PRZEGLĄD STATYSTYCZNY
R. LVIII – ZESZYT 1-2 – 2011

JOANNA KISIELIŃSKA

# DOKŁADNA METODA BOOTSTRAPOWA I JEJ ZASTOSOWANIE DO ESTYMACJI WARIANCJI

## 1. WPROWADZENIE

Dana jest zmienna losowa $X$ o nieznanym rozkładzie $F$. Interesuje nas parametr rozkładu, który oznaczymy jako $\theta$. Jeśli parametru nie można wyznaczyć bezpośrednio, konieczne jest pobranie próby losowej oraz dobranie odpowiedniego estymatora parametru $\theta$. Estymator jest statystyką określoną na przestrzeni prób. Próbę losową oznaczymy jako $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, jej realizację jako $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, zaś estymator parametru $\theta$ jako $\hat{\theta} = t(\mathbf{X})$.

Efron [4] zaproponował metodę, którą nazwał bootstrap polegającą na losowaniu z uzyskanej próby (próby pierwotnej) $\mathbf{x}$, kolejnych prób wtórnych (prób bootstrapowych) o liczebności $n$. Losowanie odbywa się ze zwracaniem przy założeniu jednakowych prawdopodobieństw równych $1/n$ wylosowania, każdej z wartości $x_i$, dla $i$=1, …, $n$. Generowany jest w ten sposób rozkład $\hat{F}$, zwany rozkładem bootstrapowym z próby.

Próba bootstrapowa oznaczana jest jako $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_n^*)$, a dowolna jej realizacja jako $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$. Statystyka $\hat{\theta}$ dla próby bootstrapowej oznaczana jest jako $\hat{\theta}* = t(\mathbf{X}*)$.

Istotą metody jest aproksymacja rozkładu statystyki $\hat{\theta}$, rozkładem statystyki bootstrapowej $\hat{\theta}*$. Efron [4] zaproponował trzy metody wyznaczania rozkładu $\hat{\theta}*$:

1. przeprowadzenie właściwych rozważań teoretycznych,
2. zastosowanie aproksymacji Monte Carlo,
3. aproksymację rozkładu $\hat{\theta}*$ poprzez rozwinięcie funkcji $t$ w szereg Taylora.

Jeśli wybrane zostanie rozwiązanie 2 konieczne jest określenie liczby losowanych prób bootstrapowych $N$. Efron [5] opierając się na bootstrapowej wariancji stwierdza, że wystarczy niewielka liczba losowań, aby otrzymać wystarczającą dokładność. Z taką oceną nie zgadzają się Booth i Sarkar [1], którzy zastosowali aproksymację rozkładu względnej wariancji bootstrapowej rozkładem normalnym. Pozwoliło to na oszacowanie $N$ dla określonego poziomu błędu na założonym poziomie ufności. Okazało się, że uzyskanie błędu poniżej 10% przy poziomie ufności 0,95 wymaga przyjęcia $N$ około 800. Domański i Pruska [3] (str. 261) stwierdzają, że przyjmuje się $N \geqslant 1000$.

Prowadząc rozważania nad metodą bootstrapową zadać można sobie pytanie, czy losowanie próby bootstrapowej $\mathbf{X}^*$, z wcześniej już uzyskanej próby pierwotnej $\mathbf{X}$ jest konieczne? Losowanie próby jest niezbędne, jeśli zbadanie całej populacji nie jest

możliwe lub jest zbyt kosztowne. Posługiwanie się próbą zamiast populacją ma swoje istotne implikacje będące w obszarze zainteresowań statystyki matematycznej.

Zwróćmy uwagę, że podstawową własnością próby jest jej skończony rozmiar. Określony dla niej rozkład bootstrapowy $\hat{F}$ jest prostym rozkładem dyskretnym. Rozkład dowolniej statystyki określonej dla $n$ dyskretnych zmiennych losowych o skończonej liczbie realizacji, nie musi być szacowany, ponieważ może być po prostu wyznaczony. Kwestią otwartą pozostaje jedynie, jak dużego nakładu obliczeń wymaga takie podejście, o czym mowa będzie dalej.

Rozkład bootstrapowy jest w istocie rozkładem empirycznym określonym dystrybuantą[1)]:

$$F_n(t) = \frac{\#\ \{1 \leqslant i \leqslant n : X_i \leqslant t\}}{n} \text{ dla } t \in R^1 \tag{1}$$

gdzie $X_1, X_2, \ldots, X_n$ jest ciągiem ciągłych i niezależnych zmiennych losowych o jednakowych rozkładach określonych dystrybuantą $F$.

Z twierdzenia Gliwienki-Cantelliego (podstawowego twierdzenia statystyki matematycznej) wynika, że:

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| \tag{2}$$

dąży do 0 z prawdopodobieństwem 1.

Twierdzenie to upoważnia do wnioskowania statystycznego – formułowania stwierdzeń dotyczących populacji na podstawie próby.

## 2. DOKŁADNA METODA BOOTSTRAPOWA

Załóżmy, że z populacji opisanej zmienną losową $X$ wylosowano $n$ elementową próbę pierwotną $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Ponieważ dla niektórych $i \neq j$ może zachodzić $x_i = x_j$, zredukujmy[2)] rozmiar wylosowanej próby do $k$ różnych wartości. Prawdopodobieństwa $p_i$ wylosowania z próby realizacji $x_i$, gdzie $i$=1, 2,..., $k$ nie muszą być wówczas jednakowe (jak to ma miejsce dla próby bootstrapowej).

W próbie **x**, każda wartość $x_i$ jest realizacją zmiennej losowej $X$ o rozkładzie $F$. Zgodnie z twierdzeniem Gliwienki-Cantelliego, rozkład $F$ może być przybliżony rozkładem empirycznym $F_n$.

Wprowadźmy pojecie dyskretnej zmiennej losowej próby, którą oznaczyć można jako $X^D$. Zbiorem realizacji tej zmiennej jest pierwotna próba losowa $\{x_1, x_2, \ldots, x_k\}$. Poszczególne wartości przyjmowane są z prawdopodobieństwami $p_i^{\mathrm{D}} = P\left(X^{\mathrm{D}} = x_i\right) = p_i$[3)], dla $i$= 1, 2, ... , $k$. Rozkład ten jest równoważny rozkładowi empirycznemu

[1)] Zapis zgodnie z Zieliński [9] str. 11, gdzie # oznacza liczebność.

[2)] Redukcja ta nie jest konieczna, lecz wskazana, ponieważ pozwala na zmniejszenie wymiaru problemu.

[3)] Prawdopodobieństwa te byłyby jednakowe i równe $1/n$ gdyby nie przeprowadzono redukcji wymiaru próby do $k$ różnych wartości.

i oznaczony zostanie jako $F^D$. W takim przypadku $n$ elementową wtórną próbę losową można zapisać jako $\mathbf{X}^{\mathrm{D}} = \left(X_1^{\mathrm{D}}, X_2^{\mathrm{D}}, \cdots, X_n^{\mathrm{D}}\right)$, a dowolną jej realizację jako $\mathbf{x}^{\mathrm{D}} = \left(x_1^{\mathrm{D}}, x_2^{\mathrm{D}}, \cdots, x_n^{\mathrm{D}}\right)$. Zmienne $X_i^{\mathrm{D}}$, dla $i$= 1, 2, … , $n$ mają rozkłady $F^D$. Statystykę $\hat{\theta}*$ dla próby bootstrapowej oznaczyć można wówczas jako $\hat{\theta}^{\mathrm{D}} = t\left(\mathbf{X}^{\mathrm{D}}\right)$. Rozkład statystyki $\hat{\theta}$ przybliżać będziemy rozkładem statystyki $\hat{\theta}^{\mathrm{D}}$. Zauważmy, że problemy związane z ewentualnym obciążeniem, zgodnością i efektywnością estymatora dotyczą estymatora $\hat{\theta}$. W dokładnej metodzie bootstrapowej wartość estymatora $\hat{\theta}^{\mathrm{D}}$ jest obliczana, a nie szacowana na podstawie próby. Metoda nie wprowadza więc dodatkowego obciążenia.

Dla pojedynczej $b$-tej realizacji wtórnej próby $\mathbf{x}^{\mathrm{D}b} = \left(x_1^{\mathrm{D}b}, x_2^{\mathrm{D}b}, \cdots, x_n^{\mathrm{D}b}\right)$ należy obliczyć wartość $\hat{\theta}^{\mathrm{D}}(b) = t\left(\mathbf{x}^{\mathrm{D}b}\right) = t\left(x_1^{\mathrm{D}b}, x_2^{\mathrm{D}b}, \cdots, x_n^{\mathrm{D}b}\right)$ oraz prawdopodobieństwo jej wylosowania. Ze względu na redukcję wymiaru $n$ elementowej próby do $k$ różnych wartości, prawdopodobieństwa wylosowania poszczególnych prób bootstrapowych nie są już jednakowe. Prawdopodobieństwo wylosowania $b$-tej próby jest równe $p^{Db} = P\left(\hat{\theta}^{\mathrm{D}} = \hat{\theta}^{\mathrm{D}}(b)\right) = \prod_{i=1}^{n} p_i^{\mathrm{D}b}$, gdzie $p_i^{\mathrm{D}b} = P\left(X^{\mathrm{D}} = x_i^{\mathrm{D}b}\right)$. Liczba możliwych realizacji wtórnej próby jest równa $B = k^n$. Poprawnie napisany algorytm powinien zapewnić spełnienie warunku: $\sum_{b=1}^{B} p^{\mathrm{D}b} = 1$.

Mając wyznaczone wartości $\hat{\theta}^{\mathrm{D}}(b)$ oraz prawdopodobieństwa $p^{Db}$, dla $b$=1, 2,. . . , $B$ można określić rozkład estymatora $\hat{\theta}^{\mathrm{D}}$. Estymator ten, będąc funkcją dyskretnych zmiennych losowych o jednakowym rozkładzie $F^D$, jest również zmienną losową dyskretną. Rozkład estymatora pozwala budować przedziały ufności, czy testować hipotezy statystyczne.

Jeżeli znany jest rozkład estymatora można wyznaczyć jego wartość oczekiwaną i odchylenie standardowe. Wartość oczekiwana jest bootstrapowym oszacowaniem parametru $\theta$, zaś odchylenie standardowe błędem tego szacunku. Wartość oczekiwana i odchylenie standardowe obliczyć należy tak jak dla zmiennej dyskretnej. Oznaczając ocenę parametru jako $\hat{\theta}^{D}(\cdot)$, a błąd jako $s_{\hat{\theta}^{\mathrm{D}}}^{\mathrm{D}}$ [4)] otrzymujemy:

$$\hat{\theta}^{\mathrm{D}}(\cdot) = \sum_{b=1}^{B} \hat{\theta}^{\mathrm{D}}(b) \cdot p^{\mathrm{D}b}, \quad s_{\hat{\theta}^{\mathrm{D}}}^{\mathrm{D}} = \sqrt{\sum_{b=1}^{B} \left(\hat{\theta}^{\mathrm{D}}(b) - \hat{\theta}^{\mathrm{D}}(\cdot)\right)^2 \cdot p^{\mathrm{D}b}} \tag{3}$$

Chcąc zastosować metodę dokładnego bootstrapu warto zastanowić się, jak ma się liczba wszystkich możliwych prób wtórnych $B$ wobec zalecanej liczby prób wtórnych w klasycznej metodzie bootstrapowej. Np. dla $k$=10 i $n$=20 otrzymujemy $B = 10^{20}$,

[4)] Formuła (3) obowiązuje gdy brane są pod uwagę wszystkie próby bootstrapowe. Jeśli rozważany jest jedynie ich podzbiór należałoby dokonać korekty o czynnik $\sqrt{\frac{n}{n-1}}$.

co jest liczbą bardzo dużą, znacznie większą niż przykładowe $N$=1000. Jak pokazane zostanie dalej tak dużą liczbę powtórzeń można aktualnie wygenerować. Pionierska praca Efrona pochodzi z roku 1979. W tamtym okresie wykonanie tej liczby obliczeń w sensownym czasie było niemożliwe. Ponieważ nie można było zbadać całej „populacji" określonej próbą pierwotną, konieczne było pobieranie z „próby traktowanej jako populację" kolejnych prób – prób wtórnych.

W metodzie bootstrapowej ciąg uzyskanych wartości estymatora porządkowany jest w kolejności od najmniejszego do największego, co pozwala np. wyznaczyć granice przedziałów ufności metodą percentyli (Efron, Tibshirani [6]). W dokładnej metodzie bootstrapowej teoretycznie można również tak postąpić. Liczba możliwych realizacji statystyki $\hat{\theta}^{\mathrm{D}}$ jest wprawdzie bardzo duża, ale po pierwsze, część realizacji z pewnością będzie się powtarzać, a po drugie i tak wskazane jest pogrupowanie rezultatów w histogramie. Utworzenie histogramu będzie dla dużych problemów niezbędne[5)], ale wiązać się będzie z utratą pewnej informacji. Podkreślić należy, że mimo tego dokładną metodą bootstrapową można uzyskać bardzo dokładne oszacowanie granic przedziałów ufności, ponieważ szerokość przedziałów w histogramie nie musi być jednakowa. W zakresach wymagających podania dokładnych prawdopodobieństw (czy dystrybuanty), szerokość przedziału może być bardzo, niemal dowolnie mała.

Najprostszym sposobem wygenerowania wszystkich wtórnych prób dla rozkładu dyskretnego będzie rekurencyjne pobieranie kolejnych elementów z próby. Jeśli wystąpiły powtórzenia, zbiór z którego pobierane są wartości redukuje się do wymiaru $k$. Tak skonstruowany algorytm zaliczyć można do kategorii brute force. Algorytmy tego typu są uznawane za nieefektywne.

Liczbę generowanych realizacji prób bootstrapowych można zredukować, ponieważ w próbie wtórnej niektóre wartości będą się powtarzać – losowanie odbywa się z powtórzeniami. Problem taki dla przypadku gdy prawdopodobieństwo wylosowania każdego elementu próby jest jednakowe i elementy są jednakowe, przedstawiony jest w książce Fellera [7] w rozdziale II.5 str. 38[6)]. Dalsze rozważania prowadzone będą dla przypadku, gdy prawdopodobieństwa te nie są jednakowe, a losowane elementy mogą być różne.

Każdą $n$ elementową wtórną $b$-tą próbę bootstrapową losowaną ze zbioru $k$ elementowego można zapisać jako:

$$\mathbf{x}^{\mathrm{D}b} = (a_{1b} \times x_1, a_{2b} \times x_2, \cdots, a_{kb} \times x_k) \tag{4}$$

gdzie każde $a_{ib} \geqslant 0$, dla $i$= 1, 2, …, $k$ jest liczbą wystąpień w próbie $b$-tej $i$-tego elementu próby pierwotnej. Liczby te muszą spełniać warunek:

$$\sum_{i=1}^{k} a_{ib} = n \tag{5}$$

[5)] Przykładowo liczba realizacji $10^{20}$ wymaga ponad 80 eksbibajtów komórek pamięci.

[6)] Jak zauważa Booth [2] w prezentacji elektronicznej: Monte Carlo and the boothstrap.

przy czym część z nich może być równa 0. Jeżeli dla wybranego $i$ zachodzi $a_{ib}$=0, oznacza to, że element $x_i$ nie wystąpił w $b$-tej próbie wtórnej.

Prawdopodobieństwo wylosowania pojedynczej próby określonej (4) jest równe:

$$p^{\mathrm{D}b} = \prod_{i=1}^{k} \left(p_i^{\mathrm{D}b}\right)^{a_{ib}} \tag{6}$$

Pozostaje jeszcze określenie ile prób można wygenerować dla danego ciągu liczb $a_{ib} \geqslant 0$, dla $i$= 1, 2, ..., $k$, spełniających warunek (5). Trzeba uwzględnić wszystkie $m_b$ elementowe kombinacje, gdzie $m_b = \#\{a_{ib} \neq 0 : i = 1, 2, ..., k\}$ ze zbioru $k$ elementowego. Kombinacje należy następnie permutować, ale jedynie na pozycjach, dla których współczynniki $a_{ib}$ są unikalne.

## 3. ROZKŁAD GRANICZNY ESTYMATORA WARIANCJI Z PRÓBY BOOTSTRAPOWEJ

Dokładna metoda bootstrapowa wykorzystana zostanie do szacowania wariancji pewnej zmiennej losowej o nieznanym rozkładzie. Weryfikację poprawności metody można przeprowadzić porównując rozkład estymatora wariancji z rozkładem granicznym, który stosować można gdy próba jest duża ($n \geqslant 30$).

Pobierając $n$ elementową próbę losową określamy dyskretną zmienną losową próby $X^D$ o zbiorze realizacji $\{x_1, x_2, \ldots, x_k\}$ i rozkładzie prawdopodobieństwa określonym wartościami $p_i = \mathrm{P}(X^D = x_i)$, dla $i$= 1,2, ... , $k$, przy czym $\sum_{i=1}^{k} p_i = 1$.

Wartość oczekiwana $\mu^D$, odchylenie standardowe $\sigma^{\mathrm{D}}$ oraz czwarty moment centralny $\mu_4^{\mathrm{D}}$ zmiennej $X^D$ są równe odpowiednio:

$$\mu^{\mathrm{D}} = \sum_{i=1}^{k} x_i \cdot p_i, \quad \sigma^{\mathrm{D}} = \sqrt{\sum_{i=1}^{k} (x_i - \mu^{\mathrm{D}})^2 \cdot p_i}, \quad \mu_4^{\mathrm{D}} = \sum_{i=1}^{k} \left(x_i - \mu^{\mathrm{D}}\right)^4 \cdot p_i \tag{7}$$

Wiadomo, że rozkład wariancji z próby $S^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ pobranej z populacji o rozkładzie normalnym N($\mu$,$\sigma$) jest rozkładem asymptotycznie normalnym $\mathrm{N}\left(\frac{n-1}{n} \cdot \sigma^2, \sqrt{2 \cdot \frac{\sigma^4}{n}}\right)$.

Rozkład graniczny wariancji z próby $S^2$ pobranej z populacji o dowolnym rozkładzie z parametrami $\mu$ i $\sigma$ będzie też rozkładem normalnym (wariancja jest uśrednionym

kwadratem odchyleń od wartości przeciętnych). Wartość oczekiwana rozkładu granicznego oraz jego wariancja są określone wzorem[7)]:

$$\mathrm{E}S^2 = \frac{n-1}{n}\cdot\sigma^2, \quad \mathrm{D}^2S^2 = \frac{\mu_4-\sigma^4}{n} - \frac{2\cdot\left(\mu_4-2\cdot\sigma^4\right)}{n^2} + \frac{\mu_4-3\cdot\sigma^4}{n^3} \tag{8}$$

gdzie: $\mu_4$ jest momentem centralnym rzędu czwartego rozpatrywanego rozkładu.

Rozkłady, o których mowa powyżej są rozkładami granicznymi dla estymatora $S^2$, a nie estymatora $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$. Ponieważ $\hat{S}^2 = \frac{n}{n-1}S^2$, należy dokonać korekty parametrów rozkładów granicznych. Wartość oczekiwaną estymatora $S^2$ należy pomnożyć[8)] przez $\frac{n}{n-1}$, a wariancję przez $\left(\frac{n}{n-1}\right)^2$.

Rozkłady graniczne nieobciążonego estymatora wariancji oznaczone zostaną jako GV1 (rozkład graniczny nie wymagający normalności rozkładu zmiennej losowej w populacji) i GV2 (wymagający normalności rozkładu) są więc następujące:

$$\mathrm{GV1}: \mathrm{N}\left(\left(\sigma^{\mathrm{D}}\right)^2, \frac{n}{n-1}\cdot\sqrt{\frac{\mu_4^{\mathrm{D}}-\left(\sigma^{\mathrm{D}}\right)^4}{n} - \frac{2\cdot\left(\mu_4^{\mathrm{D}}-2\cdot\left(\sigma^{\mathrm{D}}\right)^4\right)}{n^2} + \frac{\mu_4^{\mathrm{D}}-3\cdot\left(\sigma^{\mathrm{D}}\right)^4}{n^3}}\right)$$

$$\mathrm{GV2}: \mathrm{N}\left(\left(\sigma^{\mathrm{D}}\right)^2, \frac{n}{n-1}\cdot\sqrt{\frac{2\cdot\left(\sigma^{\mathrm{D}}\right)^4}{n}}\right) \tag{9}$$

## 4. WYNIKI

Zakładamy, że dana jest zmienna losowa $X$ o nieznanym rozkładzie, którego wariancję chcemy oszacować. Pobieramy $n$ elementową próbę losową i zakładamy postać estymatora wariancji $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$. Przykładowy rozkład empiryczny próby przedstawiony jest w tabeli 1. Niech rozkład ten reprezentuje zmienna losowa $X^D$. Wartość oczekiwana i wariancja zmiennej $X^D$ są odpowiednio równe $\mu^D$= 5,174 oraz $\left(\sigma^{\mathrm{D}}\right)^2$= 3,989724. Rozkład estymatora wariancji zmiennej losowej $X$ przybliżamy rozkładem statystyki bootstrapowej. Pobranie próby losowej można interpretować jako dyskretyzację ciągłej zmiennej $X$. Stosując metodę bootstrapową rozkład estymatora pewnego parametru rozkładu zmiennej $X$, przybliżamy rozkładem estymatora dyskretnej zmiennej $X^D$ reprezentującej próbę. Różnica miedzy klasyczną metodą, a dokładną

[7)] Smirnow, Dunin-Barkowski [8] str. 237.

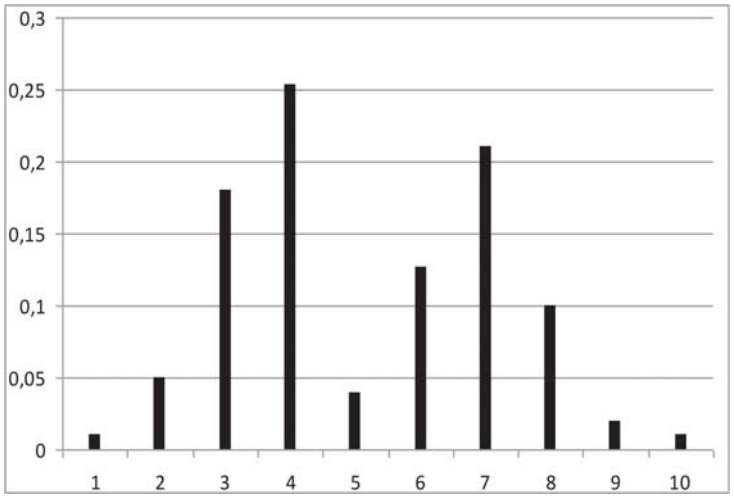
[8)] Jeżeli $n$ jest bardzo duże iloraz $\frac{n}{n-1}$ jest praktycznie równy jeden. Dla próby 30 elementowej (a dla tej liczebności próby można już stosować rozkłady graniczne) jest równy 1,0345, zaś jego kwadrat 1,0702 – o tyle więc należałoby skorygować wariancję.

polega jedynie na liczbie wziętych pod uwagę prób wtórnych. W metodzie klasycznej losowany jest pewien podzbiór możliwych prób wtórnych, w metodzie dokładnej zaś brane są pod uwagę wszystkie próby wtórne. W klasycznej metodzie bootstrapowej zamiast populacji prób wtórnych badamy więc jedynie pewien jej podzbiór. W dokładnej metodzie bootstrapowej bierzemy pod uwagę całą populację prób wtórnych. Mając do dyspozycji całą populację rozkład, bądź jego parametry nie muszą być szacowane, ponieważ mogą być obliczane – problem na tym etapie jest zagadnieniem z zakresu statystyki opisowej, a nie matematycznej.

Tabela 1

Rozkład zmiennej losowej $X^D$ reprezentującej próbę pierwotną

| $x_i$ | $p_i$ |
|---|---|
| 1 | 0,010 |
| 2 | 0,050 |
| 3 | 0,180 |
| 4 | 0,253 |
| 5 | 0,040 |
| 6 | 0,127 |
| 7 | 0,210 |
| 8 | 0,100 |
| 9 | 0,020 |
| 10 | 0,010 |




Źródło: Opracowanie własne

W tabeli 2 przedstawiono stablicowany rozkład estymatora wariancji zmiennej losowej $X^D$ (stanowiący przybliżenie rozkładu estymatora wariancji zmiennej ciągłej $X$), wyznaczony metodą dokładnego bootstrapu i oznaczony jako DBV, oraz dwa normalne rozkłady graniczne dla wariancji GV1 i GV2 określone wzorem (9).

Rozkłady podano w dwóch wariantach. W pierwszym założono $n$=20, w drugim $n$=30. Podkreślić należy, że w przypadku stosowania metod bootstrapowych wartość $n$ wynika bezpośrednio z liczebności próby. Założenie dwóch wartości dla $n$ traktować należy jedynie jako eksperyment symulacyjny mający na celu zbadanie ewentualnych podobieństw i różnic między faktycznym rozkładem estymatora a rozkładami granicznymi.

Pewnego komentarza wymaga jeszcze sposób doboru szerokości przedziału tablicowania. W ogólnym przypadku może być ona dowolnie mała. Dla rozkładów granicznych (ciągłych), dla każdego dowolnie małego przedziału istnieje większe od 0

Tabela 2

Rozkład estymatora wariancji zmiennej losowej $X^D$ uzyskany dokładną metodą bootstrapową oraz rozkłady graniczne

| | *n*=20 | | | *n*=30 | | |
|---|---|---|---|---|---|---|
| | DBV | GV1 | GV2 | DBV | GV1 | GV2 |
| **Średnia** | 3,98972 | 3,98972 | 3,98972 | 3,98972 | 3,98972 | 3,98972 |
| **Odchylenie st.** | 0,91790 | 0,91790 | 1,32806 | 0,73650 | 0,73650 | 1,06566 |
| **Przedziały** | DBV | GV1 | GV2 | DBV | GV1 | GV2 |
| **<0,00; 0,25)** | 2,77E-08 | 2,31E-05 | 0,002432 | 8,73E-12 | 1,91E-07 | 0,000225 |
| **<0,25; 0,50)** | 9,47E-07 | 4,87E-05 | 0,001867 | 1,52E-09 | 8,87E-07 | 0,000304 |
| **<0,50; 0,75)** | 5,73E-06 | 1,36E-04 | 0,003057 | 3,14E-08 | 4,36E-06 | 0,000654 |
| **<0,75; 1,00)** | 2,91E-05 | 0,000355 | 0,004832 | 3,66E-07 | 1,92E-05 | 0,001329 |
| **<1,00; 1,25)** | 9,40E-05 | 0,000856 | 0,007372 | 2,84E-06 | 7,50E-05 | 0,002560 |
| **<1,25; 1,50)** | 0,000361 | 0,001921 | 0,010858 | 2,03E-05 | 0,000262 | 0,004666 |
| **<1,50; 1,75)** | 0,001463 | 0,004003 | 0,015437 | 0,000110 | 0,000817 | 0,008051 |
| **<1,75; 2,00)** | 0,003616 | 0,007749 | 0,021185 | 0,000607 | 0,002272 | 0,013153 |
| **<2,00; 2,25)** | 0,009544 | 0,013933 | 0,028064 | 0,002482 | 0,005634 | 0,020342 |
| **<2,25; 2,50)** | 0,021507 | 0,023274 | 0,035886 | 0,008270 | 0,012467 | 0,029782 |
| **<2,50; 2,75)** | 0,037907 | 0,036112 | 0,044296 | 0,022408 | 0,024610 | 0,041280 |
| **<2,75; 3,00)** | 0,063318 | 0,052051 | 0,052778 | 0,045941 | 0,043341 | 0,054165 |
| **<3,00; 3,25)** | 0,073398 | 0,069692 | 0,060702 | 0,076109 | 0,068095 | 0,067284 |
| **<3,25; 3,50)** | 0,097465 | 0,086681 | 0,067391 | 0,110997 | 0,095448 | 0,079124 |
| **<3,50; 3,75)** | 0,121196 | 0,100148 | 0,072221 | 0,124357 | 0,119359 | 0,088088 |
| **<3,75; 4,00)** | 0,102042 | 0,107484 | 0,074710 | 0,139018 | 0,133161 | 0,092839 |
| **<4,00; 4,25)** | 0,102419 | 0,107159 | 0,074601 | 0,126058 | 0,132538 | 0,092630 |
| **<4,25; 4,50)** | 0,093798 | 0,099241 | 0,071907 | 0,107751 | 0,117691 | 0,087495 |
| **<4,50; 4,75)** | 0,074792 | 0,085377 | 0,066904 | 0,085236 | 0,093235 | 0,078239 |
| **<4,75; 5,00)** | 0,062212 | 0,068230 | 0,060088 | 0,059421 | 0,065896 | 0,066232 |
| **<5,00; 5,25)** | 0,040503 | 0,050651 | 0,052093 | 0,038703 | 0,041549 | 0,053079 |
| **<5,25; 5,50)** | 0,032387 | 0,034929 | 0,043594 | 0,024514 | 0,023373 | 0,040270 |
| **<5,50; 5,75)** | 0,024813 | 0,022375 | 0,035215 | 0,013139 | 0,011729 | 0,028923 |
| **<5,75; 6,00)** | 0,013530 | 0,013314 | 0,027459 | 0,007679 | 0,005251 | 0,019666 |
| **<6,00; 6,25)** | 0,009445 | 0,007359 | 0,020668 | 0,003776 | 0,002097 | 0,012659 |
| **<6,25; 6,50)** | 0,006109 | 0,003779 | 0,015017 | 0,001875 | 0,000747 | 0,007714 |
| **<6,50; 6,75)** | 0,003479 | 0,001802 | 0,010532 | 0,000887 | 0,000238 | 0,004450 |
| **<6,75; 7,00)** | 0,002140 | 0,000799 | 0,007130 | 0,000377 | 6,74E-05 | 0,002430 |
| **<7,00; 7,25)** | 0,001051 | 0,000329 | 0,004659 | 0,000159 | 1,7E-05 | 0,001257 |
| **<7,25; 7,50)** | 0,000650 | 1,26E-04 | 0,002939 | 6,56E-05 | 3,85E-06 | 0,000615 |
| **<7,50; +∞)** | 0,000724 | 4,46E-05 | 0,001790 | 3,83E-05 | 7,74E-07 | 0,000285 |

Źródło: Badania własne

prawdopodobieństwo, że zmienna losowa przyjmuje wartości z tego przedziału. Dla rozkładu dyskretnego, a takim jest bootstrapowy (zarówno klasyczny jak i dokładny) rozkład estymatora wariancji zmiennej $X^D$, tak nie jest. Im mniejsza szerokość, tym większej liczbie przedziałów będzie odpowiadało prawdopodobieństwo równe 0.

W tabeli 2 podano również wartości oczekiwane estymatorów wariancji zmiennej $X^D$ oraz ich odchylenia standardowe. Są to wartości dokładne – obliczone. Podane w tabeli rozkłady są tablicowane dla przedziałów. Obliczenie na ich podstawie parametrów rozkładów prowadzić należy jak dla danych pogrupowanych, wobec czego wyniki różnić się mogą od wartości dokładnych.

Wartości oczekiwane obydwu rozkładów granicznych GV1 i GV2 są z założenia równe wariancji próby. W przypadku rozkładu DBV również otrzymano wartość oczekiwaną równą wariancji próby, co potwierdza poprawność użytego algorytmu. Dokładna metoda bootstrapowa nie wprowadza dodatkowego obciążenia estymatora (w przeciwieństwie do metody bootstrapowej z losowaniem prób, gdzie obciążenie jest możliwe).

Na uwagę zasługuje fakt, że odchylenie standardowe rozkładu DBV jest równe z dokładnością do piątego miejsca po przecinku odchyleniu standardowemu rozkładu GV1. Potwierdza to poprawność algorytmu realizującego dokładny bootstrap. W przypadku rozkładu granicznego GV2 odchylenie jest zawyżone. Zawyżenie to wynika z niespełnienia warunku normalności. Różnica jest wyraźna i wskazuje na konieczność obliczania wariancji estymatora ze wzoru (8).

Na wykresie 1 przedstawiono rozkłady estymatorów wariancji zmiennej $X^D$ dla $n$=20 oraz $n$=30. Wynika z nich, że rozkład GV2 (rozkład graniczny wymagający założenia normalności) nie stanowi właściwego przybliżenia rozkładu DBV dla rozpatrywanej zmiennej losowej. Oznacza to, że w przypadku estymatorów wariancji niespełnienie warunku normalności nie może być ignorowane. Nie można wówczas stosować rozkładu granicznego, wymagającego założenia, że próba pierwotna ma rozkład normalny. Rozkład GV1 natomiast jest przybliżeniem dobrym.

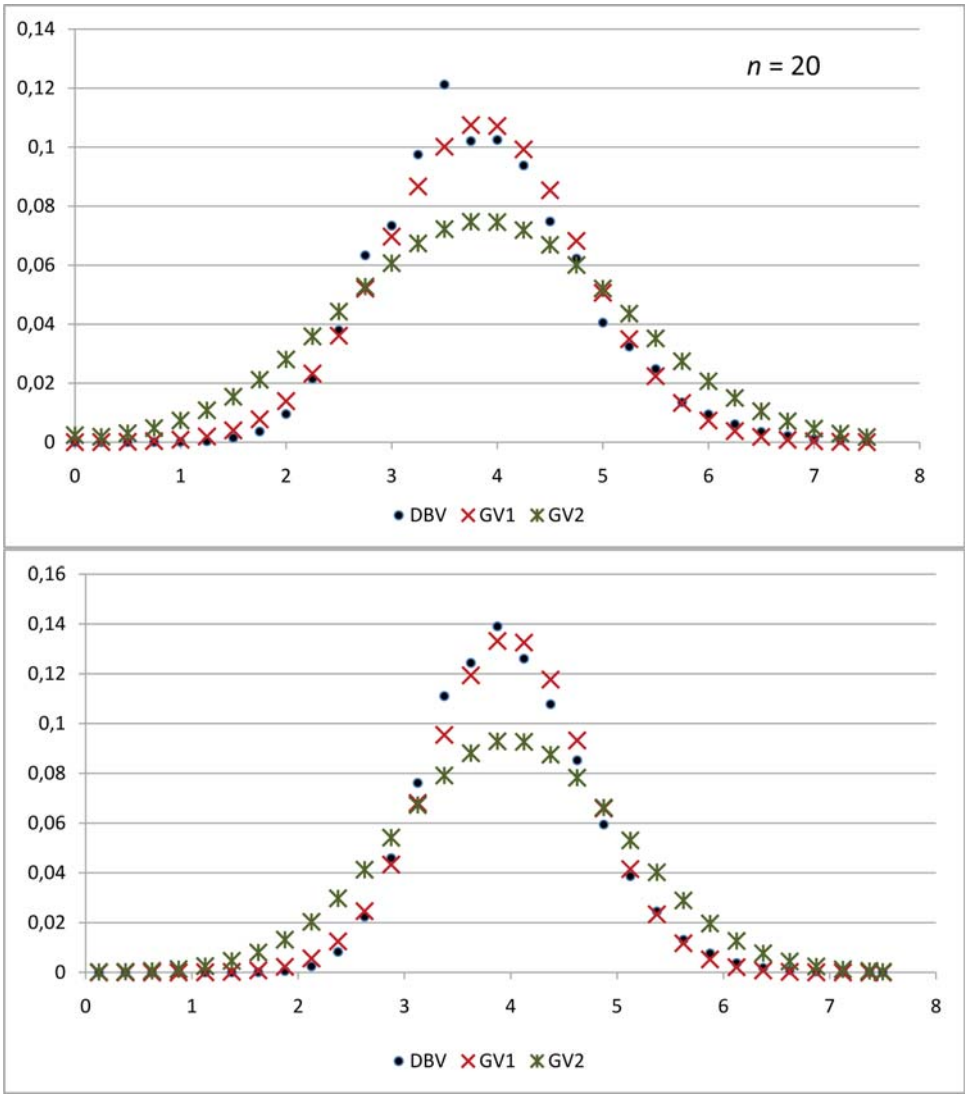
Zauważmy ponadto, że rozkład DBV jest minimalnie asymetryczny, podczas gdy rozkład graniczny jest oczywiście symetryczny.

Zgodność rozkładów GV1 i DBV potwierdził test zgodności Pearsona, zarówno dla $n$=30 jak dla $n$=20.

W tabeli 3 przedstawiono przedziały ufności dla wariancji wyznaczone przy użyciu dokładnej metody bootstrapowej DBV, rozkładu granicznego nie wymagającego założenia normalności próby GV1 oraz wymagającego założenia normalności GV2.

Porównując rozstępy poszczególnych przedziałów stwierdzamy, że najprecyzyjniejszego oszacowania dokonano przy pomocy dokładnej metody bootstrapowej (i jest to oszacowanie dokładne), następnie rozkładu granicznego GV1 (z wyjątkiem $n$=20 i 1-$\alpha$ = 0,99, dla których przedział ufności rozkładu GV1 był węższy od DBV). Najszersze przedziały ufności ma rozkład GV2.

Wykres 1. Rozkłady estymatorów wariancji DBV, GV1 i GV2




Źródło: Badania własne

Przesunięcie w lewo przedziałów dla rozkładu GV1 wobec DBV wynika z asymetrii drugiego z nich. Przedstawione obliczenia wskazują, że rozkład GV1 jest dobrym przybliżeniem rozkładu DBV.

W celu dokładniejszej prezentacji metody przedstawione zostaną wyniki estymacji dla małej próby obejmującej wartości {1, 2, 3, 4, 5} gdzie $n = k$=5 przy założeniu jednakowych prawdopodobieństw wylosowania każdego elementu, równych 0,2. Rozkład ten reprezentuje dyskretna zmienna losowa $X^D$ o wartości oczekiwanej i wariancji

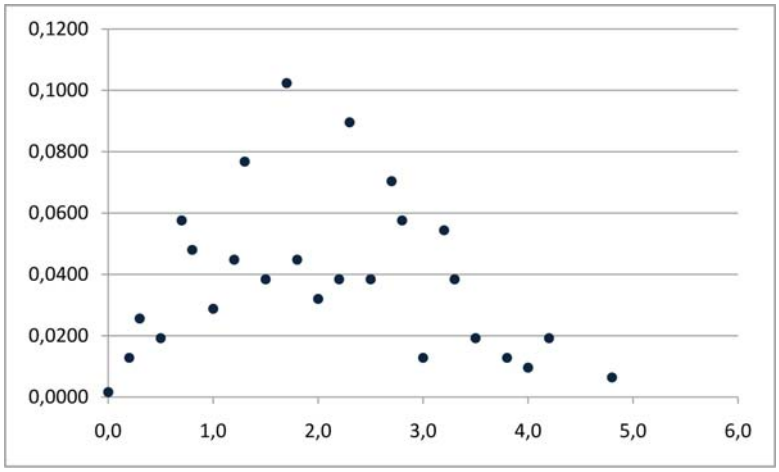
Tabela 3
Przedziały ufności dla wariancji wyznaczone dokładną metodą bootstrapową oraz przy użyciu rozkładów granicznych

| Poziom ufności | | *n*=20 | | | *n*=30 | | |
|---|---|---|---|---|---|---|---|
| | | DBV | GV1 | GV2 | DBV | GV1 | GV2 |
| **1-$\alpha$ = 0,95** | **granice** | 2,3685 | 2,1907 | 1,3868 | 2,6715 | 2,5462 | 1,9011 |
| | | 5,9565 | 5,7888 | 6,5927 | 5,0845 | 5,4332 | 6,0784 |
| | **rozstęp** | 3,5880 | 3,5981 | 5,2059 | 2,4130 | 2,8870 | 4,1773 |
| **1-$\alpha$ = 0,99** | **granice** | 1,9575 | 1,6254 | 0,5689 | 2,3265 | 2,0926 | 1,2448 |
| | | 6,6945 | 6,3541 | 7,4106 | 6,1185 | 5,8868 | 6,7347 |
| | **rozstęp** | 4,7370 | 4,7287 | 6,8417 | 3,7920 | 3,7942 | 5,4899 |

Źródło: Badania własne

odpowiednio równych $\mu^D$= 3, oraz $\left(\sigma^{\mathrm{D}}\right)^2$= 2. Wartość oczekiwana nieobciążonego estymatora wariancji jest równa 2, a jego wariancja 0,96 (zgodnie ze wzorem (8) po korekcie odpowiednio o czynniki 5/4 i $(5/4)^2$). Z 5 elementowej próby pierwotnej można wylosować $5^5$=3125 prób wtórnych, przy czym prawdopodobieństwo wylosowanie każdej z nich jest w zadanych warunkach jednakowe i wynosi 1/3125. Estymator wariancji dla przyjętej próby pierwotnej przybiera jedynie 26 różnych wartości. W tabeli 4 oraz na wykresie 2 przedstawiony został rozkład nieobciążonego estymatora wariancji wyznaczonego dokładną metodą bootstrapową. Wartość oczekiwana i wariancja tego rozkładu są dokładnie równe 2 oraz 0,96, czego należało oczekiwać. Warto dodać, że dla tak małej próby łatwo wykonać niezbędne kalkulacje z poziomu arkusza kalkulacyjnego programu Excel, bez konieczności używania specjalnego oprogramowania.

Wykres 2. Rozkłady estymatora wariancji dla próby obejmującej wartości {1, 2, 3, 4, 5}




Źródło: Badania własne

Tabela 4

Rozkład estymatora wariancji uzyskany dokładną metodą bootstrapową dla 5-cio elementowej próby obejmującej wartości {1, 2, 3, 4, 5}

| Wartości | Prawdopodobieństwa | Wartości | Prawdopodobieństwa |
|---|---|---|---|
| 0,0 | 0,0016 | 2,2 | 0,0384 |
| 0,2 | 0,0128 | 2,3 | 0,0896 |
| 0,3 | 0,0256 | 2,5 | 0,0384 |
| 0,5 | 0,0192 | 2,7 | 0,0704 |
| 0,7 | 0,0576 | 2,8 | 0,0576 |
| 0,8 | 0,0480 | 3,0 | 0,0128 |
| 1,0 | 0,0288 | 3,2 | 0,0544 |
| 1,2 | 0,0448 | 3,3 | 0,0384 |
| 1,3 | 0,0768 | 3,5 | 0,0192 |
| 1,5 | 0,0384 | 3,8 | 0,0128 |
| 1,7 | 0,1024 | 4,0 | 0,0096 |
| 1,8 | 0,0448 | 4,2 | 0,0192 |
| 2,0 | 0,0320 | 4,8 | 0,0064 |

Źródło: Badania własne

## 5. PODSUMOWANIE

W artykule omówiono dokładną metodę bootstrapową, którą można wykorzystać do szacowania estymatorów parametrów zmiennych losowych o nieznanym rozkładzie. Metoda pozwala wyznaczyć oszacowanie dowolnego parametru, błąd tego oszacowania, rozkład estymatora, czy przedziały ufności. Tradycyjnie zadanie takie realizowane jest przy pomocy metody bootstrapowej, która polega na losowaniu prób wtórnych z pierwotnej próby losowej. Losowanie próby stosowane jest w statystyce, jeśli nie może być zbadana cała populacja, lub badanie całej populacji jest zbyt kłopotliwe. Próba pierwotna jest po pierwsze skończona, a po drugie znany jest jej rozkład – jest to rozkład empiryczny. Zamiast wtórnie próbkować próbę pierwotną można wygenerować automatycznie całą przestrzeń prób wtórnych i wyznaczyć dla niej wartości statystyki będącej estymatorem poszukiwanego parametru.

W artykule przedstawiono propozycję algorytmu realizującego to zadanie. Poprawność metody sprawdzono na przykładzie wariancji. Pokazano, że wartość oczekiwana estymatora obliczonego dokładną metodą bootstapową jest równa dokładnie wariancji próby. Metoda nie wprowadza więc obciążenia wynikającego z wtórnego próbkowania jak ma to miejsce w klasycznym bootstrapie. Wniosek ten jest jedynie potwierdzeniem

oczywistego faktu, że badając całą populacje uzyskujemy wyniki dokładniejsze niż jeśli bierzemy pod uwagę próbę losową.

Rozkład nieobciążonego estymatora wariancji wyznaczony dokładną metodą bootstrapową porównano z rozkładem granicznym dla wariancji, nie wymagającym założenia normalności próby pierwotnej. Podobieństwo obydwu rozkładów wskazuje, na możliwość przybliżenia rozkładu dokładnego rozkładem granicznym.

Badania pokazały, że w przypadku estymatora obciążonego nie można pominąć kwestii normalności próby pierwotnej. Rozkład graniczny estymatora wariancji, jeśli próba nie ma rozkładu normalnego, nie może być przybliżony rozkładem $N\left(\sigma^2, \frac{n-1}{n}\sqrt{\frac{2\cdot\sigma^4}{n}}\right)$. Rozkładem granicznym jest rozkład

$$\mathrm{N}\left(\sigma^2, \frac{n}{n-1}\sqrt{\frac{\mu_4-\sigma^4}{n}-\frac{2\cdot(\mu_4-2\cdot\sigma^4)}{n^2}+\frac{\mu_4-3\cdot\sigma^4}{n^3}}\right).$$

Przeprowadzone eksperymenty symulacyjne polegające na estymacji parametrów dla różnych rozmiarów próby pierwotnej pokazały, że w przypadku prób małych ($n \leqslant 15$ i $k = n$) czas niezbędny do wygenerowania całej przestrzeni prób wtórnych jest krótki (poniżej 10 sek. na komputerze średniej klasy). Oznacza to, że nie ma wówczas potrzeby przeprowadzać wtórnego losowania próby. Dla prób większych, czas ten jest zdecydowanie dłuższy i wymaga kilku godzin obliczeń (dla $n$=20 i $k = n$ obliczenia trwały 5 godz. 30 min.) Wzrost wymiarowości problemu powoduje bardzo silne wydłużenie czasu obliczeń. Z drugiej strony pamiętać należy, że metoda bootstrapowa stosowana jest w przypadku małych prób – dla prób dużych można wykorzystać graniczne rozkłady estymatorów.

Biorąc jednak pod uwagę postęp w technice komputerowej, dokładny bootstrap będzie mógł być w przyszłości stosowany również dla prób większych.

*Szkoła Główna Gospodarstwa Wiejskiego*

LITERATURA

[1] Boot J.G., Sarkar S. [1998]: *Monte Carlo Approximation of Bootstrap Variances*. The American Statistician. Nr. 52, Assue 4. 354-357.

[2] Boot J.G.: Monte Carlo and the boothstrap. Prezentacja elektroniczna: http://www.stat.ufl.edu/~jbooth/documents/talks/bootse.pdf (2010.1.3).

[3] Domański C., Pruska K. [2000]: *Nieklasyczne metody statystyczne*. Polskie Wydawnictwo Ekonomiczne, Warszawa.

[4] Efron B. [1979]: Bootstrap methods: another look at the jackknife. *The Annals of Statistics*. Vol. 7, No. 1, 1-26.

[5] Efron B. [1987]: Better Bootstrap Confidence Intervals (with discussion). *Journal of the American Statistical Association*. Vol. 82, No. 397, 171-185.

[6] Efron B., Tibshirani R.J. [1993]: *An introduction to the Bootstrap.* Chapman & Hall. London.

[7] Feller W. [1950]: *An introduction to probability theory and its application*. John Wiley & Sons. New York, London, Sydney.
[8] Smirnow N.W., Dunin-Barkowski I.W. [1973]: *Kurs rachunku prawdopodobieństwa i statystyki matematycznej dla zastosowań technicznych*. Państwowe Wydawnictwo Naukowe, Warszawa.
[9] Zieliński R. [2004]: *Siedem wykładów wprowadzających do statystyki matematycznej*. Publikacja elektroniczna: http://www.impan.pl/~rziel/7ALL.pdf (2009.08.09)

## DOKŁADNA METODA BOOTSTRAPOWA I JEJ ZASTOSOWANIE DO ESTYMACJI WARIANCJI

### Streszczenie

W artykule przedstawiono dokładną metodę bootstrapową, którą można wykorzystać do szacowania estymatorów parametrów zmiennych losowych o nieznanym rozkładzie. Metoda pozwala wyznaczyć oszacowanie dowolnego parametru, błąd tego oszacowania, rozkład estymatora, czy przedziały ufności. Tradycyjnie zadanie takie realizowane jest przy pomocy metody bootstrapowej, która polega na wtórnym próbkowaniu analizowanej, pierwotnej próby losowej. Losowanie próby stosowane jest w statystyce, jeśli nie może być zbadana cała populacja, lub badanie całej populacji jest zbyt kłopotliwe. Próba pierwotna jest po pierwsze skończona, a po drugie znany jest jej rozkład – jest to rozkład empiryczny. Zamiast wtórnie próbkować próbę pierwotną można wygenerować automatycznie całą przestrzeń prób wtórnych i wyznaczyć dla niej wartości statystyki będącej estymatorem poszukiwanego parametru. W artykule przedstawiono propozycję algorytmu realizującego metodę dokładnego bootstrapu, którego poprawność sprawdzono na przykładzie nieobciążonego estymatora wariancji. Pokazano, że wartość oczekiwana estymatora obliczonego dokładną metodą bootstapową jest równa dokładnie wariancji próby. Metoda nie wprowadza więc obciążenia wynikającego z wtórnego próbkowania jak ma to może mieć miejsce w klasycznym bootstrapie. Rozkład estymatora wyznaczony dokładną metodą bootstrapową porównano z rozkładem granicznym estymatora wariancji, nie wymagającym założenia normalności próby pierwotnej. Badania pokazały, że w przypadku wariancji nie można pominąć kwestii normalności próby pierwotnej.

## EXACT BOOTSTRAP METHOD AND IT'S APPLICATION IN ESTIMATION OF VARIANCE

### Summary

The article presents the exact bootstrap method, which can be used to estimate the parameters of the estimators of random variables with unknown distribution. The method allows to determine an estimate of any parameter, the error of estimation, the distribution of the estimator and confidence intervals. Traditionally this task is carried out using the bootstrap method, which consists of resampling of the original sample. Random sampling is necessary if examining the entire population data is impossible or too costly. Note that the fundamental sample property is of finite size and we know its distribution – it is the empirical distribution. Rather than driving a resample, we can generate automatically the entire resample space and calculate the values of a statistic which is looking for a parameter estimator. This article describes a method for performing exact algorithm for bootstrapping, which correctness was verified on an example of the unbiased estimator of variance. It is shown that the expected value of the

estimator calculated with exact bootstrap is exactly equal the variance of the sample. The method does not introduce bias of the resampling, as it may be for the classic bootstrap. The distribution of the estimator determined by the exact bootstrap compared with the limit distribution for estimator of variance, which does not require assumptions of normality of the original sample. Research has shown that if we estimate variance we cannot ignore the issue of normality of the original sample.