

J. MIKIEWICZ (Wrocław)

O POZIOMACH UFNOŚCI W TAKSONOMII WROCŁAWSKIEJ

§ 1. Zagadnienie i cel pracy. Metoda taksonomiczna, zwana wrocławską⁽¹⁾, ma na celu graficzne przedstawienie na płaszczyźnie pokrewieństwa, lub inaczej wzajemnego podobieństwa indywiduów, rozważanego na tle odległości w przestrzeni metrycznej, wielowymiarowej, w której każda współrzędna oznacza pewną przyrodniczą cechę pomiarową. Ideę takiego ujęcia podobieństwa przyrodniczego widzimy już u J. Czekanowskiego [3]. Metoda wrocławska spełnia postawione wyżej zadanie o tyle lepiej od metody diagraficznej J. Czekanowskiego (patrz [3] i [4]), że ta ostatnia pozwala jedynie na liniowe uporządkowanie badanych indywiduów, podczas gdy taksonomia wrocławska daje uporządkowanie dendrytowe, które lepiej odzwierciedla wzajemne położenie indywiduów w wielowymiarowej przestrzeni.

Jednakże przyrodnik żąda od metody taksonomicznej jeszcze czegoś innego. Jeśli mianowicie np. w biologii możemy badać konkretne osobniki, to chcemy na podstawie ich znajomości wyciągać wnioski dotyczące gatunków, grup, krótko mówiąc populacji, które one reprezentują. W szczególności chcemy wyznaczyć dendryt przedstawiający wzajemne pokrewieństwo tych populacji. I tutaj powstają trudności, gdyż metoda taksonomiczna musi wówczas uwzględniać stochastyczność zjawisk. Celem niniejszej pracy jest badanie pokrewieństw międzygatunkowych wrocławską metodą taksonomiczną, tak żeby można było wypowiadać zdania z określonym prawdopodobieństwem o dendrytach między populacjami.

W dalszym ciągu, by uniknąć komplikacji, będziemy mówić tylko o gatunkach rozumianych jako szczepy genetycznie czyste, tzn. nie podlegające krzyżowaniu z innymi gatunkami. W tej sytuacji przynależność osobnika do danego gatunku nie budzi wątpliwości, a więc dla każdej z cech pomiarowych branych pod uwagę i dla każdego gatunku istnieje określony rozkład prawdopodobieństwa. Korzystając z centralnego twierdzenia granicznego rachunku prawdopodobieństwa można tu rozumować następująco:

⁽¹⁾ Patrz w tej sprawie [5] i [6]. Metoda ta, zwana popularnie „metodą dendrytów”, opiera się na twierdzeniu podanym w [5].

20954111 / 1963/64 8.4

Na wytworzenie danego osobnika (fenotypu) składa się, prócz zespołu cech określonego dokładnie przez dziedziczność (gatunek), również bardzo wiele drobnych, losowych impulsów, działających podczas jego rozwoju, dzięki czemu dostatecznie liczny zbiór osobników, wybranych losowo z populacji tego samego gatunku, tworzy rozkład prawdopodobieństwa zbliżony do normalnego. Rozumowanie to bywa często potwierdzane przez doświadczenie. Stąd też wydaje się naturalne przyjąć wartość oczekiwaną rozkładu danego gatunku za liczbę właściwą danemu gatunkowi (oczywiście w określonym środowisku), lub po prostu za liczbę gatunkową. Jeśli będziemy uwzględniać większą liczbę cech pomiarowych, otrzymamy zamiast liczby gatunkowej wektor gatunkowy, będący środkiem ciężkości rozkładu zmiennej losowej wielowymiarowej, czyli inaczej środkiem ciężkości łącznego rozkładu cech pomiarowych. Tu należy podkreślić, iż cechy pomiarowe gatunków biologicznych są na ogół zależne stochastycznie, stąd zaś powstaje konieczność uwzględnienia tej zależności we wzorach.

Gdybyśmy znali wspomniane wektory gatunkowe dla wszystkich gatunków rozpatrywanych, to dendryt gatunków, sporządzony metodami taksonomii wrocławskiej, byłby już jednoznacznie wyznaczony. Wektorów tych jednak nie znamy; możemy je jedynie oceniać na podstawie prób losowych, tzn. zbiorów osobników wybranych losowo z populacji gatunkowych i pomierzonych. W niniejszej pracy będziemy się starali znaleźć metody statystycznego wnioskowania na podstawie prób losowych, o dwóch grupach zjawisk:

1° O topologicznej strukturze dendrytu gatunków.

2° O rzeczywistych odległościach gatunków w dendrycie.

Dla wyjaśnienia określam, iż dwa dendryty, w sensie podanym w pracach [5] i [6], są równe topologicznie, jeśli przez operację ściągania i rozciągania ich członów (odległości), czyli przez homeomorfizm, można je doprowadzić do nakrycia. Z dwóch podanych wyżej zagadnień istotne jest pierwsze. Zobaczymy dalej, iż odpowiedź na drugie jest ściśle związana z odpowiedzią na pierwsze. Odpowiedź na pierwsze wiąże się z pewnym uogólnieniem metody przedziałów ufności, w sensie Jerzego Spławy Neymana (patrz [1] i [2]). Chodzi tu mianowicie o znajdowanie prawdopodobieństwa topologicznej równości dendrytów, bądź przynależności dendrytu do określonej topologicznie rodziny dendrytów sąsiednich.

Budowanie dendrytu metodą podaną w [5] (jako jedną z możliwych) polega, jak wiadomo, na łączeniu każdego z punktów rozpatrywanego zbioru z najbliższym, przez co tworzymy dendryt pierwszego rzędu, następnie na łączeniu każdego z tych dendrytów z najbliższym, przez co tworzymy dendryt drugiego rzędu itd. Jeśli uwzględnimy, że empiryczne odległości są zmiennymi losowymi, to zauważymy, iż każdy krok tej metody, polegający na wybraniu odległości najkrótszej, jest

zależny od losowego uporządkowania wszystkich odległości wychodzących z danego punktu (reprezentującego gatunek). Stąd pewna zmiana uporządkowania odległości może prowadzić do dendrytu różnego topologicznie, który nazywamy sąsiednim względem poprzedniego.

Łączny rozkład wszystkich odległości międzygatunkowych, lub krótko — łączny rozkład tablicy Czekanowskiego, zależy oczywiście od metryki przestrzeni cech, a więc od wzoru na odległość w tej przestrzeni, który staje się tu funkcją transformującą wektory losowe w odległości. Ta sama metryka określa rzeczywiste, tj. populacyjne odległości międzygatunkowe. W pracy tej będziemy się zajmować jedynie odległościami określonymi wzorem $d = \sum_{i=1}^n a_i |d_i|$, gdzie d_i są różnicami składowych wektorów gatunkowych, zaś a_i są dodatnimi współczynnikami normującymi skalę współrzędnych. Odległości te metryzują przestrzeń cech.

Intuicyjny sens omówionych wyżej poziomów ufności jest następujący: Załóżmy, że odległości międzygatunkowe, wychodzące z jednego punktu, są między sobą różne, a więc można teoretycznie, przez pobranie odpowiednio licznej próby łącznej, czyli odpowiednio licznej próby z każdej z populacji, tak zmniejszyć wariancję poszczególnych odległości, by z żądanym z góry prawdopodobieństwem otrzymać z próby dendryt równy topologicznie prawdziwemu. Jest jasne, że w praktyce może się to okazać niewykonalne; natomiast jest zawsze wykonalne, jeśli idzie o przynależność dendrytu z próby do określonej rodziny dendrytów. Świadczy to jedynie o pewnym faście przyrodniczym, mianowicie o tym, że pewne odległości w dendrycie rzeczywistym są niemal równe. Nawet wówczas, gdy możemy uzyskać, na zadanym z góry poziomie ufności, dendryt równy topologicznie prawdziwemu, warto podwyższyć poziom ufności, by otrzymać w ten sposób kilkoelementową rodzinę dendrytów sąsiednich, gdyż pozwala to lepiej zorientować się we wzajemnym położeniu (konstelacji) gatunków w rozpatrywanej wielowymiarowej przestrzeni cech, a tym samym lepiej poznać ich pokrewieństwa wzajemne.

W pracy omówimy najpierw ogólne zasady otrzymywania rozkładu pojedynczej odległości z próby, następnie łączny rozkład tablicy Czekanowskiego, której geometryczne własności określimy jako wielościanny „A”, a następnie podamy geometryczne odpowiedniki topologicznych własności dendrytów, które określimy jako wielościanny „B”. To pozwoli nam wykryć rodzinę dendrytów na zadanym poziomie ufności. Poważną trudność stanowi symbolika, z uwagi na wielką liczbę użytych pojęć. Stąd np. niektóre symbole występujące w twierdzeniach i ich dowodach mają tam inne znaczenie niż w pozostałym tekście.

Pragnę podziękować profesorowi dr Stefanowi Zubrzyckiemu za cenne uwagi i pomoc w opracowaniu ostatecznej wersji niniejszej pracy.

§ 2. Teoretyczna a empiryczna tablica Czekanowskiego. Tablice, o których mowa, wprowadził do badań biometrycznych J. Czekanowski (patrz np. [4]). Z matematycznego punktu widzenia są to macierze kwadratowe, w których numeracja zarówno kolumn, jak i wierszy odpowiada numeracji pewnego zbioru przedmiotów, między którymi określa się odległości (w sposób podany w § 1). Każdy element tych macierzy, czyli odległość d_{pq} , oznacza odległość między p -tym a q -tym przedmiotem; oczywiście $d_{pp} = 0$ oraz $d_{pq} = d_{qp}$. W niniejszej pracy będziemy rozróżniać tablice Czekanowskiego: teoretyczną i empiryczną. Tablica teoretyczna jest macierzą odległości między wektorami gatunkowymi, czyli przeciętnymi populacyj (patrz § 1), tablica empiryczna jest macierzą odległości ocenianych na podstawie próby, a więc jej elementy są funkcjami realizacji zmiennych losowych. Jeśli do zbudowania empirycznej tablicy Czekanowskiego użyjemy jednej próby łącznej, tzn. zawierającej pewną ilość elementów wybranych losowo i niezależnie z każdego z gatunków, to dla m gatunków otrzymamy oczywiście $k = \frac{1}{2}m(m-1)$ takich empirycznych odległości, czyli elementów d_{pq} . W dalszym ciągu będziemy się zajmować zmiennymi losowymi, nie zaś ich realizacjami; dlatego elementy d_{pq} będziemy traktować jako zmienne losowe, nie zaś konkretne liczby⁽²⁾. Jak zobaczymy w dalszej części pracy, zmienne d_{pq} są wzajemnie zależne stochastycznie, przy czym ta zależność wynika z faktu, iż w różnych elementach d występują te same zmienne losowe z próby.

Przyjęcie zasady, że zmienne losowe nie powtarzają się w odległościach empirycznych jest mniej korzystne dla metody obszarów ufności, omówionej w § 6 pracy.

Zaopatrzymy gatunki numerami r , gdzie $1 \leq r \leq m$, natomiast cechy numerami i , gdzie $1 \leq i \leq n$. Wobec tego, jeśli z r -tego gatunku pobierzemy l_r elementów próby, to j -ty osobnik, gdzie $1 \leq j \leq l_r$, będzie przez nas rozpatrywany jako wektor losowy n -wymiarowy $X_j^r = (X_{1,j}^r, X_{2,j}^r, \dots, X_{n,j}^r)$. W wyrażeniu tym, po prawej stronie równości, pierwszy wskaźnik na dole przy X oznacza numer cechy, drugi zaś numer osobnika r -tego gatunku.

Będziemy dalej zakładać, że X_j^r ma rozkład normalny:

$$(1) \quad f(x) = [|\mathbf{A}_r|(2\pi)^n]^{-1/2} \exp[-\frac{1}{2}(x - \mu^r)\mathbf{A}_r^{-1}(x - \mu^r)'].$$

We wzorze tym macierz kowariancji $\mathbf{A}_r = \Sigma_r \mathbf{P}_r \Sigma_r$, gdzie Σ_r jest macierzą diagonalną o elementach $\sigma_1^r, \sigma_2^r, \dots, \sigma_n^r$, czyli dyspersjach cech na

⁽²⁾ Małymi literami będziemy w dalszym ciągu oznaczać liczby stałe bądź zmienne rzeczywiste; wielkimi — zmienne losowe — z wyjątkiem d i s , które będą także oznaczać zmienne losowe.

przekątnej, zaś P_r jest macierzą korelacji cech o elementach $|c_{i_1, i_2}| < 1$ oraz jedynek na przekątnej. $|A_r|$ oznacza wyznacznik macierzy A_r .

Oznaczmy:

$$(2) \quad EX_j^r = (EX_{1,j}^r, EX_{2,j}^r, \dots, EX_{n,j}^r) = \mu^r = (\mu_1^r, \mu_2^r, \dots, \mu_n^r).$$

Odległość teoretyczną między gatunkami p oraz q , gdzie $1 \leq p, q \leq m$, określmy wobec tego, zgodnie z § 1, w sposób następujący:

$$(3) \quad \delta^{pq} = \sum_{i=1}^n a_i |\mu_i^p - \mu_i^q|.$$

Przyjmijmy, dla celów analitycznych, zamiast $|\mu_i^p - \mu_i^q|$ wyrażenie $\varepsilon_i^{p|q}(\mu_i^p - \mu_i^q)$, gdzie $\varepsilon_i^{p|q} = \text{sign}(\mu_i^p - \mu_i^q)$. Wówczas, jak widać, także nie musimy dbać o porządek wskaźników p, q w δ^{pq} , jeśli uwzględnimy związek

$$(4) \quad \delta^{pq} = \sum a_i \varepsilon_i^{p|q}(\mu_i^p - \mu_i^q) = \delta^{qp} = \sum a_i \varepsilon_i^{q|p}(\mu_i^q - \mu_i^p).$$

Dla zespołu m gatunków otrzymamy w ten sposób macierz elementów $\varepsilon_i^{p|q} = \varepsilon_{i,\kappa}$, gdzie wskaźniki podwójne, dotyczące par gatunków, przenumerowujemy na pojedyncze κ , przy czym $1 \leq \kappa \leq k = \frac{1}{2}m(m-1)$, gdyż tyle właśnie jest odległości łączących między sobą m elementów zbioru.

$$(5) \quad E = \begin{bmatrix} \varepsilon_{11} & \dots & \varepsilon_{1k} \\ \dots & \dots & \dots \\ \varepsilon_{n1} & \dots & \varepsilon_{nk} \end{bmatrix}.$$

Jak wiadomo, dla rozkładu normalnego średnia arytmetyczna jest z wielu względów optymalnym estymatorem wartości oczekiwanej w populacji. Niech zatem estymatorem wartości μ_i^r będzie

$$\bar{X}_i^r = \frac{1}{l_r} \sum_{j=1}^{l_r} X_{ij}^r;$$

stąd optymalnym estymatorem wartości δ^{pq} będzie

$$(6) \quad d^{pq} = \sum_{i=1}^n a_i \varepsilon_i^{p|q} (\bar{X}_i^p - \bar{X}_i^q).$$

Założenie nieujemności, konieczne dla odległości teoretycznych, dla ich estymatorów wydaje się zbędne. Z drugiej strony, normalność estymatorów jest potrzebna do uzyskania ilorazu Studenta (patrz § 3). Podstawiając $\mu^p - \mu^q = v^{p|q} = v_\kappa$, otrzymujemy z (6)

$$(7) \quad Ed_\kappa = \sum_{i=1}^n a_i \varepsilon_i^{p|q} v_i^{p|q} = \sum_{i=1}^n a_i \varepsilon_{i,\kappa} v_{i,\kappa} = \delta_\kappa.$$

Podobnie do macierzy E utworzymy macierz Z , również o wymiarach $n \times k$

$$(8) \quad Z = \begin{bmatrix} Z_{11} & \dots & Z_{1k} \\ \dots & \dots & \dots \\ Z_{n1} & \dots & Z_{nk} \end{bmatrix},$$

w której wektor kolumnowy $Z_x = Z^{p|q} = \bar{X}^p - \bar{X}^q$. Zwróćmy uwagę, że wskaźniki p, q są tu, podobnie jak przy współczynnikach ε , nieprzemienne, podczas gdy przy odległościach empirycznych d , podobnie jak to wynika dla δ z (4), są one przemienne.

Zatem znajomość macierzy E , która jest charakterystyką wzajemnego położenia danego zbioru gatunków w przestrzeni cech, pozwala skonstruować zespół optymalnych estymatorów o rozkładzie normalnym dla oceny odległości międzygatunkowych.

Wprowadźmy teraz zmienne losowe $Y_j^{p|q}$ podług formuły

$$(9) \quad Y_j^{p|q} = \sum_{i=1}^n a_i \varepsilon_i^{p|q} X_{ij}^p.$$

Oczywiście, $E Y_j^{p|q} = \bar{\mu}^{p|q} = \sum_{i=1}^n a_i \varepsilon_i^{p|q} \mu_i^p$. Zmienne te są rozłożone normalnie i mają określoną wariancję $\omega^{2p|q}$, której związek z macierzą A podaje lemat 3.2. Zatem $E(Y_j^{p|q} - \bar{\mu}^{p|q})^2 = \omega^{2p|q}$. Stąd, że $\varepsilon_i^{p|q} = -\varepsilon_i^{q|p}$ wynika podstawowy wzór

$$(10) \quad \delta^{pq} = \bar{\mu}^{p|q} + \bar{\mu}^{q|p}.$$

Dla zmiennych Y otrzymujemy

$$\begin{aligned} \text{LEMAT 2.1.} \quad & \text{Jeśli } \bar{Y}^{p|q} = \frac{1}{l_p} \sum_{j=1}^{l_p} Y_j^{p|q}, \text{ oraz } \bar{Y}^{q|p} = \frac{1}{l_q} \sum_{j=1}^{l_q} Y_j^{q|p}, \text{ to } d^{pq} = \\ & = \bar{Y}^{p|q} + \bar{Y}^{q|p}, \text{ przy czym } d^{pq} \text{ ma rozkład normalny } N\left(\delta^{pq}, \sqrt{\frac{\omega^{2p|q}}{l_p} + \frac{\omega^{2q|p}}{l_q}}\right). \end{aligned}$$

Dowód.

$$\begin{aligned} d^{pq} &= \sum_i a_i \varepsilon_i^{p|q} (\bar{X}_i^p - \bar{X}_i^q) = \sum_i a_i \varepsilon_i^{p|q} \left(\frac{1}{l_p} \sum_{j=1}^{l_p} X_{ij}^p - \frac{1}{l_q} \sum_{j=1}^{l_q} X_{ij}^q \right) = \\ &= \frac{1}{l_p} \sum_i a_i \varepsilon_i^{p|q} \sum_j X_{ij}^p + \frac{1}{l_q} \sum_i a_i \varepsilon_i^{q|p} \sum_j X_{ij}^q = \frac{1}{l_p} \sum_j \sum_i a_i \varepsilon_i^{p|q} X_{ij}^p + \\ &+ \frac{1}{l_q} \sum_j \sum_i a_i \varepsilon_i^{q|p} X_{ij}^q = \frac{1}{l_p} \sum_j Y_j^{p|q} + \frac{1}{l_q} \sum_j Y_j^{q|p} = \bar{Y}^{p|q} + \bar{Y}^{q|p}. \end{aligned}$$

Na podstawie (6) $E\bar{d}^{pq} = \delta^{pq}$, ponieważ zaś wariancję $\bar{Y}^{p|q}$ jest $\frac{\omega^{2p|q}}{l_p}$, a wariancję $\bar{Y}^{q|p}$ jest $\frac{\omega^{2q|p}}{l_q}$, to wariancję \bar{d}^{pq} jest $\frac{\omega^{2p|q}}{l_p} + \frac{\omega^{2q|p}}{l_q}$; c. n. d.

Nową trudność w rozpatrywanej tutaj metodzie stanowi fakt, iż w badaniach biometrycznych nie znamy na ogół macierzy E określonej przez (5). Znajomość średnich gatunkowych \bar{X}^r pozwala nam jednak zbudować macierz empiryczną \bar{E} , w sposób analogiczny

$$(11) \quad \bar{\varepsilon}_i^{p|q} = \text{sign} Z_i^{p|q},$$

a stąd

$$\bar{E} = \begin{bmatrix} \bar{\varepsilon}_{11} & \dots & \bar{\varepsilon}_{1k} \\ \dots & \dots & \dots \\ \bar{\varepsilon}_{n1} & \dots & \bar{\varepsilon}_{nk} \end{bmatrix}.$$

Na ogół $\bar{E} \neq E$, choć oczywiście dla każdego r , przy $l_r \rightarrow \infty$, prawdopodobieństwo $P(\varepsilon = \bar{\varepsilon})$ zbliża się do jedności dla każdego $\varepsilon_{i,x}$.

Przyjęcie współczynników $\bar{\varepsilon}_{i,x}$ otrzymanych z tej samej próby, zamiast $\varepsilon_{i,x}$, jest jednak równoznaczne z przyjęciem estymatorów

$$(12) \quad \bar{d}^{pq} = \sum_{i=1}^n a_i |\bar{X}_i^p - \bar{X}_i^q|,$$

zamiast estymatorów d^{pq} . Estymatory te nie mają rozkładu normalnego. Ich rozkład jest konsekwencją naszej nieznajomości prawdziwych współczynników $\varepsilon_i^{p|q}$, lub inaczej „nierozróżnialności” wskaźników p, q . Twierdzenia 1 i 2, podane niżej, wykorzystamy w § 3, we wzorach służących do oceny wartości δ^{pq} , za pomocą estymatorów \bar{d}^{pq} .

Oznaczmy przede wszystkim

$$E\bar{d}^{pq} = E\bar{d}_x = \bar{\delta}^{pq} = \sum_{i=1}^n a_i E\bar{Z}_{i,x} = \sum_{i=1}^n a_i \bar{v}_{i,x}, \quad \text{gdzie} \quad \bar{Z}_{i,x} = |Z_{i,x}|.$$

Poza tym, w dalszej części pracy $\Phi(x)$ będzie oznaczać dystrybuantę normalną, tj. rozkład $N(0, 1)$, zaś $\varphi(x) = \Phi'(x)$ — odpowiadającą jej gęstość. Rozważymy najpierw jednowymiarową zmienną normalną Z , taką, że $EZ = v$ oraz $E(Z - v)^2 = \sigma^2$; wprowadzimy funkcję ψ .

LEMAT 2.2. Jeśli zmienna losowa Z jest rozłożona normalnie $N(v, \sigma)$, to zmienna losowa $\bar{Z} = |Z|$ ma wartość oczekiwaną $\bar{v} = v + 2\psi(v, \sigma)$, gdzie

$$\psi(v, \sigma) = \sigma \varphi\left(\frac{v}{\sigma}\right) - v \Phi\left(-\frac{v}{\sigma}\right) > 0 \quad (3).$$

(3) Patrz w związku z tym F. C. Leone, *Technometrics*, vol. 3, No 4, 1961.

Dowód.

$$\begin{aligned}
 E\bar{Z} &= \int_{-\infty}^{\infty} \frac{1}{\sigma} |x| \varphi\left(\frac{x-\nu}{\sigma}\right) dx = \int_0^{\infty} \frac{1}{\sigma} x \varphi\left(\frac{x-\nu}{\sigma}\right) dx - \int_{-\infty}^0 \frac{1}{\sigma} x \varphi\left(\frac{x-\nu}{\sigma}\right) dx = \\
 &= \nu - 2 \int_{-\infty}^0 \frac{1}{\sigma} x \varphi\left(\frac{x-\nu}{\sigma}\right) dx = \nu - 2 \int_{-\infty}^{-\nu/\sigma} (\sigma z + \nu) \varphi(z) dz = \\
 &= \nu - 2 \left[\sigma \int_{-\infty}^{-\nu/\sigma} z \varphi(z) dz + \nu \int_{-\infty}^{-\nu/\sigma} \varphi(z) dz \right] = \\
 &= \nu - 2 \left[-\varphi\left(-\frac{\nu}{\sigma}\right) + \nu \Phi\left(\frac{\nu}{\sigma}\right) \right] = \nu + 2 \left[\sigma \varphi\left(\frac{\nu}{\sigma}\right) - \nu \Phi\left(-\frac{\nu}{\sigma}\right) \right].
 \end{aligned}$$

Pokażemy jeszcze metodą analityczną, że $\psi(\nu, \sigma) = \sigma \varphi\left(\frac{\nu}{\sigma}\right) - \nu \Phi\left(-\frac{\nu}{\sigma}\right) > 0$. Ponieważ $\lim_{\nu \rightarrow \infty} \nu \Phi\left(-\frac{\nu}{\sigma}\right) = 0$, gdyż $\frac{\Phi'(-\nu/\sigma)}{(\nu^{-1})'} = \frac{\nu^2}{\sigma} \varphi\left(\frac{\nu}{\sigma}\right) \rightarrow 0$, gdy $\nu \rightarrow \infty$, więc $\lim_{\nu \rightarrow \infty} \psi(\nu, \sigma) = 0$. Natomiast, jak łatwo się przekonać, $\psi'(\nu, \sigma) = -\Phi\left(-\frac{\nu}{\sigma}\right) < 0$, co dowodzi, że funkcja ψ jest dodatnia przy wszelkim rzeczywistym ν , c. n. d.

Jak widzimy, zmienna \bar{Z} ma średnią o $2\psi(\nu, \sigma)$ większą niż Z . Gdy $\nu > 0$, wartość ta jest mała i przy $\nu \rightarrow \infty$ zbliża się asymptotycznie do zera. Dla $\nu \rightarrow -\infty$ zbliża się asymptotycznie do wartości $-\nu$, a więc tu różnica między $E\bar{Z} = \bar{\nu}$ a $EZ = \nu$ jest duża i zbliża się asymptotycznie do wartości $2|\nu|$.

Wariancja zmiennej \bar{Z} jest zawsze mniejsza niż wariancja zmiennej Z . Wynika to stąd, iż $\bar{\sigma}^2 = E(\bar{Z} - \bar{\nu})^2 = E\bar{Z}^2 - \bar{\nu}^2 = EZ^2 - \nu^2 < EZ^2 - \nu^2 = \sigma^2$. Stąd łatwo otrzymujemy dla zmiennej losowej \bar{d} związek

$$(13) \quad E\bar{d} > Ed,$$

wynikający z definicji (6) i (7). Natomiast wariancja zmiennej \bar{d} nie musi być mniejsza niż wariancja zmiennej d . Dalsze własności średniej i wariancji zmiennych \bar{d} podają następujące dwa twierdzenia:

TWIERDZENIE 1. *Jeśli macierz kowariancji A składowych Z_i odległości empirycznej \bar{d} składa się z samych elementów dodatnich⁽⁴⁾, to wariancja statystyki \bar{d} jest mniejsza niż wariancja statystyki d , czyli*

$$E(\bar{d} - \bar{\delta})^2 < E(d - \delta)^2.$$

⁽⁴⁾ Jeśli macierze A_r składają się z samych elementów dodatnich, to i macierze A_{pq} (patrz lemat 2.3) mają tę własność. W badaniach biometrycznych można na ogół tak dobrać cechy, by wszystkie były między sobą dodatnio skorelowane.

Dowód.

$$\begin{aligned} E(\bar{d} - \bar{\delta})^2 &= E\left(\sum a_i \bar{Z}_i - \sum a_i \bar{v}_i\right)^2 = \sum_{i_1 i_2} a_{i_1} a_{i_2} E(\bar{Z}_{i_1} - \bar{v}_{i_1})(\bar{Z}_{i_2} - \bar{v}_{i_2}) = \\ &= \sum_i a_i \bar{\sigma}_i + \sum_{i_1 \neq i_2} a_{i_1} a_{i_2} E(\bar{Z}_{i_1} - \bar{v}_{i_1})(\bar{Z}_{i_2} - \bar{v}_{i_2}). \end{aligned}$$

Z drugiej strony,

$$E(d - \delta)^2 = \sum_{i_1 i_2} a_{i_1} a_{i_2} E(Z_{i_1} - v_{i_1})(Z_{i_2} - v_{i_2}) = \sum_{i=1}^n a_i^2 \sigma_i^2 + \sum_{i_1 \neq i_2} a_{i_1} a_{i_2} a_{i_1 i_2},$$

gdzie $a_{i_1 i_2}$ jest elementem macierzy A określonej w tezie, przy czym z założenia każdy ze składników drugiej sumy jest dodatni. Z uwagi na to, że jak podaliśmy poprzednio, dla każdego i zachodzi $\bar{\sigma}_i < \sigma_i$, wystarczy pokazać, iż dla każdej pary wskaźników $i_1 \neq i_2$ zachodzi nierówność

$$E(\bar{Z}_{i_1} - \bar{v}_{i_1})(\bar{Z}_{i_2} - \bar{v}_{i_2}) < E(Z_{i_1} - v_{i_1})(Z_{i_2} - v_{i_2}) = a_{i_1 i_2}.$$

Założymy w celu uproszczenia rachunków, że $v_{i_1} = v_{i_2} = 0$. W przypadku ogólnym rachunki bardzo się komplikują, jednakże dowód przebiega analogicznie. Wskaźniki i_1, i_2 zastąpimy przez 1, 2. Znajdujemy, że

$$\begin{aligned} E(\bar{Z}_1 - \bar{v}_1)(\bar{Z}_2 - \bar{v}_2) &= E(\bar{Z}_1 \bar{Z}_2) - \bar{v}_1 \bar{v}_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |z_1| |z_2| f(z_1, z_2) dz_1 dz_2 - \bar{v}_1 \bar{v}_2 = \\ &= 2 \int_0^{\infty} \int_0^{\infty} z_1 z_2 f(z_1, z_2) dz_1 dz_2 + 2 \int_0^{\infty} \int_0^{\infty} z_1 z_2 f(z_1, -z_2) dz_1 dz_2 - \bar{v}_1 \bar{v}_2, \end{aligned}$$

gdyż dzięki symetrii gęstości normalnej

$$f(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{z_1^2}{\sigma_1^2} - 2\frac{\rho z_1 z_2}{\sigma_1\sigma_2} + \frac{z_2^2}{\sigma_2^2}\right)\right],$$

mamy związek

$$\begin{aligned} \int_0^{\infty} \int_{-\infty}^0 |z_1 z_2| f(z_1, z_2) dz_1 dz_2 &= \int_{-\infty}^0 \int_0^{\infty} |z_1 z_2| f(z_1, z_2) dz_1 dz_2 = \\ &= \int_0^{\infty} \int_0^{\infty} z_1 z_2 f(z_1, -z_2) dz_1 dz_2. \end{aligned}$$

Ponieważ

$$EZ_1 Z_2 = 2 \int_0^{\infty} \int_0^{\infty} z_1 z_2 f(z_1, z_2) dz_1 dz_2 - 2 \int_0^{\infty} \int_0^{\infty} z_1 z_2 f(z_1, -z_2) dz_1 dz_2,$$

wystarczy wykazać, że zachodzi nierówność

$$4 \int_0^{\infty} \int_0^{\infty} z_1 z_2 f(z_1, -z_2) dz_1 dz_2 < \bar{v}_1 \bar{v}_2.$$

Gdy położymy $k = 2(1 - \varrho^2)$, całka powyższa da się zmajoryzować:

$$\begin{aligned}
& \int_0^\infty \int_0^\infty xy \exp \left[-\frac{1}{k} \left(\frac{x^2}{\sigma_1^2} + 2 \frac{\varrho xy}{\sigma_1 \sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right] dx dy = \\
& = \int_0^\infty x \int_0^\infty y \exp \left\{ -\frac{1}{k\sigma_2^2} \left[\left(y + \frac{\varrho\sigma_2}{\sigma_1} x \right)^2 + \frac{\sigma_2^2 x^2}{\sigma_1^2} - \frac{\varrho^2 \sigma_2^2}{\sigma_1^2} x^2 \right] \right\} dy dx = \\
& = \int_0^\infty x \exp \left[-(1 - \varrho^2) \frac{\sigma_2^2 x^2}{k\sigma_2^2 \sigma_1^2} \right] \int_0^\infty y \exp \left[-\frac{1}{k\sigma_2^2} \left(y + \frac{\varrho\sigma_2}{\sigma_1} x \right)^2 \right] dy dx = \\
& = - \int_0^\infty x \exp \left[-(1 - \varrho^2) \frac{x^2}{k\sigma_1^2} \right] \int_0^\infty \frac{-2y}{k\sigma_2^2} \exp \left[-\frac{1}{k\sigma_2^2} \left(y + \frac{\varrho\sigma_2}{\sigma_1} x \right)^2 \right] dy \frac{k\sigma_2^2}{2} dx = \\
& = - \int_0^\infty x \exp \left(-\frac{x^2}{2\sigma_1^2} \right) \left\{ \exp \left[-\frac{1}{k\sigma_2^2} \left(y + \frac{\varrho\sigma_2}{\sigma_1} x \right)^2 \right] \right\}_0^\infty + 2 \frac{\varrho\sigma_2 x}{k\sigma_2^2 \sigma_1} \times \\
& \quad \times \int_0^\infty \exp \left[-\frac{1}{k\sigma_2^2} \left(y + \frac{\varrho\sigma_2}{\sigma_1} x \right)^2 \right] dy \left\} \frac{k\sigma_2^2}{2} dx = \\
& = - \int_0^\infty x \exp \left(-\frac{x^2}{2\sigma_1^2} \right) \left\{ -\exp \left(-\frac{1}{k\sigma_2^2} \frac{\varrho^2 \sigma_2^2 x^2}{\sigma_1^2} \right) + 2 \frac{\varrho x \sqrt{2\pi}}{\sigma_1 k^{1/2}} (1 - \Phi) \right\} \frac{k\sigma_2^2}{2} dx = \\
& = \int_0^\infty x \exp \left(-\frac{x^2}{2\sigma_1^2} \right) \left\{ \frac{k\sigma_2^2}{2} \exp \left(-\frac{\varrho^2 x^2}{k\sigma_1^2} \right) + c \varrho x (\Phi - 1) \right\} dx = \\
& = \int_0^\infty \frac{k\sigma_2^2}{2} x \exp \left[-\left(\frac{1}{2\sigma_1^2} + \frac{\varrho^2}{k\sigma_1^2} \right) x^2 \right] dx + \int_0^\infty c \varrho x^2 \exp \left(\frac{-x^2}{2\sigma_1^2} \right) (\Phi - 1) dx = \\
& = \int_0^\infty (1 - \varrho^2)^2 \sigma_2^2 x \exp \left[-\frac{x^2}{2(1 - \varrho^2)\sigma_1^2} \right] dx + \int_0^\infty c \varrho x^2 \exp \left(\frac{-x^2}{2\sigma_1^2} \right) (\Phi - 1) dx = \\
& = -(1 - \varrho^2)^2 \sigma_1^2 \sigma_2^2 \int_0^\infty \frac{-x}{(1 - \varrho^2)\sigma_1^2} \exp \left[-\frac{x^2}{2(1 - \varrho^2)\sigma_1^2} \right] dx + \\
& \quad + \int_0^\infty c \varrho x^2 \exp \left(-\frac{x^2}{2\sigma_1^2} \right) (\Phi - 1) dx = \\
& = (1 - \varrho^2)^2 \sigma_1^2 \sigma_2^2 + \int_0^\infty c \varrho x^2 \exp \left(\frac{-x^2}{2\sigma_1^2} \right) (\Phi - 1) dx < (1 - \varrho^2)^2 \sigma_1^2 \sigma_2^2, \text{ gdzie } c > 0.
\end{aligned}$$

Druga całka znika, jeśli położymy zamiast Φ liczbę 1, przez co majoryzujemy wartość tej całki. Stąd, podstawiając otrzymaną wartość, dostajemy nierówność

$$4 \int_0^\infty \int_0^\infty z_1 z_2 f(z_1, -z_2) dz_1 dz_2 < \frac{4(1-\varrho^2)^2 \sigma_1^2 \sigma_2^2}{2\pi \sigma_1 \sigma_2 \sqrt{1-\varrho^2}} = \frac{2}{\pi} (1-\varrho^2)^{3/2} \sigma_1 \sigma_2.$$

Ponieważ na podstawie lematu 2.2

$$\bar{v}_1 \bar{v}_2 = \sigma_1 \frac{2}{\sqrt{2\pi}} \sigma_2 \frac{2}{\sqrt{2\pi}} = \frac{2}{\pi} \sigma_1 \sigma_2,$$

twierdzenie zostało udowodnione.

Uwaga. Należy tu zwrócić uwagę, że gdy $\varrho = 0$, w ostatniej nierówności należy położyć równość, gdyż druga całka, którą majoryzujemy, wówczas znika; stąd wówczas $E(\bar{Z}_1 - \bar{v}_1)(\bar{Z}_2 - \bar{v}_2) = 0$. W przypadku gdy $\varrho < 0$, nie można jej tak zmajoryzować.

TWIERDZENIE 2. Niech Z_i będą zmiennymi losowymi o rozkładzie normalnym $N(v_i, \sigma_i)$, a rozkład o dystrybuancie

$$\bar{F}(v_i) = \begin{cases} 0, & \text{gdy } v_i < -c_i, \\ \frac{v_i + c_i}{2c_i} & \text{gdy } -c_i < v_i < c_i, \\ 1, & \text{gdy } v_i > c_i > 0 \end{cases}$$

niech będzie rozkładem a priori parametrów v_i ; niech $R_0 = E(E\bar{d} - E\bar{d}|0)$ ⁽⁵⁾ oznacza oczekiwaną różnicę oczekiwanych wartości średniej statystyki absolutnej $\bar{d} = \sum_{i=1}^n \bar{Z}_i$, oraz średniej zwyczajnej $d = \sum_{i=1}^n Z_i$, pod warunkiem „0”, czyli pod warunkiem rozkładu a priori $\bar{F}(v_i)$ parametrów v_i , a $R_x = E(E\bar{d} - E\bar{d}|x)$ — taką samą oczekiwaną różnicę, lecz pod warunkiem „x”, gdzie $x = (x_1, x_2, \dots, x_n)$, czyli pod warunkiem rozkładów a priori parametrów v_i określonych przez:

$$\bar{F}(v_i) = \Phi\left(\frac{v_i - x_i}{\sigma_i}\right),$$

gdzie dyspersje σ_i są równe dyspersjom odpowiednich składowych Z_i statystyki \bar{d} . Wówczas R_0 daje się wyrazić wzorem

$$R_0 = \sum_{i=1}^n \frac{\sigma_i^2 + c_i^2}{2c_i} \left[\Phi\left(\frac{c_i}{\sigma_i}\right) - \Phi\left(-\frac{c_i}{\sigma_i}\right) \right] + \sum_{i=1}^n \sigma_i \varphi\left(\frac{c_i}{\sigma_i}\right).$$

(5) W omawianej tu metodzie nie badamy oczekiwanej straty kwadratowej, jak to się zwykle czyni w teorii funkcji decyzyjnych, w celu porównania dobroci estymatorów; byłoby to bowiem w tym przypadku trudne analitycznie. W § 3 wyjaśnimy znaczenie twierdzeń 1 i 2 dla porównania dobroci estymatorów \bar{d} i \bar{d} .

Stąd można napisać dla każdego i $R_0 = O(c_i)$, ponieważ drugi czynnik pierwszej sumy dąży szybko do 1, a druga suma do 0. Natomiast R_x daje się oszacować od góry funkcją

$$R_x < 2 \sum_{i=1}^n \left[\frac{\sigma_i}{\sqrt{2}} \varphi \left(\frac{x_i}{\sigma_i \sqrt{2}} \right) + \sigma_i \varphi \left(\frac{x_i}{\sigma_i} \right) - \frac{x_i}{\sigma_i} \Phi \left(-\frac{x_i}{\sigma_i} \right) \right].$$

Dla każdego składnika R_{x_i} tej sumy mamy $R_{x_i} = O(x_i^{-\theta})$, gdzie θ dodatnie.
Dowód.

1^o Obliczenie R_0 . Wartość wyrażenia $E(d|\nu) = \sum_{i=1}^n \nu_i$ wynika z definicji statystyki d , dla której przyjmujemy tu i ujemne wartości ν_i . Natomiast z lematu 2.2 wynika:

$$E(\bar{d}|\nu) = E \sum_{i=1}^n (\bar{Z}_i | \nu_i) = \sum_{i=1}^n E(\bar{Z}_i | \nu_i) = \sum_{i=1}^n \nu_i + 2 \sum_{i=1}^n \psi(\nu_i, \sigma_i).$$

Stąd znajdujemy:

$$R_0 = \int_{-c_1}^{c_1} \dots \int_{-c_n}^{c_n} 2 \sum_{i=1}^n \psi(\nu_i, \sigma_i) \prod_{i=1}^n (2c_i)^{-1} d\nu_i = \sum_{i=1}^n \frac{1}{c_i} \int_{-c_i}^{c_i} \psi(\nu_i, \sigma_i) d\nu_i.$$

Obliczamy całkę

$$\begin{aligned} \int_{-c}^c \psi(\nu, \sigma) d\nu &= \\ &= \sigma \int_{-c}^c \varphi \left(\frac{\nu}{\sigma} \right) d\nu - \int_{-c}^c \nu \Phi \left(-\frac{\nu}{\sigma} \right) d\nu = \sigma^2 \int_{-c/\sigma}^{c/\sigma} \varphi(y) dy + \sigma^2 \int_{-c/\sigma}^{c/\sigma} y \Phi(y) dy = \\ &= \sigma^2 \left[\Phi \left(\frac{c}{\sigma} \right) - \Phi \left(-\frac{c}{\sigma} \right) \right] + \frac{c^2}{2} \Phi \left(\frac{c}{\sigma} \right) - \frac{c^2}{2} \Phi \left(-\frac{c}{\sigma} \right) - \frac{\sigma^2}{2} \int_{-c/\sigma}^{c/\sigma} y^2 \varphi(y) dy = \\ &= \left(\sigma^2 + \frac{c^2}{2} \right) \left[\Phi \left(\frac{c}{\sigma} \right) - \Phi \left(-\frac{c}{\sigma} \right) \right] - \frac{\sigma^2}{2} \int_{-c/\sigma}^{c/\sigma} \varphi(y) dy + \frac{\sigma^2}{2} \int_{-c/\sigma}^{c/\sigma} [\varphi(y) dy - y^2 \varphi(y)] dy = \\ &= \frac{\sigma^2 + c^2}{2} \left[\Phi \left(\frac{c}{\sigma} \right) - \Phi \left(-\frac{c}{\sigma} \right) \right] + \frac{\sigma^2}{2} y \varphi(y) \Big|_{-c/\sigma}^{c/\sigma}. \end{aligned}$$

Podstawiając obliczoną całkę w wyrażenie na R_0 , otrzymujemy wzór wyszczególniony w tezie.

2^o Obliczenie R_x : Wyrażenia $E(d|\nu)$, oraz $E(\bar{d}|\nu)$ omówiliśmy przy obliczeniu R_0 . Stąd znajdujemy analogicznie

$$\begin{aligned} R_x &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} 2 \sum_{i=1}^n \psi(\nu_i, \sigma_i) \prod_{i=1}^n d\Phi\left(\frac{\nu_i - x_i}{\sigma_i}\right) = \\ &= 2 \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{1}{\sigma_i} \psi(\nu_i, \sigma_i) \varphi\left(\frac{\nu_i - x}{\sigma_i}\right) d\nu_i = \\ &= 2 \sum_{i=1}^n \int_{-\infty}^{\infty} \left[\varphi\left(\frac{\nu_i}{\sigma_i}\right) - \frac{\nu_i}{\sigma_i} \Phi\left(-\frac{\nu_i}{\sigma_i}\right) \right] \varphi\left(\frac{\nu_i - x_i}{\sigma_i}\right) d\nu_i. \end{aligned}$$

Jak wiemy na podstawie lematu 2.2, wartość funkcji zawartej w nawiasach prostokątnych jest zawsze dodatnia, przy czym dla $\nu > 0$ wartość funkcji $\nu\Phi\left(-\frac{\nu}{\sigma}\right)$ jest mała i dodatnia, natomiast dla $\nu < 0$ możemy napisać

$$\sigma\varphi\left(\frac{\nu}{\sigma}\right) - \nu\Phi\left(-\frac{\nu}{\sigma}\right) = \sigma\varphi\left(\frac{\nu}{\sigma}\right) - \nu + \nu\Phi\left(\frac{\nu}{\sigma}\right),$$

gdzie znowu funkcja $\nu\Phi\left(\frac{\nu}{\sigma}\right)$ jest mała co do bezwzględnej wartości i ujemna.

Stąd otrzymujemy aproksymację

$$R_x < 2 \sum_{i=1}^n \int_{-\infty}^{\infty} \varphi\left(\frac{\nu_i}{\sigma_i}\right) \varphi\left(\frac{\nu_i - x_i}{\sigma_i}\right) d\nu_i - 2 \sum_{i=1}^n \int_{-\infty}^0 \frac{\nu_i}{\sigma_i} \varphi\left(\frac{\nu_i - x_i}{\sigma_i}\right) d\nu_i.$$

Obie całki obliczamy już łatwo:

$$\begin{aligned} &\int_{-\infty}^{\infty} \varphi\left(\frac{\nu}{\sigma}\right) \varphi\left(\frac{\nu - x}{\sigma}\right) d\nu = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2} [\nu^2 + (\nu - x)^2]\right\} d\nu = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{\sigma^2} (\nu^2 - \nu x + \frac{1}{2}x^2)\right] d\nu = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{\sigma^2} [(\nu - \frac{1}{2}x)^2 + \frac{1}{4}x^2]\right\} d\nu = \\ &= \frac{\sigma}{\sqrt{4\pi}} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma\sqrt{2}}\right)^2\right] \int_{-\infty}^{\infty} \frac{\sqrt{2}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(\nu - \frac{1}{2}x\right)^2\right] d\nu = \frac{\sigma}{\sqrt{2}} \varphi\left(\frac{x}{\sigma\sqrt{2}}\right). \end{aligned}$$

$$\begin{aligned}
& \int_{-\infty}^0 \frac{v}{\sigma} \varphi\left(\frac{v-x}{\sigma}\right) dv = \\
& = -\sigma \int_{-\infty}^0 \frac{-v}{\sigma^2} \varphi\left(\frac{v-x}{\sigma}\right) dv = -\sigma \left[\int_{-\infty}^0 \frac{x-v}{\sigma^2} \varphi\left(\frac{v-x}{\sigma}\right) dv - \int_{-\infty}^0 \frac{x}{\sigma^2} \varphi\left(\frac{v-x}{\sigma}\right) dv \right] = \\
& = -\sigma \left[\varphi\left(\frac{v-x}{\sigma}\right) \Big|_{-\infty}^0 - \frac{x}{\sigma^2} \Phi\left(\frac{v-x}{\sigma}\right) \Big|_{-\infty}^0 \right] = -\sigma \varphi\left(\frac{x}{\sigma}\right) + \frac{x}{\sigma} \Phi\left(-\frac{x}{\sigma}\right).
\end{aligned}$$

Podstawiając obliczone całki w powyższą nierówność, otrzymujemy nierówność wyszczególnioną w tezie.

§ 3. Normalizacja i rozkład a posteriori wartości oczekiwanych.

Rozważmy dwie metody wyznaczania empirycznej tablicy Czekanowskiego, oparte na dwóch sposobach normowania tablicy teoretycznej. Możliwość istnienia różnych modeli, którą należy wyraźnie podkreślić, pochodzi stąd, iż odległość między dwoma punktami w przestrzeni cech, jak wspomnieliśmy w § 1, ma charakter konwencyonalny, a przy tym zależy nie tylko od metryki, którą przyjęliśmy w tej przestrzeni, ale też od skali poszczególnych współrzędnych. Jest jasne, że mierząc długość płatków kwiatowych w milimetrach, zaś długość całej rośliny, przypuśćmy, w metrach, byłoby bardzo nieestosowne dodawać, w celu wyliczenia odległości dwóch osobników, wielkości tak różnych rzędów. Należy zatem wprowadzić jakąś normalizację cech. Nie jest jednak jasne, jaką normalizację należy tu wprowadzić.

Dwie normalizacje, jakie tu rozważymy, to jest jedną opartą na średniej z wszystkich badanych gatunków, a drugą na ich wspólnej dyspersji, mają swój sens przyrodniczy, którego tu jednak nie będziemy dyskutować. Wybór drugiej normalizacji jest podyktowany przydatnością jej dla całej naszej metody, co zobaczymy w dalszym ciągu pracy⁽⁶⁾. Omówimy więc obie normalizacje, czyniąc jednak w obu przypadkach istotnie ograniczające założenie, że macierze kowariancji wszystkich rozpatrywanych gatunków są jednakowe:

$$(14) \quad A_1 = A_2 = \dots = A_m = A.$$

Odrzucenie tego założenia (np. po przeprowadzeniu i niepomyślnym wyniku testowania hipotezy równości wariancji i korelacji w danym, konkretnym zagadnieniu przyrodniczym), komplikuje dość poważnie rachunki, gdyż wówczas zmienne Y mają dla różnych wskaźników różne dyspersje, wobec czego potrzebny iloraz (25) nie będzie już miał

⁽⁶⁾ Inna jeszcze normalizacja stosowana przez przyrodników, mianowicie na rozstępie elementów skrajnych z próby, jest mniej uzasadniona, gdyż wariancja tego rozstępu jest duża.

rozkładu t Studenta, lecz, z odpowiednimi współczynnikami, rozkład d Sukhatme, zwany też rozkładem Behrensa-Fishera (patrz np. [7]). Wówczas wzory, które przedstawiamy w §§ 3 i 5 niniejszej pracy, musiałyby ulec znacznej modyfikacji i komplikacji. Poza tym, posługiwanie się tym rozkładem jest bardziej skomplikowane, a tablice statystyki d mało rozpowszechnione.

Wprowadziwszy założenie (14) omówimy naprzód ogólne własności normalizacji. Otóż zmienna losowa $a_i Z_i^{p|q}$, omówiona przy (8), ma rozkład normalny $N(a_i v_{i,x}, a_i \sigma_i \beta_x)$, gdzie $\beta_x^2 = \beta^{2pq} = l_p^{-1} + l_q^{-1}$.

LEMAT 3.1. *Normalizacja cech nie zmienia korelacji między cechami.*

Dowód.

$$\begin{aligned} \varrho'_{i_1 i_2} &= \frac{E[a_{i_1}(X_{i_1}^r - \mu_i^r) a_{i_2}(X_{i_2}^r - \mu_{i_2}^r)]}{\sqrt{E[a_{i_1}(X_{i_1}^r - \mu_{i_1}^r)]^2 E[a_{i_2}(X_{i_2}^r - \mu_{i_2}^r)]^2}} = \\ &= \frac{E(X_{i_1}^r - \mu_{i_1}^r)(X_{i_2}^r - \mu_{i_2}^r)}{\sqrt{E(X_{i_1}^r - \mu_{i_1}^r)^2 E(X_{i_2}^r - \mu_{i_2}^r)^2}} = \varrho_{i_1 i_2}. \end{aligned}$$

LEMAT 3.2. *Przy założeniu (14) zmienne losowe $Y^{p|q}$ i $Y^{q|p}$ mają tę samą wariancję: $E(Y^{p|q} - \bar{\mu}^{p|q})^2 = E(Y^{q|p} - \bar{\mu}^{q|p})^2 = \omega_x^2 = \omega^{2pq}$.*

Dowód.

$$\begin{aligned} E(Y^{p|q} - \bar{\mu}^{p|q})^2 &= E\left[\sum_{i=1}^n a_i \varepsilon_i^{p|q} (X_{ij}^p - \mu_i^p)\right]^2 = \\ &= E\left[\sum_{i_1, i_2} a_{i_1} a_{i_2} \varepsilon_{i_1}^{p|q} \varepsilon_{i_2}^{p|q} (X_{i_1 j}^p - \mu_{i_1}^p)(X_{i_2 j}^p - \mu_{i_2}^p)\right] = \\ &= \sum_{i_1, i_2} a_{i_1} a_{i_2} \varepsilon_{i_1}^{p|q} \varepsilon_{i_2}^{p|q} E(X_{i_1 j}^p - \mu_{i_1}^p)(X_{i_2 j}^p - \mu_{i_2}^p) = \\ &= \sum_{i_1, i_2} a_{i_1} a_{i_2} \varepsilon_{i_1}^{p|q} \varepsilon_{i_2}^{p|q} \lambda_{i_1 i_2} = \sum_{i_1, i_2} a_{i_1} a_{i_2} \varepsilon_{i_1}^{q|p} \varepsilon_{i_2}^{q|p} \lambda_{i_1 i_2} = E(Y^{q|p} - \bar{\mu}^{q|p})^2, \end{aligned}$$

przy czym $\lambda_{i_1 i_2}$ należy do macierzy A , c. n. d.

LEMAT 3.3. *Przy założeniu (14) macierz kowariancji A_{pq} zmiennych $a_i Z_i^{p|q}$ wyraża się wzorem macierzowym $A_{pq} = \beta^{2pq} \Theta A \Theta$, gdzie macierz Θ jest macierzą przekątną o elementach a_1, a_2, \dots, a_n na głównej przekątnej, a zerowych poza nią.*

Dowód. Opieramy się tu na prawie dodawania niezależnych zmiennych normalnych wielowymiarowych (patrz np. [1] i [2]). Zmienne $Z_i^{p|q}$ mają macierz kowariancji $\beta^{2pq} A$. Mnożenie ich przez współczynniki a_i nie zmienia korelacji, jak to wynika z lematu 3.1, natomiast powoduje mnożenie macierzy dyspersji Σ przez macierz Θ . Zatem $A_{pq} = (\Theta \Sigma) \times (\beta^{2pq} P)(\Theta \Sigma) = \beta^{2pq} \Theta (\Sigma P \Sigma) \Theta = \beta^{2pq} \Theta A \Theta$, c. n. d.

Uwaga. Z twierdzenia 1 wynika, że $\beta^{2pq} \bar{\omega}^{2pq} = E(\bar{d}^{pq} - \bar{\delta}^{pq})^2 \leq \sum_{i_1, i_2} a_{i_1 i_2} = \iota A_{pq} \iota'$ w symbolice macierzowej, gdzie $\iota = (1, 1, \dots, 1)$ jest wektorem n -wymiarowym, oraz gdy $a_{i_1 i_2} \geq 0$ przy wszelkich i_1, i_2 .

a) Normalizacja na średnią. Normalizację tę uzyskujemy przyjmując $a_i = k(\sum_{x=1}^k Z_{i,x})^{-1}$, przy czym zastrzegamy, że $\sum_{i=1}^k Z_{i,x} \neq 0$ ⁽⁷⁾. Ma ona zagwarantować, by wszystkie składniki odległości empirycznych d średnio w jednakowym stopniu wpływały na wartość tych odległości. Mamy bowiem wtedy związek $\sum_{x=1}^k a_i Z_{i,x} = k$ oraz przy ι określonym jak wyżej mamy związek macierzowy, którego uzasadnienie jest natychmiastowe:

$$a Z \left(\frac{1}{k} \iota \right)' = n.$$

Wzór ten może służyć do sprawdzania obliczeń.

Ta metoda normowania nie wartościuje cech ze względu na ich dyskryminacyjność⁽⁸⁾, tylko zrównuje skale wielkości w poszczególnych cechach. Przed obliczaniem wartości d tą metodą należałoby jednak zbadać dyskryminacyjność cech metodą, którą omówimy przy normalizacji na dyspersję. Dzięki temu usunie się cechy zbyt obciążające statystyki d_x swą dużą zmiennością.

b) Normalizacja na dyspersję. Przy tym normowaniu przyjmujemy $a_i = (s_i)^{-1}$, przy czym wyjaśnimy znaczenie symbolu s_i . Dzięki założeniu (14) wariancje wszystkich gatunków w i -tej cesze są równe $\sigma_i^2 = \lambda_{ii} \epsilon A$. Estymatorami tych wariancji są statystyki

$$(15) \quad s_{i,r}^2 = \frac{1}{l_r} \sum_{j=1}^{l_r} (X_{ij}^r - \bar{X}_i^r)^2.$$

Łącząc je, otrzymamy jako estymator wartości σ_i standardowe odchylenie ogólne

$$(16) \quad s_i^* = \sqrt{\sum_{r=1}^m l_r s_{i,r}^2 / \sum_{r=1}^m l_r}.$$

Te właśnie statystyki, złożone z wielu elementów próby, a więc mające bardzo małą wariancję, posłużą nam do normowania. Mo-

⁽⁷⁾ Zastąpienie stałych a średnimi, które są zmiennymi losowymi, budzi niewątpliwie zastrzeżenia; można jednak uważać, iż te realizacje statystyk o małej wariancji przyjęliśmy „raz na zawsze”.

⁽⁸⁾ Zagadnienie dyskryminacyjności, czyli diagnostyczności danej cechy, czy zespołu cech, wśród cech należących do danego zbioru, nie jest jeszcze dostatecznie wyjaśnione w statystyce.

głoby być przedmiotem dyskusji, czy nie lepiej zamiast s_i^* brać wartość $(s_i^*)^2$ (a zatem nie wyciągać pierwiastka kwadratowego) co poważnie zmniejsza błąd obliczeniowy. My jednak, ze względów praktycznych, użyjemy do tego celu statystyki

$$(17) \quad s_i = \sqrt{\sum_{r=1}^m l_r s_{i,r}^2 / \left(\sum_{r=1}^m l_r - m \right)}.$$

Jeśli bowiem przyjmiemy $a_i = s_i^{-1}$, to wyrażenie

$$(18) \quad a_i(Z_{i,x} - v_{i,x})\beta_x^{-1}$$

będzie miało rozkład t Studenta z $\gamma = \sum_{r=1}^m l_r - m$ stopniami swobody; wynika to z lematu 3.4 podanego na końcu niniejszego paragrafu.

Wyrażenie (18) daje nam od razu kryterium dyskryminacyjności i -tej cechy względem pary gatunków p, q . Mamy bowiem, przy zadanym poziomie ufności $1 - \alpha$ i realizacji $z_{i,x}$ zmiennej $Z_{i,x}$, nierówność

$$(19) \quad v_{i,x} > z_{i,x} - \frac{c_z}{a_i \beta_x},$$

przy czym c_z określone jest przez równość $\alpha = P(t_\gamma > c_z)$. Warunek ten jest wprawdzie dostateczny do odrzucenia cechy, lecz nie jest konieczny. Należy bowiem rozróżnić dyskryminacyjność bezwzględną, odnoszącą się do jednej odległości empirycznej $Z_{i,x}$, którą badamy za pomocą (19), oraz względną, określoną przez istotność różnicy dwóch odległości empirycznych Z_{x_1}, Z_{x_2} , lub ogólnie d_{x_1}, d_{x_2} , co omówimy w § 6. Dyskryminacyjność bezwzględna i względna, oraz kryterium (19), mają wielkie znaczenie dla metody poziomów ufności dendrytów, jak to zobaczymy w związku z twierdzeniami 6 i 7, w §§ 6 i 7.

Znajomość rozkładu wyrażenia (18) daje nam rozkład a priori potrzebny w twierdzeniu 2. W próbach bowiem o liczności przekraczającej 30, rozkład wyrażenia (18) można aproksymować przez normalny $N(0, 1)$. Natomiast w zagadnieniach empirycznych, dotyczących poziomów ufności w taksonomii wrocławskiej, liczności prób gatunkowych l_r muszą, z przyczyn wyłuszczonych w § 6, być duże, co najmniej po 10 elementów każda. Stąd widać, iż liczba stopni swobody γ w wyrażeniu (18) będzie na ogół znacznie przewyższać 30, co daje gwarancję dobrej aproksymacji.

Stąd, że wyrażenie (18) ma rozkład normalny $N(0, 1)$, oraz z lematu 3.5 wynika, iż wyrażenie $a_i Z_{i,x} \beta_x^{-1}$ ma w przybliżeniu wartość oczekiwaną $E(a_i Z_{i,x} \beta_x^{-1}) \approx \frac{v_{i,x}}{\sigma_i \beta_x}$, a więc pod tym warunkiem rozkład

normalny $N\left(\frac{\nu_{i,x}}{\sigma_i \beta_x}, 1\right)$, a wyrażenie $a_i Z_{i,x}$ — rozkład $N\left(\frac{\nu_{i,x}}{\sigma_i}, \beta_x\right)$. Niech teraz $S(x|\tau)$ oznacza gęstość warunkową zmiennej losowej, pod warunkiem τ . Jeśli dystrybuanta $F(\tau)$ rozkładu a priori parametru τ jest określona przez

$$(20) \quad F(\tau) = \begin{cases} 0, & \text{gdy } \tau < -c, \\ \frac{\tau+c}{2c}, & \text{gdy } -c < \tau < c, \\ 1, & \text{gdy } \tau > c > 0, \end{cases}$$

to gdy S jest funkcją zmiennej τ , całkowaną na prostej, otrzymujemy z twierdzenia Bayesa wzór na gęstość w rozkładzie a posteriori parametru

$$(21) \quad f(\tau|x) = \lim_{c \rightarrow \infty} \left(\frac{1}{2c} S(x|\tau) \Big/ \int_{-c}^c \frac{1}{2c} S(x|\tau) d\tau \right) = \bar{S}(\tau|x).$$

Wyrażenie (21) daje zatem rozkład a posteriori, gdy x jest realizacją zmiennej losowej X . Jeśli teraz przyjmiemy $\tau = \frac{\nu_{i,x}}{\sigma_i}$, to ze względu na powyższe rozumowanie będziemy $\tau_{i,x}$ uważać za średnią zmiennej normalnej $a_i Z_{i,x}$, a więc

$$S\left(a_i z_{i,x} \Big| \frac{\nu_{i,x}}{\sigma_i}\right) = \varphi\left(\frac{a_i z_{i,x} - \tau_{i,x}}{\beta_x}\right),$$

stąd zaś

$$\bar{S}(\tau_{i,x}|a_i z_{i,x}) = \varphi\left(\frac{\tau_{i,x} - a_i z_{i,x}}{\beta_x}\right),$$

czyli rozkład a posteriori jest $N(a_i z_{i,x}, \beta_x)$. Wyrażenia $a_i z_{i,x}$ oraz β_x grają rolę x_i oraz σ_i użytych w twierdzeniu 2. Stąd znajdujemy wzór umożliwiający oszacowanie od góry wartości oczekiwanej R_x , różnicy oczekiwanych wartości statystyk \bar{d} i d , pod warunkiem „z”, czyli pod warunkiem rozkładów a posteriori opartych na realizacji statystyk Z_1, Z_2, \dots, Z_n

$$(22) \quad R_{z,x} < 2 \sum_{i=1}^n \left[\frac{\beta_x}{\sqrt{2}} \varphi\left(\frac{a_i z_{i,x}}{\beta_x \sqrt{2}}\right) + \beta_x \varphi\left(\frac{a_i z_{i,x}}{\beta_x}\right) - \frac{a_i z_{i,x}}{\beta_x} \Phi\left(-\frac{a_i z_{i,x}}{\beta_x}\right) \right].$$

Oprócz tego wyznaczmy z tej samej próby standardowe odchylenia $\bar{s}_{p|q}$, zmiennych $\bar{Y}^{p|q}$, obliczanych przy pomocy (9) z macierzy E wzorem

$$(23) \quad \bar{s}_{p|q} = \sqrt{\frac{1}{l_p} \sum_{j=1}^n (\bar{Y}_j^{p|q} - \bar{\bar{Y}}^{p|q})^2}.$$

Gdybyśmy znali teoretyczną macierz znaków E , obliczylibyśmy zmienne $Y^{p|q}$, a stąd statystyki $s_{p|q}$, w ten sam sposób. Statystyka $\bar{s}_{p|q}^2$ jest estymatorem wariancji $\bar{\omega}^{2pq}$, zmiennej $\bar{Y}^{p|q}$, a $s_{p|q}^2$ jest estymatorem wariancji ω^{2pq} , zmiennej $Y^{p|q}$. Stąd statystyka \bar{d}^{pq} ma wariancję $\beta^{2pq}\bar{\omega}^{2pq}$, a statystyka d^{pq} wariancję $\beta^{2pq}\omega^{2pq}$.

Gdybyśmy chcieli zbadać na podstawie pobranej próby losowej w jakim stopniu nasz estymator \bar{d} ocenia dobrze wartość δ , musielibyśmy wyznaczyć jeszcze wariancję \bar{s}_r^2 , zmiennej U^r , obliczonej ze wzoru $U^r = \sum_{i=1}^n a_i \bar{X}_i^r$. Wariancja \bar{s}_r^2 tej zmiennej jest estymatorem wartości $\bar{\omega}_r^2$ i obliczamy ją ze wzoru analogicznego do (23). Gdy X_i są dodatnio skorelowane, zachodzi nierówność $\bar{\omega}^{2p,q} \leq \bar{\omega}_p^2 = \bar{\omega}_q^2$, $\beta^{2pq}\bar{\omega}^{2p,q} = \sum a_i$ (patrz lemat 3.3); stąd, gdy R^* oznacza prawą stronę (22), mamy

$$(24) \quad \frac{\sqrt{l_p + l_q R_{z,x}^*}}{E\sqrt{l_p \bar{s}_{p,q}^2 + l_q \bar{s}_{q,p}^2}} > \frac{\sqrt{l_p + l_q R_{z,x}^*}}{E\sqrt{l_p \bar{s}_p^2 + l_q \bar{s}_p^2}} > \frac{\sqrt{l_p + l_q R_{z,x}}}{E\sqrt{l_p \bar{s}_p^2 + l_q \bar{s}_q^2}}.$$

Gdy iloraz po lewej stronie podwójnej nierówności (24), który łatwo obliczyć, jest mały, powiedzmy mniejszy od jedności, to do estymowania wartości δ można użyć statystyki \bar{d} zamiast d , godząc się na niewielkie obciążenie dodatnie.

Streszczając wywody przypominamy, iż brak informacji co do macierzy E daje nam oczekiwaną różnicę R_0 między obliczaną statystyką \bar{d} a „prawidłowo” obliczaną d . Gdy jednak przez pobranie próby losowej otrzymamy pewną informację w postaci rozkładów a posteriori, to na tę różnicę otrzymamy wartość R_z oszacowaną wzorem (22). Gdy ta różnica jest mniejsza od jednej dyspersji, to estymator \bar{d} dobrze ocenia wartość δ . Wzrost liczebności próby łącznej zwiększa o tyle informację, że wtedy gdy $l_p \rightarrow \infty$ i $l_q \rightarrow \infty$, R_z dąży do zera szybciej, niż dowolna potęga naturalna l_p , l_q , czyli $R_z = o(l_p^{-\theta}) = o(l_q^{-\theta})$.

Do wyznaczania poziomów ufności dendrytów potrzebny będzie jeszcze iloraz studentowski

$$(25) \quad \frac{\bar{d}^{pq} - \delta^{pq}}{\sqrt{l_p \bar{s}_{p|q}^2 + l_q \bar{s}_{q|p}^2}} \sqrt{l_p l_q \left(1 - \frac{2}{l_p + l_q}\right)}.$$

Ma on rozkład t Studenta z $\gamma^{pq} = l_p + l_q - 2$ stopniami swobody. Wynika to bezpośrednio z lematu 3.4, który podajemy poniżej wraz z lematem 3.5.

LEMAT 3.4. *Gdy zmienna losowa Θ ma rozkład normalny $N(\nu, \beta\sigma)$, gdzie $\beta > 0$, a zmienna losowa H jest połączeniem wariancyj z prób losowych pobranych z m populacyj normalnych o tej samej dyspersji σ i o li-*

czeknościach l_r ($1 \leq r \leq m$) (jej pierwiastek jest więc postaci (16)), to iloraz $\frac{\Theta - \nu}{\sqrt{H}} \sqrt{\frac{\sum l_r - m}{\beta^2 \sum l_r}}$ ma rozkład t Studenta z $\gamma = \sum_{r=1}^m l_r - m$ stopniami swobody.

Dowód. Rozkład Studenta jest określony dla ilorazu dwóch zmiennych losowych, przy czym licznik ma rozkład normalny $N(0, \sigma)$, a mianownik jest postaci $\sqrt{\frac{1}{\gamma} \chi_\gamma^2}$, gdzie χ_γ^2 ma rozkład χ^2 z γ stopniami swobody, a więc jest sumą γ kwadratów zmiennych normalnych $N(0, \sigma)$. Stąd i z określenia zmiennej Θ wynika, iż licznik musi być postaci $(\Theta - \nu)\beta^{-1}$, a mianownik, gdy \sqrt{H} jest wyrażone wzorem (16), postaci $\sqrt{H \sum l_r / (\sum l_r - m)}$, c. n. d.

LEMAT 3.5. Jeśli zmienne losowe Θ oraz H są określone jak w lemacie 3.4, to iloraz $\frac{\Theta}{\sqrt{H}} \sqrt{\frac{\sum l_r - m}{\beta^2 \sum l_r}}$ ma w granicy, tzn. gdy $\sum l_r - m \rightarrow \infty$, rozkład normalny $N\left(\frac{\nu}{\beta\sigma}, 1\right)$.

Dowód. Iloraz wyszczególniony w tezie lematu można napisać

$$\frac{\Theta - \nu}{\sqrt{H}} \sqrt{\frac{\sum l_r - m}{\beta^2 \sum l_r}} + \frac{\nu}{\beta \sqrt{H \sum l_r / (\sum l_r - m)}}.$$

W tej sumie pierwszy składnik dąży, gdy $\sum l_r - m \rightarrow \infty$, do rozkładu $N(0, 1)$, co wynika z lematu 4.3; natomiast mianownik drugiego składnika dąży do liczby $\beta\sigma$, skąd na podstawie twierdzenia Śluckiego (patrz np. [2]) cały ten składnik dąży do wartości $\nu/\beta\sigma$. Suma więc tych dwóch składników daje w rezultacie, w granicy, rozkład $N(\nu/\beta\sigma, 1)$, c. n. d.

§ 4. Geometryczne własności teoretycznej tablicy Czekanowskiego. W tym paragrafie zajmujemy się geometryczną interpretacją teoretycznej tablicy Czekanowskiego. Składa się ona, jak już mówiliśmy, z $k = \frac{m(m-1)}{2}$

odległości, łączących punkty m -elementowego zbioru punktów w n -wymiarowej przestrzeni euklidesowej, według metryki (3). Stąd wynika fakt, że odległości te nie są niezależnymi zmiennymi rzeczywistymi, lecz że istnieją między nimi pewne związki uwarunkowane metryką. Rozważmy mianowicie przestrzeń euklidesową k -wymiarową \mathcal{D}_k , w której każda ze wspomnianych odległości będzie współrzędną punktu tej przestrzeni. Teoretyczną tablicę Czekanowskiego można interpretować w tej przestrzeni jako punkt. Zbadajmy, jaki zbiór w przestrzeni \mathcal{D}_k tworzą wszystkie możliwe tablice Czekanowskiego, określone przez zbiór m gatunków

w n -wymiarowej przestrzeni cech, z uwzględnieniem metryki (3). Zbiór taki oznaczmy przez \mathcal{A}_{mn} .

Zauważmy najpierw, że wszystkie odległości są nieujemne, wobec czego wystarczy rozpatrywać jedynie dodatni sektor układu współrzędnych, tzn. zbiór $(\zeta = (\zeta_1, \zeta_2, \dots, \zeta_k): \zeta_1 \geq 0, \zeta_2 \geq 0, \dots, \zeta_k \geq 0)$, przy czym przez \mathcal{U}_{mn} oznaczmy taki zbiór o najmniejszej liczbie wymiarów, zawierający \mathcal{A}_{mn} . Łatwo zauważyć, że wymiar tego sektora jest dokładnie równy k , $\dim \mathcal{U}_{mn} = k$; stąd można oznaczyć $\mathcal{U}_{mn} = \mathcal{U}_k$. Zbadajmy własności zbioru $\mathcal{A}_{mn} \subset \mathcal{U}_{mn}$.

Każda współrzędna δ_x ($1 \leq x \leq k$) wektora $\delta \in \mathcal{U}_k$, czyli tablicy Czekanowskiego, ma zmienność ograniczoną przez $k-1$ pozostałych zmiennych δ_x . Równanie, które określałoby tę zależność, powstałoby z pełnego układu równań typu (3). Zobaczymy nieco dalej, iż \mathcal{A}_{mn} jest pewnym wielościanem, który nazwiemy „wielościanem typu A ”.

Warto zwrócić uwagę, iż w trójwymiarowej przestrzeni \mathcal{U}_3 , czyli dla $m = 3$, współzależność trzech odległości, która sprowadza się do znanego prawa trójkąta

$$(26) \quad \zeta_1 + \zeta_2 \geq \zeta_3, \quad \zeta_2 + \zeta_3 \geq \zeta_1, \quad \zeta_3 + \zeta_1 \geq \zeta_2,$$

przedstawia się jako nieograniczony ostrosłup trójsieczny, oparty wierzchołkiem o początek układu współrzędnych, a krawędziami o przekątne ścian dodatniego sektora układu współrzędnych. Łatwo pokazać, że tak jest, gdyż szukany zbiór jest iloczynem mnogościowym trzech półprzestrzeni określonych nierównościami (26). Równania krawędzi znajdujemy kładąc $\zeta_3 = 0$, a stąd $\zeta_2 \geq \zeta_1$ i $\zeta_1 \geq \zeta_2$, czyli $\zeta_1 = \zeta_2$, a następnie $\zeta_2 = 0$, skąd $\zeta_1 = \zeta_3$, oraz $\zeta_1 = 0$, skąd $\zeta_1 = \zeta_3$. Podobnie gdy dwie zmienne są zerami, to i trzecia jest zerem.

Wielościan \mathcal{A}_{mn} w przypadku metryki (3) można przedstawić w postaci parametrycznej równaniem macierzowym

$$(27) \quad \mathbf{x} \mathbf{F}_i = \delta.$$

W równaniu tym \mathbf{x} oznacza wektor rzeczywisty, $m \times n$ -wymiarowy, postaci: $\mathbf{x} = (x_1^1, \dots, x_n^1, x_1^2, \dots, x_n^2, \dots, x_1^m, \dots, x_n^m)$. Grupy n -elementowe reprezentują tu wektory gatunkowe. $\delta = (\delta_1, \delta_2, \dots, \delta_k)$ jest wektorem reprezentującym teoretyczną tablicę Czekanowskiego. Macierz \mathbf{F}_i składa się z elementów $-1, 0, 1$. Gdy odpowiednia różnica dwóch x -ów, wynikająca ze wzoru (27), staje się ujemna, odpowiednie dwa elementy macierzy zmieniają znak; stąd zależność tej macierzy od wektora \mathbf{x} , wyrażona indeksem. Całe przekształcenie (27) jest równoważne zespołowi równań (3) bez współczynników a ; współrzędne x uważamy tu za unormowane.

Macierz Γ dzieli się w sposób naturalny na bloki

$$(28) \quad \Gamma = \begin{bmatrix} E_{11} & \dots & E_{1,m-1} \\ \dots & \dots & \dots \\ E_{m1} & \dots & E_{m,m-1} \end{bmatrix}.$$

Jak widać z określenia przekształcenia (27), wszystkie bloki mają liczbę wierszy równą n , natomiast co do liczby kolumn obowiązuje reguła: liczba kolumn w bloku E_{ij} równa się $m-j$. Wskaźniki i, j mają tu inne znaczenie niż w § 2 i odnoszą się tylko do omawianych bloków. Bloki naprzekątne, czyli typu E_{ii} , składają się z samych elementów ε , tzn. -1 , lub 1 ; bloki, dla których $j > i$, składają się z samych zer; w blokach, dla których $i > j$, $(i-1)$ -sza kolumna składa się z elementów ε o znakach przeciwnych niż odpowiednia kolumna bloku naprzekątnego o tym samym wskaźniku i , a pozostałe kolumny są złożone z zer. To określenie wyznacza oczywiście również ogólną liczbę wierszy i kolumn w macierzy Γ .

Zmieniając, zgodnie z (27), znaki elementów ε w blokach naprzekątnych, możemy je tak dobrać, by rząd macierzy Γ_ε był maksymalny; odpowiada to pewnemu obszarowi zmienności wektora ε . Jak wiadomo (patrz np. [1] i [2]), ten maksymalny rząd równa się wymiarowi wielościanu \mathcal{A}_{mn} . Stąd, znając maksymalny rząd macierzy Γ przy m gatunkach i n cechach, znamy wymiar wielościanu \mathcal{A}_{mn} . Twierdzenie 3 podaje oszacowanie tego wymiaru poprzez oszacowanie maksymalnego rzędu macierzy Γ .

TWIERDZENIE 3. *Wymiar zbioru \mathcal{A}_{mn} przy metryce (3) daje się oszacować, w zależności od wielkości liczby cech n i liczby gatunków m , w sposób następujący:*

- 1° gdy $m-1 \leq n$, to $\dim \mathcal{A}_{mn} = k = \frac{1}{2}m(m-1)$;
- 2° gdy $\frac{1}{2}(m-1) \leq n < m-1$, to $mn - \frac{1}{2}n(n+1) \leq \dim \mathcal{A}_{mn} \leq k$;
- 3° gdy $n < \frac{1}{2}(m-1)$, to $mn - \frac{1}{2}n(n+1) \leq \dim \mathcal{A}_{mn} \leq mn < k$.

Zakładamy tu ogólnie, że m i n są liczbami naturalnymi.

Dowód. Rozpatrzmy kolejno trzy przypadki wyszczególnione w tezie, określając w każdym przypadku liczbę ograniczającą wymiar zbioru \mathcal{A}_{mn} : od dołu — M_d , oraz od góry — M_g . Liczbę M_d określimy jako sumę maksymalnych rzędów bloków naprzekątnych macierzy Γ , a więc bloków $E_{1,1}, E_{2,2}, \dots$, we wzorze (28), zaś $M_g = \min[mn, \frac{1}{2}m \times (m-1)]$, mniejszy z rozmiarów macierzy Γ .

Ad 1°. Gdy $m-1 \leq n$, to można zawsze za mniejszy z rozmiarów bloków naprzekątnych przyjąć liczbę kolumn w tych blokach. Stąd $M_d = \sum_{i=1}^{m-1} i = \frac{1}{2}m(m-1)$. Również, zgodnie z definicją, $M_g = \frac{1}{2}m(m-1)$.

Ad 2°. Gdy $n < m-1$, to mamy $m-1-n$ bloków naprzekątnych o mniejszym rozmiarze n wierszy oraz pozostałe n bloków o mniejszym rozmiarze równym liczbie kolumn. Stąd $M_d = n(m-1-n) + \sum_{i=1}^n i = mn - \frac{1}{2}n(n+1)$. Oprócz tego, jeśli $\frac{1}{2}(m-1) \leq n$, to oczywiście $M_d = \frac{1}{2}m(m-1)$.

Ad 3°. Gdy $n < \frac{1}{2}(m-1)$, to $M_d = mn$, a także $mn < \frac{1}{2}m \times (m-1) = k$. Ponieważ wtedy także $n < m-1$, więc dolne ograniczenie jest takie jak w przypadku 2°.

Dla zakończenia dowodu wykazemy jeszcze, że przy wszelkich całkowitych m, n , nierówność $mn - \frac{1}{2}n(n+1) \leq \frac{1}{2}m(m-1)$ jest zawsze spełniona. Niech bowiem $\tau = n-m$; wówczas $m(m+\tau) - \frac{1}{2}(m+\tau) \times (m+\tau+1) = \frac{1}{2}(m+\tau)[m-(\tau+1)] = \frac{1}{2}[m^2 - m - \tau(\tau+1)]$. Ponieważ jednak $\tau(\tau+1) \geq 0$, przy czym równość zachodzi tylko dla $\tau = -1$, oraz $\tau = 0$, to nierówność jest spełniona. Tym samym udowodniliśmy twierdzenie.

Biorąc pod uwagę fakt, iż zbiór \mathcal{A}_{mn} , który jest opisany analitycznie wzorem (27) w postaci parametrycznej, da się opisać analitycznie układem nierówności, możemy powiedzieć, iż \mathcal{A}_{mn} jest wielościanem nieograniczonym, zawartym w dodatnim sektorze \mathcal{U}_k k -wymiarowego układu współrzędnych kartezjańskich. Jeśli idzie o związek macierzy I' z macierzą E , określoną przez (5), to widać, że każda omówiona przy (27) zmiana znaków macierzy I' daje inną możliwą macierz E ; wśród nich są też macierze empiryczne \bar{E} . Na odwrót, macierz E określa odpowiadającą jej macierz I' . Do tego zagadnienia powrócimy jeszcze w § 5.

Jak widzimy (pomijając przypadek 2°), w przypadku 3°, gdy $n < \frac{1}{2} \times (m-1)$, wymiar wielościanu \mathcal{A}_{mn} jest mniejszy niż wymiar k sektora \mathcal{U}_k , do którego należy. Jednak nawet wówczas, gdy jest on również wymiaru k , nie wypełnia całego sektora \mathcal{U}_k . Wynika to z następującego twierdzenia:

TWIERDZENIE 4. *Przy wszelkich naturalnych m, n , jeśli przez mr_k oznaczmy miarę k -wymiarową zbioru ($k = \frac{1}{2}m(m-1)$), zachodzi nierówność*

$$mr_k(\mathcal{U}_k \setminus \mathcal{A}_{mn}) \neq 0.$$

Dowód. W przypadku, gdy $m > n+1$, twierdzenie jest oczywiste. W przypadku, gdy $m \leq n+1$, oznaczmy przez $\mathcal{U}_{i,3}$ trójwymiarowe sektory dodatnie, czyli zbiory typu (ζ : $\zeta_{\kappa_1} = 0, \dots, \zeta_{\kappa_{i-1}} = 0, \zeta_{\kappa_i} \geq 0, \zeta_{\kappa_{i+1}} \geq 0, \zeta_{\kappa_{i+2}} \geq 0$). Każdy taki sektor można traktować jako rzut przestrzeni \mathcal{U}_k . Lecz sektory $\mathcal{U}_{i,3}$ zawierają w sobie rzuty wielościanu \mathcal{A}_{mn} , które oznaczmy $\mathcal{A}_{i,3}$. Rzuty te są, jak już wiemy, trójkątami nieograniczonymi, określonymi prawem trójkąta (26). Stąd $mr_3(\mathcal{U}_{i,3} \setminus \mathcal{A}_{i,3}) \neq 0$ dla każdego i ,

gdzie i jest numerem dowolnej kombinacji trzech indeksów spośród k . Jeśli teraz będziemy tworzyć iloczyny kartezjańskie zbiorów $\mathcal{U}_{i,3} \setminus \mathcal{A}_{i,3}$, przy czym wskaźnik i będzie przebiegał wszystkie trójki, to ich iloczyn katezjański $\bar{\mathcal{A}}_k$, należący oczywiście do \mathcal{U}_k , będzie miary k różnej od 0 i $\bar{\mathcal{A}}_k \subset \mathcal{U}_k \setminus \mathcal{A}_{mn}$; c. n. d.

§ 5. Własności rozkładu empirycznej tablicy Czekanowskiego.

Obecnie zajmujemy się łącznym rozkładem empirycznych odległości międzygatunkowych d_1, d_2, \dots, d_k , pod warunkiem, że ustalona jest macierz znaków E , określona przez (5). Jak powiedzieliśmy w § 4, ustalenie macierzy E odpowiada ustaleniu znaków macierzy Γ'_i określonej przez (27), co odpowiada w konsekwencji ustaleniu przynależności wektora ξ do pewnego obszaru liniowości funkcji $\xi \Gamma'_i$. Wynika to stąd, że chcemy zachować warunek nieujemności odległości teoretycznych. Jeśli dla odległości empirycznych odrzucimy ten warunek, to możemy we wzorze (27) zastąpić δ przez $d = (d_1, d_2, \dots, d_k)$, a wektor rzeczywisty ξ przez wektor losowy $\mathcal{X} = (\bar{X}_1^1, \dots, \bar{X}_n^1, \bar{X}_1^2, \dots, \bar{X}_n^2, \dots, \bar{X}_1^m, \dots, \bar{X}_n^m)$, ustalając przy tym macierz Γ_E . Stąd mamy

LEMAT 5.1. *Łączny rozkład empirycznej tablicy Czekanowskiego, czyli wektora d , jest normalny, jeśli wszystkie wchodzące w jej budowę gatunki mają rozkłady normalne.*

Dowód wynika z twierdzenia, że każda transformacja liniowa zmiennych normalnych jest też rozłożona normalnie (patrz w tej sprawie [1], [2], [7]).

W § 3 (lemat 3.2) podano formułę obliczania wariancji $\beta^{2pq} \omega^{2pq}$ zmiennej losowej d^{pq} . Analogicznie można obliczyć kowariancje odległości d^{pq_1}, d^{pq_2} , wychodzących z jednego gatunku (empirycznie: wspólnego punktu losowego):

LEMAT 5.2. *Kowariancja odległości d^{pq_1}, d^{pq_2} , wychodzących z jednego gatunku, wyraża się wzorem*

$$l_p^{-1} \omega^{p|q_1 q_2} = l_p^{-1} \eta^{p|q_1} \Lambda (\eta^{p|q_2})' = l_p^{-1} \eta^{p|q_2} \Lambda (\eta^{p|q_1})',$$

gdzie wektor $\eta^{p|q_i} = (a_1 \varepsilon_1^{p|q_i}, a_2 \varepsilon_2^{p|q_i}, \dots, a_n \varepsilon_n^{p|q_i})$.

Dowód.

$$\begin{aligned} E(d^{pq_1} - \delta^{pq_1})(d^{pq_2} - \delta^{pq_2}) &= \\ &= E[(\bar{Y}^{p|q_1} - \bar{\mu}^{p|q_1}) + (\bar{Y}^{q_1|p} - \bar{\mu}^{q_1|p})][(\bar{Y}^{p|q_2} - \bar{\mu}^{p|q_2}) + (\bar{Y}^{q_2|p} - \bar{\mu}^{q_2|p})] = \\ &= E(\bar{Y}^{p|q_1} - \bar{\mu}^{p|q_1})(\bar{Y}^{p|q_2} - \bar{\mu}^{p|q_2}), \end{aligned}$$

gdyż trzy dalsze kowariancje znikają, z uwagi na niezależność odpowiednich par zmiennych losowych. Stąd otrzymujemy dalej

$$\begin{aligned} E(\bar{Y}^{p|q_1} - \bar{\mu}^{p|q_1})(\bar{Y}^{p|q_2} - \bar{\mu}^{p|q_2}) &= \\ &= E\left[\sum_{i=1}^n a_i \varepsilon_i^{p|q_1}(\bar{X}^p - \mu_i^p)\right]\left[\sum_{i=1}^n a_i \varepsilon_i^{p|q_2}(\bar{X}_i^p - \mu_i^p)\right] = \\ &= \sum_{i_1, i_2} [a_{i_1} \varepsilon_{i_1}^{p|q_1} a_{i_2} \varepsilon_{i_2}^{p|q_2} E(\bar{X}_{i_1}^p - \mu_{i_1}^p)(\bar{X}_{i_2}^p - \mu_{i_2}^p)] = l_p^{-1} \eta^{p|q_1} A (\eta^{p|q_2})'. \end{aligned}$$

Czynnik liczebności próby l_p^{-1} powstaje tu analogicznie jak w lemacie 3.3. Przemienność $\eta^{p|q_1}$, $\eta^{p|q_2}$ wynika z symetryczności macierzy A . c. n. d.

Lematy 3.2 i 5.2 pozwalają więc obliczać ze wspólnej macierzy kowariancji cech A , macierzy znaków E , oraz wektora normującego a , macierz kowariancji łącznego rozkładu odległości empirycznych \mathfrak{L} . Zbierając te wyniki mamy więc

$$\begin{aligned} E(\bar{d}^{pq} - \delta^{pq})^2 &= \beta^{2pq} \eta^{p|q} A (\eta^{p|q})' = \beta^{2pq} \omega^{2pq}; \\ (29) \quad E(\bar{d}^{p_1 q_1} - \delta^{p_1 q_1})(\bar{d}^{p_2 q_2} - \delta^{p_2 q_2}) &= l_p^{-1} \eta^{p_1 | q_1} A (\eta^{p_2 | q_2})' = l_p^{-1} \omega^{p_1 q_1 q_2}; \\ E(\bar{d}^{p_1 q_1} - \delta^{p_1 q_1})(\bar{d}^{p_2 q_2} - \delta^{p_2 q_2}) &= 0; \quad p_1, q_1 \neq p_2, q_2. \end{aligned}$$

Macierz \mathfrak{L} jest rozmiarów $k \times k$ i ma elementy na przekątnej pomnożone przez liczby β^{2pq} , zaś pozostałe elementy pomnożone przez liczby l_r^{-1} , gdzie wskaźnik r odpowiada gatunkowi, z którego dana para odległości wychodzi. Można też rozważać elementarną macierz \mathfrak{L}_e dla odległości elementarnych, czyli powstałych z prób jednoelementowych. W tej macierzy czynniki l_r i β^{pq} nie występują, a zamiast β^{pq} jest $\sqrt{2}$.

Jak łatwo zauważyć, kowariancje typu $\omega^{p_1 q_1 q_2}$ mogą być ujemne, nawet gdy macierz A składa się z samych dodatnich elementów. By się o tym przekonać, wystarczy przyjąć we wzorze z lematu 5.2 $\lambda_{i_1 i_2} > 0$ oraz $\varepsilon_{i_1}^{p_1 | q_1} = 1$ i $\varepsilon_{i_2}^{p_2 | q_2} = -1$ przy każdym i .

Jeśli oznaczymy przez \bar{A} macierz kowariancji wektora \mathfrak{X} , to z uwagi na wzór $\zeta = \mathfrak{r}' \mathbf{I}'_E$, przekształcający formę kwadratową $\mathfrak{r}' \bar{A}^{-1} \mathfrak{r}$, otrzymujemy związek $\mathfrak{L} = \mathbf{I}'_E \bar{A} \mathbf{I}'_E$. Stąd widać, iż w \mathfrak{L} można wyróżnić $m-1$ minorów głównych, odpowiadających kolumnom macierzy blokowej (28). Kowariancje zerowe z trzeciej równości (29) znajdują się tylko poza tymi minorami i to po $\frac{1}{2}(m-2)(m-3)$ w każdym wierszu i kolumnie macierzy \mathfrak{L} .

TWIERDZENIE 5. *Gdy gatunki mają rozkłady normalne z warunkiem (14), gęstość łącznego rozkładu tablicy Czekanowskiego wyraża się wzorem*

$$(30) \quad f(\zeta) = [|\mathfrak{L}|(2\pi)^n]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\zeta - \delta)\mathfrak{L}^{-1}(\zeta - \delta)'\right],$$

przy czym elementy macierzy \mathfrak{L} wyrażają się wzorami (29), a $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_k)$ oznacza zmienny wektor rzeczywisty k -wymiarowy.

Dowód. Korzystamy tu z lematów 3.2, 5.1, 5.2, oraz z tego, że rozkład normalny jest określony przez swe pierwsze i drugie momenty.

Wniosek. Rząd rozkładu (30) równa się rzędowi macierzy przekształcenia liniowego Γ_E , wektora losowego \mathfrak{X} , ten zaś jest niewiekszy od wymiaru zbioru \mathcal{A}_{mn} , który określiliśmy w twierdzeniu 3.

§6. Poziomy ufnosci dendrytów. W niniejszym paragrafie założymy, podobnie jak w § 3, normalność rozkładów gatunkowych z warunkiem (14), oraz znajomość teoretycznej macierzy znaków E . Jak już pokazaliśmy w § 3, każda zmienna d_x ma wtedy rozkład normalny $N(\delta_x, \beta_x \omega_x)$, natomiast każda zmienna $s_x^2 = s_{pq}^2 = l_p s_{p|q}^2 + l_q s_{q|p}^2$ (patrz wzór (25)), ma rozkład χ^2 z $\gamma_x = l_p + l_q - 2$ stopniami swobody. Stąd, gdy przyjmiemy $v^{pq} = \sqrt{l_p l_q \left(1 - \frac{2}{l_p + l_q}\right)} = v_x$, zmienna

$$(31) \quad \mathcal{E}_x = \frac{d_x - \delta_x}{s_x} v_x$$

ma rozkład t Studenta z γ_x stopniami swobody. Oznaczmy przez $1 - \alpha$ poziom ufnosci, który ustalamy z góry dla dendrytów. Jest on równy prawdopodobieństwu, z jakim zachodzi relacja $\delta \in \mathcal{S}_x$, gdzie obszar losowy \mathcal{S}_x zależy od wektora losowego \mathfrak{X} , wprowadzonego na początku § 5 i należy do przestrzeni euklidesowej k -wymiarowej \mathcal{X}_k , zmiennych \mathcal{E}_x ($1 \leq x \leq k$).

Na podstawie twierdzenia 6 otrzymujemy, przyjmując za c_x ($1 \leq x \leq k$) takie liczby nieujemne, że

$$(32) \quad P(|\mathcal{E}_x| \geq c_x) = \alpha_x,$$

nierówność

$$(33) \quad 1 - \alpha = 1 - \sum_{x=1}^k \alpha_x < P(|\mathcal{E}_1| < c_1, |\mathcal{E}_2| < c_2, \dots, |\mathcal{E}_k| < c_k).$$

Stąd, jeśli w przestrzeni \mathcal{X}_k umieścimy początek układu współrzędnych w punkcie $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k)$, to zgodnie z metodą obszarów ufnosci Neymana będzie

$$\mathcal{S}_x = (\xi \in \mathcal{X}_k: |\xi_1| < c_1, |\xi_2| < c_2, \dots, |\xi_k| < c_k),$$

przy czym składowe wektora \mathcal{E} są obliczone zgodnie ze wzorem (31). W praktyce będziemy się starać, by wartości c_x były równe, tworząc w przestrzeni \mathcal{X}_k kostkę k -wymiarową \mathcal{S}_c o krawędzi $2c$.

Rozważmy teraz układ nierówności zachodzących między k zmiennymi rzeczywistymi $\zeta_x = \zeta^{pq}$ w przestrzeni euklidesowej k -wymiarowej

\mathcal{D}_k , przy czym wskaźnik $\kappa = 1, 2, \dots, k$ odpowiada pojedynczej numeracji z § 2, a pq — przemiennej numeracji podwójnej z tegoż paragrafu; $p, q = 1, 2, \dots, m$. Układ ten określimy indukcyjnie:

$$\begin{aligned}
 &1. \quad \zeta^{1q_1} < \zeta^{1j}, \quad 1 \leq j \leq m; \quad j \neq 1, q_1. \\
 &2. \quad \zeta^{i'q_2} < \zeta^{ij}, \quad q_2' = 1, q_1; \quad i = 1, q_1; \quad i \neq q_2', \\
 (34) \quad &1 \leq q_2 \leq m; \quad 1 \leq j \leq m; \\
 &q_2 \neq 1, q_1; \quad j \neq 1, q_1, q_2. \\
 &\dots\dots\dots \\
 &h. \quad \zeta^{q_h'q_h} < \zeta^{ij}, \quad q_h' = 1, q_1, \dots, q_{h-1}; \quad i = 1, q_1, \dots, q_{h-1}; \quad i \neq q_h', \\
 &1 \leq q_h \leq m; \quad 1 \leq j \leq m; \\
 &q_h \neq 1, q_1, \dots, q_{h-1}; \quad j \neq 1, q_1, \dots, q_{h-1}, q_h.
 \end{aligned}$$

Ten układ nierówności jest w przestrzeni \mathcal{D}_k zmiennych ζ odpowiednikiem określonej struktury topologicznej dendrytu zbudowanego metodą wrocławską⁽⁹⁾. Układ (34) interpretujemy krótko w ten sposób, że wychodząc z dowolnego punktu (tutaj gatunku), dołączamy do już zbudowanego dendrytu punkt najbliższy.

System nierówności (34) wyznacza w przestrzeni \mathcal{D}_k wielościan wypukły; jest on bowiem iloczynem mnogościowym półprzestrzeni. Taki elementarny wielościan będziemy nazywać „wielościanem typu B” i będziemy oznaczać przez \mathcal{B} . Jeśli więc zachodzi relacja $\delta \in \mathcal{B} \subset \mathcal{D}_k$, oznacza to, iż wektor δ , przedstawiający teoretyczną tablicę Czekanowskiego, wyznacza dendryt o określonej strukturze topologicznej.

Wielościanów \mathcal{B} jest tyle, ile różnych topologicznie dendrytów łączących m punktów w przestrzeni o wymiarze większym niż jeden. Łatwo zauważyć, że rodzina wszystkich możliwych wielościanów \mathcal{B} pokrywa całą przestrzeń \mathcal{D}_k . Jeśli bowiem zbiór wszystkich różnych topologicznie dendrytów pokrywa wszystkie połączenia m punktów, czyli całą tablicę Czekanowskiego, to odpowiadający mu zbiór systemów nierówności (34) zawiera wszystkie możliwe nierówności, a więc wszystkie możliwe półprzestrzenie, na jakie dzielimy przestrzeń \mathcal{D}_k .

Niech zatem \mathcal{C} oznacza sumę mnogościową pewnej liczby wielościanów \mathcal{B} , a \mathcal{S}_x obszar ufności, tzn. taki obszar w przestrzeni \mathcal{D}_k , że pokrywa on prawdziwą tablicę Czekanowskiego, czyli wektor δ , ze z góry

⁽⁹⁾ Układ (34) odpowiada metodzie budowania dendrytów podanej przez M. Warmusa, równoważnej metodzie podanej w [5]. Tę metodę przyjąłem ze względu na możliwość przejrzystszego zapisania jej układem nierówności. O metodzie prof. Warmusa poinformował mnie prof. St. Zubrzycki.

zadany prawdpodobieństwem. Do każdego obszaru ufności \mathcal{S}_x można zatem dobrać najmniejszy wielościan \mathcal{C} (oznaczymy go przez \mathcal{C}_x) taki, że $\mathcal{S}_x \subset \mathcal{C}_x \subset \mathcal{D}_k$. W tym miejscu należy zaznaczyć, iż na skutek przyjęcia estymatorów d podług definicji (6), otrzymane przez nas w drodze losowania obszary ufności \mathcal{S}_x pokrywają całą przestrzeń \mathcal{D}_k , natomiast wektor δ , jak pokazaliśmy w § 4, zawarty jest w wielościanie \mathcal{A}_{mn} stanowiącym część przestrzeni \mathcal{D}_k . Ponieważ każdemu wielościanowi \mathcal{C} odpowiada określona rodzina \mathcal{R} dendrytów różnych topologicznie, mamy:

Określenie. Prawdpodobieństwo, że prawdziwy dendryt Δ należy do określonej rodziny \mathcal{R}_x dendrytów różnych topologicznie, jest to prawdpodobieństwo, że wektor δ należy do obszaru ufności \mathcal{S}_x , zawartego w określonym wielościanie \mathcal{C}_x , należącym do przestrzeni euklidesowej, k -wymiarowej \mathcal{D}_k , a więc

$$(35) \quad P(\Delta \in \mathcal{R}_x) \stackrel{\text{def}}{=} P(\delta \in \mathcal{S}_x \subset \mathcal{C}_x \subset \mathcal{D}_k).$$

W przypadku, gdy \mathcal{C}_x ogranicza się do jednego tylko wielościanu \mathcal{B} , symbol $P(\Delta \in \mathcal{R}_x)$ zastępujemy symbolem $P(\Delta \stackrel{\text{top}}{=} D_x)$, gdzie D_x jest to dendryt określony topologicznie przez dany wielościan \mathcal{B} zawierający obszar ufności \mathcal{S}_x , czyli krótko mówiąc, jest to dendryt zbudowany na średnich z próby.

W celu wyznaczenia określonej przez (35) rodziny \mathcal{R}_x , dokonamy afinicznej transformacji przestrzeni \mathcal{D}_k zmiennych $\zeta_1, \zeta_2, \dots, \zeta_k$ na przestrzeń \mathcal{X}_k zmiennych $\xi_1, \xi_2, \dots, \xi_k$, podług wzoru

$$(36) \quad \xi_x = \frac{\zeta_x - d_x}{s_x} v_x \quad (1 \leq x \leq k).$$

Dzięki tej transformacji, obszar ufności \mathcal{S}_x , zawarty w przestrzeni \mathcal{D}_k , którego nie umiemy określić, przejdzie w kostkę \mathcal{S}_c określoną przy (33), natomiast każdy wielościan \mathcal{B} przejdzie w wielościan, który oznaczmy \mathcal{B}^* .

Wielościan \mathcal{B} jest określony przez układ nierówności postaci (34), wobec czego składa się z hiperpłaszczyzn ortogonalnych względem jednej z płaszczyzn, zawierających wszystkie możliwe pary osi współrzędnych. Transformacja, przeprowadzająca za pomocą wzorów (36) \mathcal{B} w \mathcal{B}^* , nie zmienia tej ortogonalności, wobec czego wystarczy rozpatrywać transformacje, według wzorów (36), prostych ograniczających półpłaszczyzny należące do wielościanu \mathcal{B} , w płaszczyznach odpowiednich par osi współrzędnych. Nierówność $\zeta_{x_1} > \zeta_{x_2}$ przejdzie zatem w nierówność

$$(37) \quad \xi_{x_2} < \frac{s_{x_1} v_{x_2}}{s_{x_2} v_{x_1}} \xi_{x_1} + (d_{x_1} - d_{x_2}) \frac{v_{x_2}}{s_{x_2}}.$$

Ponieważ w płaszczyznach par osi współrzędnych, należących do przestrzeni \mathcal{X}_k , kostkę \mathcal{S}_c reprezentuje kwadrat o środku w początku układu współrzędnych i o krawędzi długości $2c$, równoległej do osi współrzędnych, więc wystarczy znaleźć warunek, jaki muszą spełniać współczynniki b_{12} i h_{12} prostej

$$(38) \quad \xi_{x_2} = b_{12} \xi_{x_1} + h_{12},$$

by przechodziła ona na lewo i w górę od tegoż kwadratu, nie przecinając go. Elementarne rozumowanie pokazuje, że jest to warunek

$$(39) \quad h_{12} - (b_{12} + 1)c \geq 0.$$

Kładąc w (37), w miejsce znaku nierówności, znak równości, stwierdzamy, iż otrzymana prosta musi spełniać warunek (39). Każda para liczb ξ_{x_1} , ξ_{x_2} , tworząca punkt leżący na lewo i w górę od tej prostej, spełnia tym bardziej nierówność (39). Stąd, podstawiając w (39) odpowiednie wyrażenia z (37), otrzymamy podstawową nierówność

$$(40) \quad c \leq \frac{d_{x_1} - d_{x_2}}{s_{x_1}/v_{x_1} + s_{x_2}/v_{x_2}} = T_{12}.$$

Nierówność ta, w języku wrocławskiej metody konstruowania dendrytów oznacza, iż na poziomie ufności $1 - \alpha$, gdzie α , zgodnie z (33), określa stałą c , zachodzi nierówność $\delta_{x_1} > \delta_{x_2}$, dla prawdziwych odległości międzygatunkowych o wskaźnikach pojedynczych x_1 i x_2 . Jeśli zatem chcemy zbudować dendryt podług zasady określonej układem (34), musimy w każdym przypadku porównywać odległości empiryczne za pomocą wzoru (40). W niektórych przypadkach, dla danego wiersza układu (34), odległość wybrana będzie jednoznacznie wyznaczona nierównościami; w innych, gdy nierówności (40) nie wszędzie zachodzą, będzie do wyboru kilka ewentualności. W takim przypadku przyjęcie każdej z ewentualności tworzy inną strukturę topologiczną dendrytu. Nieoznaczoność zatem jednej nierówności w układzie (34) tworzy parę dendrytów sąsiednich. W ten sposób, po wyczerpaniu wszystkich ewentualności, otrzymamy szukaną rodzinę \mathcal{D}_x dendrytów sąsiednich, do której należy prawdziwy dendryt Δ , z prawdopodobieństwem większym niż $1 - \alpha$. W skrajnych przypadkach dostaniemy: albo — w najlepszym razie — jeden dendryt równy topologicznie dendrytowi prawdziwemu, co zapisujemy $P(\Delta \stackrel{\text{top}}{=} D_x) > 1 - \alpha$, albo — w najgorszym — rodzinę złożoną z wszystkich możliwych dendrytów różnych topologicznie, zbudowanych na tablicy Czekanowskiego o danym rozmiarze.

Uwaga I. Jeśli $s_{x_1} \neq s_{x_2}$, to wyrażenie T_{12} z (40) jest większe niż gdybyśmy brali zwykłą odległość półpłaszczyzny, należącej w rozpatrywanej płaszczyźnie do \mathcal{B}^* , od początku współrzędnych. Wynika to ze zna-

nej nierówności, zachodzącej między średnią arytmetyczną a geometryczną. Dzięki temu można otrzymać większe c , czyli wyższy poziom ufności.

Uwaga II. Łatwo wykazać zbieżność stochastyczną metody powyższej wraz ze wzrostem liczności l , prób gatunkowych do nieskończoności. Wówczas bowiem licznik ilorazu T_{12} w (40) coraz mniej się waha wokół średniej $\delta_{x_1} - \delta_{x_2}$, a w mianowniku to samo dotyczy liczników s_{x_1} i s_{x_2} . Ponieważ jednak mianowniki v_{x_1} i v_{x_2} dążą wtedy do nieskończoności, iloraz T dąży też do nieskończoności. To zaś pozwala zwiększać nieograniczenie parametr c , co jak widać z (33) daje poziom ufności 1.

§ 7. Nieznajomość macierzy znaków. W paragrafach 5 i 6 zakładaliśmy znajomość teoretycznej macierzy znaków E ; już jednak powiedzieliśmy w § 2, iż w praktyce zwykle jej nie znamy. W tych więc przypadkach musimy wnioskować o poziomie ufności rodziny dendrytów na podstawie samej tylko łącznej próby losowej \mathcal{X} , z której otrzymujemy macierze Z i \bar{E} .

Jako metoda weryfikacji narzuca się tu przede wszystkim szukanie łącznego poziomu ufności dla rodziny dendrytów \mathcal{R}_x , oraz macierzy \bar{E} . W tym celu możemy rozszerzyć zastosowanie nierówności zawartej w tezie twierdzenia 6. Weźmy bowiem pod uwagę wzór (19). Jeśli położymy w tym wzorze $c_z = c_{i,x} = \alpha_i \beta_x z_{i,x}$, to prawdopodobieństwo określone rozkładem Studenta $\bar{\alpha}_{i,x} = P(t > c_{i,x})$ będzie poziomem ufności odpowiedniego $\varepsilon_{i,x}$. Stąd, biorąc pod uwagę, że wszystkich ε -ów jest $n \times k$, otrzymujemy w sumie prawdopodobieństwo $\sum_{i,x} \bar{\alpha}_{i,x}$, przy czym suma ta ma $n \times k$ składników, natomiast dla weryfikacji rodziny \mathcal{R}_x dendrytów otrzymujemy w sumie $\sum_{i=1}^k \alpha_i$. Wobec tego, stosując twierdzenie 6, otrzymamy wzór na ogólny poziom ufności dendrytu

$$(41) \quad P(\Delta \in \mathcal{R}_x) \geq 1 - \sum_{i=1}^k \alpha_i - \sum_{i,x} \bar{\alpha}_{i,x}.$$

Metoda ta, jak widać z twierdzenia 6, szacuje szukane prawdopodobieństwo po lewej stronie nierówności (41) w sposób nader rozrzutny. Aby więc po prawej stronie (41) otrzymać np. 0,95, należałoby — wobec wielkiej liczby występujących tam wartości α — wziąć próby z poszczególnych gatunków bardzo liczne, powiedzmy po kilkaset elementów. Ze względu na trudności praktyczne z tym związane proponuję inny sposób postępowania.

Wykorzystamy tu, zamiast macierzy E , macierz empiryczną \bar{E} , uważając ją, przeciwnie niż dotąd, za ustaloną dla danej konstelacji gatunków. Wiedząc na podstawie wzoru (18) o zbieżności stochastycznej \bar{E} do E , możemy przypuszczać, że obie macierze nie będą się wiele różniły

między sobą. Otrzymane w ten sposób estymatory \bar{d} będą w pewnych przypadkach obciążone, choć nadal rozłożone normalnie. Lecz jednocześnie, stosując w danej próbie obliczoną z niej macierz \bar{E} , zastępujemy statystyki d_x przez \bar{d}_x . Wskutek tego otrzymujemy (por. § 2) zamiast estymatorów wartości δ_x , estymatory wartości $\bar{\delta}_x$. W naszym zagadnieniu nie interesują nas prawdziwe wartości np. δ_{x_1} i δ_{x_2} , lecz jedynie prawdopodobieństwo nierówności $\delta_{x_1} > \delta_{x_2}$. Gdy chcemy wnioskować o prawdopodobieństwie tej nierówności na podstawie znanych nam wielkości \bar{d}_{x_1} i \bar{d}_{x_2} , to uwzględniając twierdzenie 7, pod warunkiem że zastosowaliśmy normalizację na dyspersję omówioną w § 3, musimy układ nierówności (34) zastąpić przez układ (42). Ograniczenia wskaźników są tu takie same jak w (34):

[illegible]

Przekształcając teraz przestrzeń \mathcal{D}_k na przestrzeń \mathcal{X}_k zgodnie ze wzorem (36), przy czym zastępujemy d przez \bar{d} oraz s przez \bar{s} (patrz § 3), otrzymujemy analogon nierówności (40),

$$(43) \quad c \leq \frac{\bar{d}_{x_1} - \bar{d}_{x_2} - 2(n-1)\psi(0)}{\bar{s}_{x_1}/v_{x_1} + \bar{s}_{x_2}/v_{x_2}} = \bar{T}_{12}.$$

Uzasadnieniem probabilistycznego sensu tej nierówności jest fakt, iż przy obliczaniu występujących tu czterech statystyk zastępujemy jedynie macierz E przez \bar{E} , którą uważamy za ustaloną. W ten sposób rozkład ilorazu (31) pozostaje niezmienny.

Tak więc na skutek nieznanowości macierzy E , nierówność (40) pogarsza się o składnik $2(n-1)\psi(0)$, który, jak widać z dowodu twierdzenia 7, odejmujemy od licznika z nadmiarem. Jest jasne, że wobec tego odległości międzypopulacyjne, które byłyby praktycznie rozróżnialne, stają się niekiedy nierozróżnialne. Niekiedy para odległości może się w ten sposób stać nawet nierozróżnialna teoretycznie; dzieje się tak wówczas, gdy wartość oczekiwana licznika w ilorazie \bar{T}_{12} nierówności (43) jest ujemna. Należy jednak zwrócić uwagę, iż wartość, którą tu odejmujemy, nie zależy od liczności prób, wobec czego, gdy tylko wartość oczekiwana licznika jest dodatnia, wyrażenie \bar{T}_{12} równie szybko dąży do nieskończoności jak T_{12} ze wzoru (40).

W przypadku teoretycznej czy praktycznej nierozróżnialności, jedynym wyjściem jest oczywiście rozszerzenie rodziny dendrytów, którą weryfikujemy.

Na zakończenie podajemy twierdzenie 6, które wykorzystaliśmy w §§ 6 i 7 oraz twierdzenie 7, które wykorzystaliśmy w bieżącym paragrafie.

TWIERDZENIE 6. Niech X oznacza n -wymiarowy wektor losowy, przy czym składowe mogą być zależne albo nie, a ich łączny rozkład ciągły lub skokowy. Oznaczmy przez $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$ odpowiednie pola zdarzeń, czyli takie zbiory jednowymiarowe, że $P(X_i \in \mathcal{A}_i) = 1$. Oznaczmy odpowiednio przez $\alpha_1, \alpha_2, \dots, \alpha_n$ prawdopodobieństwa zdarzeń $X_1 \in \mathcal{X}_1 \subset \mathcal{A}_1, X_2 \in \mathcal{X}_2 \subset \mathcal{A}_2, \dots, X_n \in \mathcal{X}_n \subset \mathcal{A}_n$, gdzie $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n$ są zbiorami o mierze dodatniej. Niech dalej: $\bar{\mathcal{X}}_1 = \mathcal{A}_1 \setminus \mathcal{X}_1, \bar{\mathcal{X}}_2 = \mathcal{A}_2 \setminus \mathcal{X}_2, \dots, \bar{\mathcal{X}}_n = \mathcal{A}_n \setminus \mathcal{X}_n$. Jeśli położymy $\bar{\mathcal{X}}_1 \times \bar{\mathcal{X}}_2 \times \dots \times \bar{\mathcal{X}}_n = \bar{\mathcal{X}} \subset \mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_n$, to zachodzi nierówność $P(X \in \bar{\mathcal{X}}) \geq 1 - \sum_{i=1}^n \alpha_i$.

Dowód. Prawdopodobieństwo tu rozważane ma własności miary Lebesgue'a na zbiorach n -wymiarowych, wskutek czego prawdopodobieństwo rozpostarte na określonym iloczynie kartezjańskim n zbiorów jednowymiarowych jest n -krotną całką Lebesgue'a na tym iloczynie. Załóżmy najpierw, że prawdopodobieństwo na wszelkich zbiorach typu

$$(*) \quad \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_k} \times \bar{\mathcal{X}}_{i_{k+1}} \times \dots \times \bar{\mathcal{X}}_{i_n},$$

gdzie liczby i_1, \dots, i_n są dowolną permutacją liczb $1 \dots n$, a $k > 1$, jest równe zero. Wówczas

$$\begin{aligned} P(X \notin \bar{\mathcal{X}}) &= \sum_{i=1}^n \int_{\bar{\mathcal{X}}_1} \dots \int_{\bar{\mathcal{X}}_i} \dots \int_{\bar{\mathcal{X}}_n} dP(x_1, \dots, x_n) = \\ &= \sum_{i=1}^n \int_{\mathcal{A}_1} \dots \int_{\bar{\mathcal{X}}_i} \dots \int_{\mathcal{A}_n} dP(x_1, \dots, x_n) = \sum_{i=1}^n \alpha_i, \end{aligned}$$

gdyż zbiory typu $\bar{\mathcal{X}}_1 \times \dots \times \mathcal{X}_n \times \dots \times \bar{\mathcal{X}}_n$ są rozłączne. Stąd, ponieważ $P(X \in \bar{\mathcal{X}}) + P(X \notin \bar{\mathcal{X}}) = 1$, $P(X \in \bar{\mathcal{X}}) = 1 - \sum \alpha_i$. Jeśli jednak prawdopodobieństwo na zbiorach typu (*) jest dodatnie, to odejmując od jedności wartość $\sum \alpha_i$, odejmujemy wielokrotnie wartości prawdopodobieństwa na tych zbiorach. Tyłokrotnie mianowicie, ile razy dany zbiór wchodzi do zbiorów $\mathcal{A}_1 \times \dots \times \mathcal{X}_i \times \dots \times \mathcal{A}_n$, to jest k -krotnie. Stąd prawdopodobieństwo na tych zbiorach odejmujemy $k-1$ razy za dużo. To dowodzi twierdzenia.

TWIERDZENIE 7. Jeżeli $\bar{\delta}_1 - \bar{\delta}_2 > 2(n-1)\psi(0)$, to $\delta_1 > \delta_2$, przy czym zakładamy, że $E(Z_{i,\kappa} - v_{i,\kappa})^2 = 1$. Użyte tu oznaczenia zostały wprowadzone w § 2.

Dowód. Niech

$$(**) \quad \delta_1 = \sum v_{i,1} \quad \text{oraz} \quad \delta_2 = \sum v_{i,2},$$

przy czym niech $v_{ij} \geq 0$. W n -wymiarowej, kartezjańskiej przestrzeni \mathcal{X}_n , określonej wartości δ odpowiada część hiperpłaszczyzny, zawarta w dodatnim sektorze układu współrzędnych, a określona równaniami (**). Oznaczmy tę część hiperpłaszczyzny przez S_δ . Nazwijmy rangą punktu $p = (p_1, p_2, \dots, p_n)$, należącego do rozpatrywanej przestrzeni, liczbę $\sum_{i=1}^n p_i$, a więc $\mathcal{R}(p) = \sum p_i$. Stąd $\mathcal{R}(p \in S_\delta) = \delta$. Jeśli zastępujemy δ przez $\bar{\delta}$, to zbiór S_δ przechodzi na zbiór $S_{\bar{\delta}}^*$ przez transformację homeomorficzną $p_i^* = p_i + 2\psi(p_i, \sigma_i)$, przy czym $p^* = (p_1^*, p_2^*, \dots, p_n^*) \in S_{\bar{\delta}}^*$, a funkcja ψ jest określona jak w lemacie 2.2.

W przypadku, gdy przestrzeń \mathcal{X}_n jest tak znormalizowana, że $\sigma_1 = \sigma_2 = \dots = \sigma_n = 1$ ⁽¹⁰⁾, hiperpowierzchnie S_δ^* stają się osiowo symetryczne względem przekątnej dodatniego sektora układu, czyli względem prostej $p = \iota t$, gdzie $\iota = (1, 1, \dots, 1)$. Wykażemy teraz, że przy tej normalizacji

$$(**) \quad \min_{p^* \in S_\delta^*} \mathcal{R}(p^*) = \delta + 2n\psi(\delta/n).$$

Weźmy pod uwagę funkcję $\mathcal{R}(p^*) - \mathcal{R}(p) = 2 \sum \psi(p_i)$, gdy $\sigma_i = 1$ ($1 \leq i \leq n$), oraz zbadajmy jej ekstremum z warunkiem ubocznym $\mathcal{R}(p) = \delta = \text{const}$. Otrzymujemy

$$\frac{\partial}{\partial p_i} \left[2 \sum \psi(p_i) + \lambda \left(\sum p_i - \delta \right) \right] = -2\Phi(-p_i) + \lambda = 0;$$

λ oznacza tu mnożnik Lagrange'a. Stąd $-p_i = \Phi^{-1}(\lambda/2)$, wobec czego $\lambda = 2\Phi(-\delta/n)$, $\Phi(-\bar{p}_i) = \Phi(-\delta/n)$, czyli $\bar{p}_i = \delta/n$. Dla skrajnego punktu $\bar{p} \in S$, to znaczy typu $\bar{p} = (0, \dots, 0, \delta, 0, \dots, 0)$, mamy, jak łatwo widzieć,

$$\mathcal{R}(\bar{p}^*) = \delta + 2(n-1)\psi(0) + 2\psi(\delta) > \delta + 2n\psi(\delta/n) = \mathcal{R}(\bar{p}^* \in S_\delta^*),$$

gdyż $(n-1)[\psi(0) - \psi(\delta/n)] > \psi(\delta/n) - \psi(\delta)$, z uwagi na to, że funkcja ψ jest malejąca (patrz lemat 2.2). Dla każdego punktu pośredniego wartość $\mathcal{R}(p^*)$ jest pośrednia. To daje nam związek (**). W granicy otrzymujemy $\mathcal{R}(\bar{p}^* \in S_\delta^*) = \delta$, a $\mathcal{R}(\bar{\bar{p}}^* \in S_\delta^*) = \delta + 2(n-1)\psi(0)$, gdy $\delta \rightarrow \infty$, co daje wtedy różnicę rang $2(n-1)\psi(0)$. Jak wynika z przebiegu funkcji ψ , ta aproksymacja jest już dobra dla $\delta/n \geq 3$. Dla mniejszych δ różnica rang jest mniejsza, gdyż, jeśli $n > 1$, to

$$2(n-1)\psi(0) > 2(n-1)[\psi(0) - \psi(\delta/n)] + 2[\psi(\delta) - \psi(\delta/n)],$$

zgodnie z tym, co powiedzieliśmy poprzednio. To kończy dowód twierdzenia.

⁽¹⁰⁾ Jest tak wówczas, gdy przestrzeń cech znormalizujemy na dyspersję, jak to omówiliśmy w § 3.

Uwaga. Gdy żadna składowa p_i nie zbliża się zbytnio do zera, czyli gdy dyskryminacyjność bezwzględna we wszystkich cechach jest duża (patrz § 3), to wspomniana różnica rang jest mała.

§ 8. Ogólne wnioski. Dokonamy teraz krótkiego przeglądu treści niniejszej pracy. Po wprowadzającym w zagadnienie paragrafie 1 następuje § 2, który podaje formalne podstawy działań na wektorach losowych wyjściowych X' , w celu tworzenia z nich statystyk, z których w § 5 budujemy rozkład empirycznej tablicy Czekanowskiego. Twierdzenia 1 i 2, które podajemy na końcu tego paragrafu, służą w następnym § 3 do wyjaśnienia niektórych własności estymatorów \bar{d} , którymi w zagadnieniach praktycznych zmuszeni jesteśmy zastępować statystyki d . Ważny tutaj wzór (22), który opisuje zbieżność statystyki \bar{d} do d , zakłada normalizację na dyspersję, którą omawiamy także w § 3. Również twierdzenie 7, zasadnicze dla zastosowań praktycznych metody badania poziomów ufności dendrytów, zakłada tę samą normalizację na dyspersję. To wskazuje na szczególną przydatność tej normalizacji.

Następny z kolei § 4 pozornie odbiega od zasadniczego tematu, gdyż omawia podstawowe własności teoretycznej tablicy Czekanowskiego; jednak, jak to widzimy w § 5, zagadnienie to jest ściśle związane z zagadnieniem rozkładu empirycznej tablicy Czekanowskiego. Twierdzenie 5 z § 5, ma, jak to zaraz wyjaśnimy, szczególne znaczenie przy obliczaniu poziomów ufności dendrytów. Ten ostatni problem omawiamy w §§ 6 i 7. Podana tam metoda szacowania od dołu szukanego prawdopodobieństwa opiera się na twierdzeniach 6 i 7, zamieszczonych na końcu § 7.

Znaczenie twierdzenia 3 między innymi polega na tym, iż w przypadku $n < \frac{1}{2}(m-1)$ (pomijamy nie wyjaśniony przypadek przeciwny), rząd rozkładu empirycznej tablicy Czekanowskiego jest mniejszy niż wymiar kostki \mathcal{S}_c (patrz § 6), który jest zawsze równy $k = \frac{1}{2}m(m-1)$. Wówczas opłaca się zwiększyć liczbę rozpatrywanych cech tak, by powyższa nierówność nie zachodziła, gdyż wówczas rząd rozkładu i wymiar kostki \mathcal{S}_c zrównają się, dzięki czemu przy tym samym poziomie ufności można uzyskać więcej informacji o konstelacji gatunków.

Metoda przedstawiona w §§ 6 i 7 daje oszacowanie szukanego prawdopodobieństwa od dołu i to w sposób dość rozrzutny. Jak bowiem pokazaliśmy w § 5, rozkład empirycznej tablicy Czekanowskiego jest przy naszych założeniach normalny, ale zależny. Idealne rozwiązanie powinno wskazywać taką transformację zmiennych losowych, która czyniłaby ten rozkład niezależnym, co umożliwiałoby już dokładne obliczenie wartości prawdopodobieństwa określonego wzorem (35). Jednakże dwa powody skłoniły mnie do pominięcia tego zagadnienia: Po pierwsze — wielkie trudności analityczne. Znalezienie, zresztą samo nader uciążliwe,

transformacji ortogonalnej, podające wartości własne tablicy Czekanowskiego, nie rozwiązuje jeszcze zagadnienia. Aby bowiem uzyskać rozkład określonej postaci analitycznej (jak np. iloraz Studenta), należałoby następnie poszczególne zmienne podzielić przez odpowiednie wartości własne. Jednakże te wartości, będące oczywiście funkcjami wariancji i kowariancji z próby, same byłyby między sobą zależne (usunięcie tej nowej zależności byłoby już bardzo trudne). Po drugie — zachodzi tu pewien dylemat statystyczny. Często mianowicie mamy w statystyce do wyboru dwie ewentualności: albo dążyć do maksymalnej dokładności, nie bacząc na wielki nakład pracy rachunkowej, albo przeciwnie — korzystać z metod przybliżonych, mających własności asymptotyczne, tak by móc niedokładność kompensować pobieraniem odpowiednio liczniejszych próbek.

Otóż sędzę, iż metoda przedstawiona w §§ 6 i 7, a należąca do metod tego drugiego rodzaju, jest w zagadnieniach biometrycznych korzystniejsza. Łatwiej jest bowiem na ogół mierzyć większą liczbę osobników z poszczególnych gatunków, niż wykonywać wielogodzinne, nader uciążliwe rachunki.

Rachunki potrzebne w omawianej metodzie istotnie są elementarne. Po dokonaniu pomiarów i otrzymaniu w ten sposób realizacji wektorów X' , znajdujemy macierz \bar{E} , a stąd wartości $\bar{y}^{p|q}$, z których już łatwo otrzymujemy realizacje zmiennych \bar{d}^{ap} (patrz § 2). Za pomocą tejże macierzy \bar{E} znajdujemy zmienne $\bar{s}_{p|q}$ (patrz wzór (23)), a stąd zmienne \bar{s}_{pq} (patrz początek § 6). To pozwala już utworzyć ilorazy (40) albo (44), w ilości określonej liczbą $m(m-2)$, do uzyskania dendrytu pierwszego rzędu. Dla dendrytów dalszych rzędów odpowiednie liczby będą oczywiście znacznie mniejsze.

We wszystkich tych obliczeniach występują jedynie cztery arytmetyczne działania, oraz podnoszenie do kwadratu i pierwiastkowanie kwadratowe. Ilość potrzebnych działań jest tu rzeczywiście wielka, lecz w dobie maszyn matematycznych nie jest to wada istotna. Zresztą np. dla 10-ciu gatunków sytuacja jest następująca: Liczba statystyk \bar{d}_x , oraz \bar{s}_x , jest $k = 45$. Jeśli przyjmiemy poziom ufności $1 - \alpha > 0,95$, to dla poszczególnych odległości musimy przyjąć poziomy ufności rzędu 0,999 (patrz tw. 6). Dla średnich gatunkowych, np. dziesięcioelementowych, jest to możliwe do osiągnięcia. Dla znalezienia zatem dendrytu pierwszego rzędu z 10-ciu gatunków należy obliczyć 90 wartości statystyk \bar{d}_x i \bar{s}_x , a z nich macierz T elementów T (patrz § 6) o rozmiarach $10 \times 8 = 80$ elementów. Jak stąd widzimy, obliczenia te mieszczą się w granicach możliwości jednego badacza, nie dysponującego maszynami cyfrowymi.

Na zakończenie warto zwrócić uwagę na zastosowania tej metody. Powstała ona z myślą o badaniach biometrycznych — stąd wprowa-

dzono tu pojęcie gatunku — lecz oczywiście możliwości stosowania jej są o wiele szersze. Zwróćmy mianowicie uwagę na badania antropologiczno-socjologiczne, gdzie pojęcie „gatunku” daje się natychmiast zastąpić pojęciem grupy społecznej, takiej że przynależność osobników do niej nie budzi wątpliwości. Tak więc można np. szukać dendrytu narodów na danym poziomie ufności, ze względu na wyniki różnych testów psychologicznych lub innych, przy czym należy wybrać losowo z każdego branego pod uwagę narodu zadaną z góry liczbę osobników. Inną dziedziną zastosowań może być fizyka. Pojęcie „gatunku” możemy tu bowiem zastąpić pojęciem punktu w przestrzeni, a rozkłady gatunkowe mogą być konsekwencją błędów pomiarów. W fizyce ciała stałego możemy zastąpić „gatunek” pojęciem molekuly; rozkłady gatunkowe byłyby tu wynikiem ruchów cieplnych. Przykłady powyższe nie wyczerpują oczywiście możliwości zastosowań omówionej tu metody.

Prace cytowane

- [1] H. Cramér, *Mathematical methods of statistics*, Princeton 1946.
- [2] — *Metody matematyczne w statystyce*, Warszawa 1958.
- [3] J. Czekanowski, *Zarys metod statystycznych w zastosowaniu do antropologii*, PTNW 1913.
- [4] — *Zarys antropologii Polski*, Lwów 1930.
- [5] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus, S. Zubrzycki, *Sur la liaison et la division des points d'un ensemble fini*, *Colloq. Math.* 2 (1951), str. 282-285.
- [6] — *Taksonomia wrocławska*, *Przegląd Antropologiczny* 17 (1951), str. 193-211.
- [7] C. Radhakrishna Rao, *Advanced statistical methods in biometric research*, New York 1950.

Praca wpłynęła 17. 9. 1961

J. MIKIEWICZ (Wrocław)

O POZIOMACH UFNOŚCI W TAKSONOMII WROCŁAWSKIEJ

STRESZCZENIE

Celem pracy jest dostarczenie badaczom w dziedzinie nauk przyrodniczych i społecznych statystycznej metody konstruowania dendrytów wrocławskich. Podczas gdy metoda podana przez autorów taksonomii wrocławskiej pozwala wyznaczyć najkrótszy dendryt wiążący określony zbiór przedmiotów, praca niniejsza dostarcza metody statystycznego wnioskowania o dendrycie wiążącym „gatunki”, rozumiane tu jako populacje o rozkładzie normalnym. „Dendryt gatunków”, to dendryt wiążący środki ciężkości rozkładów poszczególnych gatunków. „Gatunki” mogą tu oznaczać

nie tylko populacje biologiczne, ale również grupy społeczne (np. narody), populacje powstałe na skutek błędów pomiarów, lub inne. Praca pozwala 1° wnioskować na zadanym z góry poziomie ufności o topologicznej strukturze dendrytu gatunków, 2° oceniać prawdziwe odległości istniejące w dendrycie gatunków. W pracy bada się również związane z tym zagadnienie własności teoretycznej tablicy Czekanowskiego (dla odległości będących liczbami rzeczywistymi), oraz własności rozkładu empirycznej tablicy Czekanowskiego (dla odległości z próby). Rachunki potrzebne przy stosowaniu omówionej metody są elementarne i nadają się dobrze do obliczania na maszynach cyfrowych.

Я Н М И К Е В И Ч (Вроцлав)

О ДОВЕРИТЕЛЬНОМ УРОВНЕ ВО ВРОЦЛАВСКОЙ ТАКСОНОМИИ

РЕЗЮМЕ

Вроцлавская таксономия является методом представления на плоскости взаимного расположения конечного множества точек в многомерном метрическом пространстве ([5], [6]). Этот метод обычно называется „методом дендритов”. С биометрической точки зрения этот метод сводится к графическому представлению взаимного родства, или, иначе говоря, сходства определенного множества индивидов, причем эти сходства понимаются как функции расстояния в многомерном пространстве, в котором каждая координата обозначает некоторый естественный таксономический признак. Идею такого подхода к естественному сходству находим уже у Чекановского [3]. Чекановским был также дан метод графического представления этих сходств, называемый диаграммой Чекановского ([3], [4]). Но этот метод давал лишь упорядочение индивидов на прямой. Вроцлавская таксономия обладает тем преимуществом, что позволяет подлучить дендритное упорядочение.

Вроцлавская таксономия, представленная в работах [5], [6], не является статистическим методом. Автор настоящей работы стремится дать статистический метод, разрешающий построить с заданной вероятностью дендрит, растянутый на совокупностях, названных здесь „видами”. Под „видом” здесь подразумевается совокупность индивидов с нормальным распределением. Во избежание осложнений принимаем условие, что эти совокупности генетически чисты, то есть не скрещиваются с индивидами других совокупностей. Тогда нормальность распределения данной биологической совокупности обоснована центральной теоремой. Разумеется, рассматриваемая в настоящей работе статистическая модель может найти приложение также к иным естественным и общественным явлениям, таким как общественные группы (напр. определение дендрита народов) или совокупности точек, возникших вследствие погрешности измерения.

Центры тяжести распределений отдельных совокупностей назовем видowymi векторами. Если бы эти векторы были известны кратчайший дендрит, построенный по вроцлавским методам, был бы однозначно определен. Но на практике эти векторы неизвестны и можно их только оценивать на основании случайных выборок, взятых из частных совокупностей. В настоящей работе даются методы, разрешающие сделать статистические выводы относительно: 1° топологической структуры дендрита видов и 2° истинных расстояний между видами существующими в дендрите. Под топологически равными дендритами подразумеваем дендриты, связанные гомеоморфизмом. Решение второго вопроса тесно

связано с решением первого. Решение первого вопроса связано с обобщением метода доверительных интервалов Неймана. Дается топологическое равенство истинного и выборочного дендритов или принадлежность истинного дендрита к топологически определенному семейству смежных дендритов. Каждый из возможных методов построения вроцлавского дендрита — будь это метод, приведенный в [5] или метод Мечислава Вармуса (неопубликованный) — является рядом шагов, сводящихся к выбору из данного множества расстояний кратчайшего из них. Так как эмпирические расстояния являются случайными переменными, каждый шаг метода зависит от случайного упорядочения расстояний в определенном множестве. Отсюда определенное изменение упорядочения ведет к топологически отличному дендриту, который называем смежным с предыдущим.

В работе рассматривается сначала теоретическая и эмпирическая таблица Чекановского. Теоретическая таблица Чекановского — это множество всех расстояний, соединяющих m точек n -мерного пространства. Конечно, значения этих расстояний зависят от принятой метрики. В настоящей работе принята

метрика, определенная формулой $d = \sum_{i=1}^n a_i |d_i|$, причем d_i являются разностями компонент видовых векторов, a_i — положительными коэффициентами, нормирующими масштабы координат.

В § 4 доказано, что множество всех возможных таблиц Чекановского образует в пространстве всех $k = \frac{1}{2}m(m-1)$ расстояний определенный многогранник (тип А). Его размеры, в зависимости от m и n , определяет теорема 3. Эмпирическая таблица Чекановского является множеством всех межвидовых расстояний, полученных из суммарной выборки, то есть из множества выборок частных видовых совокупностей. При определенных предположениях эмпирическая таблица Чекановского имеет нормальное распределение. Порядок этого распределения в таком случае не больше размера многогранника типа А.

Топологически определенный дендрит эквивалентен некоторой системе неравенств между k расстояниями, следовательно, ему соответствует в k -мерном пространстве некоторый выпуклый многогранник (тип В). Семейству дендритов соответствует множественная сумма таких многогранников. Эквивалентом принадлежности истинного дендрита к топологически определенному семейству дендритов является в этом пространстве принадлежность точки, представляющей теоретическую таблицу Чекановского, к определенной сумме многогранников типа В.

В § 6 представлен способ нахождения множества, окружающего вектор, представляющий эмпирическую таблицу Чекановского и содержащий с заданной вероятностью истинную (теоретическую) таблицу Чекановского. Этот метод основан на частном Студента. Так как это множество всегда принадлежит определенной сумме многогранников типа В, то связанная с ним вероятность является требуемым доверительным уровнем, для которого истинный дендрит принадлежит к определенному семейству дендритов.

В работе обсуждается способ применения обсуждаемой теории при нахождении дендритов, растянутых на видовых векторах с заданной вероятностью. Несмотря на необходимость выполнения значительного количества расчетов, этот метод удобен на практике, так как требует лишь знания четырех арифметических действий, возведения в квадратную степень и извлечения квадратного корня. Автор считает, что выполнение этих расчетов может во многих случаях оказаться выгодным для научных или других целей, особенно в условиях применения счетных машин.

J. MIKIEWICZ (Wrocław)

ON LEVELS OF CONFIDENCE IN WROCLAW TAXONOMY

SUMMARY

Wrocław taxonomy is a method of demonstrating on a plane the arrangement (constellation) of a finite set of points in a multi-dimensional metric space (see papers [5] and [6]; the method in question is popularly termed the „dendrite method”). From the biometric point of view this means a graphic representation of the affinity or — in other words — the similarities of a certain set of individuals, those similarities being understood as functions of the distances in a multi-dimensional space in which each coordinate denotes a certain biological measurement feature. This conception of biological similarity is to be found in J. Czekanowski [3]. He also introduced a method of graphical representation of those similarities which is called the Czekanowski diagram (see [3] and [4]). This method, however, only gives an arrangement of the individuals on a straight line. Wrocław taxonomy is better in so far as it gives a dendrite arrangement.

Wrocław taxonomy, presented in papers [5] and [6], is not a statistical method. The object of the present paper is to give a statistical method permitting the construction, with a probability given a priori, of a dendrite spanned over populations which are called here species. By a species we understand a population of individuals with a normal distribution. To avoid complications we assume that the individuals are genetically pure, i. e. not crossed with individuals from other populations. In this situation the normality of the distribution of a given biological population follows from the central limit theorem. The statistical model considered in this paper can of course be applied also to other biological and social phenomena, such as social groups (e. g. the determination of the dendrite of nations), or populations of points arising from measurement errors.

The centres of gravity of the distributions of the individual populations will be called the species vectors. If those vectors were known, the shortest dendrite constructed by the Wrocław methods would be uniquely determined. In practice, however, we do not know those vectors and can only estimate them on the basis of random samples taken from the individual populations. This paper presents methods of statistical inference regarding 1° the topological structure of the dendrite of species, 2° the real distances existing in the dendrite of species. By topologically equal dendrites we understand dendrites bound by a homeomorphism. The answer to the second question is closely connected with the answer to the first question. The answer to the first question is connected with the generalization of J. Neyman's method of confidence intervals. Namely we consider the probability of topological equality of the dendrites — the real one and the one from the sample, or the probability of the real dendrite belonging to a topologically defined family of neighbouring dendrites. Each of the possible methods of constructing the Wrocław dendrite, either the one presented in [5] or the method (unpublished) of M. Warmus, is a sequence of steps consisting in the choice of the shortest distance from a certain set of distances. If we take into consideration the fact that empirical distances are random variables, we shall observe that each step of the method is dependent on the random arrangement of the distances in the set in question. Hence a definite change in the arrangement of the distances leads to a topologically different dendrite, which we term neighbouring with respect to the preceding one.

In the paper we begin by considering the theoretical and empirical tables of Czekanowski. A theoretical table of Czekanowski is the set of all the distances connecting m points in an n -dimensional space. The values of those distances depend of course

on the metric adopted. In this paper we assume a metric defined by the formula $d = \sum_{i=1}^n a_i |d_i|$ where d_i are the differences of the component species vectors and a_i are the positive coefficients norming the scales of the coordinates. In § 4 it has been shown that the set of all the possible tables of Czekanowski forms in the space of all $k = \frac{1}{2}m(m-1)$ distances a certain polyhedron (type A). Its dimension in terms of m and n is given in theorem 3. An empirical table of Czekanowski is the set of all inter-species distances obtained from a joint sample, i. e. from the set of samples from the individual species populations. Under particular assumptions, an empirical table of Czekanowski has a normal distribution. The order of that distribution is then not greater than the dimension of the type A polyhedron.

A dendrite defined topologically is equivalent to a certain system of inequalities holding between k distances, and thus its counterpart in the k -dimensional space is a certain convex polyhedron (type B). A family of dendrites corresponds to the set-theoretical sum of such polyhedra. The fact that a real dendrite belongs to a topologically defined family of dendrites has its counterpart, in the space in question, in the fact that a point representing the theoretical table of Czekanowski belongs to a certain sum of B type polyhedra.

§ 6 contains a method of finding the set surrounding the vector representing the empirical table of Czekanowski and containing, with a probability given a priori, the real (theoretical) table of Czekanowski. The method is based on the Student ratio. Since the set in question always belongs to a certain sum of B type polyhedra, the probability connected with it is the required confidence level with which a real dendrite belongs to a certain family of dendrites.

The paper then discusses a method of practical application of the above theory to seeking dendrites spanned on species vectors with a probability given a priori. In spite of a large amount of calculation needed, the method is easy in practice since it only requires the knowledge of the four operations of arithmetic, of squaring and of extracting the square root. The author believes that it will be worth while to conduct these calculations in a great many cases for scientific or other purposes, particularly with the use of electronic computers.