PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 3-4 – 2009

SABINA DENKOWSKA, KAMIL FIJOREK, MARCIN SALAMAGA, ANDRZEJ SOKOŁOWSKI

EMPIRYCZNA OCENA MOCY TESTÓW DLA WIELU WARIANCJI

1. WPROWADZENIE

Testy dla wielu wariancji nie są zbyt popularne wśród testów statystycznych. Może dlatego, że rzadko wykorzystywane są jako procedura docelowa. Raczej służą do weryfikowania założenia o równości wariancji wymaganego przez inne procedury – przede wszystkim analizę wariancji. Mogą być też wykorzystywane do oceny jednorodności źródeł danych, które chcemy połączyć dla uzyskania lepszych ocen wariancji.

Testy dla wielu wariancji też zazwyczaj same wymagają spełnienia pewnych założeń. Najczęściej chodzi tu o założenie normalności rozkładu w populacjach lub przynajmniej rozkładu tego samego typu.

Oczywiście, zagadnienie porównywania rzeczywistego prawdopodobieństwa błędu pierwszego rodzaju oraz mocy testów dla wielu wariancji było już podejmowane w literaturze statystycznej. Chyba najbardziej wyczerpującym doświadczeniem były symulacje, których wyniki przedstawili Conover, Johnson, Johnson (1981)[1]. Poddali oni analizie m.in. następujące testy: Neymana i Pearsona (1931), Bartletta (1937), Cochrana (1941), Scheffego (1959), Levene'a (1960), Browna i Forsythe'a (1974).

Zasadniczym celem artykułu jest zbadanie zachowania się wybranych testów dla wielu wariancji w warunkach prawdziwości i nieprawdziwości założenia o normalności rozkładu, małych i dużych prób, prób równolicznych i prób o zróżnicowanych liczebnościach. Badania te przeprowadzono metodami symulacyjnymi.

Z bogatego zbioru testów dla wielu wariancji do analizy wybrano testy wyróżnione w badaniach Conovera, Johnsona, Johnsona z 1981 r. (test Levene'a i test Flingera-Killeena), test Browna-Forsythe'a, który jest często stosowany ze względu na dostępność w komputerowych pakietach statystycznych oraz relatywnie nową propozycję O'Briena z 1981 r.[2]

2. WYBRANE TESTY JEDNORODNOŚCI WARIANCJI

Założenie jednorodności (równości) wariancji w obrębie grup jest jednym z wymogów ważnych procedur wnioskowania statystycznego. Spełnienie tego założenia jest wymagane m.in. w przypadku stosowania jednoczynnikowej analizy wariancji ANOVA.

[1] Por. [2].
[2] Por. [1], s. 1.

Do bardziej znanych testów jednorodności wariancji obok testów Hartleya, Cochrana i Bartletta należy procedura Levene'a[3].

W teście homogeniczności wariancji weryfikujemy następujące hipotezy statystyczne:

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$

$H_1: \sigma_i^2 \neq \sigma_j^2$ przynajmniej dla jednej pary (i, j).

Założenia procedury Levene'a (1960) są następujące:

1. Cecha X ma w i-tej populacji rozkład normalny $N(\mu_i, \sigma_i^2)$, gdzie $i = 1, 2, \ldots, k$,
2. Wartości przeciętne μ_i w k populacjach są nieznane.

W teście Levene'a dla każdej obserwacji X_{ij} wyznaczane jest odchylenie absolutne względem wartości średniej $\bar{X}_i$ w i-tej grupie:

$$Z_{ij} = |X_{ij} - \bar{X}_i|, \; j = 1, 2, \ldots, n_i, \; i = 1, 2, \ldots, k. \tag{1}$$

Statystyka testowa zaproponowana przez Levene'a ma postać[4]:

$$W = \frac{N-1}{k-1} \frac{\sum_{i=1}^{k} n_i (\bar{Z}_i - \bar{Z})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2}, \tag{2}$$

przy czym: $N = \sum_{i=1}^{k} n_i$, natomiast $\bar{Z}_i$ oraz $\bar{Z}$ oznaczają odpowiednio średnią wewnątrzgrupową i średnią międzygrupową. Przy założeniu prawdziwości hipotezy zerowej statystyka W ma rozkład Fishera o $k-1$ oraz $N-k$ stopniach swobody. Jeżeli na poziomie istotności α empiryczna wartość statystyki W przekracza wartość krytyczną $F_{\alpha;k-1;N-k}$, to hipotezę zerową odrzucamy. Test Levene'a został potem zmodyfikowany przez Browna i Forsythe'a (1974) poprzez zastąpienie średniej wewnątrzgrupowej $\bar{X}_i$ w przekształceniu (1) medianą lub średnią uciętą. Autorzy modyfikacji testu Levene'a zalecają przede wszystkim stosowanie procedury z medianą grupową, gdyż zapewnia ona, ich zdaniem, test o dość wysokiej mocy, który jest odporny na brak normalności w rozkładzie danych.

Kolejną procedurą zaproponowaną do testowania jednorodności wariancji w k-populacjach jest test Flignera-Killeena (1974). Ta procedura wymaga utworzenia rankingu bezwzględnych wartości cechy X_{ij}. Uporządkowanym niemalejąco wartościom $|X_{ij}|$ przypisuje się następnie wskaźniki a_{Ni} obliczane następująco[5]:

$$a_{Ni} = \Phi^{-1}\left(\frac{1}{2} + \frac{i}{2(N+1)}\right). \tag{3}$$

[3] Por. [5], s. 278-292.
[4] Por. [7], s. 49.
[5] Por. [6], s. 3.

Warto zaznaczyć, że Conover, Johnson, Johnson (1981) zaproponowali rangowanie wartości bezwzględnych cechy X_{ij} po uprzednim odjęciu median grupowych.

Statystyka testowa w procedurze Flignera-Killeena (1974) ma postać:

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i\left(\bar{A}_i - \bar{a}\right)^2}{V^2}, \tag{4}$$

przy czym $\bar{A}_i$ jest średnią wewnątrzgrupową, $\bar{a}$ jest średnią międzygrupową oraz V^2 jest wariancją międzygrupową. Statystyka (4) ma asymptotyczny rozkład chi-kwadrat o $k-1$ stopniach swobody. Jeżeli na poziomie istotności α empiryczna wartość statystyki χ^2 przekracza wartość krytyczną $\chi^2_{\alpha,k-1}$, to hipotezę zerową (głoszącą równość wariancji w k populacjach) odrzucamy.

Nieco inne podejście w testowaniu jednorodności wariancji zaproponował O'Brien (1979, 1981). Procedura O'Briena wykorzystuje jednoczynnikową analizę wariancji ANOVA, która jest stosowana dla przekształconych wartości cechy X_{ij} zgodnie ze wzorem[6]:

$$Z_{ij} = \frac{n_i(n_i - 1{,}5)\left(X_{ij} - \bar{X}_i\right)^2 - 0{,}5\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2}{(n_i - 1)(n_i - 2)}, \tag{5}$$

$$j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, k,$$

przy czym:
n_i – liczebność i-tej grupy,
$\bar{X}_i$ – wartości średniej w i-tej grupie.

Do weryfikacji hipotezy o równości wariancji stosowany jest potem test F Fishera:

$$F = \frac{(N-k)\sum_{i=1}^{k} n_i\left(\bar{Z}_i - \bar{Z}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \bar{Z}_i\right)^2}, \tag{6}$$

przy czym: $N = \sum_{i=1}^{k} n_i$, natomiast $\bar{Z}_i$ oraz $\bar{Z}$ oznaczają odpowiednio średnią wewnątrzgrupową i międzygrupową.

Statystyka (6) ma rozkład F o $k-1$ i $N-k$ stopniach swobody. Jeżeli empiryczna wartość statystyki (6) przekracza wartość krytyczną $F_{\alpha;k-1;N-k}$, to hipotezę zerową o równości wariancji w k populacjach odrzucamy.

[6] Por. [1], s. 3.

3. METODOLOGIA BADAŃ SYMULACYJNYCH

Celem przeprowadzenia badań symulacyjnych było określenie zachowania się wybranych testów jednorodności wariancji rozumianego jako zdolność (lub jej brak) do utrzymania nominalnego prawdopodobieństwa popełnienia błędu I rodzaju oraz zdolność do uzyskiwania wysokiego prawdopodobieństwa odrzucenia fałszywej hipotezy zerowej (moc testu) w warunkach zmieniających się liczebności prób oraz zmieniających się konfiguracji wielkości wariancji w poszczególnych próbach. W badaniu rozważono cztery testy jednorodności wariancji: Levene'a, Browna-Forsythe'a, Flignera-Killeena oraz O'Briena.

Wykorzystanie symulacji komputerowych jest podyktowane praktyczną niemożliwością analitycznego wyznaczenia krzywych mocy dla wybranych testów jednorodności wariancji.

W przeprowadzonym badaniu rozważono trzy rozkłady prawdopodobieństwa (mechanizmy generujące próby):

– rozkład normalny,
– rozkład logarytmiczno-normalny,
– rozkład jednostajny.

W przypadku testów jednorodności wariancji wielkość wartości oczekiwanej nie ma wpływu na wyniki analizy. W badaniach przyjęto arbitralnie dla rozkładu logarytmiczno-normalnego oraz normalnego wartość oczekiwaną równą 10. W przypadku rozkładu jednostajnego arbitralnie przyjęto wartość 0 jako dolną granicę rozkładu. Uzyskanie pożądanych wartości odchylenia standardowego było możliwe poprzez odpowiednie sterowanie górną granicą przedziału, na którym jest określony rozkład jednostajny.

W przeprowadzonej symulacji założono, że obserwacje pochodzą z czterech populacji. Ponadto zbudowano trzy mechanizmy określania wielkości odchylenia standardowego w każdej z czterech populacji (oznaczonych dolnymi subskryptami):

– $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\sigma_4 = 1{,}0 + 3^*$ (współczynnik wzrostu odchylenia standardowego),

– $\sigma_1 = \sigma_2 = 1$, $\sigma_3 = \sigma_4 = 1{,}0 + 3^*$ (współczynnik wzrostu odchylenia standardowego),

– $\sigma_i = 1{,}0 + (i - 1)^*$ (współczynnik wzrostu odchylenia standardowego), $i = 1, 2, 3, 4$.

Współczynnik wzrostu odchylenia standardowego przyjmuje wartości od 0 do 0,35 z krokiem 0,025.

W badaniu założono następujące sześć możliwych konfiguracji liczebności prób: (10, 10, 10, 10), (40, 40, 40, 40), (5, 10, 15, 20), (20, 15, 10, 5), (30, 40, 50, 60), (60, 50, 40, 30). Wykorzystanie zarówno prób równolicznych, jak i nierównolicznych jest podyktowane chęcią zbadania wpływu tego czynnika na zachowanie się rozważanych testów statystycznych.

Ponadto uwzględnienie liczebności prób (5, 10, 15, 20), (20, 15, 10, 5), (30, 40, 50, 60), (60, 50, 40, 30) pozwala zbadać łączny wpływ wielkości odchylenia standardowego w danej próbie oraz liczebności tej próby na rozważane testy jednorodności wariancji.

Wszystkie symulacje przeprowadzono w środowisku statystycznym R (www.r-project.org) w oparciu o autorskie skrypty obliczeniowe.

4. WYNIKI BADAŃ

4.1. KONTROLA BŁĘDU PIERWSZEGO RODZAJU

W sytuacji prawdziwości hipotezy zerowej badania symulacyjne pozwoliły oszacować rozmiary rozpatrywanych testów[7] jednorodności wariancji. W celu oszacowania prawdopodobieństwa popełnienia błędu I rodzaju dziesięć tysięcy razy generowano cztery próby o różnej liczności z wybranych rozkładów i za każdym razem badano, która z procedur nie rozpozna stanu faktycznego, czyli równości wariancji we wszystkich grupach. W tabelach 1-3 przedstawione są wyniki otrzymane w zależności od liczebności podgrup.

Ważnym założeniem większości testów jednorodności wariancji jest założenie mówiące o normalności populacji, z których losujemy próby. W prowadzonych badaniach symulacyjnych szacowano prawdopodobieństwo błędu I rodzaju w sytuacji, gdy założenie o normalności jest spełnione (patrz: tabela 1), ale również badano wrażliwość tych procedur na niespełnienie tego założenia. Tabela 2 prezentuje wyniki dla prób generowanych z rozkładu lognormalnego, a tabela 3 z rozkładu jednostajnego.

Tabela 1

Oceny prawdopodobieństw odrzucenia prawdziwej hipotezy zerowej – równe odchylenia standardowe, rozkłady normalne

Liczebności grup				TEST			
I	II	III	IV	Levene'a	B-F	F-K	O'Briena
10	10	10	10	**0,0694**	0,0346	0,0358	0,0389
40	40	40	40	**0,0547**	0,0430	0,0424	0,0453
5	10	15	20	**0,0814**	0,0368	0,0379	**0,0608**
30	40	50	60	**0,0616**	**0,0501**	**0,0511**	**0,0524**

Źródło: opracowanie własne.

W tabeli 1 przedstawiono wyniki uzyskane, gdy próby były losowane z populacji normalnych o równych wariancjach. Najgorzej wypadł test Levene'a, który dla wszystkich rozpatrywanych wariantów liczebności przekroczył przyjęty na wstępie badań poziom istotności 0,05. Szczególnie wysokie prawdopodobieństwo błędnego odrzucenia prawdziwej hipotezy zerowej można zauważyć dla małych, nierównolicznych prób. Procedura ta nie wypada dobrze również w przypadku dużych, ale nierównolicznych próbek. Najlepiej w tej części badań wypadają procedury Browna-Forsythe'a oraz Flignera-Killeena. Obie zapewniają kontrolę błędu I rodzaju na poziomie 0,05, nieznacznie przekraczając to prawdopodobieństwo w przypadku dużych, nierównolicznych prób. Pomiędzy procedurami Browna-Forsythe'a oraz Flignera-Killeena uplasowała się

[7] Patrz [4], s. 23.

procedura O'Briena, która zapewnia kontrolę błędu I rodzaju dla prób równolicznych, również małych. W przypadku próbek o różnej liczności oszacowane prawdopodobieństwa, mimo iż większe od 0,05, są jednak znacznie niższe od prawdopodobieństw otrzymanych przy zastosowaniu testu Levene'a.

Tabele 2 i 3 przedstawiają wyniki w przypadku, gdy niespełnione są założenia o normalności rozkładów.

W tabeli 2 przedstawione są wyniki symulacji prowadzonych dla prób generowanych z rozkładu lognormalnego. Zaskakują bardzo złe wyniki testu Levene'a. Oszacowane prawdopodobieństwa popełnienia błędu I rodzaju przekraczają 0,3! Okazuje się, że zastosowanie tego testu do prób z rozkładu lognormalnego powoduje odrzucenie co najmniej 30% prawdziwych hipotez zerowych. Najbardziej odporny okazał się test Browna-Forsythe'a, który jedynie w sytuacji małych, nierównolicznych prób przekroczył założony na wstępie poziom istotności 0,05. W miarę odporny okazał się również test O'Briana, który kontrolę prawdopodobieństwa odrzucenia hipotezy prawdziwej zapewniał w przypadku dużych prób. Niestety, prawdopodobieństwo błędu I rodzaju dwukrotnie przekracza dopuszczalny poziom istotności w przypadku zastosowania tego testu do małych nierównolicznych próbek. Test Flignera-Killeena we wszystkich rozpatrywanych wariantach liczebności ponad dwukrotnie przekraczał α.

Tabela 2

Oceny prawdopodobieństw odrzucenia prawdziwej hipotezy zerowej – równe odchylenia standardowe, rozkłady lognormalne

Liczebności grup				TEST			
I	II	III	IV	Levene'a	B-F	F-K	O'Briena
10	10	10	10	**0,3438**	0,039	**0,1205**	**0,0584**
40	40	40	40	**0,3013**	0,0428	**0,1533**	0,0344
5	10	15	20	**0,3328**	**0,057**	**0,1175**	**0,1044**
30	40	50	60	**0,3059**	0,0485	**0,1654**	0,0405

Źródło: opracowanie własne.

Tabela 3 przedstawia wyniki uzyskane dla prób losowanych z rozkładu jednostajnego. I w tym przypadku najgorzej wypadła procedura Levene'a, która dla wszystkich czterech wariantów liczebności dała oszacowania przekraczające 0,05. Uzyskane przez tę procedurę wyniki są zbliżone do rezultatów uzyskanych dla rozkładu normalnego (por. tabela 1) i potwierdza się spostrzeżenie, że procedura ta nie wypada dobrze w przypadku małych, nierównolicznych prób, jak również w przypadku dużych, ale nierównolicznych prób. W sytuacji rozkładu prostokątnego zarówno procedura Browna-Forsythe'a, jak i Flignera-Killeena wypadły w badaniach bardzo dobrze, uzyskując prawdopodobieństwa błędnej decyzji poniżej 0,05. Dla testu O'Briena, podobnie jak w przypadku rozkładu normalnego, otrzymano nieznaczne przeszacowanie prawdopodobieństwa błędu I rodzaju w przypadku prób nierównolicznych.

Tabela 3

Oceny prawdopodobieństw odrzucenia prawdziwej hipotezy zerowej – równe odchylenia standardowe, rozkłady jednostajne

Liczebności grup				TEST			
I	II	III	IV	Levene'a	B-F	F-K	O'Briena
10	10	10	10	**0,0709**	0,025	0,0232	0,0411
40	40	40	40	**0,0562**	0,033	0,0332	0,0471
5	10	15	20	**0,097**	0,0374	0,0309	**0,0665**
30	40	50	60	**0,0607**	0,0389	0,0379	**0,0523**

Źródło: opracowanie własne.

W prowadzonych badaniach rozważano trzy rozkłady, z których generowano próby losowe. Okazało się, że wyniki jakie uzyskano dla rozkładu jednostajnego tylko nieznacznie różnią się od wyników otrzymanych dla rozkładu normalnego, który stanowi kluczowe założenie większości procedur służących do badania jednorodności wariancji. Natomiast bardzo istotne różnice wystąpiły w przypadku zastosowania testów do prób wygenerowanych z rozkładu lognormalnego. W przypadku gdy $\sigma \geq 1$ rozkład lognormalny charakteryzuje się silną prawostronną asymetrią i wydaje się, że właśnie to odejście od symetrii jest powodem tak znacznego pogorszenia oszacowań błędów I rodzaju procedur Levene'a i Flignera-Killeena.

Podsumowując część badań dotyczącą kontroli błędu I rodzaju możemy stwierdzić, że pod tym względem zdecydowanie najgorzej wypadła procedura Levene'a. Okazała się „nieodporna" w przypadku nierównych lub nielicznych prób, nawet tych losowanych z rozkładu normalnego. W przypadku rozkładu asymetrycznego jakim jest rozkład lognormalny, wyniki tej procedury pogarszają się tak drastycznie, że jej stosowanie wydaje się wręcz niedopuszczalne. Najlepiej spośród rozważanych testów wypadł test Browna-Forsythe'a. Test ten okazał się znacznie bardziej od innych procedur odporny na nierównoliczność prób, czy też odchylenia od normalności. Stosunkowo dobrze wypadł on również w przypadku silnej asymetrii rozkładu lognormalnego, przekraczając zaledwie o 0,007 założony poziom istotności dla prób małych i nierównolicznych. Z pozostałych testów ciekawie wypada test O'Briena, dla którego oszacowane prawdopodobieństwa testowe oscylują koło 0,05, nieznacznie przekraczając tę wartość dla prób nierównolicznych. W przypadku silnej asymetrii rozkładu lognormalnego test ten okazał się nieodporny tylko w przypadku małych prób, w przeciwieństwie do testu Flignera-Killeena, dla którego wszystkie oszacowane prawdopodobieństwa odrzucenia prawdziwej hipotezy zerowej okazały się ponad dwukrotnie większe od przyjętego poziomu istotności.

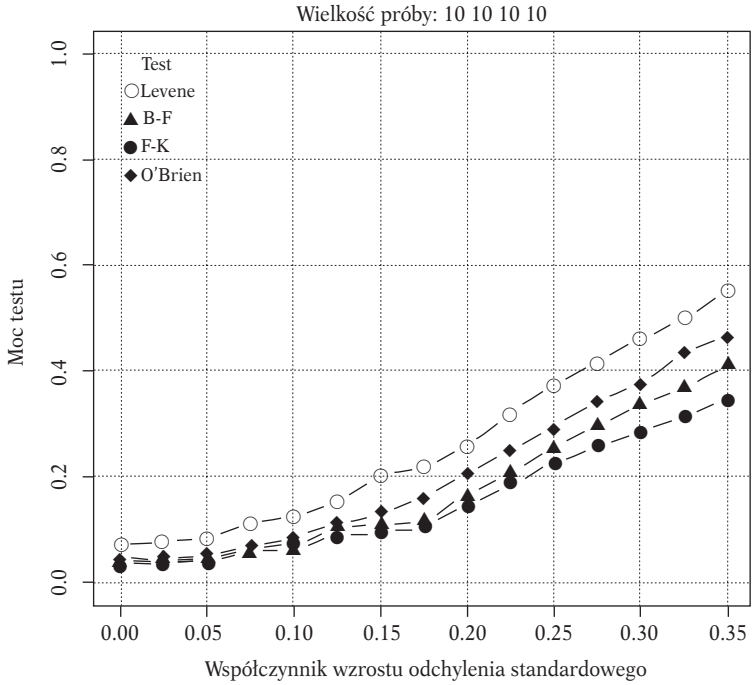
4.2. BADANIE MOCY PROCEDUR

Mocą testu nazywamy prawdopodobieństwo podjęcia słusznej decyzji polegającej na odrzuceniu fałszywej hipotezy zerowej. Druga część badań miała na celu porów-

nanie mocy wybranych testów służących do badania równości wielu wariancji. W tym przypadku tysiąc razy generowano cztery próby o różnej liczebności z wybranych rozkładów i za każdym razem badano, która z procedur rozpozna stan faktyczny, czyli brak równości wariancji we wszystkich grupach. Dla każdej z procedur szacowano prawdopodobieństwa wykrycia niejednorodności wariancji.

Istotnym założeniem wielu testów jednorodności wariancji jest założenie dotyczące normalności rozkładów, z których pobierane są próby. Niestety, w praktyce często założenie to nie jest spełnione, a mimo to procedury są stosowane. W badaniach sprawdzano więc moc procedur nie tylko w sytuacji, gdy próby są generowane z rozkładów normalnych. Generowano również próby z rozkładu jednostajnego i lognormalnego. Tak więc symulacje miały na celu również ocenę odporności wybranych testów w przypadku niespełnienia założenia o normalności.

Na wstępie, w celu porównania mocy procedur w sytuacji spełnionych założeń, eksperymenty przeprowadzono dla prób generowanych z rozkładów normalnych. Wszystkie cztery rozpatrywane testy osiągały porównywalne rezultaty w przypadku dużych prób. Oczywiście, prawdopodobieństwa rozpoznania niejednorodności wariancji rosły wraz ze wzrostem różnic pomiędzy wariancjami. Jednak wykrywalność niejednorodności wariancji okazała się wyraźnie lepsza, gdy rosła wariancja tylko w jednej lub dwóch grupach. Dla obu tych sytuacji oszacowane prawdopodobieństwa różniły się nieznacznie.

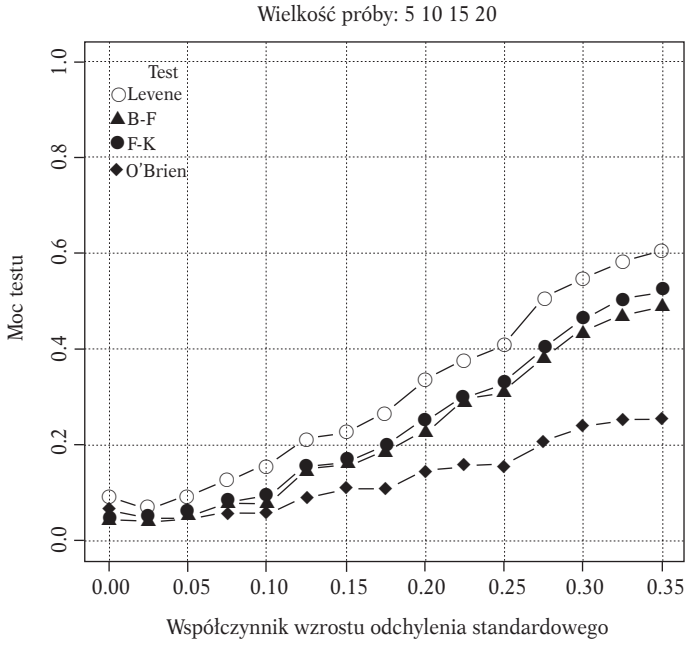


Rysunek 1. Oceny mocy testów dla prób z rozkładu normalnego. Odchylenia standardowe wynoszą odpowiednio: $\sigma_1 = \sigma_2 = \sigma_3 = 1$ i $\sigma_4 = 1 + 3^*$ (współczynnik wzrostu odchylenia standardowego)

Źródło: opracowanie własne.

Wyraźne różnice pomiędzy testami dało się zaobserwować w wykrywalności niejednorodności wariancji w przypadku małych prób. W przypadku prób równolicznych 10-elementowych zdecydowanie najlepszym rozpoznaniem niejednorodności wykazała się procedura Levene'a. Procedura ta jednak nie zapewnia kontroli błędu I rodzaju na założonym poziomie istotności, co stanowi jej poważny mankament. Dla pozostałych trzech procedur różnice w oszacowanych prawdopodobieństwach były nieznaczne, ale wyraźnie gorsze od prawdopodobieństw otrzymanych dla procedury Levene'a. Tylko w sytuacji, gdy wariancja rosła w jednej grupie (patrz: rys. 1) różnice miedzy testami były wyraźniejsze i spośród testów kontrolujących prawdopodobieństwo błędu I rodzaju najlepiej wypadł test O'Briena, trochę gorzej test Browna-Forsythe'a, a najgorzej test Flignera-Killeena. Oczywiście, różnice pomiędzy testami rosły wraz ze wzrostem wariancji w grupie.

W przypadku małych nierównolicznych prób okazało się, iż istotny wpływ na rozpoznanie niejednorodności wariancji przez różne testy ma fakt, czy mniejszym próbom w eksperymencie odpowiadały mniejsze czy większe wariancje. I tak np. procedura O'Briena była najlepsza, gdy większym liczebnościom odpowiadały mniejsze wariancje, podczas gdy w sytuacji odwrotnej wypadała ona najgorzej w rankingu testów jednorodności wariancji (patrz: rys. 2).



Rysunek 2. Oceny mocy testów dla prób z rozkładu normalnego. Odchylenia standardowe w poszczególnych grupach wynoszą odpowiednio: $\sigma_1 = \sigma_2 = 1$ i $\sigma_3 = \sigma_4 = 1 + 3^*$ (współczynnik wzrostu odchylenia standardowego)

Źródło: opracowanie własne.

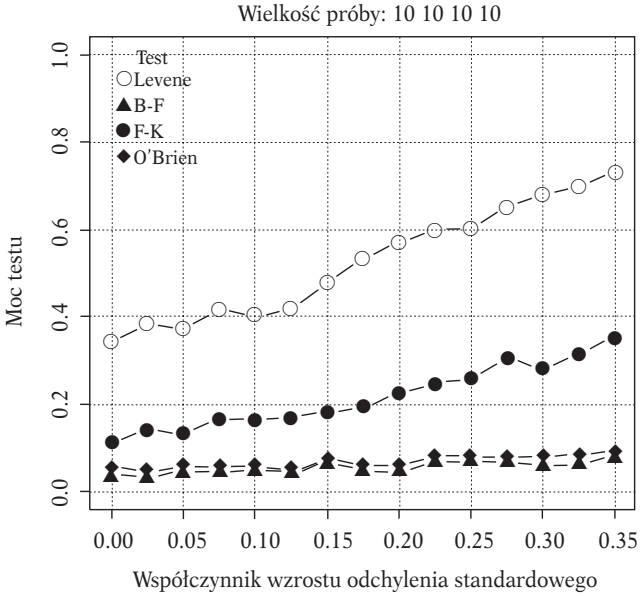
Podsumowując, eksperymenty symulacyjne przeprowadzone na próbach generowanych z rozkładów normalnych, można stwierdzić, iż w przypadku dużych prób wszystkie badane procedury miały zbliżone wyniki rozpoznania niejednorodności podgrup. W przypadku małych prób sytuacja była bardziej skomplikowana. Najlepsze prawdopodobieństwa rozpoznania niejednorodności miał test Levene'a, który jednak nie zapewniał kontroli błędu I rodzaju na założonym poziomie istotności.

W przypadku procedury O'Briena wyniki, w porównaniu do pozostałych testów, wahały się od bardzo dobrych do bardzo słabych, a zależało to od tego, czy liczniejszym próbom odpowiadały mniejsze czy większe wariancje. Wydaje się, że taka „niestabilność" dyskwalifikuje procedurę O'Briena w badaniach praktycznych w przypadku małych prób.

Dobre, stabilne wyniki miała procedura Browna-Forsythe'a, która w przeciwieństwie do procedury Levene'a kontrolowała prawdopodobieństwo odrzucenia prawdziwej hipotezy zerowej na przyjętym w eksperymencie poziomie istotności 0,05.

Ponieważ w badaniach praktycznych nie mamy możliwości rozpoznania, czy mniejszym próbom odpowiada większa czy mniejsza wariancja (chyba, że oprzemy się na obserwacjach dla prób) wydaje się więc, że najbezpieczniej w przypadku małych, nierównolicznych prób decyzję o niejednorodności wariancji przeprowadzać za pomocą np. testu Browna-Forsythe'a.

Kolejne eksperymenty symulacyjne polegały na losowaniu prób z rozkładu lognormalnego i jednostajnego o zadanych parametrach i badaniu wrażliwości testów na odejście od założenia o normalności.



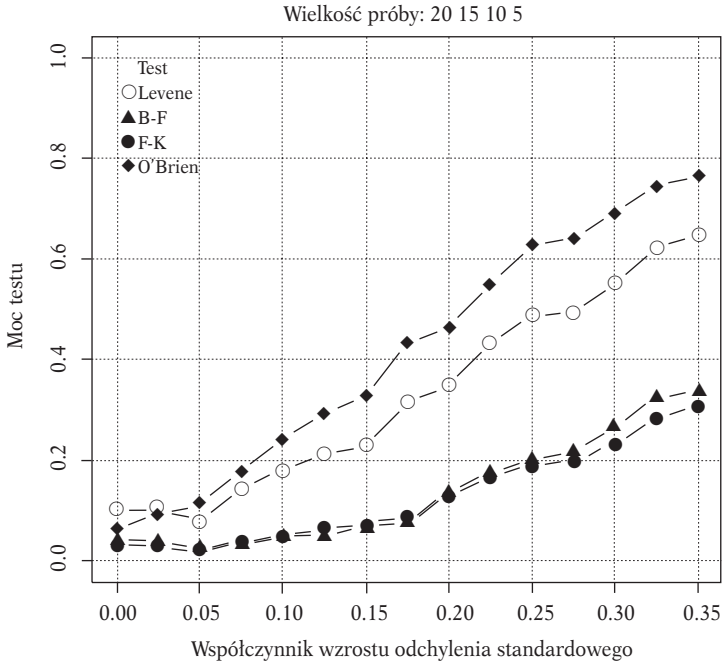
Rysunek 3. Oceny mocy testów dla prób z rozkładu lognormalnego. Odchylenia standardowe wynoszą odpowiednio dla *i*-tej grupy: $\sigma_4 = 1 + (i - 1)^*$ (współczynnik wzrostu odchylenia standardowego)

Źródło: opracowanie własne.

Interesująco, ale zarazem niepokojąco wypadły badania, w których próby były generowane z rozkładu lognormalnego. Zdecydowanie najlepiej niejednorodność wariancji wykrywała procedura Levene'a (por. rys. 3). W przypadku małych, nierównolicznych prób z rozkładu lognormalnego okazało się, że istotna jest zależność pomiędzy liczebnością próby i wielkością wariancji. I tak, w przypadku gdy liczniejszej próbie odpowiada większa wariancja (bez znaczenia jest fakt, czy zmiany wariancji zachodziły w jednej, dwóch czy w trzech grupach) procedura Flignera-Killeena wydaje się jedynym rozwiązaniem. Natomiast gdy mniejszej próbie odpowiada większa wariancja, wyraźnie poprawia się moc pozostałych procedur (Browna-Forsythe'a i O'Briena).

W przypadku dużych prób generowanych z rozkładów jednostajnych wszystkie testy uzyskały prawdopodobieństwa rozpoznania niejednorodności wariancji co najmniej tak dobre, jak w analogicznym eksperymencie z próbami generowanymi z rozkładu normalnego. Wśród rozważanych testów najlepiej wypadła procedura O'Briena.

Procedura O'Briena również w przypadku małych prób dała najlepsze oszacowania prawdopodobieństw rozpoznania niejednorodności wariancji. Szczególne duży kontrast w stosunku do pozostałych procedur pojawiał się wtedy, gdy próby były równoliczne lub gdy mniejszej próbie odpowiadała większa wariancja (rys. 4).



Rysunek 4. Oceny mocy testów dla prób z rozkładu jednostajnego, w przypadku gdy mniejszej próbie odpowiada większa wariancja. Odchylenia standardowe wynoszą dla *i*-tej grupy odpowiednio: $\sigma_4 = 1 + (i - 1)^*$ (współczynnik wzrostu odchylenia standardowego)

Źródło: opracowanie własne.

W najgorszym przypadku, gdy najmniejszej próbie odpowiadała najmniejsza wariancja, procedura O'Briena miała wyniki zbliżone do wyników pozostałych procedur: Browna-Forsythe'a i Flignera-Killeena. Najlepiej w tym eksperymencie wypadła procedura Levene'a, należy jednak wspomnieć, że również w tym eksperymencie test ten dał prawie dwukrotne przeszacowanie przyjętego na wstępie poziomu istotności.

5. PODSUMOWANIE

Przeprowadzone badania symulacyjne potwierdziły, że nie ma jednego najlepszego, czy też „uniwersalnego" testu jednorodności wariancji. Zalecenia autorów, wynikające z przeprowadzonych badań, odnośnie wyboru najwłaściwszego testu jednorodności w zależności od rozważanych czynników, takich jak m.in. rozkład, czy liczebności prób, przedstawione są w tabelach 4-5. W kolumnie „zalecenia" wymienione są procedury w zalecanej kolejności stosowania.

Tabela 4

Zalecenia odnośnie wyboru testu jednorodności wariancji w przypadku losowania prób z rozkładów co najmniej zbliżonych do rozkładu normalnego lub do rozkładu jednostajnego

Liczebności prób		Zalecenia (rozkłady zbliżone do rozkładu normalnego)	Zalecenia (rozkłady zbliżone do rozkładu jednostajnego)
Małe	Równoliczne	1. Test O'Briena 2. Test Browna-Forsythe'a 3. Test Flignera-Killeena 4. Test Levene'a (p = 0,0694)	1. Test O'Briena 2. Test Levene'a (p = 0,0709) 3. Test Browna-Forsythe'a 4. Test Flignera-Killeena
	Nierównoliczne	1. Test Browna-Forsythe'a 2. Test Flignera-Killeena 3. Test Levene'a (p = 0,0814) 4. Test O'Briena (p = 0,0608)	1. Test O'Briena (p = 0,0665) 2. Test Levene'a (p = 0,097) 3. Test Browna-Forsythe'a 4. Test Flignera-Killeena
Duże	Równoliczne	1. Test O'Briena 2. Test Levene'a (p = 0,0547) 3. Test Browna-Forsythe'a 4. Test Flignera-Killeena	1. Test O'Briena 2. Test Flignera-Killeena 3. Test Levene'a (p = 0,0562) 4. Test Browna-Forsythe'a
	Nierównoliczne	1. Test O'Briena (p = 0,0524) 2. Test Levene'a (p = 0,0616) 3. Test Browna-Forsythe'a (p = 0,0501) 4. Test Flignera-Killeena (p = 0,0511)	1. Test O'Briena (p = 0,0523) 2. Test Flignera-Killeena 3. Test Levene'a (p = 0,0607) 4. Test Browna-Forsythe'a

Źródło: opracowanie własne (w nawiasach obok nazw procedur podano oszacowane dla tych procedur prawdopodobieństwa popełnienia błędu I rodzaju większe od 0,05).

Część z zalecanych testów nie zapewnia kontroli błędu I rodzaju. Ponieważ dla rozpatrywanych procedur różnice w oszacowaniach prawdopodobieństw błędu I rodzaju były znaczne w zależności od rodzaju procedury, czy też rozważanej sytuacji badawczej, w nawiasie obok nazwy procedury podano oszacowane prawdopodobieństwa popełnienia błędu I rodzaju, wtedy gdy jest ono większe od 0,05. Kolejność zalecanych testów zależała od kilku czynników. Oczywiście, autorzy starali się wyróżnić testy o największej mocy, zapewniające kontrolę błędu I rodzaju. W niektórych sytuacjach badawczych wybór najlepszych procedur był szczególnie trudny, gdy np. procedury dobrze rozpoznające niejednorodność wariancji, miały zbyt wysokie prawdopodobieństwa błędu I rodzaju. Prawdopodobieństwa popełnienia błędu I rodzaju zostały więc podane, by badacz korzystający z tego opracowania miał orientację, na jaki błąd I rodzaju się naraża dokonując wyboru procedury i ewentualnie rozważył, czy jest to dopuszczalne w prowadzonych przez niego badaniach. Jeśli nie dopuszcza takiego poziomu błędu I rodzaju, może rozważyć zastosowanie testu na „odpowiednio" niższym poziomie istotności lub też zastosować kolejny test z listy zaleceń.

Tabela 5

Zalecenia odnośnie wyboru testu jednorodności wariancji w przypadku losowania prób z rozkładów o wyraźnej asymetrii umiarkowanej

Liczebności prób		Zalecenia
Małe	Równoliczne	Test Flignera-Killeena (p = 0,1205)
	Nierównoliczne	Żaden z testów nie zasługuje na rekomendację, ale „najmniejszym złem" wydaje się test Flignera-Killeena (p = 0,1175).
Duże	Równoliczne i nierównoliczne	1. Test Flignera-Killeena (p = 0,1700) 2. Test Browna-Forsythe'a

Źródło: opracowanie własne (w nawiasach obok nazw procedur podano oszacowane dla tych procedur prawdopodobieństwa popełnienia błędu I rodzaju większe od 0,05).

W tabeli 4 podane są zalecenia w sytuacji, gdy próby są losowane z rozkładów normalnych (lub rozkładów co najmniej zbliżonych do normalnych) oraz dla prób losowanych z rozkładów lognormalnych. W badaniach rozważano odchylenia standardowe, dla których rozkład lognormalny charakteryzuje się wyraźną asymetrią prawostronną. Wydaje się, że zalecenia z tej części badań można odnieść nie tylko do rozkładu lognormalnego, ale do szerszej klasy rozkładów umiarkowanie asymetrycznych.

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] Abdi H., [2007], *O'Brien test for homogeneity of variance*, [w:] N.J. Salkind (red.), Encyclopedia of Measurement and Statistics, Thousand Oaks (CA), Sage, pp. 701-704.

[2] Conover W.J., Johnson M.E., Johnson M.M., [1981], *A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data*, Technometrics, Vol. 23, No. 4, s. 351-361.
[3] Domański Cz., [1986], *Teoretyczne podstawy testów nieparametrycznych i ich zastosowanie w naukach ekonomiczno-społecznych*, Acta Universitatis Lodziensis, Wyd. UŁ, s. 233-240.
[4] Domański Cz., [1990], *Testy statystyczne*, PWE.
[5] Domański Cz., Pruska K., [2000], *Nieklasyczne metody statystyczne*, PWE.
[6] Flinger M.A., Killeen T.J., [1976], *Distribution-Free Two-Sample Tests for Scale*, „Journal of the American Statistical Association", 71, 210-213.
[7] Lehmann E.L., (1959), *Testing Statistical Hypothesis*, New York, John Wiley.
[8] Levene, H., [1960], *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. (red.), Stanford University Press, s. 278-292.
[9] Xu-Feng N., [2004], *Statistical Procedures for Testing Homogeneity of Water Quality Parameters*, Department of Statistics Florida State University Tallahassee.
[10] Zieliński W., [1999], *Wybrane testy statystyczne*, Fundacja „Rozwój SGGW", Warszawa.

Praca wpłynęła do redakcji w kwietniu 2009 r.

EMPIRYCZNA OCENA MOCY TESTÓW DLA WIELU WARIANCJI

Streszczenie

Testy dla wielu wariancji stosuje się zwykle do weryfikowania założenia o równości wariancji wymaganego przez inne procedury – przede wszystkim analizę wariancji. Mogą być też wykorzystywane do oceny jednorodności źródeł danych, które chcemy połączyć dla uzyskania lepszych ocen wariancji.

Testy dla wielu wariancji też zazwyczaj same wymagają spełnienia pewnych założeń. Najczęściej chodzi tu o założenie normalności rozkładu w populacjach lub przynajmniej rozkładu tego samego typu.

Celem artykułu jest zbadanie zachowania się wybranych testów dla wielu wariancji w warunkach prawdziwości i nieprawdziwości założenia o normalności rozkładu, małych i dużych prób, prób równolicznych i prób o zróżnicowanych liczebnościach. Badania te przeprowadzono metodami symulacyjnymi. W badaniu rozważono cztery testy jednorodności wariancji: Levene'a, Browna-Forsythe'a, Flignera-Killeena oraz O'Briena.

Słowa kluczowe: testy dla wielu wariancji, symulacje Monte Carlo, test Levene'a, test Browna-Forsythe'a, test Flignera-Killeena, test O'Briena.

A COMPARATIVE STUDY OF TESTS POWER FOR HOMOGENEITY OF VARIANCES

Summary

Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. Tests for homogeneity of variances can be used to verify that assumption and for pooling of data from different sources to yield an improved estimated variance. Tests for homogeneity of variances can be used usually under assumption of normal distributions or nearly normal distributions. In this paper some tests for homogeneity of variances are examined under the null hypothesis and under the alternative, for various sample sizes, for various symmetric and asymmetric distributions. Monte Carlo simulations has been used for this. In this paper the following procedures have been analyzed: Levene'a test, Brown-Forsythe test, Fligner-Killeen test and O'Brien test.

Key words: tests for homogeneity of variances, Monte Carlo, Levene's test, Brown-Forsythe test, Fligner-Killeen test, O'Brien test.