Jacek Stelmach
Grzegorz Kończak
Uniwersytet Ekonomiczny w Katowicach

O PORÓWNYWANIU DWÓCH POPULACJI WIELOWYMIAROWYCH Z WYKORZYSTANIEM OBJĘTOŚCI ELIPSOID UFNOŚCI

Wprowadzenie

Statystyka dostarcza wielu różnorodnych metod pozwalających na porównanie dwóch populacji. Samo określenie porównanie dwóch populacji może dotyczyć porównywania np. wartości przeciętnych, wskaźników struktury, wariancji lub współczynników korelacji. W takim przypadku mamy do czynienia z wnioskowaniem parametrycznym. Do odpowiedzi na tak postawione zagadnienia wykorzystujemy testy parametryczne, jak np. test t, test z, test równości wskaźników struktury lub współczynników korelacji. Często dokonując porównań populacji, stawiamy hipotezę o jednakowych rozkładach w tych populacjach. W takim przypadku odwołujemy się do testów nieparametrycznych, jak np. test serii lub test Manna-Whitneya. W opracowaniu przedstawiono propozycję porównania rozkładów dwóch populacji wielowymiarowych. Podstawą porównań są objętości elipsoid ufności dla dwóch populacji oraz dla populacji będącej sumą mnogościową analizowanych populacji. W rozważaniach nie przyjmowano założenia o postaci rozkładu. Takie podejście nie daje możliwości wyznaczenia dokładnego rozkładu rozważanej statystyki i odczytania wartości krytycznej testu z tablic. Z tego powodu w analizach wykorzystano testy permutacyjne. Ze względu na możliwość porównania własności rozważanej procedury z testem T^2 Hotellinga w analizach symulacyjnych skoncentrowano się na porównaniu populacji o wielowymiarowych rozkładach normalnych.

1. Wybrane metody porównania dwóch populacji

1.1. Porównanie populacji jednowymiarowych

Do najczęściej wykorzystywanych testów pozwalających na porównanie dwóch populacji należy zaliczyć test *t*. Pozwala on na porównanie wartości oczekiwanych w dwóch populacjach na podstawie pobranych niezależnie prób. Należy on do grupy testów parametrycznych. Związane są z tym określone założenia. Próby powinny być pobrane z populacji o rozkładach normalnych, a w przypadku prób o dużych liczebnościach jest dopuszczalne, aby rozkład był zbliżony do normalnego. Dodatkowo, jeżeli znane jest odchylenie standardowe rozkładu σ, to jest wykorzystywany test, który w literaturze zwykle jest określany jako test *z* [Kanji 2006]. Jeżeli dysponujemy próbkami z dwóch populacji o rozkładach zero-jedynkowych, to dla porównania możemy wykorzystać test równości wskaźników struktury. Test dla równości wariancji (test *F*) pozwala na nieco inne spojrzenie na porównanie dwóch populacji. W teście jest wykorzystywana statystyka *F*. Pozwala on na porównanie wariancji w dwóch populacjach. Podobnie jak w przypadku testu *t* zakłada się, że próby pochodzą z populacji o rozkładach normalnych.

Inną grupę testów stanowią testy nieparametryczne. Do najważniejszych testów nieparametrycznych pozwalających na porównanie dwóch populacji należy zaliczyć test Manna-Whitneya i test serii Walda-Wolfowitza [Blalock 1974]. Oba testy nie wymagają spełnienia ostrych założeń, jednak jako testy nieparametryczne charakteryzują się niewielką mocą. Powszechnie stosowane testy parametryczne wymagają spełnienia ostrych założeń dotyczących postaci rozkładów. Testy nieparametryczne charakteryzują się natomiast niewielką mocą. Alternatywnym rozwiązaniem jest stosowanie testów permutacyjnych [Efron i Tibshirani 1993]. Testy te nie wymagają spełnienia założenia normalności rozkładów, co jest niezbędne przy testach parametrycznych, a mimo to charakteryzują się podobną mocą.

1.2. Porównanie populacji wielowymiarowych

Porównując populacje wielowymiarowe napotykamy problemy, które nie występowały przy podobnej analizie dla populacji jednowymiarowych. Dla pewnych zmiennych (współrzędnych wektora) wartości oczekiwane mogą być jednakowe, a dla innych mogą się znacznie różnić. Ważne w takim przypadku będzie nie tylko stwierdzenie, że wektory wartości oczekiwanych nie są jedna-

kowe, ale również wskazanie, dla których zmiennych występują różnice w wartościach oczekiwanych. Poza różnicami w wartościach oczekiwanych mogą występować różnice w macierzach wariancji-kowariancji, co w praktyce przekłada się na inne wielowymiarowe kształty populacji.

Dla porównania wektorów wartości przeciętnych w dwóch populacjach wielowymiarowych można wykorzystać np. test T^2 Hotellinga. Statystyka ta jest uogólnieniem statystyki t wykorzystywanej dla porównania populacji ze względu na jedną zmienną mierzalną. Dla stosowania tego testu jest wymagane spełnienie założenia, że próby pobrano z populacji o wielowymiarowych rozkładach normalnych [Rencher 2002].

Przyjmijmy, że pobrano dwie niezależne próby $\mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1n_1} : N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ oraz $\mathbf{x}_{21}, \mathbf{x}_{22}, ..., \mathbf{x}_{2n_2} : N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Zakładając, że macierze kowariancji są nieznane, ale jednakowe ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$) do weryfikacji hipotezy o równości wektorów średnich:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

wykorzystywana jest statystyka [Rencher 2002]:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{pl}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{1}$$

gdzie:

$$\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T + \sum_{i=1}^{n_2} (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)^T \right) \tag{2}$$

jest nieobciążonym estymatorem wariancji $\boldsymbol{\Sigma}$.

Statystyka (1) ma rozkład Hotellinga. Wartości krytyczne dla tej statystyki wyznaczamy wykorzystując fakt, że statystyka:

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \tag{3}$$

ma rozkład F o p oraz n_1+n_2-p-1 stopniach swobody. Weryfikując hipotezę o równości wektorów wartości oczekiwanych, obszar krytyczny jest prawo-

stronny. Statystyka (1) może być wykorzystana do weryfikacji hipotezy o jednakowych wektorach wartości przeciętnych, gdy jest spełnione założenie wielowymiarowej normalności rozważanych wektorów. Nieco zmodyfikowana postać statystyki może być wykorzystana, jeżeli np. znana jest postać macierzy kowariancji $\mathbf{\Sigma}$.

A.C. Rencher przedstawia również test dla porównania wariancji dla dwóch populacji wielowymiarowych [Rencher 2002]. Do weryfikacji hipotezy o identyczności macierzy wariancji $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ jest wykorzystywana statystyka:

$$M = \frac{|\mathbf{S}_1|^{v_1/2} + |\mathbf{S}_2|^{v_2/2}}{|\mathbf{S}_{pl}|^{(v_1+v_2)/2}}, \tag{4}$$

gdzie $v_i = n_i - 1$, a $\mathbf{S}_i$ jest macierzą kowariancji dla i-tej próbki ($i = 1, 2$).

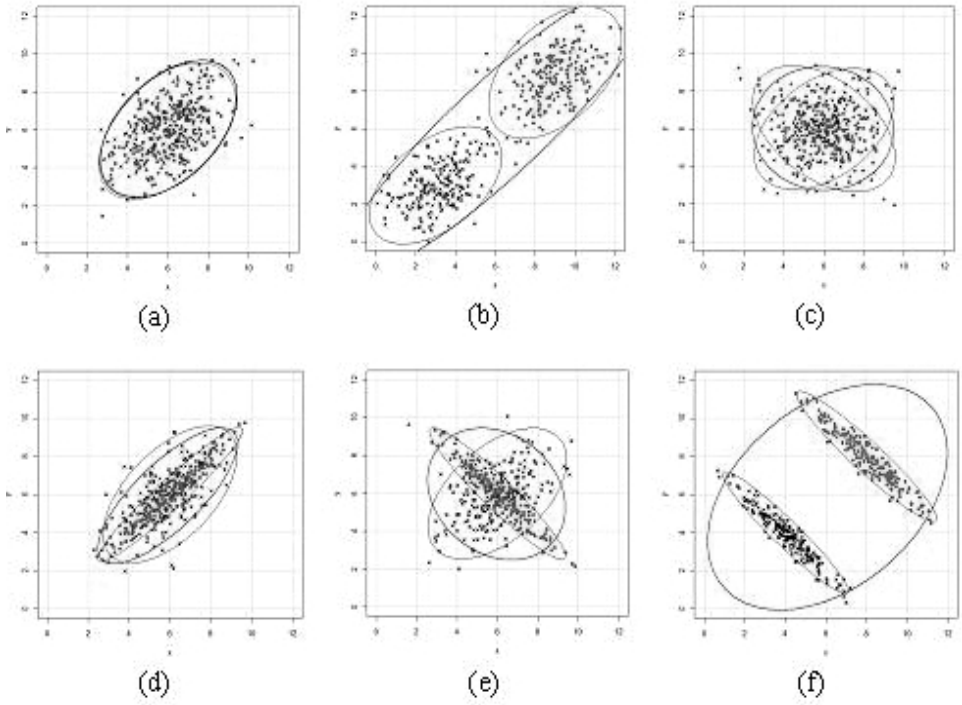
W praktyce badań statystycznych bardzo często nie ma podstaw do przyjęcia założenia o normalności rozkładów. W takich przypadkach niezbędne są narzędzia pozwalające na porównanie dwóch populacji wielowymiarowych nie przyjmując dodatkowych założeń np. odnośnie do postaci rozkładów.

2. Objętość elipsoid ufności – porównanie dwóch populacji wielowymiarowych

Rozważmy dwie populacje o dwuwymiarowych rozkładach normalnych. W kolejnych częściach rysunku 1 przedstawiono wykresy rozrzutu dla próbek wylosowanych z dwóch populacji o rozkładach normalnych. Parametry rozkładów normalnych były identyczne tylko w przypadku przedstawionym na rysunku 1a. W kolejnych częściach rysunku populacje różniły się wektorami wartości oczekiwanych i/lub macierzami kowariancji. Odpowiednie przypadki zostały schematycznie przedstawione na rysunkach 1a-f. Na tych rysunkach zostały również przedstawione elipsoidy ufności dla obu próbek oraz dla próby będącej mnogościową sumą tych prób. Tylko dla przypadku a) rozkłady w populacjach są jednakowe. W przypadkach b-f rozkłady nie są identyczne. Krótka charakterystyka rozważanych rozkładów została przedstawiona poniżej:

a) rozkłady identyczne (jednakowe wartości oczekiwane, jednakowe macierze wariancji),
b) różne wartości oczekiwane, jednakowe macierze wariancji,

c) jednakowe wartości oczekiwane, różne macierze wariancji (inne kształty),
d) jednakowe wartości oczekiwane, różne macierze wariancji (inny rozrzut),
e) jednakowe wartości oczekiwane, różne macierze wariancji (inny kształt i rozrzut)
f) różne wartości oczekiwane, jednakowe macierze wariancji.



Rys. 1. Dwie populacje – wzajemne położenie

Jak łatwo zauważyć, w przypadku gdy próbki pochodzą z populacji o różnych rozkładach, elipsoida ufności dla sumy prób charakteryzuje się większym polem (ogólnie: większą objętością) niż elipsoida dla każdej z dwóch prób. Dla przedstawionych graficznie 6 przypadków wzajemnego położenia tylko w pierwszym przypadku rozkłady w dwóch populacjach są jednakowe. Zastosowanie testu T^2 Hotellinga pozwoli wskazać na różnice w populacjach w przypadkach b) i f). W pozostałych wektory wartości przeciętnych są takie same, ale inne są macierze wariancji. Proponowany test powinien skutecznie

wskazywać występowanie różnic w rozkładach we wszystkich pięciu przypadkach (b-f), a jedynie w przypadku a) prowadzić do stwierdzenia braku podstaw do odrzucenia hipotezy. Do porównania rozkładów zostanie wykorzystana statystyka:

$$T = \frac{\max\{volA, volB, vol(A \cup B)\}}{\min\{volA, volB, vol(A \cup B)\}} \tag{5}$$

gdzie:

vol(A), vol (B) oznaczają objętość elipsoidy (w przypadku dwuwymiarowym pole elipsy) ufności otrzymaną na podstawie próby odpowiednio z populacji *A* oraz *B*,

$vol(A \cup B)$ oznacza objętość elipsoidy (w przypadku dwuwymiarowym pole elipsy) ufności otrzymaną na podstawie połączonych prób.

Statystyka (5) może przyjmować wartości z przedziału $[1, +\infty)$. W przypadku gdy rozkłady populacji będą identyczne, wartości statystyki będą bliskie wartości 1. Duże wartości statystyki *T* będą świadczyły o występujących różnicach w rozkładach populacji. Obszar krytyczny w proponowanym teście jest prawostronny.

3. Testy permutacyjne

Weryfikując hipotezę H_0, musimy znać wartości krytyczne, które określą obszar odrzucenia tej hipotezy. W rozważanym przypadku, rozkład statystyki testowej jest nieznany – a więc także nie jest on stablicowany. Nie jest więc możliwe odczytanie wartości krytycznych. W takich sytuacjach bardzo pomocne będzie zastosowanie testów permutacyjnych [Good 1994, s. 176]. Testy te nie wymagają żadnej wiedzy o rozkładzie wykorzystywanej statystyki. W eksperymencie zastosowano test permutacyjny, w którym z uwagi na czas obliczeń ograniczono się do $N = 1000$ permutacji (co dla większości przypadków jest ilością wystarczającą) – [Hesterberg et al. 2003, s. 18-60], polegających na utworzeniu podgrup przez losowanie bez zwracania z sumy mnogościowej prób pochodzących z populacji *A* i *B*.

Wartość *p-value* testu permutacyjnego wyznaczono z zależności:

$$ASL = \frac{card\{i \in \{1,2,...,N\} : T^* > T_i\}}{N}, \tag{6}$$

gdzie:

T^* – wartość statystyki obliczonej dla próby pierwotnej,

T_i – wartość statystyk obliczonych dla prób permutacyjnych.

Wartości *p-value* mniejsze od przyjętego poziomu istotności α prowadzą do odrzucenia hipotezy o identyczności rozkładów, a większe od α prowadzą do stwierdzenia o braku podstaw do odrzucenia hipotezy o identyczności rozkładów.

4. Analiza symulacyjna

Wszystkie analizy symulacyjne i obliczenia wykonano w programie R (http://www.r-project.org). W symulacjach rozważano populacje 5-wymiarowe. Dane o wektorach wartości oczekiwanych $\boldsymbol{\mu}_A$ i $\boldsymbol{\mu}_B$ oraz macierzach wariancji-kowariancji $\boldsymbol{\Sigma}_A$ i $\boldsymbol{\Sigma}_B$ populacji przedstawia tabela 1. Kolejne wiersze tabeli (a-f) korespondują z prezentacją graficzną na rysunku 1.

Tabela 1

Wektory wartości przeciętnych i macierze wariancji-kowariancji przypadków testowych

Przypadek	$\boldsymbol{\mu}_A$	$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_A$	$\boldsymbol{\Sigma}_B$
a)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\mathbf{I}$	$\mathbf{I}$
b)	[0, 0, 0, 0, 0]	[1, 1, 1, 1, 1]	$\mathbf{I}$	$\mathbf{I}$
c)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_2$
d)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\mathbf{I}$	$1{,}7\mathbf{I}$
e)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\mathbf{I}$	$\boldsymbol{\Sigma}_3$
f)	[0, 0, 0, 0, 0]	[1, 1, 1, 1, 1]	$\boldsymbol{\Sigma}_2$	$\boldsymbol{\Sigma}_2$

Macierze występujące w tabeli 1 są zadane następującymi wzorami:

I – macierz jednostkowa o wymiarach 5 x 5.

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0{,}6 & 0{,}6 & 0 & 0 \\ 0{,}6 & 1 & 0{,}6 & 0{,}6 & 0 \\ 0{,}6 & 0{,}6 & 1 & 0{,}6 & 0 \\ 0 & 0{,}6 & 0{,}6 & 1 & 0{,}6 \\ 0 & 0 & 0 & 0{,}6 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -0{,}6 & 0{,}6 & 0 & 0 \\ -0{,}6 & 1 & -0{,}6 & 0{,}6 & 0 \\ 0{,}6 & -0{,}6 & 1 & -0{,}6 & 0 \\ 0 & 0{,}6 & -0{,}6 & 1 & -0{,}6 \\ 0 & 0 & 0 & -0{,}6 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0{,}7 & 0{,}2 & 0 & 0 \\ 0{,}7 & 1 & 0{,}7 & 0{,}2 & 0 \\ 0{,}2 & 0{,}7 & 1 & 0{,}7 & 0{,}2 \\ 0 & 0{,}2 & 0{,}7 & 1 & 0{,}7 \\ 0 & 0 & 0{,}2 & 0{,}7 & 1 \end{bmatrix}$$

Dla każdego z przedstawionych w tabeli 1 przypadków rozważano próbki o liczebnościach $n_1 = n_2 = 10$, 20, 30 i 50. Dla tych przypadków generowano 1000-krotnie próby z populacji A oraz B. Następnie przyjmując poziom istotności $\alpha = 0{,}05$ oraz wykorzystując statystykę (5), przeprowadzano test permutacyjny. Na tej podstawie wyznaczano oceny prawdopodobieństw odrzucenia hipotezy H_0.

Niezależnie od powyżej opisanych symulacji wykonano analizę Monte Carlo pozwalającą porównać własności rozważanego testu z klasycznym testem T^2 Hotellinga. W tym celu porównywano dwie populacje o dwuwymiarowych rozkładach normalnych o parametrach $\boldsymbol{\mu}_A = [0;0]$, $\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ oraz

$\boldsymbol{\mu}_B = [x; x]$, $\boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, gdzie x = 0,1; 0,2; …;2. W symulacjach uwzględniono próbki o liczebnościach $n_1 = n_2 = 10$, 20, 30 i 50. Na podstawie $N = 1000$ symulacji dla każdej takiej pary prób przeprowadzono test permutacyjny oraz T^2 Hotellinga. W obu przypadkach wyznaczono liczbę odrzuceń hipotezy o równości wektorów wartości przeciętnych. Na tej podstawie otrzymano oceny prawdopodobieństw odrzucenia hipotezy H_0.

5. Wyniki

W analizach uwzględniono objętości elipsoid ufności pokrywających wielowymiarową przestrzeń z badanymi obserwacjami – z 95% prawdopodobieństwem, wykorzystując procedury pakietu R. Wyniki analiz (dane porównywanych populacji w tabeli 1) umieszczono w tabeli 2. Przedstawiono w niej oceny prawdopodobieństw odrzucenia hipotezy o identyczności rozkładów 5-wymiarowych populacji (dla rozkładów identycznych – przypadek testowy a) oznacza błąd pierwszego rodzaju, dla pozostałych, w których symulowano różnice w rozkładach – moc testu) – z poziomem istotności $\alpha = 0.05$.

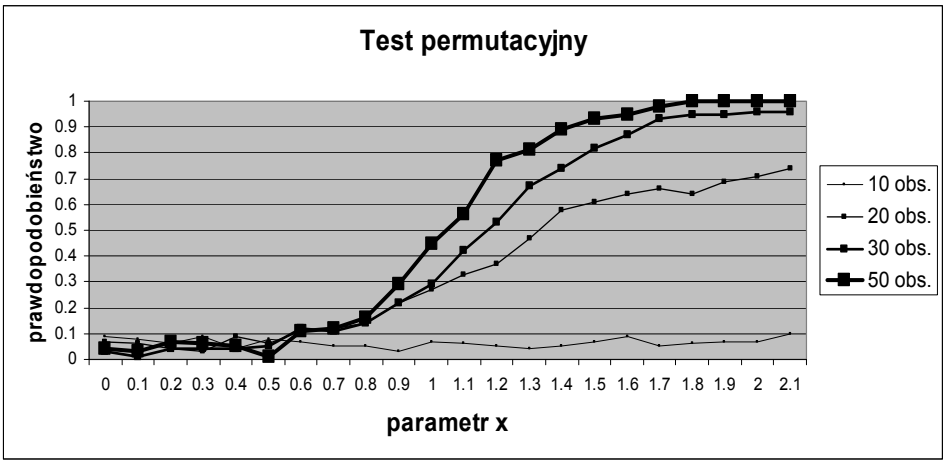
Tabela 2

Oceny prawdopodobieństw odrzucenia hipotezy o identyczności rozkładów

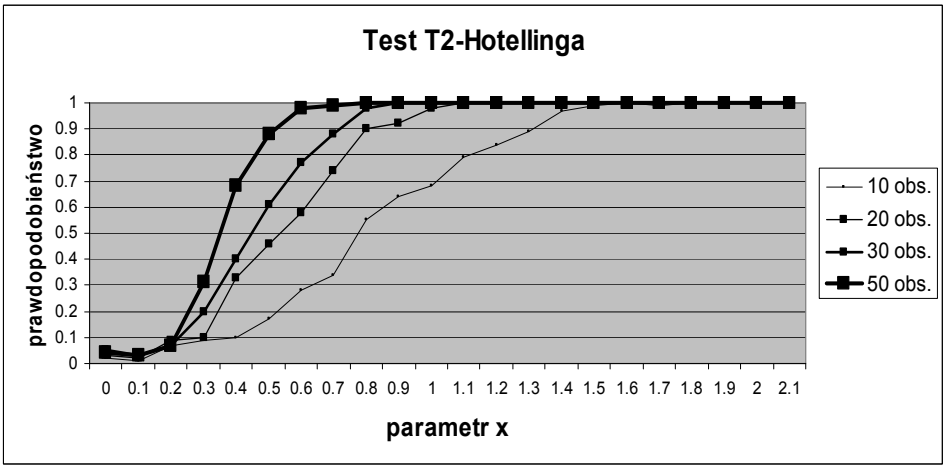
Przypadek	Liczebność próby			
	10	20	30	50
a)	0,036	0,037	0,048	0,037
b)	0,120	0,312	0,337	0,397
c)	0,897	1,000	1,000	1,000
d)	0,095	0,361	0,626	0,890
e)	0,486	0,969	1,000	1,000
f)	0,213	0,615	0,825	0,953

Dla przypadku identycznych rozkładów populacji (wariant a) rozmiar proponowanego testu jest nieco mniejszy od przyjętego poziomu istotności α. We wszystkich przypadkach, gdy rozkłady nie są identyczne, test dla wszystkich

rozważanych liczebności skutecznie wskazuje na występujące różnice. Szczególnie dobrze jest to widoczne dla prób o liczebności przynajmniej 30. Przeciętny błąd oceny szacowanych prawdopodobieństw we wszystkich analizowanych sytuacjach jest mniejszy od 0,016.



Rys. 2. Oceny prawdopodobieństwa odrzucenia hipotezy zerowej dla testu permutacyjnego w zależności od różnicy w wartościach współrzędnych *x*



Rys. 3. Oceny prawdopodobieństwa odrzucenia hipotezy zerowej dla testu T^2-Hotellinga w zależności od różnicy w wartościach współrzędnych *x*

Na rysunkach 2 i 3 zawarto wyniki analizy Monte Carlo – porównawczo test permutacyjny ze statystyką testową jak w (5) oraz test T^2 Hotellinga dla dwuwymiarowych prób o rozkładach różniących się wektorami średnich. Różnicę w wartościach współrzędnych tych wektorów opisuje parametr x wykresu leżący na osi odciętych. Badanie przeprowadzono dla różnych liczebności symulowanych prób. Analizując otrzymane w tej części wyniki, można zauważyć, że test T^2 Hotellinga skuteczniej odróżnia populacje różniące się wektorami wartości średnich. Porównań dokonano dla populacji o rozkładach normalnych, bo tylko wówczas jest uprawnione porównanie tych dwóch testów. Stosowanie testu T^2 Hotellinga wymaga spełnienia założenia o normalności rozkładu w badanych populacjach. Zaletą proponowanego testu permutacyjnego jest fakt, że może on być stosowany dla populacji o dowolnych rozkładach.

Podsumowanie

Proponowana statystyka testowa, oparta na analizie objętości elipsoid ufności obejmujących badane próby pozwala na weryfikację hipotez o identyczności rozkładów, także w przypadkach, w których test T^2 Hotellinga z uwagi na równość wartości średnich z badanych prób nie doprowadzi do odrzucenia hipotezy. Dodatkowo wykorzystanie testów permutacyjnych zwalnia z weryfikacji założenia o zgodności badanych rozkładów z rozkładem normalnym wielowymiarowym i nie wymaga tablicowania proponowanej statystyki testowej. Przeprowadzone badania symulacyjne wykazały zadowalające prawdopodobieństwo odrzucenia hipotezy zerowej dla rozkładów różniących się macierzą kowariancji czy wektorem wartości przeciętnych już dla prób o liczebności powyżej 30 obserwacji, a wysokie prawdopodobieństwo – dla liczebności ponad 50 obserwacji.

Przeprowadzona analiza Monte Carlo, porównująca moc testów permutacyjnego i T^2 Hotellinga, przeprowadzona dla rozkładów dwuwymiarowych normalnych, różniących się tylko wektorem wartości średnich (a więc dla rozkładów, do których jest predystynowany test parametryczny T^2 Hotellinga) wykazała większą moc testu parametrycznego. Niemniej jednak test permutacyjny cechował się porównywalną wielkością błędu pierwszego rodzaju, a zdolność rozpoznawania różniących się populacji osiągał dla różnicy wektorów wartości średnich na poziomie [1.0; 1.0].

Literatura

Blalock H.M. (1974): *Statystyka dla socjologów*. PWN, Warszawa.
Efron B., Tibshirani R. (1993): *An Introduction to the Bootstrap*. Chapman & Hall, New York.
Good P.I. (1994): *Permutation Tests: A Practical Guide for Testing Hypotheses*. Springer-Verlag, New York.
Hesterberg T., Monaghan S., Moore D.S., Clipson A., Epstein R. (2003): *The Practice of Business Statistics.* W.H. Freeman and Company, New York.
Kanji G.K. (2006): *100 Statistical Tests*. Sage Publications, London.
Rencher A.C. (2001): *Methods of Multivariate Analysis*. John Wiley & Sons, New York.

ON THE COMPARISON OF TWO MULTIDIMENSIONAL POPULATIONS USING THE CONFIDENCE ELLIPSOID VOLUMES

Summary

A comparison of two populations seems to be interesting and very common statistical problem. The most often way is to verify the hypothesis concerned the equality of certain, characteristic parameter i.e. mean, standard deviation or fraction with parametric or non-parametric tests. The authors propose to compare the distribution of two populations – comparing the confidence ellipsoid volumes. Since their distribution is unknown – permutation tests were applied. A Monte-Carlo simulation let to compare power of these tests with T^2 Hotelling tests. Proposed methods can be used, when the assumptions for parametric tests couldn't be verified.