Listy Biometryczne - Biometrical Letters
Vol.28 (1991), No.1, 45-68

ANALIZA PODTABLIC KONTYNGENCJI ZA POMOCĄ MODELU LOGARYTMICZNO-LINIOWEGO

Paweł Krajewski

Instytut Genetyki Roślin PAN,
Strzeszyńska 34, 60-479 Poznań

Streszczenie

W pracy przedstawione są metody wyjaśniania nieadekwatności modelu logarytmiczno-liniowego wielowymiarowej tablicy kontyngencji za pomocą testowania zgodności modeli implikowanych. Podane są procedury testowania indywidualnego oraz jednoczesnego modeli implikowanych. Dyskutowany jest problem wyboru rodziny modeli implikowanych, w szczególności dla klasy tzw. modeli rozkładalnych. Przedstawiona jest przykładowa analiza tablicy kontyngencji uzyskanej w badaniach dotyczących odporności odmian rzepaku na choroby grzybowe.

1. WSTĘP

Analiza wielowymiarowych tablic kontyngencji za pomocą modelu logarytmiczno-liniowego służy do badania zależności (interakcji) wielu zmiennych dyskretnych obserwowanych w pewnej populacji. Podstawy teoretyczne tej analizy przedstawione są m.in. w monografiach Bishop, Fienberga i Hollanda (1975), Habermana (1978), Placketta (1981) i Wickensa (1989), przeglądowych pracach Imrey'a, Kocha i Stokes (1981,1982) oraz w pracy Czajki i Krajewskiego (1986).

W niniejszej pracy przedstawiamy procedury mające zastosowanie w sytuacji, gdy pewien model logarytmiczno-liniowy tablicy kontyngencji zostanie uznany za nieadekwatny do sytuacji zaobserwowanej. Celem analizy statystycznej jest wtedy wyjaśnienie nieadekwatności poprzez weryfikację modeli bardziej szczegółowych. Metody opisane w pracy są właściwe dla tablic generowanych przez dowolną liczbę zmiennych. Przedstawiona dyskusja obejmuje całą klasę hierarchicznych modeli logarytmiczno-liniowych. W swej warstwie teoretycznej metody bazują na pracach Gabriela (1966) oraz Goodmana (1971) dotyczących tablic dwuwymiarowych i modelu niezależności.

Słowa kluczowe: tablice kontyngencji, model logarytmiczno-liniowy, model implikowany, testowanie jednoczesne,

2. TABLICA KONTYNGENCJI I MODEL LOGARYTMICZNO-LINIOWY

S-wymiarowa tablica kontyngencji powstaje w wyniku obserwacji S zmiennych dyskretnych, które oznaczamy dalej przez $A_1,\dots,A_S$. Wśród tych zmiennych wyróżniamy czynniki, to znaczy zmienne które a priori dzielą badaną populację na warstwy, oraz zmienne losowe. Odpowiednio do takiego rozróżnienia dzielimy zbiór numerów zmiennych $\mathcal{F} = \{1,2,\dots,S\}$ na dwa podzbiory: zbiór numerów czynników $\mathcal{F}_1$ i zbiór numerów zmiennych losowych $\mathcal{F}_2$. W dalszych rozważaniach bierzemy pod uwagę tylko takie sytuacje, w których $\mathcal{F}_2 \neq \emptyset$, gdyż przy $\mathcal{F}_2=\emptyset$ struktura populacji ze względu na obserwowane zmienne, określona przez liczebności prób pobranych z poszczególnych warstw, jest całkowicie znana przed przeprowadzeniem doświadczenia i nie może podlegać analizie statystycznej. Nie wykluczamy natomiast sytuacji, w których $\mathcal{F}_1=\emptyset$; tablica powstaje wtedy w wyniku obserwacji S zmiennych losowych w jednej populacji.

Oznaczmy dalej przez $T(\alpha)$, gdzie α jest dowolnym podzbiorem zbioru $\mathcal{F}$, liczbę podklas tablicy generowanej przez zmienne A_s, $s \in \alpha$. Liczba ta wynosi

$$T(\alpha) = \prod_{s\in\alpha} I_s ,$$

gdzie I_s oznacza liczbę kategorii zmiennej A_s. Przy takim oznaczeniu, liczba wszystkich podklas tablicy zapisana może być jako $T = T(\mathcal{F})$, zaś liczba warstw wyznaczonych przez czynniki jako $T_F = T(\mathcal{F}_1)$. Niech $\mathbf{y} = (y_1,\dots,y_T)'$ oraz $\mathbf{p} = (p_1,\dots,p_T)'$ oznaczają odpowiednio wektor liczebności obserwowanych podklas tablicy oraz wektor nieznanych prawdopodobieństw zdarzeń polegających na zaliczeniu elementu próby do poszczególnych podklas (identyfikując podklasę za pomocą jednego wskaźnika t (=1,2,...,T) zakładamy, iż podklasy są uporządkowane kolejno według kategorii pierwszej, drugiej,..., S-tej zmiennej). Zakładamy, że tablica jest kompletna, co oznacza iż $p_t > 0$, t=1,...,T. Mamy również:

$$\mathbf{c}_h'\mathbf{y} = N_h , \quad \mathbf{c}_h'\mathbf{p} = 1 , \quad h = 1,\dots,T_F ,$$

gdzie N_h oznacza liczebność próby pobranej z h-tej warstwy, natomiast $\mathbf{c}_h = (c_{h1},\dots,c_{hT})'$, przy czym

$$c_{ht} = \begin{cases} 1, & \text{jeżeli t-ta podklasa należy do h-tej warstwy,} \\ 0, & \text{w przeciwnym wypadku.} \end{cases}$$

Sposób uzyskania tablicy kontyngencji powoduje, iż wektor $\mathbf{y}$ stanowi obserwację zmiennej losowej o rozkładzie iloczynowo-wielomianowym, to znaczy rozkładzie określonym funkcją prawdopodobieństwa będącą iloczynem T_F funkcji opisujących rozkład wielomianowy (por. np. Bishop i in., 1975, str.63). Wektor wartości oczekiwanych zmiennej o takim rozkładzie to wektor $\mathbf{m} = (m_1,\dots,m_T)'$, przy czym $m_t = N_h p_t$, jeżeli t-ta podklasa należy do h-tej warstwy. Spełniony jest oczywiście warunek $\mathbf{c}_h'\mathbf{m} = N_h$, h=1,...,$T_F$.

Jeżeli $\mathcal{F}_1=\emptyset$, to rozkład iloczynowo-wielomianowy sprowadza się do rozkładu wielomianowego.

Wprowadźmy teraz pewne oznaczenia niezbędne do zapisania modelu logarytmiczno-liniowego tablicy kontyngencji. Niech Ω oznacza zbiór (rodzinę) wszystkich 0,1,2,...,S-elementowych uporządkowanych podzbiorów zbioru $\mathcal{F}$. Niech dla dowolnych elementów α, β zbioru Ω takich, że $\alpha \subseteq \beta$ i $\beta \neq \emptyset$, $\mathbf{X}_\alpha^\beta$ oznacza macierz postaci

$$\mathbf{X}_\alpha^\beta = \bigotimes_{s\in\beta} \mathbf{B}_s \quad , \tag{2.1}$$

gdzie $\otimes$ oznacza operator iloczynu kroneckerowskiego, natomiast

$$\mathbf{B}_s = \begin{cases} \mathbf{I}_{I_s}, & s \in \alpha, \\ \mathbf{1}_{I_s}, & s \in \beta\setminus\alpha, \end{cases}$$

przy czym $\mathbf{I}_{I_s}$ i $\mathbf{1}_{I_s}$ oznaczają odpowiednio macierz jednostkową stopnia I_s oraz I_s-wymiarowy wektor jedynek (w wypadku gdy $\beta=\mathcal{F}$ dla oznaczenia macierzy zdefiniowanej przez (2.1) używać będziemy symbolu $\mathbf{X}_\alpha$).

Przy wprowadzonych oznaczeniach, model logarytmiczno-liniowy tablicy kontyngencji można zapisać w postaci

$$\mathrm{Log}\ \mathbf{m} = \sum_{\alpha\in\mathcal{M}} \mathbf{X}_\alpha \mathbf{u}_\alpha \quad , \tag{2.2}$$

gdzie $\mathcal{M} \subseteq \Omega$, a $\mathbf{u}_\alpha$ oznacza wektor nieznanych parametrów związanych z zależnościami zmiennych losowych (bądź wpływami lub interakcjami czynników) A_s, $s \in \alpha$ (przy $\mathcal{M} = \Omega$ zapis (2.2) przedstawia tzw. model nasycony). Ograniczając rozważania do klasy modeli hierarchicznych możemy podać równoważne przedstawienie modelu w postaci

$$\mathrm{Log}\ \mathbf{m} = \sum_{\alpha\in\mathcal{G}(\mathcal{M})} \mathbf{X}_\alpha \mathbf{u}_\alpha \quad , \tag{2.3}$$

gdzie $\mathcal{G}(\mathcal{M})$ oznacza zbiór elementów maksymalnych zbioru $\mathcal{M}$ ze względu na relację zawierania, lub, oznaczając przez $\mathbf{X}_{\mathcal{G}(\mathcal{M})}$ macierz dzieloną złożoną z macierzy $\mathbf{X}_\alpha$, $\alpha \in \mathcal{G}(\mathcal{M})$, w postaci

$$\mathrm{Log}\ \mathbf{m} \in \mathbb{R}(\mathbf{X}_{\mathcal{G}(\mathcal{M})}) \quad , \tag{2.4}$$

gdzie $\mathbb{R}(\mathbf{X})$ oznacza przestrzeń rozpiętą na kolumnach macierzy $\mathbf{X}$. Zbiór $\mathcal{G}(\mathcal{M})$ nazywany jest klasą generującą modelu. W dalszym ciągu symbolu $\mathcal{M}$ będziemy używać także dla określenia samego modelu logarytmiczno-liniowego będącego przedmiotem rozważań; posługiwać się też będziemy sformułowaniem "model o klasie generującej $\mathcal{G}(\mathcal{M})$". Ograniczamy przy tym rozważania do modeli spełniających warunki

$$\bigcup_{\alpha\in\mathcal{G}(\mathcal{M})} \alpha = \mathcal{F} \quad , \tag{2.5}$$

oraz

$$\mathcal{F}_1 \in \mathcal{M} \quad . \tag{2.6}$$

Pierwszy z nich oznacza, że w modelu $\mathcal{M}$ występują parametry dotyczące wszystkich obserwowanych zmiennych losowych. Drugi związany jest z koniecznością uwzględnienia w modelu struktury prób losowych pobranych z poszczególnych warstw.

Przypomnijmy teraz podstawowe, znane z literatury fakty dotyczące analizy statystycznej modelu logarytmiczno-liniowego. Ocenę największej wiarygodności $\hat{\mathbf{m}}$ wektora liczebności oczekiwanych $\mathbf{m}$ przy prawdziwości modelu $\mathcal{M}$ otrzymuje się rozwiązując układ równań

$$X'_{\alpha}\mathbf{m} = X'_{\alpha}\mathbf{y} \quad , \quad \alpha \in \mathcal{G}(\mathcal{M}) \quad . \tag{2.7}$$

Jedną z metod numerycznych prowadzących do rozwiązania układu (2.7) jest metoda iteracyjnego dopasowania Deminga-Stephana. Test zgodności modelu $\mathcal{M}$ oparty jest na statystyce ilorazowej postaci

$$G^2(\mathcal{M}) = 2\mathbf{y}'(\mathrm{Log}\ \mathbf{y} - \mathrm{Log}\ \hat{\mathbf{m}}) \quad , \tag{2.8}$$

która ma asymptotycznie rozkład χ^2 o liczbie stopni swobody

$$v = T - \sum_{\alpha\in\mathcal{M}} \prod_{s\in\alpha} (I_s - 1) - 1 \quad . \tag{2.9}$$

Uznajemy więc, że model $\mathcal{M}$ jest zgodny, gdy

$$P\{\chi^2(v) > G^2(\mathcal{M})\} > \alpha \quad , \tag{2.10}$$

gdzie $\chi^2(v)$ oznacza zmienną losową o rozkładzie χ^2 z v stopniami swobody, natomiast α jest zadanym poziomem istotności, lub, równoważnie, wtedy gdy

$$G^2(\mathcal{M}) < \chi^2_{\alpha}(v) \ , \tag{2.11}$$

gdzie $\chi^2_{\alpha}(v)$ oznacza odpowiednią wartość krytyczną rozkładu χ^2.

Rozważmy na koniec tego paragrafu pewne założenia związane z wprowadzonym modelem probabilistycznym tablicy, które okażą się istotne w odniesieniu do modeli rozpatrywanych w paragrafie 3. Pierwsze z tych założeń mówi iż podklasy tablicy, definiowane poprzez zbiory kategorii poszczególnych zmiennych, są ustalone a priori. Drugie założenie dotyczy wielkości N_h, które traktowane są jako znane stałe. W większości sytuacji praktycznych założenia te są spełnione. Jest tak wtedy, gdy przed przeprowadzeniem losowania znamy wszystkie możliwe stany elementów populacji ze względu na każdą zmienną A_s, $s \in \mathcal{F}$, oraz jeżeli wielkość próby może być ustalona z góry. Często jednak możemy mieć wątpliwość co do spełnienia rozważanych warunków. Szczególnie w pewnych badaniach typu "obserwacyjnego" (np. badania populacji naturalnych, badania medyczne), próba pobierana jest w sposób sekwencyjny i proces ten, z różnych względów, jest w pewnym momencie przerywany. Liczebność próby nie jest wtedy stałą (jest zmienną losową), a także nie mamy pewności, czy w próbie reprezentowane są wszystkie kategorie badanych zmiennych. Pomimo to, w sytuacjach takich stosuje się zarówno model iloczynowo-wielomianowy, jak i metodę analizy struktury tablicy za pomocą modelu logarytmiczno-liniowego. Jest to słuszne dopóty, dopóki wnioskowanie w takich sytuacjach traktuje się jako warunkowe, to znaczy dotyczące "zaobserwowanych" kategorii zmiennych przy uzyskanej wielkości próby.

3. MODELOWANIE PODTABLIC

3.1. *Określenie modelu implikowanego*

Oznaczmy całą tablicę generowaną przez zmienne $A_1,\dots,A_S$ symbolem $\mathcal{T}$. Niech $\mathcal{K}_s$, $s \in \mathcal{S}$, oznacza pewien K_s-elementowy podzbiór zbioru kategorii zmiennej A_s. Załóżymy, iż $\mathcal{K}_s$ składa się przynajmniej z dwóch elementów. Przez $\mathcal{T}_1$ oznaczmy podtablicę tablicy $\mathcal{T}$ utworzoną z podklas odpowiadających elementom iloczynu kartezjańskiego $\mathcal{K}_1 \times \dots \times \mathcal{K}_S$. Jeżeli wprowadzimy oznaczenie

$$T_1(\alpha) = \prod_{s\in\alpha} K_s ,$$

to liczbę podklas $\mathcal{T}_1$ zapiszemy jako $T_1 = T_1(\mathcal{S})$.

Dla przeprowadzenia rozważań dotyczących modelu tablicy $\mathcal{T}_1$, oznaczmy przez $\mathbf{W}$ macierz o elementach równych 0 lub 1 "wybierającą" z wektora $\mathbf{m}$ liczebności oczekiwane podklas należących do tablicy $\mathcal{T}_1$. Macierz taką można zapisać w postaci

$$\mathbf{W} = \bigotimes_{s\in\mathcal{S}} \mathbf{V}_s , \tag{3.1.1}$$

gdzie z kolei $\mathbf{V}_s$, $s \in \mathcal{S}$, jest macierzą stopnia $K_s \times I_s$, której wiersze to wektory jednostkowe odpowiadające kategoriom zmiennej A_s należącym do zbioru $\mathcal{K}_s$, a więc macierzą "wybierającą" odpowiednie kategorie danej zmiennej. Używając macierzy $\mathbf{W}$, można wektory liczebności oczekiwanych i obserwowanych tablicy $\mathcal{T}_1$ zapisać odpowiednio w postaci $\mathbf{Wm}$ oraz $\mathbf{Wy}$. Jeżeli dodatkowo zdefiniujemy macierz $\bar{\mathbf{W}}$, jako taką która wybiera z wektora $\mathbf{m}$ liczebności oczekiwane podklas nie należących do $\mathcal{T}_1$, to można zauważyć, iż wektor $\mathbf{Um}$, gdzie

$$\mathbf{U} = \begin{bmatrix} \mathbf{W} \\ \bar{\mathbf{W}} \end{bmatrix} ,$$

składa się z liczebności oczekiwanych tablicy $\mathcal{T}$ uporządkowanych w ten sposób, iż pierwsze T_1 składowych to liczebności podklas $\mathcal{T}_1$, zaś pozostałe $T-T_1$ składowych to liczebności podklas nie należących do $\mathcal{T}_1$.

Zastanówmy się dalej nad strukturą wektora $\mathbf{Wm}$ przy prawdziwości dowolnego modelu $\mathcal{M}$ tablicy $\mathcal{T}$. Wykorzystując zapis (2.3) mamy

$$\operatorname{Log} \mathbf{Wm} = \mathbf{W} \operatorname{Log} \mathbf{m} = \sum_{\alpha\in\mathcal{G}(\mathcal{M})} \mathbf{WX}_\alpha \mathbf{u}_\alpha . \tag{3.1.2}$$

Zgodnie z (3.1.1) oraz (2.1)

$$\mathbf{WX}_\alpha = \bigotimes_{s\in\mathcal{S}} \mathbf{V}_s \bigotimes_{s\in\mathcal{S}} \mathbf{B}_s ,$$

gdzie

$$B_s = \begin{cases} I_{I_s}, & s\in\alpha, \\ 1_{I_s}, & s\in\mathcal{F}\setminus\alpha. \end{cases}$$

Korzystając z własności iloczynu kroneckerowskiego oraz zauważając, iż $V_s 1_{I_s} = 1_{K_s}$, otrzymujemy

$$WX_\alpha = \bigotimes_{s\in\mathcal{F}} C_s,$$

gdzie

$$C_s = \begin{cases} V_s, & s\in\alpha, \\ 1_{K_s}, & s\in\mathcal{F}\setminus\alpha. \end{cases}$$

Ze względu na tożsamości

$$V_s = I_{K_s} V_s, \quad 1_{K_s} = 1_{K_s} 1,$$

możemy dalej zapisać, iż

$$WX_\alpha = \bigotimes_{s\in\mathcal{F}} D_s \bigotimes_{s\in\alpha} V_s,$$

gdzie

$$D_s = \begin{cases} I_{K_s}, & s\in\alpha, \\ 1_{K_s}, & s\in\mathcal{F}\setminus\alpha, \end{cases}$$

lub

$$WX_\alpha = Z_\alpha W_\alpha,$$

gdzie

$$W_\alpha = \bigotimes_{s\in\alpha} V_s$$

jest macierzą wybierającą spośród kombinacji kategorii zmiennych A_s, $s\in\alpha$, te, które należą do podtablicy $\mathcal{T}_1$, natomiast Z_α jest postaci analogicznej do X_α (różnica między tymi macierzami tkwi tylko w rozmiarach definiujących je wektorów 1 i macierzy I). Wracając więc do (3.1.2), mamy

$$\mathrm{Log}\, Wm = \sum_{\alpha\in\mathcal{G}(\mathcal{M})} Z_\alpha W_\alpha u_\alpha$$

lub inaczej

$$\mathrm{Log}\, Wm = \sum_{\alpha\in\mathcal{G}(\mathcal{M})} Z_\alpha w_\alpha, \tag{3.1.3}$$

gdzie $w_\alpha = W_\alpha u_\alpha$, $\alpha\in\mathcal{G}(\mathcal{M})$, jest pewnym wektorem z przestrzeni $\mathbb{R}^{T_1(\alpha)}$. Zapis (3.1.3) oznacza, iż model analogiczny do modelu $\mathcal{M}$ jest prawdziwy dla podtablicy $\mathcal{T}_1$ będącej fragmentem tablicy $\mathcal{T}$. Inaczej mówiąc, prawdziwość dowolnego modelu $\mathcal{M}$ dla tablicy $\mathcal{T}$ implikuje prawdziwość analogicznego modelu dla podtablicy $\mathcal{T}_1$. Sformułowanie "model analogiczny" rozumieć tu należy jako model tablicy o innej liczbie podklas, jednak opisujący tę samą strukturę wektora liczebności oczekiwanych. Stosując symbol $\mathcal{M}[\mathcal{T}]$ dla oznaczenia modelu $\mathcal{M}$ dla tablicy $\mathcal{T}$, wykazany powyżej związek zapiszemy w postaci implikacji

$$\mathcal{M}[\mathcal{T}] \Rightarrow \mathcal{M}[\mathcal{T}_1] \quad , \tag{3.1.4}$$

która uzasadnia nazywanie $\mathcal{M}[\mathcal{T}_1]$ modelem implikowanym przez $\mathcal{M}[\mathcal{T}]$.

Zauważmy dalej, iż postać (3.1.3) modelu $\mathcal{M}[\mathcal{T}_1]$ można zapisać równoważnie jako

$$\begin{cases} \text{Log } \mathbf{Wm} \in \mathbb{R}(\mathbf{Z}_{\mathcal{G}(\mathcal{M})}) \quad , \\ \text{Log } \bar{\mathbf{W}}\mathbf{m} \in \mathbb{R}^{T-T_1} \quad , \end{cases} \tag{3.1.5}$$

gdzie $\mathbf{Z}_{\mathcal{G}(\mathcal{M})}$ jest macierzą dzieloną złożoną z macierzy $\mathbf{Z}_\alpha$, $\alpha\in\mathcal{G}(\mathcal{M})$. Taki zapis koresponduje z postacią (2.4) modelu ogólnego $\mathcal{M}[\mathcal{T}]$ i wskazuje, że model implikowany $\mathcal{M}[\mathcal{T}_1]$ możemy rozumieć jako model całej tablicy $\mathcal{T}$ nakładający efektywne ograniczenia tylko na liczebności oczekiwane podklas podtablicy $\mathcal{T}_1$.

Należy w tym miejscu poczynić ważną uwagę związaną ze stosowalnością, w rozpatrywanej tu sytuacji, modelu probabilistycznego podanego w paragrafie 2. Ponieważ $\mathbf{Wy}$ jest wektorem obserwowanych liczebności podklas tablicy $\mathcal{T}_1$, to $N_1 = \mathbf{1}'_{T_1}\mathbf{Wy}$ jest równe liczebności w pobranej próbie losowej tych elementów, które zaliczone zostały do podklas tablicy $\mathcal{T}$ należących do $\mathcal{T}_1$. Jeżeli $\mathcal{P}_1\neq\emptyset$ i tablica $\mathcal{T}_1$ jest wybrana w ten sposób, iż obejmuje wszystkie podklasy tablicy $\mathcal{T}$ odpowiadające pewnym wybranym kombinacjom kategorii czynników A_s, $s\in\mathcal{P}_1$, to wartość N_1 jest stałą. Jest ona równa liczebności jednej (lub sumarycznej liczebności kilku) z niezależnych prób składających się na całą pobraną próbę losową. Wtedy też odpowiedni model iloczynowo-wielomianowy może być stosowany do obserwowanej podtablicy kontyngencji. Jeżeli wymienione warunki nie są spełnione, a więc gdy $\mathcal{P}_1=\emptyset$, lub gdy podtablica $\mathcal{T}_1$ nie jest wybrana w określony powyżej sposób, to N_1 jest zmienną losową i odpowiednie modele probabilistyczne nie mają zastosowania. Wspomnijmy jednak uwagę kończącą paragraf 2. Zgodnie z nią, fakty dotyczące tablicy $\mathcal{T}$ możemy stosować w odniesieniu do podtablicy $\mathcal{T}_1$, jeżeli założymy iż N_1 jest stałą, równą wartości zaobserwowanej. Widać jednak, iż skutkiem tego jest pewna "warunkowość" wniosków wypowiadanych o tablicy $\mathcal{T}_1$.

Zajmijmy się dalej wyznaczeniem oceny największej wiarygodności $\tilde{\mathbf{m}}$ wektora liczebności oczekiwanych $\mathbf{m}$ przy prawdziwości modelu $\mathcal{M}[\mathcal{T}_1]$. Wykorzystamy w tym celu postać (3.1.5) modelu implikowanego. Stosując wzór (2.7) wnioskujemy iż $\tilde{\mathbf{m}}$ otrzymamy jako rozwiązanie układu równań

$$\begin{bmatrix} \mathbf{Z}'_{\mathcal{G}(\mathcal{M})} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{T-T_1} \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \bar{\mathbf{W}} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{Z}'_{\mathcal{G}(\mathcal{M})} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{T-T_1} \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \bar{\mathbf{W}} \end{bmatrix} \mathbf{y} \quad .$$

Układ powyższy można przekształcić do postaci

$$\begin{cases} \mathbf{Z}'_{\mathcal{G}(\mathcal{M})}\mathbf{Wm} = \mathbf{Z}'_{\mathcal{G}(\mathcal{M})}\mathbf{Wy} \quad , \\ \bar{\mathbf{W}}\mathbf{m} = \bar{\mathbf{W}}\mathbf{y} \quad . \end{cases} \tag{3.1.6}$$

Widoczne jest więc, że wektor $W\tilde{m}$ ocen liczebności oczekiwanych podklas należących do $\mathcal{T}_1$ można wyznaczyć zakładając dla Wm prawdziwość modelu $\mathcal{M}$ i postępując tak jak gdyby Wy było całą tablicą kontyngencji. Natomiast oceny oczekiwanych liczebności podklas nie należących do $\mathcal{T}_1$ są równe odpowiednim liczebnościom obserwowanym.

Powyższe wnioski dotyczące oceny $\tilde{m}$ wykorzystamy teraz do wyznaczenia statystyki ilorazowej służącej do testowania zgodności modelu $\mathcal{M}[\mathcal{T}_1]$. Zgodnie z (2.8), mamy

$$G^2(\mathcal{M}[\mathcal{T}_1]) = 2y'(\mathrm{Log}\ y - \mathrm{Log}\ \tilde{m}) = 2y'U'U(\mathrm{Log}\ y - \mathrm{Log}\ \tilde{m})$$
$$= 2y'W'[\mathrm{Log}\ Wy + \mathrm{Log}\ \overline{W}y - \mathrm{Log}\ W\tilde{m} - \mathrm{Log}\ \overline{W\tilde{m}}] .$$

Biorąc pod uwagę (3.1.6), otrzymujemy

$$G^2(\mathcal{M}[\mathcal{T}_1]) = 2y'W'(\mathrm{Log}\ Wy - \mathrm{Log}\ W\tilde{m}) . \qquad (3.1.7)$$

Statystyka ta, przy prawdziwości modelu $\mathcal{M}[\mathcal{T}_1]$, ma asymptotycznie rozkład χ^2 o liczbie stopni swobody v_1, przy czym, zgodnie ze wzorem (2.9),

$$v_1 = T_1 - \sum_{\alpha\in\mathcal{M}} \prod_{s\in\alpha} (K_s-1) - 1 . \qquad (3.1.8)$$

W świetle powyższych rozważań interesujące jest znalezienie relacji pomiędzy statystykami służącymi do badania zgodności modeli $\mathcal{M}[\mathcal{T}]$ i $\mathcal{M}[\mathcal{T}_1]$. Biorąc pod uwagę implikację (3.1.4) oraz monotoniczność statystyki G^2 ze względu na zawieranie przestrzeni definiujących modele logarytmiczno-liniowe (Bishop i in., 1975, str.523), mamy

$$G^2(\mathcal{M}[\mathcal{T}]) \geq G^2(\mathcal{M}[\mathcal{T}_1]) . \qquad (3.1.9)$$

Wymienione statystyki mogą służyć do testowania zgodności $\mathcal{M}[\mathcal{T}]$ pod warunkiem iż prawdziwy jest $\mathcal{M}[\mathcal{T}_1]$, bowiem

$$G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1]) = G^2(\mathcal{M}[\mathcal{T}]) - G^2(\mathcal{M}[\mathcal{T}_1]) . \qquad (3.1.10)$$

Zapisując (3.1.10) w postaci

$$G^2(\mathcal{M}[\mathcal{T}]) = G^2(\mathcal{M}[\mathcal{T}_1]) + G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1]) \qquad (3.1.11)$$

wnioskujemy, iż statystykę testującą zgodność modelu $\mathcal{M}$ w tablicy $\mathcal{T}$ można rozłożyć na dwa składniki, z których jeden opisuje zgodność modelu $\mathcal{M}$ w dowolnej podtablicy $\mathcal{T}_1$, a drugi zgodność $\mathcal{M}$ w $\mathcal{T}$ pod warunkiem, iż $\mathcal{M}$ jest prawdziwy w $\mathcal{T}_1$. Stosując (3.1.11) należy pamiętać, iż w ogólności statystyka $G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1])$ ma asymptotycznie rozkład χ^2 o liczbie stopni swobody $v-v_1$ tylko przy prawdziwości modelu $\mathcal{M}[\mathcal{T}]$ (Bishop i in., 1975, str.524). W praktyce służyć ona może do oceny poprawy zgodności modelu logarytmiczno-liniowego wynikającej z przejścia od "prostszego" (w sensie liczby parametrów) modelu $\mathcal{M}[\mathcal{T}]$ do "bardziej skomplikowanego" modelu $\mathcal{M}[\mathcal{T}_1]$.

3.2. Badanie zgodności wielu modeli implikowanych

Z praktycznego punktu widzenia interesujące jest rozpatrzenie sytuacji w której, po stwierdzeniu nieadekwatności modelu $\mathcal{M}[\mathcal{T}]$, przeprowadza się testowanie zgodności tego modelu dla wielu podtablic tablicy $\mathcal{T}$. Podtablice mogą być wyróżnione a priori lub wybierane w trakcie obliczeń. Poszczególne modele implikowane można testować w sposób indywidualny, to znaczy za pomocą zwykłej procedury opartej na statystyce G^2 postaci (3.1.7), przy ewentualnym wykorzystaniu dekompozycji (3.1.11). Pozwala to na kontrolę prawdopodobieństwa popełnienia błędu I rodzaju w każdym z przeprowadzonych testów. Często jednak wszystkie modele implikowane będące przedmiotem zainteresowania traktuje się jako rodzinę modeli. Niezbędne jest wtedy zastosowanie procedur testowania jednoczesnego. Procedury te umożliwiają kontrolę prawdopodobieństwa błędu polegającego na stwierdzeniu nieadekwatności przynajmniej jednego prawdziwego modelu. Podaniem takich procedur zajmiemy się w dalszym ciągu tego paragrafu. Na wstępie spróbujemy pokazać, w jaki sposób pewne relacje pomiędzy zbiorami podklas wyróżnionych podtablic wpływają na wybór procedury testowej oraz na związki pomiędzy zastosowanymi testami.

Rozpatrzmy najpierw wypadek, w którym dla dwóch spośród wyróżnionych podtablic, powiedzmy dla $\mathcal{T}_1$ i $\mathcal{T}_2$, prawdziwa jest (w sensie zawierania zbiorów podklas) relacja $\mathcal{T}_2 \subset \mathcal{T}_1$. Stosując rozumowanie analogiczne do przedstawionego w paragrafie 3.1 można pokazać, iż $\mathcal{M}[\mathcal{T}_2]$ jest modelem implikowanym przez $\mathcal{M}[\mathcal{T}_1]$, a więc uzasadnione jest rozważanie testu zgodności $\mathcal{M}[\mathcal{T}_1]$ pod warunkiem prawdziwości $\mathcal{M}[\mathcal{T}_2]$. Odpowiednia statystyka testowa ma (analogiczną do (3.1.10)) postać

$$G^2(\mathcal{M}[\mathcal{T}_1]|\mathcal{M}[\mathcal{T}_2]) = G^2(\mathcal{M}[\mathcal{T}_1]) - G^2(\mathcal{M}[\mathcal{T}_2]) \,. \tag{3.2.1}$$

Ze względu na relację $\mathcal{T}_2 \subset \mathcal{T}_1 \subset \mathcal{T}$, uzasadnione jest też w tym wypadku wykorzystanie, wynikającej z (3.1.11) i (3.2.1), dekompozycji statystyki $G^2(\mathcal{M}[\mathcal{T}])$ postaci

$$G^2(\mathcal{M}[\mathcal{T}]) = G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1]) + G^2(\mathcal{M}[\mathcal{T}_1]|\mathcal{M}[\mathcal{T}_2]) + G^2(\mathcal{M}[\mathcal{T}_2]) \quad .$$

Sugeruje ona możliwość postępowania hierarchicznego pozwalającego ocenić poprawę zgodności przy przejściu od modelu $\mathcal{M}[\mathcal{T}]$ do $\mathcal{M}[\mathcal{T}_1]$ oraz od $\mathcal{M}[\mathcal{T}_1]$ do $\mathcal{M}[\mathcal{T}_2]$.

Przedstawione powyżej uwagi przenoszą się na sytuację ogólną, w której R wyróżnionych tablic spełnia warunek

$$\mathcal{T}_R \subset \mathcal{T}_{R-1} \subset \ldots \subset \mathcal{T}_1 \,.$$

Zajmijmy się z kolei wypadkiem, w którym dwie spośród wyróżnionych tablic, oznaczone znowu przez $\mathcal{T}_1$ i $\mathcal{T}_2$, są rozłączne $(\mathcal{T}_1 \cap \mathcal{T}_2 = \emptyset)$. Dodatkowo rozważmy model tablicy $\mathcal{T}$ powstający przez koniunkcję modeli $\mathcal{M}[\mathcal{T}_1]$ i $\mathcal{M}[\mathcal{T}_2]$, a więc model stanowiący iż ograniczenia nakładane na liczebności oczekiwane przez model $\mathcal{M}$ są spełnione wewnątrz obu podtablic

$\mathcal{T}_1$ i $\mathcal{T}_2$. Oznaczmy taki model symbolem $\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]$. Przeprowadzając rozumowanie podobne do przedstawionego w paragrafie 3.1 można pokazać, iż ocena $\tilde{\mathbf{m}}_{12}$ wektora $\mathbf{m}$ przy prawdziwości modelu $\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]$ spełnia związki

$$\mathbf{W}_1\tilde{\mathbf{m}}_{12} = \mathbf{W}_1\tilde{\mathbf{m}}_1 \quad ,$$

$$\mathbf{W}_2\tilde{\mathbf{m}}_{12} = \mathbf{W}_2\tilde{\mathbf{m}}_2 \quad , \tag{3.2.2}$$

$$\bar{\mathbf{W}}\tilde{\mathbf{m}}_{12} = \bar{\mathbf{W}}\mathbf{y} \quad ,$$

przy czym $\mathbf{W}_1$, $\mathbf{W}_2$ oraz $\bar{\mathbf{W}}$ oznaczają odpowiednio macierze wybierające liczebności podklas należących do $\mathcal{T}_1$, należących do $\mathcal{T}_2$ oraz nie należących ani do $\mathcal{T}_1$ ani do $\mathcal{T}_2$, natomiast $\tilde{\mathbf{m}}_1$, $\tilde{\mathbf{m}}_2$ oznaczają oceny wektora liczebności oczekiwanych przy prawdziwości odpowiednio $\mathcal{M}[\mathcal{T}_1]$ i $\mathcal{M}[\mathcal{T}_2]$. Wzory (3.2.2) wyrażają fakt, że przy prawdziwości $\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]$ oceny liczebności oczekiwanych podklas należących do $\mathcal{T}_1$ są takie same jak przy prawdziwości $\mathcal{M}[\mathcal{T}_1]$ oraz oceny liczebności oczekiwanych podklas należących do $\mathcal{T}_2$ są takie same jak przy $\mathcal{M}[\mathcal{T}_2]$. Dla pozostałych podklas oceny te są równe liczebnościom obserwowanym. Przechodząc do wyznaczenia statystyki $G^2(\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2])$, otrzymujemy z (2.8)

$$G^2(\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]) = 2\mathbf{y}'[\mathrm{Log}\ \mathbf{y} - \mathrm{Log}\ \tilde{\mathbf{m}}_{12}]$$

$$= 2\mathbf{y}'\mathbf{U}'\mathbf{U}\ [\mathrm{Log}\ \mathbf{y} - \mathrm{Log}\ \tilde{\mathbf{m}}_{12}] \quad ,$$

gdzie $\mathbf{U}' = [\mathbf{W}_1' : \mathbf{W}_2' : \bar{\mathbf{W}}']$. Stąd

$$G^2(\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]) = 2\mathbf{y}'\mathbf{W}_1'[\mathrm{Log}\ \mathbf{W}_1\mathbf{y} - \mathrm{Log}\ \mathbf{W}_1\tilde{\mathbf{m}}_{12}]$$

$$+ 2\mathbf{y}'\mathbf{W}_2'[\mathrm{Log}\ \mathbf{W}_2\mathbf{y} - \mathrm{Log}\ \mathbf{W}_2\tilde{\mathbf{m}}_{12}]$$

$$+ 2\mathbf{y}'\bar{\mathbf{W}}'[\mathrm{Log}\ \bar{\mathbf{W}}\mathbf{y} - \mathrm{Log}\ \bar{\mathbf{W}}\tilde{\mathbf{m}}_{12}] \quad ,$$

zaś po uwzględnieniu (3.2.2) i (3.1.7)

$$G^2(\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]) = G^2(\mathcal{M}[\mathcal{T}_1]) + G^2(\mathcal{M}[\mathcal{T}_2]) \quad . \tag{3.2.3}$$

Tak więc statystyka służąca do testowania zgodności $\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]$ jest sumą statystyk służących do testowania $\mathcal{M}[\mathcal{T}_1]$ i $\mathcal{M}[\mathcal{T}_2]$. Przy tym, $G^2(\mathcal{M}[\mathcal{T}_1])$ i $G^2(\mathcal{M}[\mathcal{T}_2])$ są niezależne, jeżeli $\mathcal{T}_1$ i $\mathcal{T}_2$ należą do różnych "warstw" tablicy $\mathcal{T}$ wyznaczonych przez pewien czynnik A_s, $s \in \mathcal{F}_1$; w przeciwnym wypadku są one asymptotycznie niezależne (Goodman, 1968; Gillula, 1985). Wzór (3.2.3) sugeruje możliwość łatwego uzupełnienia testów zgodności modeli $\mathcal{M}[\mathcal{T}_1]$ i $\mathcal{M}[\mathcal{T}_2]$ testem zgodności modelu $\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]$ przy użyciu statystyki opartej na v_1+v_2 stopniach swobody.

Kontynuując rozpatrywanie tablic rozłącznych, zauważmy iż $\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]$ jest również modelem implikowanym przez $\mathcal{M}[\mathcal{T}]$. Stosując (3.1.11) otrzymujemy

$$G^2(\mathcal{M}[\mathcal{T}]) = G^2(\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]) + G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]) \quad ,$$

oraz uwzględniając (3.2.3)

$$G^2(\mathcal{M}[\mathcal{T}]) = G^2(\mathcal{M}[\mathcal{T}_1]) + G^2(\mathcal{M}[\mathcal{T}_2]) + G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1,\mathcal{T}_2]) \quad . \tag{3.2.4}$$

Interpretacja takiej dekompozycji jest analogiczna do interpretacji dekompozycji opisanej wzorem (3.1.11).

Podane fakty można uogólnić na sytuację, w której R wyróżnionych tablic spełnia warunek $\mathcal{T}_r \cap \mathcal{T}_{r'} = \emptyset$, $r,r'= 1,\dots,R$, $r\neq r'$. Oznaczając symbolem $\mathcal{M}[\mathcal{T}_1,\dots,\mathcal{T}_R]$ model powstający przez koniunkcję warunków nakładanych przez modele $\mathcal{M}[\mathcal{T}_r]$, $r=1,\dots,R$, dochodzimy zgodnie z (3.2.3) i (3.2.4) do wzoru

$$G^2(\mathcal{M}[\mathcal{T}]) = G^2(\mathcal{M}[\mathcal{T}_1,\dots,\mathcal{T}_R]) + G^2(\mathcal{M}[\mathcal{T}]|\mathcal{M}[\mathcal{T}_1,\dots,\mathcal{T}_R]), \tag{3.2.5}$$

gdzie

$$G^2(\mathcal{M}[\mathcal{T}_1,\dots,\mathcal{T}_R]) = \sum_{r=1}^{R} G^2(\mathcal{M}[\mathcal{T}_r]) \quad . \tag{3.2.6}$$

Przejdźmy do podania procedur odpowiednich dla jednoczesnego testowania zgodności rodziny modeli implikowanych $\{\mathcal{M}[\mathcal{T}_r];\ r=1,\dots,R\}$. Rodzinę tę, uzupełnioną modelem $\mathcal{M}[\mathcal{T}_{R+1}] \equiv \mathcal{M}[\mathcal{T}]$, oznaczymy symbolem Ψ. Ograniczymy się przy tym do sytuacji, w której koniunkcja warunków nakładanych na wektor $\mathbf{m}$ przez wszystkie modele $\mathcal{M}[\mathcal{T}_r]$, $r=1,\dots,R$, jest równoważna modelowi $\mathcal{M}[\mathcal{T}]$, co zanotujemy jako

$$\mathcal{M}[\mathcal{T}] \Longleftrightarrow \mathcal{M}[\mathcal{T}_1,\dots,\mathcal{T}_R] \quad . \tag{3.2.7}$$

Jakkolwiek nie jest to warunek konieczny zastosowania procedur testowania jednoczesnego, to jednak w praktyce wybór Ψ winien być zgodny z (3.2.7). Wynika to z faktu, iż celem analizy modeli implikowanych jest wyjaśnienie nieadekwatności modelu $\mathcal{M}[\mathcal{T}]$. Poniżej podajemy dwa jednoczesne testy zgodności rodziny modeli Ψ otrzymane przez uogólnienie wyników Gabriela (1966) oraz Goodmana (1971), przy wykorzystaniu ogólnej teorii testów jednoczesnych rozpatrywanej m.in. przez Gabriela (1969) i Millera (1981) (patrz również Caliński i in., 1979).

Pierwszy z testów, oznaczany dalej symbolem STP1, jest oparty na statystyce ilorazowej G^2 określonej przez (3.1.7). Przy ustalonym poziomie istotności α, zaleca on uznać dowolny model $\mathcal{M}[\mathcal{T}_r] \in \Psi$ za nieadekwatny wtedy, gdy

$$G^2(\mathcal{M}[\mathcal{T}_r]) > \chi^2_\alpha(v) \tag{3.2.8}$$

(por. warunek (2.10)). Test ten jest (zgodnie z terminologią Gabriela, 1969) konsekwentny, co oznacza że adekwatność modelu $\mathcal{M}[\mathcal{T}]$ pociąga adekwatność wszystkich modeli implikowanych (lub że nieadekwatność któregokolwiek z modeli implikowanych pociąga nieadekwatność modelu $\mathcal{M}[\mathcal{T}]$). Własność ta wynika bezpośrednio z monotoniczności statystyki G^2. Prawdopodobieństwo stwierdzenia nieadekwatności co najmniej jednego prawdziwego modelu $\mathcal{M}[\mathcal{T}_r]$ wynosi α, gdy zgodny jest model $\mathcal{M}[\mathcal{T}]$ (por. twierdzenie 5.1, Caliński i in., 1979). Natomiast prawdopodobieństwo α_r stwierdzenia nieadekwatności dowolnego prawdziwego modelu $\mathcal{M}[\mathcal{T}_r]$ można wyznaczyć ze wzoru

$$\alpha_r = P\{\chi^2(v_r) > \chi^2_\alpha(v)\} \quad , \tag{3.2.9}$$

gdzie $\chi^2(v_r)$ jest zmienną losową χ^2 o liczbie stopni swobody v_r odpowiadającej modelowi $\mathcal{M}[\mathcal{T}_r]$. Z własności rozkładu χ^2 oraz z tego, że dla dowolnego $\mathcal{M}[\mathcal{T}_r] \in \Psi$ zachodzi nierówność $v_r \le v$, wynika iż $\alpha_r \le \alpha$.

Drugi jednoczesny test zgodności rodziny modeli Ψ jest oparty na tzw. zasadzie Roya unii i przekroju i będzie oznaczany symbolem STP2. Dla jego określenia niezbędne jest wprowadzenie pojęcia tzw. modelu minimalnego rodziny Ψ. Modelem minimalnym nazywamy taki model $\mathcal{M}[\mathcal{T}_r] \in \Psi$, który nie implikuje żadnych innych modeli z rodziny Ψ. Oznaczmy rodzinę modeli minimalnych symbolem Ψ_{min}. Zgodnie z testem STP2 dowolny model $\mathcal{M}[\mathcal{T}_r] \in \Psi$ uznamy za nieadekwatny, gdy

$$G^2_{max}(\mathcal{M}[\mathcal{T}_r]) > \xi \quad , \tag{3.2.10}$$

gdzie

$$G^2_{max}(\mathcal{M}[\mathcal{T}_r]) = \max_{\mathcal{T}' \subset \mathcal{T}_r} (G^2(\mathcal{M}[\mathcal{T}']); \mathcal{M}[\mathcal{T}'] \in \Psi_{min}) \ . \tag{3.2.11}$$

(zauważmy, że dla $\mathcal{M}[\mathcal{T}_r] \in \Psi_{min}$, $G^2_{max}(\mathcal{M}[\mathcal{T}_r]) = G^2(\mathcal{M}[\mathcal{T}_r])$). Stała ξ jest wyznaczona tak, aby spełniona była dla przyjętego poziomu istotności α i przy prawdziwości modelu $\mathcal{M}[\mathcal{T}]$ równość

$$P\{G^2_{max}(\mathcal{M}[\mathcal{T}]) > \xi\} = \alpha \tag{3.2.12}$$

($G^2_{max}(\mathcal{M}[\mathcal{T}])$ oznacza tutaj zmienną losową).

Nietrudno wykazać, iż procedura STP2 jest konsekwentna oraz, dodatkowo, harmonijna (wg terminologii Gabriela, 1969), co oznacza iż nieadekwatność dowolnego modelu rodziny Ψ (w szczególności - modelu $\mathcal{M}[\mathcal{T}]$) pociąga nieadekwatność przynajmniej jednego z modeli implikowanych przez ten model. Wadą procedury jest fakt, iż w ogólnym wypadku stała ξ nie może być wyznaczona w sposób dokładny, gdyż wymagałoby to znajomości dokładnego rozkładu zmiennej losowej występującej we wzorze (3.2.12). Standardowym sposobem postępowania w takich sytuacjach jest zastąpienie ξ pewną wartością $\tilde{\xi} > \xi$, taką dla której prawdopodobieństwo wymienione w (3.2.12) jest nie większe od α. Do wyznaczenia $\tilde{\xi}$ można zastosować nierówność Bonferroniego (Miller, 1981, str.8; Alt, 1982) w sposób analogiczny do tego, w jaki została ona zastosowana do konstrukcji testów jednoczesnych w jednozmiennej (Gabriel, 1969) lub wielozmiennej (Caliński i in., 1979) analizie wariancji. Zgodnie z (3.2.11) oraz wspomnianą nierównością mamy bowiem przy prawdziwości modelu $\mathcal{M}[\mathcal{T}]$

$$P\{G^2_{max}(\mathcal{M}[\mathcal{T}]) > \xi\} = P\{\max_{\mathcal{T}'}(G^2(\mathcal{M}[\mathcal{T}']);\ \mathcal{M}[\mathcal{T}'] \in \Psi_{min}) > \xi\}$$

$$= 1 - P\{\bigcap_{\mathcal{M}[\mathcal{T}']\in\Psi_{min}} (G^2(\mathcal{M}[\mathcal{T}']) < \xi)\}$$

$$\le \sum_{\mathcal{M}[\mathcal{T}']\in\Psi_{min}} P\{G^2(\mathcal{M}[\mathcal{T}']) > \xi\}.$$

Jeżeli dalej założymy, że liczba stopni swobody związana z każdym z modeli minimalnych jest taka sama i wynosi v_{min}, to otrzymujemy

$$P\{G^2_{max}(\mathcal{M}[\mathcal{T}]) > \xi\} \le \sum_{\mathcal{M}[\mathcal{T}']\in\Psi_{min}} P\{\chi^2(v_{min}) > \xi\}$$

$$= g\ P\{\chi^2(v_{min}) > \xi\},$$

gdzie g oznacza liczbę modeli minimalnych rodziny Ψ. Jedną z możliwości wyboru ξ jest przyjęcie $\tilde{\xi} = \chi^2_{\alpha/g}(v_{min})$. Wtedy

$$P\{G^2_{max}(\mathcal{M}[\mathcal{T}]) > \tilde{\xi}\} \le g\,\frac{\alpha}{g} = \alpha \quad ,$$

a więc żądana nierówność jest spełniona.

Należy zaznaczyć, iż wartość krytyczna ξ procedury STP2 może być wyznaczona dokładnie wtedy, gdy statystyki $G^2(\mathcal{M}[\mathcal{T}'])$, $\mathcal{M}[\mathcal{T}'] \in \Psi_{min}$, są niezależnymi zmiennymi losowymi (por. komentarz do wzoru 3.2.3). Przy wprowadzonych powyżej założeniach otrzymujemy wówczas

$$P\{G^2_{max}(\mathcal{M}[\mathcal{T}]) > \xi\} = 1 - \prod_{\mathcal{M}[\mathcal{T}']\in\Psi_{min}} P\{G^2(\mathcal{M}[\mathcal{T}']) < \xi\}$$

$$= 1 - [1 - P\{\chi^2(v_{min}) > \xi\}]^g \quad .$$

Tak więc warunek (3.2.12) jest spełniony przez wartość ξ wyznaczoną tak, aby

$$P\{\chi^2(v_{min}) > \xi\} = 1-(1-\alpha)^{1/g} \quad . \tag{3.2.13}$$

Wartość krytyczna wyznaczona zgodnie z tym warunkiem jest oczywiście mniejsza od wartości oznaczonej uprzednio przez $\tilde{\xi}$. Stosując metody analityczne można jednak wykazać iż dla małych wartości α i dużych wartości g

$$1 - (1-\alpha)^{1/g} \approx \frac{\alpha}{g} \quad .$$

Tak więc, w sytuacji niezależności statystyk odpowiadających modelom minimalnym, dobrym wyborem wartości krytycznej dla procedury STP2 jest wybór $\tilde{\xi} = \chi^2_{\alpha/g}(v_{min})$. Należy mieć przy tym świadomość, iż wybór taki odpowiada pewnemu przybliżeniu, a nie oszacowaniu z góry (jak to miało miejsce w sytuacji ogólnej), dokładnej wartości krytycznej spełniającej warunek (3.2.12).

Podajmy teraz pewne uwagi dotyczące porównania opisanych procedur testowania jednoczesnego STP1 i STP2. Procedura STP2, będąc procedurą harmonijną, zapewnia iż wnioskowanie przeprowadzone za jej pomocą prowadzi zawsze do znalezienia chociaż jednej podtablicy odpowiedzialnej za stwierdzenie nieadekwatności modelu $\mathcal{M}[\mathcal{T}]$. Zgodnie z ogólną teorią testów jednoczesnych, STP2 jest procedurą lepszą niż STP1, gdyż jest bardziej rozkładająca lub, inaczej, mocniejsza względem modeli minimalnych rodziny Ψ (por. twierdzenie 6.4, Caliński i in., 1979). Z drugiej strony, zastosowanie w STP2 oszacowania górnego $\tilde{\xi}$ wartości krytycznej ξ powoduje iż jest to procedura konserwatywna.

Aby prowadzić dalsze porównania STP1 i STP2 zastanówmy się jak w praktyce powinien być dokonywany wybór modeli tworzących rodzinę Ψ. Zauważmy, iż warunek (3.2.7) jest spełniony w szczególności przez rodzinę wszystkich modeli implikowanych przez $\mathcal{M}[\mathcal{T}]$. Jednoczesne badanie zgodności modelu $\mathcal{M}$ dla wszystkich podtablic tablicy $\mathcal{T}$ nie jest jednak korzystne. Pomijając nawet kwestię ilości niezbędnych obliczeń, postępowanie takie nie jest konieczne ze względu na konsekwentność używanych procedur testowych. W większości sytuacji praktycznych okazuje się, że wyjaśnienie nieadekwatności modelu $\mathcal{M}$ można przeprowadzić wybierając rodzinę Ψ w ten sposób, aby składała się ona tylko z modeli minimalnych o takiej samej liczbie stopni swobody.

Ważnym przypadkiem szczególnym jest tutaj sytuacja, w której wyjaśnianie nieadekwatności modelu $\mathcal{M}[\mathcal{T}]$ przeprowadza się ze względu na jedną, wybraną zmienną, powiedzmy A_q. Wybór rodziny modeli minimalnych o takiej samej liczbie stopni swobody jest wtedy równoważny wyborowi rodziny podtablic tablicy $\mathcal{T}$ odpowiadających podzbiorom kategorii A_q o równej liczebności l $(2 \le l \le I_q-1)$. Liczba wszystkich modeli tego typu jest równa

$$R = \binom{I_q}{l}.$$

Dobór liczby l, nazywanej dalej rzędem modeli implikowanych, zależy od konkretnej sytuacji eksperymentalnej; wartość $l = 2$ odpowiada wyborowi par kategorii, wartość $l = 3$ - wyborowi trójek kategorii, itd. Liczbę modeli implikowanych można dodatkowo ograniczyć jeżeli wśród kategorii zmiennej A_q istnieje kategoria "kontrolna", to znaczy taka, która winna wystąpić w każdej z rozpatrywanych podtablic. Wtedy liczba wszystkich modeli rozpatrywanego typu wynosi

$$R = \binom{I_q-1}{l-1}.$$

Zagadnienie wyboru wartości l w odniesieniu do pewnej klasy modeli nienasyconych jest również poruszone w paragrafie 3.3. Przykłady praktycznych podejść do problemu pokazane są w paragrafie 4.

W związku z przeprowadzonymi porównaniami obu procedur zauważmy iż wartość krytyczna $\tilde{\xi}$ dla STP2 zależy od wyboru modeli minimalnych rodziny Ψ. Tak więc STP2, a nie STP1, powinna być stosowana, gdy zainteresowanie ogranicza się do niektórych modeli implikowanych. Zastosowanie STP1 jest wówczas uzasadnione tylko wtedy, gdy wybór modeli minimalnych nie zezwala na wyznaczenie wartości krytycznej procedury STP2 (nie są to modele o takiej samej liczbie stopni swobody). Poza tym, biorąc pod uwagę liczbę niezbędnych operacji, zastosowanie procedury STP1 wydaje się korzystniejsze wtedy, gdy jakkolwiek badamy zgodność niewielu modeli wybranych, to jednak "potencjalnie" przedmiotem zainteresowania jest bardzo liczna rodzina Ψ modeli implikowanych przez $\mathcal{M}[\mathcal{T}]$. W takiej sytuacji STP1 przewiduje wyznaczanie wartości G^2 tylko dla modeli wybranych. Procedura STP2 wymaga natomiast znajomości wartości G^2 dla wszystkich modeli minimalnych rodziny Ψ. Prowadzić to może do znacznej ilości obliczeń.

3.3. Modele rozkładalne

W dotychczasowych rozważaniach tego rozdziału obowiązywało założenie, iż każdy ze zbiorów $\mathcal{K}_s$ kategorii generujących podtablicę $\mathcal{T}_1 \subset \mathcal{T}$ jest przynajmniej dwuelementowy. Zapewnia ono, iż tablica $\mathcal{T}_1$ jest rzeczywiście tablicą S-wymiarową. Interpretacja modelu $\mathcal{M}[\mathcal{T}_1]$ jest wtedy dokładnie taka, jak interpretacja modelu ogólnego $\mathcal{M}[\mathcal{T}]$. Okazuje się jednak, iż dla pewnych modeli logarytmiczno-liniowych tablicy $\mathcal{T}$ szczególnie interesujące jest rozpatrywanie modeli implikowanych podtablic o mniejszej liczbie wymiarów. Problemem tym zajmiemy się w niniejszym paragrafie.

Załóżmy, że podtablica $\mathcal{T}_1$ tablicy $\mathcal{T}$ wybrana jest w ten sposób, iż $K_s=1$ dla $s \in \gamma \subset \mathcal{P}$. Z uwagi na istniejącą zawsze możliwość formalnej zmiany numeracji zmiennych przyjmiemy, że $\gamma = \{1,\dots,P\}$, P<S. Tablica $\mathcal{T}_1$ składa się więc z podklas odpowiadających pewnej kombinacji kategorii pierwszych P zmiennych. Macierz $\mathbf{W}$, wybierająca z wektora $\mathbf{m}$ liczebności oczekiwane podklas tablicy $\mathcal{T}_1$, może być wyznaczona zgodnie z ogólnym wzorem (3.1.1), przy czym teraz

$$\mathbf{V}_s = \mathbf{e}'_{I_s,i_s} \quad , \qquad s\in\gamma \quad ,$$

gdzie $\mathbf{e}_{I_s,i_s}$ oznacza wektor jednostkowy wymiaru I_s posiadający jedynkę na i_s-tym miejscu (i_s oznacza numer wybranej kategorii zmiennej A_s). Dla dowolnego modelu $\mathcal{M}$ postaci (2.3), model implikowany tablicy $\mathcal{T}_1$ można zapisać w formie (3.1.3). Jednak, ponieważ $K_s=1$ dla $s\in\gamma$, macierze występujące w (3.1.3) są postaci

$$\mathbf{Z}_\alpha = \bigotimes_{s\in\mathcal{P}\setminus\gamma} \mathbf{B}_s \quad ,$$

gdzie

$$\mathbf{B}_s = \begin{cases} \mathbf{I}_{K_s}, & s\in\alpha, \\ \mathbf{1}_{K_s}, & s\in\mathcal{P}\setminus\alpha, \end{cases}$$

co w oznaczeniach paragrafu 2 można zapisać jako

$$\mathbf{Z}_\alpha = \mathbf{Z}_{\alpha\setminus\gamma}^{\mathcal{P}\setminus\gamma} \quad .$$

Tak więc model implikowany $\mathcal{M}[\mathcal{T}\ ]$ jest postaci

$$\mathrm{Log}\ \mathbf{Wm} = \sum_{\alpha\in\mathcal{G}(\mathcal{M})} \widehat{Z^{\mathcal{T}\setminus\gamma}_{\alpha\setminus\gamma}}\, w_{\alpha\setminus\gamma} \ . \tag{3.3.1}$$

Zauważmy, iż model (3.3.1) może być w odniesieniu do $\mathcal{T}_1$ modelem nasyconym. Jest tak wtedy, gdy dla pewnego $\alpha_o \in \mathcal{G}(\mathcal{M})$ zachodzi równość

$$\widehat{Z^{\mathcal{T}\setminus\gamma}_{\alpha_o\setminus\gamma}} = I_{T(\mathcal{T}\setminus\gamma)} ,$$

a więc wtedy, gdy $\alpha_o\setminus\gamma = \mathcal{T}\setminus\gamma$. Jeżeli na przykład $S=3$ i $\gamma = \{1\}$, to sytuacja taka ma miejsce dla modelu o klasie generującej $\mathcal{G}(\mathcal{M}) = \{13,23\}$. Co więcej, różne modele ogólne mogą implikować takie same modele $\mathcal{M}[\mathcal{T}_1]$. Prostym przykładem tego, znowu przy $S=3$ i $\gamma = \{1\}$, są modele $\mathcal{M}_1$ i $\mathcal{M}_2$ o klasach generujących $\mathcal{G}(\mathcal{M}_1) = \{12,13\}$ i $\mathcal{G}(\mathcal{M}_2) = \{2,13\}$. W obu wypadkach, w zapisie modelu implikowanego (3.3.1) wystąpią macierze Z_2^{23} i Z_3^{23}.

Rozważania powyższe wskazują, iż rozpatrywanie modelu implikowanego $\mathcal{M}[\mathcal{T}_1]$ w sytuacji, gdy $\mathcal{T}_1$ nie jest tablicą S-wymiarową, ma sens praktyczny tylko dla niektórych modeli $\mathcal{M}$. Formalnie jednak dla dowolnego modelu $\mathcal{M}$ rodzinę modeli implikowanych uzupełnić można o wszystkie modele tego typu. Jeżeli bowiem $\mathcal{T}_1$ jest tablicą taką, że $\mathcal{M}[\mathcal{T}_1]$ jest modelem nasyconym, to implikacja (3.1.4) spełniona jest w sposób oczywisty.

Załóżmy teraz, iż, jak poprzednio, $K_s=1$ dla $s\in\gamma$ oraz, dodatkowo, że $K_s = I_s$ dla $s \in \mathcal{T}\setminus\gamma$. Tak wybrana podtablica dotyczy jednej kombinacji kategorii zmiennych A_s, $s\in\gamma$, oraz wszystkich kategorii pozostałych zmiennych. Oznaczmy tę podtablicę symbolem $\mathcal{T}_{i_1\ldots i_P}$, gdzie i_s, $s\in\gamma$, jest numerem interesującej kategorii zmiennej A_s, a więc $i_s = 1,\ldots,I_s$. Liczba różnych podtablic $\mathcal{T}_{i_1\ldots i_P}$ wynosi $T(\gamma)$, przy czym są to tablice rozłączne. Przy wprowadzonych oznaczeniach, zanotujmy następującą definicję.

Definicja 3.3.1. Model $\mathcal{M}$ tablicy S-wymiarowej $\mathcal{T}$ nazywamy γ-rozkładalnym, jeżeli

$$\mathcal{M}[\mathcal{T}] \Longleftrightarrow \mathcal{M}[\mathcal{T}_{i_1\ldots i_P};\ i_s=1,\ldots,I_s,\ s=1,\ldots,P] \ .$$

Nietrudno znaleźć uzasadnienie praktyczne tej definicji. Ponieważ rozpatrywane tablice $\mathcal{T}_{i_1\ldots i_P}$ są rozłączne, to, zgodnie z (3.2.3), dla modelu γ-rozkładalnego prawdziwa jest równość

$$G^2(\mathcal{M}[\mathcal{T}]) = \sum_{i_1=1}^{I_1}\ldots\sum_{i_P=1}^{I_P} G^2(\mathcal{M}[\mathcal{T}_{i_1\ldots i_P}]). \tag{3.3.2}$$

Statystykę G^2 dla modelu ogólnego $\mathcal{M}[\mathcal{T}]$ można więc rozłożyć na $T(\gamma)$ (asymptotycznie) niezależnych składników odpowiednich dla testowania zgodności $\mathcal{M}$ w pewnych podtablicach generowanych przez zmienne A_s, $s\in\mathcal{T}\setminus\gamma$. Definicja 3.3.1 wskazuje również, iż w wypadku odrzucenia adekwatności $\mathcal{M}[\mathcal{T}]$ podane w paragrafie 3.2 procedury testowania jednoczesnego korzystnie jest zastosować w stosunku do $T(\gamma)$-elementowej rodziny modeli

$$\Psi = \{\mathcal{M}[\mathcal{T}_{i_1 \ldots i_P}];\ i_s=1,\ldots,I_s,\ s=1,\ldots,P\}.$$

Rodzina ta składa się tylko z modeli minimalnych o takiej samej liczbie stopni swobody, a więc właściwą procedurą jest tutaj STP2. Jeżeli nieadekwatność $\mathcal{M}[\mathcal{T}]$ wyjaśniana jest ze względu na jedną zmienną A_q (przypadek szczególny dyskutowany już w paragrafie 3.2.), to Ψ jest rodziną modeli implikowanych rzędu $l = 1$.

Podamy obecnie twierdzenie identyfikujące modele rozkładalne.

Twierdzenie 3.3.1. Model $\mathcal{M}$ o klasie generującej $\mathcal{G}(\mathcal{M})$ jest γ-rozkładalny wtedy i tylko wtedy gdy $\gamma\subset\alpha$ dla każdego $\alpha\in\mathcal{G}(\mathcal{M})$.

Dowód. Oznaczmy przez $W_{i_1 \ldots i_P}$ macierz wybierającą z wektora $\mathbf{m}$ liczebności oczekiwane podklas należących do tablicy $\mathcal{T}_{i_1 \ldots i_P}$, $i_s = 1,\ldots,I_s$. Ponieważ dla każdej tablicy $\mathcal{T}_{i_1 \ldots i_P}$ mamy $J_s = I_s$ dla $s\in\mathcal{P}\setminus\gamma$, to dla dowolnego modelu $\mathcal{M}$ model implikowany $\mathcal{M}[\mathcal{T}_{i_1 \ldots i_P}]$ można, zgodnie z (3.3.1), zapisać w postaci

$$\operatorname{Log} W_{i_1 \ldots i_P}\mathbf{m} = \sum_{\alpha\in\mathcal{G}(\mathcal{M})} X^{\mathcal{P}\setminus\gamma}_{\alpha\setminus\gamma}\, w^{(i_1 \ldots i_P)}_{\alpha\setminus\gamma}, \tag{3.3.3}$$

gdzie

$$X^{\mathcal{P}\setminus\gamma}_{\alpha\setminus\gamma} = \bigotimes_{s\in\mathcal{P}\setminus\gamma} B_s$$

i gdzie

$$B_s = \begin{cases} I_{I_s}, & s\in\alpha,\\ \mathbf{1}_{I_s}, & s\in\mathcal{P}\setminus\alpha.\end{cases}$$

Z drugiej strony, jeżeli $\mathcal{M}'$ jest pewnym zbiorem zawartym w $\Omega(\mathcal{P}\setminus\gamma)$, i założymy, iż dla każdej kombinacji $(i_1 \ldots i_P)$, $i_s = 1,\ldots,I_s$, prawdziwy jest warunek

$$\operatorname{Log} W_{i_1 \ldots i_P}\mathbf{m} = \sum_{\beta\in\mathcal{M}'} X^{S\setminus\gamma}_{\beta}\, w^{(i_1 \ldots i_P)}_{\beta},$$

to można pokazać, że dla pewnych wektorów $\tilde{u}_\delta$

$$\operatorname{Log}\mathbf{m} = \sum_{\delta\in\mathcal{G}'} X_\delta\, \tilde{u}_\delta, \tag{3.3.4}$$

gdzie $\mathcal{G}'$ jest klasą generującą zbioru $\{\delta : \delta = \beta\cup\gamma,\ \beta\in\mathcal{M}'\}$. Jeżeli przyjęty dla $\mathcal{T}_{i_1 \ldots i_P}$ warunek jest postaci (3.3.3), to występujący w (3.3.4) zbiór $\mathcal{G}'$ ma postać $\{\delta : \delta = (\alpha\setminus\gamma)\cup\gamma,\ \alpha\in\mathcal{G}(\mathcal{M})\}$. Jest on równy zbiorowi $\mathcal{G}(\mathcal{M})$ wtedy i tylko wtedy, gdy $(\alpha\setminus\gamma)\cup\gamma = \alpha$ dla $\alpha\in\mathcal{G}(\mathcal{M})$, lub, równoważnie, gdy $\gamma\subset\alpha$ dla $\alpha\in\mathcal{G}(\mathcal{M})$. □

Podajmy teraz pewne przykłady modeli rozkładalnych. Przy S=3 modelem {1}-rozkładalnym jest model o klasie generującej {12,13}. Jest on prawdziwy wtedy i tylko wtedy, gdy dla każdej z I_1 dwuwymiarowych tablic generowanych przez A_2 i A_3 prawdziwy jest model o klasie generującej {2,3}. Przy S=4, rozważmy model o klasie generującej {123,124}. Zgodnie z twierdzeniem 3.3.1 model ten jest {1}-, {2}- oraz {12}-rozkładalny. Jego prawdziwość jest więc równoważna prawdziwości modelu {23,24} dla każdej z I_1 tablic trójwymiarowych generowanych przez A_2, A_3, A_4, prawdziwości modelu {13,14} dla każdej z I_2 tablic generowanych przez A_1, A_3 i A_4, oraz prawdziwości modelu {3,4} dla każdej z I_1I_2 tablic dwuwymiarowych generowanych przez A_3 i A_4. Sposób zastosowania powyższych własności modelu w trakcie analizy zależy od określenia zmiennych A_1, A_2, A_3 i A_4 i ich roli w eksperymencie.

Kończąc ten paragraf zanotujmy fakt, iż modele nazwane tutaj rozkładalnymi były rozważane przez Bishop i in. (1975, str.105), jednak tylko z punktu widzenia wyznaczania ocen liczebności oczekiwanych. W świetle przedstawionych dotychczas rozważań jest oczywiste, iż jeżeli model $\mathcal{M}$ jest γ-rozkładalny, to oceny liczebności oczekiwanych podklas tablicy $\mathcal{T}$ przy prawdziwości $\mathcal{M}$ można wyznaczyć dla każdej z podtablic $\mathcal{T}_{i_1 \ldots i_p}$ niezależnie, zakładając prawdziwość modelu $\mathcal{M}[\mathcal{T}_{i_1 \ldots i_p}]$.

4. PRZYKŁAD ZASTOSOWANIA

W paragrafie tym dokonamy analizy wyników doświadczenia przeprowadzonego w Pracowni Genetyki Odporności Instytutu Genetyki Roślin PAN w celu oceny odporności odmian rzepaku na chorobę powodowaną przez grzyb Sclerotinia Sclerotiorum (Lib.) de Bary (dane niepublikowane). W doświadczeniu brało udział pięć odmian. Obserwacje były dokonywane w trzech terminach, natomiast stopień porażenia roślin był mierzony w skali pięciostopniowej. Wyniki doświadczenia, przedstawione w Tabeli 1, stanowią trójwymiarową tablicę kontyngencji. Generujące tę tablicę zmienne to: A_1 - odmiana (5 kategorii), A_2 - termin (3 kategorie) i A_3 - stopień porażenia roślin (5 kategorii). Zmienne A_1 i A_2 są czynnikami, natomiast A_3 jest zmienną losową.

Analizę statystyczną rozpatrywanej tablicy rozpocznijmy od testowania zgodności (na poziomie 0.05) modelu $\mathcal{M}$ o klasie generującej {12,13,23}, który w tym wypadku opisuje brak interakcji czynników A_1 i A_2 pod względem wpływu na rozkład zmiennej A_3. Liczba stopni swobody dla tego modelu wynosi 32. Obliczenia wskazują, że $G^2(\mathcal{M}) = 53.67$, a więc model $\mathcal{M}$ nie jest adekwatny. Do wyjaśnienia tej nieadekwatności zmierza dalsza analiza polegająca na testowaniu zgodności modeli implikowanych przez $\mathcal{M}$.

Tok postępowania przy wyjaśnianiu nieadekwatności modelu nienasyconego zależy od interpretacji tego modelu w konkretnej sytuacji eksperymentalnej. W naszym wypadku model $\mathcal{M}$, opisujący brak interakcji $A_1 \times A_2$, korzystnie jest interpretować na dwa sposoby. Po pierwsze, jako model stwierdzający, że wpływ terminu (A_2) na stopień porażenia roślin

Tab. 1. Liczebności roślin pięciu odmian rzepaku obserwowanych w trzech terminach i sklasyfikowanych ze względu na stopień porażenia przez Scerotinia Sclerotiorum

Stopień porażenia roślin (%)		Odmiana														
		Primor(P)			Quinta(Q)			Jampol(J)			Jet-Neuf(N)			Górczański(G)		
		Termin			Termin			Termin			Termin			Termin		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
0	(I)	7	0	0	0	1	0	0	0	0	8	11	4	1	0	0
1-5	(II)	16	9	8	23	6	1	15	0	0	31	20	19	26	22	17
6-25	(III)	16	4	1	14	13	1	22	11	1	1	7	6	12	5	6
26-75	(IV)	1	0	0	3	0	13	2	0	4	0	0	6	1	0	5
76-100	(V)	0	27	31	0	20	25	1	29	35	0	2	5	0	13	12

(opisany przez zmienną losową A_3) jest taki sam dla każdej odmiany (A_1). Po drugie, jako model opisujący sytuację taką, w której niejednorodność odmian pod względem stopnia porażenia roślin jest taka sama we wszystkich terminach. Poniżej przeprowadzamy analizę pewnych modeli implikowanych przez $\mathcal{M}$, bazującą na takich dwóch interpretacjach.

Przyjmując pierwsze z wymienionych znaczeń modelu $\mathcal{M}$, zastosujemy do wyjaśnienia nieadekwatności tego modelu procedury testowania jednoczesnego opisane w paragrafie 3.2. Spróbujmy za ich pomocą znaleźć taką odmianę, dla której wpływ terminu na stopień porażenia roślin jest różny od tego wpływu dla pozostałych odmian. Można to uczynić badając zgodność $\mathcal{M}$ we wszystkich podtablicach tablicy oryginalnej odpowiadających czwórkom odmian, to znaczy przeprowadzając testowanie zgodności rodziny wszystkich modeli implikowanych rzędu 4 wybranych ze względu na zmienną A_1. Wyniki obliczeń podaje Tabela 2. W ich świetle należałoby formalnie stwierdzić

Tab.2. Wyniki testowania zgodności modelu opisującego brak interakcji $A_1 \times A_2$ dla wszystkich podtablic odpowiadających czwórkom odmian

Odmiany	Odmiana pominięta	Wartość statystyki G^2
Q,J,N,G	P	32.45
P,J,N,G	Q	43.27
P,Q,N,G	J	41.91
P,Q,J,G	N	41.84
P,Q,J,N	G	40.73
Wartość krytyczna dla testowania jednoczesnego procedurą STP2 na poziomie 0.05		42.98

nieadekwatność modelu $\mathcal{M}$ dla podtablicy odpowiadającej odmianom P, J, N i G. Nie może to jednak być wniosek zbyt kategoryczny, gdyż wartość G^2 dla tej podtablicy jest bliska wartości krytycznej. Dodatkowo, zbliżone wartości G^2 otrzymano także dla trzech innych podtablic wymienionych w Tabeli 2. Bardziej interesujący wydaje się fakt, iż wartość G^2 dla podtablicy związanej z odmianami Q, J, N i G jest znacznie mniejsza od wartości krytycznej. Należy zatem przypuszczać, iż odmianą, dla której wpływ terminu na stopień porażenia roślin jest różny od pozostałych, jest odmiana P. W odniesieniu do modelu ogólnego oznacza to, iż odmiana P okazała się najbardziej odpowiedzialna za istnienie interakcji $A \times A$ w całej tablicy kontyngencji.

Zastosowany powyżej wybór rodziny modeli implikowanych nie jest jedynym możliwym wyborem. Do wyjaśnienia przyczyn nieadekwatności modelu $\mathcal{M}$ możemy także dążyć poprzez testowanie zgodności tego modelu dla podtablic dotyczących wszystkich par odmian (modele rzędu 2). Postępowanie takie odpowiada porównywaniu odmian (każdej odmiany z każdą) pod względem wpływu terminu na stopień porażenia roślin. Analizując wyniki obliczeń przedstawione w Tabeli 3 stwierdzamy znaczne podobieństwo odmian P i G oraz odmian Q, J i N.

Tab. 3. Wartość statystyki G^2 dla porównań odmian pod względem wpływu terminu na rozkład zmiennej A_3 opisującej stopień porażenia roślin

	Odmiana P	Odmiana Q	Odmiana J	Odmiana N
Q	20.69			
J	18.56	3.91		
N	23.59	5.19	2.25	
G	5.08	17.21	17.43	12.56
Wartość krytyczna dla testowania jednoczesnego procedurą STP2 na poziomie 0.05				23.11

Przejdźmy z kolei do próby wyjaśnienia nieadekwatności modelu $\mathcal{M}$ korzystając z drugiej interpretacji tego modelu. Będziemy zmierzać do wskazania takiego terminu, w którym niejednorodność odmian ma charakter inny niż w pozostałych terminach, lub inaczej, do znalezienia terminów podobnych pod względem niejednorodności odmian. We wnioskowaniu nie skorzystamy z procedur testowania jednoczesnego. Pokażemy natomiast sposób praktycznego wykorzystania dekompozycji statystyki $G^2(\mathcal{M})$ określonej wzorem (3.2.11). Uzasadnione jest przy tym ograniczenie się do rozpatrzenia dwóch podtablic oryginalnej tablicy $\mathcal{T}$: podtablicy $\mathcal{T}_1$ odpowiadającej terminom 1 i 2 oraz podtablicy $\mathcal{T}_2$ odpowiadającej terminom 2 i 3. Rezultaty odpowiednich obliczeń dogodnie jest przedstawić w sposób wskazany w Tabeli 4. Pokazuje ona, iż przejście od opisu badanej tablicy kontyngencji za pomocą modelu $\mathcal{M}[\mathcal{T}]$ (postulującego jednakowe zróżnicowanie odmian we wszystkich terminach) do opisu tej tablicy za pomocą modelu $\mathcal{M}[\mathcal{T}_1]$ (postulującego jednakowe zróżnicowanie odmian w terminach 1 i 2) związane jest z poprawą adekwatności modelu odpowiadającej wartości 23.73 statystyki warunkowej. Model $\mathcal{M}[\mathcal{T}_1]$ nie jest jednak adekwatny, jeżeli utrzymamy przyjęty wcześniej poziom istotności 0.05. Podobne rozumowanie można przeprowadzić w odniesieniu do modelu $\mathcal{M}[\mathcal{T}_2]$. Tutaj poprawa adekwatności opisana jest przez wartość statystyki warunkowej równą 41.61, przy czym model $\mathcal{M}[\mathcal{T}_2]$ należy uznać za adekwatny do sytuacji zaobserwowanej. Wnioskujemy więc, że struktura analizowanej tablicy kontyngencji może być opisana za pomocą modelu $\mathcal{M}[\mathcal{T}_2]$ postulującego brak interakcji A_1xA_2 w podtablicy $\mathcal{T}_2$.

Tab. 4. Analiza zgodności modelu $\mathcal{M}$ dla tablicy oryginalnej $\mathcal{T}$ oraz podtablic $\mathcal{T}_1$ (terminy 1 i 2) i $\mathcal{T}_2$ (terminy 2 i 3)

Zgodność modelu	Wartość statystyki G^2	Liczba stopni swobody	Wyliczony poziom istotności
$\mathcal{M}[\mathcal{T}]$	53.67	32	0.010
$\mathcal{M}[\mathcal{T}_1]$	29.94	16	0.018
$\mathcal{M}[\mathcal{T}]\mid\mathcal{M}[\mathcal{T}_1]$	23.73	16	
$\mathcal{M}[\mathcal{T}_2]$	12.06	16	0.740
$\mathcal{M}[\mathcal{T}]\mid\mathcal{M}[\mathcal{T}_2]$	41.61	16	

Podsumujmy uzyskane wnioski dotyczące wyjaśnienia nieadekwatności modelu stwierdzającego brak interakcji $A_1 \times A_2$ w rozpatrywanej tablicy kontyngencji. Z pierwszej części przeprowadzonej analizy wynika, iż model ten nie jest adekwatny dlatego, że wpływ terminu na stopień porażenia roślin odmian Q,J,N i G jest inny niż dla odmiany P. Pokazaliśmy także, że najbardziej zbliżoną do odmiany P jest odmiana G. Druga część analizy wskazała, iż nieadekwatność modelu wynika z tego, że zróżnicowanie odmian w terminie 1 jest inne niż w terminie 2. Jeżeli obserwacje byłyby dokonywane tylko w terminach 2 i 3, to strukturę odpowiedniej tablicy kontyngencji można byłoby opisać w sposób adekwatny za pomocą modelu stwierdzającego brak interakcji.

Ostatni z zanotowanych powyżej wniosków wskazuje, że dla tablicy opisującej stopień porażenia roślin pięciu odmian w terminach drugim i trzecim można przeprowadzić dalszą analizę polegającą na testowaniu zgodności modeli logarytmiczno-liniowych o mniejszej liczbie parametrów. W Tabeli 5 są przedstawione wyniki testowania modelu opisującego jednorodność odmian pod względem stopnia porażenia roślin (model $\mathcal{M}_1$) oraz modelu opisującego brak wpływu terminu na stopień porażenia roślin (model $\mathcal{M}_2$). Oba te modele nie są adekwatne. Nieadekwatność $\mathcal{M}_1$ oznacza, że zróżnicowanie odmian w terminie drugim i trzecim jest istotne. Nieadekwatność $\mathcal{M}_2$ oznacza, że istotny jest wpływ terminu obserwacji na stopień porażenia roślin.

Na zakończenie spróbujmy stwierdzoną powyżej nieadekwatność modelu $\mathcal{M}_1$ o klasie generującej {12,23} wyjaśnić poprzez rozbicie wartości G^2 dla tego modelu na składniki związane z terminami pierwszym i drugim. Możemy to uczynić, ponieważ model $\mathcal{M}_1$ jest modelem {2}-rozkładalnym. Wyniki obliczeń, przedstawione w Tabeli 6, wskazują że zróżnicowanie odmian pod względem stopnia porażenia roślin jest istotne w obu terminach.

Tab. 5. Wyniki testowania zgodności wybranych modeli nienasyconych tablicy kontyngencji odpowiadającej dwu ostatnim terminom obserwacji

Model	Klasa generująca	Wartość statystyki ^{2}	Liczba stopni swobody	Wyliczony poziom istotności
$\mathcal{M}$	{12,13,23}	12.06	16	0.740
$\mathcal{M}_1$	{12,23}	233.53	32	0.000
$\mathcal{M}_2$	{12,13}	75.10	20	0.000

Tab. 6. Wartości statystyki G^2 dla testowania zgodności modelu opisującego brak zróżnicowania odmian w terminach drugim i trzecim

Termin	Wartość statystyki G^2
1	113.22
2	120.31
Razem	233.53
Wartość krytyczna dla testowania jednoczesnego procedurą STP2 na poziomie 0.05	28.85

LITERATURA

Alt F.B. (1982). Bonferroni inequalities and intervals. *Encyclopedia of statistical sciences.* Vol. 2. (Kotz S., Johnson N.L., Eds.). Wiley, New York.

Bishop Yvonne M.M., Fienberg S.E. i Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* MIT Press, Cambridge, Massachusetts.

Caliński T., Dyczkowski A. i Sitek Mirosława (1979). Procedury testów jednoczesnych w wielozmiennej analizie wariancji. *Matematyka Stosowana* **XIV**, 5-31.

Czajka S. i Krajewski P. (1986). Podstawy analizy statystycznej wielowymiarowych tablic kontyngencji przy użyciu modelu logarytmiczno-liniowego. *Listy Biometryczne* **25**, 17-36.

Gabriel K.R. (1966). Simultaneous test procedures for multiple comparisons on categorical data. *J. Amer. Statist. Ass.* **61**, 1081-1096.

Gabriel K.R. (1969). Simultaneous test procedures - some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224-250.

Gilula Z. (1985). On the analysis of heterogeneity among populations. *J. R. Statist. Soc.* **47**, 76-83.

Goodman L.A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Ass.* **63**, 1091-1131.

Goodman L.A. (1971). A simple simultaneous test procedure for quasi-independence in contingency tables. *Appl. Statist.* **20**, 165-177.

Haberman S.J. (1978). *Analysis of qualitative data.* Vol.1: Introductory topics, Vol.2: New developments. Academic Press, New York.

Imrey P.B., Koch G.G. i Stokes Maura E. (1981). Categorical Data Analysis: some reflections on the log-linear model and logistic regression. Part I: Historical and methodological overview. *Int. Statist. Rev.* **49**, 265-283.

Imrey P.B., Koch G.G. i Stokes Maura E. (1982). Categorical Data Analysis: some reflections on the log-linear model and logistic regression. Part II: Data analysis. *Int. Statist. Rev.* **50**, 35-63.

Miller R.G. (1981). *Simultaneous statistical inference.* 2nd Edition. Springer Verlag, New York.

Plackett R.L. (1981). *The analysis of categorical data.* 2nd edition. Griffin, London.

Wickens T.D. (1989). *Multiway contingency tables analysis for the social sciences.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Praca wpłynęła 12 listopada 1990

ANALYSIS OF CONTINGENCY SUBTABLES USING THE LOG-LINEAR MODEL

Summary

The paper describes methods of explaining the inadequacy of a log-linear model of a multidimensional contingency table by testing the fit of implied models. Some individual and simultaneous testing procedures are proposed. The problem of the proper choice of the family of implied models is discussed, in particular for the class of the so called decomposable models. The theory is illustrated with an example of the analysis of a contingency table obtained in the study on rape varieties resistance.

Key words: contingency tables, log-linear model, implied model, simultaneous testing