DOI: 10.5604/01.3001.0014.1223

PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 4 – 2016

JOANNA KISIELIŃSKA[1]

ROZKŁADY WYBRANYCH BOOTSTRAPOWYCH ESTYMATORÓW MEDIANY ORAZ ZASTOSOWANIE DOKŁADNEJ METODY PERCENTYLI DO JEJ PRZEDZIAŁOWEGO SZACOWANIA

1. WPROWADZENIE

Do szacowania parametrów rozkładu zmiennej losowej w przypadku, gdy rozkład ten nie jest znany, można zastosować metodę bootstrapową. Metoda, zaproponowana przez Efrona (1979), polega na wtórnym próbkowaniu próby losowej (nazywanej dalej pierwotną) i przybliżaniu rozkładu estymatora parametru rozkładem określonym dla prób wtórnych (bootstrapowym rozkładem estymatora). Liczba możliwych do wylosowania (ze zwracaniem) ze skończonej próby pierwotnej, prób wtórnych jest skończona i równa liczbie wariacji z powtórzeniami. Jeśli próba pierwotna liczy n elementów, prób wtórnych jest n^n. Wiele z nich różni się jedynie kolejnością elementów – prób wtórnych zawierających różne elementy jest $\binom{2n-1}{n}$ (Fisher, Hall, 1991).

Klasyczną metodę bootstrapową można w uproszczeniu interpretować jako losowanie ze zwracaniem N-elementowej próby ze zbioru wszystkich prób wtórnych, liczącego $B = n^n$ elementów. W przypadku małych prób pierwotnych może się wiec zdarzyć, że liczba losowanych prób wtórnych N będzie większa od B i liczbę losowanych prób wtórnych trzeba ograniczyć. Ze względu na skończoną liczbę prób wtórnych możliwe jest ponadto wielokrotne wylosowanie tej samej próby, z kolei inna może nie być wylosowana ani razu. Określając rozkład estymatora bootstrapowego warto więc wykorzystywać wszystkie możliwe próby wtórne. Metodę bazującą na wszystkich próbach wtórnych nazywa się dokładną metodą bootsrapową (*exact bootstrap method*). Metoda przedstawiona została między innymi przez Fishera, Halla (1991), Kisielińską (2013). Na potrzebę stosowania metody dokładnej zwracają uwagę Hutson, Ernst (2000). Zauważają bowiem, że metoda ta pozwala wyeliminować błędy wynikające z losowania prób wtórnych.

Należy zwrócić uwagę, że dokładna metoda bootstrapowa nazywa się dokładną jedynie dlatego, że wykorzystuje wszystkie możliwe próby wtórne. Kwestia dokładności szacunku parametru badanej populacji względem jego wartości faktycznej dotyczy próby pierwotnej i zasad jej pobierania.

[1] Szkoła Główna Gospodarstwa Wiejskiego, Wydział Nauk Ekonomicznych, Katedra Ekonomiki Rolnictwa i Międzynarodowych Stosunków Gospodarczych, ul. Nowoursynowska 166, 02-787 Warszawa, Polska, e-mail: joanna_kisielinska@sggw.pl.

Parametrem najczęściej wykorzystywanym do charakterystyki poziomu zjawiska jest wartość oczekiwana. Jednak w przypadku, gdy rozkład cechuje asymetria, warto zastosować medianę, której wybrane metody punktowego i przedziałowego szacowania są przedmiotem niniejszego artykułu. Evans i inni (2006) przedstawili wzory pozwalające wyznaczyć rozkłady bootstrapowych estymatorów statystyk pozycyjnych o dowolnej randze. Mogą być one zastosowane dla określenia rozkładu estymatora w postaci mediany z próby o nieparzystej liczbie obserwacji (estymator standardowy). Nie można ich jednak użyć, do wyznaczenia rozkładu estymatora w postaci mediany z próby oraz estymatora standardowego dla próby o parzystej liczbie obserwacji. W pracy zaproponowane zostaną formuły określające rozkład estymatora standardowego (dla próby o parzystej liczbie elementów). Dla estymatora w postaci mediany z próby o parzystej liczbie elementów pokazane zostanie, że nie jest możliwe podanie wzorów ogólnych (obowiązujących dla każdej próby) określających ten rozkład. W pracy pokazane zostanie, jak mimo tego rozkład estymatora wyznaczyć.

W artykule zwrócono uwagę na istotne własności rozkładu standardowego estymatora mediany (zarówno dla próby o nieparzystej jak i parzystej liczbie elementów), które pozwalają stworzyć gotowe tablice oraz z góry określić, które elementy próby pierwotnej są granicami przedziałów ufności. W przypadku estymatora w postaci mediany z próby o parzystej liczbie elementów, nie jest to możliwe i dla każdej próby należy przedziały wyznaczać indywidualne.

Podkreślić należy, że przedziały ufności dla trzech wziętych pod uwagę estymatorów wyznaczane były metodą percentyli zastosowaną dla wszystkich możliwych prób wtórnych (wykorzystano dokładną metodę bootstrapu). Metodę taką można nazwać dokładną metodą percentyli przez analogię do dokładnej metody bootstrapowej. W pracy pokazane zostanie, że w przypadku standardowego estymatora mediany określanie granic przedziałów ufności dokładną metodą percentyli jest znacznie prostsze niż zastosowanie metody klasycznej (z losowaniem).

Ilustracją rozważań teoretycznych są badania symulacyjne przeprowadzone dla wybranych rozkładów. W przypadku próby o nieparzystej liczbie obserwacji wyznaczone zostały granice przedziałów ufności dokładną metodą percentyli oraz zbadany został wpływ liczebności próby oraz założonego poziomu ufności na dokładność oszacowania. Dla prób o parzystej liczbie obserwacji dodatkowo porównane zostaną wyniki estymacji przedziałowej uzyskane dwoma estymatorami – standardowym i w postaci mediany z próby.

2. WYBRANE ESTYMATORY MEDIANY

Rozważania prowadzone będą dla zmiennej losowej X określonej dla populacji o nieznanym rozkładzie F o medianie $M_{0,5}$. Z populacji pobrano prostą próbę losową $(X_1, X_2,\ldots, X_n)$, której niemalejąco uporządkowana realizacja oznaczona zostanie jako $(x_{(1)}, x_{(2)},\ldots, x_{(n)})$. Statystyka pozycyjna o randze k oznaczona będzie jako $X_{(k)}$.

Przyjmujemy, że medianą z próby o nieparzystej liczbie obserwacji jest statystyka:

$$Me = X_{\left(\frac{n+1}{2}\right)}, \tag{1}$$

zaś z próby o parzystej liczbie obserwacji statystyka:

$$Me = \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)} \right). \tag{2}$$

Mediana z próby jest medianowo-nieobciążonym estymatorem mediany dla n nieparzystego. Jak udowodnił Zieliński (1995), w przypadku parzystego n, mediana próby może znajdować się dowolnie daleko od szacowanej mediany rozkładu, ponieważ dla każdej stałej C istnieje rozkład F, dla którego odległość oszacowania od faktycznej wartości mediany będzie większa od C. Z twierdzenia tego nie wynika jednak, że ma to miejsce dla każdego rozkładu.

Z drugiej strony wiadomo (Domański, Pruska, 2000), że granicznym rozkładem mediany z próby wylosowanej z populacji o rozkładzie z gęstością f taką, że $f(M_{0,5}) > 0$ jest rozkład normalny:

$$Me \sim \mathrm{N}\left(M_{0,5}, \frac{1}{2\sqrt{n} f\left(M_{0,5}\right)} \right), \tag{3}$$

gdzie f jest funkcją gęstości badanej zmiennej losowej w populacji. Oznacza to, że mediana z próby jest asymptotycznie nieobciążonym estymatorem mediany populacji.

Dla n parzystego medianowo-nieobciążonym estymatorem mediany jest natomiast estymator standardowy (Pekasiewicz, 2015) określony jako:

$$Me^{st} = X_{\left(\frac{n}{2} + I_{(0,5)}(u)\right)}, \tag{4}$$

gdzie u jest wartością zmiennej losowej o rozkładzie jednostajnym na przedziale [0,1], a $I_{(0,5)}(u)$ indykatorem przyjmującym wartość 1 lub 0, w zależności od tego, czy u jest większe bądź równe 0,5, czy mniejsze.

3. ROZKŁADY BOOTSTRAPOWYCH ESTYMATORÓW MEDIANY

W metodzie bootstrapowej, z prostej próby pierwotnej losowane są ze zwracaniem próby wtórne, które oznaczone zostaną jako $\left(X_1^*, X_2^*, ..., X_n^*\right)$. Uporządkowany zbiór możliwych realizacji każdej ze zmiennych X_i^* to $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$. Jeżeli w próbie pierwotnej nie występują powtórzenia, prawdopodobieństwo wylosowania każdej wartości $x_{(i)}$ jest jednakowe i równe $\frac{1}{n}$. Statystyka pozycyjna próby wtórnej o randze k oznaczona będzie jako $X^*_{(k)}$.

Dla każdej próby wtórnej można wyznaczyć medianę według reguły (1) dla n nieparzystego oraz reguł (2) lub (4) dla n parzystego. Bootstrapowymi realizacjami estymatora mediany, w przypadku wykorzystania formuł (1) i (4), mogą być jedynie elementy próby pierwotnej. Dla estymatora określonego formułą (2) mogą być to nie tylko elementy próby pierwotnej, ale również średnie arytmetyczne wszystkich dwuelementowych kombinacji z niej wylosowanych.

3.1. PRÓBA O NIEPARZYSTEJ LICZBIE OBSERWACJI

W przypadku próby o nieparzystej liczbie elementów bootstrapowym estymatorem mediany jest statystyka pozycyjna $X^*_{(m)}$ o randze $m = \frac{n+1}{2}$. Rozkład tego estymatora jest określony formułą (por. Efron, 1979; Pekasiewicz, 2015):

$$\begin{cases} \text{dla } l = 2,3,...,n-1: \\ P\left(X^*_{(m)} = x_{(l)}\right) = \sum_{j=0}^{m-1} \left(\frac{l-1}{n}\right)^j \left(1 - \frac{l-1}{n}\right)^{n-j} \binom{n}{j} - \sum_{j=0}^{m-1} \left(\frac{l}{n}\right)^j \left(1 - \frac{l}{n}\right)^{n-j} \binom{n}{j}, \\ \text{dla } l = 1, n: \\ P\left(X^*_{(m)} = x_{(l)}\right) = \sum_{j=m}^{n} \left(\frac{1}{n}\right)^j \left(\frac{n-1}{n}\right)^{n-j} \binom{n}{j}. \end{cases} \tag{5}$$

Rozkład ten można przedstawić również następująco (na podstawie Evans i inni, 2006):

$$\begin{cases} \text{dla } l = 2,3,\ldots,n-1: \\ P\left(X^*_{(m)} = x_{(l)}\right) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-m} \left(\frac{l-1}{n}\right)^i \left(\frac{1}{n}\right)^{n-i-j} \left(\frac{n-l}{n}\right)^j \binom{n}{i, j, n-i-j}, \\ \text{dla } l = 1, n: \\ P\left(X^*_{(m)} = x_{(l)}\right) = \sum_{j=m}^{n} \left(\frac{1}{n}\right)^j \left(\frac{n-1}{n}\right)^{n-j} \binom{n}{j}. \end{cases} \tag{6}$$

Rozkład określony formułą (5) lub (6) ma dwie istotne własności. Po pierwsze prawdopodobieństwa $P\left(X^*_{(m)} = x_{(l)}\right)$ nie zależą od wartości obserwacji wchodzących w skład próby, lecz jedynie od jej rozmiaru, a wobec tego są jednakowe dla wszystkich

n elementowych prób pierwotnych. Z własności tej wynika możliwość stworzenia gotowych tablic zawierających rozkłady estymatora dla różnych n[2]. Można ponadto pokazać, że dla $l = 1, 2,\ldots, n$:

$$P\left(X_{(m)}^{*} = x_{(l)}\right) = P\left(X_{(m)}^{*} = x_{(n-l+1)}\right). \tag{7}$$

3.2. PRÓBA O PARZYSTEJ LICZBIE OBSERWACJI

Rozważania prowadzone będą dla dwóch bootstrapowych estymatorów mediany. Pierwszy z nich określony regułą (2), oznaczony zostanie jako $X_{(m_1,m_1+1)}^{*a}$, drugi natomiast dany regułą (4) jako $X_{(m_1,m_1+1)}^{*b}$, gdzie $m_1 = \dfrac{n}{2}$.

Jak wspomniano wcześniej, w przypadku estymatora $X_{(m_1,m_1+1)}^{*a}$, medianą w próbach wtórnych mogą być wszystkie elementy próby oraz średnie arytmetyczne wszystkich kombinacji dwuelementowych z próby. Rozważmy dwie próby czteroelementowe: (1; 2; 3; 4) oraz (1; 2; 4; 7). W przypadku próby pierwszej medianą w próbach wtórnych mogą być wartości 1; 1,5; 2; 2,5; 3; 3,5; 4 natomiast dla próby drugiej 1; 1,5; 2; 2,5; 3, 4; 4,5; 5; 6; 8. Bootstrapowy estymator mediany $X_{(m_1,m_1+1)}^{*a}$ dla próby pierwszej ma 7 realizacji, zaś dla drugiej 10. Prawdopodobieństwa poszczególnych realizacji mediany będą więc różne dla różnych prób, ponieważ średnie arytmetyczne różnych kombinacji mogą być takie same, a ponadto mogą być równe elementom próby pierwotnej.

Następny problem wynika z kolejności możliwych wartości mediany w próbach wtórnych. Dla próby (1; 2; 2,4; 5) średnia arytmetyczna elementu pierwszego i trzeciego jest mniejsza od elementu drugiego, natomiast dla próby (1; 2; 4; 5) jest od niego większa.

Przedstawione problemy powodują, że nie jest możliwe podanie ogólnej formuły na prawdopodobieństwa dla poszczególnych realizacji estymatora. Można natomiast wyznaczyć prawdopodobieństwa, że na pozycjach m_1 i m_1+1 w próbach wtórnych występuje dowolny l-ty element próby, lub dwa dowolne jej elementy: l_1 i l_2, przy czym $l_1 < l_2$.

Prawdopodobieństwo, że w uporządkowanej próbie wtórnej l-ty element próby pierwotnej występuje na co najmniej dwóch środowych pozycjach m_1 i m_1+1 jest następujące:

[2] Na stronie http://joanna_kisielinska.users.sggw.pl podano tablice rozkładu bootstrapowego estymatora mediany dla liczebności nieparzystych z przedziału 3–201.

$$\begin{cases} \text{dla } l = 2,\dots,n-1 \\ P\left(\left(X^*_{(m_1)} = x_{(l)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l)}\right)\right) = \sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{l-1}{n}\right)^i \left(\frac{1}{n}\right)^{n-i-j}\left(\frac{n-l}{n}\right)^j \binom{n}{i, n-i-j, j}, \\ \text{dla } l = 1, n \\ P\left(\left(X^*_{(m_1)} = x_{(l)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l)}\right)\right) = \sum_{i=m_1+1}^{n}\left(\frac{1}{n}\right)^i \left(\frac{n-1}{n}\right)^{n-i}\binom{n}{i}. \end{cases} \tag{8}$$

Można pokazać, że

$$P\left(\left(X^*_{(m_1)} = x_{(l)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l)}\right)\right) = P\left(\left(X^*_{(m_1)} = x_{(n-l+1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(n-l+1)}\right)\right).$$

Prawdopodobieństwo, że w uporządkowanej próbie wtórnej pierwszy element uporządkowanej próby pierwotnej występuje dokładnie m_1 razy, a element l-ty występuje przynajmniej raz na pozycji m_1+1 jest równe:

$$P\left(\left(X^*_{(m_1)} = x_{(1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l)}\right)\right) = \sum_{i=0}^{m_1-1}\left(\frac{1}{n}\right)^{n-i}\left(\frac{n-l}{n}\right)^i \binom{n}{m_1, n-m_1-i, i}, \tag{9}$$

dla l = 2, 3, ..., n − 1.

Prawdopodobieństwo, że w uporządkowanej próbie wtórnej n-ty element uporządkowanej próby pierwotnej występuje dokładnie m_1 razy, a element l-ty występuje przynajmniej raz na pozycji m_1 jest równe:

$$P\left(\left(X^*_{(m_1)} = x_{(l)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(n)}\right)\right) = \sum_{i=0}^{m_1-1}\left(\frac{l-1}{n}\right)^i \left(\frac{1}{n}\right)^{n-i}\binom{n}{i, n-m_1-i, m_1}, \tag{10}$$

przy czym $P\left(\left(X^*_{(m_1)} = x_{(1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l)}\right)\right) = P\left(\left(X^*_{(m_1)} = x_{(n-l+1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(n)}\right)\right)$, dla l = 2, 3, …, n − 1.

Prawdopodobieństwo, że w próbie wtórnej pierwszy oraz n-ty element uporządkowanej próby pierwotnej występują dokładnie po m_1 razy jest równe:

$$P\left(X^*_{(m_1)} = x_{(1)}\right) = P\left(X^*_{(m_1+1)} = x_{(n)}\right) = \left(\frac{1}{n}\right)^n \binom{n}{m_1}. \tag{11}$$

Prawdopodobieństwo, że w uporządkowanej próbie wtórnej element l_1 znajduje się przynajmniej raz na pozycji m_1, a element l_2 przynajmniej raz na pozycji m_1+1 jest równe:

$$
\begin{aligned}
&P\left(\left(X^*_{(m_1)} = x_{(l_1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l_2)}\right)\right) = \\
&= \sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{l_1-1}{n}\right)^i \left(\frac{1}{n}\right)^{n-i-j}\left(\frac{n-l_2}{n}\right)^j \binom{n}{i, m_1-i, m_1-j, j}, \\
&\forall\, (l_1, l_2) : l_1 < l_2 \in \{2,3,\ldots,n-2\} \times \{3,4,\ldots,n-1\},
\end{aligned}
\tag{12}
$$

przy czym

$$
P\left(\left(X^*_{(m_1)} = x_{(l_1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(l_2)}\right)\right) = P\left(\left(X^*_{(m_1)} = x_{(n-l_2+1)}\right) \wedge \left(X^*_{(m_1+1)} = x_{(n-l_1+1)}\right)\right).
$$

Aby wyznaczyć rozkład estymatora $X^{*a}_{(m_1, m_1+1)}$ należy określić wszystkie możliwe jego realizacje i je uporządkować, a następnie ze wzorów od (8) do (12) obliczyć prawdopodobieństwa ich wystąpienia[3]. Prawdopodobieństwa te należy w pewnych przypadkach zsumować. Ponieważ nie jest z góry znana liczba oraz kolejność wystąpienia uporządkowanych realizacji estymatora $X^{*a}_{(m_1, m_1+1)}$, nie można podać ogólnej postaci jego rozkładu – prawdopodobieństwa dla poszczególnych realizacji estymatora zależą od elementów próby.

Realizacjami estymatora $X^{*b}_{(m_1, m_1+1)}$ są jedynie elementy próby. Prawdopodobieństwa dane wzorami od (8) do (12) należy podzielić na pół i przypisać je elementom występującym na pozycji m_1 i m_1+1. Ostatecznie prawdopodobieństwa dla poszczególnych realizacji estymatora określone są następująco:

$$
\begin{cases}
P\left(X^{*b}_{(m_1, m_1+1)} = x_{(1)}\right) = P\left(X^{*}_{(m_1, m_1+1)} = x_{(n)}\right) = \displaystyle\sum_{i=m_1+1}^{n}\left(\frac{1}{n}\right)^i\left(\frac{n-1}{n}\right)^{n-i}\binom{n}{i} + \\
+\dfrac{1}{2}\displaystyle\sum_{l=2}^{n-1}\sum_{i=0}^{m_1-1}\left(\frac{1}{n}\right)^{n-i}\left(\frac{n-l}{n}\right)^{i}\binom{n}{m_1, n-m_1-i, i} + \dfrac{1}{2}\left(\frac{1}{n}\right)^n\binom{n}{m_1},
\end{cases}
\tag{13}
$$

$$
\begin{cases}
P\left(X^{*b}_{(m_1, m_1+1)} = x_{(2)}\right) = P\left(X^{*}_{(m_1, m_1+1)} = x_{(n-1)}\right) = \displaystyle\sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{1}{n}\right)^{n-j}\left(\frac{n-2}{n}\right)^{j}\binom{n}{i, n-i-j, j} + \\
+\dfrac{1}{2}\displaystyle\sum_{i=0}^{m_1-1}\left(\frac{1}{n}\right)^{n-i}\left(\frac{n-2}{n}\right)^{i}\binom{n}{m_1, n-m_1-i, i} + \dfrac{1}{2}\displaystyle\sum_{i=0}^{m_1-1}\left(\frac{1}{n}\right)^{n}\binom{n}{i, n-m_1-i, m_1} + \\
+\dfrac{1}{2}\displaystyle\sum_{l=3}^{n-1}\sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{1}{n}\right)^{n-j}\left(\frac{n-l}{n}\right)^{j}\binom{n}{i, m_1-i, m_1-j, j},
\end{cases}
\tag{14}
$$

[3] Prawdopodobieństwa te umieszczono na stronie http://joanna_kisielinska.users.sggw.pl dla liczebności parzystych z przedziału 4–200.

$$\begin{cases} P\left(X^{*b}_{(m_1,m_1+1)} = x_{(l)}\right) = P\left(X^{*b}_{(m_1,m_1+1)} = x_{(n-l+1)}\right) = \\ = \sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{l-1}{n}\right)^i\left(\frac{1}{n}\right)^{n-i-j}\left(\frac{n-l}{n}\right)^j\binom{n}{i,n-i-j,j} + \\ + \frac{1}{2}\sum_{i=0}^{m_1-1}\left(\frac{1}{n}\right)^{n-i}\left(\frac{n-l}{n}\right)^i\binom{n}{m_1,n-m_1-i,i} + \frac{1}{2}\sum_{i=0}^{m_1-1}\left(\frac{l-1}{n}\right)^i\left(\frac{1}{n}\right)^{n-i}\binom{n}{i,n-m_1-i,m_1} + \\ + \frac{1}{2}\sum_{l_1=2}^{l-1}\sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{l_1-1}{n}\right)^i\left(\frac{1}{n}\right)^{n-i-j}\left(\frac{n-l}{n}\right)^j\binom{n}{i,m_1-i,m_1-j,j} + \\ + \frac{1}{2}\sum_{l_2=l+1}^{n-1}\sum_{i=0}^{m_1-1}\sum_{j=0}^{m_1-1}\left(\frac{l-1}{n}\right)^i\left(\frac{1}{n}\right)^{n-i-j}\left(\frac{n-l_2}{n}\right)^j\binom{n}{i,m_1-i,m_1-j,j}, \quad \text{dla } l = 3,4,\ldots,m_1. \end{cases} \tag{15}$$

Rozkład dany formułami od (13) do (15) ma dwie istotne własności, które cechowały rozkład estymatora mediany dla próby o nieparzystej liczbie obserwacji. Tak więc $P\left(X^{*b}_{(m_1,m_1+1)} = x_{(l)}\right)$, dla $l = 1, 2, \ldots, n$ są jednakowe dla wszystkich prób o zadanej liczebności (co umożliwia opracowanie gotowych tablic[4]) oraz spełniają warunek (7).

4. SZACOWANIE MEDIANY DOKŁADNĄ METODĄ PERCENTYLI

Do wyznaczenia przedziałów ufności dla mediany można użyć metody percentyli, opisanej miedzy innymi w pracach: Domański, Pruska (2000), Wilcox (2001), Pekasiewicz (2015). Podkreślić należy, że losowanie wtórnych prób bootstrapowych z całej ich populacji (o liczebności n^n) nie jest tu potrzebne. Dla estymatorów $X^*_{(m)}$ i $X^{*b}_{(m_1,m_1+1)}$ można z góry określić, które elementy próby (pierwotnej) są granicami przedziałów dla zadanych poziomów ufności, ponieważ prawdopodobieństwa poszczególnych ich realizacji nie zależą od elementów próby, a jedynie od jej rozmiaru. W przypadku estymatora $X^{*a}_{(m_1,m_1+1)}$ wyznaczenie granic wymaga dodatkowych obliczeń, związanych z wyznaczaniem rozkładu (prawdopodobieństwa poszczególnych jego realizacji zależą od elementów próby) i w tym przypadku losowanie prób wtórnych może być uzasadnione, aczkolwiek niekonieczne.

Przewaga dokładnej metody percentyli nad metodą klasyczną (z losowaniem) dla estymatorów $X^*_{(m)}$ i $X^{*b}_{(m_1,m_1+1)}$ wynika przede wszystkim z jej prostoty. Wiadomo, że stosując dokładną metoda bootstrapową nie wprowadza się dodatkowych błędów, wynikających z wtórnego próbkowania (Hutson, Ernst, 2000). W przypadku metody klasycznej liczba losowanych prób wtórnych jest duża, a wobec tego ewentualne błędy będą niewielkie.

Wyznaczając granice przedziałów ufności dla mediany należy pamiętać, że rozkłady jej bootstrapowych estymatorów dla skończonego n są dyskretne, a wobec tego

[4] Na stronie http://joanna_kisielinska.users.sggw.pl podano tablice rozkładu bootstrapowego standardowego estymatora mediany dla liczebności parzystych z przedziału 4–200.

ich dystrybuanty nie przyjmują wszystkich wartości z przedziału [0,1]. Dla zadanego poziomu ufności $1-\alpha$, lewą granicą przedziału ufności mediany będzie największa realizacja estymatora spełniająca warunek:

$$F\left(x_{(d)}^{*}\right) \leq \frac{\alpha}{2}, \tag{16}$$

gdzie F jest dystrybuantą wybranego estymatora mediany. Granicą prawą przedziału jest natomiast realizacja najmniejsza, taka że:

$$F\left(x_{(g)}^{*}\right) \geq 1-\frac{\alpha}{2}. \tag{17}$$

Brak równości we wzorach (16) i (17), wynikający z dyskretności rozkładu estymatora, powoduje, że rzeczywisty poziom istotności, który oznaczony zostanie jako α_f, będzie zwykle mniejszy od założonego α.

4.1. PRÓBA O NIEPARZYSTEJ LICZBIE OBSERWACJI

Jak wspomniano, realizacjami estymatora $X_{(m)}^{*}$ są jedynie elementy próby pierwotnej, wobec czego we wzorach (16) i (17) zamiast $x_{(d)}^{*}$ i $x_{(g)}^{*}$ należałoby wstawić $x_{(d)}$ i $x_{(g)}$. Ponieważ jego rozkład określony formułami (5) lub (6) nie zależy od elementów próby, a jedynie od jej liczebności, pozycje d i g określające granice przedziałów ufności można z góry określić dla dowolnych prób o zadanej liczebności[5]. Ze względu na własność (7) zachodzi ponadto: $g = n - d + 1$.

Znając granice przedziałów ufności można wyznaczyć faktyczny poziom istotności, który spełnia warunki: $F\left(x_{(d)}^{*}\right) = \frac{\alpha_f}{2}$ oraz $F\left(x_{(g)}^{*}\right) = 1-\frac{\alpha_f}{2}$.

Połowa długości przedziału ufności jest miarą dokładności oszacowania mediany. Dla małych prób jest ona średnią odległością granic przedziału od mediany estymatora. Z własności (7) wynika bowiem, że granice przedziałów ufności nie będą równo oddalone od mediany estymatora, jeżeli próba nie spełnia następującego warunku:

$$\forall l \in [1, m-1] : x_{(m)} - x_{(l)} = x_{(n-l+1)} - x_{(m)}, \tag{18}$$

Dla prób dużych, miarą dokładności może być natomiast wariancja lub odchylenie standardowe estymatora, ponieważ wartość oczekiwana staje się wówczas bliska medianie (rozkładem granicznym mediany jest rozkład normalny dany formułą (3)). Wariancję lub odchylenie standardowe bootstrapowego estymatora mediany można wyznaczyć, ponieważ znany jest jego rozkład (określony formułą (5 lub 6). Można także wykorzystać wzory podane w pracy Maritz, Jarrett (1978).

---

[5] Numery obserwacji stanowiące granice przedziałów ufności dla liczebności nieparzystych z przedziału 3–201 i wybranych poziomów ufności, podano na stronie http://joanna_kisielinska.users.sggw.pl.

4.2. PRÓBA O PARZYSTEJ LICZBIE OBSERWACJI

Jak pokazano, nie można podać ogólnej postaci rozkładu estymatora $X^{*a}_{(m_1,m_1+1)}$, ponieważ zależy on od elementów próby. W konsekwencji tego nie jest możliwe określenie z góry granic przedziałów ufności. Dla każdej małej próby trzeba wyznaczać je indywidualnie. Ponieważ prawdopodobieństwo odpowiadające dowolnej l-tej realizacji estymatora $X^{*a}_{(m_1,m_1+1)}$ może być różne od prawdopodobieństwa dla realizacji o pozycji $n-l+1$, prawdopodobieństwa $\alpha_{fd} = P\left(X^{*a}_{(m_1,m_1+1)} < x^*_{(d)}\right)$ i $\alpha_{fg} = P\left(X^{*a}_{(m_1,m_1+1)} > x^*_{(g)}\right)$ mogą być różne. Prawdopodobieństwo α_{fd} jest faktycznym lewostronnym poziomem istotności, zaś α_{fg} prawostronnym.

Z własności estymatora $X^{*b}_{(m_1,m_1+1)}$ (tak jak dla estymatora $X^*_{(m)}$) wynika, że można z góry określić pozycje granic przedziałów ufności dla wszystkich prób o zadanej liczebności[6], przy czym między numerami obserwacji stanowiącymi jego granice zachodzi zależność:

$$g = n - d + 1.$$

5. WYNIKI BADAŃ

Badania symulacyjne przeprowadzono dla populacji o wybranych sześciu rozkładach – czterech asymetrycznych i dwóch symetrycznych. Rozkłady te oznaczone zostały jako $F(20,10)$, $F(2,10)$, $CHI(2)$, $CHI(10)$, $N(5,1)$ i $N(5,4)$.

Dla różnych liczebności losowano 1000 razy po n liczb pseudolosowych z przedziału [0,1], które traktowano jako dystrybuantę rozkładów. Zastosowano jednakowe wartości dystrybuant dla wszystkich rozkładów, co pozwala uzyskać lepszą porównywalność wyników. Dobór taki uniezależnia wyniki obliczeń dla poszczególnych rozkładów od jakości generatora liczb pseudolosowych. Mając wylosowanych n wartości dystrybuanty, dla każdego z wziętego pod uwagę rozkładów, określano próby losowe. Dla 1000 wylosowanych w ten sposób prób wyznaczono rozkłady bootstrapowych estymatorów mediany, co pozwoliło na jej przedziałowe oszacowanie.

Wszystkie obliczenia wykonane zostały w Excelu z wykorzystaniem języka VBA for Application.

5.1. ESTYMACJA MEDIANY DOKŁADNĄ METODĄ BOOTSTRAPOWĄ DLA PRÓBY O NIEPARZYSTEJ LICZBIE OBSERWACJI

Badania przeprowadzono dla prób o liczebnościach od 11 do 201. Liczebności dobrano tak, aby obejmowały zarówno małe jak i duże próby. Aby możliwa była prezentacja wyników liczebności zwiększano o 10.

---

[6] Numery obserwacji stanowiące granice przedziałów ufności dla liczebności parzystych z przedziału 4–200 i wybranych poziomów ufności, podano na stronie http://joanna_kisielinska.users.sggw.pl.

Wyniki estymacji przedziałowej mediany dokładną metodą percentyli dla wybranych rozkładów przy użyciu estymatora $X^*_{(m)}$ przedstawiono w tabelach 1 i 2.

Tabela 1.

Lewa granica przedziału ufności oraz faktyczny i zliczeniowy poziom istotności uzyskany dokładną metodą percentyli dla *n* nieparzystego

n	α = 0,05			α = 0,04			α = 0,03			α = 0,02			α = 0,01		
	d	α_f	α_o	d	α_f	α_o	d	α_f	α_o	d	α_f	α_o	d	α_f	α_o
11	2	0,014	0,012	2	0,014	0,012	2	0,014	0,012	2	0,014	0,012	1	0,000	0,000
21	6	0,037	0,031	6	0,037	0,031	5	0,009	0,004	5	0,009	0,004	5	0,009	0,004
31	10	0,039	0,039	10	0,039	0,039	9	0,014	0,019	9	0,014	0,019	8	0,004	0,011
41	14	0,036	0,024	14	0,036	0,024	13	0,015	0,016	13	0,015	0,016	12	0,005	0,012
51	18	0,031	0,019	18	0,031	0,019	17	0,014	0,004	17	0,014	0,004	16	0,006	0,000
61	22	0,026	0,016	22	0,026	0,016	22	0,026	0,016	21	0,012	0,012	20	0,005	0,008
71	27	0,040	0,034	27	0,040	0,034	26	0,021	0,024	25	0,011	0,016	24	0,005	0,008
81	31	0,032	0,027	31	0,032	0,027	30	0,017	0,021	30	0,017	0,021	29	0,009	0,018
91	36	0,043	0,034	35	0,025	0,022	35	0,025	0,022	34	0,014	0,014	33	0,007	0,007
101	40	0,034	0,030	40	0,034	0,030	39	0,020	0,015	38	0,011	0,011	37	0,006	0,008
111	45	0,044	0,020	44	0,027	0,017	44	0,027	0,017	43	0,016	0,011	42	0,009	0,006
121	49	0,034	0,034	49	0,034	0,034	48	0,021	0,026	47	0,013	0,013	46	0,007	0,009
131	54	0,042	0,028	53	0,027	0,014	53	0,027	0,014	52	0,017	0,003	50	0,006	0,001
141	58	0,034	0,029	58	0,034	0,029	57	0,021	0,020	56	0,013	0,012	55	0,008	0,009
151	63	0,040	0,035	62	0,027	0,027	62	0,027	0,027	61	0,017	0,019	59	0,006	0,001
161	68	0,047	0,047	67	0,032	0,029	66	0,021	0,021	65	0,014	0,015	64	0,008	0,012
171	72	0,037	0,019	72	0,037	0,019	71	0,025	0,012	70	0,017	0,008	68	0,007	0,007
181	77	0,043	0,028	76	0,030	0,021	76	0,030	0,021	74	0,013	0,015	73	0,009	0,012
191	82	0,049	0,045	81	0,035	0,032	80	0,024	0,022	79	0,016	0,013	77	0,007	0,008
201	86	0,040	0,034	86	0,040	0,034	85	0,028	0,024	84	0,019	0,021	82	0,008	0,011

Uwagi: d – numer obserwacji, która jest dolną granicą przedziału ufności, α_f – faktyczny poziom ufności, α_o – udział przedziałów niezawierających mediany rozkładu wyznaczony z 1000 prób.

Źródło: badania własne.

W tabeli 1 przedstawiono numery obserwacji stanowiące lewą granice przedziałów ufności oraz faktyczne poziomy istotności α_f. Mając numer obserwacji stanowiący lewą granicę przedziału, można dla dowolnej próby losowej natychmiast (bez konieczności losowania prób wtórnych) wyznaczyć przedziały ufności. Ponieważ rozkład estymatora jest dyskretny, faktyczne poziomy ufności $1-\alpha_f$ są wyższe od założonych. W tabeli 1 podano ponadto udziały przedziałów niezawierających mediany rozkładu α_o (zliczeniowe poziomy istotności). Udziały te były zwykle mniejsze od α_f. Ze względu na sposób losowania prób udziały przedziałów niezawierających mediany były jednakowe dla wszystkich rozkładów. Wartości te nie stanowią dokładnej informacji o faktycznych poziomach ufności, można je jedynie traktować jako sprawdziany poprawności działania zastosowanego generatora liczb pseudolosowych.

Tabela 2.

Średnie szerokości połowy przedziałów ufności dla $\alpha = 0{,}05$ oraz średnie oszacowania odchylenia standardowego wyznaczone z 1000 prób dla n nieparzystego

n	F(20,10) $M_{0,5} = 1{,}0349$ $\sigma = 0{,}8539$		F(2,10) $M_{0,5} = 0{,}7435$ $\sigma = 1{,}6137$		CHI(2) $M_{0,5} = 1{,}3863$ $\sigma = 2{,}0000$		CHI(10) $M_{0,5} = 9{,}3418$ $\sigma = 4{,}4721$		N(5,1) $M_{0,5} = 5$ $\sigma = 1{,}000$		N(5,4) $M_{0,5} = 5$ $\sigma = 4{,}0000$	
	d	σ_b	d	σ_b	d	σ_b	d	σ_b	d	σ_b	d	σ_b
11	0,6997	0,2657	1,1761	0,4575	1,8108	0,7132	4,5660	1,7577	1,0511	0,4017	4,2045	1,6067
21	0,3814	0,1764	0,6177	0,2909	1,0304	0,4833	2,7045	1,2451	0,6265	0,2871	2,5060	1,1484
31	0,2929	0,1444	0,4705	0,2355	0,7964	0,3966	2,1108	1,0347	0,4902	0,2394	1,9610	0,9576
41	0,2592	0,1238	0,4142	0,2000	0,7071	0,3402	1,8851	0,8966	0,4385	0,2080	1,7541	0,8320
51	0,2400	0,1129	0,3831	0,1820	0,6556	0,3107	1,7491	0,8213	0,4068	0,1906	1,6271	0,7624
61	0,2229	0,1007	0,3562	0,1620	0,6103	0,2773	1,6252	0,7334	0,3773	0,1700	1,5090	0,6800
71	0,1898	0,0927	0,3026	0,1490	0,5205	0,2554	1,3899	0,6759	0,3229	0,1566	1,2915	0,6264
81	0,1826	0,0855	0,2904	0,1367	0,5005	0,2353	1,3408	0,6272	0,3120	0,1458	1,2480	0,5833
91	0,1621	0,0808	0,2574	0,1291	0,4443	0,2224	1,1927	0,5934	0,2778	0,1380	1,1112	0,5522
101	0,1631	0,0774	0,2594	0,1237	0,4477	0,2131	1,1991	0,5675	0,2788	0,1318	1,1153	0,5272
111	0,1461	0,0729	0,2322	0,1163	0,4014	0,2008	1,0760	0,5358	0,2502	0,1245	1,0010	0,4981
121	0,1466	0,0698	0,2330	0,1114	0,4028	0,1923	1,0793	0,5128	0,2510	0,1191	1,0039	0,4764
131	0,1349	0,0668	0,2142	0,1065	0,3706	0,1840	0,9940	0,4915	0,2312	0,1142	0,9249	0,4569
141	0,1351	0,0638	0,2146	0,1018	0,3713	0,1759	0,9952	0,4695	0,2313	0,1090	0,9254	0,4361
151	0,1274	0,0625	0,2023	0,0995	0,3503	0,1722	0,9401	0,4605	0,2187	0,1070	0,8748	0,4282
161	0,1184	0,0597	0,1878	0,0950	0,3254	0,1644	0,8741	0,4402	0,2034	0,1024	0,8136	0,4096
171	0,1213	0,0587	0,1925	0,0934	0,3335	0,1616	0,8945	0,4325	0,2080	0,1006	0,8320	0,4022
181	0,1133	0,0564	0,1797	0,0898	0,3115	0,1555	0,8369	0,4165	0,1948	0,0969	0,7790	0,3876
191	0,1066	0,0546	0,1691	0,0867	0,2933	0,1503	0,7884	0,4036	0,1835	0,0940	0,7341	0,3759
201	0,1086	0,0534	0,1722	0,0848	0,2986	0,1470	0,8021	0,3939	0,1866	0,0916	0,7465	0,3665

Uwagi: d – średnie szerokości połowy przedziałów ufności dla $\alpha = 0{,}05$, σ_b – średnie oszacowania odchylenia standardowego estymatora bootstrapowego.

Źródło: badania własne.

Tabela 2 zawiera średnie szerokości połowy przedziałów ufności dla α = 0,05 oraz średnie oszacowania odchylenia standardowego estymatora wyznaczone z 1000 prób dla wybranych rozkładów. Oczywiste jest, że wraz ze wzrostem liczebności próby obydwie miary dokładności szacunku zmniejszają się, czyli oszacowania są coraz bardziej precyzyjne. Zauważyć należy, że z wyjątkiem n = 11 połowy długości przedziałów ufności dla α = 0,05 są mniej więcej dwukrotnie większe niż odchylenie standardowe. Uzyskane wyniki nie stoją więc w sprzeczności z wynikami Wilcoxa (2001), na które powołuje się Pekasiewicz (2015) (str. 53), że rozkład mediany z próby jest zbieżny do rozkładu normalnego dla prób już 20-elementowych.

5.2. ESTYMACJA MEDIANY DOKŁADNĄ METODĄ BOOTSTRAPOWĄ DLA PRÓBY O PARZYSTEJ LICZBIE OBSERWACJI

Badania wykonano dla prób o liczebnościach od 10 do 200 (zwiększając liczebności o 10). Symulacje przeprowadzono używając estymatorów $X^{*a}_{(m_1,m_1+1)}$ i $X^{*b}_{(m_1,m_1+1)}$. Dla obydwu estymatorów zastosowano takie same zestawy 1000 prób losowych dla każdej z liczebności.

W tabelach 3, 4 i 5 przedstawiono wyniki estymacji przedziałowej mediany dokładną metodą percentyli dla wybranych rozkładów. Tabela 3 zawiera dla estymatora $X^{*a}_{(m_1,m_1+1)}$ średnie faktyczne poziomy istotności $\overline{\alpha}_{fd}$ i $\overline{\alpha}_{fg}$ obliczone z 1000 prób. Pamiętać bowiem należy, że w ogólnym przypadku wartości α_{fd} i α_{fg} są różne dla poszczególnych prób i rozkładów, choć badania pokazały, że różnice te były niewielkie. Bardzo małe są również różnice pomiędzy $\overline{\alpha}_{fd}$ i $\overline{\alpha}_{fg}$.

W przypadku estymatora $X^{*b}_{(m_1,m_1+1)}$ podano numery obserwacji będące lewymi granicami przedziałów ufności, oraz połowę faktycznego poziomu istotności $\frac{\alpha_f}{2}$. Porównując $\frac{\alpha_f}{2}$ z $\overline{\alpha}_{fd}$ i $\overline{\alpha}_{fg}$ stwierdzić można, że z wyjątkiem nielicznych przypadków $\frac{\alpha_f}{2}$ jest mniejsze od $\overline{\alpha}_{fd}$ i $\overline{\alpha}_{fg}$, co jest konsekwencją znacznie mniejszej liczby realizacji estymatora $X^{*b}_{(m_1,m_1+1)}$ od $X^{*a}_{(m_1,m_1+1)}$. Na podkreślenie zasługuje fakt, że różnice pomiędzy $\frac{\alpha_f}{2}$ i $\overline{\alpha}_{fd}$ oraz $\overline{\alpha}_{fg}$ są zwykle małe.

W tabeli 4 podano średnie szerokości połowy przedziałów ufności dla α= 0,05 oraz średnie oszacowania odchylenia standardowego estymatora wyznaczone z 1000 prób dla estymatora $X^{*a}_{(m_1,m_1+1)}$, zaś w tabeli 5 dla estymatora $X^{*b}_{(m_1,m_1+1)}$. Podkreślić należy, że połowy długości przedziałów ufności oraz odchylenia standardowe uzyskane estymatorem $X^{*a}_{(m_1,m_1+1)}$ są dla wszystkich badanych rozkładów i wszystkich liczebności prób mniejsze, niż uzyskane estymatorem $X^{*b}_{(m_1,m_1+1)}$. Można więc stwierdzić, że dla badanych rozkładów estymator $X^{*b}_{(m_1,m_1+1)}$ ma mniejszą wariancję i pozwala uzyskać dokładniejsze przedziałowe oszacowanie mediany niż estymator $X^{*a}_{(m_1,m_1+1)}$.

Tabela 3.

Faktyczne poziomy istotności dla estymatora $X^{*a}_{(m_1,m_1+1)}$ oraz lewa granica przedziału ufności dla estymatora $X^{*b}_{(m_1,m_1+1)}$ uzyskane dokładną metodą percentyli dla n parzystego

n	$\alpha = 0,05$				$\alpha = 0,04$				$\alpha = 0,03$				$\alpha = 0,02$				$\alpha = 0,01$			
	$X^{*a}_{(m_1,m_1+1)}$		$X^{*b}_{(m_1,m_1+1)}$		$X^{*a}_{(m_1,m_1+1)}$		$X^{*b}_{(m_1,m_1+1)}$		$X^{*a}_{(m_1,m_1+1)}$		$X^{*b}_{(m_1,m_1+1)}$		$X^{*a}_{(m_1,m_1+1)}$		$X^{*b}_{(m_1,m_1+1)}$		$X^{*a}_{(m_1,m_1+1)}$		$X^{*b}_{(m_1,m_1+1)}$	
	$\overline{\alpha}_{fd}$	$\overline{\alpha}_{fg}$	d	$\frac{\alpha_f}{2}$	$\overline{\alpha}_{fd}$	$\overline{\alpha}_{fg}$	d	$\frac{\alpha_f}{2}$	$\overline{\alpha}_{fd}$	$\overline{\alpha}_{fg}$	d	$\frac{\alpha_f}{2}$	$\overline{\alpha}_{fd}$	$\overline{\alpha}_{fg}$	d	$\frac{\alpha_f}{2}$	$\overline{\alpha}_{fd}$	$\overline{\alpha}_{fg}$	d	$\frac{\alpha_f}{2}$
10	0,0195	0,0195	2	0,0196	0,0195	0,0195	2	0,0196	0,0070	0,0070	1	0,0009	0,0070	0,0070	1	0,0009	0,0011	0,0012	1	0,0009
20	0,0208	0,0208	5	0,0089	0,0177	0,0171	5	0,0089	0,0109	0,0112	5	0,0089	0,0092	0,0092	5	0,0089	0,0046	0,0047	4	0,0016
30	0,0226	0,0226	9	0,0117	0,0171	0,0174	9	0,0117	0,0138	0,0136	9	0,0117	0,0077	0,0077	8	0,0035	0,0042	0,0043	8	0,0035
40	0,0206	0,0205	13	0,0115	0,0182	0,0183	13	0,0115	0,0132	0,0136	13	0,0115	0,0083	0,0083	12	0,0043	0,0045	0,0046	12	0,0043
50	0,0238	0,0239	18	0,0222	0,0175	0,0175	17	0,0104	0,0132	0,0130	17	0,0104	0,0078	0,0078	16	0,0044	0,0046	0,0047	16	0,0044
60	0,0222	0,0222	22	0,0181	0,0194	0,0194	22	0,0181	0,0141	0,0143	21	0,0089	0,0095	0,0096	21	0,0089	0,0045	0,0046	20	0,0041
70	0,0224	0,0224	26	0,0146	0,0181	0,0179	26	0,0146	0,0119	0,0119	26	0,0146	0,0085	0,0087	25	0,0075	0,0042	0,0042	24	0,0036
80	0,0234	0,0232	31	0,0211	0,0178	0,0177	30	0,0118	0,0132	0,0135	30	0,0118	0,0096	0,0096	29	0,0062	0,0045	0,0047	28	0,0031
90	0,0240	0,0239	35	0,0166	0,0186	0,0187	35	0,0166	0,0141	0,0141	34	0,0094	0,0090	0,0088	34	0,0094	0,0041	0,0042	32	0,0026
100	0,0234	0,0235	40	0,0219	0,0190	0,0190	39	0,0131	0,0140	0,0140	39	0,0131	0,0091	0,0089	38	0,0076	0,0046	0,0047	37	0,0042
110	0,0241	0,0242	44	0,0172	0,0187	0,0188	44	0,0172	0,0141	0,0143	43	0,0104	0,0089	0,0089	42	0,0061	0,0047	0,0047	41	0,0034
120	0,0238	0,0235	49	0,0216	0,0190	0,0189	48	0,0136	0,0145	0,0144	48	0,0136	0,0093	0,0094	47	0,0083	0,0044	0,0042	46	0,0048
130	0,0233	0,0233	53	0,0170	0,0188	0,0187	53	0,0170	0,0144	0,0143	52	0,0107	0,0094	0,0093	51	0,0066	0,0045	0,0045	50	0,0039
140	0,0233	0,0230	58	0,0207	0,0184	0,0184	57	0,0135	0,0143	0,0143	57	0,0135	0,0094	0,0094	56	0,0085	0,0045	0,0045	54	0,0031
150	0,0220	0,0220	63	0,0245	0,0187	0,0187	62	0,0164	0,0145	0,0145	61	0,0106	0,0094	0,0094	60	0,0067	0,0046	0,0046	59	0,0042
160	0,0232	0,0232	67	0,0194	0,0175	0,0175	67	0,0194	0,0142	0,0142	66	0,0130	0,0093	0,0093	65	0,0084	0,0047	0,0047	63	0,0033
170	0,0239	0,0239	72	0,0226	0,0187	0,0187	71	0,0154	0,0138	0,0138	70	0,0103	0,0091	0,0091	69	0,0067	0,0047	0,0047	68	0,0043
180	0,0235	0,0235	76	0,0180	0,0190	0,0191	76	0,0180	0,0140	0,0140	75	0,0123	0,0093	0,0093	74	0,0082	0,0047	0,0047	72	0,0034
190	0,0237	0,0237	81	0,0207	0,0188	0,0188	80	0,0144	0,0140	0,0141	80	0,0144	0,0088	0,0088	79	0,0098	0,0047	0,0047	77	0,0042
200	0,0243	0,0243	86	0,0234	0,0189	0,0189	85	0,0166	0,0145	0,0145	84	0,0115	0,0093	0,0093	83	0,0078	0,0046	0,0046	81	0,0034

Uwagi: $\overline{\alpha}_{fd}$ – średni faktyczny poziom istotności lewostronny, $\overline{\alpha}_{fg}$ – średni faktyczny poziom istotności prawostronny, d – numer obserwacji, która jest dolną granicą przedziału ufności, α_f – połowa faktycznego poziom istotności.

Źródło: badania własne.

Tabela 4.

Średnie szerokości połowy przedziałów ufności estymatora $X^{*a}_{(m_1,m_1+1)}$ dla $\alpha = 0{,}05$ oraz średnie oszacowania odchylenia standardowego wyznaczone z 1000 prób dla n parzystego

n	F(20,10) $M_{0,5} = 1{,}0349$ $\sigma = 0{,}8539$		F(2,10) $M_{0,5} = 0{,}7435$ $\sigma = 1{,}6137$		CHI(2) $M_{0,5} = 1{,}3863$ $\sigma = 2{,}0000$		CHI(10) $M_{0,5} = 9{,}3418$ $\sigma = 4{,}4721$		N(5,1) $M_{0,5} = 5$ $\sigma = 1{,}000$		N(5,4) $M_{0,5} = 5$ $\sigma = 4{,}0000$	
	d	σ_b	d	σ_b	d	σ_b	d	σ_b	d	σ_b	d	σ_b
10	0,5217	0,2763	0,8681	0,4813	1,3657	0,7292	3,4855	1,7866	0,8042	0,4087	3,2169	1,6348
20	0,3281	0,1842	0,5333	0,3046	0,8941	0,5050	2,3394	1,2951	0,5438	0,2977	2,1751	1,1908
30	0,2579	0,1469	0,4141	0,2401	0,7023	0,4033	1,8598	1,0487	0,4302	0,2422	1,7207	0,9688
40	0,2337	0,1220	0,3728	0,1966	0,6375	0,3345	1,6926	0,8858	0,3941	0,2063	1,5765	0,8253
50	0,2051	0,1102	0,3278	0,1776	0,5615	0,3032	1,4949	0,8009	0,3467	0,1857	1,3869	0,7430
60	0,1818	0,0962	0,2882	0,1536	0,4972	0,2641	1,3366	0,7060	0,3118	0,1646	1,2470	0,6586
70	0,1727	0,0927	0,2747	0,1483	0,4733	0,2550	1,2677	0,6798	0,2950	0,1582	1,1798	0,6329
80	0,1652	0,0860	0,2639	0,1387	0,4540	0,2376	1,2080	0,6248	0,2798	0,1440	1,1193	0,5761
90	0,1483	0,0797	0,2355	0,1271	0,4072	0,2194	1,0932	0,5866	0,2545	0,1365	1,0179	0,5460
100	0,1393	0,0742	0,2214	0,1187	0,3825	0,2044	1,0257	0,5441	0,2387	0,1264	0,9549	0,5055
110	0,1354	0,0747	0,2154	0,1195	0,3722	0,2060	0,9970	0,5479	0,2317	0,1271	0,9268	0,5083
120	0,1321	0,0692	0,2103	0,1102	0,3633	0,1905	0,9710	0,5099	0,2253	0,1187	0,9013	0,4748
130	0,1280	0,0679	0,2031	0,1082	0,3517	0,1869	0,9441	0,4998	0,2197	0,1162	0,8788	0,4650
140	0,1217	0,0639	0,1936	0,1019	0,3349	0,1760	0,8959	0,4695	0,2080	0,1090	0,8321	0,4360
150	0,1184	0,0622	0,1879	0,0991	0,3256	0,1714	0,8737	0,4581	0,2032	0,1065	0,8129	0,4259
160	0,1125	0,0600	0,1784	0,0955	0,3093	0,1653	0,8314	0,4427	0,1935	0,1030	0,7742	0,4119
170	0,1111	0,0584	0,1763	0,0930	0,3056	0,1609	0,8206	0,4307	0,1909	0,1001	0,7637	0,4005
180	0,1053	0,0554	0,1670	0,0881	0,2895	0,1526	0,7779	0,4085	0,1810	0,0950	0,7240	0,3799
190	0,1031	0,0547	0,1634	0,0870	0,2835	0,1507	0,7621	0,4039	0,1774	0,0939	0,7095	0,3757
200	0,1009	0,0533	0,1600	0,0847	0,2776	0,1467	0,7458	0,3931	0,1735	0,0914	0,6941	0,3656

Uwagi: d – średnie szerokości połowy przedziałów ufności dla $\alpha = 0{,}05$, σ_b – średnie oszacowania odchylenia standardowego.

Źródło: badania własne.

Tabela 5.

Średnie szerokości połowy przedziałów ufności estymatora $X^{*b}_{(m_1,m_1+1)}$ dla $\alpha = 0,05$ oraz średnie oszacowania odchylenia standardowego wyznaczone z 1000 prób dla n parzystego

n	F(20,10) $M_{0,5} = 1,0349$ $\sigma = 0,8539$		F(2,10) $M_{0,5} = 0,7435$ $\sigma = 1,6137$		CHI(2) $M_{0,5} = 1,3863$ $\sigma = 2,0000$		CHI(10) $M_{0,5} = 9,3418$ $\sigma = 4,4721$		N(5,1) $M_{0,5} = 5$ $\sigma = 1,000$		N(5,4) $M_{0,5} = 5$ $\sigma = 4,0000$	
	d	σ_b	d	σ_b	d	σ_b	d	σ_b	d	σ_b	d	σ_b
10	0,6594	0,3108	1,1047	0,5438	1,7056	0,8179	4,3241	1,9970	0,9980	0,4567	3,9919	1,8268
20	0,4605	0,1963	0,7455	0,3249	1,2316	0,5382	3,2441	1,3788	0,7568	0,3169	3,0270	1,2676
30	0,3424	0,1536	0,5526	0,2514	0,9293	0,4220	2,4481	1,0962	0,5671	0,2531	2,2685	1,0123
40	0,3108	0,1261	0,4986	0,2033	0,8453	0,3459	2,2435	0,9154	0,5214	0,2132	2,0858	0,8528
50	0,2299	0,1132	0,3674	0,1825	0,6287	0,3116	1,6743	0,8226	0,3889	0,1908	1,5554	0,7631
60	0,2081	0,0984	0,3306	0,1572	0,5690	0,2703	1,5271	0,7224	0,3560	0,1685	1,4241	0,6739
70	0,2057	0,0946	0,3274	0,1514	0,5631	0,2602	1,5076	0,6934	0,3509	0,1614	1,4037	0,6456
80	0,1830	0,0875	0,2925	0,1413	0,5027	0,2419	1,3377	0,6359	0,3100	0,1466	1,2399	0,5863
90	0,1740	0,0810	0,2765	0,1292	0,4773	0,2229	1,2796	0,5959	0,2978	0,1386	1,1912	0,5546
100	0,1509	0,0753	0,2399	0,1204	0,4142	0,2073	1,1103	0,5518	0,2583	0,1282	1,0333	0,5126
110	0,1560	0,0757	0,2481	0,1211	0,4284	0,2087	1,1469	0,5550	0,2666	0,1287	1,0662	0,5149
120	0,1428	0,0701	0,2275	0,1116	0,3928	0,1928	1,0489	0,5161	0,2434	0,1201	0,9734	0,4805
130	0,1448	0,0686	0,2298	0,1094	0,3975	0,1890	1,0675	0,5053	0,2485	0,1175	0,9942	0,4701
140	0,1353	0,0645	0,2153	0,1030	0,3722	0,1779	0,9957	0,4745	0,2312	0,1101	0,9248	0,4406
150	0,1234	0,0628	0,1958	0,1000	0,3391	0,1730	0,9101	0,4626	0,2117	0,1075	0,8468	0,4300
160	0,1244	0,0606	0,1974	0,0964	0,3420	0,1669	0,9185	0,4468	0,2138	0,1039	0,8551	0,4157
170	0,1177	0,0589	0,1868	0,0938	0,3237	0,1623	0,8689	0,4344	0,2021	0,1010	0,8085	0,4040
180	0,1177	0,0558	0,1867	0,0888	0,3236	0,1538	0,8690	0,4118	0,2022	0,0958	0,8088	0,3830
190	0,1124	0,0551	0,1783	0,0877	0,3091	0,1519	0,8307	0,4070	0,1933	0,0947	0,7733	0,3786
200	0,1063	0,0537	0,1686	0,0853	0,2923	0,1478	0,7851	0,3960	0,1827	0,0921	0,7307	0,3683

Uwagi: d – średnie szerokości połowy przedziałów ufności dla α=0,05, σ_b – średnie oszacowania odchylenia standardowego.

Źródło: badania własne.

W przypadku estymatora $X^{*a}_{(m_1,m_1+1)}$ wzrostowi liczebności próby dla wszystkich badanych rozkładów i wszystkich wziętych pod uwagę liczebności towarzyszy zmniejszanie średniej szerokości przedziałów ufności. Jeśli chodzi o średnie odchylenie standardowe, to jedynie dla liczebności 110 zaobserwowano wzrost jego wartości wobec odchylenia dla n = 100. Dla estymatora $X^{*b}_{(m_1,m_1+1)}$ wzrost średniej długości przedziału wystąpił dla n = 110, 130 i 160 dla wszystkich wziętych pod uwagę rozkładów, zaś dla rozkładów *CHI*(10), *N*(5,1) i *N*(5,4) jeszcze dla n = 180. Wzrost średniego odchylenia standardowego zaobserwowano jedynie dla n = 110.

Relacja połowy średniej długości przedziału ufności dla α = 0,05 do średniego odchylenia standardowego estymatora $X^{*a}_{(m_1,m_1+1)}$ jest równa 1,88, a dla estymatora $X^{*b}_{(m_1,m_1+1)}$ zaś 2,11. Może to wskazywać, że rozkład estymatora $X^{*a}_{(m_1,m_1+1)}$ jest bliższy rozkładowi normalnemu niż estymatora $X^{*b}_{(m_1,m_1+1)}$.

Pruska (2007) badała miedzy innymi wpływ zwiększania liczebności próby o jeden, dla prób 20 i 40 elementowych, na oszacowanie odchylenia standardowego estymatora bootstrapowego w postaci mediany z próby. Z badań tych wynikało, że zwiększenie liczebności (do liczebności nieparzystej) skutkowało dla większości badanych rozkładów zmniejszaniem odchylenia standardowego.

Oszacowania odchylenia standardowego bootstrapowych estymatorów mediany w postaci mediany z próby przedstawione w tabelach 2, 4 i 5, nie wskazują na tak jednoznaczne wyniki. Dla rozkładu *F*(20,10) odchylenia standardowe, dla prób o liczebnościach nieparzystych o jeden większych niż o liczebnościach parzystych, były w 10 na 20 przypadków większe, dla *F*(2,10), *CHI*(2) i *CHI*(10) w 11, dla *N*(5,1) i *N*(5,4) w 13. Połowy szerokości przedziałów ufności zaś, były we wszystkich przypadkach większe dla prób o nieparzystej (większej o jeden) liczebności, co może wynikać ze znacznie większej liczby realizacji estymatora $X^{*a}_{(m_1,m_1+1)}$ od estymatora $X^{*}_{(m)}$. Oparte na podobnych zasadach porównanie odchyleń standardowych estymatorów $X^{*}_{(m)}$ i $X^{*b}_{(m_1,m_1+1)}$ dało wyniki przeciwne. Oszacowane odchylenia standardowe estymatora $X^{*}_{(m)}$ były mniejsze w większości przypadków niż estymatora $X^{*b}_{(m_1,m_1+1)}$, podobnie jak połowy długości przedziałów ufności.

6. PODSUMOWANIE

W artykule pokazano, że dokładna metoda percentyli może być wykorzystana do szacowania mediany. Stosując estymator standardowy, dla prób o wybranych nieparzystych i parzystych liczebnościach, wyznaczono numery obserwacji stanowiące granice przedziałów ufności. Pokazano, że nie jest w tym przypadku potrzebne losowanie prób wtórnych.

Dla wziętych pod uwagę rozkładów i liczebności prób, estymator w postaci mediany z próby o parzystej liczbie obserwacji dawał prawidłowe oszacowania przedziałów ufności, które zawierały faktyczną medianę rozkładu ze zliczeniowym poziomem ufności zbliżonym do założonego.

Estymator w postaci mediany z próby o parzystej liczbie obserwacji ma mniejszą wariancję od estymatora standardowego i dawał dokładniejsze oszacowania mediany (krótsze przedziały ufności). Zastosowanie jego wymaga jednak znacznie większego nakładu obliczeń.

Zwiększenie liczebności próby o jeden (z parzystej do nieparzystej liczebności) skutkowało wzrostem odchylenia standardowego i długości przydziałów ufności w połowie lub więcej przypadków dla estymatora w postaci mediany z próby, zaś zwykle zmniejszeniem tych wielkości w przypadku estymatora standardowego.

LITERATURA

Baszczyńska A., Pekasiewicz D., (2010), Selected MethodS of Interval Estimation of the Median. The Analysis of Accuracy of Estimation, *Acta Universitatis Lodziensis. Folia Oeconomica*, 235, 21–29.

Domański C., Pruska K., (2000), *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Efron B., (1979), Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7 (1), 1–26.

Evans D. L., Leemis L. M., Drew J. H., (2006), The Distribution of Order Statistics for Discrete Random Variables With Applications of Bootstrapping, *Journal on Computing*, 18 (1), 19–30.

Fisher N. I., Hall P., (1991), Bootstrap Algorithms for Small Samples, *Journal of Statistical Planning and Inference*, 27, 157–169.

Hutson A. D., Ernst M. D., (2000), The Exact Bootstrap Mean and Variance of an *L*-estimator, *Journal of the Royal Statistical Society: Series B*, 62 (1), 89–94.

Kisielińska J., (2013), The Exact Bootstrap Method Shown on the Example of the Mean and Variance Estimation, *Computational Statistics*, 28 (3), 1061–1077.

Maritz J. S., Jarrett R. G., (1978), A Note on Estimating the Variance of the Sample Median, *Journal of the American Statistical Association*, 73 (361), 194–196.

Pekasiewicz D., (2015), *Statystyki pozycyjne w procedurach estymacji i ich zastosowania w badaniach ekonomicznych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Pruska K., (2007), Estimation of Bias and Variance of Sample Median by Jackknife and Bootstrap Methods, *Acta Universitatis Lodziensis, Folia Oeconomica*, 206, 67–78.

Wilcox R. R., (2001), Fundamentals of Modern Statistical Methods, Springer, New York.

Zieliński R., (1995), Estimating Median and Other Quantiles in Nonparametric Models, *Applicationes Mathematicae*, 23 (3), 363–370.

ROZKŁADY WYBRANYCH BOOTSTRAPOWYCH ESTYMATORÓW MEDIANY ORAZ ZASTOSOWANIE DOKŁADNEJ METODY PERCENTYLI DO JEJ PRZEDZIAŁOWEGO SZACOWANIA

Streszczenie

W artykule przedstawiono wyniki zastosowania dokładnej metody percentyli do szacowania mediany. Dla próby o nieparzystej liczbie obserwacji zastosowano estymator w postaci mediany z próby (estymator standardowy), zaś dla próby o parzystej liczbie obserwacji estymator w postaci mediany z próby oraz estymator standardowy. Przedstawiono formuły określające rozkład estymatorów standardowych oraz opisano w jaki sposób można wyznaczyć rozkład estymatora w postaci mediany z próby o parzystej liczbie obserwacji. Badania przeprowadzono dla wybranych rozkładów i wybranych liczebności próby. Oszacowano odchylenia standardowe wziętych pod uwagę estymatorów oraz przedziały ufności wyznaczone przy ich pomocy. W pracy pokazano, że użycie dokładnej metody percentyli dla estymatorów standardowych jest znacznie prostsze niż stosowanie metody klasycznej z losowaniem prób wtórnych.

Słowa kluczowe: dokładna metoda precentyli, estymacja mediany

DISTRIBUTION OF SELECTED BOOTSTRAP MEDIAN ESTIMATORS AND APPLICATION OF THE EXACT PERCENTILE METHOD FOR ITS INTERVAL ESTIMATING

A b s t r a c t

The article presents the results of the use of exact percentiles methods for estimating median. For a sample with an odd number of observations used estimator in the form of the sample median (standard estimator), and for a sample with an even number of observations, used estimator in the form of the sample median and standard estimator. It presents a formula defining the distribution of standard estimator and describes how you can determine the distribution of the estimator in the form of the sample median for sample with an even number of observations. The study was conducted for the selected distributions and selected sample size. It was estimated standard deviation and confidence intervals for selected estimators. The study shows, that the use of exact percentiles method for standard estimators is much simpler than using the classical method with the resampling.

Keywords: exact percentiles method, median estimating