ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
Seria III: MATEMATYKA STOSOWANA XV (1979)

STANISŁAW GNOT, HENRYK MATEJ, TEOFIL SZULGA (Wrocław)

Test dokładny dla testowania hipotezy o równowadze Hardy'ego–Weinberga

(*Praca przyjęta do druku 4.04.1978*)

1. Podstawowe pojęcia i definicje w genetyce. Informacja genetyczna żywych organizmów zapisana jest w materiale genetycznym zawartym w jądrze komórki. Nosicielem tej informacji jest kwas dezoksyrybonukleinowy (DNA), związek chemiczny w formie długich nici, zbudowany z jednostek zwanych *nukleotydami*. Sekwencja nukleotydów stanowi kod informacyjny. W jądrze materiał genetyczny podzielony jest na specyficzne zbiory zwane *chromosomami*. W komórkach człowieka chromosomy są tworami parzystymi, przy czym liczba par jest stała i wynosi 23. W tej liczbie mieści się 22 pary chromosomów nie związanych z płcią oraz jedna para chromosomów płciowych. Parzystość chromosomów wynika z połączenia się dwu rodzicielskich komórek zwanych *gametami*.

Jednostką funkcjonalną materiału genetycznego jest *gen*. Jest to niewielki odcinek DNA, a jego miejsce (pozycja) w chromosomie nazywa się *locusem*. Geny związane z danym miejscem mogą występować w różnych postaciach (*allelach*). Ze względu na parzystość chromosomów pojęcie „locus" obejmuje dwa identyczne miejsca na parze chromosomów. Wynika z tego, że osobnik określony jest przez parę genów w jednym locus. Osobnik posiadający dwa identyczne allele w danym locus nazywa się *homozygotą*, natomiast osobnik, u którego allele są różne, nazywa się *heterozygotą*. Całkowite genetyczne ukonstytuowanie organizmu określa jego *genotyp*. Termin genotyp używany jest także w odniesieniu do większej liczby loci. Produktem aktywności genotypu są cechy osobnicze, nazywane *fenotypem*. Jeżeli u heterozygoty produkty obu alleli ujawniają się w fenotypie, mówimy o *kodominacji*. Jeżeli natomiast ujawni się tylko produkt jednego allelu, gen taki nazywamy *dominującym*, a gen nieujawniony — genem *recesywnym*. Rozpatrzmy dla przykładu jeden locus w chromosomie z allelami A_1 i A_2. W tym przypadku możliwymi genotypami są: $A_1 \times A_1$, $A_1 \times A_2$, $A_2 \times A_2$. Jeżeli gen A_2 jest genem recesywnym, wówczas genotypy $A_1 \times A_1$ i $A_1 \times A_2$ określają taki sam fenotyp. W wielu układach genetycznych liczba możliwych alleli jest większa od 2.

Rozważmy jedno miejsce w chromosomie u indywiduów tworzących pewną populację π. Przypuśćmy, że w locus tym mogą pojawić się allele $A_1, A_2, \ldots, A_m$.

Niech G_{ij} będzie genotypem powstałym przez połączenie gamet z allelami A_i oraz A_j ($G_{ij} \equiv G_{ji}$). Oznaczmy przez p_{ij} częstość (frakcję) genotypu G_{ij} w badanej populacji ($p_{ij} \equiv p_{ji}, p_{ij} > 0$). Parametry

$$t_i = p_{ii} + (1/2) \sum_{i \neq j} p_{ij}, \quad i = 1, 2, \ldots, m, \tag{1}$$

nazywamy *częstościami* genów $A_1, A_2, \ldots, A_m$, odpowiednio. Struktura genetyczna badanej populacji może być opisywana zarówno w terminach częstości genotypów, jak też częstości genów. Związek pomiędzy częstościami genów i genotypów dany wzorem (1) wygodnie przedstawić jest w formie macierzowej. Niech $\mathbf{A}$ będzie macierzą częstości genotypowych określoną w następujący sposób:

$$\mathbf{A} = \begin{bmatrix} p_{11} & (1/2)p_{12} & \ldots & (1/2)p_{1m} \\ (1/2)p_{21} & p_{22} & & (1/2)p_{2m} \\ \vdots & \vdots & & \vdots \\ (1/2)p_{m1} & (1/2)p_{m2} & \ldots & p_{mm} \end{bmatrix}.$$

Macierz $\mathbf{A}$ jest macierzą symetryczną o elementach sumujących się do jedności, tzn.

$$\mathbf{1'A1} = 1,$$

gdzie $\mathbf{1}$ jest wektorem kolumną, złożonym z samych jedynek. Wektor częstości genów $\mathbf{t} = (t_1, t_2, \ldots, t_m)'$ wyraża się wzorem:

$$\mathbf{t} = \mathbf{A1}.$$

W dalszym ciągu rozważać będziemy populacje, w których spełnione są następujące założenia:

a) w populacji nie występują czynniki oddziaływające na jej strukturę genetyczną, takie jak: mutacja, selekcja, migracja,

b) populacja zawiera jednakową liczbę osobników męskich i żeńskich,

c) rozkład częstości genotypów (genów), męskich i żeńskich jest taki sam,

d) poszczególne generacje populacji nie zachodzą na siebie,

e) każdy z dwóch genów może pojawić się w wyprodukowanej przez osobnika gamecie z prawdopodobieństwem 1/2 (jest to tzw. pierwsze prawo Mendla).

Przyjęte założenia umożliwiają badanie dynamicznej i statystycznej struktury populacji przy pomocy modeli matematycznych. Warto zaznaczyć, że drobne odstępstwa od większości z tych założeń nie wpływają zasadniczo na zmianę rozważanych modeli.

2. Losowa asocjacja gamet. Prawo Hardy'ego–Weinberga. Mówimy, że w populacji π z wektorem częstości genów $\mathbf{t}$ spełnione jest *założenie losowej asocjacji gamet*, jeżeli częstości genotypów w pierwszej generacji π_1 spełniają warunki:

$$p_{ij} = 2t_i t_j, \quad i = j, \; i = 1, 2, \ldots, m,$$

$$p_{ii} = t_i^2, \quad i = 1, 2, \ldots, m,$$

lub w notacji macierzowej

$$\mathbf{A}_1 = \mathbf{tt'} = \mathbf{A11'A}.$$

Tutaj $\mathbf{A}$ i $\mathbf{A}_1$ są macierzami częstości genotypów w populacji π i π_1, odpowiednio. Rozważmy pewną wyjściową populację π_0 z macierzą częstości genotypów $\mathbf{A}_0$ i z wektorem częstości genów $\mathbf{t}_0 = \mathbf{A}_0\mathbf{1}$. Załóżmy, że w populacji π_0 założenie losowej asocjacji gamet jest spełnione, tzn. w pierwszej generacji π_1 macierz częstości genotypów jest postaci

$$\mathbf{A}_1 = \mathbf{t}_0\mathbf{t}_0' = \mathbf{A}_0\mathbf{1}\mathbf{1}'\mathbf{A}_0.$$

Wektorem częstości genów w populacji π_1 jest

$$\mathbf{t}_1 = \mathbf{A}_1\mathbf{1} = \mathbf{t}_0\mathbf{t}_0'\mathbf{1} = \mathbf{t}_0,$$

a macierzą częstości genotypów w drugiej generacji π_2 jest

$$\mathbf{A}_2 = \mathbf{t}_1\mathbf{t}_1' = \mathbf{A}_1.$$

W konsekwencji otrzymujemy następujące wnioski:

(i) w populacji z losową asocjacją gamet wektor częstości genów $\mathbf{t}$ jest w każdej generacji taki sam jak w populacji wyjściowej (jest to tzw. prawo równowagi Hardy'ego–Weinberga),

(ii) w populacji założenie losowej asocjacji gamet jest spełnione wtedy i tylko wtedy, gdy macierz $\mathbf{A}$ częstości genotypów spełnia warunek:

$$\mathbf{A} = \mathbf{t}\mathbf{t}' = \mathbf{A}\mathbf{1}\mathbf{1}'\mathbf{A}. \tag{2}$$

Wiele problemów rozważanych w genetyce populacyjnej rozwiązywanych jest przy założeniu losowej asocjacji gamet. Testowanie tej hipotezy opiera się zazwyczaj na statystyce χ^2 Pearsona, jednakże w przypadkach niewielkiej liczby obserwacji test χ^2 powinien być zastąpiony testem dokładnym. W pracy podany jest opis jednostajnie najmocniejszego testu nieobciążonego dla modeli genetycznych z dwoma allelami. Test konstruuje się na podstawie ogólnej teorii testowania hipotez liniowych dla rodzin rozkładów wykładniczych rozwiniętej przez Birnbauma [1], Lehmanna [3], Truaxa [5], Truaxa i Matthesa [4]. Podany jest też przykład zastosowania tego testu dla układu MN grup krwi.

3. Postać kanoniczna hipotezy liniowej dla rodziny wykładniczej rozkładów prawdopodobieństwa. Rozważmy rodzinę wykładniczą rozkładów prawdopodobieństwa wektora losowego $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)'$ postaci:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = W(\boldsymbol{\theta})H(\mathbf{x})\exp\Big\{\sum_{i=1}^{k}\theta_i x_i\Big\} = W(\boldsymbol{\theta})H(\mathbf{x})\exp\{\boldsymbol{\theta}'\mathbf{x}\}, \quad \boldsymbol{\theta} \in \Omega, \tag{3}$$

względem pewnej σ-skończonej miary μ. Tutaj $\boldsymbol{\theta}'\mathbf{x}$ jest iloczynem skalarnym wektorów $\boldsymbol{\theta}$ i $\mathbf{x}$, Ω jest tzw. naturalną przestrzenią parametrów, tzn.

$$\Omega = \Big\{\boldsymbol{\theta} \in \mathscr{R}^k\colon \int_{\mathscr{X}} \exp\{\boldsymbol{\theta}'\mathbf{x}\}\, d\mu(x) < \infty\Big\},$$

a $\mathscr{X}$ jest przestrzenią realizacji wektora losowego $\mathbf{x}$. Niech Ω_0 będzie r-wymiarową podprzestrzenią liniową przestrzeni $\mathscr{R}^k$ $(0 < r \leqslant k)$. Rozważmy hipotezę liniową

$$H\colon \boldsymbol{\theta} \in \Omega_0 \cap \Omega.$$

Niech $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_r$ będzie bazą w przestrzeni Ω_0, a $\mathbf{p}_{r+1}, \mathbf{p}_{r+2}, \ldots, \mathbf{p}_k$ niech będzie uzupełnieniem bazy w Ω_0 do bazy w $\mathscr{R}^k$. Każdy wektor $\boldsymbol{\theta} \in \Omega$ można w sposób jednoznaczny przedstawić w postaci:

$$\boldsymbol{\theta} = \sum_{i=1}^{k} \xi_i \mathbf{p}_i.$$

Niech $\xi = (\xi_1, \xi_2, \ldots, \xi_k)$ będzie wektorem współczynników kombinacji $\sum_{i=1}^{k} \xi_i \mathbf{p}_i$ i niech $\mathbf{P}$ będzie macierzą, której kolumnami są wektory $\mathbf{p}_i$, $i = 1, 2, \ldots, k$. Wówczas

$$\boldsymbol{\theta} = \mathbf{P}'\xi \quad \text{oraz} \quad \xi = \mathbf{P}'^{-1}\boldsymbol{\theta}.$$

Rozważmy wektor losowy $\mathbf{T} = \mathbf{PX}$. Na mocy (3) rozkład prawdopodobieństwa wektora $\mathbf{T}$ jest postaci:

$$g_\xi(\mathbf{t}) = W(\mathbf{P}'\xi)H(\mathbf{P}^{-1}\mathbf{t})\exp\{\xi'\mathbf{t}\}.$$

Ponieważ dla $\boldsymbol{\theta} \in \Omega_0$ $\xi_{r+1} = \xi_{r+2} = \ldots = \xi_k = 0$, hipoteza H przyjmuje następującą postać kanoniczną:

$$H\colon\ \xi_{r+1} = \xi_{r+2} = \ldots = \xi_k = 0.$$

W przypadku, gdy $r = k-1$, tzn. gdy hipoteza H dotyczy tylko jednego parametru i jest postaci:

$$H\colon\ \xi_k = 0,$$

rozwiązaniem problemu testowania H jest jednostajnie najmocniejszy test nieobciążony, który jest określony za pomocą następującej funkcji krytycznej:

$$\varphi(t_k|t_1, t_2, \ldots, t_{k-1}) = \begin{cases} 1, & \text{gdy } t_k > C_1 \text{ lub } t_k < C_2, \\ \gamma_i, & \text{gdy } t_k = C_i,\ i = 1, 2, \\ 0, & \text{gdy } C_1 < t_k < C_2, \end{cases}$$

z funkcjami C_i i γ_i zależnymi od $t_1, t_2, \ldots, t_{k-1}$ i wyznaczonymi z warunków:

$$E_{\xi_k=0}[\varphi(T_k|t_1, t_2, \ldots, t_{k-1})] = \alpha$$

oraz

$$E_{\xi_k=0}[T_k\varphi(T_k|t_1, t_2, \ldots, t_{k-1})] = \alpha E_{\xi_k=0}(T_k|t_1, t_2, \ldots, t_{k-1})$$

(por. Lehmann [2]). Aby wyznaczyć funkcje C_i i γ_i, konieczna jest znajomość warunkowego rozkładu prawdopodobieństwa T_k, przy warunkach $T_1 = t_1$, $T_2 = t_2, \ldots, T_{k-1} = t_{k-1}$ oraz przy prawdziwości hipotezy H: $\xi_k = 0$.

7. Testowanie hipotezy o równowadze Hardy'ego–Weinberga dla układu z dwoma allelami. Rozważmy układ genetyczny z dwoma allelami A_1 i A_2. Niech g_{ij} będzie częstością osobników powstałych przez połączenie gamety męskiej A_i z gametą żeńską A_j, $i, j = 1, 2$. Przy tych oznaczeniach $p_{11} = g_{11}$, $p_{12} = g_{12} + g_{21}$ i $p_{22} = g_{22}$ są częstościami genotypów G_{11}, G_{12} i G_{22} odpowiednio. Rozkład prawdopodo-

bieństwa liczby X_{ij} osobników powstałych przez połączenie gamety męskiej A_i z żeńską A_j w próbie n elementowej jest rozkładem wielomianowym postaci:

$$(n!/x_{11}!\,x_{12}!\,x_{21}!\,x_{22}!)\cdot g_{11}^{x_{11}}g_{12}^{x_{12}}g_{21}^{x_{21}}g_{22}^{x_{22}}, \quad \sum_{i,j} x_{ij} = n, \quad \sum_{i,j} g_{ij} = 1.$$

W przypadku chromosomów nie związanych z płcią naturalnym wydaje się założenie

$$g_{12} = g_{21} \quad \text{lub (co jest równoważne)} \quad p_{12} = 2g_{12}.$$

Przy tym założeniu zgodnie ze wzorem (2) założenie losowej asocjacji gamet jest spełnione wtedy i tylko wtedy, gdy macierz częstości genotypów

$$\mathbf{A} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

jest macierzą rzędu jeden lub, co jest równoważne, wtedy i tylko wtedy, gdy wyznacznik macierzy $\mathbf{A}$ jest równy zeru. Hipoteza o losowej asocjacji gamet przyjmuje zatem postać:

$$H\colon\ g_{11}g_{22} = g_{12}^2.$$

Przyjmując oznaczenia $\theta_1 = \ln(g_{11}/g_{12})$, $\theta_2 = \ln(g_{22}/g_{12})$ otrzymujemy:

$$H\colon\ \theta_1+\theta_2 = 0,$$

a rozkład prawdopodobieństwa wektora losowego $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})'$ przyjmuje postać:

$$H(\mathbf{x})g_{21}^n \exp(x_{11}\theta_1 + x_{22}\theta_2).$$

Bazą w przestrzeni $\Omega_0 = \{(\theta_1, \theta_2)'\colon \theta_1+\theta_2 = 0\}$ jest wektor $\mathbf{p}_1 = (1, -1)'$, który wraz z wektorem $\mathbf{p}_2 = (1, 0)'$ tworzy bazę w $\mathscr{R}^2$. Przyjmując

$$\mathbf{P} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{T} = P(X_{11}, X_{22})' = (X_{11}-X_{22}, X_{11})' = (T_1, T_2)'.$$

Rozkład prawdopodobieństwa wektora $\mathbf{T} = (T_1, T_2)'$ jest postaci:

$$W(\boldsymbol{\xi})H(\mathbf{t})\exp\{\xi_1 t_1 + \xi_2 t_2\},$$

gdzie $\boldsymbol{\xi} = \mathbf{P}'^{-1}(\theta_1, \theta_2)'$. Ponieważ

$$\mathbf{P}'^{-1} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix},$$

mamy: $\xi_1 = -\theta_2 = \ln(g_{12}/g_{22})$, $\xi_2 = \theta_1+\theta_2 = \ln(g_{11}g_{22}/g_{12}^2)$, a hipoteza H przyjmuje następującą postać kanoniczną:

$$H\colon\ \xi_2 = 0.$$

Jednostajnie najmocniejszym testem nieobciążonym dla testowania H, zgodnie z rozważaniami rozdziału poprzedniego, jest

$$\varphi(t_2|T_1 = t_1) = \begin{cases} 1, & \text{gdy } t_2 < C_1(t_1) \text{ lub } t_2 > C_2(t_1), \\ \gamma_i(t_1), & \text{gdy } t_2 = C_i(t_1),\ i = 1, 2, \\ 0, & \text{gdy } C_1(t_1) < t_2 < C_2(t_1). \end{cases}$$

Ponieważ $X_{11}-X_{22}=t_1$ i $X_{11}+X_{22}=n$, zmienna losowa $T_2=X_{11}$ może przyjmować wartości z przedziału $\langle n_0, n_1\rangle$, gdzie $n_0=\max\{0, t_1\}$ i $n_1=[(n+t_1)/2]$. Funkcje C_i i γ_i wyznacza się z warunków:

$$\sum_{t_2=n_0}^{C_1}\Pr_{\xi_2=0}\{T_2=t_2|T_1=t_1\}+\sum_{i=1}^{2}\gamma_i\Pr_{\xi_2=0}\{T_2=C_i|T_1=t_1\}+$$

$$+\sum_{t_2=C_2}^{n_1}\Pr_{\xi_2=0}\{T_2=t_2|T_1=t_1\}=\alpha,$$

$$\sum_{t_2=n_0}^{C_1}t_2\Pr_{\xi_2=0}\{T_2=t_2|T_1=t_1\}+\sum_{i=1}^{2}\gamma_i t_2\Pr_{\xi_2=0}\{T_2=C_i|T_1=t_1\}+$$

$$+\sum_{t_2=C_2}^{n_1}t_2\Pr_{\xi_2=0}\{T_2=t_2|T_1=t_1\}=\alpha\sum_{t_2=n_0}^{n_1}t_2\Pr_{\xi_2=0}\{T_2=t_2|T_1=t_1\}.$$

W celu znalezienia warunkowego rozkładu prawdopodobieństwa zmiennej losowej T_2 przy warunku $T_1=t_1$ i przy założeniu losowej asocjacji gamet, zauważmy, że:

$$\Pr\{T_2=t_2|T_1=t_1\}=\Pr\{X_{11}=t_2, X_{11}-X_{22}=t_1\}/\Pr\{X_{11}-X_{22}=t_1\}=$$

$$=\Pr\{X_{11}=t_2, X_{22}=t_2-t_1\}/\sum_{t_2=n_0}^{n_1}\Pr\{X_{11}=t_2, X_{22}=t_2-t_1\}.$$

Łączny rozkład prawdopodobieństwa zmiennych X_{11} i X_{22} jest rozkładem wielomianowym postaci:

$$\Pr\{X_{11}=t_2, X_{22}=t_2-t_1\}=[n!/t_2!(t_2-t_1)!(n-2t_2+t_1)!]g_{11}^{t_2}g_{22}^{t_2-t_1}(2g_{12})^{n-2t_2+t_1}=$$

$$=g_{12}^{n}(g_{12}/g_{22})^{t_1}2^{n-2t_2+t_1}H_{t_1}(t_2)\varrho^{t_2},$$

gdzie

$$H_{t_1}(t_2)=n!/t_2!(t_2-t_1)!(n-2t_2+t_1)! \quad \text{oraz} \quad \varrho=g_{11}g_{22}/g_{12}^2.$$

Z powyższych rozważań wynika, że

$$\Pr\{T_2=t_2|T_1=t_1\}=(1/4)^{t_2}H_{t_1}(t_2)\varrho^{t_2}/\sum_{t_2=n_0}^{n_1}(1/4)^{t_2}H_{t_1}(t_2)\varrho^{t_2}.$$

Przy założeniu losowej asocjacji gamet $\varrho=1$ rozkład warunkowy zmiennej T_2 przy warunku $T_1=t_2$ przyjmuje następującą postać:

$$\Pr\{T_2=t_2|T_1=t_1\}=(1/4)^{t_2}H_{t_1}(t_2)/\sum_{t_2=n_0}^{n_1}(1/4)^{t_2}H_{t_1}(t_2). \tag{4}$$

Przykład. Rozważmy układ *MN* grup krwi. Przypuśćmy, że w 60-elementowej próbie zaobserwowano następujące liczebności poszczególnych fenotypów: *MM*–24, *MN*–26, *NN*–10. Zgodnie z oznaczeniami przyjętymi w rozdziale 4 mamy: $X_{11}=24$, $X_{12}+X_{21}=26$, $X_{22}=10$, $T_1=14$, $n_0=14$ i $n_1=37$. Rozkładem warun-

kowym prawdopodobieństwa zmiennej losowej $T_2 = X_{11}$, obliczonym przy warunkach $T_1 = 14$ i $\varrho = 1$ na mocy wzoru (4) jest:

t_2	14	15	16	17	18	19	20	21	22
$\Pr\{T_2 = t_2 \mid T_1 = 14\}$	.000	.000	.000	.001	.008	.029	.075	.144	.202
	23	24	25	26	27	28	29	30	31
	.212	.167	.099	.044	.014	.003	.001	.000	.000
	32	33	34	35	36	37			
	.000	.000	.000	.000	.000	.000			

Przyjmując poziom istotności $\alpha = .05$ znajdujemy $C_1 = 19$, $C_2 = 26$. Ponieważ zaobserwowana wartość $T_2 = 24$ zawarta jest w przedziale (C_1, C_2), nie ma podstaw, aby odrzucić hipotezę H.

Prace cytowane

[1] A. Birnbaum, *Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests*, Ann Statist. 25 (1954), str. 21–36.

[2] E. L. Lehmann, *Testing Statistical Hypotheses*, John Wiley, New York 1959.

[3] — *Significance level and power*, Ann. Statist. 29 (1959), str. 1167–1176.

[4] T. K. Matthes, D. R. Truax, *Tests of composite hypotheses for multivariate exponential family*, Ann. Statist. 38 (1967), str. 681–697.

[5] D. R. Truax, *Multidecision problems for the multivariate exponential family*, Stanford Technical Report, No 32 (1955).