoport aránya 27,44%. A kapott mintákban az idősek arányának átlaga 33,55%, ez a több, mint hat százalékpontos növekedés jóval jelentősebb, mint a fenti esetek eltérése. Vagyis jelentős torzító hatás lép fel a harmadik korcsoport arányában.

A legszélsőségesebb mintákban előfordult 27,14%-os arány, illetve 42,21%-os arány is. A kapott minták alsó kvartilise 31,16%, míg a minták egynegyedében 36,18% felett van az időskorúak hányada. A populációs arány az első percentilisek között található, átlagosan huszonöt mintából huszonnégy esetben az idősek felülreprezentáltak. A medián 33,17%, azaz a minták felében ez alatt van, a másik felében ezen érték felett van a harmadik korcsoport aránya. Az eloszlás kissé jobbra ferdül, a nagyobb értékek jobban szórtak, mint a mediántól balra levők. A sűrűségfüggvény laposabb, mint a normális eloszlás sűrűségfüggvényénye. Itt is azt mondhatjuk, hogy viszonylag nagy az értékek változékonysága.

Eddig csak a marginálisok szerinti illeszkedést vizsgáltuk. A cellák megoszlásának illeszkedését a nem és korcsoport kategóriakombinációkkal kapott hat értékű változóra alkalmazott khi-négyzet tesztekkel ellenőriztem. A teszteredmények a száz esetből 31 esetben mutattak szignifikáns eltérést a szokásos 5%-os szinten. A részletesebb jellemzés érdekében az alábbi táblázat mutatja a száz teszt elért valószínűségi szintjének (a továbbiakban: p értékének) eloszlását:

20. táblázat: A p-k eloszlása a khi-négyzet tesztekből - háztartáson belül Kish-kulcsos választás

	Percentilis	Szélsőértékek		
1%	0,000	Legkisebbek		
5%	0,003	0,000		
10%	0,0095	0,000	Minta elemszáma	100
25%	0,031	0,000		
		0,001	Átlag	0,2516
50%	0,138		Szórás	0,2641
		Legnagyobbak	Variancia	0,0698
75%	0,4395	0,824	Ferdeség	0,9895
90%	0,7195	0,853	Csűcsosság	2,7562
95%	0,791	0,895		
99%	0,8965	0,898		

A minták p-értékeinek fele 0,138 alatt van; és mintegy negyedük esetén szignifikáns a populációtól vett eltérés 0,031-ös szinten.

Eredményeink szerint torzító hatás lép fel a Kish-kulcsot alkalmazó választáskor. Nagy valószínűséggel igaz, hogy ezt a torzítást maga a mintavétel módja okozza: az, hogy minden háztartásból választunk, és hogy a háztartásokon belül pedig Kish-módszerrel. Vélhetően a populáció háztartásszerkezete olyan, hogy ezt a módszert alkalmazva, a fenti torzítások jönnek létre a mintában. A Kish-kulcs eredeti célja a közelítően azonos háztartáson belüli kiválasztási valószínűségek biztosítása volt. Felmerülhet a kérdés, hogy milyen eredményt kapnánk, ha megfelelő eszközökkel valóban egyenlő esélyt biztosítanánk adott háztartás tagjainak a mintába kerüléshez. A következő fejezetben ezt a problémát elemzem.

7.4. A kiválasztás szimulációja - random választás

Számítógépes véletlenszám-generálás segítségével megoldható, hogy tökéletesen random módon[14] válasszunk ki háztartásonként egy kérdezettet. Száz ilyen mintaválasztást szimuláltam egymás után. Egy ilyen minta elkészítésekor először a háztartások közül választottam ismét véletlen módon tíz százalékos mintát. A háztartásmintából ezután háztartásonként választottam ki random módon egy személyt. A mintanagyság a fenti, Kish-kulcsot használó esetben pontosan ennyi volt, így könnyen összevethetők az eredmények. A kapott minták nem/korcsoport eloszlását először ismét a marginálisok mentén jellemzem. A Függelék 1. táblázata a férfiak arányának eloszlását jellemzi. A Kish-kulcsos kiválasztáshoz képest most a populációs férfiaránytól (45,39%) kissé még távolabbi értékeket kaptunk, az arány eloszlásának kvartilisei, az arány átlaga még kisebb, mint a Kish-módszernél volt. A férfiak még inkább alulreprezentáltak a mintákban. Ez a mintavételre visszatérve azt jelenti, hogy a random választási módszer is torzítja a férfiak arányát a mintában

A fiatal korosztály arányának megoszlása a mintákban a Függelék 2. táblájában kerül részletesen jellemzésre. A Kish-kulcs mintáiban általában populációs számarányuk alatt képviseltették magukat a fiatalok, a random kiválasztás esetén ez a torzítás kissé megnövekszik. Az átlagos fiatalarány, vagy a 100 minta arányszámainak negyedelőpontjai még kisebb értéket vesznek fel, mint a Kish-kulcs esetében.

A középső korosztály arányának eloszlása a Függelék 3. táblájában található. Összevetve az eloszlás jellemzőit a Kish-kulccsal kapott arányok eloszlásának jellemzőivel, csaknem azonos értékeket találunk. A random kiválasztás is alulreprezentálja a második korosztályt.

Az idősebbek a populációs hányaduknál várhatóan megint nagyobb számarányt képviselnek. A Függelék 4. táblázata tartalmazza az arány eloszlásának részletes mutatóit. A populációban a harmadik korcsoport aránya 27,44%. A kapott mintákban az idősek arányának átlaga 33,23%, ez a csaknem hat százalékpontos növekedés megegyezik a Kish-kulcsnál kapott értékkel. Vagyis jelentős torzító hatás lép fel a véletlen választáskor is.

A cellák megoszlásának illeszkedését ismét khi-négyzet tesztekkel ellenőriztem. A teszteredmények a száz esetből 36 esetben mutattak szignifikáns eltérést a szokásos 5%-os szinten. A részletesebb jellemzés érdekében a Függelék 5. táblázata mutatja a száz teszt elért valószínűségi szintjének eloszlását. A minták p-értékeinek fele 0,0975 alatt van; és mintegy negyedük esetén szignifikáns a populációtól vett eltérés 0,025-ös szinten. A Kish-kulcsot alkalmazva kissé nagyobb értékeket kaptunk, de az eltérés nem jelentős.

Eredményeink szerint torzító hatás lép fel a véletlen választáskor is. Az összehasonlítás szerint a Kish-kulccsal kapott minták reprezentativitása gyakorlatilag megegyezik a véletlen módszerrel kapott mintákéval, bár kis eltérést találtunk a Kish-kulcs javára. A szimuláció alkalmas a mintavételi hibák jó részének kiszűrésére, a szimulációkat sorozatosan végezve pedig várhatóan érvényes következtetéseket tehetünk a minták jellegzetességeire - mert a véletlen hatása tűnik így el. A minták nem/kor szerinti rossz illeszkedése ezért vélhetően magának a mintavételi módnak a következménye. A populáció háztartásszerkezete olyan, hogy ezt a módszert alkalmazva, a fenti torzítások jönnek létre a mintában.

7.5. A populációt változtatva

Kish (65) értékelése szerint több felmérésben is demográfiailag reprezentatív mintát kapott a kulcsot alkalmazva. A fenti, mai magyar adatok ennek az ellenkezőjéről tanúskodnak, Mint írtam, ennek oka, hogy a Kish-kulccsal kapott minta reprezentativitása erősen függ a populációs háztartásstruktúrától. Ezt bizonyítva megismételtem az előző fejezetbeli véletlen választásos elemzést egy másik felmérésen. Ez a felmérés országosan reprezentatív *személyi mintát* produkált, tehát a személyek megközelítőleg egyforma valószínűséggel kerültek a mintába. (A fenti elemzések reprezentatív *háztartásmintán* alapultak, azaz a háztartásoknak volt egyenlő esélye a mintába kerülésre.) Itt nem volt háztartáskiválasztási lépcsőfok, tehát közvetlen személykiválasztás valósult meg. Emiatt a háztartások (felnőtt) taglétszámukkal arányos valószínűséggel kerültek a mintába. Például itt az egyszemélyes háztartások nem a háztartások közti arányukban kerülnek a mintába, hanem az egyszemélyes háztartásokban élők lakosságszámhoz viszonyított arányában. Használjuk fel a mintabeli személyek háztartástársairól rendelkezésre álló nem- és kor-információkat, készítsünk pszeudomintát ezen háztartástagok szerepeltetésével. Ha most ismét

[14] Tulajdonképpen a számítógépes véletlenszám generálás is csak "pszeudorandom", de ettől most eltekinthetünk.

az így kialakított pszeudomintát tekintem populációnak, ennek a populációnak az összetétele lényegesen különbözik a valódi populációétól az előbb tárgyalt, háztartásokra vonatkozó nem egyenlő bekerülési valószínűségek miatt.

Elemzésemben az Országos Lakossági Egészségfelmérés (OLEF-2000) adatbázisát használtam. A mintát a Központi Választási és Népességnyilvántartó Hivatal címlistájából választották. A mintaválasztáskor a nagy települések mintába jutását előre biztosították, majd megye és településnagyság-kategória szerint rétegezve, véletlen csoportos személyi mintát vettek. 5534 személy adatait tartalmazza a bázis. A Család felmérés korábbi elemzésében már alkalmazott módon a személyi mintából olyan pszeudomintát készítettem, mely az eredeti személyekből és háztartásuk tagjaiból áll. Ezek közül elhanyagolható nagyságrendben, 11 esetben hiányzott valamelyik háztartástag koráról vagy neméről az információ, ezeket a könnyebb elemzés érdekében kizártam az elemzésből. Így 5531 háztartás 16960 lakójának adatait használtam fel. Közülük 13275-en töltötték be a 18-adik életévüket.

Mint fentebb már megjegyeztem, az OLEF2000 háztartásmintájának összetétele különbözik a populációétól, különbözik a Háztartási Panel mintájában találttól is. Említettük, hogy a Kish-kulccsal vett minta reprezentativitása erősen függ a populáció háztartásszerkezetétől. Az OLEF2000 háztartásszerkezetének különbözősége várhatóan éppen a kulcs eredményességének növekedése irányába hat. Mint populáció, éppen azokat a tulajdonságait veszítette el, amik az előbbi populáció esetében a minta összetételét torzították. Például a Háztartási Panelben 21%-os az egyszemélyes háztartások aránya, az OLEF2000-ben ezzel szemben csupán 13%-s (az 1996-os mikrocenzus szerint a populációs arány 26%-os). A pszeudomintát, mint populációt használva fel egy mintaválasztási szimulációhoz, a kapott minta nem- és kor arányai várhatóan közelebb lesznek a pszeudominta összetételéhez, mint az előbbi vizsgálatban.

Mintavételi szimulációkat végzek, azaz úgy kezelem a pszeudomintát, mintha egy virtuális populáció lenne. A továbbiakban ezért a "pszeudominta" megnevezést "populáció"-ra váltom. Ezen a populáción tesztelem a véletlen kiválasztás használatának hatását a minta nem- és korösszetételére. Hipotézisem tehát az, hogy a háztartáson belüli véletlen kiválasztás itt kisebb torzulást okoz, a kapott minták kevésbé térnek el a populációtól, mint a Háztartási Panel alkotta populáció esetén.

Száz mintaválasztást szimuláltam egymás után. Egy minta elkészítésekor először a háztartások közül választottam véletlen módon öt százalékos mintát, ez 275 háztartás kiválasztását jelenti. A háztartásmintából ezután háztartásonként választottam ki random módon egy személyt. A kapott minták nem/korcsoport eloszlását viszonyítottam a populációs arányokhoz. Khi-négyzet tesztet végeztem, a hat cella gyakoriságának illeszkedését tesztelve. A száz esetből 26 esetben 5%-os szinten szignifikáns eltérést, szemben a Háztartási Panel alkotta populáció esetén talált 36 szignifikáns eltéréssel. A 100 teszt elért valószínűségi szintjének (p értékének) eloszlását mutatja az alábbi táblázat:

21. táblázat: p-értékek eloszlása háztartáson belüli random választáskor

	Percentilis	Szélsőértékek		
1%	0,0005	Legkisebbek		
5%	0,006	0,000		
10%	0,0105	0,001	Minta elemszáma	100
25%	0,0445	0,002		
		0,002	Átlag	0,2251
50%	0,1705		Szórás	0,2258
		Legnagyobbak	Variancia	0,0510
75%	0,36	0,775	Ferdeség	1,1889
90%	0,585	0,804	Csűcsosság	3,6760
95%	0,6745	0,892		
99%	0,904	0,916		

A Háztartási Panel alkotta populáció esetén kapott eloszlás minden sarokszáma alacsonyabb ezeknél az értékeknél (lásd a Függelék 5. tábláját). Ott az átlag 0,1948 volt, most 0,2251. Ott 0,0975 volt a medián, most 0,1705. Az alsó kvartilis most 0,0445, az előbb 0,025; a felső kvartilis most 0,36, az előbb 0,2915.

Eredményeinket úgy interpretálhatjuk, hogy azok alátámasztják a fent megfogalmazott hipotézist: az aktuális populációnkban kisebb hibát okoz a kiválasztási módszer, mint a Magyar Háztartás Panel populációjában.

8. Súlyozás

Tarjányi József, a Tárki kutatója -szíves közlése szerint- több helyen módosította a Tárki felméréseiben használt Kish-kulcsot. Tapasztalatai szerint a módosítás javított a minta összetételén, de nem szüntette meg a szisztematikus torzításokat. Ezért ma is a javított kulcsot használják, és súlyozással egyenlítik ki az eltéréseket. A súlyokat ő a megvalósult minta jellemzőihez igazítva állapítja meg, a populációs arányokhoz igazítva azokat. Kish viszont csupán a háztartásnagysággal történő súlyozást említette. A kétfajta súlyozás más célokból történik. A továbbiakban a súlyozás módszeréről, a súlyok típusairól és a súlyozás korlátairól írok.

8.1. A súlyozásról általában

A súlyozás széles körben alkalmazott módszer a felmérések elemzésénél, a becslések megadásánál[15]. A súlyozás konkrét metódusa szerint a mintabeli személyek kapott jellemzőjét egy súlyszámmal szorozzuk. Ennek interpretációja például 2-vel történő szorzás esetén: az adott egyén válaszával két személyt képvisel. Rövid bemutatásához vegyünk példának egy olyan esetet, amikor egy változó populációs totálját szeretnénk megbecsülni. Ilyen változó lehet az adott időszak alatt egyénenként kiadott hálapénz összege. Szeretnénk megbecsülni az adott időszak alatt az ország egészségügyi ellátórendszerében mozgó hálapénz összmennyiségét (jelölje X ezt a mennyiséget). Ha a W súlyokat használjuk, a becslést a következő alakban kapjuk:

$\widehat{X} = \sum_u W(u)x(u)$

ahol

u	a felmérés kérdezettje
$x(u)$	az u kérdezett válasza, az általa kiadott hálapénz
$W(u)$	az u kérdezett súlya

A súlyok interpretációja: az u kérdezett nem 1 tagját, hanem $W(u)$ tagját képviseli a populációnak. A súlyok összege a populáció (pontosabban a keret) nagyságával egyezik meg.

8.1.1. Az alapsúly

Az elemzés során akalmazott súly gyakran egyszerre több célt szolgál. Alkalmazható a súlyozás a mintába kerülési valószínűségek eltéréseinek korigálására, használhatának célja lehet ezenkívül a válaszmegtagadások hatásának ellentételezése, vagy a minta bizonyos jellemzők szerinti reprezentativitásának javítása (Botman et al. 2000.).Elsődlegesen a mintába kerülési esélyek korrigálásához használnak súlyozást, ekkor az adott személy súlya a kiválasztási valószínűségének reciproka. A mintavételi hibától eltekintve minden más hibától mentes hipotetikus szituációban (ahol nem fordul elő válaszmegtagadás stb) az alapsúly már önmagában torzítatlan becslést eredményez. Az előbbi példán bemutatva, a kiválasztási valószínűségek reciprokát használva súlyként, a hálapénz összmennyiségének becslése torzítatlan, azaz várható értéke a tényleges összmennyiség:

$$E(\widehat{X}) = E(\sum_{u=1..n} W_a(u)x(u)) = \sum_{u=1..n} W_a(u)E(x(u)) = = \sum_{u=1..n} W_a(u)\left(\frac{\sum_{v=1..N} x(v)}{N}\right) = \left(\frac{\sum_{v=1..N} x(v)}{N}\right)\sum_{u=1..n} W_a(u) = = \left(\frac{\sum_{v=1..N} x(v)}{N}\right)\sum_{u=1..n} \frac{1}{P(u)} = \left(\frac{\sum_{v=1..N} x(v)}{N}\right) N = \sum_{v=1..N} x(v) = X$$

ahol

$W_a(u)$	az alapsúly értéke az u mintabeli személyre
n	a mintanagyság
N	a populáció nagysága
$x(u)$	az u mintabeli személy által kiadott hálapénz

Az átlag becslése egyszerű véletlen mintavétel esetén, azaz súlyok használata nélkül a következő:

$\widehat{\overline{X}} = \frac{\sum_u x(u)}{n}$

Az átlag becslésének varanciája ekkor:

[15] A fejezet megírásában felhasznált források: Kish, 1965 és Levy, Lemeshow, 1999.

$Var(\overline{X}) = \left(\frac{\sigma_x^2}{n}\right)\left(\frac{N-n}{N}\right)$ [16]

ahol

σ_x^2 az X varianciája a populációban.

Egyszerű véletlen mintavételkor az átlag varianciája lineáris függvénye a populációs varianciának, ezért egyszerűen számolható. Egyenlőtlen kiválasztási valószínűségek esetén alkalmazott súlyozáskor a súlyozott átlag $\left(\widehat{\overline{X}} = \frac{\sum_u W(u)x(u)}{\sum_u W(u)}\right)$ varianciája nem lineáris függvénye a populációs varianciának. Az átlag becslésének varianciája a nem lineáris függvény miatt igen bonyolultan számítható (ez a matematikai nehézség általában minden paraméter becslésének varianciája számításakor fennáll, ha összetett mintavételi eljárást -rétegzést, több lépcsős csoportos mintavételt, utólagos rétegzést- alkalmaztunk). Több, jó közelítést adó számítást közöl a szakirodalom (Carlson, 1998, Levy, Lemeshow, 1999.). Ezek két általános kategóriába sorolhatók. Az egyik típus a nemlineáris függvény Taylor-sorba fejtésének első tagjaként kapott lineáris közelítést alkalmazza. A másik kategóriába az ún. ismételgető technikák (*replication techniques*) tartoznak. Ezek az eljárások a mintából vett almintákból egyenként kiszámolják a paraméter becslését, és ezen becslések egymás közötti szóródását használják a becslés varianciájának számításához. Az ilyen típusú leggyakoribb technikák közé tartozik az úgynevezett *jackknife*, vagy a *bootstrap* eljárás. A 8.3. fejezet becsléseinek varianciáját a Stata programcsomaggal számoltam, ez a program a Taylor-soros lineáris közelítést használja.

Az becslések varianciája általában nagyobb, ha a mintavétel egyenlőtlen kiválasztási valószínűségekkel történt. A súlyok nagyobb változékonysága nagyobb varianciát okoz. A felmérés költségvonzataival magyarázható, miért alkalmaznak a gyakorlatban mégis gyakran ilyen módszert az egyszerű véletlen mintavétel helyett. Az utóbbi nehezebben kivitelezhető, nagyobb logisztikai igényekkel jár, tehát drágább.

8.1.2. Rétegzés, utólagos rétegezés

Rétegzett mintavételi eljárás alkalmazásakor a keretet diszjunkt rétegekre osztjuk, és mindegyik rétegből előre meghatározott nagyságú mintát veszünk. A rétegeken belül egymástól függetlenül választunk mintát. Ha a populációs arányoknak megfelelően képezzük a rétegek nagyságát, előre biztosítható a minta rétegjellemzők szerinti reprezentativitása, így potenciálisan megbízhatóbb becsléseket kaphatunk alkalmazásával. A megbízhatóság növelése a becslés varianciájának csökkenéséből következik. Például az átlag becslése rétegzett minta esetén:

$\overline{X}_r = \frac{\sum_{h=1..L} N_h \overline{x}_h}{N}$

ahol

$\overline{X}_r$	átlag becslése a rétegzett mintával
N	a populáció létszáma
L	a rétegek száma
N_h	a populáció létszáma a h. rétegben
$\overline{x}_h$	mintaátlag a h. rétegben $\left(\frac{\sum_{i=1..n_h} x_{h,i}}{n_h}\right)$
$x_{h,i}$	a h. rétegbeli i. személy válasza
n_h	mintanagyság a h. rétegben

Ha a mintarétegek aránya a rétegek populációs arányával egyezik meg, az átlag becslésének varianciája a következő:

$Var(\overline{X}_r) = \left(\frac{N-n}{N^2}\right)\sum_{h=1..L}\left(\frac{N_h^2}{N_h-1}\right)\left(\frac{\sigma_{hx}^2}{n}\right)$

ahol

σ_{hx}^2	X varianciája a h. rétegen belül
n	a mintanagyság

Vessük össze ezt a varianciát azzal, amit egyszerű véletlen mintavétel alkalmazása esetén kaptunk volna!

A populációs variancia a rétegek közötti és a rétegeken belüli varianciára bontható:

[16] Felmérésből származó, tehát speciális tulajdonsággal bíró becslések varianciája nem minden statisztikai programcsomaggal számítható korrekt módon. A legtöbb program még az egyszerű véletlen mintavétel esetén számolt átlag varianciáját is túlbecsli, az $\frac{N-n}{N} - es$ szorzótényező figyelembe nem vétele miatt (Levy, Lemeshow, 1999).

$\sigma_x^2 = \sigma_{kx}^2 + \sigma_{bx}^2$
ahol

$\sigma_{bx}^2 = \frac{\sum_{h=1..L} N_h \sigma_{hx}^2}{N}$ a rétegeken belüli varianciák súlyozott összege

$\sigma_{kx}^2 = \frac{\sum_{h=1..L} N_h \left(\overline{X}_h - \overline{X}\right)^2}{N}$ a rétegek közötti variancia

itt

$\overline{X}_h$ a h. rétegen belüli átlaga az X-nek

$\overline{X}$ X populációs átlaga

A rétegzett mintával kapott varianciája elég nagy populációs réteglétszámok esetén közelítőleg kifejezhető az X rétegek közötti varianciájával:

$Var(\overline{X}_r) \approx \left(\frac{\sigma_{kx}^2}{n}\right)\left(\frac{N-n}{N}\right)$

Az egyszerű véletlen mintavétellel kapott mintából számolt variancia, mint korábban láttuk:

$Var(\overline{X}) = \left(\frac{\sigma_x^2}{n}\right)\left(\frac{N-n}{N-1}\right)$

Ugyanakkora minták esetén a rétegzett mintából és az egyszerű véletlen mintából kapott variancia hányadosa ezért:

$\frac{Var(\overline{X})}{Var(\overline{X}_r)} = \frac{\sigma_{kx}^2 + \sigma_{bx}^2}{\sigma_{bx}^2} = 1 + \frac{\sigma_{kx}^2}{\sigma_{bx}^2} \geq 1$

Tehát a populációs rétegarányokat használó rétegzett mintával legfeljebb akkora varianciát kapunk, mint egyszerű véletlen mintavételkor. Ha a rétegek nem különböznek egymástól, semmit vagy igen keveset nyerünk a rétegzéssel. Annál jobban növelhető ez a nyereség, minél nagyobb a rétegek közöti variancia a rétegeken belülihez képest, azaz minél távolabb vannak a rétegek egymástól, minél homogénebbek. A megvalósítás szempontjából drágább rétegzett minta alkalmazását akkor érdemes megfontolni, ha ismerünk olyan, a vizsgálni kívánt populációs jellemzővel szoros kapcsolatban álló rétegző változót, aminek a populációs arányait is tudottak.

A gyakorlatban sokszor nehezen kivitelezhető rétegzett minta létrehozása. A leggyakoribb kényelmetlenség, hogy a rétegzési szempont szerinti besorolás nem ismert előre egyénenként, csak a személyes megkeresés után dönthető el, melyik rétegbe tartozik az egyén. Ez lényeges költségnövekedést okoz, hiszen tulajdonképpen egy szélesebb körű előfelmérést tesz szükségessé. Az utólagos rétegzés ezt a hátrányát küszöböli ki a módszernek, sokszor megtarva annak előnyeit. Az utólagos rétegzéskor utólag illesztjük a mintánk arányait a fontosnak tartott jellemzők szerinti populációs arányokhoz, az alulreprezentált rétegeket nagyobb súllyal vonva be az elemzésbe. Belátható, hogy elegendően nagy minta esetén az utólag rétegzett mintából számolt sztenderd hiba nem sokkal nagyobb, mint amekkora a rétegzett mintából kapott sztenderd hiba értéke lenne. A populációs arányoknak megfelelő nagyságú rétegek alkalmazásakor az utólag rétegzett mintából számolt mintaátlag becsült varianciája:

$\widehat{Var}(\overline{X}_{ur}) = \left(\frac{N-n}{nN}\right) \sum_{h=1..L} \frac{N_h}{N} s_{hx}^2 + \left(\frac{1}{n^2}\right) \sum_{h=1..L} \left(\frac{N-n_h}{N}\right) s_{hx}^2$

ahol

$\overline{X}_{ur}$ mintaátlag utólag rétegzett mintában

$s_{hx}^2 = \sum_{i=1..n_h} \left(x_{h,i} - \overline{x}_h\right)^2 / (n_h - 1)$ X becsült varianciája a h. rétegen belül

$x_{h,i}$ a h. rétegbeli i. személy válasza

$\overline{x}_h$ a mintaátlag a h. rétegben

Az összeg első tagja megegyezik a populációs arányoknak megfelelően rétegzett, n elemű mintából becsült varianciával. A második tag növeli ezt a varianciát, de ennek a tagnak az értéke nagy n esetén kellően kicsiny.

Mint korábban láttuk, a becslés varianciája megegyezik az átlagos négyzetes hibának és a torzítás négyzetének a különbségével. Ha torzítatlan becslésekről van szó, a variancia megegyezik az átlagos négyzetes hibával, így a fenti variancia-összehasonlítás az átlagos négyzetes hibák, azaz a becslési pontosságok összehasonlításának feleltethető meg. Tehát torzítalan becslések esetén a populációs arányoknak megfelelően rétegzett mintából legalább olyan pontos becslés nyerhető, mint egyszerű véletlen mintából. Ugyanakkor az utólagos becslés pontossága is közel van ahhoz, mint amit rétegzett mintából kaptunk volna.

Láttuk, hogy az alapsúly használatával torzítatlan becslést kapunk. Az alapsúlyt és az utólagos rétegezést egyszerre alkalmazó becslési eljárás a következő pozitívumokat eredményezi:

- torzítatlan

- általában pontosabb becslést eredményez

Az alapsúly az utólagos rétegzéssel a következő módon kombinálható. Ha ismert egy változó Y populációs totálja, és Y előismereteink szerint összefügg a becsülendő X-szel, akkor az Y mintából becsült értékének és valódi értékének hányadosával érdemes súlyozni az X becslését. Ez továbbra is torzítatlan becslést ad, ha X és Y becslése torzítatlan, és kisebb hibát eredményez. A gyakorlatban ezt a tételt használja ki az, aki a nem-kor összetételre súlyoz, ismerve azok keretbeli megoszlását. A súly képzése az alapsúllyal kombinálva történik. Végezzünk a példánkban nem és korcsoport szerint az utólagos rétegezés. Jelölje b a két változó egyik kategória-kombinációját, $Y(b)$ a hivatalos becslést az adott kategória létszámára (pl. a Központi Statisztikai Hivatal adata), $\widehat{Y}(b)$ a felmérésből az alapsúllyal becsült gyakoriságot a b-re:

$\widehat{Y}(b)=\sum_u W_a(u)I_b(u)$

ahol $I_b(u) =$ 1, ha az u a b kategóriában van

0 máskor

Most az utólagos rétegzés súlya:

$W_{ur}(b) = Y(b)/\widehat{Y}(b)$

A kapott súly az alapsúly és az utólagos rétegzés súlyának szorzata, a becslés az X populációs együttjára:

$\widehat{X} = \sum_u W_a(u)W_{ur}(u)x(u)$

Amikor a megvalósult minta arányait megváltoztatjuk az utólagos rétegezéssel, egy lépésben próbáljuk ellensúlyozni azt a sok tényezőt, ami a hibát okozta. Ezek között a tényezők között lehet a keret megválasztásának hibája, a minta véletlen jellegéből fakadó mintavételi hiba, vagy a szisztematikus válaszmegtagadás okozta arányeltolódás. Ezek hatása egyébként sem nem választható szét, ha csupán a minta összetételét ismerjük.

8.1.3. A mintavételi módszer hatása a becslés varinciájára

Egyenlőtlen kiválasztási valószínűségek esetén általában megnövekszik a becslések varianciája. Ugyanígy a variancia növekedését eredményezi a csoportos mintavételi módszer. Ezzel szemben ezek a mintavételi módszerek általában olcsóbban kivitelezhetők. Tehát a statisztikai hatékonyság potenciális növekedése áll szemben az anyagi költségmegtakarítással. A 8.2. fejezetben leírt rétegzéskor viszont csökken a becslések varianciája, ám ez a módszer bizonyos költségnövekedést jelent a felmérés kivitelezésében. A Kish által kifejlesztett tervezés-hatás együttható (*design effect*, Kish, 1965) a mintavételi módszernek betudható varianciaváltozás mértéke. Számításakor a becslés adott mintavételnek megfelelő varianciáját osztjuk ugyanezen paraméter azonos mintanagyság, de egyszerű véletlen mintavétel esetén kapott becslésének varianciájával. Ha az együttható 1, az alkalmazott mintavételi módszernek nem volt hatása a varianciára. 1-nél nagyobb tervezés-hatás együttható azt jelzi, hogy a mintavételi eljárás a becslési variancia növekedését eredményezte. A tervezés-hatás együttható használható a hatékony mintanagyság meghatározására is. Például egy összetett mintavételi módszer 1,5-ös tervezés-hatás együttható és 1500-as mintanagyság mellett becslési pontosság szempontjából egy 1000 fős mintán alkalmazott egyszerű véletlen mintavétellel egyezik meg (Carlson ,1998). 500 megfigyelést vesztünk az összetett eljárás esetleges költségmegtakarításának ellentételezéseként. A 8.3. fejezetben a tervezés-hatás együtthatót is megadom a mintavételi szimulációból kapott becslések számításakor. Ugyanott részletesebben bemutatom az együttható számítását.

8.1.4. A súlyozás hatása a becslés varianciájára

A súlyozás alkalmazásával gyakran megnő a becslések megbízhatósági intervalluma, a súlyok nagyobb szórása nagyobb sztenderd hibát okoz (Kish 1965) Mint láttuk, az utólagos rétegzés is növelheti a becslések megbízhatósági intervallumát. A kutató döntésétől függ, vajon elfogadja-e a potenciálisan nagy torzítás lehetőségét, vagy súlyozást alkalmaz, ezzel esetleg jelentősen növelve a megbízhatósági intervallumokat, azaz csökkentve a becslések megbízhatóságát.

8.2. Súlyozás a Kish-kulccsal vett minta elemzésekor

8.2.1. Az alapsúly

Ha a háztartás minden felnőtt tagjának egyforma esélye van a mintába kerülésre (ilyen - a már tárgyalt kis hibáktól eltekintve- a Kish-kulcs, ilyen az legutolsó vagy a legközelebbi születésnap módszere), a torzítatlan becslés érdekében a kérdezettek a háztartásuk felnőttjeinek számával súlyozandók. A Kish-kulcs első publikálásakor Leslie Kish már érintette a kulcs használatával választott minta súlyozásának problematikáját (Kish, 1949). A korabeli Egyesült Államokbeli háztartásszerkezetet vette alapul, melynek jellemző vonása volt a 2 felnőtt tagú háztartások magas aránya. A nem túl variábilis háztartásszerkezet miatt a háztartásnagyságra épülő súlyok is viszonylag homogének. Kish számításai szerint ezen homogenitás pozitív következménye a súlyozás után egyébként megnövő variancia nem számottevő, 10% körüli növekedése. Másik pozitív következményként említi a súlyozott és súlyozatlan becslések nem túl nagy különbségét. (Olyan változók becslése esetén persze, amelyek nem korrelálnak pozitívan a háztartásnagysággal). Ez a nem túl nagy különbség pontosabban a becslések varianciájához viszonyított kis különbségét jelenti. Kish beszámolójában említi, hogy több kutatás sokéves tapasztalata szerint a súlyozott és súlyozatlan becslések különbsége annyira elhanyagolható, hogy munkamódszerük szerint, miután néhány tucat változón tesztelik a súlyozás jelentéktelen hatását, a teljes elemzésből elhagyják a súlyokat (Kish, 1965).

Kish a fenti pozitívumokat a korabeli Egyesült Államokbeli, tehát háztartásszerkezet következményeként vezette le, így azok nem feltétlenül általánosíthatóak. Az alábbi elemzés mai magyar adatokat véve ugyanezen szempontok szerint értékeli a Kish-kulcs használata után alkalmazott súlyozás okozta hibanövekedést és becslésváltozást. Ismét a Magyar Háztartás Panel adatbázisát használtam. Az elemzés célja a súlyok eloszlásának vizsgálata, illetve annak eldöntése, változik-e számottevően a jellemzők becslése. Több súly vizsgálatát is végrehajtottam: az alapsúlyt, az alapsúlyozást követő utólagos rétegzést, és gyakorlati módszertanban gyakran alkalmazott "puszta" utólagos rétegzést is bevontam az elemzésbe.

Mint korábban is, a Háztartás Panel mintáját pszeudopopulációnak használva, a mintavételi szimuláció során a háztartások közül 10%-os véletlen mintát vettem, majd a kulcs alkalmazásával választottam ki a kérdezendőt a háztartás tagjai közül. Az Kish-módszer szerint a kulcs 8 tábláját sorrendben 16, 8, 8, 16, 16, 8, 8, 16%-os arányban kell kiosztani, véletlenszerűen. Ideális esetben, ha a háztartásminta elég nagy ahhoz, hogy a megfelelő arányban osszuk ki a 8 tábla mindegyikét, a Kish-kulcs a 4. táblázatban leírt háztartáson belüli kiválasztási valószínűségeket produkálja. A szimulációban csak 199 háztartás került be a mintába, így ezt a kiosztási arányt nem sikerült pontosan teljesítenem (lásd a Függelék 6. tábláját). Ezért a szimulációban az eredeti, egyenletes kiválasztási valószínűségeket is csak közelítőleg sikerült elérni (lásd a 22. táblát).

22. táblázat: Háztartáson belüli kiválasztási valószínűségek a szimulációs vizsgálatban

	Háztartástagok száma					
Tag sorszáma	1	2	3	4	5	6 vagy több
1	1	0,516	0,342	0,256	0,171	0,171
2		0,484	0,337	0,256	0,171	0,171
3			0,321	0,246	0,251	0,171
4				0,241	0,166	0,166
5					0,241	0,161
6						0,161
7 vagy több	0	0	0	0	0	0

Az alapsúlyt a mintába kerülési valószínűség reciprokaként kaphatjuk meg. A háztartások kiválasztásának esélye azonos volt minden háztartásra, a minta háztartásainak és a (pszeudo)populáció háztartásainak számának hányadosaként kaphatjuk meg. Most az egyén mintába kerülési valószínűsége ezen érték és a táblázatbeli érték szorzataként áll elő. Az alapsúlyok összege közelítőleg a pszeudopopuláció létszáma. A könnyebb jellemzés érdekében a normált változat eloszlását vizsgáltam (lásd a Függelék 7. tábláját.) A normálást úgy hajtottam végre, hogy a súly értékeit az összegükkel osztottam, hogy a súlyok átlaga 1 legyen, így könnyebben interpretálhatók az értékek. A normált alapsúly értéktartománya 0,484 és 2,006 között van, szórása 0,409. Az érdeklődésünk középpontjában levő korcsoport- és nem kategóriákban a normált alapsúly átlagát a 23. táblázat mutatja. Látható, hogy a fiatal férfiak rétege kapja a

legnagyobb súlyt, az idős nőké pedig a legkisebbet. Korkategóriánként: a fiatalok kapják a legnagyobb súlyt, az idősek a legkisebbet. A nők kisebb súlyt kapnak, mint a férfiak. Ezek a számok a fentebb már leírt tendenciát tükrözik, a Kish-kulcs és általában a háztartáson belüli véletlen mintavétel önmagában az idősek, a nők felülreprezentáltságát okozza.

23. táblázat: A normált alapsűly átlaga a nem/korcsoport kategóriákra

	nem		
korcsoport	férfi	nő	O2sszesen
18-39	1,170	1,116	1,142
40-59	1,005	1,062	1,037
60+	0,880	0,813	0,832
O2sszesen	1,042	0,972	1

8.2.2. Utólagos rétegezés

Utólagos rétegzéskor a minta korcsoport és nem szerinti arányait állítottam be a (pszeudo)populáció összetételének megfelelően. Mint a 8.2. fejezetben láttuk, a mintából az alapsúllyal becsült nem/korcsoport kategóriák (rétegek) létszámát arányítjuk a populációban meglevőhöz, tehát nem a mintabeli arányokat használjuk közvetlenül. Az alapsúly és ezen arány szorzata adja az új súlyt. Az alapsúly alkalmazása nélkül, csupán a minta arányaira épülő utólagos rétegzést a következő fejezetben vizsgálom.

A szimulált mintavételi szituációnkban az alapsúllyal korrigált értékek már torzítatlan becslést biztosítanak. Az utólagos rétegzéstől a becslés nagyobb pontosságát várjuk, segítségével a minta bizonyos jellemzők szerinti véletlen ingadozásait kívánjuk ellensúlyozni. Az utólagos rétegezés végleges súlyának kialakításához használt szorzóegyüttható - a réteglétszámok populációs és mintabeli létszámának hányadosa - mintánként változik. Az alapsúly már a mintavétel előtt adott, mert a kiválasztási valószínűség függvénye, tehát minden mintára állandó. A rétegzés szorzóegyütthatója a minta véletlen, nem szisztematikus ingadozását korrigálja, értékéből a minta (adott jellemzők szerinti) átlagos vagy szélsőséges voltára következtethetünk.

A populációs és a mintából becsült réteglétszámok arányát a 24. táblázat mutatja. A végleges súly kialakításakor ezekkel a számokkal szorzódik az alapsúly. A táblázat szerint miután korrigáltuk a mintát az alapsúllyal, az idős nők aránya még mindig magasabb populációs arányoknál, de az idős férfiaké már alacsonyabb. A középkorúak alapsúlya nem különbözött lényegesen 1-től, de 1-nél nagyobb volt, most a rétegzéssel alapsúlyuknál nagyobb súlyt kapnak a középkorúak. A legnagyobb rétegzési súllyal az idős és középkorú férfiak szerepelnek, a fiataloké csak utánuk következik. Mivel a középkorú férfiak normált alapsúlyának átlaga alig különbözik 1-től, és most a rétegzésnél magas, arra következtethetünk, hogy a mintánkban a várhatónál alacsonyabb a középkorú férfiak aránya.A fiatalok súlya nagyobb mint 1, de jóval alapsúlyuk alatt marad. Azaz arányuknak az alapsúllyal történt nagyobb korrekciója után még egy kisebb aránynövelés szükséges.

24. táblázat: A populációs és a becsült réteglétszámok aránya

	nem	
korcsoport	férfi	nő
18-39	1,010	1,036
40-59	1,185	1,051
60+	1,202	0,808

Ezen értékek és az alapsúly szorzataként adott végleges súly normált alakjának rétegenkénti átlaga megegyezik a következő fejezet 23. táblájában közölt értékekkel. A Függelék 8. táblázata jellemzi az utólagos rétegzés végleges súlyának normált változatának eloszlását. A súly tartománya kissé megnyílt az alapsúlyéhoz képest: most 0,383 és 2,192 között mozognak az értékek, a szórás 0,409-ről 0,437-re nőtt.

8.2.3. Utólagos rétegezés, alapsúly nélkül

Elterjedt gyakorlat az alapsúly - azaz a kiválasztási valószínűségek - figyelembe vétele nélkül hozzáigazítani a minta bizonyos jellemzőit a populációs arányokhoz. Az alapsúly használatakor a mintába kerülés egyenlőtlen esélyeit kívánjuk korrigálni, torzítatlan becslések elérése érdekében. Ha nem vesszük figyelembe, hogy a mintába különböző valószínűséggel kerültek be a személyek, a potenciális torzítás esélye

annak ellenére fennáll, hogy kiegyenlítettük például a minta nem- és korösszetételét. Az alábbiakban így pusztán nem- és korcsoport szerint illesztem a minta arányait a (pszeudo)populációéhoz. Az alkalmazott súly minden egyénre a hozzá tartozó réteg populációs arányának és a mintabeli aránynak a hányadosa, így a súly rétegenként azonos. A 23. táblázat szerint a fiatalok és a középkorú férfiak a populációs arányuknál kisebb számban képviseltetik magukat, az idős nők súlya viszont 2/3 körül van, ők a populációs létszámuknál nagyobb arányban szerepeltek a mintában.

25. táblázat: Rétegenkénti sűlyok, alapsűly használata nélkül

	nem		
korcsoport	férfi	nő	O2sszes
18-39	1,159	1,134	1,146
40-59	1,167	1,094	1,126
60+	1,036	0,644	0,758
O2sszes	1,129	0,913	1

A súly eloszlásáról közöl részleteket a Függelék 9. táblázata. A súly értéktartománya szűkebb, mint fent - 0,644 és 1,167 közötti. A szórása is alacsony, 0,213. Mivel ez a súly rétegenként állandó, és az alapsúllyal kombinált utólagos rétegezés rétegenkénti átlaga éppen ez a súly, a két súly varianciájának viszonya előre meghatározható: az utóbbi varianciája mindig legalább akkora, mint az előzőé. Az alapsúllyal kombinált utólagos rétegezés súlyának varianciájának rétegek közötti variancia-része éppen az alapsúly nélküli rétegzés-súly varianciája. Ez a súly tehát kevésbé változékony, mint a két fenti, de összevetésükkor nem feledkezhetünk meg róla, hogy ez a súly nem biztosít torzítatlan becslést. A különböző súlyokkal számolt becslések alakulása a következő fejezet témája.

8.3. Becslések a Kish-kulccsal vett mintából

Az előző fejezetben a Magyar Háztartási Panel-ből, mint pszeudopopulációból vett minta súlyozását vizsgáltam. Három súlyozási módszert hasonlítottam össze: a kiválasztási valószínűségek reciprokaként előálló alapsúlyt, az alapsúly és az utólagos rétegzés kombinációjaként előálló súlyt, és az egylépéses utólagos rétegzést. A súlyok összevetésekor azok eloszlásának jellemzéséből indultam ki. Releváns összehasonlítási dimenzió ezen kívül a súlyokkal kapott becslések összevetése, ezt végzem most. Mivel a Panel minden háztartásának minden lakója töltött ki kérdőíveket, az az előny is fennáll, hogy a valódi (pszeudo)populációs paraméterekkel is összevethetők a becslések.

A Stata programcsomagot használtam a becslések számítására. A Stata korrekt módon számolja a becslés varianciáját az alapsúlyt alkalmazva (egyenlőtlen kiválasztási valószínűségek esetén), ezenkívül képes figyelembe venni a rétegzés és a csoportos mintavétel hatását is. Ugyanezen esetekre a tervezés--hatás együtthatót is kiszámítja (lásd a következőkben). Utólagos rétegzésre azonban a pontbecsléseken túl nincs felkészítve[17]. Ezért a becslések megbízhatósági intervallumát csak a súlyozás nélküli és az alapsúlyos esetben adtam meg. Hangsúlyozandó, hogy a két megadott megbízhatósági intervallum közül csak az alapsúllyal számolt értékelhető korrektnek. Ugyanis a súlyozás nélküli becslés esetén a becslés varianciájának számításakor nem vesszük figyelembe az egyenlőtlen kiválasztási valószínűségeket. Úgy számoljuk a megbízhatósági intervallumot, mintha egyszerű véletlen mintát vettünk volna, azonos kiválasztási esélyekkel. Az alapsúllyal számolt becslésekhez a tervezés-hatás együtthatót is megadtam. A tervezés-hatás együttható (THE) a Θ paraméter becslésének varianciájára ebben az esetben a következő módon számítható:

$$\text{THE}=\frac{V\left(\widehat{\Theta}\right)}{V_{evm}\left(\widehat{\Theta}_{evm}\right)}$$

ahol $V\left(\widehat{\Theta}\right)$ az alapsúllyal számolt becslés varianciája (Taylor-sorfejtéssel lineárisan közelítve), $V_{evm}\left(\widehat{\Theta}_{evm}\right)$ pedig egy egyszerű véletlen mintavételt alkalmazó hipotetikus felmérés esetén a $\Theta-$ra adott becslés varianciája. Számítása akkor, amikor Θ az Y populációs összege (mint a 8. fejezet példájában az országos szinten kiadott hálapénz összmennyisége):

[17] A leggyakrabban használt statisztikai programcsomagok közül a Stata, a SUDAAN, a WesVar, a PC Carp van felkészítve korrekt felmérésstatisztikai számításokra. A SAS és az SPSS fejlesztési tervei közé tartozik az ilyen típusú modulok beépítése (Carlson, 1998).

$$V_{evm}\left(\widehat{\Theta}_{evm}\right) = \frac{\sum_u w(u)}{N-1} \sum_u w(u) \left(y(u) - \frac{\sum_u w(u)y(u)}{\sum_u w(u)} \right)^2$$

ahol

$w(u)$ az u személy súlya (ez a súly azonos a számlálóban használt alapsúllyal)

$y(u)$ az u személy által kiadott hálapénz

A becslések összevetésekor szem előtt kell tartanunk, hogy a példaként kiemelt változók jelentősen függhetnek egymástól, ezért függ egymástól a becslésük pontossága is. Példaként: a súlyozás nélküli becslés becsli leginkább felül a legfeljebb 8 általános iskolai osztályt végzettek arányát, és ugyanez a becslési módszer adja a legmagasabb becslést a számítógéppel nem rendelkezők arányára is.

Ha a populációs paramétertől való távolságuk szerint vetjük össze a becsléseket, nem találjuk egyik módszert sem jobbnak a többinél, sőt inkább úgy tűnik, egyformán teljesítenek. Meglepően magas azon jellemzőknek a száma, amiket mindegyik becslés azonos irányban becsül rosszul, például a háztartások átlagos havi kiadását és az SZDSZ-szavazók arányát mindegyik eljárás alulbecsli. Általánosabban: megfigyelhető, hogy a becslések kevésbé térnek el egymástól, mint a valódi paramétertől.

Azonban az abszolút eltérések összevetése önmagában nem elegendő, érdemes az eltérések populációs értékhez viszonyított arányát is figyelembe venni. A többkategóriás változók kategóriáinak válaszarányának becslései mutatják a legnagyobb instabilitást, például a legfeljebb 8 általános iskolai osztályt végzettek arányának súlyozás nélküli becslése 40%-kal magasabb, mint a valódi arány. Az ilyen típusú, kategóriák arányaira vonatkozó becslésektől azonban elméletileg sem várhatunk olyan pontosságot, mint a jellemzők átlagaira vonatkozó becslésektől. A változók átlagára, pl. a nettó jövedelem átlagára vonatkozó becslések eltérése a valódi átlagtól 8-10% körüli. A megbízhatósági intervallumokkal kapcsolatban említésre méltó, hogy a súlyozás nélküli becslés megbízhatósági intervalluma több esetben nem fedi le a valódi paramétert, így a háztartások átlagos kiadásának vagy a 8 iskolai osztályt végzettek arányának becslésekor. Az alapsúllyal számolt becslés tervezés-hatás együtthatójának legmagasabb értéke 1,31 (az egyetemet végzettek arányának becslésekor). Ezt interpretálhatjuk úgy, hogy ennél a becslésnél a Kish-kulcs használata - nem egyenlő kiválasztási valószínűségekkel - a becslés varianciájának 30%-os emelkedését okozta.

26. táblázat: Akülönböző sűlyok alkalmazásával kapott becslések összevetése

Valódi érték	Sűlyozás nélkül		Alapsűly			Alapsűly+ utólagos r.	Utólagos rétegzés
	Becslés	95%-os m. i.	Becslés	95%-os m. i.	THE	Becslés	Becslés
			Nettó jövedelem átlaga				
19316	18323	[16620;20018]	17688	[15970;19407]	1,07	17882	18909
			Iskolai végzettség - százalékos arány				
			legfeljebb 8 ált.				
14,8	20,3	[15,1;26,8]	18,3	[13,1;25,0]	1,11	17,3	16,6
			szakmunkásképző				
24,2	21,9	[16,5;28,5]	20,8	[15,2;27,8]	1,14	20,7	21,7
			középiskola				
25,4	21,4	[16,1;27,9]	22,8	[16,8;30,2]	1,22	23,4	23,8
			főiskola., technikum				
27,8	27,3	[21,3;34,2]	27,6	[21,2;35,0]	1,15	27,8	28,4
			egyetem				
7,8	9,1	[5,7;14,2]	10,5	[6,4;16,8]	1,31	10,8	9,5
	Hány forint lenne szükséges havonta a családjának a megfelelő színvonalú megélhetéshez? -átlag						
84906	71668	[64777;78560]	81337	[73398;89277]	82603	75722	

Valódi érték	Súlyozás nélkül		Alapsúly			Alapsúly+ utólagos r.	Utólagos rétegzés
	Becslés	95%-os m. i.	Becslés	95%-os m. i.	THE	Becslés	Becslés
Háztartás átlagos havi kiadása - átlag							
34997	31142	[28737;33547]	33778	[31211;36344]	1,14	34036	32373
Autóval nem rendelkeznek - százalékos arány							
54,8	58,8	[51,8;65,5]	52,8	[45,2;60,3]	1,18	51,9	55,6
PC-vel nem rendelkeznek - százalékos arány							
88,1	89,5	[84,3;93,0]	88,2	[82,2;92,4]	1,23	87,6	88,1
Nem műveltek földet az elmúlt évben - százalékos arány							
52,4	58,3	[51,3;65,0]	54,2	[46,5;61,6]	1,18	53,0	57,9
Mennyire elégedett életpályájával (0:egyáltalán nem,...,10:teljesen) - pontátlag							
5,57	5,58	[5,23;5,93]	5,68	[5,26;6,09]	1,28	5,67	5,56
Biztosan/valószínűleg elmenne szavazni, ha most vasárnap lennének a parlamenti választások - százalékarány							
66,4	65,7	[58,2;72,5]	66,6	[58,6;73,7]	1,14	66,5	64,2
Valódi érték	Súlyozás nélkül		Alapsúly			Alapsúly+ utólagos r.	Utólagos rétegzés
	Becslés	95%-os m. i.	Becslés	95%-os m. i.	THE	Becslés	Becslés
Ha (mégis) elmenne választani, melyik pártra szavazna - százalékarány							
MDF							
10,3	11,9	[6,4;21,0]	11,0	[5,7;20,2]	1,06	11,6	11,9
SZDSZ							
16,8	14,3	[8,2;23,7]	13,7	[7,6;23,5]	1,09	13,0	13,7
FKGP							
18,1	19,1	[11,9;29,1]	20,4	[12,4;31,7]	1,20	20,3	18,6
MSZP							
25,7	23,8	[15,8;34,3]	25,9	[16,9;37,5]	1,17	26,1	24,4
FIDESZ							
10,4	10,7	[5,6;20,0]	8,8	[4,4;19,4]	1,15	8,7	11,4

A becslések összehasonlításából nem vonhattunk le olyan típusú következtetést,ami valamelyik becslési mód előnyét bizonyítaná a többivel szemben. A fenti példa inkább modell típusú elemzés, nem volt célom gyakorlati bizonyítékát adni a korábbi elméleti megfontolásoknak. Példaként: bár a puszta utólagos rétegzést használó becslés nem mutatkozott pontatlanabbnak, mint a többi, mégis igaz rá, hogy potenciálisan torzított becslése a paraméternek, míg az alapsúllyal számoló becslés torzítatlan.

9. A Kish-kulcs megváltoztatása

Mint az 7.3. fejezetbeli szimulációs vizsgálatokból kiderült, a Kish-kulcs használatával vett minta nem/-korcsoport megoszlása lényegesen eltér a (pszeudo)populációs arányoktól. Az előző fejezetben ezen súlyozással próbáltunk meg javítani. Az alapsúly alkalmazásával, azaz a kiválasztási valószínűségek kiegyenlítésével a minta összetétele már várhatóan reprodukálja a populációs összetételt. Sőt,ezt az illeszkedést nem csak a nem/korcsoport jellemzők, hanem minden más jellemző esetén is elvárhatjuk.

Mégis, felmerülhet az igény, hogy a minta valamilyen jellemzők szerinti összetétele nyersen, azaz súlyok nélkül is közel legyen a populációéhoz. Az alábbiakban a kulcs módosítására teszünk kísérletet, olyan céllal, hogy a minta a nem/korcsoport jellemzők szerint minél közelebb legyen a populációs paraméterekhez. Ez a fajta reprezentativitás-javító módszer nem ismeretlen a szakirodalomban, mint korábban említettem, Kish maga is variálta a kulcs beállításait. Ezenkívül a 6. fejezetben, a Hagen-Collier módszer (Lavrakas 1993) kapcsán idéztem Lavrakast, aki azt ajánlja, hogy a gyakorlatban a korábban létrejött minták nem és kor szerinti eltolódásainak megfelelően állítsuk be a Hagen-Collier 4 táblájának megfelelő arányát. Ugyanis a tapasztalatok szerint amikor egyenlő arányban elosztva használják a négy verziót, a

középkorúak illetve a férfiak alulreprezentáltak lesznek a mintában.

A Kish-kulcs módosításakor a mintavételi eljárás alapvető vonásait állandónak tételezzük fel. Így nem változtatunk azon, hogy a háztartások azonos valószínűséggel kerülnek a mintába, s hogy minden háztartásból pontosan egy személyt választunk ki. A kulcsból azt a jellemzőt is megtartjuk, hogy a háztartás tagjait nemük és koruk szerint sorba tesszük, és egy, a kérdőívhez csatolt táblából keresheti ki a kérdezőbizos, hogy a sor hányadik tagját kérdezze. A módosítás csak erre a táblára terjed ki.

Felhasználva a populáció háztartásszekezetének ismeretét, egy ilyen táblából egyértelműen meghatározhatjuk a minta várható nem- és korcsoportszerkezetét. A feladatat tehát olyan táblák elkészítése, amik használata esetén a minta nem- és korcsoport összetétele megegyezik, vagy a "lehető legközelebb van" a populációs megoszláshoz. A feladat még egy paramétert tartalmaz, ez a kérdőívhez csatolt táblák típusainak száma. Minél több táblafajtát használhatunk, annál finomabb valószínűségeloszlás jön létre, s vélhetően annál közelebb hozható a minta a populációhoz. A Kish kulcs 5. fejezetbeli eredetije 8 táblát használ fel, de ezeket különböző arányban kiosztva a kérdőívek között, pl. a kérdőívek 1/12-éhez a B1-es tábla van csatolva. Ezért az alábbi módosításban 12-ben rögzítem a táblatípusok számát.

Az alábbi táblázat a módosított tábla meghatározásra váró ismeretlen számadatait tartalmazza:

27. táblázat: A módosítással meghatározandó vij ismeretlenek

A tábla aránya	A tábla jelzete	Ha a felnőttek száma: 1	2	3	4	5	6 vagy több
		A kiválasztott sorszáma:					
1/12	1.	1	v21	v31	v41	v51	v61
1/12	2.	1	v22	v32	v42	v52	v62
1/12	3.	1	v23	v33	v43	v53	v63
1/12	4.	1	v24	v34	v44	v54	v64
1/12	5.	1	v25	v35	v45	v55	v65
1/12	6.	1	v26	v36	v46	v56	v66
1/12	7.	1	v27	v37	v47	v57	v67
1/12	8.	1	v28	v38	v48	v58	v68
1/12	9.	1	v29	v39	v49	v59	v69
1/12	10.	1	v210	v310	v410	v510	v610
1/12	11.	1	v211	v311	v411	v511	v611
1/12	12.	1	v212	v312	v412	v512	v612

Az ismeretlen vij-k segítségével egyértelműen meghatározható, hogy m létszámú háztartásnak a nem/-kor szerinti sorban l. tagja mekkora valószínűséggel kerül kiválasztásra. Ha $l > 6$, ez a valószínűség 0. Más esetben ez a valószínűség 12-ed része annak, ahányszor a 27. táblázat m. oszlopában az l szerepel. Tehát a kulcs nmódosításának feladata egyszerűsíthető: nem a táblákat kell meghatározni, csak azokat a valószínűséget (12-edekben), amekkora eséllyel a táblák használatával az m. tagú háztartások l. tagja kiválasztásra kerül (minden m-re és minden l-re). Jelöljük a meghatározandó valószínűségeket pij-vel (i tagú háztartás sorban j. tagjához tartozó kiválasztási valószínűség). Az ismeretlenek tehát: $p21$, $p22$, $p31$, $p32$, $p33$, $p41$, $p42$, $p43$, $p44$, $p51$, $p52$, $p53$, $p54$, $p55$, $p61$, $p62$, $p63$, $p64$, $p65$, $p66$, összesen 20 ismeretlen.

A feladat megoldásához szükség van arra, hogy tudjuk, miután pij valószínűséggel kiválasztottuk egy i tagú háztartás sorban j. tagját, ő milyen valószínűséggel férfi/nő, milyen valószínűséggel fiatal/középkorú/idős. Ez 6 érték, egy 3×2-es keresztáblával adjuk meg. Az ilyen keresztáblák megadása a populáció ismeretét előfeltételezi. A feladat input paramétere tehát összesen 21 3×2-es tábla, jelölve: $a11$, $a21$, .. , $a66$.

Az i tagú háztartásban lakó i. kérdezett kiválasztási valószínűsége a háztartásának kiválasztási valószínűségének és a háztartáson belül az egyén kiválasztási valószínűségének szorzata. Az utóbbi érték a pij-vel jelölt ismeretlen paraméter, az előbbi érték viszont a populáció háztartásszerkezetének függvénye. Mivel a háztartások azonos valószínűséggel kerülnek a mintába, az i létszámú háztartás kiválasztási valószínűsége az i tagú háztartások populációs arányával egyezik meg. Jelölje ezt az értéket bi, a feladat input paramétere tehát a fentieken kívül még: $b1$, $b2$, $b3$, $b4$, $b5$, $b6$.

Most már megadható a mintában várható adott nemű, adott korcsoporthoz tartozó személyek aránya, ez hat érték: $c1..c6$. $c1$ az első korcsoporthoz tartozó férfiak arányát jelöli, $c6$ a harmadik korcsoporthoz tartozó nők arányát. ci-k megadhatók bi-k és pij-k függvényében:

$c1 = b1 \times a11[1,1] + b2 \times (p21 \times a21[1,1] + p22 \times a22[1,1]) +$
$+b3 \times (p31 \times a31[1,1] + p32 \times a32[1,1] + p33 \times a33[1,1]) +$
$+b4 \times (p41 \times a41[1,1] + p42 \times a42[1,1] + p43 \times a43[1,1] + p44 \times a44[1,1]) +$
$+b5 \times (p51 \times a51[1,1] + p52 \times a52[1,1] + p53 \times a53[1,1] + p54 \times a54[1,1] +$
$+p55 \times a55[1,1]) + b6 \times (p61 \times a61[1,1] + p62 \times a62[1,1] + p63 \times a63[1,1] +$
$+p64 \times a64[1,1] + p65 \times a65[1,1] + p66 \times a66[1,1])$

A $c2..c6$ hasonlóan áll elő az aij mátrixok megfelelő elemének behelyettesítésével, lásd a Függelék 10. táblázatát.

A célként jelölt populációs nem/korcsoport megoszlás egy 3×2-es táblával megadható. A tábla értékeit jelölje $t1..t6$, ahol $t1$ a férfiak 1-es (fiatal) korcsoportjának populációs aránya, $t6$ a nők 3. korcsoportjának populációbeli aránya. Keressük pij-k függvényében ci-k azon értékeit, amelyeknek értéke a megfelelő ti értékével egyezik meg, ekkor a minta nem/korcsoport megoszlása pontosan reprodukálja a populációét.

A feladat megfogalmazása tehát: adottak $t1..t6$ skalárok, adottak $a21..a66$ mátrixok. Keressük azon $p21...p66$ skalárokat, amik kielégítik az alábbi egyenletet:

$|t1 - c1| + |t2 - c2| + |t3 - c3| + |t4 - c4| + |t5 - c5| + |t6 - c6| = 0$

az alábbi feltételekkel:

$p21 + p22 = 1$
$p31 + p32 + p33 = 1$
$p41 + p42 + p43 + p44 = 1$
$p51 + p52 + p53 + p54 + p55 = 1$
$p61 + p62 + p63 + p64 + p65 + p66 = 1$
$pij \geq 0 \; \forall \; i, j - re$
$pij = kij/12 \; \forall \; i, j - re$, ahol kij egész szám.

Az Microsoft Excell Solver bővítményprogramját használtam a feladat megoldásának keresésére.[18] A feladat tulajdonképpen egy sokváltozós nemlineáris egyenlet egészértékű megoldását keresi, korlátozó feltételek mellett.

Az input paraméterek megadásakor a korábban is használt Magyar Háztartás Panel mintáját használom populációként. A populációs nem/korcsoport arányok, a ti-k értéke:

28. táblázat: A kulcs-módosítás input paraméterei

	férfi	nő
1. korcsoport	t1=0,1979	t4=0,2051
2. korcsoport	t2=0,1466	t5=0,1760
3. korcsoport	t3=0,1094	t6=0,1650

Az egyenletnek nincs olyan megoldása, amely a feltételeket kielégítené.

Tehát 12 táblát használva nem reprodukálható pontosan a populáció nem/korcsoport megoszlása. Keressük ehelyett azokat a táblákat, amelyek használata esetén a populációhóz "legközelebb levő" mintát kapunk. A legközelebb levő fogalmát operacionalizáljuk a khi-négyzet statisztikához hasonlóan: összegezzük a 6 cellára a mintából várt és a populációs arányok négyzetes eltérésének és a várt aránynak a hányadosát. Ez az érték N elemű, pontosan a várt arányokat ($N * ci$) adó mintát feltételezve a következő összefüggésben áll a khi-négyzet statisztikával:

$f(c1..c6) := \sum_{i=1..6} \frac{(c_i - t_i)^2}{t_i} = \sum_{i=1..6} \frac{1}{N} \frac{(N*c_i - N*t_i)^2}{N*t_i} = \frac{1}{N}\kappa^2$

Ebből következik, hogy az f függvény értékét joggal feleltetem meg a produkált arányok és a populációs arányok eltérésének mértékének, hiszen hagyományosan így mérjük két eloszlás különbözőségét. Ám a függvény értékéből nem helyes a khi-négyzet statisztika elért valószínűségi szintjére következtetni, hiszen a statisztika egy valódi minta és a mögötte álló eloszlás paramétereit veti össze, míg én tulajdonképpen két eloszlást vetek össze a **várható minta** és a populáció összehasonlításakor.

Most a feladat megfogalmazása:

$f(c1..c6) \rightarrow \min .$
$p21 + p22 = 1$

[18] A Microsoft Excel Solver az Általános Redukált Gradiens (GRG2) nemlineáris optimalizálási eljárást használja, amelyet Leon Lasdon (University of Texas at Austin) és Allan Waren (Cleveland State University) fejlesztett ki. A lineáris és az egész értékű problémákra a változókat korlátozó szimplex módszert, valamint a "branch-and-bound" (elágazás és korlátozás) módszert használja; ez utóbbit John Watson és Dan Fylstra (Frontline Systems, Inc.) valósította meg.

$p31 + p32 + p33 = 1$
$p41 + p42 + p43 + p44 = 1$
$p51 + p52 + p53 + p54 + p55 = 1$
$p61 + p62 + p63 + p64 + p65 + p66 = 1$
$pij \geq 0 \ \forall \ i, j - re$
$pij = kij/12 \ \forall \ i, j - re$, ahol kij egész szám.

Ez egy nemlineáris egészértékű optimalizálási probléma, megszorító feltételekkel.

Kiszámítható, hogy a Kish-kulcs alkalmazása esetén milyen ci-ket kapunk, azaz milyen a mintabeli várható aránya az adott nemű, adott korcsoporthoz tartozó személyeknek A ci-k számításához ilyenkor a 4. táblázatban adott háztartáson belüli kiválasztási valószínűségeket használjuk fel. Az alábbi tábla mutatja a várható nem- és korcsoport arányokat:

29. táblázat: Várható mintaösszetétel Kish-kulcs alkalmazásakor

	férfi	nő
1. korcsoport	c1=0,1727	c4=0,1950
2. korcsoport	c2=0,1292	c5=0,1700
3. korcsoport	c3=0,1187	c6=0,2140

Megfigyelhetjük az idős korosztály felülreprezentáltságát, főként az idős nőkét. A másik két korosztály a kellőnél kisebb számban szerepel, főként a fiatal és a középkorú férfiak.

Ezeket az értékeket f-be helyettesítve a függvény értéke 0,021573301. Ha találunk olyan pontot, ahol az f ennél kisebb értéket vesz fel, akkor a Kish-kulcsnál jobb nem/korcsoport illeszkedést produkáló kiválasztási eljárást definiálhatunk.

Az f függvénynek létezik elvi optimuma a megadott tartományon. f ugyanis folytonos függvény, a megadott tartomány zárt és korlátos. Weierstrass tétele szerint folytonos függvénynek van legkisebb és legnagyobb értéke korlátos és zárt intervallumon. Az elvi optimum megtalálása azonban nem egyszerű matematikai probléma.

A Solver algoritmikus megoldási módszere nem garantálja a valódi optimum megtalálását, ugyanis konvergencia típusú megállási kritériumokat alkalmaz a véges idejű futás biztosítása érdekében. A talált optimum mindenesetre jóval kedvezőbb beállítást eredményez, mint a Kish-táblák által adott 0,021573301. A talált optimum 0,013180972, ennek 1,637-szerese a Kish-sel kapott. Az optimum helye:

k21=8 k22=4
k31=2 k32=0 k33=10
k41=1 k42=4 k43=0 k44=7
k51= 0 k52=12 k53=0 k54=0 k55=0
k61=0 k62=12 k63=0 k64=0 k65=0 k66=0

Ezen értékeket interpretálva elmondható, hogy a várt tendenciák tapasztalhatók. A kétszemélyes háztartásokból 2/3 valószínűséggel férfit választunk, vélhetően a nők Kish-kulcsos felülreprezentáltságát korrigálva. Ugyanígy a három tagú háztartásokból a várhatóan fiatal lány 3. tagot választjuk 5/6 valószínűséggel, korábbi alulreprezentáltságukat javítva. Az optimum esetén adódó nem- és korcsoport arányok:

30. táblázat: A módosítás esetén várható minta összetétele

	férfi	nő
1. korcsoport	c1=0,1868	c4=0,2030
2. korcsoport	c2=0,1280	c5=0,1593
3. korcsoport	c3=0,1306	c6=0,1921

Láthatóan javul a fiatal korosztályok illeszkedése, viszont a középkorúak és az idős férfiak részaránya távolabb kerültek a populációstól, mint a Kish esetén kapott. Az idős nők aránya viszont javult, a populációéhoz közeledve.

Felvethető az a kérdés, hogy ha nem köt minket a táblatípusok számának korlátja (ami az előbb 12 volt), mennyire javítható ez az optimum. A táblák számának növelése nyilván költségnövekedést eredményez - nagyobb adminisztrációt, nagyobb nyomdaköltséget igényelhet. Azonkívül a minta nagysága maga is felső korlátja a kioszthatóó táblák számának. Az elvi probléma azonban végiggondolásra érdemes.

Az optimumkeresés módosulása az egészértékűségi feltétel elhagyásával formalizálható:

$f(c1..c6) \rightarrow \min$.
$p21 + p22 = 1$

$p31 + p32 + p33 = 1$
$p41 + p42 + p43 + p44 = 1$
$p51 + p52 + p53 + p54 + p55 = 1$
$p61 + p62 + p63 + p64 + p65 + p66 = 1$
$pij \geq 0 \; \forall \; i, j - re$

A Solver által talált optimum értéke 0,013135139. Ez az érték 0,9965-szerese az egészértékűségi feltétel mellett adott értéknek, nem eredményezett sokkal jobb eredményt a feltételek elhagyása. Az optimumot adó pij-kből számolt várható nem/korcsoport arányok a következők:

31. táblázat: Pozitivitás feltétel melleti optimum

	férfi	nő
1. korcsoport	c1=0,1875	c4=0,2020
2. korcsoport	c2=0,1293	c5=0,1580
3. korcsoport	c3=0,1307	c6=0,1920

Az arányok első két tizedesjegye nem változott az egészértékűségi feltétel mellett kapott eredménytől. A $pijk$-k a következők:

p21=0,6513 p22=0,3487
p31=0,2467 p32=0,0000 p33=0,7533
p41=0,0000 p42=0,4060 p43=0,0000 p44=0,5940
p51=0,0000 p52=0,8296 p53=0,0000 p54=0,1714 p55=0,0000
p61=0,0000 p62=1,0000 p63=0,0000 p64=0,0000 p65=0,0000 p66=0,0000

A pij értékek is igen közel állnak az egészértékű probléma optimális megoldásához. Most is, mint az egészértékű probléma esetén, előálltak 0 valószínűségek. Korábban a mintaválasztási eljárásokkal szembeni alapvető követelmények között soroltam fel a pozitív kiválasztási valószínűségek feltételét. A keret megváltoztatását jelenti a 0 értékű kiválasztási esély feltételezése, hiszen ezzel szűkítjük a kiválaszthatók körét. Ugyanezeket a problémákat variálhatjuk úgy is, hogy a pozitivitásfeltétel helyett más alsó határt szabunk meg. Az utóbbi, egészértékfeltétel nélküli optimalitási feladatot megoldottam 1/100-os alsó határ megkövetelésével is. A megoldás erős stabilitást mutatott: a korábbi 0 értékek 1/100-ra változtak, a korábbi nemnulla értékek ennek megfelelően enyhén csökkentek. Az f értéke kissé megnőtt, de csupán néhány százalékkal. A 0 megoldásokat kiküszöbölendő, a 12 táblát feltételező feladatban az egészértékű feltétel mellett ilyenkor az 1/12-es alsó korlátot kell megadni. A megoldás itt már várhatóan jobban romlik, az 1/12-es érték jelentős megszorítás. A kapott optimum 0,013938589, ami a korábbihoz képest 5,74 százalékpontos növekedés, tehát nem jelentős az optimum romlása. A kapott pij értékek:

k21=8 k22=4
k31=2 k32=1 k33=9
k41=1 k42=3 k43=1 k44=7
k51=1 k52=8 k53=1 k54=1 k55=1
k61=1 k62=7 k63=1 k64=1 k65=1 k66=1

Ezeket az értékeket felhasználva a mintavételi eljárás beállítására, várhatóan az alábbi nem és korcsoport arányokat kapjuk:

32. táblázat: Pozitivitási és egészértékűségi feltétel melleti optimum

	férfi	nő
1. korcsoport	c1=0,1856	c4=0,1990
2. korcsoport	c2=0,1281	c5=0,1616
3. korcsoport	c3=0,1311	c6=0,1942

A táblázatbeli arányok két tizedesjegyre kerekítve megegyeznek az egészértékűségi és nemnegativitási feltétel nélküli (tehát a legkevesebb megszorítást tartalmazó) modell optimuma által adott kerekített arányokkal. Azaz ez a megoldás nem ad lényegesen különböző eredményt és nem is sokkal rosszabb az f értékét tekintve.

A kapott pij értékekkel a következő Kish-tábla alakítható ki:

33. táblázat: Módosított Kish-tábla

A tábla aránya	A tábla jelzete	Ha a felnőttek száma: 1	2	3	4	5	6 vagy több
		A kiválasztott sorszáma:					
1/12	1.	1	1	1	1	1	1
1/12	2.	1	1	1	2	2	2
1/12	3.	1	1	2	2	2	2
1/12	4.	1	1	3	2	2	2
1/12	5.	1	1	3	3	2	2
1/12	6.	1	1	3	4	2	2
1/12	7.	1	1	3	4	2	2
1/12	8.	1	1	3	4	2	2
1/12	9.	1	2	3	4	2	3
1/12	10.	1	2	3	4	3	4
1/12	11.	1	2	3	4	4	5
1/12	12.	1	2	3	4	5	6

Fontos megjegyezni, hogy a Kish-kulcs módosításakor kizárólag a nem/korcsoport szerinti illeszkedés javítását tartottuk szem előtt. Bár speciális kutatói cél lehet ez a fajta reprezentativitás, a kulcs módosítása nem garantálja feltétlenül a nyers, súlyozatlan becslések javulását. Ugyanakkor ha a torzítatlan becslés érdekében a kiválasztási valószínűségek reciprokával, az alapsúllyal szeretnénk súlyozni, a súlyok nagyobb változékonyságával kell számolnunk. Ugyanis, míg a Kish-kulcs egyenletes kiválasztási valószínűségeket biztosított háztartáson belül, a módosított változatok sok esetben igen kicsi háztartáson belüli kiválasztási valószínűséget javasolnak, s ezek invertálás után igen nagy súlyértékeket eredményeznek.

10. Összefoglaló

Dolgozatom legfontosabb megállapításai a következők:

- Nincs bizonyíték arra, hogy a Kish-kulcs magasabb visszautasítási arányt okozna más, funkcionálisan hasonló eljárással összevetve.
- A kulccsal vett minták magyarországi minták nem és kor jellemzők szerint lényegesen eltérnek a populációtól. Ennek oka az, hogy a kulcsot alkalmazva a minta az aktuális háztartásszerkezet függvénye.
- Ezzel szemben a Kish-módszer a szimulációban jól teljesíti eredeti célját: a háztartáson belüli random választás eredményével csaknem megegyező mintákat produkál.
- A szimulációs becslésben alkalmazott súlyok eltérő, de nem egyértelműen jellemezhető becsléseket eredményeztek. Mégis szem előtt tartandó, hogy várható tulajdonságaik szerint, sok minta átlagában, összevethetőek.
- A kulcs eredményesen javítható, ha célunk a minta bizonyos jellemzők szerinti reprezentativitásának javítása.

Azt gondolom, hogy sikerült bizonyítanom a kulccsal kapott minta összefüggését a populációs háztartásszerkezettel. Az általam felkutatott források nem bizonyították ezt a kapcsolatot, sőt, általában nem is említették ezt a tényezőt. Másik eredményem, a kulcs módosítása önmagában is értékelhető. De fontos szempont, hogy a megoldásra alkalmazott algoritmus sokkal szélesebb körben is alkalmazható. Használatával meghatározhatunk más, nem a kor és nem jellemzők szerinti reprezentativitást javító optimális kulcs-variánst is. Sőt hasonló problémára vezet bármely, jól definiált mintavételi eljárás optimalizálása is.

Természetesen vetődnek fel további, még nyitva álló kérdések. Elemzésemet csak a nem és kor szerinti reprezentativitásra szorítottam meg. Vizsgálandó lenne a Kish-kulccsal kapott minták más, lényeges jellemzők szerinti illeszkedése. A kulcs-módosítás eredményeként kapott eljárás tesztelése is szükséges

lenne, ugyanis a nem és kor szerinti illeszkedés javulása nem feltétlenül jelenti a minta más jellemzők szerinti javulását. A módosított kulcs használatával kapott mintavételi valószínűségek megváltozása is további vizsgálatokra szorul. Továbbá a kulcs módosítását csupán a Magyar Háztartás Panelből vezettem le, a megnyugtató érvényességhez szükséges lenne ugyanezt az optimalizáló feladatot a teljes magyar háztartásstruktúra jellemzőiből kiindulva végrehajtani.

11. Függelék

1. táblázat: Férfiak arányának eloszlása - háztartáson belüli random választás, 100 minta

	Percentilis	Szélsőértékek		
1%	0,3417	Legkisebbek		
5%	0,3668	0,3367		
10%	0,3719	0,3467	Minta elemszáma	100
25%	0,3920	0,3467		
		0,3518	Átlag	0,4162
50%	0,4121		Szórás	0,0366
		Legnagyobbak	Variancia	0,0013
75%	0,4397	0,4824	Ferdeség	0,3477
90%	0,4724	0,4925	Csűcsosság	2,6586
95%	0,4799	0,5025		
99%	0,5050	0,5075		

2. táblázat: Fiatalok arányának eloszlása - háztartáson belüli random választás, 100 minta

	Percentilis	Szélsőértékek		
1%	0,2839	Legkisebbek		
5%	0,3040	0,2714		
10%	0,3216	0,2965	Minta elemszáma	100
25%	0,3367	0,3015		
		0,3015	Átlag	0,3619
50%	0,33618		Szórás	0,0345
		Legnagyobbak	Variancia	0,0012
75%	0,3819	0,4271	Ferdeség	0,0451
90%	0,4095	0,4271	Csűcsosság	2,5483
95%	0,4221	0,4322		
99%	0,4347	0,4372		

3. táblázat: Középkorűak arányának eloszlása - háztartáson belüli random választás, 100 minta

	Percentilis	Szélsőértékek		
1%	0,2211	Legkisebbek		
5%	0,2563	0,2211		
10%	0,2638	0,2211	Minta elemszáma	100
25%	0,2864	0,2462		
		0,2513	Átlag	0,3059
50%	0,3015		Szórás	0,0345
		Legnagyobbak	Variancia	0,0012
75%	0,3241	0,3819	Ferdeség	0,3246
90%	0,3518	0,3869	Csűcsosság	3,353
95%	0,3693	0,3920		
99%	03970	0,4020		

4. táblázat: Idősebbek arányának eloszlása - háztartáson belüli random választás, 100 minta

	Percentilis	Szélsőértékek		
1%	0,2563	Legkisebbek		
5%	0,2814	0,2563		
10%	0,2965	0,2563	Minta elemszáma	100
25%	0,3116	0,2714		
		0,2814	Átlag	0,3323
50%	0,3317		Szórás	0,0284
		Legnagyobbak	Variancia	0,0008
75%	0,3518	0,3768	Ferdeség	-0,2182
90%	0,3668	0,3819	Csűcsosság	3,0145
95%	0,3744	0,3869		
99%	0,3970	0,4070		

5. táblázat: A p-k eloszlása a khi-négyzet tesztekből - háztartáson belül random választás

	Percentilis	Szélsőértékek		
1%	0,000	Legkisebbek		
5%	0,0055	0,000		
10%	0,012	0,000	Minta elemszáma	100
25%	0,025	0,000		
		0,004	Átlag	0,1948
50%	0,0975		Szórás	0,2273
		Legnagyobbak	Variancia	0,0517
75%	0,2915	0,771	Ferdeség	0,13358
90%	0,6045	0,797	Csűcsosság	3,6481
95%	0,663	0,814		
99%	0,8315	0,849		

6. táblázat: A Kish-táblák kiosztási aránya

Tábla száma	Gyakoriság	Százalékban	Ideális százalékarány
1	34	17,09	16
2	17	8,54	8
3	17	8,54	8
4	34	17,09	16
5	33	16,58	16
6	16	8,04	
7	16	8,04	
8	32	16,08	
O2sszesen	199	100	

7. táblázat: A normált alapsűly eloszlása

	Percentilis	Szélsőértékek		
1%	0,484	Legkisebbek		
5%	0,484	0,484		
10%	0,484	0,484	Minta elemszáma	199
25%	0,484	0,484		
		0,484	Átlag	1
50%	0,992		Szórás	0,409
		Legnagyobbak	Variancia	0,170
75%	1,416	2,006	Ferdeség	0,563
90%	1,504	2,006	Csűcsosság	2,897
95%	1,888	2,006		
99%	2,006	2,006		

8. táblázat: A normált sűly eloszlása utólagos rétegzés után

	Percentilis	Szélsőértékek		
1%	0,383	Legkisebbek		
5%	0,383	0,383		
10%	0,383	0,383	Minta elemszáma	199
25%	0,570	0,383		
		0,383	Átlag	1
50%	1,008		Szórás	0,437
		Legnagyobbak	Variancia	0,191
75%	1,138	2,066	Ferdeség	0,544
90%	1,644	2,066	Csűcsosság	3,082
95%	1,869	2,192		
99%	2,192	2,192		

9. táblázat: A normált sűly eloszlása, puszta utólagos rétegzés használatakor

	Percentilis	Szélsőértékek		
1%	0,644	Legkisebbek		
5%	0,644	0,644		
10%	0,644	0,644	Minta elemszáma	199
25%	0,644	0,644		
		0,644	Átlag	1
50%	1,094		Szórás	0,213
		Legnagyobbak	Variancia	0,045
75%	1,159	1,167	Ferdeség	-1,016
90%	1,167	1,167	Csűcsosság	2,162
95%	1,167	1,167		
99%	1,167	1,167		

10.táblázat: ci-k a Kish-módosi1tás feladatban, i=2..5

$$c2 = b1 \times a11[2,1] + b2 \times (p21 \times a21[2,1] + p22 \times a22[2,1])+$$
$$+b3 \times (p31 \times a31[2,1] + p32 \times a32[2,1] + p33 \times a33[2,1])+$$
$$+b4 \times (p41 \times a41[2,1] + p42 \times a42[2,1] + p43 \times a43[2,1] + p44 \times a44[2,1])+$$
$$+b5 \times (p51 \times a51[2,1] + p52 \times a52[2,1] + p53 \times a53[2,1] + p54 \times a54[2,1]+$$
$$+p55 \times a55[2,1]) + b6 \times (p61 \times a61[2,1] + p62 \times a62[2,1] + p63 \times a63[2,1]+$$
$$+p64 \times a64[2,1] + p65 \times a65[2,1] + p66 \times a66[2,1])$$
$$c3 = b1 \times a11[3,1] + b2 \times (p21 \times a21[3,1] + p22 \times a22[3,1])+$$
$$+b3 \times (p31 \times a31[3,1] + p32 \times a32[3,1] + p33 \times a33[3,1])+$$
$$+b4 \times (p41 \times a41[3,1] + p42 \times a42[3,1] + p43 \times a43[3,1] + p44 \times a44[3,1])+$$
$$+b5 \times (p51 \times a51[3,1] + p52 \times a52[3,1] + p53 \times a53[3,1] + p54 \times a54[3,1]+$$
$$+p55 \times a55[3,1]) + b6 \times (p61 \times a61[3,1] + p62 \times a62[3,1] + p63 \times a63[3,1]+$$
$$+p64 \times a64[3,1] + p65 \times a65[3,1] + p66 \times a66[3,1])$$
$$c4 = b1 \times a11[1,2] + b2 \times (p21 \times a21[1,2] + p22 \times a22[1,2])+$$
$$+b3 \times (p31 \times a31[1,2] + p32 \times a32[1,2] + p33 \times a33[1,2])+$$
$$+b4 \times (p41 \times a41[1,2] + p42 \times a42[1,2] + p43 \times a43[1,2] + p44 \times a44[1,2])+$$
$$+b5 \times (p51 \times a51[1,2] + p52 \times a52[1,2] + p53 \times a53[1,2] + p54 \times a54[1,2]+$$
$$+p55 \times a55[1,2]) + b6 \times (p61 \times a61[1,2] + p62 \times a62[1,2] + p63 \times a63[1,2]+$$
$$+p64 \times a64[1,2] + p65 \times a65[1,2] + p66 \times a66[1,2])$$
$$c5 = b1 \times a11[2,2] + b2 \times (p21 \times a21[2,2] + p22 \times a22[2,2])+$$
$$+b3 \times (p31 \times a31[2,2] + p32 \times a32[2,2] + p33 \times a33[2,2])+$$
$$+b4 \times (p41 \times a41[2,2] + p42 \times a42[2,2] + p43 \times a43[2,2] + p44 \times a44[2,2])+$$
$$+b5 \times (p51 \times a51[2,2] + p52 \times a52[2,2] + p53 \times a53[2,2] + p54 \times a54[2,2]+$$
$$+p55 \times a55[2,2]) + b6 \times (p61 \times a61[2,2] + p62 \times a62[2,2] + p63 \times a63[2,2]+$$
$$+p64 \times a64[2,2] + p65 \times a65[2,2] + p66 \times a66[2,2])$$

$$
\begin{aligned}
c6 = {} & b1 \times a11[3,2] + b2 \times (p21 \times a21[3,2] + p22 \times a22[3,2]) + \\
& + b3 \times (p31 \times a31[3,2] + p32 \times a32[3,2] + p33 \times a33[3,2]) + \\
& + b4 \times (p41 \times a41[3,2] + p42 \times a42[3,2] + p43 \times a43[3,2] + p44 \times a44[3,2]) + \\
& + b5 \times (p51 \times a51[3,2] + p52 \times a52[3,2] + p53 \times a53[3,2] + p54 \times a54[3,2] + \\
& + p55 \times a55[3,2]) + b6 \times (p61 \times a61[3,2] + p62 \times a62[3,2] + p63 \times a63[3,2] + \\
& + p64 \times a64[3,2] + p65 \times a65[3,2] + p66 \times a66[3,2])
\end{aligned}
$$

12. Hivatkozások

Binson, D., Canchola, J.A., Catania, J.A. (2000). Random Selection in a telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods. Journal of Official Statistics, 16., 53-59.

Botman S.L., Moore T.F., Moriarity C.L., and Parsons V.L. Design and Estimati.on for the National Health Interview Survey, 1995-2004. National Center for Health Statistics. Vital and Health Statistics 2(130) 2000.

Carlson, B. L. (1998). Software for Statistical Analysis of Sample Survey Data. http://www.fas.harvard.edu/~stats/survey-soft/blc-eob.html

Encyclopedia of the Social Sciences.(1930). E.R. Selgiman & A. Johnson, New York, London.

Groves, R.M., (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons, Inc.

Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L., és Waksberg, J. (szerk.). (1988). Telephone Survey Methodology. New York: John Wiley and Sons, Inc.

Kennedy, J.M., (1993). A Comparison of Telephone Survey Respondent Selection Procedures. (http://www.indiana.edu/ csr/aapor93.html)

Kish, L.(1949). A Procedure for Objective Respondent Selection Within the Household. Journal of the American Statistical Association, 380-387.

Kish, L.(1965). Survey Sampling. New York: John Wiley and Sons, Inc.

Lavrakas, P.J., (1993). Telephone survey methods: Sampling, selection and supervision. Applied Social Research Methods Series, Vol 7. Sage

Levy, P.S., Lemeshow, S., (1999). Sampling of Populations. New York: John Wiley and Sons, Inc.

Magyar Statisztikai Évkönyv 1993, Központi Statisztikai Hivatal, Budapest, 1994.

Magyar Statisztikai Évkönyv 1999, Központi Statisztikai Hivatal, Budapest, 2000.

Mikrocenzus, 1996 Főbb eredmények, Központi Statisztikai Hivatal, Budapest, 1996.

Mikrocenzus, 1996 A népesség és a lakások jellemzői, Központi Statisztikai Hivatal, Budapest, 1996.

Oldendick, R.W., Bishop, G.G., Sorenson, S.B., és Tuchfarber, A.J. (1988). A Comparison of the Kish and Last Birthday Methods of Respondent Selection in Telephone Surveys. Journal of Official Statistics,4,307-318.

Piazza, T., Respondent selection for CATI/CAPI (Equivalent to the Kish Method). http://srcweb.berkeley.edu:4229/res/rsel.html

Särndal, C-E., Swensson, B., Wretman, J., (1997). Model Assisted Survey Sampling. Springer-Verlag New York, Inc.

Tarjányi József. Módszertani problémák a telefonos közvélemény-kutatásokban. http://www.c3.hu/scripta/scripta0/replika/1920/14vita.htm

Warwick, D.P., Lininger, C.A., (1975). The Sample Survey: Theory and Practice. McGraw-Hill, Inc.

13. Felhasznált adatbázisok

- *ISSP 1994. Család II.* A felmérés adatait a TÁRKI Társadalomtudományi Adatbank bocsátotta rendelkezésemre. A kutatást vezető intézmény a TÁRKI. Az adatfelvétel időpontja 1994. május-június. Az 1990-es KSH Népszámlálás lakásmintájából készült országos reprezentatív minta. Módszere: kérdőív kérdezőbiztosokkal. Esetszám: 1500

- *Magyar Háztartás Panel IV. hullám* A felmérés adatait a TÁRKI Társadalomtudományi Adatbank bocsátotta rendelkezésemre. A kutatást vezető intézmények a BKE Szociológiai Tanszék és

a TÁRKI. Az adatfelvétel időpontja 1995. Országosan reprezentatív háztartás minta. Módszere: kérdőíves adatfelvétel. Esetszám: 8043

- *OLEF-2000.* A kutatást az Egészségfejlesztési Kutatóintézet vezette, a terepmunka kivitelezője a Magyar Gallup Intézet volt. Az adatfelvétel időpontja 2000. november-december. A Központi Népességnyilvántartó és Választási Hivatal címlistájából készült országos reprezentatív minta. Módszere: kérdőív kérdezőbiztosokkal. Esetszám: 5534