Infinite Allele Model with Fluctuating Population Size and the Effective Size of Population

Masaru Iizuka (飯塚 勝)

Division of Mathematics, Kyushu Dental College
2-6-1 Manazuru, Kokurakita-ku, Kitakyushu 803-8580, Japan

Abstract

The effective size of population has played an important role in population genetics. We consider the infinite allele model whose population size is a stationary stochastic process. For this model, the effects of fluctuation of population size on the effective size are investigated. The effective size is not smaller than the harmonic mean of population size and it is not larger than the arithmetic mean of population size. The effective size is equal to the harmonic mean if the fluctuation of population size is uncorrelated. To see explicitly the effects of the fluctuation of population size on the effective size, we introduce a two-valued Markov chain model. Results on the two-valued Markov chain model are compared with those of the Wright-Fisher model with fluctuating population size.

This is a joint work with H. Tachida and H. Matsuda (Kyushu University)

1. Infinite allele model

An infinite dimensional diffusion process \( \{ x(t) \}_{t \geq 0} = \{(x_1(t), x_2(t), \ldots) \}_{t \geq 0} \) on \( \{(x_1, x_2, \ldots) \in \mathbb{R}^\infty | x_i \geq 0, \sum_{i=1}^\infty x_i = 1 \} \) with the infinitesimal generator

\[
A = \frac{1}{2} \sum_{i,j=1}^\infty \frac{x_i (\delta_{ij} - x_j)}{N} \frac{\partial^2}{\partial x_i \partial x_j} - \sum_{i=1}^\infty u x_i \frac{\partial}{\partial x_i}
\]

is referred to as the infinite allele model in population genetics (Ewens (1979)). Here, \( x_i(t) \) is the relative frequency (gene frequency) of the \( i \)-th type (allele) \( A_i \) at time \( t \), \( N \) is the size of haploid population and \( u \) is the mutation rate per generation \( (N > 0, 0 < u < 1) \).

The genetic diversity of this population can be measured by the average heterozygosity (the probability that two genes sampled randomly from the population are different alleles)

\[
H(t) = 1 - \sum_{i=1}^\infty E[x_i(t)^2].
\]

By Kolmogorov forward equation, we have

\[
\frac{dH(t)}{dt} = -\frac{1}{N} H(t) + 2u \{1 - H(t)\},
\]
which has the solution
\[ H(t) = \frac{2Nu}{1 + 2Nu} + (H(0) - \frac{2Nu}{1 + 2Nu})e^{-(2u + \frac{1}{N})t} \]
and
\[ H(\infty) = \lim_{t \to \infty} H(t) = \frac{2Nu}{1 + 2Nu}. \]
By this formula, the population size \( N \) can be expressed as
\[ N = \frac{H(\infty)}{2u(1 - H(\infty))}. \] (2)

2. Infinite allele model with fluctuating population size

There is a lot of ecological data to the effect that the numbers of individuals in natural populations fluctuate considerably in each epoch and from generation to generation (Andrewartha and Birch (1954), Elton and Nicholson (1942) and Odum (1959)). The variations in population size are influenced by such factors as climate, the abundance of available resources, fluctuation in prey-predator balance, competition with other species using the same habitat (Nicholson (1957)).

Fluctuation of population size is not independent from generation to generation in general as in the case of stochastic selection (Gillespie and Guess (1978), Iizuka (1987), Iizuka and Matsuda (1982), Seno and Shiga (1984) and Takahata et al. (1975)).

Let \( N(t) \) be the size of haploid population at time \( t \). In this paper, we assume that \( \{N(t)\}_{t \geq 0} \) is a stationary process that is independent of the diffusion process \( \{x(t)\}_{t \geq 0} \) defined by Eq.(1). An infinite dimensional diffusion process in a random environment with the infinitesimal generator
\[ L = \frac{1}{2} \sum_{i,j=1}^{\infty} \frac{x_{i}(\delta_{ij} - x_{j})}{N(t)} \frac{\partial^{2}}{\partial x_{i} \partial x_{j}} - \sum_{1 \leq i < \infty} u \frac{\partial}{\partial x_{i}} \]
is referred to as the infinite allele model with fluctuating population size. For this model, the average heterozygosity \( H(t) \) satisfies
\[ \frac{dH(t)}{dt} = -\frac{1}{N(t)}H(t) + 2u\{1 - H(t)\}. \]
The solution of this differential equation is
\[ H(t) = 2u \int_{0}^{t} e^{\int_{s}^{t} (2u + \frac{1}{N(s)})ds} ds + H(0)e^{\int_{0}^{t} (2u + \frac{1}{N(s)})ds}. \]
Let
\[ N_{h} = E\left[ \frac{1}{N(t)} \right]^{-1} \]
be the harmonic mean of \( N(t) \). By applying Jensen's inequality for \( f(x) = e^{-x} \) \((x > 0)\), we have
\[ E[H(t)] \geq \frac{2N_{h}u}{1 + 2N_{h}u} \{1 - e^{-(2u + \frac{1}{N_{h}})t}\} + H(0)e^{-(2u + \frac{1}{N_{h}})t}. \]
and
\[ E[H(\infty)] \geq \frac{2N_{h}u}{1 + 2N_{h}u}. \]  
(3)

By Eq.(2), the effective size of population \( N_{e} \) can be defined as
\[ N_{e} = \frac{1}{2u} \frac{E[H(\infty)]}{1 - E[H(\infty)]}, \]  
(4)
and we have
\[ N_{e} \geq N_{h}, \]  
(5)
by Inequality (3). Inequality (5) shows that the effective size of population \( N_{e} \) is not smaller than the harmonic mean \( N_{h} \).

Let
\[ N_{a} = E[N(t)] \]
be the arithmetic mean of \( N(t) \). By applying Jensen’s inequality again for \( g(x) = -e^{-\frac{1}{x}} \) \((x > \frac{1}{2})\), we have
\[ E[H(t)] \leq \frac{2N_{a}u}{1 + 2N_{a}u} \{1 - e^{-(2u + \frac{1}{N_{a}})t} + H(0)e^{-(2u + \frac{1}{N_{a}})t}\} \]
and
\[ E[H(\infty)] \leq \frac{2N_{a}u}{1 + 2N_{a}u}. \]  
(6)
Then we have
\[ N_{e} \leq N_{a}, \]  
(7)
by Inequality (6). Inequality (7) shows that the effective size of population \( N_{e} \) is not larger than the arithmetic mean \( N_{a} \).

3. Two-valued Markov chain model

To see how the effective size of population \( N_{e} \) depends on the probability law of \( \{N(t)\}_{-\infty < t < \infty} \) and how \( N_{e} \) is different from \( N_{h} \) and \( N_{a} \), we consider a special case of two-valued Markov chain for \( \{N(t)\}_{-\infty < t < \infty} \). Let \( \{N(t)\}_{-\infty < t < \infty} \) be a Markov chain on \( \{N_{1}, N_{2}\} \) such that
\[ P(N(t + \Delta t) \neq N(t)|N(t) = N_{i}) = \gamma_{i}\Delta t, \]
\[ P(N(t + \Delta t) = N(t)|N(t) = N_{i}) = 1 - \gamma_{i}\Delta \quad (\Delta t \downarrow 0) \]
\((N_{1} < N_{2})\). Here, \( \gamma_{i} \) is the jump rate from the state \( N_{i} \) to \( N_{j} \) \((j \neq i)\). Since
\[ \text{Cov}[N(t_{1}), N(t_{2})] = (N_{2} - N_{1})^{2}r(1 - r)e^{-2\gamma|t_{1} - t_{2}|}, \]
where
\[ r = \frac{\gamma_{1}}{\gamma_{1} + \gamma_{2}} \]
and
\[ \gamma = \frac{\gamma_{1} + \gamma_{2}}{2}, \]
$N(t_1)$ and $N(t_2)$ are positively autocorrelated if $0 < \gamma < \infty$ and they are uncorrelated if $\gamma = \infty$ ($t_1 \neq t_2$).

In this case, the average heterozygosity $H(t)$ satisfies

$$\frac{dH(t)}{dt} = F_i(H(t))$$

if $N(t) = N_i$, where

$$F_i(x) = -\frac{x}{N_i} + 2u(1-x) = -\frac{2u}{H_i}(x - H_i), \quad (8)$$

and

$$H_i = \frac{2N_iu}{1+2N_iu}.$$  

Note that $H_1 \leq H(t) \leq H_2$ if $H_1 \leq H(0) \leq H_2$.

By Matsuda and Ishii (1981), the stationary probability density function of $H(t)$, $p(x)$, is given by

$$p(x) = C\{\frac{1}{F_2(x)} - \frac{1}{F_1(x)}\}exp\{-\int[\frac{\gamma_1}{F_1(x)} + \frac{\gamma_2}{F_2(x)}]dx\},$$

$(H_1 < x < H_2)$, where $C$ is the normalization constant. By this formula, we have

$$p(x) = \frac{\alpha_1 + \alpha_2}{B(\alpha_1, \alpha_2)(\alpha_2H_1 + \alpha_1H_2)(H_2 - H_1)^{\alpha_2 - 1}x(x - H_1)^{\alpha_1 - 1}(H_2 - x)^{\alpha_2 - 1}}$$

$(H_1 < x < H_2)$, where

$$\alpha_i = \frac{N_i\gamma_i}{1+2N_iu},$$

and

$$B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt \quad (p, q > 0).$$

Then we have

$$E[H(\infty)] = \int_{H_1}^{H_2} xp(x)dx = \frac{(1-r)H_1 + rH_2 + \frac{2}{u}H_1H_2}{1 + \frac{2}{u}(rH_1 + (1-r)H_2)}.$$  

By Eq.(4), we have

$$N_e = N_h + \frac{(N_2 - N_h)(N_h - N_1)}{2(\gamma + u)N_1N_2 + N_h}, \quad (9)$$

since

$$\frac{1}{N_h} = \frac{1 - r}{N_1} + \frac{r}{N_2}.$$  

By this formula, $N_e = N_h$ if and only if $\{N(t)\}_{-\infty<t<\infty}$ is uncorrelated ($\gamma = \infty$). In the literature, it is said that the effective size is equal to the harmonic mean of the population size when population sizes are not constant (Crow (1954), Nei et al. (1975) and Wright (1938)). Our result shows, however, that this is true if and only if stochastic fluctuations of population sizes are uncorrelated. This property was pointed out by Iizuka (1999) for the Wright-Fisher model with fluctuating population size.

Introducing a measure of autocorrelation

$$V = \int_0^\infty Cov[\frac{1}{N(0)}, \frac{1}{N(t)}]dt = \frac{r(1-r)}{2\gamma}(\frac{1}{N_1} - \frac{1}{N_2})^2,$$
we can express \( N_e \) as

\[
N_e = N_h\left(1 + \frac{N_h V}{1 + \frac{u}{\gamma} + \frac{1}{N_1\gamma_1 + N_2\gamma_2}}\right).
\] (10)

To see how \( N_e \) is larger than \( N_h \), we consider the ratio of the effective size to the harmonic mean

\[
\beta = \frac{N_e}{N_h} = 1 + \frac{(1-b)^2}{b} \frac{r(1-r)}{1 + 2(1 - r + br)N_2(\gamma + u)},
\]

where

\[
b = \frac{N_1}{N_2} \quad (0 < b < 1).
\]

Numerical examples for the case of \( \gamma_1 = \gamma_2 \) and \( N_2u = 1 \) are presented in Table 1.

<table>
<thead>
<tr>
<th>( b )</th>
<th>2( N_2\gamma )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>10</td>
<td>1.29</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>1.96</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>2.26</td>
</tr>
<tr>
<td>0.01</td>
<td>10</td>
<td>4.74</td>
</tr>
<tr>
<td>0.01</td>
<td>1</td>
<td>13.19</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>16.75</td>
</tr>
<tr>
<td>0.001</td>
<td>10</td>
<td>39.35</td>
</tr>
<tr>
<td>0.001</td>
<td>1</td>
<td>125.69</td>
</tr>
<tr>
<td>0.001</td>
<td>0.1</td>
<td>161.93</td>
</tr>
</tbody>
</table>

The ratio \( \beta \) of the effective size \( N_e \) to the harmonic mean \( N_h \) in the case of \( \gamma_1 = \gamma_2 \) and \( N_2u = 1 \).

Table 1 shows that ratio \( \beta \) can be very large if \( N_1 \) is much smaller than \( N_2 \) and the jump rate \( \gamma \) is not large.

Noting that

\[
N_a = (1-r)N_1 + rN_2,
\]

and

\[
N_1\gamma_1 + N_2\gamma_2 = \frac{N_a - 1}{N_h V},
\]

Eq.(10) can be expressed as

\[
N_e = N_h\left(1 + \frac{N_h V}{1 + \frac{u}{\gamma} + N_h V\frac{N_a - 1}{N_h}}\right) = cN_a + (1 - c)N_h,
\] (11)

where

\[
c = \frac{N_h V}{N_h V + (1 + \frac{u}{\gamma})(\frac{N_a}{N_h} - 1)}
\]

\((0 < c < 1)\).

By Eq.(11), we have

\[N_h < N_e < N_a\]
if $\gamma < \infty$. That is, the effective size is larger than the harmonic mean and smaller than the arithmetic mean of $N(t)$ if $\gamma$ is finite.

Finally, we consider the asymptotic behavior of $N_e$. By Eq.(11), we have

$$N_e \to N_h$$

if $N_h V \to 0$ (weak autocorrelation),

$$N_e \to c_\alpha N_a + (1 - c_\alpha)N_h$$

if $N_h V \to \alpha$ with $0 < \alpha < \infty$ (moderate autocorrelation) and

$$N_e \to N_a$$

if $N_h V \to \infty$ (strong autocorrelation), assuming that $0 < b, r < 1$ and $0 \leq \frac{u}{\gamma} < \infty$. Here, we put

$$c_\alpha = \frac{\alpha}{\alpha + (1 + \frac{u}{\gamma})(\frac{N_a}{N_h} - 1)}$$

$(0 < c_\alpha < 1)$. Note that weak autocorrelation, moderate autocorrelation and strong autocorrelation correspond to $N_1 \gamma_1 + N_2 \gamma_2 \to \infty$, $N_1 \gamma_1 + N_2 \gamma_2 \to \Gamma$ with $0 < \Gamma < \infty$ and $N_1 \gamma_1 + N_2 \gamma_2 \to 0$, respectively.

4. $K$ allele diffusion model with fluctuating population size

For $K$ allele diffusion model, the results of the previous sections hold with minor modifications ($2 \leq K < \infty$). We summarize the results on $K$ allele diffusion model with fluctuating population size in this section.

A $K$-1 dimensional diffusion process \{$x^{(K)}(t)\}_{t \geq 0} = \{(x_1(t), x_2(t), \ldots, x_{K-1}(t))\}_{t \geq 0}$ on

$\{(x_1, x_2, \ldots, x_{K-1}) \in \mathbb{R}^{K-1}| x_i \geq 0, \sum_{i=1}^{K-1} x_i \leq 1\}$

with the infinitesimal generator

$$L_K = \frac{1}{2} \sum_{i,j=1}^{K-1} \frac{x_i (x_i - x_j)}{N(t)} \frac{\partial^2}{\partial x_i \partial x_j} - \sum_{i=1}^{K-1} \frac{u}{K-1} (1 - K x_i) \frac{\partial}{\partial x_i}$$

is referred to as the $K$ allele diffusion model with fluctuating population size. For this model, the average heterozygosity,

$$H(t) = 1 - \sum_{i=1}^{K} E[x_i(t)^2],$$

satisfies

$$\frac{dH(t)}{dt} = -\frac{1}{N(t)} H(t) + 2u \{1 - \frac{K}{K-1} H(t)\},$$

and the effective size of population can be defined by

$$N_e = \frac{1}{2u (1 - \frac{K}{K-1} E[H(\infty)])},$$
where we put
\[ x_K(t) = 1 - \sum_{i=1}^{K-1} x_i(t). \]

Inequalities (5) and (7) hold for this model again.

For the two-valued Markov chain model, Eq.(8) is modified as
\[ F_i(x) = -\frac{x}{N_i} + 2u(1 - \frac{K}{K-1}x) = \frac{2u}{H_i}(x - H_i), \]
with
\[ H_i = \frac{2N_i u}{1 + 2N_i \frac{K}{K-1} u}. \]

Eqs.(9) and (11) are modified as
\[ N_e = N_h + \frac{(N_2 - N_h)(N_h - N_1)}{2(\gamma + \frac{K}{K-1} u) N_1 N_2 + N_h}, \]
and
\[ N_e = N_h(1 + \frac{N_h V}{1 + \frac{K}{K-1} \frac{u}{\gamma} + N_h V \frac{N_j}{N_a - N_h}}) = c^K N_a + (1 - c^K) N_h, \]
where
\[ c^K = \frac{N_h V}{N_h V + (1 + \frac{K}{K-1} \frac{u}{\gamma})(\frac{N_a}{N_h} - 1)} \]
\((0 < c^K < 1)\).

For the asymptotic behavior of \(N_e\), we have
\[ N_e \rightarrow N_h \]
if \(N_h V \rightarrow 0\),
\[ N_e \rightarrow c^K_{\alpha} N_a + (1 - c^K_{\alpha}) N_h \]
if \(N_h V \rightarrow \alpha\) with \(0 < \alpha < \infty\) and
\[ N_e \rightarrow N_a \]
if \(N_h V \rightarrow \infty\), assuming that \(0 < b, r < 1\) and \(0 \leq \frac{u}{\gamma} < \infty\). Here, we put
\[ c^K_{\alpha} = \frac{\alpha}{\alpha + (1 + \frac{K}{K-1} \frac{u}{\gamma})(\frac{N_a}{N_h} - 1)} \]
\((0 < c^K_{\alpha} < 1)\).

5. Comparison with the Wright-Fisher model with fluctuating population size

The Wright-Fisher model with fluctuating population size is analyzed by Iizuka (1999). In this section, we compare the results of this model with those of the models in the previous sections.
Let \( N^{(k)} \) be the size of haploid population in generation \( k \). In this section, we assume that \( \{N^{(k)}\}_{k=0,\pm 1,\pm 2,...} \) is a stationary Markov chain on \( \{N_1, N_2\} \) such that

\[
P(N^{(k+1)} \neq N^{(k)} | N^{(k)} = N_i) = q_i,
\]

\[
P(N^{(k+1)} = N^{(k)} | N^{(k)} = N_i) = 1 - q_i
\]

and

\[
P(N(0) = N_i) = p_i^{(st)},
\]

where \( 1 < N_1 < N_2 < \infty, 0 \leq q_i \leq 1, q_1 + q_2 > 0 \) and

\[
(p_1^{(st)}, p_2^{(st)}) = \left( \frac{q_2}{q_1 + q_2}, \frac{q_1}{q_1 + q_2} \right)
\]

is the stationary distribution of the Markov chain. The parameter \( q_i \) is the jump probability from the state \( N_i \) to \( N_j \) \((i, j = 1, 2 \text{ and } j \neq i)\).

Note that the covariance of \( N^{(k)} \) and \( N^{(k+j)} \) is

\[
Cov[N^{(k)}, N^{(k+j)}] = \frac{q_1 q_2 (1 - q_1 - q_2)^j}{(q_1 + q_2)^2} (N_2 - N_1)^2.
\]

This formula means that constant population size, periodic change per generation, positively autocorrelated fluctuation, uncorrelated fluctuation, and negatively autocorrelated fluctuation correspond to \( q_1 q_2 = 0, q_1 = q_2 = 1, 0 < q_1 + q_2 < 1 \) with \( q_1 q_2 \neq 0, q_1 + q_2 = 1 \) with \( q_1 q_2 \neq 0, \) and \( 1 < q_1 + q_2 < 2 \), respectively. The harmonic mean of \( N^{(k)} \) is

\[
N_h^* = E\left[ \frac{1}{N^{(k)}} \right]^{-1} = \left( \frac{p_1^{(st)}}{N_1} + \frac{p_2^{(st)}}{N_2} \right)^{-1} = (q_1 + q_2) \left( \frac{q_2}{N_1} + \frac{q_1}{N_2} \right)^{-1}
\]

For given \( \{N^{(k)}\}_{k=0,\pm 1,\pm 2,...} \), we consider a haploid population with two alleles \( A_1 \) and \( A_2 \). The population size in generation \( k \) is \( N^{(k)} \). Let \( Z^{(k)} \) be the number of allele \( A_1 \) in generation \( k \). Then \( \{Z^{(k)}\}_{k=0,1,2,...} \) is a discrete time Markov chain with

\[
P(Z^{(k+1)} = j | Z^{(k)} = i) = \binom{N^{(k)}}{j} \left( \frac{i}{N^{(k)}} \right)^j \left( 1 - \frac{i}{N^{(k)}} \right)^{N^{(k+1)} - j},
\]

\( 0 \leq i \leq N^{(k)}, 0 \leq j \leq N^{(k+1)} \). Let \( x^{(k)} = \frac{Z^{(k)}}{N^{(k)}} \) be the gene frequency of allele \( A_1 \) in generation \( k \). The process \( \{x^{(k)}\}_{k=0,1,2,...} \) is referred to as the Wright-Fisher model with fluctuating population size. Note that there is no mutation in this model.

The effective size \( N_e^* \) can be defined by

\[
1 - \frac{1}{N_e^*} = \lim_{k \to \infty} \left\{ E\left[ \prod_{j=0}^{k-1} \left( 1 - \frac{1}{N(j)} \right) \right] \right\}^{\frac{1}{k}},
\]

and we have

\[
N_e^* = \frac{1}{1 - \alpha_+},
\]

where \( \alpha_+ \) is the largest solution to

\[
\alpha^2 - \left\{ (1 - \frac{1}{N_1})(1 - q_1) + (1 - \frac{1}{N_2})(1 - q_2) \right\} \alpha + (1 - \frac{1}{N_1})(1 - \frac{1}{N_2})(1 - q_1 - q_2) = 0.
\]
(see Iizuka (1999)).

It is shown that $N_e^* > N_h^*$ if $0 < q_1 + q_2 < 1$, $N_e^* = N_h^*$ if $q_1 + q_2 = 1$ and $N_e^* < N_h^*$ if $1 < q_1 + q_2 < 2$. The property that the effective size is equal to the harmonic mean if and only if $\{N^{(k)}\}_{k=0,1,2,...}$ is uncorrelated stochastic process is consistent with the result for the infinite allele model with fluctuating population size. For the infinite allele model with fluctuating population size, however, the case of negative autocorrelation does not exist. The reason is that the stochastic effect of fluctuation of population size is incorporated into the ordinary differential equation for $H(t)$ in Section 2 and this is a stochastic ordinary differential equation model (see Iizuka (1987)).

The Wright-Fisher model with fluctuating population size is more fundamental than the infinite allele model and $K$ allele diffusion model with fluctuating population size since the stochastic effect of fluctuation of population size is introduced after diffusion approximation for the later models. It seems not to be easy, however, to incorporate the effect of mutation into the Wright-Fisher model with fluctuating population size. It is easy to incorporate the effect of mutation into the infinite allele model and $K$ allele diffusion model with fluctuating population size since the differential equation for $H(t)$ is linear as far as there is no selection as we have seen in Section 2 and Section 3.

This research was partially supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan.

References


