Anna B a r t k o w i a k (Wrocław)

REALIZACJA NUMERYCZNA NIEKTÓRYCH ZAGADNIEŃ ANALIZY WARIANCJI

1. Wstęp

Tematyka analizy wariancji jest niezmiernie szeroka i zróżnicowana. Istnieje kilka monografii z tej dziedziny, w każdym niemal podręczniku statystyki matematycznej znajduje się przynajmniej jeden rozdział poświęcony temu tematowi. Analiza wariancji znajduje coraz to szerszy zakres zastosowań w najróżnorodniejszych dziedzinach.

Obliczenia związane z analizą wariancji są raczej żmudne i czasochłonne, a przy tym dość różnorodne i zróżnicowane. Wobec coraz to zwiększających się możliwości korzystania z usług maszyn cyfrowych pojawia się problem: jak programować obliczenia z tej dziedziny. Oczywiście byłoby bardzo nierozsądne programować każde zagadnienie oddzielnie: byłoby to niepotrzebną stratą pracy i środków. Wobec tego należy skomplikowane obliczenia rozłożyć na prostsze zadania elementarne, które można by realizować za pomocą odpowiednio opracowanych algorytmów. Algorytmy te powinny być stosunkowo proste,

a jednocześnie dość uniwersalne, tj. takie, żeby można było z nich komponować różne warianty obliczeń.

W dalszym ciągu podamy wstępny podział obliczeń analizy wariancji na określone rodzaje, wyodrębnimy pewne etapy obliczeń jako zadania elementarne oraz omówimy, w jaki sposób można realizować wyodrębnione zadania elementarne.

2. Zasadnicze rodzaje obliczeń analizy wariancji oraz zagadnienia elementarne rozpatrywane przy tych obliczeniach

Wykonywane współcześnie obliczenia analizy wariancji można podzielić na kategorie (rodzaje) powstałe wskutek rozpatrywania następujących czynników:

A. Liczba obserwowanych zmiennych wynikowych. Rozróżniamy tu obliczenia jednozmienne i wielozmienne.

B. Schemat doświadczenia ortogonalny lub nieortogonalny.

C. Parametry modelu stałe, losowe lub mieszane.

W zagadnieniach praktycznych spotykamy się z różnymi kombinacjami wymienionych wyżej czynników.

W każdej z tych kategorii wykonuje się obliczenia według następujących punktów, nazywanych przez nas dalej zadaniami prostymi:

1. Sformułowanie matematyczne rozważanego modelu.
2. Przygotowanie danych w postaci odpowiednich tablic.
3. Obliczanie zmienności czyli sum kwadratów (surowych lub lub poprawionych).
4. Testowanie hipotez ogólnych.

5. Testowanie hipotez szczegółowych.

6. Konstruowanie przedziałów ufności dla wybranych funkcji parametrycznych (w szczególności dla średnich).

W dalszym ciągu omówimy bardziej szczegółowo podobieństwa między modelami wielozmiennymi i jednozmiennymi, algorytmy obliczające sumy kwadratów dla modeli jednozmiennych z dowolną liczbą czynników oraz testowanie hipotez ogólnych w modelach z czynnikami stałymi.

3. Wielozmienna analiza wariancji jako uogólnienie obliczeń jednozmiennych

Podstawowym równaniem w jednozmiennym modelu analizy wariancji jest następujące równanie (por. Searle [13], Ahrens [1], Morrison [11]):

$$\underset{n\times 1}{y} = \underset{n\times t}{X}\,\underset{t\times 1}{b} + \underset{n\times 1}{e}, \tag{1}$$

gdzie y jest wektorem obserwacji rozważanej cechy y u n osobników, X jest macierzą doświadczenia, b jest wektorem parametrów, e - wektorem błędów.

Weryfikowanie hipotez parametrycznych w tym modelu odbywa się za pomocą testu stosunku wiarogodności, sprowadzającego się do obliczeniu statystyki F jako stosunku dwóch zmienności podzielonych przez odpowiednie liczby stopni swobody:

$$F = \frac{SS_H}{df_H} : \frac{SS_E}{df_E}. \tag{2}$$

Zmienności SS_H i SS_E są formami kwadratowymi obserwowanych wartości cechy y:

(3) $$SS_H = \underline{y}' Q_H \underline{y}, \quad SS_E = \underline{y}' Q_E \underline{y},$$

gdzie Q_H, Q_E są macierzami kwadratowymi o wymiarze $n \times n$, wynikającymi z rozważanego modelu.

W przypadku obserwacji wielozmiennych u każdego z n obserwowanych osobników mierzymy p cech, skąd otrzymujemy macierz obserwacji:

(4) $$\underset{p\times n}{Y'} = \begin{vmatrix} y_{11} & \cdots & y_{1n} \\ y_{21} & \cdots & y_{2n} \\ \vdots & & \\ y_{p1} & \cdots & y_{pn} \end{vmatrix} = (\tilde{\underline{y}}_1, \tilde{\underline{y}}_2, \ldots, \tilde{\underline{y}}_n)$$

gdzie

(5) $$\tilde{\underline{y}}_i' = (y_{1i}, y_{2i}, \ldots, y_{pi})$$

oznacza wektor obserwacji p cech u i-tego osobnika.

Wprowadźmy definicję macierzy H i E w następujący sposób:

(6) $$\underset{p\times p}{H} = \underset{p\times n}{Y'} \underset{n\times n}{Q_H} Y, \qquad \underset{p\times p}{E} = \underset{p\times n}{Y'} \underset{n\times n}{Q_E} Y.$$

Testy wielozmiennej analizy wariancji są obliczane na podstawie pierwiastków charakterystycznych równania macierzowego

(7) $$|H - \lambda E| = 0.$$

Z powyższego wynika, że jeśli znamy macierze Q_H i Q_E wyznaczające zmienności dla hipotezy i błędu w modelu jednozmiennym, to wykorzystując równanie (6) potrafimy wyznaczyć macierze H i E będące podstawą testów wielozmiennych.

Zauważmy jednak, że macierze Q_H i Q_E są wymiaru $n \times n$, a więc trudno jest pamiętać je in extenso w pamięci operacyjnej maszyny cyfrowej. Np. przy liczebności $n = 200$ każda z tych

macierzy składa się z 40000 elementów, a przy n = 800 każda z tych macierzy liczy 640000 elementów.

Jednak w najczęściej obliczanych modelach jednozmiennych nie ma potrzeby pamiętania macierzy Q_H i Q_E in extenso, wystarcza pamiętać jedynie formy kwadratowe $\underline{y}'Q_H\underline{y}$ i $\underline{y}'Q_E\underline{y}$ w postaci wyrażeń algebraicznych będących formami biliniowymi obserwowanych wartości $y_1, y_2, \ldots, y_n$:

$$SS_H = \underline{y}'Q_H\underline{y} = \sum_{j,k=1}^{n} q_{Hjk}y_jy_k,$$

(8)

$$SS_E = \underline{y}'Q_E\underline{y} = \sum_{j,k=1}^{n} q_{Ejk}y_jy_k$$

We wzorze powyższym symbole y_j, y_k oznaczają wartości rozważanej cechy zaobserwowane u j-tego i k-tego osobnika.

Zamieniając iloczyn skalarów y_jy_k na iloczyn wektorów $\underline{\tilde{y}}_j\underline{\tilde{y}}_k'$, gdzie $\underline{\tilde{y}}_j$, $\underline{\tilde{y}}_k$ są opisane wzorem (5), otrzymujemy następujące wzory na macierze H i E:

$$(9) \qquad H = \sum_{j,k}^{n} q_{Hjk}\underline{\tilde{y}}_j\underline{\tilde{y}}_k', \qquad E = \sum_{j,k}^{n} q_{Ejk}\underline{\tilde{y}}_j\underline{\tilde{y}}_k'.$$

W szczególności jeśli formy kwadratowe rozważane w modelu jednozmiennym są funkcjami kwadratowymi zmiennych obliczanych jako formy liniowe wektorów obserwacji, czyli są postaci:

$$(10) \qquad \underline{y}'Q_H\underline{y} = \sum_{j=1}^{a} c_j\left[\sum_{k=1}^{n} a_{kj}y_k\right]^2, \qquad \underline{y}'Q_E\underline{y} = \sum_{j=1}^{b} b_j\left[\sum_{k=1}^{n} b_{kj}y_k\right]^2,$$

otrzymujemy poprzez zamianę kwadratów skalarów na odpowiedni iloczyn wektorowy następujące wzory dla modelu wielozmiennego:

(11)
$$H = \sum_{j=1}^{a} c_j \Big[\sum_{k=1}^{n} a_{kj}\tilde{\underline{y}}_k\Big]\Big[\sum_{k=1}^{n} a_{kj}\tilde{\underline{y}}_k\Big]' \qquad E = \sum_{j=1}^{b} d_j \Big[\sum_{k=1}^{n} b_{kj}\tilde{\underline{y}}_k\Big]\Big[\sum_{k=1}^{n} b_{kj}\tilde{\underline{y}}_k\Big]'.$$

Tak więc jeśli potrafimy testować jakąś hipotezę w modelu jednozmiennym, to potrafimy ją testować również w modelu wielozmiennym.

PRZYKŁAD 1. Analiza wariancji z pojedynczą klasyfikacją. Obserwowane wartości cechy y oznaczamy symbolem y_{jl}, gdzie j oznacza numer klasy ($j = 1, 2, \ldots, J$), a l - numer osobnika w danej klasie.

Testem hipotezy o braku zróżnicowania między rozważanymi klasami jest statystyka F obliczana na podstawie zmienności międzyklasowej i wewnątrzklasowej wyznaczanych następująco:

$$\underline{y}'Q_H\underline{y} = \sum_{j=1}^{J} n_j(\bar{y}_{j\cdot}-\bar{y}_{\cdot\cdot})^2 \quad \text{(zmienność między klasami)},$$

$$\underline{y}'Q_G\underline{y} = \sum_{j=1}^{J}\sum_{l=1}^{n_j} (y_{jk}-\bar{y}_{j\cdot})^2 \quad \text{(zmienność wewnątrz klas)}.$$

W przypadku wielozmiennym, skalar y_{jl} zostaje zastąpiony wektorem $\tilde{\underline{y}}_{jl}$, gdzie $\tilde{\underline{y}}'_{jl} = \{y_{1jl}, y_{2jl}, \ldots, y_{pjl}\}$, czyli obserwowane wielkości są zapisywane za pomocą trzech wskaźników, z których pierwszy oznacza numer cechy, drugi - numer klasy, a trzeci - numer osobnika w tej klasie.

Stąd otrzymujemy następujące wzory na macierze H i E:

$$H = \sum_{j=1}^{J} n_j(\bar{\underline{y}}_{j\cdot}-\bar{\underline{y}}_{\cdot\cdot})(\bar{\underline{y}}_{j\cdot}-\bar{\underline{y}}_{\cdot\cdot})', \qquad E = \sum_{j=1}^{J}\sum_{l=1}^{n_j} (\tilde{\underline{y}}_{jl}-\bar{\underline{y}}_{j\cdot})(\tilde{\underline{y}}_{jl}-\bar{\underline{y}}_{j\cdot})'.$$

Elementy macierzy $H = \{h_{rs}\}$ i $E = \{e_{rs}\}$, $r, s = 1, 2, \ldots, p$ przyjmują następującą postać:

$$h_{rs} = \sum_{j=1}^{J} (\bar{y}_{rj.} - \bar{y}_{r..})(\bar{y}_{sj} - \bar{y}_{s..}),$$

$$e_{rs} = \sum \sum (y_{rjl} - \bar{y}_{rj.})(y_{sjl} - \bar{y}_{sj.}).$$

Morrison [11] podaje tabele, zawierające explicite wzory na elementy macierzy H i E dla podstawowych ortogonalnych układów eksperymentalnych. Dalsze przykłady testowania hipotez wielozmiennych na zasadzie analogii z hipotezami dla modeli jednozmiennych można znaleźć w książce Ahrensa [1].

4. Jednozmienne modele ortogonalne. Obliczanie zmienności odpowiadających efektom głównym rozważanych czynników i ich interakcjom.

Ogólny model jednozmienny zapisujemy w postaci równania (1), w którym t oznacza liczbę parametrów rozważanego modelu. Macierz $X = \{x_{ij}\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, t$ jest dana w postaci zero-jedynkowej; przy czym $x_{ij} = 0$ oznacza, że j-ty parametr nie ma wpływu na obserwację u i-tego osobnika. Aby ocenić efekt określonej kombinacji parametrów dobrze jest zaplanować doświadczenie w ten sposób, aby każda kombinacja czynników wystąpiła w odpowiedniej liczbie powtórzeń. Jeśli liczba powtórzeń jest taka sama dla każdej kombinacji rozważanych czynników, to mamy do czynienia z doświadczeniem orto-

gonalnym. Obliczenia dla takiego doświadczenia, chociaż żmudne przy większej liczbie czynników, są jednak znacznie uproszczone w porównaniu z ilością obliczeń niezbędnych przy doświadczeniach nieortognalnych.

Przypuśćmy, że rozważamy doświadczenie w układzie krzyżowym z trzema czynnikami, oznaczanymi umownie literami A, B, C. Czynniki te występują odpowiednio na I, J, K poziomach. Każda kombinacja $\langle i,j,k\rangle$, $i = 1, 2,\ldots, I$; $j = 1, 2,\ldots, J$; $k = 1, 2,\ldots, K$ występuje w L powtórzeniach. Przy tych oznaczeniach każdą z obserwowanych w doświadczeniu wartości zmiennej y można zapisać w postaci:

$$(12)\qquad y_{ijkl} = \mu+A_i+B_j+C_k+(AB)_{ij}+(AC)_{ik}+(BC)_{jk}+(ABC)_{ijk}+e_{ijkl},$$

przy czym symbole μ, A_i, B_j, C_k, $(AB)_{ij}$, $(AC)_{ik}$, $(BC)_{jk}$, $(ABC)_{ijk}$ oznaczają parametry modelu i są interpretowane jako średnie generalne, efekty główne poszczególnych poziomów badanych czynników, interakcje podwójne i interakcja potrójna poszczególnych poziomów badanych czynników. Parametry ze wskaźnikami spełniają określone warunki uboczne (restrykcje), jak to podaje np. Scheffé [12] lub Searle [13]. Symbol e_{ijkl} oznacza tzw. błąd lub niedopasowanie modelu.

Zmienność resztową rozważanego modelu (12) obliczamy ze wzoru:

$$SS_E = SS_T-SS_A-SS_B-SS_C-SS_{AB}-SS_{AC}-SS_{BC}-SS_{ABC},$$

gdzie:

$$SS_T = \sum_{i,j,k,l} (y_{ijkl}-\bar{y}....)^2,$$

$$SS_A = \sum_{i,j,k,l} (\hat{A}_i)^2 = \sum_{i,j,k,l} (\bar{y}_i...-\bar{y}....)^2 =$$

$$= JKL \sum_i (\bar{y}_{i\cdots} - \bar{y}_{\cdots\cdot})^2,$$

$$SS_B = \sum_{i,j,k,l} (\hat{B}_j)^2 = \sum_{i,j,k,l} (\bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdots\cdot})^2 =$$

$$= IKL \sum_j (\bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdots})^2,$$

$$SS_C = \sum_{i,j,k,l} (\hat{C}_k)^2 = \sum_{i,j,k,l} (\bar{y}_{\cdot\cdot k\cdot} - \bar{y}_{\cdots\cdot})^2 =$$

$$= IJL \sum_k (\bar{y}_{\cdot\cdot k\cdot} - \bar{y}_{\cdots\cdot})^2,$$

$$SS_{AB} = \sum_{i,j,k,l} (\widehat{(AB)}_{ij})^2 = \sum_{i,j,k,l} (\bar{y}_{ij\cdot\cdot} - \bar{y}_{i\cdots} - \bar{y}_{\cdot j\cdot\cdot} + \bar{y}_{\cdots\cdot})^2,$$

$$SS_{AC} = \sum_{i,j,k,l} (\widehat{(AC)}_{ik})^2 = \sum_{i,j,k,l} (\bar{y}_{i\cdot k\cdot} - \bar{y}_{i\cdots} - \bar{y}_{\cdot\cdot k\cdot} + \bar{y}_{\cdots\cdot})^2,$$

$$SS_{BC} = \sum_{i,j,k,l} (\widehat{(BC)}_{jk})^2 = \sum_{i,j,k,l} (\bar{y}_{\cdot jk\cdot} - \bar{y}_{\cdot j\cdot\cdot} - \bar{y}_{\cdot\cdot k\cdot} + \bar{y}_{\cdots\cdot})^2,$$

$$SS_{ABC} = \sum_{i,j,k,l} (\widehat{(ABC)}_{ijk})^2 = \sum_{i,j,k,l} (\bar{y}_{ijk\cdot} - \bar{y}_{ij\cdot\cdot} - \bar{y}_{i\cdot k\cdot} -$$

$$- \bar{y}_{\cdot jk\cdot} + \bar{y}_{i\cdots} + \bar{y}_{\cdot j\cdot\cdot} + \bar{y}_{\cdot\cdot k\cdot} - \bar{y}_{\cdots\cdot})^2.$$

Ogólnie przy rozważaniu doświadczenia z K czynnikami można wyodrębnić 2^K źródeł zmienności wyrażających wpływ efektów głównych oraz różnych interakcji.

Przypatrując się wzorom na zmienności wypisane wyżej stwierdzamy, że każda ze zmienności może być scharakteryzowana odpowiednim ciągiem literowym. Kodując występowanie określonej litery w danym ciągu literowym symbolem "1", natomiast jej brak symbolem "0" otrzymujemy ciąg zero-jedynkowy, który z kolei można odczytywać jako rozwinięcie dwójkowe pewnej

liczby całkowitej. W ten sposób każda zmienność może być w sposób jednoznaczny identyfikowana odpowiadającą jej liczbą całkowitą.

W doświadczeniu z powtórzeniami wprowadzamy dodatkowy (fikcyjny) czynnik oznaczający powtórzenia. W rozważanym przykładzie z czynnikami A, B, C dodatkowy czynnik oznaczający powtórzenia, otrzyma nazwę D. W tym przypadku poszczególne zmienności będą oznaczane liczbami całkowitymi od 1 do 2^4-1, jak to pokazuje tablica 1.

Powyższy sposób porządkowania zmienności został zaproponowany przez Yatesa.

Tablica 1. Przykład kodowania zmienności w doświadczeniu z K = 4 czynnikami

Nazwa zmienności	Układ literowy	Układ zerojedynkowy	Liczba całkowita
SS_A	␣ ␣ ␣ A	0001	1
SS_B	␣ ␣ B ␣	0010	2
SS_{AB}	␣ ␣ B A	0011	3
SS_C	␣ C ␣ ␣	0100	4
SS_{AC}	␣ C ␣ A	0101	5
SS_{BC}	␣ C B ␣	0110	6
SS_{ABC}	␣ C B A	0111	7
SS_D	D ␣ ␣ ␣	1000	8
SS_{AD}	D ␣ ␣ A	1001	9
SS_{BD}	D ␣ B ␣	1010	10
SS_{ABD}	D ␣ B A	1011	11
SS_{CD}	D C ␣ ␣	1100	12
SS_{ACD}	D C ␣ A	1101	13
SS_{BCD}	D C B ␣	1110	14
SS_{ABCD}	D C B A	1111	15

Można podać ogólną regułę na tworzenie składników sumy $SS_{\langle KOMB\rangle}$, gdzie $\langle KOMB\rangle$ oznacza dowolną kombinację literową rozważanych czynników (por. Scheffé [12]). Niech np. rozważaną kombinacją liter będzie kombinacja ABD. Dla danych postaci $\{y_{ijkl}\}$ zmienność wywołana interakcją potrójną ABD jest postaci:

$$SS_{ABD} = \sum_{i,j,k,l} (\bar{y}_{ij\cdot 1} - \bar{y}_{ij\cdot\cdot} - \bar{y}_{i\cdot\cdot 1} - \bar{y}_{\cdot j\cdot 1} + \bar{y}_{i\cdot\cdot\cdot} + \bar{y}_{\cdot j\cdot\cdot} + \bar{y}_{\cdot\cdot\cdot 1} - \bar{y}_{\cdot\cdot\cdot\cdot})^2.$$

Dla danej kombinacji $\langle KOMB\rangle$ składającej się z s liter kolejne składniki $SS_{\langle KOMB\rangle}$ można tworzyć kolejno w s+1 etapach poprzez następujące czynności:

a) Najpierw tworzymy średnią brzegową taką, że litery nie występujące w danej kombinacji zostają zastąpione kropką, natomiast litery w tej kombinacji występujące zostają zastąpione wskaźnikami sumowania. Jest to pierwszy etap czynności.

b) Następnie dla j = 1, 2,..., s tworzy się dalsze elementy według następującej zasady:

W elemencie utworzonym w pierwszym etapie (opisanym w punkcie a)) j wskaźników zostaje zastąpionych kropką, a powstała w ten sposób średnia brzegowa zostaje pomnożona przez czynnik $(-1)^j$. Rozważając wszystkie możliwe kombinacje opuszczanych wskaźników otrzymujemy dla każdej wartości j $\binom{s}{j}$ dalszych średnich brzegowych.

Po wykonaniu czynności opisanych w punkcie a) i b) otrzymujemy w sumie 2^s składników jednego kwadratowego wyrazu sumy $SS_{\langle KOMB\rangle}$.

Hartley [9], [10] podał algorytm, obliczający dla doświadczenia z K czynnikami zmienności odpowiadające dowolnej kombinacji rozważanych czynników.

Zasadniczymi elementami algorytmu Hartley'a są trzy operatory stosowane do tablicy obserwacji, nazwane przez niego Σ, Δ oraz $(\ )^2$.

$\sum_i$, operator sumujący, sumuje poprzez wszystkie poziomy wskaźnika i (i = 1, 2,..., I) pozostawiając pozostałe wskaźniki na tych samych poziomach.

Δ_i, operator różnicujący, mnoży poszczególne wyrazy przez I i odejmuje od tak otrzymanej liczby rezultat $\sum_i$.

$(...)^2$, operator obliczający średnie kwadraty, tworzy sumę kwadratów dla wszystkich elementów stojących wewnątrz nawiasu i dzieli otrzymaną sumę przez liczbę sumowanych elementów.

Aby otrzymać zmienność $SS_{\langle KOMB \rangle}$ odpowiadającą interakcji s-tego rzędu między czynnikami opisanymi daną s-literową kombinacją KOMB należy:

a) zastosować do tablicy danych operator Δ dla wskaźników odpowiadających literom występującym w kombinacji KOMB;

b) zastosować operator Σ do pozostałych wskaźników (uwaga: operator Σ stosujemy dopiero po wykonaniu wszystkich operacji Δ);

c) do otrzymanych wyrazów zastosować operator $(\ )^2$.

W rezultacie działań wymienionych w punktach a), b), c) otrzymujemy wielkość $N \cdot SS_{\langle KOMB \rangle}$, gdzie N jest całkowitą liczbą obserwacji (np. dla omawianego przedtem przykładu N = IJKL).

PRZYKŁAD 4.1. Chcąc obliczyć dla rozważanego przedtem doświadczenia z czynnikami A, B, C, D zmienność SS_A postępujemy następująco:

1. Stosując operator Δ do danych y_{ijkl}, $i = 1,\ldots, I$, $j = 1,\ldots, J$, $k = 1,\ldots, K$, $l = 1,\ldots, L$, otrzymujemy w rezultacie na miejscu każdej wartości y_{ijkl} odpowiednią różnicę:

$$\Delta_i \{y_{ijkl}\} \Rightarrow \{Iy_{ijkl} - y_{.jkl}\}.$$

2. Stosując operatory $\sum_j \sum_k \sum_l$ do tablicy różnic wyznaczonych w punkcie 1 otrzymujemy z odpowiednim mnożnikiem różnice między średnimi brzegowymi przy różnych poziomach czynnika A a średnią generalną:

$$\sum_j \sum_k \sum_l \{Iy_{ijkl} - y_{.jkl}\} \Rightarrow \{Iy_{i...} - y_{....}\} = \{IJK(\bar{y}_{i...} - \bar{y}_{....})\}.$$

Otrzymany zbiór składa się z I elementów.

3. Stosując do otrzymanego w punkcie 2 zbioru operator obliczający średnie kwadraty otrzymujemy z dokładnością do stałego współczynnika N szukaną zmienność SS_A:

$$(\{IJKL(\bar{y}_{i...} - \bar{y}_{....})^2\})^2 \Rightarrow NSS_A.$$

PRZYKŁAD 4.2. Chcąc obliczyć zmienność wywołaną interakcją czynników A, C postępujemy następująco:

1. Stosując operatory Δ_i i Δ_k do danych y_{ijkl}, $i = 1,\ldots, I$, $j = 1,\ldots, J$, $k = 1,\ldots, K$, $l = 1,\ldots, L$, otrzymujemy na miejscu każdej wartości y_{ijkl} odpowiednią różnicę składającą się z czterech składników:

$$\underset{i}{\Delta}\,\underset{k}{\Delta}\,\{y_{ijkl}\} \Rightarrow \{IKy_{ijkl} - Ky_{.jkl} - Iy_{ij.l} + y_{.j.l}\}.$$

2. Stosując do tablicy wyników otrzymanych w punkcie 1 obliczeń operatory sumowania $\underset{j}{\Sigma}$ i $\underset{l}{\Sigma}$ otrzymujemy tablicę średnich brzegowych:

$$\underset{i}{\Sigma}\,\underset{j}{\Sigma}\,\{IKy_{ijkl} - Ky_{.jkl} - Iy_{ij.l} + y_{.j.l}\} \Rightarrow$$
$$\Rightarrow \{IKy_{i.k.} - Ky_{..k.} - Iy_{i...} + y_{....}\} =$$
$$= \{IJKL(\bar{y}_{i.k.} - \bar{y}_{..k.} - \bar{y}_{i...} + \bar{y}_{....})\}.$$

Otrzymana w rezultacie działania operatorów $\underset{j}{\Sigma}$ i $\underset{l}{\Sigma}$ tablica zawiera IK elementów.

3. Stosując do otrzymanych w punkcie 2 wyników operator obliczający średnie kwadraty otrzymujemy szukaną zmienność SS_{AC} pomnożoną przez stały współczynnik:

$$(\{IJKL(\bar{y}_{i.k.} - \bar{y}_{..k.} - \bar{y}_{i...} + \bar{y}_{....}\})^2 \Rightarrow$$

$$\Rightarrow (IJKL)^2 \sum_{i=1}^{I} \sum_{k=1}^{K} (\bar{y}_{i.k.} - \bar{y}_{..k.} - \bar{y}_{i...} + \bar{y}_{....})^2 =$$

$$= N \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} (\bar{y}_{i.k.} - \bar{y}_{..k.} - \bar{y}_{i...} + \bar{y}_{....})^2 = NSS_{AC}.$$

W pakiecie podprogramów Scientific Subroutine Package IBM oraz w bibliotece podprogramów m.c. serii Odra 1300 znajdują się trzy podprogramy realizujące w języku FORTRAN omawiany algorytm Hartley'a. Są to procedury AVDAT, AVCAL i MEANQ [5].

Podprogram AVDAT wczytuje dane y_{ijk} w kolejności leksykograficznej i umieszcza je w odpowiednich miejscach jednowymiarowej tablicy danych.

Podprogram AVCAL oblicza za pomocą operatorów "Σ" i "Δ"

odpowiednie różnice i umieszcza je w tej samej tablicy danych.

Podprogram MEANQ oblicza za pomocą operatora "suma kwadratów" odpowiednie sumy, odpowiadające im stopnie swobody oraz średnie kwadraty.

W ten sposób dla doświadczenia z dowolną liczbą czynników otrzymujemy zmienności odpowiadające dowolnej kombinacji tych czynników.

Kolejność czynników jest określona tablicą LEVEL. Np. dla doświadczenia z trzema czynnikami A, B, C występującymi odpowiednio na I = 2, J = 2 i K = 3 poziomach po nadaniu elementom tablicy LEVEL wartości LEVEL(1) = 2, LEVEL(2) = 2, LEVEL(3) = 3 należy wczytać dane y_{ijk} w następującej kolejności:

y_{111} y_{211} y_{121} y_{221} y_{112} y_{212} y_{122} y_{222} y_{113} y_{213} y_{123} y_{223}.

Otrzymamy wtedy tablicę sum SS zawierającą poprawione zmienności w następującym porządku:

SS(1) = SS_A, SS(2) = SS_B, SS(3) = SS_{AB}, SS(4) = SS_C,
SS(5) = SS_{AC}, SS(6) = SS_{BC}, SS(7) = SS_{BC}.

Gdybyśmy natomiast nadali elementom tablicy LEVEL wartości LEVEL(1) = 3, LEVEL(2) = 2, LEVEL(3) = 2, oraz wczytali dane y_{ijk} w kolejności:

y_{111} y_{112} y_{113} y_{121} y_{122} y_{123} y_{211} y_{212} y_{213} y_{221} y_{222} y_{223},

to otrzymalibyśmy tablicę SS zawierającą poprawione zmienności w następującym porządku:

SS(1) = SS_C, SS(2) = SS_B, SS(3) = SS_{BC}, SS(4) = SS_A,
SS(5) = SS_{AC}, SS(6) = SS_{AB}, SS(7) = SS_{ABC}.

W omawianym pakiecie podprogramów m.c. Odra 1305 nie ma podprogramu obliczającego średnie brzegowe. Oryginalne dane w czasie działania podprogramu AVCAL zostają zniszczone, i tym samym nie ma możliwości korzystania z nich celem obliczenia średnich.

Gower [8] zaproponował inny, znacznie ogólniejszy algorytm umożliwiający otrzymanie zmienności odpowiadających dowolnej kombinacji czynników, oraz opublikował procedury realizujące ten algorytm w języku ALGOL 60.

Nieco uproszczoną, ale za to znacznie szybszą wersję tego algorytmu w postaci trzech procedur o nazwach means, squares, searchmeans opracowała Bartkowiak [4].

Procedura means wczytuje dane i umieszcza je w sposób rozrzucony w jednowymiarowej tablicy będącej podstawą do dalszych obliczeń. W czasie wczytywania danych są obliczane jednocześnie wszystkie możliwe średnie brzegowe.

Procedura squares oblicza sumy kwadratów dla zmienności odpowiadających efektom głównym i interakcjom badanych czynników. Najpierw oblicza się sumy kwadratów tzw. "niepoprawione", które następnie zostają przekształcone w tzw. "poprawione" sumy kwadratów. Jednocześnie ze zmiennościami zostają obliczone odpowiadające im stopnie swobody.

Procedura searchmeans, trzecia z omawianego zestawu, umożliwia wydobycie z obliczanej tablicy w odpowiedniej kolejności poszczególnych średnich brzegowych oraz wydrukowanie ich w odpowiedniej kolejności, z możliwością wydrukowania również odchylenia standardowego tych średnich oraz przedziału ufności.

5. Testowanie hipotez statystycznych

Mając obliczone tzw. poprawione sumy kwadratów, czyli zmienności odpowiadające rozmaitym kombinacjom rozważanych K czynników łatwo sformalizować testowanie hipotez dotyczących efektów głównych oraz interakcji rozważanych czynników dla następujących schematów doświadczalnych:

a) doświadczenie z K czynnikami w układzie krzyżowym bez powtórzeń;

b) doświadczenie z K czynnikami w układzie krzyżowym z powtórzeniami;

c) doświadczenie z K czynnikami w układzie krzyżowym i powtórzeniach w blokach;

d) doświadczenie z K czynnikami w układzie hierarchicznym rozszczepianych poletek doświadczalnych (metoda podbloków losowanych, termin angielski: split-split-plot).

Implementację programu, realizujące cztery wymienione schematy doświadczalne dla modeli o czynnikach stałych zrealizowała Bartkowiak [3]. Testowanie hipotez dla modeli mieszanych, tj. takich, w których niektóre parametry uważa się za losowe, nastręcza różne trudności formalne i jest tematem wielu publikacji, których nie będziemy tu omawiać.

Warto w tym miejscu wspomnieć, że przy wykonywaniu obliczeń na maszynie cyfrowej przyjmowanie lub odrzucanie testowanych hipotez odbywa się nie na podstawie tzw. wartości krytycznych (tablicowych) rozkładu F Snedecora, odczytanych dla zadeklarowanego poziomu istotności α, lecz wyznacza się

wręcz dla obliczonej wartości statystyki F odpowiednie prawdopodobieństwo przekroczenia tej wartości w rozkładzie F. Jeśli prawdopodobieństwo to jest małe, to odrzucamy testowaną hipotezę.

Wobec powyższego otrzymujący wyniki obliczeń nie jest zmuszony do odszukiwania odpowiednich wartości krytycznych w specjalnych tablicach statystycznych.

6. Jednozmienne modele nieortogonalne

Obliczenia dla nieortogonalnych modeli analizy wariancji są znacznie żmudniejsze i dopiero w ostatnich latach dzięki możliwościom korzystania z usług maszyn cyfrowych stały się bardziej dostępne. W stosunkowo częstym użyciu jest dobrze opisany szczególny przypadek dla K = 2, czyli analiza wariancji z dwoma czynnikami w układzie krzyżowym lub hierarchicznym z powtórzeniami lub bez (por. Scheffé [12], Searle [13], Bartkowiak [3]).

Dużo programów na m.c. Odra 1204 dla różnych układów doświadczeń opublikował Ośrodek Poznański w wydawanej przez siebie serii Roczników A.R. pod tytułem Algorytmy Biometryczne i Statystyczne (dotychczas ukazało się 6 zeszytów z tej serii). W szczególności warto wspomnieć w tym miejscu realizację algorytmu analizy wariancji dla losowego nieortogonalnego modelu hierarchicznego opracowaną przez Calińskiego i Kalę [6], analizę wariancji dla nieortogonalnych klasyfikacji krzyżowych bez interakcji metodą Reesa opracowaną przez

Cerankę i Mejzę, oraz analizę wariancji dla klasyfikacji krzyżowych metodą Bocka opracowaną przez Baksalarego, Kalę i Katulską [2].

Bibliografia

[1] Ahrens H. i Läuter J., Mehrdimensionale Varianzanalyse, Berlin 1974.

[2] Baksalary J., Kala R. i Katulska K., Analiza wariancji dla klasyfikacji krzyżowych metodą Bocka, ABS-51, Algorytmy Biometryczne i Statystyczne, zeszyt 6, Poznań 1977, str. 3-32.

[3] Bartkowiak A., Opis merytoryczny programów statystycznych opracowanych w Instytucie Informatyki Uniwersytetu Wrocławskiego, Wydawnictwa Uniwersyteckie, Wrocław 1978.

[4] Bartkowiak A., Calculation of all marginal means from an n-way table, Algorithm 38, Evaluation of corrected sums of squares for analysis of variance in a factorial design with n factors, Algorithm 39, Search for marginal means at given factor levels in an n-way table containing data scores and all marginal means, Algorithm 40, Zastosowania Matematyki, XIV, 4 (1975), 647-662.

[5] Biblioteka podprogramów statystycznych, Dokumentacja, Zakład Informatyki i Cybernetyki Instytutu Matematyki U.Ł., Ośrodek Obliczeniowy, Łódź - maj 1975.

[6] Caliński T. i Kala R., Analiza wariancji dla losowego nieortogonalnego modelu hierarchicznego, ABS-25, Algorytmy Biometryczne i Statystyczne, zeszyt 3, Poznań 1974, str. 75-91.

[7] Ceranka B. i Mejza I., Analiza wariancji dla nieortogonalnych klasyfikacji krzyżowych bez interakcji metodą Reesa, ABS-46, Algorytmy Biometryczne i Statystyczne, zeszyt 6, Poznań 1977, str. 3-32.

[8] Gower J.C., Analysis of variance for a factorial table, Algorithm AS-19, Applied Statistics 18 (1969), 199-202.

[9] Hartley H.O., A plan for programming analysis of variance for general purpose computers, Biometrics 12 (1956), 110-112.

[10] Hartley H.O., Analysis of variance, in: Mathematical methods for digital computers, ed. by A. Ralston and H. Wilf, Wiley 1962, Chapter 20.

[11] Morrison, D.F., Multivariate statistical method, New York 1967.

[12] Scheffé H., The analysis of variance, New York 1959.

[13] Searle S.R., Linear models, New York 1971.