Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

DIFFUSION PROCESS CALCULATIONS FOR MUTANT GENES IN NONSTAYIONARY POPULATIONS

BY RUZONG FAN AND KENNETH LANGE[1]

University of Michigan

Diffusion process approximations were introduced into population genetics by Fisher and Wright and perfected by Kimura. Contrary to popular scientific opinion, these pioneers did not solve all of the interesting modeling problems. For instance, none of them has much to say about the stochastic dynamics of recessive disease genes. They are also more or less silent on the stochastic aspects of evolution in the presence of exponential population growth. The current paper uses Itô's formula to derive an infinite hierarchy of integral equations satisfied by the moments of a diffusion process. These integral equations can be converted into an infinite hierarchy of ordinary differential equations and solved either exactly or numerically. We illustrate some of the possibilities for dominant, neutral, and recessive models of inheritance by computing the moments of gene frequencies in the presence of exponential population growth.

1. Introduction. The evolutionary forces governing the distribution and dynamics of human genetic diseases can be modeled in a variety of ways. The earliest and most understandable models are deterministic (Cavalli-Sforza and Bodmer 1971, Crow and Kimura 1970, Ewens 1979, Nagylaki 1992). Later models attempt to capture the more subtle stochastic effects that inevitably come into play. For autosomal dominant or X-linked diseases, branching process models are ideal (Fisher 1930, Haldane 1927, Harris 1989, Skellam 1949). By viewing each new disease mutation as the progenitor of an independently evolving clan of deleterious gene carriers, one can answer a host of interesting population genetic questions (Gladstien and Lange 1978a, Gladstien and Lange 1978b, Lange and Gladstien 1980, Lange 1982). We have recently extended these branching process models to include exponential growth of the surrounding population of normal individuals (Lange and Fan 1997, Fan and Lange 1998).

For recessive diseases, selection occurs when carrier individuals from the same or different clans mate. Thus, the independence assumption of the branching process paradigm breaks down. Although the alternative Wright-Fisher model of evolution eschews the dubious assumption of independently evolving clans, it has yielded, contrary to popular scientific opinion, little insight into the balance between selection and mutation for recessive diseases (Crow and

[1]Research supported in part by USPHS grant GM53275.
AMS 1991 *subject classifications.* Primary 60J25, 60J60, 60J65, 60J70, 62P10; secondary 92D10.
Key words and phrases. Brownian motion, diffusion processes, Itô integral, mutant genes, population genetics, stochastic differential equations.

Kimura 1970, Ewens 1979). Even the enormously prolific Kimura was largely silent on the question of recessive diseases and the impact of population growth on their dynamics.

The present paper crafts some new calculational tools for the Wright-Fisher model. Following the lead of Kimura (Crow and Kimura 1970), we immediately pass to the diffusion approximation of the Wright-Fisher model. This puts the considerable machinery of stochastic integration at our disposal (Chung and Williams 1990). Within this framework, we derive infinite hierarchies of integral and ordinary differential equations for the moments of a diffusion process. When the infinitesimal mean of the diffusion process is linear in its spatial variable and the infinitesimal variance is quadratic, these equations can often be solved exactly (Fan *et al.* 1998). When the infinitesimal mean and variance are polynomials in their spatial variables, we show how the hierarchy of differential equations can still be solved numerically. These results are of independent interest quite apart from their applications in population genetics. Our solutions specifically incorporate time inhomogeneities such as growth of the surrounding normal population.

With these introductory comments in mind, Section 2 briefly reviews the Wright-Fisher model and its classical diffusion approximation to the frequency of a neutral or deleterious gene. Section 3 derives via Itô's formula the aforementioned infinite hierarchies of integral and differential equations for the moments of a diffusion process. Sections 4 and 5 then discuss exact moment calculation for the neutral and dominant versions of the Wright-Fisher model. Section 6 describes the numerical techniques used for the recessive disease moments featured in the examples of Section 7. Finally, our concluding discussion suggests limitations of the genetic models and raises open problems about the rigor of the numerical methods.

2. Wright-Fisher genetic model. The Wright-Fisher model for the evolution of a deleterious or neutral gene postulates (a) discrete generations, (b) finite population size, (c) no immigration, and (d) formation of gametes by random binomial sampling. In assumption (d), each current population member contributes to an infinite pool of potential gametes in proportion to his or her fitness. Mutation from the normal allele a to the deleterious allele b takes place at this stage with mutation rate η; backmutation is not permitted. In the neutral model we neglect mutation and treat the two alleles symmetrically. Once the pool of potential gametes is formed, actual gametes are sampled randomly. Although the three genotypes occur in the usual Hardy-Weinberg proportions just after gamete sampling, selection causes allele frequencies to change over time.

If we let $w_{a/a}$, $w_{a/b}$, and $w_{b/b}$ denote the average fitnesses of the three genotypes a/a, a/b and b/b of an autosomally determined trait, then for a dominant disease we may suppose that $w_{a/a} = 1$ and $w_{a/b} = w_{b/b} = f < 1$. For a neutral trait, $w_{a/a} = w_{a/b} = w_{b/b} = 1$, and for a recessive disease $w_{a/a} = w_{a/b} = 1$ and $w_{b/b} = f < 1$. For our purposes, the population size N_m at generation m need not be constant. The primary object of study in this paper is the frequency X_m of allele b at generation m. This frequency is the ratio of the total number Y_m of b alleles to the total number of genes $2N_m$. The Wright-Fisher model specifies that Y_m is binomially distributed with $2N_m$ trials and success probability $p(X_{m-1})$ determined by the proportion $p(X_{m-1})$ of b alleles in the pool of potential gametes for generation m. In passing to a diffusion approximation, we take one generation as the unit of time and substitute

$$\begin{aligned}
\mu(m, x_m) &= \mathrm{E}\,(X_{m+1} - X_m \mid X_m = x_m) \\
&= p(x_m) - x_m \\
\sigma^2(m, x_m) &= \mathrm{Var}(X_{m+1} - X_m \mid X_m = x_m) \\
&= \frac{p(x_m)[1 - p(x_m)]}{2N_{m+1}}
\end{aligned}$$

for the infinitesimal mean $\mu(t, x)$ and variance $\sigma^2(t, x)$ of the diffusion process evaluated at time $t = m$ and position $x = x_m$ (Crow and Kimura 1970, Ewens 1979). Given the assumption of exponential growth in a human population, the population size at generation m is $N_m = N_0 e^{cm}$ for some growth rate c.

Under neutral evolution, the gamete success probability is $p(x) = x$. This formula for $p(x)$ entails no systematic tendency for either allele to expand at the expense of the other allele. For a dominant disease, $p(x) = \eta + fx$, which implies an equilibrium frequency of $x_\infty = \frac{\eta}{1-f}$ in the corresponding deterministic model. Finally, for a recessive disease, $p(x) = \eta + x - (1-f)x^2$, which implies an equilibrium frequency of $x_\infty = \sqrt{\frac{\eta}{1-f}}$. These formulas and the approximations made in deriving them are discussed in the references (Crow and Kimura 1970, Ewens 1979, Fan and Lange 1998, Lange 1997). Most population geneticists substitute $p(x) = x$ in the infinitesimal variance $\sigma^2(t, x)$. This action is justified for neutral and recessive inheritance, but less so for dominant inheritance where the allele frequency x is typically on the order of magnitude of the mutation rate η.

For the neutral and dominant models, the infinitesimal mean $\mu(x, t)$ is linear in x, and the infinitesimal variance $\sigma^2(t, x)$ is quadratic. As we shall see, these facts enable one to calculate the moments of the diffusion approximation X_t to the discrete process X_n exactly. For the recessive model, $\mu(t, x)$ is unfortunately

quadratic in x. This increase in the degree of $\mu(t,x)$ hinders exact calculation of moments. However, we will show how to compute moments numerically even in this harder case.

3. Moment equations for diffusion processes. We consider a general diffusion process generated by a stochastic differential equation. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space with a right-continuous increasing family $(\mathcal{F}_t)_{t\geq 0}$ of sub σ-fields of $\mathcal{F}$, each of which contains all P-null sets. If B_t is standard Brownian motion, X_0 is an $\mathcal{F}_0$-measurable random variable independent of B_t, and $\sigma(s,x)$ and $\mu(s,x)$ are sufficiently smooth functions, then the solution X_t to the stochastic differential equation

$$X_t = X_0 + \int_0^t \sigma(s, X_s) dB_s + \int_0^t \mu(s, X_s) ds, \tag{3.1}$$

exists and is an $\mathcal{F}_t$-adapted continuous process. The smoothness assumptions on $\sigma(s,x)$ and $\mu(s,x)$ are Lipschitz conditions that need not concern us here (Chung and Williams 1990). We will require that X_t be square integrable and, indeed, possess any higher order moments mentioned below. The stochastic integral involved in equation (3.1) can be either Itô's or Stratonovich's integral. Itô's integral can be transformed to Stratonovich's by a change of variables and vice versa. We prefer Itô's integral because it allows us to use the infinitesimal means and variances directly.

If we apply Itô's formula to equation (3.1) with the transformation function $f(x) = e^{iux}$ (Chung and Williams 1990), then we find that

$$\begin{aligned} e^{iuX_t} = e^{iuX_0} &+ iu \int_0^t e^{iuX_s} \sigma(s, X_s) dB_s + iu \int_0^t e^{iuX_s} \mu(s, X_s) ds \\ &- \frac{u^2}{2} \int_0^t e^{iuX_s} \sigma^2(s, X_s) ds. \end{aligned}$$

Taking expectations now yields the equation

$$\begin{aligned} \mathrm{E}\left(e^{iuX_t}\right) = \mathrm{E}\left(e^{iuX_0}\right) &+ iu \int_0^t \mathrm{E}\left[e^{iuX_s} \mu(s, X_s)\right] ds \\ &- \frac{u^2}{2} \int_0^t \mathrm{E}\left[e^{iuX_s} \sigma^2(s, X_s)\right] ds \end{aligned} \tag{3.2}$$

for the characteristic function of X_t. Repeatedly differentiating equation (3.2) with respect to u and evaluating the results at $u = 0$ produces the hierarchy of integral equations

$$\mathrm{E}(X_t) = \mathrm{E}(X_0) + \int_0^t \mathrm{E}\left[\mu(s, X_s)\right] ds \tag{3.3}$$

$$\begin{aligned} \mathrm{E}(X_t^n) &= \mathrm{E}(X_0^n) + n\int_0^t \mathrm{E}\left[X_s^{n-1}\mu(s,X_s)\right] ds \\ &\quad + \frac{n(n-1)}{2}\int_0^t \mathrm{E}\left[X_s^{n-2}\sigma^2(s,X_s)\right] ds, \quad n \geq 2. \end{aligned} \tag{3.4}$$

Finally differentiating equations (3.3) and (3.4) with respect to t gives the corresponding hierarchy of differential equations

$$\frac{d}{dt}\mathrm{E}(X_t) = \mathrm{E}[\mu(t,X_t)] \tag{3.5}$$

$$\begin{aligned} \frac{d}{dt}\mathrm{E}(X_t^n) &= n\mathrm{E}\left[X_t^{n-1}\mu(t,X_t)\right] \\ &\quad + \frac{n(n-1)}{2}\mathrm{E}\left[X_t^{n-2}\sigma^2(t,X_t)\right], \quad n \geq 2. \end{aligned} \tag{3.6}$$

In some cases, these differential equations are tractable analytically. When they are intractable analytically, they may be tractable numerically.

Equation (3.4) for the second moment of X_t amounts to

$$\mathrm{E}(X_t^2) - \mathrm{E}(X_0^2) = 2\int_0^t \mathrm{E}\left[X_s\mu(s,X_s)\right]ds + \int_0^t \mathrm{E}[\sigma^2(s,X_s)]ds. \tag{3.7}$$

To recast this as an equation for the variance (Fan *et al.* 1998), note that equation (3.3) entails $d\,\mathrm{E}(X_t) = \mathrm{E}[\mu(t,X_t)]dt$. Hence, the fundamental theorem of calculus implies

$$\begin{aligned} \mathrm{E}(X_t)^2 - \mathrm{E}(X_0)^2 &= 2\int_0^t \mathrm{E}(X_s)d\,\mathrm{E}(X_s) \\ &= 2\int_0^t \mathrm{E}(X_s)\,\mathrm{E}[\mu(s,X_s)]ds. \end{aligned}$$

Subtracting this identity from equation (3.7) gives the promised variance equation

$$\begin{aligned} &\mathrm{Var}(X_t) - \mathrm{Var}(X_0) \\ &= \int_0^t \mathrm{E}[\sigma^2(s,X_s)]ds + 2\int_0^t \mathrm{Cov}[X_s,\mu(s,X_s)]\,ds. \end{aligned} \tag{3.8}$$

If $\mu(t,x)$ is linear in x and $\sigma^2(t,x)$ is quadratic in x, then the differential equations (3.5) and (3.6) take the form

$$y'(t) = f(t) + g(t)y(t). \tag{3.9}$$

In many cases of interest, the solution

$$y(t) = e^{\int_0^t g(s)ds}\left(y(0) + \int_0^t f(s)e^{-\int_0^s g(u)du}ds\right). \tag{3.10}$$

can be explicitly calculated (Fan *et al.* 1998). This is certainly true for the neutral and dominant Wright-Fisher models. However, for the recessive Wright-Fisher model, the infinitesimal mean $\mu(t,x)$ is quadratic in x.

If $\mu(t,x)$ is quadratic in x, then the hierarchy of moment equations (3.3) and (3.4) is coupled in the sense that lower order moment functions depend on higher order moment functions. This fact prevents one from solving the equations recursively. However, numerical progress can be made if we view the hierarchy of equations as a single infinite-dimensional differential equation. For the sake of concreteness, suppose that

$$\begin{aligned} \mu(t,x) &= \mu_0(t) + \mu_1(t)x + \mu_2(t)x^2 \\ \sigma^2(t,x) &= \sigma_0(t) + \sigma_1(t)x + \sigma_2(t)x^2, \end{aligned} \tag{3.11}$$

and let $M(t)$ be the infinite-dimensional column vector whose nth entry is $m_n(t) = \mathrm{E}(X_t^n)$ for $0 \le n < \infty$. Then the hierarchy of equations (3.5) and (3.6) can be written as the single differential equation

$$\frac{d}{dt}M(t) = A(t)M(t), \tag{3.12}$$

where the infinite-dimensional matrix $A(t) = A_1(t) + A_2(t)$ is the sum of the two infinite-dimensional matrices

$$A_1(t) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \mu_0(t) & \mu_1(t) & \mu_2(t) & 0 & 0 & 0 & \cdots \\ 0 & 2\mu_0(t) & 2\mu_1(t) & 2\mu_2(t) & 0 & 0 & \cdots \\ 0 & 0 & 3\mu_0(t) & 3\mu_1(t) & 3\mu_2(t) & 0 & \cdots \\ 0 & 0 & 0 & 4\mu_0(t) & 4\mu_1(t) & 4\mu_2(t) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$A_2(s) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \sigma_0(t) & \sigma_1(t) & \sigma_2(t) & 0 & 0 & \cdots \\ 0 & \frac{3\cdot 2}{2}\sigma_0(t) & \frac{3\cdot 2}{2}\sigma_1(t) & \frac{3\cdot 2}{2}\sigma_2(t) & 0 & \cdots \\ 0 & 0 & \frac{4\cdot 3}{2}\sigma_0(t) & \frac{4\cdot 3}{2}\sigma_1(t) & \frac{4\cdot 3}{2}\sigma_2(t) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

If $\mu(t,x)$ and $\sigma^2(t,x)$ are general polynomials in x rather than quadratics, then the same differential equation (3.12) holds provided we modify $A(t)$ in the obvious manner. For the sake of simplicity, we confine our attention to the quadratic case.

4. Neutral model moments. In the neutral genetic model with exponential population growth, the infinitesimal mean and variance are

$$\mu(t,x) = 0 \qquad \sigma^2(t,x) = \frac{x(1-x)}{2N_0e^{ct}}.$$

Equation (3.3) makes it evident that $m_1(t) = \mathrm{E}(X_t) = \mathrm{E}(X_0)$ for all $t \geq 0$. In view of equation (3.6), the nth moment $m_n(t)$ of X_t satisfies

$$\frac{d}{dt}m_n(t) = -\frac{n(n-1)[m_n(t) - m_{n-1}(t)]}{4N_0e^{ct}}. \tag{4.1}$$

To solve equation (4.1), we make the change of variables $s = e^{-ct} - 1$ and $\nu_n(s) = m_n(t)$. Because $ds = -ce^{-ct}dt$, this substitution yields Kimura's (1955) equation

$$\frac{d}{ds}\nu_n(s) = \frac{n(n-1)[\nu_n(s) - \nu_{n-1}(s)]}{4N_0c} \tag{4.2}$$

for a stationary neutral process.

Now consider the trial solution

$$\nu_n(s) = \sum_{i=0}^{n-1} c_{ni}e^{\lambda_i s}. \tag{4.3}$$

The fact that this sum has upper limit $n-1$ will determine the eigenvalues λ_i. If we differentiate the trial solution (4.3), compare the result to equation (4.2), and equate coefficients of $e^{\lambda_i s}$, then we find that

$$\lambda_i c_{ni} = \frac{n(n-1)}{4N_0c}(c_{ni} - c_{n-1,i}).$$

This determines c_{ni} as

$$c_{ni} = \frac{n(n-1)}{n(n-1) - 4N_0c\lambda_i}c_{n-1,i} \tag{4.4}$$

unless $n(n-1) = 4N_0c\lambda_i$. We want this exceptional condition to occur when $i = n-1$ because then the requirement $c_{n-1,n-1} = 0$ imposes no constraint on $c_{n,n-1}$. Thus, we take

$$\lambda_{n-1} = \frac{n(n-1)}{4N_0c}.$$

The coefficients c_{ni} can now be found by invoking the initial conditions. Clearly, $c_{10} = m_1(0)$. Suppose we know the coefficients $c_{n-1,0}, \ldots, c_{n-1,n-2}$ determining $\nu_{n-1}(s)$. Then the coefficients $c_{n0}, \ldots, c_{n,n-2}$ can be computed via the recurrence relation (4.4). The final coefficient $c_{n,n-1}$ is determined by the initial condition $m_n(0) = \sum_{i=0}^{n-1} c_{ni}$. These considerations allow us to calculate, for example,

$$\begin{aligned}
\nu_2(s) &= m_1(0) + [m_2(0) - m_1(0)]e^{\frac{s}{2N_0c}} \\
\nu_3(s) &= m_1(0) + \frac{3}{2}[m_2(0) - m_1(0)]e^{\frac{s}{2N_0c}} \\
&\quad + \frac{1}{2}[2m_3(0) - 3m_2(0) + m_1(0)]e^{\frac{3s}{2N_0c}} \\
\nu_4(s) &= m_1(0) + \frac{9}{5}[m_2(0) - m_1(0)]e^{\frac{s}{2N_0c}} \\
&\quad + [2m_3(0) - 3m_2(0) + m_1(0)]e^{\frac{3s}{2N_0c}} \\
&\quad + \frac{1}{5}[5m_4(0) - 10m_3(0) + 6m_2(0) - m_1(0)]e^{\frac{3s}{N_0c}}.
\end{aligned}$$

Kimura (1955) gives explicit expressions for the coefficients c_{ni} when $X_0 = p$ is constant and $m_n(0) = p^n$ for all n.

Similar reasoning enables one to find not only the moments but also the density function $f(t,x)$ of the neutral Wright-Fisher process X_t. It is well known that $f(t,x)$ satisfies the Fokker-Planck or Kolmogorov forward equation (Feller 1951)

$$\frac{\partial}{\partial t} f(t,x) = \frac{1}{4N_0 e^{ct}} \frac{\partial^2}{\partial x^2}\big[x(1-x)f(t,x)\big].$$

The change of variables $s = 1 - e^{-ct}$ and $g(s,x) = f(t,x)$ transforms this partial differential equation into the corresponding partial differential equation

$$\frac{\partial}{\partial s} g(s,x) = \frac{1}{4N_0 c} \frac{\partial^2}{\partial x^2}\big[x(1-x)g(s,x)\big].$$

for neutral evolution in a stationary population. Crow and Kimura (1970) explain how $g(s,x)$ and therefore $f(t,x)$ can be expanded in terms of appropriate eigenfunctions.

5. Dominant model moments. For a dominant disease, the infinitesimal mean $\mu(t,x) = \eta - (1-f)x$. Hence, the differential equation (3.5) for the first moment $m_1(t) = E(X_t)$ becomes

$$\frac{d}{dt} m_1(t) = \eta - (1-f)m_1(t)$$

with solution

$$m_1(t) = \left[m_1(0) - \frac{\eta}{1-f}\right] e^{-(1-f)t} + \frac{\eta}{1-f}.$$

Because the infinitesimal variance is $\sigma^2(t,x) = \frac{(\eta+fx)(1-\eta-fx)}{2N_0e^{ct}}$ under exponential growth, the differential equation (3.6) for the second moment $m_2(t)$ amounts to

$$\begin{aligned}\frac{d}{dt}m_2(t) &= 2\eta m_1(t) - 2(1-f)m_2(t) \\ &\quad + \frac{1}{2N_0e^{ct}}\left[\eta(1-\eta) + f(1-2\eta)m_1(t) - f^2m_2(t)\right].\end{aligned}$$

Obviously, this differential equation is the special case of equation (3.9) with $y(t) = m_2(t)$ and

$$\begin{aligned}f(t) &= 2\eta m_1(t) + \frac{1}{2N_0e^{ct}}\left[\eta(1-\eta) + f(1-2\eta)m_1(t)\right] \\ g(t) &= -2(1-f) - \frac{f^2}{2N_0e^{ct}}.\end{aligned}$$

The solution (3.10) to the differential equation (3.9) includes the integral

$$\int_0^t g(s)ds = -2(1-f)t - \frac{f^2}{2cN_0}\left(1 - e^{-ct}\right)$$

and the integral $\int_0^t f(s)e^{-\int_0^s g(u)du}ds$. The latter integral decomposes as a linear combination of terms of the kind

$$\int_0^t e^{-\alpha s - \beta e^{-cs}}ds = \frac{1}{c}\int_0^{ct} e^{-\frac{\alpha}{c}u - \beta e^{-u}}du$$

for $\beta = \frac{f^2}{2cN_0}$ and various choices of the constant α. One can evaluate each such term via the special function

$$\begin{aligned}I_{\alpha,\beta}(t) &= \int_0^t e^{-\alpha s - \beta e^{-s}}ds \\ &= \sum_{k=0}^{\infty}\int_0^t e^{-\alpha s}\frac{(-\beta e^{-s})^k}{k!}ds \\ &= \sum_{k=0}^{\infty}\frac{(-\beta)^k}{k!(\alpha+k)}\left[1 - e^{-(\alpha+k)t}\right].\end{aligned}$$

Because β is small in practice, the above series converges quickly. Somewhat tedious algebra shows that

$$\begin{aligned}
&\int_0^t f(s)e^{-\int_0^s g(u)du}ds \\
&= \frac{e^\beta}{c}\left\{2\eta\left[m_1(0) - \frac{\eta}{1-f}\right] I_{\frac{f-1}{c},\beta}(ct) + \frac{2\eta^2}{1-f} I_{\frac{2(f-1)}{c},\beta}(ct)\right. \\
&\quad + \frac{f(1-2\eta)}{2N_0}\left[m_1(0) - \frac{\eta}{1-f}\right] I_{\frac{f-1}{c}+1,\beta}(ct) \\
&\quad \left. + \frac{\eta(1-\eta-f\eta)}{2N_0(1-f)} I_{\frac{2(f-1)}{c}+1,\beta}(ct)\right\}.
\end{aligned}$$

6. Numerical methods. Over a short time interval dt, the differential equation (3.12) entails the Euler approximation

$$M(t+dt) \approx [I + dtA(t)]M(t), \tag{6.1}$$

where I is the identity matrix. If we partition the interval $[0,t]$ into n subintervals $[i\delta_n, (i+1)\delta_n]$ for $i = 0, \ldots, n-1$ and $\delta_n = \frac{t}{n}$, then the approximation (6.1) propagates into Euler's method

$$M(t) \approx \prod_{i=0}^{n-1} [I + \delta_n A(i\delta_n)]M(0) \tag{6.2}$$

of solving for $M(t)$. With luck, the expression on the left of (6.2) will tend to $M(t)$ as n tends to ∞. Mathematical justification of limits of this type belongs to the province of product integration (Dollard and Friedman 1979, Gill and Johansen 1990). If $A(t) = A$ does not depend on the time parameter t, then the product integral

$$M(t) \;=\; \prod_{s=0}^{t} e^{A(s)ds} M(0) \;=\; e^{tA} M(0)$$

coincides with multiplication by a matrix exponential.

Making the theory of product integration rigorous in the current context is difficult because $M(t)$ and $A(t)$ are infinite dimensional and $A(t)$ is unbounded. For practical purposes, we truncate $M(t)$ and $A(t)$ to their first k rows and columns and carry out all computations with the resulting finite-dimensional versions $M_k(t)$ and $A_k(t)$ of $M(t)$ and $A(t)$. The sparsity of the matrices $M_k(t)$ in our genetic examples obviously decreases both the computational complexity

and the storage requirements of the matrix-vector multiplications implied by formula (6.2).

In the case of neutral and dominant genes, where we have analytic results, the truncated system is easy to solve numerically. Unfortunately, for recessive genes with high initial frequencies, the truncated system poses more of a numerical challenge. In addition to Euler's method, we have tried a standard fourth-order Runge-Kutta scheme (Birkhoff and Rota 1978, Press *et al.* 1992) and the power series method sketched below. To achieve stable solutions, all three methods require short steps ($n \geq 20000$) and many moments ($k \geq 500$) for initial gene frequencies in excess of .005. We enhance the stability of each method by instituting three safeguards. First, we perform all computations in double precision. Second, at the end of each step, we reset all negative entries $m_j(t)$ of $M_k(t)$ to 0. In our models all moments are nonnegative, so presumably this tactic helps. Third, at the end of each step, we also exploit Hölder's moment inequality $m_j(t)^{1/j} \leq m_{j+1}(t)^{1/(j+1)}$ by replacing $m_{j+1}(t)$ by $\max\{m_{j+1}(t), m_j(t)^{(j+1)/j}\}$ recursively for $j = 1, 2, \cdots, k-1$.

To explain the series method, note that for a recessive disease in the presence of exponential population growth, the quadratic expressions (3.11) for the infinitesimal mean and variance have coefficients

$$\begin{bmatrix} \mu_0(t) \\ \mu_1(t) \\ \mu_2(t) \end{bmatrix} = \begin{bmatrix} \eta \\ 0 \\ -(1-f) \end{bmatrix}, \qquad \begin{bmatrix} \sigma_0(t) \\ \sigma_1(t) \\ \sigma_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{e^{-ct}}{2N_0} \\ -\frac{e^{-ct}}{2N_0} \end{bmatrix}.$$

These expressions imply that the matrix $A(t)$ can be expanded in the power series

$$A(t) = B_1 + B_2 e^{-ct} = \sum_{i=0}^{\infty} C_i t^i,$$

where $C_0 = B_1 + B_2$ and $C_i = \frac{(-c)^i}{i!} B_2$, $i \geq 1$, are constant matrices. If we formally expand the moment vector $M(t) = \sum_{i=0}^{\infty} D_i t^i$ in a similar power series and differentiate term by term, then equating coefficients of t^j in the differential equation (3.12) yields the recurrence

$$D_{j+1} = \frac{1}{j+1} \sum_{i=0}^{j} C_i D_{j-i}.$$

Together with the initial condition $D_0 = M(0)$, this gives an effective method of computing the series expansion of $M(t)$ (Apostol 1969). For t reasonably small, we can terminate the expansion after a few terms, say five, and still retain a

good approximation to $M(t)$. For larger t, we choose n large and approximate $M(t/n)$, use this as an initial value to approximate $M(2t/n)$, and so forth, until we finally recover $M(t)$. Once again we must operate on truncated vectors and matrices.

The fact that all three solution methods ultimately provide similar answers increases our overall confidence that we can compute the moment vector $M(t)$ accurately. However, we are still far from fully understanding the analytic and numerical behavior of the moment differential equations. More research is clearly needed.

7. Examples. To illustrate the theory, we now turn to some concrete examples based on the demographic history of Finland [Hästbacka *et al.* (1992)]. Finland was settled around 2000 years ago by a founding population of about 1000 people. Given a current Finnish population of 5 million people and a generation time of 25 years, this implies an exponential growth rate of $c = \frac{1}{80}\ln(5000) = .1065$ per generation. In Figures 1 through 5, we consider the evolution of the b allele in three simple biallelic genetic models. For the sake of completeness, our graphs extrapolate 20 generations into the future.

Figure 1 plots the coefficient of variation $\kappa_2(t)^{\frac{1}{2}}/\kappa_1(t)$, skewness $\kappa_3(t)/\kappa_2(t)^{\frac{3}{2}}$, and kurtosis $\kappa_4(t)/\kappa_2(t)^2$ of the frequency X_t of the b allele under the neutral Wright-Fisher model. Here $\kappa_j(t)$ is the jth cumulant of X_t at time t. For a normally distributed random variable, skewness and kurtosis are both 0. Because neither selection nor mutation operate in this model, the mean of X_t and its deterministic analog remain fixed at our chosen initial value $X_0 = .1$. The figure makes it evident that the variance first increases sharply and later flattens out. This behavior confirms our intuition that most of the stochastic effects take place in the early generations when the population size is small. The nontrivial skewness and kurtosis that develop during this period are eventually frozen into place by the exponential growth of the population.

Figure 2 plots the mean of X_t and its deterministic analog for a dominant disease allele with a mutation rate $\eta = 10^{-6}$, a fitness $f = .9$, and an initial gene frequency $X_0 = 5 \times 10^{-4}$. The upper and lower bands in this Figure are the curves $\max\{0, \kappa_1(t) \pm 2\kappa_2(t)^{\frac{1}{2}}\}$. The initial gene frequency corresponds to one affected person among the 1000 founders. With such a high fitness, the effects of the affected founder persist for many generations. Eventually, however, the balance between selection and mutation asserts itself, and the mean approaches its low equilibrium level. Figure 3 shows that in the process a quasi-stochastic equilibrium develops with decreasing variance and low skewness and kurtosis. The large skewness and kurtosis seen in early generations presumably reflect the nonnegligible probability that the affected population founder generates a

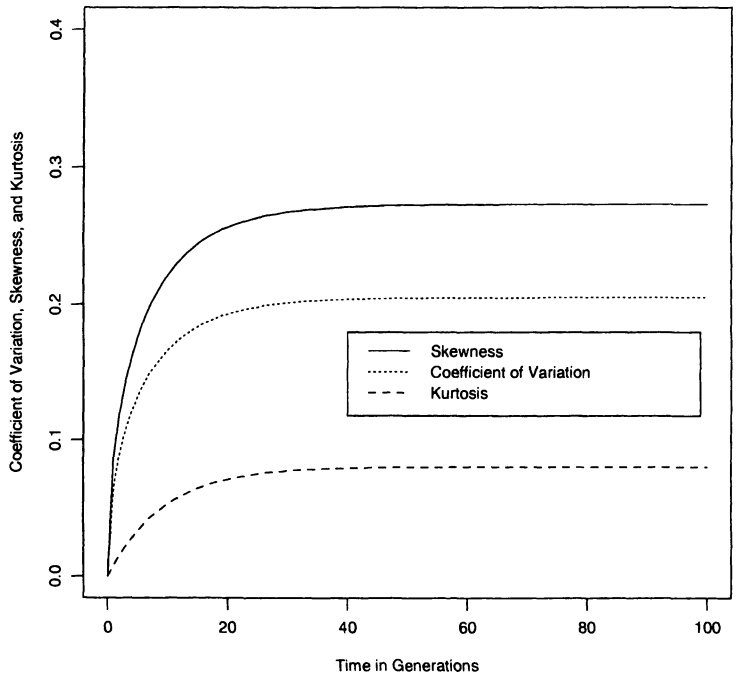

FIG. 1. *Coefficient of Variation, Skewness, and Kurtosis of the Frequency of a Neutral Gene when* $X_0 = .1$.

very large clan of affected descendants.

Finally, Figures 4 and 5 depict the dynamics of a recessive disease with a mutation rate $\eta = 10^{-6}$, a fitness $f = .5$, and an initial gene frequency $X_0 = .005$. Figure 4 displays good agreement between the mean of X_t and its deterministic analog. Under the pressure of selection, both are slowly tending to the deterministic equilibrium. Just as with a neutral gene, there are large stochastic effects in early generations that persist for the duration of Figures 4 and 5. The interesting behavior of the kurtosis curve in Figure 5 is difficult to rationalize. Possibly it is a numerical artifact, but our calculations appear stable when the step size is small enough and sufficiently many moments are taken into account.

8. Discussion. The diffusion process models familiar to population geneticists almost invariably assume a stationary population (Crow and Kimura 1970, Ewens 1979, Feller 1951). However, human populations tend to display exponential growth with episodes of decline brought on by famine, plague, and war.

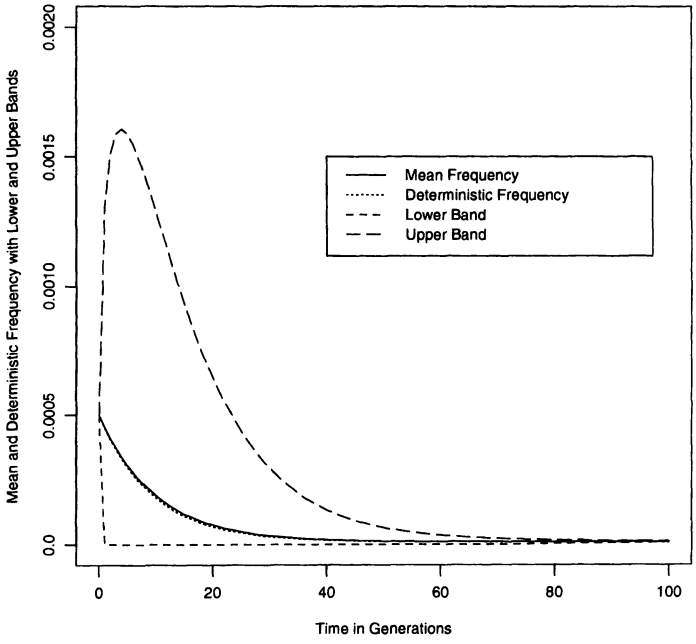

FIG. 2. *Mean and Deterministic Value for the Frequency of a Dominant Gene with* $\eta = 10^{-6}$, $f = .90$, *and* $X_0 = 5 \times 10^{-4}$.

In the current paper, we extend to exponentially growing populations some of the moment calculations for gene frequencies previously carried out under the stationary Wright-Fisher model. We also fill in some mathematical gaps in the treatment of recessive diseases.

We view diffusion process models as complementary to branching process models for neutral and dominant disease genes. As emphasized in our introduction, branching processes are poor vehicles for modeling recessive diseases. The Wright-Fisher model overcomes this defect, but at the cost of introducing diffusion approximations and a specific sampling framework for generating gametes. The binomial sampling premise of the Wright-Fisher effectively requires that each parent produces a Poisson number of children. This offspring assumption probably underestimates the variance in the number of children per parent.

The example featured in Figures 2 and 3 clearly illustrates how quickly the deterministic balance between selection and mutation is reached for a dominant disease. The frequency of the disease allele rapidly tends to its deterministic equilibrium with little stochastic variation left in later generations. A low mu-

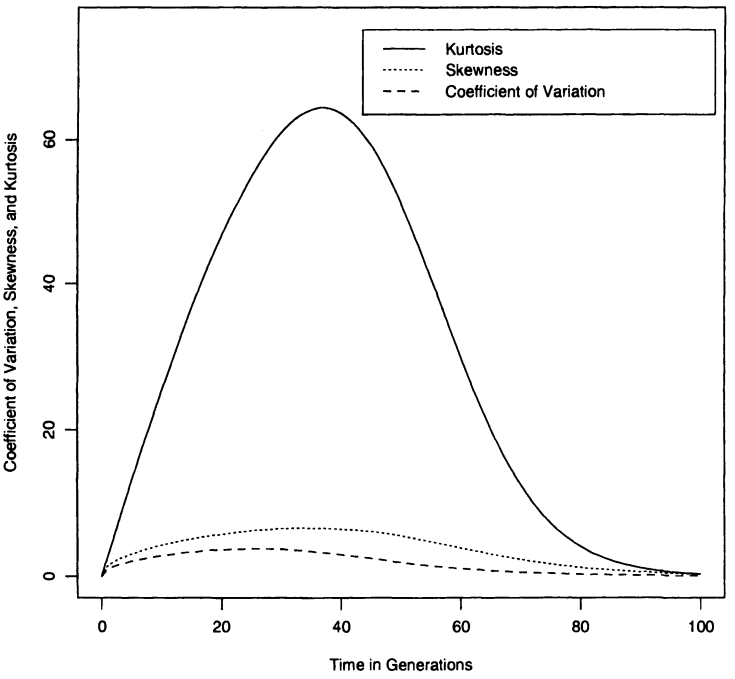

FIG. 3. *Coefficient of Variation, Skewness, and Kurtosis of the Frequency of a Dominant Gene when* $\eta = 10^{-6}$, $f = .90$, *and* $X_0 = 5 \times 10^{-4}$.

tation rate and a high fitness retard the approach of a dominant gene to its deterministic equilibrium. It is noteworthy that cumulants (means, variances, skewness, and kurtosis) calculated by diffusion methods and by branching process methods are comparable for dominant diseases. Indeed, the cumulant results from the two branching process models in Lange and Fan (1997) appear to bracket the cumulants results from the diffusion process model. Obviously, this check is impossible for a recessive disease.

Neutral and recessive genes operate on an entirely different time scale than dominant genes. Stochastic fluctuations are considerable in a small population, and the large variance that develops in early generations persists for many generations to come. This suggests that predictions from the standard deterministic models be treated with extreme caution. Although population geneticists are well aware of this fact, it constantly needs to be reiterated for geneticists lacking relevant training.

Our methods for calculating the moments of a diffusion process should be

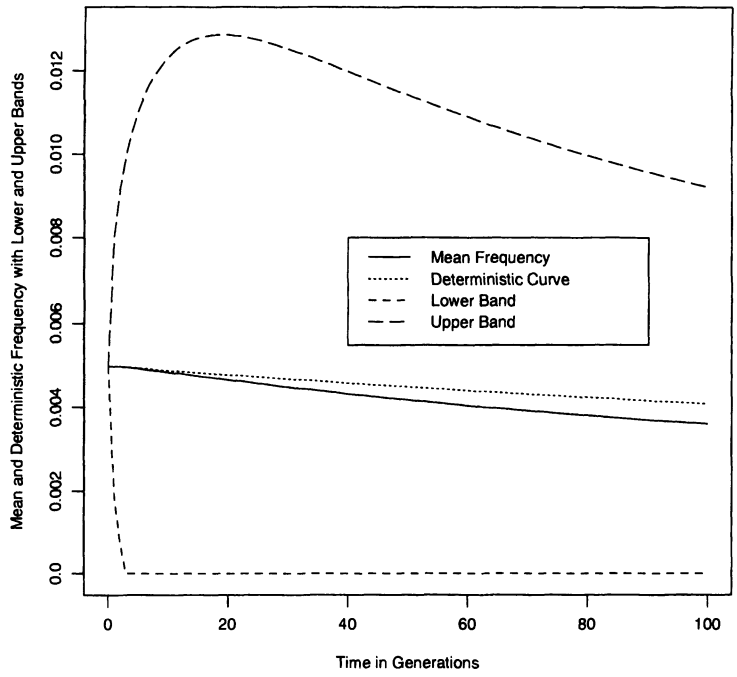

FIG. 4. *Mean and Deterministic Value for the Frequency of a Recessive Gene with* $\eta = 10^{-6}$, $f = .5$, *and* $X_0 = .005$.

of general interest. The versatility of the methods in the face of time inhomogeneities and polynomial dependence of the infinitesimal means and variances is a major advantage. Putting our numerical methods on a firmer theoretical foundation is clearly the next order of business. The techniques of functional analysis such as the Hille-Yosida theorem for continuous semigroups of operators offer one line of attack (Yosida 1980).

Extensions of the moment calculations to multivariate diffusion processes are also worth pursuing. In previous papers (Lange and Fan 1997, Fan and Lange 1998), we have set down branching process models that illuminate some of the issues in haplotype mapping of disease genes. To extend these calculations to recessive diseases, we must contend with multivariate versions of the Wright-Fisher model. The necessary particle types are intact and recombined chromosomes that carry ancestral disease mutations.

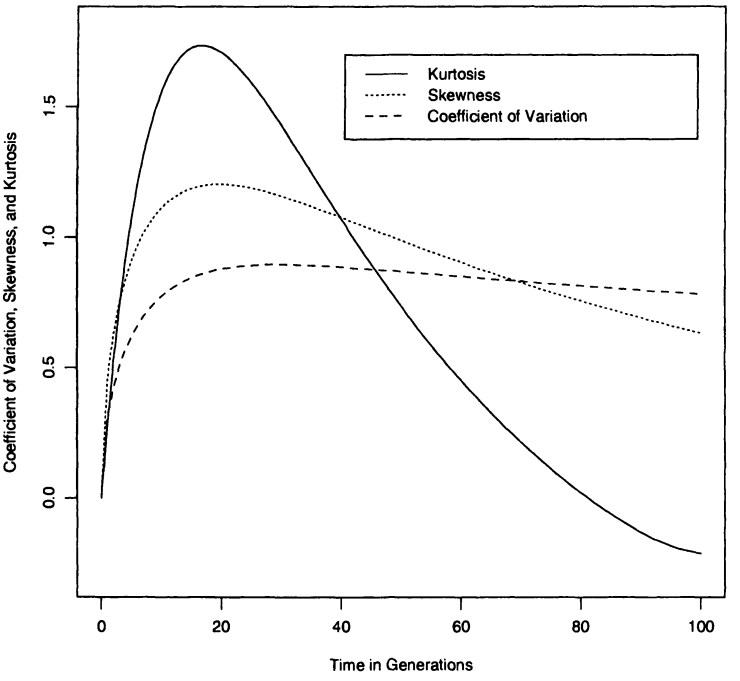

FIG. 5. *Coefficient of Variation, Skewness, and Kurtosis of the Frequency of a Recessive Gene when* $\eta = 10^{-6}$, $f = .5$, *and* $X_0 = 0.005$.

REFERENCES

APOSTOL, T.M. (1969). *Calculus, Vol. II.* 2nd ed., pp 217–221. Wiley, New York.

BIRKHOFF, G. and ROTA, G-C. (1978). *Ordinary Differential Equations.* 3rd ed. Wiley, New York.

CAVALLI-SFORZA, L.L. and BODMER, W.F. (1971). *The Genetics of Human Populations.* Freeman, San Francisco.

CHUNG, K.L. and WILLIAMS, R.J. (1990). *Introduction to Stochastic Integration.* 2nd ed., p 94. Birkhäuser, Boston.

CROW, J.F. and KIMURA, M. (1970). *An Introduction to Population Genetics Theory.* pp 367–432. Harper & Row, New York.

DOLLARD, J.D. and FRIEDMAN, C.N. (1979). *Product Integration with Application to Differential Equations.* Addison-Wesley, Reading, MA.

EWENS, W.J. (1979). *Mathematical Population Genetics.* pp 138–175. Springer-Verlag, New York.

FAN, R.Z. and LANGE, K. (1998). Models for haplotype evolution in a nonstationary population. *Theor. Pop. Biol.* (in press).

FAN, R.Z., LANGE, K. and PENA, E. (1998). A note on variation in point processes and damage models (submitted).

FELLER, W. (1951). Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* pp 227–246. University of California Press,

Berkeley, CA.

FISHER, R.A. (1930). The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh* **50** 205–220.

GILL, R.D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Annals Stat.* **18** 1501–1555.

GLADSTIEN, K. and LANGE, K. (1978a). Number of people and number of generations affected by a single deleterious mutation. *Theor. Pop. Biol.* **14** 313–321.

GLADSTIEN, K. and LANGE, K. (1978b). Equilibrium distributions for deleterious genes in large stationary populations. *Theor. Pop. Biol.* **14** 322–328.

HALDANE, J.B.S. (1927). A mathematical theory of natural and artificial selection, Part V: Selection and mutation. *Proc. Camb. Phil. Soc.* **23** 838–844.

HARRIS, T.E. (1989). *The Theory of Branching Processes.* Dover, New York, p 99.

HÄSTBACKA, J., de la CHAPELLE, A., KAITILA, I., SISTONEN, P., WEAVER, A. and LANDER, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2** 204–211.

KIMURA, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* **41** 144–50.

LANGE, K. (1982). Calculation of the equilibrium distribution for a deleterious gene by the finite Fourier transform. *Biometrics* **38** 79–86.

LANGE, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis.* Springer-Verlag, New York.

LANGE, K. and FAN, R.Z. (1997). Branching process models for mutant genes in nonstationary populations. *Theor. Pop. Biol.* **51** 118–133.

LANGE, K. and GLADSTIEN, K. (1980) Further characterization of the long-run population distribution of a deleterious gene. *Theor. Pop. Biol.* **18** 31–43.

NAGYLAKI, T. (1992). *Introduction to Theoretical Population Genetics.* Springer-Verlag, Berlin.

PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. and FLANNERY, B.P. (1992). *Numerical Recipes in Fortran: The Art of Scientific Computing.* 2nd ed. Cambridge University Press, Cambridge.

SKELLAM, J.G. (1949). The probability distribution of gene-differences in relation to selection, mutation, and random extinction. *Proc. Camb. Phil. Soc.* **45** 364-367.

YOSIDA, K. (1980). *Functional Analysis.* 6th ed. Springer-Verlag, New York.

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
THE UNIVERSITY OF MICHIGAN
ANN ARBOR, MI 48109-2029
RFAN@UMICH.EDU

DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF MICHIGAN
ANN ARBOR, MI 48109-2029
KLANGE@UMICH.EDU