A mintavételi elrendezésnek megfelelő variancia-becslés

Készítette:

Németh Renáta

alkalmazott matematikus szak

Témavezető:

Rudas Tamás

ELTE Szociológiai Intézet, Statisztika Tanszék

Eötvös Loránd Tudományegyetem

Természettudományi Kar

Budapest, 2002.

Tartalomjegyzék

1. Előszó

A mintavétellel végzett felmérés napjainkra igen elterjedt információszerzési eszközzé vált. Piackutatások, közvéleménykutatások, kormányzati megrendelésre vagy tudományos céllal készült felmérések tízezreit végzik évente szerte a világon.

A matematikai statisztika sztenderd feltevése a megfigyelések függetlensége és azonos eloszlása. A felmérések mintájának esetén ugyanakkor ez a feltevés a legritkább esetben teljesül. Ennek legkézenfekvőbb oka a populációk véges elemszáma. Véges elemszám esetén visszatevés nélkül végezve a mintavételt a megfigyeléseink nyilván nem függetlenek. A feltevés nem teljesülésének másik oka a felmérések mintavételi elrendezésének (*sampling design*) nagyfokú összetettsége. A gyakorlatban szinte soha nem alkalmazzák a független, azonos eloszlású megfigyeléseket adó visszatevéses egyszerű véletlen mintavételt.

A mintavételi elrendezések összetettségének szükségessége igen sok tényezőre vezethető vissza. Az elrendezés kialakításakor figyelembe kell venni a célpopuláció jellemzőit, pl. földrajzi elhelyezkedését. Figyelembe kell venni a célváltozókat, pl. egy nemzetiségi összehasonlító kutatás esetén a kis létszámú kisebbség tagjait nagyobb valószínűséggel érdemes a mintába venni. Lényegesek az adminisztratív korlátok, pl. nem végezhető egyszerű véletlen mintavétel, ha nem áll rendelkezésre nyilvántartás a célpopulációról. Meghatározó szempont a mintavétel költségvonzata, így pl. olcsóbbak a földrajzilag kevéssé szórt mintát eredményező módszerek. A költségeket ugyanakkor érdemes hatásossági szempontokkal együtt vizsgálni, az olcsóbb eszköz akkor preferálható egy drágábbal szemben, ha az alkalmazásával kapott eredmények megbízhatósága nem csökken jelentősen. A mintavételi elrendezés kialakítása ezen szempontok együttes figyelembevételével történik.

A felmérésből származó becslésekkel szemben támasztott alapvető követelmény azok megbízhatóságának mérhetősége. A megbízhatóság leggyakrabban használt mértéke a becslés varianciája. Általában ez a variancia ismeretlen, és csak a mintából becsülhető. A variancia becslésének problematikája képezi ezen dolgozat tárgyát.

A felmérésből számolt becslés varianciája a becslés módján kívül a mintavételi elrendezésnek is függvénye. A felmérések kiértékelésénél jelenleg leggyakrabban alkalmazott módszerek, illetve az értékelésnél használt szoftverek (így a magyarországi közvélemény-

kutatók széles körében használt SPSS is) azzal a fent említett feltevéssel élnek, miszerint a mintaelemek független módon és azonos valószínűséggel kerültek a mintába. Ezeknek a sztenderd eljárásoknak a használata téves varancia-becsléshez, és így érvénytelen következtetésekhez vezethet. A dolgozat célja a mintavételi elrendezésnek megfelelő, ahhoz illeszkedő (*design-adjusted*) variancia-becslési módszerek áttekintése.

Az utóbbi ötven évben intenzív tudományos tevékenység folyt ezen a területen. A XIX. század végétől már rendszeresen végzett felmérések elméleti hátterét adó felmérés-statisztika kialakulása Neyman 1934-es, nagy hatású tanulmányához köthető. Egyéb megfontolások mellett ebben a cikkben került sor elsőként a valószínűségi mintavételnek az abban az időben széles körben alkalmazott kvótás mintavétellel (*purposive selection*) szembeni előnyének az alátámasztására, illetve annak a felismerésnek a bizonyítására, hogy a mintából kapott becslések érvényességének nem előfeltétele az azonos kiválasztási valószínűségek megléte.

Az alábbi gyakran idézett művek egy-egy mérföldkövet jelentettek a felmérés-statisztika történetében, dolgozatunk megírásában az utolsó két munkára közvetlenül is támaszkodtunk:

- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society,* 97, 558-625.
- Yates, F. (1949). *Sampling Methods for Censuses and Surveys.* London: Griffin.
- Deming, W. E. (1950). *Some Theory of Sampling.* New York: Dover.
- Hansen, M. H., Hurwitz, W. N., Madow, W. G. (1953). *Sampling Survey Methods and Theory* I., II. New York: Wiley
- Cohran, W. G. (1977). *Sampling Techniques.* New York: Wiley
- Wolter, K. M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag
- Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag

A matematikai eredmények számítástechnikai megvalósításaként több statisztikai szoftverben található már mintavételi elrendezésnek megfelelő elemzésre alkalmas modul. A legelterjedtebbek közé sorolható CLAN, GES, SAS, Stata, SUDAAN és WesVar PC újabb verziói képesek a dolgozatban sorravett módszerek közül néhány alkalmazására, de az általuk lefedett mintavételi elrendezések és becslések köre korlátozott (Carlson, 1998).

A dolgozat tehát valamely θ paraméter $\hat{\theta}$ becslésének $V\left(\hat{\theta}\right)$ varianciájának, illetve a variancia $\hat{V}\left(\hat{\theta}\right)$ becslésének számítási módszereit veszi számba, különböző mintavételi elrendezések esetén. A 3.1. fejezetben azt az esetet tárgyaljuk, amikor $\hat{\theta}$ π-becslés. Két torzítatlan becslést adunk $V\left(\hat{\theta}\right)$-ra, a Horvitz-Thomson és a Yates-Grundy becslést. A mintavételi praxisban gyakran használt néhány elrendezésre expliciten is megadjuk ezeket a becsléseket a 3.3. és a 3.4. fejezetekben, a különböző elrendezések hatásosságának összevetésére törekedve. Ugyan (3.18) szerint a π-becslés alkalmazása mellett általában egyik elrendezés sem jobb a másiknál, a hatásosság összevetése bizonyos előfeltevések mellet mégis informatív lehet. Abban az esetben, amikor $\hat{\theta}$ más π-becsléseknek nemlineáris függvénye, $\hat{\theta}$ eloszlása matematikaliag nehezen kezelhető, így $V\left(\hat{\theta}\right)$ legtöbbször nem adható meg egzakt módon. Az 4.1. fejezetben a nemlineáris függvény lineáris approximációja segítségével adunk közelítő becslést $V\left(\hat{\theta}\right)$-ra. Az 4.2. fejezet a komplex mintavételi elrendezések, ill. komplex becslések esetén gyakran alkalmazott ismételgető technikák néhány változatának (random csoportok, jackknife, bootstrap) számbavétele. A korábbi fejezetek eredményeinek általánosabb felhasználhatóságára mutat néhány példát a 5. fejezet, míg az utolsó rész hazai felmérés adatainak felhasználásával mutatja be a mintavételi elrendezésnek megfelelő variancia-becslés alkalmazását.

1.1. Megjegyzés. *1. A mintavétellel kapcsolatos angol nyelvű szakkifejezések magyarra fordításáról legtöbb esetben még nem jött létre konszenzus, ezért ezek első előfordulásakor megadtuk az eredeti angol terminust is.*

2. A dolgozat egyfajta áttekintést ad, tematikája igen széles kört ölel fel, és inkább az alkalmazásra koncentrál. Emiatt a tételeket legtöbbször bizonyítás nélkül közöljük. Kivételt képeznek a fontos alaptételek tipikus, több esetre adaptálható bizonyításai.

2. Bevezetés

Legyen az U N elemű *véges populáció*

$$U = \{1, \ldots, k, \ldots, N\} \stackrel{def}{=} \{u_1, \ldots, u_k, \ldots u_N\}, \; N < \infty,$$

az egyszerűség kedvéért a továbbiakban az u_k elemet az indexével, k-val azonosítjuk.

Legyen az s minta a populáció tetszőleges részhalmaza: $s \subseteq U$. $n(s)$ jelöli a mintanagyságot.

2.1. Megjegyzés. *A dolgozatban visszatevés nélküli mintavétellel foglalkozunk. Visszatevéses mintavétel esetén a mintát nem halmazként, hanem rendezett n-esként definiálnánk.*

Legyen $\mathbf{S}$ az U hatványhalmaza, $\mathbf{S} = Pow\,(U)\,, |\mathbf{S}| = 2^N$. Jelölje $\mathbf{A}$ az $\mathbf{S}$ által generált σ-algebrát, $\mathbf{A} = Pow\,(\mathbf{S})\,$. Jelölje $p(\cdot)$ az ($\mathbf{S}$,$\mathbf{A}$) mérhető téren értelmezett valószínűségi mértéket. Rögzített $p(\cdot)$ mellett $p(s)$ adja meg az s kiválasztásának valószínűségét: az s mintát tekinthetjük az S halmazértékű valószínűségi változó kimenetének, az S eloszlását $p(\cdot)$ adja meg:

$$P(S = s) \stackrel{def}{=} p(s), \; s \in \mathbf{S}.$$

Mivel $p(\cdot)$ valószínűségi mérték $(\mathbf{S}, \mathbf{A})$-n, $p(s) \geq 0, \forall s \in \mathbf{S}$,illetve $\sum_{s \in \mathbf{S}} p(s) = 1$. A $p(\cdot)$ függvényt nevezzük *mintavételi elrendezésnek.*

Rögzített $p(\cdot)$ mellett definiáljuk az alábbi *mintaelem indikátorokat*:

$$k \in U : I_k \stackrel{def}{=} \begin{cases} 1, \text{ ha } k \in S \\ 0 \text{ egyébként} \end{cases}$$

A mintaelem indikátorok segítségével határozhatók meg az *elsőrendű kiválasztási valószínűségek*:

$$\pi_k \stackrel{def}{=} \sum_{s \ni k} p(s) = P(I_k = 1).$$

A továbbiakban feltesszük, hogy $\pi_k > 0, \forall k \in U$. A *másodrendű kiválasztási valószínűségek:*

$$\pi_{kl} \stackrel{def}{=} \sum_{s \ni k \& l} p(s) = P(I_k = 1 \wedge I_l = 1).$$

2.2. Állítás. *Tetszőleges $p(\cdot)$ mintavételi elrendezés esetén a mintaelem indikátorok várható értékére és kovarianciájára:*

$$E(I_k) = \pi_k,$$

$$C(I_k, I_l) = \pi_{kl} - \pi_k \pi_l, \forall k, l \in U.$$

2.3. Definíció. *A kovarianciákra külön jelölést vezetünk be:*

$$\triangle_{kl} \stackrel{def}{=} C(I_k, I_l).$$

Legyen $y : U \to \mathbf{R}$ az U elemeinek valamely jellemzője, jelölje y_k $(k \in U)$ a jellemző értékét a k. elem esetében. $\theta = \theta(y_1, \ldots, y_N)$ a populációs *paraméter*. Becslése: $\hat{\theta} = \hat{\theta}(S, y_1, \ldots, y_N)$. A becslés várható értéke és varianciája rögzített $p(\cdot)$ mintavételi elrendezés mellett:

$$E(\hat{\theta}) = \sum_{s \in \mathbf{S}} p(s) \hat{\theta}(s),$$

$$V(\hat{\theta}) = \sum_{s \in \mathbf{S}} p(s) \left(\hat{\theta}(s) - E(\hat{\theta}) \right)^2.$$

3. Az összeg becslése

3.1. A π-becslés

3.1. Definíció. *A populációs összeg:* $t \stackrel{def}{=} \sum_U y_k$.

Ahol, mint a továbbiakban is, az egyszerűség kedvéért a $\sum_U y_k$ jelölést alkalmazzuk a $\sum_{k \in U} y_k$ kifejezésre.

3.2. Definíció. *Az összeg $\pi-$becslése (Horvitz, Thomson, 1952):*

$$\hat{t}_\pi \stackrel{def}{=} \sum_s \frac{y_k}{\pi_k}.$$

Megjegyzés. *Korábban feltettük, hogy $\pi_k > 0, \forall k \in U$.*

Jelölés. $\check{y}_k \stackrel{def}{=} \frac{y_k}{\pi_k}$, *elnevezés: az y π-kiterjesztettje.*

3.3. Tétel. *A $\pi-$becslés az összeg torzítatlan becslése, varianciája:*

$$V(\hat{t}_\pi) = \sum \sum_U \triangle_{kl} \check{y}_k \check{y}_l,$$

ahol $\triangle_{kl}$ *a (2.3)-ben definiált. Ha* $\pi_{kl} > 0$ *minden* $k,l \in U$*-ra, a* $V(\hat{t}_\pi)$*-ra a következő torzítatlan becslés adható (Horvitz-Thomson becslés: Horvitz, Thomson, 1952):*

$$\hat{V}(\hat{t}_\pi) \stackrel{def}{=} \sum\sum_s \check{\triangle}_{kl}\check{y}_k\check{y}_l,$$

ahol $\check{\triangle}_{kl} = \frac{\triangle_{kl}}{\pi_{kl}}$.

Bizonyítás. A torzítatlanság könnyen belátható a mintaelem-indikátorok bevezetésével:

$$E\left(\hat{t}_\pi\right) = E\left(\sum_s \frac{y_k}{\pi_k}\right) = E\left(\sum_U I_k \frac{y_k}{\pi_k}\right) = E\left(\sum_U I_k \frac{y_k}{E(I_k)}\right) = t.$$

A becslés varianciája hasonlóan adódik:

$$\begin{aligned} V(\hat{t}_\pi) &= V(\textstyle\sum_U I_k\check{y}_k) = \sum_U V\left(I_k\right)\check{y}_k^2 + \sum_{k\neq l}\sum_U C(I_k,I_l)\check{y}_k\check{y}_l \\ &= \textstyle\sum_U \triangle_{kk}\check{y}_k^2 + \sum_{k\neq l}\sum_U \triangle_{kl}\check{y}_k\check{y}_l = \sum\sum_U \triangle_{kl}\check{y}_k\check{y}_l. \end{aligned}$$

Végül $\hat{V}(\hat{t}_\pi)$ torzítatlansága a $\pi_{kl} > 0, \forall k,l \in U$ feltétel mellett:

$$E\left(\hat{V}(\hat{t}_\pi)\right) = E\left(\sum\sum_U I_kI_l\check{\triangle}_{kl}\check{y}_k\check{y}_l\right) = \sum\sum_U \triangle_{kl}\check{y}_k\check{y}_l = V(\hat{t}_\pi),$$

felhasználva, hogy $E\left(I_kI_l\check{\triangle}_{kl}\right) = \pi_{kl}\check{\triangle}_{kl} = \triangle_{kl}$.

3.4. Definíció. *Azt mondjuk, hogy* $p(\cdot)$ *rögzített mintanagyságot adó a mintavételi elrendezés, ha* $\forall s : p(s) > 0 \to |s| = n$ *fennáll.*

Rögzített mintanagyságot adó elrendezés esetén a $\pi-$becslés varianciájának az alábbi tétel alapján más torzítatlan becslése is ismert.

3.5. Tétel. *Ha* $p(\cdot)$ *rögzített mintanagyságot adó elrendezés, akkor a (3.3)-ban definiált* $V(\hat{t}_\pi)$ *felírható a következő alakban:*

$$V(\hat{t}_\pi) = -\tfrac{1}{2}\sum\sum_U \triangle_{kl}\left(\check{y}_k - \check{y}_l\right)^2.$$

A $\pi_{kl} > 0, \forall k,l \in U$ *feltétel mellett* $V(\hat{t}_\pi)$*-re a következő torzítatlan becslés adható (Yates-Grundy becslés: Yates, Grundy (1953)):*

$$\hat{V}(\hat{t}_\pi) \stackrel{def}{=} -\tfrac{1}{2}\sum\sum_s \check{\triangle}_{kl}\left(\check{y}_k - \check{y}_l\right)^2.$$

Bizonyítás. A $V(\hat{t}_\pi)$ két felírásának ekvivalenciája könnyen belátható a négyzetes tag felbontásával:

$$V(\hat{t}_\pi) = -\tfrac{1}{2}\sum\sum_U \triangle_{kl}\left(\check{y}_k - \check{y}_l\right)^2 = \sum\sum_U \triangle_{kl}\check{y}_k\check{y}_l - \sum\sum_U \triangle_{kl}\check{y}_k^2,$$

ahol a második tag kiesik, mert

$$\sum\sum_U \triangle_{kl}\check{y}_k^2 = \sum_{k\in U}\check{y}_k^2 \sum_{l\in U}\triangle_{kl},$$

és a rögzített mintanagyság kihasználásával

$$\sum_{l\in U}\triangle_{kl} = \sum_{l\in U}\pi_{kl} - \sum_{l\in U}\pi_k\pi_l = n\pi_k - n\pi_k = 0.$$

A $\hat{V}(\hat{t}_\pi)$ torzítatlan volta könnyen adódik a

$$\hat{V}(\hat{t}_\pi) = -\tfrac{1}{2}\sum\sum_U I_k I_l \check{\triangle}_{kl}\left(\check{y}_k - \check{y}_l\right)^2$$

átalakítással, és a már használt $E\left(I_k I_l \check{\triangle}_{kl}\right) = \pi_{kl}\check{\triangle}_{kl} = \triangle_{kl}$ azonosság alkalmazásával.

3.6. Megjegyzés. *Mind a Horvitz-Thomson, mind a Yates-Grundy becsléshez található olyan mintavételi elrendezés és olyan $\{y_1, \ldots, y_N\}$ populáció, melyek esetén negatív értéket adnak a variancia becslésére. A Yates-Grundy becslés nemnegatív voltának könnyen belátható elégséges feltétele: $\triangle_{kl} = \pi_{kl} - \pi_k\pi_l < 0$, $\forall k, l \in U$.*

3.1.1. A π-becsléssel kapcsolatos optimalitási tételek

Az alábbiakban a π-becslés optimalitásával kapcsolatos néhány fontos tételt közlünk. Jelentőségüket az adja, hogy segítségükkel a π-becslés, ill. varianciájának Horvitz-Thomson becslése könnyebben elhelyezhető más becslések között; a bizonyítások bemutatása a közvetlenebb interpretálhatóságot szolgálja. A fejezet utolsó tételének (3.19) következménye a dolgozat egészén végigvonuló azon meggondolást formalizálja, miszerint nem létezik minden helyzetben legjobbnak ítélhető elrendezés, és mindig csak az adott előismeretek függvényében dönthetünk egyik vagy másik elrendezés mellett.

Az alábbiakban $y : U \to \mathbf{R}$ jelöli az U elemeinek valamely jellemzőjét, Y az összes ilyen jellemző osztályát, $t(y)$ a becsülendő populációs összeget, $\hat{t}(s, y)$ ennek az s mintából számolt becslését, $\hat{t}_\pi(s, y)$ a π-becslést.

3.7. Definíció. *Jelölje D a t összeg torzítatlan becsléseinek osztályát, $p(\cdot)$ legyen rögzített elrendezés. Azt mondjuk, hogy $\hat{t} \in D$ jobb becslés, mint $\hat{t}' \in D$, ha $\forall y : V_p\left(\hat{t}, y\right) \leq V_p\left(\hat{t}', y\right)$, és legalább egy $y \in Y$-ra szigorú egyenlőtlenség teljesül. A becslések D osztályán belül a $\hat{t}$ megengedett becslés, ha az osztály egyik tagja sem jobb nála.*

3.8. Tétel. *A $\hat{t}_\pi$ a $\forall k \in U : \pi_k > 0$ feltételt teljesítő bármely $p(\cdot)$ mellett megengedett D-ben (Godambe, Joshi, 1965).*

Bizonyítás. *Indirekt módon tegyük fel, hogy van jobb $\hat{t} \in D$. Erre:*

$\hat{t}(s,y) = \hat{t}_\pi(s,y) + \left(\hat{t}(s,y) - \hat{t}_\pi(s,y\right) = \hat{t}_\pi(s,y) + h(s,y)$,

a becslések torzítatlanságából

(i) $\sum_{s \in S} h(s,y)p(s) = 0$,

és $\sum_{s \in S} \hat{t}^2(s,y)p(s) \leq \sum_{s \in S} \hat{t}_\pi^2(s,y)p(s)$, $\forall y \in Y$, *amiből*

(ii) $\sum_{s \in S} h^2(s,y)p(s) \leq -2 \sum_{s \in S} h(s,y)\hat{t}_\pi(s,y)p(s)$, $\forall y \in Y$.

Legyen $Y_k \subset Y$ olyan, hogy minden $y \in Y_k$ függvénynek éppen k nemnulla értéke van.

Lemma. *Ha $\forall s \in S, y \in Y_k : h(s,y)p(s) = 0$, akkor $\forall s \in S, y \in Y_{k+1} : h(s,y)p(s) = 0$.*

A lemma bizonyításához legyen $y' \in Y_{k+1,}$ erre (i) és (ii) miatt teljesül

(iii) $\sum_{i=0}^{k+1} \sum_{s \in S_i} h(s,y')p(s) = 0$, *és*

(iv) $\sum_{i=0}^{k+1} \sum_{s \in S_i} h^2(s,y')p(s) \leq -2 \sum_{i=0}^{k+1} \sum_{s \in S_i} h(s,y^{'})\hat{t}_\pi(s,y')p(s)$,

ahol $S_i \subset S$ olyan, hogy minden $s \in S_i$ mintának éppen i nemnulla y' értéke van. $\forall s \in S, \forall y \in Y_k : h(s,y)p(s) = 0$ teljesüléséből következik, hogy $\forall s \in S_i$, $i = 1, \ldots, k$: $h(s,y')p(s) = 0$. Ebből (iii) és (iv) felhasználásával

(v) $\sum_{s \in S_{k+1}} h(s,y')p(s) = 0$, *és*

(vi) $\sum_{s \in S_{k+1}} h^2(s,y')p(s) \leq -2 \sum_{s \in S_{k+1}} h(s,y^{'})\hat{t}_\pi(s,y')p(s)$.

(vii) *Mivel $\hat{t}_\pi(s,y')$ minden $s \in S_{k+1}$-re konstans (jel.: c_c), ezért*

$(viii)$ $\sum_{s \in S_{k+1}} h^2(s,y')p(s) \leq -2c_0 \sum_{s \in S_{k+1}} h(s,y^{'})p(s)$.

(v) *és* $(viii)$ *következményeként $\forall s \in S_{k+1} : h(s,y')p(s) = 0$. Mivel $\bigcup_{i=0}^{k+1} S_i = S$, teljesül $\forall s \in S : h(s,y')p(s) = 0$ is, tehát a lemmát bizonyítottuk. Most (i)-ből és (ii)-ből következik, hogy $\forall s \in S, \forall y \in Y_0 : h(s,y)p(s) = 0$, amiből a lemmát induktívan felhasználva következik, hogy $\forall s \in S, \forall y \in Y : h(s,y)p(s) = 0$. Ezzel a tételt beláttuk.*

3.9. Megjegyzés. *A tétel tartalmas, azaz van olyan, a $\forall k \in U : \pi_k > 0$ feltételt teljesítő*

$p(\cdot)$, *amely mellett D-ben van nem megengedett becslés.*

A tétel bizonyításakor csupán $\hat{t}_\pi$ torzítatlanságát, és (vii)-et implikáló tulajdonságát használtuk ki. (vii)-nek elégséges feltétele az, hogy $\hat{t}_\pi(s,y)$ olyan, hogy s-nek csak a nemnulla y értékkel rendelkező elemeitől függ, azaz ha s_1 és s_2 olyan, hogy ha $\forall k \in s_1 \cup s_2 \setminus (s_1 \cap s_2) : y_k = 0$, akkor $\hat{t}_\pi(s_1,y) = \hat{t}_\pi(s_2,y)$. Így a fenti tétel általánosabban is megfogalmazható.

3.10. Tétel. *Legyen $g(s,y)$ az $S \times Y$-on értelmezett függvény, amely (s,y)-tól csak azon (k,y_k) értékeken keresztül függ, melyekre $k \in s$ és $y_k \neq 0$. Ekkor g torzítatlan megengedett becslése a $G(y) = \sum_{s \in S} g(s,y)p(s)$ paraméternek a $G(y)$ torzítatlan becsléseinek osztályában.*

A (3.10)-ból közvetlenül adódik az alábbi

3.11. Következmény. *A Horvitz-Thomson becslés a $\forall k \in U : \pi_k > 0$, és a $\pi_{kl} > 0$, $\forall k,l \in U$ feltételt teljesítő bármely $p(\cdot)$ mellett megengedhető a $V_p\left(\hat{t}_\pi, y\right)$ torzítatlan becsléseinek osztályában.*

3.12. Megjegyzés. *A Yates-Grundy becslésre nem teljesül a (3.10) tétel feltétele.*

A fenti tételek a π-becslés, ill. a variancia Horvitz-Thomson becslésének torzítatlan becslésosztályban való megengedettségéről szóltak. Felmerülhet a kérdés, vajon nem hatásosak-e ezek a becslések. Az alábbiak erre a kérdésre adnak nemleges választ.

3.13. Tétel. *Legyen $p(\cdot)$ nem-cenzus elrendezés, azaz $\exists k \in U : \pi_k \neq 1$. Legyen $G(y)$ paraméterfüggvény az y-nak tényleges függvénye , azaz*
$\exists k \in U : (\pi_k \neq 1 \wedge \exists y \exists y' : [y_k \neq y'_k \wedge (\forall i \neq k : y_i = y'_i) \wedge G(y) \neq G(y')])$.
Tegyük még fel, hogy a becslések körét megszorítjuk a kizárólag a mintától függő becslések körére, vagyis azokra a $\hat{G}(s,y)$ $S \times Y$-on értelmezett függvényekre, amely (s,y)-tól csak azon (k,y_k) értékeken keresztül függnek, melyekre $k \in s$. Ekkor a $G(y)$ torzítatlan becsléseinek osztályában nincs hatásos becslés.

Bizonyítás. *(Basu, 1971) Legyen $\hat{G}(s,y)$ a paraméter torzítatlan becslése. Rögzítsük a tetszőleges $y_0 \in Y$ függvényt. Vegyük a $\hat{G}^*(s,y) = \hat{G}(s,y) - \hat{G}(s,y_0) + G(y_0)$ becslést.*

Nyilván $\forall y \in Y : E_p\left[\hat{G}^*(s,y)\right] = E_p\left[\hat{G}(s,y)\right] = G(y)$. *Továbbá* $\forall s \in S : \hat{G}^*(s,y_0) = G(y_0)$. *Ezért* $V_p\left[\hat{G}^*(s,y_0)\right] = 0$. *Azaz* $\forall y \in Y$ *esetén van 0 varianciájú torzítatlan becslés. Így* $\hat{G}(s,y)$ *akkor hatásos, ha* $V_p\left[\hat{G}(s,y)\right] = 0, \forall y \in Y$. *Ebből* $p(\cdot)$ *nem-cenzus tulajdonságát,* $G(y)$ *tényleges függvény voltát, és* $\hat{G}(s,y)$*-nek kizárólag a mintától való függését felhasználva a tétel következik.*

3.14. Következmény. [1] *Bármely* $p(\cdot)$ *nem-cenzus elrendezés mellett az összeg torzítatlan, kizárólag a mintától függő becsléseinek osztályában nincs hatásos becslés.*

3.15. Következmény. [2] *Bármely, a* $\forall k \in U : \pi_k > 0$ *feltételt teljesítő* $p(\cdot)$ *nem-cenzus elrendezés mellett a* $V_p\left(\hat{t}_\pi, y\right)$ *torzítatlan, kizárólag a mintától függő becsléseinek osztályában nincs hatásos becslés.*

Eddig adott $p(\cdot)$ elrendezés mellett értékeltünk különböző becsléseket. Megfordítva, az összeg π-becslésének teljesítményét is összevethetjük különböző elrendezések mellett.

3.16. Tétel. *(Godambe, Joshi, 1965) Legyen* $p(\cdot)$ *rögzített mintanagyságot adó elrendezés,* π_k, $k \in U$ *a hozzá tartozó elsőrendű kiválasztási valószínűségek. Ekkor* $p(\cdot)$*-hez nem konstruálható olyan nem rögzített mintanagyságot adó, azonos* π_k *kiválasztási valószínűségeket implikáló* $p'(\cdot)$ *elrendezés, amely mellett* $\forall y \in Y : V_{p'}\left(\hat{t}_\pi, y\right) \leq V_p\left(\hat{t}_\pi, y\right)$.

Bizonyítás. *Jelölje* $E_{p'}(n(s))$ *a várható mintanagyságot,* $E_{p'}(n(s)) = \sum_U \pi_k$. *Jelölje* π_{kl} *ill.* π'_{kl} *a megfelelő másodrendő kiválasztási valószínűségeket. Könnyen belátható, hogy*
(i) $\sum_{k\in U}\sum_{k'\neq k\in U}\pi'(k,k') = E^2_{p'}(n(s)) - E_{p'}(n(s)) + Var_{p'}(n(s))$, *illetve, hogy a* π*-becslés tetszőleges* $p_0(\cdot)$ *melletti varianciája felbontható, mint*
(ii) $V_{p_0}\left(\hat{t}_\pi, y\right) = \sum_U y_k^2/\pi_{0k} + \sum_{k\in U}\sum_{k'\neq k\in U}\left[\pi_0(k,k')/\left(\pi_{0k}\pi_{0k'}\right)\right] y_k y_{k'} - t^2(y)$.
Ha a tétel nem teljesülne, akkor a két variancia egyenlőtlenségéből, az $y \in Y : y_k = \pi_k$ *függvényt és* $p(\cdot)$*-t ill.* $p'(\cdot)$*-t* (ii)*-be helyettesítve, a*
$\sum_{k\in U}\sum_{k'\neq k\in U}\pi(k,k') \geq \sum_{k\in U}\sum_{k'\neq k\in U}\pi'(k,k')$
egyenlőtlenséget kapnánk, ami (i)*-nek ellentmond.*

[1] A következmény ismert tétel (Godambe, Joshi, 1965), de a (3.13) következményeként történő levezetése a szakdolgozat szerzőjétől származik.

[2] A következmény ismert tétel (Godambe, Joshi, 1965), de a (3.13) következményeként történő levezetése a szakdolgozat szerzőjétől származik.

3.17. Megjegyzés. [3] *$p'(\cdot)$ nem rögzített mintanagyságot adó voltát az utolsó lépésben használtuk ki, amiből látható, hogy rögzített mintanagyságot adó $p'(\cdot)$-re a $V_{p'}\left(\hat{t}_\pi, y\right) = V_p\left(\hat{t}_\pi, y\right)$ egyenlőséget vezethetnénk le. Elmondható, hogy az azonos, rögzített mintanagyságot adó elrendezéseket elsőrendű kiválasztási valószínűségeik a π-becslés varianciájának erejéig meghatározzák.*

A tétel a rögzített mintanagyságot adó elrendezések előnyben részesítését támaszthatja alá, ennek a preferenciának a gyakorlati okaira a Bernoulli-mintavételnél visszatérünk.
Az (3.16) tételnél általánosabb a következő

3.18. Tétel. [4] *Legyen $p(\cdot)$ rögzített mintanagyságot adó elrendezés. Ekkor minden nem rögzített, de vele azonos várható mintanagyságot adó $p'(\cdot)$ elrendezésre*
$\exists y \in Y : V_{p'}\left(\hat{t}_\pi, y\right) > V_p\left(\hat{t}_\pi, y\right)$, illetve minden rögzített, vele azonos várható mintanagyságot adó $p''(\cdot)$ elrendezésre vagy $\forall y \in Y : V_{p''}\left(\hat{t}_\pi, y\right) = V_p\left(\hat{t}_\pi, y\right)$, vagy $\exists y \in Y : V_{p''}\left(\hat{t}_\pi, y\right) > V_p\left(\hat{t}_\pi, y\right)$.
Bizonyítás. *A $(\forall k \in U : \pi_k = \pi'_k)$, ill. a $(\forall k \in U : \pi_k = \pi''_k)$ feltétel teljesülése esetén a tétel teljesülése a (3.16)-ból következik. Egyébként $p(\cdot)$-t tekintve $\exists i \in U : \pi_i > \pi'_i$. Legyen y olyan, hogy $y_i \neq 0$, és $\forall k \neq i : y_k = 0$. Ekkor a korábbi tétel bizonyításának (ii) felbontásából $V_{p'}\left(\hat{t}_\pi, y\right) = \frac{y_i^2}{\pi'_i} - y_i^2 > \frac{y_i^2}{\pi_i} - y_i^2 = V_p\left(\hat{t}_\pi, y\right)$. Ugyanez a meggondolás bizonyítja a tételt $p'(\cdot)$ helyett $p''(\cdot)$-re is.*

3.19. Következmény. *Az összeg π-becslésének rögzítése mellett értelemszerűen definiálható elrendezések megengedettsége. Legyen $p(\cdot)$ és $p'(\cdot)$ két elrendezés, a várható mintanagyságok $n = E_p(n_s)$ és $n' = E_{p'}(n_s)$. Azt mondjuk, hogy a $p(\cdot)$ elrendezés jobb, mint a $p'(\cdot)$, ha (1) $n \leq n'$ és (2) $\forall y \in Y : V_p(\hat{t}_\pi, y) \leq V_{p'}(\hat{t}_\pi, y)$, és (1) vagy legalább egy y-ra (2) szigorú egyenlőtlenséggel teljesül. A $p(\cdot)$ elrendezés megengedett, ha nincs nála jobb. A tétel következményeként elmondható, hogy minden rögzített mintanagyságú elrendezés megengedett a vele azonos várható mintanagyságot adó elrendezések osztályán.*

[3] A szakdolgozat szerzőjétől..

[4] A tétel bizonyítása, ill. (3.16)-ból történő levezetése a szakdolgozat szerzőjétől származik. A tételt Chaudhuri (1988) bizonyítás nélkül közli az optimalitási tételkört áttekintő dolgozatában.

Az utóbbi tétel szerint tehát általánosságban egyik elrendezés sem preferálható az összes többivel szemben. Ugyanakkor bizonyos speciális esetekben érdemleges összehasonlításokat végezhetünk, mint ahogy a 3.3. és a 3.4. fejezetben látni fogjuk. Ugyanis a populációra vonatkozó előismeretek birtokában, vagyis Y-t leszűkítve egy adott elrendezésnél gyakran jobb elrendezés található.

3.1.2. Példa: egyszerű véletlen mintavétel

A π-becslés alkalmazásának példájaként tekintsük az egyik legegyszerűbb elrendezést.

3.20. Definíció. *Egyszerű véletlen mintavételről (simple random sampling, jel: SI) beszélünk, ha a mintavételi elrendezés:*

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}}, \text{ ha } |s| = n \\ 0 \text{ egyébként} \end{cases}.$$

SI elrendezés esetén a π-becslésre egyszerűen adódik:

$$\hat{t}_\pi = N\overline{y}_s,$$

ahol $\overline{y}_s = \sum_s y_k/n$ a mintaátlag. Jelölje $\overline{y}_U \stackrel{def}{=} \sum_U y_k/N$ a populációs átlagot, $f \stackrel{def}{=} \frac{n}{N}$ a mintaarányt, és S^2_{yU} a populációs varianciát:

$$S^2_{yU} \stackrel{def}{=} \frac{1}{N-1}\sum_U (y_k - \overline{y}_U)^2.$$

A mintaelem-indikátorok kovarianciájára egyszerűen adódik:

$$\triangle_{kl} = -\frac{f(1-f)}{N-1}, \forall k,l \in U, k \neq l.$$

Mivel az SI rögzített mintanagyságú elrendezés, a $\hat{t}_\pi$ varianciája 3.5 szerint felírható, mint:

3.21. Állítás. $V_{SI}(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{yU}$.
(3.5) felhasználásával a variancia Yates-Grundy becslése az alábbi alakban áll elő:

$$\hat{V}_{SI}(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{ys},$$

ahol $S^2_{ys} \stackrel{def}{=} \frac{1}{n-1}\sum_s (y_k - \overline{y}_s)^2$.

3.22. Megjegyzés. *Az SI elrendezés esetén a variancia Horvitz-Thomson és Yates-Grundy becslései megegyeznek.*

Alkalmazás. Az egyszerű véletlen mintavétel szerepe fontos, hiszen az elmélet kiindulópontja, de a gyakorlatban ritkán alkalmazzák. Ennek egyik oka az, hogy alkalmazása a populáció elemeinek már a mintavétel előtti azonosítását, felsorolását feltételezi. Ez sokszor nem megvalósítható, ilyenkor vagy az elemek azonosítható csoportjainak kiválasztása előzi meg az elemek kiválasztását (lásd később a csoportos és a kétlépcsős mintavételt), vagy a populáció azonosíthatóságát nem igénylő szisztematikus mintavételt alkalmazzák. Ugyanakkor az SI-t azonosítható populáció esetén is ritkán használják, ennek legfőbb oka a magas költségigénye. Mivel $p(s)$ minden n elemű mintára azonos, földrajzilag igen szórt mintát eredményezhet, ami pl. a kérdezők utazási költségeit emeli. Olcsóbb lehetőséget jelent a később tárgyalandó csoportos mintavétel. Végül megemlítjük a hatásosság szempontját: az SI azonos kiválasztási valószínűségeket eredményez ($\forall k : \pi_k = \frac{n}{N}$), ám megfelelően megválasztott, nem azonos kiválasztási valószínűségekkel sok esetben kisebb becslési variancia érhető el, lásd a Poisson mintavétel vagy a rétegezés esetén tárgyalt optimalitási problémát.

3.2. Az elrendezés hatása

Az SI elrendezés viszonyítási alapként szolgál más mintavételi elrendezések esetén kapott $\pi-$becslések varianciájának megítéléséhez.

3.23. Definíció. *Legyen az SI elrendezéshez tartozó minta n elemű. Jelöljön $p(\cdot)$ egy másik mintavételi elrendezést, $\hat{t}_\pi$ pedig az elrendezés esetén kapott $\pi-$becslést. Az összevetés érvényességéhez tegyük fel, hogy a várható mintanagyság $p(\cdot)$-re is $\sum_U \pi_k = n$. Ekkor az elrendezés hatásának (design effect) nevezzük az alábbi mennyiséget:*

$$deff(p) \stackrel{def}{=} \frac{V_p(\hat{t}_\pi)}{V_{SI}(N\overline{y}_s)} = \frac{\sum\sum_U \triangle_{kl}\check{y}_k\check{y}_l}{N^2\frac{1-f}{n}S^2_{yU}}.$$

A $deff$ a $p(\cdot)$ elrendezésnek a becslési varianciára gyakorolt hatását méri. Az 1-et meghaladó értékű $deff$ úgy interpretálható, hogy a $p(\cdot)$ alkalmazásával vesztünk a becslés megbízhatóságából az SI-hez képest.

Az általunk felkutatott források ugyan nem említik, de érdemes a $deff$-et kapcsolatba hozni a (3.18) tétellel. A tétel következménye szerint a π-becslés mellett az azonos várható mintanagyságot adó elrendezések között minden rögzített mintanagyságot adó elrendezés

megengedett. Így rögzített mintanagyságot adó $p(\cdot)$ mellett a $deff(p)$ y-tól függően az 1 mindkét oldalán felvesz értékeket, vagy degenerált esetben értéke konstans 1. Szintén a következmény szerint, nem rögzített mintanagyságot adó $p(\cdot)$ viszont biztos, hogy legalább egy y-ra kevésbé hatásos, mint az SI. Tehát általában nem található az SI-nél jobb elrendezés, de, mint később látni fogjuk, bizonyos előismeretek birtokában az y-t specifikálva jobbnak ítélhetjük egyiket a másiknál.

A következőkben néhány igen gyakran alkalmazott mintavételi elrendezést veszünk sorra, a legtöbb felmérés ezekre az elrendezésekre, vagy ezek kombinációival kapott elrendezésekre épül. A felsorolt esetekre kiszámítjuk az összeg π-becslésének $V(\hat{t}_\pi)$ varianciáját és a $\hat{V}(\hat{t}_\pi)$ becslést, speciális esetekben összevetve az egyes elrendezések hatásosságát.

3.3. Közvetlen elemkiválasztásra épülő elrendezések

A közvetlen elemkiválasztásra épülő elrendezések közül a SI elrendezést korábban már vizsgáltuk. Egy másik, egyszerű példa a Bernoulli mintavétel.

3.3.1. Bernoulli mintavétel

3.24. Definíció. *A $p(\cdot)$ Bernoulli mintavétel (jel.: BE), ha $\exists\ 0 < \pi < 1$, hogy*

$$p(s) = \pi^{n(s)} (1-\pi)^{N-n(s)},\ s \in \mathbf{S}.$$

A BE egy megvalósítása: legyenek $\varepsilon_1, \ldots, \varepsilon_i, \ldots \varepsilon_N$ független azonos eloszlású valószínűségi változók, $\varepsilon_i \sim U(0,1)$. A $k \in U$ kiválasztásáról a következő módon döntünk: $k \in s \Leftrightarrow \varepsilon_k < \pi$.

3.25. Állítás. *Az $I_1 \ldots I_N$ mintaelem indikátorok függetlenek és azonos binomiális eloszlásúak: $P(I_k = 1) = \pi$, $\forall k \in U$.*

Az összeg (3.2) szerinti π-becslése:

$$\hat{t}_\pi = \frac{1}{\pi} \sum_s y_k.$$

A becslés varianciája (3.3)-ból:

$$V_{BE}\left(\hat{t}_\pi\right) = \left(\frac{1}{\pi} - 1\right) \sum_U y_k^2.$$

A $V_{BE}\left(\hat{t}_\pi\right)$ torzítatlan becslése (3.3) alapján:

$\hat{V}_{BE}\left(\hat{t}_\pi\right) = \frac{1}{\pi}\left(\frac{1}{\pi}-1\right)\sum_s y_k^2$.

A várható mintanagyságot $n = N\pi$-vel jelölve a variancia az alábbi alakban írható fel:

$V_{BE}\left(\hat{t}_\pi\right) = N^2\left(\frac{1-f}{n}\right)S_{yU}^2\left(1-\frac{1}{N}+\left(\frac{\bar{y}_U}{S_{yU}}\right)^2\right)$,

amiből az elrendezés hatása (3.21) felhasználásával:

$deff(BE) = \frac{V_{BE}\left(\hat{t}_\pi\right)}{V_{SI}(N\overline{y}_s)} = \left(1-\frac{1}{N}+\left(\frac{\bar{y}_U}{S_{yU}}\right)^2\right)$,

ami általában 1-nél nagyobb érték, tehát azonos mintanagyság mellett az egyszerű véletlen mintavétel várhatóan megbízhatóbb variancia-becslést ad.

Alkalmazás. A BE nem rögzített mintanagyságú elrendezés. Ez a felmérések gyakorlatában nagy hátrányt jelent a pontosan nem kalkulálható költségek miatt. Mégis érdemes említést tenni róla, hiszen az előre tervezett, rögzített mintanagyság ténylegesen a gyakorlatban sem valósítható meg; és az ezt okozó (egy lehetséges modell szerint azonos, π valószínűséggel fellépő) válaszmegtagadás éppen a BE mintavétellel modellezhető jól.

3.3.2. Poisson mintavétel

Az egyszerű véletlen kiválasztás és a Bernoulli mintavétel azonos kiválasztási valószínűségeket eredményez. A gyakorlatban nagyobb hatásosságuk miatt gyakran nem azonos kiválasztási valószínűségeket adó elrendezést alkalmaznak, ilyen a Poisson mintavétel is.

3.26. Definíció. *A $p(\cdot)$ Poisson mintavétel (jel.: PO), ha*

$p(s) = \prod_{k\in s}\pi_k\prod_{k\in U\setminus s}(1-\pi_k)$, $s\in\mathbf{S}$.

A Poisson mintavétel tulajdonságai:

(i) $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$,

(ii) $V_{PO}\left(\hat{t}_\pi\right) = \sum_U\left(\frac{1}{\pi_k}-1\right)y_k^2$,

(iii) a Horvitz-Thomson becslés: $\hat{V}_{PO}\left(\hat{t}_\pi\right) = \sum_s(1-\pi_k)\check{y}_k^2$.

Mint a Bernoulli mintavétel esetén, a mintanagyság most sem rögzített. Rögzített várható mintanagyság, $n = \sum_U \pi_k$ mellett adódik a kérdés: milyen π_k elosztás mellett optimális az elrendezés, azaz milyen elosztás minimalizálja a $V_{PO}\left(\hat{t}_\pi\right)$ varianciát? A feladat ekvivalens a

$$\left(\sum_U y_k^2/\pi_k\right)\left(\sum_U \pi_k\right) \to \min$$

optimalizálási problémával. A Cauchy-Schwartz egyenlőtlenségből

$$\left(\sum_U y_k^2/\pi_k\right)\left(\sum_U \pi_k\right) \geq \left(\sum_U y_k\right)^2,$$

itt az egyenlőség $y_k/\pi_k = \lambda$ esetben teljesül. Mivel $n = \sum_U \pi_k$, az eredmény:

$$\pi_k = n y_k / \sum_U y_k$$

feltéve, hogy $y_k \leq \sum_U y_k/n$, $\forall k$.

Alkalmazás. Mivel az y_k-k ismeretlenek, az optimumot adó π_k elosztás a gyakorlatban nem valósítható meg. Ugyanakkor kihasználható ez a tétel, ha rendelkezésre áll az U elemeiről bizonyos előre ismert x_k segédinformáció, ami hozzávetőlegesen arányos y_k-val: $y_k/x_k \approx c$. Ebben az esetben a $\pi_k = n x_k / \sum_U x_k$ $(x_k \leq \sum_U x_k/n)$ választással a kapott π-becslés varianciája kicsi lesz, tehát megfelelően hatásos stratégiát (elrendezés-becslés párt) konstruálhatunk. Példaként: Magyarországon az U azonosítására a Központi Nyilvántartó és Választási Hivatal címlistája szolgál, mely tartalmazza a lakosok születési évét és lakhelyét, így előre ismert segédinformációként szolgálhat a populáció tagjainak kora, vagy lakhelyének lélekszáma.

3.3.3. Szisztematikus mintavétel

A 3.1.2. fejezetben az SI elrendezés kapcsán említettük annak kivitelezési nehézségeit. A szisztematikus mintavétel alkalmazása lényegesen egyszerűbb, ezért széles körben használt mintavételi eljárás.

Legyen adott az $a \in \mathbf{N}$, $a < N$ kiválasztási köz. Legyen $n \stackrel{def}{=} \left\lfloor \frac{N}{a} \right\rfloor$, ekkor $N = na + c$, $0 \leq c < a$. A mintavételi eljárás:

(i) 1 és a között azonos $1/a$ kiválasztási valószínűséggel véletlen számot választunk, legyen ez r.

(ii) A minta a következő:

$$s \stackrel{def}{=} \{k : k = r + (j-1)\, a \leq N; j = 1, \ldots, n_s\} = s_r$$

Jelölje $\mathbf{S}_{SY} = \{s_1, \ldots, s_a\} \subseteq \mathbf{S}$ az a mellett pozitív valószínűséggel fellépő minták halmazát.

3.27. Definíció. *Az így definiált eljárás elnevezése szisztematikus mintavétel (jel.: SY). A megfelelő $p(\cdot)$ elrendezés:*

$$p(s) = \begin{cases} 1/a, \text{ ha } s \in \boldsymbol{S}_{SY} \\ 0 \text{ egyébként} \end{cases}$$

Az első- és másodrendő mintaelem indikátorok:

$$\pi_k = 1/a$$

$$\pi_{kl} = \begin{cases} 1/a, \text{ ha } \exists j \in \mathbf{N} : k = l + ja \\ 0 \text{ egyébként} \end{cases}$$

Az SY elrendezés tulajdonságai:

(i) $\hat{t}_\pi = a \sum_s y_k$,

(ii) $V_{SY}\left(\hat{t}_\pi\right) = a \sum_{r=1}^{a} \left(t_{s_r} - \overline{t}\right)^2$, ahol $t_{s_r} = \sum_{s_r} y_k$, $s_r \in \mathbf{S}_{SY}$ az s_r-beli mintaösszeg, és $\overline{t} = \sum_{r=1}^{a} \frac{t_{s_r}}{a}$ a mintaösszegek átlaga.

3.28. Megjegyzés. *Mivel a $\pi_{kl} > 0, \forall k, l \in U$ feltétel nem teljesül, sem a Horvitz-Thomson, sem a Yates-Grundy torzítatlan variancia-becslés nem alkalmazható. Sőt, az SY esetén egyáltalán nem létezik torzítatlan $\hat{V}_{SY}\left(\hat{t}_\pi\right)$ variancia-becslés, mivel nincs információnk az U rendezéséről. Az irodalom torzított, de a varianciát várhatóan inkábbb felülbecslő becsléseket ismertet, vagy az elrendezés különböző módosításait javasolja a $\pi_{kl} > 0, \forall k, l \in U$ feltétel teljesüléséhez (Särndal et al, 1992, Wolter, 1985). Ha az U rendezése ismert, a becslés elvégezhető, pl. ha véletlen rendezést feltételezünk, az SY az egyszerű véletlen mintavétellel ekvivalens, és a variancia becslése is $\hat{V}_{SI}(\hat{t}_\pi)$-val egyezik.*

A $V(\hat{t}_\pi)$ varianciára kapott felírásból látható, hogy a becslés annál megbízhatóbb, minél kisebb a mintaösszegek közötti eltérés. Mivel az $\mathbf{S}_{SY}$ az U elemeinek sorrendjének függvénye, a mintaösszegek értéke, így maga a becslés megbízhatósága is az U rendezésétől függ.

Alkalmazás. A mintaösszegek közötti eltérés csökkenthető, ha az U rendezését az y_k értékek szerinti rendezéssel oldjuk meg. A gyakorlatban természetesen ezek az y_k-k ismeretlenek, de gyakran rendelkezésre áll olyan x_k segédinformáció az U elemeiről, amely segítheti az y_k értékek szerinti rendezést, pl. az $\frac{x_k}{y_k} \sim c$ feltétellel, ahol c konstans. A segédinformáció használatával lényegesen csökkenthető az SY elrendezés összeg-becslésének varianciája. A magyarországi gyakorlatban, ahol az U-t, mint említettük, gyakran a Központi Nyilvántartó és Választási Hivatal címlistájában felsorolt személyek képezik, ilyen segédinformáció lehet a lakóhelyül szolgáló település mérete. Az SY tervezésekor az U-t településméret szerint rendezik, feltételezve, hogy az összefüggésben áll a megismerni kívánt y jellemzővel, ezért az így kapott $\mathbf{S}_{SY}$-beli minták nem térnek el nagyon egymástól y szerint sem, vagyis megbízható összeg-becslést adnak.

A fentiek alapján az SY elrendezés hatásosabb, ha az $\mathbf{S}_{SY}$-beli minták kevéssé különülnek el. A következő tétel ezt a meggondolást számszerűsíti.

3.29. Tétel. *Tegyük fel, hogy $N = an$, a egész. Ekkor*

$$V_{SY}\left(\hat{t}_\pi\right) = \frac{N^2 S_{yU}^2}{n}\left[(1-f) + (n-1)\,\delta\right], \text{ ebből}$$

$$deff(SY) = \frac{V_{SY}}{V_{SI}} = 1 + \frac{n-1}{1-\frac{1}{a}}\delta,$$

ahol δ a homogenitás mértéke, a csoportokon belüli (sum of squares within, SSW) és azok közötti (sum of squares between, SSB) négyzetes eltéréssel definiált mennyiség:

$$\delta \overset{def}{=} 1 - \frac{N-1}{N-a}\frac{SSW}{SST} = 1 - \frac{N-1}{N-a}\frac{\sum_{r=1}^{a}\sum_{s_r}\left(y_k - \overline{y}_{s_r}\right)^2}{\sum_U (y_k - \overline{y}_U)^2}.$$

A δ lehetséges értékei $-\frac{a-1}{N-a}$ és 1 között vannak. Láthatóan az SY hatékonyabb, mint az SI, ha a $\delta < 0$, azaz ha kismértékű a minták így definiált homogenitása.

Alkalmazás. A szisztematikus mintavétel kézenfekvő megoldás akkor, amikor nincs megfelelő nyilvántartás az U elemeiről, pl. ha egy kórház betegeinek kartonjai közül, egy folyamatosan növekvő adatbázisból választunk mintát. Mivel nem ismert a populáció elemszáma, az SI elrendezés ekkor nem valósítható meg. Ugyanakkor a folyamatosan bejövő kartonok közül könnyen kiválasztható minden a., az a értékét a kívánt mintanagyság szerint a korábbi tapasztalatok alapján megadva. Ha a ciklusú periodicitás tapasztalható a bejövő kartonok y_k értékeiben, a nagy δ kevéssé hatásos becslést implikál. Ha viszont y_k

és a betegfelvétel időpontja között összefüggés van, akkor - mivel a lista dátum szerint rendezett, - hatásosabb becslést kapunk.

3.3.4. Rétegezés

Mint korábban a 3.1.2. fejezetben említettük, az SI elrendezés hátránya több elrendezéssel szemben kisebb hatásosságában jeletkezik. A rétegzés az SI kivitelezési egyszerűségét kombinálja a potenciálisan nagyobb hatásossággal. Elsődleges célja, hogy az y szempontjából heterogén populációt homogén rétegekre bontsuk, amelyeken belül kisebb varianciájú becslés nyerhető viszonylag kis mintából is. Más esetekben az egyes részpopulációk eltérő viselkedése (pl. válaszmegtagadás), vagy a róluk rendelkezésre álló eltérő előzetes ismeretek motiválhatják a rétegenként eltérő elrendezést. Használata az U elemeiről meglévő előzetes segédinformáció meglétét előfeltételezi: az U minden eleméről tudnunk kell, hogy mely réteghez tartozik.

3.30. Definíció. *Legyen $U_1, \ldots, U_h, \ldots, U_H$ az U egy partíciója, az U_h osztályokat nevezzük rétegeknek. Rétegezésnek (jel.: ST) nevezzük azt a $p(\cdot)$ elrendezést, amelynek esetén a mintaválasztás rétegenként egymástól független módon történik, az U_h osztályra a $p_h(\cdot)$ elrendezést alkalmazva.*

Az s mintára: $s = s_1 \cup \ldots \cup s_H$, továbbá a rétegenkénti függetlenség miatt: $p(s) = p_1(s_1) \cdots p_H(s_H)$. Az elrendezés tulajdonságaiból könnyen adódnak az alábbiak:

3.31. Tétel. *Rétegezéskor az összeg $\pi-$becslése:*

$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi}$,

ahol $\hat{t}_{h\pi}$ a $t_{h\pi}$ π-becslése, $t_{h\pi} = \sum_{U_h} y_k$. A becslés varianciája a következő alakban áll elő:

$V_{ST}(\hat{t}_\pi) = \sum_{h=1}^{H} V_h(\hat{t}_{h\pi})$,

ahol $V_h(\hat{t}_{h\pi})$ a $\hat{t}_{h\pi}$ becslés varianciája. A variancia torzítatlan becslése:

$\hat{V}_{ST}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}_h(\hat{t}_{h\pi})$,

feltéve, hogy létezik torzítatlan $\hat{V}_h(\hat{t}_{h\pi})$ becslés minden h-ra.

3.32. Példa. *Alkalmazzuk eredményeinket arra az estre, amikor a rétegeken belül mindenhol SI elrendezést használunk (jel.: STSI). Jelölje n_h a h. rétegen belüli mintanagyságot, legyen* $\bar{y}_{s_h} = \sum_{s_h} y_k/n_h$, $f_h = \frac{n_h}{N_h}$,
$S^2_{yU_h} = \frac{1}{N_h-1}\sum_{U_h}(y_k - \bar{y}_{U_h})^2$, $S^2_{ys_h} = \frac{1}{n_h-1}\sum_{s_h}(y_k - \bar{y}_{s_h})^2$. *Ekkor*

(i) $\hat{t}_\pi = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$,

(ii) $V_{STSI}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S^2_{yU_h} = \sum_{h=1}^{H} N_h^2 S^2_{yU_h}/n_h - \sum_{h=1}^{H} N_h S^2_{yU_h}$,

(iii) $\hat{V}_{STSI}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S^2_{ys_h}$.

Alkalmazás. Az $n_1, \ldots, n_h, \ldots, n_H$ rétegenkénti mintanagyságok meghatározása pl. a $V_{ST}(\hat{t}_\pi)$ variancia minimalizálásának feltétele mellett történhet (*Neyman allokáció*); a 5.1. fejezet ennek a problémának komplexebb változatát mutatja be, amikor a variancia minimalizálása adott költség-korlát figyelembevételével történik. Az allokáció meghatározása ezen kívül történhet az y_k-val feltételezésünk szerint erősen összefüggő x_k segédinformáció előzetesen ismert rétegenkénti jellemzőinek figyelembevételével. Az utóbbi stratégiát gyakran alkalmazzák a gyakorlatban, pl. ennek speciális esete az, amikor a mintát megyék szerint rétegezik, az n_h mintanagyságokat a megyénkénti népességszámmal arányosan megállapítva. Ez az *arányos allokáció*, mely a következőképpen definiálható:

$$n_h = nN_h/N,$$

ahol $N_h = |U_h|$. Belátható, hogy arányos allokáció mellett, amikor a $p_h(\cdot)$ rétegenkénti elrendezések mindegyike SI, a π-becslés varianciája a következőképpen áll elő:

$$V_{STSI,\,p}(\hat{t}_\pi) = N\left(\tfrac{1}{n} - \tfrac{1}{N}\right)\sum_{h=1}^{H} N_h S^2_{yU_h},$$

Rögzített n mellett összevetve az SI és az arányos allokációval működő STSI elrendezést:

$$V_{SI}\left(N\overline{y}_s\right) - V_{STSI,p}(\hat{t}_\pi) =$$
$$= \tfrac{N^3}{N-1}\left(\tfrac{1}{n} - \tfrac{1}{N}\right)\left[\sum_{h=1}^{H}\tfrac{N_h}{N}\left(\overline{y}_{U_h} - \overline{y}_U\right)^2 - \tfrac{1}{N}\sum_{h=1}^{H}\left(1 - \tfrac{N_h}{N}\right)S^2_{yU_h}\right].$$

Látható, hogy elméletileg elképzelhető az STSI hátránya az SI-vel szemben - ha a rétegenkénti $\overline{y}_{U_h}$ átlagok azonosak vagy közel azonosak. Ugyanakkor az y szempontjából jól elkülönült és viszonylag homogén rétegek megválasztásával az SI-nél jóval hatásosabb

elrendezés valósítható meg. Mivel az adminisztráció vagy egyéb költségek szempontjából az SI és az ST nem különböznek lényegesen, általában érdemes az utóbbit választani.

3.4. Közvetett kiválasztásra épülő módszerek

A gyakorlatban ritkán alkalmazzák a fenti, elemkiválasztásra épülő elrendezéseket. Ez elsősorban megvalósíthatósági vagy gazdaságossági szempontokra vezethető vissza. Megvalósításuk akadálya, hogy több országban nem áll rendelkezésre az U populáció elemeinek közvetlen elérését lehetővé tevő nyilvántartás. Ehelyett az elemek csoportjait nyilvántartó listát használhatnak, és a csoportokból vesznek mintát. Pl. Nagy-Britanniában a háztartások címlistáját alkalmazzák háztartás-választásra, majd a háztartás tagjai közül a terepmunka során választják ki a kérdezendő személyt. Másik fontos érv a közvetlen elemkiválasztás ellen annak költségvonzata: nagyon szétszórt mintát eredményezhet, ami költségessé teszi a terepmunkát. A csoportos és többlépcsős elrendezések ilyen esetben is jobb megoldást nyújtanak.

3.4.1. Egylépcsős csoportos mintavétel

Legyen $U_1, \ldots, U_h, \ldots, U_{N_I}$ az U populáció egy partíciója, az U_h osztályokat nevezzük csoportoknak. A csoportok halmazát reprezentáljuk az $U_I = \{1, \ldots, N_I\}$ indexhalmazzal. (Az I index a mintavétel első lépcsőjére utal.) A csoportok nagyságát jelölje N_i, $N = \sum_{i \in U_I} N_i$.

3.33. Definíció. *Egylépcsős csoportos mintavétel (jel.: C) alkalmazásakor a $p_I(\cdot)$ elrendezésnek megfelelően U_I-ből s_I mintát választunk, majd a kiválasztott csoportok minden elemét bevonjuk a végső s mintába: $s = \cup_{i \in s_I} U_i$.*

A $p_I(\cdot)$-nek megfelelő, csoportokra vonatkozó első- és másodrendű mintaelem-indikátorokra:

$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I)$, $i \in U_I$,

$\pi_{Iij} = \sum_{s_I \ni i,j} p_I(s_I)$, $i, j \in U_I$.

Jelölje t_i az i csoportbeli összeget: $t_i = \sum_{U_i} y_k$. Jelölje $\check{t}_i = t_i/\pi_{Ii}$ a csoportbeli összeg π-kiterjesztettjét. A (2.3) szerint jelölje az indikátorok kovarianciáját $\triangle_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ii}$, $\check{\triangle}_{Iij} = \triangle_{Iij}/\pi_{Iij}$ ezek π-kiterjesztettjét. Ekkor a (3.2) és (3.3) alkalmazásával közvetlenül adódik az alábbi tétel.

3.34. Tétel. *Csoportos mintavétel esetén az összeg π-becslése:*

$$\hat{t}_\pi = \sum_{s_I} \check{t}_i.$$

A becslés varianciája:

$$V_C(\hat{t}_\pi) = \sum\sum_{U_I} \triangle_{Iij}\check{t}_i\check{t}_j,$$

míg a variancia Horvitz-Thomson becslése:

$$\hat{V}_C(\hat{t}_\pi) = \sum\sum_{s_I} \check{\triangle}_{Iij}\check{t}_i\check{t}_j.$$

Amennyiben $p_I(\cdot)$ rögzített mintanagyságot adó elrendezés, a variancia és annak Yates-Grundy becslése előáll, mint

$$V_C(\hat{t}_\pi) = -\tfrac{1}{2}\sum\sum_{U_I} \triangle_{Iij}\left(\check{t}_i - \check{t}_j\right)^2,$$

$$\hat{V}_C(\hat{t}_\pi) = -\tfrac{1}{2}\sum\sum_{s_I} \check{\triangle}_{Iij}\left(\check{t}_i - \check{t}_j\right)^2.$$

Alkalmazás. A variancia utóbbi előállítása alapján elmondható, hogy ha a $\check{t}_i = \frac{t_i}{\pi_{Ii}}$ mennyiségek közel azonosak, nagyon kicsi a variancia. Tehát a $\pi_{Ii} \sim t_i$ választással, ha ez kivitelezhető, a csoportos mintavétel igen hatásos lehet. Egy másik megközelítésben ha a π_{Ii}-k azonosak, akkor hatásosabb becslést a $t_i = N_i\bar{y}_{U_i}$ csoportösszegek azonossága esetén kapnánk, vagyis ha az $\bar{y}_{U_i}$ csoportátlagok N_i^{-1}-vel megközelítően arányosak. Azonban a gyakorlatban ez ritkán fordul elő.

Speciálisan legyen a $p_I(\cdot)$ elrendezés egyszerű véletlen mintavétel (jel.: SIC), tehát válasszunk visszatevés nélkül, azonos kiválasztási valószínűségekkel az N_I-ből n_I csoportot. Jelölje $\overline{N} = N/N_I$ az átlagos csoportnagyságot. Ekkor az elrendezés hatása:

$$deff(SIC) = \frac{V_{SIC}}{V_{SI}} = 1 + \frac{N - N_I}{N_I - 1}\delta + \frac{Cov}{\overline{N}S_{yU}^2}, \tag{3–1}$$

ahol a δ a csoportok homogenitásának mértéke, a (3.29)-beli definíciót a csoportokra alkalmazva: $\delta = 1 - \frac{N-1}{N-N_I} \frac{\sum_{U_I} \sum_{U_i} \left(y_k - \overline{y}_{U_i}\right)^2}{\sum_U \left(y_k - \overline{y}_U\right)^2}$; Cov pedig az N_i csoportnagyság és az $N_i \overline{y}_{U_i}^2$ kovarianciája: $Cov = \frac{1}{N_I - 1} \sum_{U_I} \left(N_i - \overline{N}\right) N_i \overline{y}_{U_i}^2$.

Alkalmazás. Speciálisan azonos csoportnagyságok esetén (3–1) a $deff(SIC) = 1 + \frac{N-N_I}{N_I-1}\delta \approx 1 + \left(\overline{N} - 1\right)\delta$ formulára egyszerűsödik: csupán δ előjelétől függ az SIC-nek a SI-hez viszonyított hatásossága. A gyakorlati alkalmazásokban a csoportok (pl. háztartások) inkább homogének, tagjaiknak hasonlósága nagyobb, mint a véletlen módon kijelölt csoportoké lenne. Tehát pozitív δ feltételezhetünk, amiből a SIC hátránya következik. Ha a csoportnagyság nem állandó, akkor $Cov \neq 0$, és mivel Cov előjele a gyakorlatban inkább pozitív (a csoportnagyság pozitív összefüggésben áll $N_i \overline{y}_{U_i}^2$-vel), a SIC hátránya ilyen esetben még jelentősebb.
Említettük a közvetett kiválasztási módszerek, így a csoportos mintavétel költségekben jelentkező előnyét a közvetlen kiválasztásra épülő módszerekkel szemben. Ugyanakkor a fentiek szerint jelentős hátránya jelentkezhet a becslési hatásosság csökkenésében.

3.4.2. Kétlépcsős mintavétel

A becslési variancia a csoportos mintavétel esetén nyilván csökkenthető a mintába kerülő csoportok számának növelésével, de így a költségek is növekednek. Ésszerűbb megoldás a csoportok számának növelése mellett a csoportokon belül egy második mintavételi lépcsőt alkalmazni.

Vegyük az U egy $U_1, \ldots, U_i, \ldots, U_{N_I}$ partícióját, $|U_i| = N_i$. A csoportok halmazát reprezentáljuk az $U_I = \{1, \ldots, N_I\}$ indexhalmazzal.

3.35. Definíció. *Kétlépcsős mintavétel:*

(i) Első lépcső: A $p_I(\cdot)$ elrendezésnek megfelelően választjuk $s_I \subset U_I$ mintát.

(ii) Második lépcső: A $p_i(\cdot|s_I)$ elrendezésnek megfelelően választjuk az $s_i \subset U_i$ mintát, minden $i \in s_I-$re.

A továbbiakban szűkítjük az általános definíciót az invariancia és a függetlenség tulajdonságának megkövetelésével.

3.36. Definíció. *A második lépcső elrendezése invariáns, ha* $\forall i, \forall s_I \ni i : p_i(\cdot|s_I) = p_i(\cdot)$.

3.37. Definíció. *A második lépcső elrendezése független, ha*

$$\forall s_I : P\left(\cup_{i\in s_I} s_i | s_I\right) = \prod_{i\in s_I} P\left(s_i|s_I\right).$$

Az első lépcső $p_I(\cdot)$ elrendezéséhez tartozó első- és másodrendű kiválasztási valószínűségeket jelölje π_{Ii}, π_{Iij}, a második lépcső $p_i(\cdot)$ elrendezéséhez tartozó feltételes kiválasztási valószínűségeket pedig $\pi_{k|i}$, $\pi_{kl|i}$. Ezekkel a jelölésekkel az invariancia és a függetlenség tulajdonságából adódik az alábbi állítás.

3.38. Állítás. $\pi_k = \pi_{Ii}\pi_{k|i}$, *ha* $k \in U_i$, *továbbá*

$$\pi_{kl=}\begin{cases} \pi_{Ii}\pi_{k|i}, \text{ ha } k = l \in U_i \\ \pi_{Ii}\pi_{kl|i}, \text{ ha } k\&l \in U_i,\ k \neq l \\ \pi_{Iij}\pi_{k|i}\pi_{l|j}, \text{ ha } k \in U_i \wedge l \in U_j\ (i \neq j) \end{cases}$$

Tegyük fel, hogy az s_I minta realizálódott az első lépcsőben. Legyen $\check{y}_{k|i} = y_k/\pi_{k|i}$. A t_i-vel jelölt U_i-beli összeg $p_i(\cdot)$-nek megfelelő π-becslése:

$$\hat{t}_{i\pi} = \sum_{s_i} \check{y}_{k|i}.$$

$\hat{t}_{i\pi}$ torzítatlan becslés t_i-re. A becslés varianciája:

$$V_i = \sum\sum_{U_i} \triangle_{kl|i}\check{y}_{k|i}\check{y}_{l|i},$$

mely torzítatlan módon becsülhető az alábbi statisztikával:

$$\hat{V}_i = \sum\sum_{s_i} \check{\triangle}_{kl|i}\check{y}_{k|i}\check{y}_{l|i}.$$

A fentiek segítségével, (3.2)-t és (3.3)-t felhasználva kapjuk:

3.39. Tétel. *Kétlépcsős mintavétel esetén az összeg* π*-becslése:*

$$\hat{t}_\pi = \sum_{s_I} \hat{t}_{i\pi}/\pi_{Ii},$$

ahol $\hat{t}_{i\pi}$ *a* t_i*-vel jelölt* U_i*-beli összeg* $p_i(\cdot)$*-nek megfelelő* π*-becslése. A becslés varianciája szétbontható az első és a második lépcsőnek megfelelő komponensre:*

$$V_{2st}(\hat{t}_\pi) = V_I + V_{II},$$

ahol a $\check{t}_i \overset{def}{=} t_i/\pi_{Ii}$ *jelöléssel:*

$$V_I = \sum\sum_{U_I} \triangle_{Iij}\check{t}_i\check{t}_j,$$

$$V_{II} = \sum_{U_I} V_i/\pi_{Ii}.$$

A komponensek torzítatlan becslése:

$$\hat{V}_I = \sum\sum_{s_I} \check{\triangle}_{Iij}\frac{\hat{t}_{i\pi}}{\pi_{Ii}}\frac{\hat{t}_{j\pi}}{\pi_{Ij}} - \sum_{s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\hat{V}_i,$$

$$\hat{V}_{II} = \sum_{s_I}\hat{V}_i/\pi_{Ii}^2.$$

Bizonyítás. *Az invariancia és a függetlenség felhasználásával adódik:*

1. $\hat{t}_\pi = \sum_s y_k/\pi_k = \sum_{s_I}\sum_{s_i} y_k/\left(\pi_{Ii}\pi_{k|i}\right) = \sum_{s_I}\left(\sum_{s_i}\check{y}_{k|i}\right)/\pi_{Ii}$

   $= \sum_{s_I}\hat{t}_{i\pi}/\pi_{Ii}$

2. $V_{2st}(\hat{t}_\pi) = V_{p_I}\left[E\left(\hat{t}_\pi|s_I\right)\right] + E_{p_I}\left[V\left(\hat{t}_\pi|s_I\right)\right]$

   $= V_{p_I}\left(\sum_{s_i}\check{t}_i\right) + E_{p_I}\left(\sum_{s_i}V_i/\pi_{Ii}^2\right) = \sum\sum_{U_I}\triangle_{Iij}\check{t}_i\check{t}_j + \sum_{U_I}V_i/\pi_{Ii}$

3. $E\left(\hat{V}_I\right) = E_{p_I}\left[E\left(\hat{V}_I|s_I\right)\right]$

   $= E_{p_I}\left(\sum\sum_{s_I}\check{\triangle}_{Iij}\frac{t_{i\pi}}{\pi_{Ii}}\frac{t_{j\pi}}{\pi_{Ij}} + \sum_{s_I}\check{\triangle}_{Iii}V_i/\pi_{Ii}^2 - \sum_{s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)V_i\right)$

   $= \sum\sum_{U_I}\triangle_{Iij}\frac{t_{i\pi}}{\pi_{Ii}}\frac{t_{j\pi}}{\pi_{Ij}} + \sum_{U_I}\triangle_{Iii}V_i/\pi_{Ii}^2 - \sum_{U_I}\left(\frac{1}{\pi_{Ii}}-1\right)V_i$

4. $E\left(\hat{V}_{II}\right) = E_{p_I}\left[\sum_{s_I}E\left(\hat{V}_i|s_I\right)/\pi_{Ii}^2\right] = E_{p_I}\left(\sum_{s_I}V_i/\pi_{Ii}^2\right) = \sum_{U_I}V_i/\pi_{Ii}$

3.40. Megjegyzés. *1. $s_I = U_I$ feltétel mellett speciális eseteként a rétegezést kapjuk.*
2. $\forall i : s_i = U_i$ feltétel mellett az egylépcsős csoportos mintavételt kapjuk a kétlépcsős mintavétel speciális eseteként. Ekkor $V_{II} = 0$.

Alkalmazás. A mintavétel második lépcsőjéhez kötődő V_{II} variancia-komponens növeli a becslés varianciáját. Első olvasásra ezért úgy tűnhet, hogy az egylépcsős csoportos mintavétel megbízhatóbb becsléseket ad. Ez akkor igaz, ha a csoportok száma azonos. Ezzel szemben a kétlépcsős módszer előnye éppen az, hogy ugyanakkora minta mellett több csoport kiválasztására van lehetőség, és V_I csökkenése gyakran nem csak kompenzálja, de túl is haladja a V_{II} értékét.
Az elrendezés tervezésénél fontos szempont lehet annak ismerete, hogy a mintavétel egyes lépcsői milyen arányban járulnak hozzá a becslés teljes varianciájához. Ilyenkor pl. egy teszt-jellegű előfelmérésből becsülhető a V_I és V_{II} értéke.

4. Általános variancia-becslési módszerek

A 3. fejezetben az összeg π-becslésének varianciáját és a variancia Horvitz-Thomson ill. Yates-Grundy becslését mutattuk be néhány gyakran alkalmazott elrendezés esetén. Ha a θ, vagy a $\hat{\theta}$ az ott tárgyaltnál "komplexebb módon" áll elő (pl. θ több populációs összeg nemlineáris függvénye, lásd az 4.1.3. fejezetet), vagy ha maga az elrendezés rendkívül összetett, gyakran nem vagy csak nehezen állítható elő közvetlen módon kedvező tulajdonságokkal bíró variancia-becslés. A jelen fejezet két olyan általánosan alkalmazott módszert mutat be, melyek közvetett módon adnak becsléseket. Az első módszer, a linearizáció a komplex θ egy egyszerűbb közelítésén alapul. A módszerek másik típusát a rendkívül széles körben alkalmazott ismételgető technikák adják. Alapötletük az, hogy adott becslés varianciájával kapcsolatos információkhoz juthatunk a mintavétel kvázi-ismétléseiből kapott becslések varianciájából.

4.1. Linearizáció

4.1.1. π-becslés magasabb dimenzióban

A 3.1. fejezetben az y jellemző összegére definiáltuk a π-becslést. Legyen most az y vektorértékű függvény: $y \stackrel{def}{=} (y_1, \ldots, y_q)'$. Jelölje y_{jk} a j. koordináta értékét $k \in U$-ra. A t-re keresünk becslést

$$t \stackrel{def}{=} (t_1, \ldots, t_j, \ldots, t_q)',$$

ahol $t_j \stackrel{def}{=} \sum_U y_{jk}$. $s \subset U$ minta esetén a t paraméter π-becsléséhez koordinátánként alkalmazzuk a már ismert egydimenziós π-becslést:

$$\hat{t}_\pi \stackrel{def}{=} (\hat{t}_{1\pi}, \ldots, \hat{t}_{j\pi}, \ldots, \hat{t}_{q\pi})',$$

ahol $\hat{t}_{j\pi} \stackrel{def}{=} \sum_s \check{y}_{jk}$. $\hat{t}_\pi$ torzítatlan becslés t-re. A becslés varianciája, és a variancia becslése könnyen adódik az egydimenziós esetből. A

$$V(\hat{t}_\pi) = E\left\{\left(\hat{t}_\pi - t\right)\left(\hat{t}_\pi - t\right)'\right\}$$

kovarianciamátrix (j, j') cellája:

$$C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum\sum_{U} \triangle_{kl} \check{y}_{jk} \check{y}_{j'l}. \tag{4–1}$$

A $V(\hat{t}_\pi)$ mátrix torzítatlan becslése $\hat{V}(\hat{t}_\pi)$, melynek $(j, j\prime)$ cellája:

$$\hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum\sum_{s} \check{\triangle}_{kl} \check{y}_{jk} \check{y}_{j'l}. \tag{4–2}$$

4.1.2. π-becslés lineáris paraméterfüggvény esetén

Tegyük most fel, hogy a θ paraméter több összeg lineáris függvénye:

$\theta = a_0 + \sum_{j=1}^{q} a_j t_j$,

ahol $t_j = \sum_U y_{jk}$, $j = 1, \ldots, q$. Az 4.1.1. fejezetbeli eredmények felhasználásával adódik, hogy a

$\hat{\theta} = a_0 + \sum_{j=1}^{q} a_j \hat{t}_{j\pi}$

torzítatlan becslés θ-ra, ahol $\hat{t}_{j\pi}$ a t_j összeg π-becslése. A variancia:

$V(\hat{\theta}) = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_j a_{j'} C\left(\hat{t}_{j\pi}, \hat{t}_{j'\pi}\right)$,

a $C\left(\hat{t}_{j\pi}, \hat{t}_{j'\pi}\right)$ kovarianciák az (4–1)-ben adottak. A varancia becslése

$V(\hat{\theta}) = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_j a_{j'} \hat{C}\left(\hat{t}_{j\pi}, \hat{t}_{j'\pi}\right)$,

a kovarianciák (4–2)-ben adott becslésével.

4.1.3. Az általános eset

Tegyük fel, hogy a θ paraméter több összeg függvénye:

$\theta = f\left(t_1, \ldots, t_q\right)$,

ahol most f tetszőleges valós függvény. A gyakorlatban sokszor előforduló példa két ismeretlen populációs összeg arányának becslése (pl. adott pártra szavazók aránya a biztos szavazók között):

$\theta = f\left(t_1, t_2\right) = \frac{t_1}{t_2}$.

Természetes módon adódik θ becslésére:

$$\hat{\theta} = f\left(\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi}\right).$$

A $\hat{\theta}$ statisztikai tulajdonságainak (torzítás, variancia) levezetése nem okoz gondot, ha f lineáris függvény (lásd az 4.1.2. alfejezetet). Ugyanakkor nemlineáris függvények esetén a becslés torzításának és varianciának gyakran nehézkes az egzakt megadása. A probléma gyakran alkalmazott megoldása a nemlineáris függvény lineáris közelítésén alapul. A közelítést a Taylor-sorfejtés lineáris szeletével végezzük.

Rögzítsük az U véges populációt, $|U| = N$. A $\hat{\theta}$ becslés az s minta függvénye, legyen ennek mérete $n(s)$. Tegyük fel, hogy a q-dimenziós Euklideszi téren értelmezett f valós függvény másodrendű parciális deriváltjai folytonosak a $t = (t_1, \ldots, t_q)$ pont $\hat{t}_\pi$-t tartalmazó nyílt környezetében. Ekkor a Taylor-képlet Lagrange-féle maradéktagos alakja:

$$f(\hat{t}_\pi) = f(t) + \sum_{j=1}^{q} a_j \left(\hat{t}_{j\pi} - t_j\right) + R_{n(s)}\left(\hat{t}_\pi, t\right),$$

itt az $n(s)$ index a mintanagyságtól való függést hangsúlyozza. A maradéktag kifejtése

$$R_{n(s)}\left(\hat{t}_\pi, t\right) = \sum_{j=1}^{p} \sum_{i=1}^{p} \frac{1}{2} \frac{\partial^2 f(t_0)}{\partial \hat{t}_{j\pi} \partial \hat{t}_{i\pi}} \left(\hat{t}_{j\pi} - t_j\right)\left(\hat{t}_{i\pi} - t_i\right),$$

ahol t_0 értéke $\hat{t}_\pi$ és t között van, továbbá

$$a_j \stackrel{def}{=} \frac{\partial f}{\partial \hat{t}_{j\pi}}\Big|_{(\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi}) = (t_1, \ldots, t_q)}. \tag{4–3}$$

A másod- és magasabbrendű tagokat elhagyva az alábbi közelítéshez jutunk:

$$\hat{\theta} \approx \hat{\theta}_0 \stackrel{def}{=} \theta + \sum_{j=1}^{q} a_j \left(\hat{t}_{j\pi} - t_j\right). \tag{4–4}$$

Elég nagy minták esetén a lineáris közelítés kielégítő eredményt ad, a közelítés kevésbé jól működik nagyon ferde eloszlású y esetén (Wolter, 1985, Särndal et al, 1992). Most vezessük be az alábbi rövidítést:

$$u_k \stackrel{def}{=} \sum_{j=1}^{q} a_j y_{jk}, \tag{4–5}$$

és legyen $\check{u}_k \stackrel{def}{=} u_k/\pi_k$. Ekkor a $V\left(\hat{\theta}\right)$ variancia közelítése (AV=*approximate variance*):

$$AV\left(\hat{\theta}\right) = V\left(\hat{\theta}_0\right) = V\left(\sum_{j=1}^{q} a_j \hat{t}_{j\pi}\right) = V\left(\sum_{s} \check{u}_k\right) = \sum\sum_{U} \triangle_{kl} \check{u}_k \check{u}_l. \tag{4–6}$$

Az u_k-k értéke az $a_1, \ldots, a_q$ mennyiségek függvénye, amiket viszont az ismeretlen $t_1, \ldots, t_q$ paraméterekkel definiáltunk. Tehát az u_k-k ismeretlen mennyiségek. Becsüljük az a_j-ket oly módon, hogy az őket definiáló (4–3) kifejezésbe az ismeretlen t_j-k helyébe azok π-becslését helyettesítjük. Ezen becslések felhasználásával definiáljuk az u_k-k becslését:

$$\hat{u}_k = \sum_{j=1}^{q} \hat{a}_j y_{jk}. \tag{4–7}$$

Így végül eljutunk a keresett $V\left(\hat{\theta}\right)$-becsléshez:

$$\hat{V}\left(\hat{\theta}\right) = \sum \sum_{s} \check{\triangle}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}. \tag{4–8}$$

Általános esetben $\hat{V}\left(\hat{\theta}\right)$ $V\left(\hat{\theta}\right)$-nak nem torzítatlan becslése. Ugyanakkor legtöbbször az $\hat{u}_k$-k konzisztens becslések az u_k-kra, és - mivel konzisztens becslés függvénye általában maga is konzisztens - legtöbb esetben $\hat{V}\left(\hat{\theta}\right)$ is konzisztens becslés $AV\left(\hat{\theta}\right)$-ra. $\hat{V}\left(\hat{\theta}\right)$-nak $V\left(\hat{\theta}\right)$ becsléseként történő használatát $V\left(\hat{\theta}\right)$ és $AV\left(\hat{\theta}\right)$ elég nagy minták esetén várható kis különbsége indokolja.

Az ismertetett elsőrendű közelítés nagy minták esetén szimulációk és valós felmérések tapasztalatai szerint legtöbbször kielégítő eredményt ad (Wolter, 1985, Särndal et al, 1992). Ugyanakkor mások megkérdőjelezik a módszer feltétel nélkül történő használatának megalapozottságát, elsősorban a $\hat{\theta}_0$-nak a $\hat{\theta}$-hoz történő konvergenciájának (midőn $n(s) \to \infty$), illetve a konvergencia sebességének feltételeit elemezve (Wolter, 1985).

4.1.4. Példa: az arány becslése

Térjünk vissza a korábban említett példára: adjunk becslést a linearizáció módszerével két ismeretlen populációs összeg arányára, R-re:

$$R = \frac{t_y}{t_z} = \frac{\sum_U y_k}{\sum_U z_k}.$$

Az adódó nemlineáris becslés:

$$\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}.$$

A szükséges parciális deriváltak:

$$\frac{\partial \hat{R}}{\partial \hat{t}_{y\pi}} = \frac{1}{\hat{t}_{z\pi}}; \; \frac{\partial \hat{R}}{\partial \hat{t}_{z\pi}} = -\frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}^2}.$$

Ebből

$$a_1 = \frac{\partial \hat{R}}{\partial \hat{t}_{y\pi}}\Big|_{(t_y,t_z)} = \frac{1}{t_z};\ a_2 = \frac{\partial \hat{R}}{\partial \hat{t}_{z\pi}}\Big|_{(t_y,t_z)} = -\frac{t_y}{t_z^2}.$$

Most (4–5) felhasználásával:

$$u_k = a_1 y_k + a_2 z_k = \frac{1}{t_z}\left(y_k - R z_k\right),$$

ebből (4–7)-et alkalmazva:

$$\hat{u}_k = \frac{1}{\hat{t}_{z\pi}}\left(y_k - \hat{R} z_k\right).$$

Mindezek felhasználásával kapjuk

4.1. Tétel. *Linearizációt alkalmazva, az* $\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}$ *közelítése:*

$$\hat{R} \approx \hat{R}_0 = R + \frac{1}{t_z}\sum_s \frac{y_k - R z_k}{\pi_k}.$$

A $\hat{R}$ *becslés az* R *közelítőleg torzítatlan becslése, varianciája (4–6)-ból megközelítőleg:*

$$AV\left(\hat{R}\right) = V\left(\hat{R}_0\right) = \frac{1}{t_z^2}\sum\sum_U \triangle_{kl}\frac{y_k - R z_k}{\pi_k}\frac{y_l - R z_l}{\pi_l},$$

míg (4–8)-ból a variancia becslése:

$$\hat{V}\left(\hat{R}\right) = \frac{1}{\hat{t}_{z\pi}^2}\sum\sum_s \check{\triangle}_{kl}\frac{y_k - \hat{R} z_k}{\pi_k}\frac{y_l - \hat{R} z_l}{\pi_l}.$$

4.2. Ismételgető technikák

Ez a fejezet néhány, a becslési variancia közelítésére alkalmas módszert mutat be. Gyakran ismételgető technikáknak (*replication techniques*) is nevezik őket, mert a mintavétel többszöri (kvázi-)megismétlésén alapulnak. Előnyük, hogy problémák széles körére alkalmazhatók, ezenkívül - szemben a korábban látott módszerekkel - nem igénylik a másodrendű kiválasztási valószínűségek, a π_{kl}-ek potenciálisan nehéz meghatározását.

Hátrányuk, hogy az ezen technikák segítségével kapott variancia-becslések bonyolultak, tulajdonságaik legtöbbször nehezen levezethetők, sokszor csak szimulációs teszteken alapuló eredmények állnak rendelkezésre. Az ismételgetős technikák komputer-intenzív módszerek, kivitelezésük nagy mennyiségű számítást igényel.

Az alábbi eredményeket a következő alfejezetekben gyakran használjuk majd. A becslési variancia és az almintákból kapott becslések változékonyságának összefüggését

írják le, ezt a kapcsolatot használják ki az ismételgető technikák a variancia becslésének megadására.

Legyen $\hat{\theta}$ a θ paraméter s mintából kapott becslése. Vegyük az s minta A számú almintáját, mindegyikből számítsuk ki a θ becslését, legyenek ezek a statisztikák $\hat{\theta}_1, \ldots, \hat{\theta}_a, \ldots, \hat{\theta}_A$. Átlagukként előállítható a következő alternatív becslés:

$\hat{\theta}^* = \frac{1}{A} \sum_{a=1}^{A} \hat{\theta}_a$.

Definiáljuk az alábbi két statisztikát:

$$\hat{V}_1 = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_a - \hat{\theta}^*\right)^2, \tag{4–9}$$

$$\hat{V}_2 = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_a - \hat{\theta}\right)^2. \tag{4–10}$$

Az (4–9)-ben adott statisztika $V\left(\hat{\theta}^*\right)$ becsléseként történő alkalmazása az alábbiakkal támasztható alá:

4.2. Állítás. $E\left(\hat{V}_1\right) =$
$= V\left(\hat{\theta}^*\right) - \frac{1}{A(A-1)} \sum_{a=1}^{A} \sum_{\substack{b=1 \\ b \neq a}}^{A} C\left(\hat{\theta}_a, \hat{\theta}_b\right) + \frac{1}{A(A-1)} \sum_{a=1}^{A} \left[E\left(\hat{\theta}_a\right) - E\left(\hat{\theta}^*\right)\right]^2$.

A felbontás szerint, amikor a $\hat{\theta}_a$-k párontként korrelálatlanok, azonos várható értékkel, a $\hat{V}_1$ torzítatlan becslés $V\left(\hat{\theta}^*\right)$-ra. A (4–10)-ben adott statisztika, mint $V\left(\hat{\theta}^*\right)$-becslés pedig a következő meggondolással indokolható:

$\sum_{a=1}^{A} \left(\hat{\theta}_a - \widehat{\theta}\right)^2 = \sum_{a=1}^{A} \left(\hat{\theta}_a - \widehat{\theta}^*\right)^2 + A\left(\hat{\theta} - \widehat{\theta}^*\right)^2$,

amiből következik, hogy

$$\hat{V}_1 \leq \hat{V}_2, \tag{4–11}$$

azaz $\hat{V}_2$ elfogadható, ha a konzervatív (a varianciát inkább felülbecslő) becslést preferáljuk. Ha feltesszük, hogy $V\left(\hat{\theta}^*\right)$ és $V\left(\hat{\theta}\right)$ közel azonosak, akkor mindkét becslés alkalmazható $V\left(\hat{\theta}\right)$ becslésére.

Az alább következő variancia-becslések mindegyike az (4–9)-ben, vagy az (4–10)-ben adotthoz hasonló alakban áll elő. A sorra vett legtöbb esetben nem teljesül, hogy a $\hat{\theta}_a$-k párontként korrelálatlanok, így mindkét statisztika torzított becslést ad.

4.2.1. Random csoportok

Független random csoportok

Célunk az U véges populáció θ paraméterének becslése. A becslés a következő algoritmust követi:

(i) Az s_1 minta kiválasztása a $p(\cdot)$ elrendezésnek megfelelően.

(ii) Az s_1 visszatétele után az s_2 minta kiválasztása, megint a $p(\cdot)$ elrendezésnek megfelelően.

(iii) Az első két lépés ismétlése addig, amíg az $s_1, \ldots, s_A$ mintákat ki nem választjuk. Ezeket a mintákat nevezzük random csoportoknak.

(iv) Az A random csoportból ugyazon becslő függvényt alkalmazva θ becsléseként a $\hat{\theta}_a$, $a = 1, \ldots, A$ becsléseket kapjuk.

Definiáljuk θ-ra a következő becslést:

$$\widehat{\theta}^* \stackrel{def}{=} \sum_{a=1}^{A} \hat{\theta}_a / A.$$

Ekkor $V\left(\widehat{\theta}^*\right)$-ra az (4–9)-hez hasonló torzítatlan becslés adható:

$$\hat{V}\left(\widehat{\theta}^*\right) = \sum_{a=1}^{A} \left(\hat{\theta}_a - \widehat{\theta}^*\right)^2 / A\,(A-1).$$

A $\widehat{\theta}^*$ statisztika a θ becsléseként használatos, míg a becslés $V\left(\widehat{\theta}^*\right)$ varianciájának $\hat{V}\left(\widehat{\theta}^*\right)$ az un. *random csoport becslése.* A legtöbb alkalmazásban θ megegyezik a $\hat{\theta}_a$-k közös várható értékével, μ-vel, így

$$E\left(\widehat{\theta}^*\right) = \mu = \theta,$$

vagy legalábbis $\widehat{\theta}^*$ aszimptotikusan torzítatlan becslés θ-ra, és torzítása nagy minták esetén elhanyagolható.

Kézenfekvő lenne egy másik, alternatív módszerként θ-t az összevont mintából becsülni. A $\hat{\theta}_a$-khoz tartozó becslő függvényt az $s = \cup_{a=1}^{A} s_a$ összevont mintára alkalmazva kapjuk $\hat{\theta}$-t. $\hat{\theta}$ varianciájának becslésére alkalmazzuk a korábban $V\left(\widehat{\theta}^*\right)$ becsléseként definiált statisztikát:

$$\hat{V}_1\left(\hat{\theta}\right) \stackrel{def}{=} \sum_{a=1}^{A} \left(\hat{\theta}_a - \widehat{\theta}^*\right)^2 / A\,(A-1)\,.$$

Nemlineáris becslések esetén $\hat{\theta}$ és $\widehat{\theta^*}$ általában nem egyeznek meg, így a $\hat{V}_1\left(\hat{\theta}\right)$ nem torzítatlan becslés $V\left(\widehat{\theta^*}\right)$-ra. Ugyanakkor $\hat{V}_1\left(\hat{\theta}\right)$ torzítása általában elhanyagolható nagyságrendű (Wolter, 1985).

Egy másik, a (4–10)-hez hasonló becslés a

$$\hat{V}_2\left(\hat{\theta}\right) \stackrel{def}{=} \sum_{a=1}^{A}\left(\hat{\theta}_a - \widehat{\theta}\right)^2 / A\left(A-1\right).$$

(4–11)-ből következően

$$\hat{V}_1\left(\hat{\theta}\right) \leq \hat{V}_2\left(\hat{\theta}\right),$$

azaz $\hat{V}_2\left(\hat{\theta}\right)$ elfogadható, ha a konzervatív becslést preferáljuk. Ugyanakkor belátható, hogy általában $\hat{V}_1\left(\hat{\theta}\right)$ és $\hat{V}_2\left(\hat{\theta}\right)$ különbsége kicsi. Wolter (1985) szerint nehezen eldönthető, melyik becslés ajánlható inkább; mindenesetre megmutatható, hogy $\hat{V}_1\left(\hat{\theta}\right)$ legfeljebb akkora torzításhoz vezet, mint $\hat{V}_2\left(\hat{\theta}\right)$, és varianciájuk megegyezik.

Alkalmazás. A független random csoportok módszere a gyakorlatban igen költséges megoldás, hiszen több minta kiválasztását igényli. A kiválasztott csoportok A számának elég nagynak kell lenni a becslés kis varianciája érdekében, ám A-t gyakorlati okokból csupán 5-10 körüli értéknek választják[5]. Ezen kívül a mintánkénti becslések függetlenségének a követelménye nehezen teljesíthető adott kérdezői apparátus, adott adatfeldolgozási eljárás mellett.

Nem független random csoportok

A nem független random csoportok módszere egyetlen minta esetére adaptálja a korábbi eredményeket, a minta egy partícióját használja fel a variancia becslésére.

Vegyük a $p(\cdot)$ elrendezéssel kapott s minta egy partícióját: $s = \cup_{a=1}^{A} s_a$, az s_a részmintákat nevezzük random csoportoknak. Az s felosztását úgy végezzük, hogy a kapott random csoportokhoz lényegében ugyanaz a $p(\cdot)$ elrendezés tartozzon. Ez utóbbi feltétel sok esetben nehezen teljesíthető. Egyszerűbb példák: ha $p(\cdot)$ SI elrendezés, a random csoportok kiválasztása is egyszerű véletlen visszatevés nélküli mintavétellel megy; SY elrendezés esetén az s is szisztematikusan kerül felosztásra; rétegzés esetén rétegenként az

[5] A független random csoportok módszerét használta pl. a KSH 1973-ban végzett jövedelem-felmérése, ők öt csoportot választottak.

eredeti $p_h(\cdot)$, $(h = 1 \dots H)$ elrendezéssel választjuk ki megfelelő arányban az s_a-k elemeit, így mindegyik random csoport maga is rétegzett lesz; ha $p(\cdot)$ többlépcsős mintavétel, akkor az első lépcső csoportjait osztjuk A random csoportba az első lépcső $p_I(\cdot)$ elrendezésének megfelelően.

Hasonlóan a független random csoportok esetéhez, jelölje $\hat{\theta}$ a θ paraméter s-ből számolt becslését, míg $\hat{\theta}_a$ az s_a mintából ugyanazon becslő függvénnyel számolt becslését, a $\hat{\theta}_a$-król feltesszük, hogy torzítatlan vagy legalábbis közel torzíttalan becslések. Legyen $\widehat{\theta}^* \stackrel{def}{=} \sum_{a=1}^{A} \hat{\theta}_a / A$. Becsüljük a $V\left(\widehat{\theta}\right)$-t az alábbi statisztikával:

$$\hat{V}_1\left(\widehat{\theta}\right) = \sum_{a=1}^{A} \left(\hat{\theta}_a - \widehat{\theta}^*\right)^2 / A\,(A-1)\,,$$

vagy az alábbi módon:

$$\hat{V}_2\left(\widehat{\theta}\right) = \sum_{a=1}^{A} \left(\hat{\theta}_a - \widehat{\theta}\right)^2 / A\,(A-1)\,.$$

A két statisztika képzése most is az (4–9) és (4–10) formulákhoz hasonló, $\hat{V}_1\left(\hat{\theta}\right) \leq \hat{V}_2\left(\hat{\theta}\right)$ itt is fennáll. Mivel a $\hat{\theta}_a$-k nem függetlenek, a $\hat{V}_1\left(\widehat{\theta}\right)$ nem torzítatlan becslés sem $V\left(\widehat{\theta}\right)$-ra, sem $V\left(\widehat{\theta}^*\right)$-ra. Az alábbi tétel $\hat{V}_1\left(\widehat{\theta}\right)$-t jellemzi:

4.3. Tétel. *Legyen* $E\left(\hat{\theta}_a\right) = \mu_a$, *legyen* $E\left(\widehat{\theta}^*\right) = \sum_{a=1}^{A} \mu_a / A \stackrel{def}{=} \bar{\mu}$.

Ekkor a $V\left(\widehat{\theta}^*\right)$ *becslésének várható értékére:*

$$E\left(\hat{V}_1\left(\widehat{\theta}\right)\right) = V\left(\widehat{\theta}^*\right) + \tfrac{1}{A(A-1)} \sum_{a=1}^{A} (\mu_a - \bar{\mu})^2 - 2 \sum_{a=1}^{A} \sum_{b>a}^{A} Cov\left(\hat{\theta}_a, \hat{\theta}_b\right) / \left(A\,(A-1)\right).$$

Ha a random csoportok elemszáma megegyezik, azaz $|s_a| = m$, $a = 1, \dots, A$, *akkor*

$$E\left(\hat{V}_1\left(\widehat{\theta}\right)\right) - V\left(\widehat{\theta}^*\right) = -Cov\left(\hat{\theta}_1, \hat{\theta}_2\right).$$

A tétel alapján elmondható, hogy elég nagy minta esetén $\hat{V}_1\left(\widehat{\theta}\right)$, mint $V\left(\widehat{\theta}^*\right)$-becslés torzítása elhanyagolható és negatív, mivel a $2 \sum_{a=1}^{A} \sum_{b>a}^{A} Cov\left(\hat{\theta}_a, \hat{\theta}_b\right) / \left(A\,(A-1)\right)$ mennyiség ilyenkor relatíve kicsi és negatív. Wolter (1985) szerint a torzítás csökken a csoportok számának (vagy méretének) a növelésével, és mértéke a legtöbb esetben elhanyagolható.

Alkalmazás. Mivel a módszer egyetlen minta elemeiből képez random csoportokat, költsége csupán viszonylag nagy számítási igényében jelentkezik.

4.2.2. Jackknife

A jackknife módszer alkalmazási területe igen széles. Első alkalmazásaiban bizonyos torzított becslés-osztályok torzításának csökkentésére használták, később javaslatok születtek variancia-becslési alkalmazásaira is. A következőkben a jackknife véges populáció feltételezése melletti variancia-becslési alkalmazására szorítkozunk.

Legyen $p(\cdot)$ közvetlen elemkiválasztásra épülő elrendezés, de nem rétegezés. A nem független random csoportok módszeréhez hasonlóan, osszuk a $p(\cdot)$ elrendezéssel kapott s mintát A számú nem független, m méretű csoportra ($m = n/A$), az 4.2.1. fejezetben leírt módon. Jelölje $\hat{\theta}$ a θ paraméter s-ből számolt becslését, míg $\hat{\theta}_{(a)}$ az a. csoport ($a = 1 \dots A$) elhagyásával kapott mintából ugyanazon becslő függvénnyel számolt becslést. Legyen

$$\hat{\theta}_a \stackrel{def}{=} A\hat{\theta} - (A-1)\,\hat{\theta}_{(a)}, \tag{4–12}$$

az a. pszeudoérték. Ekkor a θ paraméter jackknife-becslése:

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a, \tag{4–13}$$

míg a jackknife variancia-becslésre az (4–9) és (4–10) formulákhoz hasonlóan ismét két alternatíva írható fel:

$$\hat{V}_{JK1} = \frac{1}{A\,(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2, \tag{4–14}$$

$$\hat{V}_{JK2} = \frac{1}{A\,(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}\right)^2. \tag{4–15}$$

Mindkét statisztika használatos $V\left(\hat{\theta}\right)$ és $V\left(\hat{\theta}_{JK}\right)$ becslése is, (4–11)-ből következően fennáll a $\hat{V}_{JK1} \leq \hat{V}_{JK2}$ összefüggés. $\hat{V}_{JK1}$ (4.2)-ből következően torzítatlan becslés $V\left(\hat{\theta}_{JK}\right)$-ra, amennyiben a $\hat{\theta}_a$-k páronként korrelálatlan azonos várható értékű valószínűségi változók, ám ténylegesen legtöbbször nem azok.

4.4. Példa. *Vegyük azt az egyszerű esetet, mikor $\hat{\theta}$ π-becslés. Legyen pl. θ az összeg, legyen $\hat{\theta}$ az összeg π-becslése, legyen az elrendezés SI . Ekkor*

$\hat{\theta} = \hat{t}_\pi = N\overline{y}_s$,

$\hat{\theta}_{\pi(a)} = [N/(n-m)] \sum_{s-s_a} y_k = N\overline{y}_{s-s_a}$,

$\hat{\theta}_a = \frac{N}{m} \sum_{s_a} y_k = N\overline{y}_{s_a}$,

$\hat{\theta}_{JK} = N\overline{y}_s = \hat{\theta}$,

azaz a jackknife-becslés megegyezik a π-becslésssel, ezért az (4–14) és (4–15) formulák ekvivalensek, mégpedig az alábbi becslést adják:

$\hat{V}_{JK} = \frac{N^2}{A(A-1)} \sum_{a=1}^{A} \left(\overline{y}_{s_a} - \overline{y}_s\right)^2$.

Ebből a becslés torzítása (3.21) felhasználásával:

$E\left(\hat{V}_{JK}\right) - V\left(\hat{\theta}\right) = NS_{yU}^2$.

A jackknife variancia-becslés torzítása könnyen megszüntethető a $\hat{V}_{JK}$ statisztika $(1 - \frac{n}{N})$-nel, az un. véges populáció korrekciós együtthatóval (fpc: finite population correction) történő szorzásával.

A fenti klasszikus módszer közvetlen elemkiválasztásra épülő elrendezésre működik jól, de a rétegezés esetében a módszer egy másik variánsa alkalmazandó. Tegyük fel, hogy az s mintának a h. rétegbe tartozó része $(h = 1 \ldots H)$ A_h random módon csoportra van osztva. Összesen A csoport van, $A = \sum_{h=1}^{H} A_h$. Mint korábban, legyen $\hat{\theta}$ a θ paraméter s-ből számolt becslés. Jelölje $\hat{\theta}_{(ha)}$ az s mintából, a h. réteg a. csoportjának elhagyása után ugyanazon becslő függvénnyel számolt θ-becslést. Ekkor a $V\left(\hat{\theta}\right)$ jackknife variancia-becslése a következő:

$\hat{V}_{JK3} \stackrel{def}{=} \sum_{h=1}^{H} \left[(A_h - 1)/A_h\right] \sum_{a=1}^{A_h} \left[\hat{\theta}_{(ha)} - \hat{\theta}\right]^2$.

4.5. Példa. *Vegyük azt az esetet, amikor a rétegeken belül SI elrendezést alkalmazunk, és a θ összegre π-becslést adunk: $\hat{\theta} = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$. Ekkor a fentiek a következő alakban állnak elő:*

$\hat{\theta}_{(ha)} = N_1\bar{y}_{s_1} + \ldots + N_{h-1}\bar{y}_{s_{h-1}} + N_h\bar{y}_{s_h - s_{h_a}} + N_{h+1}\bar{y}_{s_{h+1}} + \ldots + N_H\bar{y}_{s_H}$,

$\hat{V}_{JK3} = \sum_{h=1}^{H} \left[N_h^2/(A_h(A_h-1))\right] \sum_{a=1}^{A_h} \left[\bar{y}_{s_{ha}} - \bar{y}_{s_h}\right]^2$.

Az $A_h = n_h$ esetben, tehát amikor egyelemű csoportokat hagyunk el, a becslési formula a következő alakra egyszerűsödik:

$\hat{V}_{JK3} = \sum_{h=1}^{H} N_h^2 S_{ys_h}^2/n_h$,

amit tagonként $\left(1 - \frac{n_h}{N_h}\right)$*-val, a véges populáció korrekciós együtthatóval szorozva éppen a Horvitz-Thomson becslést kapjuk.*

Kétlépcsős mintavételi elrendezés esetén a jackknife technikát az első lépcső elemeire alkalmazzuk. A 3.4.2. fejezet jelöléseivel, legyen s_I az első lépcsőben kapott minta, n_I számú csoporttal. Partícionáljuk random módon s_I-t A osztályra, mindegyik osztályban m csoporttal. Jelölje $\hat{\theta}$ a θ paraméter s-ből számolt becslését, legyen $\hat{\theta}_{(a)}$ az a becslés, amit az a. osztály elhagyásával kapott mintából számolunk. A θ paraméter $\hat{\theta}_{JK}$ jackknife-becslése, továbbá a $V\left(\hat{\theta}\right)$ ill. $V\left(\hat{\theta}_{JK}\right)$ varianciák $\hat{V}_{JK1}$ és $\hat{V}_{JK2}$ becslései közvetlenül kaphatók, az (4–12)-(4–15) formulákkal megegyezően.

Alkalmazás. Mint a fenti példák mutatták, a π-becslés és általában a lineáris becslések esetén a jackknife variancia-becslés tulajdonságai kedvezőek és könnyen levezethetők. Ugyanakkor a jackknife előnye nem itt jelentkezik, hiszen a lineáris becsléseknél a tradícionális formulák is jól működnek. A jackknife-módszer elsődleges alkalmazási területe a más módszerekkel nehezen megközelíthető nemlineáris becslések variancia-becslése. A komplex, nemlineáris becslések esetén történő jackknife-használatra szimulációs teszteken kívül kevés egzakt eredmény létezik, tulajdonképpen a jackknife nemlineáris becslésekre történő alkalmazását a lineáris esetben tapasztalt kedvező működése igazolhatja.
A jackknife alkalmazásakor döntést kell hozni a csoportok A számának értékéről. Az A növelése a becslés varianciájának csökkenését eredményezi, ugyanakkor lényegesen növelheti a számítási igényt. E két szempont közötti kompromisszum a konkrét alkalmazás ismeretében hozható meg.

4.2.3. Bootstrap

A jackknife-hoz hasonlóan a bootstrap alkalmazásai is szélesebb területet ölelnek fel. Több lehetséges variancia-becslési alkalmazása közül itt csak egyet említünk. Tegyük fel, hogy az $s \subset U$ mintához a $p(\cdot)$ visszatevés nélküli elrendezés alkalmazásával jutottunk. A $V\left(\hat{\theta}\right)$ becslését keressük, az eljárás a következő:

(i) Az s segítségével egy U^* mesterséges populációt kunstruálunk.

(ii) U^*-ból a $p(\cdot)$ elrendezés mellett független minták egy sorozatát választjuk:

$s_1, \ldots, s_a, \ldots, s_A$. A függetlenség érdekében minden kiválasztott mintát visszateszünk U^*-ba a következő minta kivétele előtt. Az s_a-kból az $\hat{\theta}$-hoz hasonló módon $\hat{\theta}_a^*$ becslést számítunk.

(iii) $V\left(\hat{\theta}\right)$ becslését a $\hat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a^*$ átlagot felhasználva kapjuk:

$\hat{V}_{BS} = \frac{1}{A-1}\sum_{a=1}^{A}\left(\hat{\theta}_a^* - \hat{\theta}^*\right)^2$.

4.6. Megjegyzés. *Más alkalmazásokban nem a variancia becslése a cél, hanem a $\hat{\theta}$ becslés konfidencia-intervallumának megadása. Ekkor a $\hat{\theta}_a^*$-k empirikus eloszlásának felhasználásával közvetlenül juthatunk a megbízhatósági intervallumhoz.*

Az U^* populáció konstrukciója úgy történhet, hogy a $k \in s$ elemet $1/\pi_k$ példányban vesszük fel, úgy, hogy minden példányhoz azonosan az eredeti y_k érték tartozzon. (Az egyszerűség kedvéért feltéve, hogy $1/\pi_k$ egész szám.)

4.3. A módszerek összevetése

Wolter (1985) a következő szempontokat ajánlja a variancia-becslési módszerek összevetésére:

- a becslés statisztikai tulajdonságai: a becslés torzítása; varianciája; milyen megbízhatósággal adható segítségével a becslésre konfidencia-intervallum ($\sim$ ismert-e a $\frac{\hat{\theta}-\theta}{\sqrt{\hat{V}(\hat{\theta})}}$ statisztika aszimptotikus eloszlása).
- adminisztratív szempontok: a szükséges számítások mennyisége, szakember- és időigénye. Egy átlagos felmérés kereteit messze meghaladja a minden célparamétertípusra külön történő nagy pontosságú becslés konstruálása. Ezért bár kevésbé pontos, de egyszerű, azaz hatékony módszerek a kívánatosak.
- a módszer flexibilitása: az általa lefedett elrendezések és $\hat{\theta}$ becslések köre.

Wolter az itt tárgyaltak közül a linearizációt, a random csoportok ill. a jackknife módszerét veti össze, öt különböző tanulmány szimulációs eredményeire hivatkozva. Konklúziója szerint a linearizáció teljesít a legjobban a torzítás és a variancia szempontjából,

míg a random csoportok és a jackknife előnyösebb, ha a konfidencia-intervallum konstruálhatóságából indulunk ki. Flexibilitás szempontjából nincs lényeges különbség a módszerek között. A számítási költségeket tekintve a jackknife a legdrágább, a többi módszer gyakorlatilag azonos költséget igényel.

5. Felhasználás: optimális elrendezések

A korábbi fejezetekben különböző elrendezésekhez adtuk meg a θ paraméter becslésének varianciáját, és a variancia becslését. A jelen fejezetben ezeknek az eredményeknek két fontos alkalmazási területét vizsgáljuk. Ez első alkalmazás a költséghatékony elrendezések problémája, amikor a költségek és a becslés varianciájával definiált hatásosság együttes figyelembevétel keresünk optimális elrendezést. A másik alkalmazási terület a gyakorlatban általában rendelkezésre álló segédinformációk kihasználásával foglalkozik, a becslések hatásosságának növelése érdekében. Teljességre nem törekedve, az alkalmazások közül csak néhány példát mutatunk.

5.1. Költséghatékonyság

A megengedettségi tételek között szereplő (3.18) tétel tulajdonképpen az itt sorra kerülő problémák speciális esetére ad választ. Amikor ugyanis a felmérés költsége minden $k \in U$-ra azonos, akkor az adott várható mintanagyság mellett történő megengedett elrendezés keresése egy adott várható költség mellett történő keresésnek felel meg. Az általános, k-nként változó költséget megengedő költséghatékonysági optimalizálási problémák tipikus kérdésfeltevése lehet a következő: Hogyan allokáljuk a mintát rétegezés esetén a rétegek között ahhoz, hogy a költségek rögzített felső korlátja mellett leghatásosabb becslést konstruálhassunk? Fordítva: a kívánt becslési megbízhatóság mely legolcsóbb allokációval érhető el? Más elrendezések esetén hasonlóan állnak elő költséghatékonysági problémák, nyilván ezek célváltozói nem a rétegenkénti mintanagyságok, ehelyett pl. kétlépcsős elrendezésnél az $\frac{n_I}{N_I}$ érték, a mintába kerülő csoportok optimális aránya a meghatározandó. Az alábbiakban a rétegezés esetét vizsgáljuk.

Tegyük fel, hogy adottak a rétegeken belül alkalmazott elrendezések. Az összeget

kívánjuk becsülni, és erre a π-becslést használjuk. Célunk az n elemű minta allokálása, azaz az n_h, $h = 1 \dots H$ mintanagyságok meghatározása. Tegyük fel, hogy a rögzített elrendezés olyan, hogy

$$V_{ST}\left(\hat{t}_\pi\right) = \sum_{h=1}^{H} A_h/n_h + B = V,$$

ahol A_h *és* B nem függ n_h-tól. Ilyen elrendezés a (3.32) példa alapján az STSI is. Tegyük fel továbbá, hogy a felmérés költségeire az alábbi összefüggés teljesül:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h, \tag{5–1}$$

azaz adott a c_0 allokációtól független alapköltség, és c_h a költsége egyetlen h. rétegbeli elem mintába vonásának. A Cauchy-Schwartz egyenlőtlenségből

$$\left(\sum_{h=1}^{H} A_h/n_h\right)\left(\sum_{h=1}^{H} n_h c_h\right) \geq \left[\sum_{h=1}^{H} \left(A_h c_h\right)^{1/2}\right]^2,$$

itt az egyenlőség feltétele:

$$\left(\frac{n_h c_h}{A_h/n_h}\right)^{1/2} = \text{konstans}.$$

Ennek segítségével a példaként említett két kérdés mindegyike választ nyer.

5.1. Állítás. *Adott C költség mellett a varianciát minimalizáló elosztás:*

$$n_h = (C - c_0)\left(A_h/c_h\right)^{1/2} / \sum_{h=1}^{H} \left(A_h c_h\right)^{1/2}.$$

Adott V varianciát legkisebb költséggel megvalósító allokáció:

$$n_h = \left(A_h/c_h\right)^{1/2} \left(\sum_{h=1}^{H} \left(A_h c_h\right)^{1/2}\right) / \left(V - B\right).$$

5.2. Megjegyzés. *Itt a mintanagyságra nem tettünk megkötést, ugyanakkor nyilván az $n_h \leq N_h$ feltétel teljesülése elengedhetetlen. A korlátozó feltételekkel bővített problémára később visszatérünk.*

Alkalmazás. STSI elrendezés esetén rögzített C költség mellett az optimális allokáció előáll, mint

$$n_h = (C - c_0)\,\frac{N_h S_{yU_h}/c_h^{1/2}}{\sum_{h=1}^{H} N_h S_{yU_h} c_h^{1/2}}, \tag{5–2}$$

azaz adott réteg esetén minél kisebb a c_h költség, annál nagyobb mintát érdemes belőle venni, illetve ha nagy a rétegbeli szórás, nagy az optimális mintanagyság is.

Általánosabb optimalitási problémák bevezetése előtt definiáljuk a következő konvex programozási feladatot.

$$f(x_1, \ldots, x_J) = \sum_{j=1}^{J} \frac{q_j}{x_j} \to \min$$

az alábbi feltételek mellett:

$$\sum_{j=1}^{J} Q_{ij} x_j \leq Q_{i0},\ i = 1, \ldots, I$$

$$x_{j0} \leq x_j,\ j = 1, \ldots, J.$$

Itt $x_j \in \mathbf{R}$, míg a konstansokra: $q_j > 0$, $x_{j0} \geq 0$, $Q_{i0} > 0$, $Q_{ij} \geq 0$, $\forall i, j$. A programozási feladatnak létezik megoldása, méghozzá analitikusan megadható megoldása (Särndal et al, 1992). Több költséghatékonysági probléma vezethető vissza erre a programozási feladatra, ezáltal garantált a megoldhatóságuk. Az alábbiakban két ilyen problémát vezetünk elő.

1. probléma Adottak az $y_1, \ldots, y_i, \ldots, y_I$ jellemzők, a $t_i = \sum_U y_{ik}$ összegeket kívánjuk becsülni, π-becsléssel. Az elrendezés STSI, a (3.32) alapján az i. összeg π-becslése: $\hat{t}_{i\pi} = \sum_{h=1}^{H} N_h \bar{y}_{is_h}$, ahol $\bar{y}_{is_h}$ az y_i mintaátlaga a h. rétegben. Az i. variancia:

$$V_i = V_{STSI}\left(\hat{t}_{i\pi}\right) = B_i + \sum_{h=1}^{H} \frac{A_{ih}}{n_h},$$

ahol $B_i = -\sum_{h=1}^{H} N_h S_{ih}^2$, és $A_{ih} = N_h^2 S_{ih}^2$, S_{ih}^2 az y_i varianciája a h. rétegben. A becslések megbízhatóságával szemben támasztott követelményünk:

$$V_i \leq V_{0i},\ i = 1, \ldots, I.$$

A rétegenkénti mintanagyságokra vonatkozó megkötés:

$$1 \leq n_h \leq N_h.$$

Feladat: a varianciákra és a rétegenkénti mintanagyságokra vonatkozó feltételeket teljesítő, a (5–1)-ben adott költségfüggvényt minimalizáló allokáció megtalálása.
A probléma közvetlenül visszavezethető a programozási feladatra, a
$(J, j, q_j, x_j, Q_{ij,} Q_{i0}) = (H, h, c_h, 1/n_h, A_{ih}, V_{0i} - B_i)$ szereposztással.

2. probléma Az y jellemző összegének π-becslését adjuk, STSI elrendezés mellett, de nem csak a teljes populációra vonatkozó, hanem a rétegenkénti $\hat{t}_{i\pi}$ becslés kívánt

megbízhatóságát is elérő allokációt keresünk, amely a (5–1)-ben adott költségfüggvény mellett a legolcsóbban megvalósítható. Ennek megfelelően a korlátozó feltételek:

$$V_{STSI}\left(\hat{t}_\pi\right) \le V_0,$$

$$N_h^2\left(\tfrac{1}{n_h} - \tfrac{1}{N_h}\right) S_{yU_h}^2 \le V_{0h},\ h = 1, \dots, H,$$

ahol V_0, V_{0h} konstansok.

Algebrai átalakítások után a probléma visszavezethető a programozási feladatra, a $(J, j, x_j) = (H, h, 1/n_h)$ megfeleltetéssel, ill. a $Q_{1j} = N_h^2 S_{yU_h}^2$, $Q_{10} = V_0 + \sum_{h=1}^{H} N_h S_{yU_h}^2$, ill. $i \ge 2 : Q_{ij} = 1\ (i = j)$, $Q_{ij} = 0\ (i \ne j)$, $Q_{i0} = V_{0h}/N_h^2 S_{yU_h}^2 + 1/N_h$ szereposztással.

5.3. Megjegyzés. *A programozási probléma a valós számhalmaz egy konvex halmazán keresett optimumot, így az $1/n_h$ értékeknek az x_j célváltozónak történő megfeleltetése nem teljesen korrekt megoldás. A megoldásul kapott x_j-hez legközelebb eső $1/n_h$ választással azonban kielégítő közelítéssel megoldhatjuk az allokációs feladatot.*

5.2. A segédinformáció kihasználása

A különböző elrendezések mellett kapott becslési varianciákkal kapcsolatos korábbi eredmények alkalmazhatók általánosabb modellek esetére is, amikor az ismeretlen, y_k-k függvényeként előálló paraméter becslését bizonyos segédinformációk segítik. A fentiekben már említettünk olyan eseteket, amikor az előre ismert segédinformáció kihasználásával hatásosabb becslésekhez juthatunk. A Poisson-mintavétel esetén az elvi, de a gyakorlatban nem megvalósítható optimális elrendezés kiválasztási valószínűségeire adott megoldás $\pi_k = ny_k/\sum_U y_k$ volt. Ehelyett olyan x_k segédinformáció használatát ajánlottuk, ami hozzávetőlegesen arányos y_k-val: $y_k/x_k \approx c$. Ebben az esetben a $\pi_k = nx_k/\sum_U x_k$ $(x_k \le \sum_U x_k/n)$ választással a kapott π-becslés varianciája kicsi lesz. Egy másik segédinformáció-alkalmazás a szisztematikus mintavétel esete, ahol az y_k-val hozzávetőlegesen arányos x_k segédinformáció szerint rendezett populáción lényegesen kisebb az összegbecslés varianciája.

Harmadik példa a rétegezés volt, ami maga is bizonyos segédinformáció meglétét előfeltételezi: az U minden eleméről tudnunk kell, hogy mely réteghez tartozik. További

segédinformációt kihasználva rétegezés esetén a 5.1. fejezet eredményeit alkalmazva definiálhatunk optimális, vagy közel optimális allokációt. Tegyük fel, hogy az x segédinformáció y-nal jól korrelál, és rétegenként ismert az x S_{xU_h} szórása. A (5–2) formulát alkalmazva y helyett x-re, és konstans c_h költséget feltételezve az x-optimális allokáció előáll, mint

$$n_h = n\frac{N_h S_{xU_h}}{\sum_{h=1}^{H} N_h S_{xU_h}}.$$

Tökéletes, vagy közel tökéletes korreláció esetén optimális, vagy majdnem optimális allokációt kapunk. Ha a segédinformációról a rétegenkénti összegek értéke áll rendelkezésre, az összeggel arányos allokáció alkalmazható:

$$n_h = n\frac{\sum_{U_h} x_k}{\sum_U x_k},$$

feltéve, hogy x pozitív. Az allokáció akkor van közel az optimálishoz, ha y és x jól korrelál, és x rétegenkénti összege arányos annak rétegenkénti szórásával.

5.2.1. A hányados-becslés

A linearizáció alkalmazásaként, az arány (4.1)-ben adott becsléséből vezethető le az összeg gyakran használt becslése, a hányados-becslés (*ratio-estimator*). A hányados-becslés lényeges előnye, hogy sok esetben hatásosabb becslést biztosít, mint a π-becslés, viszont bizonyos előre adott segédinformáció meglétét feltételezi.

A $t_y = \sum_U y_k$ összegre adunk becslést. Tegyük fel, hogy a $t_z = \sum_U z_k$ összeg értéke ismert. A $t_y = t_z \frac{t_y}{t_z}$ átalakításból kiindulva definiáljuk az alábbi becslést:

$$\hat{t}_{yra} = t_z \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}.$$

A (4.1)-et felhasználva adódnak:

$$AV\left(\hat{t}_{yra}\right) = \sum\sum_U \triangle_{kl} \frac{y_k - Rz_k}{\pi_k} \frac{y_l - Rz_l}{\pi_l}.$$

Az $AV\left(\hat{t}_{yra}\right)$ értéke nulla, ha $y_k - Rz_k$ nulla minden k-ra. Ha ez nem is érhető el a gyakorlatban, mindenesetre igen kicsi a variancia, ha ezek a különbségek kicsik, azaz az (y_k, z_k) pontok egy origón áthaladó R meredekségű egyeneshez közel helyezkednek el.

Alkalmazás. SI elrendezés esetén a hányados-becslés közelítő varianciája felírható, mint

$$AV_{SI}\left(\hat{t}_{yra}\right) = N^2 \frac{1-f}{n}\left(S_{yU}^2 + R^2 S_{zU}^2 - 2RS_{zyU}\right),$$

ahol $R = \frac{t_y}{t_z}$, S_{yU}^2 és S_{zU}^2 a populációs varianciák, S_{zyU} a populációs kovariancia. Az összeg π-becslése SI elrendezés esetén $\hat{t}_{y\pi} = N\bar{y}_s$, ennél a becslésnél nem használjuk ki a z_k-król rendelkezésre álló populációs, ill. mintabeli információkat. A hányados-becslés gyakran hatásosabb módszer, mint a π-becslés. Az $AV_{SI}\left(\hat{t}_{yra}\right) \leq V_{SI}\left(\hat{t}_{y\pi}\right)$ egyenlőtlenség teljesülésének szükséges és elégséges feltétele:

$$r \geq \frac{1}{2}\left(\frac{cv_{zU}}{cv_{yU}}\right),$$

ahol r a z és y korrelációja, $cv_{zU} = S_{zU}/\bar{z}_U$ és $cv_{yU} = S_{yU}/\overline{y}_U$ az un. variációs együtthatók. Ha ezek értéke közel azonos, akkor 0,5-öt meghaladó értékű korreláció esetén a segédinformáció használata hatásosabb becslést eredményez.

Alkalmazás. Az alábbi példa tanulsága az, hogy bizonyos esetekben az ismert paramétert is érdemesebb a mintából becsülni, mert így hatásosabb becsléhez juthatunk. A populációs átlag, $\bar{y}_U = \frac{1}{N}\sum_U y_k$ becslése visszavezethető az arány becslésére. Ha N értéke ismert, a π-becslés az $\widehat{\bar{y}}_{U\pi} = \frac{1}{N}\sum_s \frac{y_k}{\pi_k}$ alakban áll elő, a (3.3) tétel alapján

$$V\left(\widehat{\bar{y}}_{U\pi}\right) = \frac{1}{N^2}\sum\sum_U \triangle_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Másik lehetőség, hogy az N-et is a mintából becsüljük, akár ismert az értéke, akár nem. Az (4.1) tételt felhasználva

$$\widehat{\bar{y}}_{Ura} = \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} = \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}},$$

$$AV\left(\widehat{\bar{y}}_{Ura}\right) = \frac{1}{N^2}\sum\sum_U \triangle_{kl}\frac{y_k-\bar{y}_U}{\pi_k}\frac{y_l-\bar{y}_U}{\pi_l}.$$

Az $AV\left(\widehat{\bar{y}}_{Ura}\right) < V\left(\widehat{\bar{y}}_{U\pi}\right)$ egyenlőtlenség teljesül, ha pl. az alábbi feltételek bármelyike teljesül:

(i) A mintanagyság nem rögzített, pl. BE vagy PO elrendezés esetén. Ilyenkor a hányados-becslésnek mind a nevezője, mind a számlálója a mintanagyságnak megfelelően változik, míg a π-becslésnek csak a számlálója reagál a mintanagyság változására, ezért az utóbbi varianciája nagyobb.

(ii) Az π_k-k nem korrelálnak az y_k értékekkel. Ekkor ha a mintába a nagy y_k értékű, de kis π_k-hoz tartozó elem bekerül, a π-becslés igen nagyra nő, míg a hányados-becslés számlálójának növekedését a nevező növekedése némileg kompenzálja.

A példák arra utalnak, hogy a hányados-becslés előnye egyfajta adaptivitás, a mintával szemben mutatott érzékenység.

A hányados-becslés kedvező tulajdonságait a linearizáció módszerét felhasználva támaszthattuk alá. A következőkben a linearizációt és más, korábban ismertetett variancia-becslési módszereket alkalmazva hasonlítunk össze néhány Magyarországon gyakran alkalmazott becslést és elrendezést.

6. Gyakorlati alkalmazás

A mintavételnek megfelelő variancia-becslési módszerek alkalmazásaként a mai, hazai felmérés-statisztikai praxis néhány valóban létező kérdésére kerestünk választ. Két mintavételi elrendezést vizsgáltunk. Az első esetben háztartásokat választunk visszatevés nélkül, egyszerű véletlen (SI) mintavétellel, és a mintába került háztartások minden tagját bevonjuk a mintába. A minta nagysága ekkor nem rögzített, és az elsőrendű kiválasztási valószínűségek azonosak: $\pi_k = \frac{n_I}{N_I}$, $\forall k \in U$, ahol n_I a mintába került háztartások előre rögzített száma, míg N_I a populáció háztartásainak számát jelöli. Ezt az elrendezést jelölje d_I. A másik esetben a kiválasztott háztartásból csak egy személy kerül a mintába, és a háztartáson belül szintén SI elrendezést alkalmazunk a személy kiválasztására. A minta n nagysága itt rögzített, és az elsőrendű kiválasztási valószínűségek a személyek háztartásának nagyságával fordítottan arányosak: $\pi_k = \frac{1}{|h_k|}\frac{n_I}{N_I}$, $\forall k \in U$, ahol h_k jelöli a k. elem háztartását. Az elrendezést jelöljük d_{II}-vel. Mindkét elrendezés lényegében megfeleltethető a mai hazai mintavételi gyakorlatban gyakran alkalmazott egy-egy eljárásnak, azzal a különbséggel, hogy a gyakorlatban a háztartások kiválasztása inkább több lépcsőben, és rétegezéssel történik, de a háztartások mintába kerülési esélye lényegében ott is azonos. A gyakorlatban a háztartások kiválasztására országos címlista áll rendelkezésre, a háztartáson belül személykiválasztásra pedig valamely, a véletlen kiválasztást imitáló módszert alkalmaznak, pl. azt a személyt választják, akinek a felkeresés időpontjához legközelebb esik a születésnapja.

Mindkét elrendezés esetén három különböző becslési módszer alkalmaztunk az összeg becslésére. Az első becslés (e_1) egyszerű π-becslés. A második becslés (e_2) a nagyon széles

körben alkalmazott un. utólagos rétegezéssel kapott becslés, amikoris a minta valamely ismérv szerinti megoszlását az ismert populációs megoszláshoz illesztik, súlyozással. Az utólagos rétegezés célja a becslés torzításának és varianciájának csökkentése. Vizsgálatunkban az életkort választottuk rétegezési ismérvnek, tehát a mintába került személyekhez tartozó súly a személy korcsoportjának populációbeli létszámának és ezen h. korcsoport mintabeli létszámának a hányadosa: $\frac{N_h}{n_h}$. Ez a becslési eljárás nem veszi figyelembe az elsőrendű kiválasztási valószínűségeket, így a d_{II} esetében alkalmazva nem feltétlenül ad torzítatlan becslést. Megjegyezzük, hogy ezt a problémát a felmérések kivitelezői általában nem veszik figyelembe. Példaként említjük a KSH egyik felmérését, amikor a mintát a d_{II}-höz hasonló elrendezés szerint vették, a becsléseket pedig az e_2-nek megfelelően, több ismérv szerinti utólagos rétegzéssel végezték, a mintába kerülési valószínűségek háztartásnagyságtól való függését figyelmen kívül hagyva (Egészségi Állapot Felvétel, 1994).

A harmadik becslés, e_3 éppen ez utóbbi probléma miatt került kipróbálásra, itt az utólagos rétegzést a kiválasztási valószínűségekkel kombinálva alkalmaztuk.. A h. korcsoportba tartozó személy súlya ekkor a következőképpen áll elő: $\frac{1}{\pi_k}\frac{N_h}{\sum_{s_h}\pi_k}$, ahol s_h a h. korcsoporthoz tartozó mintabeli személyek halmaza. Vagyis a súly a π-becslésnél alkalmazott $\frac{1}{\pi_k}$ tényező, és az utólagos rétegzés módosított súlyának szorzata, ahol a módosított súly nevezőjében a h. korcsoportnak nem a mintabeli gyakorisága, hanem π-becsléssel becsült populációs gyakorisága található. Az e_3 torzítatlan becslés, a hányados-becslés egyfajta továbbfejlesztésének tekinthető, mint az 1. melléklet (d_I, e_2) részének linearizációval kapcsolatos számításainál láthatjuk, tulajdonképpen h számú hányadosbecslés összegeként áll elő. Bár a becslést a hazai gyakorlatban nem alkalmazzák, nemzetközi szinten elfogadottnak tekinthető (Botman et al, 2000).

Kutatási kérdéseink a következők voltak:

(i) d_I és d_{II} összevetése. Melyik elrendezés jobb, azaz melyik esetben nagyobb a becslés varianciája, ha figyelembe vesszük a felmérés költségét is, és azonos költségű, ezért eltérő nagyságú minták mellett vizsgáljuk a két elrendezést?

(ii) e_1, e_2 és e_3 összevetése. Mekkora torzítást okoz e_2 esetén az mintába kerülési esélyek egyenlőtlenségének figyelembe nem vétele? Kisebb-e e_3 varianciája, mint e_1-é, aho-

gyan azt a hányados-becslésnél elmondottak alapján esetleg elvárhatnánk?

(iii) A variancia-becslési módszerek: a Horvitz-Thomson, a linearizáció, a nem független random csoportok, a jackknife és a bootstrap összevetése különböző stratégiák (d_I / d_{II}, ill. e_1 / e_2 / e_3 kombinációi) esetén. Mekkora a variancia-becslések torzítása? Milyen megbízhatóak, azaz mekkora a varianciájuk? A segítségükkel képzett adott szintű konfidencia-intervallum valóban az adott valószínűséggel fedi-e a becsült paramétert?

(iv) A fenti problémák eloszlás-érzékenysége. Az y jellemző összegének becslésének, illetve a variancia-becsléseknek a tulajdonságai lényegesen eltérnek-e különböző y-ok esetén?

A kérdések megválaszolására a Monte Carlo típusú szimulációs vizsgálatot találtuk alkalmasnak. A szimuláció során pszeudopopulációként a Tárki adatbázisát, a Magyar Háztartási Panel IV. hullám[6] felmérését használtuk. Az adatbázis 4397 személy adatait tartalmazza. A (*iv*) kutatási kérdés megválaszolása érdekében két jellemzőt vizsgáltunk. Az egyik jellemző (y_1) a havi nettó jövedelem volt, ennek populációs összegét becsültük. A másik jellemzőnek a politikai szimpátiát választottuk, a populációs összeg a biztosan az MSZP-re[7] szavazni szándékozók száma, itt a jellemző (y_2) 0/1 értékű indikátor-változó. Az adatbázist leszűkítettük azon személyekre, akik esetében mindkét információ szerepelt, így 1954 háztartás 3979 lakója alkotta a pszeudopopulációt.

A szimuláció során 500-szor egymás után, visszatevéssel választottunk mintát. A d_I esetében ezt a mintát $n_I = 150$ háztartás alkotta. Gyakorlati tapasztalatokra építve a felmérés költségfüggvényét a következő módon határoztuk meg: $c(s) = c_0 + n_I c_1 + n c_1$, ahol n_I a mintába került háztartások száma, n a mintabeli személyek várható száma. A két elrendezés ésszerű összevetését biztosítva a d_{II} mintanagyságát úgy határoztuk meg, hogy az a d_I várható költségével azonos költséggel járjon. Ennek megfelelően a d_{II}

[6] *Magyar Háztartás Panel IV. hullám* A felmérés adatait a TÁRKI Társadalomtudományi Adatbank bocsátotta rendelkezésünkre. A kutatást vezető intézmények a BKE Szociológiai Tanszék és a TÁRKI. Az adatfelvétel időpontja 1995. Országosan reprezentatív háztartás minta. Módszere: kérdőíves adatfelvétel.

[7] Választásunk oka, hogy az adatbázisban az MSZP támogatottsága a legnagyobb.

esetén a mintanagyság $n = n_I = 225$. A két elrendezés szerint egymás után végzett mintavételek során minden minta esetén meghatároztuk y_1 és y_2 összegének becslését mind a három becslés szerint, s ha megadható volt, a becslések várható értékét és varianciáját is. Az elrendezések és becslések kombinációi öt stratégiát eredményeznek, mivel (d_I, e_2) és (d_I, e_3) ugyanazon stratégiát határozzák meg. Minden becslésre kiszámítottuk - ha a számítás elvégezhető volt - a becslési variancia becslésének Horvitz-Thomson (HT), linearizációval kapott (L), random csoportokkal (RG), jackknife módszerrel (JK) illetve bootstrap módszerrel (BS) számolt becslését. Yates-Grundy becslés egyik elrendezés esetén sem adható, hiszen d_I nem rögzített mintanagyságú, d_{II} esetén pedig nem teljesül a $\pi_{kl} > 0$, $\forall k,l \in U$ feltétel. Az RG esetén $A = 5$ csoportra osztottuk a mintát. A JK módszert d_I esetén egy-egy háztartást elhagyva, míg d_{II} esetén egy-egy személyt elhagyva végeztük. A JK és az RG becslésnek, mint láttuk, két változata létezik, az összevetés céljával mindkét változatot kiszámítottuk. A bootstrap módszer alkalmazásakor a mintából készített virtuális populációból 500-szor egymás után vettünk mintát. Az 1. Mellékletben közöljük ezeknek a számításoknak a nemtriviális részeit. A szimulációt Stata programcsomagban implementáltuk, a programok a dolgozat végén, a 2. mellékletben találhatók, elektronikus változatuk a cs.elte.hu/~nmthrnt/index.htm honlapon hozzáférhető.

Eredmények

A szimuláció eredményei az 1. és 2. táblázatokban találhatók. A táblázatban t jelöli a paramétert, $\bar{\hat{t}}$ a paraméter becsléseinek átlagát, $s^2\left(\hat{t}\right)$ pedig varianciáját. A következő sorok a variancia-becsléseket jellemzik, pl. $\overline{\hat{V}}_{HT}(\hat{t})$ a variancia szimulációval kapott 500 HT-becslésének átlaga, $s^2\left(\hat{V}_{HT}\right)$ pedig az 500 érték szórásnégyzete, a $(t \in CI_{HT})$-vel jelzett sorban található értékek pedig azt jelzik, a szimulált minták hány százalékában esett a t paraméter a HT-becsléssel képzett, normális eloszlást feltételező konfidencia-intervallumba, a $\hat{t} \pm 1,96 \cdot \hat{V}_{HT}$ értékek közé.

Az (i) kutatási kérdésre válaszolva azt láthatjuk, hogy a d_{II} általában kevésbé hatékonynak ítélhető, mint d_I. Az e_1 esetén, amikor a becslés varianciája szimuláció nélkül is meghatározható volt, a d_{II} esetén kapott variancia y_1-re 21%-kal nagyobb, mint d_I esetén ($2,65 \cdot 10^{13}$ vs. $2,18 \cdot 10^{13}$); míg az y_2-re a varianciák nem különböznek jelentősen. Az y_1-et

tekintve e_2 esetén is jobbnak ítélhető d_{II}, akár a becsült varianciákat, akár a becslések szórásnégyzetét vesszük figyelembe . Az y_2 esetén ugyan az (d_{II}, e_2) esetén kisebb becsült varianciákat találunk, mint az (d_I, e_2) esetén, de ez előbbi stratégia torzított becsléshez vezet, ezért ésszerűbb az utóbbit a torzítatlan (d_{II}, e_3)-mal összevetni, ez az összevetés már nem mutat jelentős különbséget. Vélhetően a probléma érzékeny az y eloszlására, hiszen az összevetésekben y_1 ill. y_2 esetén nem teljesen azonos következtetésekre jutottunk. Mindent figyelembe véve, inkább a kevesebb háztartást, de több személyt mintába vonó d_I választása tűnik indokoltabbnak.

A (ii) kérdésben a becslések összehasonlítását tűztük ki célul. A $\bar{\hat{t}}$ eredmények alapján elmondható, hogy a d_{II} esetén a kiválasztási valószínűségeket figyelembe nem vevő e_2 valóban enyhén torzítottnak ítélhető. A másik problémára, miszerint megbízhatóbb becsléseket eredményez-e a π-becslés utólagos rétegezéssel kombinálva, nem adható általános válasz. Az y_1 esetén egyértelműen igennel felelhetünk: (d_I, e_1) vs. (d_I, e_2), ill. (d_{II}, e_1) vs. (d_{II}, e_3) összevetése szerint szinte kivétel nélkül minden variancia-becslés átlaga, és a becslések szórásnégyzete is nagyobb értéket mutat az e_1 esetén. Ugyanakkor y_2-re nem egyértelmű a kép. Az eredmények alapján az y eloszlásától függ, hogy érdemes-e utólagos rétegezést végezni.

A (iii) kérdés a variancia-becslési módszerek tulajdonságaira vonatkozott. Elmondható, hogy a JK és RG becslések két változata gyakorlatilag minden szempontból megegyező eredményeket hozott, tehát nincs köztük jelentős különbség. A becslések pontosságával kapcsolatban: azokat az eseteket tekintve, amikor a becslés valós varianciája megadható volt, általában a RG torzítása volt a legjelentősebb, míg átlagosan a BS volt a legközelebb a valódi értékhez.. A többi esetben a valós érték ismeretének hiányában $s^2\left(\hat{t}\right)$-hez viszonyíthatunk, így d_I esetén a linearizáció torzítása a legkisebb, d_{II}-nél nem állapítható meg jelentős különbség a becslések között, de megemlítendő, hogy itt a linearizáció nem szerepelt. Fontos észrevétel, hogy a variancia-becslések általában pozitív irányban torzítanak, ami azt jelenti, hogy segítségükkel inkább konzervatív becsléseket kapunk. Az összevetés másik kiindulópontja a variancia-becslések megbízhatósága. Az s^2 értékeket alapul véve elmondható, hogy - amikor megadható - a HT és a L becslés varianciája a legkisebb, más esetben a BS ítélhető a legjobbnak. Egyértelműen a legkevésbé

megbízhatónak ítélhető az RG, amely a legszélsőségesebb esetekben a többi variancia-becslésnél akár nagyságrenddel nagyobb varianciát is produkál. Az RG-n kívüli becslések varianciáinak különbsége legfeljebb 20-40%-os. Megemlítjük, hogy Wolter (1985) által hivatkozott szimulációs vizsgálatok eredménye szerint is az L az előnyösebb, a J-vel ill. a J-vel és RG-vel összevetve.

A névlegesen 95%-os konfidencia-intervallum is az RG esetében működik a legrosszabb módon. Bár a (d_{II}, e_2) stratégia esetén nem kaptunk a többi d_{II} stratégiánál lényegesen rosszabb eredményt, itt szem előtt kell tartani, hogy, mivel a becslés torzított, a konfidencia-intervallum elhelyezkedését a variancia-becslés esetleges hibáján kívül ez a torzítás is befolyásolja. Érdemes megjegyezni, hogy általában minden variancia-becslés esetén a mintáknak kevesebb, mint 95%-ában fedte az intervallum a paramétert. Hasonló következtetésre jutottak a Wolter (1985) által hivatkozott szimulációs vizsgálatok.

A (d_I, e_1) és (d_{II}, e_1) stratégiák esetében ismert a becslés varianciájának valós értéke, itt érdemes részletesebben is megvizsgálni a variancia-becslések megoszlását. A megoszlások ábrázolása box-plot diagrammal történt. A dolgozat végén található 2. melléklet diagrammjai a 25. és a 75. percentilist $\left(x_{[25]}, x_{[75]}\right)$ dobozzal ábrázolják, amit a medián oszt ketté. A dobozból kinyúló talp alsó határa az a legkisebb érték, ami nagyobb-egyenlő, mint $x_{[25]} - 1,5\left(x_{[75]} - x_{[25]}\right)$, míg a talp felső határa az a legnagyobb érték, ami kisebb-egyenlő, mint $x_{[25]} + 1,5\left(x_{[75]} - x_{[25]}\right)$. A két talpon kívül eső szélsőértékeket egyenként ábrázolja a diagramm. Az 1. és 2. ábra az y_1 esetében szemlélteti a két stratégia variancia-becsléseinek megoszlását. Az ábrákon a variancia valós értékét is jeleztük. Látható, hogy az e_1 esetén a HT és a BS, míg az e_2 esetén a JK és a BS eloszlása nagyon hasonló. Az RG mindkét esetben rendkívül magas szélsőértékeket is produkál, de a valós értéknél lényegesen nagyobb becslések valamennyi becslésnél megfigyelhetők. Ugyanakkor a becslések mediánja a valós érték alatt van.

A 2. melléklet 3. és 4. ábrája az y_2 esetében szemlélteti a két stratégia variancia-becsléseinek megoszlását. Látható, hogy az e_1 esetén a HT és a BS, míg az e_2 esetén a JK és a BS eloszlása ismét nagyon hasonló. Az RG mindkét esetben rendkívül magas szélsőértékeket is produkál, de most szélsőértékek a többi becslésnél nem találhatók, tehát megoszlásuk kevésbé nyújtott. A becslések mediánja ismét inkább a valós érték alatt van.

A (iv) kérdésre, miszerint különbözik-e az $(i)-(iii)$ problémákra adott válasz y_1 és y_2 esetén, a fentiekben már megadtuk a választ. Mindhárom esetben fontos volt az y szerepe, ezért elmondható, hogy a gyakorlatban mind az elrendezések, mind az összeg-becslések ill. annak variancia-becslései közötti választásnál érdemes figyelembe venni a felmérésünk y célváltozóját. Az y-ra vonatkozó előismereteket figyelembe véve dönthetünk a megfelelő mintavételi és elemzési stratégiáról.

Néhány Monte Carlo vizsgálat alapján nyilván nem lehet határozott következetetéseket levonni, vagy általános javaslatokat megfogalmazni. Ugyanakkor, mivel a különböző feltételek mellett kapott eredmények mutatnak némi egyezést, óvatos következtetések megfogalmazhatók. A variancia-becslési módszerek vizsgálatával kapcsolatban összefoglalóan elmondhatjuk, hogy a random csoportok módszere ítélhető a legrosszabbnak, a többi módszer nem különbözik lényegesen egymástól egyik szempontból sem. Ugyanakkor ennek a módszernek a teljesítménye nyilván függ a random csoportok számától. Wolter (1985) szerint a csoportok számának növelésével csökken a becslés varianciája, de torzítása nő. A Horvitz-Thomson becslés és a linearizáció módszere a szimuláció eredményei szerint mind várható értékét, mind varianciáját, mind konfidencia-intervallum képzési lehetőségét tekintve kissé jobbnak ítélhető társainál . Bár ezek a módszerek csak bizonyos elrendezés/becslés kombináció esetén állnak rendelkezésre, alkalmazásuk ezekben az esetekben javasolható.

1. táblázat. A szimuláció eredményei, d_I

	y_1		y_2	
	(d_I, e_1)	(d_I, e_2)	(d_I, e_1)	(d_I, e_2)
t	$7,40 \cdot 10^7$		418	
$\bar{\hat{t}}$	$7,42 \cdot 10^7$	$7,49 \cdot 10^7$	423	414
$V(\hat{t})$	$2,18 \cdot 10^{13}$	–	7327	–
$s^2\left(\hat{t}\right)$	$2,02 \cdot 10^{13}$	$1,72 \cdot 10^{13}$	6812	6843
$\overline{\hat{V}}_{HT}(\hat{t})$	$2,22 \cdot 10^{13}$	–	7430	–
$\overline{\hat{V}}_{L}(\hat{t})$	–	$1,91 \cdot 10^{13}$	–	6892
$\overline{\hat{V}}_{JK1}(\hat{t})$	–	$2,12 \cdot 10^{13}$	–	7636
$\overline{\hat{V}}_{JK2}(\hat{t})$	–	$2,12 \cdot 10^{13}$	–	7636
$\overline{\hat{V}}_{RG1}(\hat{t})$	$2,49 \cdot 10^{13}$	$2,08 \cdot 10^{13}$	8117	7592
$\overline{\hat{V}}_{RG2}(\hat{t})$	–	$2,10 \cdot 10^{13}$	–	7660
$\overline{\hat{V}}_{BS}(\hat{t})$	$2,22 \cdot 10^{13}$	$2,02 \cdot 10^{13}$	7433	8201
$s^2\left(\hat{V}_{HT}\right)$	$2,19 \cdot 10^{26}$	–	$3,10 \cdot 10^6$	–
$s^2\left(\hat{V}_{L}\right)$	–	$2,01 \cdot 10^{26}$	–	$2,50 \cdot 10^6$
$s^2\left(\hat{V}_{JK1}\right)$	–	$2,48 \cdot 10^{26}$	–	$3,10 \cdot 10^6$
$s^2\left(\hat{V}_{JK2}\right)$	–	$2,48 \cdot 10^{26}$	–	$3,10 \cdot 10^6$
$s^2\left(\hat{V}_{RG1}\right)$	$6,04 \cdot 10^{26}$	$4,63 \cdot 10^{26}$	$35,6 \cdot 10^6$	$32,4 \cdot 10^6$
$s^2\left(\hat{V}_{RG2}\right)$	–	$4,73 \cdot 10^{26}$	–	$32,5 \cdot 10^6$
$s^2\left(\hat{V}_{BS}\right)$	$2,27 \cdot 10^{26}$	$2,44 \cdot 10^{26}$	$3,30 \cdot 10^6$	$3,50 \cdot 10^6$
$t \in CI_{HT}$	95,6%	–	96,0%	–
$t \in CI_{L}$	–	96,8%	–	93,2%
$t \in CI_{JK1}$	–	97,2%	–	94,6%
$t \in CI_{JK2}$	–	97,2%	–	94,6%
$t \in CI_{RG1}$	91,8%	92,0%	90,0%	88,4%
$t \in CI_{RG2}$	–	92,0%	–	88,6%
$t \in CI_{BS}$	95,2%	97,0%	95,6%	95,4%

2. táblázat. A szimuláció eredményei, d_{II}

	y_1			y_2		
	(d_{II}, e_1)	(d_{II}, e_2)	(d_{II}, e_3)	(d_{II}, e_1)	(d_{II}, e_2)	(d_{II}, e_3)
t	$7,40 \cdot 10^7$			418		
$\bar{\hat{t}}$	$7,41 \cdot 10^7$	$7,69 \cdot 10^7$	$7,41 \cdot 10^7$	411	405	411
$V(\hat{t})$	$2,65 \cdot 10^{13}$	–	–	7206	–	–
$s^2\left(\hat{t}\right)$	$3,02 \cdot 10^{13}$	$2,30 \cdot 10^3$	$2,52 \cdot 10^{13}$	7311	6213	7047
$\overline{\hat{V}}_{JK1}(\hat{t})$	$2,89 \cdot 10^{13}$	$2,42 \cdot 10^3$	$2,54 \cdot 10^{13}$	7675	6318	7568
$\overline{\hat{V}}_{JK2}(\hat{t})$	–	$2,42 \cdot 10^3$	$2,54 \cdot 10^{13}$	–	6318	7568
$\overline{\hat{V}}_{RG1}(\hat{t})$	$2,91 \cdot 10^{13}$	$2,58 \cdot 10^3$	$2,67 \cdot 10^{13}$	7446	6396	7692
$\overline{\hat{V}}_{RG2}(\hat{t})$	–	$2,61 \cdot 10^3$	$2,72 \cdot 10^{13}$	–	6465	7798
$\overline{\hat{V}}_{BS}(\hat{t})$	$2,75 \cdot 10^{13}$	$2,17 \cdot 10^3$	$2,29 \cdot 10^{13}$	7287	6400	7569
$s^2\left(\hat{V}_{JK1}\right)$	$2,07 \cdot 10^{26}$	$2,22 \cdot 10^{26}$	$1,84 \cdot 10^{26}$	$3,80 \cdot 10^6$	$1,32 \cdot 10^6$	$3,10 \cdot 10^6$
$s^2\left(\hat{V}_{JK2}\right)$	–	$2,22 \cdot 10^{26}$	$1,84 \cdot 10^{26}$	–	$1,32 \cdot 10^6$	$3,10 \cdot 10^6$
$s^2\left(\hat{V}_{RG1}\right)$	$6,28 \cdot 10^{26}$	$7,26 \cdot 10^{26}$	$7,70 \cdot 10^{26}$	$34,3 \cdot 10^6$	$21,9 \cdot 10^6$	$34,1 \cdot 10^6$
$s^2\left(\hat{V}_{RG2}\right)$	–	$7,45 \cdot 10^{26}$	$8,16 \cdot 10^{26}$	–	$22,1 \cdot 10^6$	$34,6 \cdot 10^6$
$s^2\left(\hat{V}_{BS}\right)$	$2,02 \cdot 10^{26}$	$1,98 \cdot 10^{26}$	$1,67 \cdot 10^{26}$	$3,61 \cdot 10^6$	$1,38 \cdot 10^6$	$2,83 \cdot 10^6$
$t \in CI_{JK1}$	92,8%	94,4%	93,8%	93,8%	92,8%	93,8%
$t \in CI_{JK2}$	–	94,4%	93,8%	–	92,8%	93,8%
$t \in CI_{RG1}$	85,8%	88,0%	85,0%	86,4%	85,6%	88,0%
$t \in CI_{RG2}$	–	89,0%	85,6%	–	85,6%	88,6%
$t \in CI_{BS}$	91,8%	93,0%	92,0%	93,6%	92,8%	93,8%

6.1. 1. Melléklet

A (d_I, e_1) stratégia

A d_I esetén $n_I = 150$, $N_I = 1954$.

• $\hat{t}_{e_1} = N_I \bar{t}_{s_I} = N_I \sum_{s_I} \frac{t_i}{n_I} = N_I \sum_{U_i \in s_I} \frac{\sum_{k \in U_i} y_k}{n_I}$, ahol s_I a mintába került háztartásoknak, $U_I = \{U_1, \ldots, U_i, \ldots, U_{N_I}\}$ a populáció háztartásainak halmaza. A becslés π-becslés, ezért torzítatlan.

• $V\left(\hat{t}_{e_1}\right) = N_I^2 \frac{1-f_I}{n_I} S_{tU_I}^2$, ahol $f_I = \frac{n_I}{N_I}$, $S_{tU_I}^2 = \frac{1}{N_I - 1} \sum_{U_I} \left(t_i - \bar{t}_{U_I}\right)^2$, $\bar{t}_{U_I} = \frac{1}{N_I} \sum_{U_I} t_i$.

• $\hat{V}_{HT}\left(\hat{t}_{e_1}\right) = N_I^2 \frac{1-f_I}{n_I} S_{ts_I}^2$, ahol $S_{ts_I}^2 = \frac{1}{n_I - 1} \sum_{s_I} \left(t_i - \bar{t}_{s_I}\right)^2$. A HT becslés torzítatlan.

• L: A módszernek itt nincs értelme, hiszen a π-becslés lineáris becslés.

• JK: Elvégzése felesleges, mert könnyen beláthatóan $\hat{V}_{JK}\left(\hat{t}_{e_1}\right) = \frac{1}{1-f_I} \hat{V}_{HT}\left(\hat{t}_{e_1}\right)$.

• RG: A háztartásokat $A = 5$ csoportba soroltuk. $\hat{t}_a = N_I \bar{t}_{s_a}$, $a = 1 \ldots A$, ahol s_a a mintabeli háztartások a csoportba eső elemeinek halmaza. $\hat{t}_{RG} = \frac{1}{A} \sum_{a=1}^{A} \hat{t}_a = N_I \sum_{U_i \in s_I} \frac{t_i}{|s_a| A} = \hat{t}_{e_1}$, így az RG1 és RG2 azonos becslést ad.
$\hat{V}_{RG}\left(\hat{t}_{e_1}\right) = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{t}_a - \hat{t}_{RG}\right)^2$.

• BS: A d_I esetében az U^*-ba az s_I elemeit 13 példányban vesszük fel, $13 \approx \frac{1}{P(U_i \in s_I)} = \frac{N_I}{n_I}$. $A = 500$ mintát veszünk, d_I elrendezés szerint, az egyes mintákat a következő minta kivétele előtt a virtuális populációba visszahelyezve. A minták mindegyike 150 háztartást tartalmaz. Ezeket a mintákat jelölje s_a^*, $a = 1 \ldots A$. Ekkor $\hat{t}_a^* = N_I \bar{t}_{s_a^*}$, $\hat{t}_{BS} = \frac{1}{A} \sum_{a=1}^{A} \hat{t}_a^*$, $\hat{V}_{BS}\left(\hat{t}_{e_1}\right) = \frac{1}{A-1} \sum_{a=1}^{A} \left(\hat{t}_a^* - \hat{t}_{BS}\right)^2$. Könnyen belátható, hogy
$E\left(\hat{V}_{BS}\left(\hat{t}_{e_1}\right) | s\right) = \frac{N_I}{N_I - 1} \frac{n_{I-1}}{n_I} \hat{V}_{HT}\left(\hat{t}_{e_1}\right)$,
ezért a végső BS becslést ezzel a korrekcióval adtuk meg:
$\hat{V}_{BS}\left(\hat{t}_{e_1}\right) = \frac{1}{(A-1)} \sum_{a=1}^{A} \left(\hat{t}_a^* - \hat{t}_{BS}\right)^2 \frac{N_I - 1}{N_I} \frac{n_I}{n_I - 1}$,
így ez a BS-becslés, akárcsak a HT-becslés, torzítatlan.

A (d_I, e_2) stratégia

A korcsoportok: 16-39, 40-59, 60+, ezek definiálják az U_h rétegeket, $h = 1 \ldots H$. A minta e rétegek szerinti partíciója: $s = s_1 \cup \ldots \cup s_h \cup \ldots \cup s_H$.

• $\hat{t}_{e_2} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{s_h} y_k$. E becslés torzítatlan, de varianciája közvetlenül nem megadható.

• HT: nem értelmezhető, mert e_2 nem π-becslés.

• L: Az összeg és becslése felírható a következő alakban: $\hat{t}_{e_2} = \sum_{h=1}^{H} N_h \frac{\sum_{s_h} y_k \frac{N_I}{n_I}}{\sum_{s_h} 1 \frac{N_I}{n_I}} = \sum_{h=1}^{H} N_h \frac{\hat{t}_h}{\hat{N}_h} = f(\hat{t}_1, \ldots, \hat{t}_H, \hat{N}_1, \ldots, \hat{N}_H)$, $t = f(t_t, \ldots, t_H, N_1, \ldots, N_H)$. Itt $\hat{t}_h$ és $\hat{N}_h$ is felfogható összegek π-becsléseként: $t_h = \sum_U y_{hk}$, ahol $y_{hk} = y_k$, ha $k \in U_h$, egyébként 0. Ugyanígy $N_h = \sum_U I_{hk}$, ahol $I_{hk} = 1$, ha $k \in U_h$, egyébként 0.
A szükséges parciális deriváltak: $\frac{\partial \hat{t}_{e_2}}{\partial \hat{t}_h} = N_h \frac{1}{\hat{N}_h}$, $\frac{\partial \hat{t}_{e_2}}{\partial \hat{N}_h} = -N_h \frac{\hat{t}_h}{\hat{N}_h^2}$. Ebből $\hat{u}_k = \sum_{h=1}^{H} y_{hk} \frac{N_h}{\hat{N}_h} - \sum_{h=1}^{H} I_{hk} \frac{\hat{t}_h}{\hat{N}_h^2} N_h$. A variancia approximációja
$AV(\hat{t}_{e_2}) = V\left(\sum_s u_k \frac{N_I}{n_I}\right) = \sum \sum_U \triangle_{kl} u_k u_l \left(\frac{N_I}{n_I}\right)^2$,ebből

$\hat{V}_L\left(\hat{t}_{e_2}\right) = \sum\sum_s \check{\triangle}_{kl} u_k u_l \left(\frac{N_I}{n_I}\right)^2 =$

$= \sum_{k\in s}\sum_{l\in s} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \left(\sum_h \left(c_h y_{hk} - d_h I_{hk}\right)\right)\left(\sum_h \left(c_h y_{hl} - d_h I_{hl}\right)\right)$, ahol $c_h = \frac{N_h}{n_h}$, $d_h = N_h \frac{t_{s_h}}{n_h^2}$, és $\pi_k = \frac{n_I}{N_I}$, $\forall k$, továbbá $\pi_{kl} = \pi_k$, ha k és l ugyanazon háztartás tagjai, egyébként $\pi_{kl} = \frac{n_I(n_I-1)}{N_I(N_I-1)}$.

• JK: Az $s_a \in s_I$, $(a = 1 \ldots A,\ A = n_I)$ háztartás elhagyásával:

$\hat{t}_{(a)} = \sum_{h=1}^{H} \frac{N_h}{|s_h \setminus s_a|} \sum_{s_h \setminus s_a} y_k$. $\hat{t}_a = n_I \hat{t}_{e_2} - (n_I - 1)\,\hat{t}_{(a)}$, $\hat{t}_{JK} = \frac{1}{n_I}\sum_{a=1}^{A} \hat{t}_a$, $\hat{V}_{JK1}\left(\hat{t}_{e_2}\right) = \frac{1}{n_I(n_I-1)} \sum_{a=1}^{A} \left(\hat{t}_a - \hat{t}_{JK}\right)^2$, $\hat{V}_{JK2}\left(\hat{t}_{e_2}\right) = \frac{1}{n_I(n_I-1)} \sum_{a=1}^{A} \left(\hat{t}_a - \hat{t}_{e_2}\right)^2$.

• RG: $\hat{t}_a = \sum_{h=1}^{H} \frac{N_h}{|s_a \cap s_h|} \sum_{s_a \cap s_h} y_k$, $\hat{t}_{RG} = \frac{1}{A}\sum_{a=1}^{A} \hat{t}_a$,

$\hat{V}_{RG1}\left(\hat{t}_{e_2}\right) = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{t}_a - \hat{t}_{RG}\right)^2$, $\hat{V}_{RG2}\left(\hat{t}_{e_2}\right) = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\hat{t}_a - \hat{t}_{e_2}\right)^2$.

• BS: $\hat{t}_a^* = \sum_{h=1}^{H} \frac{N_h}{n_h^*} \sum_{s_h^*} y_k$, ahol n_h^* a h. réteg gyakorisága az a. bootstrap-mintában. Ebből $\hat{t}_{BS}$ és $\hat{V}_{BS}\left(\hat{t}_{e_2}\right)$ már adódik.

A (d_{II}, e_1) stratégia

A jelölések, mint fent. d_{II} esetén $n_I = 225$, $n_I = n$.

• $\hat{t}_{e_1} = \sum_s y_k \left|h_k\right| \frac{N_I}{n_I}$, ahol h_k a k. személy háztartása. A becslés π-becslés, ezért torzítatlan.

• $V\left(\hat{t}_{e_1}\right) = \sum\sum_U \left(\pi_{kl} - \pi_k\pi_l\right) \check{y}_k \check{y}_l =$

$= \sum_{k\in U}\left[\sum_{l\in U, l\neq k, h_l = h_k}\left(-\frac{n_I^2}{|h_k|^2 N_I^2}\check{y}_k\check{y}_l\right) + \left(\frac{n_I}{|h_k|N_I} - \frac{n_I^2}{|h_k|^2 N_I^2}\right)\check{y}_k^2 + \right.$

$\left. + \sum_{l\in U, h_l\neq h_k}\left(\frac{n_I}{|h_k|N_I}\frac{\left(n_I-1\right)}{|h_l|\left(N_I-1\right)} - \frac{n_I^2}{|h_k||h_l|N_I^2}\right)\check{y}_k\check{y}_l\right]$, ahol h_k jelöli a h. személy háztartását.

• HT: Mivel a $\pi_{kl} > 0$, $\forall k, l \in U$ feltétel nem teljesül, ezért HT-becslés nem adható.

• L: A módszernek itt nincs értelme, hiszen a π-becslés lineáris becslés.

• JK: A d_{II} esetében egy-egy személyt hagyunk el, tehát $|s_a| = 1$.

$\hat{t}_{(a)} = \sum_{s\setminus s_a} y_k \left|h_k\right| \frac{N_I}{n_I - 1}$, ebből már $\hat{t}_a$, $\hat{t}_{JK}$ és $\hat{V}_{JK}\left(\hat{t}_{e_1}\right)$ adódik. Mivel $\hat{t}_{JK} = \hat{t}_{e_1}$, a jackknife becslés két változata egybeesik.

• RG: A d_{II} esetében a mintabeli személyeket osztjuk 5 csoportra. $\hat{t}_a = \sum_{s_a} y_k \left|h_k\right| \frac{N_I}{n_I/5}$, $a = 1 \ldots 5$, $A = 5$. Ebből $\hat{t}_{RG}$ és $\hat{V}_{RG}\left(\hat{t}_{e_1}\right)$ már adódik. Mivel $\hat{t}_{RG} = \hat{t}_{e_1}$, az RG becslés két változata egybeesik.

• BS: A d_{II} esetében az U^*-ba az s_I elemeit 9 példányban vesszük fel, $9 = \left\lceil \frac{1}{P(U_i \in s_I)} \right\rceil = \left\lceil \frac{N_I}{n_I} \right\rceil$. Az s_a^*, $a = 1 \ldots A$ mintát d_{II} szerint vesszük, n_I háztartást választva, mégpedig úgy, hogy a kiválasztott háztartásnak mindig azt a tagját vonjuk a mintába, aki az eredeti

s mintában is szerepelt. Most $\hat{t}_a^* = \sum_{s_a^*} y_k \, |h_k| \, \frac{N_I(U^*)}{n_I}$, $a = 1 \ldots A$, $A = 500$, ahol $N_I(U^*)$ a virtuális populáció háztartásainak száma. Ebből $\hat{t}_{BS}$ és $\hat{V}_{BS}\left(\hat{t}_{e_1}\right)$ már adódik.

A (d_{II}, e_2) stratégia

• $\hat{t}_{e_2} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{s_h} y_k$. A becslés várhatóan nem torzítatlan, torzítása és varianciája a szimulációs eredményekből becsülhető.

• HT: Mivel a $\pi_{kl} > 0$, $\forall k, l \in U$ feltétel nem teljesül, ezért HT-becslés nem adható.

• L: Nem adható, mert a $\pi_{kl} > 0$, $\forall k, l \in U$ feltétel nem teljesül.

• JK: $\hat{t}_{(a)} = \sum_{h=1}^{H} \frac{N_h}{|s_h \setminus s_a|} \sum_{s_h \setminus s_a} y_k$, ebből már $\hat{t}_a$, $\hat{t}_{JK}$ és $\hat{V}_{JK1}\left(\hat{t}_{e_2}\right), \hat{V}_{JK2}\left(\hat{t}_{e_2}\right)$ adódik. Mivel $\hat{t}_{JK} \neq \hat{t}_{e_2}$, a jackknife becslés két változata nem esik egybe.

• RG: $\hat{t}_a = \sum_{h=1}^{H} \frac{N_h}{|s_a \cap s_h|} \sum_{s_a \cap s_h} y_k$. Ebből $\hat{t}_{RG}$ és $\hat{V}_{RG1}\left(\hat{t}_{e2}\right), \hat{V}_{RG2}\left(\hat{t}_{e2}\right)$ már adódik. Mivel $\hat{t}_{RG} \neq \hat{t}_{e_1}$, az RG becslés két változata nem esik egybe.

• BS: $\hat{t}_a^* = \sum_{h=1}^{H} \frac{N_h}{n_h^*} \sum_{s_h^*} y_k$, ahol n_h^* a h. réteg gyakorisága az a. bootstrap-mintában. Ebből $\hat{t}_{BS}$ és $\hat{V}_{BS}\left(\hat{t}_{e_2}\right)$ már adódik.

A (d_{II}, e_3) stratégia

• $\hat{t}_{e_3} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_{h\pi}} \hat{t}_{h\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} \frac{N_{h(k)}}{\sum_{s_{h(k)}} \frac{1}{\pi_k}}$, ahol $h(k)$ a k. személy korcsoportjának indexe, továbbá $\pi_k = \frac{1}{|h_k|} \frac{n_I}{N_I}$. A becslés torzítatlan, de nemlineáris, így varianciája közvetlenül nem megadható.

• HT, L: Mivel a $\pi_{kl} > 0$, $\forall k, l \in U$ feltétel nem teljesül, ezért HT ill. L-becslés nem adható.

• JK: $\hat{t}_{(a)} = \sum_{s \setminus s_a} \frac{y_k}{\pi'_k} \frac{N_{h(k)}}{\sum_{s_{h(k)} \setminus s_a} \frac{1}{\pi'_k}}$, ahol $\pi'_k = \frac{1}{|h_k|} \frac{n_I - 1}{N_I}$. Ebből már $\hat{t}_a$, $\hat{t}_{JK}$ és $\hat{V}_{JK1}\left(\hat{t}_{e_3}\right)$, $\hat{V}_{JK2}\left(\hat{t}_{e_3}\right)$ adódik. Mivel $\hat{t}_{JK} \neq \hat{t}_{e_3}$, a jackknife becslés két változata nem esik egybe.

• RG: $\hat{t}_{(a)} = \sum_{s_a} \frac{y_k}{\pi'_k} \frac{N_{h(k)}}{\sum_{s_{h(k)} \cap s_a} \frac{1}{\pi'_l}}$, ahol $\pi'_k = \frac{1}{|h_k|} \frac{n_I / 5}{N_I}$. Ebből már $\hat{t}_a$, $\hat{t}_{RG}$ és $\hat{V}_{RG1}\left(\hat{t}_{e_3}\right)$, $\hat{V}_{RG2}\left(\hat{t}_{e_3}\right)$ adódik. Mivel $\hat{t}_{JK} \neq \hat{t}_{e_3}$, az RG becslés két változata nem esik egybe.

• BS: $\hat{t}_{(a)} = \sum_{s_a^*} \frac{y_k}{\pi'_k} \frac{N_{h(k)}}{\sum_{s^*_{h(k)}} \frac{1}{\pi'_k}}$, ahol $\pi'_k = \frac{1}{|h_k|} \left\lceil \frac{n_I}{N_I} \right\rceil$, és $s^*_{h(k)}$ a k. személy korcsoportjához tartozók halmaza az a bootstrap-mintában. Ebből $\hat{t}_{BS}$ és $\hat{V}_{BS}\left(\hat{t}_{e_3}\right)$ már adódik.

7. Irodalom

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. In: V. P. Godambe, D. A. Sprott. *Foundations of Statistical Inference.* Toronto, Holt, Rinehart and Winston 203-242.

Bean, J. A. (1975). Distribution and properties of variance estimators for complex multistage probability samples. *Vital and Health Statistics* 2. (65)

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

Carlson, B. L. (1998). Software for Statistical Analysis of Sample Survey Data. In: Armitage, P., Colton, T. (szerk.) *Encyclopedia of Biostatistics.* Chichester, Wiley

Chaudhuri, A. (1988). Optimality of sampling strategies. In: P. R. Krishnaiah és C. R. Rao (szerk.)- *Handbook of Statistics*, Vol. 6. Amsterdam: North-Holland, 47-89.

Egészségi Állapot Felvétel, 1994. - Életmód, kockázati tényezők. (1996). Központi Statisztikai Hivatal, Budapest.

Godambe, V. P., Joshi, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations, I. *Annals of Mathematical Statisics,* 36, 1707-1722.

Horvitz, D. G., Thomson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Joshi, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations, III. *Annals of Mathematical Statisics,* 36, 1730-1742.

Levy, P. S., Lemeshow, S. (1999). *Sampling of Populations.* Wiley.

Rao, J. N. K. (1988). Variance estimation in sample surveys. In: P. R. Krishnaiah és C. R. Rao (szerk.)- *Handbook of Statistics*, Vol. 6. Amsterdam: North-Holland, 427-447.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling.* New York, Springer-Verlag

Wolter, K. M. (1985). *Introduction to Variance Estimation.* New York, Springer-Verlag

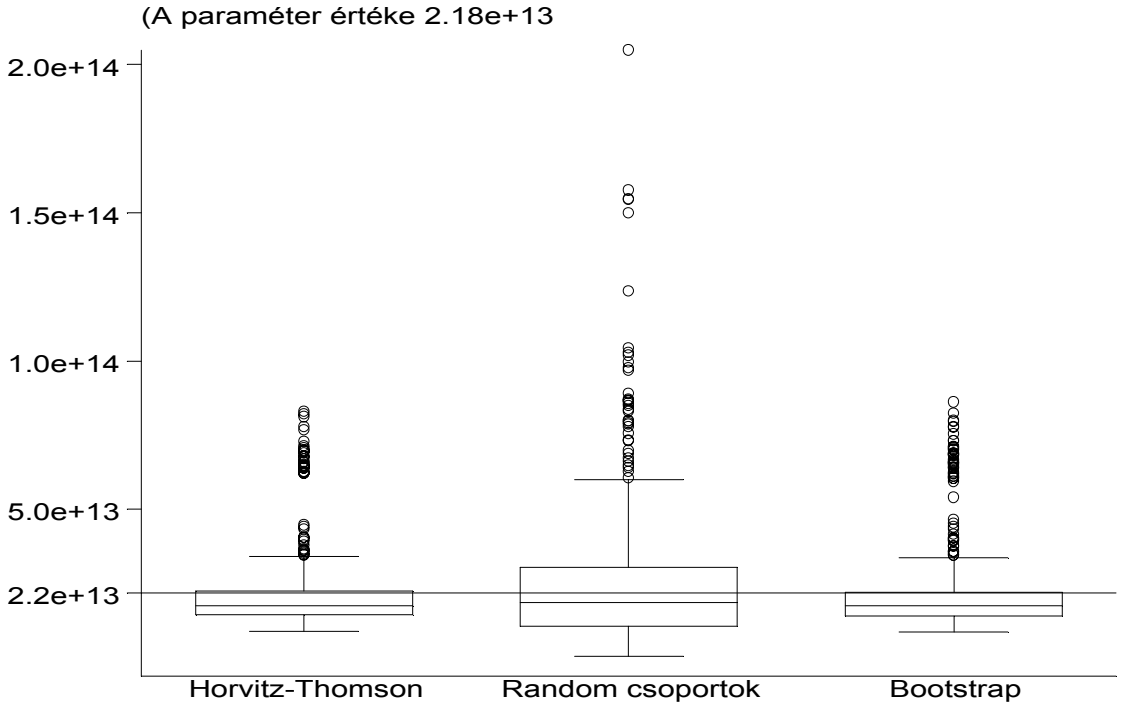
Yates, F., Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B,* 15, 235-261.

2. melléklet

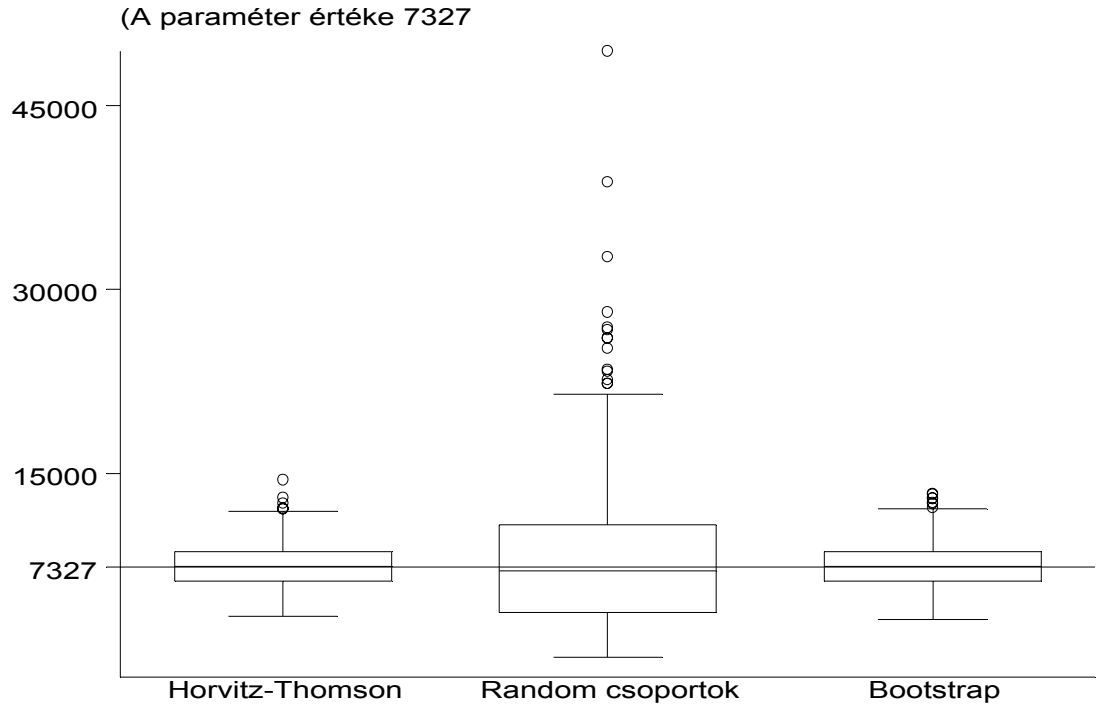
Tartalomjegyzék

Szimulációs eredmények – grafikonok

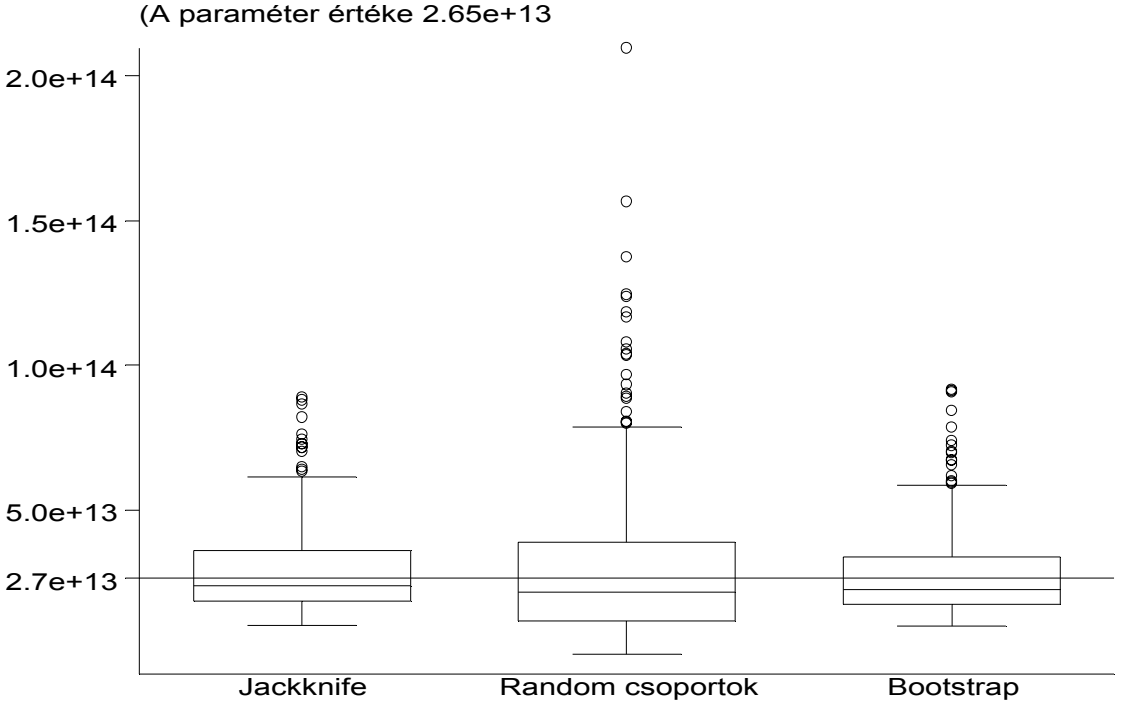
1. ábra y_1 változó, (d_I, e_1) stratégia



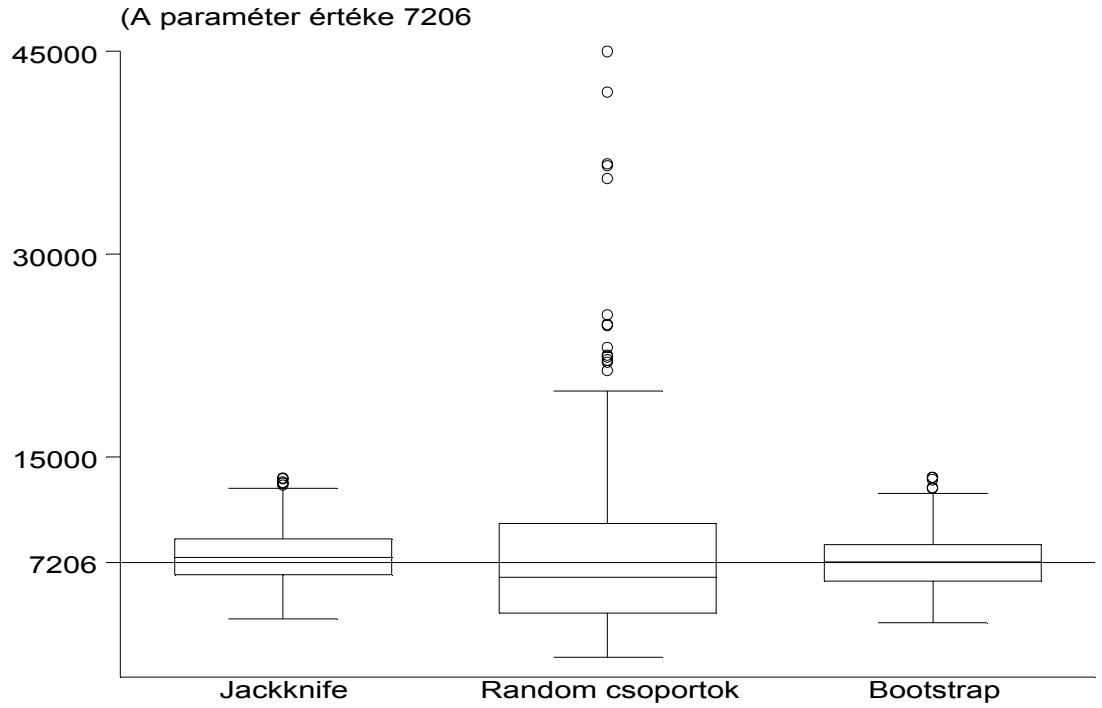
2. ábra y_2 változó, (d_I, e_1) stratégia



3. ábra y_1 változó, (d_{II}, e_1) stratégia



4. ábra y_2 változó, (d_{II}, e_1) stratégia



Szimulációs programok, Stata-ban implementálva

Előkészítés

```
clear
log using c:\nemeth\sim_1.log, replace
set memory 100m
set more off
set log linesize 120
use "C:\nemeth\d08_0.dta", clear

**********előkészítés
***változók előkészítése
   rename HAZON95 haztart
   lab var haztart "háztartás sorszáma"
   rename E4ANEME nem
   gen korcsop=.
   for num 1/3 \ num 55 35 -1: recode korcsop .=X if Y<E4ASZUL
   lab var korcsop "korcsoport"
   lab def korcsop 1 "16-39" 2 "40-59" 3 "60+"
   lab val korcsop korcsop
 ***1. paraméter: összesített havi nettó jövedelem országos szinten
   *a változó generálása
     gen jov=E4JOBECS
```

```
      recode jov 9=.
      lab var jov "havi nettó jövedelme"
 ***2. paraméter: a biztos MSZP-szavazók száma
   *a változó generálása
      gen mszp=1 if  E4XPAR94==4&E4XVALAM==1
      recode mszp .=0 if E4XPAR94~=.|E4XVALAM~=.
      lab def mszp 0 nem 1 igen
      lab val mszp mszp
      lab var mszp "biztos mszp szavazó"
***az adatbázis szűrése azokra, akikről minden információ rendelkezésre áll
   keep if mszp~=.&jov~=.
   save c:\nemeth\d08_1.dta, replace
**********
log close
clear
```

(d_1, e_1) stratégia

```
clear
log using c:\nemeth\sim_11a.log, replace
set memory 500m
set more off
set log linesize 120
use "C:\nemeth\d08_1.dta", clear

**1. paraméter
   *a paraméter valós értéke
     egen par=sum(jov)
     tab par

**********design I.

***a szükséges konstansok:
   *a háztartások száma a populációban
     quietly tab haztart
     scalar NI=r(r)
     di NI
   *a kiválasztott háztartások száma legyen 150.
     scalar nI=150
***a háztartásonkénti összeg számítása
   sort haztart
   egen ti=sum(jov), by(haztart)
***háztartások nagysága
   gen konst=1
   sort haztart
   egen haztnagy=sum(konst), by(haztart)
   lab var haztnagy "háztartás nagyság"


****1. becslés (pi-becslés)

****a populáció háztartásokra redukálása
   preserve
     sort haztart
     quietly by haztart: keep if _n==1
***a becslés varianciájának számítása
```

```
   **háztartások közötti populációs variancia számítása (S^2(tUI))
     quietly sum ti
     scalar S2U=r(Var)*NI/(NI-1)
   **a becslés varianciája:
    scalar V=NI^2*((1-nI/NI)/nI)*S2U
    di V
*szimuláció: 500 ismétlés
   program define sim11
     **a paraméter becslése
          gen unif=uniform()
          sort unif
                  gen minta=1 if _n<151
                  quietly sum ti if minta==1
                  scalar e`1'=NI*r(mean)
         **Yates-Grundy becslés nem adható, a jackknife-becslés megegyezik a HT-val, linearizációnak nincs
értelme a lineáris paraméterfüggvény miatt
          **a variancia Horvitz-Thomson-becslése
                  *háztartások közötti mintabeli variancia számítása (S^2(tsI))
              quietly sum ti if minta==1
              scalar S2s=r(Var)*NI/(NI-1)
        *a becslés varianciája:
              scalar h`1'=NI^2*((1-nI/NI)/nI)*S2s
          **a variancia random csoportok-becslése (A=5)
              gen unif2=uniform()
              sort minta unif2
              gen csoport=1 if _n<31
              for num 2/5 \ num 61 91 121 151: recode csoport .=X if _n<Y
              sort csoport
              egen ta=mean(ti), by(csoport)
              replace ta=. if csoport==.
              replace ta=NI*ta
              replace ta=(1/(5*(5-1)))*(ta-e`1')^2
              scalar r`1'=ta[1]+ta[31]+ta[61]+ta[91]+ta[121]
          **a variancia bootstrap-becslése (A=500 minta a virtuális populációból, visszatevéssel)
                  *a virtuális populáció konstruálása, [NI/nI]=13-szorosára duzzasztva a mintát, így 1950 elemet
kapunk:
                   sort minta
                   gen tivirt=.
                   quietly for num 1/150: replace tivirt=ti[X] if mod(_n,150)==mod(X,150)
                   replace tivirt=. if 1950<_n
              *500 minta visszatevéssel
                   quietly for num 1/500: gen uni=uniform()\sort uni\gen bsminta=1 if _n<151\quietly sum tivirt if
bsminta==1\scalar beX=NI*r(mean)\drop uni bsminta
                  *korrigált bootstrap-becslés a varianciára
                   scalar beatlag=be1
                   quietly for num 2/500:scalar beatlag=beatlag+beX
                   scalar beatlag=beatlag/500
                   quietly for num 1/500: scalar beX=(1/(500-1))*(beX-beatlag)^2
          scalar b`1'=0
                   quietly for num 1/500: scalar b`1'=b`1'+beX
                   *korrekció
                     scalar b`1'=b`1'*(NI-1)*nI/(NI*(nI-1))
         drop unif minta unif2 csoport ta tivirt
   end
  gen est11=.
  gen ht11=.
```

```
 gen rg11=.
 gen bs11=.
 for num 1/500: di X \ quietly sim11 X \ quietly sort CASE \ quietly replace est11=eX if _n==X \ quietly replace
ht11=hX if _n==X \ quietly replace rg11=rX if _n==X \ quietly replace bs11=bX if _n==X

 save c:\nemeth\d08_s11a.dta, replace

 restore

**********
**********
*a becslések eloszlásának jellemzői

        *design I, 1. becslés
        use c:\nemeth\d08_s11a.dta, clear
                sum est11 ht11 rg11 bs11, det
                *95%-os konfidencia-intervallum
                  for var ht11 rg11 bs11: gen Xci=cond(par<est11+1.96*sqrt(X)&est11-
1.96*sqrt(X)<par,1,0)\replace Xci=. if X==.|est11==.\sum Xci

log close
clear
```

(d_I, e_2) stratégia

```
clear
log using c:\nemeth\sim_12a.log, replace
set memory 500m
set more off
set log linesize 120
use "C:\nemeth\d08_1.dta", clear


**1. paraméter
   *a paraméter valós értéke
    egen par=sum(jov)
    tab par

**********design I.

***a szükséges konstansok:
   *a háztartások száma a populációban
    quietly tab haztart
    scalar NI=r(r)
    di NI
   *a kiválasztott háztartások száma legyen 150.
    scalar nI=150
***a háztartásonkénti összeg számítása
   sort haztart
   egen ti=sum(jov), by(haztart)
***háztartások nagysága
   gen konst=1
   sort haztart
   egen haztnagy=sum(konst), by(haztart)
   lab var haztnagy "háztartás nagyság"
```

```
****2. becslés (utólagos rétegezés korcsoportokra)
   *a populációs réteggyakoriság: Nh
    sort korcsop
    egen Nh=sum(konst), by(korcsop)
***a becslés nemlineáris, ezért variaciája és torzítása közvetlenül nem megadható
*szimuláció: 500 ismétlés
   program define sim12
     **a paraméter becslése
                set seed `1'
        gen unif=uniform()
                sort haztart
                quietly by haztart: replace unif=. if _n~=1
        sort unif
                gen mintas=1 if _n<151
                sort haztart
                egen minta=min(mintas), by(haztart)
                quietly sum minta
                scalar n=r(sum)
                *mintabeli réteggyakoriság: nh
                 sort minta korcsop
                 egen nh=sum(konst) if minta==1, by(korcsop)
                *súlyozott változó
                 gen sulyjov=jov*Nh/nh
                *a becslés
                quietly sum sulyjov if minta==1
                scalar e`1'=r(sum)
        **Horvitz-Thomson vagy Yates-Grundy becslés nem adható
        **a variancia linearizációs becslése
                sort korcsop
                egen tsh=sum(jov) if minta==1, by(korcsop)
                for num 1/3: gen iX=1 if korcsop==X&minta==1 \ recode iX .=0 if minta==1
                for num 1/3: gen yX=jov if korcsop==X&minta==1 \ recode yX .=0 if minta==1
                gen ch=Nh/nh
                gen dh=tsh*Nh/nh^2
                gen v=ch*(y1+y2+y3)-dh*(i1+i2+i3)
                sort haztart
                egen vhazt=sum(v), by(haztart)
                replace vhazt=. if minta~=1
                egen vossz =sum(v) if minta==1
                gen seged1=(1-(nI*(NI-1))/((nI-1)*NI))*v*(vossz-vhazt)
                egen seged2=sum(seged1) if minta==1
                gen seged3=(1-nI/NI)*v*vhazt
                egen seged4=sum(seged3) if minta==1
                sort minta
                scalar l`1'=seged2[1]+seged4[1]
        **a variancia jackknife-becslése
                quietly local i=1
                quietly sort minta haztart
                quietly while minta[`i']<. {
                 sort minta haztart
                 egen nha=sum(minta) if haztart~=haztart[`i'], by(korcsop)
                 replace nha=. if minta~=1
                 gen sulyjova=jov*Nh/nha
                 quietly sum sulyjova
                 scalar ta0`i'=r(sum)
```

```
  drop nha sulyjova
  local i=`i'+1
  sort minta haztart
 }
 quietly gen ta0=.
 sort minta haztart
 quietly sum minta
 local n=r(sum)
 quietly for num 1/`n': replace ta0=ta0X if _n==X
 *pszeudoértékek:
  gen ta=nI*e`1'*minta-(nI-1)*ta0
  replace ta=. if mintas~=1
  *a paraméter jackknife-becslése:
  egen tj=mean(ta)
 *a variancia 1. jackknife-becslése:
  gen seged5=(ta-tj)^2
  egen seged6=sum(seged5)
  sort mintas
  scalar j1`1'=seged6[1]/(nI*(nI-1))
 *a variancia 2. jackknife-becslése:
  gen seged7=(ta-e`1')^2
  egen seged8=sum(seged7)
  sort mintas
  scalar j2`1'=seged8[1]/(nI*(nI-1))
**a variancia random csoportok-becslése (A=5)
 gen unif2=uniform()
 sort mintas unif2
 gen csop=1 if _n<31
 for num 2/5 \ num 61 91 121 151: recode csop .=X if _n<Y
      egen csoport=min(csop), by(haztart)
 *csoportbeli réteggyakoriság: csh
       egen csh=sum(konst) if minta==1, by(korcsop csoport)
      *súlyozott változó
       gen sulycsop=jov*Nh/csh
      *a becslés
       for num 1/5: quietly sum sulycsop if csoport==X\ scalar taX=r(sum)
       scalar csest=(ta1+ta2+ta3+ta4+ta5)/5
      scalar r1`1'=0
      for num 1/5: scalar r1`1'=r1`1'+(taX-csest)^2
      scalar r2`1'=0
      for num 1/5: scalar r2`1'=r2`1'+(taX-e`1')^2
 scalar r1`1'=(1/(5*(5-1)))*r1`1'
      scalar r2`1'=(1/(5*(5-1)))*r2`1'
**a variancia bootstrap-becslése (A=500 minta a virtuális populációból, visszatevéssel)
      *a virtuális populáció konstruálása, [NI/nI]=13-szorosára duzzasztva a mintát:
      local tol=_N+1
      local ig=13*`n'
      if _N<`ig' {
            for num `tol'/`ig': set obs X
       }
       sort minta haztart
       gen jovvirt=.
       gen korcvirt=.
       gen haztbs=.
       gen htbs=.
       quietly for num 1/`n': replace jovvirt=jov[X] if mod(_n,`n')==mod(X,`n')
```

```
                quietly for num 1/`n': replace htbs=mintas[X] if mod(_n,`n')==mod(X,`n')
                quietly for num 1/`n': replace haztbs=haztart[X] if mod(_n,`n')==mod(X,`n')
                gen hanyadik=int(_n/(`n'+1))
                quietly for num 1/`n': replace korcvirt=korcsop[X] if mod(_n,`n')==mod(X,`n')
                replace jovvirt=. if `ig'<_n
                replace htbs=. if `ig'<_n
                replace haztbs=. if `ig'<_n
                replace hanyadik=. if `ig'<_n
                replace korcvirt=. if `ig'<_n
        *500 minta visszatevéssel
                quietly for num 1/500: set seed X\ gen uni=uniform()\sort htbs uni\gen bshminta=1 if
_n<151\egen bsminta=min(bshminta), by(haztbs hanyadik)\*mintabeli réteggyakoriság: nhbs\sort bsminta korcvirt\
egen nhbs=sum(konst) if bsminta==1, by(korcvirt)\ *súlyozott változó\ gen sulyjovb=jovvirt*Nh/nhbs \ *a becslés \
quietly sum sulyjovb if bsminta==1\scalar beX=r(sum)\drop uni nhbs sulyjovb bshminta bsminta
                scalar beatlag=be1
                quietly for num 2/500:scalar beatlag=beatlag+beX
                scalar beatlag=beatlag/500
                quietly for num 1/500: scalar beX=(1/(500-1))*(beX-beatlag)^2
        scalar b`1'=0
                quietly for num 1/500: scalar b`1'=b`1'+beX
                sort CASE
                drop if _n>3979
         drop unif-hanyadik
   end
  gen est12=.
  gen lin12=.
  gen jk112=.
  gen jk212=.
  gen rg112=.
  gen rg212=.
  gen bs12=.
  for num 1/500: di X \ quietly sim12 X \ quietly sort CASE \ quietly replace est12=eX if _n==X \ quietly replace
lin12=lX if _n==X \quietly replace jk112=j1X if _n==X \quietly replace jk212=j2X if _n==X \ quietly replace
rg112=r1X if _n==X \ quietly replace rg212=r2X if _n==X \ quietly replace bs12=bX if _n==X

 save c:\nemeth\d08_s12a.dta, replace

**********
*a becslések eloszlásának jellemzői

        *design I, 1. becslés
        use c:\nemeth\d08_s12a.dta, clear
                sum est12 lin12 jk112 jk212 rg112 rg212 bs12, det
                *95%-os konfidencia-intervallum
                 for var lin12 jk112 jk212 rg112 rg212 bs12: gen Xci=cond(par<est12+1.96*sqrt(X)&est12-
1.96*sqrt(X)<par,1,0)\replace Xci=. if X==.|est12==.\sum Xci

log close
clear
***************
```

(d_{II}, e_1) stratégia

```
clear
log using c:\nemeth\sim_21a.log, replace
set memory 500m
```

```
set more off
set log linesize 120
use "C:\nemeth\d08_1.dta", clear


**1. paraméter
   *a paraméter valós értéke
    egen par=sum(jov)
    tab par

**********design II.

***a szükséges konstansok:
   *a háztartások száma a populációban
    quietly tab haztart
    scalar NI=r(r)
    di NI
   *a kiválasztott háztartások száma legyen 225.
    scalar nI=225
***háztartások nagysága
   gen konst=1
   sort haztart
   egen haztnagy=sum(konst), by(haztart)
   lab var haztnagy "háztartás nagyság"

****1. becslés (pi-becslés)
*a becslés varianciája megadható:
        gen jovv=jov*haztnagy*NI/nI
        gen tag2=jovv^2*(nI/(haztnagy*NI)-(nI/(haztnagy*NI))^2)
        egen htossz=sum(jovv), by(haztart)
        gen tag1=-jovv*(nI/(haztnagy*NI))^2*(htossz-jovv)
        gen jovs2=jovv*(1/haztnagy)*((nI*(nI-1))/(NI*(NI-1))-(nI)^2/(NI)^2)
        egen jovs3=sum(jovs2)
        egen jovs4=sum(jovs2), by(haztart)
        gen tag3=(jovs3-jovs4)*jovv*(1/haztnagy)
        gen ossz=tag1+tag2+tag3
        egen var=sum(ossz)
        tab var
*szimuláció: 500 ismétlés
   program define sim21
      **a paraméter becslése
            set seed `1'
        gen unif=uniform()
            sort haztart unif
            gen minta=.
            quietly by haztart: replace minta=1 if _n==1
            gen unif2=uniform()
        sort minta unif2
            replace minta=. if 225<_n
            *a becslés
            quietly sum jovv if minta==1
            scalar e`1'=r(sum)
            ***a másodrendű kiválasztási valószínűségek nemnulla volta nem teljesül, ezért Horvitz-Thomson,
Yates-Grundy, ill. linearizált becslés nem adható
         **a variancia jackknife-becslése
            egen sj1=sum(jovv) if minta==1
```

```
                replace sj1=sj1*nI/(nI-1)
                gen ta0=sj1-jovv*nI/(nI-1)
                *pszeudoértékek:
                  gen ta=nI*e`1'*minta-(nI-1)*ta0
                *a variancia 1. jackknife-becslése (megegyezik 2.-kal):
                  gen seged5=(ta-e`1')^2
                  egen seged6=sum(seged5)
                  sort minta
                  scalar j`1'=seged6[1]/(nI*(nI-1))
          **a variancia random csoportok-becslése (A=5)
             gen unif3=uniform()
             sort minta unif3
             gen csop=1 if _n<46
             for num 2/5 \ num 91 136 181 226: recode csop .=X if _n<Y
                *a becslés
                  for num 1/5: quietly sum jovv if csop==X\ scalar taX=r(sum)*5
                scalar r`1'=0
                for num 1/5: scalar r`1'=r`1'+(taX-e`1')^2
             scalar r`1'=(1/(5*(5-1)))*r`1'
          **a variancia bootstrap-becslése (A=500 minta a virtuális populációból, visszatevéssel)
                *a virtuális populáció konstruálása:
                gen jovvir=jovv if minta==1
                  egen jovvirt=min(jovvir), by(haztart)
                local tol=_N+1
                quietly sum haztnagy if minta==1
                local n=r(sum)
                local ig=`n'*round(NI/nI,1)
                if _N<`ig' {
                        for num `tol'/`ig': set obs X
                 }
                  sort minta haztart
                  gen haztbs=.
                  gen htbs=.
                  quietly for num 1/`n': replace jovvirt=jovvirt[X] if mod(_n,`n')==mod(X,`n')
                  quietly for num 1/`n': replace htbs=minta[X] if mod(_n,`n')==mod(X,`n')
                  replace jovvirt=. if `ig'<_n
                  replace htbs=. if `ig'<_n
                  quietly sum htbs
                  scalar NIbs=r(N)
                  replace jovvirt=jovvirt*NIbs/NI
             *500 minta visszatevéssel
                  quietly for num 1/500: set seed X\ gen uni=uniform()\sort htbs uni\gen bshminta=1 if _n<226\*a
becslés \ quietly sum jovvirt if bshminta==1\scalar beX=r(sum)\drop uni bshminta
                  scalar beatlag=be1
                  quietly for num 2/500:scalar beatlag=beatlag+beX
                  scalar beatlag=beatlag/500
                  quietly for num 1/500: scalar beX=(1/(500-1))*(beX-beatlag)^2
       scalar b`1'=0
                  quietly for num 1/500: scalar b`1'=b`1'+beX
                  sort CASE
                  drop if _n>3979
        drop unif-htbs
  end
 gen est21=.
 gen rg21=.
 gen jk21=.
```

```
  gen bs21=.
  for num 1/500: di X \ quietly sim21 X \ quietly sort CASE \ quietly replace est21=eX if _n==X \ quietly replace
rg21=rX if _n==X \quietly replace jk21=jX if _n==X \ quietly replace bs21=bX if _n==X

 save c:\nemeth\d08_s21a.dta, replace

**********
*a becslések eloszlásának jellemzői

        *design I, 1. becslés
        use c:\nemeth\d08_s21a.dta, clear
                sum est21 jk21 rg21 bs21, det
                *95%-os konfidencia-intervallum
                  for var jk21 rg21 bs21: gen Xci=cond(par<est21+1.96*sqrt(X)&est21-
1.96*sqrt(X)<par,1,0)\replace Xci=. if X==.|est21==.\sum Xci

log close
clear
***************
```

(d_{II}, e_2) stratégia

```
clear
log using c:\nemeth\sim_22a.log, replace
set memory 500m
set more off
set log linesize 120
use "C:\nemeth\d08_1.dta", clear


**1. paraméter
   *a paraméter valós értéke
    egen par=sum(jov)
    tab par

**********design II.

***a szükséges konstansok:
   *a háztartások száma a populációban
    quietly tab haztart
    scalar NI=r(r)
    di NI
   *a kiválasztott háztartások száma legyen 225.
    scalar nI=225
***háztartások nagysága
   gen konst=1
   sort haztart
   egen haztnagy=sum(konst), by(haztart)
   lab var haztnagy "háztartás nagyság"

****2. becslés (utólagos rétegezés korcsoportokra)
   *a populációs réteggyakoriság: Nh
    sort korcsop
    egen Nh=sum(konst), by(korcsop)
*a becslés varianciája nem adható meg.
*szimuláció: 500 ismétlés
```

```
program define sim22
      **a paraméter becslése
                  set seed `1'
        gen unif=uniform()
                  sort haztart unif
                  gen minta=.
                  quietly by haztart: replace minta=1 if _n==1
                  gen unif2=uniform()
        sort minta unif2
                  replace minta=. if 225<_n
                  *mintabeli réteggyakoriság: nh
                   sort minta korcsop
                   egen nh=sum(konst) if minta==1, by(korcsop)
                  *súlyozott változó
                   gen sulyjov=jov*Nh/nh
                  *a becslés
                  quietly sum sulyjov if minta==1
                  scalar e`1'=r(sum)
                  ***a másodrendű kiválasztási valószínűségek nemnulla volta nem teljesül, ezért Horvitz-Thomson,
Yates-Grundy, ill. linearizált becslés nem adható
          **a variancia jackknife-becslése
                  quietly local i=1
                  quietly sort minta haztart
                  quietly while minta[`i']<. {
                   sort minta haztart
                   egen nha=sum(minta) if haztart~=haztart[`i'], by(korcsop)
                   sort minta haztart
                   replace nha=. if minta~=1|haztart==haztart[`i']
                   gen sulyjova=jov*Nh/nha
                   quietly sum sulyjova
                   scalar ta0`i'=r(sum)
                   drop nha sulyjova
                   local i=`i'+1
                   sort minta haztart
                  }
                  quietly gen ta0=.
                  sort minta haztart
                  quietly sum minta
                  local n=r(sum)
                  quietly for num 1/`n': replace ta0=ta0X if _n==X
                  *pszeudoértékek:
                   gen ta=nI*e`1'*minta-(nI-1)*ta0
                   *a paraméter jackknife-becslése:
                   egen tj=mean(ta)
                  *a variancia 1. jackknife-becslése:
                   gen seged5=(ta-tj)^2
                   egen seged6=sum(seged5)
                   sort minta
                   scalar j1`1'=seged6[1]/(nI*(nI-1))
                  *a variancia 2. jackknife-becslése:
                   gen seged7=(ta-e`1')^2
                   egen seged8=sum(seged7)
                   sort minta
                   scalar j2`1'=seged8[1]/(nI*(nI-1))
          **a variancia random csoportok-becslése (A=5)
             gen unif3=uniform()
```

```
            sort minta unif3
            gen csop=1 if _n<46
            for num 2/5 \ num 91 136 181 226: recode csop .=X if _n<Y
            *csoportbeli réteggyakoriság: csh
                  egen csh=sum(konst) if minta==1, by(korcsop csop)
                 *súlyozott változó
                  gen sulycsop=jov*Nh/csh
                 *a becslés
                  for num 1/5: quietly sum sulycsop if csop==X\ scalar taX=r(sum)
                  scalar csest=(ta1+ta2+ta3+ta4+ta5)/5
                 scalar r1`1'=0
                 for num 1/5: scalar r1`1'=r1`1'+(taX-csest)^2
                 scalar r2`1'=0
                 for num 1/5: scalar r2`1'=r2`1'+(taX-e`1')^2
            scalar r1`1'=(1/(5*(5-1)))*r1`1'
                  scalar r2`1'=(1/(5*(5-1)))*r2`1'
          **a variancia bootstrap-becslése (A=500 minta a virtuális populációból, visszatevéssel)
                  *a virtuális populáció konstruálása:
                  gen jovvir=jov if minta==1
                   egen jovvirt=min(jovvir), by(haztart)
                  gen korcvir=korcsop if minta==1
                   egen korcvirt=min(korcvir), by(haztart)
                  local tol=_N+1
                  quietly sum haztnagy if minta==1
                  local n=r(sum)
                  local ig=`n'*round(NI/nI,1)
                  if _N<`ig' {
                           for num `tol'/`ig': set obs X
                   }
                   sort minta haztart
                   gen htbs=.
                   quietly for num 1/`n': replace jovvirt=jovvirt[X] if mod(_n,`n')==mod(X,`n')
                   quietly for num 1/`n': replace htbs=minta[X] if mod(_n,`n')==mod(X,`n')
                   quietly for num 1/`n': replace korcvirt=korcsop[X] if mod(_n,`n')==mod(X,`n')
                   replace jovvirt=. if `ig'<_n
                   replace htbs=. if `ig'<_n
                   replace korcvirt=. if `ig'<_n
                   quietly sum htbs
                   scalar NIbs=r(N)
            *500 minta visszatevéssel
                   quietly for num 1/500: set seed X\ gen uni=uniform()\sort htbs uni\gen bshminta=1 if
_n<226\*mintabeli réteggyakoriság: nhbs\sort bshminta korcvirt\ egen nhbs=sum(konst) if bshminta==1,
by(korcvirt)\ *súlyozott változó\ gen sulyjovb=jovvirt*Nh/nhbs \ *a becslés \ quietly sum sulyjovb if
bshminta==1\scalar beX=r(sum)\drop uni nhbs sulyjovb bshminta
                   scalar beatlag=be1
                   quietly for num 2/500:scalar beatlag=beatlag+beX
                   scalar beatlag=beatlag/500
                   quietly for num 1/500: scalar beX=(1/(500-1))*(beX-beatlag)^2
         scalar b`1'=0
                   quietly for num 1/500: scalar b`1'=b`1'+beX
                   sort CASE
                   drop if _n>3979
          drop unif-htbs
   end
  gen est22=.
  gen rg122=.
```

```
  gen rg222=.
  gen jk222=.
  gen jk122=.
  gen bs22=.
  for num 1/500: di X \ quietly sim22 X \ quietly sort CASE \ quietly replace est22=eX if _n==X \ quietly replace
rg122=r1X if _n==X\ quietly replace rg222=r2X if _n==X \quietly replace jk122=j1X if _n==X\quietly replace
jk222=j2X if _n==X \ quietly replace bs22=bX if _n==X

 save c:\nemeth\d08_s22a.dta, replace

**********
*a becslések eloszlásának jellemzői

        *design I, 1. becslés
        use c:\nemeth\d08_s22a.dta, clear
                sum est22 jk122 jk222 rg122 rg222 bs22, det
                *95%-os konfidencia-intervallum
                  for var jk122 jk222 rg122 rg222 bs22: gen Xci=cond(par<est22+1.96*sqrt(X)&est22-
1.96*sqrt(X)<par,1,0)\replace Xci=. if X==.|est22==.\sum Xci

log close
clear
****************
```

(d_{II}, e_3) stratégia

```
clear
log using c:\nemeth\sim_23a.log, replace
set memory 500m
set more off
set log linesize 120
use "C:\nemeth\d08_1.dta", clear


**1. paraméter
   *a paraméter valós értéke
    egen par=sum(jov)
    tab par

**********design II.

***a szükséges konstansok:
   *a háztartások száma a populációban
    quietly tab haztart
    scalar NI=r(r)
    di NI
   *a kiválasztott háztartások száma legyen 225.
    scalar nI=225
***háztartások nagysága
   gen konst=1
   sort haztart
   egen haztnagy=sum(konst), by(haztart)
   lab var haztnagy "háztartás nagyság"

****2. becslés (pi-becslés, rétegezéssel korcsoportokra)
   *a populációs réteggyakoriság: Nh
```

```
    sort korcsop
    egen Nh=sum(konst), by(korcsop)
*a becslés varianciája nem adható meg.
*szimuláció: 500 ismétlés
  program define sim23
      **a paraméter becslése
            set seed `1'
            gen unif=uniform()
            sort haztart unif
            gen minta=.
            quietly by haztart: replace minta=1 if _n==1
            gen unif2=uniform()
            sort minta unif2
            replace minta=. if 225<_n
            gen pik=(1/haztnagy)*(nI/NI) if minta==1
            gen pikrec=1/pik if minta==1
            *réteggyakoriság súlyozott becslése
              sort minta korcsop
              egen Nhest=sum(pikrec) if minta==1, by(korcsop)
            *súlyozott változó
              gen sulyjov=jov*Nh/(pik*Nhest) if minta==1
            *a becslés
            quietly sum sulyjov if minta==1
            scalar e`1'=r(sum)
            ***a másodrendű kiválasztási valószínűségek nemnulla volta nem teljesül, ezért Horvitz-Thomson,
Yates-Grundy, ill. linearizált becslés nem adható
        **a variancia jackknife-becslése
            gen pikj=pik*(nI-1)/nI if minta==1
            gen pikjrec=1/pikj
            quietly local i=1
            quietly sort minta haztart
            quietly while minta[`i']<. {
              sort minta haztart
              egen nha=sum(pikjrec) if haztart~=haztart[`i'], by(korcsop)
              sort minta haztart
              replace nha=. if minta~=1|haztart==haztart[`i']
              gen sulyjova=jov*Nh/(pikj*nha)
              quietly sum sulyjova
              scalar ta0`i'=r(sum)
              drop nha sulyjova
              local i=`i'+1
              sort minta haztart
            }
            quietly gen ta0=.
            sort minta haztart
            quietly sum minta
            local n=r(sum)
            quietly for num 1/`n': replace ta0=ta0X if _n==X
            *pszeudoértékek:
              gen ta=nI*e`1'*minta-(nI-1)*ta0
              *a paraméter jackknife-becslése:
              egen tj=mean(ta)
            *a variancia 1. jackknife-becslése:
              gen seged5=(ta-tj)^2
              egen seged6=sum(seged5)
              sort minta
```

```
          scalar j1`1'=seged6[1]/(nI*(nI-1))
         *a variancia 2. jackknife-becslése:
          gen seged7=(ta-e`1')^2
          egen seged8=sum(seged7)
          sort minta
          scalar j2`1'=seged8[1]/(nI*(nI-1))
**a variancia random csoportok-becslése (A=5)
        gen pikrg=pik*45/nI
        gen pikrgrec=1/pikrg
   gen unif3=uniform()
   sort minta unif3
   gen csop=1 if _n<46
   for num 2/5 \ num 91 136 181 226: recode csop .=X if _n<Y
   *csoportbeli becsült réteggyakoriság: csh
         egen csh=sum(pikrgrec) if minta==1, by(korcsop csop)
        *súlyozott változó
         gen sulycsop=jov*Nh/(pikrg*csh)
        *a becslés
         for num 1/5: quietly sum sulycsop if csop==X\ scalar taX=r(sum)
         scalar csest=(ta1+ta2+ta3+ta4+ta5)/5
        scalar r1`1'=0
        for num 1/5: scalar r1`1'=r1`1'+(taX-csest)^2
        scalar r2`1'=0
        for num 1/5: scalar r2`1'=r2`1'+(taX-e`1')^2
   scalar r1`1'=(1/(5*(5-1)))*r1`1'
        scalar r2`1'=(1/(5*(5-1)))*r2`1'
**a variancia bootstrap-becslése (A=500 minta a virtuális populációból, visszatevéssel)
        *a virtuális populáció konstruálása:
        gen jovvir=jov if minta==1
         egen jovvirt=min(jovvir), by(haztart)
        gen korcvir=korcsop if minta==1
         egen korcvirt=min(korcvir), by(haztart)
        gen pikb=pik/((nI/NI)*round(NI/nI,1)) if minta==1
         egen pikvirt=min(pikb), by(haztart)
        gen pikrecb=1/pikb if minta==1
         egen pikrvirt=min(pikrecb), by(haztart)
        local tol=_N+1
        quietly sum haztnagy if minta==1
        local n=r(sum)
        local ig=`n'*round(NI/nI,1)
        if _N<`ig' {
                for num `tol'/`ig': set obs X
         }
         sort minta haztart
         gen htbs=.
         quietly for num 1/`n': replace jovvirt=jovvirt[X] if mod(_n,`n')==mod(X,`n')
         quietly for num 1/`n': replace htbs=minta[X] if mod(_n,`n')==mod(X,`n')
         quietly for num 1/`n': replace korcvirt=korcsop[X] if mod(_n,`n')==mod(X,`n')
         quietly for num 1/`n': replace pikvirt=pikvirt[X] if mod(_n,`n')==mod(X,`n')
         quietly for num 1/`n': replace pikrvirt=pikrvirt[X] if mod(_n,`n')==mod(X,`n')
         replace jovvirt=. if `ig'<_n
         replace htbs=. if `ig'<_n
         replace korcvirt=. if `ig'<_n
         replace pikvirt=. if `ig'<_n
         replace pikrvirt=. if `ig'<_n
   *500 minta visszatevéssel
```

```
                    quietly for num 1/500: set seed X\ gen uni=uniform()\sort htbs uni\gen bshminta=1 if
_n<226\*réteggyakoriság súlyozott becslése\sort bshminta korcvirt\egen Nhestb=sum(pikrvirt) if bshminta==1,
by(korcvirt)\*súlyozott változó\gen sulyjovb=jovvirt*Nh/(pikvirt*Nhestb) if bshminta==1\*a becslés\quietly sum
sulyjovb if bshminta==1\scalar beX=r(sum)\drop uni Nhestb sulyjovb bshminta
                    scalar beatlag=be1
                    quietly for num 2/500:scalar beatlag=beatlag+beX
                    scalar beatlag=beatlag/500
                    quietly for num 1/500: scalar beX=(1/(500-1))*(beX-beatlag)^2
          scalar b`1'=0
                    quietly for num 1/500: scalar b`1'=b`1'+beX
                    sort CASE
                    drop if _n>3979
          drop unif-htbs
   end
  gen est23=.
  gen rg123=.
  gen rg223=.
  gen jk223=.
  gen jk123=.
  gen bs23=.
  for num 1/500: di X \ quietly sim23 X \ quietly sort CASE \ quietly replace est23=eX if _n==X \ quietly replace
rg123=r1X if _n==X\ quietly replace rg223=r2X if _n==X \quietly replace jk123=j1X if _n==X\quietly replace
jk223=j2X if _n==X \ quietly replace bs23=bX if _n==X

 save c:\nemeth\d08_s23a.dta, replace

**********
*a becslések eloszlásának jellemzői

          *design I, 1. becslés
          use c:\nemeth\d08_s23a.dta, clear
                    sum est23 jk123 jk223 rg123 rg223 bs23, det
                    *95%-os konfidencia-intervallum
                     for var jk123 jk223 rg123 rg223 bs23: gen Xci=cond(par<est23+1.96*sqrt(X)&est23-
1.96*sqrt(X)<par,1,0)\replace Xci=. if X==.|est23==.\sum Xci

log close
clear
**********
```