ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XI, 4 (1970)

J. MIKIEWICZ (Wrocław)

DENDRYTOWE OBSZARY UFNOŚCI

1. WPROWADZENIE

Praca niniejsza przedstawia ulepszenie metody statystycznej podanej w pracy [7]. Zawiera ona również przykład praktycznego zastosowania tej metody.

Taksonomia wrocławska (p. [4] i [5]) ma na celu graficzne przedstawienie wzajemnego podobieństwa rozważanego zbioru przedmiotów. Posługuje się w tym celu najkrótszym dendrytem, tj. najkrótszym połączeniem odcinkami wszystkich punktów reprezentujących badane przedmioty w przestrzeni cech. Dendryt daje się przedstawić na płaszczyźnie i ma odzwierciedlać wzajemne położenie punktów w wielowymiarowej przestrzeni.

Sens takiego dendrytu wiąże się z określeniem podobieństwa przyrodniczego między przedmiotami. Jeśli bowiem za miarę podobieństwa, lub ściślej za miarę niepodobieństwa, uznać odległość pary punktów reprezentujących dwa przedmioty w przestrzeni cech, to definicja odległości punktów będzie jednocześnie definicją podobieństwa przedmiotów i w istotny sposób wpłynie na topologiczną strukturę najkrótszego dendrytu.

Taksonomia wrocławska musi również uwzględnić sytuację, w której porządkowane przedmioty mają cechy zmieniające się losowo, np. gdy są to losowo wybrane reprezentanty populacyj przyrodniczych lub gdy pomiar wartości cechy obarczony jest błędem losowym. Taki przypadek omówiony jest w zakończeniu pracy. Warto może wspomnieć, że zdaniem profesora J. Różyckiego, który użyczył nam materiałów pomiarowych, potrzeba takiej metody już się zarysowała w dyskusjach naukowych na Politechnice Wrocławskiej. Interesuje nas w tych przypadkach najkrótszy dendryt łączący środki ciężkości populacyj.

Praca [7] podaje metodę przyporządkowania próbie losowej zbioru (rodziny) dendrytów, do którego należy prawdziwy najkrótszy dendryt międzypopulacyjny, z większym niż z góry dane prawdopodobieństwem.

Uważamy tutaj dendryty za *równe*, jeśli składają się z odcinków łączących te same wierzchołki. Praca [7] przyjmuje jako „odległość teoretyczną" (tj. międzypopulacyjną) wzór (3), str. 394, a także pewne założenia ograniczające klasę rozważanych populacyj, jak ich normalność i równość macierzy kowariancyj.

Praca niniejsza, przyjmując również formułę odległości teoretycznej (3), wprowadza nową formułę (4) „odległości empirycznej" tzn. z próby. Dzięki tej formule oraz twierdzeniom pomocniczym § 4, można z odległości empirycznych, dla populacyj klasy bardzo ogólnej, utworzyć ilorazy Studenta (12), a stąd znaleźć obszary ufności dla odległości teoretycznych.

Rozumowania podobne do przedstawionych w [7] prowadzą nas do podstawowego wzoru (18), będącego niewielką modyfikacją wzoru (43) z [7]. Sposób korzystania z tego wzoru w praktyce podajemy w § 5. Praca niniejsza daje więc uogólnienie metody z pracy [7] oraz — dla praktyków — opis jej stosowania.

Tu wspomnę jeszcze o dwóch zagadnieniach, które powstają w związku z rozwiązaniem, a z których drugie omawia szerzej § 3. Otóż zbiór wszystkich odległości między pewną liczbą punktów w dowolnej przestrzeni nazywamy *tablicą Czekanowskiego*. Wyznacza ona jednoznacznie najkrótszy dendryt (przypadek równych odległości, zachodzący dla odległości empirycznych z prawdopodobieństwem 0, wykluczamy z naszych rozważań). Jednakże zależy ona od przyjętej definicji odległości. W niniejszej pracy, podobnie jak w [7], przyjmujemy odległość

$$\delta^{(pq)} = \sum_{i=1}^{n} |x_i^{(p)} - x_i^{(q)}|,$$

gdzie $x_i^{(r)}$ oznacza wartość i-tej cechy u r-tego osobnika lub przedmiotu, który bierzemy pod uwagę. Wydaje się, że podobieństwo przyrodnicze cechuje przechodniość, którą można wyrazić matematycznie prawem trójkąta. Odległość powyższa jest szczególnym przypadkiem obszernej klasy odległości określonych wzorem

$$\delta_{M,t}^{(pq)} = \sqrt[t]{\sum_{i=1}^{n} |x_i^{(p)} - x_i^{(q)}|^t}, \qquad t \geqslant 1.$$

Nazwiemy je *odległościami Minkowskiego*, gdyż ze znanej nierówności (Minkowskiego) wynika dla nich prawo trójkąta:

$$\delta_{M,t}^{(pq)} \leqslant \delta_{M,t}^{(pr)} + \delta_{M,t}^{(qr)}$$

(por. [6], rozdz. VIII).

Można oczywiście przyjąć różne inne definicje odległości. I tak np. J. Czekanowski (p. [3], str. 179, i dalej) używał w swych pracach współ-

czynnika korelacji, obliczanego dla zbioru par uporządkowanych $(x_i^{(p)}, x_i^{(q)})$ do określenia odległości między osobnikami p-tym a q-tym. Takie określenie odległości w przestrzeni cech jest interesujące, jednakże z powodu związanych z nią trudności rachunkowych, nie będziemy jej rozważać w niniejszej pracy.

Z punktu widzenia naszych dalszych rozważań, spośród odległości Minkowskiego odległość $\delta_{M,1}^{(pq)}$ jest najdogodniejsza, gdyż dla niej najłatwiej jest uwzględnić statystyczne własności związanych z nią odległości empirycznych, którymi operujemy w naszej metodzie, a które są estymatorami odległości teoretycznych. To uzasadnia nasz wybór.

Metoda kostkowa prowadzi do wzorów rachunkowo prostszych niż metoda Hotellinga zastosowana do tablicy Czekanowskiego, a mniej jest rozrzutna niż metoda uogólnionej nierówności Czebyszewa (patrz § 3).

2. ZAGADNIENIE I JEGO ROZWIĄZANIE

2.1. Najkrótszy dendryt. Podamy tutaj formalną definicję dendrytu, która algebraizuje pojęcie równości topologicznej dendrytów, omawiane już w [7].

Niech $\mathscr{M}$ będzie skończonym zbiorem elementów $a_1, a_2, \ldots, a_m$, $\mathscr{U}$ zaś zbiorem wszystkich nieuporządkowanych par $u = (a_p, a_q)$ elementów zbioru $\mathscr{M}$. Parę (a_p, a_q) będziemy również nazywali *odcinkiem*, a elementy a_p i a_q *końcami* tego odcinka. Dowolny podzbiór $\mathscr{G}$ zbioru $\mathscr{U}$ nazywamy *grafem*. O grafie $\mathscr{G}$ mówimy, że jest *rozpięty* na zbiorze końców odcinków składających się na niego, a końce te nazywać będziemy *wierzchołkami* grafu. Graf $\mathscr{G}$ nazywać będziemy *spójnym*, gdy dla każdych dwóch jego wierzchołków a_p i a_q istnieje w $\mathscr{G}$ łańcuch wiążący a_p z a_q czyli gdy istnieje zbiór takich odcinków $(b_1, b_2), (b_2, b_3), \ldots, (b_{k-1}, b_k)$ należących do $\mathscr{G}$, że $a_p = b_1$ i $a_q = b_k$. Łańcuch złożony z różnych odcinków i taki, że $b_1 = b_k$, nazywa się *cyklem*. Graf spójny bez cykli nazywa się *dendrytem*.

Niech z każdą parą $u = (a_p, a_q)$ związana będzie nieujemna liczba $\delta(u) = \delta^{pq}$, zwana także dalej długością odcinka u. Przyjmujemy przy tym, że $\delta^{pp} = 0$ dla $p = 1, 2, \ldots, m$. Tablica tych odległości bywa nazywana tablicą Czekanowskiego. *Długością* grafu $\mathscr{G}$ nazywamy liczbę $\sum_{u \in \mathscr{G}} \delta(u)$.

W literaturze rozpatrywane było zagadnienie wyznaczania na podstawie tablicy Czekanowskiego najkrótszego grafu spójnego, rozpiętego na $\mathscr{M}$ (zob. [4] i [9]). Wiadomo, że najkrótszy graf spójny rozpięty na $\mathscr{M}$ jest dendrytem, a więc *najkrótszy graf spójny rozpięty na $\mathscr{M}$ jest najkrótszym dendrytem rozpiętym na $\mathscr{M}$*. Znane są dwa algorytmy wyznaczania najkrótszego dendrytu rozpiętego na $\mathscr{M}$: jeden opisany w [4] i drugi

zaproponowany przez M. Warmusa (por. [7] s. 27). Oba algorytmy są również opisane w pracy [9].

2.2. Zagadnienie niniejszej pracy. Niech teraz elementy zbioru $\mathcal{M}$ o liczności m ($m > 1$) będą populacjami statystycznymi, z których każda scharakteryzowana jest przez pewien rozkład prawdopodobieństwa n cech ($n \geqslant 1$). Przyjmujemy, że cechy te zostały już uprzednio w pewien sposób unormowane, np. na średnią lub dyspersję (tzn. przez podzielenie danej cechy we wszystkich populacjach przez średnią średnich albo — w drugim przypadku — średnią dyspersji populacyjnych; p. [7], § 3, a także str. 31). Załóżymy także, że populacje te dają się podzielić na l warstw każda ($l > 1$). Element próby z j-tej warstwy r-tej populacji jest wektorem losowym o składowych $X_{1j}^r, \ldots, X_{nj}^r$.

Piszemy $\mu_{ij}^r = EX_{ij}^r$, czyli że μ_{ij}^r jest średnią i-tej cechy w j-tej warstwie r-tej populacji.

Zakładać będziemy, że dla wszelkich par p, q, gdzie $p, q = 1, 2, \ldots, m$, mamy (pionowa kreska oznacza uporządkowanie pary wskaźników)

$$v_i^{p|q} = \mu_{ij}^p - \mu_{ij}^q = \text{const} \tag{1}$$

dla $j = 1, 2, \ldots, l$ oraz

$$\lambda_{\alpha\beta}^r = E(X_{\alpha j}^r - \mu_{\alpha j}^r)(X_{\beta j}^r - \mu_{\beta j}^r) = \text{const} \tag{2}$$

dla $j = 1, 2, \ldots, l$ $(\alpha, \beta = 1, 2, \ldots, n)$.

A więc, mniej dokładnie, zakładamy tu równe przesunięcie, bez rozciągania, w odpowiednich warstwach wszystkich populacyj.

Wprowadzony tu model warstwowy może mieć dwojaki sens praktyczny. Z jednej strony, gdy mamy do czynienia z jednolitymi populacjami, podział na warstwy ma charakter czysto teoretyczny i jego celem jest wtedy ustalenie obiektywnej zasady wzajemnego przyporządkowania sobie elementów prób z różnych populacyj w regule postępowania wyrażonej wzorami (4) i (5). Stąd, w zgodzie z (1) i (2), podział populacyj na warstwy może być wówczas dokonany w oparciu o dowolną cechę niezależną od cech rozpatrywanych; a więc zasada wzajemnego przyporządkowania wyrażona w (4) może być loteryjna. Wówczas oczywiście warstwy danej populacji nie różnią się między sobą i założenia (1) i (2) są spełnione automatycznie. Z drugiej strony, warstwy mogą być odrębnymi populacjami. Widzimy to np. w przykładzie podanym w § 5, gdzie jest mowa o kamieniołomach, w których poziome warstwy skalne mają zwykle nieco różne własności. W tym wypadku założenia (1) i (2) są najprostszymi założeniami o własnościach tych populacyj. Założenia bardziej ogólne wymagałyby dalszej rozbudowy teorii.

Odległość między populacją p-tą i q-tą określamy przez

$$\delta^{pq} = \sum_{i=1}^{n} |v_i^{p|q}|. \tag{3}$$

Tak określone δ^{pq} nazywać będziemy w dalszym ciągu *odległościami teoretycznymi*, a cały ich układ *teoretyczną tablicą Czekanowskiego*. Odległości δ^{pq} w rozpatrywanym przez nas zadaniu są ustalonymi liczbami. Interesuje nas odpowiadający im najkrótszy dendryt Δ rozpięty na populacjach $a_1, \ldots, a_m$. Dla jednoznaczności dendrytu Δ wystarczy, jeśli długości różnych odcinków są różne (można nawet osłabić nieco ten warunek). W dalszym ciągu zakładać będziemy, że dendryt Δ jest wyznaczony jednoznacznie.

Odległości δ^{pq} w praktyce nie znamy, a o dendrycie Δ możemy wnioskować tylko na podstawie próby. W tym celu posłużymy się odległościami empirycznymi, tj. obliczonymi na podstawie próbek. Weźmy mianowicie pod uwagę próbki z populacyj, zawierające po jednym elemencie z każdej warstwy każdej z tych populacyj. Wartość i-tej cechy takiego elementu oznaczmy przez X_{ij}^r. Otrzymany w ten sposób zbiór elementów próbkowych będziemy dalej nazywać *próbą łączną* i oznaczać przez $\mathfrak{X}$.

Statystykę

$$D_j^{pq} = \sum_{i=1}^{n} |X_{ij}^p - X_{ij}^q| \tag{4}$$

będziemy nazywać *warstwową empiryczną odległością* populacyj a_p i a_q, obliczoną na podstawie elementów próby z j-tych warstw tych populacyj, a statystykę

$$\bar{D}^{pq} = \sum_{j=1}^{l} D_j^{pq} \tag{5}$$

po prostu *empiryczną odległością* populacyj a_p i a_q. Z naszych założeń wynika, że statystyki D_j^{pq} mają jednakowe pierwsze i drugie momenty dla $j = 1, 2, \ldots, l$. Zauważmy przy okazji, że postępowanie to jest podobne do metody zmiennych połączonych, stosowanej w doświadczalnictwie (por. [8], str. 138-140).

Definicja (5) odległości empirycznych jest konwencjonalna. Przyjmujemy ją jednak ze względu na jej prostotę analityczną, a także inne korzystne własności, które omówimy na początku § 3.

2.3. Bliższe sformułowanie zagadnienia. Tablica Czekanowskiego, zarówno teoretyczna, jak i empiryczna, jest symetryczna i ma na głównej przekątnej same zera. Jest ona w zupełności wyznaczona przez $k = \frac{1}{2}m(m-1)$ elementów położonych powyżej głównej przekątnej. Jeśli pola tablicy Czekanowskiego nad główną przekątną ponumerujemy kolejno od 1 do k, to tablicę Czekanowskiego możemy identyfikować z określonym punktem k-wymiarowej przestrzeni euklidesowej $\mathscr{E}^k$, a mianowicie

tablicę Czekanowskiego odległości teoretycznych δ^{pq} będziemy identyfikować z punktem

$$(6) \qquad (\delta^{12}, \ldots, \delta^{1m}, \delta^{23}, \ldots, \delta^{2m}, \ldots, \delta^{m-1,m}) = (\delta_1, \ldots, \delta_k) \in \mathscr{E}^k.$$

Podobnie tablicę odległości empirycznych $\bar{D}^{pq}$ będziemy identyfikować z wektorem losowym

$$(\bar{D}^{12}, \ldots, \bar{D}^{1m}, \bar{D}^{23}, \ldots, \bar{D}^{2m}, \ldots, \bar{D}^{m-1,m}) = (\bar{D}_1, \ldots, \bar{D}_k) \in \mathscr{E}^k.$$

O tym, z jakich odcinków składa się dendryt Δ, decyduje pewna liczba porównań długości tych odcinków. *Algorytm Warmusa*, który te porównania w pewien sposób systematyzuje, polega na tym, że wybieramy dowolny element zbioru $\mathscr{M}$ i jako pierwszy odcinek budowanego dendrytu Δ bierzemy najkrótszy odcinek, którego jednym końcem jest wybrany element. Następnie do już zbudowanej części dendrytu Δ dołączamy najkrótszy odcinek, którego tylko jeden koniec jest wierzchołkiem już zbudowanej części dendrytu Δ. Postępowanie to powtarzamy aż do powstania dendrytu rozpiętego na całym zbiorze $\mathscr{M}$.

Gdy dana jest teoretyczna tablicy Czekanowskiego, algorytm ten pozwala skonstruować dendryt Δ. Z drugiej strony, jeżeli dany jest dowolny dendryt Δ, algorytm ten pozwala zobaczyć, jakie warunki musi spełniać tablica Czekanowskiego, by ten dendryt okazał się najkrótszy.

Na to, by Δ był najkrótszym dendrytem rozpiętym na zbiorze $\mathscr{M}$ potrzeba i wystarcza, żeby elementy $\delta^{pq} = \delta_\varkappa$ $(1 \leqslant \varkappa \leqslant k)$ tablicy Czekanowskiego spełniały pewien układ nierówności. Dla dendrytu Δ rozpiętego na $\mathscr{M}$ oznaczmy przez $\mathscr{B}_\Delta$ zbiór tych wszystkich tablic Czekanowskiego $(\delta_1, \ldots, \delta_k) \in \mathscr{E}^k$, dla których Δ okazuje się najkrótszym dendrytem.

Zbiór $\mathscr{B}_\Delta$ jest wypukłym stożkiem, tzn. takim zbiorem, który wraz z punktem $(\delta_1, \ldots, \delta_k)$ zawiera punkt $(\varepsilon\delta_1, \ldots, \varepsilon\delta_k)$ dla dowolnego $\varepsilon > 0$. Zbiorów $\mathscr{B}_\Delta$ jest oczywiście tyle, ile jest różnych dendrytów rozpiętych na zbiorze $\mathscr{M}$. Mnogościowa suma wszystkich zbiorów $\mathscr{B}_\Delta$ jest zbiorem $\mathscr{A}$ wszystkich punktów przestrzeni $\mathscr{E}^k$ o nieujemnych współrzędnych.

Zbiór $\mathscr{A}_{mn}$ wszelkich możliwych tablic Czekanowskiego, jakie możemy otrzymać zgodnie z wzorem (3), jest częścią właściwą zbioru $\mathscr{A}$, gdyż konsekwencją wzoru (3) jest m. in. warunek trójkąta, będący istotnym ograniczeniem. Zbiór $\mathscr{A}_{mn}$ jest również stożkiem.

Zgodnie z podanymi definicjami warunek $(\delta_1, \ldots, \delta_k) \in \mathscr{B}_\Delta$, jest równoważny z tym, że Δ jest najkrótszym dendrytem rozpiętym na $\mathscr{M}$. Nasze zagadnienie możemy więc sformułować inaczej w sposób następujący: Mamy zdefiniować przepis na przyporządkowywanie próbom losowym sum mnogościowych pewnej liczby wielościanów $\mathscr{B}_\Delta$ tak, aby można było oszacować od dołu prawdopodobieństwo tego, że — orze-

kając według tego przepisu, iż przyporządkowany próbie $\mathfrak{X}$ zbiór $\mathfrak{C}_{\mathfrak{X}}$ pokrywa teoretyczną tablicę Czekanowskiego — otrzymamy orzeczenie zgodne ze stanem faktycznym. Zbiorowi $\mathfrak{C}_{\mathfrak{X}}$ odpowiada oczywiście wzajemnie jednoznacznie pewna rodzina $\mathfrak{R}_{\mathfrak{X}}$ dendrytów rozpiętych na $\mathcal{M}$.

2.4. Rozwiązanie w przypadku populacyj normalnych i dostatecznie wzajemnie odległych. Załóżmy obecnie, że w rozpatrywanych przez nas populacjach cechy mają łączny rozkład normalny.

Niech

$$\lambda_{\alpha\beta}^{pq} = \lambda_{\alpha\beta}^{p} + \lambda_{\alpha\beta}^{q}, \qquad \alpha, \beta = 1, 2, \ldots, n, \tag{7}$$

gdzie składniki sumy są określone wzorem (2). Przyjmijmy także, że

$$D_{ij}^{pq} = |X_{ij}^{p} - X_{ij}^{q}|, \qquad \theta_i^{pq} = ED_{ij}^{pq}, \tag{8}$$

$$\theta^{pq} = \sum_{i=1}^{n} \theta_i^{pq} = ED^{pq}.$$

W pracy [7] (lemat 2.2) pokazano, że jeżeli rozkład prawdopodobieństwa zmiennej losowej Y jest normalny z wartością oczekiwaną a i dyspersją b, to

$$E|Y| = |a| + 2\psi(|a|, b) > |a|,$$

gdzie

$$\psi(a, b) = b\varphi\left(\frac{a}{b}\right) - a\Phi\left(-\frac{a}{b}\right),$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(x) = \int_{-\infty}^{x} \varphi(t)\, dt.$$

Wobec tego mamy

$$\theta_i^{pq} = |v_i^{p|q}| + 2\psi(|v_i^{p|q}|, \sqrt{\lambda_{ii}^{pq}}) > v_i^{pq}, \tag{9}$$

gdzie $v_i^{p|q} = \mu_{ij}^{p} - \mu_{ij}^{q}$ dla $j = 1, 2, \ldots, l$.

Stąd wynika, że D_j^{pq} jest obciążonym estymatorem odległości teoretycznej $\delta^{pq} = \sum_i |v_i^{p|q}|$, a mianowicie średnio biorąc przecenia δ^{pq}.

Z przyjętych założeń oraz z własności rozkładu normalnego wynika, że jeśli

$$|v_i^{p|q}| / \sqrt{\lambda_{ii}^{pq}} \geqslant 3, \tag{10}$$

to zmienne losowe D_{ij}^{pq} mają dla $j = 1, 2, \ldots, l$ rozkłady prawdopodobieństwa jednakowe i bardzo bliskie rozkładowi normalnemu. Wynika stąd natychmiast również jedno z założeń twierdzenia II (które sfor-

mułujemy i udowodnimy w § 4) wyrażające się nierównością

$$(n-1)\,\delta_{\varkappa}^2 \geqslant \sum_{i=1}^{n} \lambda_{ii,\varkappa}. \tag{11}$$

Jeszcze bliższe normalności i jednakowe są wtedy rozkłady prawdopodobieństwa statystyk D_j^{pq}, o ile tylko nie zachodzi osobliwość rozkładów prawdopodobieństwa jednocześnie obu wektorów losowych $(X_{1j}^p, X_{2j}^p, \ldots, X_{nj}^p)$ i $(X_{1j}^q, X_{2j}^q, \ldots, X_{nj}^q)$. Wobec tego iloraz

$$\Xi_{\varkappa} = \frac{\bar{D}_{\varkappa} - \theta_{\varkappa}}{S_{\varkappa}}\, v, \tag{12}$$

gdzie $S_{\varkappa}^2 = l^{-1}\sum (D_{j\varkappa} - \bar{D}_{\varkappa})^2$ oraz $v^2 = l-1$, ma rozkład prawdopodobieństwa praktycznie nie różniący się od rozkładu Studenta z v^2 stopniami swobody.

Przyjmijmy poziom ufności $1-\alpha$, gdzie $\alpha > 0$. Możemy wówczas znaleźć taką liczbę $c > 0$, że $P(|\Xi_{\varkappa}| > c) = \alpha_{\varkappa}$ i $\sum_{\varkappa=1}^{k} \alpha_{\varkappa} = \alpha$. Na mocy lematu 1 z § 4 będziemy wówczas mieli nierówność

$$P(|\Xi_1| \leqslant c, |\Xi_2| \leqslant c, \ldots, |\Xi_k| \leqslant c) \geqslant 1 - \sum_{\varkappa=1}^{k} \alpha_{\varkappa}. \tag{13}$$

Stąd, biorąc pod uwagę parę nierówności $|\Xi_{\varkappa_1}| \leqslant c$, $|\Xi_{\varkappa_2}| \leqslant c$, otrzymamy dla $\theta_{\varkappa_1}$ i $\theta_{\varkappa_2}$ prostokąt ufności opisany przez następujące nierówności:

$$\begin{aligned} \bar{D}_{\varkappa_1} - c\frac{S_{\varkappa_1}}{v} &\leqslant \theta_{\varkappa_1} \leqslant \bar{D}_{\varkappa_1} + c\frac{S_{\varkappa_1}}{v}, \\ \bar{D}_{\varkappa_2} - c\frac{S_{\varkappa_2}}{v} &\leqslant \theta_{\varkappa_2} \leqslant \bar{D}_{\varkappa_2} + c\frac{S_{\varkappa_2}}{v}. \end{aligned} \tag{14}$$

Przypuśćmy teraz, że cały ten prostokąt leży wewnątrz obszaru

$$\theta_{\varkappa_1} > \theta_{\varkappa_2} + g_{\varkappa_1\varkappa_2}\sqrt{\frac{2(n-1)}{\pi-2}}, \tag{15}$$

gdzie dla dowolnej pary wskaźników $\varkappa_1, \varkappa_2$ określamy

$$g_{\varkappa_1\varkappa_2} = \max(g_{\varkappa_1}, g_{\varkappa_2}),$$

a $g_{\varkappa}$ są dla dowolnego $\varkappa$ określone wzorem

$$g_{\varkappa} = S_{\varkappa}\sqrt{\frac{l}{y_{l-1}(\beta_{\varkappa})}}, \tag{16}$$

w którym $y_{l-1}(\beta_\varkappa)$ jest taką liczbą, że dla zmiennej losowej χ^2_{l-1} o rozkładzie chi-kwadrat z $l-1$ stopniami swobody mamy

$$P[\chi^2_{l-1} < y_{l-1}(\beta_\varkappa)] = \beta_\varkappa. \tag{17}$$

Przy uwzględnieniu (14), nierówność (15) jest spełniona wówczas, gdy

$$T_{\varkappa_1\varkappa_2} = \frac{\bar{D}_{\varkappa_1} - \bar{D}_{\varkappa_2} - g_{\varkappa_1\varkappa_2}\sqrt{2(n-1)/(\pi-2)}}{S_{\varkappa_1}+S_{\varkappa_2}}\, v \geqslant c. \tag{18}$$

Rola składnika zawierającego $g_{\varkappa_1\varkappa_2}$ jest następująca: Jak wiemy z (9), $\theta_\varkappa$ są większe od $\delta_\varkappa$. Gdy więc dwie wartości $\theta_{\varkappa_1}$ i $\theta_{\varkappa_2}$ są bliskie, zachodzi możliwość, że jednocześnie $\theta_{\varkappa_1} > \theta_{\varkappa_2}$ i $\delta_{\varkappa_1} < \delta_{\varkappa_2}$. Niech

$$\sigma^2_{i,\varkappa} = E(D_{ij,\varkappa} - \theta_{i,\varkappa})^2, \tag{19}$$
$$\sigma^2_\varkappa = E(D_{j,\varkappa} - \theta_\varkappa)^2.$$

W celu wyciągania należytych wniosków co do nierówności zachodzących między odległościami teoretycznymi $\delta_\varkappa$, posłużymy się twierdzeniem II (p. str. 410). W założeniu tego twierdzenia występują trzy warunki, które — spełnione dla pary nierówności empirycznych o wskaźnikach $\varkappa_1$ i $\varkappa_2$ — pociągają za sobą nierówność $\delta_{\varkappa_1} > \delta_{\varkappa_2}$ bądź $\delta_{\varkappa_1} < \delta_{\varkappa_2}$. Z tych warunków nierówność (11) wynika bezpośrednio z (10) i uważamy w związku z tym, że jest spełniona; nieco dalej powiemy o jej weryfikowaniu w praktyce. Nierówności $\sum\sigma^2_{i,\varkappa} \leqslant g^2_\varkappa$ otrzymujemy zakładając, że cechy są nieskorelowane i dobierając liczby $g_\varkappa$ z wzoru (16) w ten sposób, by stosując również tu rozumowanie opisane przy wzorze (13), otrzymać nierówność

$$P(\sigma_1 \leqslant g_1, \sigma_2 \leqslant g_2, \ldots, \sigma_k \leqslant g_k) \geqslant 1 - \sum_{\varkappa=1}^{k} \beta_\varkappa. \tag{20}$$

Na poziomie zatem ufności $1-\Sigma\beta_\varkappa$ zachodzą łącznie wszystkie nierówności

$$\sum_{i=1}^{n} \sigma^2_{i,\varkappa} \leqslant g^2_\varkappa,$$

dla wszelkich $\varkappa = 1, 2, \ldots, k$. Liczby $g_\varkappa$ dobieramy przy tym, zgodnie z (16), możliwie najmniejsze, z uwagi na nierówność (18), z której wynika potrzebna nam nierówność (15), będąca trzecim założeniem twierdzenia II.

Zastosujemy teraz lemat 1 zarówno do prawdopodobieństwa $\alpha = \sum_\varkappa \alpha_\varkappa$, jak i $\beta = \sum_\varkappa \beta_\varkappa$. Wybierając liczby $g_\varkappa$ według wzoru (16) badamy, czy spełniona jest nierówność (18) i w przypadku pomyślnym uważamy nie-

równość (15) zachodzącą między parametrami θ_{x_1} i θ_{x_2} za istotną na poziomie ufności co najmniej $1-\alpha-\beta$. Uważamy wówczas za spełnione na tym poziomie ufności wszystkie założenia twierdzenia II, a wobec tego uważamy za spełnioną i tezę, czyli uważamy nierówność między odległościami δ_{x_1} i δ_{x_2} za istotną na poziomie ufności co najmniej $1-\alpha-\beta$.

Otrzymany tą drogą układ orzeczeń o nierównościach zachodzących między odległościami teoretycznymi wyznacza rodzinę $\mathfrak{C}_{\mathfrak{x}}$ wielościanów $\mathscr{B}_{\Delta}$, zawierającą z ufnością co najmniej $1-\alpha-\beta$ teoretyczną tablicę Czekanowskiego, a wraz z $\mathfrak{C}_{\mathfrak{x}}$ wyznaczona jest rodzina ufności $\mathfrak{R}_{\mathfrak{x}}$.

Praktycznie rodzinę $\mathfrak{R}_{\mathfrak{x}}$ wyznaczamy budując najkrótszy dendryt rozpięty na zbiorze $\mathscr{M}$ (np. metodą Warmusa), uwzględniając wszystkie warianty, do jakich prowadzą orzeczenia typu $\delta_{x_1} < \delta_{x_2}$ lub $\delta_{x_1} > \delta_{x_2}$ Oczywiście, jeśli nie wystąpi żadne orzeczenie tego typu, otrzymamy jeden tylko dendryt, natomiast w drugim skrajnym przypadku, gdy wszystkie orzeczenia są alternatywne — rodzina $\mathfrak{R}_{\mathfrak{x}}$ składa się z wszystkich możliwych dendrytów rozpiętych na zbiorze $\mathscr{M}$.

Tak więc na podstawie łącznej próby $\mathfrak{X}$ otrzymujemy dendrytową rodzinę ufności $\mathfrak{R}_{\mathfrak{x}}$ na poziomie ufności co najmniej $1-\alpha-\beta$, co zapiszemy wzorem

$$P(\Delta \in \mathfrak{R}_{\mathfrak{x}}) \geqslant 1-\alpha-\beta. \tag{21}$$

Oszacowanie to jest asymptotycznie dokładnym rozwiązaniem zadania, które sobie postawiliśmy, jeśli założymy podzielność każdej z populacyj na dowolną ilość warstw, spełniających odpowiednio warunki (1) i (2), oraz możność pobrania jednoelementowej próbki z każdej warstwy (patrz wyjaśnienie dot. warunków (1) i (2)). Gdy bowiem liczba warstw $l \to \infty$, statystyki występujące we wzorze (18) zbliżają się do swych średnich, a współczynnik v dąży wówczas wraz z l do nieskończoności, dzięki czemu do nieskończoności dąży cały iloraz $T_{x_1 x_2}$. Pozwala to zwiększać nieograniczenie parametr c, przy nie zwiększających się liczbach g_x zgodnie z wzorem (16). Dzięki temu w granicy prawdopodobieństwo po prawej stronie wzoru (21) równe jest jedności.

Podobnie możemy zwiększyć poziom ufności (21), z jakim otrzymamy wynik, przez powiększenie liczebności w *wiązkach*, które omówimy w punkcie 2.6.

2.5. Weryfikowanie założeń. W przedstawionym wyżej rozwiązaniu korzystamy z założeń (10), które w praktyce musimy weryfikować statystycznie, oraz przy wykorzystaniu wzoru (16) wnioskujemy o wielkości $\sum_{i=1}^{n} \sigma_{i,x}^2$ na podstawie statystyk S_x^2, zawierających wariancje i kowariancje branych pod uwagę składowych. Przypominamy, że założenia (10) przyjmujemy, chcąc otrzymać ilorazy (12), mające z praktyczną dokładnością rozkłady Studenta, a także jako warunki zachodzenia nie-

równości (11). W celu weryfikacji tych założeń, możemy w oparciu o lemat 2 w § 4 posłużyć się jednostronnymi przedziałami ufności dla stosunku $|v_i^{p|q}|/\sqrt{\lambda_{ii}^{pq}}$, określonymi przy pomocy wzoru

$$\frac{|v_i^{p|q}|}{\sqrt{\lambda_{ii}^{pq}}} \geqslant \frac{|\bar{X}_i^{p|q}|}{S_{i,pq}} \sqrt{y_{l-1}(\eta')/l} + x\left(\frac{\eta}{2}\right)/\sqrt{l}, \tag{22}$$

gdzie

$$X_{ij}^{p|q} = X_{ij}^{p} - X_{ii}^{q}, \qquad \bar{X}_i^{p|q} = l^{-1} \sum_j X_{ij}^{p|q},$$

natomiast

$$S_{i,pq}^2 = l^{-1} \sum_{j=1}^{l} (X_{ij}^{p|q} - \bar{X}_i^{p|q})^2.$$

Liczbę $y_{l-1}(\eta')$ określamy podobnie jak w (17), rozpatrując zmienną losową chi-kwadrat z $l-1$ stopniami swobody,

$$P[\chi_{l-1}^2 < y_{l-1}(\eta')] = \eta',$$

oraz podobnie określamy $x(\eta)$, rozpatrując zmienną losową normalną $N(0, 1)$.

Przedziały te pokrywają prawdziwą wartość stosunku $|v_i^{p|q}|/\sqrt{\lambda_{ii}^{pq}}$ z prawdopodobieństwem większym niż $(1-\eta)(1-\eta')$ (z uwagi na niezależność omawianych statystyk — patrz lemat 2). Dla $\eta = \eta' = 0{,}05$ mamy

$$(1-\eta)(1-\eta') \approx 0{,}90.$$

Wydaje się, że dla celów praktycznych taki wybór poziomu ufności jest wystarczający. Jeśli weźmiemy pod uwagę, że statystyka D_j^{pq} powstaje przez dodanie wielu składników, możemy praktycznie przyjąć, że ma ona rozkład normalny z dostatecznym przybliżeniem, gdy prawe strony we wzorze (22) są nie mniejsze niż 2 i kwestionować tylko te cechy, dla których tak nie jest. Ze względu na nierówność (11), która zachodzi nawet wówczas, gdy znacznie zmniejszymy prawą stronę (10), musimy uważać, by dla większych składowych prawe strony (22) były nie mniejsze niż 2.

Kryterium (22) służy zatem do wstępnej selekcji materiału pod względem dyskryminacyjności cech (por. przepis podany w § 5). O ile pewne cechy z punktu widzenia tego kryterium są niedyskryminacyjne, a chcemy je uwzględnić, możemy je uczynić dyskryminacyjnymi przez pobieranie odpowiednio licznych prób z każdej warstwy rozpatrywanych populacyj i utworzenie wiązek.

Co do wyboru liczb $g_\varkappa$ według wzoru (16), należy zwracać uwagę by cechy były słabo skorelowane. Zgodnie bowiem z twierdzeniem I

i uwagami do tego twierdzenia, tylko wówczas można przyjąć, że statystyki

$$lS_{\varkappa}^2 \Big/ \sum_{i=1}^{n} \sigma_{i,\varkappa}^2$$

mają praktycznie rozkład chi-kwadrat z $l-1$ stopniami swobody i w związku z tym na przyjętym poziomie ufności możemy się spodziewać nierówności

$$\sum_{i=1}^{n} \sigma_{i,\varkappa} \leqslant g_{\varkappa}^2 \qquad (1 \leqslant \varkappa \leqslant k)$$

należącej do założeń twierdzenia II.

2.6. Uogólnienie rozwiązania. Gdy 1° wiemy, że rozkłady wektorów $(X_{1j}^r, \ldots, X_{nj}^r)$ dla $r = 1, 2, \ldots, m$ są normalne, lecz nie są dostatecznie odległe, wobec czego nie możemy się upewnić z pomocą kryterium (22), że statystyki D_j^{pq} są w przybliżeniu normalne lub gdy 2° wiemy, że rozkłady tych wektorów nie są normalne, możemy z każdej warstwy r-tej populacji pobrać u_r-elementowe próby. Taki τ-ty element, pobrany z j-tej warstwy r-tej populacji oznaczymy przez $(X_{1j\tau}^r, \ldots, X_{nj\tau}^r)$. Z elementów j-tej warstwy możemy utworzyć wiązkę

$$\bar{X}_{ij}^r = \frac{1}{u_r} \sum_{\tau=1}^{u_r} X_{ij\tau}^r. \tag{23}$$

Z takich wiązek dopiero będziemy tworzyć statystyki

$$D_j^{pq} = \sum_{i=1}^{n} |\bar{X}_{ij}^p - \bar{X}_{ij}^q|, \tag{24}$$

analogicznie jak w (4). Jeśli wektory $(\bar{X}_{1j}^p, \ldots, \bar{X}_{nj}^p)$ i $(\bar{X}_{1j}^q, \ldots, \bar{X}_{nj}^q)$ powstały z prób dostatecznie licznych, możemy uzyskać, na podstawie twierdzenia Lindeberga i Levy'ego dla zmiennych losowych wielowymiarowych (p. np. [1]), przybliżoną normalność łącznego rozkładu różnicy tych wektorów. Na tej samej drodze uzyskujemy następnie przybliżoną normalność statystyk $D_{j,\varkappa}$ wymaganą w ilorazie (12).

Rozkład wspomnianej różnicy wektorów losowych (wiązek) powinno się jeszcze sprawdzić z pomocą kryterium (22). Wówczas korzystając ze statystyk (24), tak jak poprzednio korzystaliśmy z D_j^{pq}, możemy już dalej postępować analogicznie, jak w rozwiązaniu szczególnym, podanym wyżej. Łatwo bowiem zauważyć, że otrzymawszy rozkłady bliskie normalnym, możemy stosować również omówione w § 4 lematy i twierdzenia.

3. DYSKUSJA ROZWIĄZANIA

Już we wstępie wspomnieliśmy o dwóch grupach zagadnień powstających w związku z powyższym rozwiązaniem. Pierwsza, dotycząca pojęcia odległości w przestrzeni cech, została tam omówiona nieco szerzej. Obecnie zajmiemy się grupą zagadnień probabilistycznych.

W rozwiązaniu przedstawionym w § 2 posłużyliśmy się odległościami empirycznymi (4), jako estymatorami odległości teoretycznych $\delta_{M,1}$, nie twierdząc przy tym, że są one optymalne ze statystycznego punktu widzenia. W [7] rozważa się jeszcze dwa inne estymatory odległości $\delta_{M,1}$: odległość „epsylonową" d^{pq} oraz „moduło-średniową" $\bar{d}^{pq}$ (1). Określamy je następująco (2):

$$
\begin{aligned}
d^{pq} &= \sum_{i=1}^{n} \varepsilon_i^{p|q}(\bar{X}_i^p - \bar{X}_i^q), \qquad \text{gdzie } \varepsilon_i^{p|q} = \operatorname{sign}(\mu_i^p - \mu_i^q), \\
\bar{d}^{pq} &= \sum_{i=1}^{n} |\bar{X}_i^p - \bar{X}_i^q|.
\end{aligned} \tag{25}
$$

Pierwsze momenty tych odległości wyrażają się następująco:

$$
\begin{aligned}
Ed^{pq} &= \delta^{pq} = \sum_i |\nu_i^{p|q}|, \\
E\bar{d}^{pq} &= \bar{\delta}^{pq} = \delta^{pq} + 2\sum_i \psi\left(|\nu_i^{p|q}|, \sqrt{\frac{\lambda_{ii}^p}{l_p} + \frac{\lambda_{ii}^q}{l_q}}\right).
\end{aligned} \tag{26}
$$

Lemat 3 pozwala porównać wariancje odległości $\bar{D}$ i $\bar{d}$. Wariancja d (patrz lemat 2.1 pracy [7]) jest

$$
E(d^{pq} - \delta^{pq})^2 = \omega_{pq}^2 = \frac{\omega_{p|q}^2}{l_p} + \frac{\omega_{q|p}^2}{l_q}, \qquad \text{gdzie } \omega_{p|q}^2 = \sum_{i_1 i_2} \varepsilon_{i_1}^{p|q} \varepsilon_{i_2}^{p|q} \lambda_{i_1 i_2}^p. \tag{27}
$$

Porównując omówione wzory stwierdzamy, iż d ma rozkład dokładnie normalny, lecz wymaga niedogodnych w praktyce założeń (równość macierzy kowariancyj, znajomość macierzy znaków $\varepsilon^{p|q}$ — patrz [7]). Odległość $\bar{d}$ ma na ogół mniejszą wariancję od obu pozostałych, jest asymptotycznie normalna i nieobciążona. Mimo tych zalet $\bar{d}$ nie nadają się do naszych celów, gdyż nie można z nich utworzyć ilorazu (12) o rozkładzie Studenta.

---

(1) W celu zachowania zgodności oznaczeń z pracą [7] odstąpiliśmy tu od zasady oznaczania wielkimi literami zmiennych losowych, a małymi — ich realizacji.

(2) Zakłada się pobranie różnolicznych prób, tzn. l_r-elementowej próby z r-tej populacji, $r = 1, 2, \ldots, m$. Stąd $\bar{X}^p$ i $\bar{X}^q$ oznaczają w [7] średnie z prób różnolicznych pobranych odpowiednio z populacyj p-tej i q-tej.

Przechodząc do dyskusji omawianej poprzednio metody kostkowej opartej o lemat 1, porównamy ją z *metodą elipsoidową* ([3]). Metoda ta opiera się na uogólnionej nierówności Czebyszewa. Jeśli X będzie oznaczać dowolny wektor losowy, a $\xi(X)$ nieujemną funkcję tego wektora, to ogólna postać nierówności Czebyszewa wyrazi się wzorem

$$P[\xi(X) \geqslant c] \leqslant E\xi(X)/c.$$

Podstawiając

$$\xi(X) = \sum_{\varkappa=1}^{k} \Xi_{\varkappa}^2,$$

gdzie $\Xi_{\varkappa}$ są określone wzorem (12), oraz c^2 zamiast c, otrzymamy stąd

$$P(R \geqslant c) \leqslant c^{-2} k \frac{l-1}{l-3}. \tag{28}$$

We wzorze tym $R = \sqrt{\Sigma \Xi_{\varkappa}^2}$ jest losowym promieniem sfery o środku w początku układu współrzędnych; po prawej stronie nierówności mamy sumę wariancyj rozkładów Studenta. Równość $R = c$ wyznacza nam granice obszaru ufności dla wektora θ (por. wzór (13)) na poziomie ufności określonym przez c. Obszar ten jest ograniczony elipsoidą

$$\sum_{\varkappa=1}^{k} \frac{(p_{\varkappa} - D_{\varkappa})^2}{S_{\varkappa}^2} = c^2.$$

Podobnie jak w metodzie poprzedniej, weźmiemy pod uwagę dowolną parę zmiennych losowych $\Xi_{\varkappa_1}$ i $\Xi_{\varkappa_2}$. Dla pary wartości oczekiwanych $\theta_{\varkappa_1}, \theta_{\varkappa_2}$ obszarem ufności jest elipsa. Gdy chcemy zatem wykonać dowolny krok metody, zakładając chwilowo, że interesują nas te wartości oczekiwane, możemy go wykonać na danym poziomie ufności, jeśli elipsa ta leży całkowicie po jednej stronie prostej $p_{\varkappa_1} = p_{\varkappa_2}$. Jeśli zatem ma być $\theta_{\varkappa_1} > \theta_{\varkappa_2}$ na danym poziomie ufności, to musi zajść nierówność będąca analogonem kryterium (40) pracy [7]:

$$\frac{\bar{D}_{\varkappa_1} - \bar{D}_{\varkappa_2}}{\sqrt{S_{\varkappa_1}^2 + S_{\varkappa_2}^2}} v \geqslant c. \tag{29}$$

Podobnie jak poprzednio, chcąc uzyskać kryterium odległości teoretycznych $\delta_{\varkappa_1}, \delta_{\varkappa_2}$ na podstawie twierdzenia II, zastępujemy nierówność $\theta_{\varkappa_1} > \theta_{\varkappa_2}$ przez

$$\theta_{\varkappa_1} > \theta_{\varkappa_2} + g_{\varkappa_1 \varkappa_2} \sqrt{\frac{2(n-1)}{\pi - 2}},$$

a stąd otrzymujemy kryterium analogiczne do (18).

([3]) O metodzie elipsoidowej w nieco innym sensie powiemy nieco dalej.

Porównajmy omawianą metodę z metodą kostkową. Mamy m populacyj, a więc k odległości (por. § 2), przy czym z każdej populacji wybieramy po l elementów. Jeśli przyjmiemy poziom ufności $1-\alpha$, to:

1. Dla nierówności Czebyszewa z wariancją Studenta (28) musi być

$$k\frac{l-1}{l-3}\,c^{-2} \leqslant \alpha.$$

2. Dla metody kostkowej mamy $\alpha_\varkappa = \alpha/k$ (gdy dla wszystkich odległości z osobna przyjmiemy równe prawdopodobieństwa), a więc $P(|\Xi_\varkappa| > c) = \alpha/k$.

Oznaczając dystrybuantę Studenta z $l-1$ stopniami swobody przez $s_{l-1}(x)$ otrzymujemy stąd

$$2ks_{l-1}(-c) \leqslant \alpha.$$

Załóżmy, że $S_{\varkappa_1} = S_{\varkappa_2}$. Mianownik w ilorazie (18) będzie więc większy $\sqrt{2}$ razy od mianownika w (29), czyli że osiągniemy ten sam próg biorąc w metodzie Czebyszewa $c/\sqrt{2}$, gdzie c użyliśmy w metodzie kostkowej. A więc ostatecznie porównujemy

$$2k\frac{l-1}{l-3}\,c^{-2} \quad \text{oraz} \quad 2ks_{l-1}(-c).$$

Ponieważ dla $l \geqslant 5$ jest $1 \leqslant (l-1)/(l-3) \leqslant 2$, wystarczy więc, że porównamy funkcje c^{-2} i $s_{l-1}(-c)$. Dla $c = 5$ mamy dla metody kostkowej $s_{l-1}(-c) < 0{,}01$, gdy $l > 2$, oraz $c^{-2} = 1/25 = 0{,}04$.

Dla rosnącego c i liczby stopni swobody $l-1 \geqslant 2$, $s_{l-1}(-c)$ maleją szybciej niż c^{-2}, co wynika ze wzoru na funkcje $s_{l-1}(x)$. Widzimy tu wielką przewagę metody kostkowej; należy jednak zwrócić uwagę, że w metodzie tej musimy znać rozkłady poszczególnych współrzędnych (tj. rozkłady Studenta), podczas gdy w metodzie elipsoidowej nie jest to konieczne, gdyż wystarczy tu znać wariancje rozkładów poszczególnych współrzędnych ([4]).

Zwrócimy jeszcze uwagę na metodę uogólnionego ilorazu Studenta, która pozwoliłaby nam wyznaczać najmniejszą rodzinę $\mathfrak{R}$ dendrytów, spełniającą warunek określony przez (21). Znajdując bowiem najmniejsze obszary ufności przy danym z góry prawdopodobieństwie, znajdujemy w ten sposób na tym poziomie ufności najmniejszą rodzinę dendrytów (pomijamy tu problem nierówności (15) zakładającej zbyt duży próg rozróżnialności). Założeniem koniecznym i wystarczającym stosowalności tej metody jest normalność empirycznej tablicy Czekanowskiego. Podobnie jak w § 2 (opis przy wzorze (23)) uzyskaliśmy przy bardzo ogólnych

([4]) Rozkłady przeto mogą tu być dowolne, byleśmy tylko umieli je unormować na wariancję (jak np. w [7], § 3).

założeniach przybliżoną normalność rozkładu pojedynczej odległości empirycznej, możemy, rozumując w ten sam sposób, uzyskać na podstawie twierdzenia Lindeberga i Levy'ego dla zmiennych losowych wielowymiarowych (patrz np. [1], § 24.7) przybliżoną normalność łącznego rozkładu zespołu odległości empirycznych. Znalezienie wzorów na momenty tego rozkładu, przy określonych założeniach co do zmiennych składowych, nie jest łatwe (w pracy [7] uzyskaliśmy te wzory dość prosto dla odległości „epsylonowych" d^{pq}). Ich znajomość nie jest wszakże potrzebna przy stosowaniu omawianej metody.

Obszarami ufności w tej metodzie będą również elipsoidy, przy czym parametr c będzie przeciwobrazem nałożonego prawdopodobieństwa podług rozkładu T^2 Hotellinga. Jej poważną wadą jest jednak ta okoliczność, że rozważana elipsoida jest położona dowolnie względem układu współrzędnych, wobec czego analogon kryteriów (18) i (29) wymaga rozwiązania równania sekularnego (lub zastosowania metod przybliżonych jego rozwiązania). Wobec tego metody omówione poprzednio, choć przybliżone, ale posiadające własności asymptotyczne i nie wymagające pracochłonnych rachunków, wydają się bardziej użyteczne w praktyce. Łatwiej jest bowiem niekiedy pomierzyć większą liczbę osobników z poszczególnych populacyj niż wykonywać skomplikowane i pracochłonne rachunki, nawet przy zastosowaniu maszyn cyfrowych.

Ostatecznie więc stwierdzimy ogólnie, że w zależności od warunków zadania jedna z omówionych metod może się okazać najkorzystniejszą.

4. LEMATY I TWIERDZENIA

Podane lematy i twierdzenia wykorzystaliśmy już w poprzednich paragrafach. Z wyjątkiem lematu 1 wszystkie one zakładają normalność rozkładów cech we wszystkich populacjach, a niektóre — normalność łącznego rozkładu cech. Tę normalność albo zakładamy (rozwiązanie szczegółowe opisane w punkcie 2.4), albo realizujemy w przybliżeniu przez tworzenie wiązek (uogólnienie rozwiązania w punkcie 2.6).

Lematu 3 nie wykorzystujemy bezpośrednio. Z twierdzenia I korzystamy w dowodzie lematu 3, z niego zaś w dowodzie twierdzenia II. To ostatnie twierdzenie jest podstawą metody opisanej w § 2. Oba twierdzenia są uogólnieniami wyników pracy [7] (odpowiednio twierdzeń 1 i 7).

LEMAT 1. *Jeśli* $Z_1, Z_2, \ldots, Z_k$ *oznaczają dowolne podzbiory mierzalne zbioru liczb rzeczywistych, a* $Z'_1, Z'_2, \ldots, Z'_k$ *ich dopełnienia, to dla dowolnych zmiennych losowych* $\Xi_1, \Xi_2, \ldots, \Xi_k$ *spełniona jest nierówność*

$$P(\Xi_1 \epsilon Z'_1, \Xi \epsilon Z'_2, \ldots, \Xi \epsilon Z'_k) \geqslant 1 - \sum_{\varkappa=1}^{k} \alpha_\varkappa,$$

gdzie $\alpha_\varkappa = P(\Xi_\varkappa \epsilon Z_\varkappa)$.

Lemat ten jest konsekwencją praw de Morgana. W pracy [7] nosi on nazwę twierdzenia 6 i jest tam udowodniony metodami analizy matematycznej.

LEMAT 2. *Niech*

$$\bar{X} = \frac{1}{l}\sum_{j=1}^{l} X_j$$

oznacza średnią z l-elementowej próby losowej z populacji normalnej $N(\mu, \sqrt{\lambda})$, *a*

$$S^2 = \frac{1}{l}\sum_{j=1}^{l}(X_j - \bar{X})^2$$

wariancję z tejże próby. Niech $x(\eta)$ *oznacza przeciwobraz prawdopodobieństwa* η *względem unormowanej dystrybuanty normalnej, a* $y_{l-1}(\eta')$ *przeciwobraz prawdopodobieństwa* η' *względem dystrybuanty rozkładu* χ^2 *z* $l-1$ *stopniami swobody. Wtedy nierówność*

$$\frac{|\mu|}{\sqrt{\lambda}} > \frac{|\bar{X}|}{S}\sqrt{\frac{y_{l-1}(\eta')}{l}} + \frac{x(\eta/2)}{\sqrt{l}}$$

określa jednostronny przedział ufności dla ilorazu $|\mu|/\sqrt{\lambda}$, *pokrywający ten iloraz z prawdopodobieństwem większym niż* $(1-\eta)(1-\eta')$.

Dowód. Załóżmy, że $\mu = 0$. Wówczas na poziomie ufności $1-\eta$ zachodzi nierówność

$$\bar{X}\sqrt{\frac{l}{\lambda}} < -x(\eta),$$

a dla zmiennej losowej $|\bar{X}|$ na tymże poziomie ufności zachodzi nierówność

$$|\bar{X}|\sqrt{\frac{l}{\lambda}} < -x(\eta/2).$$

Dla $\mu \neq 0$ zachodzi z prawdopodobieństwem większym niż $1-\eta$ nierówność

$$(|\bar{X}| - |\mu|)\sqrt{\frac{l}{\lambda}} < -x(\eta/2)$$

lub równoważna jej nierówność

$$\frac{|\mu|}{\sqrt{\lambda}} > \frac{|\bar{X}|}{\sqrt{\lambda}} + \frac{x(\eta/2)}{\sqrt{l}}.$$

Zauważmy, że zgodnie z założeniami lematu, na poziomie ufności $1-\eta'$ zachodzi nierówność

$$\sqrt{\lambda} < S\sqrt{\frac{l}{y_{l-1}(\eta')}}.$$

Z ostatnich dwu nierówności wynika teza lematu na mocy znanego faktu niezależności statystyk $\bar{X}$ i S.

TWIERDZENIE I. *Wartość bezwzględna kowariancji pary zmiennych losowych X, Y o łącznym rozkładzie normalnym jest nie mniejsza od bezwzględnej wartości kowariancji ich wartości bezwzględnych*

$$|E(X-EX)(Y-EY)| \geqslant |E(|X|-E|X|)(|Y|-E|Y|)|.$$

Gdy rozkład jest nieosobliwy, równość zachodzi wtedy i tylko wtedy, gdy zmienne losowe X, Y są niezależne.

Dowód. Niech $f(x, y, \varrho)$ oznacza łączną gęstość zmiennych losowych X, Y z nieujemnymi średnimi ν_1 i ν_2 i współczynnikiem korelacji ϱ. Oznaczmy całki w poszczególnych ćwiartkach układu współrzędnych w przypadku $\varrho_1 \geqslant 0$, w sposób następujący:

$$(*)\qquad \begin{aligned} \mu_1 &= \int_0^\infty \int_0^\infty |xy| f(x, y, \varrho_1)\,dxdy, \\ \mu_2 &= \int_0^\infty \int_{-\infty}^0 |xy| f(x, y, \varrho_1)\,dxdy, \\ \mu_3 &= \int_{-\infty}^0 \int_{-\infty}^0 |xy| f(x, y, \varrho_1)\,dxdy, \\ \mu_4 &= \int_{-\infty}^0 \int_0^\infty |xy| f(x, y, \varrho_1)\,dxdy. \end{aligned}$$

Jak łatwo widzieć, $\operatorname{Cov}(X, Y) = \mu_1-\mu_2+\mu_3-\mu_4-\nu_1\nu_2$, a

$$\operatorname{Cov}(|X|, |Y|) = \mu_1+\mu_2+\mu_3+\mu_4-\theta_1\theta_2,$$

gdzie θ_1, θ_2 są średnimi zmiennych $|X|$, $|Y|$.

Niech

$$\Delta = \operatorname{Cov}(X, Y)-\operatorname{Cov}(|X|, |Y|) = \theta_1\theta_2-\nu_1\nu_2-2(\mu_2+\mu_4).$$

Pokażemy najpierw, że jeśli $\nu_1 \geqslant 0$ i $\nu_2 \geqslant 0$, to $\Delta > 0$, gdy $\operatorname{Cov}(X, Y) > 0$.

Weźmy w tym celu pod uwagę taką gęstość normalną $f(x, y, \varrho_2)$, że $\varrho_2 > 0$. Wystarczy pokazać (patrz wzór (9)), że $2\nu_1\psi_2+2\nu_2\psi_1+4\psi_1\psi_2 >$

$> 2(\mu_2+\mu_4)$, a otrzymamy stąd, że $(\nu_2+\psi_2)\psi_1 > \mu_2$ oraz $(\nu_1+\psi_1)\psi_2 > \mu_4$. Mamy więc, stosując zwykłe oznaczenia:

$$\mu_2 = \int_0^{\infty}\int_0^{-\infty} xyf(x,y)\,dxdy = \int_0^{-\infty} xf_1(x)\int_0^{\infty} yf(y|x)\,dydx <$$

$$< \int_0^{-\infty} xf_1(x)\int_0^{\infty} yf(y|0)\,dydx = \bar{\nu}_2(0)\int_0^{-\infty} xf_1(x)\,dx =$$

$$= \bar{\nu}_2(0)\psi_1 < (\nu_2+\psi_2)\psi_1.$$

Pierwsza nierówność jest tu uzasadniona faktem, iż funkcja

$$\bar{\nu}_2(x) = \int_0^{\infty} yf(y|x)\,dy = v(x)+\bar{\psi}_2(x)$$

jest dla $\varrho > 0$ rosnąca względem x, gdyż

$$v(x) = \begin{cases} 0, & \text{gdy } \nu_2(x) \leqslant 0, \\ \nu_2(x), & \text{gdy } \nu_2(x) > 0, \end{cases}$$

a $\nu_2(x)$, jako średnia warunkowa, jest rosnąca, oraz $\bar{\psi}_2(x)$ oznacza funkcję $\psi(|\nu_2(x)|)$; $\theta = \nu+\psi(\nu,\sigma)$ jest dla stałej σ i $\nu > 0$ rosnącą funkcją ν (patrz dowód lematu 2.2 w [7]).

Drugą nierówność uzasadniamy następująco: $\bar{\nu}_2(0) \leqslant v_2(\nu_1)+\bar{\psi}_2(\nu_1)$, a $v_2(\nu_1) = \nu_2$ oraz $\bar{\psi}_2(\nu_1) < \psi_2$, gdyż po lewej stronie występuje wariancja warunkowa, mniejsza od brzegowej występującej po prawej stronie.

Dowód nierówności $(\nu_1+\psi_1)\psi_2 > \mu_4$, gdy $\varrho > 0$, przebiega podobnie. Zatem $\Delta > 0$, gdy $\nu_1 > 0$, $\nu_2 > 0$ oraz $\varrho > 0$.

Przypadek $\varrho = 0$, w którym zachodzi $\Delta = 0$, jest trywialny, gdyż funkcje zmiennych losowych niezależnych są niezależne stochastycznie.

W przypadku $\nu_1 > 0$, $\nu_2 > 0$ oraz $\varrho < 0$, oznaczmy całki (*) odpowiednio przez $\mu_1', \mu_2', \mu_3', \mu_4'$. Szacowanie podobne do przedstawionego wyżej (dla przypadku $\varrho > 0$) pokazuje, że $\mu_2' > \mu_2$. Podobnie dowodzimy, że $\mu_4' > \mu_4$, co daje nam w tym przypadku nierówność $\Delta < 0$.

Podobnie jak wyżej, wykazujemy też, iż suma $\mathrm{Cov}(X, Y)+\mathrm{Cov}(|X|, |Y|)$ jest większa od zera, gdy $\varrho > 0$, oraz mniejsza od zera, gdy $\varrho < 0$.

Na koniec, w przypadkach $\nu_1\nu_2 < 0$ oraz $\nu_1 < 0$ i $\nu_2 < 0$, sprowadzamy zagadnienie do przedstawionego wyżej, przez zmianę znaków; w pierwszym przypadku jednej zmiennej, a w drugim — obu. Rozważanie geometryczne pokazuje, że nie zmienia to wartości $\mathrm{Cov}(|X|, |Y|)$, c. n. d.

Teza udowodnionego twierdzenia wskazuje na to, że stosując w praktyce proponowane tutaj metody trzeba dobierać cechy słabo skorelowane. Ponieważ składowe D_{ij}^{pq} odległości empirycznych są słabiej skorelowane niż cechy wyjściowe, więc suma kowariancji z losowymi zna-

kami, o rozkładzie granicznie normalnym, tworząca wraz z wariancją $S_\varkappa^2$ odległości empirycznej rozkład złożony, mało wpływa na wariancję tej statystyki, nie zmieniając jej średniej $\sigma_\varkappa^2$ (np. dla $|\varrho| < 0{,}5$ i dla pięciu cech wzrost tej wariancji jest rzędu 1/20). W dowodzie twierdzenia II zobaczymy, że dzięki stosowanej tam majoryzacji wystarczy nieco powiększyć wielkości $g_\varkappa$.

LEMAT 3. *Wariancje σ_{pq}^2 i $\bar{\omega}_{pq}^2$ odległości empirycznych*(5) *D^{pq} oraz $\bar{d}^{pq}$ można oszacować od góry jak następuje*:

$$\sigma_{pq}^2 < \sum_{i_1 i_2} |\lambda_{i_1 i_2}^{pq}|, \qquad \bar{\omega}_{pq}^2 < \sum_{i_1 i_2} \left| \frac{\lambda_{i_1 i_2}^{p}}{l_p} + \frac{\lambda_{i_1 i_2}^{q}}{l_q} \right|.$$

Dowód. Znajdziemy wariancje dla omawianych odległości w jednej współrzędnej. Mamy:

$$\sigma_{i,pq}^2 = E(D_{ij}^{pq} - \theta_i^{pq})^2 = E(D_{ij}^{pq})^2 - (\theta_i^{pq})^2 = (\nu_i^{p|q})^2 + \lambda_{ii}^{pq} -$$
$$- [|\nu_i^{p|q}| + 2\psi(|\nu_i^{p|q}|, \sqrt{\lambda_{ii}^{pq}})]^2 = \lambda_{ii}^{pq} - 4[|\nu_i^{p|q}|\psi(|\nu_i^{p|q}|, \sqrt{\lambda_{ii}^{pq}}) +$$
$$+ \psi^2(|\nu_i^{p|q}|, \sqrt{\lambda_{ii}^{pq}})] < \lambda_{ii}^{pq},$$

$$\bar{\omega}_{i,pq}^2 = E(\bar{d}_i^{pq} - \bar{\delta}_i^{pq})^2 = E(\bar{d}_i^{pq})^2 - (\bar{\delta}_i^{pq})^2 =$$
$$= (\nu_i^{p|q})^2 + \frac{\lambda_{ii}^{p}}{l_p} + \frac{\lambda_{ii}^{q}}{l_q} - \left[|\nu_i^{p|q}| 2\psi\left(|\nu_i^{p|q}|, \sqrt{\frac{\lambda_{ii}^{p}}{l_p} + \frac{\lambda_{ii}^{q}}{l_q}}\right)\right]^2 =$$
$$= \frac{\lambda_{ii}^{p}}{l_p} + \frac{\lambda_{ii}^{q}}{l_q} - 4\left[|\nu_i^{p|q}| \psi\left(|\nu_i^{p|q}|, \sqrt{\frac{\lambda_{ii}^{p}}{l_p} + \frac{\lambda_{ii}^{q}}{l_q}}\right) + \psi^2\left(|\nu_i^{p|q}|, \sqrt{\frac{\lambda_{ii}^{p}}{l_p} + \frac{\lambda_{ii}^{q}}{l_q}}\right)\right] <$$
$$< \frac{\lambda_{ii}^{p}}{l_p} + \frac{\lambda_{ii}^{q}}{l_q}.$$

Na podstawie tych równości oraz twierdzenia I, znajdujemy oszacowanie dla wariancji odległości empirycznych D i $\bar{d}$ podane w tezie lematu.

W następnym twierdzeniu użyjemy różnych symboli wprowadzonych w § 2. Dla uproszczenia zapisu wskaźniki $\varkappa_1, \varkappa_2$ numeracji pojedynczej $\varkappa = 1, 2, \ldots, k$ zastąpimy wskaźnikami 1, 2.

TWIERDZENIE II. *Jeśli dla pewnych liczb dodatnich g_1, g_2 spełnione są nierówności*

$$\sum_{i=1}^{n} \sigma_{i,1}^2 \leqslant g_1^2, \qquad \sum_{i=1}^{n} \sigma_{i,2}^2 \leqslant g_2^2,$$

$$\delta_1^2(n-1) \geqslant \sum_{i=1}^{n} \lambda_{ii,1}, \qquad \delta_2^2(n-1) \geqslant \sum_{i=1}^{n} \lambda_{ii,2}$$

(5) Odległość D^{pq} określona została wzorem (4), a odległość $\bar{d}^{pq}$ wzorem (25).

oraz

$$\theta_1 - \theta_2 > g_{12}\sqrt{2(n-1)/(\pi-2)}, \quad \text{gdzie } g_{12} = \max(g_1, g_2),$$

to

$$\delta_1 > \delta_2.$$

Dowód. Niech

$$\delta = \sum_{i=1}^{n} v_i,$$

przy czym zakładamy, że $v_i \geqslant 0$ $(0 \leqslant i \leqslant n)$. W n-wymiarowej kartezjańskiej przestrzeni K_n, ustalonej wartości δ odpowiada część hiperpłaszczyzny zawarta w dodatnim sektorze układu współrzędnych, a określona tym równaniem. Oznaczamy tę część hiperpłaszczyzny przez S_δ.

Jeśli zastępujemy prawdziwą wartość odległości δ przez średnią θ statystyki $\bar{D}$, to zgodnie z wzorem (9) zbiór punktów $p \in S_\delta$ przejdzie na zbiór punktów $p^* \in S_\delta^*$ podług formuły $p_i^* = p_i + 2\psi(p_i, \sqrt{\lambda_{ii}})$.

Wprowadźmy funkcję

$$\Psi(p) = \sum_{i=1}^{n} (p_i^* - p).$$

Jeżeli zachodzi nierówność $\theta_1 - \theta_2 > \sup[\Psi(p_1) - \Psi(p_2)]$, gdzie $p_1 \in S$ i $p_2 \in S_\delta$, a stała $\bar{g}^2 = \sum_{i=1}^{n} \lambda_{ii}$, to $\delta_1 > \delta_2$.

Zbadajmy to supremum. W tym celu znajdziemy najpierw ekstremum funkcji $\Psi(p)$ z warunkiem ubocznym $p \in S_\delta$ przy ustalonych wartościach $\lambda_{11}, \lambda_{22}, \ldots, \lambda_{nn}$. Ponieważ

$$\frac{\partial}{\partial p_i}\psi(p_i, \sqrt{\lambda_{ii}}) = -\Phi\left(-\frac{p_i}{\sqrt{\lambda_{ii}}}\right),$$

więc otrzymujemy metodą Lagrange'a,

$$\frac{\partial}{\partial p_i}\left[\Psi + \tau\left(\sum p_i - \delta\right)\right] = -2\Phi\left(-\frac{p_i}{\sqrt{\lambda_{ii}}}\right) + \tau = 0,$$

a stąd $-p_i/\sqrt{\lambda_{ii}} = \Phi^{-1}(\tau/2)$, czyli

$$\delta = -\sum \sqrt{\lambda_{ii}}\,\Phi^{-1}(\tau/2),$$

skąd $\tau = 2\Phi(-\delta/\sum\sqrt{\lambda_{ii}})$. Punkt ekstremum ma więc współrzędne

$$\bar{p}_i = \delta\sqrt{\lambda_{ii}}/\sum_j \sqrt{\lambda_{jj}}.$$

Dla sprawdzenia, czy jest to minimum, zbadajmy pochodną funkcji Ψ w dowolnym kierunku, poczynając od punktu $\bar{p}$ wewnątrz hiperpowierzchni $\sum p_i = \delta$. Ten ostatni warunek spełnia prosta, której współczynniki kierunkowe $a_1, a_2, \ldots, a_n$ sumują się do zera. Stąd szukana prosta daje się przedstawić w postaci parametrycznej równaniami $x_i = a_i t + \delta\sqrt{\lambda_{ii}}/\sum_j \sqrt{\lambda_{jj}}$, wobec czego szukana pochodna jest

$$\partial\Psi/\partial t = -2\sum_i a_i \Phi\left(-\frac{a_i}{\sqrt{\lambda_{ii}}}t - \frac{\delta}{\sum\sqrt{\lambda_{jj}}}\right), \qquad \sum a_i = 0.$$

Gdy $t = 0$, $\partial\Psi/\partial t = 0$. Gdy $t > 0$, to łatwo znaleźć takie $a_1, a_2, \ldots, a_n$, że $\partial\Psi/\partial t > 0$. Widać także, że przy ustalonych, dowolnych a_i, $\partial\Psi/\partial t$ jest funkcją monotoniczną; jest więc dla $t > 0$ dodatnia. Supremum w zbiorze S_δ należy więc szukać na brzegach tego zbioru, ograniczonego płaszczyznami, przechodzącymi przez pary osi układu współrzędnych; brzegi te są zatem wielokątem o wierzchołkach w punktach $(0, 0, \ldots, \delta_i, \ldots, 0)$, co należy czytać w ten sposób, że wartość δ stoi w tym wektorze na i-tym miejscu.

Zbadajmy pochodne na bokach tego wielokąta. Dowolny punkt boku łączącego i-ty wierzchołek z j-tym jest $(0, \ldots, 0, \delta_i t, 0, \ldots, 0, \delta_j(1-t), 0, \ldots, 0)$, gdzie $0 \leqslant t \leqslant 1$. Pochodna względem t ma więc postać następującą:

$$\partial\Psi/\partial t = 2\delta\left[\Phi\left(\frac{t-1}{\sqrt{\lambda_{ii}}}\delta\right) - \Phi\left(-\frac{\delta t}{\sqrt{\lambda_{jj}}}\right)\right].$$

Jak widać, dla małych t pochodna ta jest ujemna, a dla większych dodatnia, ma więc punkt zerowy. Maksimum należy zatem szukać na wierzchołkach wspomnianego wielokąta.

Zbadajmy teraz ekstremum warunkowe funkcji Ψ na wierzchołku S_δ z warunkiem ubocznym $\sum\lambda_{ii} = \bar{g}^2$, gdzie $\bar{g}$ jest dowolną liczbą dodatnią. Ponieważ pochodna funkcji Ψ względem λ_{ii} jest skomplikowana, podstawimy $y_i = 1/\sqrt{\lambda_{ii}}$. Badamy pierwszy wierzchołek, co nie zmniejsza ogólności rozważań.

Naszym warunkiem ubocznym jest więc

$$f = y_1^{-2} + \sum_{i=2}^{n} y_i^{-2} - \bar{g}^2 = 0.$$

Zgodnie z metodą Lagrange'a otrzymujemy wtedy

$$\frac{\partial}{\partial y_1}[2\psi(\delta, 1/y_1) + \tau f] = -\frac{1}{y_1^2}\varphi(\delta y_1) - 2\tau y_1^{-3} = 0,$$

$$\frac{\partial}{\partial y_i}[2\psi(0, 1/y_i) + \tau f] = -\frac{1}{y_i^2\sqrt{2\pi}} - 2\tau y_i^{-3} = 0,$$

skąd $\tau = -\bar{y}/\sqrt{8\pi}$, gdzie $\bar{y} = y_i, 2 \leqslant i \leqslant n$. Wobec tego $y_1^{-2} = \bar{g}^2 - (n-1)\bar{y}^{-2}$, a więc

$$\bar{y}^2 = \frac{n-1}{\bar{g}^2 - y_1^{-2}},$$

co po podstawieniu do wyrażenia na τ daje

$$\tau = -\sqrt{\frac{n-1}{8\pi(\bar{g}^2 - y_1^{-2})}}.$$

Podstawiając tę wartość w równanie pochodnej względem y_1 otrzymamy ostatecznie

$$-y_1^{-2}\varphi(\delta y_1) + 2y_1^{-3}\sqrt{\frac{n-1}{8\pi(\bar{g}^2 - y_1^{-2})}} = 0.$$

Wykażemy teraz, że dla $\delta \geqslant \bar{g}/\sqrt{n-1}$ równanie to jest nieprawdziwe i że wówczas lewa strona jest większa od zera.

Kładąc w miejsce równości znak $\leqslant$ i podstawiając $y_1^2 = z$, otrzymujemy po uproszczeniach

$$(**) \qquad e^{\delta^2 z} \leqslant \frac{1}{n-1}(\bar{g}^2 z - 1).$$

Istnieje tu warunek $z \geqslant 1/\bar{g}^2$; pochodna lewej strony (**) równa jest $\delta^2 e^{\delta^2 z}$, a stąd styczną w punkcie $z = 1/\bar{g}^2$ jest $\delta^2 e^{\delta^2/\bar{g}^2} z + e^{\delta^2/\bar{g}^2}(1 - \delta^2/\bar{g}^2)$. Prosta ta przy założeniu $\delta > \bar{g}/\sqrt{n-1}$ ma współczynnik kierunkowy większy niż $(n-1)^{-1}\bar{g}^2$, gdy $n \geqslant 2$, oraz przecina oś z w punkcie $z_0 = 1/\bar{g}^2 - 1/\delta^2 < 1/\bar{g}^2$, co dowodzi, że nierówność (**) jest nieprawdziwa.

Widzimy więc, że funkcja Ψ nie ma ekstremum warunkowego na dowolnym wierzchołku i rośnie, gdy $y_1 \to \infty$. Stąd oraz z równości ekstremalnych y_i znajdujemy, że

$$\sup_{\delta > \bar{g}/\sqrt{n-1}} \Psi = 2(n-1)\psi(0, \bar{g}/\sqrt{n-1}).$$

Oczywiście

$$\sup_{\delta > \bar{g}/\sqrt{n-1}} \Psi > \sup_{\delta > \bar{g}/\sqrt{n-1}} [\Psi(p_1) - \Psi(p_2)],$$

przy czym łatwo zauważyć, że równość jest osiągana tylko dla $\delta \to \infty$. Łatwo także zauważyć, podstawiając $u_i = \delta_i/\sqrt{\lambda_{ii}}$, że

$$\sup_{(\delta, g')} \Psi < \sup_{(\delta, \bar{g})} \Psi,$$

jeśli $g' < \bar{g}$, przy czym $(g')^2 = \sum_{i=1}^{n} \lambda'_{ii}$, podobnie jak $\bar{g}^2$.

W dotychczasowym rozumowaniu wykorzystaliśmy drugie z założeń twierdzenia; obecnie wykorzystamy pierwsze, a mianowicie że

$$g_1^2 \geqslant \sum_{i=1}^{n} \sigma_{i1}^2 \quad \text{i} \quad g_2^2 \geqslant \sum_{i=1}^{n} \sigma_{i2}^2.$$

Na tej podstawie możemy przyjąć za majoranty wartości σ_{i1} i σ_{i2} odpowiednio $g_1/\sqrt{n-1}$ i $g_2/\sqrt{n-1}$. Dalej, przechodząc do wartości λ_{ii} zauważmy, że na podstawie wzorów lematu 3 mamy

$$\sigma_i^2 \geqslant \lambda_{ii} - 4\psi^2(0, \sqrt{\lambda_{ii}}) = \lambda_{ii}\left(1 - \frac{2}{\pi}\right),$$

gdyż funkcja $v\psi(v, \sqrt{\lambda}) + \psi^2(v, \sqrt{\lambda})$ jest malejąca względem $v \geqslant 0$. Stąd otrzymujemy majorantę wartości λ_{ii} w postaci

$$\frac{\pi}{\pi - 2}\sigma_i^2.$$

Wstawiając wartość

$$\sqrt{\frac{\pi}{\pi-2}} \cdot \frac{g}{\sqrt{n-1}}$$

zamiast $\bar{g}/\sqrt{n-1}$ w $\sup \Psi$ mamy ostatecznie

$$\sup_{\delta > \bar{g}/\sqrt{n-1}} \Psi \leqslant 2(n-1)\psi\left(0, g\sqrt{\frac{\pi}{(\pi-2)(n-1)}}\right) = g\sqrt{\frac{2(n-1)}{\pi-2}}.$$

Biorąc $\max(g_1, g_2)$, otrzymujemy tezę twierdzenia.

5. PRZYKŁAD

W poprzednich paragrafach omówiliśmy teoretyczne podstawy metody wyznaczania, przy założonym poziomie ufności $1-\alpha-\beta$, rodziny dendrytów $\mathfrak{R}$, do której z prawdopodobieństwem większym niż $1-\alpha-\beta$ należy najkrótszy dendryt Δ, rozpięty na zbiorze elementów będących populacjami statystycznymi o pewnych szczególnych własnościach. Porządkowane elementy będziemy także nazywali *obiektami statystycznymi*. A oto krótki opis postępowania przy stosowaniu metody w praktyce.

1. Wybieramy interesujące nas obiekty statystyczne oraz cechy, względem których będziemy badali pokrewieństwa obiektów. Musimy przy tym zwrócić uwagę, by wybrać cechy dostatecznie silnie dyskry-

minujące, tzn. takie, względem których rozpatrywane obiekty są dostatecznie odległe (przy danej liczbie elementów w próbce pobranej z każdej populacji). Cechy muszą być także związane z interesującymi nas właściwościami porządkowanych obiektów. Jeśli badane populacje są podzielone na warstwy, to we wszystkich warstwach muszą być spełnione warunki (1) i (2) ze str. 394. W przypadku populacyj jednorodnych wystarczy podać umowną zasadę przyporządkowywania fikcyjnym warstwom poszczególnych elementów próby.

2. Z każdej warstwy każdej populacji pobieramy jednoelementową próbkę i mierzymy w niej interesujące nas cechy. Gdy tworzymy wiązki (p. str. 402), z każdej warstwy r-tej populacji pobieramy próbki u_r-elementowe. Obiekty, cechy oraz warstwy numerujemy jak w § 2.

3. Za pomocą kryterium (22) sprawdzamy dyskryminacyjność cech w poszczególnych parach rozpatrywanych obiektów. Wystarczy, by dla pary obiektów p, q prawa strona równania (22) dla większych różnic $\bar{x}_i^p - \bar{x}_i^q$ $(i = 1, 2, \ldots, n)$ była nie mniejsza niż 2. Cechy i populacje niedyskryminujące odrzucamy (np. gdy stwierdzimy, że dla żadnej z cech obiekty p i q nie są dostatecznie odległe według kryterium (22), powiemy, że w warunkach danego eksperymentu i przy zastosowaniu opisanej metody obiekty te są statystycznie nierozróżnialne i przynajmniej jeden z nich można pominąć).

4. W przypadku wielowarstwowych populacyj sprawdzamy (jeśli nie wiemy tego skądinąd), czy nie występuje rozciąganie od warstwy do warstwy. Dla sprawdzenia hipotezy równości wariancyj w warstwach można zastosować np. test Bartletta.

5. Normujemy wszystkie populacje w poszczególnych cechach na średnią albo na dyspersję (ewentualnie normujemy wszystkie warstwy populacyj).

6. Tworzymy statystyki (5) i (12), tzn. odległości empiryczne i ich średnie kwadratowe odchylenia, oraz za pomocą tych ostatnich i wzoru (16) znajdujemy (na poziomie ufności $1-\beta_\varkappa$) liczby $g_\varkappa$, a stąd $k(k-1)/2$ liczb $g_{\varkappa_1\varkappa_2}$.

7. Z wzoru (18) obliczamy $k(k-1)/2$ ilorazów $t_{\varkappa_1\varkappa_2}$.

8. Przy założonych prawdopodobieństwach $\alpha_\varkappa$ znajdujemy dendrytową rodzinę ufności $\mathfrak{R}$. W tym celu, wychodząc od dowolnego elementu zbioru $\mathscr{M}$, przy każdym kroku metody Warmusa wyznaczamy te odcinki, które można przyłączyć z uwagi na ich istotną mniejszość od wszystkich innych jeszcze nie przyłączonych. Jeżeli dany krok konstrukcji dendrytu można wykonać na kilka sposobów, to każdy z nich poprowadzi na ogół do różnych dendrytów. Uwzględniając te różne możliwości skonstruujemy całą rodzinę dendrytów $\mathfrak{R}$.

Zwróćmy jeszcze uwagę na fakt, że w proponowanej metodzie można się ograniczyć do wykonania tylko pierwszego kroku metody Warmusa. Wystarczy wówczas rozpatrywać jedynie zbiór $m-1$ odległości wybranego obiektu od pozostałych obiektów. Postępujemy tak wtedy, gdy do danego obiektu szukamy tylko obiektu najbliższego (np. w poszukiwaniu najlepszego materiału zastępczego). W tym przypadku, z uwagi na mniejszą liczbę porównywanych odległości, będzie tylko $2(m-1)$ prawdopodobieństw $\alpha_\varkappa$ i $\beta_\varkappa$.

Obecnie zilustrujemy opisane postępowanie na materiale pomiarowym udostępnionym przez prof. dr J. Różyckiego, a pochodzącym z prowadzonej przez niego Katedry Budowy Dróg Politechniki Wrocławskiej. Autor korzystał również z informacji udzielanych mu życzliwie przez pracownika tej katedry doc. dra B. Stypułkowskiego. Zastosowanie metody omówimy według opisanego schematu postępowania, zaczynając od łącznego omówienia punktów 1 i 2.

Kamieniołomy nadsyłają do Katedry próbki materiałów kamiennych. W Katedrze bada się te próbki pod względem przydatności do budowy dróg (w omawianym przykładzie przedmiotem badań jest tłuczeń drogowy). Bada się w tym celu różne własności mechaniczne kamienia, takie jak wytrzymałość, nasiąkliwość wodą i ścieralność wyznaczaną różnymi metodami. Celem badania jest syntetyczna ocena badanego materiału. Metoda przedstawiona w tej pracy może być przydatna dla ilościowego ujęcia podobieństw między materiałami z różnych kamieniołomów, a także podobieństw do konwencjonalnie wyznaczonego „idealnego materiału".

Z dostępnych materiałów wybrano 4 kamieniołomy i 3 badane cechy (mające szczególne znaczenie przy ocenie tłucznia). Wyniki pomiarów przedstawia tablica I. Dane tej tablicy nie były niestety wynikiem badań przeprowadzonych z uwzględnieniem potrzeb opisanej tu metody. Nie wiemy np. jakiego typu są rozkłady omawianych cech w omawianych populacjach. Biorąc pod uwagę ingerencję drobnych zdarzeń losowych w pomiarach dokonywanych na jednorodnym materiale skalnym, można się spodziewać normalności tych rozkładów. Dla niektórych złoży (np. skał osadowych) jest przypuszczalnie możliwy podział na warstwy, odpowiadający rzeczywistemu modelowi warstwowemu, wspomnianemu w wyjaśnieniach na str. 394. Widzimy, iż materiał zawarty w tablicy I może mieć jedynie wartość ilustracyjną, nie pozwala bowiem na wysnuwanie jakichś teoretycznych wniosków. Jest on zresztą zbyt szczupły. Politechnika Wrocławska nie posiada jednak wyników pomiarowych zawierających powtórzenie prób więcej niż trzykrotne, gdyż nie przewidują tego normy techniczne, według których próby te zostały przeprowadzone. Gdybyśmy pobrali wieloelementowe próbki (wiązki), moglibyśmy otrzymać nieliczną rodzinę dendrytów na rozsądnym poziomie

ufności i wykorzystać ją przy wydawaniu orzeczeń o wartości użytkowej materiałów kamiennych.

Przyjętą numerację przedstawia tablica I. Symbole cech oznaczają:

a — wartość w $^0/_0$ zużycia kruszywa odsianego, o prawidłowym kształcie ziaren i średnicy 40-63 mm, w młynie typu Los Angeles;

b — wytrzymałość na ściskanie w kG/cm^2 (kostki mokrej wymiarów 5 cm $\times$ 5 cm $\times$ 5 cm określenie w PN-54/0-04110);

c — nasiąkliwość wagową wyrażną w $^0/_0$.

Przechodząc do omówienia punktu 3 uczynimy ogólną uwagę, iż w zasadzie wystarczy zbadać za pomocą kryterium (22) odległości najmniejsze, gdyż odległości istotnie od nich większe tym bardziej je spełniają. Zbadajmy więc przykładowo odległość (1,2) w cesze a i odległość (1,4) w cesze c. Mamy

$$s^2_{a,12} = [(1{,}3-1{,}5)^2+(2{,}7-1{,}5)^2+(0{,}6-1{,}5)^2]/3 \approx 0{,}76,$$

skąd

$$v^{1|2}_a/\sqrt{\lambda^{12}_{aa}} > \frac{1{,}5}{0{,}87}\sqrt{\frac{1}{3}0{,}1} - 1{,}65/\sqrt{3} < 0;$$

$$s^2_{c,14} = [(25{,}5-28{,}3)^2+(32{,}0-28{,}3)^2+(27{,}5-28{,}3)^2]/3 \approx 7$$

stąd

$$v^{4|1}_c/\sqrt{\lambda^{14}_{cc}} > \frac{28{,}3}{2{,}65}\sqrt{\frac{1}{3}0{,}1} - 1{,}65/\sqrt{3} \approx 1.$$

Obliczenia podobne do przedstawionych tutaj pokazują małą dyskryminacyjność cech w rozpatrywanym przez nas materiale zawartym w tablicy I. W celu uzyskania wyników zadowalających należałoby więc pobrać próbki o wiele liczniejsze (po 30 lub więcej powtórzeń z każdej odmiany kamienia) i utworzyć z nich wiązki.

Punkt 4, z uwagi na brak danych, pomijamy. Co do punktu 5 to zastosujemy bardzo łatwe w liczeniu przybliżone unormowanie cech na średnią. Uzyskamy je przez podzielenie cechy b przez 100 oraz pomnożenie cechy c przez 50 (czyli 100/2). Cechę a pozostawiamy niezmienioną. W tej więc normalizacji szukamy najłatwiejszych dzielników, zbliżonych wartością do średniej średnich (p. str. 394).

Z uwagi na szczupłość materiału i niemożność zastosowania normalnej procedury (np. określenie nie za wielkich wartości $g_\varkappa$ na rozsądnym poziomie ufności jest tu niemożliwe, gdyż, jak zobaczymy niżej, ilorazy $t_{\varkappa_1\varkappa_2}$ są na ogół nieistotne, a statystyki $S^2_\varkappa$ mają rozkład χ^2 tylko z dwoma stopniami swobody, pomijając fakt, iż rozkład ten w rzeczywistości dość się różni od χ^2), omówimy teraz w skrócie punkty 6, 7 i 8 w oparciu o załączoną tablicę II odległości empirycznych, obliczonych z tablicy I.

TABLICA I. Tłuczeń drogowy — dane

Kamieniołomy	1 Gracze bazalt			2 Kowalskie bazalt			3 Strzelin granit			4 Kieleckie kwarcyt		
Lp.	1	2	3	1	2	3	1	2	3	1	2	3
a	9,8	10,6	8,8	8,5	7,9	8,2	15,7	15,4	15,5	22,8	23,3	23,0
b	2010	1940	1810	2050	2060	2410	1800	1330	1650	2000	2130	1860
unorm.	20,1	19,4	18,1	20,5	20,6	24,1	18,0	13,3	16,5	20,0	21,3	18,6
c	0,36	0,24	0,19	0,20	0,33	0,17	0,43	0,43	0,44	0,87	0,88	0,74
unorm.	18,0	12,0	9,5	10,0	16,5	8,5	21,5	21,5	22,0	43,5	44,0	37,0

Ponieważ przykład niniejszy służy jedynie jako ilustracja metody, ustalimy wartości $g_\varkappa$ arbitralnie, biorąc za nie największe realizacje bezwzględnej wartości różnicy odległości warstwowej i średniej

$$g_\varkappa = \max_j |D_{j\varkappa} - \bar{D}_\varkappa| = \max_j \Delta_{j\varkappa}.$$

Tak ustalimy następujące wartości $g'_\varkappa = g_\varkappa \sqrt{2(3-1)/(\pi-2)}$:

$$g'_{12} = 2{,}1, \quad g'_{13} = 11, \quad g'_{14} = 7{,}7,$$
$$g'_{23} = 9{,}8, \quad g'_{24} = 6{,}3, \quad g'_{34} = 13{,}1.$$

Stąd łatwo już wyliczyć ilorazy t ze wzoru (18):

$$t_{(1,2),(1,3)} < 0, \quad t_{(1,3),(1,4)} \approx 2{,}7, \quad t_{(1,4),(2,4)} < 0,$$
$$t_{(1,2),(1,4)} \approx 9, \quad t_{(1,3),(2,3)} < 0, \quad t_{(1,4),(3,4)} < 0,$$
$$t_{(1,2),(2,3)} \approx 1{,}4, \quad t_{(1,3),(2,4)} \approx 3{,}9, \quad t_{(2,3),(2,4)} \approx 3{,}2,$$
$$t_{(1,2),(2,4)} \approx 16, \quad t_{(1,3),(3,4)} \approx 0{,}1, \quad t_{(2,3),(3,4)} < 0,$$
$$t_{(1,2),(3,4)} \approx 2{,}1, \quad t_{(1,4),(2,3)} \approx 1{,}8, \quad t_{(2,4),(3,4)} \approx 0{,}4.$$

Wartości istotne na poziomie ufności 0,9 zostały pogrubione, przy czym wygrubiono parę liczb oznaczającą odległość istotnie mniejszą, a więc tę, którą wybieramy wykonując krok w budowie dendrytu.

Na tej podstawie łatwo sprawdzimy, iż wychodząc z elementu nr 1 (Gracze) można wykonać pierwszy krok w budowie dendrytu algorytmem Warmusa na dwa sposoby: 1. (1,2); 2. (1,3).

Dwa pierwsze kroki można już wykonać na cztery sposoby:

1. (1,2), (1,3); 2. (1.2), (2,3);
3. (1,3), (2,3); 4. (1,3), (3,4);

TABLICA II. Tłuczeń drogowy — odległości empiryczne

Kamieniołomy	12 Gracze-Kowalskie			13 Gracze-Strzelin			14 Gracze-Kieleckie			23 Kowalskie-Strzelin			24 Kowalskie-Kieleckie			34 Strzelin-Kieleckie		
L.p.	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
a	1,3	2,7	0,6	5,9	4,8	6,7	13,0	12,7	14,2	7,2	7,5	7,3	14,3	15,4	14,8	7,1	7,9	7,5
b	0,4	1,2	6,0	2,1	6,1	1,6	0,1	1,9	0,5	2,5	7,3	7,6	0,5	0,7	5,5	2,0	8,0	2,1
c	8,0	4,5	1,0	3,5	9,5	12,5	25,5	32,0	27,5	11,5	5,0	13,5	33,5	27,5	28,5	22,0	22,5	15,0
D^{pq}	9,7	8,4	7,6	11,5	20,4	20,8	38,6	46,6	42,2	21,2	19,8	28,4	48,3	43,6	48,8	31,1	38,4	24,6
$\overline{D}^{pq}$		8,57			17,56			42,47			23,17			46,9			31,37	
Δ^{pq}	**1,13**	0,17	0,97	**6,06**	2,84	3,24	3,87	**4,13**	0,27	1,93	3,33	**5,27**	1,4	**3,3**	1,9	0,27	**7,03**	6,77
s^{pq}		0,84			4,30			3,27			3,76			2,34			5,62	

trzy pierwsze kroki, tworzące już szukaną rodzinę dendrytów $\mathfrak{R}$ w tym przypadku, można wykonać na większą ilość sposobów.

Gdybyśmy pobrali liczniejsze próbki, uzyskana rodzina dendrytów oczywiście wydatnie by się zmniejszyła; być może udałoby się w rozważanym tu przypadku uzyskać jedyny dendryt odmian tłucznia. Otrzymane dendryty możemy nałożyć na siebie i znaleźć w ten sposób wielokrotne połączenia obiektów (a więc „najpewniejsze").

Na zakończenie pragnę zwrócić uwagę na nieco inny sposób zastosowania omawianej tu metody, ważny w praktyce. Załóżmy mianowicie, że spośród odmian tłucznia chcemy wybrać najlepszy pod względem pewnego zespołu cech użytkowych (np. w budownictwie drogowym). W tym celu obieramy w przestrzeni cech punkt określający „tłuczeń idealny". Punkt ten można oczywiście potraktować jako wektor losowy o zerowej macierzy kowariancyj; metoda obejmuje i ten przypadek.

Wybór elementu najbliższego danemu jest równoznaczny z wykonaniem pierwszego kroku w budowie dendrytu. Stąd już widać, że i w tym przypadku możemy porównywać z sobą odległości empiryczne za pomocą ilorazu T (wzór (18)), znajdując na założonym poziomie ufności tłuczeń najbliższy idealnemu, czyli najlepszy. Tutaj żądany poziom ufności można uzyskać posiadając nawet mniej liczne próbki, gdyż wystarczy tu rozpatrywać przestrzeń samych odcinków (odległości) wychodząc z danego punktu.

Prace cytowane

[1] H. Cramér, *Metody matematyczne w statystyce*, Warszawa 1958.

[2] J. Czekanowski, *Zarys metod statystycznych w zastosowaniu do antropologii*, PTNW 1913.

[3] — *Zarys antropologii Polski*, Lwów 1930.

[4] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus i S. Zubrzycki, *Sur la liaison et la division des points d'un ensemble fini*, Colloq. Math. 2 (1951), str. 282-285.

[5] — *Taksonomia wrocławska*, Przegląd Antropologiczny 17 (1951), str. 193-211.

[6] S. Hartman i J. Mikusiński, *Teoria miary i całki Lebesgue'a*, Warszawa 1957.

[7] J. Mikiewicz, *O poziomach ufności w taksonomii wrocławskiej*, Zastosow. Matem. 7 (1963), str. 1-40.

[8] W. Oktaba, *Elementy statystyki matematycznej i metodyka doświadczalnictwa*, Łódź — Warszawa 1962.

[9] R. C. Prim, *Shortest connection and some generalizations*, The Bell System Techn. Journ. 36 (1957), No 6.

Praca wpłynęła 15. 3. 1965
Wersja zmieniona 23. 9. 1968

Я. МИКЕВИЧ (Вроцлав)

ДОВЕРИТЕЛЬНЫЕ ОБЛАСТИ ДЛЯ ДЕНДРИТОВ

РЕЗЮМЕ

Настоящая работа содержит улучшение методов автора, опубликованных раньше в статье [7], предлагающей проверку статистической достоверности дендрита, конструированного методом вроцлавской таксономии. Здесь автор вводит новое определение эмпирических расстояний (формула (4)) между исследованными популяциями. Это дает возможность ввести частные Студента (12) и в дальнейшем построить доверительные области для неизвестных теоретических расстояний. Формула (18) указывает критический уровень достоверности сравниваемых эмпирических расстояний. Если в конструкции кратчайшего дендрита сравниваемые расстояния существенно не различны, дендритное упорядочение не удовлетворяет свойством однозначности. Это приводит к конструкции семейства дендритов, которое соответствует суммарной случайной выборке из исследованных популяций. Кроме того, получается оценка снизу вероятности того, что верным окажется решение о принадлежности неизвестного теоретического дендрита к полученному семейству дендритов.

В заключении приводится пример, указывающий применение методов автора к результатам измерений технических свойств камня для постройки дорог.

J. MIKIEWICZ (Wrocław)

CONFIDENCE REGIONS FOR DENDRITES

SUMMARY

This paper presents an improvement of a previous method, published by the author in [7], for the verification of the statistical significance of a dendrite constructed according to the rules of the so called Wrocław taxonomy. The author introduces here a new definition of the empirical distances (see formula (4)) between the populations to be arranged. These allow for the construction of Student's quotients (12) and so for the determination of the confidence regions for the unknown theoretical distances. Formula (18) gives a significance limit for the difference in the compared empirical distances. If one constructs the shortest dendrite, and some of the compared distances do not significantly differ one from another, the dendrite arrangement will not be unique. This leads to a family of dendrites which corresponds to the joint sample composed of observations taken from all studied populations. A lower bound is also obtained for the probability of no mistake in the sentence that the unknown theoretical dendrite belongs to the constructed family of dendrites.

In the last section there is given an example of the application of the proposed method to the measurements of technical characteristics of different kinds of stone material for road construction.