Nagysággal arányos valószínűség szerinti mintavétel másodrendű bekerülési valószínűségekkel, visszatevés nélkül*

Mihályffy László,
a Központi Statisztikai Hivatal ny. statisztikai főtanácsadója
E-mail: Laszlo.Mihalyffy@ksh.hu

A szerző adott első- és másodrendű bekerülési valószínűségek ismeretében olyan mintavételi módszert mutat be, amelynek eredményeként a sokaság bármely eleme mintába kerülésének, illetve bármely két eleme együttes mintába kerülésének a valószínűsége megegyezik a megfelelő adott valószínűséggel.

TÁRGYSZÓ:
Mintavétel nagysággal arányos valószínűséggel.
Horvitz–Thompson-becslés.
Szórásnégyzet-becslés.

DOI: 10.20311/stat2016.03.hu0233

* A szerző köszönetét fejezi ki a lektornak, akinek hasznos észrevételei sokat segítettek a dolgozat színvonalának javításában.

A tanulmányban értékösszegek szórásnégyzete becslésének feladatával foglalkozunk nagysággal arányos valószínűséggel kiválasztott, rögzített elemszámú minták esetén. Terminológiai szempontból megjegyezzük, hogy a „nagysággal arányos valószínűség szerinti kiválasztás" helyett a dolgozat egészében használhatnánk a „nem egyenlő valószínűség szerinti kiválasztás" kifejezést is, a két fogalom között a gyakorlat szempontjából nincs nagy különbség. Olyan mintavételi feladatok megoldására adunk egyszerű eljárást, amelyekben a célsokaság egyes elemeihez tartozó bekerülési valószínűségek mellett a sokaságból kiválasztható elempárok mintába kerülésének együttes valószínűsége is ismert. Ez azt jelenti, hogy az elsőrendű, vagyis az egyes elemekhez tartozó bekerülési valószínűségek ismeretében meg kell határozni a másodrendű-, az elempárokhoz tartozó együttes bekerülési valószínűségek alkalmas rendszerét, a kétféle valószínűségek közötti összefüggés alapján, (lásd később). Ennek az utóbbi részfeladatnak a megoldása bizonyos esetekben igen egyszerű, erre majd mutatunk példát. Az általános esetre nézve egyelőre nincs javaslatunk.

Az első- és másodrendű bekerülési valószínűségekre épülő mintavételi eljárásunk leírását tanulmányunk *1.* fejezete tartalmazza, ezt követi egy numerikus példával illusztrált alkalmazás bemutatása a *2.* fejezetben. A *3.* fejezetben javasolt új eljárásunk tulajdonságaival foglalkozunk, ennek keretében keresünk választ arra a kérdésre, hogy milyen körülmények között előnyös ennek alkalmazása a nagysággal arányos valószínűség szerinti mintavétel körébe tartozó hasonló célú módszerekhez képest.

A dolgozatban a következő jelöléseket alkalmazzuk.

$U = \{1, 2, \ldots, N\}$: N elemű véges sokaság;

$s = \{i_1, i_2, \ldots, i_n\}$: n elemű minta az U sokaságból;

U^*: az U sokaságból kiválasztható n elemű minták halmaza;

$C = N!/((N-n)!n!)$: a különböző n elemű mintáknak a száma;

x_i: indikátorváltozó, $x_i = 1$, ha $i \in s$, egyébként $x_i = 0$, $i = 1, 2, \ldots, N$;

$(x_1, x_2, \ldots, x_N)$: az s minta alternatív jelölése $\left(\sum_{k=1}^{N} x_k = n \right)$;

$p(s)$: valószínűségi függvény, pozitív minden $s \in U^*$ mintára, $\sum_{s \in U^*} p(s) = 1$;

$p(s) \propto \Phi(s)$: mintavételi terv, $p(s)$ hozzárendelése egy konkrét függvénytípushoz;

$H = -\sum_{s \in U^*} p(s) \log p(s)$: a mintavételi terv entrópiája;

π_i: az $i \in U$ elem bekerülésének valószínűsége egy $s \in U^*$ mintába;

π_{ij}: az $i, j, i \neq j$ elemek együttes bekerülésének valószínűsége egy $s \in U^*$ mintába;

$p_j = \pi_j / n$: a j elem kiválasztásának valószínűsége a sokaságból $j = 1, 2, \ldots, N$;

πps-minta : nagysággal arányos valószínűséggel visszatevés nélkül kiválasztott minta;

pps-minta : nagysággal arányos valószínűséggel visszatevéssel kiválasztott minta.

A π_i és a π_{ij} valószínűségeket első-, illetve másodrendű bekerülési valószínűségeknek is nevezzük. A bemutatott jelölések mellett szükségünk lesz még a πps-mintákra vonatkozó következő összefüggésekre is:

$$\pi_1 + \pi_2 + , ..., + \pi_N = n , \qquad /1/$$

$$\sum_{j \neq i}^{N} \pi_{ij} = (n-1)\, \pi_i , \; i = 1, 2, \ldots, N , \qquad /2/$$

$$\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i . \qquad /3/$$

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i \in s} \sum_{j \in s,\, j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \qquad /4/$$

$$\hat{V}(\hat{Y}_{pps}) = \frac{1}{n(n-1)} \sum_{j \in s} \left(\frac{y_j}{p_j} - \hat{Y}_{pps} \right)^2 . \qquad /5/$$

A /3/ összefüggés a *Horvitz–Thompson*- [1952] becslőfüggvény az $Y = \sum_{k=1}^{N} y_k$ alakú értékösszegek becslésére. $\hat{Y}_{HT}$ szórásnégyzete a

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} y_i y_j$$

kifejezéssel egyenlő, ennek a mintából származó becslése *Sen* [1953], valamint *Grundy–Yates* [1953] nevéhez fűződik. (A /4/ képletben a minta elemeit azonosítójuk szerinti növekvő sorrendben kell figyelembe venni.) Az /5/ képlet $V\left(\hat{Y}_{HT}\right)$ megfelelője *pps*-minták esetén.

Szembetűnő a *πps*- és a *pps*-mintákhoz tartozó szórásnégyzet-becslések közötti különbség. A /4/ képlet alkalmazását megnehezítik a benne szereplő π_{ij} másodrendű bekerülési valószínűségek, ezek értékét ugyanis a mintavételi terv általában csak közvetett módon határozza meg, kiszámításuk vagy becslésük meglehetősen számításigényes. A *pps*-mintához tartozó $\hat{V}\left(\hat{Y}_{pps}\right)$ becslés viszont csak az egyelemű mintákhoz tartozó $p_1, p_2, ..., p_N$ valószínűségektől függ $\left(\sum_{k=1}^{N} p_k = 1\right)$, kiszámítása tehát rendkívül egyszerű. A *πps* mintavételi eljárásokkal kapcsolatos kutatásokat nagyrészt az motiválja, hogy a szórásnégyzetek becslésére minél egyszerűbb módszert találjanak.

1. Mintavétel másodrendű bekerülési valószínűségek segítségével

Tegyük fel, hogy rendelkezésünkre állnak az /1/ és a /2/ összefüggéseket kielégítő első- és másodrendű bekerülési valószínűségek. Tegyük fel azt is, hogy ezek valamennyien 0 és 1 közé esnek, továbbá, hogy $i \neq j$ esetén $\pi_{ij} \leq \pi_i\pi_j$; az utóbbi feltétel biztosítja, hogy a /4/ képlettel becsült szórásnégyzet nem negatív legyen. Az említett $p_i = \pi_i/n$ jelölést alkalmazva, definiáljuk a következő algoritmust.

1. lépés. Az $U = \{1, 2, \ldots, N\}$ sokaságból p_i valószínűséggel válasszuk ki az *i* elemet.

2. lépés. Az $U \setminus \{i\}$ „redukált” sokaságból válasszunk ki $n-1$ elemet nagysággal arányos valószínűséggel, a π_{i1}/π_i, π_{i2}/π_i, …, $\pi_{i,\,i-1}/\pi_i$, $\pi_{i,\,i+1}/\pi_i$, …, π_{iN}/π_i valószínűségek felhasználásával. Jelöljük a kiválasztott elemeket i_2-vel, i_3-mal, …, i_n-nel. Az eljárás befejeződött, az eredmény az *n* elemű $s = \{i, i_2, i_3, \ldots, i_n\}$ minta.

Megjegyzés. A 2. lépésben az $n-1$ elem kiválasztását célszerűen a randomizált szisztematikus mintavétel módszerével végezzük el (*Hartley–Rao* [1962)], mivel az ismert eljárások között technikailag ez a legegyszerűbb, és ügyes alkalmazás esetén a legkevésbé műveletigényes.[1] A módszer részletes leírása megtalálható a Függelékben.

1. Tétel. A bemutatott algoritmus alkalmazása mellett az *U* sokaság bármely *i* eleme mintába kerülésének a valószínűsége π_i. Annak a valószínűsége, hogy a sokaság bármely két eleme, *i* és *j*, $i \neq j$ együtt kerüljön a mintába, π_{ij}-vel egyenlő.

Bizonyítás. Annak valószínűsége, hogy az 1. lépésben az *i* elem kerül a mintába, $p_i = \pi_i / n$. Ha az 1. lépésben p_j valószínűséggel a sokaságból a *j* elemet választjuk, $j \neq i$, akkor az *i* elem mintába kerülésének $P(i \mid j)$ feltételes valószínűsége a π_{ji}/π_j kifejezéssel egyenlő, vagyis azzal a valószínűséggel, amellyel az $U \setminus \{j\}$ redukált sokaság *i* eleme bekerül valamelyik $(n-1)$ elemű mintába. Mivel a $P(i \mid i)$ kifejezéshez csak az 1 értéket lehet hozzárendelni, és az 1. lépéshez tartozó választási lehetőségek teljes eseményrendszert alkotnak, a teljes valószínűség tétele szerint

$$P(i) = \sum_{j=1}^{N} p_j P(i \mid j) = \sum_{j \neq i}^{N} (\pi_j / n)\pi_{ji}/\pi_j + p_i .$$

Mivel $\pi_{ji} = \pi_{ij}$, a $p_i = \pi_i / n$, a $P(i)$ valószínűséget a /2/ egyenlőség miatt a következőképpen is írhatjuk:

$$P(i) = \sum_{j \neq i}^{N} \frac{\pi_{ij}}{n} + \frac{\pi_i}{n} = \frac{(n-1)\pi_i}{n} + \frac{\pi_i}{n} = \pi_i ;$$

ezzel a tétel első állítását bizonyítottuk. Tekintsük most a sokaság *i* és *j* elemeit, $i \neq j$, és vizsgáljuk ezek együttes bekerülési valószínűségét. Két dolgot kell figyelembe vennünk. Egyrészt, hogy az algoritmus működése során bármely *n* elemű minta *n* különböző helyzetben fordulhat elő annak függvényében, hogy melyik elemét választjuk ki az 1. lépésben, és a megfigyelt bekerülési valószínűségek értékének pontosan $1/n$-ed része képződik minden egyes előfordulásnál. Másrészt, minden olyan mintának, amely tartalmazza az *i* és a *j* elemet, elő kell fordulnia olyan hely-

[1] A randomizált szisztematikus kiválasztás első lépésében a sokaság elemeit véletlen sorrendbe kell rendezni. Ez nagy elemszámú sokaság esetén műveletigényes részfeladat, de az átrendezést nem kell minden egyes mintavétel előtt megismételni.

zetben is, amikor az i elemet az algoritmus 1. lépésében választjuk ki $p_i = \pi_i / n$ valószínűséggel, és ezt megszorozzuk a j elem $P(j|i)$ feltételes valószínűségével, ami nem más, mint j bekerülési valószínűsége az $U \setminus \{j\}$ redukált sokaság $(n-1)$ elemű mintáiba. A $p_i \pi_{ij} / \pi_i = \pi_{ij} / n$ érték a keresett másodrendű bekerülési valószínűségnek a speciális esetből adódó része, a teljes érték pedig $n \times \pi_{ij} / n$.

Ad hoc kifejezéssel élve, eljárásunkat *p_ij*-módszernek fogjuk nevezni. Kivételes esetektől eltekintve, a gyakorlati alkalmazásokban a *p_ij*-módszert meg kell előznie egy olyan algoritmusnak, amely az elsőrendű bekerülési valószínűségek adott – az /1/ feltételt kielégítő – rendszere mellett a másodrendű bekerülési valószínűségek egy konzisztens rendszerét állítja elő.

2. Példa az alkalmazásra

Tegyük fel, hogy adottak az /1/ feltételt kielégítő $\pi_1, \pi_2, \ldots, \pi_N$ elsőrendű bekerülési valószínűségek, és végezzük el a következő műveleteket. Legyen

$$i = 1, 2, ..., N \text{ esetén } p_i = \pi_i / n, \qquad /6/$$

$$\tau = \sum_{i=1}^{N} \frac{p_i}{1 - 2p_i}, \qquad /7/$$

$$i = 1, 2, ..., N \text{ esetén } u_i = \frac{n-1}{n(1+\tau)} \frac{1}{1-2p_i}, \qquad /8/$$

$$i, j = 1, 2, ..., N,\ i \neq j \text{ esetén } x_{ij} = u_i + u_j,\ x_{11} = x_{22} =, ..., = x_{NN} = 0, \qquad /9/$$

$$\pi_{ij} = x_{ij} \pi_i \pi_j,\ i, j = 1, 2, ..., N,\ i \neq j. \qquad /10/$$

A /6/–/10/ összefüggésekkel meghatározott másodrendű bekerülési valószínűségekkel több helyen is találkozhatunk az irodalomban. Az $n = 2$ esetben *Brewer* [1963], *Rao* [1965], illetve *Durbin* [1967] módszerében szerepelnek. A $0 < \pi_i < 1$, $i = 1, 2, ..., N$ és az /1/ feltételek mellett a /6/–/10/ képletekkel meghatározott π_{ij} értékek minden esetben pozitívok, de az $n > 2$ esetben előfordulhat, hogy $x_{ij} > 1$

bizonyos indexpárokra, és akkor nem teljesül a $\pi_{ij} \le \pi_i \pi_j$ feltétel. Ha az előző feltételek mellett $i = 1, 2, ..., N$ esetén megköveteljük az $np_i < 1/2$ egyenlőtlenség teljesülését is, akkor a /6/–/10/ feltételekkel meghatározott π_{ij} valószínűségekre a $\pi_{ij} \le \pi_i \pi_j$ egyenlőtlenség is teljesül, $\pi_{ij} = 0$ miatt akkor is, ha $i = j$. Ekkor tehát a π_{ij} bekerülési valószínűségek minden rájuk vonatkozó feltételt teljesítenek, és a /4/ szórásnégyzet-becslés minden esetben nem negatív értéteket szolgáltat.

Mindezt a következő numerikus példával illusztráljuk. Legyen $N = 7$, és legyenek az elsőrendű bekerülési valószínűségek

$$0{,}48,\ 0{,}29,\ 0{,}49,\ 0{,}48,\ 0{,}41,\ 0{,}37,\ 0{,}48; \qquad /11/$$

ezek összege $n = 3$, tehát háromelemű minták kiválasztása a feladat. Jelöljük π-vel azt a vektort, amelynek komponensei az elsőrendű bekerülési valószínűségek. A /6/–/10/ képletek felhasználásával az $\mathbf{X} = (x_{ij})_{N \times N}$ és a $\mathbf{\Pi} = (\pi_{ij})_{N \times N}$ mátrixokra a következőket kapjuk:

$$\mathbf{X} = \begin{vmatrix} 0 & 0{,}7466 & 0{,}8142 & 0{,}8102 & 0{,}7842 & 0{,}7708 & 0{,}8102 \\ 0{,}7466 & 0 & 0{,}7506 & 0{,}7466 & 0{,}7206 & 0{,}7072 & 0{,}7466 \\ 0{,}8142 & 0{,}7506 & 0 & 0{,}8142 & 0{,}7882 & 0{,}7748 & 0{,}8142 \\ 0{,}8102 & 0{,}7466 & 0{,}8142 & 0 & 0{,}7842 & 0{,}7708 & 0{,}8102 \\ 0{,}7842 & 0{,}7206 & 0{,}7882 & 0{,}7842 & 0 & 0{,}7448 & 0{,}7842 \\ 0{,}7708 & 0{,}7072 & 0{,}7748 & 0{,}7708 & 0{,}7448 & 0 & 0{,}7708 \\ 0{,}8102 & 0{,}7466 & 0{,}8142 & 0{,}8102 & 0{,}7842 & 0{,}7708 & 0 \end{vmatrix}$$

$$\mathbf{\Pi} = \begin{vmatrix} 0 & 0{,}1039 & 0{,}1915 & 0{,}1867 & 0{,}1543 & 0{,}1369 & 0{,}1867 \\ 0{,}1039 & 0 & 0{,}1067 & 0{,}1039 & 0{,}0857 & 0{,}0759 & 0{,}1039 \\ 0{,}1915 & 0{,}1067 & 0 & 0{,}1915 & 0{,}1584 & 0{,}1405 & 0{,}1915 \\ 0{,}1867 & 0{,}1039 & 0{,}1915 & 0 & 0{,}1543 & 0{,}1369 & 0{,}1867 \\ 0{,}1543 & 0{,}0857 & 0{,}1584 & 0{,}1543 & 0 & 0{,}1130 & 0{,}1543 \\ 0{,}1369 & 0{,}0759 & 0{,}1405 & 0{,}1369 & 0{,}1130 & 0 & 0{,}1369 \\ 0{,}1867 & 0{,}1039 & 0{,}1915 & 0{,}1867 & 0{,}1543 & 0{,}1369 & 0 \end{vmatrix} \qquad /12/$$

Könnyen ellenőrizhető, hogy a $\boldsymbol{\pi}$ vektor és a /12/ képlettel meghatározott $\mathbf{\Pi}$ mátrix kielégíti a /2/ feltételt $n = 3$ értéke mellett. A következőkben bemutatjuk egy minta kiválasztását a $\mathbf{\Pi}$ mátrix segítségével az 1. fejezetben ismertetett algoritmus alapján.

Az első lépésben az $N = 7$ elemű sokaság egy *i* indexű elemét és ezzel együtt a $\mathbf{\Pi}$ mátrix *i*-edik sorát kell kiválasztanunk. A második lépésben ezután a kiválasztott sor elemeinek π_{ij} valószínűségéből π_i -vel való osztással elsőrendű bekerülési valószínűségeket képezünk, melyekkel $n-1$ elemű mintákat választhatunk ki – *n* esetünkben 3 – abból a sokaságból, amelyet az adott hételemű sokaságból az *i* indexű elem kihagyásával kapunk. Ennek alapja a /2/ egyenlőség, amelyet most a következő alakba írhatunk:

$$\sum_{j=1,\, j\neq i}^{7} \pi_{ij}/\pi_i = 3-1 = 2\,. \qquad /2'/$$

Segédeszközként mind az első, mind pedig a második lépésben a randomizált szisztematikus eljárást használjuk. Ennél a módszernél a sokaság elemeit véletlen sorrendbe kell rendezni; feltesszük, hogy a /11/ szerinti felsorolás már ezt a sorrendet tükrözi.

Az algoritmus első lépésben a hételemű sokaságból csak egyet kell kiválasztanunk a p_j valószínűségekkel, ezek értéke most /6/ szerint 0,16, 0,29/3, 0,49/3, 0,16, 0,41/3, 0,37/3, 0,16. A Függelék alapján ezekből a következő kumulált összegeket képezzük: 0,16, 0,257, 0,42, 0,58, 0,717, 0,84, 1,0 (kerekített értékek). A (0, 1) intervallumon egyenletes eloszlású változót generáló program a 0,1443637 értéket eredményezte, ez 0 és 0,16 közé esik, így $i = 1$ eleme a mintának, a további mintaelemeket pedig a $\mathbf{\Pi}$ mátrix első sorának segítségével határozzuk meg.

$\mathbf{\Pi}$ első sorából a 0 elemet elhagyjuk, a további értékek

0,1039, 0,1915, 0,1867, 0,1543, 0,1369 és 0,1867

pedig rendre a sokaság 2, 3, 4, 5, 6, illetve 7 indexű elemeihez tartoznak. Ha ezeket π_1 -gyel osztjuk és összegezzük, akkor az előzők és /2'/ szerint kettőt kapunk eredményül. Mivel ez egyrészt a sokaságbeli elemek nagyságának összege, másrészt pedig kételemű mintát kell kiválasztanunk, a randomizált szisztematikus módszert alkalmazva az egységnyi lépéshosszt kell használnunk. Az eljárás szerint az egyes elemekhez tartozó valószínűségekből képzett kumulált részletösszegekre a következő értékeket kapjuk:

Megnevezés	Elem indexe					
	2	3	4	5	6	7
Kumulált valószínűség	0,2165	0,6155	1,0044	1,3259	1.6111	2,0000

Az eljárás kezdő értékét egy, a lépésköznél nem nagyobb véletlen szám határozza meg, erre a véletlenszám-generátor a $k_1 = 0{,}4915$ értéket eredményezte. A következe-

ző (és egyben utolsó) „kereső” érték egyenlő a $k_1 +$ lépésköz összeggel, azaz $k_2 = 1{,}4915$. Mivel $0{,}2165 < k_1 \leq 0{,}6155$ és $1{,}3259 < k_2 \leq 1{,}6111$, a keresett minta második eleme $i_2 = 3$, harmadik eleme pedig $i_3 = 6$. A minta tehát a sokaság 1, 3 és 6 indexű elemeiből áll.

3. A *p_ij*-módszer tulajdonságai

A πps mintavételi eljárások irodalma rendkívül terjedelmes, logikusan felvethető a kérdés, hogy a *p_ij*-módszer alkalmazása milyen körülmények között előnyösebb, mint a tekintett témakörben 1962 óta megjelenteké, és egyáltalán milyen értelemben tekinthető újnak. A kérdésre adható válasz érdekében *Bondesson* [2012] cikkéből célszerű kiindulnunk.

Az említett tanulmány címe (On sampling with prescribed second-order inclusion probabilities – Mintavétel az előírt másodrendű bekerülési valószínűségekkel) majdnem azonos a jelen dolgozat címével, és hasonló a helyzet a két publikáció tárgyát illetően is. A hasonlóságok mellett azonban jelentős különbségek is vannak a két írásban a kitűzött célok és elért eredmények között.

Bondesson célja a másodrendű bekerülési valószínűségek előzetes kijelölésével egyrészt az volt, hogy bizonyos becsült értékösszegek szórásnégyzete adott körülmények között minimális legyen, másrészt pedig, hogy a mintavételi terv entrópiája – lásd a definíciót a jelölések között – maximális legyen. Az entrópia maximalizálására irányuló törekvés a πps -mintákkal, mintavételi tervekkel kapcsolatos kutatásokban az 1990-es évek közepétől figyelhető meg (*Chen–Dempster–Liu* [1994], *Soofi* [1994]). A kutatásokat az motiválja, hogy magas entrópia esetén a nem tipikus adottságokkal rendelkező minták esélye a kiválasztásra viszonylag kisebb. A *p_ij*-módszer kidolgozásának viszont az volt a célja, hogy amennyiben rendelkezésünkre áll az első- és a másodrendű bekerülési valószínűségeknek az /1/–/2/ feltételeket kielégítő konzisztens rendszere, akkor a lehető legegyszerűbb módon adjunk meg egy olyan mintavételi eljárást, amelynél a mintából származó becslésekre a /4/ Sen–Grundy–Yates-formula alkalmazható. A kétféle megközelítést egymás mellé helyezve a következőket mondhatjuk.

Bondesson számára imperatívusz volt célja eléréséhez a másodrendű feltételes Poisson-mintavétel[2] használata, ami tekintélyes matematikai apparátus alkalmazását

[2] A másodrendű feltételes Poisson-mintavétel bonyolultabb módszer, mint az (egyszerű) feltételes Poisson-mintavétel.

és ugyancsak tekintélyes gépidő felhasználását jelentette. Módszere a Gibbs-mintavételnek köszönhetően maximális hatékonyságot eredményezett az entrópia-maximalizáló mintavételi eljárások körében, az alkalmazhatóság felső korlátját a sokaság $N = 250$ elemszáma közelében lehet megvonni. Ugyanez az elemszám a lényegesen kevésbé ambiciózus célkitűzés jegyében kidolgozott *p_ij*-módszer esetén nem problematikus, bár megfelelő szoftver készítése esetén bizonyos takarékossági szempontok figyelembe vételére ösztönözhet.

Mindebből azt a következtetést lehet levonni, hogy lehetnek olyan esetek, amikor indokolt a magas entrópiára való törekvés, és érdemes az ezzel kapcsolatos áldozatot meghozni, de lehetnek olyan esetek is, amikor kisebb az entrópia jelentősége. Az utóbbi esetben célszerű lehet az egyszerűbb és lényegesen kevésbe számításigényes *p_ij*-módszer alkalmazása. Olyan értékeléssel vagy összehasonlítással, hogy a magas entrópiáról való lemondás milyen következményekkel jár, egyelőre ritkán találkozunk.

Annak feltételezésével, hogy rendelkezésünkre áll egy algoritmus vagy szoftver, amely az elsőrendű bekerülési valószínűségek ismeretében előállítja a másodrendű bekerülési valószínűségek konzisztens rendszerét, összehasonlítottuk a *p_ij*-módszert a mintavétel módjának, valamint a szórásnégyzet-becslés lehetőségének a szempontjából a következő standard πps mintavételi módszerekkel:

– feltételes Poisson-mintavétel (*Hájek* [1964], [1981]; *Chen–Dempster–Liu* [1994]);
– *Sampford*-féle [1967] mintavétel;
– *Sunter* [1986] szekvenciális módszere.

Szórásbecslés céljára mind a *p_ij*-módszer, mind pedig az utóbbi három eljárás egzakt másodrendű bekerülési valószínűségeket használ, ezek a Sampford-mintavételnél és kedvező esetben a *p_ij*-módszernél is zárt alakban, a másik két módszernél rekurzív, illetve iteratív számítás eredményeként állnak rendelkezésre. E tekintetben tehát nincs lényeges különbség az említett módszerek között.

A mintavétel módja szerint a *p_ij*- és a Sunter-féle szekvenciális módszert a szekvenciális eljárások közé, a Sampford- és a feltételes Poisson-mintavételt pedig az elfogadó-elutasító stratégiát alkalmazó módszerek közé soroljuk. Tekintsük először a két szekvenciális eljárást. Mind a kettő a sokaság elemeinek rendezésével kezdődik, a *p_ij*-módszernél – amely a randomizált szisztematikus metódus változatának tekinthető – véletlen sorrendre, a Sunter módszerénél pedig a bekerülési valószínűség nagysága szerint csökkenő sorrendre van szükség. Nem szükséges minden egyes mintavételnél újra rendezni a sokasági elemeket. A *p_ij*-módszernél először N elemből választunk egyet, majd ezután a randomizált szisztematikus kiválasztás szabályai szerint $N-1$ elemből $(n-1)$ elemet; egy kezdeti véletlen számra és egy

$n-1$ elemű számtani sorozat meghatározására van szükség. Ennek minden egyes eleménél az azt közre fogó két sokaságbeli elem közül a kisebb kerül a mintába. Sunter módszerénél a sokaság elemei a kialakított sorrend szerint egymás után vesznek részt egy Bernoulli-kísérletben, amelynek kimenetele szerint vagy bekerülnek a mintába, vagy nem. Végeredményben mindkét mintavételnél $O(N)$ számú összehasonlításra van szükség.

Tekintsük most a Sampford- és a feltételes Poisson-mintavételt. Mint említettük, ezek az elfogadás-elvetés stratégiáját alkalmazzák, ami azt jelenti, hogy egymás után n elemű mintákat generálnak, amíg végre csupa különböző elemekből álló mintát nem találnak. Nyilvánvaló, hogy ez sokkal hosszabb számítási időt igényel, mint a sokaság elemeinek egymás utáni megfigyelése, összehasonlítása. Következésképpen a *p_ij*-módszernek és Sunter szekvenciális módszerének a számításigénye nagyságrendben egyenlő, míg a Sampford-mintavételnek és a feltételes Poisson-mintavételnek a számításigénye ennél nagyobb.

A πps mintavételi módszereket tanulmányozva azt láthatjuk, hogy a másodrendű bekerülési valószínűségek kezelése valamilyen formában mindig része a mintavételi tervnek. A Sampford-mintavétel esetében például egy részprogram a π_{ij} valószínűségeket analitikus formában állítja elő a π_i valószínűségekből. A Bondesson által megoldott feladatban, amelyben a másodrendű bekerülési valószínűségek bemenő adatok voltak, az alkalmasan választott mintavételi eljárás – a másodrendű Poisson-mintavétel – valószínűségi függvényének $N \times (N-1)$ számú paraméterét iteratív módszerrel úgy kellett meghatározni, hogy a sokaságbeli elempárok mintába kerülésének valószínűsége a bemenő adatokkal egyezzen meg. Kis pontatlanságot megengedve azt mondhatjuk, hogy a *p_ij*-módszer abban különbözik a többi πps mintavételi tervtől, hogy nem a mintavételi terv határozza meg a másodrendű bekerülési valószínűségeket, hanem ez utóbbiak a mintavételi tervet.

Függelék

Randomizált szisztematikus mintavétel. A sokaság N elemét véletlen sorrendbe rendezzük, és a nagyságukat reprezentáló a_i mennyiségekből kumulált összegeket képezünk a következőképpen: $t_1 = a_1$, $t_2 = t_1 + a_2$, $t_3 = t_2 + a_3$, ..., $T = t_N = t_{N-1} + a_N$. A d lépésközt a $d = T/n$ összefüggéssel definiáljuk, ahol n a minta elemszáma. Választunk egy valós értékű $k_1 < d$ kezdő értéket, és képezzük a k_1, $k_2 = k_1 + d$, $k_3 = k_2 + d$, $k_4 = k_3 + d$, ... sorozatot. A mintába azok a v elemek kerülnek, amelyekhez van a k. sorozatnak egy olyan k_l eleme, amelyre fennáll a $t_{v-1} < k_l \le t_v$ összefüggés (előfordulhat, hogy $t_0 = 0$). A v elem az $a_v = t_v - t_{v-1}$ nagysággal arányos valószínűséggel kerül a mintába. A sokaságbeli elemek nagyságát jellemző a_i mennyiségek lehetnek a π_i bekerülési valószínűségek is.

Irodalom

BONDESSON, L. [2012]: On sampling with prescribed second-order inclusion probabilities. *Scandinavian Journal of Statistics*. Vol. 39. Issue 4. pp. 813–829. http://dx.doi.org/10.1111/j.1467-9469.2012.00808.x

BREWER, K. W. R. [1963]: A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*. Vol. 5. Issue. 1. pp. 5–13. http://dx.doi.org/10.1111/j.1467-842X.1963.tb00132.x

BREWER, K. R. W. – DONADIO, M. E. [2003]: The high entropy variance of the Horvitz–Thompson estimator. *Survey Methodology*. Vol. 29. No. 2. pp. 189–196.

CHEN, X. H. – DEMPSTER, A. P. – LIU, J. S. [1994]: Weighted finite population sampling to maximize entropy. *Biometrika*. Vol. 81. Issue 3. pp. 457–469. http://dx.doi.org/10.1093/biomet/81.3.457

DURBIN, J. [1967]: Design of multi-stage surveys for estimation of sampling error. *Applied Statistics*. Vol. 16. No. 2. pp. 152–164. http://dx.doi.org/10.2307/2985777

HÁJEK, J. [1964]: Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics.* Vol. 35. No. 4. pp. 1491–1528. http://dx.doi.org/10.1214/aoms/1177700375

HÁJEK, J. [1981]: *Sampling from a Finite Population*. Marcel Dekker. New York.

HARTLEY, B. G. – RAO, J. N. K. [1962]: Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics.* Vol. 33. No. 2. pp. 350–374. http://dx.doi.org/10.1214/aoms/1177704564

HORVITZ, D. G. – THOMPSON, D. J. [1952]: A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association.* Vol. 47. Issue 260. pp. 663–685. http://dx.doi.org/10.1080/01621459.1952.10483446

RAO, J. N. K. [1965]: On two simple schemes of unequal probability sampling without replacement. *Journal of Indian Statistical Association*. Vol. 3. pp. 173–180.

SAMPFORD, M. R. [1967]: On sampling without replacement with unequal probabilities of selection. *Biometrika*. Vol. 54. No. 3–4. pp. 499–513. http://dx.doi.org/10.2307/2335041

SEN, A. R. [1953]: On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*. Vol. 5. No. 2. pp. 119–127.

SOOFI, E. S. [1994]: Capturing the intangible concept of information. *Journal of the American Statistical Association*. Vol. 89. Issue 428. pp. 1243–1254. http://dx.doi.org/10.1080/01621459.1994.10476865

SUNTER, A. B. [1977]: List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*. Vol. 26. No. 3. pp. 261–268. http://dx.doi.org/10.2307/2346966

SUNTER, A. B. [1986]: Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*. Vol. 54. No. 1. pp. 33–50. http://dx.doi.org/10.2307/1403257

YATES, F. – GRUNDY, P. M. [1953]: Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*. Vol. 15. No. 2. pp. 253–261.

Summary

A simple method for selecting a sample of fixed size from a finite universe with probability proportional to size and without replacement is introduced in the paper provided that besides the first order inclusion probabilities a consistent set of the second order inclusion probabilities is also given.

Helyreigazítás

A „Nagysággal arányos valószínűség szerinti mintavétel másodrendű bekerülési valószínűségekkel, visszatevés nélkül” című tanulmányhoz (*Statisztikai Szemle*. 94. évf. 3. sz. 233–245. old. DOI: 10.20311/stat2016.03.hu0233).

Az 1. tétel második állításának a bizonyítása túlságosan szűkszavú, pontatlan. A bizonyítás pontos megfogalmazása a következő.

A j és az i sokaságbeli elemek együttes bekerülésének valószínűsége vizsgálatánál azt kell észrevennünk, hogy az algoritmus 1. lépésében egy sokaságbeli elemet, éspedig a j elemet választjuk ki, a 2. lépésben pedig tulajdonképpen a $(j, 1)$, …, $(j, j-1)$, $(j, j+1)$, …, (j, N) elempárok közül választunk ki $n-1$ számút a π_{kl} elemekből álló mátrix j-edik sorában. A bizonyítás első részében láttuk, hogy a $(\pi_j / n)\, \pi_{ji} / \pi_j$ valószínűség annak a valószínűségnek a része, összetevője, amellyel az i elem bekerül a teljes sokaságnak egy n elemű mintájába; ugyanez a kifejezés része annak a valószínűségnek is, amellyel a (j, i) elempárt tartalmazza egy n elemű minta. Ha az algoritmus 1. lépését arra korlátoznánk, hogy a teljes sokaságnak csak egy rögzített j elemét lehet kijelölni, akkor a (j, i) elempár bekerülési valószínűsége π_{ji} / n lenne. Az algoritmus azonban – módosítás, korlátozás hiányában – minden n elemű mintát pontosan n-szer állít elő, annak függvényében, hogy melyik elemét választjuk ki az 1. lépésben. Ennek következtében a (j, i) – és ezzel együtt az (i, j) – elempár bekerülési valószínűsége $n \times \pi_{ji} / n = \pi_{ji}$. Ezzel az állítással kapcsolatban meg kell még jegyeznünk a következőt: az elemek egy i_1, i_2, …, i_n mintája meghatározza a belőle kiválasztható összes (i_k, i_l) elempárt, $k \neq l$, az utóbbiak közül pedig bármely $n-1$ darab meghatározza az elemekből álló n elemű mintát, valamint az összes többi elempárt, valamennyit a $\{\pi_{ij}\}_{N \times N}$ mátrixnak megfelelő bekerülési valószínűséggel, feltéve, hogy a párokban szereplő $2(n-1)$ számú i_k azonosító tartalmazza a minta elemeinek i_1, i_2, …, i_n azonosítóit. A szükséges konzisztenciát a mátrix tulajdonságai biztosítják.

Mihályffy László,
a Központi Statisztikai Hivatal ny. statisztikai főtanácsadója
E-mail: Laszlo.Mihalyffy@ksh.hu