

Grzegorz Kończak

Akademia Ekonomiczna w Katowicach

O TEŚCIE RÓWNOŚCI WARIANCJI WYKORZYSTUJĄCYM METODĘ *BOOTSTRAP*

1. Wstęp

Podczas przeprowadzania różnych analiz statystycznych niezbędne jest spełnienie założenia o równości wariancji dla wielu zmiennych. Założenie to w praktyce jest sprawdzane z wykorzystaniem takich testów statystycznych, jak np. test Bartletta, Cochрана–Coksa lub Hartleya. Do najczęściej stosowanych w statystyce metod oceny wariancji estymatora na podstawie analizy prób należy metoda *bootstrap*. Jest to nieparametryczna metoda wykorzystująca analizy symulacyjne. Tą metodą możemy aproksymować rozkłady estymatorów i statystyk testowych. Na tej podstawie mogą być konstruowane różnorodne testy statystyczne, wykorzystujące rezultaty z wielokrotnie generowanych prób (podprób) z wylosowanej próby, która jest traktowana jako nowa populacja. Otrzymywane w ten sposób ciągi wartości rozważanych statystyk stanowią podstawę do podjęcia decyzji o przyjęciu lub odrzuceniu odpowiedniej hipotezy.

W artykule zaproponowano test równości wariancji wykorzystujący metodę *bootstrap*. Na podstawie uzyskanych rezultatów z próby, zgodnie z założeniami wspomnianej metody, generowane są próby bootstrapowe. Wykorzystując rezultaty pochodzące z tych prób, uzyskiwane są wartości statystyki pozwalającej na ocenę jednorodności wariancji. Ostateczna decyzja jest podejmowana na podstawie uzyskanego rozkładu otrzymanych wartości w N seriach generowań prób bootstrapowych. Rozważania analityczne uzupełniono symulacjami komputerowymi. Przedstawiono porównania testu Bartletta równości wariancji oraz testu wykorzystującego metodę *bootstrap*. Zwrócono uwagę na przypadki, kiedy może być uzasadnione zastosowanie w analizach testu opartego na metodzie *bootstrap*.

2. Postawienie problemu

Zakładamy, że analizowana zmienna X ma ciągły rozkład w k badanych populacjach. Wartości oczekiwane i odchylenia standardowe tej zmiennej w badanych populacjach oznaczamy przez $E(X_i) = \mu_i$ oraz $D(X_i) = \sigma_i$ dla $i = 1, 2, \dots, k$.

Interesuje nas weryfikacja hipotezy o równości wszystkich wariancji, co możemy zapisać następująco:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2, \quad (1)$$

wobec hipotezy alternatywnej $H_1 : \sigma_i^2 \neq \sigma_j^2$, dla pewnych $i \neq j$.

Stosowane klasycznie metody weryfikacji hipotezy o równości wielu wariancji wymagają spełnienia założenia o normalności rozkładu każdej ze zmiennych. Przy spełnieniu tego założenia do weryfikacji hipotezy (1) możemy wykorzystać np. test Bartletta, Hartleya, a również Cochрана (por. [Kendall, Stuart 1963]). W wielu praktycznych zastosowaniach nie możemy przyjąć wspomnianego założenia i wówczas nie jest możliwe stosowanie wspomnianych testów. Osłabienie założenia normalności badanych rozkładów można uzyskać, wykorzystując nieparametryczne metody wnioskowania statystycznego.

Założmy, że z badanych k populacji o dowolnych ciągłych rozkładach pobieramy próbki losowe o liczebnościach $n_1, n_2, \dots, n_k$. Oznaczmy przez x_{ij} j -ty pomiar dla i -tej zmiennej ($i = 1, 2, \dots, k$ oraz $j = 1, 2, \dots, n_i$). Rezultaty uzyskane w poszczególnych próbach możemy wobec powyższego zapisać następująco:

1. próba: $x_{11}, x_{12}, \dots, x_{1n_1}$,

2. próba: $x_{21}, x_{22}, \dots, x_{2n_2}$,

....

k -ta próba: $x_{k1}, x_{k2}, \dots, x_{kn_k}$.

Wariancję z próby dla i -tej zmiennej zapiszemy następująco:

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad i = 1, 2, \dots, k, \quad (2)$$

gdzie:
$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad i = 1, 2, \dots, k. \quad (3)$$

Wartości wariancji (2) mogą być wykorzystane do weryfikacji hipotezy (1) za pomocą testów Bartletta oraz, jeśli liczebności we wszystkich próbkach są jednakowe, czyli gdy $n_1 = n_2 = \dots = n_k$, za pomocą testów Hartleya i Cochрана, jeśli spełnione jest wcześniej wspomniane założenie normalności rozkładów. W praktyce badań statystycznych często spotykamy się z sytuacjami, gdy wspomnianego założenia nie można przyjąć. W takich przypadkach jednym z możliwych rozwiązań jest wykorzystanie metod nieparametrycznych, które nie wymagają spełnienia założenia normalności rozkładu badanych zmiennych. W dalszej części zostanie

zaproponowany test wykorzystujący nieparametryczne metody statystyczne, nie wymagające spełnienia założenia o normalności rozkładów analizowanych zmiennych losowych.

3. Konstrukcja testu wykorzystująca metodę *bootstrap*

Założmy, że badana jest populacja ze względu na zmienną losową X o rozkładzie F . Przyjmijmy, że z populacji pobierana jest próba losowa $(X_1, X_2, \dots, X_n)$, a $(x_1, x_2, \dots, x_n)$ jest realizacją tej próby. Wówczas rozkład prawdopodobieństwa

$$P(X_B = x_k) = \frac{1}{n}, k = 1, 2, \dots, n \quad (4)$$

nosi nazwę rozkładu bootstrapowego z próby (por. np. [Efron, Tibshirani 1993; Domański, Pruska 2000]). Oznaczmy ten rozkład przez $\hat{F}$. Próbą bootstrapową nazywamy wektor losowy $\mathbf{x}^* = (X_1^*, X_2^*, \dots, X_N^*)$, taki że X_i^* ($i = 1, 2, \dots, N$) są niezależnymi zmiennymi losowymi o rozkładzie $\hat{F}$ (4), a N jest ustaloną liczbą naturalną. Realizację tego wektora losowego oznaczmy przez $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_N^*)$. Metody bootstrapowe pozwalają na szacowanie nieznanymi parametrów populacji oraz umożliwiają weryfikację hipotez, a także ocenę wariancji estymatorów. W oparciu o te metody można również przeprowadzać weryfikacje hipotez statystycznych.

Zakładamy, że z k populacji o rozkładach ciągłych pobrano próby o liczebnościach $n_1, n_2, \dots, n_k$. Dla tak uzyskanej próby wyznaczana jest wartość statystyki jak w teście Hartleya

$$T = \frac{\max S_i^2}{\min S_i^2}, \quad (5)$$

gdzie S_i^2 jest zadana (2), a

$$S^2 = \frac{1}{n-k} \sum_{i=1}^k (n_i - 1) S_i^2, \quad n = \sum_{i=1}^k n_i.$$

Odpowiednio duże wartości statystyki T świadczą przeciw hipotezie (1). Jeżeli rozważane zmienne losowe mają rozkłady normalne, wartości krytyczne można odczytać z tablic wartości krytycznych testu Hartleya. Jeżeli nie jest spełnione założenie normalności rozkładu, pozostaje ustalić wartość krytyczną pozwalającą na podjęcie decyzji o braku podstaw lub o odrzuceniu hipotezy o równości wariancji przy ustalonym poziomie istotności α . W tym celu na podstawie uzyskanej próby wyznaczane są oceny wariancji S_i^2 ($i = 1, 2, \dots, k$). Następnie wartości z próby podlegają następującej transformacji:

$$\tilde{x}_{ij} = \frac{x_{ij}}{S_i}, \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n_i). \quad (6)$$

Po takiej transformacji wariancje dla wszystkich zmiennych są jednakowe. Jeżeli tak określone próby będą w dalszej części traktowane jako populacje, z których będą losowane próby bootstrapowe, to dla zmiennych określonych (6) prawdziwa jest H_0 (por. np. [Chou 1996]).

W celu określenia wartości krytycznych statystyki (5) ze zmodyfikowanej próby losowane są próby bootstrapowe. Każdą spośród k tak zmodyfikowanych prób traktujemy jako sztuczną populację i ze wszystkich prób dokonujemy N -krotnego ($N \geq 1000$) losowania zwrotnego. Dla i -tej próby ($i = 1, 2, \dots, k$) każdorazowo losujemy n_i elementów, tworząc w ten sposób próbę bootstrapową. Dla każdej takiej próby bootstrapowej wyznaczamy średnią i wariancję na podstawie (3) i (2). Tym sposobem dla każdej zmiennej X_i ($i = 1, 2, \dots, k$) otrzymujemy N -elementowy zbiór wartości wariancji bootstrapowych.

$$U_i = \{S_{i1}^{2*}, S_{i2}^{2*}, \dots, S_{iN}^{2*}\}, \quad (7)$$

gdzie

$$S_{is}^{2*} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\tilde{x}_{ijs}^* - \bar{\tilde{x}}_{is}^*)^2 \quad (8)$$

jest wariancją s -tej próby bootstrapowej dla i -tej zmiennej oraz

$$\bar{\tilde{x}}_{is}^* = \frac{1}{n_i} \sum_{j=1}^{n_i} \tilde{x}_{ijs}^* \quad (9)$$

jest średnią s -tej próby bootstrapowej dla j -tej zmiennej, a $\tilde{x}_{ijs}^*$ jest j -tą obserwacją w s -tej próbie bootstrapowej dla i -tej zmiennej ($i = 1, 2, \dots, k$; $s = 1, 2, \dots, N$; $j = 1, 2, \dots, n_i$).

Każdorazowo po wylosowaniu próby bootstrapowej wyznaczana jest wartość statystyki

$$T_s^* = \frac{\max_{is} S_{is}^{2*}}{\min_{is} S_{is}^{2*}}, \quad (s = 1, 2, \dots, N; i = 1, 2, \dots, k). \quad (10)$$

Na podstawie uzyskanych wyników wyznaczamy kwantyle statystyki T^* rzędu $1 - \alpha$, które stanowią podstawę do podjęcia decyzji o ewentualnym odrzuceniu hipotezy H_0 (prawostronny obszar odrzucenia hipotezy).

4. Porównanie klasycznych metod i proponowanej metody (symulacje komputerowe)

W przeprowadzonych symulacjach komputerowych rozważano zmienne losowe o trzech typach rozkładów ciągłych: jednostajnych, trójkątnych oraz beta. W żadnym z tych przypadków nie jest możliwe zastosowanie do weryfikacji hipotezy (1) testów Bartletta, Hartleya i Cochрана ze względu na niespełnienie założenia o normalności rozkładów badanych zmiennych. Parametry analizowanych zmiennych

losowych przedstawiono w tab. 1. Opis symulacji i wyniki przedstawiono w pkt. 4.2. Po przeprowadzeniu tych symulacji dokonano porównania proponowanej metody z klasycznym testem Bartletta przy założeniu normalności rozkładu badanej zmiennej (pkt 4.3).

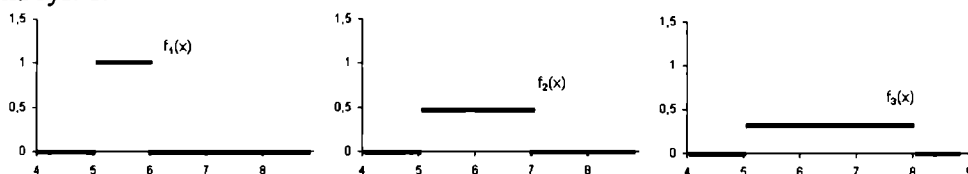
4.1. Charakterystyka badanych rozkładów

W kolejnych etapach symulacji komputerowych rozważano zmienne losowe o rozkładach jednostajnych, trójkątnych oraz beta. W każdym z tych przypadków poddano analizie trzy zmienne losowe X , Y i Z , różniące się wariancjami. Dla poszczególnych typów rozkładów zmienne były następujące:

- Rozkłady jednostajne zadane były następującymi funkcjami gęstości:

$$f(x) = \begin{cases} \frac{1}{c-5}, & x \in [5, c], \\ 0, & x \notin [5, c], \end{cases} \quad \text{gdzie dla zmiennych } X, Y \text{ i } Z \text{ przyjmuje wartości } 6, 7 \text{ i } 8.$$

Funkcje gęstości $f_1(x)$, $f_2(x)$ i $f_3(x)$ zmiennych losowych X , Y i Z przedstawiono na rys. 1.



Rys. 1. Funkcje gęstości zmiennych losowych o rozkładach jednostajnych

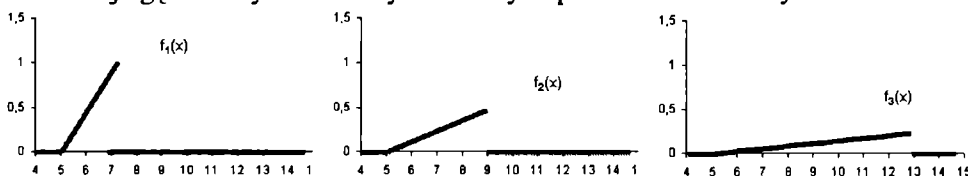
Źródło: obliczenia własne.

- Zmienne losowe o rozkładach trójkątnych zadane były następująco:

$$f(x) = \begin{cases} \frac{2}{(c-5)^2}(x-5), & x \in [5, c], \\ 0, & x \notin [5, c], \end{cases}$$

gdzie dla zmiennych X , Y i Z przyjmuje wartości 7, 9 i 13.

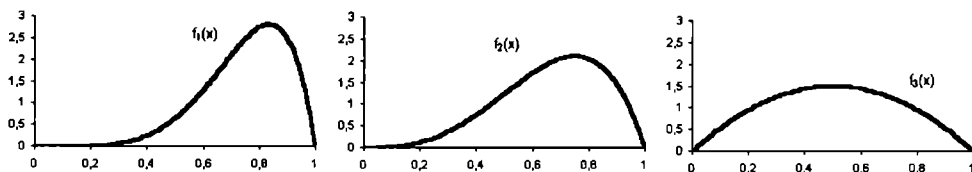
Funkcje gęstości tych zmiennych losowych przedstawiono na rys. 2.



Rys. 2. Funkcje gęstości zmiennych losowych o rozkładach trójkątnych

Źródło: obliczenia własne.

- W trzeciej części symulacji rozważano zmienne losowe o rozkładach beta $B(6,2)$, $B(4,2)$ oraz $B(2,2)$. Odpowiednie funkcje gęstości zostały przedstawione na rys. 3.



Rys. 3. Funkcje gęstości zmiennych losowych o rozkładach beta
Źródło: obliczenia własne.

Wariancje rozważanych zmiennych losowych zostały przedstawione w tab. 1. Największe dysproporcje wariancji występowały w przypadku rozkładów trójkątnych, a najmniejsze w przypadku rozkładów beta. W symulacjach skorzystano z procedur Brandta [1998].

Tabela 1. Podstawowe informacje o analizowanych rozkładach zmiennych losowych

Rozkłady	X		Y		Z		$\frac{D^2 Z}{D^2 X}$
	Rozkład	$D^2 X$		$D^2 Y$		$D^2 Z$	
Jednostajne	J[5,6]	1/12	J[5,7]	2/12	J[5,8]	3/12	3
Trójkątne	T[5,7]	2/9	T[5,9]	8/9	T[5,13]	32/9	16
Beta	B(6,2)	1/48	B(4,2)	2/63	B(2,2)	1/20	2,4

Źródło: obliczenia własne.

4.2. Przebieg symulacji komputerowych

Symulacje komputerowe zostały przeprowadzone dla każdego z wymienionych wcześniej typów rozkładów zmiennych losowych. W każdym przebiegu poddawano analizie trzy zmienne losowe X_1 , X_2 i X_3 o identycznych bądź różnych rozkładach, według następującego schematu (w każdym przebiegu symulacji analizowane zmienne były tej samej klasy):

I przebieg: wszystkie trzy zmienne losowe miały jednakowe rozkłady.

II przebieg: dwie zmienne losowe miały jednakowe rozkłady, a trzecia zmienna losowa charakteryzowała się większą wariancją.

III przebieg: wszystkie trzy zmienne losowe charakteryzowały się różnymi wariancjami spełniającymi warunek: $D^2 X_1 < D^2 X_2 < D^2 X_3$.

Schemat doboru zmiennych losowych w procesie symulacji przedstawiono w tab. 2.

Tabela 2. Schemat układu zmiennych losowych w kolejnych etapach symulacji

Przebieg symulacji	Zmienna losowa		
	X_1	X_2	X_3
I	X	X	X
II	X	X	Y
III	X	Y	Z

Źródło: opracowanie własne.

Dla każdego z rozważanych przypadków w pierwszym przebiegu symulacji wariancje wszystkich trzech zmiennych losowych były jednakowe, czyli hipoteza (1) powinna być odrzucona z prawdopodobieństwem zbliżonym do przyjętego poziomu istotności α . W analizach przyjmowano poziomy istotności 0,1 i 0,05. We wszystkich symulacjach losowo próbkę o liczebności $n_1 = n_2 = n_3 = 10$, przy liczbie powtórzeń $N = 1000$.

4.2.1. Wyniki symulacji

Rezultaty przeprowadzonych symulacji komputerowych przedstawiono w tab. 3 (dla poziomu istotności $\alpha = 0,1$) i w tab. 4 (dla poziomu istotności $\alpha = 0,05$). W obu przypadkach dla I przebiegu hipoteza H_0 była prawdziwa, a dla dwóch pozostałych przebiegów fałszywa. Największe różnice wariancji występowały dla III przebiegu symulacji.

Tabela 3. Częstość odrzucenia H_0 ($\alpha = 0,1$)

Przebieg symulacji	Rozkłady		
	jednostajne	trójkątne	beta
I	0,076	0,112	0,093
II	0,743	0,693	0,146
III	0,968	0,952	0,246

Źródło: symulacje komputerowe.

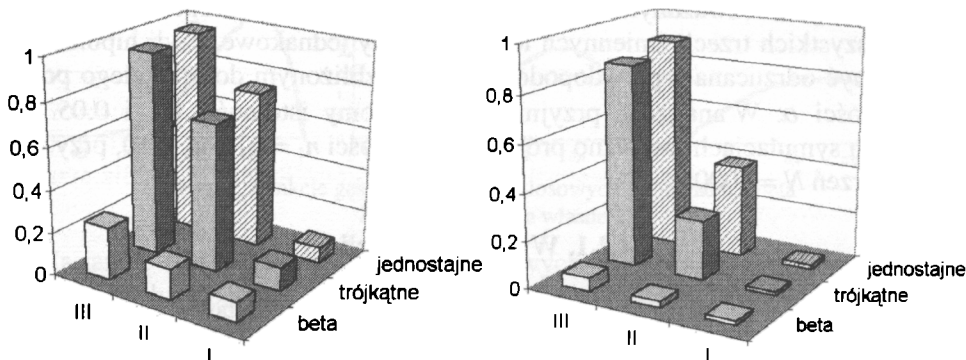
Przy poziomie istotności $\alpha = 0,1$, przy założeniu prawdziwości H_0 , częstości odrzucenia tej hipotezy były zbliżone do poziomu istotności. We wszystkich przebiegach wraz ze wzrostem dysproporcji wariancji zmiennych następował znaczny wzrost częstości odrzuceń H_0 . Najmniej zauważalny wzrost występował dla zmiennych o rozkładzie beta, jednak w tym przypadku dysproporcje wariancji zmiennych losowych były najmniejsze.

Tabela 4. Częstość odrzucenia H_0 ($\alpha = 0,05$)

Przebieg symulacji	Rozkłady		
	jednostajne	trójkątne	beta
I	0,017	0,014	0,017
II	0,392	0,253	0,026
III	0,903	0,861	0,057

Źródło: symulacje komputerowe.

Przy poziomie istotności $\alpha = 0,05$ uzyskano rezultaty zbliżone do poprzedniego przypadku. Należy jednak zwrócić uwagę na stosunkowo niewielkie częstości odrzucenia hipotezy H_0 dla II i III przebiegu symulacji dla rozkładów beta.



Rys. 4. Częstość odrzucenia hipotezy H_0 dla $\alpha = 0,1$ (po lewej) oraz $\alpha = 0,05$ (po prawej)

Źródło: obliczenia własne.

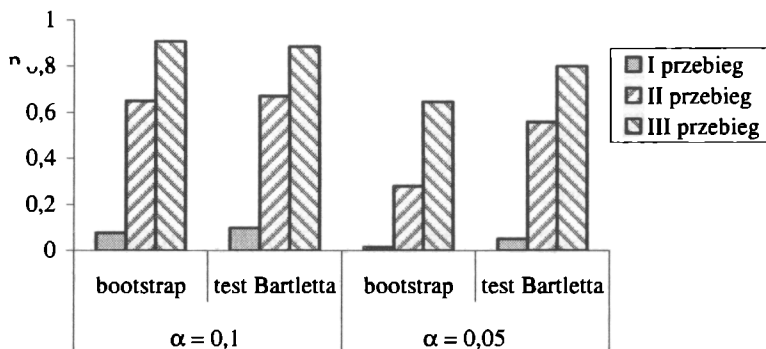
4.2.2. Porównanie proponowanej metody i klasycznego testu Bartletta

W ostatnim etapie symulacji przeprowadzono porównanie proponowanej metody oraz klasycznego testu Bartletta. W tym przypadku przyjęto założenie normalności rozkładów. Rozważano następujące rozkłady zmiennych losowych X : $N(10,1)$, Y : $N(10,2)$ oraz Z : $N(10,3)$. W poszczególnych przebiegach zmienne losowe były przyporządkowane według schematu zamieszczonego w tab. 2. Wyniki symulacji przedstawiono w tab. 5.

Tabela 5. Częstość odrzucenia H_0

Przebieg symulacji	$\alpha = 0,1$		$\alpha = 0,05$	
	<i>bootstrap</i>	test Bartletta	<i>bootstrap</i>	test Bartletta
I	0,078	0,0985	0,015	0,0512
II	0,652	0,6725	0,280	0,5588
III	0,908	0,8862	0,647	0,8001

Źródło: symulacje komputerowe.



Rys. 5. Oceny prawdopodobieństw odrzucenia H_0 dla proponowanego testu i testu Bartletta

Źródło: obliczenia własne.

Porównanie proponowanej metody z testem Bartletta pozwala stwierdzić, że w przypadku jednakowych wariancji obserwowanych zmiennych (I przebieg) proponowany test jest bardziej zachowawczy niż test Bartletta, szczególnie przy poziomie istotności 0,05. W przypadku gdy nie wszystkie wariancje są jednakowe (II i III przebieg), rezultaty uzyskiwane z wykorzystaniem proponowanego testu są zbliżone do wyników uzyskiwanych klasycznym testem. Rezultaty analiz symulacyjnych zostały również przedstawione na rys. 5.

5. Wnioski

Metody statystyczne mogą być stosowane tylko wówczas, gdy spełnione są wszystkie wymagane założenia. Przy testowaniu hipotez statystycznych bardzo często musi być spełnione założenie normalności rozkładu testowanych zmiennych. Jeśli to założenie nie jest spełnione, nie ma podstaw do stosowania tej metody. Klasycznie stosowane testy równości wariancji wymagają, by rozkłady badanych zmiennych były normalne.

W artykule przedstawiono propozycję testu równości wariancji dla k zmiennych. Ponieważ w proponowanym teście wykorzystano metodę *bootstrap*, można było osłabić wspomniane wyżej założenia, wymagając jedynie ciągłości rozkładów badanych zmiennych. Jak wykazały symulacje komputerowe, proponowany test pozwala przeprowadzić weryfikację hipotezy o równości wariancji wielu zmiennych o rozkładach ciągłych. Dla rozważanych rozkładów, gdy prawdziwa była hipoteza H_0 o równości wariancji, proponowany test był zbyt zachowawczy, tzn. hipoteza H_0 była odrzucana rzadziej niż wynikałoby to z przyjętego poziomu istotności. W miarę różnicowania się wariancji zmiennych znacznie częściej odrzucana była hipoteza H_0 .

Przeprowadzono symulacyjne porównanie proponowanego testu oraz klasycznego testu Bartletta dla równości wariancji dla zmiennych o rozkładach normalnych. Podobnie jak poprzednio, proponowany test był zbyt zachowawczy, a w przypadku różnych wariancji test proponowany był nieco słabszy od klasycznego testu Bartletta. Jeśli jednak nie są spełnione założenia dotyczące normalności rozkładu zmiennych, to proponowany test daje możliwość przeprowadzenia weryfikacji hipotezy głoszącej równość wariancji.

Literatura

- Brandt S. (1998), *Analiza danych*, Wydawnictwo Naukowe PWN, Warszawa.
Chou P.H. (1996), *Using Bootstrap to Test Mean-Variance Efficiency of a Given Portfolio*, Economics Working Paper Archive at WUSTL, nr 9609002.

- Domański Cz., Pruska K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.
- Efron B., Tibshirani R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- Kendall M.G., Stuart A. (1963), *The Advanced Theory of Statistics*, vol. I, II, Charles Griffin, Londyn.

ON THE BOOTSTRAP TEST OF HOMOGENITY VARIANCES

Summary

The assumption of homogeneity of variance is supposed in various statistical methods such as ANOVA, discriminant analysis, cluster analysis, etc. The paper presents a test for homogeneity of variance using bootstrap methods that does not rely on normal distribution assumption. There are presented analytical results and the results of Monte Carlo study. There are compared the proposed test and classical Hartley's homogeneity variance test.