Listy Biometryczne
Nr 34-36 (1972)

Wiesław W a g n e r (Poznań)

O TESTOWANIU RÓWNOŚCI K MACIERZY KOWARIANCJI

1. Wstęp

W doświadczeniach rolniczych i przyrodniczych obserwuje się zwykle więcej niż jedną cechę na każdej jednostce doświadczalnej. Stąd wyniki z takich doświadczeń mają przeważnie charakter wielowymiarowy. Ponieważ cechy te są najczęściej skorelowane, więc analizowanie każdej z obserwowanych cech oddzielnie może okazać się w wielu przypadkach mało efektywne. Jeśli badacza interesuje kompleksowe ujęcie obserwowanych cech, wówczas naturalnym staje się przejście od jednozmiennej analizy wariancji (ang. ANOVA) do wielozmiennej analizy wariancji (ang. MANOVA), badającej wpływ różnych źródeł zmienności na ogólną zmienność wielu cech. Istotnym problemem staje się wykorzystanie metod wielozmiennej analizy, które umożliwiają uzyskanie informacji dotyczących p skorelowanych zmiennych w jednej lub więcej populacjach, z których pobrano losowe próby. Przy wielu metodach przyjmuje się założenie, że te zmienne mają p-wymiarowe rozkłady normalne o tych samych macierzach kowariancji. Jednak taka jednorodność nie zawsze zachodzi i w związku z tym konieczne jest sprawdzenie hipotez dotyczących równości macierzy kowariancji.

2. Jednokierunkowa klasyfikacja obserwacji

Wyobraźmy sobie, że porównujemy k-populacji pod względem p różnych cech. Z każdej z tych populacji pobrano po jednej próbie losowej. Oznaczmy liczebność próby z g-tej populacji przez N_g' (g =

= 1, 2, ..., k), a sumę wszystkich liczebności, to znaczy liczbę wszystkich jednostek doświadczalnych przez N':

$$N' = N_1' + N_2' \ldots + N_k'.$$

Na każdej jednostce doświadczalnej obserwowanych jest p cech. Mamy wówczas N' obserwacji p-wymiarowych sklasyfikowanych według jednego kryterium - przynależności do jednej z k populacji. Oznaczmy j-tą obserwację ($j = 1, \ldots, N_g'$) i-tej cechy ($i = 1, \ldots, p$) w próbie pobranej w g-tej populacji ($g = 1, \ldots, k$) przez x_{gji}. Przy tych oznaczeniach macierz obserwacji dla g-tej próby można zapisać w postaci

$$\underline{X}_g = \begin{bmatrix} x_{g11} & x_{g12} & \cdots & x_{g1p} \\ x_{g21} & x_{g22} & \cdots & x_{g2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{gN_g'1} & x_{gN_g'2} & \cdots & x_{gN_g'p} \end{bmatrix}, \tag{1}$$

gdzie N_g' wierszy odpowiada kolejnym obserwacjom, a p kolumn odpowiada kolejnym cechom. Macierz

$$\underline{X} = (\underline{X}_1', \underline{X}_2', \ldots, \underline{X}_k')', \tag{2}$$

typu $N' \times p$, złożona z podmacierzy określonych w (1), jest wówczas macierzą wszystkich p-wymiarowych obserwacji. Zakładamy, że wiersze macierzy $\underline{X}_g$ są wzajemnie niezależnymi p-wymiarowymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(\underline{\mu}_g, \underline{\Sigma}_g)$, gdzie wektor wartości średnich

$$\underline{\mu}_g = \left[\mu_{g1}, \mu_{g2}, \ldots, \mu_{gp}\right]', \tag{3}$$

a macierz kowariancji

$$\underline{\Sigma}_g = \begin{bmatrix} \sigma_{g11} & \sigma_{g12} & \cdots & \sigma_{g1p} \\ \sigma_{g21} & \sigma_{g22} & \cdots & \sigma_{g2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{gp1} & \sigma_{gp2} & & \sigma_{gpp} \end{bmatrix} \quad (g = 1, 2, \ldots, k). \tag{4}$$

Przyjmując takie założenie dopuszczamy istnienie różnych wektorów wartości średnich i różnych macierzy kowariancji dla różnych populacji.

Przy wyborze odpowiedniej metody do analizy obserwacji wielowymiarowych sklasyfikowanych według jednego kryterium - według po-

pulacji, interesuje nas to, czy dopuszczalne jest przyjęcie upraszczającego założenia, że wszystkie macierze $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ są jednakowe.

3. Rozkład statystyki M

Hipotezę zerową o jednorodności między populacjami w sensie równości macierzy kowariancji możemy zapisać w postaci:

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k .$$

Przy sprawdzaniu tej hipotezy stosuje się statystykę M Bartletta postaci

$$M = N \ln |\underline{S}| - \sum_{g=1}^{k} N_g \ln |\underline{S}_g| \tag{5}$$

gdzie $N = N_1 + N_2 + \ldots + N_k$, $N_g = N_g' - 1$, macierz $\underline{S}_g$ typu $p \times p$ jest nieobciążonym estymatorem macierzy kowariancji dla g-tej populacji o N_g stopniach swobody, a $N\underline{S} = N_1\underline{S}_1 + N_2\underline{S}_2 + \ldots + N_k\underline{S}_k$.

Anderson [1] wyprowadził kryterium M modyfikując test oparty na ilorazie wiarogodności.

Kulback [5] natomiast pokazał, że kryterium M jest statystyką minimalnej informacji dyskryminacyjnej dla hipotezy zerowej i wykazuje, że przy hipotezie alternatywnej o niejednorodności macierzy kowariancji przyjmuje ona wartość minimalną.

W oparciu o wyniki Boxa [2] Korin [3], [4] zbudował tablice 5-procentowych punktów istotności dla rozkładu statystyki M uwzględniając różne wartości parametrów k, p oraz N_0, gdzie $N_0 = N_1 = N_2 = \ldots = N_k$. Podane przez Korina tablice [4] obejmują jednak zbyt mały zakres zmieniającej się wielkości N_0. Często zachodzi potrzeba odczytania punktu istotności dla większej liczby obserwacji N_0. Wówczas z konieczności musimy szukać przybliżonych punktów istotności dla M. Jeśli spełniony jest warunek $N_1 = N_2 = \ldots = N_k = N_0$ to możemy je znaleźć używając jednego z następujących przybliżeń Boxa [2] dla rozkładu M:

(a) Przybliżenie rozkładem χ^2. Tutaj korzystamy z tego, że statystyka $(1 - A_1)\cdot M$ ma w przybliżeniu rozkład chi-kwadrat o f_1 stopniach swobody, tzn.

$$M \sim \chi^2_{f_1}/(1 - A_1), \tag{6}$$

gdzie liczba stopni swobody

$$f_1 = \tfrac{1}{2} p(p+1)(k-1), \tag{7}$$

natomiast

$$A_1 = \frac{2p^2 + 3p - 1}{6(p+1)kN_0} . \tag{8}$$

(b) Przybliżenie rozkładem F. Tutaj wykorzystuje się to, że statystyka M/b ma w przybliżeniu rozkład F o f_1 i f_2 stopniach swobody, tzn.

$$M \sim b\, F_{f_1;\, f_2}, \tag{9}$$

gdzie f_1 i A_1 są określone w (7) i (8), zaś

$$f_2 = \text{entier}\, \frac{f_1 + 2}{A_2 - A_1^2} , \tag{10}$$

przy czym

$$A_2 = \frac{(p-1)(p+2)(k^2+k+1)}{6k^2N_0^2} , \tag{11}$$

natomiast

$$b = \frac{f_1}{1 - A_1 - f_1/f_2} \tag{12}$$

Obliczenie wartości statystyki M poprzedzone jest znalezieniem macierzy

$$\underline{S} = N^{-1}(N_1\underline{S}_1 + N_2\underline{S}_2 + \ldots + N_k\underline{S}_k), \tag{13}$$

gdzie macierz $\underline{S}_g$ jest macierzą sum kwadratów i iloczynów odchyleń dla próby z j-tej populacji (g = 1, ..., k), to znaczy

$$\underline{S}_g = \begin{bmatrix} s_{g11} & s_{g12} & \cdots & s_{g1p} \\ s_{g21} & s_{g22} & \cdots & s_{g2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{gp1} & s_{gp2} & \cdots & s_{gpp} \end{bmatrix} \tag{14}$$

gdzie

$$s_{gii'} = \frac{1}{N_g} \sum_{j=1}^{N_g} x_{gji}\, x_{gji'} - \frac{1}{N_g^2} T_{gi}\, T_{gi'},$$

przy czym

$T_{gi} = \sum_{j=1}^{N_g} x_{gji}$ jest sumą wszystkich obserwacji dotyczących i-tej cechy (i, i' = 1, ..., p) dla próby z g-tej populacji.

Wartość funkcji testowej M porównuje się z wartością krytycz-

ną odczytaną z tablic Korina [4] dla poziomu istotności 5%. W przypadku większej liczby obserwacji można posłużyć się przybliżeniem rozkładem χ^2 lub F (patrz wzory (6),(9)). W przypadku, gdy wartość M przekroczy wartość krytyczną przy ustalonym poziomie istotności, wówczas odrzucamy hipotezę zerową o równości macierzy kowariancyjnych.

4. Przykład

Jako przykład niech posłuży analiza danych pochodzących z badań hodowlanych nad słonecznikiem oleistym, przeprowadzonych w SHB Borowo. Na dane te składają się obserwacje dotyczące 5 cech u 7 rodów słonecznika, z których wylosowano po 28 roślin. Korzystając ze wzorów (13) i (14) wyliczono macierze $\underline{S}$ i $\underline{S}_g$ ($g = 1, \ldots, 7$), a następnie uzyskano wartość statystyki M równą 26,2056.

Ponieważ tablice [4] nie zawierają wartości krytycznych dla poziomu istotności 5% i parametrów $N_0 = 27$, $p = 5$ i $k = 7$, dlatego też zostały one wyznaczone według wzorów (6) i (9). Wyniki zestawiono w tablicy 1.

T a b l i c a 1. Wartości krytyczne na poziomie 5% dla parametrów $k = 7$, $p = 5$, $N_0 = 27$

Przybliżenie	χ^2 wg wzoru (6)	122,352
Przybliżenie	F wg wzoru (9)	124,593

Wartość krytyczną dla rozkładów χ^2 i F przy poziomie istotności 5% dla odpowiedniej liczby stopni swobody odczytano z tablic [6]. Ponieważ wartość M jest znacznie mniejsza od wartości krytycznych podanych w tablicy 1, więc nie mamy podstaw do odrzucenia hipotezy zerowej o jednorodności macierzy kowariancyjnych.

Obliczenia do tego przykładu wykonano według programu ułożonego w języku ODRA ALGOL na emc ODRA 1204.

Literatura cytowana

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, New York 1958.

[2] Box, G. E. P., A general distribution theory for a class of likelihood criteria, Biometrika, 36 (1949), str. 317-346.

[3] Korin, B. P., On the distribution of a statistic used for testing a covariance matrix, Biometrika, 55 (1968), str. 171-179.

[4] Korin, B. P., On testing the equality of k covariance matrices, Biometrika, 56 (1969), str. 216-219.

[5] Kulback, S., Information Theory and Statistics, New York 1959.

[6] Pearson, E. S. and Hartley, H. O., Biometrika Tables for Statisticians, vol. 1, Cambridge 1966.