O METODĂ DE STABILIRE A VOLUMULUI OPTIM PENTRU UN EŞANTION

ŞTEFAN ŞTEFĂNESCU

INTENŢII ŞI NOTAŢII

În această lucrare vom sugera utilizarea unei tehnici de simulare stocastică pe calculator pentru a stabili gradul de reprezentativitate al unui eşantion de volum dat ce va fi selectat aleator. Reprezentativitatea eşantionului va fi determinată în raport cu o caracteristică aleasă X a populaţiei studiate. Se va estima probabilitatea ca valorile h (X) să aparţină unui interval de încredere precizat, unde h este o funcţie specificată (de exemplu media variabilei X). Prin procedeul de simulare propus poate fi determinat volumul optim al unui eşantion atunci când se cere o anume acurateţe rezultatelor.

În cadrul acestui studiu vom adopta următoarele **notaţii**:

Mulţimea $W \equiv \{w_1, w_2, w_3, \ldots, w_m\}$ semnifică o populaţie cu m „indivizi" w_i, $1 \le i \le m$.

Prin p este desemnat un parametru (sau grup de parametri) ce caracterizează populaţia W.

Variabila X defineşte o caracteristică a populaţiei studiate (de exemplu, vârstă, sex, studii, naţionalitate etc.). În general, variabila X este asociată răspunsului dat de membrii populaţiei W la o anumita întrebare din chestionar .

Valorile $x_1, x_2, x_3, \ldots, x_{m-1}, x_m$ sunt realizările variabilei X . Astfel prin x_i, $1 \le i \le m$, vom desemna codul de răspuns la întrebarea X dat de cel de al i -lea individ al populaţiei W.

Submulţimea $E \subset W$, $E \equiv \{e_1, e_2, e_3, \ldots, e_{n-1}, e_n\}$ caracterizează un eşantion de volum n cu „indivizii" e_j, $1 \le j \le n$.

Valorile $x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_{n-1}^{(k)}, x_n^{(k)}$ reprezintă codurile răspunsurilor la întrebarea X date de cei n indivizi ai eşantionului $E^{(k)} \subset W$. Dacă individul $e_j^{(k)}$ din eşantionul $E^{(k)}$, $1 \le j \le n$, este cel de al i -lea element w_i al populaţiei W atunci, în mod evident avem $x_j^{(k)} = x_i$.

Variabila s reprezintă numărul de „simulări" ce sunt efectuate în procesul de estimare.

$p^{(k)}$ este o estimaţie a parametrului p, valoare ce este obţinută numai prin utilizarea informaţiei din eşantionul $E^{(k)}$.

Prin Λ_E este desemnată variabila (aleatoare) ale cărei realizări sunt estimaţiile $p^{(k)}$ rezultate prin luarea în considerare a tuturor celor n eşantioanele $E^{(k)}$, $1 \le k \le \nu$, de volum n ce se pot alcătui cu cei m membri ai populaţiei W.

Vom nota prin $q = q_{n,p,d}$ probabilitatea ca valorile $p^{(k)}$, $1 \le k \le n$, să aparţină intervalului [p - d, p + d], d fiind un număr real, ales arbitrar.

EŞANTIONARE ŞI REPREZENTATIVITATE

Este cunoscut faptul că cei n „indivizi" $e_1, e_2, \ldots, e_n$ ai unui eşantion E, $E \subset W$, pot fi „aleşi" în diferite moduri din cei m indivizi $w_1, w_2, \ldots, w_m$ ai populaţiei W (a se vedea monografia lui Kish, 1963). În acest sens vom reaminti mai multe tehnici de eşantionare. Astfel, membrii e_j ai eşantionului E pot fi selectaţi din întreaga populaţie W după un algoritm determinist sau în mod aleator. În cazul unei selecţii aleatoare evidenţiem două posibile variante de obţinere a eşantionului E: toate elementele din populaţia W au aceeaşi şansă de a fi alese, sau vom accepta probabilităţi diferite de selectare ce depind efectiv de fiecare element w_i, $1 \le i \le m$.

În funcţie de tipurile de analize ce urmează a se desfăşura, „indivizii" populaţiei sunt grupaţi după anumite criterii (de exemplu, integrarea indivizilor pe gospodării). În acest context, elementele eşantionului pot fi astfel de grupe.

Din motive de eficienţă, în practică se utilizează adesea divizarea populaţiei pe „straturi", caz în care eşantionarea se va desfăşura diferenţiat, separat pentru fiecare strat în parte. Rezultatele finale vor fi însă obţinute prin ponderarea corespunzătoare a rezultatelor parţiale ce au fost iniţial deduse pe straturi.

Pe lângă aceste proceduri, există şi posibilitatea desfăşurării gradate, pe etape, a procesului de eşantionare. Astfel, eşantionul iniţial E_1, de volum n_1 relativ mare, este obţinut cu intenţia depistării unei eventuale structurări în populaţia W (dacă se urmăreşte, de exemplu, repartizarea geografică a unor categorii). Intenţionându-se utilizarea eşantionului E_1 într-o operaţiune de „tatonare", la selectarea elementelor sale nu vor fi alocate prea multe resurse. Ulterior vor fi obţinute eşantioanele E_2, E_3, … de volume n_2, respectiv n_3, de regulă cu mult mai mici. Aceste eşantioane sunt însă „mai specializate", alegerea elementelor lor fiind „direcţionată" în raport cu scopul urmărit în cercetare.

Din categoria eşantioanelor nealeatoare remarcăm eşantioanele fixe (tip panel) sau acelea obţinute prin procedura cotelor (a se vedea: Kish, 1963; Raj, 1968; Rotariu, Iluţ, 1997).

În practică, în vederea asigurării unei eficienţe sporite a procedurii de selectare, la construcţia unui eşantion se folosesc simultan mai multe tehnici de

eşantionare. Astfel, metodele aleatoare de selectare se vor îmbina cu procedee deterministe. Menționăm în acest context eşantionarea aleatoare multistratificată folosită în studiul „Dimensiuni ale sărăciei" (Cătălin Zamfir, 1995, p. 12-14), caz în care sunt însă utilizate şi procedee nealeatoare. O asemenea modalitate este adesea întrebuințată la construirea unui eşantion național (Rotariu, Iluț, 1997, p. 148-152).

Eşantioanele E sunt construite în vederea studierii, cu un efort minim şi în timp util, a unei populații numeroase W, greu de inventariat în totalitate. Eşantionul E, ca submulțime a populației W, $E \subset W$, va conține însă o informație trunchiată în raport cu populația W. Aşadar, aprecierea comportamentului întregii populații W pe baza informației prezente în eşantionul E va fi supusă unor erori mai mari sau mai mici.

În această situație, problema principală constă în a aprecia, cât mai exact posibil, mărimea acestor erori. Erorile depind, în mod evident, de algoritmul concret folosit la selectarea elementelor eşantionului, ca şi de volumul acestuia. Dimensiunea erorilor defineşte gradul de reprezentativitate al eşantionului respectiv.

Pentru a putea fi măsurat, gradul de reprezentativitate al eşantionului E se stabileşte în raport cu o caracteristică aleasă X (sau un grup de caracteristici) a populației W. Exprimarea riguroasă, prin formule, a gradului de reprezentativitate al lui E este soluționată din punct de vedere teoretic pentru principalele procedee de alegere a eşantionului. De multe ori însă, sunt precizate formule aproximative ce au rezultat prin impunerea unor condiții suplimentare, adesea simplificatoare. Chiar şi în această ultimă situație se poate ajunge la expresii complicate, impuse în fond de complexitatea algoritmilor stocastici şi determinişti ce intervin la selectarea elementelor eşantionului în cauză. În plus, evaluarea efectivă a formulelor deduse este greoaie din punct de vedere tehnic datorită multiplelor relații de natură combinatorială, cât şi datorită termenilor aleatori ce intervin în respectiva formulă.

În cazul în care la obținerea unui eşantion E sunt folosite combinat mai multe procedee de selecție, practic este imposibil de a stabili din punct de vedere teoretic formule exacte care să măsoare gradul de reprezentativitate al lui E.

În cele ce urmează vom estima gradul de reprezentativitate al eşantionului E de volum n utilizând pentru aceasta o tehnică de simulare stocastică. O astfel de procedură este deosebit de flexibilă, putând fi adaptată cu uşurință diverselor metode concrete, unele sofisticate, de selectare a eşantionului E.

Numai cu un simplu efort de programare pe calculator, fără a se interveni efectiv „în teren", pot fi făcute analize utile privind tehnica de eşantionare ce trebuie adoptată. În plus, pentru orice metodă de selectare fixată, comparând rezultatele simulărilor efectuate cu eşantioane de diferite volume, se

va putea stabili în final dimensiunea optimă a eşantionului ce va trebui construit într-o situaţie concretă.

SOLUŢIONAREA UNEI PROBLEME CONCRETE

Vom ilustra aspectele amintite pe un exemplu ce apare adesea în practică, anume eşantionul simplu aleator.

Aşadar vom accepta :

Ipoteza 1. Cele n elemente e_1, e_2, …, e_n ale eşantionului E sunt n „indivizi" din populaţia W (şi nu „grupuri" de indivizi) . Elementele lui E sunt alese în mod aleator din populaţia $W \equiv \{w_1, w_2, \ldots, w_m\}$, fiecare dintre acestea având aceeaşi şansă de a fi selectate (nici un individ w_i, $1 \le i \le m$, nu este privilegiat).

În final, va trebui să stabilim reprezentativitatea eşantionului E în raport cu o caracteristică X a populaţiei W. Pentru simplificarea expunerii vom presupune:

Ipoteza 2. Variabila X este o variabilă dihotomică (binomială) ce ia valorile 1 („da") sau 0 („nu"). În plus, volumul m al populaţiei este considerat finit, eventual foarte mare.

Vom nota prin p probabilitatea de a obţine un răspuns afirmativ la întrebarea X presupunând ca au fost intervievaţi toţi membrii populaţiei W.

Cu notaţiile anterioare, ştiind răspunsul x_i dat de individul w_i la „întrebarea" X avem

$$p = [x_1 + x_2 + x_3 + \ldots + x_{m-1} + x_m] / m \qquad (1)$$

În practică suntem în imposibilitate de a trece în revistă răspunsurile la întrebarea X a tuturor membrilor populaţiei W. Aşadar, evaluarea numărului de răspunsuri afirmative 1 din populaţia W nu poate fi realizată prin intermediul formulei (1).

De fapt, populaţia W depinde de parametrul p ce trebuie însă estimat utilizându-se numai informaţia deţinută în eşantionul E. Acest aspect ar putea fi generalizat, în locul parametrului p considerându-se o mulţime de astfel de parametri.

Având răspunsurile $x_1^{(E)}$, $x_2^{(E)}$, …, $x_n^{(E)}$ la întrebarea X pentru cei n membri ai eşantionului E, atunci parametrul p va fi estimat de valoarea

$$p^{(E)} = [x_1^{(E)} + x_2^{(E)} + x_3^{(E)} + \ldots + x_{n-1}^{(E)} + x_n^{(E)}] / n \qquad (2)$$

Prezenţa în eşantionul E a unei informaţii „trunchiate" în raport cu răspunsul efectiv al tuturor membrilor populaţiei W la întrebarea X, va antrena în mod obligatoriu apariţia unei erori ε măsurată prin „diferenţa de răspuns" dintre populaţia W şi eşantionul E, adică

$$\varepsilon^{(E)} = | p - p^{(E)} | \qquad (3)$$

Gradul de „reprezentativitate" al eşantionului E de volum n ar putea fi stabilit în raport cu mărimea erorii ε .

Problema nu poate fi însă abordată în această manieră deoarece practic cantitatea p nu este cunoscută, efectiv fiind dedusă numai valoarea $p^{(E)}$.

În plus, considerând un alt eşantion $E^{(k)}$ de acelaşi volum n, cu răspunsurile $x_1^{(k)}$, $x_2^{(k)}$, …, respectiv $x_n^{(k)}$, la întrebarea X, vom obține o alta estimație $p^{(k)}$ a parametrului necunoscut p, unde evident

$$p^{(k)} = [x_1^{(k)} + x_2^{(k)} + x_3^{(k)} + \ldots + x_{n-1}^{(k)} + x_k^{(k)}] / n \qquad (4)$$

eroarea estimării fiind de această dată $\varepsilon^{(k)}$,

$$\varepsilon^{(k)} = | p - p^{(k)} | \qquad (5)$$

Presupunând că avem posibilitatea de a enumera efectiv toate cele ν eşantioane $E^{(k)}$ de volum n, $1 \le k \le \nu$, ce se pot construi cu cei m indivizi ai populației W, prin procedura descrisă vom obține ν estimații (diferite) $p^{(k)}$, $1 \le k \le \nu$, fiecare dintre acestea sugerând o posibilă valoare pentru parametrul p.

Aşadar vom aprecia gradul de reprezentativitate al eşantioanelor E de volum n prin repartiția valorilor $e^{(k)}$, $1 \le k \le \nu$.

Precizarea repartiției valorilor $e^{(k)}$, $1 \le k \le \nu$, este imposibil de realizat în realitate, fiind mult mai uşor de a se opera cu probabilitățile $q = q_{n,p,d}$ determinate astfel ca valorile $p^{(k)}$, $1 \le k \le \nu$, să aparțină unui interval [p - d, p + d] precizat anterior. Capetele acestui interval sunt alese arbitrar prin specificarea numărului real d.

Formalizând, dacă vom desemna prin Λ_E variabila (aleatoare) ale cărei realizări sunt cele n estimații $p^{(k)}$ rezultate prin luarea în considerare a tuturor eşantioanelor $E^{(k)}$, $1 \le k \le \nu$, de volum n ce se pot alcătui cu cei m membri ai populației W, atunci.

$$q = q_{n,p,d} = \Pr(\Lambda_E \in [p - d, p + d]) \qquad (6)$$

Prin urmare, reprezentativitatea eşantioanelor E de volum n va fi dictată de şirul de probabilități $q_{n,p,d}$, şir indexat după indicii p şi d (volumul n este presupus fix).

În literatura de specialitate probabilitățile $q_{n,p,d}$ au fost efectiv calculate în cazul unor proceduri simple de alegere a membrilor eşantionului. În situația combinării diverselor proceduri de selectare (metode aleatoare utilizate împreună cu metode deterministe), nu au fost deduse formule exacte ce dau valorile probabilităților $q_{n,p,d}$. În plus, determinarea valorilor $q_{n,p,d}$ presupune utilizarea unui aparat matematic sofisticat.

Toate aceste neajunsuri pot fi înlăturate prin aplicarea unei tehnici de simulare stocastică pe calculator, fapt ce va fi explicat în continuare.

Menționăm că analiza prezentată poate fi preluată fără modificări în cazul unei populații infinite.

PROCEDURA DE SIMULARE

Pentru înțelegerea detaliată a problematicii simulării stocastice recomandăm parcurgerea cărții „Stochastic modelling" (Nelson, 1995), unde tratarea multiplelor aspecte se face mai mult din punct de vedere practic, cu multe exemple.

În prezenta lucrare nu vom intra în detalii tehnice. Vom urmări să sugerăm modul intuitiv în care se realizează efectiv algoritmul de simulare. În final vom face o analiză a rezultatelor simulărilor şi vom trage anumite concluzii ce sunt valabile şi pentru alte tipuri de eşantioane.

În cazul unei populații W numeroase este practic imposibilă enumerarea tuturor celor n eşantioane $E^{(k)}$, $1 \le k \le \nu$, de volum n, aceasta în vederea estimării probabilităților $q_{n,p,d}$. De regulă, valoarea ν este „astronomică" (ν reprezintă combinări de m obiecte luate câte n).

Această dificultate va fi înlăturată luând în considerare numai s astfel de eşantioane, unde s este de regulă ales mult mai mic decât ν. În noua situație, probabilitatea reală $q_{n,p,d}$ va fi estimată prin valoarea $q^{(s)} = q_{n,p,d}{}^{(s)}$ considerându-se numai cele s realizări $p^{(1)}, p^{(2)}, \ldots, p^{(s)}$ ale variabilei Λ_E, adică

$$q^{(s)} = q_{n,p,d}{}^{(s)} = \mu / s \qquad (7)$$

unde μ este numărul de cazuri în care valorile $p^{(1)}, p^{(2)}, \ldots, p^{(s)}$ aparțin intervalului $[p - d, p + d]$.

Formula (7) s-a obținut aproximând probabilitatea $q_{n,p,d} = Pr\,(p - d \le \Lambda_E \le p + d)$ prin raportul $q^{(s)} = \mu / s$ dintre numărul m de „cazuri favorabile" şi numărul s de „cazuri posibile" (rezultat clasic ce apare în următoarele lucrări: Iosifescu, ş.a., 1985; Klimov, 1986; Koroliouk, ş.a., 1983; Popoulis, 1990).

În concluzie, algoritmul **AS** de obținere a estimației $q^{(s)}$ (formula (7)) a probabilității q (formula (6)) şi care utilizează s paşi de simulare, are următoarea structură :

Algoritmul AS (simulare stocastică).

Pas 0. Precizează valorile parametrilor : m, n, p, d, s .

Pas 1. Luând în considerare populația W se generează s eşantioane $E^{(1)}, E^{(2)}, \ldots, E^{(s)}$ (simplu)

aleatoare, toate având acelaşi volum n .

Pas 2. Se calculează estimațiile $p^{(1)}, p^{(2)}, \ldots, p^{(s)}$ după formula (4).

Pas 3. Evaluează numărul μ, adică, câte valori $p^{(1)}, p^{(2)}, \ldots, p^{(s)}$ aparțin intervalului $[p - d, p + d]$.

Pas 4. Se estimează $q^{(s)}$ utilizându-se relația $q^{(s)} = \mu / s$ (formula (7)).

Pas 5. Se editează valoarea estimată $q^{(s)}$.

Valorile $q^{(s)}$ astfel obținute vor caracteriza gradul de reprezentativitate al eşantioanelor E de volum n (ce sunt selectate simplu aleator dintr-o populație W cu m „indivizi").

Remarcă. Pentru generarea eşantioanelor (simplu) aleatoare (Pasul 1 al algoritmului **AS**), ca şi pentru obținerea estimației $q^{(s)}$ se poate folosi produsul soft SPSS sau se poate rescrie algoritmul **AS** folosind orice alt limbaj de programare (de exemplu limbajul BASIC). Pentru această lucrare am preferat scrierea algoritmului **AS** în limbajul PASCAL .

Menționăm faptul că limbajele de programare (ca şi SPSS-ul) au subrutine specializate în generarea de valori aleatoare (uniform repartizate pe intervalul [0, 1]).

ANALIZA REZULTATELOR SIMULĂRII

Formula (7) poate fi rescrisă şi interpretată în sensul

$$q = q_{n,p,d} = \Pr (p^{(E)} - d \le p \le p^{(E)} + d) \qquad (8)$$

unde $p^{(E)}$ este o estimare a răspunsurilor populației W considerându-se eşantioane E de volum n (a se vedea formula (2)).

Pentru orice eşantion E de volum n din populația W putem calcula exact modul de răspuns $p^{(E)}$ al indivizilor eşantionului la întrebarea X. Aproximând probabilitatea q prin valoarea $q^{(s)}$ dată de algoritmul **AS**, din formula (8) rezultă

$$q^{(s)} \approx \Pr (p^{(E)} - d \le p \le p^{(E)} + d) \qquad (9)$$

Aşadar, cu o şansă de (aproximativ) $q^{(s)}$ procente, indicatorul p al răspunsurilor afirmative la întrebarea X la nivelul întregii populații W se încadrează în intervalul $[p^{(E)} - d, p^{(E)} + d]$, interval ce este de această dată cunoscut.

Cum X este o variabilă dihotomică a cărei medie Med (X) este chiar p, reinterpretăm formula (9) în sensul

$$q^{(s)} \approx \Pr (p^{(E)} - d \le \text{Med}\,(X) \le p^{(E)} + d) \qquad (10)$$

estimația $q^{(s)}$ fiind produsă de algoritmul **AS** .

În plus, valoarea $p^{(E)}$ poate fi privită ca medie a răspunsurilor la întrebarea X, răspunsuri colectate numai de la indivizii eşantionului E. Deci $p^{(E)}$ este „media pe eşantion" în raport cu caracteristica X.

Prin rularea algoritmului **AS** s-au obținut rezultatele precizate în Tabelele 1-4. În acest context s-au efectuat 30.000 de simulări pentru eşantioane de diverse volume n, $n \in \{100, 400, 700, 1000, 1300, 1600\}$, şi diferite caracteristici X, $p \in \{0{,}05, 0{,}1, 0{,}2, \ldots, 0{,}9, 0{,}95\}$. Intervalul de încredere are lungimea 2 d, variabila $d \in \{0{,}005, 0{,}025\}$, fapt ce semnifică „precizii" diferite atribuite rezultatelor.

Studiul a fost făcut pentru populaţii cu un număr variabil de indivizi, m $\in$ {50.000, 1.000.000}.

Comparând Tabelele 1 şi 2, respectiv Tabelele 3 şi 4, concludem că o mărire semnificativă a populaţiei (de 20 de ori, adică o creştere de la 50.000 de persoane la 1.000.000 de persoane) nu antrenează o creştere palpabilă a gradului de reprezentativitate a diverselor eşantioane (ale căror volume variază). Eşantioanele studiate se raportează la caracteristici X distincte (diferite valori ale parametrului p).

Cum era şi de aşteptat, o mărire a preciziei estimaţiei (micşorarea intervalului de încredere [p - d, p + d]) se realizează printr-o micşorare a şansei de apartenenţă la intervalul respectiv (adică o creştere a riscului de neapartenenţă la intervalul de încredere). Acest lucru reiese experimental comparând Tabelele 1 cu 3, respectiv din analizarea concomitenta a Tabelelor 2 şi 4.

În plus, mărirea volumului eşantionului E antrenează în mod obligatoriu o creştere a reprezentativităţii sale .

Astfel, pentru fiecare linie a oricărui Tabel 1-4, şirul probabilităţilor $q^{(s)} = q_{n,p,d}^{(s)}$ este crescător în raport cu n. Trebuie precizat faptul că valorile estimaţiilor $q_{n,p,d}^{(s)}$ ce se regăsesc pe aceeaşi linie a unuia dintre Tabelele 1-4 au parametrii p şi d menţinuţi constanţi. Poziţionarea elementelor $q_{n,p,d}^{(s)}$ în cadrul unei linii fixate din Tabelele 1-4 este dictată de indicele n (ce reprezintă volumul eşantionului E).

Tabelul nr.1

Estimaţia experimentală $q_{n,p,d}^{(s)}$ a probabilităţii ca media pe eşantion în raport cu caracteristica X să aparţină intervalului [p - 0,025, p + 0,025], pentru eşantioane de diferite volume n (s = 30000, m = 1000000, d = 0.025)

p	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
0,050	0,7533	0,9786	0,9972	0,9996	0,9999	1,0000
0,100	0,5945	0,9036	0,9724	0,9917	0,9974	0,9991
0,200	0,4659	0,7833	0,9010	0,9511	0,9752	0,9859
0,300	0,4200	0,7293	0,8523	0,9171	0,9521	0,9723
0,400	0,3870	0,6927	0,8251	0,8947	0,9345	0,9587
0,500	0,3835	0,6986	0,8124	0,8896	0,9272	0,9560
0,600	0,3899	0,6886	0,8210	0,8906	0,9319	0,9571
0,700	0,4156	0,7213	0,8490	0,9150	0,9506	0,9701
0,800	0,4671	0,7918	0,9047	0,9522	0,9751	0,9871
0,900	0,5899	0,9034	0,9723	0,9917	0,9972	0,9992
0,950	0,7536	0,9778	0,9975	0,9997	1,0000	1,0000

Tabelul nr. 2

Estimația experimentală $q_{n,p,d}^{(s)}$ a probabilității ca media pe eşantion în raport cu caracteristica X să aparțină intervalului [p - 0,025, p + 0,025] pentru eşantioane de diferite volume n (s = 30000, m = 50000, d = 0,025).

p	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
0,050	0,7521	0,9778	0,9972	0,9997	0,9999	1,0000
0,100	0,5993	0,9033	0,9732	0,9926	0,9977	0,9994
0,200	0,4721	0,7860	0,9014	0,9523	0,9760	0,9875
0,300	0,4143	0,7288	0,8505	0,9151	0,9502	0,9699
0,400	0,3908	0,6888	0,8220	0,8953	0,9351	0,9592
0,500	0,3853	0,7085	0,8177	0,8918	0,9284	0,9574
0,600	0,3879	0,6886	0,8228	0,8935	0,9341	0,9582
0,700	0.4194	0,7229	0,8501	0,9174	0,9503	0,9702
0,800	0,4666	0,7875	0,8992	0,9512	0,9761	0,9883
0,900	0,5922	0,8996	0,9722	0,9920	0,9976	0,9994
0,950	0,7568	0,9799	0,9974	0,9998	0,9999	1.0000

Tabelul nr.3

Estimația experimentală $q_{n,p,d}^{(s)}$ a probabilității ca media pe eşantion în raport cu caracteristica X să aparțină intervalului [p - 0,005, p + 0,005] pentru eşantioane de diferite volume n (s = 30000, m = 1000000, d = 0,005) .

p	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
0,050	0,1816	0,3486	0,4490	0,5285	0,5909	0,6336
0,100	0,1301	0,2594	0,3398	0,3951	0,4493	0,4957
0,200	0,0997	0,1959	0,2562	0,3072	0,3451	0,3845
0,300	0,0880	0,1755	0,2285	0,2724	0,3077	0,3403
0,400	0,0823	0,1598	0,2153	0,2553	0,2892	0,3210
0,500	0,0785	0,1942	0,2054	0,2725	0,2877	0,3323
0,600	0,0812	0,1635	0,2127	0,2478	0,2870	0,3215
0,700	0,0879	0,1726	0,2310	0,2691	0,3083	0,3405
0,800	0,1021	0,1975	0,2629	0,3089	0,3450	0,3855
0,900	0,1292	0,2562	0,3389	0,4000	0,4540	0,4920
0,950	0,1780	0,3457	0,4538	0,5266	0,5897	0,6377

Tabelul nr.4

Estimaţia experimentală $q_{n,p,d}^{(s)}$ a probabilităţii ca media pe eşantion în raport cu caracteristica X să aparţină intervalului [p - 0,005, p + 0,005] pentru eşantioane de diferite volume n (s = 30000, m = 50000, d = 0,005).

p	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
0,050	0,1831	0,3502	0,4600	0,5310	0,5886	0,6369
0,100	0,1303	0,2614	0,3418	0,4006	0,4527	0,4958
0,200	0,0969	0,1941	0,2609	0,3078	0,3486	0,3897
0,300	0,0854	0,1717	0,2242	0,2658	0,3039	0,3365
0,400	0,0855	0,1625	0,2096	0,2532	0,2843	0,3123
0,500	0,0796	0,1975	0,2036	0,2698	0,2816	0,3280
0,600	0,0778	0,1621	0,2150	0,2542	0,2881	0,3165
0,700	0,0887	0,1709	0,2280	0,2684	0,3027	0,3395
0,800	0,0971	0,1964	0,2610	0,3017	0,3502	0,3853
0,900	0,1339	0,2604	0,3372	0,4008	0,4524	0,4953
0,950	0,1802	0,3454	0,4528	0,5252	0,5894	0,6350

Remarcă. Datorită dihotomiei caracteristicii X avem o „simetrie” a rezultatelor în raport cu valorile luate de parametrul p. Astfel, răspunsurile afirmative la întrebarea X sintetizate prin valoarea p pot fi interpretate ca răspunsuri negative la aceeaşi întrebarea X unde, de această dată, s-a considerat 1 - p drept parametru. Această observaţie poate fi validată urmărind efectiv rezultatele din Tabelele 1-4. Astfel se constată că $q_{n,p,d}^{(s)} \approx q_{n,1-p,d}^{(s)}$.

Utilizarea algoritmului **AS** permite obţinerea unor tabele cu estimaţiile probabilităţilor $q_{n,p,d}$ în raport cu orice valori ale parametrilor n, p, d, fapt ilustrat prin Tabelele 5 şi 6 . Asemenea tabele definesc de fapt gradul de reprezentativitate al unui eşantion E în raport cu cei trei parametri n, p, d.

Tabelele 5 şi 6 permit stabilirea volumului n al eşantionului E pentru o probabilitate $q_{n,p,d}$ data, valorile parametrilor p şi d fiind presupuse cunoscute.

Prin urmare, utilizarea unor tabele de tipul Tabelelor 5 şi 6 conduce la o dimensionare optimă a volumului unui eşantion E, în ipoteza unui grad de reprezentativitate acceptat.

Tabelul nr.5

Listarea perechilor de valori $q_{n,p,d}^{(s)}$ (n), unde la calculul valorii $q_{n,p,dS(s)}$ s-au utilizat s = 100.000 eşantioane E de volum n (eşantioanele E sunt extrase aleator dintr-o populaţie cu m = 2.000.000 indivizi; p = 0,05, d = 0,01).

0,2181 (50); 0,3294 (100); 0,4077 (150); 0,4690 (200); 0,5206 (250);
0,5637 (300); 0,6004 (350); 0,6345 (400); 0,6628 (450); 0,6868 (500);
0,7120 (550); 0,7345 (600); 0,7544 (650); 0,7734 (700); 0,7896 (750);
0,8049 (800); 0,8183 (850); 0,8291 (900); 0,8413 (950); 0,8517 (1000);
0,8624 (1050); 0,8722 (1100); 0,8809 (1150); 0,8890 (1200); 0,8960 (1250); 0,9031 (1300);
0,9088 (1350); 0,9155 (1400); 0,9210 (1450); 0,9262 (1500); 0,9307 (1550); 0,9349 (1600);
0,9392 (1650); 0,9427 (1700); 0,9458 (1750); 0,9492 (1800); 0,9520 (1850); 0,9556 (1900);
0,9587 (1950); 0,9607 (2000);
0,9616 (2050); 0,9641 (2100); 0,9661 (2150); 0,9674 (2200); 0,9698 (2250); 0,9715 (2300);
0,9735 (2350); 0.9752 (2400); 0.9769 (2450); 0.9779 (2500); 0,9793 (2550); 0,9804 (2600);
0,9815 (2650); 0,9826 (2700); 0,9837 (2750); 0,9848 (2800); 0,9857 (2850); 0,9864 (2900);
0,9873 (2950); 0,9876 (3000); 0,9885 (3050); 0,9890 (3100); 0,9897 (3150); 0,9901 (3200);
0,9903 (3250); 0,9909 (3300); 0,9915 (3350); 0,9923 (3400); 0,9928 (3450); 0,9931 (3500);
0,9934 (3550); 0,9937 (3600); 0,9942 (3650); 0,9946 (3700); 0,9947 (3750); 0,9950 (3800);
0,9953 (3850); 0,9955 (3900); 0,9958 (3950); 0,9961 (4000);
0,9968 (4050); 0,9971 (4100); 0,9972 (4150); 0,9974 (4200); 0,9975 (4250); 0,9975 (4300);
0,9977 (4350); 0,9978 (4400); 0,9978 (4450); 0,9980 (4500); 0,9982 (4550); 0,9983 (4600);
0,9984 (4650); 0,9985 (4700); 0,9986 (4750); 0,9986 (4800); 0,9986 (4850); 0,9987 (4900);
0,9987 (4950); 0,9988 (5000);
0,9993 (5500); 0,9997 (6000);

Tabelul nr.6

Listarea perechilor de valori $q_{n,p,d}^{(s)}$ (n), unde la calculul valorii $q_{n,p,d}^{(s)}$ s-au utilizat s = 100.000 eşantioane E de volum n (eşantioanele E sunt extrase aleator dintr-o populaţie cu m = 2.000.000 indivizi; $p \in \{0,05, 0,10, 0,15, \ldots . 0,45, 0,50\}$, d = 0,02).

6.a p = 0,05

0,6153 (100); 0,7994 (200); 0,8875 (300); 0,9349 (400); 0,9610 (500);
0,9767 (600); 0,9855 (700); 0,9907 (800); 0,9943 (900); 0,9963 (1000);
0,9977 (1100); 0,9983 (1200); 0,9990 (1300); 0,9993 (1400); 0,9997 (1500); 0,9998 (1600);
0,9999 (1700); 0,9999 (1800); 0,9999 (1900); 1,0000 (2000); 1,0000 (2100); 1,0000 (2200);
1,0000 (2300);

6.b p = 0,10

0,4771 (100); 0,6460 (200); 0,7474 (300); 0,8133 (400); 0,8606 (500);
0,8960 (600); 0,9227 (700); 0,9408 (800); 0,9556 (900); 0,9649 (1000);
0,9730 (1100); 0,9799 (1200); 0,9840 (1300); 0,9876 (1400); 0,9902 (1500); 0,9925 (1600);
0,9941 (1700); 0,9956 (1800); 0,9960 (1900); 0,9969 (2000); 0,9977 (2100); 0,9979 (2200);
0,9984 (2300); 0,9986 (2400); 0,9989 (2500); 0,9992 (2600); 0,9994 (2700); 0,9995 (2800);
0,9995 (2900); 0,9997 (3000);
0,9998 (3100); 0,9998 (3200); 0,9999 (3300); 0,9999 (3400); 0,9999 (3500);
0,9999 (3600); 0,9999 (3700); 0,9999 (3800); 0,9999 (3900); 1,0000 (4000);
1,0000 (4100); 1,0000 (4200); 1,0000 (4300);

6.c p = 0,15

0,4202 (100); 0,5671 (200); 0,6637 (300); 0,7332 (400); 0,7889 (500);
0,8297 (600); 0,8610 (700); 0,8864 (800); 0,9058 (900); 0,9217 (1000);
0,9367 (1100); 0,9478 (1200); 0,9571 (1300); 0,9644 (1400); 0,9698 (1500); 0,9752 (1600);
0,9789 (1700); 0,9826 (1800); 0,9858 (1900); 0,9882 (2000); 0,9898 (2100); 0,9914 (2200);
0,9928 (2300); 0,9940 (2400); 0,9952 (2500); 0,9959 (2600); 0,9969 (2700); 0,9972 (2800);
0,9979 (2900); 0,9981 (3000);
0,9981 (3100); 0,9985 (3200); 0,9987 (3300); 0,9989 (3400); 0,9991 (3500);
0,9992 (3600); 0,9993 (3700); 0,9993 (3800); 0,9995 (3900); 0,9995 (4000);
0,9997 (4100); 0,9997 (4200); 0,9998 (4300); 0,9998 (4400); 0,9998 (4500);

6.d p = 0,20

0,3762 (100); 0,5156 (200); 0,6114 (300); 0,6815 (400); 0,7349 (500);
0,7780 (600); 0,8134 (700); 0,8403 (800); 0,8655 (900); 0,8864 (1000);
0,9024 (1100); 0,9149 (1200); 0,9281 (1300); 0,9377 (1400); 0,9465 (1500); 0,9536 (1600);
0,9603 (1700); 0,9650 (1800); 0,9696 (1900); 0,9739 (2000); 0,9776 (2100); 0,9802 (2200);
0,9832 (2300); 0,9852 (2400); 0,9873 (2500); 0,9891 (2600); 0,9907 (2700); 0,9920 (2800);
0,9930 (2900); 0,9940 (3000);
0,9946 (3100); 0,9952 (3200); 0,9959 (3300); 0,9964 (3400); 0,9968 (3500);
0,9974 (3600); 0,9978 (3700); 0,9980 (3800); 0,9983 (3900); 0,9986 (4000);
0,9985 (4100); 0,9987 (4200); 0,9990 (4300); 0,9991 (4400); 0,9992 (4500);

6.e p = 0,25

0,4357 (100); 0,5365 (200); 0,6128 (300); 0,6740 (400); 0,7227 (500);
0,7606 (600); 0,7935 (700); 0,8228 (800); 0,8470 (900); 0,8649 (1000);
0,8822 (1100); 0,8977 (1200); 0,9106 (1300); 0,9213 (1400); 0,9305 (1500); 0,9379 (1600);
0,9459 (1700); 0,9528 (1800); 0,9575 (1900); 0,9624 (2000); 0,9667 (2100); 0,9702 (2200);
0,9739 (2300); 0,9774 (2400); 0,9798 (2500); 0,9825 (2600); 0,9843 (2700); 0,9860 (2800);
0,9875 (2900); 0,9891 (3000);
0,9902 (3100); 0,9913 (3200); 0,9922 (3300); 0,9928 (3400); 0,9940 (3500);
0,9944 (3600); 0,9950 (3700); 0,9956 (3800); 0,9960 (3900); 0,9964 (4000);
0,9970 (4100); 0,9974 (4200); 0,9977 (4300); 0,9980 (4400); 0,9982 (4500);

6.f p = 0,30

0,3339 (100); 0,4576 (200); 0,5467 (300); 0,6144 (400); 0,6677 (500);
0,7140 (600); 0,7504 (700); 0,7826 (800); 0,8088 (900); 0,8319 (1000);
0,8511 (1100); 0,8685 (1200); 0,8835 (1300); 0,8967 (1400); 0,9086 (1500); 0,9197 (1600);
0,9281 (1700); 0,9359 (1800); 0,9433 (1900); 0,9496 (2000); 0,9548 (2100); 0,9600 (2200);
0,9636 (2300); 0,9671 (2400); 0,9709 (2500); 0,9739 (2600); 0,9767 (2700); 0,9793 (2800);
0,9814 (2900); 0,9832 (3000);
0,9847 (3100); 0,9861 (3200); 0,9875 (3300); 0,9889 (3400); 0,9902 (3500);
0,9913 (3600); 0,9922 (3700); 0,9931 (3800); 0,9937 (3900); 0,9946 (4000);
0,9950 (4100); 0,9954 (4200); 0,9958 (4300); 0,9962 (4400); 0,9966 (4500);

6.g p = 0,35

0,3256 (100); 0,4458 (200); 0,5320 (300); 0,6003 (400); 0,6505 (500);
0,6969 (600); 0,7320 (700); 0,7630 (800); 0,7911 (900); 0,8130 (1000);
0,8352 (1100); 0,8531 (1200); 0,8697 (1300); 0,8841 (1400); 0,8971 (1500); 0,9070 (1600);
0,9177 (1700); 0,9256 (1800); 0,9334 (1900); 0,9402 (2000); 0,9461 (2100); 0,9513 (2200);
0,9566 (2300); 0,9610 (2400); 0,9643 (2500); 0,9678 (2600); 0,9714 (2700); 0,9741 (2800);
0,9763 (2900); 0,9788 (3000);
0,9809 (3100); 0,9829 (3200); 0,9846 (3300); 0,9861 (3400); 0,9876 (3500);
0,9882 (3600); 0,9895 (3700); 0,9904 (3800); 0,9915 (3900); 0,9920 (4000);
0,9928 (4100); 0,9936 (4200); 0,9942 (4300); 0,9947 (4400); 0,9951 (4500);

6.h p = 0,40

0,3145 (100); 0,4364 (200); 0,5201 (300); 0,5867 (400); 0,6390 (500);
0,6825 (600); 0,7172 (700); 0,7510 (800); 0,7789 (900); 0,8034 (1000);
0,8243 (1100); 0,8433 (1200); 0,8594 (1300); 0,8737 (1400); 0,8873 (1500); 0,8983 (1600);
0,9084 (1700); 0,9182 (1800); 0,9262 (1900); 0,9327 (2000); 0,9392 (2100); 0,9441 (2200);
0,9495 (2300); 0,9537 (2400); 0,9584 (2500); 0,9622 (2600); 0,9654 (2700); 0,9689 (2800);
0,9717 (2900); 0,9744 (3000);
0,9772 (3100); 0,9790 (3200); 0,9806 (3300); 0,9826 (3400); 0,9842 (3500);
0,9852 (3600); 0,9867 (3700); 0,9877 (3800); 0,9888 (3900); 0,9896 (4000);
0,9908 (4100); 0,9916 (4200); 0,9926 (4300); 0,9932 (4400); 0,9939 (4500);

6.i p = 0,45

0,3120 (100); 0,4286 (200); 0,5107 (300); 0,5747 (400); 0,6287 (500);
0,6736 (600); 0,7122 (700); 0,7436 (800); 0,7723 (900); 0,7947 (1000);
0,8174 (1100); 0,8353 (1200); 0,8517 (1300); 0,8668 (1400); 0,8806 (1500); 0,8925 (1600);
0,9022 (1700); 0,9114 (1800); 0,9205 (1900); 0,9274 (2000); 0,9341 (2100); 0,9411 (2200);
0,9466 (2300); 0,9515 (2400); 0,9558 (2500); 0,9598 (2600); 0,9634 (2700); 0,9666 (2800);
0,9693 (2900); 0,9722 (3000);
0,9756 (3100); 0,9775 (3200); 0,9798 (3300); 0,9814 (3400); 0,9833 (3500);
0,9843 (3600); 0,9859 (3700); 0,9868 (3800); 0,9881 (3900); 0,9889 (4000);
0,9898 (4100); 0,9908 (4200); 0,9916 (4300); 0,9923 (4400); 0,9929 (4500);

6.j p = 0,50

0,3816 (100); 0,4771 (200); 0,5480 (300); 0,6087 (400); 0,6533 (500);
0,6944 (600); 0,7294 (700); 0,7567 (800); 0,7819 (900); 0,8051 (1000);
0,8249 (1100); 0,8427 (1200); 0,8596 (1300); 0,8735 (1400); 0,8857 (1500); 0,8966 (1600);
0,9073 (1700); 0,9162 (1800); 0,9241 (1900); 0,9322 (2000); 0,9376 (2100); 0,9433 (2200);
0,9475 (2300); 0,9525 (2400); 0,9574 (2500); 0,9622 (2600); 0,9646 (2700); 0,9678 (2800);
0,9707 (2900); 0,9733 (3000);
0,9758 (3100); 0,9775 (3200); 0,9794 (3300); 0,9812 (3400); 0,9828 (3500);
0,9853 (3600); 0,9864 (3700); 0,9874 (3800); 0,9889 (3900); 0,9898 (4000);
0,9898 (4100); 0,9908 (4200); 0,9916 (4300); 0,9924 (4400); 0,9931 (4500);

Analizând Tabelele 5 şi 6 constatăm că la o creştere liniară a volumului n al eşantionului E nu corespunde tot o creştere liniară a indicatorului $q_{n,p,d}$ ce exprimă gradul de reprezentativitate al lui E.

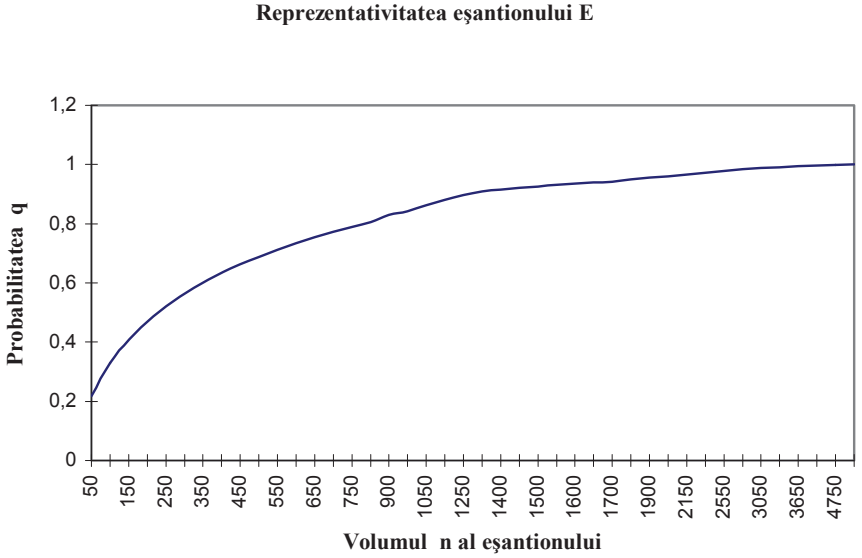
Astfel se observă că indicatorul $q_{n,p,d}$ are mai întâi o creştere foarte bruscă, după care creşterea sa devine cu mult mai lentă, aceasta începând cu o

valoare n_0 atribuită volumului eşantionului. În fapt, valoarea n_0 defineşte „volumul optim" al eşantionului E deoarece o creştere a volumului lui E peste pragul n_0 nu aduce modificări semnificative ale probabilităţilor $q_{n,p,d}$.

Aspectul menţionat este sugerat sugestiv de Figura 1 unde este reprezentat graficul funcţiei $q_{n,p,d}^{(s)}$ al cărei argument este volumul n al eşantionului E. În reprezentarea grafică amintită, parametrii p, d, s sunt fixaţi. Precizăm faptul că în acest caz valorile $q_{n,p,d}^{(s)}$ au fost preluate din Tabelul 5, pentru p = 0,05, d = 0,01, s = 100.000, m = 2.000.000 .

Figura 1. Graficul estimaţiei $q_{n,p,d}^{(s)}$ în raport cu volumul n al eşantionului

(p = 0,05, d = 0,01, s = 100000, m = 2000000; datele preluate din Tabelul 5).



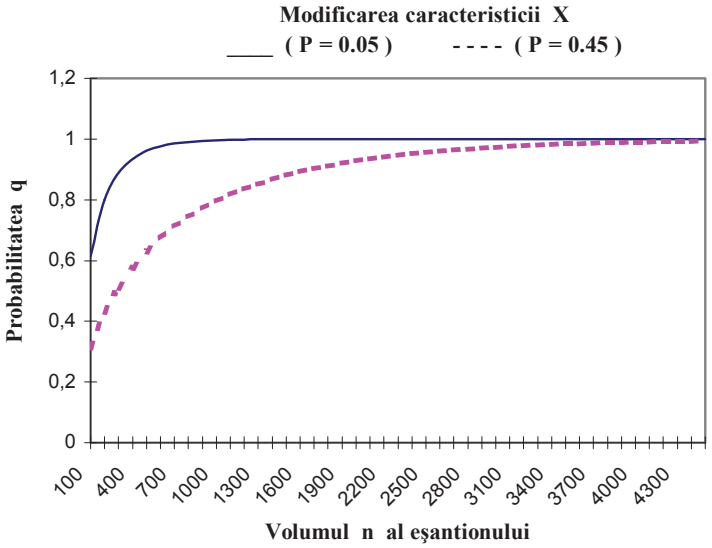
Menţionăm că o reorientare spre o altă caracteristică X a populaţiei W (de exemplu, modificarea valorii parametrului p), caracteristică în raport cu care s-a stabilit de fapt gradul de reprezentativitate al eşantionului E, conduce inevitabil la un alt volum optim n_0 pentru E.

Această afirmaţie este justificată clar prin studierea comparativă a datelor din Tabelele 6.a - 6.j. Astfel, trecerea la o nouă valoare a parametrului

p, $0 \le p \le 1$, păstrează alura curbei din Figura 1 (mai întâi o creştere rapidă urmată de o plafonare), fără însă ca noua curbă să se suprapună peste cea veche.

Un exemplu concludent este prezentat în Figura 2 unde sunt reprezentate simultan graficele funcțiilor $f_1(n) = q_{n,\ 0,05,\ 0,02}^{(100000)}$ şi $f_2(n) = q_{n,\ 0,45,\ 0,02}^{(100000)}$, selectarea simplu aleatoare a eşantionului E făcându-se dintr-o populație de doua milioane de indivizi.

Figura 2. Graficul estimației $q_{n,p,d}^{(s)}$ în funcție de volumul n al eşantionului atunci când caracteristica X se modifică; variante : p = 0.05, p = 0.45 (cazul m = 2000000, s = 100000,
d = 0.02; datele preluate din Tabelele 6.a şi 6.i).



ÎMBUNĂTĂȚIREA ESTIMAȚIILOR $Q_{N,P,D}^{(S)}$ ALE PROBABILITĂȚII $Q_{N,P,D}$

Algoritmul **AS** ar putea fi apelat de un număr r de ori, la fiecare apelare t, $1 \le t \le r$, listându-se valorile estimațiilor $q_{n,p,d}^{(s)}$. Vom desemna prin $q_{n,p,d,t}^{(s)}$ valoarea $q_{n,p,d}^{(s)}$ obținută la a t -a rulare a algoritmului **AS**. Bineînțeles,

estimaţiile $q_{n,p,d,t}^{(s)}$ vor fi diferite de la o apelare la alta a algoritmului **AS**, fapt ilustrat în Tabelul 7 (valorile $q_{n,p,d,t}^{(s)}$ se modifică în raport cu indicele t ce dă numărul apelării).

Vom nota prin $\alpha_{n,p,d,r}$, $\beta_{n,p,d,r}$ media, respectiv abaterea standard, a celor r valori $q_{n,p,d,t}^{(s)}$, $1 \leq t \leq r$, obţinute succesiv după r rulări (apelări) ale algoritmului **AS**. Vom putea evalua mărimea fluctuaţiile estimaţiilor $q_{n,p,d,t}^{(s)}$ în raport cu diversele apelări t ale algoritmului **AS** prin determinarea abaterii standard $\beta_{n,p,d,r}$ a şirului de valori $q_{n,p,d,t}^{(s)}$, $1 \leq t \leq r$. Aceste aspecte sunt subliniate şi de rezultatele prezentate în Tabelele 7 şi 8 (pentru care s-au efectuat s = 10000, respectiv s = 90000 de simulări).

Tabelul nr.7

Apelarea de r = 10 ori a algoritmului de simulare **AS** în vederea comparării diverselor estimaţii $q_{n,p,d,t}^{(s)}$, $1 \leq t \leq r$, ale probabilităţii $q_n = Pr(\Lambda_E \in [0{,}045, 0{,}055])$ pentru eşantioane de diferite volume n (s = 10000, m = 1000000, p = 0,05, d = 0,005) .

t	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
1	0,1832	0,3452	0,4598	0,5248	0,5988	0,6391
2	0,1813	0,3436	0,4567	0,5213	0,5867	0,6346
3	0,1848	0,3462	0,4551	0,5349	0,5968	0,6427
4	0,1783	0,3500	0,4647	0,5343	0,5885	0,6335
5	0,1791	0,3472	0,4494	0,5225	0,5938	0,6393
6	0,1740	0,3556	0,4567	0,5307	0,5974	0,6428
7	0,1852	0,3479	0,4630	0,5339	0,6003	0,6495
8	0,1849	0,3453	0,4582	0,5337	0,5919	0,6348
9	0,1809	0,3451	0,4523	0,5187	0,5872	0,6295
10	0,1827	0,3551	0,4547	0,5305	0,5870	0,6368
$a_{n,p,d,r}$	0,18144	0,34812	0,45706	0,52853	0,59284	0,63826
$b_{n,p,d,r}$	0,003362796	0,003985173	0,004398454	0,005806901	0,005019004	0,005434556

Prin mărirea numărului s de simulări va creşte acurateţea estimaţiilor $q_{n,p,d}^{(s)}$ privind probabilitatea ca media pe eşantion $p^{(E)}$, la caracteristica X, să aparţină intervalului [p - d, p + d].

Acest lucru este justificat şi de faptul că valorile $q_{n,p,d}^{(s)}$ sunt determinate ca raport dintre numărul μ al „cazurilor favorabile" şi numărul s al „cazurilor posibile" (formula (7)). Avem egalitatea $q_{n,p,d} = \mu / s$ dacă $s = \nu$, adică în situaţia în care se iau în considerare toate eşantioanele E de acelaşi volum n ce pot fi construite cu cei m indivizi din populaţia W.

Ilustrarea acestor aspecte reiese clar prin compararea Tabelelor 7 şi 8 . O mărire a numărului s de simulări de la s = 10.000 la s = 90.000 conduce la o

micşorare (de aproximativ 3 ori) a abaterilor standard $\beta_{n,p,d,r}$ faţă de media $\alpha_{n,p,d,r}$ ale celor r estimaţii $q_{n,p,d,t}^{(s)}$, $1 \le t \le r$.

Tabelul nr.8

Apelarea de r = 10 ori a algoritmului de simulare **AS** în vederea comparării diverselor estimaţii $q_{n,p,d,t}^{(s)}$, $1 \le t \le r$, ale probabilităţii $q_n = Pr(\Lambda_E \in [0{,}045, 0{,}055])$ pentru eşantioane de diferite volume n (s = 90000, m = 1000000, p = 0,05, d = 0,005)

t	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
1	0,1802	0,3488	0,4545	0,5298	0,5942	0,6409
2	0,1794	0,3459	0,4587	0,5284	0,5915	0,6407
3	0,1786	0,3506	0,4595	0,5315	0,5924	0,6397
4	0,1803	0,3468	0,4562	0,5296	0,5901	0,6387
5	0,1818	0,3443	0,4548	0,5250	0,5892	0,6374
6	0,1806	0,3478	0,4566	0,5306	0,5934	0,6404
7	0,1804	0,3470	0,4575	0,5272	0,5918	0,6369
8	0,1821	0,3466	0,4585	0,5290	0,5920	0,6374
9	0,1817	0,3472	0,4565	0,5298	0,5939	0,6397
10	0,1794	0,3492	0,4549	0,5268	0,5887	0,6375
$\alpha_{n,p,d,r}$	0,18045	0,34742	0,45677	0,52877	0,59172	0,63893
$\beta_{n,p,d,r}$	0,001088347	0,001692808	0,00166316	0,001853672	0,001800444	0,001458115

Remarcă. Este cunoscut următorul rezultat teoretic (Gentle, 1998; Nelson, 1995): o mărire a numărului de simulări (independente) de un număr de λ^2 ori antrenează o creştere a acurateţii rezultatelor simulării de numai λ ori (abaterea standard a rezultatelor se micşorează de λ ori).

Această observaţie a fost deja confirmată experimental prin rezultatele prezentate în Tabelele 7-8 (a se compara abaterile standard $\beta_{n,p,d,10}$ listate în aceste tabele).

O mărire a numărului r de apelări ale Algoritmului **AS** va conduce în mod evident la rezultate mai precise. În Tabelele 7 - 8 analiza era făcută pentru r = 10 apelări ale algoritmului de simulare **AS**. Cu aceleaşi valori ale parametrilor n, p, d, s reluăm această analiză, de această dată utilizându-se rezultatele obţinute din r = 50 de rulări ale procedurii **AS**. Editarea sintetică a rezultatelor astfel obţinute este prezentate în Tabelul 9 (9.a şi 9.b). O trecere de la s = 10.000 de simulări (Tabelul 9.a) la s = 90.000 simulări (de 9 ori mai mult; Tabelul 9.b) va antrena, din punct de vedere teoretic, o reducere a abaterii standard de $\lambda = 9^{1/2} = 3$ ori. Rezultatul teoretic menţionat este validat experimental prin compararea valorilor $\beta_{n,p,d,50}$ din Tabelele 9.a şi 9.b.

Tabelul nr.9

Mediile $\alpha_{n,p,d,r}$ şi abaterile standard $\beta_{n,p,d,r}$ ale rezultatelor obţinute din r = 50 de rulări ale Algoritmului **AS** (în varianta: m = 1000000, p = 0,05, d = 0,005).

9.a Varianta s = 10.000 simulări

α, β	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
$\alpha_{n,p,d,50}$	0,17993	0,34806	0,45624	0,52901	0,59159	0,64014
$\beta_{n,p,d,50}$	0,00407	0,00526	0,00486	0,00515	0,00516	0,00526

9.b Varianta s = 90.000 simulări

α,β	n = 100	n = 400	n = 700	n = 1000	n = 1300	n = 1600
$\alpha_{n,p,d,50}$	0,17991	0,34772	0,45604	0,52906	0,59170	0,63970
$\beta_{n,p,d,50}$	0,00155	0,00158	0,00157	0,00208	0,00159	0,00161

CONCLUZII ŞI POSIBILE EXTENSII

Alternativa simulării stocastice

Procedura simulării stocastice poate fi adaptată cu uşurinţă şi în cazul unor metode complexe de selectare a eşantionului E. Numai printr-un simplu efort de programare pe calculator, fără a se face investigaţii efective „în teren", pot fi determinate eficienţa şi dezavantajele unei tehnici de eşantionare ce se intenţionează a fi folosită.

Fixându-ne asupra unei proceduri de selectare a eşantionului, prin compararea rezultatelor simulărilor stocastice, se va putea aprecia în final dimensiunea optimă a eşantionului în raport cu resursele existente într-o situaţie concretă.

Concluziile analizei prezentate sunt, în general, valabile şi pentru alte tipuri de eşantioane.

În acest context, sintetizăm unele dintre rezultatele calitative ce au fost justificate experimental:

- În cadrul procesului de eşantionare, creşterea volumului eşantionului E afectează în mod direct gradul de reprezentativitate al acestuia. Trebuie însă precizat că mărirea numărului indivizilor populaţiei W nu modifică semnificativ acest grad de reprezentativitate. Ţinându-se seamă de acest ultim aspect, cu erori neglijabile, populaţii infinite pot fi asimilate cu populaţii finite ce au un număr relativ mare de indivizi.

- La o creştere liniară a volumului eşantionului, creşterea reprezentativităţii sale nu mai este liniară. În această situaţie asistăm la o creştere a reprezentativităţii, mai întâi după o „pantă abruptă", urmând apoi o „pantă deosebit de lină". Aparent paradoxal, acest fapt ne conduce la eşantioane de volum relativ mic (de ordinul miilor) ce sunt însă reprezentative pentru

populații deosebit de numeroase (de ordinul miliardelor), aspect ce a fost validat experimental în prezentul studiu.

- Tehnica simulării stocastice nu presupune cunoştinţe matematice deosebite sau efectuarea de raționamente sofisticate ce conduc la calcule complexe. În plus, algoritmul de simulare poate fi adaptat uşor diverselor variante de lucru.

- Mărirea numărului de „simulări" afectează direct, dar nu proporțional, acuratețea rezultatelor.

- Dintre rezultatele menționate în această lucrare, multe au deja o justificare teoretică. Avantajul simulării stocastice îl constituie însă ilustrarea cantitativă a aspectelor respective.

O observație generală

Dacă se urmăresc comparativ rezultatele simulărilor, acestea vor fi mereu altele datorită utilizării în simulările respective a unor şiruri diferite de numere aleatoare. Eroarea poate fi însă controlată, ea având o interpretare statistică.

Menționăm faptul că rezultatele simulărilor pot fi afectate şi de erori datorate imperfecțiunilor generatorilor de numere aleatoare ce sunt folosiți de diversele produse soft prezente pe piață.

În cazul în care se cer precizii foarte mari privind rezultatele, pe lângă mărirea numărului de simulări, este obligatorie efectuarea unei analize calitative a şirului de valori aleatoare produse de generatorul de numere întâmplătoare ce va fi întrebuințat. De regulă, un astfel de generator produce (prin metode matematice, şi nu numai) valori pseudoaleatoare, ce sunt uniform repartizate într-un interval (de obicei în intervalul [0, 1]; a se vedea, de exemplu, Gentle, 1998).

Menționăm faptul că există o mare varietate de proceduri pentru generarea şirurilor de valori întâmplătoare ce sunt uniform repartizate pe un interval. În acest sens, recomandăm lucrările lui Niederreiter (1995), Yarmolik şi Demidenko (1988) în care sunt prezentate analize aprofundate privind calitățile statistice ale numerelor pseudoaleatoare produse de diverse tipuri de algoritmi matematici de generare a şirurilor de numere întâmplătoare.

Sugestii

Remarcăm câteva posibile extinderi ce urmează a fi abordate prin tehnica simulării stocastice: caracteristica X nu este neapărat dihotomică şi poate fi chiar şi o variabilă continuă, operarea simulată cu mai multe caracteristici, influența unor caracteristici corelate, stabilirea gradului de eroare impus de diferite proceduri de selectare şi precizarea eşantionului optim, determinarea intervalelor de încredere $[d_1, d_2]$ pentru o probabilitate $q_{n, X}$ fixată.

BIBLIOGRAFIE

Septimiu Chelcea, *Cunoaşterea vieţii sociale - Chestionarul şi interviul în ancheta sociologică*, Editura Institutului Naţional de Informaţii, Bucureşti, 1996.

Floyd J. Fowler Jr., *Survey research methods*, Applied Social Research Methods Series, vol. 1, SAGE Publications, London, 1993.

Victor E. McGee, *Principles of statistics - Traditional and bayesian*, The Century Psychology Series, New York.

James E. Gentle, *Random number generation and Monte Carlo methods*, Statistics and Computing, Springer Verlag, New York, 1998.

Christian Gourieroux, *Theorie des sondages*, Collection „Economie et statistiques avancées", Economica, Paris, 1981.

Harry Joe, *Multivariate dependence measures and data analysis*, Computational statistics & Data analysis, vol. 16, Nr.3, 1993.

Marius Iosifescu, Costache Moineagu, Vladimir Trebici, Emiliana Ursianu, *Mică enciclopedie de statistică*, Editura Ştiinţifică şi Enciclopedică, Bucureşti, 1985.

Leslie Kish, *Survey sampling*, John Wiley & Sons, New York, 1963, (prima ediţie).

G. Klimov, *Probability theory and mathematical statistics*, MIR Publishers, Moscova, 1986.

V. Koroliouk, N. Portenko, A. Skorokhod, A. Tourbine, *Aide-mémoire des théorie des probabilités et de statistique mathématique*, MIR, Moscova, 1983.

Barry L. Nelson, *Stochastic modeling*, McGraw Hill, New York, 1995.

Harald Niederreiter, *Pseudorandom vector generation by the multiple-recursive matrix method*, Mathematics of Computation, 64 (1995).

Andrei Novak, *Sondarea opiniei publice*, Editura Studenţească, Bucureşti, 1996.

Athanasios Papoulis, *Probability & Statistics*, Prentice Hall, Englewood Cliffs, 1990.

Traian Rotariu, Petre Iluţ, *Ancheta sociologică şi sondajul de opinie*, Editura POLIROM, Iaşi, 1997.

Des Raj, *Sampling theory*, TATA McGraw-Hill, Bombay, 1968.

V.N. Yarmolik, S.N. Demidenko, *Generation and application of pseudorandom sequences for random testing*, John Wiley and Sons, New York, 1988.

Cătălin Zamfir (coordonator), *Dimensiuni ale sărăciei*.,Editura Expert, Bucureşti, 1995.

*** *SPSS Base 7,5 for Windows*, User's guide, 1997.