

Tehnici de esantionare cu probabilitati neegale

Prof.dr. Vergil VOINEAGU, conf.dr. Tudorel ANDREI
Catedra de Statistica si Previziune Economica, A.S.E. Bucuresti

În cadrul acestui articol sunt prezentate principalele aspecte teoretice ale esantionarii cu probabilitati neegale. Sunt descrise avantajele acestei tehnici, comparativ cu tehnicile de esantionare cu probabilitati egale. Sunt descrise câteva tehnici de implementare a unei tehnici de esantionare cu probabilitati neegale.

Cuvinte cheie: esantionare, volum, variabila, estimare, probabilitati, estimatori.

Esantionarea cu probabilitati neegale fara revenire

Se considera cazul în care esantionul este de volum $n > 1$, iar alocarea unitatilor în esantion se realizeaza cu tehnica esantionarii cu revenire. Probabilitatea ca o unitate sa fie inclusa în esantion este calculata pe baza relatiei:

$$p_i = P(\text{unitatea } i \text{ este alocata}) = \frac{X_i}{\sum_{i=1}^N X_i}, \quad [1]$$

unde x este variabila pentru care estimam o serie de parametri pentru caracterizarea tendintei centrale sau gradului de dispersare. Cum valorile acestei variabile nu sunt cunoscute, atunci pentru definirea acestor probabilitati se considera realizarile altei variabile, corelata cu aceasta.

Probabilitatea ca o unitate sa fie inclusa într-un esantion de volum n este notata prin p_i . Aceasta este determinata prin relatia:

$$p_i = P(u_i \in E_n) = 1 - P(u_i \neq u_{i_j}, j = \overline{1, n}) \quad [2]$$

$$= 1 - (1 - p_i)^n.$$

Pentru cazul în care $n=1$ se obtine egalitatea $p_i = p_i$. În cazul acestei tehnici de esantionare fiecare extragere este independenta de cele precedente, întrucât structura populatiei nu se schimba de la o alocare la alta. Se noteaza prin q_i numarul de aparitii ale unitatii u_i în cadrul unui esantion de volum n . Variabila q_i are urmatoarele caracteristici: $E(q_i) = np_i$, $\text{var}(q_i) = np_i(1 - p_i)$, iar $\text{cov}(q_i, q_j) = -np_i p_j$, $i \neq j$. Aceste trei caracteristici sunt evidente, întrucât

$$(q_1, q_2, \dots, q_N) \rightarrow B(n; p_1, p_2, \dots, p_N).$$

Un estimator natural pentru volumul total al caracteristicii X este: $\hat{T} = \frac{1}{n} \sum_{i \in E_n} \frac{X_i}{p_i} q_i$. [3]

Proprietatea 1. Estimatorul mai sus definit este nedeplasat. Pentru a demonstra aceasta proprietate se aplica operatorul de medie termenilor egalitatii [3]:

$$E(\hat{T}) = \frac{1}{n} \sum_{i \in E} E\left(\frac{X_i}{p_i}\right) np_i = \sum_{i \in E} \left(\sum_{k=1}^N \frac{X_k}{p_k} p_k\right) p_i = T. \quad [4]$$

Proprietatea 2. Varianta estimatorului în cazul esantionarii cu probabilitati neegale este mai mica decât în cazul esantionarii cu probabilitati egale. Pentru a demonstra aceasta proprietate se tine seama de expresiile celor doua variante. Astfel,

- pentru cazul esantionarii simple cu probabilitati egale, cu revenire s-a obtinut:

$$\text{var}(\hat{T}) = \frac{N^2}{n} \sum_{i=1}^N (X_i - N \cdot T)^2 = \frac{1}{n} \left[\sum_{i=1}^N \frac{X_i^2}{1/N} - T^2 \right] \quad [5]$$

- pentru cazul esantionarii cu probabilitati neegale:

$$\text{var}(\hat{T}) = \frac{1}{n} \sum_{i=1}^N \left(\frac{X_i}{p_i} - T\right)^2 p_i = \frac{1}{n} \left[\sum_{i=1}^N \frac{X_i^2}{p_i} - T^2 \right] \quad [6]$$

Verificarea relatiei de ordine dintre cele doua variante revine la a compara termenii

$$\sum_{i=1}^N \frac{X_i^2}{p_i} \text{ si } \sum_{i=1}^N \frac{X_i^2}{1/N}.$$

Se obtine apoi forma echivalenta a inegalitatii variantelor:

$$\sum_{i=1}^N X_i^2 \left(\frac{1}{p_i} - \frac{1}{1/N} \right) < 0.$$

Ultima inegalitate este îndeplinita, daca pentru valori mari ale variabilei X sunt valabile inegalitatile $\frac{1}{N} < p_i$, iar pentru valori mici, inegalitatile sunt de sens contrar. În acest caz regasim fundamentul sondajului cu probabilitati neegale, ca valorile carac-

teristicii sa fie corelate pozitiv cu probabilitatile de alocare în esantion.

Varianta estimatorului definita prin relatia [6] se scrie sub forma echivalenta:

$$\text{var}(\hat{T}) = \frac{1}{n} \sum_{i=1}^N \sum_{j=1}^N \left(\frac{X_i}{p_i} - \frac{X_j}{p_j} \right)^2 p_i p_j. \quad [7]$$

Un estimator al variantei volumului total al caracteristicii este:

$$\hat{\text{var}}(\hat{T}) = \frac{1}{n(n-1)} \sum_{i \in E_n} \left(\frac{x_i}{p_i} - \hat{T} \right)^2. \quad [8]$$

$$E(\hat{\text{var}}(\hat{T})) = \frac{1}{n(n-1)} E \left(\sum_{i \in E_n} \left(\frac{x_i}{p_i} - T \right)^2 \right) - \frac{1}{n-1} E((\hat{T} - T)^2) = \frac{1}{n(n-1)} \sum_{i \in E_n} \text{var}(\hat{T}) - \frac{1}{n-1} \text{var}(\hat{T}) = \text{var}(\hat{T}).$$

Se prezinta, în cele ce urmeaza, *metoda frecventelor cumulate* pentru definirea unui esantion, ce are la baza tehnica esantionarii simple cu probabilitati neegale, cu revenire. Pentru aplicarea acestei tehnici se parcurg urmatoarele etape:

- Se genereaza un sir de numere aleatoare folosind un generator al repartitiei uniforme pe intervalul [0, 1]. Consideram sirul de numere $(u_1, u_2, \dots, u_n)$;

- Folosind realizările unei variabile aleatoare, pentru care dispunem de seria de date, se calculeaza probabilitatile $(p_1, p_2, \dots, p_N)$;

- Este inclusa în esantion unitatea statistica pentru care este satisfacuta relatia de ordine: $C(p_{i-1}) < u_k \leq C(p_i)$, $k = \overline{1, n}$, unde

$C(p_i) = \sum_{j=1}^i p_j$ reprezinta seria probabilitatilor

cumulate crescator.

Exemplul 1. Pentru 5 magazine de vânzare a produselor alimentare se cunosc datele din tabelul 1. Se considera cazul în care, pentru estimarea volumului total al vânzarilor, în situatia în care nu se cunosc vânzarile pentru fiecare magazin, se organizeaza un sondaj statistic cu dimensiunea esantionului $n=1$.

$$\begin{aligned} \text{var}(\hat{T}) &= \frac{1}{5} [(500 - 2000)^2 + (1200 - 2000)^2 + (795 - 2000)^2 + (2305 - 2000)^2 + (5200 - 2000)^2] \\ &= 2935010. \end{aligned}$$

Rezultatele obtinute sunt putin performante, chiar daca estimatorul este nedeplasat.

Proprietatea 3. Estimatorul variantei volumului total al caracteristicii este un estimator nedeplasat. Pentru a demonstra proprietatea se tine seama ca [8] se scrie sub forma echivalenta:

$$\hat{\text{var}}(\hat{T}) = \frac{1}{n(n-1)} \sum_{i \in E_n} \left(\frac{x_i}{p_i} - T \right)^2 - \frac{1}{n-1} (\hat{T} - T)^2 \quad [9]$$

Aplicând operatorul de medie egalitatii [9] se obtin succesiv rezultatele urmatoare:

Tabelul 1.

Magazin	Nr. de angajati	Vânzari (mld. lei)
A	2	100
B	5	240
C	3	159
D	10	461
E	20	1 040
Total	40	2 000

Pentru fiecare magazin se cunoaste numarul de angajati. Se va estima volumul total al vânzarilor folosind tehnica simpla de esantionare si esantionarea cu probabilitati inegale. Pentru cazul din urma se considera ca volumul vânzarilor este determinat în mod direct de numarul de angajati. Pentru a caracteriza performantele estimatorilor în conditiile folosirii uneia din cele doua tehnici de esantionare se calculeaza indicatorii: $B(\hat{T})$, $\text{var}(\hat{T})$, $EMP(\hat{T})$.

- **Esantionarea simpla.** Estimatiile pentru volumul total al vânzarilor sunt trecute în tabelul 1. Se observa ca, în situatia folosirii acestei metode de esantionare, estimatorul folosit este nedeplasat, deci $E(\hat{T}) = 2000$, iar varianta este:

Dealtfel, pentru datele de mai sus, se obtine o valoare foarte mare a coeficien-

tului de variatie a estimatorului, egal cu 85.66%. Pentru a diminua varianta estimatorului se recurge la o alta metoda de esantionare, care tine seama de importanta fiecărei unitati în cadrul populatiei. Pentru a caracteriza dimensiunea fiecărei unitati se considera realizarile pentru variabila numar de angajati.

- **Esantionarea cu probabilitati neegale.** Se calculeaza pentru fiecare unitate pro-

babilitatea de a accede în esantion, prin relatia: $p_i = \frac{n_i}{\sum_j n_j}$.

Rezultatele sunt prezentate în tabelul 2. În cazul acestei metode de esantionare cele 5 estimatii, precum si deplasarea acestora, sunt trecute în tabelul 2. Urmărind rezultatele de mai sus se obtine, pentru fiecare esantion, marimea deplasării, iar pe ansamblu se evalueaza deplasarea estimatorului, aceasta fiind zero. Astfel,

$$B(\hat{T}) = [2000 \cdot \frac{1}{20} + 1920 \cdot \frac{1}{8} + 2120 \cdot \frac{3}{40} + 1844 \cdot \frac{1}{4} + 2080 \cdot \frac{1}{2}] - 2000 = 0.$$

Se calculeaza varianta estimatorului

$$\begin{aligned} \text{var}(\hat{T}) &= (2000 - 2000)^2 \cdot \frac{1}{20} + (1920 - 2000)^2 \cdot \frac{1}{8} + (2120 - 2000)^2 \cdot \frac{3}{40} + (1844 - 2000)^2 \cdot \frac{1}{4} \\ &+ (2080 - 2000)^2 \cdot \frac{1}{2} = 9280. \end{aligned}$$

Tabelul 2.

	n_i	$p_i = n_i / \sum_i n_i$	Esan- tionul	$\hat{T} = \sum_{i \in E} \frac{x_i}{p_i}$	$B(\hat{T}) = \hat{T} - T$
A	2	2/40	100	$100 \cdot (40/2) = 2000$	0
B	5	5/40	240	$240 \cdot (40/5) = 1920$	-80
C	3	3/40	159	$159 \cdot (40/3) = 2120$	120
D	10	10/40	461	$461 \cdot (40/10) = 1844$	-156
E	20	20/40	1 040	$1040 \cdot (40/20) = 2080$	80
Total	40	1	2 000 -		-

În acest caz varianta estimatorului este cu mult mai mica decât cel al esantionării simple. Deplasările estimatorilor în condițiile aplicării celor două tehnici de esantionare sunt reprezentate grafic în figura 1.

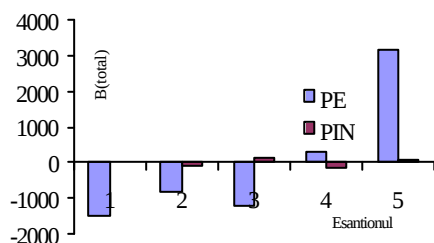


Fig. 1. Deplasarea estimatorilor cele două tipuri de esantionari

Se observa o mai mare stabilitate a estimatiilor în jurul valorii reale, în cazul tehnicilor de esantionare cu probabilitati neegale.

Esantionarea cu probabilitati neegale, cu revenire

Tehnica de esantionare cu revenire, mai sus prezentata, ofera o serie de avantaje în privinta definirii estimatorilor si proprietăților acestora, dar rezultatele obtinute sunt mai puțin eficiente comparativ cu tehnica de esantionare fara revenire. În cazul acestei tehnici, probabilitatile de alegere a unitatilor simple în cadrul esantionului se modifica de la o etapa la alta în functie de unitatile care au fost deja incluse în cadrul esantionului până la o anumita etapa.

Se noteaza si în acest caz prin p_i probabilitatea ca unitatea i sa fie inclusa în cadrul esantionului. Definim prin $p_{i(j)}$ probabilitatea de a selecta la a doua extragere unitatea i , stiind ca la prima s-a extras unitatea j . Aceasta se calculeaza pe baza relatiei: $p_{i(j)} = \frac{p_i}{1 - p_j}$.

Se calculeaza probabilitatea de a alege la prima extragere unitatea i , iar la a doua unitatea j pe baza relatiei: $p_{(i-1, j-2)} = p_i \frac{p_j}{1-p_j}$.

$$P(i, j) = P(i-1, j-2) + P(j-1, i-2) = p_i \frac{p_j}{1-p_j} + p_j \frac{p_i}{1-p_i}. \quad [10]$$

Pentru un esantion de volum 2 se calculeaza probabilitatea ca o unitate sa fie inclusa în cadrul esantionului pe baza relatiei: $p_i = \sum_{\substack{j=1 \\ j \neq i}}^N p_{ij}$. [11]

Pentru un esantion de volum n sunt valabile urmatoarele doua egalitati:

$$\sum_{i=1}^N p_i = n \text{ si } \sum_{\substack{i, j \\ i \neq j}} p_{ij} = n(n-1). \quad [12]$$

Ultimele doua relatii sunt evidente, tinând seama de formula de calcul a probabilitatilor ce intervin în aceste relatii. Se defineste variabila binara Z_i , astfel:

$$Z_i = \begin{cases} 1, & u_i \in E_n \\ 0, & u_i \notin E_n \end{cases}$$

$$P(u_1-1, u_2-2, \dots, u_n-n) = p_1 \frac{p_2}{1-p_1} \frac{p_3}{1-(p_1+p_2)} \dots \frac{p_n}{1-(p_1+p_2+\dots+p_{n-1})}. \quad [13]$$

Exemplul 2. Pentru datele din exemplul 1 se calculeaza p_{ij} . Rezultatele sunt pre-

Se determina atunci probabilitatea, ca dupa doua extrageri, unitatile incluse în cadrul esantionului sa fie i si j :

Cele doua probabilitati care intervin în cadrul relatiilor [12] sunt determinate pe baza formulelor: $p_i = E(Z_i)$, si $p_{ij} = E(Z_i Z_j)$. În aceste conditii se demonstreaza egalitatile din [12]:

• Pentru prima relatie se obtine succesiv: $\sum_{i=1}^N p_i = \sum_{i=1}^N E(Z_i) = E(\sum_{i=1}^N Z_i) = n$.

• Pentru a doua relatie se stabilesc urmatoarele rezultate:

$$\begin{aligned} \sum_{ij=1, i \neq j}^N p_{ij} &= \sum_{ij=1, i \neq j}^N E(Z_i Z_j) = \\ E(\sum_{i, j, i \neq j}^N Z_i Z_j) &= E(\sum_i Z_i \sum_{j \neq i} Z_j) = n(n-1). \end{aligned}$$

În general, pentru un esantion de volum n se defineste probabilitatea:

zentate în cadrul tabelului 3. Urmărind datele din acest tabel, se observa ca $\sum_{i=1}^5 p_i = 2$.

Tabelul 3. Calcularea probabilitatilor p_{ij} si p_i .

	A	B	C	D	E	p_i
A	-	0.013722	0.008001	0.029825	0.076316	0.127864
B	0.013722	-	0.020849	0.077381	0.196429	0.308381
C	0.008001	0.020849	-	0.04527	0.115541	0.189662
D	0.029825	0.077381	0.04527	-	0.416667	0.569142
E	0.076316	0.196429	0.115541	0.416667	-	0.804952
p_i	0.127864	0.308381	0.189662	0.569142	0.804952	2

Se defineste pentru volumul total al caracteristicii urmatorul estimator:

$$\hat{T}_{HT} = \sum_{i \in E_n} \frac{X_i}{p_i} = \sum_{i=1}^N \frac{X_i}{p_i} Z_i. \quad [14]$$

$$E(\hat{T}_{HT}) = \sum_{i=1}^N E\left(\frac{X_i}{p_i}\right) E(Z_i) = \sum_{i=1}^N E(X_i) = N \cdot E(X_i) = T. \quad [15]$$

Proprietatea 2. Varianta estimatorului [14] este reprezentata prin expresia:

$$\text{var}(\hat{T}) = \sum_{i, j, i < j} \left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2 (p_i p_j - p_{ij}). \quad [16]$$

Proprietatea 1. Estimatorul Horvitz-Thompson este un estimator nedeplasat. Daca se aplica operatorul de medie estimatorului H-T se obtine:

Pentru a demonstra aceasta relatie se aplica operatorul de varianta celor doi termeni ai egalitatii [16]. Se obtine rezultatul urmator:

$$\begin{aligned}
\text{var}(\hat{T}_{HT}) &= \text{var}\left(\sum_{i=1}^N \frac{X_i}{p_i} Z_i\right) = \sum_{i=1}^N \text{var}\left(\frac{X_i}{p_i} Z_i\right) + \sum_{i=1}^N \sum_{j>i}^N \text{cov}\left(\frac{X_i}{p_i} Z_i, \frac{X_j}{p_j} Z_j\right) \\
&= \sum_{i=1}^N \frac{X_i^2}{p_i^2} \text{var}(Z_i) + \sum_{i=1}^N \sum_{j>i}^N \frac{X_i X_j}{p_i p_j} \text{cov}(Z_i, Z_j) \\
&= \sum_{i=1}^N \frac{X_i^2}{p_i^2} (p_i - p_i^2) + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{X_i X_j}{p_i p_j} (p_{ij} - p_i p_j) \\
&= \sum_{i,j,i<j}^N \left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2 (p_i p_j - p_{ij})
\end{aligned}$$

Un estimator natural al variantei volumului total al caracteristicii este:

$$\hat{\text{var}}(\hat{T}) = \sum_{i,j=1,i<j}^n \left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2 \frac{p_i p_j - p_{ij}}{p_{ij}}. \quad [17]$$

Cazul în care probabilitatile sunt egale nu este decât un caz particular al tehnicii de esantionare cu probabilitati neegale, fapt care se demonstreaza. Dealtfel, în acest caz sunt valabile urmatoarele:

$$p_i = 1/N, p_{ij} = P(Z_i = 1, Z_j = 1) = \frac{n(n-1)}{N(N-1)}.$$

Particularizând [14] pentru acest caz se obtine: $\hat{T}_{HT} = \sum_{i \in E_n} \frac{X_i}{n/N} = N \bar{X}$. [18]

În mod asemanator se procedeaza în cazul variantei. Pentru cazul în care probabilitatile sunt calculate în raport cu repar-

$$C(p_{i-1}) < u + (i-1) \leq C(p_i), i = \overline{2, N}. \quad [19]$$

Se verifica aceasta relatie, pâna la alocarea tuturor unitatilor esantionului.

Exemplul 3. Consideram o populatie constituita din 7 unitati elementare. Probabilitatile de introducere a fiecărei unitati în esantion si seria cumulata a probabilitatilor sunt trecute în tabelul 4.

Tabelul 4.

Unitate	p_i	p_i	$C(p_i)$
a			
1	0.050	0.100	0.100
2	0.075	0.150	0.250
3	0.155	0.310	0.560
4	0.265	0.530	1.090
5	0.275	0.550	1.640
6	0.150	0.300	1.940
7	0.030	0.060	2.000

tizarea pe unitati a valorilor unei variabile corelata cu prima, atunci $p_i = Y_i / \sum Y_i$. Se obtin în aceasta situatie urmatoarele rezultate:

- volumul total al caracteristicii X este

$$\text{estimat prin } \hat{T}_{HT} = \sum_{i \in E_n} \frac{X_i}{Y_i / \sum Y_k};$$

- varianta acestui estimator este estimata

$$\text{prin } \hat{\text{var}}(\hat{T}) = \frac{1}{n} \sum_{i,j \in E_n, i < j} \left(\frac{X_i}{Y_i} - \frac{X_j}{Y_j}\right)^2 Y_i Y_j.$$

Prezentam în cele ce urmeaza *esantionarea sistematica* în conditiile folosirii metodei de esantionare cu probabilitati neegale fara revenire. Reamintim ca, $p_i = n \cdot p_i$, iar pentru calcularea probabilitatilor p_i se recurge la urmatoarea formula de calcul: $p_i = \frac{Y_i}{\sum Y_i}$.

Pentru generarea unui esantion de volum stabilit se parcurg urmatoarele etape:

- se genereaza un numar aleator $u \in U[0,1]$;
- se calculeaza frecventele cumulate crescator $C(p_i) = \sum_{h=1}^i p_h, i = \overline{1, N}$;
- unitatea u_i este inclusa în esantion, daca este satisfacuta relatia:

Folosind un generator de numere aleatoare se obtine numarul $u = 0.382$. Pentru a constitui un esantion de volum 2 se parcurg mai multe iteratii. Se obtin succesiv rezultatele urmatoare: i) $0.382 \in (0.100 \ 0.250]$, deci a doua unitate este inclusa în esantion; ii) $1.382 \in (1.090 \ 1.640]$, deci a patra unitate este inclusa în esantion.

Bibliografie

1. Andrei, T., Stancu, S. *Statistica-teorie si aplicatii-editia a II-a*. Editura All, Bucuresti, 2000.
2. Mihoc, Gh., Urseanu, V. *Sondaje si estimatii statistice-teorie si aplicatii*. Editura Tehnica, Bucuresti, 1977
3. Sharon, L., L. *Sampling: Design and Analysis*. Brooks/Cole Publishing Company. London. 1999.