

ASUPRA UNOR TEHNICI DE RANDOMIZARE A RĂSPUNSURILOR

¹*Dr., conf. univ., Andrei POȘTARU*
²*Dr., conf. univ., Nicolai PRODAN*

^{1, 2}*Universitatea de Stat din Moldova,*
Republica Moldova, Chișinău, str. A. Mateevici
tel. (+373) 22 577 401, www.usm.md

Abstract

Techniques to organize surveys designed to improve the quality of information collected

Key words: *surveys, sensitive question, anonymity in survey, randomized answer, related question, forced answer*

JEL CLASSIFICATION: C-61

Introducere

Se știe că rezultatele sondajelor sociologice de multe ori nu reflectă adecvat realitatea într-o problemă sau alta. Cauzele rezidă în modul de organizare a lor, dar mai ales în nedorința oamenilor de a răspunde sincer la întrebări, dacă acestea au un caracter intim, sau dacă țin de anumite vicii, sau, mai ales, dacă se referă la încălcarea unor legi. În articolul dat noi ne vom opri asupra unor tehnici de organizare a sondajelor, care au menirea să sporească calitatea informației colectate.

1. Tehnica clasică a răspunsurilor randomizate

Tehnica clasică a răspunsurilor randomizate, care mai este numită și tehnica „întrebării legate” (Related Question Tehnique), a fost propusă de către Stenley Warner în 1965[1]. Ea are la bază principiul “întâmplării dirijate” și din punct de vedere tehnic este simplă.

Presupunem că fiecare persoană dintr-o populație (statistică) aparține grupului A sau grupului B. De regulă, A este un grup de persoane, afectate de o caracteristică negativă (de exemplu sunt consumatori de droguri etc.). Se pune problema estimării proporției persoanelor care aparțin grupului A. Pentru aceasta din populația dată se extrage o selecție de n persoane (respondenți) (în general, conform schemei cu întoarcere). Fiecare respondent primește o fișă cu două afirmații:

1. ”Eu aparțin grupului A” și 2. ”Eu nu aparțin grupului A” (sau „Eu aparțin grupului B”).

Cu ajutorul unui dispozitiv de randomizare („randomizator”), de exemplu, cu ajutorul unui zar, sau a unei rulete etc., respondentul alege din cele două afirmații una: prima afirmație este aleasă cu probabilitatea p , iar a doua cu probabilitatea $1-p$. De exemplu, respondentul aruncă zarul de două ori și dacă cade suma de 7 puncte, atunci alege afirmația 1. Afirmația fiind aleasă, respondentul își exprimă acordul sau dezacordul cu ea prin „da” sau „nu”. Interviewerul, la rândul său, înregistrează răspunsul, fără a cunoaște la care dintre afirmații se referă respondentul. Cunoscând numărul total de răspunsuri “da” ale respondenților poate fi determinată proporția π_A a persoanelor ce aparțin grupului A. Într-adevăr, să introducem evenimentele aleatoare:

$$C = \{\text{un respondent ales la întâmplare reacționează la afirmația aleasă cu “da”}\};$$
$$H_i = \{\text{un respondent ales la întâmplare alege afirmația } i\}, i=1,2.$$

Conform formulei probabilității totale

$$P(C) = P(H_1)P(C|H_1) + P(H_2)P(C|H_2) \text{ sau } P(C) = p\pi_A + (1-p)\pi_B,$$

unde π_A și π_B sunt proporțiile persoanelor ce aparțin grupelor A și B, respectiv. Dacă notăm prin θ proporția respondenților care răspund “da”, atunci

$$\theta = p\pi_A + (1-p)\pi_B. \quad (1)$$

Prin urmare,

$$\pi_A = \frac{\theta - (1-p)\pi_B}{p}. \quad (2)$$

$$\text{Deoarece } B = A^c, \pi_A = \frac{\theta - (1-p)(1-\pi_A)}{p},$$

deci

$$\pi_A = \frac{\theta - (1-p)}{2p-1}, p \neq 0,5. \quad (3)$$

Relația (3) sugerează pentru π_A următorul estimator:

$$\hat{\pi}_A = \frac{S_n - (1-p)}{2p-1}, p \neq 0,5, \quad (4)$$

unde S_n este numărul total de răspunsuri „da”.

Să calculăm dispersia acestui estimator (care este o caracteristică bună a „calității” lui). Se știe că S_n urmează legea binomială de repartiție cu parametrii (n, θ) . Deci

$$D(\hat{\pi}_A) = D\left(\frac{S_n}{n}\right) = D\left(\frac{S_n}{n(2p-1)}\right) = \frac{n\theta(1-\theta)}{n^2(2p-1)^2} = \frac{\theta(1-\theta)}{n(2p-1)^2}$$

Înlocuind aici θ prin expresia $\theta = p\pi_A + (1-p)(1-\pi)$, care rezultă din (3),

obținem

$$\frac{\theta(1-\theta)}{n(2p-1)^2} = \frac{[p\pi_A + (1-p)(1-\pi_A)][p + \pi_A(1-2p)]}{n(2p-1)^2} = \frac{\pi_A^2(2p-1)^2 + \pi_A p(2p-1) + p(1-p) - \pi_A(2p-1)(1-p)}{n(2p-1)^2} = \frac{\pi_A(2p-1)(p-1+p) - \pi_A^2(2p-1)^2 + p(1-p)}{n(2p-1)^2} = \frac{\pi_A(1-\pi_A)}{n} + \frac{p(1-p)}{n(2p-1)^2}.$$

$$\text{Astfel, } D(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{Dn} + \frac{p(1-p)}{n(2p-1)^2}. \quad (5)$$

În cazul interviului “direct”, când întrebarea se pune direct, este evident că

$$D(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n}.$$

Observăm că aici lipsește al doilea termen din (5), care poate fi interpretat ca penalitate (sau plată), cauzată de intervierea indirectă. Este evident, că penalitatea este cu atât mai neimportantă, cu cât p este mai aproape de 0 sau 1.

Probabilitatea π_A poate fi estimată și cu ajutorul unui interval de încredere. În [2] noi deducem un interval de încredere pentru probabilitatea θ și, ținând cont de legătura dintre θ și π_A , pentru π_A construim intervalul

$$\frac{1}{2p-1} \left[\frac{S_n}{n} - (1-p) + u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_n(1-S_n)}{n}} \right] < \pi_A < \frac{1}{2p-1} \left[\frac{S_n}{n} - (1-p) - u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_n(1-S_n)}{n}} \right]$$

pentru $p < 1/2$; dacă $p > 1/2$, atunci intervalul este

$$\frac{1}{2p-1} \left[\frac{S_n}{n} - (1-p) - u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_n(1-S_n)}{n}} \right] < \pi_A < \frac{1}{2p-1} \left[\frac{S_n}{n} - (1-p) + u_{1-\frac{\alpha}{2}} \sqrt{\frac{S_n(1-S_n)}{n}} \right]$$

Tot în [2] se propun și alte procedee pentru calculul valorii aproximative a lui π_A , când se cunoaște numărul total de “da”. Se aduc rezultatele unor aplicații.

Metoda lui Warner are două neajunsuri. În primul rând, crește dispersia estimatorului, deoarece pe lângă dispersia proporției de selecție $D(\frac{S_n}{n})$, se adaugă un termen, legat de procedura de randomizare. În al doilea rând, prezența a două afirmații va conduce la creșterea erorii măsurătorii.

2. Tehnica întrebării libere (Unrelated Question Technique)

Această tehnică [3, 4] propune ca în calitate de afirmație/întrebare alternativă să fie folosită o afirmație /întrebare, ce nu are legătură cu cea “delicată”. De exemplu: 1.”Eu consum droguri;” 2.”Eu sunt abonat la ziare”.

Metoda propusă sporește sinceritatea răspunsurilor, comparativ cu “metoda răspunsului legat” și totodată permite să fie diminuată dispersia erorii suplimentare a selecției. Există două variante ale metodei: cu proporție necunoscută și cu proporție cunoscută a întrebării neutre. Pentru determinarea estimatorului probabilității (proporției) π_A selecția se împarte în două subselecții și ambele sunt examinate conform tehnicii întrebării libere în raport cu aceeași afirmație neutră, dar randomizarea are probabilități diferite, p_1 și p_2 . Ca rezultat pentru proporțiile de “da” obținem doi estimatori:

$$\hat{\theta}_1 = p_1 \pi_A + (1 - p_1) \pi_B, \quad (1)$$

$$\hat{\theta}_2 = p_2 \pi_A + (1 - p_2) \pi_B. \quad (2)$$

Înmulțind (1) și (2) cu $(1-p_2)$ și $(1-p_1)$, respectiv, și scăzându-le, pentru π_A obținem estimatorul

$$\hat{\pi}_A = \frac{\hat{\theta}_1(1-p_2) - \hat{\theta}_2(1-p_1)}{p_1 - p_2}, \quad p_1 \neq p_2.$$

Dispersia estimatorului proporției π_A poate fi determinată cu ușurință:

$$D(\hat{\pi}_A) = \frac{1}{(p_1 - p_2)^2} \left(\frac{\hat{\theta}_1(1-\hat{\theta}_1)(1-p_2)^2}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)(1-p_1)^2}{n_2} \right).$$

Observăm că în cazul dat nu este necesar să cunoaștem proporția π_B .

Atunci când se cunoaște proporția π_B a afirmației neutre, formulele de mai sus pot fi simplificate: cum $\theta = p\pi_A + (1-p)\pi_B$, deducem:

$$\hat{\pi}_A = \frac{\hat{\theta} - (1-p)\pi_B}{p}, \quad D(\hat{\pi}_A) = \frac{\hat{\theta}(1-\hat{\theta})}{np^2}.$$

În acest caz modelul poate fi aplicat dacă cunoaștem proporția π_B sau dacă aceasta poate fi calculată statistic. Pentru întrebări neutre cu proporție cunoscută a răspunsului poate servi, de

exemplu, luna de naștere: ”Eu m-am născut în luna iunie”, $\pi_B = \frac{30}{365}$; sau felul de a fi stângaci: ”Eu sunt stângaci”, $\pi_B=0,12$. Dificultatea în acest caz constă în faptul că repartiția răspunsurilor în populația statistică poate fi diferită de cea din selecția X_n . Mai mult, să găsim întrebări cu repartiții (proporții) cunoscute, este o problemă dificilă.

3. Tehnica răspunsului forțat (Forced Response Technique)

Această tehnică constă în aceea că în loc de întrebarea neutră respondentul este ”condus” în mod ”forțat” spre formularea unui răspuns ”în direcția” întrebării delicate. De exemplu, pot fi propuse următoarele două afirmații: 1.”Eu consum droguri”; 2.”Spuneți (răspundeți) „da”.

Se ia, de exemplu, $p=1/2$: se aruncă moneda simetrică și dacă cade stema, atunci se răspunde la prima afirmație, dacă cade banul, atunci se răspunde ”da”. Astfel, atunci când cade banul respondentul trebuie să răspundă ”da”(adică în direcția răspunsului delicat), indiferent, consumă sau nu consumă el droguri; dacă însă cade stema, atunci el trebuie să răspundă sincer (”da” sau „nu”).

Este de așteptat că o jumătate dintre respondenți vor răspunde la a doua afirmație, adică cu ”da”, deoarece în jumătate de cazuri va cădea ”banul”, cealaltă jumătate vor răspunde sincer la prima afirmație. Prin urmare, dacă, de exemplu, în total 60% de respondenți vor răspunde afirmativ, adică cu ”da”, atunci, având în vedere că 50% au răspuns afirmativ din cauza că moneda a căzut cu banul, vom putea trage concluzia: 10% este proporția celor din selecție care consumă droguri (am ținut cont de subselecția, determinată de fața cu ”banul”). Această tehnică (a răspunsului forțat) permite cercetătorului să estimeze amploarea caracteristicii A, de care este interesat, fiind scutit de căutarea unei întrebări neutre. Mai mult, cu ajutorul acestei tehnici pentru estimatorul proporției π_A putem obține o dispersie care este doar de două ori mai mare decât dispersia de selecție din cazul întrebării directe. Este evident, că în cazul acestei tehnici, pentru proporția răspunsurilor ”da” putem aplica formula

$$\theta = p\pi_A + (1 - p)\pi_B \text{ cu } \pi_B = 1: \theta = p\pi_A + 1 - p,$$

de unde $\pi_A = \frac{\theta - (1-p)}{p}$. Prin urmare, $\hat{\pi}_A = \frac{S_n - (1-p)}{p}$ este un estimator de verosimilitate maxima nedepasat pentru probabilitatea π_A . Dispersia lui este: $D(\hat{\pi}_A) = \frac{\theta(1-\theta)}{np^2}$. Înlocuind aici θ prin expresia respectivă deducem cu ușurință:

$$\theta(1 - \theta) = (p\pi_A + 1 - p)(p - \pi_A) = p^2\pi_A(1 - \pi_A) + p(1 - p)(1 - \pi_A).$$

Deci dispersia estimatorului $\hat{\pi}_A$ poate fi scrisă astfel:

$$D(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{(1-p)(1-\pi_A)}{np}.$$

4. Tehnica lui Mangat

În lucrarea [5] fiecare interviuat din selecția de volum n folosește două randomizatoare. Randomizatorul R_1 îi permite respondentului să obțină două afirmații: 1.”Eu aparțin grupului A” (cu probabilitatea T) și 2.”Utilizați randomizatorul R_2 ”(cu probabilitatea 1-T). Randomizatorul R_2 este analog cu randomizatorul lui Warner: el îi prezintă respondentului două afirmații: 1.”Eu aparțin grupului A”(cu probabilitatea p) și 2.”Eu nu aparțin grupului A”(cu probabilitatea (1-p)). Constatăm cu ușurință, că pentru probabilitatea θ_1 a răspunsului ”da”avem formula

$$\theta_1 = T\pi_A + (1 - T)[p\pi_A + (1 - p)(1 - \pi_A)]. \quad (4)$$

De aici

$$\pi_A = \frac{\theta_1 - (1-T)(1-p)}{2p-1+2T(1-p)}.$$

Prin urmare, în acest caz pentru probabilitatea π_A estimatorul de verosimilitate maximă este

$$\hat{\pi}_A = \frac{\frac{S_n}{n} - (1-T)(1-p)}{2p-1+2T(1-p)},$$

unde S_n este numărul răspunsurilor “da”. Evident,

$$D(\hat{\pi}_A) = \frac{D\left(\frac{S_n}{n}\right)}{[2p-1+T(1-p)]^2} = \frac{\theta_1(1-\theta_2)}{n[2p-1+T(1-p)]^2}.$$

Înlocuind aici θ_1 din (4) deducem

$$D(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{(1-T)(1-p)[1-(1-T)(1-p)]}{[2p-1+2T(1-p)]^2}.$$

5. Tehnica întrebărilor încrucișate

În situația când proporția întrebării neutre nu se cunoaște Moors a propus [6] ca selecția X să fie împărțită în două subselecții X_{n_1} și X_{n_2} ($n_1 + n_2 = n$); X_{n_1} se va examina conform tehnicii întrebării libere, iar X_{n_2} va fi folosită pentru estimarea proporției π_B (a întrebării neutre), conform tehnicii întrebării directe. Este de așteptat ca această tehnică să conducă la diminuarea dispersiei estimatorului proporției π_A . Examinând X_{n_1} și X_{n_2} obținem relațiile

$$\theta_1 = p_1\pi_A + (1-p_1)\pi_B, \quad \theta_2 = \pi_B$$

și, respectiv, estimatorii de verosimilitate maximă

$$\hat{\pi}_A = \frac{\hat{\theta}_1 - (1-p_1)\hat{\pi}_B}{p_1}, \quad \hat{\pi}_B = \hat{\theta}_2.$$

Cum $\hat{\theta}_1 = \frac{S_{n_1}}{n_1}$ și $\hat{\theta}_2 = \frac{S_{n_2}}{n_2}$ (S_{n_i} este numărul de “da” în selecția X_{n_i}), pentru estimatorul proporției π_A avem:

$$D(\pi_A) = \frac{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{(1-p_1)^2(1-\pi_B)\pi_B}{n_2}}{p_1^2}.$$

Abordarea lui Moors și-a găsit dezvoltare în 1973 [7], când a fost propusă o tehnică cu două alternative neutre a_1 și a_2 : selecția X_n se împarte în două subselecții, X_{n_1} și X_{n_2} ; X_{n_1} este examinată conform tehnicii întrebării libere, adică se examinează întrebarea delicată cu una dintre alternativele neutre, de exemplu, cu a_1 , dar prin întrebare directă se examinează și alternativa a_2 pentru a afla proporția ei în X_{n_1} ; în X_{n_2} , conform tehnicii întrebării libere, este examinată întrebarea delicată cu alternativa neutră a_2 și, totodată, prin întrebare directă este examinată alternativa neutră a_1 pentru a afla proporția ei în X_{n_2} . Această tehnică poate fi descrisă de relațiile:

$$\theta_1 = p\pi_A + (1-p)\pi_{B_2}, \quad \theta_2 = p\pi_A + (1-p)\pi_{B_1}.$$

De aici deducem estimatorii

$$\hat{\pi}_A^{(1)} = \frac{\hat{\theta}_1 - (1-p)\hat{\pi}_{B_2}}{p}, \quad \hat{\pi}_A^{(2)} = \frac{\hat{\theta}_2 - (1-p)\hat{\pi}_{B_1}}{p}.$$

Aici $\hat{\theta}_1$ și $\hat{\theta}_2$ sunt proporțiile răspunsurilor de “da” în cadrul tehnicii de randomizare pentru subselecțiile X_{n_1} și, respectiv, X_{n_2} ; $\hat{\pi}_{B_1}$ și $\hat{\pi}_{B_2}$ sunt probabilitățile de răspunsuri afirmative la întrebarea directă în selecția X_{n_2} și, respectiv, la întrebarea directă în selecția X_{n_1} ; p este probabilitatea ca respondentul să primească afirmația (întrebarea) delicată. Pentru dispersia de selecție, evident, avem:

$$D(\hat{\pi}_A^{(1)}) = \frac{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{(1-p)^2\pi_{B_2}(1-\pi_{B_2})}{n_2}}{p^2},$$

$$D(\hat{\pi}_A^{(2)}) = \frac{\frac{\theta_2(1-\theta_2)}{n_2} + \frac{(1-p)^2\pi_{B_1}(1-\pi_{B_1})}{n_1}}{p^2}.$$

$\hat{\pi}_A^{(1)}$ și $\hat{\pi}_A^{(2)}$ sunt estimatori independenți (și nedeplasați) ai parametrului π_A ; pentru π_A vom considera un estimator nou:

$$\hat{\pi}_A = w\hat{\pi}_A^{(1)} + (1-w)\hat{\pi}_A^{(2)}, \quad \text{unde} \quad w = \frac{D(\hat{\pi}_A^{(2)})}{D(\hat{\pi}_A^{(1)}) + D(\hat{\pi}_A^{(2)})}.$$

Deci,

$$\hat{\pi}_A = \frac{D(\hat{\pi}_A^{(2)})\hat{\pi}_A^{(1)}}{D(\hat{\pi}_A^{(1)}) + D(\hat{\pi}_A^{(2)})} + \frac{D(\hat{\pi}_A^{(1)})\hat{\pi}_A^{(2)}}{D(\hat{\pi}_A^{(1)}) + D(\hat{\pi}_A^{(2)})}.$$

Acest estimator este nedeplasat (pentru parametrul π_A); și are dispersia mai mică decât fiecare dintre dispersiile estimatorilor $\hat{\pi}_A^{(1)}$ și $\hat{\pi}_A^{(2)}$. Într-adevăr, să calculăm dispersia lui $\hat{\pi}_A$.

Pentru simplitatea scrierii vom nota: $D(\hat{\pi}^{(i)})$ prin $D_i, i=1,2$.

$$D(\hat{\pi}_A) = \frac{D_2^2}{(D_1 + D_2)^2} D_1 + \frac{D_1^2}{(D_1 + D_2)^2} D_2 = \frac{D_1 D_2 (D_1 + D_2)}{(D_1 + D_2)^2} = \frac{D_1 D_2}{D_1 + D_2}$$

Calculăm diferența

$$\frac{D_1 D_2}{D_1 + D_2} - D_1 = \frac{D_1 D_2 - D_1^2 - D_1 D_2}{D_1 + D_2} = -\frac{D_1^2}{D_1 + D_2} < 0.$$

De aici rezultă că $D(\hat{\pi}_A) < D(\hat{\pi}_A^{(1)})$. În mod analog constatăm că $D(\hat{\pi}_A) < D(\hat{\pi}_A^{(2)})$.

Remarcă. În clasa estimatorilor lui π_A de forma $w\hat{\pi}_A^{(1)} + (1-w)\hat{\pi}_A^{(2)}$ minimumul dispersiei este atins pentru $w = \frac{D_2}{D_1 + D_2}$. Într-adevăr, să examinăm funcția $f(x) = D(x\hat{\pi}_A^{(1)} + (1-x)\hat{\pi}_A^{(2)}) = x^2 D_1 +$

$(1-x)^2 D_2$, $f'(x) = 2x D_1 - 2(1-x) D_2 = 2x(D_1 + D_2) - 2D_2$. Punctul $x = \frac{D_2}{D_1 + D_2}$ este punctul de minimum, ceea ce trebuia de demonstrat.

În [8] se propune să luăm $w = \frac{1}{D_1 + D_2}$, ceea ce nu are o argumentare rațională.

CONCLUZII

Au fost examinate câteva tehnici de organizare a sondajelor, menite să contribuie la creșterea veridicității rezultatelor. Tehnicile de randomizare evoluează, apar tehnici noi. Sperăm că în cercetările sociologice, care se efectuează în Republica Moldova, își vor găsi aplicație și tehnicile, despre care vorbim în materialul de față.

BIBLIOGRAFIE

- [1.] Warner S.L. Randomized response: a survey technique for eliminating evasive answer bias. În: J. Am. Statist. Assoc. 1965. Vol. 60, p. 63-69.
- [2.] Poștaru A., Prodan N., Benderschi O. Studiarea probabilității π în metoda răspunsurilor randomizate. În: Studia Universitatis, Seria Științe exacte și Economice, nr.7 (87), Chișinău CEP USM, 2015, p. 31-35.
- [3.] Horvitz D. G., Shah B. V., Simmons W. R. The unrelated question randomised response Model. În: Proceedings of the Social Statistics Section, ASA, 1967. Vol. 326. p. 65- 72.
- [4.] Greeberg B., Abul-Ela A. –L., Simmons W., Horvitz W. The unrelated question randomized response model: theoretical framework. În: J. Am. Statist. Assoc. 1969. Vol. 64, p. 525-530.
- [5.] Mangat N. S., Singh R. An alternative randomized response procedure. În: Biometrika. 1990. Vol. 77, p. 439-442.
- [6.] Moors J. J. A. Optimization of the unrelated question randomized response model. În: J. Am. Statist. Assoc. 1971. Vol.66, p. 627-629.
- [7.] Folsom R. E., Greenberg B. G., Horvitz D. G., Abernathy J. R. În: J. Am. Statist. Assoc. 1973. Vol. 68, p. 525-530.
- [8.] Калинин К. О. Исследование социально приемлемого поведения в России: Мониторинг общественного мнения 1(119). Январь – февраль. 2014.