METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XII/2, 2011, str. 191–201

DOKŁADNA METODA BOOTSTRAPOWA NA PRZYKŁADZIE ESTYMACJI ŚREDNIEJ

Joanna Kisielińska
Katedra Ekonomiki Rolnictwa i Międzynarodowych Stosunków Gospodarczych
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: joanna_kisielinska@sggw.pl

Streszczenie: Metoda bootstrapowa polega na wtórnym próbkowaniu pierwotnej próby losowej pobranej z populacji o nieznanym rozkładzie. W artykule pokazano, że wtórne próbkowanie nie jest konieczne, jeśli rozmiar próby nie jest zbyt duży. Możliwe jest wówczas automatyczne wygenerowanie wszystkich prób wtórnych i obliczenie wszystkich realizacji wybranego estymatora. Metodę dokładnego bootstrapu zastosowano do oszacowania średniej. Losowanie próby może być interpretowane jako dyskretyzacja ciągłej zmiennej losowej. Biorąc pod uwagę postęp w technice komputerowej, można mieć nadzieję, że znaczenie dyskretnych zmiennych losowych w statystyce będzie coraz większy.

Słowa kluczowe: dokładna metoda bootstrapowa, nieparametryczna estymacja średniej

WPROWADZENIE

Jednym z podstawowych problemów statystyki matematycznej jest estymacja parametrów. Zakładamy, że dana jest zmienna losowa X o rozkładzie prawdopodobieństwa F. Chcemy określić parametr tego rozkładu, który oznaczmy jako θ. Jeżeli parametru nie można wyznaczyć należy pobrać próbę losową oraz dobrać odpowiedni estymator. Próba losowa może być interpretowana jako ciąg zmiennych losowych o jednakowych rozkładach. Wprowadźmy oznaczenia: $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ - n elementowa próba losowa, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ – pewna realizacja próby losowej, zaś $\hat{\theta} = t(\mathbf{X})$ - estymator parametru θ. Zgodnie z tym zapisem estymator jest statystyką określona na przestrzeni prób.

W pewnych przypadkach przyjmując określony rozkład F zmiennej X, można wyznaczyć rozkład statystyki będącej estymatorem parametru. Często jednak rozkład F nie jest znany, bądź znajomość jego nie wystarcza do wyznaczenia rozkładu estymatora. Można wówczas zastosować metodę bootstrapową, zaproponowaną przez Efrona [Efron 1979]. Metoda ta polega na losowaniu z pierwotnej próby losowej **x**, prób wtórnych o liczebnościach n. Próby wtórne nazywamy próbami bootstrapowymi. Losowanie próby wtórnej odbywa się ze zwracaniem, przy założeniu jednakowego prawdopodobieństwa wylosowania każdego elementu próby pierwotnej. Dla próby n - elementowej prawdopodobieństwo to jest równe $1/n$. Rozkład w ten sposób określony jest równoważny rozkładowi empirycznemu i nazywany jest rozkładem bootstrapowym oznaczanym jako $\hat{F}$.

Wprowadźmy oznaczenie: $\mathbf{X}^* = \left(X_1^*, X_2^*, \cdots, X_n^*\right)$ - n elementowa wtórna próba bootstrapowa, $\mathbf{x}^* = \left(x_1^*, x_2^*, \cdots, x_n^*\right)$ - dowolna jej realizacja oraz $\hat{\theta}^* = t(\mathbf{X}^*)$ - estymator $\hat{\theta}$ dla próby bootstrapowej.

Istotą metody bootstrapowej jest aproksymacja rozkładu statystyki $\hat{\theta}$, rozkładem statystyki $\hat{\theta}^*$. Metodę bootstrapową stosujemy bowiem wówczas, gdy określenie rozkładu statystyki $\hat{\theta}$ nie jest możliwe. Rozkład statystyki $\hat{\theta}^*$ natomiast może być wyznaczony np. metodą Monte Carlo (inne metody przedstawiono np. w pracach [Efron, Tibshirani 1993], [Domański, Pruska 2000]).

Oznaczmy liczbę wylosowanych prób wtórnych jako N. Ciąg prób wtórnych można zapisać jako: $\mathbf{x}^{*1}$, $\mathbf{x}^{*2}$, … , $\mathbf{x}^{*N}$. Każda z prób pozwala wyznaczyć pojedynczą realizację statystyki $\hat{\theta}^*$. Dla określonej próby pierwotnej realizacja statystyki $\hat{\theta}^*$ dla b-tej próby wtórnej może być zapisana jako:

$$\hat{\theta}^*(b) = t\left(\mathbf{x}^{*b}\right) \tag{1}$$

Oszacowaniem bootstrapowym parametru θ będzie wówczas:

$$\hat{\theta}^*(\bullet) = \frac{1}{N}\sum_{b=1}^{N}\hat{\theta}^*(b) \tag{2}$$

Oszacowaniem standardowego błędu szacunku parametru θ będzie odchylenie standardowe postaci:

$$s^* = \sqrt{\frac{1}{N}\sum_{b=1}^{N}\left(\hat{\theta}^*(b) - \hat{\theta}^*(\bullet)\right)^2} \tag{3}$$

lub:

$$\hat{s}^* = \sqrt{\frac{1}{N-1}\sum_{b=1}^{N}\left(\hat{\theta}^*(b) - \hat{\theta}^*(\bullet)\right)^2} \tag{4}$$

w zależności od tego, czy N obejmuje wszystkie możliwe próby (3), czy jest jedynie pewnym ich podzbiorem (4).

Liczba wszystkich możliwych prób wtórnych jest określona liczbą sposobów, na które spośród n wartości (z próby pierwotnej) można wylosować także n wartości (próby wtórne). Problem taki znany jest z kombinatoryki i wiadomo, że ze zbioru n – elementowego n wartości można wylosować na $B = n^n$ sposobów[1]. Wartość B określa rozmiar przestrzeni prób wtórnych. Prawdopodobieństwa wylosowania pojedynczej wtórnej próby są jednakowe i równe $1/B$.

Podkreślić należy, że przestrzeń prób wtórnych jest przestrzenią skończoną o rozmiarze B i taki jest rozmiar dokładnej (idealnej) próby bootstrapowej. Jeżeli nie jest on zbyt duży, można wyznaczyć wszystkie realizacje estymatora. Jeśli nie jest możliwe wygenerowanie całej przestrzeni prób wtórnych ponieważ n^n jest zbyt duże, konieczne jest losowanie, czyli zastosowanie klasycznego bootstrapu. Warto jednak zwrócić uwagę, że jeżeli używany jest estymator postaci sumy n niezależnych zmiennych losowych o jednakowych rozkładach i skończonej średniej i wariancji, z centralnego twierdzenia granicznego wynika, że dla dużych n estymator będzie miał rozkład asymptotycznie normalny.

Metodę bootstrapową wykorzystującą całą przestrzeń prób wtórnych nazwać można metodą dokładnego bootstrapu. Stwierdzenie, że metoda jest dokładna dotyczy jedynie próbkowania wtórnego. Jeśli zaś chodzi o próbę pierwotną to jej adekwatność względem pierwotnej zmiennej X, opieramy na twierdzeniu Gliwienki-Cantelliego.

DOKŁADNA METODA BOOTSTRAPOWA

Niech z populacji opisanej zmienną losową X o nieznanym rozkładzie prawdopodobieństwa F wylosowana zostanie n elementowa próba (pierwotna) $\mathbf{x} = (x_1, x_2, \dots , x_n)$. Jeżeli dla niektórych $i \neq j$ zachodzi $x_i = x_j$, możliwe jest zredukowanie[2] rozmiaru próby do k różnych wartości. Prawdopodobieństwa p_i uzyskania realizacji x_i, dla i=1, 2, …, k nie muszą być wówczas jednakowe dla wszystkich i. W przypadku klasycznego bootstrapu prawdopodobieństwa p_i są jednakowe i równe $1/n$.

Zauważmy, że próbę pierwotną wraz z prawdopodobieństwami uzyskania poszczególnych realizacji interpretować można jako pewną dyskretną zmienną losową o liczbie realizacji równej k (lub n jeśli w próbie nie wystąpiły powtórzenia). Zmienną tą nazwać można zmienną losową próby, ponieważ jej zbiór realizacji oraz rozkład wynikają bezpośrednio z pobranej pierwotnej próby

[1] Wartość ta określa liczbę n elementowych wariacji z powtórzeniami ze zbioru n - elementowego.

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

losowej. Zmienną tą oznaczmy jako X^{D}, a jej rozkład prawdopodobieństwa jako F^{D}. Rozkład ten jest następujący:

$$p_i^{\mathrm{D}} = P\left(X^{\mathrm{D}} = x_i\right) = p_i \text{, dla } i = 1, 2, \dots, k. \tag{5}$$

Rozkład zmiennej losowej X^{D} jest równoważny rozkładowi bootstrapowemu $\hat{F}$. Próbę wtórną oznaczyć można jako $\mathbf{X}^{\mathrm{D}} = \left(X_1^{\mathrm{D}}, X_2^{\mathrm{D}}, \cdots, X_n^{\mathrm{D}}\right)$, statystykę $\hat{\theta}^*$ zaś jako $\hat{\theta}^{\mathrm{D}} = t\left(\mathbf{X}^{\mathrm{D}}\right)$. Rozkład statystyki $\hat{\theta}$ może być przybliżony rozkładem statystyki $\hat{\theta}^{\mathrm{D}}$. Podkreślić należy, że ewentualne obciążenie, problem zgodności i efektywności dotyczy estymatora $\hat{\theta}$, ponieważ w dokładnej metodzie bootstrapowej realizacje estymatora $\hat{\theta}^{\mathrm{D}}$ są obliczane dla całej populacji prób wtórnych, a nie szacowana na podstawie próby. Metoda nie może więc wprowadzać żadnego dodatkowego obciążenia. Problem na tym etapie jest problemem statystyki opisowej, a nie matematycznej.

Pojedyncza b-tą realizację próby wtórnej $\mathbf{x}^{\mathrm{D}b} = \left(x_1^{\mathrm{D}b}, x_2^{\mathrm{D}b}, \cdots, x_n^{\mathrm{D}b}\right)$ pozwala obliczyć wartość $\hat{\theta}^{\mathrm{D}}(b) = t\left(\mathbf{x}^{\mathrm{D}b}\right)$ stanowiącej pojedynczą realizację estymatora $\hat{\theta}^{\mathrm{D}}$. Prawdopodobieństwo wylosowania b-tej próby wtórnej jest równe:

$$p^{Db} = P\left(\hat{\theta}^{\mathrm{D}} = \hat{\theta}^{\mathrm{D}}(b)\right) = \prod_{i=1}^{n} p_i^{\mathrm{D}b} \tag{6}$$

gdzie $p_i^{\mathrm{D}b} = P\left(X^{\mathrm{D}} = x_i^{\mathrm{D}b}\right)$. Wymiar przestrzeni prób wtórnych jest równy $B = k^n$, a wobec tego musi być spełniony warunek: $\sum_{b=1}^{B} p^{\mathrm{D}b} = 1$.

Formuła (6) określa rozkład estymatora $\hat{\theta}^{\mathrm{D}}$, którym przybliżamy rozkład estymatora $\hat{\theta}$. Jest to rozkład dyskretny o skończonej, choć w niektórych przypadkach bardzo dużej liczbie realizacji. Rozkład ten wykorzystać można do punktowej lub przedziałowej oceny parametru θ, bądź testowania hipotez.

Zauważmy, że metoda bootstrapowa w istocie polega na przybliżeniu nieznanego ciągłego rozkładu pewnej zmiennej losowej $\hat{\theta}$, dyskretną zmienną losową $\hat{\theta}^{\mathrm{D}}$ o rozkładzie, który można wygenerować na podstawie próby. Losując próbę przeprowadzamy tak naprawdę dyskretyzację pewnego ciągłego zjawiska. Ciągłą zmienną losową X staramy się przybliżyć ciągiem jej realizacji $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Dyskretyzacja problemów ciągłych stosowana jest powszechnie w naukach technicznych czy problemach numerycznych. Jako przykład podać można problem optymalizacji dynamicznej, którego szczególnym przypadkiem jest zadanie sterowania optymalnego. Jedynie w nielicznych przypadkach możliwe jest znalezienie funkcji czasu stanowiącej rozwiązanie, wobec czego zadania rozwiązuje się stosując dyskretyzację.

Wygenerowanie wszystkich prób wtórnych może być zrealizowane jako rekurencyjne pobieranie kolejnych elementów z n-elementowej próby. Jeśli wystąpiły powtórzenia, zbiór z którego pobierane są wartości redukuje się do wymiaru k. Okazuje się jednak, że liczbę generowanych realizacji prób bootstrapowych można znacznie zredukować, ponieważ w próbie wtórnej część wartości będzie się powtarzać - losowanie odbywa się z powtórzeniami. Algorytm pozwalający wygenerować wszystkie próby wtórne przy założeniu jednakowych prawdopodobieństw wylosowania każdego elementu próby pierwotnej przedstawiony jest w pracy [Fisher and Hall 1991]. Na stronie http://mors.sggw.waw.pl/~jkisielinska umieszczony został skompilowany program napisany w języku C++ generujący wszystkie realizacje estymatora średniej, dla n elementowej próby wtórnej zredukowanej do k różnych wartości. Estymator średniej dany jest wzorem:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

Jeżeli próba jest duża ($n \geq 30$) uzyskany rozkład może być porównany z rozkładem granicznym. Jeżeli założymy, że próba została pobrana z populacji o dowolnym rozkładzie o wartości oczekiwanej μ i odchyleniu standardowym σ, estymator średniej $\overline{X}$ ma rozkład normalny z parametrami:

$$\mu_{\overline{X}} = \mu; \sigma_{\overline{X}} = \sigma / \sqrt{n} \tag{8}$$

Stosując dokładną metodę bootstrapową rozkład zmiennej losowej X przybliżamy rozkładem dyskretnej zmiennej losową X^{D}, której wartość oczekiwaną μ^{D} oraz odchylenie standardowe σ^{D} obliczmy według wzorów:

$$\mu^{\mathrm{D}} = \sum_{i=1}^{k} x_i \cdot p_i, \sigma^{\mathrm{D}} = \sqrt{\sum_{i=1}^{k} \left(x_i - \mu^{\mathrm{D}}\right)^2 \cdot p_i} \tag{9}$$

Wobec tego normalny, graniczny rozkład estymatora $\overline{X}$ (który oznaczymy jako GA) jest następujący:

$$\mathrm{GA} : \mathrm{N}\left(\mu^{\mathrm{D}}, \sigma^{\mathrm{D}} / \sqrt{n}\right), \tag{10}$$

WYNIKI OBLICZEŃ

Dokładna metoda bootstrapowa wykorzystana zostanie do estymacji średniej dla przykładowej próby losowej przedstawionej w tabeli 1. Przyjęto, że prawdopodobieństwa wylosowania poszczególnych realizacji nie są jednakowe. Takie podejście w pewnym stopniu uzasadnia wykorzystanie rozkładu z tabeli 1 dla różnych liczebności próby - obliczenia wykonane zostaną dla przykładowych trzech rozmiarów: n=10, n=20 i n=30. Podkreślić należy, że w przypadku stosowania metod bootstrapowych wartość n wynika bezpośrednio z liczebności próby. Założenie różnych wartości dla n traktować należy jedynie jako eksperyment symulacyjny.

Zakładamy, że dana jest n elementowa próba pobrana z nieznanego rozkładu prawdopodobieństwa F, którą reprezentuje dyskretna zmienna losowa X^D. Rozkład prawdopodobieństwa zmiennej X^D przedstawiony jest w tabeli 1. Wartość oczekiwana i odchylenie standardowe X^D są odpowiednio równe μ^D = 5,17400, oraz σ^D = 1,99743.

Tabela 1. Rozkład zmiennej losowej X^D

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

Źródło: opracowanie własne

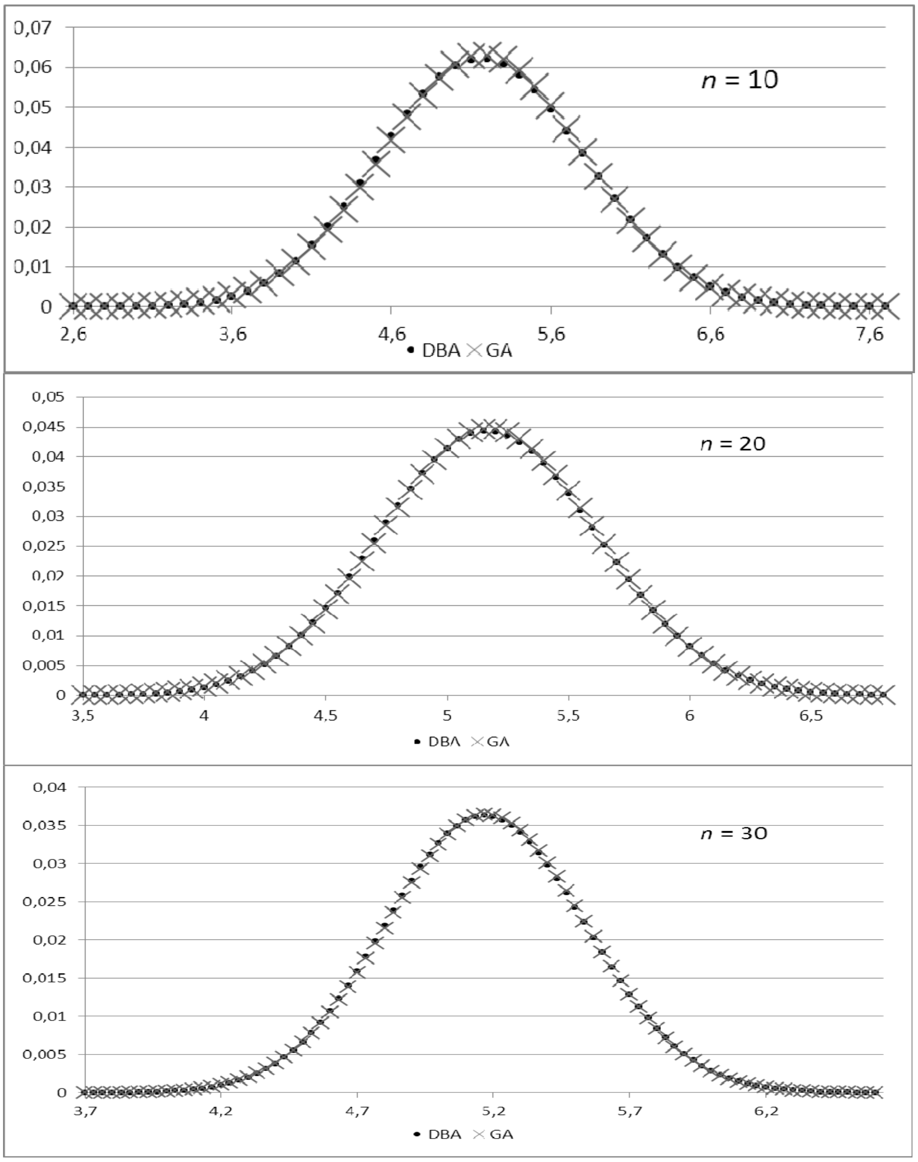
Na rysunku 1 przedstawiono rozkład estymatora średniej wyznaczony metodą dokładnego bootstrapu (oznaczony jako DBA) oraz rozkład graniczny GA określony wzorem (10), w przedziale odległym od średniej o 4 odchylenia standardowe. Dokładny rozkład bootstrapowy dla średniej podany został jako prawdopodobieństwo przyjęcia przez nią poszczególnych wartości (rozkład estymatora średniej jest rozkładem dyskretnym). W przypadku rozkładów granicznych natomiast jest to prawdopodobieństwo przyjęcia wartości z przedziału (od połowy przedziału między wartościami z lewej, do połowy z prawej strony). Dla trzech przykładowych rozmiarów próby wykresy są niemal identyczne.

W celu ilustracji rzędu podobieństwa, w tabeli 2 podano rozkłady estymatora średniej DBA i GA dla trzech przykładowych rozmiarów próby. Z uwagi na ograniczoną ilość miejsca wybrano po 21 realizacji skoncentrowanych wokół średniej. Wartości prawdopodobieństw pokrywają się z dokładnością niemal do trzeciego miejsca po przecinku (a co najmniej drugiego). Im większa próba, tym różnice między rozkładami są mniejsze. Można więc wnioskować, że w przypadku średniej stosowanie rozkładów granicznych jest w pełni uzasadnione, nawet jeśli liczebność próby jest mniejsza od 30. Fakt ten potwierdzony został przeprowadzonymi eksperymentami symulacyjnymi dla innych rozkładów, niż dany w tabeli 1. Wykorzystując udostępniony program czytelnik może sam przeprowadzić podobne eksperymenty.

W tabeli 3 przedstawione zostały parametry rozkładów estymatora średniej – wartość oczekiwana i odchylenie standardowe, natomiast w tabeli 4 przedziały ufności wyznaczone na podstawie uzyskanych rozkładów.

Jeśli chodzi o parametry rozkładów są one niemal jednakowe (z dokładnością do 5-ciu miejsc po przecinku dla trzech liczebności). Dla rozkładu DBA parametry obliczone zostały na podstawie wszystkich wygenerowanych realizacji – są to więc wartości dokładne. Wartości oczekiwane rozkładów granicznych GA dla trzech liczebności próby są równe wartości oczekiwanej dla próby (co wynika bezpośrednio z teorii). W przypadku rozkładu DBA również otrzymano wartość oczekiwaną równą wartości oczekiwanej dla próby. Rezultat ten potwierdza poprawność użytego algorytmu. Podkreślić należy, że dokładna metoda bootstrapowa nie wprowadza dodatkowego obciążenia estymatora (w przeciwieństwie do metody bootstrapowej z losowaniem prób).

Rysunek 1. Rozkłady estymatorów średniej DBA i GA dla przykładowych trzech liczebności próby



Źródło: obliczenia własne

Tabela 2. Rozkłady estymatorów średniej DBA i GA dla przykładowych trzech liczebności próby

n=10			n=20			n=30		
$\overline{X}$	prawdop.		$\overline{X}$	prawdop.		$\overline{X}$	prawdop.	
	DBA	GA		DBA	GA		DBA	GA
4,20	0,01993	0,01926	4,65	0,02288	0,02245	4,8333	0,02385	0,02357
4,30	0,02514	0,02427	4,70	0,02588	0,02543	4,8667	0,02584	0,02556
4,40	0,03083	0,02983	4,75	0,02891	0,02846	4,9000	0,02775	0,02749
4,50	0,03679	0,03575	4,80	0,03187	0,03145	4,9333	0,02954	0,02932
4,60	0,04275	0,04179	4,85	0,03468	0,03432	4,9667	0,03120	0,03102
4,70	0,04840	0,04764	4,90	0,03728	0,03699	5,0000	0,03267	0,03253
4,80	0,05343	0,05297	4,95	0,03956	0,03937	5,0333	0,03392	0,03384
4,90	0,05755	0,05744	5,00	0,04146	0,04138	5,0667	0,03493	0,03491
5,00	0,06048	0,06075	5,05	0,04292	0,04295	5,1000	0,03568	0,03571
5,10	0,06205	0,06266	5,10	0,04389	0,04403	5,1333	0,03614	0,03623
5,20	0,06217	0,06304	5,15	0,04433	0,04457	5,1667	0,03630	0,03645
5,30	0,06083	0,06185	5,20	0,04424	0,04456	5,2000	0,03617	0,03636
5,40	0,05813	0,05919	5,25	0,04360	0,04400	5,2333	0,03575	0,03597
5,50	0,05426	0,05524	5,30	0,04246	0,04290	5,2667	0,03504	0,03530
5,60	0,04948	0,05028	5,35	0,04086	0,04131	5,3000	0,03407	0,03434
5,70	0,04408	0,04464	5,40	0,03884	0,03928	5,3333	0,03286	0,03314
5,80	0,03834	0,03865	5,45	0,03648	0,03689	5,3667	0,03143	0,03171
5,90	0,03256	0,03264	5,50	0,03385	0,03421	5,4000	0,02982	0,03009
6,00	0,02700	0,02688	5,55	0,03103	0,03133	5,4333	0,02807	0,02831
6,10	0,02185	0,02159	5,60	0,02811	0,02834	5,4667	0,02620	0,02642
6,20	0,01725	0,01691	5,65	0,02516	0,02531	5,5000	0,02427	0,02445

Źródło: obliczenia własne

Tabela 3. Wartość oczekiwana i odchylenie standardowe rozkładów estymatora średniej

rozmiar próby	n=10		n=20		n=30	
rozkład estymatora	DBA	GA	DBA	GA	DBA	GA
średnia	5,17400	5,17400	5,17400	5,17400	5,17400	5,17400
odchylenie standardowe	0,63164	0,63164	0,44664	0,44664	0,36468	0,36468

Źródło: obliczenia własne

W przypadku przedziałów ufności dla rozkładów DBA i GA występują pewne różnice - największe dla n=10, zarówno dla poziomu ufności 0,95 jak i 0,99. Przedziały obliczone z rozkładu granicznego są zwykle nieco węższe, niż wyznaczone metodą dokładnego bootstrapu. Podkreślić należy, że dla ustalonego rozmiaru próby n nie jest możliwe uzyskanie dowolnie dużej dokładności oszacowania przedziałów ufności, ponieważ dokładny rozkład bootstrapowy DBA jest rozkładem dyskretnym, a wobec tego cechuje go pewna ziarnistość. Wraz ze wzrostem rozmiaru próby różnice są coraz mniejsze. Dla $n = 20$ granice przedziałów ufności rozkładów DBA i GA różnią się na drugim miejscu po przecinku, dla $n = 30$ na trzecim.

Tabela 4. Przedziały ufności dla średniej wyznaczone na podstawie uzyskanych rozkładów estymatora

rozmiar próby		n=10		n=20		n=30	
rozkład estymatora		DBA	GA	DBA	GA	DBA	GA
$1-\alpha = 0,95$	granice	3,9000	3,9360	4,2750	4,2986	4,4500	4,4592
		6,5000	6,4120	6,0750	6,0494	5,9000	5,8888
	rozstęp	2,6000	2,4760	1,8000	1,7508	1,4500	1,4295
$1-\alpha = 0,99$	granice	3,6000	3,5470	4,0500	4,0235	4,2333	4,2346
		6,9000	6,8010	6,3500	6,3245	6,1333	6,1134
	rozstęp	3,3000	3,2540	2,3000	2,3009	1,9000	1,8787

Źródło: obliczenia własne

WNIOSKI

W artykule zwrócono uwagę, że ze względu na postęp w technice komputerowej metoda bootstrapowa polegająca na wtórnym próbkowaniu pierwotnej próby losowej może być zastąpiona dokładną metodę bootstrapową. W przypadku metody dokładnego bootstrapu generowane są wszystkie możliwe próby wtórne, co pozwala wyznaczyć wszystkie realizacje statystyki stanowiącej estymator poszukiwanego parametru. W klasycznej metodzie boostrapowej realizacje estymatora są losowane, czego skutkiem może być wylosowanie tej samej próby wtórnej (rozkład estymatora bootstrapowego jest bowiem rozkładem dyskretnym). Podkreślić należy, że próbę losujemy jeśli nie możemy zbadać całej populacji lub badanie całej populacji jest zbyt kłopotliwe bądź kosztowne. Jeśli możliwe jest wygenerowanie całej przestrzeni prób wtórnych nie ma potrzeby wtórnego losowania.

Ponieważ czas niezbędny do wygenerowania całej przestrzeni prób wtórnych zależy nie tylko od rozmiaru próby, ale również rodzaju estymatora, trudno podać wartość n, która stanowiła by wartość graniczną (aktualnie maksymalną) dla metody dokładnej. Orientacyjnie można przyjąć, że jest to w przypadku $k = n$ mniej więcej 20. Dla większych prób czas obliczeń jest aktualnie zbyt długi, co nie oznacza, że w przyszłości tak będzie. Moc obliczeniowa komputerów rośnie bardzo szybko i to co dziś zajmuje wiele godzin, w przyszłości trwać może ułamek sekundy.

Metodę dokładnego bootstrapu wykorzystano do estymacji średniej. Pokazano, że wartości oczekiwane estymatorów są równe dokładnie średniej. Metoda nie wprowadza więc obciążenia wynikającego z wtórnego próbkowania jak może to mieć miejsce w klasycznym bootstrapie. Rozkłady estymatorów wyznaczonych dokładną metodą bootstrapową porównano z rozkładami granicznymi. Podobieństwo rozkładów wskazuje, na możliwość przybliżenia rozkładu „dokładnego" rozkładem granicznym.

Jakie są konsekwencje przedstawionej w artykule możliwości wygenerowania pełnej informacji zawartej w próbie? Kwestia podstawowa to znacznie większa swoboda w konstrukcji estymatorów. Nie ma bowiem potrzeby zakładania postaci rozkładu, aby możliwe było określenie rozkładu estymatora. Dążyć należy, aby estymator był nieobciążony, zgodny i efektywny. Do badania tych kwestii metoda dokładnego boostrapu może być również przydatna.

Podkreślić należy, że pobranie próby losowej może być traktowane jako zastąpienie ciągłej zmiennej losowej o nieznanym rozkładzie zmienną dyskretną o rozkładzie znanym – rozkładzie bootstrapowym (równoważnym rozkładowi empirycznemu). Przekształcenia zmiennych dyskretnych są prostsze od przekształceń zmiennych ciągłych, ponieważ rozkłady statystyk zmiennych dyskretnych mogą być obliczane automatycznie. W rzeczywistości ze względu na skończoną dokładność wszelkich pomiarów możemy obserwować i tak jedynie zmienne dyskretne. Można przypuszczać, że wraz z rosnącą mocą komputerów ich znaczenie w statystyce będzie coraz większy.

BIBLIOGRAFIA

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. Vol. 7, No. 1, 1-26.

Efron B., Tibshirani R.J. (1993) An introduction to the Bootstrap. Chapman & Hall. London.

Fisher N.I., Hall P. (1991) Bootstrap algorithms for small samples. Journal of Statistical Planning and Inference. Vol. 27, 157-169.

THE EXACT BOOTSTRAP METHOD ON THE EXAMPLE OF THE MEAN

Abstract: The bootstrap method is based on resampling of the original random sample drawn from a population with an unknown distribution. In the article it was shown that resampling is unnecessary if the sample size is not too large. It is possible to automatically generate all possible resamples and calculate all realizations of the required estimator. The method was used to estimate mean. Random sampling may be interpreted as discretization of a continuous variable. Because of the progress in computer technology we may hope that the role of discrete variables in statistics will increase.

Key words: exact bootstrap method, nonparametric mean estimation