STUDIA UNIVERSITATIS MOLDAVIAE, 2015, nr.7(87)
Seria "Ştiinţe exacte şi economice" ISSN 1857-2073 ISSN online 2345-1033 p.31-35

STUDIEREA PROBABIBILITĂŢII π

ÎN METODA RĂSPUNSURILOR RANDOMIZATE

Andrei POŞTARU, Nicolae PRODAN, Olga BENDERSCHI

Universitatea de Stat din Moldova

Este examinată probabilitatea π de a obţine răspunsuri sincere în sondaje. Sunt expuse câteva metode de estimare a lui π: metoda intervalelor de încredere, metoda verosimilităţii maxime, verificarea ipotezelor statistice simple etc.

Cuvinte-cheie: *sondaje, probabilitate, statistică, estimaţie, ipoteze statistice, interval de încredere, răspunsuri randomizate.*

RANDOMIZED RESPONSE METHOD: ESTIMATIONS OF THE PROBILITY π

The probability π to get honest answers in surveys is considered. There are presented several methods for π estimating: confidence interval method, maximum likelihood estimation, statistical hypotheses testing etc.

Keywords: *surveys, probability, statistics, estimation, statistical hypotheses, confidence interval, randomized response.*

Introducere

Rezultatele cercetărilor sociologice, de regulă, conţin erori, uneori foarte grave, care pot reduce la zero veridicitatea lor. Oamenii nu sunt sinceri în răspunsurile lor la întrebări intime, la fel şi la întrebări ce ţin de consumul de droguri, violenţa în familie, eschivarea de la achitarea corectă a impozitelor etc. Astfel, problema sporirii sincerităţii răspunsurilor este actuală [1].

1. Modelul de sondaj al lui Stanley Warner

Stanley Warner [3] a propus un model de sondaj care sporeşte substanţial veridicitatea răspunsurilor. Una dintre variantele posibile constă în următoarele. Presupunem că se efectuează un sondaj pentru a determina amploarea în societate a unui fenomen *C* (negativ). Să notăm prin π probabilitatea ca un individ luat la întâmplare să fie afectat de fenomenul *C*. Pentru evaluarea acestei probabilităţi se poate lua un eşantion de *N* indivizi şi dacă *k* dintre ei sunt afectaţi de fenomenul *C*, atunci putem spune $\pi \approx \frac{k}{N}$. Spre regret, numărul *k* este necunoscut. Prin întrebări directe este imposibil să aflăm valoarea adevărată a lui *k*. Se propune următorul procedeu.

Fiecărui dintre *N* respondenţi i se înmânează o fişă cu două întrebări:

1. Este adevărat că sunteţi afectat de fenomenul *C*?
2. Este adevărat că nu sunteţi afectat de fenomenul *C*?

Fiecare respondent va răspunde doar cu „da" sau „nu" şi doar la una dintre aceste întrebări, întrebare „aleasă" la întâmplare. Cum se poate organiza aceasta?

De exemplu, odată cu fişa fiecărui respondent i se înmânează şi un zar pe care urmează sa-l arunce în sus, o singură dată. Acei care obţin, de exemplu, cel mult două puncte, vor răspunde la prima întrebare, iar cei care obţin trei puncte şi mai multe vor răspunde la întrebarea a doua. Totodată, nimeni nu trebuie să cunoască rezultatul aruncării şi, în consecinţă, nici la care din întrebare se dă răspuns. Astfel, fiecare respondent răspunde doar la una din întrebări: la prima întrebare cu probabilitatea $p = \frac{1}{3}$, iar la a doua cu probabilitatea $1 - p = \frac{2}{3}$. Bineînţeles, putem alege şi alte probabilităţi, în loc de zar utilizând alte mecanisme (tabele de numere aleatoare etc.). Cazurile $p = 1$ şi $p = 0$ conduc la cazul de sondaj clasic.

În condiţiile în care este pus respondentul acum, el ar trebui (cel puţin din punct de vedere psihologic) să se simtă în afara oricărui „pericol" venit din cauza unui răspuns sincer.

Într-adevăr, ambele răspunsuri posibile, şi „da" şi „nu", pot avea o interpretare „pozitivă": răspunsul „da" poate fi atribuit celei de a doua întrebări, iar „nu" primei întrebări.

Astfel, putem spera că în urma unui astfel de sondaj se vor obţine rezultate mult mai veridice.

În [2] pentru probabilitatea π, de care suntem interesaţi, noi am stabilit formula

$$\pi = \frac{\frac{n}{N} - 1 + p}{2p - 1}; \quad \left(p \neq \frac{1}{2}\right). \tag{1}$$

Aici N este numărul respondenţilor, n este numărul răspunsurilor de „da" date de ei, iar p reprezintă probabilitatea cu care un respondent răspunde la prima întrebare. Tot acolo, în [2], este examinată probabilitatea π ca funcţie de n, n fiind interpretat ca o variabilă aleatoare.

În prezenta lucrare este studiată în continuare probabilitatea π, sunt expuse câteva metode de estimare, este testată o ipoteză statistică privind valoarea lui π.

Numărul n de „da" poate fi privit ca o variabilă aleatoare. Se constată cu uşurinţă că n are repartiţie binomială $n \sim B(N, \psi)$, unde $\psi = 1 - p + (2p - 1)\pi$, şi, deci, $Mn = N(1 - p + (2p - 1)\pi)$; M este simbolul valorii medii.

Astfel, pentru probabilitatea π obţinem formula (exactă)

$$\pi = \frac{\frac{Mn}{N} - 1 + p}{2p - 1}, \tag{2}$$

formulă ce diferă de (1) prin aceea că n este înlocuit cu valoarea lui medie.

În continuare vom examina rolul lui ψ la stabilirea valorii probabilităţii π. Amintim că ψ reprezintă probabilitatea răspunsului „da" al unui respondent şi este o funcţie de π.

Mai întâi menţionăm că: a) dacă $p < \frac{1}{2}$, atunci $\psi = 1 - p + (2p - 1)\pi \in [p, 1 - p]$;

b) dacă $p > \frac{1}{2}$, atunci $\psi = 1 - p + (2p - 1)\pi \in [1 - p, p]$.

2. Unele metode de evaluare a probabilităţii π

Vom expune unele metode de evaluare a lui π, atunci când se cunoaşte numărul răspunsurilor de „da" obţinute.

Metoda 1. Presupunem că se efectuează un sondaj în care se obţin n^* de „da". Dacă $p < \frac{1}{2}$ şi $n^* \in [Np, N(1 - p)]$, sau dacă $p > \frac{1}{2}$ şi $n^* \in [N(1 - p), Np]$, atunci (în ambele cazuri) drept estimaţie a funcţiei parametrice $\psi(\pi) = 1 - p + (2p - 1)\pi$ putem lua statistica $\hat{\psi} = \frac{n^*}{N}$. Prin urmare, în acest caz pentru probabilitatea π obţinem estimaţia $\hat{\pi} = \frac{\hat{\psi} - 1 + p}{2p - 1}$.

Metoda 2. Pentru parametrul ψ, ca probabilitate, putem construi un interval de încredere.

Se ştie că frecvenţa relativă $\frac{n^*}{N}$ este o estimaţie eficientă a probabilităţii ψ. Alegând un prag de semnificaţie α, construim intervalul de încredere

$$\left(\frac{n^*}{N} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{n^*}{N}\left(1 - \frac{n^*}{N}\right)}{N}}, \frac{n^*}{N} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{n^*}{N}\left(1 - \frac{n^*}{N}\right)}{N}} \right)$$

(aici $u_{1-\frac{\alpha}{2}}$ este „$1 - \frac{\alpha}{2}$ - cuantila" repartiţiei normale $N(0,1)$).

Acest interval acoperă valoarea parametrului ψ cu probabilitatea 1 – α. Mai important pentru noi este să construim un interval de încredere pentru probabilitatea π. Alegând un prag de semnificaţie α şi ţinând cont de legătura dintre π şi ψ, putem scrie:

$$\frac{n^*}{N} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}} < 1-p+(2p-1)\pi < \frac{n^*}{N} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}},$$

adică

$$\frac{n^*}{N} + p - 1 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}} < (2p-1)\pi < \frac{n^*}{N} + p - 1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}}.$$

În funcţie de valoarea lui *p*, conchidem:

a) dacă $p < \frac{1}{2}$, atunci

$$\frac{1}{2p-1}\left[\frac{n^*}{N} + p - 1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}}\right] < \pi < \frac{1}{2p-1}\left[\frac{n^*}{N} + p - 1 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}}\right];$$

b) dacă $p > \frac{1}{2}$, atunci

$$\frac{1}{2p-1}\left[\frac{n^*}{N} + p - 1 - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}}\right] < \pi < \frac{1}{2p-1}\left[\frac{n^*}{N} + p - 1 + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\frac{n^*}{N}\left(1-\frac{n^*}{N}\right)}{N}}\right].$$

Metoda 3. Pentru π putem construi o estimaţie de verosimilitate maximă.

Amintim că ψ este funcţie de probabilitatea π: $\psi = 1-p+(2p-1)\pi$. Evident, $P(n=n^*) = C_N^{n^*}\psi^{n^*}(1-\psi)^{N-n^*}$.

Conform teoremei Moivre–Laplace,

$$P(n=n^*) \approx \frac{1}{\sqrt{N\psi(1-\psi)}} \cdot \varphi\left(\frac{n^* - N\psi}{\sqrt{N\psi(1-\psi)}}\right),$$

astfel probabilitatea $P(n=n^*)$ este o funcţie g(π).

Ca estimaţie a lui π putem lua acea valoare π^* a lui π, pentru care probabilitatea $P(n=n^*)$ este maximală. Deoarece funcţia *g*(π), ca funcţie de π, este destul de complicată şi greu de cercetat la maximum, se va proceda în felul următor: în intervalul (0,1) cu un pas oarecare *h* (de exemplu, cu *h* = 0,01) examinăm pe rând valorile lui π, calculând pentru fiecare din ele valoarea funcţiei *g*(π). Valoarea π^* va fi acea valoare a lui π, pentru care *g*(π) ia valoarea maximă.

Metoda 4. Amintim că ψ reprezintă probabilitatea ca un respondent să răspundă cu „da"; $\psi = 1-p+(2p-1)\pi$ şi deci $\pi = \frac{\psi-(1-p)}{2p-1}$. Observăm că legătura dintre ψ şi π este bijectivă.

În continuare, pentru a determina valoarea (aproximativă) a lui ψ şi, prin urmare, şi a lui π, problema respectivă este formulată ca o problemă de verificare a unei ipoteze statistice privind parametrul repartiţiei binomiale (am menţionat deja că ψ este parametrul unei repartiţii binomiale).

Într-adevăr, problema privind determinarea probabilității ψ poate fi privită ca o problemă de verificare a ipotezelor, anume:

Să se verifice ipoteza nulă

H_0: $\psi = \psi_0$

cu alternativa:

H_1: $\psi > \psi_0$, $\psi_0 \in (0,1)$.

Pentru realizarea practică a acestei idei vom formula următoarea serie de probleme:

Să se verifice ipoteza nulă H_0: $\psi = kh$ cu alternativa H_1: $\psi > kh$ $\left(k = 1, 2, ..., \left[\frac{1}{h}\right]\right)$ sau, pe scurt:

H_0: $\psi = kh$,

H_1: $\psi > kh$.

Aici $h \in (0,1)$ este un număr mic (de exemplu, $h = 0{,}01$). La rezolvarea acestei serii de probleme ne vom conduce de următorul algoritm:

Se ia $k = 1$ şi, aplicând un test statistic corespunzător (care urmează să fie descris), se rezolvă problema:

H_0: $\psi = h$,

H_1: $\psi > h$.

Dacă ipoteza H_0 este acceptată, atunci problema privind valoarea probabilității ψ este rezolvată: $\psi = h$. Dacă ipoteza H_0 este respinsă şi, prin urmare, este acceptată alternativa H_1 ($\psi > h$), atunci se formulează problema:

H_0: $\psi = 2h$,

H_1: $\psi > 2h$.

Se aplică testul menționat şi se rezolvă această problemă. Dacă ipoteza nulă H_0 este acceptată, atunci problema despre valoarea probabilității ψ este rezolvată: $\psi = 2h$. În caz contrar este acceptată alternativa H_1, se ia $k = 3$ şi se testează ipoteza H_0: $\psi = 3h$ cu alternativa H_1: $\psi > 3h$ etc.

Sunt posibile două variante de încheiere a acestui algoritm:

a) pentru un k_0 oarecare ipoteza H_0 este acceptată şi valoarea probabilității ψ este stabilită: $\psi = k_0 h$;

b) toate ipotezele H_0: $\psi = kh$, $k = 1, 2, ..., \left[\frac{1}{h}\right]$ sunt respinse şi, prin urmare, seria noastră de probleme se încheie cu acceptarea ipotezei H_1: $\psi > \left[\frac{1}{h}\right] \cdot h$.

În acest caz, problema privind valoarea probabilității π încă nu este rezolvată, doar se poate constata că $\psi > \left[\frac{1}{h}\right] \cdot h$, adică $|1 - \psi| < h$. Deci, putem considera că $\psi \approx 1$, dacă h este suficient de mic. Dar putem micşora „pasul", luând, de exemplu, un nou „pas" $h^* = 0{,}01h$ şi trecând la rezolvarea unei serii noi de probleme:

H_0: $\psi = kh^*$,

H_1: $\psi > kh^*$, $k = 1, 2, ..., \left[\frac{1}{h^*}\right]$.

Şi de data aceasta vom aplica algoritmul expus mai sus. Din nou putem avea cazurile a) şi b), menționate deja. În cazul b) vom lua $\psi \approx 1$ sau vom continua testarea cu un pas mai mic, rezolvând o serie nouă de probleme.

În continuare vom descrie testul statistic care, în algoritmul expus mai sus, se aplică la verificarea ipotezelor de tipul:

H_0: $\psi = \psi_0$

cu alternativa

H_1: $\psi > \psi_0$, $\psi_0 \in (0,1)$.

Fie că n^* dintre cei N respondenţi au răspuns „da". Conform teoremei integrale Moivre-Laplace, dacă ipoteza H_0 este adevărată, atunci pentru valori mari ale lui N ($N > 50$) repartiţia variabilei aleatoare

$$Z = \frac{n^* - N\psi_0}{\sqrt{N\psi_0(1-\psi_0)}}, \text{ sau, echivalent, a lui } Z = \frac{\frac{n^*}{N} - N\psi_0}{\sqrt{\frac{\psi_0(1-\psi_0)}{N}}}$$

este aproximativ normală cu parametrii 0 şi 1.

Această variabilă aleatoare este luată în calitate de statistică a testului; regiunea critică a testului statistic pentru pragul de semnificaţie α este mulţimea valorilor lui Z, determinate de inegalitatea $z > u_{1-\alpha}$, unde $u_{1-\alpha}$ este „1-α – cuantila" funcţiei de repartiţie $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}u^2} du$ ": $\Phi(u_{1-\alpha}) = 1 - \alpha$.

Metoda 5. La efectuarea a N experimente Bernoulli cu probabilitatea succesului egală cu $\psi = 1 - p + (2p-1)\pi$ în fiecare experiment, cel mai probabil număr de succese n_0 este partea întreagă a numărului fracţionar $(N-1)(1-p+(2p-1)\pi)$; în caz contrar n_0 are două valori:

$$(N-1)(1-p+(2p-1)\pi);\ (N-1)(1-p+(2p-1)\pi)+1.$$

Aceasta ne poate permite să estimăm probabilitatea π. Într-adevăr, fie că am efectuat un sondaj cu N respondenţi şi n^* dintre aceştia au răspuns „da". Cunoscând numărul n^* vom calcula succesiv valorile numărului n_0 pentru π = h, $2h$, ..., eventual până la $\left[\frac{1}{h}\right].h$, oprindu-ne la valoarea π = kh, pentru care $n_0 = n^*$. În consecinţă, considerăm π ≈ kh.

3. Aplicaţii şi concluzii

Metodele descrise mai sus au fost testate pe un eşantion de 80 de studenţi de la USM. Au fost abordate trei fenomene: 1) mituirea profesorilor; 2) dezamăgirea în alegerea specialităţii (facultăţii); 3) copierea la examen. În toate cazurile s-a luat $p = \frac{1}{3}$, adică la prima din cele două întrebări formulate în fişe a răspuns, în medie, fiecare al treilea dintre cei chestionaţi. Numărul răspunsurilor de „da" în legătură cu cele trei fenomene este, respectiv: 1) $n_1 = 45$; 2) $n_2 = 37$; 3) $n_3 = 85$.

Se verifică cu uşurinţă că aceste rezultate satisfac condiţiile care permit să aplicăm metodele de estimare a probabilităţii π, expuse mai sus. Astfel, 4 din aceste metode au condus la următoarele valori aproximative (sau intervale de încredere pentru π):

Metoda 1. $\pi_1 \approx 0{,}3125$; $\pi_2 \approx 0{,}6125$; $\pi_3 \approx 0{,}6875$.

Metoda 2. $\pi_1 \in (0{,}0039;\ 0{,}636)$; $\pi_2 \in (0{,}2887; 0{,}9285)$; $\pi_3 \in (0{,}3667;\ 0{,}9884)$.

Metoda 3. $\pi_1 \approx 0{,}30$; $\pi_2 \approx 0{,}61$; $\pi_3 \approx 0{,}70$.

Metoda 5. $\pi_1 \in (0{,}297;\ 0{,}353)$; $\pi_2 \in (0{,}593;\ 0{,}629)$; $\pi_3 \in (0{,}67;\ 0{,}70)$.

Trebuie să recunoaştem că volumul eşantionului este destul de modest şi sondajul nu s-a efectuat în toate grupele concomitent. Rolul sondajului este unul mai mult ilustrativ. Noi nu pretindem că rezultatele lui reflectă exact situaţia reală în problemele respective. Totodată, trebuie să menţionăm faptul că referitor la mită în învăţământ rezultate similare cu ale noastre am întâlnit în unele materiale oficiale.

Bibliografie:

1. NATHAN, G. Bibliographie de la métode des résponses randomisées (1965-1987). En: *Statistique Canada: Tehniques d'enquête*, 1988, no.14, p.351-365.
2. POŞTARU, A., PRODAN, N., BENDERSCHI, O. Asupra metodei răspunsurilor randomizate. În: *Materialele Conferinţei internaţionale „Modelarea Matematică, Optimizare şi Tehnologii Informaţionale"*. Ediţia a IV-a. Chişinău: Evrica, 2014. Vol.I, p.184-190. ISNB 978-9975-62-365-0
3. WARNER, S. Randomizes response: a survey technique for eliminating evasive answer bias. In: *J. Am. Statist. Assoc.*, 1965, no.60, p.63-69.

Prezentat la 01.10.2015