Estymacja parametrów

Krzysztof Topolski

Wrocław, 14 stycznia 2020

Genetyka populacyjna

Genetyka populacyjna zajmuje się badaniem częstości występowania poszczególnych alleli oraz genotypów w populacji. Bada także zmiany tych częstości spowodowane doborem naturalnym oraz innymi mechanizmami ewolucji. Opis struktury genetycznej badanej populacji nie jest prosta lista genotypów ale raczej częstością z jaką wystepuja w populacji poszczególne allele oraz genotypy.

Formalizm

Załóżmy, że w wybranym locusie mogą występować dwa allele, które będziemy oznaczać przez A_1 i A_2. To założenie prowadzi w konsekwencji do występowania w diploidalnej populacji trzech genotypów A_1A_1, A_2A_2 oraz A_1A_2, których częstości występowania oznaczymy przez x_{ij}, $i, j \in \{1, 2\}$.

Genotyp:	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Częstość:	x_{11}	x_{12}	x_{22}

Ponieważ mamy do czynienia z częstościamu więc

$$x_{11} + x_{12} + x_{22} = 1.$$

Formalizm

Załóżmy, że w wybranym locusie mogą występować dwa allele, które będziemy oznaczać przez A_1 i A_2. To założenie prowadzi w konsekwencji do występowania w diploidalnej populacji trzech genotypów A_1A_1, A_2A_2 oraz A_1A_2, których częstości występowania oznaczymy przez x_{ij}, $i, j \in \{1, 2\}$.

Genotyp:	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Częstość:	x_{11}	x_{12}	x_{22}

Ponieważ mamy do czynienia z częstościamu więc

$$x_{11} + x_{12} + x_{22} = 1.$$

Związek pomiędzy częstością allei i częstościa genotypów

Częstości alleli odgrywają ważna role w genetyce populacyjnej. Znając częstości poszczególnych genotypów, bez trudu możemy wyznaczyć częstości alleli. Częstość allelu A_1 jest równa

$$p = x_{11} + \frac{1}{2}x_{12},$$

natomiast czestość allelu A_2

$$q = 1 - p = x_{22} + \frac{1}{2}x_{12}.$$

Związek pomiędzy częstością allei i częstościa genotypów

O częstości allelu A_1 możemy myśleć na dwa sposoby. Pierwszy z nich to po prostu częstość występowania allelu A_1 w całej populacji.
O czestości allelu A_1 możemy także mysleć jak o prawdopodobieństwie zdarzenia polegającego na tym, że losowo wybrany z populacji allel jest typu A_1.

Związek pomiędzy częstością allei i częstościa genotypów

Proces losowego wyboru allelu można podzielić na dwa etapy. Pierwszy z nich to losowy wybór osobnika z całej populacji natomiast drugi etap to losowy wybór allelu z genotypu tak wybranego osobnika. Ponieważ mamy do czynienia z trzema możliwymi genotypami, takie spojrzenie prowadzi do następujacego wzoru wyrażajacego częstość allelu A_1

$$p = 1 \cdot x_{11} + \frac{1}{2} \cdot x_{12} + 0 \cdot x_{22} = x_{11} + \frac{1}{2} x_{12}.$$

Związek pomiędzy częstością allei i częstościa genotypów

Większość locii posiada więcej niż dwa możliwe allele. W przypadku gdy mamy do czynienia z n allelami $A_1, A_2, ..., A_n$ to oznaczając czestości możliwych w tej sytuacji genotypów A_iA_j gdzie $i, j \in \{1, 2, ..., n\}$ przez x_{ij}, to częstość allelu A_i, $i = 1, 2, ..., n$ możemy zapisać wzorem

$$p_i = x_{ii} + \frac{1}{2}\sum_{j=1}^{i-1} x_{ji} + \frac{1}{2}\sum_{j=i+1}^{n} x_{ij}.$$

Prawo Hardyego-Weinberga

Pierwszym ważnym rezultatem teoretycznym w genetyce populacyjnej jest prawo Hardy'ego-Weinberga. Przy spełnieniu przez populację pewnych założeń, prawo to pozwala na podstawie częstości poszczególnych alleli wyliczyć częstość poszczególnych genotypów.

Prawo Hardyego-Weinberga

Rozpatrzmy pojedynczy locus o k możliwych allelach $A_1, ..., A_k$. Powiemy, że populacja jest w warunkach *równowagi Hardyego-Weinberga* jeśli dwa allele w zadanym locus losowo wybranego osobnika populacji są stochastycznie niezależne i jednakowo prawdopodobne. To znaczy, że jeśli w całej populacji allele $A_1, ..., A_k$ występują z częstościami odpowiednio $p_1, ..., p_k$, to uporządkowana para alleli (A_i, A_j) zostanie zaobserwowana w zadanym locus u losowo wybranego osobnika z populacji z prawdopodobieństwem równym $p_i p_j$.

Prawo Hardyego-Weinberga

Założeni o pozostawaniu populacji w warunkach równowagi Hardyego-Weinberga pojawia się w genetyce bardzo często. Uzasadnieniem tego założenia może być fakt, iż populacja nie będąca w warunkach równowagi Hardyego-Weinberga już w wyniku jednego losowego kojarzenia osiąga równowagę Hardyego-Wainberga. Zademonstrujemy to w przypadku locus o dwóch allelach A i a.

Prawo Hardyego-Weinberga

Każdy osobnik populacji, ze względu na ten locus jest jednego z trzech nieuporządkowanych genotypów AA, Aa lub aa. Oznaczmy proporcje występowania tych genotypów w populacji odpowiednio przez P, $2Q$ i R.
Zakładamy, że nie mam różnic w tych proporcjach pomiędzy płciami. Tak więc P, Q i R są dowolnymi liczbami spełniającymi warunki

$$P,\, 2Q,\, R \in [0,\, 1] \quad \text{oraz} \quad P + 2Q + R = 1.$$

Prawo Hardyego-Weinberga

W populacji o liczebności N mamy

1. $(2P + 2Q)N$ alleli A
2. $(2R + 2Q)N$ alleli a
3. przy ogólnej liczbie wszystkich alleli równej $2N$.

Tak więc częstość alleli A i a w całej populacji jest równa odpowiednio $p_A = P + Q$ i $p_a = R + Q$.

Zgodnie z prawami Mendla prawdopodobieństwa genotypu potomka, pod warunkiem genotypu rodziców, są równe

Genotyp rodziców	Częstość	AA	Aa	aa
$AA \times AA$	P^2	1	-	-
$AA \times Aa$	$P2Q$	1/2	1/2	-
$AA \times aa$	PR	-	1	-
$Aa \times AA$	$2QP$	1/2	1/2	-
$Aa \times Aa$	$4Q^2$	1/4	1/2	1/4
$Aa \times aa$	$2QR$	-	1/2	1/2
$aa \times AA$	RP	-	1	-
$aa \times Aa$	$R2Q$	-	1/2	1/2
$aa \times aa$	R^2	-	-	1

Tabela 1. *Uporządkowane pary możliwych rodziców, ich częstość przy założeniu losowego kojarzenia oraz warunkowe prawdopodobieństwa genotyp ich potomków.*

Oznaczmy prawdopodobieństwa zdarzeń polegających na tym, że potomek jest genotypu AA, Aa oraz aa odpowiednio przez P_1, $2Q_1$ i R_1 wtedy

$$\begin{aligned} P_1 &= P^2 + \frac{1}{2}P2Q + \frac{1}{2}2QP + \frac{1}{4}(2Q)^2 = (P+Q)^2, \\ 2Q_1 &= \frac{1}{2}P2Q + PR + \frac{1}{2}2QP + \frac{1}{2}(2Q)^2 + \frac{1}{2}2QR + RP + \frac{1}{2}R2Q \\ &= 2(P+Q)(R+Q), \\ R_1 &= \frac{1}{4}(2Q)^2 + \frac{1}{2}2QR + \frac{1}{2}R2Q + R^2 = (Q+R)^2. \end{aligned}$$

Zauważmy, że z równań otrzymujemy następujące równości

$$P_1 = p_A^2,$$

$$2Q_1 = 2p_A p_a,$$

$$R_1 = p_a^2.$$

Innymi słowy, prawdopodobieństwo genotypu potomka jest takie samo jak prawdopodobieństwo genotypu losowo wybranej osoby z populacji będącej w równowadze Hardyego-Weinberga.

Z równań (1)-(2) możemy wywnioskować, że

$$Q_1^2 = P_1 R_1.$$

Jeśli równość tego typu jest prawdziwa dla początkowych częstości

$$P,\ Q,\ R, \quad (Q^2 = PR)$$

to ponieważ

$$P + 2Q + R = 1,$$

więc

$$\begin{aligned} P_1 &= (P+Q)^2 = P^2 + 2PQ + Q^2 = P^2 + 2PQ + PR = P, \\ R_1 &= (R+Q)^2 = Q^2 + 2RQ + R^2 = PR + 2RQ + R^2 = R, \\ Q_1 &= (P+Q)(Q+R) = \sqrt{P_1 R_1} = \sqrt{PR} = Q. \end{aligned}$$

Prawo Hardyego-Weinberga

Tak więc częstości P_1, Q_1, R_1 osiągnięte w populacji w drugiej generacji nie ulegną zmianie w następnych generacjach, powstających przez losowe krzyżowanie. W tym sensie mówimy o równowadze Hardyego-Weinberga.

Prawo Hardyego-Wainberga

Jeśli patrzymy na proces kojarzenia nie z punktu widzenia osobników ale z punktu widzenia komórek rozrodczych to wyprowadzenie prawa Hardyego-Weinberga staje sie jeszcze prostsze.

Prawo Hardyego-Wainberga

Zygota otrzymuje jeden allel od ojca i jeden od matki zgodnie z nastepujacym schematem:

Allel od ojca	**Allel od matki**	**Prawdopodobieństwo**
A	**A**	$p \times p = p^2$
A	**a**	$p \times q = pq$
a	**A**	$q \times p = pq$
a	**a**	$q \times q = q^2$

Prawo Hardyego-Wainberga

Założyliśmy tutaj, że zdarzenie otrzymania, na przykład **A** od ojca i **a** od matki są zdarzeniami niezależnymi i stąd prawdopodobieństwo takiego zdarzenia jest równe iloczynowi prawdopodobieństwa, otrzymania **A** od ojca, które jest równe **p**, oraz prawdopodobieństwa otrzymania **a** od matki, równego q.

Prawo Hardyego-Wainberga

Taka sytuacja zachodzi gdy kojarzenie jest losowe. Zauważmy, że genotyp **Aa** u potomka powstaje w dwóch sytuacjach, allel **A** od ojca i allel **a** od matki oraz allel **a** od ojca i allel **A** od matki. Stąd prawdopodobieństwo, że losowo wybrany osobnik z populacji potomków ma genotyp **Aa** jest równe $pq + qp = 2pq$ i genotypy **AA**, **Aa** i **aa** występuja w populacji potomków z częstościami odpowiednio p^2, $2pq$, i q^2.

Prawo Hardyego-Wainberga a rzeczywistość

Jak łatwo sobie wyobrazić, założenie o losowym kojarzeniu nie musi być spełnione, na przykład gdy allele maja wpływ na czas kwitnięcia u roślin lub na wybór partnera u zwierząt. Założenie o losowym kojarzeniu może być naruszone także w sytuacji gdy allele nie maja wpływu na dobór, na przykład w sytuacji gdy kojarzenie w obrębie krewnych staje się częste co miało miejsce w przypadku arystokracji.

Prawo Hardyego-Wainberga a rzeczywistość

Populacja, dla której możemy założyć zachodzenie prawa Hardyego-Weinberga musi spełniać następujące warunki:

1. organizmy są diploidalne
2. rozmnażają się płciowo
3. pokolenia nie zachodzą na siebie
4. osobniki kojarzą się losowo
5. populacja jest nieskończona
6. nie ma migracji
7. nie ma mutacji
8. dobór naturalny nie wpływa na badany locus.

Uwaga 1.

W praktyce możemy założyć, że populacji, w której występują czynniki wymienione w punktach (6)-(8) ale ich działanie się znosi będzie spełniała prawo Hardyego-Weinberga. Prawo Hardyego-Weinberga można rozszerzyć na dowolną liczbę alleli. W przypadku trzech alleli A_1, A_2, A_3 występujących z częstością odpowiednio

$$p_1, \quad p_2, \quad \text{oraz} \quad p_3,$$

czestośći występowania w populacji genotypów

$$A_1A_1, \quad A_2A_2, \quad A_3A_3, \quad A_1A_2, \quad A_1A_3, \quad A_2A_3$$

są równe odpowiednio

$$p^2, \quad p_2^2, \quad p_3^2, \quad 2p_1p_2, \quad 2p_1p_3, \quad 2p_2p_3.$$

Uwaga 2.

Może być wiele przyczyn dla który populacja nie znajduje się w warunkach równowagi Hardy'ego-Weinberga.

1. Na populacje mogą działać siły selekcji.
2. Dobór nie jest losowy.
3. Populacja nie jest jednorodna, na przykład składa się z dwóch populacji znajdujących się w warunkach równowagi Hardy'ego-Weinberga, ale o różnych czestościach alleli.
   W takiej sytuacji obserwujemy tak zwany efekt Wahlunda, czyli więcej homozygot i mniej heterozygot niż można by przewidywać na podstawie prawa Hardyego-Weinberga.
4. Poszczególne genotypy wchodzą do próby, na podstawie której sprawdzamy zachodzenie prawa Hardy'ego-Weinberga w populacji, z niejednakowymi prawdopodobieństwami.

Estymacja częstości alleli

Z diploidalnej populacji losujemy niezależnie n osobników i na podstawe tej próby losowej chcemy dokonać estymacji czestości allelu A w tej populacji.

PRZYKŁAD.

Przyjmijmy, że w próbie losowej o liczebności n zaobserwowaliśmy następujące liczebności poszczególnych genotypów.

AA	**Aa**	**aa**
n_{AA}	n_{Aa}	n_{aa}
49	26	25

Estymacja częstości alleli

Z diploidalnej populacji losujemy niezależnie n osobników i na podstawe tej próby losowej chcemy dokonać estymacji czestości allelu A w tej populacji.

PRZYKŁAD.

Przyjmijmy, że w próbie losowej o liczebności n zaobserwowaliśmy następujące liczebności poszczególnych genotypów.

AA	**Aa**	**aa**
n_{AA}	n_{Aa}	n_{aa}
49	26	25

Estymacja częstości alleli

Na podstawie tych danych częstość allelu A w tej populacji możemy estymować na przykład w następujacy sposób

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n} = \frac{98 + 26}{200} = 0,62.$$

Estymacja częstości alleli

Możemy też skorzystać z prawa Hardyego-Weinberga, na mocy którego częstość genotypu **AA**, p_{AA}, jest równa p_A^2 i dokonać estymacji p_A jedynie na podstawie n_{AA}.

$$\tilde{p}_A = \sqrt{\frac{n_{AA}}{n}} = \sqrt{0,49} = 0,7.$$

Estymacja częstości alleli

Możemy również skorzystać z faktu iż na mocy prawa Hardyego-Weinberga $p_A = 1 - p_a = 1 - \sqrt{p_{aa}}$ co prowadzi do estymatora

$$\hat{p}_A = 1 - \sqrt{\frac{n_{aa}}{n}} = 1 - \sqrt{\frac{25}{100}} = 0,5.$$

Estymacja metodą najwiekszej wiarogodnosci

Estymatorem, który jest najcześciej używanym w zagadnieniu estymacji częstości alleli jest estymator otrzymany metodą największej wiarogodności.

Estymacja metodą najwiekszej wiarogodności

Rozkład czestości genotypów jest rozkładem wielomainowym i przy założeniu zachodzenia dla badanej loci prawa Hardyego-Weinberga prawdopodobieństwo zaobserwowania w n elementowej, próbie losowej n_{AA}, n_{Aa} i n_{aa} osobników o genotypie odpowiednio, AA, Aa i aa jest równe

$$\binom{n}{n_{AA}}\binom{n-n_{AA}}{n_{Aa}}\binom{n-n_{AA}-n_{Aa}}{n_{aa}} p_A^{2n_{AA}}[2p_A(1-p_A)]^{n_{Aa}}[1-p_A]^{2n_{aa}},$$

gdzie p_A oznacza częstość allelu A w populacji.

Estymacja metodą najwiekszej wiarogodności

Stąd funkcja wiarogodności ma postać

$$\ln L(p) = \ln C + (2n_{AA} + n_{Aa}) \ln p_A + (n_{Aa} + 2n_{aa}) \ln(1 - p_A).$$

Maksimum funkcja wiarogodności osiąga dla

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n_{AA} + 2n_{Aa} + 2n_{aa}} = \frac{2n_{AA} + n_{Aa}}{2n}.$$