

Probleme ale tehnicilor de sondaje de tip cluster

Prof.dr. Vergil VOINEAGU, prep. Dan PELE, Mircea-Alexandru VOINEAGU
Catedra de Statistica si Previziune Economica, A.S.E. Bucuresti

Cercetarea printr-un sondaj statistic a caracteristicilor unei populatii simple poate fi dificila, din câteva motive naturale: inexistenta unei baze de sondaj formata din date corecte, actualizate si imposibilitatea construirii unei asemenea baze de sondaj; costurile prea mari pentru intervievarea unor esantioane relativ mari, cu o larga raspândire geografica. De foarte multe ori, este mai avantajos sa consideram populatia tinta ca fiind formata din grupuri (clusters), fiecare grup continând la rândul sau un numar de unitati de observatie. În general, populatia este organizata în mod natural în astfel de grupuri, cum ar fi familiile, scolile, etc.

Cuvinte cheie: sondaj, cluster, esantion, estimator, populatie.

Sa presupunem, de exemplu, ca vrem sa estimam numarul total de biciclete deținute de membrii unei comunitati de 1000 de familii. Putem alege un esantion simplu aleator de 400 de familii sau putem împarti comunitatea în grupe de câte 20 de familii si sa esantionam fiecare familie (sau sa subesantionam câteva dintre familii) aparținând fiecăruia din cele 20 de grupuri alese aleator dntre cele 500 care formeaza comunitatea.

În general, sondajul de tip cluster va conferi mai putina precizie decât un sondaj simplu aleator efectuat pe acelasi esantion. Sondajul grupat pare destul de asemanator cu cel stratificat; un cluster, ca si un strat, este format printr-o grupare a membrilor populatiei. Procesul de selectie a unitatilor din esantion este însa esentialmente diferit în cele doua cazuri (anexa 1).

Daca în comparatie cu sondajul aleator simplu, sondajul stratificat creste precizia estimatiei, sondajul grupat realizeaza în general contrariul.

Membrii aceluiasi grup au în general caracteristici apropiate în comparatie cu unitatile selectate aleator din populatia totala. Esantionând fiecare unitate din grup, se repeta partial aceeasi informatie, în loc sa se obtina noi date si informatii, ceea ce duce la scaderea preciziei estimatorilor.

Cu toate acestea, sondajul de tip cluster este destul de des utilizat în practica pentru ca de obicei este mai ieftin si mult mai

convenabil de esantionat în grupuri decât în populatia totala. În esenta, exista doua tipuri de sondaj grupat:

- într-o etapa, când fiecare element din grupul esantionat este luat în considerare;
- în doua etape, când se realizeaza o subselectie în cadrul grupelor (recomandata când volumul grupelor este mare). De obicei, grupele au volume inegale, dar exista cazuri când se lucreaza cu grupe de volume egale (de exemplu în cazul blocurilor cu acelasi numar de apartamente). În cele ce urmeaza vom lucra sub urmatoarele ipoteze: volumul total al populatiei este cunoscut si grupurile si unitatile din grupa sunt selectate cu aceeasi probabilitate.

Notatii:

Fie $U = \{u_1, \dots, u_n\}$ populatie de n grupuri; S desemneaza esantionul de grupuri extras din U iar S_i este esantionul de unitati asociat grupului i , $\forall i = \overline{1, N}$.

Notam y_{ij} - variabila de interes pentru unitatea j din grupul i .

Notatii pentru grupuri

N – numarul de grupuri în populatie;

M_i – volumul grupului “ i ”;

$K = \sum_{i=1}^N M_i$ - numarul total de unitati în populatie;

$$t_i = \sum_{j=1}^{M_i} y_{ij} - \text{totalul în grupul "i"};$$

$$t = \sum_{i=1}^N t_i = \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij} - \text{totalul populației};$$

$$s_t^2 = \sum_{i=1}^N \frac{\left(t_i - \frac{t}{N}\right)^2}{N-1} - \text{varianța inter-grupe.}$$

Notatii pentru unitati

$$\bar{Y}_u = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{y_{ij}}{K} - \text{media populației}$$

$$\bar{y}_{iu} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i} - \text{media pe grupul "i"};$$

$$S^2 = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iu})^2}{K-1} - \text{varianța globală};$$

$$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iu})^2}{M_i-1} - \text{varianța grupeii "i"};$$

$$S_{intra}^2 = \frac{1}{N} \sum_{i=1}^N S_i^2 - \text{varianța intra-grupe.}$$

Notatii pentru esantion

n - numărul de grupuri din esantion;

m_i - numărul de unitati din grupul "i" în esantion;

$$\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i} - \text{media (estimată) pentru}$$

grupul "i";

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} - \text{totalul estimat pentru grupul "i"};$$

$$\hat{t}_{unb} = \sum_{i \in S} \frac{N}{n} \hat{t}_i - \text{estimator nedeplasat}$$

pentru totalul populației;

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2 - \text{varianța estimată}$$

inter-grupe;

$$s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i-1} - \text{varianța estimată în}$$

grupul "i";

$$s_{intra}^2 = \frac{1}{n} \sum_{i \in S} s_i^2 - \text{varianța estimată intra-grupe.}$$

Sondaj de tip cluster într-o etapa

Pentru acest tip de sondaj, dintr-un grup care compune populația, fie toate elementele sunt cuprinse în esantion, fie nici unul. Sondajul grupat într-o etapa este utilizat în cercetările în care costul esantionării unitatilor este neglijabil în comparație cu cel al esantionării grupelor.

Din populația de N grupuri, cel de-al "i"-lea conține M_i unitati. Extragem un esantion simplu aleator de n grupuri și măsurăm valoarea caracteristicii de interes pentru fiecare unitate din cele n grupuri.

Deci în acest caz $M_i = m_i$.

• Grupe de volume egale

Sa presupunem ca $M_i = m_i = M$. Sondajele care privesc populația nu intra de obicei în acest model, dar sondajele din industrie sau agricultura îl pot utiliza cu succes. Astfel, luăm un esantion simplu aleator de n observatii $\{t_u / i \in S\}$; t_i - totalul în grupul i . Într-o cercetare de householding pentru a estima venitul pentru o familie de două persoane, y_{ij} sunt veniturile individuale, t_i - venitul mediu în familia i , $\bar{t}_u$ - venitul mediu pe familie, $\bar{y}_u$ - venitul mediu pe o persoana. Pentru a estima venitul total t , folosim estimatorul:

$$\hat{t} = \frac{N}{n} \sum_{i \in S} t_i.$$

Din rezultatele pentru sondajul aleator simplu avem:

- media: $E(\hat{t}) = t \Rightarrow \hat{t}$ este un estimator nedeplasat pentru t ;

- varianța: $V(\hat{t}) = N^2 \left(1 - \frac{n}{N} \right) \frac{S_t^2}{n}$;

- eroarea

$$\text{standard: } SE(\hat{t}) = N \sqrt{\left(1 - \frac{n}{N} \right) \frac{s_t^2}{n}}.$$

Pentru a estima media $\bar{y}_u$, folosim estimatorul:

$$\hat{y} = \frac{\hat{t}}{NM}, \text{ obținând:}$$

- media: $E(\hat{y}) = \bar{y}_u \Rightarrow \hat{y}$ este un estimator nedeplasat pentru $\bar{y}_u$;

- varianta: $V(\hat{y}) = \left(1 - \frac{n}{N}\right) \frac{S_t^2}{nM^2}$;

- eroarea

standard: $SE(\hat{y}) = \frac{1}{M} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}}$.

Proprietate:

Fie o populatie formata din NM unitati si fie variabilele de selectie y_{ij} , $i = 1, N, j = 1, M$. Fie

$$\text{Din calculul direct} \Rightarrow (NM - 1)S^2 = N(M - 1)S_{\text{intra}}^2 + \frac{N - 1}{M} S_t^2.$$

$$\text{Atunci avem: } V(\hat{t}) < V(\hat{t}^*) \Leftrightarrow S_t^2 < MS^2 \Leftrightarrow S^2 < S_{\text{intra}}^2.$$

Deci, pentru ca selectia dintr-o populatie grupata sa fie preferabila selectiei dintr-o populatie simpla trebuie ca grupele sa fie mai eterogene decât populatia în ansamblu, ceea ce e greu de întâlnit în practica. De aceea se considera ca în general, prin gruparea populatiei se pierde ceva din precizia estimatorului.

$$\hat{t}^* = NM \frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^m y_{ij} \text{ estimatorul totalului}$$

într-o selectie aleatoare fara revenire si $\hat{t}$ estimatorul totalului pentru selectia grupata într-o etapa.

$$\text{Atunci } V(\hat{t}) < V(\hat{t}^*) \Leftrightarrow S^2 < S_{\text{intra}}^2.$$

Demonstratie:

Din proprietatile sondajului simplu aleator avem: $V(\hat{t}^*) = NM \frac{1}{n} (N - n) S^2$.

Definim $deff = \frac{V(\hat{t})}{V(\hat{t}^*)}$ - efect al modului de proiectare (design effect).

Rezulta ca daca $deff < 1$, atunci este preferabil sondajul grupat.

Sa consideram doua populatii artificiale, fiecare având trei straturi cu câte trei elemente pe fiecare strat.

Cluster	Populatia A			Populatia B		
1	10	20	30	9	10	11
2	11	20	32	17	20	20
3	9	17	31	31	32	30

Cele doua populatii au aceleasi elemente, deci avem $\bar{y}_u = 20$ si $s^2 = 84,5$, comune pentru populatiile A si B. În populatia A,

cea mai mare variabilitate o gasim în interiorul grupelor, în vreme ce pentru populatia B faptul e valabil între grupe.

Cluster	Populatia A		Populatia B	
1	20	100	10	1
2	21	111	19	3
3	19	114	31	1
	$S_{\text{intra}}^2 = 111,6$		$S_{\text{intra}}^2 = 1,6$	

Observam urmatoarele:

- populatia A:

$$deff < 1 \Leftrightarrow S^2 < S_{intra}^2 \Leftrightarrow 84,5 < 111,6 \Rightarrow$$

este recomandabil sondajul grupat;

- populatia B:

$$deff > 1 \Leftrightarrow S^2 > S_{intra}^2 \Leftrightarrow 84,5 > 1,6 \Rightarrow$$

este recomandabil sondajul simplu.

• Grupe de volume inegale

În cercetarile sociale grupele sunt rareori de volume egale. Spre exemplu, într-unul din cele mai timpurii sondaje, în SUA în 1937, Enumerative Chek Census a ales un esantion de 2% dintre drumurile postale si chestionarele au fost distribuite tuturor familiilor pe fiecare ruta postala aleasa. Cum rutele postale au un numar diferit de familii, marimea grupelor poate varia foarte mult.

Sondajul într-o etapa cu grupe de volume inegale este o generalizare a celui cu grupe de volume egale; în acest caz avem doua moduri de estimare a totalului si mediei populatiei: folosind estimatori nedeplasati sau estimatori de tip raport.

Estimatie nedeplasata

Un estimator nedeplasat pentru totalul t este:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} t_i \quad \text{cu eroarea standard}$$

$$SE(\hat{t}_{unb}) = N \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}}.$$

Proprietate Eroarea medie patratica a estimatorului $\hat{t}_r$ este mai mica decât dispersia estimatorului nedeplasat

$$E(\hat{t}_r - t)^2 = N^2 \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{i=1}^N (t_i - \bar{y}_u M_i)^2.$$

Demonstratie Avem: $V(\hat{t}) = N^2 \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) S_t^2 = N^2 \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \sum_{i=1}^N (t_i - \bar{y}_i)^2$

Pe de alta parte:

Diferenta între cele doua tipuri de sondaj grupat într-o etapa este ca variatia totalurilor t_i ale grupurilor individuale e de asteptat sa fie mai mare atunci când grupele au marimi diferite. De regula ne asteptam ca totalurile t_i si volumele M_i sa fie corelate: t_i sa fie mai mare când M_i este mare si t_i sa fie mai mic daca volumul M_i este mic. Deci s_t^2 este mai mare atunci când grupele au volume diferite fata de cazul când au volume egale. Din aceasta cauza variatia poate fi mai mare si prin urmare putem obtine o precizie mai mica.

Putem defini un estimator nedeplasat pentru $\bar{y}_u$ si variatia sa.

Fie $K = \sum_{i=1}^N M_i$ numarul total al unitatilor

populatiei. Atunci avem $\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K}$ cu

$$\text{eroarea standard } SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{K}.$$

Estimatorul de tip raport

Vom lua în considerare cazul când ne asteptam ca totalurile t_i sa fie corelate cu M_i si folosind estimatia de tip raport, cantitatile M_i apar ca variabile auxiliare.

Definim estimatorii de tip raport pentru medie si total astfel:

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i}, \quad \hat{t}_r = K \hat{\bar{y}}_r, \quad \text{unde } K = \sum_{i=1}^N M_i.$$

Eroarea medie patratica a estimatorului de tip raport pentru total este:

$$\hat{t} = \frac{N}{n} \sum_{i \in S} \sum_{j \in S_i} y_{ij} \quad \text{atunci când variabilitatea}$$

dintre mediile grupelor este mai mica decât cea dintre totalurile grupelor.

$$E(\hat{t}_r - t)^2 = N^2 \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{i=1}^N M_i^2 (\bar{y}_{iu} - \bar{y}_u)^2$$

Inegalitatea $E(\hat{t}_r - t)^2 < V(\hat{t})$ este echivalenta cu:

$$\sum_{i=1}^N M_i^2 (\bar{y}_{iu} - \bar{y}_u)^2 < \sum_{i=1}^N (t_i - \bar{y}_u)^2.$$

Aceasta conditie este frecvent întâlnita în practica, pentru grupe care variaza mult ca dimensiune.

Observatie Pentru a putea lucra cu estimatorul de tip raport pentru totalul $\hat{t}_r$, este nevoie cunoasterea numarului total K de unitati de populatie.

Sondajul de tip cluster în doua etape

În cazul sondajului grupat într-o etapa, examinam toate elementele din interiorul unui grup din esantion. Însa în multe situatii elementele unui grup pot fi atât de asemanatoare încât cercetând toate unitatile unui grup se pierd resurse: financiare, de timp, etc. În aceste situatii, este mai mult mai ieftin sa elaboram un subesantion în interiorul fiecarui grup esantionat:

- selectam un esantion simplu aleator de volum n din populatia de N grupuri;

- selectam un esantion simplu aleator de volum m_i din fiecare grup i ales în prima etapa (ANEXA2).

Estimatorii pentru t si $\bar{y}_u$ sunt analogi celor din sondajul într-o singura etapa, dar formula variantei este mult mai complicata.

Vom estima totalul pe grupuri prin:

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i$$

iar un estimator nedeplasat pentru totalul populatiei va fi:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} \hat{t}_i = \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i.$$

Dispersia lui $\hat{t}$ are doua componente:

- variabilitatea între grupe;
- variabilitatea unitatilor în interiorul grupelor.

Proprietate Pentru sondajul grupat în doua etape avem:

$$V(\hat{t}_{unb}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^N \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i}.$$

Demonstratie: în [1] (demonstratia se bazeaza pe folosirea notiunii de medie conditionata).

Pentru a estima dispersia $V(\hat{t}_{unb})$, fie

$$s_t^2 = \frac{\sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2}{n-1}$$

$$\hat{V}(\hat{t}_{unb}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}.$$

Evident, eroarea standard a lui $\hat{t}_{unb}$ va fi:

$$SE(\hat{y}_{unb}) = \sqrt{\hat{V}(\hat{t}_{unb})}.$$

Daca este cunoscut numarul total al elementelor populatiei K , putem estima media populatiei prin:

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K}$$

cu eroarea standard a lui $\hat{t}_{unb}$ va fi:

$$SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{K}.$$

Ca si în sondajul într-o singura etapa cu grupe de volume inegale, ambele componente ale dispersiei pot fi afectate de variatiile numarului unitatilor dintr-un grup (M_i) si de variabilitatea $\bar{y}_i$.

Estimatorul de tip raport

Pentru a estima media sau totalul populatiei putem folosi urmatorii estimatori de tip raport:

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i}$$

$$\hat{t}_r = K \hat{\bar{y}}_r = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i} K.$$

Proprietate.

Eroarea medie patratica a estimatorului $\hat{t}_r$ este aproximata perin expresia:

$$E(\hat{t}_r - t)^2 \approx N^2 \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{i=1}^N M_i^2 (\bar{y}_{iu} - \bar{y}_u)^2 + \frac{N}{n} \sum_{i=1}^N M_i^2 \frac{1}{m_i} \left(1 - \frac{m_i}{M_i}\right) \cdot \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iu})^2.$$

Demonstratie: Presupunem ca $\frac{\sum_{i=1}^n M_i}{n}$ este un bun estimator al volumului mediu al grupelor a.î.: $K \approx N \frac{1}{n} \sum_{i=1}^n M_i$.

Cu notatia $\bar{z}_i = \frac{y_i}{m_i} \quad \forall i = \overline{1, n}$ avem

$$\begin{aligned} \text{Rezulta ca: } E(\hat{t}_r - t)^2 &\approx V\left(\frac{N}{n} \sum_{i=1}^n M_i (\bar{z}_i - \bar{y}_u)\right) = \\ &= N^2 \cdot \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{i=1}^N [M_i (\bar{y}_{iu} - \bar{y}_u)]^2 + \frac{N}{n} \sum_{i=1}^N M_i^2 \frac{1}{m_i} \left(1 - \frac{m_i}{M_i}\right) \cdot \\ &\quad \cdot \frac{1}{M_i - 1} \sum_{j=1}^{M_i} [(y_{ij} - \bar{y}_{iu}) - (\bar{y}_{iu} - \bar{y}_u)]^2. \end{aligned}$$

$$\hat{t}_r - t \approx \frac{N}{n} \sum_{i=1}^n M_i (\bar{z}_i - \bar{y}_u).$$

Aceasta variabila aleatoare are forma estimatorului totalului într-o populatie grupata, dar cu variabilele de selectie $(y_{ij} - \bar{y}_u), j = \overline{1, m}, i = \overline{1, n}$.

Observatie Estimatorul mediei $\hat{\bar{y}}_r = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i}$, spre deosebire de $\bar{y} = \frac{1}{K} \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i$ nu

necesita cunoasterea numarului total de unitati din populatie.

Un estimator pentru dispersia lui $\hat{\bar{y}}_r$ este ([1]):

$$V(\hat{\bar{y}}_r) = \frac{1}{\bar{M}} \left[\left(1 - \frac{n}{N}\right) \frac{s_r^2}{n} + \frac{1}{nN} \sum_{i \in S} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i} \right]$$

unde $s_r^2 = \frac{\sum_{i \in S} (M_i \bar{y}_i - M_i \hat{\bar{y}}_r)^2}{n-1}$, iar $\bar{M}$ este volumul mediu al grupelor.

Exemplu

Sa presupunem ca vrem sa estimam, într-o crescatorie de iepuri (perfect sanatosi) numarul mediu de picioare. Crescatoria are

doua cladiri: una cu 30 de exemplare si alta cu 10 exemplare.

Selectam o cladire cu probabilitatea 0,5, apoi selectam aleator doi indivizi din acea cladire si folosim $\hat{y}_{unb}$ pentru a estima numarul mediu de picioare per individ. Presupunem ca am selectat prima cladire. În mod deloc surprinzator, fiecare individ are patru picioare, deci $\hat{t}_I = 30 \cdot 4 = 120$.

Atunci un estimator nedeplasat pentru total

va fi: $\hat{t}_{unb} = \frac{2}{1} \cdot \hat{t}_I = 240$, deci numarul mediu de picioare pe individ este $\hat{y}_{unb} = \frac{240}{40} = 6$!

$$\hat{V}(\hat{t}_{unb}) = 2^2 \left(1 - \frac{1}{2} \right) \frac{S_t^2}{n} + \frac{2}{1} \sum_{i \in S} \left(1 - \frac{m_i}{M_i} \right) M_i^2 \frac{S_i^2}{m_i} = \frac{1}{2} \cdot 4 \cdot 3200 = 64000.$$

Aplicând însa estimatorul de tip raport obținem în ambele cazuri $\hat{y}_r = 4$, un rezultat corect, cu $V(\hat{y}_r) = 0$.

În general, estimatorul nedeplasat al totalului populatiei este inefficient daca volumele grupelor sunt inegale si t_i sunt proportionale cu M_i . Dispersia lui $\hat{t}_{unb}$ depinde varianta lui t_i si aceasta varianta poate fi foarte mare daca M_i – urile sunt diferite. Estimatorul de tip raport, însa, în general

Daca selectam a doua cladire,

$$\hat{t}_{II} = 10 \cdot 4 = 40, \quad \hat{t}_{unb} = \frac{2}{1} \cdot \hat{t}_{II} = 80 \quad \text{si}$$

$$\hat{y}_{unb} = \frac{80}{40} = 2 !$$

Evident ca n-am obtinut niste estimari bune, desi sunt nedeplasate. Calitatea slaba a lor este reflectata însa de nivelul ridicat al dispersiei:

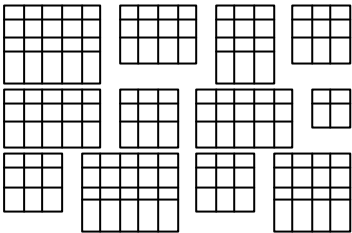
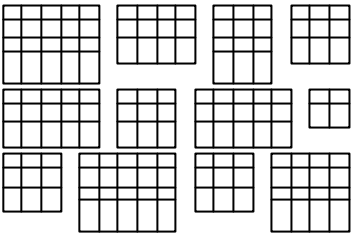
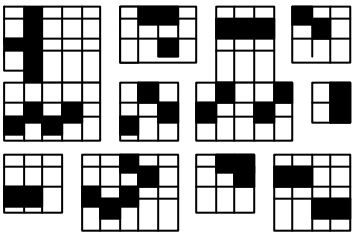
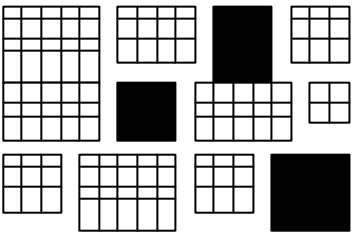
da performante bune atunci când t_i este direct proportional cu M_i , $\forall i = \overline{1, N}$.

Bibliografie

1. W.Cochran – Sampling Technics, 3rd Edition, Wiley, New York, 1977
2. M.Dumitrescu – Sondaje statistice si aplicatii, Editura Tehnica, 2000
3. S.Lohr – Sampling: Design and Analysis, Duxbury Press, 1999

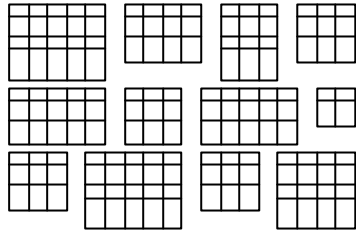
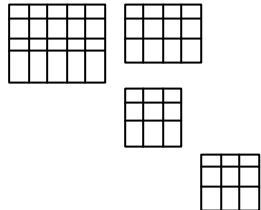
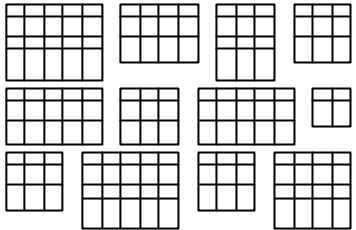
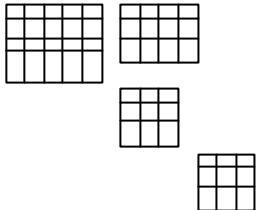
ANEXA 1

Asemanari si deosebiri între sondajul stratificat si sondajul grupat

Sondaj stratificat	Sondaj grupat
Fiecare element al populatiei se afla în exact un singur strat	Fiecare element al populatiei se afla în exact un singur cluster
Populatie de H straturi; stratul h are n_h elemente: 	Sondaj grupat într-o etapa; populatie de N grupe: 
Se considera un esantion simplu aleator din <i>fiecare</i> strat: 	Se considera un esantion simplu aleator de grupuri; se observa toate elementele din interiorul unui grup esantionat: 
Varianta estimatorului $\bar{y}_u$ depinde de variabilitatea valorilor în interiorul straturilor.	Grupul este unitatea de sondaj; cu cât s esantioneaza mai multe grupuri, cu atât varianta este mai mica. În cazul estimatorului $\bar{y}_u$ depinde în primul rând de variabilitatea între mediile grupelor.
Pentru o precizie mai mare elementele din interiorul fiecarui strat trebuie sa aiba valori apropiate, dar mediile straturilor trebuie sa difere unele de altele, pe cât este posibil.	Pentru o precizie mai mare elementele din interiorul fiecarui grup trebuie sa fie eterogene, iar mediile grupelor trebuie sa aiba valori apropiate una de alta.

ANEXA 2

Diferenta dintre sondajul grupat într-o etapa si cel în doua etape

O etapa	Doua etape
Populatie de N grupe:  Se considera un esantion simplu aleator de n grupe:  Se observa fiecare unitate din interiorul fiecarui grup esantionat: 	Populatie de N grupe:  Se considera un esantion simplu aleator de n grupe:  Se considera un esantion simplu aleator de m_i unitati pentru fiecare grup i esantionat: 