Metody Ilościowe w Badaniach Ekonomicznych Received: 02.08.2021
Tom XXII/2, 2021, pp. 51-60 Accepted: 23.08.2021

PRZEDZIAŁ UFNOŚCI DLA ODSETKA PYTAŃ DRAŻLIWYCH

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

Streszczenie: Jednym z problemów w badaniach ekonomiczno-społecznych jest oszacowanie odsetka odpowiedzi na pytania drażliwe. Pytania drażliwe są takimi pytaniami odpowiedzi na które respondent może nie udzielić rzetelnej odpowiedzi. W pracy podano konstrukcję dokładnego przedziału ufności dla tego odsetka i porównano zaproponowany przedział z przedziałami asymptotycznymi.

Słowa kluczowe: pytania drażliwe, model nierandomizowanych odpowiedzi, dokładny przedział ufności

JEL classification: C83, C99

WSTĘP

W badaniach ekonomiczno-społecznych zainteresowani jesteśmy określeniem skali zjawisk, które nie muszą być zgodne z prawem, być wstydliwymi czy też społecznie nieakceptowalnymi. Na przykład takim pytaniem może być pytanie o dawanie łapówek, o oszustwo podatkowe, o nie heteronormatywność czy też o stosowanie przemocy wobec dzieci.

Formalnie, niech Y będzie zmienną losową o rozkładzie dwupunktowym:

$$P\{Y = 1\} = \pi = 1 - P\{Y = 0\}.$$

Liczba $\pi \in (0, 1)$ jest prawdopodobieństwem odpowiedzi *TAK* na pytanie drażliwe. Estymujemy prawdopodobieństwo π.

Niestety, zmienne losowe $Y_1, \ldots, Y_n$ z wiadomych powodów nie są obserwowalne. Dlatego też odpowiedź na pytanie drażliwe jest „ukrywana". Odpowiedzi na pytanie drażliwe są „ukrywane" przez zadawanie pytania „neutralnego", czyli takiego pytania, na które respondent zawsze odpowie zgodnie

https://doi.org/10.22630/MIBE.2021.22.2.5

z prawdą. Pytanie neutralne ma dwie odpowiedzi *TAK* lub *NIE*. W ankiecie stawiane są jednocześnie dwa pytania: drażliwe i neutralne, ale ankietowany odpowiada tylko na jedno z nich. Wynikiem tak skonstruowanej ankiety jest jedna odpowiedź *TAK* lub *NIE*, przy czym ankieter ma nie wiedzieć na które pytanie odpowiada ankietowany. Odpowiedź Z na ankietę jest zmienną losową o rozkładzie dwupunktowym

$$P\{Z = 1\} = \varrho = 1 - P\{Z = 0\}.$$

Prawdopodobieństwo ϱ jest uzależnione od π oraz **znanego** q, gdzie q jest prawdopodobieństwem odpowiedzi *TAK* na pytanie neutralne. Forma tej zależności jest uzależniona od sposobu „ukrywania" pytania drażliwego.

Jednym z pierwszych mechanizmów ukrywania jest model randomizowanych odpowiedzi (Warner 1965, Greenberg et.al. 1969, Boruch 1971, Fox & Tracy 1986). W zależności od pewnego mechanizmu losowego (rzut kostka, losowanie kart) ankietowany odpowiada bądź na pytanie drażliwe bądź pytanie neutralne. Prawdopodobieństwo r wyboru pytania jest znane. Wynik losowania pytania nie jest znany ankieterowi.

Innym modelem jest model nierandomizowanych odpowiedzi (Yu et.al. 2008, Tian 2014, Groenitz 2014). Ankietowany odpowiada na pytanie drażliwe oraz pytanie neutralne. Odpowiedzią na całą ankietę jest *TAK* gdy odpowiedzi na oba pytania są jednakowe.

PRZEDZIAŁ UFNOŚCI

W modelu nierandomizowanych odpowiedzi zadawane są dwa pytania: drażliwe i neutralne. Na każde z tych pytań można odpowiedzieć *TAK* lub *NIE*. Odpowiedź na całą ankietę jest *TAK*, jeżeli na oba zadane pytania jest taka sama odpowiedź. To znaczy, jeżeli przez Z oznaczymy zmienną losową będącą wynikiem ankiety, to

$$Z = \begin{cases} 1, & \text{jeżeli } \{Y = 1 \& Q = 1\} \text{ lub } \{Y = 0 \& Q = 0\} \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

Tutaj Y oznacza odpowiedź na pytanie drażliwe, zaś Q na pytanie neutralne.

Zmienna losowa Y jest zmienną losową o rozkładzie dwupunktowym

$$P\{Y = 1\} = \pi = 1 - P\{Y = 0\}.$$

Prawdopodobieństwo π jest nieznanym prawdopodobieństwem uzyskania pozytywnej odpowiedzi na pytanie drażliwe. Zadanie polega na estymacji tego prawdopodobieństwa. A dokładniej, chcemy zbudować przedział ufności dla tego prawdopodobieństwa.

O zmiennej losowej Q zakładamy, że ma rozkład dwupunktowy

$$P\{Q=1\}=q=1-P\{Q=0\}.$$

Ponadto zakładamy, że prawdopodobieństwo q jest znane oraz, że zmienne losowe Y oraz Q są niezależne.

Z konstrukcji ankiety wynika, że zmienna losowa Z (odpowiedź na ankietę) również jest zmienną losową o rozkładzie dwupunktowym

$$P\{Z=1\}=\varrho=1-P\{Z=0\},$$

gdzie $\varrho=q\pi+(1-q)(1-\pi)=\pi(2q-1)+(1-q)$.

Niech $Z_1,\ldots,Z_n$ będą wynikami n ankiet oraz niech $Z=\sum_{i=1}^n Z_i$. Zmienna losowa ma rozkład dwumianowy z parametrami n oraz ϱ. Estymatorem Największej Wiarogodności (ENW) prawdopodobieństwa sukcesu ϱ jest $\bar{Z}=\frac{1}{n}Z$. Łatwo zauważyć, że ENW interesującego nas prawdopodobieństwa π jest

$$\hat{\pi}=\frac{\bar{Z}-1+q}{2q-1}.$$

Estymator $\hat{\pi}$ jest nieobciążonym estymatorem prawdopodobieństwa π o wariancji

$$Var\hat{\pi}=\frac{\pi(1-\pi)}{n}+\frac{q(1-q)}{n(2q-1)^2}.$$

Nieobciążonym estymatorem wariancji estymatora $\hat{\pi}$ jest

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

Tian (2014) zaproponował asymptotyczne przedziały ufności dla prawdopodobieństwa π. Ich konstrukcja oparta jest na znanym twierdzeniu de Moivra-Laplace'a. Zgodnie z tym twierdzeniem

$$\hat{\varrho}\sim AN\left(\varrho,\frac{\varrho(1-\varrho)}{n}\right).$$

Zastosowanie tego twierdzenia do estymatora $\hat{\pi}$ daje

$$\hat{\pi}\sim AN\left(\pi,\frac{\pi(1-\pi)}{n}+\frac{(1-q)q}{n(2q-1)^2}\right).$$

Na podstawie tego asymptotycznego przybliżenia budowany jest przedział ufności dla π, przy czym robione jest to na dwa sposoby.

W pierwszym sposobie zakłada się, że

$$\frac{\hat{\pi}-\pi}{\sqrt{\widehat{Var\hat{\pi}}}} \sim N(0,1).$$

A zatem (przyjmujemy poziom ufności δ), przedział #1 uzyskujemy jako rozwiązanie względem π nierówności

$$\left|\frac{\hat{\pi}-\pi}{\sqrt{\widehat{Var\hat{\pi}}}}\right| \leq u_{\frac{1+\delta}{2}},$$

gdzie u_q oznacza kwantyl rzędu q standardowego rozkładu normalnego $N(0,1)$. Przedział ma postać

$$\left(\hat{\pi}-u_{\frac{1+\delta}{2}}\sqrt{\widehat{Var\hat{\pi}}};\ \hat{\pi}+u_{\frac{1+\delta}{2}}\sqrt{\widehat{Var\hat{\pi}}}\right)$$

Drugi sposób polega na tym by w powyższym przybliżeniu nie stosować estymatora wariancji estymatora $\hat{\pi}$, tylko po prostu jego wariancję

$$\frac{\hat{\pi}-\pi}{\sqrt{Var\hat{\pi}}} \sim N(0,1)$$

A zatem (przyjmujemy poziom ufności δ), przedział #2 uzyskujemy jako rozwiązanie względem π nierówności

$$\left|\frac{\hat{\pi}-\pi}{\sqrt{Var\hat{\pi}}}\right| \leq u_{\frac{1+\delta}{2}},$$

a dokładniej z nierówności

$$\left(\frac{\hat{\pi}-\pi}{\sqrt{Var\hat{\pi}}}\right)^2 \leq u^2_{\frac{1+\delta}{2}},$$

Rozwiązując powyższą nierówność względem π otrzymujemy przedział #2

$$\frac{2n\hat{\pi}+u^2_{\frac{1+\delta}{2}} \pm u_{\frac{1+\delta}{2}}\sqrt{4n\hat{\pi}(1-\hat{\pi})+\frac{4n(1-q)q+u^2_{\frac{1+\delta}{2}}}{(1-2q)^2}}}{2\left(n+u^2_{\frac{1+\delta}{2}}\right)}$$

Powyższe przedziały są budowane dla „dużych” prób. Z tego powodu przedziały te są przybliżone. Ich zalety i wady (w ogólnym ujęciu) są szczegółowo omówione w monografii Zielińskiego (2010).

Clopper i Pearson (1934) podali konstrukcję dokładnego przedziału ufności dla prawdopodobieństwa sukcesu

$$\varrho \in (\varrho_L(Z), \varrho_U(Z)),$$

gdzie

$$\varrho_L(Z) = \begin{cases} 0 & \text{dla } Z = 0, \\ B^{-1}\left(n - Z + 1, Z; \frac{1+\delta}{2}\right) & \text{dla } Z > 0, \end{cases}$$

$$\varrho_U(Z) = \begin{cases} 1 & \text{dla } Z = n, \\ B^{-1}\left(n - Z, Z + 1; \frac{1-\delta}{2}\right) & \text{dla } Z < n. \end{cases}$$

oraz $B^{-1}(\cdot, \cdot; \cdot)$ jest kwantylem rozkładu Beta. Ponieważ $\varrho = q\pi + (1-q)(1-\pi)$, więc łatwo uzyskujemy przedział ufności #3 dla prawdopodobieństwa π:

$$\begin{cases} \frac{1}{2q-1}\left((\varrho_L(Z), \varrho_U(Z)) - (1-q)\right), & \text{gdy } q > 0.5, \\ \frac{1}{2q-1}\left((\varrho_U(Z), \varrho_L(Z)) - (1-q)\right), & \text{gdy } q < 0.5. \end{cases}$$

Należy zauważyć, że przedział ten nie jest asymptotyczny.

Ze względu na swoją konstrukcję wszystkie powyższe przedziały mogą „wykraczać” poza naturalny dla prawdopodobieństwa przedział $(0, 1)$. W związku z tym uzyskany przedział ufności jest obcinany w zerze lub w jedynce.

PORÓWNANIE PRZEDZIAŁÓW UFNOŚCI

Przedział ufności powinien ze z góry zadanym prawdopodobieństwem pokrywać szacowany parametr. Dlatego też zbadamy jakie jest prawdopodobieństwo pokrycia odsetka pytań drażliwych przez przedziały ufności zaprezentowane w poprzednim rozdziale. Te prawdopodobieństwa oszacowane będą na podstawie badań symulacyjnych.

W symulacjach zastosowano próbę liczności $n = 1000$. Generowano dwie zmienne losowo o rozkładach dwupunktowych z prawdopodobieństwami odpowiednio π (jako odpowiedź Y na pytanie drażliwe) oraz q (jako odpowiedź Q na pytanie neutralne). Wynikiem Z ankiety jest

$$Z = Y \cdot Q + (1 - Y) \cdot (1 - Q).$$

Otrzymano w ten sposób odpowiedzi na całą ankietę $Z_1, \ldots, Z_n$. Na podstawie takiej próby budowano przedziały ufności #1, #2 oraz #3 według podanych wyżej wzorów. Procedurę symulacyjną powtórzono 10000 razy. Na poniższych wykresach przedziały te oznaczone są odpowiednio AP, WP, CP. Przyjęto poziom ufności 0.95.

Na rysunku 1. pokazane jest prawdopodobieństwo pokrycia nieznanego prawdopodobieństwa π uzyskania pozytywnej odpowiedzi na pytanie drażliwe (na osi poziomej podane jest prawdopodobieństwo π). Na wykresie (a) podane jest przeciętne ze względu na prawdopodobieństwo q pozytywnej odpowiedzi na pytanie neutralne, natomiast na wykresach (b) i (c) dla $q = 0.1$ oraz $q = 0.2$, odpowiednio. Jak widać, prawdopodobieństwa pokrycia przez przedziały AP oraz WP są mniejsze od przyjętego poziomu ufności. Zatem przedziały te nie są przedziałami ufności w myśl definicji Neymana (1934). Zastosowanie tych przedziałów do wnioskowania o prawdopodobieństwie π obarczone jest większym ryzykiem błędnego wnioskowania niż zakładane pięcioprocentowe ryzyko. Co więcej to ryzyko jest nieznane. Tej wady nie ma przedział CP.

Na rysunku 2. porównana jest długość przedziałów ufności. Na rysunku (a) podana jest długość przedziałów przeciętnie ze względu na prawdopodobieństwo q pozytywnej odpowiedzi na pytanie neutralne, natomiast na rysunkach (b) i (c) dla $q = 0.1$ oraz $q = 0.2$, odpowiednio. Przedziały AP oraz WP mają podobną długość, natomiast przedział CP jest dłuższy niż dwa pozostałe. To oczywiście jest konsekwencją tego, że prawdopodobieństwa pokrycia przez przedziały AP i WP są mniejsze niż nominalne (Rys. 1). Możemy wyprowadzić wniosek, iż te dwa przedziały nieprawidłowo szacują prawdopodobieństwo π odpowiedzi na pytanie drażliwe.

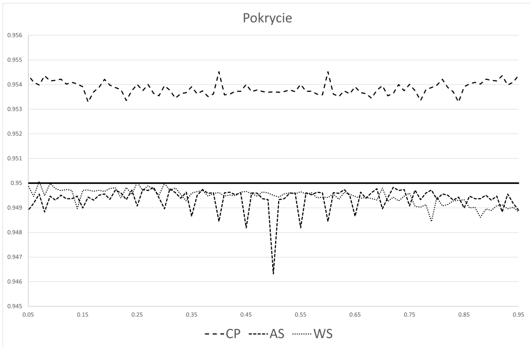
UWAGI KOŃCOWE

W pracy podjęto problem przedziałowej estymacji odsetka pozytywnych odpowiedzi na pytanie drażliwe. Ze względu na to, że jest wiele różnych modeli stosowanych do punktowego oszacowania tego odsetka w pracy zajęto się się tylko jednym z nich, a mianowicie modelem nierandomizowanych odpowiedzi zaproponowanym przez Yu i in. (2014).

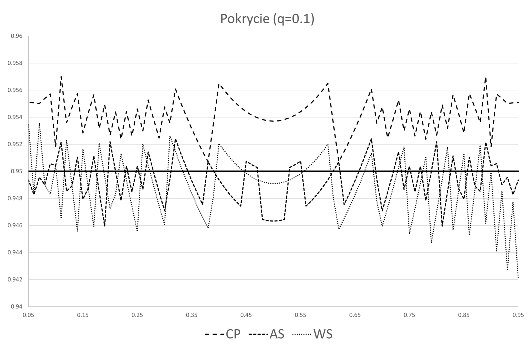
Porównane zostały trzy metody konstrukcji przedziału ufności. Dwie z nich oparte są na asymptotycznym twierdzeniu de Moivre-Laplace'a. Trzecia metoda bazuje na dokładanym przedziale ufności opartym na pomyśle Cloppera i Pearsona. Okazuje się, że pierwsze dwie metody dają przedziały nie realizujące postulatu zachowania zadanego poziomu ufności, natomiast trzecia z nich zapewnia spełnianie tego warunek. Otrzymane przedziały są co prawda szersze od asymptotycznych, ale te ostatnie mają zbyt duże (i nieznane) ryzyko błędnego wnioskowania.

Można oczekiwać, że w innych modelach stosowanych w praktyce uzyskiwania odpowiedzi na pytania drażliwe sytuacja będzie podobna. Badania nad konstrukcją przedziałów ufności dla odsetka pozytywnych odpowiedzi na pytanie drażliwe w różnych modelach są kontynuowane i będą opublikowane oddzielnie.

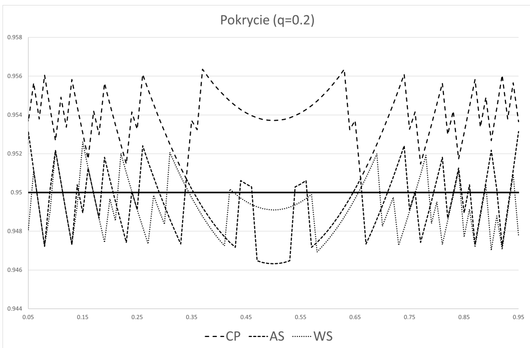
Rysunek 1. Prawdopodobieństwo pokrycia



(a)



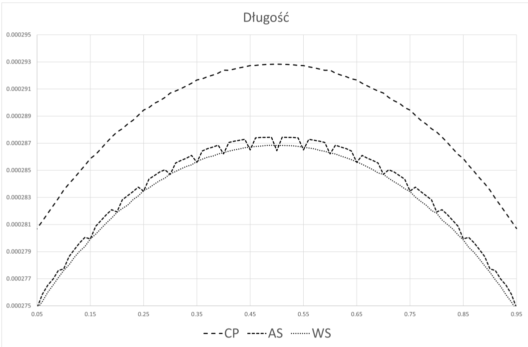
(b)



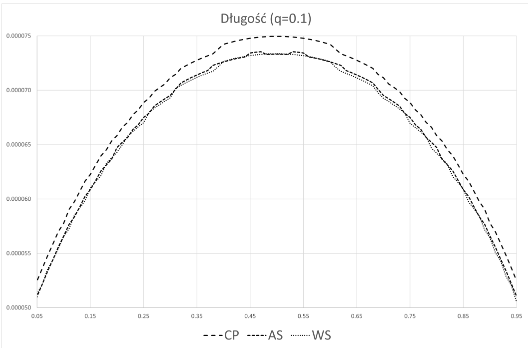
(c)

Źródło: obliczenia własne

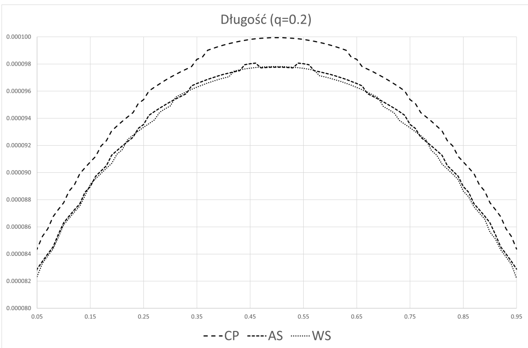
Rysunek 2. Długość przedziału ufności



(a)



(b)



(c)

Źródło: obliczenia własne

BIBLIOGRAFIA

Arnab R., Shangodoyin D. K., Arcos A. (2019) Nonrandomized Response Model for Complex Survey Designs. Statistics in Transition New Series, 20, 67-86.

Boruch R. F. (1971) Assuring Confidentiality of Responses in Social Research: A Note on Strategies. American Sociologist, 6, 308–311.

Clopper C. J., Pearson E. S. (1934) The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika, 26, 404-413.

Fox J. A., Tracy P. E. (1986) Randomised Response: A Method for Sensitive Surveys. Beverly Hills: Sage Publications.

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

Groenitz H. (2014) A New Privacy-protecting Survey Design for Multichotomous Sensitive Variables. Metrika, 77, 211-224.

Neyman J. (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97, 558-625.

Tian G.-L. (2014) A New Non-randomized Response Model: The Parallel Model. Statistica Neerlandica, 68, 293–323.

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.

Zieliński W. (2010) Estymacja wskaźnika struktury. Wydawnictwo SGGW.

CONFIDENCE INTERVAL FOR THE FRACTION OF SENSITIVE QUESTIONS

Abstract: In sociological research it is important to estimate a fraction of responses to sensitive questions. Sensitive questions are such that the respondent may not give a true answer. In the paper a confidence interval for a proportion of positive answers to a senstitive question is constructed.

Keywords: sensitive questions, nonrandomized response model, exact confidence interval

JEL classification: C83, C99