Listy Biometryczne
Vol. XXIII (1986) No.1
Biometrical Letters

PODSTAWY ANALIZY STATYSTYCZNEJ WIELOWYMIAROWYCH TABLIC KONTYNGENCJI PRZY UŻYCIU MODELU LOGARYTMICZNO-LINIOWEGO*

STANISŁAW CZAJKA, PAWEŁ KRAJEWSKI

Zakład Metod Matematycznych i Statystycznych Akademii Rolniczej w Poznaniu
Instytut Genetyki Roślin Polskiej Akademii Nauk w Poznaniu

Praca wpłynęła 5 kwietnia 1985; w wersji ostatecznej 19 grudnia 1985

Czajka S., Krajewski P., 1987. An introduction to statistical analysis of multidimensional contingency tables by a log-linear model. Listy Biometryczne XXIII z. 1. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu (Adam Mickiewicz University Press), pp.17-36, 4 tabl. ISBN 83-232-0090-4, ISSN 0458-0036.

The paper is devoted to the problem of analysis of multivariate discrete (qualitative) data. The concepts of a multidimensional contingency table and a log-linear model are described. Using parameters of the model some hypotheses about various types of dependencies among variables are formulated. Estimation of expected counts and testing the hypotheses are described. The theory is illustrated with examples from the field of plant breeding.

1. WSTĘP

W badaniach genetycznych i hodowlanych często pojawia się potrzeba określania zależności cech (zmiennych losowych) zwanych jakościowymi lub dyskretnymi. Informacje o rozkładzie cech jakościowych w próbie przedstawia się zwykle w formie tzw. tablicy kontyngencji. Analiza statystyczna tej tablicy polega na weryfikacji hipotez mówiących o braku zależności obserwowanych zmiennych losowych. Nie jest to trudne, jeżeli badaniu podlegają dwie cechy i wystarczające jest zastosowanie zwykłego testu niezależności χ^2. Sytuacja komplikuje się, gdy liczba cech przekracza dwa i przedmiotem

* Praca wykonana w ramach problemu węzłowego 09.1 koordynowanego przez Instytut Hodowli i Aklimatyzacji Roślin.

zainteresowania są nie tylko zależności rzędu pierwszego (tzn. między każdymi dwiema zmiennymi), ale i zależności rzędów wyższych. Tablicę kontyngencji nazywamy wtedy tablicą wielowymiarową. W pracy niniejszej prezentujemy pewną metodę, która pozwala na opisanie struktury takiej tablicy poprzez wyodrębnienie zależności istotnych statystycznie.

Prace poświęcone analizie statystycznej wielowymiarowych tablic kontyngencji zajmują od pewnego czasu poczesne miejsce w światowej literaturze statystycznej. Rozwijane i z powodzeniem stosowane w praktyce są metody różniące się nieco między sobą podłożem teoretycznym lub przeznaczeniem. Podstawy podejścia, które chcemy przedstawić, opracowane zostały w końcu lat pięćdziesiątych. Wtedy to zauważono, że zależność dyskretnych zmiennych losowych można badać formułując pewne funkcje parametrów łącznego rozkładu tych zmiennych podobne do efektów głównych i interakcyjnych występujących w analizie wariancji doświadczeń wieloczynnikowych. Fundamentalne prace z tego zakresu opublikował w latach 1958-1965 I.J. Good. Jego koncepcje rozwinęli Darroch (1962) i Birch (1963), którzy opisali model tablicy kontyngencji zwany logarytmiczno-liniowym. O związanych z zastosowaniem tego modelu problemach teoretycznych i praktycznych w wyczerpujący sposób traktują monografie Bishop i in. (1975) oraz Placketta (1981); godnymi uwagi są też przeglądowe artykuły Imrey'a i in. (1981,1982).

Omówienie metody rozpoczynamy od podania w punkcie 2 niniejszej pracy sposobu konstrukcji wielowymiarowej tablicy kontyngencji i jej modelu probabilistycznego. Dalej zajmujemy się określeniem postaci interesujących hipotez statystycznych, a w punkcie 4 wprowadzamy pojęcie modelu logarytmiczno-liniowego i pokazujemy jego zastosowanie do formułowania wspomnianych hipotez w przypadkach dwu- i trójwymiarowym. Uogólnienie rozważań na przypadek dowolnej liczby zmiennych przedstawione jest w punkcie 5. Problemowi testowania opisanych hipotez poświęcony jest punkt 6, przy czym ograniczamy się w nim jedynie do poglądowego przedstawienia faktów niezbędnych do zrozumienia zasad prezentowanego podejścia. Kończący pracę punkt 7 pokazuje sposób zastosowania metody na konkretnych przykładach zaczerpniętych z dziedziny hodowli roślin.

2. WIELOWYMIAROWE TABLICE KONTYNGENCJI

Rozważania zacznijmy od sprecyzowania niektórych użytych we wstępie pojęć podstawowych. Cechę (właściwość) elementów pewnej populacji nazywać będziemy jakościową, jeżeli pozwala ona podzielić badaną zbiorowość na skończoną liczbę rozłącznych klas zwanych kategoriami. Zgodnie z taką definicją, cechą jakościową jest każda cecha mierzalna, mogąca przyjmować tylko skończoną liczbę różnych wartości liczbowych; jej kategoriami są grupy utworzone z elementów pod względem tej cechy równych. Częściej jednak, mówiąc o cechach jakościowych, będziemy mieli na myśli cechy niemierzalne, których kategorie są zdefiniowane jako grupy elementów określane wspólnym, na ogół umownym i nie związanym z konkretną wartością liczbową,

mianem (nazwą). Dla takich cech przeznaczona jest przede wszystkim opisana w niniejszej pracy metoda analizy statystycznej. Nie wykorzystuje ona w żaden sposób określających kategorie wartości lub nazw; w tym sensie może być rozumiana jako analiza, wynikającej z podziału na kategorie, klasyfikacji elementów populacji generalnej.

Zauważmy dalej, że jeżeli liczba kategorii cechy ilościowej wynosi I, to umowne ponumerowanie tych kategorii prowadzi do określenia dyskretnej zmiennej losowej przyjmującej wartości 1,2,...,I. Fakt powyższy pozwala na wygodne, choć nie całkiem precyzyjne, używanie określeń „cecha" i „zmienna" wymiennie. Czynimy tak w dalszym ciągu pracy, używając też dla oznaczenia cechy jakościowej i związanej z nią dyskretnej zmiennej losowej tego samego symbolu.

Przejdźmy teraz do określenia pojęcia wielowymiarowej tablicy kontyngencji. Z tablicą taką mamy do czynienia w sytuacji eksperymentalnej, w której interesujące jest jednoczesne opisanie badanej populacji pod względem kilku cech jakościowych. Oczywiste jest, że łącznie cechy te wprowadzają podział elementów populacji na rozłączne podklasy (kombinacje kategorii), których liczba jest równa iloczynowi liczb kategorii poszczególnych zmiennych. Przez wielowymiarową tablicę kontyngencji rozumieć będziemy tablicę liczebności elementów należących do tak zdefiniowanych podklas w próbie losowej. Zgodnie z tym, gdy badaniu podlegają dwie cechy, A_1 i A_2 o liczbach kategorii odpowiednio I oraz J, tablica kontyngencji jest dwuwymiarowa i składa się z liczb x_{ij} określających obserwowane liczebności elementów należących do i-tej (i=1,...,I) kategorii cechy A_1 oraz j-tej (j=1,...,J) kategorii cechy A_2; tablicę taką oznaczać będziemy symbolem $\{x_{ij}\}$. Trzy zmienne losowe A_1, A_2 i A_3 o liczbach kategorii I, J i K generują tablicę trójwymiarową o elementach x_{ijk} (i=1,...,I; j=1,...,J; k=1,...,K), oznaczaną dalej symbolem $\{x_{ijk}\}$. Przeniesienie pojęcia tablicy kontyngencji na przypadek dowolnej liczby cech nie przedstawia żadnych trudności. Zajmijmy się więc modelem probabilistycznym tej tablicy.

Tablica kontyngencji generowana przez s cech jakościowych $A_1,...,A_s$ o liczbach kategorii $I_1,...,I_s$ składa się z rozłącznych podklas, których liczba wynosi

$$t = \prod_{q=1}^{s} I_q .$$

Załóżmy, że prawdopodobieństwo wylosowania z populacji elementu należącego do r-tej (r=1,...,t) podklasy jest nieznane i wynosi p_r, przy czym $p_r > 0$ oraz

$$\sum_{r=1}^{t} p_r = 1. \qquad (2.1)$$

Określmy zmienne losowe X_r, r=1,...,t, w ten sposób, że $X_r = x_r$, gdy w próbie o liczebności N element należący do r-tej podklasy wystąpi x_r razy.

Zgodnie z przyjętymi powyżej założeniami zmienna losowa $\underline{X} = (X_1,\ldots,X_t)'$ ma rozkład wielomianowy, tzn. dla każdego wektora $\underline{x} = (x_1,\ldots,x_t)'$ o składowych spełniających warunki $x_r = 0,1,\ldots,N$ oraz $\sum_{r=1}^{t} x_r = N$

$$P(\underline{X} = \underline{x}) = N! \prod_{r=1}^{t} \frac{p_r^{x_r}}{x_r!}.$$

Można pokazać, że wartość oczekiwana zmiennej $\underline{X}$ wynosi

$$\underline{m} = N\underline{p}.$$

Składowe tak określonego wektora $\underline{m}$ nazywać będziemy liczebnościami oczekiwanymi podklas tablicy kontyngencji. Zgodnie z przyjętymi założeniami są one nieznane i spełniają warunki: $m_r > 0$,

$$\sum_{r=1}^{t} m_r = N. \tag{2.2}$$

Parametry m_r tworzą, podobnie jak obserwacje x_r, tablicę s-wymiarową, którą będziemy nazywać tablicą liczebności oczekiwanych i zapisywać przy użyciu symboli typu $\{m_{ij}\}$ oraz $\{m_{ijk}\}$. Warunek (2.2) powoduje, że w tablicy liczebności oczekiwanych występuje t-1 wielkości niezależnych.

3. SFORMUŁOWANIE HIPOTEZ STATYSTYCZNYCH

Jak określono we wstępie do niniejszej pracy, jej celem jest podanie metody odpowiedniej dla badania struktury zależności obserwowanych zmiennych losowych. Złożoność tej struktury, a więc liczba i rodzaj interesujących hipotez, zależy bezpośrednio od liczby badanych cech. W przypadku dwuwymiarowym interesująca jest jedynie hipoteza o niezależności zmiennych A_1 i A_2 rozumianej w (zwykłym) sensie równości prawdopodobieństw p_{ij} oraz iloczynu odpowiednich prawdopodobieństw brzegowych. Hipotezę tę zapisać można też za pomocą warunku

$$\frac{m_{ij}\, m_{i'j'}}{m_{i'j}\, m_{ij'}} = 1, \quad \begin{matrix} i,i' = 1,\ldots,I, \\ j,j' = 1,\ldots,J. \end{matrix} \tag{3.1}$$

Lewa strona tego wzoru jest miarą zależności, znaną w literaturze jako iloraz iloczynów krzyżowych. Łatwo pokazać, że w zbiorze wszystkich ilorazów tego typu, możliwych do wyliczenia w tablicy dwuwymiarowej, istnieje (I-1)(J-1) ilorazów niezależnych. Na przykład, w tablicy liczebności oczekiwanych o rozmiarach 2 x 3, która ma postać

$$\begin{matrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{matrix}$$

wystarczy ograniczyć się do rozpatrzenia wielkości

$$\alpha_1 = \frac{m_{11}\, m_{22}}{m_{21}\, m_{12}}, \qquad \alpha_2 = \frac{m_{12}\, m_{23}}{m_{22}\, m_{13}}. \tag{3.2}$$

Każdy iloraz wyliczony dla innego doboru wskaźników, zmieniających się tak jak we wzorze (3.1), może być przedstawiony jako funkcja α_1 i α_2. Jeżeli fakt ten uwzględnimy formułując hipotezę o niezależności, to dla omawianego przykładu otrzymamy równoważny opisanemu przez (3.1) warunek

$$\alpha_1 = \alpha_2 = 1. \tag{3.3}$$

Zajmijmy się teraz omówieniem hipotez dotyczących struktury zależności zmiennych w przypadku trójwymiarowym. Jego cechą charakterystyczną jest możliwość badania tak zwanej zależności drugiego rzędu. Aby wyjaśnić to pojęcie zauważmy, że trójwymiarową tablicę kontyngencji o rozmiarach I x J x K traktować można jako zbiór K tablic dwuwymiarowych o rozmiarach I x J. Jeżeli stopień zależności zmiennych A_1 i A_2, mierzony w tak określonych tablicach wielkością odpowiednich ilorazów iloczynów krzyżowych, jest taki sam dla wszystkich kategorii cechy A_3, to mówimy o braku zależności drugiego rzędu zmiennych A_1, A_2 i A_3. Sytuację tę można opisać za pomocą warunku

$$\frac{m_{ijk}\, m_{i'j'k}}{m_{i'jk}\, m_{ij'k}} = \frac{m_{ijk'}\, m_{i'j'k'}}{m_{i'jk'}\, m_{ij'k'}}, \qquad \begin{matrix} i,i' = 1,\ldots,I, \\ j,j' = 1,\ldots,J, \\ k,k' = 1,\ldots,K. \end{matrix} \tag{3.4}$$

Łatwo sprawdzić, że warunek (3.4) opisuje jednocześnie stałą zależność zmiennych A_1 i A_3 dla wszystkich kategorii A_2 oraz zmiennych A_2 i A_3 dla wszystkich kategorii A_1. Brak zależności drugiego rzędu oznacza więc taką samą zależność dowolnej pary zmiennych dla wszystkich kategorii trzeciej zmiennej.

Warunek (3.4) można znacznie uprościć pamiętając o związkach zachodzących pomiędzy ilorazami iloczynów krzyżowych wyliczonymi dla różnych doborów wskaźników. I tak, w przypadku tablicy $\{m_{ijk}\}$ o rozmiarach 2 x 3 x 2 mającej postać

$$\begin{matrix} m_{111} & m_{121} & m_{131} & \qquad & m_{112} & m_{122} & m_{132} \\ m_{211} & m_{221} & m_{231} & & m_{212} & m_{222} & m_{232} \end{matrix}, \tag{3.5}$$

wystarczy rozważyć układ równości

$$\alpha_1 = \beta_1, \qquad \alpha_2 = \beta_2, \tag{3.6}$$

gdzie

$$\alpha_1 = \frac{m_{111}\, m_{221}}{m_{211}\, m_{121}}, \quad \alpha_2 = \frac{m_{121}\, m_{231}}{m_{221}\, m_{131}},$$

$$\beta_1 = \frac{m_{112}\, m_{222}}{m_{212}\, m_{122}} \quad \beta_2 = \frac{m_{122}\, m_{232}}{m_{222}\, m_{132}},$$

Zauważmy, że określenie braku zależności drugiego rzędu w tablicy trójwymiarowej jest rekurencyjne. Korzysta się w nim bezpośrednio z definicji zależności pary zmiennych w tablicy dwuwymiarowej. Wniosek ten okaże się przydatny przy uogólnianiu rozważań na przypadek większej niż 3 liczby badanych cech.

Kolejnymi hipotezami dotyczącymi struktury zależności zmiennych w przypadku trójwymiarowym są hipotezy o niezależności warunkowej. Jedna z nich, mówiąca o niezależności zmiennych A_1 i A_2 wewnątrz dowolnej, ustalonej kategorii zmiennej A_3, jest określona warunkiem

$$\frac{m_{ijk}\, m_{i'j'k}}{m_{i'jk}\, m_{ij'k}} = 1, \quad \begin{matrix} i,i' = 1,\ldots,I, \\ j,j' = 1,\ldots,J, \\ k = 1,\ldots,K. \end{matrix} \tag{3.7}$$

Jak widać, hipoteza ta specyfikuje wartości ilorazów iloczynów krzyżowych, które opisują zależność zmiennych A_1 i A_2 i które pojawiły się już we wzorze (3.4). Dla rozpatrywanej jako przykład tablicy (3.5) wzór (3.7) przyjmuje postać

$$\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1. \tag{3.8}$$

Podobne warunki mogą być spełnione w odniesieniu do zależności zmiennych A_1 i A_3 przy ustalonej kategorii A_2 lub zależności zmiennych A_2 i A_3 przy ustalonej kategorii A_1, co prowadzi do określenia hipotez o warunkowej niezależności pozostałych par zmiennych. Można też łączyć tego typu stwierdzenia i formułować hipotezy o niezależności jednej ze zmiennych względem obu pozostałych bądź o niezależności całkowitej (tzn. o warunkowej niezależności wszystkich par zmiennych). Określające te hipotezy warunki uzyskamy nakładając ograniczenia na odpowiednie ilorazy iloczynów krzyżowych wyznaczone w tablicy $\{m_{ijk}\}$.

Znajomość postaci hipotez interesujących podczas analizy tablic dwu- i trójwymiarowych pozwala na rozpatrzenie sytuacji ogólnej, w której liczba obserwowanych zmiennych jest dowolna i wynosi s. Oczywiste jest teraz, że w tablicy s-wymiarowej możliwe jest badanie szeregu zależności rzędów 1,2,...,s-1. Definicje tych zależności otrzymamy stosując rozumowanie podobne do przedstawionego w przypadku trójwymiarowym, tzn. traktując tablicę s-wymiarową jako zbiór tablic o liczbie wymiarów s-1. Posłużymy się tu przykładem tablicy czterowymiarowej o rozmiarach I x J x K x L, którą potraktujemy jako zbiór L tablic trójwymiarowych o rozmiarach I x J x K. W każdej z takich „podtablic" można określić za pomocą warunku podobnego do (3.4) zależność drugiego rzędu zmiennych A_1, A_2 i A_3. Jeżeli jest ona ta-

ka sama dla każdej kategorii zmiennej A_4, to mówimy o braku zależności rzędu trzeciego. Taka interpretacja pozwala łatwo uzyskać definicję braku zależności rzędu drugiego i pierwszego w tablicy czterowymiarowej. Zauważmy jednak, że przy tego typu rozważaniach pojawia się trudność precyzyjnego zapisu poszczególnych hipotez. Nie jest odpowiedni do tego aparat matematyczny, posługujący się tylko definicją ilorazu iloczynów krzyżowych. Już w przypadku czterech zmiennych dla zapisu hipotezy o braku zależności trzeciego rzędu konieczne jest wprowadzenie wyrażeń bardziej skomplikowanych, praktycznie powodujących nieczytelność odpowiednich wzorów. Okazuje się, że trudność tę można pokonać poprzez odwołanie się do znanych w statystyce metod modelowania liniowego, co też uczynimy w następnym punkcie.

4. MODEL LOGARYTMICZNO-LINIOWY TABLIC DWU- I TRÓJWYMIAROWYCH

Podstawą podejścia, które chcemy zaprezentować, jest idea zastąpienia występujących w rozważaniach punktu 3 ilorazów iloczynów krzyżowych ich logarytmami naturalnymi. Konsekwencją tego jest możliwość prostej parametryzacji struktury zależności wielowymiarowej tablicy kontyngencji i uproszczenia zapisów interesujących hipotez.

Rozważania szczegółowe zacznijmy od przypadku dwuwymiarowego, któremu odpowiada tablica liczebności oczekiwanych $\{m_{ij}\}$ o rozmiarach I x J. Wprowadźmy oznaczenie

$$l_{ij} = \log m_{ij}$$

i zdefiniujmy w tablicy $\{l_{ij}\}$ sumy brzegowe

$$l_{i+} = \sum_{j=1}^{J} l_{ij}, \quad l_{+j} = \sum_{i=1}^{I} l_{ij}, \quad l_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} l_{ij}$$

oraz wielkości

$$\begin{aligned} w &= \frac{1}{IJ} l_{++}, \\ w_{1(i)} &= \frac{1}{J} l_{i+} - w, \\ w_{2(j)} &= \frac{1}{I} l_{+j} - w, \\ w_{12(ij)} &= l_{ij} - w_{1(i)} - w_{2(j)} - w. \end{aligned} \tag{4.1}$$

Przy takich oznaczeniach można zapisać prawdziwą dla $i = 1,\ldots,I$, $j = 1,\ldots,J$ równość

$$\log m_{ij} = w + w_{1(i)} + w_{2(j)} + w_{12(ij)} . \tag{4.2}$$

Zapis ten nazywany jest modelem logarytmiczno-liniowym dwuwymiarowej tablicy kontyngencji. Przedstawia on logarytm naturalny wartości oczekiwanej m_{ij} jako funkcję pewnych nieznanych wielkości, zwanych dalej parametrami modelu. Z definicji (4.1) tych parametrów wynika, że

$$\sum_{i=1}^{I} w_{1(i)} = \sum_{j=1}^{J} w_{2(j)} = 0,$$

$$\sum_{i=1}^{I} w_{12(ij)} = 0, \quad j = 1,\ldots,J, \tag{4.3}$$

$$\sum_{j=1}^{J} w_{12(ij)} = 0, \quad i = 1,\ldots,I,$$

natomiast warunek

$$\sum_{i=1}^{I} \sum_{j=1}^{J} m_{ij} = N$$

implikuje zależność

$$\exp w = N / \sum_{i=1}^{I} \sum_{j=1}^{J} \exp(w_{1(i)} + w_{2(j)} w_{12(ij)}). \tag{4.4}$$

Równości (4.3) i (4.4) powodują, że liczba niezależnych parametrów modelu (4.2) wynosi IJ-1 i jest równa liczbie niezależnych elementów tablicy $\{m_{ij}\}$. W związku z tym model logarytmiczno-liniowy (4.2) nazywany jest nasyconym. Jego użyteczność wynika z faktu, że interesującą nas niezależność zmiennych losowych A_1 i A_2 można zdefiniować nakładając na parametry $w_{12(ij)}$ dodatkowe ograniczenia

$$w_{12(ij)} = 0, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J. \tag{4.5}$$

Równoważność warunków (4.5) i (3.1) łatwo wykazać dla dowolnych wartości I i J, stosując elementarne własności funkcji logarytmicznej. Bardziej celowe i pouczające wydaje się jednak ograniczenie rozważań do przykładowej tablicy o ustalonych rozmiarach, co powoduje łatwość zastosowania wprowadzonej w punkcie 2 notacji wektorowej.

Przyjmijmy więc, podobnie jak w punkcie 3, że I=2, J=3. Uwzględniając warunki (4.3), ograniczenia (4.5) zapiszemy w postaci

$$\begin{aligned} w_{12(11)} &= 0, \\ w_{12(12)} &= 0. \end{aligned} \tag{4.6}$$

Ponieważ zgodnie z definicją (4.1) mamy

$$w_{12(11)} = \frac{1}{6}(2l_{11} - 2l_{21} - l_{12} + l_{22} - l_{13} + l_{23}),$$

$$w_{12(12)} = \frac{1}{6}(-l_{11} + l_{21} + 2l_{12} - 2l_{22} - l_{13} + l_{23}),$$

układ (4.6) można zapisać w postaci równoważnej

$$\frac{1}{6}\begin{pmatrix} 2 & -2 & -1 & 1 & -1 & 1 \\ -1 & 1 & 2 & -2 & -1 & 1 \end{pmatrix} \text{Log } \underline{m} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

gdzie

$$\underline{m} = (m_{11}, m_{21}, m_{12}, m_{22}, m_{13}, m_{23})',$$

natomiast Log $\underline{m}$ jest wektorem kolumnowym, złożonym z logarytmów naturalnych składowych wektora $\underline{m}$. Proste przekształcenia pokazują, że układ powyższy jest równoważny układowi

$$\begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix} \text{Log } \underline{m} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

który może być przekształcony do postaci

$$\log \alpha_1 = 0$$
$$\log \alpha_2 = 0$$

równoważnej warunkowi (3.3).

Z rozważań powyższych wynika, że parametry $w_{12(ij)}$ modelu logarytmiczno-liniowego (będące w istocie funkcjami logarytmów ilorazów iloczynów krzyżowych) określają stopień zależności zmiennych A_1 i A_2. W związku z tym model (4.2) zapisany przy założeniu prawdziwości warunku (4.5), będący postaci

$$\log m_{ij} = w + w_{1(i)} + w_{2(j)},$$

nazwiemy modelem niezależności. Jest on, w przeciwieństwie do (4.2.), nienasycony, gdyż liczba jego niezależnych parametrów, wynosząca I+J-2, jest mniejsza niż liczba niezależnych liczebności oczekiwanych m_{ij}.

Przechodząc do przypadku trójwymiarowego zastosujmy rozumowanie podobne do zaprezentowanego w punkcie 3. Potraktujmy tablicę $\{m_{ijk}\}$ o rozmiarach I x J x K jako K tablic dwuwymiarowych i dla każdej z nich zapiszmy model nasycony

$$\log m_{ijk} = w^{(k)} + w^{(k)}_{1(i)} + w^{(k)}_{2(j)} + w^{(k)}_{12(ij)},$$

którego parametry są zdefiniowane wzorami podobnymi do (4.1). Całą tablicę trójwymiarową opiszemy modelem

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + \quad (4.7)$$
$$+ u_{123(ijk)},$$

gdzie

$$u = \frac{1}{K}\sum_{k=1}^{K} w^{(k)}, \quad u_{3(k)} = w^{(k)} - u,$$

$$u_{1(i)} = \frac{1}{K}\sum_{k=1}^{K} w_{1(i)}^{(k)}, \quad u_{13(ik)} = w_{1(i)}^{(k)} - u_{1(i)},$$

$$u_{2(j)} = \frac{1}{K}\sum_{k=1}^{K} w_{2(j)}^{(k)}, \quad u_{23(jk)} = w_{2(j)}^{(k)} - u_{2(j)}, \tag{4.8}$$

$$u_{12(ij)} = \frac{1}{K}\sum_{k=1}^{K} w_{12(ij)}^{(k)}, u_{123_{(ijk)}} = w_{12(ij)}^{(k)} - u_{12(ij)}.$$

Łatwo sprawdzić, że tak zdefiniowane wielkości spełniają związki podobne do (4.3) i (4.4). Powoduje to, że liczba niezależnych parametrów modelu (4.7) wynosi IJK-1. Tym samym (4.7) jest modelem nasyconym. Nakładając na jego parametry dodatkowe ograniczenia możemy konstruować modele nienasycone, odpowiadające poszczególnym hipotezom sformułowanym w punkcie 3. Związki tego typu można wykazać stosując postępowanie podobne do przedstawionego w przypadku dwóch zmiennych. Przykładowo, układ ograniczeń

$$u_{123(ijk)} = 0, \quad i=1,\ldots,I; \quad j=1,\ldots,J; \quad k=1,\ldots,K, \tag{4.9}$$

dla rozpatrywanej w punkcie 3 tablicy trójwymiarowej o rozmiarach 2 x 3 x x 2 ma postać

$$u_{123(111)} = 0,$$
$$u_{123(121)} = 0. \tag{4.10}$$

Korzystając z definicji (4.8) występujące powyżej parametry można przedstawić jako funkcje liniowe logarytmów wartości oczekiwanych m_{ijk}. Zastosowanie prostych przekształceń pozwala wtedy na wykazanie równoważności warunków (3.6) i (4.10). Szczegółowy zapis odpowiednich wzorów jest jednak trudny ze względu na duży rozmiar występujących w nich macierzy i wektorów, toteż pełną interpretację logarytmiczno-liniowego modelu tablicy trójwymiarowej przeprowadzimy w sposób skrócony, korzystając z faktów pokazanych w przypadku dwuwymiarowym.

U w a g a. W dalszym ciągu, dla uproszczenia zapisów, będziemy pomijać wskaźniki występujące przy parametrach modelu. Na przykład zapis $u_{12}=0$ będziemy rozumieli jako skróconą notację układu równości $u_{12(ij)} = 0$, i=1,...,I; j=1,...,J. Nie powinno też doprowadzić do nieporozumień określenie typu „grupa parametrów u_{12}" (stosowane przy odwoływaniu się do wszystkich wielkości $u_{12(ij)}$).

Zacznijmy od interpretacji warunku (4.9), który zgodnie z przyjętą notacją zapiszemy w postaci

$$u_{123} = 0. \tag{4.11}$$

Ograniczenie to, zgodnie z definicją (4.8), oznacza, że

$$w^{(k)}_{12(ij)} = \frac{1}{K} \sum_{k=1}^{K} w^{(k)}_{12(ij)},$$

a ta równość spełniona jest wtedy i tylko wtedy, gdy

$$w^{(1)}_{12(ij)} = w^{(2)}_{12(ij)} = \ldots = w^{(K)}_{12(ij)},$$

tzn. gdy parametry określające zależność zmiennych A_1 i A_2 są takie same dla każdej z K warstw tablicy trójwymiarowej. Ponieważ model (4.7) można też uzyskać traktując tablicę $\{m_{ijk}\}$ jako I dwuwymiarowych tablic o rozmiarach J x K lub jako J dwuwymiarowych tablic o rozmiarach I x K, więc warunek (4.11) opisuje sytuację, w której zależność każdej pary obserwowanych cech jest taka sama dla wszystkich kategorii cechy trzeciej, co odpowiada podanemu w punkcie 3 określeniu braku zależności drugiego rzędu.

Zapiszmy dalej warunek

$$u_{12} = u_{123} = 0. \tag{4.12}$$

Z definicji (4.8) wynika, że jest on równoważny układowi ograniczeń

$$w^{(k)}_{12(ij)} = 0,$$

który opisuje niezależność zmiennych A_1 i A_2 przy ustalonej kategorii trzeciej zmiennej A_3. Tak więc model nienasycony, odpowiadający ograniczeniu (4.12), nazwiemy modelem warunkowej niezależności zmiennych A_1 i A_2. Podobnie pokazać można, że ograniczenia

$$u_{13} = u_{123} = 0,$$

$$u_{23} = u_{123} = 0$$

dotyczą odpowiednio warunkowej niezależności zmiennych A_1 i A_3 oraz zmiennych A_2 i A_3.

Idąc dalej zapytajmy o znaczenie modelu (4.7) przy ograniczeniach

$$u_{12} = u_{13} = u_{123} = 0. \tag{4.13}$$

Jak wynika z poprzednich rozważań, opisują one sytuację, w której zmienna A_1 jest niezależna od A_2 i A_3, natomiast A_2 i A_3 są zależne. Przez analogię do (4.13) otrzymamy interpretację modeli nienasyconych odpowiadających warunkom

$$u_{12} = u_{23} = u_{123} = 0,$$

$$u_{13} = u_{23} = u_{123} = 0.$$

Wreszcie model nienasycony o parametrach spełniających ograniczenia

$$u_{12} = u_{13} = u_{23} = u_{123} = 0$$

opisuje niezależność każdej pary cech, toteż nazwiemy go modelem całkowitej niezależności. Odpowiada on najprostszej strukrurze zależności zmiennych w przypadku trójwymiarowym.

Kończąc rozważania na temat tablic trójwymiarowych zauważmy, że analizowane układy warunków nie opisują wszystkich możliwych dla tych tablic modeli nienasyconych. Ograniczyliśmy się do omówienia modeli zwanych hierarchicznymi. Mają one wyspecyfikowane wartości parametrów określających zależności pewnych podzbiorów zmiennych oraz wszystkich parametrów określających zależność tych podzbiorów z pozostałymi zmiennymi. Korzystając z definicji (4.8) można pokazać, że tylko modele hierarchiczne mają przejrzystą interpretację w terminach stałej zależności cech lub braku zależności i do nich też ograniczymy się w dalszych rozważaniach. Z podobnych względów pomijamy modele, które mają wyspecyfikowane wartości „efektów głównych" jednej lub więcej zmiennych. Ich interpretacja nie jest interesująca z punktu widzenia analizy struktury zależności cech jakościowych.

5. MODEL I JEGO ZASTOSOWANIE W PRZYPADKU WIELOWYMIAROWYCH TABLIC KONTYNGENCJI

W punkcie 3 stwierdziliśmy, że w przypadku ogólnym, tzn. gdy liczba obserwowanych cech wynosi s, struktura tablicy kontyngencji obejmuje zależności rzędów 1,2,...,s-1. Odpowiedni dla opisu tej struktury model można otrzymać przez uogólnienie rozważań podanych w punkcie 4, co prowadzi do określenia modelu logarytmiczno-liniowego tablicy wielowymiarowej. Model taki przedstawia logarytmy oczekiwanych liczebności podklas jako funkcje liniowe parametrów, których definicje łatwo uzyskać stosując postępowanie rekurencyjne, przedstawione w punkcie 3 dla tablicy trójwymiarowej. Podejście to ułatwia też interpretację poszczególnych modeli nienasyconych, powstających przez nakładanie na wybrane parametry dodatkowych ograniczeń. Na przykład, gdy obserwujemy cztery zmienne A_1, A_2, A_3 i A_4, model opisujący brak zależności trzeciego rzędu zadany warunkiem

$$u_{1234} = 0$$

interpretujemy jako odpowiadający sytuacji, w której zależność drugiego rzędu każdej trójki zmiennych (np. A_1, A_2, A_3) jest taka sama dla każdej kategorii czwartej zmiennej. W konsekwencji tego warunek

$$u_{123} = u_{1234} = 0$$

opisuje brak zależności A_1, A_2 i A_3 przy ustalonej kategorii zmiennej A_4, lub inaczej - taką samą zależność zmiennych A_1 i A_2 dla każdej kategorii A_3 przy ustalonej kategorii A_4. Nakładanie na parametry dalszych ograniczeń prowadzi do modeli warunkowej i całkowitej niezależności, których in-

terpretacja jest na ogół łatwiejsza ze względu na brak parametrów wysokich rzędów.

Niniejszy punkt kończy rozważania związane z określeniem modelu logarytmiczno-liniowego i interpretacją jego parametrów. Uzasadniły one użyteczność modelu do opisu hipotez dotyczących struktury wielowymiarowych tablic kontyngencji. Pozostawiając do omówienia w punkcie następnym metodę testowania tych hipotez, podajmy kilka uwag dotyczących wykorzystania prezentowanego podejścia w praktyce eksperymentalnej.

W większości przypadków praktycznych informacje o strukturze zależności, uzyskane poprzez testowanie jednej, wybranej hipotezy, są niewystarczające. Gdy liczba obserwowanych zmiennych przekracza 3 nie jest też celowe badanie dopasowania wszystkich dających się skonstruować hierarchicznych modeli nienasyconych. Pomijając nawet problem związany z kosztem obliczeń, wnioskowanie na podstawie uzyskanych w ten sposób informacji może okazać się uciążliwe. Aby zmniejszyć liczbę wymagających testowania hipotez najkorzystniej jest odwołać się do apriorycznej wiedzy eksperymentatora i, o ile to możliwe, wnioskowanie ograniczyć tylko do zależności uznanych przez niego za najbardziej interesujące (por. przykład 7.2). Istnieją również metody pozwalające na przeprowadzenie efektywnej analizy statystycznej w sytuacji, gdy brak jest hipotez sformułowanych a priori. Celem wnioskowania jest wtedy znalezienie optymalnego (w pewnym sensie) hierarchicznego modelu nienasyconego. Szczegółowe omówienie tych metod przekracza jednak ramy niniejszego artykułu; wśród poświęconej im literatury na wyróżnienie zasługują m.in. prace Goodmana (1971), Browna (1976) i Havránka (1984).

Wspomnijmy jeszcze, że opisywana metoda formułowania i testowania hipotez dotyczących zależności zmiennych dyskretnych znajduje też zastosowanie w przypadku, gdy dysponujemy kilkoma próbami pobranymi niezależnie z różnych populacji. Taką sytuację eksperymentalną można uważać za szczególny przypadek określonej w punktach 2 i 3 (Bishop i in., 1975). Należy jedynie założyć, że jedna z „tworzących" tablicę kontyngencji cech nie ma charakteru losowego i opisuje tylko przynależność elementu próby do badanych zbiorowości (patrz przykład 7.2). Możliwość powyższa powoduje, że metoda analizy statystycznej, wykorzystująca model logarytmiczno-liniowy, wydaje się odpowiednia dla dostatecznie szerokiej klasy problemów związanych z cechami dyskretnymi.

6. ESTYMACJA LICZEBNOŚCI OCZEKIWANYCH I TESTOWANIE HIPOTEZ

Z dotychczasowych rozważań wynika, że hipotezy dotyczące rozkładów badanych zmiennych losowych można formułować w wygodny sposób za pomocą ograniczeń nakładanych na parametry modelu logarytmiczno-liniowego. Do weryfikowania tych hipotez służy metoda będąca w pewnym sensie uogólnieniem powszechnie znanego testu niezależności, stosowanego dla tablic dwuwymiaro-

wych. Jej idea sprowadza się do wyznaczenia, przy prawdziwości testowanej hipotezy, ocen nieznanych liczebności oczekiwanych oraz obliczenia wartości statystyki, która jest miarą odległości pomiędzy tymi ocenami a liczebnościami obserwowanymi. Pozwala to wnioskować o dopasowaniu odpowiadającego weryfikowanej hipotezie modelu nienasyconego.

Wprowadźmy najpierw niezbędne oznaczenia. Podobnie jak w punkcie 4 używać będziemy symbolu „+" dla określenia operacji sumowania po wszystkich kategoriach jednej z badanych zmiennych. Oznacza to na przykład, że przy czterech cechach mających I, J, K i L kategorii

$$x_{ijk+} = \sum_{l=1}^{L} x_{ijkl}, \tag{6.1}$$

$$x_{i+++} = \sum_{j=1}^{J} x_{ij++} = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk+} = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} x_{ijkl}.$$

Zdefiniowane w ten sposób sumy tworzą w naturalny sposób tzw. konfiguracje, czyli tablice o liczbie wymiarów mniejszej niż tablica obserwowana. Konfiguracje te będziemy oznaczć dalej literą C z odpowiednimi wskaźnikami. Zgodnie z tym wzory (6.1) określają odpowiednio elementy konfiguracji trójwymiarowej $C_{123} = \{x_{ijk+}\}$ oraz jednowymiarowej $C_1 = \{x_{i+++}\}$. Oczywiście każdej konfiguracji odpowiada pewna tablica sum brzegowych liczebności oczekiwanych; w rozpatrywanym przykładzie są to odpowiednio tablice $\{m_{ijk+}\}$ i $\{m_{i+++}\}$. Możemy także każdej konfiguracji przypisać w sposób wzajemnie jednoznaczny grupę parametrów nasyconego modelu logarytmiczno-liniowego, opisujących zależności pomiędzy uwzględnionymi w konfiguracji zmiennymi; dla C_{123} są to parametry $u_{123(ijk)}$, dla C_1 - parametry $u_{1(i)}$.

Oznaczmy dalej przez M dowolny hierarchiczny model nienasycony oraz przez H_o odpowiadającą temu modelowi hipotezę. Zauważmy, że opis modelu M nie musi zawierać w sposób jawny wszystkich jego parametrów. Wystarczy podać tylko niektóre ich grupy, tworzące tzw. zbiór definiujący. Jest to taki zbiór, którego znajomość pozwala ustalić pełną postać modelu przy wykorzystaniu zasady hierarchiczności. Przykładowo, niech M dotyczy tablicy czterowymiarowej i ma postać

$$\log m_{ijkl} = u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{13} + u_{23} + u_{14} + u_{123}. \tag{6.2}$$

Zbiorem definiującym tego modelu jest zbiór $\{u_{123}, u_{14}\}$, gdyż obecność grupy parametrów u_{123} implikuje obecność grup u_{12}, u_{13}, u_{23}, u_1, u_2, u_3, natomiast obecność u_{14} implikuje obecność parametrów u_4. Łatwo sprawdzić, że dla dowolnego modelu hierarchicznego zbiór definiujący określony jest w sposób jednoznaczny.

Jak wspomnieliśmy, pierwszym etapem postępowania przy testowaniu H_o jest znalezienie ocen liczebności oczekiwanych przy założeniu, że hipoteza jest prawdziwa. Odpowiednią do tego celu metodę podał po raz pierwszy Birch (1963). Pokazał on, że konfiguracje odpowiadające parametrom tworzącym

zbiór definiujący model M stanowią minimalne statystyki dostateczne dla klasy rozkładów wielomianowych określonej przez ten model oraz, że elementy tych konfiguracji są równe estymatorom największej wiarygodności odpowiednich sum brzegowych wartości oczekiwanych. Pozwala to utworzyć układ równań liniowych, którego rozwiązaniem (przy ograniczeniach określonych przez model) są szukane estymatory oczekiwanych liczebności podklas. Przykłady takiego postępowania podajemy poniżej. Wspomnijmy jeszcze, że warunki konieczne i dostateczne stosowalności metody podał Haberman (1973). Spełnienie tych warunków zależy od liczby i rozmieszczenia w tablicy kontyngencji liczebności obserwowanych równych zeru. W większości przypadków praktycznych na to, aby rozwiązanie wspomnianego wyżej układu równań istniało i było jednoznaczne, wystarcza brak zer w konfiguracjach stanowiących minimalne statystyki dostateczne (tzw. konfiguracjach dostatecznych).

P r z y k ł a d 6.1. Załóżmy, że analizowana tablica kontyngencji jest czterowymiarowa i interesująca jest hipoteza odpowiadająca modelowi (6.2). Model ten definiuje zbiór $\{u_{123}, u_{14}\}$, a więc minimalnymi statystykami dostatecznymi są konfiguracje

$$C_{123} = \{x_{ijk+}\}, \quad C_{14} = \{x_{i++l}\}.$$

Zgodnie z przytoczonymi powyżej rezultatami Bircha prawdziwe są, dla wszystkich możliwych wartości wskaźników i, j, k, l, równości

$$\begin{aligned} \hat{m}_{ijk+} &= x_{ijk+}, \\ \hat{m}_{i++l} &= x_{i++l}. \end{aligned} \tag{6.3}$$

Występują w nich estymatory największej wiarygodności sum brzegowych liczebności oczekiwanych. Równości (6.3) tworzą układ równań liniowych, którego niewiadomymi są wielkości $\hat{m}_{ijkl}$, czyli oceny nieznanych liczebności oczekiwanych m_{ijkl}. Oceny te można wyznaczyć rozwiązując układ (6.3) przy ograniczeniach wynikających z założenia prawdziwości testowanej hipotezy mających postać

$$\log \hat{m}_{ijkl} = u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{13} + u_{23} + u_{14} + u_{123}. \tag{6.4}$$

Rozwiązanie analityczne problemu jest w tym szczególnym przypadku stosunkowo proste, pomimo nieliniowego charakteru ograniczeń (6.4). Można bowiem pokazać, że zapis (6.4) jest równoważny warunkowi

$$\hat{m}_{ijkl} = \frac{\hat{m}_{ijk+}\,\hat{m}_{i++l}}{\hat{m}_{i+++}}, \tag{6.5}$$

który łącznie z (6.3) pozwala zapisać wzór

$$\hat{m}_{ijkl} = \frac{x_{ijk+}\,x_{i++l}}{x_{i+++}}.$$

P r z y k ł a d 6.2. W przypadku, gdy interesują nas trzy zmienne losowe i sformułujemy hipotezę o braku zależności drugiego rzędu postaci

$$H_o : u_{123} = 0,$$

minimalnymi statystykami dostatecznymi są konfiguracje $C_{12} = \{x_{ij+}\}$ $C_{13} = \{x_{i+k}\}$ oraz $C_{23} = \{x_{+jk}\}$, natomiast estymatory $\hat{m}_{ijk}$ otrzymamy rozwiązując układ równań

$$\begin{aligned} \hat{m}_{ij+} &= x_{ij+} \\ \hat{m}_{i+k} &= x_{i+k} \\ \hat{m}_{+jk} &= x_{+jk} \end{aligned} \tag{6.6}$$

przy ograniczeniach

$$\log \hat{m}_{ijk} = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23}. \tag{6.7}$$

W tym przypadku rozwiązanie nie może być uzyskane metodą analityczną (nie istnieje równoważny warunkowi (6.7) zapis szukanych ocen jako funkcji estymatorów występujących we wzorach (6.6)). W celu wyliczenia wartości $\hat{m}_{ijk}$ należy zastosować procedurę iteracyjną.

Uogólniając fakty pokazane na powyższych przykładach można stwierdzić, że estymatory największej wiarygodności liczebności oczekiwanych znajduje się korzystając tylko z postaci przyjętego modelu nienasyconego i elementów konfiguracji dostatecznych. Ponieważ rozwiązanie analityczne problemu nie zawsze istnieje, przy obliczeniach korzysta się na ogół z tzw. metody iteracyjnego dopasowania Deminga-Stephana (Bishop i in., 1975).

Przejdźmy teraz do problemu wyliczenia statystyki odpowiedniej dla testowania hipotezy H_o. Teoria wskazuje, że najlepiej jest wykorzystać tu funkcję testową opartą na ilorazie wiarygodności. Ma ona postać

$$G^2 = 2\sum_{r=1}^{t} z_r ,$$

gdzie

$$z_r = \begin{cases} x_r \log \dfrac{x_r}{\hat{m}_r}, & x_r > 0 , \\ 0, & x_r = 0 , \end{cases}$$

natomiast $\hat{m}_r$, $r = 1,\ldots,t$ są estymatorami największej wiarygodności liczebności oczekiwanych przy prawdziwości H_o. Statystyka G^2 ma asymptotyczny rozkład χ^2 o liczbie stopni swobody $v = t-p-1$, gdzie p oznacza ilość niezależnych parametrów modelu nienasyconego M odpowiadającego weryfikowanej hipotezie. Wyznaczanie wielkości v ilustruje poniższy przykład.

P r z y k ł a d 6.3. W sytuacji opisanej w przykładzie 6.2 liczby niezależnych parametrów $u_{1(i)}$, $u_{2(j)}$ i $u_{3(k)}$ wynoszą odpowiednio I-1, J-1, K-1, natomiast liczby niezależnych parametrów opisujących zależności par zmiennych są równe (I-1)(J-1), (I-1)(K-1) oraz (J-1)(K-1). Wynika to z

ograniczeń stanowiących, że odpowiednie sumy parametrów modelu logarytmiczno-liniowego są równe zeru (por. wzory 4.3). Liczbą wszystkich niezależnych parametrów modelu (6.7) jest więc

$$p = (I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(K-1),$$

a liczba stopni swobody dla testowania postawionej hipotezy wynosi:

$$v = IJK - p - 1 = (I-1)(J-1)(K-1).$$

7. PRZYKŁADY ZASTOSOWAŃ

Zastosowanie przedstawionego podejścia do analizy statystycznej wielowymiarowych tablic kontyngencji pokażemy na przykładach zaczerpniętych z badań prowadzonych w Ogrodzie Botanicznym PAN. Wszystkie potrzebne obliczenia wykonano używając programu obliczeniowego przeznaczonego na maszynę cyfrową Odra 1204 w Ośrodku Obliczeniowym Akademii Rolniczej w Poznaniu.

P r z y k ł a d 7.1. W próbie liczącej 300 roślin, pobranej z pewnej populacji żyta, badano wartości trzech cech jakościowych. Były to: B_1 - pokrój rośliny w stadium krzewienia, B_2 - owłosienie osi kłosa, B_3 - owłosienie dokłosia. Cecha B_1 mierzona była w skali trójstopniowej, cechy B_2 i B_3 - w skali dwustopniowej; wartości cech kodowano za pomocą kolejnych liczb naturalnych. Uzyskane obserwacje, zebrane w trójwymiarowej tablicy kontyngencji o rozmiarach 3 x 2 x 2, prezentujemy w postaci tabeli 1.

T a b l i c a 1. Trójwymiarowa tablica kontyngencji; obserwowane liczebności roślin sklasyfikowanych ze względu na pokrój w stadium krzewienia (B_1), owłosienie osi kłosa (B_2) oraz owłosienie dokłosia (B_3)

Cecha		B_3	
B_1	B_2	1	2
1	1	4	18
	2	4	101
2	1	3	13
	2	39	76
3	1	3	1
	2	18	20

Jak wspomnieliśmy poprzednio, w przypadku trójwymiarowym możliwe jest przeprowadzenie analizy polegającej na testowaniu dopasowania wszystkich dających się skonstruować hierarchicznych modeli nienasyconych. Wyniki badań przedstawiamy w tablicy 2, która podaje postać modelu (znak + w kolumnie odpowiadającej danej grupie parametrów oznacza, że ich wartości nie są wyspecyfikowane przez hipotezę), odpowiadającą mu liczbę stopni swobody, wyliczoną wartość statystyki G^2 oraz poziom istotności, rozumiany jako prawdopodobieństwo przekroczenia wyliczonej wartości przez zmienną losową χ^2.

Wykonana analiza pozwala stwierdzić, że strukturę zależności badanych cech stosunkowo dobrze opisuje model M_1 stałej zależności par zmiennych dla różnych kategorii trzeciej zmiennej, jak i dwa modele warunkowej niezależności: zmiennych B_1 i B_2 (model M_2) oraz B_2 i B_3 (model M_4). Wyraźny wzrost wartości G^2 następuje przy testowaniu dopasowania modelu M_3, co oznacza konieczność odrzucenia hipotezy o warunkowej niezależności zmiennych B_1 i B_3. Ostateczny wniosek z analizy można wypowiedzieć na podstawie

T a b l i c a 2. Wyniki testowania dopasowania nienasyconych modeli logarytmiczno-liniowych trójwymiarowej tablicy kontyngencji (objaśnienia w tekście)

Model	Grupa parametrów							Liczba stopni swobody	Wartość statystyki G^2	Poziom istotności
	u	u_1	u_2	u_3	u_{12}	u_{13}	u_{23}			
M_1	+	+	+	+	+	+	+	2	7.21	0.027
M_2	+	+	+	+		+	+	4	9.87	0.043
M_3	+	+	+	+	+		+	4	53.65	0.000
M_4	+	+	+	+	+	+		3	7.64	0.054
M_5	+	+	+	+	+			5	53.66	0.000
M_6	+	+	+	+		+		5	9.87	0.079
M_7	+	+	+	+			+	6	55.88	0.000
M_8	+	+	+	+				7	55.89	0.000

testowania modeli M_5, M_6, M_7. Podane wyniki wskazują, że uzasadniony jest opis badanej tablicy trójwymiarowej modelem nienasyconym M_6

$$\log m_{ijk} = u + u_1 + u_2 + u_3 + u_{13},$$

a więc, że jedyną zależnością istotną statystycznie jest zależność pomiędzy pokrojem rośliny w stadium krzewienia a owłosieniem dokłosia.

P r z y k ł a d 7.2. W doświadczeniu wzięto pod uwagę trzy populacje żyta, które oznaczać będziemy symbolami Π_1, Π_2 i Π_3. Z każdej populacji pobrano próbę o liczebności 150 roślin, po czym rośliny opisano pod względem trzech cech jakościowych dotyczących liścia flagowego. Były to: A_1 - charakter liścia flagowego (3), A_2 - ustawienie na początku kłoszenia (2), A_3 - ustawienie w pełni kłoszenia (2); cyfry w nawiasach podają liczby kategorii cech. Obserwacje zamieszczamy w tablicy 3.

T a b l i c a 3 Czterowymiarowa tablica kontyngencji; obserwowane liczebności roślin pochodzących z trzech populacji żyta i sklasyfikowanych ze względu na trzy cechy jakościowe liścia flagowego

Populacja		Π_1		Π_2		Π_3	
Cecha		A_3		A_3		A_3	
A_1	A_2	1	2	1	2	1	2
1	1	25	5	4	0	2	4
	2	9	15	2	2	4	14
2	1	33	12	58	7	22	13
	2	8	30	23	27	9	43
3	1	7	2	21	3	14	2
	2	0	4	2	1	2	21

Zgodnie z uwagą podaną w punkcie 5 otrzymana tablica kontyngencji jest czterowymiarowa o rozmiarach 3 x 2 x 2 x 3, przy czym A_4 oznacza zmienną opisującą przynależność elementów próby do jednej z trzech populacji. Ponieważ ze względu na cel eksperymentu najbardziej interesujące było zba-

danie zależności cech A_2 i A_3, analizę statystyczną ograniczono do testowania hipotez wymienionych w tablicy 4. Uzyskane wyniki pozwoliły przeprowadzić następujące wnioskowanie.

T a b l i c a 4. Wyniki testowania hipotez dotyczących parametrów modelu logarytmiczno-liniowego czterowymiarowej tablicy kontyngencji (objaśnienia w tekście)

Hipoteza	Treść hipotezy	Liczby stopni swobody	Wartość statystyki G^2	Poziom istotności
H_1	$u_{1234}= 0$	4	6.34	0.175
H_2	$u_{123}= u_{1234}= 0$	6	10.88	0.092
H_3	$u_{234}= u_{1234}= 0$	6	6,96	0.325
H_4	$u_{123}= u_{234}=u_{1234}= 0$	8	10.95	0.205
H_5	$u_{23}= u_{123}= u_{234}=u_{1234}= 0$	9	121.65	0.000

Stwierdzono, że na poziomie istotności 0,05 nie ma podstaw do odrzucenia hipotezy H_1 mówiącej o braku zależności trzeciego rzędu. Fakt ten świadczy, zgodnie z interpretacją podaną w punkcie 5, o stałej zależności drugiego rzędu każdej trójki cech przy różnych kategoriach czwartej cechy. Dalsza analiza wskazuje na brak podstaw do odrzucenia hipotez o braku zależności drugiego rzędu zmiennych A_1, A_2 i A_3 (hipoteza H_2) oraz A_2, A_3 i A_4 (H_3). Co więcej, dobre dopasowanie wykazuje model nienasycony odpowiadający hipotezie H_4, mówiącej o jednoczesnym braku wymienionych zależności. Wynika z tego, że zależność cech A_2 i A_3 jest taka sama dla różnych kategorii cech A_1 i A_4. Ostatnią ze sformułowanych hipotez, dotyczącą braku zależności cech A_2 i A_3 przy ustalonych kategoriach A_1 i A_4, należy odrzucić. Konkluzją powyższych rozważań może być stwierdzenie, że zależność pomiędzy ustawieniem liścia flagowego na początku (A_2) i w pełni kłoszenia (A_3) jest istotna statystycznie, taka sama w każdej z badanych populacji i nie ma na nią wpływu charakter liścia flagowego.

LITERATURA

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. J. Roy Stat. Soc. B 25, 220-233.

Bishop, Yvonne M.M., Fienberg, S.E. i Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge.

Brown, M.B. (1976). Screening effects in multidimensional contingency tables. Appl. Statist. 25, 37-46.

Darroch, J.N. (1962). Interactions in multi-factor contingency tables. J. Roy. Stat. Soc. B 24, 162-166.

Goodman, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13, 33-61.

Haberman, S.J. (1973). Log-linear models for frequency data: sufficient statistics and likelihood equations. Annals of Statistics 1, 617-632.

Havránek, T. (1984). A procedure for model search in multidimensional contingency tables. Biometrics 40, 95-100.

Imrey, P.B., Koch, C.G. i Stokes, Maura E. (1981). Categorical data analysis: some reflections on the log-linear model and logistic regression. Part 1: Historical and methodological overview. Int. Stat. Rev. 49, 265-283.

Imrey, P.B., Koch, C.G. i Stokes, Maura E. (1982). Categorical data analysis: some reflections on the log-linear model and logistic regression. Part 2: Data analysis. Int. Stat. Rev. 50, 35-63.

Plackett, R.L. (1981). The Analysis of Categorical Data (2nd edition). Griffin, London.