METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXI/2, 2020, s. 90 – 101

Received: 14.10.2020
Accepted: 20.11.2020

# WPŁYW ASYMETRII ROZKŁADU NA DOBÓR BOOTSTRAPOWEGO ESTYMATORA KWARTYLI

**Joanna Kisielińska** https://orcid.org/0000-0003-3289-1525
Instytut Ekonomii i Finansów
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: joanna_kisielinska@sggw.edu.pl

**Streszczenie:** Badania dotyczą estymacji kwartyli (pierwszego, drugiego i trzeciego) w sytuacji gdy brak jest informacji o rozkładzie, z którego wylosowana została próba, zaś iloczyn *np* ma wartość całkowitą (gdzie *n* jest liczebnością próby, a *p* rzędem kwantyla). Jeśli *np* nie jest całkowite jako estymator kwantyla rzędu *p* wybierana jest zwykle statystyka pozycyjna rzędu [*np*]+1. Jeśli *np* nie jest całkowite rozwiązań jest znacznie więcej. W niniejszej pracy porównane zostały dwa dokładne bootstrapowe estymatory kwartyli w postaci pojedynczych statystyk pozycyjnych rzędu *np* i *np* +1. Do oceny wykorzystane zostały obciążenie i wariancja estymatora oraz szerokość przedziałów ufności i zliczeniowy poziom ufności. Przedziały ufności wyznaczone zostały dokładną metodą percentyli. Próby losowano z rozkładów o asymetrii prawo i lewostronnej oraz symetrycznych, co umożliwiło wybór estymatora najbardziej odpowiedniego w danej sytuacji.

**Słowa kluczowe:** estymacja kwantyli, estymator bootstrapowy, dokładna metoda bootstrapowa, dokładna metoda percentyli

**JEL classification:** C13, C14

## WSTĘP

Dana jest zmienna losowa $X$ o dystrybuancie $F$ i funkcji gęstości $f$. Kwantylem rzędu $p$ (lub $p$-kwantylem) rozkładu zmiennej $X$ nazywamy $\xi_p$, takie że:

$$\xi_p = \inf\{x : F(x) \geq p\}. \tag{1}$$

Bahadur [1966] pokazał, że do estymacji kwantyli mogą być wykorzystywane kwantyle z próby:

https://doi.org/10.22630/MIBE.2020.21.2.9

$$\xi_{pn} = \inf\{x : F_n(x) \geq p\}. \tag{2}$$

gdzie $F_n$ jest dystrybuantą empiryczną, a $n$ liczebnością próby.

Kwantyle próbkowe przedstawiane są (a nawet utożsamiane [Pekasiewicz 2015, 14]) ze statystykami pozycyjnymi. Statystyką pozycyjną rzędu $j$, $n$ elementowej próby ($X_{nj}$), nazywamy najmniejszy $j$-ty element próby. Funkcja gęstości tej statystyki (gdy $X$ jest zmienną ciągłą) dana jest wzorem (np. Evans, Leemis, Drew [2006, 20]):

$$f_{X_{nj}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)]^{j-1}[1-F(x)]^{n-j}. \tag{3}$$

Ze wzoru tego nie można skorzystać, jeśli nie jest znany rozkład, z którego pobrano próbę. Można natomiast zastosować metodę bootstrapową.

Jeśli do estymacji kwantyli wykorzystywane są statystyki pozycyjne, pojawia się pytanie o wartość $j$ – czyli o rząd statystyk jaki należy wybrać. W przypadku, gdy $np$ nie jest całkowite wybór jest jednoznaczny - zwykle przyjmowane jest $j = [np]+1$. Dla całkowitego $np$ natomiast możliwości jest wiele. Ogólny zapis podają Hyndman i Fan [1996, 361]:

$$\hat{\xi}_{pn} = (1-\gamma)X_{nj} + \gamma X_{n,j+1}, \tag{4}$$

gdzie $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$ dla $m \in \mathrm{R}$ i $0 \leq \gamma \leq 1$. Parametr $\gamma$ jest funkcją $j$ i $g$, gdzie $j = [pn+m]$ a $g = pn + m - j$.

W niniejszej pracy rozważania ograniczone zostaną do estymatorów kwantyli w postaci pojedynczych statystyk pozycyjnych.

Badania przedstawione w artykule dotyczą estymacji kwartyli (pierwszego, drugiego i trzeciego) w sytuacji, gdy brak jest informacji o rozkładzie, z którego wylosowana została próba, zaś iloczyn $np$ ma wartość całkowitą. Celem ich jest porównanie dwóch estymatorów kwantyli w postaci statystyk pozycyjnych rzędu $np$ i $np + 1$ dla różnych rodzajów asymetrii rozkładów, z których losowana jest próba. Estymatory porównywane będą na podstawie wartości obciążenia i wariancji estymatora oraz szerokości przedziałów ufności i zliczeniowego poziomu ufności.

Jako metodę estymacji wybrano metodę bootstrapową w wersji dokładnej, której zastosowanie w przypadku kwantyli jest prostsze, niż w wersji klasycznej. Badania symulacyjne prowadzono metodą Monte Carlo. Obliczenia wykonane zostały w Excelu z wykorzystaniem języka VBA for Application.

## METODA BADAWCZA

Zakładamy, że dana jest próba pierwotna $(x_1,\ldots,x_n)$ wylosowana z nieznanego rozkładu $F$. Z próby tej losowanych jest ze zwracaniem $B$ prób zwanych wtórnymi,

które oznaczone są przez $\left(X_1^*,...,X_n^*\right)$. Zmienne $X_i^*$ mają jednakowy rozkład - rozkład empiryczny. Efron [1979] przyjmował, że wylosowanie każdego elementu próby pierwotnej jest jednakowe i równe $1/n$. Jednak, ze względu na skończoną dokładność pomiarów, w próbie losowej elementy mogą się powtarzać mimo, że $X$ jest zmienną ciągłą. W takim przypadku rozkład empiryczny dany jest prawdopodobieństwami $p_i$ wylosowania $x_i$, dla $i=1,\ldots,k.$, gdzie $k$ jest liczbą różnych elementów próby. Suma wszystkich prawdopodobieństw $p_i$ musi być równa 1.

Na podstawie każdej wylosowanej próby wtórnej wyznaczana jest wartość estymatora. W ten sposób otrzymywane jest $B$ jego realizacji. Realizacje te określają bootstrapowy rozkład estymatora.

Liczba wszystkich możliwych do wylosowania prób wtórnych $N$ jest równa liczbie wariancji z powtórzeniami, czyli $n^n$, przy czym liczba prób różnych jest mniejsza i wynosi $\binom{2n-1}{n}$. Każda unikalna próba wtórna może być następnie permutowana, ale jedynie na pozycjach o nie powtarzających się elementach. Metoda wykorzystująca wszystkie próby wtórne zwana jest w literaturze dokładną metodą bootstrapową (exact bootstrap method). Na możliwość jej stosowania zwrócili uwagę Fisher i Hall [1991], a algorytm pozwalający na jej użycie przedstawiony został między innym w pracy Kisielińska [2014].

W przypadku, gdy estymatorem jest pojedyncza statystyka pozycyjna zastosowanie metody dokładnej jest prostsze niż klasycznej (z losowaniem), ponieważ znany jest rozkład dowolnej statystyki pozycyjnej określony na podstawie wszystkich prób wtórnych.

Bootstrapowy rozkład $j$-tej statystyki pozycyjnej dany jest formułą [Evans i in. 2006, 23] (przypadek losowania ze zwracaniem z populacji o rozkładzie dyskretnym):

$$P\left(X_{nj}^* = x_l\right) = \begin{cases} \text{dla } l = 1 \\ \sum_{w=0}^{n-j} \binom{n}{w} \left[f_n(x_1)\right]^{n-w} \left[S_n(x_2)\right]^w \\ \text{dla } l = 2,3,\ldots,k-1 \\ \sum_{u=0}^{j-1} \sum_{w=0}^{n-j} \binom{n}{u, n-u-w, w} \left[F_n(x_{l-1})\right]^u \left[f_n(x_l)\right]^{n-u-w} \left[S_n(x_{l+1})\right]^w \\ \text{dla } l = k \\ \sum_{u=0}^{j-1} \binom{n}{u} \left[F_n(x_{k-1})\right]^u \left[f_n(x_k)\right]^{n-u} \end{cases}, \tag{5}$$

gdzie: $f_n(x) = P\left(X_i^* = x\right)$, $F_n(x) = P(X_i^* \le x)$, $S_n(x) = P(X_i^* \ge x)$.

Realizacjami estymatora w postaci pojedynczej statystyki pozycyjnej mogą być jedynie elementy próby pierwotnej. Ponadto prawdopodobieństwa określone wzorem (5), w przypadku, gdy w próbie nie było powtórzeń, mają bardzo przydatną

własność – zależą jedynie od wielkości próby, a nie od jej elementów. Pozwala to na stworzenie gotowych tablic dla prób nawet bardzo dużych. Mając obliczone prawdopodobieństwa poszczególnych realizacji bez trudu można wyznaczyć wartość oczekiwaną i wariancję estymatora[1]:

$$E(X_{nj}^*) = \sum_{l=1}^{k} x_l \cdot P\left(X_{nj}^* = x_l\right), \tag{6}$$

oraz:

$$V(X_{nj}^*) = \sum_{l=1}^{k} \left(x_l - E(X_{nj}^*)\right)^2 \cdot P\left(X_{nj}^* = x_l\right). \tag{7}$$

To, że prawdopodobieństwa (5) zależą jedynie od $n$ oznacza również, że z góry wiadomo, które elementy próby pierwotnej stanowią granice przedziałów ufności. Dla zadanego poziomu ufności 1-$\alpha$, lewą granicę przedziału stanowi $x_d$, takie że $d = \inf\left\{m : \sum_{l=1}^{m} P(X_{nj}^* = x_l) \geq \alpha/2\right\}$, prawą zaś $x_g$, takie że $g = n - m = \inf\left\{m : \sum_{l=0}^{m} P(X_{nj}^* = x_{n-l}) \geq \alpha/2\right\}$.

Granice przedziałów ufności w przypadku estymatorów bootstrapowych wyznacza się metodą percentyli. W metodzie tej (opisanej np. w pracy Wilcox [2001, 88]), $B$ wartości estymatora wyznaczonych na podstawie $B$ prób wtórnych porządkujemy i wyznaczamy realizacje o numerach $[\alpha/2 \cdot B]$ oraz $B$-$[\alpha/2 \cdot B]$+1.

Metodę wyznaczania granic przedziałów ufności bazującą na rozkładzie dokładnego estymatora bootstrapowego (5) nazwać można dokładną metodą percentyli przez analogię do dokładnej metody bootstrapowej (metoda ta zastosowana do estymacji mediany została przedstawiona w pracy Kisielińska [2016, 418]).

Ponieważ w przypadku bootstrapowych estymatorów w postaci pojedynczych statystyk pozycyjnych z góry wiadomo, które elementy próby stanowią granicę przedziałów ufności, metoda w wersji dokładnej jest wygodniejsza niż w wersji klasycznej (nie wymaga sortowania).

Dodać jeszcze należy, że jeżeli w próbie pierwotnej występują powtórzenia można również bez problemu zastosować metodę dokładną - wystarczy wówczas prawdopodobieństwa dla powtarzających się realizacji zsumować.

[1] Znajomość rozkładu nie jest konieczna do wyznaczenia wartości oczekiwanej i wariancji. Wzory pozwalające miary te oliczyć dla dowolnego L-estymatora podają Hudson i Ernst [2000, 91].

## ORGANIZACJA BADAŃ

Do estymacji kwartyli – pierwszego $p$=0.25, drugiego $p$=0.5 (mediany) i trzeciego $p$=0.75, zastosowano dwa rodzaje bootstrapowych estymatorów: $E_1 = X^*_{n,np}$ oraz $E_2 = X^*_{n,np+1}$. Estymatory te porównane zostały ze względu na obciążenie, wariancję oraz szerokość przedziałów ufności i zliczeniowy poziom ufności. Szacowanie tych miar przeprowadzono metodą symulacji Monte Carlo (MC). W każdym eksperymencie losowano po $R$ prób losowych o zadanej liczebności z populacji o wybranych rozkładach. Ponieważ w eksperymentach symulacyjnych wiadomo z jakiego rozkładu losowana jest próba, bez trudu można oszacować obciążenie estymatora. Oszacowanie obciążenia i wariancji metodą MC jest następujące:

$$\begin{aligned} \hat{bias}_{MC} &= \frac{1}{R}\sum_{r=1}^{R} E^r(X^*_{nj}) - \xi_p \\ \hat{V}_{MC} &= \frac{1}{R}\sum_{r=1}^{R} V^r(X^*_{nj}) \end{aligned}, \tag{7}$$

gdzie $E^r(X^*_{nj})$ jest wartością oczekiwaną estymatora uzyskaną w $r$-tej replikacji, zaś $V^r(X^*_{nj})$ wariancją.

Podobnie szacowana jest średnia szerokość przedziałów ufności:

$$\hat{d}_{MC} = \frac{1}{R}\sum_{r=1}^{R}\left(x_g^r - x_d^r\right), \tag{8}$$

gdzie $x_g^r$ i $x_d^r$ są granicami przedziału ufności uzyskanego w $r$-tej replikacji.

Zliczeniowy poziom ufności wyznaczany jest na postawie wszystkich $R$ replikacji jako:

$$\varphi_R = \frac{\#\{[x_d^r, x_g^r] : \xi_p \in [x_d^r, x_g^r]\}}{R}. \tag{9}$$

W eksperymentach symulacyjnych przedstawionych w niniejszym artykule, losowano po $R$=2000 prób o liczebnościach od 20 do 60 ze wzrostem po 4 elementy. Liczebności dobrano tak, aby spełniony był warunek całkowitej wartości $np$ dla wszystkich kwartyli.

Generatory liczb pseudolosowych zwracają liczby losowe $\varphi_i$ z przedziału [0,1], które były traktowane jako dystrybuanta. Wówczas $i$-ty element próby można wyznaczyć jako $x_i = F^{-1}(\varphi_i)$. Dla zapewnienia porównywalności wyników przyjęto jednakowe wartości dystrybuant dla obydwu estymatorów i wszystkich rozkładów, co uniezależnia wyniki obliczeń dla poszczególnych przypadków od jakości generatora.

Próby losowano z sześciu rozkładów: dwóch o asymetrii prawostronnej: (LogNorm(1,0.75) i Gamma(2,2)), dwóch o asymetrii lewostronnej (-LogNorm(1,0.6)+5 i Gamma(1.25,2.5)+5) oraz dwóch symetrycznych (N(3,0.5) i N(3,2)).

## WYNIKI BADAŃ

W celu oszacowania miar (7) i (8) wyznaczono rozkłady estymatorów E1 i E2 dla $p = 0{,}25$, $p = 0{,}5$ i $p = 0{,}75$ i wybranych liczebności prób. W celu ilustracji właściwości rozkładów tych estymatorów sporządzono tabele 1 i 2.

Tabela 1. Prawdopodobieństwa poszczególnych realizacji estymatorów E1 i E2 dla przypadku $p = 0{,}25$ i $n = 20$

| *i* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| E1 | 0,0026 | 0,0406 | 0,1270 | 0,2002 | 0,2148 | 0,1773 | 0,1193 | 0,0672 | 0,0321 | 0,0130 |
| E2 | 0,0003 | 0,0109 | 0,0561 | 0,1285 | 0,1870 | 0,2008 | 0,1710 | 0,1198 | 0,0703 | 0,0346 |
| *i* | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| E1 | 4,4E-03 | 1,2E-03 | 2,7E-04 | 4,4E-05 | 5,2E-06 | 3,7E-07 | 1,4E-08 | 1,7E-10 | 3,3E-13 | 6,1E-18 |
| E2 | 1,4E-02 | 4,8E-03 | 1,3E-03 | 2,7E-04 | 3,9E-05 | 3,6E-06 | 1,8E-07 | 3,2E-09 | 9,5E-12 | 3,7E-16 |

Uwagi: *i* oznacza numer określający element uporządkowanej próby pierwotnej

Źródło: opracowanie własne

Tabela 2. Numery elementów próby pierwotnej stanowiące granice przedziałów w ufności dla $1\text{-}\alpha = 0{,}95$, wyznaczone dokładną metodą percentyli

| *n* | *p* = 0,25 | | | | *p* = 0,5 | | | | *p* = 0,75 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | E1 | | E2 | | E1 | | E2 | | E1 | | E2 | |
| | *xd* | *xg* | *xd* | *xg* | *xd* | *xg* | *xd* | *xg* | *xd* | *xg* | *xd* | *xg* |
| 20 | 1 | 10 | 2 | 11 | 5 | 15 | 6 | 16 | 10 | 19 | 11 | 20 |
| 24 | 2 | 12 | 3 | 13 | 6 | 18 | 7 | 19 | 12 | 22 | 13 | 23 |
| 28 | 2 | 13 | 3 | 14 | 8 | 20 | 9 | 21 | 15 | 26 | 16 | 27 |
| 32 | 3 | 14 | 4 | 15 | 10 | 22 | 11 | 23 | 18 | 29 | 19 | 30 |
| 36 | 4 | 16 | 5 | 17 | 11 | 25 | 12 | 26 | 20 | 32 | 21 | 33 |
| 40 | 5 | 17 | 5 | 18 | 13 | 27 | 14 | 28 | 23 | 36 | 24 | 36 |
| 44 | 5 | 18 | 6 | 19 | 15 | 29 | 16 | 30 | 26 | 39 | 27 | 40 |
| 48 | 6 | 19 | 7 | 21 | 16 | 32 | 17 | 33 | 28 | 42 | 30 | 43 |
| 52 | 7 | 21 | 8 | 22 | 18 | 34 | 19 | 35 | 31 | 45 | 32 | 46 |
| 56 | 8 | 22 | 8 | 23 | 20 | 36 | 21 | 37 | 34 | 49 | 35 | 49 |
| 60 | 8 | 23 | 9 | 24 | 22 | 38 | 23 | 39 | 37 | 52 | 38 | 53 |

Uwagi: Podane w tabeli wartości to kolejne element uporządkowanej próby pierwotnej stanowiące granice przedziałów ufności

Źródło: opracowanie własne

W tabeli 1 przedstawiono prawdopodobieństwa poszczególnych realizacji obydwu bootstrapowych estymatorów kwartyla pierwszego (którymi są kolejne elementy uporządkowanej próby pierwotnej) dla prób o liczebności $n = 20$. Przypomnieć należy, że wartości te są jednakowe dla wszystkich dwudziesto-elementowych prób i nie zależą od tego, z którego rozkładu próba została wylosowana. Rozkład ten ma jedynie wpływ na wartości poszczególnych realizacji – czyli elementy próby pierwotnej. Dzięki tym własnościom z góry wiadomo, które elementy próby pierwotnej są granicami przedziałów ufności. Numery te dla dwóch estymatorów kwartyli przedstawiono w tabeli 2 dla $n$ od 20 do 60 z przyrostem po 4 elementy. Wyniki te potwierdzają stwierdzenie, że w przypadku estymacji kwantyli, dokładna metoda percentyli jest znacznie prostsza do zastosowania niż klasyczna (z losowaniem prób). Dodatkowo jest z pewnością dokładniejsza – analizy prowadzone na całej populacji prób wtórnych są pewniejsze niż prowadzone jedynie na wylosowanej z niej próbie, nawet badzo licznej.

Oszacowane obciążenie estymatora E1 zastosowanego do estymacji trzech kwartyli było w przypadku wszystkich rozkładów i niemal wszystkich liczebności prób ujemne, zaś estymatora E2 dodatnie, czego należało się spodziewać.

W tabeli 3 przedstawiono porównanie oszacowanych miar, uzyskanych bootstrapowymi estymatorami kwartyli - E1 i E2 oraz wybranych sześciu rozkładów. Przedziały szacowano przyjmując dla wszystkich kwartyli poziom ufności równy 0.95.

Na przewagę jednego estymatora nad drugim wskazuje niższe co do wartości bezwzględnej oszacowane obciążenie, mniejsza oszacowana wariancja i węższe oszacowane przedziały ufności. W przypadku zliczeniowego przedziału ufności przyjęto, że korzystniejsza jest wyższa jego wartość, choć nie jest to do końca stwierdzenie poprawne. Idealna sytuacja ma miejsce gdy zliczeniowy poziom ufności jest równy założonemu. Ze względu na dyskretny charakter rozkładów estymatorów bootstrapowych przeciwdziedzina dystrybuanty nie zawiera wszystkich wartości z przedziału [0,1]. Dlatego wyznaczone przedziały są zwykle nieco szersze niż wynikałoby to z przyjętego poziomu ufności, czego konsekwencją może być zawyżony zliczeniowy poziom ufności.

Dla wszystkich kwartyli, ze względu na wartość bezwzględną obciążenia oraz wariancję w przypadku prób z rozkładów o asymetrii prawostronnej, przewagę ma estymator E1 (wyjątek stanowi obciążenie estymatora kwartyla trzeciego dla dwóch liczebności prób). W przypadku prób z rozkładów o asymetrii lewostronnej zdecydowaną przewagę wykazywał (bez wyjątków) estymator E2. Jeśli próby losowano z rozkładów symetrycznych wariancję miał zwykle mniejszą estymator E2, zaś wartość bezwzględną obciążenia dla kwartyli pierwszego i drugiego zwykle estymator E1, a dla trzeciego E2. Dla przpadku rozkładów symetrycznych uznać można, że są obydwa estymatory niemal równoważne.

Jeśli chodzi o szerokość oszacowanych przedziałów ufności to oceny są podobne do tych sformułowanych dla wariancji (choć występują nieliczne wyjątki dla niektórych liczebności prób).

W przypadku zliczeniowego poziomu ufności, dla rozkładów asymetrycznych oceny są przeciwne niż dokonane na podstawie szerokości przedziałów ufności. Dla rozkładów symetrycznych i kwartyla pierwszego oraz drugiego przewagę wykazuje estymator E2 zarówno ze względu na szerokość przedziałów jak i zliczeniowy poziom ufności. W przypadku rozkładów symetrycznych i kwartyla trzeciego węższe przedziały pozwala budować zwykle estymator E1, czemu jednak towarzyszy niższy zliczeniowy poziom ufności.

Tabela 3. Porównanie estymatorów E1 i E2 zastosowanych do estymacji kwartyli, ze względu na obciążenie, wariancję, szerokość przedziałów ufności oraz zliczeniowy poziom ufności

| $p = 0{,}25$ | LogNorm (1,0.75) | Gamma (2,2) | -LogNorm (1,0.6)+5 | -Gamma (1.25,2.5)+5 | N(3,0.5) | N(3,2) |
|---|---|---|---|---|---|---|
| obciążenie | E1 | E1 | E2 | E2 | E1/E2 | E1/E2 |
| wariancja | E1 | E1 | E2 | E2 | E2 | E2 |
| szer. przedz. ufn. | E1/E2 | E1/E2 | E2/E1 | E2/E1 | E2/E1 | E2/E1 |
| zlicz. poziom ufn. | E2/E1 | E2/E1 | E1/E2 | E1/E2 | E2/E1 | E2/E1 |

| $p = 0{,}5$ | LogNorm (1,0.75) | Gamma (2,2) | -LogNorm (1,0.6)+5 | -Gamma (1.25,2.5)+5 | N(3,0.5) | N(3,2) |
|---|---|---|---|---|---|---|
| obciążenie | E1 | E1 | E2 | E2 | E1/E2 | E1/E2 |
| wariancja | E1 | E1 | E2 | E2 | E2/E1 | E2/E1 |
| szer. przedz. ufn. | E1 | E1 | E2 | E2 | E2/E1 | E2/E1 |
| zlicz. poziom ufn. | E2/E1 | E2/E1 | E2/E1 | E2/E1 | E2/E1 | E2/E1 |

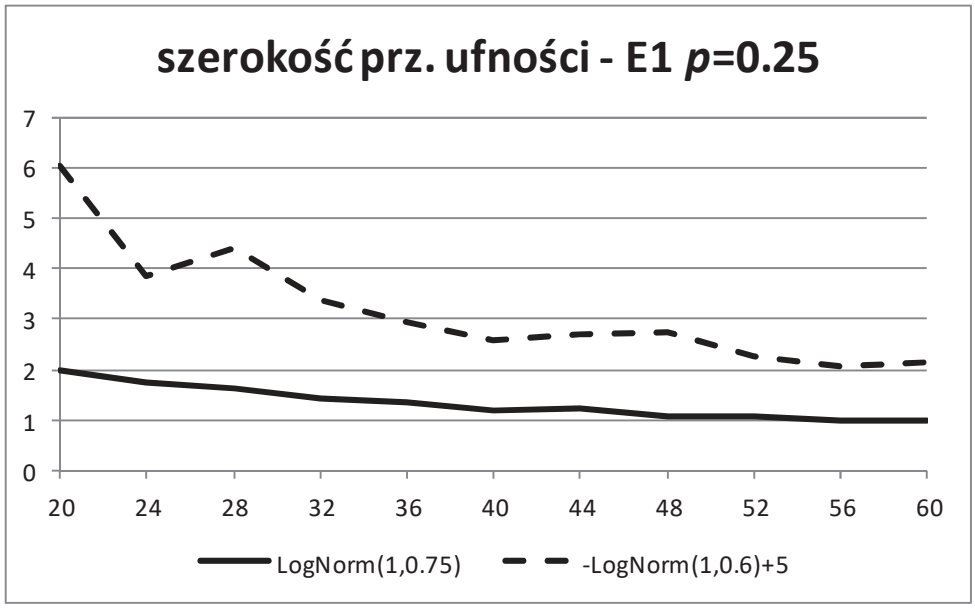
| $p = 0{,}75$ | LogNorm (1,0.75) | Gamma (2,2) | -LogNorm (1,0.6)+5 | -Gamma (1.25,2.5)+5 | N(3,0.5) | N(3,2) |
|---|---|---|---|---|---|---|
| obciążenie | E1 | E1/E2 | E2 | E2 | E2/E1 | E2/E1 |
| wariancja | E1 | E1 | E2 | E2 | E1 | E1 |
| szer. przedz. ufn. | E1/E2 | E1/E2 | E1/E2 | E2 | E1/E2 | E1/E2 |
| zlicz. poziom ufn. | E2/E1 | E2/E1 | E2/E1 | E2/E1 | E2/E1 | E2/E1 |

Uwagi: Zapis E1 lub E2 oznacza, że miara dla danego estymatora dla wszystkich liczebności miała wartość korzystniejszą. Zapis E1/E2 oznacza, że miara ta dla większej liczby przypadków była korzystniejsza dla estymatora E1 niż E2, zapis E2/E1 odwrotnie.

Źródło: opracowanie własne

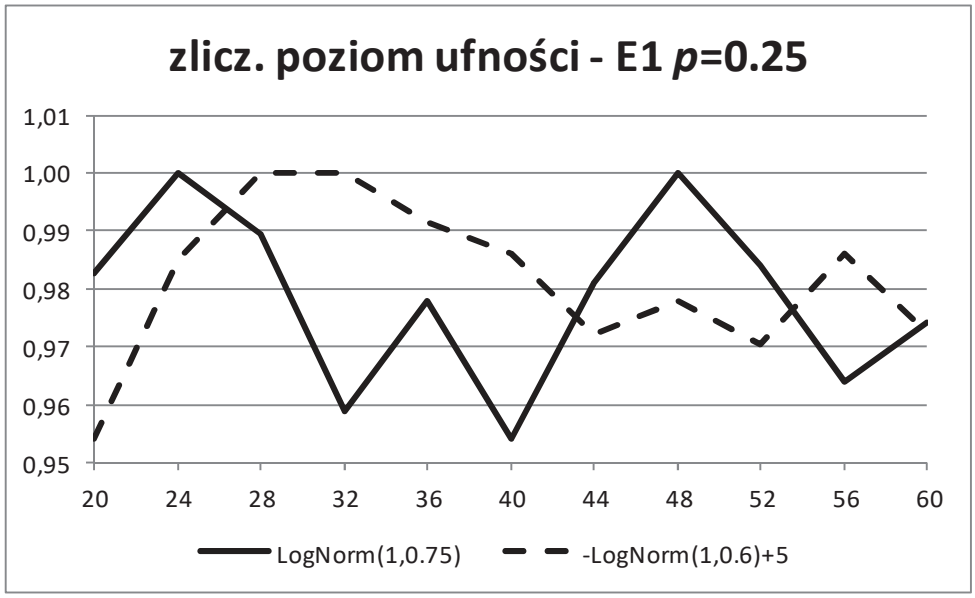
Jak wspomniano wcześniej, eksperymenty symulacyjne prowadzono dla różnych liczebności prób. Wzrostowi liczebności próby towarzyszy zwykle zmniejszanie się oszacowanych miar: obciążenia, wariancji oraz szerokości przedziałów ufności. Na rysunku 1 przedstawiono przykładowy wpływ zmian liczebności próby na szerokość 0.95 przedziałów ufności dla kwartyla pierwszego, oszacowanych za pomocą estymatora E1 dla dwóch rozkładów – jednego o asymetrii prawostronnej i drugiego o asymetrii lewostronnej. Dla pozostałych estymatorów, rozkładów i kwartyli wykresy były bardzo podobne. Podkreślić należy, że w niektórych przypadkach wzrostowi liczebności próby o 4 może towarzyszyć wzrost wykorzystanych w badaniach miar, co wynika z jakości użytego generatora liczb pseudolosowych. Jeszcze silniejszy jest wpływ jakości generatora na zliczeniowy poziom ufności. Przy zmianach liczebności prób występują bardzo silne fluktuacje tej miary, co pokazano na rysunku 2. Aby sprawdzić, czy zwiększenie liczby losowanych prób zmniejszy obserwowane fluktuacje, kilkakrotnie losowano po 4 000, 8 000 i 16 000 prób. Okazało się, że różnice w oszacowanych wartościach zliczeniowego poziomu ufności w poszczególnych eksperymentach były duże. W literaturze można znaleźć przykłady porównań różnych generatorów liczb pseudolosowych (np. [Sulewski 2019, Kozioł i Zieliński]).

Rysunek 1. Wpływ zmian liczebności próby na szerokość 0,95 przedziałów ufności kwartyla pierwszego wyznaczonych za pomocą estymatora E1




Źródło: opracowanie własne

Rysunek 2. Wpływ zmian liczebności próby na zliczeniowy poziom ufności 0,95 przedziałów ufności kwartyla pierwszego, wyznaczonych za pomocą estymatora E1




Źródło: opracowanie własne

## PODSUMOWANIE

Przeprowadzone badania dotyczyły estymacji trzech kwartyli metodą dokładnego bootstrapu dla przypadku, gdy iloczyn *np* jest całkowity. Pokazały one, że w przypadku gdy próba losowana jest z rozkładu o asymetrii prawostronnej lepsze rezultaty (ocena na podstawie obciążenia, wariancji i szerokości przedziałów ufności) uzyskuje się stosując estymator E1, czyli statystykę pozycyjną rzędu *np*. Jeśli próba pochodzi z rozkładu o asymetrii lewostronnej lepsze wyniki daje zastosowanie estymatora E2, czyli statystyki pozycyjnej rzędu *np*+1, w przypadku rozkładów symetrycznych estymatory są niemal równoważne. W literaturze przedmiotu najczęściej wykorzystywany jest estymator E1, co można tłumaczyć faktem, że częściej rozważane są rozkłady o asymetrii prawostronnej. Jednak zastosowania praktyczne mogą często dotyczyć cech o rozkładach z asymetrią lewostronną, co stanowi potwierdzenie celowości przeprowadzonych badań.

Metodę bootstrapowa stosuje się, jeśli brak jest informacji o rozkładzie, z którego wylosowano próbę. Jednak z próby można obliczyć współczynnik asymetrii, co stanowić może wskazówkę co do rodzaju asymetrii rozkładu.

W artykule pokazano, że w przypadku estymacji kwantyli zastosowanie dokładnej metody bootstrapowej oraz dokładnej metody percentyli jest znacznie prostsze, niż użycie metod tych w wersji klasycznej. Zwrócono uwagę na wpływ jakości generatorów liczb pseudolosowych na wyniki estymacji. Wpływ ten jest

szczególnie wyraźny dla zliczeniowego poziomu istotności. W przeprowadzonych eksperymentach symulacyjnych metodą Monte Carlo, nawet jeśli losowano wiele prób, zliczeniowy poziom ufności znacznie różni się dla każdego uruchomienia procedury losowania prób. Miarę to należy więc traktować jedynie orientacyjnie.

## BIBLIOGRAFIA

Bahadur R.R. (1966) A Note on Quantiles in Large Samples. The Annals of Mathematical Statistics, 37(3), 577-580.

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.

Hyndman R. J., Fan Y. (1996) Sample Quantiles in Statistical Packages. The American Statistician, 50(4), 361-365.

Kisielińska J. (2014) Szacowanie mediany przy użyciu dokładnej metody bootstrapowej. Metody Ilościowe w Badaniach Ekonomicznych, XV(3), 232-242.

Kisielińska J. (2016) Rozkłady wybranych bootstrapowych estymatorów mediany oraz zastosowanie dokładnej metody percentyli do jej przedziałowego szacowania. Przegląd Statystyczny, 63(4), 411-429.

Kozioł D., Zieliński W. A Comparison of Random Number Generators (publikacja internetowa: wojtek.zielinski.statystyka.info [dostęp: 1.10. 2020]

Sulewski P. (2019) Porównanie generatorów liczb pseudolosowych. Wiadomości Statystyczne, 64, 5-31.

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

## THE INFLUENCE OF THE ASYMMETRY OF THE DISTRIBUTION ON THE SELECTION THE BOOTSTRAP QUARTILE ESTIMATOR

**Abstract:** The research concerns the estimation of quartiles (first, second and third) in a situation where there is no information about the distribution from which the sample was drawn, and the product np is integer (where n is the sample size and p is the quantile order). If np is not an integer the order statistic of order [*np*] + 1 is chosen as the *p*-quantile estimator. If np is not integer, there are many more solutions. In this paper, two exact bootstrap quartile estimators in the form of single order statistics of *np* i *np* +1 order, were compared. The estimator's bias and variance as well as the width of the confidence intervals and the coverage probability were used for the assessment. Confidence

intervals were determined with the exact percentile method. The samples were drawn from the distributions with right and left asymmetry as well as symmetrical, which made it posiible to select the most appropriate estimator in a given situation.

**Keywords:** quartile estimation, bootstrap estimator, exact bootstrap method, exact percentile method

**JEL classification:** C13, C14