Metody Ilościowe w Badaniach Ekonomicznych
Tom XXVI/1, 2025, pp. 12–21

Received: 10.01.2025
Accepted: 05.03.2025

OPTYMALNY DOBÓR PRÓBY W MODELU KRZYŻOWYM ORAZ TRÓJKĄTNYM PYTAŃ DRAŻLIWYCH

Stanisław Jaworski https://orcid.org/0000-0002-6169-2886
Wojciech Zieliński https://orcid.org/0000-0003-0749-8764
Instytut Ekonomii i Finansów
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: stanislaw_jaworski@sggw.edu.pl
e-mail: wojciech_zielinski@sggw.edu.pl

Streszczenie: W pracy zajmujemy się optymalnym doborem próby w wybranych modelach: krzyżowym oraz trójkątnym. Kryterium optymalności polega na doborze rozmiaru próby, dla którego względna długość przedziału ufności nie przekracza ustalonej wartości.

Słowa kluczowe: pytania drażliwe, model nierandomizowanych odpowiedzi

JEL classification: C83, C99

WSTĘP

W wielu badaniach społecznych i ekonomicznych badacz chce uzyskać odpowiedź na pytania drażliwe dotykające sfer intymnych, a odpowiedzi mogą stawiać ich w złym świetle. Istnieje wiele zachowań, które nie są dobrze odbierane w społeczeństwie, a przyznanie się do nich wiąże się z lękiem o zdemaskowanie. Trudno się przyznać do problemów z alkoholem, przemocy w rodzinie czy nieprzestrzegania prawa. Nawet zapewnienie przez ankietera o pełnej anonimowości badania może nie być wystarczającym powodem do udzielania szczerych odpowiedzi. Dlatego też stosowane są pewne wybiegi, by mimo wszystko w miarę rzetelnie ocenić w populacji skalę zjawiska drażliwego. Ogólną ideą tych pomysłów jest zadawanie obok interesującego badacza pytania drażliwego pytania neutralnego niezwiązanego z badaną kwestią.

https://doi.org/10.22630/MIBE.2025.26.1.2

Pierwszy pomysł na realizację tego zadania pochodzi od Warnera [Warner 1965]. Zaproponował on stosowanie randomizacji odpowiedzi. To podejście wymaga zastosowania pewnego „urządzenia" dokonującego takiej randomimzacji. Podejście Warnera było modyfikowane na wiele sposobów. Tian i in. [2007] zaproponowali inny sposób działania nie wymagający randomizacji odpowiedzi. Zaproponowali oni dwa modele zadawania pytań: model krzyżowy oraz model trójkątny. Spośród wielu prac można przykładowo wymienić Horvitz i in. [1967], Greenberg i in. [1969], Yu i in. [2008], Tan i in. [2009], Jaworski i Zieliński [2023] oraz Jaworski [2025].

Nas interesuje przedziałowa estymacja odsetka pytań drażliwych. Chcemy znaleźć minimalną wielkość próby pozwalającą na estymację względnego odsetka przy zachowaniu pewnej anonimowości badanych osób. Anonimowość ankietowanej osoby mierzona będzie prawdopodobieństwem odgadnięcia odpowiedzi na pytanie drażliwe na podstawie znajomości całościowej odpowiedzi.

Poniżej rozważymy dwa nierandomizowane modele stosowane w badaniach drażliwych: model krzyżowy oraz model trójkątny.

Model Krzyżowy. W modelu krzyżowym niezrandomizowanych odpowiedzi zadawane są dwa pytania, drażliwe i neutralne. Jeżeli respondent może odpowiedzieć na oba twierdząco, wpisuje do ankiety *TAK*. Jeżeli na oba może odpowiedzieć przecząco, również wpisuje *TAK*. W przeciwnym przypadku wpisuje *NIE*. Ważne jest by oba pytania były tak dobrane, aby udzielane na nie odpowiedzi respondenta były realizacjami niezależnych zmiennych losowych.

Formalnie, niech Z oznacza zmienną losową o rozkładzie dwupunktowym, która przyjmuje wartość 1, gdy respondent na oba pytania udzieliłby tej samej odpowiedzi oraz 0 w przeciwnym przypadku:

$$Z = \begin{cases} 1, & \text{gdy obie odpowiedzi są takie same,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases} \tag{1}$$

Niech ρ oznacza prawdopodobieństwo przyjęcia przez zmienną Z wartości 1. Zatem

$$P_\pi(Z = 1) = \rho = q\pi + (1 - q)(1 - \pi), \tag{2}$$

gdzie π i q oznaczają prawdopodobieństwa, że respondent odpowiedziałby twierdząco odpowiednio na pytanie drażliwe oraz neutralne. Zakładamy, że prawdopodobieństwo q jest znane. By π było identyfikowale, to prawdopodobieństwo q nie może być równe 0.5. Próba niezależnych zmiennych losowych $Z_1, \ldots, Z_n$ o tym samym rozkładzie prawdopodobieństwa co Z, opisuje wyniki ankiet przeprowadzonych wśród reprezentatywnej próby n respondentów.

Estymatorem prawdodobieństwa π uzyskanym metodą momentów jest

$$\hat{\pi}_K = \frac{\bar{Z} - (1-q)}{2q-1},$$

gdzie $\bar{Z} = \left(\sum_{i=1}^{n} Z_i\right)/n$.

Model Trójkątny. W modelu trójkątnym niezrandomizowanych odpowiedzi zadawane są dwa pytania, dokładnie tak samo jak w modelu krzyżowym. Różnica polega na tym, że respondent wpisuje do ankiety *NIE*, jeżeli na oba pytania odpowiedziałby przecząco. W przeciwnym razie wpisuje *TAK*. W szczególności:

$$Z = \begin{cases} 0, & \text{jeśli obie odpowiedzi są } NIE, \\ 1, & \text{w przeciwnym przypadku.} \end{cases} \tag{3}$$

W tym modelu

$$P_\pi(Z=1) = \varrho = \pi + (1-\pi)q, \tag{4}$$

a estymatorem prawdodobieństwa π uzyskanym metodą momentów jest

$$\hat{\pi}_T = \frac{\bar{Z} - q}{1-q}.$$

PRZEDZIAŁY UFNOŚCI

Rozważamy przybliżone przedziały przedziały ufności typu Walda. Są to asymptotyczne przedziały oparte na Centralnym Twierdzeniu Granicznym, które głosi, że asymptotyczny rozkład prawdopodobieństwa estymatorów $\hat{\pi}_K$ oraz $\hat{\pi}_T$ jest rozkładem normalnym. Dokładniej, gdy liczność próby rośnie nieograniczenie, to

$$\frac{\hat{\pi}_K - \pi}{\widehat{Var\pi_K}} \to N(0,1).$$

Tutaj $\widehat{Var\pi_K}$ jest nieobciążonym estymatorem wariancji estymatora $\hat{\pi}_K$:

$$\widehat{Var\pi_K} = \frac{\bar{Z}(1-\bar{Z})}{(n-1)(2q-1)^2} = \frac{\hat{\pi}_K(1-\hat{\pi}_K)}{n-1} + \frac{q(1-q)}{(n-1)(2q-1)^2}.$$

Przyjmując poziom ufności δ i stosując standardowe metody otrzymujemy przedział ufności dla prawdopodobieństwa drażliwego

$$\left(\hat{\pi}_K - u_{(1+\delta)/2}\sqrt{\widehat{Var\pi_K}},\ \hat{\pi}_K + u_{(1+\delta)/2}\sqrt{\widehat{Var\pi_K}}\right),$$

gdzie $u_{(1+\delta)/2}$ oznacza kwantyl rozkładu $N(0,1)$.

Podobnie mamy dla modelu trójkątnego:

$$\frac{\hat{\pi}_T - \pi}{\widehat{Var\pi_T}} \to N(0,1).$$

W tym modelu

$$\widehat{Var\pi_T} = \frac{\bar{Z}(1-\bar{Z})}{(n-1)(1-q)^2} = \frac{\hat{\pi}_T(1-\hat{\pi}_T)}{(n-1)} + \frac{q(1-q)}{(n-1)(1-q)^2}$$

i asymptotyczny przedział ufności dla prawdopodobieństwa π ma postać

$$\left(\hat{\pi}_T - u_{(1+\delta)/2}\sqrt{\widehat{Var\pi_T}},\ \hat{\pi}_T + u_{(1+\delta)/2}\sqrt{\widehat{Var\pi_T}}\right).$$

OPTYMALNY ROZMIAR PRÓBY

Rozważamy względną długość przybliżonego przedziału ufności

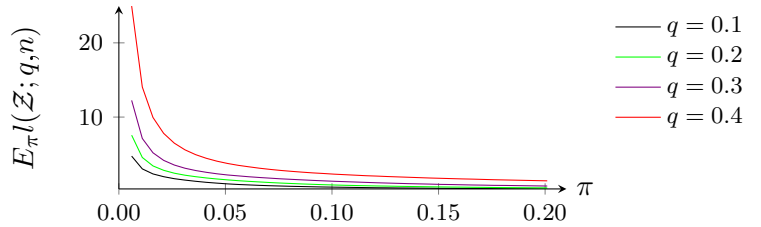
$$l(z;q,n) = \frac{\pi_G(z) - \pi_D(z)}{\pi}, \tag{5}$$

gdzie $(\pi_G(z), \pi_D(z))$ oznacza przedział ufności dla π po zaobserwowaniu wyniku z, gdzie z jest realizacją zmiennej losowej $\mathcal{Z} = \sum_i Z_i$. Naszym celem jest kontrola tej długości za pomocą rozmiaru próby. W tym celu dla danego modelu wyznaczamy

$$\begin{aligned} E_\pi \{l(\mathcal{Z};q,n)\mathbb{1}(\pi \in (\pi_G(\mathcal{Z}), \pi_D(\mathcal{Z}))\} = \\ \sum_{k=0}^{n} l(k;q,n)\mathbb{1}(\pi \in (\pi_G(k), \pi_D(k))P_\pi\{\mathcal{Z} = k\}. \end{aligned} \tag{6}$$

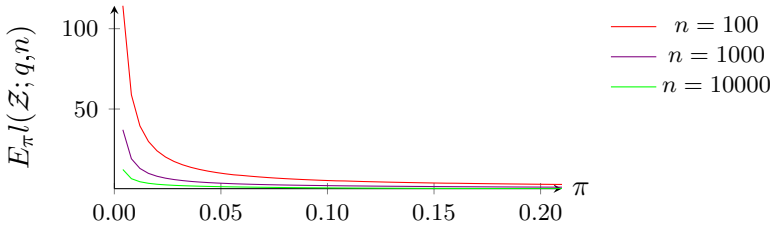
Jest to średnia względna długość przedziałów pokrywających π. Bez straty na ogólności zakładamy, że $\pi < 0.5$. Na podstawie obliczeń numerycznych można zaobsewować, że dla każdego $\pi \in (0, 0.5)$ wartość oczekiwana $E_\pi l(\mathcal{Z};q,n)$ rośnie względem q i maleje względem n (rys. 1, 2).

Rysunek 1. Średnia długość względna dla $n = 1000$ w modelu krzyżowym



Źródło: obliczenia własne

Rysunek 2. Średnia długość względna dla $n = 1000$ w modelu trójkątnym



Źródło: obliczenia własne

Przedstawione w pracy wykresy (rys. 1, 2) zostały wykonane dla modelu krzyżowego. Wykresy dla modelu trójkątnego są podobne. Obliczenia numeryczne wykonano dla

$$(n, q) \in \{100, 1000, 10000\} \times \{0.1, 0.2, 0.3, 0.4\}.$$

Stąd, przy ustalonej wielkości próby n jesteśmy zainteresowani, aby wartość parametru q była możliwie najmniejsza. Dla małych q uzyskujemy wysoką precyzję oszacowania. Z drugiej strony q nie może być zbyt małe. Na przykład $q = 0$ odpowiada przypadkowi, w którym nie zadajemy pytania neutralnego. Brak pytania neutralnego oznacza brak ochrony prywatności respondenta. Dlatego q powinno mieć wartość większą od zera. Dolne ograniczenie dla q powinno być wynikiem odpowiednio wysokiego poziomu ochrony prywatności respondenta.

Tan i in. [2009] zaproponowali, aby poziom prywatności respondenta kontrolować za pomocą następujących prawdopodobieństw warunkowych:

$$P_\pi\{Y=1|Z=1\} \text{ oraz } P_\pi\{Y=1|Z=0\},$$

które określają przypuszczalną przynależność respondenta do grupy wrażliwej, w zależności od wyniku jego ankiety. W modelu krzyżowym prawdopodobieństwa te przedstawiają się następująco (zastosowanie twierdzenia Bayesa):

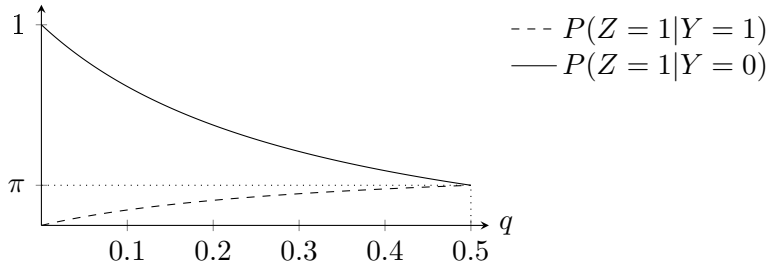
$$\begin{aligned} P_\pi\{Y=1|Z=1\} &= \frac{\pi q}{\pi q + (1-\pi)(1-q)}, \\ P_\pi\{Y=1|Z=0\} &= \frac{\pi(1-q)}{\pi(1-q)+(1-\pi)q}, \end{aligned} \tag{7}$$

a w modelu trójkątnym

$$\begin{aligned} P_\pi\{Y=1|Z=1\} &= \frac{\pi}{\pi+(1-\pi)q}, \\ P_\pi\{Y=1|Z=0\} &= 0. \end{aligned} \tag{8}$$

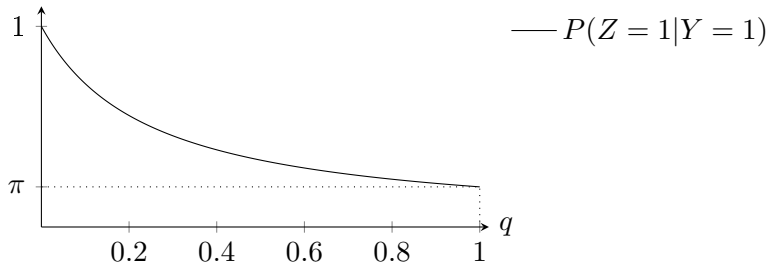
Zależność tych prawdopodobieństw od parametru q przedstawiona jest na rys. 3 dla modelu krzyżowego, a dla trójkątnego na rys. 4.

Rysunek 3. Ochrona prywatności w modelu krzyżowym



Źródło: obliczenia własne

Rysunek 4. Ochrona prywatności w modelu trójkątnym



Źródło: obliczenia własne

Z wykresów (rys. 3) oraz (rys. 4) wynika, że im mniejsza jest wartość parametru q, tym słabszy jest poziom ochrony prywatności (tym wyższe jest prawdopodobieństwo $P_\pi\{Y = 1|Z = 1\}$). Aby zachować odpowiedni poziom prywatności, musi być spełnione:

$$P_\pi\{Y = 1|Z = 1\} \leq \gamma \text{ oraz } P_\pi\{Y = 1|Z = 0\} \leq \gamma \tag{9}$$

dla z góry ustalonego $\gamma \in (0, 1)$ (odpowiednio dużego) oraz dowolnego $\pi \leq \pi_0$. Wartość π_0 odzwierciedla naszą wiedzę o odsetku osób należących do grupy wrażliwej. Na przykład z badań nad HIV w USA [NZIP 2022] wynika, że odsetek nowych zachorowań na tę chorobę maleje. Gdyby π oznaczał odsetek obecnie chorych na HIV, można by założyć, że $\pi_0 = 0.0036$, ponieważ 0.36% jest znanym poprzednio odsetkiem zarażonych tym wirusem.

Najmniejsze $q = q(\pi_0; \gamma)$, które spełnia nierówności (9), ma postać:

a) $q(\pi_0; \gamma) = \frac{\pi_0(1-\gamma)}{\gamma(1-2\pi_0)+\pi_0}$ w modelu krzyżowym,

b) $q(\pi_0; \gamma) = \frac{\pi_0(1-\gamma)}{\gamma(1-\pi_0)}$ w modelu trójkątnym.

Rozwiązanie nierówności (9) jest możliwe tylko wtedy, gdy $\gamma \geq \pi$. Zatem powinniśmy przyjąć, że $\gamma \geq \pi_0$. W przeciwnym przypadku nierówności (9) nie byłyby spełnione dla wszystkich $\pi \leq \pi_0$.

Formuły na $q(\pi_0; \gamma)$ zostały wyprowadzone w pracach Jaworski i Zieliński [2023], Jaworski [2025].

Zauważmy, że średnia długość względna rośnie nieograniczenie, gdy π zbliża się do zera. Stąd bez wiedzy o dolnym ograniczeniu parametru π nie jest możliwa kontrola długości przedziału ufności. Dlatego przyjmujemy, że $\pi \in \langle \pi_{-1}, \pi_0 \rangle$, gdzie dodatnie wartości π_{-1} oraz π_0 można określić na podstawie posiadanej wiedzy.

Uwzględniając powyższe uwagi, możemy na potrzeby tej pracy sformułować następującą definicję optymalnego rozmiaru próby.

Definicja. Zakładamy, że $\pi \in \langle \pi_{-1}, \pi_0 \rangle$, gdzie π_{-1}, π_0 są znanymi wartościami mniejszymi od 0.5. Przy ustalonym $\gamma \leq \pi_0$ oraz ustalonym d, najmniejszą liczebność próby n dla której

$$E_\pi l(Z; q(\pi_0, \gamma), n) \leq d \quad (\forall \pi \in \langle \pi_{-1}, \pi_0 \rangle)$$

nazywamy ***optymalnym rozmiarem próby***.

Przykład numeryczny. Przyjmujemy $\gamma = 0.5,\ d = 0.5$. Niech n^* oznacza optymalny rozmiar próby. W zależności od wartości π_{-1}, π_0 wyznaczony numerycznie optymalny rozmiar próby przedstawiono w tabeli 1.

Tabela 1. Optymalny rozmiar próby

Model	π_0	$\pi_{-1} = 0,1\pi_0$	n^*	$\pi_{-1} = 0,2\pi_0$	n^*
Krzyżowy	0,1	0,01	83447	0,02	22214
	0,2	0,02	64327	0,04	16726
	0,3	0,03	82676	0,06	21082
	0,4	0,04	209311	0,08	52631
Trójkątny	0,1	0,01	74083	0,02	19690
	0,2	0,02	48016	0,04	12397
	0,3	0,03	46620	0,06	12421
	0,4	0,04	67870	0,08	16593

Źródło: obliczenia własne

PODSUMOWANIE

W pracy zaprezentowano optymalny dobór wielkości próby z uwzględnieniem ochrony prywatności respondenta. Dobór próby oparto na kontroli względnej długości przydziałów ufności pokrywających szacowany prametr. Zaprezentowane podejście jest podobne do tego z prac Jaworski i Zieliński [2023] oraz Jaworski [2025], w których rozważano bezwględną długość przedziałów ufności, a nie względną. W ujęciu względnym ujawniła się konieczność kontrolowania

dolnego ograniczenia na prawdopodobieństwo π poprzez wykorzystnie dodatkowej wiedzy o odsetku osób należących do grupy wrażliwej. W przeciwnym przypadku oczekiwana precyzja, ze względu na niepraktycznie duży rozmiar próby optymalnej, mógłaby być trudna do uzyskania.

BIBLIOGRAFIA

Greenberg B. G., Abu-Ela A. A., Simons W. R., Horvitz D. G. (1969) The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association, 64, 520-539.

Horvitz D. G., Simons W. R. (1967) The Unrelated Question Randomized Response Model. Proceedings of the Social Statistics Section. American Statistical Association, 65-67.

Jaworski S. (2025) The Optimal Sample Size in Triangular Model for Sensitive Questions. Statistics in Transition, 26 (1), 221-231.

Jaworski S., Zieliński W. (2023) The Optimal Sample Size in the Crosswise Model for Sensitive Questions. Applicationes Mathematicae, 50 (1), 21-34.

NZIP (2022) Zakażenia HIV i zachorowania na AIDS w Polsce. Narodowy Instytut Zdrowia Publicznego. http://wwwold.pzh.gov.pl/oldpage/-epimeld/hiv_aids/index.htm (pobrane: 27.05.2022)

Tan M., Tian G. L., Tang M. L. (2009) Sample Surveys with Sensitive Questions: a Non-randomized Response Approach. The American Statistician, 63, 9–16.

Tian G. L., Yu J. W., Tang M. L., Geng Z. (2007). A New Nonrandomized Model for Analyzing Sensitive Questions with Binary Outcomes. Statistics in Medicine, 26, 4238–4252.

Warner S. L. (1965) Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, 63-69.

Yu J.-W., Tian G.-L., Tang M.-L. (2008) Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis. Metrika, 67, 251-263.

OPTIMAL SAMPLE SELECTION IN THE CROSSWISE AND TRIANGULAR MODELS OF SENSITIVE QUESTIONS

Abstract: In this work, we focus on the optimal selection of a sample in selected models: the crosswise and the triangular model. The optimality criterion consists in choosing the sample size for which the relative length of the confidence interval does not exceed a predetermined value.

Keywords: sensitive questions, nonrandomized response models
JEL classification: C83, C99