An alternative approach to asymptotic results on genetic composition when the population size is varying

C. C. Heyde

CSIRO Division of Mathematics and Statistics, Box 1965, P.O. Canberra City, A.C.T. 2601, Australia

Abstract. This paper is concerned with methodology for studying the long-term genetic composition of a population of haploid individuals in the case where the population size is varying. A general approach requiring a minimum of assumptions is described based on constructing martingales out of expressions for the means of the numbers of allelic types, conditional on the past. Earlier investigations were based on studying the asymptotic behaviour of the proportion of alleles of a certain type in the population. Several applications demonstrate that the approach suggested in the paper provides results which usefully complement ones obtained previously.

Key words: Allelic composition — Ultimate homozygosity — Polymorphism — Growth of conditional means — Martingale methods

1. Introduction

We consider a population whose individual members may have two possible alleles a or A at a single locus. Suppose that the stochastic processes $\{X_n, n \geq 0\}$ and $\{Y_n, n \geq 0\}$ track the numbers of a and A alleles respectively in the population and write $\{N_n = X_n + Y_n, n \geq 0\}$ for the corresponding population size process. In the case of constant N_n , classical models due to Wright–Fisher and to Moran (see e.g. Ewens (1979, Chapter 3)) have been formulated to study the changing allelic composition of the population and various generalizations dealing with varying N_n have been introduced. The asymptotic behaviour of the proportion of a alleles $Z_n = X_n / N_n$ has been the major focus and in particular whether $\{Z_n\}$ converges a.s. and, if it does, when $P(Z_n \rightarrow 0 \text{ or } 1) = 1$. See for example Seneta (1974), Heyde and Seneta (1975), Heyde (1977, 1981, 1982, 1983), Daley, Hall and Heyde (1982).

The approach via the study of the process $\{Z_n\}$ has provided comprehensive results but there have been several important deficiencies. Detailed structural assumptions have been necessary for progress and the underlying similarity of the results coming from different models lacks adequate explanation. In the (most common) case where $P(Z_n \rightarrow 0 \text{ or } 1) = 1$, only crude information is available on

the individual asymptotic behaviour of $\{X_n\}$ and $\{Y_n\}$. There has not been a simple and intuitive method for discriminating between the cases where $P(Z_n \rightarrow 0 \text{ or } 1) = 1$ and $P(Z_n \rightarrow Z) = 1$ with $P(0 < Z < 1) > 0$. Finally, in cases where extinction of the population is possible, the use of $Z_n = X_n/N_n$ raises awkward conditioning problems. These deficiencies can be largely avoided if the processes $\{X_n\}$ and $\{Y_n\}$ are studied directly and separately. In fact, it is the behaviour of the conditional means, given the past, of X_n and Y_n that is crucial. Considerable information can be obtained without the need for any concrete distributional assumptions.

2. Methods

Quite a wide variety of population models give rise to evolutionary equations for conditional means of the form

$$\begin{aligned} E(X_{n+1} | \mathcal{F}_n) &= X_n(1 + f_n) \quad \text{a.s.,} \\ E(Y_{n+1} | \mathcal{F}_n) &= Y_n(1 + g_n) \quad \text{a.s.,} \end{aligned} \quad (1)$$

where $\mathcal{F}_n$ is the (past history) σ -field generated by $\{X_k, Y_k, 0 \leq k \leq n\}$, f_n and g_n are $\mathcal{F}_n$ -measurable and $1 + f_n \geq 0$ a.s., $1 + g_n \geq 0$ a.s. In particular, all the above cited papers deal with models whose behaviour follows the prescription (1). The models of Heyde (1981, 1982, 1983) and Daley et al. (1982) based on extensions of the Moran model all give the form (1) with

$$\begin{aligned} f_n &= [(1 + \sigma)EU / (1 + \sigma Z_n) - 1] / N_n, \\ g_n &= [EV / (1 + \sigma Z_n) - 1] / N_n, \end{aligned} \quad (2)$$

while the Wright–Fisher model extension treated in Heyde (1977) gives (1) with

$$\begin{aligned} f_n &= [(1 + \sigma)E(N_{n+1} | \mathcal{F}_n) / (1 + \sigma Z_n)N_n] - 1 \\ g_n &= [E(N_{n+1} | \mathcal{F}_n) / (1 + \sigma Z_n)N_n] - 1. \end{aligned} \quad (3)$$

(In each case σ represents a selective advantage. For (3) the $\sigma = 0$ case had previously been discussed in Heyde and Seneta (1975) and the work originated with Seneta 1974.)

To analyse the system (1) we note that

$$\left\{ X_n / \prod_{j=0}^{n-1} (1 + f_j), \mathcal{F}_n, n \geq 1 \right\}, \quad \left\{ Y_n / \prod_{j=0}^{n-1} (1 + g_j), \mathcal{F}_n, n \geq 1 \right\}$$

are non-negative martingales with means EX_0 and EY_0 respectively. Consequently, the martingale convergence theorem gives

$$X_n / \prod_{j=0}^{n-1} (1 + f_j) \xrightarrow{\text{a.s.}} \alpha, \quad Y_n / \prod_{j=0}^{n-1} (1 + g_j) \xrightarrow{\text{a.s.}} \beta \quad (4)$$

for some random variables $\alpha \geq 0$ a.s. and $\beta \geq 0$ a.s. with $E\alpha \leq EX_0$, $E\beta \leq EY_0$. Of course α and/or β might be a.s. zero but they can often be identified in particular cases as we shall observe below.

The comparison of growth rates of X_n and Y_n is always crucial and we shall distinguish the cases (i) $f_n = g_n$ for all n (balanced) and (ii) $f_n \neq g_n$ for some n (unbalanced).

In the balanced case we note that

$$N_n / \prod_{j=0}^{n-1} (1 + f_j) \xrightarrow{\text{a.s.}} \alpha + \beta$$

and on the set $\{\alpha + \beta > 0\}$,

$$Z_n = X_n / N_n \xrightarrow{\text{a.s.}} \alpha / (\alpha + \beta).$$

In the unbalanced case we note that if $\prod_{j=0}^{\infty} [(1 + f_j)/(1 + g_j)] = 0$ or ∞ a.s. the growth rates of the numbers of the two alleles are intrinsically different and we expect $Z_n \rightarrow 0$ or 1 a.s.

Equations (1) with f_n and g_n given by (2) are of the balanced type if and only if $(1 + \sigma)EU = EV$ while for (3) the corresponding condition is $\sigma = 0$. In the case of (2) it is clear that $Z_n \rightarrow 0$ or 1 can be expected if $(1 + \sigma)EU \neq EV$.

The basic approach via (1) and (4) thus requires a minimum of structural assumptions, much less than for the models in the above cited papers on allelic composition of populations. In the next section we shall show how it can usefully complement results already obtained in the literature using the specific models and the more technically complex and rather less transparent approach via the study of $\{Z_n, n \geq 0\}$.

3. Applications

For our first application we shall take the setting of Daley, Hall and Heyde (1982) (henceforth known as DHH) which is a little more general than the original formulation of Heyde (1981, 1982).

Now it is established in DHH that, provided $\min(EU, EV) > 1$,

$$Z_n = X_n / N_n \xrightarrow{\text{a.s.}} Z \quad (5)$$

and

$$N_n \sim nC(Z) \quad (6)$$

as $n \rightarrow \infty$ where

$$C(Z) = [(1 + \sigma)ZEU + (1 - Z)EV - (1 + \sigma Z)] / (1 + \sigma Z).$$

We note that if $\{h_j, j \geq 0\}$ is a sequence of positive constants with $h_j \sim j^{-1} \theta$ as $j \rightarrow \infty$ then

$$\prod_{j=0}^{n-1} (1 + h_j) \sim n^{\theta} \quad (7)$$

as $n \rightarrow \infty$, since

$$\log \prod_{j=0}^{n-1} (1 + h_j) = \sum_{j=0}^{n-1} \log(1 + h_j) \sim \sum_{j=0}^{n-1} h_j \sim \theta \sum_{j=1}^{n-1} j^{-1} \sim \theta \log n,$$

the second last step being a consequence of

$$\lim_{n \rightarrow \infty} (\log n)^{-1} \sum_{j=1}^{n-1} j^{-1} (jh_j - \theta) = 0.$$

Then, from (1), (2), (5), (6) and (7), we have

$$\begin{aligned} \prod_{j=0}^{n-1} (1 + f_j) &\sim n^{a(Z)} \\ \prod_{j=0}^{n-1} (1 + g_j) &\sim n^{b(Z)} \end{aligned} \quad (8)$$

as $n \rightarrow \infty$ where

$$a(Z) = \left[\frac{(1 + \sigma)EU}{1 + \sigma Z} - 1 \right] / C(Z)$$

and

$$b(Z) = \left[\frac{EV}{1 + \sigma Z} - 1 \right] / C(Z).$$

It is clear from (8) that the intrinsic rates of increase of X_n and Y_n are different unless $(1 + \sigma)EU = EV$.

From (4) and (8),

$$n^{-a(Z)} X_n \xrightarrow{\text{a.s.}} \alpha(Z), \quad n^{-b(Z)} Y_n \xrightarrow{\text{a.s.}} \beta(Z). \quad (9)$$

However, we already know from (5) and (6) that

$$\begin{aligned} n^{-1} X_n &\xrightarrow{\text{a.s.}} ZC(Z) \\ n^{-1} Y_n &\xrightarrow{\text{a.s.}} (1 - Z)C(Z). \end{aligned} \quad (10)$$

Note that

$$a(Z) \leq 1 \Leftrightarrow (1 + \sigma)EU \leq EV \text{ or } Z = 1$$

$$b(Z) \leq 1 \Leftrightarrow (1 + \sigma)EU \geq EV \text{ or } Z = 0.$$

Thus, if $(1 + \sigma)EU = EV$ we see that (9) and (10) are identical. On the other hand, if $(1 + \sigma)EU \neq EV$, then it follows from Theorem 1 of DHH that $Z = 0$ or 1 with probability one and we obtain from (9) and (10) that if $EV > (1 + \sigma)EU$

$$n^{-(1 + \sigma)EU - 1) / (EV - 1)} X_n \rightarrow \alpha(0), \quad n^{-1} Y_n \rightarrow EV - 1 \quad \text{a.s. on } \{Z = 0\}$$

$$n^{-1} X_n \rightarrow EU - 1, \quad n^{-1} Y_n \rightarrow 0 \quad \text{a.s. on } \{Z = 1\},$$

while if $EV < (1 + \sigma)EU$,

$$\begin{aligned} n^{-1} X_n &\rightarrow 0 \\ n^{-1} Y_n &\rightarrow EV - 1 \end{aligned} \quad \text{a.s. on } \{Z = 0\},$$

$$\begin{aligned} n^{-1} X_n &\rightarrow (1 + \sigma)EU - 1 \\ n^{-(EV - (1 + \sigma)EU) / ((1 + \sigma)EU - (1 + \sigma))} Y_n &\rightarrow \beta(1) \end{aligned} \quad \text{a.s. on } \{Z = 1\}.$$

Of course $\alpha(0)$ and $\beta(1)$ could be zero but, nevertheless, differences in growth rates of X_n and Y_n are highlighted.

A particular advantage of the approach via (1) is that it allows us to straightforwardly cope with the possibility of death without replacement in the model. We shall, however, omit the details since the basic results are unchanged. The approach used previously, based on the study of $X_n/(X_n + Y_n)$, raises awkward conditioning problems if extinction is possible.

As a second application we take $N_n = c_n, \{c_n\}$ being an increasing sequence of positive constants, in the model treated by Heyde (1977) which gives rise to (3). Then,

$$1 + f_n = \frac{(1 + \sigma)c_{n+1}}{c_n + \sigma X_n} \sim \frac{(1 + \sigma)c_{n+1}}{c_n(1 + \sigma Z)} \quad \text{a.s.},$$

$$1 + g_n = \frac{c_{n+1}}{c_n + \sigma X_n} \sim \frac{c_{n+1}}{c_n(1 + \sigma Z)} \quad \text{a.s.},$$

it being known that $c_n^{-1} X_n \xrightarrow{\text{a.s.}} Z$ as $n \rightarrow \infty$. Thus, from (4),

$$\begin{aligned} [(1 + \sigma Z)/(1 + \sigma)]^n c_0 c_n^{-1} X_n &\xrightarrow{\text{a.s.}} \alpha(Z) \\ (1 + \sigma Z)^n c_0 c_n^{-1} Y_n &\xrightarrow{\text{a.s.}} \beta(Z) \end{aligned} \quad (11)$$

and the intrinsic rates of increase of X_n and Y_n are the same if and only if $\sigma = 0$.

First suppose $\sigma = 0$; we can identify $\alpha(Z) = c_0 Z$ and $\beta(Z) = c_0(1 - Z)$ and it is known (Seneta (1974)) that $P(0 < Z < 1) > 0$ if $\sum c_n^{-1} < \infty$ while $P(\{Z = 0\} \cup \{Z = 1\}) = 1$ if $\sum c_n^{-1} = \infty$.

Now suppose that $\sigma > 0$. In this case we have $P(\{Z = 0\} \cup \{Z = 1\}) = 1$ (Heyde (1977)). On $\{Z = 0\}$ we have $c_n^{-1} X_n \rightarrow 0$, $c_n^{-1} Y_n \rightarrow 1$, so that identification with (11) gives $\alpha(0) = 0$, $\beta(0) = c_0$. Also, on $\{Z = 1\}$ we have $c_n^{-1} X_n \rightarrow 1$, $c_n^{-1} Y_n \rightarrow 0$ so that, from (11), $\alpha(1) = c_0$. It is not clear whether $\beta(1) = 0$ in general but if $(1 + \sigma)^n c_n^{-1} \rightarrow \infty$ then $\beta(1) = 0$. This last result is implied by $Y_N = 0$ for some sufficiently large (random) N which holds since, if not, then $Y_n \geq 1$ for all n and hence $(1 + \sigma)^n c_n^{-1} Y_n \rightarrow \infty$ contradicting (11). Nevertheless, the difference in growth rates of X_n and Y_n is highlighted.

References

- Daley, D. J., Hall, P., Heyde, C. C.: Further results on the survival of a gene represented in a founder population. *J. Math. Biology* **14**, 355–363 (1982)
- Ewens, W. J.: *Mathematical population genetics*, Berlin–Heidelberg–New York: Springer 1979
- Heyde, C. C.: The effect of selection on genetic balance when the population size is varying. *Theoret. Population Biology* **11**, 249–251 (1977)
- Heyde, C. C.: On the survival of a gene represented in a founder population. *J. Math. Biology* **12**, 91–99 (1981)
- Heyde, C. C.: The effect of differential reproductive rates on the survival of a gene represented in a founder population. In: (Gani, J., Hannan, E. J., eds.) *Essays in statistical science. J. Applied Prob.* **19A**, 19–25 (1982)
- Heyde, C. C.: On limit theorems for gene survival. *Colloquia Mathematica János Bolyai* (1983) in press

Heyde, C. C., Seneta, E.: The genetic balance between random sampling and random population size. *J. Math. Biology* **1**, 317–320 (1975)

Seneta, E.: A note on the balance between random sampling and population size. *Genetics* **77**, 607–610 (1974)

Received February 14/Revised April 19, 1983