A rétegzett mintavételről

Galambosné Tiszberger Mónika,
a Pécsi Tudományegyetem egyetemi tanársegédje
Email: tiszbergerm@ktk.pte.hu

A mintavételes eljárások gyakorlati alkalmazása elképzelhetetlen rétegzés nélkül. A tanulmány első felében a szerző a rétegzés sajátosságait vizsgálja, kiemelten azt, hogy milyen okokból alkalmaznak a statisztikusok rétegzett mintákat. Véleménye szerint ugyanis a sokasághoz képest homogénebb rétegek kialakításának célja mellett számos egyéb megfontolásból is a rétegzés módszertanát (technikáját) használjuk fel a mintavétel során.

A tanulmány második felében a szerző a mintanagyság rétegek közötti elosztásának lehetőségeit vizsgálja. Az allokációs módszerek hatásosságát hasonlítja össze két réteget feltételezve, konkrét esetekre vetítve. Számításai mondanivalójának lényege, hogy a Neyman-féle optimális rétegzés sok, konkrétan meghatározható esetben nem javítja jelentősen a becslés hatékonyságát. Vagyis, bizonyos kiinduló feltételek alapján már a mintavételi terv kialakításának az elején támpontot kaphatunk ahhoz, hogy érdemes-e egyáltalán a rétegeken belüli szórások összegyűjtésével, becslésével foglalkozni. A számítások a gyakorlati munkához is ötleteket adhatnak a mintavétellel foglalkozó szakemberek számára. Tanulmánya befejezéseként a szerző az eredményeket egy valós mezőgazdasági példával is illusztrálja.

TÁRGYSZÓ:
Rétegzett mintavétel.

A statisztikai módszerek leglényegesebb csoportosítása aszerint történik, hogy megfigyelhetőnek tekintjük-e a egész sokaságot vagy sem. Ha a sokaságnak csak egy megfelelően kiválasztott részét tudjuk megfigyelni és az alapján vonunk le következtetéseket a sokaságra, következtetéses statisztikáról beszélünk. Ennek két ága a becsléselmélet és a hipotézisvizsgálat (*Pintér–Rappai* [2007]). E módszerek alapját tehát egy bizonyos módon kiválasztott és megfigyelt mintasokaság adatai adják.

A mintavétel elméleti megalapozása a matematikai statisztika talán legfontosabb vívmánya a gazdasági-társadalmi jelenségek megfigyelése, számbavétele terén. A statisztika tárgya a tömegjelenségek vizsgálata. Ugyanakkor a tömeget nyilvánvalóan nem lehet, vagy legalábbis nem kifizetődő, nem célszerű teljes egészében megfigyelni. A teljes sokaság megfigyelése gyakran túl költséges és hosszú időt igénylő folyamat. Ezért jó, hogy a sokaság egy megfelelően kiválasztott, viszonylag kis részéből is megbecsülhetők a kérdéses mutatók, belőle következtetések vonhatók le a teljes sokaságról. Természetesen annak, hogy megfigyeléseink nem teljes körűek, ára van. Ezt hívjuk mintavételi vagy becslési hibának. Azonban, tetszőlegesen kiválasztott megbízhatósági szint mellett a bizonytalanság is számszerűsíthető matematikai módszerekkel. Így a hibázás mértékének is tudatában lehetünk, és ez az ismeret már támpontot nyújthat a felhasználóknak az adatok minőségének megítélésében.

1. A rétegzés lehetséges okai, céljai

A statisztika tankönyvek által is bemutatott mintavételi módszerek két nagy csoportja a véletlenen alapuló és a nem véletlen eljárások. A rétegzett mintavétel a csoportos és a többlépcsős módszerekkel együtt az egyszerű véletlen eljárásokba tartozik. Részletesen nem térek ki az egyes módszerek ismertetésére, hiszen a cikknek nem célja ez. A továbbiakban csak a rétegzett mintavétel részleteivel, sajátosságaival foglalkozom, további csoportosítási lehetőséget bemutatva. (A rétegzett mintavételről más szempontok, célok szerint olvashattunk korábban *Fraller* [2011] cikkében. Jelen írás részben ennek kiegészítéseként is tekinthető.)

A rétegzés a sokaság egy vagy több alkalmasan megválasztott szempont szerinti csoportosítását jelenti. Ez az általános megfogalmazás már magában rejti a rétegzett mintavétel további csoportosítási lehetőségét. Eszerint ugyanis maga az ismérv, illetve a rétegzés történhet „bárhogyan”. Általánosságban tehát nem mondunk semmit a rétegképző ismérvek tulajdonságairól. Nyilván szükség van arra, hogy a felhaszná-

lásra szánt rétegképző ismérv(ek) szempontjából előzetes információval rendelkezzünk, és annak alapján minden sokasági elemet el tudjunk helyezni. A rétegzést, a csoportosításhoz hasonlóan, egyértelműen kell elvégezni, vagyis minden elemet pontosan be kell sorolni egy rétegbe, illetve egy elem nem tartozhat egyszerre több réteghez. A teljes sokaság nagysága ebben a szemléletben így írható fel:

$$N = N_1 + N_2 + ... + N_L = \sum_{h=1}^{L} N_h ,$$

ahol:

N – az alapsokasági elemszám;
h – a rétegek;
L – az összes réteg száma.

Felmerül a kérdés, hogy milyen tulajdonságokkal rendelkezik egy „megfelelő" rétegképző ismérv. Ez az a pont, ahol a szakirodalom nem kezd el kategóriákat képezni, hanem kétféleképpen jár el. Az egyik irány az, amikor minden fajta rétegzést „egy kalap alá" vesznek (*Ay* [1976], *Cochran* [1977], *Éltető* [1982], *Kish* [1995]), nem törődve a rétegképző ismérvek tulajdonságaival. A rétegzést mint technikát, mint eljárást kezelik és felsorolják, hogy egyébként a rétegképzés milyen ismérvek mentén, milyen kiinduló feltételek mellett valósul meg. A másik elgondolás szerint csakis az számít rétegzésnek, amikor a becsülni kívánt változóval fennálló sztochasztikus kapcsolat alapján történik a rétegzés, vagyis a sokaságot homogénebb rétegekre bontjuk a mintavétel előtt (*Marton* [1991], *Hunyadi–Vita* [2002], *Pintér–Rappai* [2007]). Ilyenkor minden más esetet figyelmen kívül hagynak. Marton ezen felül a területi rétegzést említi, de nem is rétegeknek nevezi a területi egységeket, hanem tartományoknak. Legtöbbször azonban szó sem esik más lehetőségekről.

A rétegképzés alapjától függetlenül, módszertanilag minden esetben ugyanúgy járunk el, mind a mintavétel, mind a becslések elkészítése során. A végső mintába minden rétegből kerülnek be elemek. A rétegenkénti független mintavétel következménye az, hogy a rétegeken belüli standard hiba négyzeteinek súlyozott összegeként adódik a becslés hibájának mértéke. Ez azt jelenti, hogy a külső szórást gyakorlatilag eliminálják, és csak a belső szórások mértéke fogja befolyásolni a rétegzett mintából történő becslések hibáját (természetesen a mintavételi arány és a megbízhatósági szint mellett). A rétegeken belüli kiválasztás módszere, aránya lehet csak eltérő, de mindenképpen szerepel az összes réteg. A rétegzett mintavétel módszerének típusait a következők szerint lehet összefoglalni.

1. A becsülni kívánt változóval sztochasztikus kapcsolatban álló ismérvet keresünk, és ez fogja jelenteni a rétegképzés alapját. Egy ilyen ismérv megválasztása el-

sősorban a standard hibára gyakorolt jótékony hatása miatt előnyös. Ekkor a heterogén sokaságot viszonylag homogén (homogénebb) rétegekre bontjuk. Rétegzett mintavételből nyert adatokból történő becslés esetén a rétegek közötti eltérésektől, mint már említettem, eltekintünk, és csak a belső szórás fog szerepet játszani a standard hiba számszerűsítése során. Tehát a külső szórás nagysága csökkenti a becslési hibát. Ráadásul a vizsgált mutató(k) belső szórását ebben az esetben viszonylag alacsonynak gondoljuk, hiszen éppen az volt a lényeg, hogy a rétegképző ismérvek szempontjából egymáshoz hasonló egyedek kerüljenek azonos rétegbe, így a teljes sokasághoz képest lényegesen homogénebb csoportokat kapunk. Ennek megfelelően az eliminálható külső szórás képvisel viszonylag magasabb arányt. Ebben az esetben tehát az ismérv megválasztásával hajtunk végre optimalizálást a mintavételi hiba nagyságának csökkentése érdekében. Végletes esetben akár maga a vizsgált ismérv is lehetne a kiválasztott rétegképző ismérv, de mivel erre irányul a felmérés, erről pontos információnk nyilvánvalóan nincs, esetleg valamilyen közelítő ismerettel rendelkezhetünk róla. (Erre lehet megoldás a kétfázisú mintavétel módszere (*Hunyadi* [1991])). Az esetek többségében azonban inkább valamilyen segédváltozót vagy -változókat használunk fel. (A segédváltozó felhasználásával történő rétegzés hatékonyságjavító tulajdonságáról pontosabb képet kapunk *Fraller* [2011] cikkében, eltérő allokációs technikák mellett.)

2. Erős indoka lehet a rétegzésnek az összeírási költségek optimalizálása. Előfordulhat, hogy nagyon eltérő költségekkel tudunk csak bizonyos részsokaságokat elérni, ezért eszerint is elvégezhető a rétegzés, és így a költségek bizonyos határok között maradhatnak azáltal, hogy a „drágább" elemekből kisebb mintával is megelégszünk. Tehát a mintavétel során tisztán költség alapú optimalizálás esetében fel sem merül a homogenizálás igénye.

3. Rétegzést azért is alkalmazhatunk, mert a különböző rétegekbe tartozó egyedekre másféle összeírási, megfigyelési módszert szeretnénk, vagy vagyunk kénytelenek alkalmazni. Ennek egyik oka az lehet, hogy fizikailag különböző helyen találjuk meg az egyes rétegekhez tartozó tagokat. Például egy multinacionális vállalat esetében valószínűleg más módszereket alkalmaznak egy dolgozói felmérés esetén az anyavállalatnál dolgozókra, illetve a leányvállalatok munkavállalóira, akik sok esetben nagyon távol vannak a központtól, és ebből adódóan akár eltérő kulturális jegyekkel is bírnak. Másik ilyen szempont az lehet, amikor az egyes sokasági elemek túlzottan eltérő természetűek. Ezen azt értem, hogy másféle módszerrel lehet őket hatékonyan felmérni. Egy szellemi munkát végző ember valószínűleg könnyebben tölt ki egy kérdőívet, mint egy betanított fizikai munkás. Lehetséges, hogy az utóbbi csoporttal célravezetőbb interjút készíteni, vagy összeírói segítséget adni a kérdőív kérdéseinek megválaszolásához. Ilyen példát a mezőgazdaságból is hozhatunk, ahol a gazdasági szervezetek, illetve a legnagyobb egyéni gazdaságok esetében feltételezzük, hogy rendelkeznek olyan szintű nyilvántartásokkal, könyveléssel, illetve szakér-

telemmel, amelynek alapján önállóan is ki tudják tölteni a postai úton megkapott kérdőíveket (*Galambosné Tiszberger* [2009]). Ugyanakkor az egyéni gazdaságok nagyobb részéhez, akik kisebb volumenben, kevesebb nyilvántartással dolgoznak, inkább összeírókat küldünk ki, akik segítenek – a gazdasági szervezetekével azonos tartalmú – kérdőívet pontosan kitölteni. Ilyen esetekben sem valószínű, hogy a rétegek a teljes sokaságon belül homogénebb részsokaságokat jelentenek, hiszen a mintavételkor nem vesszük figyelembe a vizsgált változókat.

4. Okként merülhet fel az a helyzet, hogy másféle információval rendelkezünk a különböző egyedekről. A mezőgazdasági statisztika egyik problémája az, hogy az egyéni gazdaságok azonosító adatairól nincs naprakész regiszter. A cenzusok alkalmával létrejön ugyan egy teljes sokaságot lefedő nyilvántartás, de annak frissítése két teljes körű összeírás között csak részben lehetséges. Ugyanakkor a cégnyilvántartás alapján a gazdasági szervezetekről viszonylag pontos, naprakész és teljes képet kapunk. A szervezeteknek tehát lehet postai úton küldeni a kérdőívet, hiszen pontos név- és címadatok vannak. Az egyéni gazdaságok esetében ez jelenleg nem megoldható, vagy legalábbis a visszaérkezési arányt lényegesen rontaná a kézbesítetlen levelek mennyisége. Módszertani szempontból ekkor is két rétegről beszélünk, viszont az összeírt változók tulajdonságait egyáltalán nem vesszük figyelembe.

5. Az is indokolhatja bizonyos rétegek külön kezelését, hogy az így képzett részsokaságra vonatkozóan is szeretnénk látni a becslés eredményeit, illetve ha külön előírások vannak bizonyos részsokaságokból elért pontosságra, hibanagyságra vonatkozóan. A magyar viszonyok között, és általában az EU többi országában is, ilyen rétegeket képez a NUTS II-es szintű területi beosztás. Magyarországon ez a régiók szintje. Tehát a régiót a legtöbb esetben már a kiinduláskor rétegképző ismérvnek tekintjük. Erre azért van szükség, hogy régiós adatokat is közölhessenek a Központi Statisztikai Hivatal kiadványai. Másképpen fogalmazva, azt szeretnénk, hogy a minta a régiókra is reprezentatív legyen. Ezt a hazai felhasználói igények, illetve az EU-s szabályozás is szükségessé teszik. Esetenként még megyei szintű adatok előállítására is sor kerül a hazai igények miatt. Az angol nyelvű szakirodalomban egyébként külön elnevezést („domain") is kapnak az ilyen jellegű rétegek. A magyar terminológiában ezt a fajta „területi" rétegzést nem különítjük el. A legtöbb téma szempontjából a területi besorolás, rétegzés ugyancsak nem a homogenitást növeli. Inkább az országos eloszlás leképeződése látszik a területi szinteken is. Vannak persze kivételes témák, amelyek esetében a földrajzi hovatartozás homogenizál, de ezek előfordulása csak esetleges.

6. Gyakori vezérelv a rétegzés mögött az is, amikor a reprezentativitást szeretnénk biztosítani bizonyos tulajdonságok mentén. Például a demográfiai jellemzők rétegképző ismérvként alkalmazása mindennapi gyakorlat a marketingkutatásoknál. Az életkor, a végzettség, és a nem olyan változók, amelyekről ismert sokasági megoszlások állnak rendelkezésre, és ezt a struktúrát szeretnénk viszontlátni a mintasokaság-

ban is. Ezzel biztosítható az, hogy a fontosnak ítélt szempontok szerint valóban reprezentatív, a sokaságra „hasonlító” mintánk legyen. Az azonban, hogy ezáltal pontosabb becslések adhatók a vizsgált változók értékeiről, korántsem biztos. Feltételezhető, hogy például a vásárlási attitűdöt befolyásolják ezek a tényezők, de nem ez az elsődleges oka a rétegzésnek.

7. Egy speciális esetként fogható fel, amikor ritka populáció – azaz olyan közösségek, társadalmi csoportok, amelyekről közvetlen mintavételi keret nem áll rendelkezésre – megfigyelése a cél. Ilyenkor úgynevezett aszimmetrikus mintarétegzést hajthatunk végre. Ezt akkor célszerű alkalmazni, ha a ritka populáció rétegképző ismérvek szerinti eloszlása ugyan ismeretlen, de azt tudjuk róla, hogy nem egyenletes. Azokat a rétegeket szükséges felülreprezentálni a mintavétel során, amelyekben a ritka populáció tagjainak előfordulása nagyobb valószínűséggel bír. Így jó eséllyel növelhető a hasznos minta elemszáma (*Kapitány* [2010]). Példa lehet erre a magas vérnyomásban szenvedők célsokasága, ahol a célszemélyek előfordulása a középkorú, illetve idősebb népesség körében valószínűleg gyakoribb. A rétegzés tényét és mikéntjét itt a célsokaság minél jobb elérése motiválja.

A rétegzés egyik hasznos eredménye az, hogy a rétegképző ismérv(ek) szempontjából a sokaság minden fontos része be fog kerülni a mintába, így ilyen értelemben reprezentálni fogja a sokaságot. Ez általánosságban igaz. Emellett az előbbi felsorolásból jól látható, hogy többféle indok, indokrendszer alapján juthatunk el a rétegzett mintavétel valamilyen formájának használatához. Mindebből azt tartom a legfontosabbnak, hogy a becslés hatásossága, pontossága szempontjából a gyakorlatban csakis az első eset az, ami biztosan javítja a becsléshez tartozó hibaszámok alakulását az egyszerű véletlen kiválasztáshoz képest. Hiszen ekkor azért és csak azért választjuk ki a rétegképző ismérve(ke)t, mert az célszerűen erős sztochasztikus kapcsolatban áll a vizsgált változóval, és ezáltal homogénebb csoportokat hoz létre a megfigyelés célja szerint. Az összes többi (2–7.) esetben más motiválja a rétegzést, és a homogénebb csoportok létrejötte csak esetleges, de semmiképpen sem törvényszerű velejáró. Véleményem szerint ebből pedig az következik, hogy lényegét tekintve külön alcsoportként kezelendő az első eset az összes többitől. Az első esetet *kapcsolati rétegzésnek* neveztem el, hiszen az ismérv és a vizsgált változó kapcsolata alapján dől el a rétegképzés kérdése. A többi esetet összefoglalóan *technikai rétegzésnek* hívom, mivel a mintavétel és a becslések elvégzése során a lépéseket és a módszertant tekintve úgy járunk el, mint az első esetben, nem feltétlenül érünk el számottevő csökkenést a mintavételi hiba nagyságában (az egyszerű véletlen kiválasztáshoz képest). Azért is tartom fontosnak egy ilyen fajta elkülönítés bevezetését, mert a szakirodalom azt sugallja, hogy a rétegzés a becslés hibáját javító művelet, a gyakorlatban viszont ez gyakran nem, vagy csak részben igaz. Látnunk kell tehát, hogy más okok, feltételek is eredményezhetnek rétegzett mintát, a hatékonyság csekély mértékű javítása mellett.

2. A minta elosztása a rétegek között

A rétegzés természetesen valamennyi mintavételi eljárással jól kombinálható. Különböző szempontjait is alkalmazhatjuk egy időben, ha a feltételek megkívánják. A gyakorlatban, a korábban említett előnyök miatt, nagyon elterjedt ez a módszer. A rétegzett mintavétel során a rétegekbe sorolt sokasági elemekből rétegenként különkülön és egymástól függetlenül végezzük el a minta kiválasztását.

A rétegzés indokától, típusától függetlenül merül fel az a kérdés, hogy milyen megoszlással, és mekkora mintát válasszunk ki az egyes rétegekből. Természetesen a maximális, teljes mintaelemszámot (n) meghatározhatjuk egyrészt a pénzügyi korlátokból, másrészt a pontossági követelmények és a sokasági szóródás (rétegenkénti belső szórások) felhasználásával. (A legnehezebb esetben a szükséges mintanagyságot egymással ellentétes irányba befolyásoló mindkét feltétel, a pontossági követelmények és a maximális mintanagyság is adott már a mintavételi terv kialakítása előtt. Például a mezőgazdasági statisztika esetében az állatállományra és a növénytermesztésre vonatkozó éves országos becsléseknél EU-s jogszabályban meghatározott hibahatáron belül kell maradni, a másik oldalról pedig a KSH költségvetése mereven meghatározza, hogy mekkora összeg áll rendelkezésre az összeírások végrehajtására.) A rétegek között egy előre adott, illetve valamilyen feltételek mentén kalkulált mintaelemszámot alapvetően háromféle módon tudunk elosztani. A háromféle módszer tárgyalása során az egyszerűbbtől haladok a bonyolultabb felé.

– *Egyenletes rétegzés.* Ebben az esetben a teljes sokaságra már valamilyen módszerrel meghatározott mintaelemszámot egyszerűen elosztjuk a rétegek számával, és minden rétegből ugyanakkora mintát veszünk ($n_h = n/L$) (n – mintaelemszám). Ez a legegyszerűbb allokációs módszer. Nem igényel előzetes tervezést, és könnyű kivitelezni. Általános feltételek mellett, ha az egyes rétegek mutatóira is kíváncsiak vagyunk, szintén jó megoldás lehet ez a fajta allokáció. Akkor célszerű alkalmazni, ha hasonló méretű rétegeink vannak, ilyenkor az arányoshoz hasonló eredményeket ad.

– *Arányos rétegzés.* Amennyiben a rétegenkénti elemszámok nagyobb eltérést mutatnak (a gyakorlatban ez a jellemzőbb), akkor az egyenletes elosztás helyett általában, a rétegnagysággal arányos vagy egyszerűen csak arányos rétegzés használata lehet célszerűbb. Ebben az esetben a rendelkezésre álló mintanagyságot az egyes rétegek nagyságával arányosan osztjuk szét a rétegek között. Ekkor a mintában a rétegek közötti megoszlás éppen a sokasági összetétellel fog megegyezni. Ez formálisan a következőt jelenti: $n_h/n = N_h/N$, vagyis $n_h/N_h = n/N$.

Ezáltal minden rétegben biztosítjuk az azonos mintavételi arányt, és egy ún. önsúlyozó mintát kapunk. Ezzel a módszerrel a nagyobb elemszámú rétegből nagyobb, a kisebb elemszámú rétegekből pedig kisebb minta kiválasztására kerül sor.

– *Optimális rétegzés.* A mintaelemszám elosztását más szempontok szerint is meghatározhatjuk. Kétféle tényező figyelembe vételével optimalizálhatjuk az elosztást: az egyes elemek összeírási költségeivel, vagy a rétegeken belüli szóródással. Bármelyiket is választjuk, nyilván a minimalizálás lesz a cél. Ezt az allokációt akkor célszerű felhasználni, ha a rétegek között jelentős eltérések vannak az összeírási költségek vagy a szóródások tekintetében (magas a külső szórás aránya a teljes szóráson belül). A költségek oldaláról talán egyszerűbb megragadni a problémát, mert a költség a teljes kérdőívre, vagyis az összes változóra együttesen vonatkozik. Viszont a szóródás esetében ki kell választanunk egy „legfontosabb" vagy egy fontos és viszonylag nagy szórással jellemezhető változót, ami mentén az optimális allokációt elvégezzük. A következő képlet mutatja a minta elosztásának elvét a szórások szerinti optimalizációra:

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} ,$$

ahol:

σ – a vizsgált vagy azzal szoros sztochasztikus kapcsolatban álló változó szórása (általában becsült).

A szórásminimalizálás vezérelte optimális rétegzést Neyman-féle optimális rétegzésnek, vagy Neyman-allokációnak is szokták nevezni. (A módszer az eljárás bevezetőjéről *Jerzy Spława-Neymanról* (1934) kapta az elnevezést.)[1]

A *költségoptimális elosztás* is egy lehetséges módszer. Mivel azonban ez az allokációs típus nem függ a rétegek jellemzőitől, hanem csakis a lekérdezés költségtényezőitől, ezért ezzel a továbbiakban nem foglalkozom.

A kérdés az, hogy bizonyos feltételek mellett melyik elosztás a hatásosabb, illetve az egyszerű véletlen (EV) kiválasztáshoz képest hogyan alakulnak a becslőfüggvé-

[1] Egyes források szerint csak később bukkantak rá, hogy Csuprov már 11 évvel korábban bemutatta a módszert (*Cochran* [1977]).

nyek varianciái (természetesen azonos sokaság, mintanagyság mellett). Ehhez először nézzük meg a *Hunyadi* [2001] által is bemutatott összefüggéseket:[2]

$$\mathrm{Deff}\left(AR\right)=\frac{\mathrm{Var}\left(\overline{y}_A\right)}{\mathrm{Var}\left(\overline{y}_{EV}\right)}\approx\frac{S_B^2}{S^2}\ .$$

Ez az összefüggés azt mutatja be, hogy az arányosan rétegzett mintából történő átlagbecslés minden esetben hatásosabb, mint az egyszerű véletlen mintából számított átlagbecslés, kivéve, ha a belső variancia megegyezik a teljes varianciával, vagyis a külső variancia nulla. Megjegyezzük, hogy ez az összefüggés nem teljesen pontos, mégis ritka, a gyakorlatban nem jellemző esetek kivételével jól mutatja a rétegzés lényegét.

Az optimálisan rétegzett mintából történő becslés az arányoshoz képest még egy tényezővel tudja javítani a becslés hatásosságát. Ez pedig a szórások eltérése saját átlaguktól, vagyis a rétegeken belüli szórások „változatossága". Ebből következik az a megállapítás is, hogy amennyiben a rétegek szórásai megegyeznek, úgy az arányos és az optimális elosztás azonos hatásosságot eredményez (de még mindig jobbak az egyszerű véletlen mintából történő becsléshez képest).

Cochran [1977] is ezt a következtetésláncot fogalmazza meg:

$$\mathrm{Var}\left(\overline{y}_{No}\right)\le\mathrm{Var}\left(\overline{y}_A\right)\le\mathrm{Var}\left(\overline{y}_{EV}\right),$$

ahol:

No – a Neyman-féle optimális,
A – az arányos elosztás,
EV – az egyszerű véletlen kiválasztás.

Az allokációs eljárások alkalmazási feltételeiről, lehetőségeiről ennél többet nem szoktunk olvasni a tankönyvekben, a szakirodalomban. Adja magát az az érdekes kérdés, hogy konkrétabban milyen paraméterek alakulásának függvénye lehet a módszerek közötti választás, illetve hogyan alakul a mintavételi hiba nagysága, vagyis a becslés hatásossága különböző kiinduló feltételek és más-más allokációs eljárások alkalmazása során, számszerűsített formában. Ez a vizsgálat áll a továbbiakban a tanulmány középpontjában.

Az arányos és az egyenletes elosztás között egyértelmű rangsorolási lehetőség nincsen. Bizonyos esetekben az egyik, másokban a másik adja a pontosabb becslés lehetőségét. Az azonban kétségtelen, hogy a három eljárás közül a Neyman-féle op-

[2] Deff-mutatóról (Design Effect) lásd *Hunyadi* [2001a] 12. old.

timális fogja minden esetben a legjobb eredményt adni, már ami a becslés standard hibáját illeti. (Az arányos és az egyenletes elosztás szélsőséges esetben is csak elérni tudja ezt.) Ennek megfelelően mind az arányos, mind az egyenletes elosztás hatásosságát továbbiakban a Neyman-féle allokációhoz képest fogom megvizsgálni.

A később bemutatandó vizsgálatok során a mért változóm a kétféle elosztás melletti standard hiba négyzeteinek hányadosa, egyfajta Deff-mutató lett ($\mathrm{Var}(\bar{y}_{No})/\mathrm{Var}(\bar{y}_A)$, illetve $\mathrm{Var}(\bar{y}_{No})/\mathrm{Var}(\bar{y}_E)$), természetesen azonos sokaságot és mintaelemszámot feltételezve ($n_A = n_{No}$, illetve $n_E = n_{No}$). (Itt az E az egyenletes elosztásra utal.) Ez annyiban tér el a szokásos terminológiától, hogy a hatásosságot nem az egyszerű véletlen kiválasztásból nyert becslés varianciájához hasonlítom. Ez a hányados minden esetben kisebb 1-nél, kivéve néhány speciális szélsőséget, amelyre a későbbiekben kitérek. Az összehasonlításhoz, az egyszerűség kedvéért, két réteget használtam fel. A hányadosokat a két réteg sokasági elemszámának hányadosa (N_1/N_2), illetve a két rétegre jellemző (becsült) szórás aránya (s_1/s_2) függvényében számítottam. A megfelelően nagy sokasági elemszámok, illetve a szórások nagyságrendje nem befolyásolja az eredményeket, ezért elegendő és általános érvényű eredményekre vezet az arányok használata. A mintavétel arányától azonban nem függetlenek az eredmények, ezért 5 százalékos kiválasztási arányt használtam fel az összes számítás során.

2.1. Neyman-féle optimális kontra arányos allokáció

A szakirodalom csak annyit említ ezzel kapcsolatban, hogy az egyenletes és az arányos elosztáshoz képest is a Neyman-féle optimális allokáció javítani fogja a mintavételi hibát (azonos mintaelemszám mellett). A korábbiakban ismertetetteknek megfelelően ez így is van. Ezt az összefüggést akkor is „érezzük”, ha egy pillantást vetünk a rétegzett mintavétel legegyszerűbb, az átlagbecsléshez tartozó hibaképletére (visszatevés nélküli, egyszerű véletlen kiválasztás a rétegeken belül).

$$s_{\bar{y}}^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right),$$

ahol:

s_h – a rétegeken belüli korrigált szórás.

A képletből látható, hogy az N_h/N arány (az egyes rétegek sokasági aránya) az allokációtól függetlenül, mindig ugyanakkora marad. Ugyanakkor a nagyobb szórás-

négyzeteket a Neyman-féle optimális elosztás esetében nagyobb mintaelemszámmal osztjuk, és amit itt nyerünk, az meghaladja a kisebb szórású rétegek kisebb arányából adódó veszteséget. Ezt végiggondolva a következő kérdés adódik: mennyivel hatásosabb a Neyman-féle optimális elosztás az arányoshoz képest? A válasz talán nem meglepő, mégsem nevezhető triviálisnak.

A kétféle elosztás azonos mintamegosztást hoz létre abban az esetben, ha a szórás a két rétegben megegyezik (nincs mire optimalizálni, nagysággal arányos, önsúlyozó elosztás jön létre). Az 1. táblázat néhány olyan (önkényesen kiválasztott) arány-kombinációk esetére mutatja meg a $\mathrm{Var}(\bar{y}_{No})/\mathrm{Var}(\bar{y}_{A})$ hányados értékét, amelynél minden esetben a nagyobb elemszámú réteghez tartozik a szórás magasabb értéke.

1. táblázat

A Neyman-féle optimális és az arányos allokáció relatív hatásossága

N_1/N_2	s_1/s_2								
	0,01	0,05	0,1	0,15	0,25	0,5	0,6	0,8	1
0,01	0,990	0,991	0,992	0,992	0,994	0,997	0,998	1,000	1,000
0,05	0,951	0,955	0,961	0,964	0,972	0,988	0,992	0,998	1,000
0,1	*0,906*	*0,914*	*0,923*	*0,931*	*0,947*	0,977	0,985	0,996	1,000
0,15	*0,865*	*0,876*	*0,889*	*0,901*	*0,923*	0,967	0,979	0,995	1,000
0,25	*0,794*	*0,810*	*0,830*	*0,849*	*0,883*	*0,950*	0,969	0,993	1,000
0,5	**0,656**	**0,684**	**0,717**	**0,749**	*0,809*	*0,922*	0,952	0,989	1,000
0,6	**0,613**	**0,644**	**0,682**	**0,719**	*0,786*	*0,914*	*0,948*	0,989	1,000
0,8	**0,542**	**0,579**	**0,624**	**0,668**	**0,749**	*0,903*	*0,942*	0,988	1,000
1	**0,484**	**0,526**	**0,578**	**0,628**	**0,721**	*0,895*	*0,938*	0,987	1,000

Megjegyzés. A nyolcadik oszlop első eleme csak a kerekítés miatt lett 1. Itt és a 2–4. táblázatokban a számok eltérő jelölései a különböző nagyságrendek szemléltetését szolgálják.

Forrás: A két réteg jellemzőinek szimulálásával végzett saját számítás.

Szeretném hangsúlyozni, hogy ezek az eredmények abban az esetben állnak fenn, ha a nagyobb elemszámú réteg rendelkezik magasabb szórással (nem relatív értelemben!). Ez olyan esetekben fordulhat elő, ahol a rétegképző ismérv nincs sztochasztikus kapcsolatban az összes változóval, mert például többcélú adatgyűjtésről van szó, és a többféle megfigyelt változó alapján kellett a rétegzést egy kompromisszumos megoldással elvégezni. Ezzel azt is kimondtam, hogy nem ez az eset lesz a tipikus, hanem inkább a későbbiekben tárgyalt másik, legalábbis a saját gazdaságstatisztikai tapasztalataim ezt mutatják. Hiszen a gazdaság legtöbb területén a kisebb méretű egységekből van viszonylag sok, a nagyobb, meghatározó méretűekből kevés. Utób-

biak esetében pedig a méret (a termelés volumene, a földterület nagysága, a forgalom alakulása) sokkal szélesebb skálán mozog, és nagyságrendjében is nagyobb. Ezért a szórás is (abszolút értelemben) magasabb.

Az 1. táblázatból érdekes következtetések vonhatóak le. Először is egyértelmű szabályszerűséget fedezünk fel, hiszen minden sor balról jobbra növekszik, és minden oszlop lefelé csökken (kivéve az utolsót). Ugyanakkor az is szembetűnő, hogy az átlóban, illetve az átló felett található értékek viszonylag magasak, 1-hez közeliek. Ez azt jelenti, hogy amiként csökken a rétegek sokasági nagyságának egymáshoz viszonyított aránya, úgy csökken az esély arra, hogy a rétegek szórásainak arányától függetlenül túl nagyot nyerjünk a Neyman-féle optimális rétegzéssel. Összességében az 1. táblázatban bal alsó részében szereplő (szürkével jelölt) számok a legérdekesebbek. Vagyis akkor tudunk jelentősen javítani a becslésünk hatásosságán, ha a kialakított egyik réteg elemszáma legfeljebb duplája a másiknak, illetve a rétegek szórásai között nagyon jelentős eltérés van. Ez az információ bizonyos szempontból új irányt adhat az olyan típusú mintavételi tervek kialakításának, ahol mennyiségi ismérvek alapján alakítjuk ki a rétegeket, és ennek a két réteget elkülönítő határértéknek azt az optimális nagyságát keressük, amely legnagyobb mértékben csökkenti a végső becslés hibáját. Hiszen már nemcsak a belső szórás minimalizálása lesz a cél, hanem a rétegek nagyságának és a szórásuknak az aránya is fontos szerepet játszhat a határvonal meghúzásakor. Azt is láthatjuk az eredményekből, hogy a Neyman-féle optimális allokáció alkalmazásával szélsőséges esetben (lásd az 1. táblázat bal alsó sarka) akár kevesebb mint a felére is csökkenthetjük a becslési hiba nagyságát, az arányos elosztáshoz képest. Természetesen ez azt is jelenti, hogy az egyszerű véletlen mintavételhez képest még nagyobb a „nyereségünk".

Másik, véleményem szerint gyakorlati szempontból még figyelemre méltóbb tanulsága a számításoknak az, hogy amennyiben nem tudunk hasonló elemszámú rétegeket létrehozni ilyen kiinduló feltételek mellett, akkor sok értelme nincs is a szórásokra való optimalizálással bajlódni, hiszen mindössze pár százalékkal tudjuk csökkenteni a becslési hiba nagyságát az eltérő méretű rétegek mellett. Ez azért is fontos szempont, mert sok esetben nem kis erőfeszítést igényel az egyes rétegekhez tartozó szórások meghatározása a tényleges összeírást megelőzően, illetve sok esetben csak közelítő becsléseket tudunk készíteni. Amennyiben elemszámában nagyon eltér egymástól a kívánatos ismérv szerint kialakított két réteg, úgy megtakaríthatjuk a szórásokról való informálódással eltöltött időt, és helyette használhatjuk a minta rétegnagysággal arányos elosztását.

Ezzel azonban a kép még korántsem teljes, hiszen meg kell vizsgálnunk a másik típusú eseteket is, vagyis, amikor a nagyobb elemszámú réteghez kisebb szórás párosul. A számítások eredményeit a 2. táblázat tartalmazza. Véleményem szerint, ez a megközelítés leginkább a kapcsolati rétegzés esetét szemlélteti. Itt tehát azt feltételezzük, hogy a rétegképző ismérv az általunk vizsgált mennyiségi változóval szto-

chasztikus kapcsolatban áll, vagyis valamilyen szempont szerint kisebb, illetve nagyobb értékeket tartalmazó rétegeket kapunk. Ekkor nagy az esélye annak, hogy a nagyobb értékekhez tartozik a magasabb szórás (extrém esetektől eltekintve), ugyanakkor kisebb az elemszám (például ez a helyzet a mezőgazdaságban vagy a kiskereskedelemben).

2. táblázat

A Neyman-féle optimális és arányos allokáció relatív hatásossága

N_1/N_2	s_1/s_2								
	1	2	5	10	20	25	50	75	100
0,01	1,000	0,990	*0,867*	0,578	**0,248**	**0,172**	**0,038**	**0,003**	*
0,05	1,000	*0,958*	0,644	0,323	0,138	0,105	**0,045**	**0,028**	**0,019**
0,1	1,000	*0,932*	0,563	0,295	0,157	0,132	**0,085**	**0,070**	**0,063**
0,15	1,000	*0,914*	0,538	0,305	0,188	0,165	**0,123**	**0,110**	**0,104**
0,25	1,000	*0,895*	0,535	0,344	0,248	0,229	**0,193**	**0,181**	**0,175**
0,5	1,000	*0,883*	0,584	0,443	0,370	0,355	**0,327**	**0,317**	**0,312**
0,6	1,000	*0,884*	0,605	0,476	0,409	0,395	**0,369**	**0,360**	**0,355**
0,8	1,000	*0,889*	0,644	0,532	0,474	0,462	**0,439**	**0,431**	**0,427**
1	1,000	*0,895*	0,676	0,578	0,526	0,516	**0,495**	**0,488**	**0,484**

* A kisebb elemszámú rétegből (N_1) az optimalizáló feltétel szerint N_1-nél nagyobb mintát lenne szükséges kiválasztani, így ez a kombináció nem értelmezhető.

Forrás: A két réteg jellemzőinek szimulálásával végzett saját számítás.

A 2. táblázatban az oszlopokon belül megfigyelt szabályszerűség már nem monoton, hanem inkább parabolikus természetű. Lefelé haladva egy pontig csökken, majd növekszik az arányszám. A sorok esetében balról jobbra haladó folyamatos csökkenést találunk.

Amennyiben tehát $N_1<N_2$ és $s_1>s_2$, akkor a 2. táblázat adataiból néhány mondatba összefoglalva, a következő következtetéseket szűrhetjük le. Ha sikerül elérnünk egy olyan rétegzést, ahol az 1. réteg szórása legalább tízszerese a 2. rétegének, akkor – az N_1/N_2 (<1) aránytól gyakorlatilag függetlenül – biztosak lehetünk abban, hogy legalább 40 százalékkal pontosabb eredményt kapunk a Neyman-féle optimális allokációval az arányos elosztáshoz képest. Ha pedig a 2. réteg legalább négyszeres elemszámmal rendelkezik az 1. réteghez képest, és a szórása ugyanakkor csak legfeljebb 5 százaléka az 1. rétegének, akkor már garantáltan több mint 75 százalékos javulást tudunk elérni a szórásokra optimalizált elosztással. Ezek az információk hasznos támpontot adhatnak a mintavételi terv előkészítése során, bonyolultabb számítások elvégzése nélkül. Azt is láthatják, hogy az 1. táblázathoz képest itt jóval alacsonyabb

értékeket találunk. Ez pedig azt jelenti, hogy a kapcsolati rétegzés, illetve a $N_1<N_2$ és $s_1>s_2$ kiinduló feltételeknek megfelelő helyzet jóval magasabb mértékű hatásosságjavításra ad lehetőséget.

2.2. Neyman-féle optimális kontra egyenletes allokáció

A Neyman-féle optimális elosztás az egyenleteshez képest is minden esetben jobb eredményeket ad, kivéve, ha ugyanaz a szórás jellemző mindkét rétegre. Itt azonban találunk egy tágabban értelmezhető szabályszerűséget is, arra az esetre ahol a kétféle allokáció pontosan ugyanazt a standard hiba négyzetet, vagyis a mintanagyság azonos elosztását eredményezi: ha $N_1/N_2 = s_2/s_1$ áll fenn, vagyis a rétegek nagyságának aránya éppen a szórásarányok reciprokával megegyező. A 4. táblázatban is szerepel egy ilyen „kitüntetett" eset (2. sor, 2. oszlop). Az összefüggés a kétféle allokáció között valamivel egyértelműbb, mint a korábbi (1. és 2. táblázatban bemutatott) esetben, ezért kevesebb értékpár bemutatásával illusztrálom a problémát.

Amikor a reláció a rétegnagyság és a szórás mértéke között azonos irányú a két réteg között, akkor a „lejtő" a 3. táblázatban is megvan, viszont amíg az 1. táblázat esetében az irányt a bal alsó–jobb felső sarok jelölte ki, addig itt a bal felső–jobb alsó sarkok által kijelölt képzeletbeli nyíl mutatja a számok növekedésének irányát.

3. táblázat

A Neyman-féle optimális és egyenletes elosztás relatív hatásossága

N_1/N_2	s_1/s_2					
	0,01	0,1	0,25	0,5	0,75	1
0,01	**0,487**	**0,488**	**0,490**	**0,492**	**0,495**	**0,497**
0,1	**0,487**	**0,496**	**0,512**	**0,537**	**0,562**	**0,587**
0,25	**0,487**	**0,510**	**0,548**	*0,610*	*0,670*	*0,725*
0,5	**0,486**	**0,532**	*0,608*	*0,724*	*0,821*	*0,895*
0,75	**0,485**	**0,555**	*0,666*	*0,821*	*0,924*	0,979
1	**0,484**	**0,578**	*0,721*	*0,895*	0,979	1,000

Forrás: A két réteg jellemzőinek szimulálásával végzett saját számítás.

Ismét az ellenétesen alakuló méretszórás-párosítás hozza az értékek érdekesebb, kevésbé egyértelmű elhelyezkedését. (Lásd a 4. táblázatot.)

4. táblázat

A Neyman-féle optimális és egyenletes elosztás relatív hatásossága

N_1/N_2	s_1/s_2						
	1	10	25	50	75	100	200
0,01	**0,497**	**0,578**	*0,680*	*0,789*	*0,733*	*	*
0,1	**0,587**	1,000	*0,796*	**0,581**	**0,494**	**0,449**	**0,380**
0,25	*0,725*	*0,825*	*0,608*	**0,520**	**0,490**	**0,474**	**0,451**
0,5	*0,895*	*0,668*	**0,546**	**0,503**	**0,488**	**0,481**	**0,470**
0,75	0,979	*0,608*	**0,526**	**0,497**	**0,488**	**0,483**	**0,476**
1	1,000	**0,578**	**0,516**	**0,495**	**0,488**	**0,484**	**0,479**

* A kisebb elemszámú rétegből (N_1) az optimalizáló feltétel, és az egyenletes elosztás szerint is N_1-nél nagyobb mintát lenne szükséges kiválasztani, így ez a kombináció nem értelmezhető.

Forrás: A két réteg jellemzőinek szimulálásával végzett saját számítás.

A legnagyobb javulást akkor érjük el, ha minél nagyobb a szórások aránya. A rétegek elemszámainak arányától szinte függetlenül, körülbelül a felére csökkenthető a becslőfüggvény varianciája magas szórásarányok mellett. Kiugróan „jó" eredményt biztosít még az az eset is, ha szélsőségesen eltérő a rétegek nagysága.

A 3. és 4. táblázatban azt is leolvashatjuk az adatokból, hogy kevés a „nagy" szám, vagyis könnyebben érünk el javulást ebben az esetben, mint amikor az arányos elosztásról térünk át az optimálisra.

3. Illusztratív példák

Kutatásaim során többek között az állatállomány alakulásának statisztikai megfigyeléséhez, az európai uniós előírásoknak megfelelő pontossági követelményekhez igazodó hatékony mintavételi módszert keresem. Ennek kapcsán lehetőségem nyílt az említett problémát néhány valós példával is illusztrálni.

Magyarországon a mezőgazdasági összeírások rendszerében az adatszolgáltatói kört két nagy csoportra bontjuk: gazdasági szervezetek és egyéni gazdaságok.[3] Előbbieket teljes körűen megfigyelik, utóbbiak esetében merül fel a mintavételi módszerek alkalmazása (*Laczka* [2007]). A gazdasági szervezetek száma az elmúlt évtize-

[3] Egyéni gazdaság: olyan háztartás, melynek mezőgazdasági tevékenysége a tartott állatok számát, illetve a művelt földterület nagyságát illetően meghalad bizonyos fizikai küszöbértéke(ke)t.

dekben nyolcezer körül alakult, míg az egyéni gazdaságok, a folyamatos csökkenő tendencia ellenére, még a 2010. évi Általános Mezőgazdasági Összeírás adatai alapján is több mint félmilliós sokaságot alkotnak.

Kutatásaim során az utolsó elérhető, nagyobb lélegzetű összeírás a 2007. évi Gazdaságszerkezeti Összeírás (GSZÖ 2007) adatait használom fel a Központi Statisztikai Hivatal (KSH) engedélyével, amelyért ezúton is köszönetet mondok. A következőkben bemutatott számítások ezen alapulnak.

Példaként a sertésállományra alkalmazható lehetőségeket fogom bemutatni. A gyakorlatban az a kérdés merül fel, hogy miként lenne érdemes rétegezni egy (vagy több) változó szempontjából az egyéni gazdaságok sokaságát. Amennyiben rendelkezésünkre állnak előzetes információk (korábbi összeírásokból), akkor akár a vizsgált változó maga is jelentheti a rétegképzés alapján. Ebben a tanulmányban felvetett elméleti probléma gyakorlati megjelenése tehát az, hogy milyen állatállomány-nagyságnál érdemes meghúzni a réteghatárt, ha két rétegbe szeretnénk sorolni az állattartással foglalkozó egyéni gazdaságokat, kifejezetten a sertésszám alapján. A réteghatár nagyságát változtatva természetesen módosulni fog a „kisebb" és a „nagyobb" állománnyal rendelkező rétegek elemszámainak az aránya, illetve a rétegekben kialakuló szórások értéke is változik.

A korábbi elméleti felvetéshez hasonlóan azt vizsgáltam meg, hogy különböző réteghatárok mellett hogyan alakul a különböző kombinációban megvalósuló N_1/N_2 és s_1/s_2[4] arányoknak megfelelően a Neyman-féle optimális rétegzés hatékonysága az egyenletes, illetve az arányos allokációkhoz képest. (Az 5. táblázat első oszlopában a réteghatár azt mutatja, hogy mekkora az a sertésállomány-nagyság, amely fölött már a „nagyok" csoportjába tartozik egy gazdaság.) Az összehasonlításokat 5 százalékos mintavételi arány mellett végeztem el. A réteghatárok kiválasztása önkényes, de a nagyságrendek változása jól látszik az adatokon. 140 sertésszám fölé menni már azért nem volt értelme, mert a „nagyok" alkotta réteg elemszáma már túlságosan alacsony lett volna.

Az 5. táblázat utolsó oszlopában a szóráshányados-mutatót (H – a vegyes kapcsolat szorosságát mérő mutatószám, a külső és a teljes szórás hányadosa) azért szerepeltettem, hogy érzékelhető legyen, milyen erősségű sztochasztikus kapcsolat áll fenn a csoportosítás (2 réteg képzése) és a vizsgált sertésállomány között. Látható, hogy 100-120 sertést tartó gazdaságok képezik azt a határvonalat, amely mentén a kialakult 2 csoport a legszorosabb sztochasztikus kapcsolatot mutatja a sertésállománnyal. ($1–H^2$ pedig azt is megmutatja, hogy milyen az arányos elosztással képzett rétegzett minta hatásossága az egyszerű véletlen kiválasztáshoz képest. Vagyis közvetve az is leolvasható, hogy az egyszerű véletlenhez képest a 100-120 sertést jelentő réteghatár adná a leghatékonyabb lehetőséget arányos elosztással.)

[4] Az 1-es index rendre a „nagyobbak" rétegét, a 2-es index pedig rendre a „kisebbek" rétegét jelenti.

5. táblázat

A sokasági adatok alakulása különböző réteghatárok mellett

Sertés-állomány (db)	N_1	N_2	N_1/N_2	s_1	s_2	s_1/s_2	Szóráshányados
1	33 620	46 217	0,7274	14,73	0,49	30,2423	0,536
2	18 250	61 587	0,2963	19,30	0,82	23,6753	0,613
3	13 028	66 809	0,1950	22,28	0,98	22,6473	0,652
4	10 010	69 827	0,1434	24,86	1,13	22,0909	0,679
5	8 239	71 598	0,1151	26,92	1,28	21,0071	0,697
6	6 998	72 839	0,0961	28,77	1,42	20,3201	0,710
9	4 803	75 034	0,0640	33,50	1,79	18,6895	0,736
10	4 104	75 733	0,0542	35,68	1,96	18,1878	0,745
20	1 505	78 332	0,0192	52,42	3,04	17,2423	0,790
30	768	79 069	0,0097	66,16	3,75	17,6502	0,814
40	474	79 363	0,0060	76,89	4,27	18,0198	0,827
50	339	79 498	0,0043	84,28	4,62	18,2433	0,834
60	252	79 585	0,0032	90,64	4,94	18,3384	0,839
70	189	79 648	0,0024	96,10	5,25	18,2976	0,843
80	149	79 688	0,0019	99,81	5,51	18,1136	0,845
90	122	79 715	0,0015	101,51	5,71	17,7663	0,847
100	94	79 743	0,0012	101,57	5,98	16,9767	0,849
120	79	79 758	0,0010	99,90	6,17	16,1953	0,849
140	66	79 771	0,0008	97,97	6,38	15,3462	0,846

Forrás: A KSH GSZÖ 2007 adatbázisa alapján saját számítás.

A tanulmány elméleti fejtegetéseihez kapcsolódva a következőkben bemutatom, hogy milyen konkrét eredmények születtek a gyakorlatban a sertésállomány példáján. A táblázat formáján változtattam, hiszen a konkrét számszerű megvalósulások nem töltenek meg egy mátrixot.

A 6. táblázat összevontan tartalmazza a kapott eredményeket. Az elméleti vizsgálatoknak megfelelően alakultak az értékek. Három adatot emeltem ki ebben a táblázatban. Az első (0,152) azt mutatja meg, hogy hol jelentkezett a legnagyobb hatékonyságjavulás a Neyman-féle optimális allokációval az arányoshoz viszonyítva. Eszerint kilenc sertésnél kell meghúzni a határt, hogy az optimális elosztás a legtöbb hozadékot nyújtsa. Egyben természetesen a Neyman-féle elosztás optimuma is ebben a pontban található. Az utolsó oszlopban a 0,498 jelöli azt a sort, ahol az optimális allokáció a legnagyobb javulást eredményezi az egyenleteshez képest. Az értékek folyamatosan csökkennek, azt sugallva, hogy egy magasabb réteghatár még alacso-

nyabb értéket hozna. Viszont mivel N_1 már az utolsó sorban is nagyon kicsi, nem lenne értelme tovább folytatni a sorozatot.

6. táblázat

A különböző allokációs eljárások egymáshoz viszonyított hatékonysága

Sertésállomány (db)	N_1/N_2	s_1/s_2	$Var(\bar{y}_{No})/Var(\bar{y}_A)$	$Var(\bar{y}_{No})/Var(\bar{y}_E)$
1	0,727	30,242	0,431	0,517
2	0,296	23,675	0,260	0,596
3	0,195	22,647	0,203	0,666
4	0,143	22,091	0,173	0,739
5	0,115	21,007	0,160	0,814
6	0,096	20,320	0,153	0,877
9	0,064	18,690	0,152	0,989
10	0,054	18,188	0,155	0,999
20	0,019	17,242	0,220	0,762
30	0,010	17,650	0,303	0,630
40	0,006	18,020	0,384	0,575
50	0,004	18,243	0,451	0,550
60	0,003	18,338	0,516	0,534
70	0,002	18,298	0,585	0,522
80	0,002	18,114	0,643	0,514
90	0,002	17,766	0,695	0,509
100	0,001	16,977	0,764	0,503
120	0,001	16,195	0,809	0,500
140	0,001	15,346	0,850	0,498

Forrás: A KSH GSZÖ 2007 adatbázisa alapján saját számítás.

A harmadik kiemelt érték (0,999) is figyelmet érdemel. Az a sajátos helyzet állt ugyanis elő, amit a korábbiakban már kiemeltem, miszerint $N_1/N_2 = s_2/s_1$, és ilyen összefüggés mellett az egyenletes elosztás a Neyman-féle optimális allokációval gyakorlatilag megegyező eredményt (hatásosságot) biztosít. n_1=n_2=1996, $Var(\bar{y}_E) = 0,0025$ az egyenletes esetben, n_1=1981, n_2=2010, $Var(\bar{y}_{No}) = 0,0025$ optimális elosztás mellett.

Ez pedig azt is jelenti a gyakorlati felhasználó számára, hogy felesleges a szórásokra történő optimalizálással „bajlódni”, mert az egyenletes elosztás is javít annyit a mintavételi hibán. Ráadásul ez igen közel esik az optimális elosztás szerinti legjobb megoldáshoz. Így a mintaelemszám fele-fele arányban történő elosztásával az ará-

nyos rétegzéshez képest, a becslőfüggvény varianciája kerekítve a 15 százalékára esik vissza, ha az 1-10 sertést tartók lesznek az egyik és a 10 feletti sertést tartók a másik réteg.

Felhívnám a figyelmet a gyakorlati eredmények még egy érdekességére. Ha a 6. táblázatban végignézzük a két utolsó oszlopot, akkor azt is észrevehetjük, hogy a sertésállomány esetében a különböző rétegzések mellett az egyenletes elosztás több esetben áll közelebb a Neyman-féle optimálishoz, mint az arányos. Illetve a minta rétegnagysággal arányos felosztásával optimális esetben sem lehet jobb eredményt biztosítani, mint a másik két allokációs módszerrel.

4. Összegzés

A rétegzett mintavétel, véleményem szerint, egy összefoglaló elméleti kategória, olyan esetekre, ahol a sokaságot még a mintavételt megelőzően diszjunkt rétegekbe soroljuk. A rétegeket valamilyen ok vagy szempont miatt megválasztott ismérv szerinti csoportosítás alakítja ki. Ilyen általános feltételek mellett, azonos eljárással, azonos becslési módszertannal hajtható végre a mintavételes adatfelvétel folyamata. A rétegképző ismérv szempontjából fontosnak tartom elkülöníteni, és külön nevesíteni azt az esetet, amikor a becsülni kívánt változóval sztochasztikus kapcsolatban álló ismérvet választunk a rétegzéshez, és ezáltal a sokasághoz képest homogénebb részsokaságokat, rétegeket kapunk. Ez a becslés standard hibáját minden esetben csökkenti az egyszerű véletlen kiválasztáshoz képest. Az elért homogenitás szintjétől függően lényeges javulás is várható. Ezt az esetet kapcsolati rétegzésnek nevezem. A többi, technikájában, módszerében az előzővel megegyező rétegzési fajtát összefoglalóan technikai rétegzésként kezelem. Ezekben az a közös, hogy a rétegzésnek nem célja a homogén csoportok kialakítása, hanem valami más motiválja az eljárás használatát (összeírási költségek minimalizálása, más összeírási módszer, más alapinformációk, reprezentáció szintje stb.).

A rétegek között a különböző allokációs technikák hatásosságát hasonlítottam össze az általános elvek bemutatásán túl, konkrétan számszerűsíthető eseteket figyelembe véve, gyakorlati példával is illusztrálva. A Neyman-féle optimális elosztást hasonlítom az arányos, illetve az egyenletes elosztáshoz, két réteget feltételezve. Gyakorlati szempontból ezt azért tartom jelentős nézőpontnak, mert az összehasonlítás eredményeit tartalmazó táblák, illetve a levont következtetések támpontot, ötletet, megoldási útvonalat adhatnak a mintavételi terv kialakításának fázisában a konkrét számokon keresztül. Ezen túl pedig a számszerűsítés mint elgondolás további irányt jelenthet az összehasonlítások elvégzéséhez.

A bemutatott eredményekből az is kiviláglik, hogy a rétegzés, illetve a mintaelemszám rétegek közötti felosztása a kiindulási feltételek függvényében rendkívül szélsőséges eredményeket is hozhat.

Irodalom

AY, J. [1976]: *A mintavételes állatösszeírások módszertani kérdései*. Kandidátusi értekezés. Marx Károly Közgazdaságtudományi Egyetem. Budapest.

COCHRAN, W. G. [1977]: *Sampling Techniques.* John Wiley and Sons, Inc. New York.

ÉLTETŐ Ö. [1982]: Mintavételi eljárások. In: *Éltető Ö. – Meszéna Gy. – Ziermann M.*: *Sztochasztikus módszerek és modellek*. Közgazdasági és Jogi Könyvkiadó. Budapest.

FRALLER G. [2011]: Szemelvények a mintavételi rétegzés területéről. *Statisztikai Szemle*. 89. évf. 4. sz. 357–378. old.

GALAMBOSNÉ TISZBERGER M. [2009]: Kisgazdaságok a magyar mezőgazdaságban. In: *Buday-Sántha A.* (szerk.): *Évkönyv 2009*. Pécsi Tudományegyetem. Pécs. 60–69. old.

HAJDU O. – PINTÉR J. – RAPPAI G. – RÉDEY K. [1994]: *Statisztika I.* Janus Pannonius Tudományegyetem. Pécs.

HUNYADI, L. [1991]: A Two-Phase Sampling Design. *Pure Mathematics and Applications*. Serial C. Vol 2. No. 1. pp. 95–111.

HUNYADI L. [2001a]: *A mintavétel alapjai.* SZÁMALK Kiadó. Budapest.

HUNYADI L. [2001b]: *Statisztikai következtetéselmélet közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

HUNYADI L. – VITA L. [2002]: *Statisztika közgazdászoknak.* Központi Statisztikai Hivatal. Budapest.

KAPITÁNY B. [2010]: Mintavételi módszerek ritka populációk esetén. *Statisztikai Szemle*. 88. évf. 7–8. sz. 739–754. old.

KISH, L. [1995]: *Survey Sampling*. John Wiley and Sons, Inc. New York.

LACZKA É. [2007]: A magyar mezőgazdaság az EU-csatlakozás körüli években, 2000–2005. *Statisztikai Szemle*. 85. évf. 1. sz. 5–20. old.

MARTON Á. [1991]: *A reprezentatív felvételek megbízhatósága*. Központi Statisztikai Hivatal. Budapest.

PINTÉR J. – RAPPAI G. [2001]: A mintavételi tervek készítésének néhány gyakorlati megfontolása. *Marketing és Menedzsment*. 35. évf. 4. sz. 4–10. old.

PINTÉR J. – RAPPAI G. (szerk.) [2007]: *Statisztika.* Pécsi Tudományegyetem. Pécs.

Summary

Sampling techniques have great relevance worldwide. They are applied in every field of life. Within the topic of sampling, stratification is a very popular method. After dealing with the nature of stratified sampling, in the first part of the study, the author examines the existence of possible

subgroups within stratification. Besides creating more homogeneous strata within the population, there are several other reasons behind the selection of the methodology (technique) of stratification.

The methods of sample size allocation within strata are analysed in the second half of the article. Efficiency of different allocation techniques is compared, assuming two strata. The results of the comparison are probably not surprising, but interesting conclusions can be drawn. Depending on the correlation between the size and the standard deviation of the strata, different results are reached. The essence of the message is that there are several cases, when the *Neyman* optimal stratification does not improve notably the efficiency of the estimation. Therefore, besides certain basic conditions, we might get information at the beginning of the sampling procedure, if it worth dealing with the standard deviation of the strata. The calculations might give ideas to the practical application for the professionals. In the final section of the article the results are illustrated by some examples from agricultural statistics.