Genet. Res., Camb. (1987), **501**, *pp.* 155–163 *With 2 text-figures Printed in Great Britain*

Estimating the proportion of neutral mutants

G. A. WATTERSON
Mathematics Department, Monash University, Clayton, Victoria, 3168, Australia

(Received 10 July 1986 and in revised form 17 February 1987)

Summary

Kimura used the heterozygosity and the number of low-frequency alleles to estimate that about 14% of mutations are selectively neutral. The method is shown to be subject to biases and to disruption due to bottleneck effects. Let deleterious alleles have selective disadvantage, s, compared with neutral alleles and let N_e denote the effective diploid population size. The estimator, $\hat{P}$, of the proportion of neutral alleles is positively biased if (roughly) $4N_e s < 25$ or if $4N_e s > 200$. In the former case, one cannot adequately detect the different influences of deleterious and neutral alleles, whereas in the latter case, deleterious alleles will rarely appear in the sample. These difficulties cause the biases in $\hat{P}$, and are likely to cause similar biases for any estimation method based solely on allele frequencies. There is substantial sampling variability in $\hat{P}$ in cases of practical interest, when data from 11 loci, or even as many as 31 loci, are pooled. If there has been a recent contraction in population size, $\hat{P}$ will be positively biased, often yielding values greater than 1 or even being infinite. But after a recent expansion in population size, the heterozygosity will not have made as quick an increase and $\hat{P}$ will be negatively biased. Population expansion alone can produce $\hat{P}$ values close to those observed by Kimura, even if all alleles are neutral. In an appendix, a new method for simulating samples of neutral and deleterious genes is described.

1. The estimator

Kimura (1983) has estimated the proportion of mutations which are selectively neutral to be about 0·14, with a standard error of 0·06. This estimate is an average taken over various species, namely, plaice, humans, monkeys and fruit flies. The method used to obtain the estimate is the following.

Let $\Phi(x)\,dx$ denote the expected number of alleles (averaged over many independent loci) whose relative frequencies are in the range x to $x+dx$, in a given population. Then the expected heterozygosity, μ_H, in the population is given approximately by

$$\mu_H \approx 1 - \int_0^1 x^2 \Phi(x)\,dx. \tag{1.1}$$

Consider a sample of $2n$ genes. The number of alleles, whose relative frequencies in the sample are below q, is expected to be μ_q, given approximately by

$$\mu_q \approx \int_{1/2n}^{q} \Phi(x)\,dx. \tag{1.2}$$

Here it is assumed that $1/2n \ll q$, but also that q is small, say $q = 0{\cdot}01$.

Suppose now that the effective population size of the species being sampled is N_e, that the total mutation rate to all types of alleles is ν_T, but that included within that rate, the mutation rate to selectively neutral alleles is ν. We define $\theta_T = 4N_e\nu_T$, and $\theta = 4N_e\nu$. Then under certain circumstances, (1.1) and (1.2) may be further approximated by

$$\mu_H \approx \theta/(\theta+1) \tag{1.3}$$

and

$$\mu_q \approx \theta_T \log(2nq). \tag{1.4}$$

The proportion of neutral mutations, among all mutations, is

$$P_{\text{neut}} = \nu/\nu_T = \theta/\theta_T. \tag{1.5}$$

Provided that (1.3) and (1.4) are adequate approximations, P_{neut} would be given approximately by

$$P_{\text{neut}} \approx \frac{\mu_H}{1-\mu_H}\,\frac{\log(2nq)}{\mu_q}. \tag{1.6}$$

Kimura suggested that μ_H be estimated by $\overline{H}$, the average of the heterozygosities in samples of genes taken from a large number, l say, of loci. Similarly, μ_q is estimated by $\overline{n}_a(x < q)$, the average of the observed numbers of alleles whose relative frequencies are less than q in the samples. Thus finally P_{neut} is estimated by $\hat{P}$ say, given by

$$\hat{P} = \frac{\overline{H}}{1-\overline{H}}\,\frac{\log(2nq)}{\overline{n}_a(x<q)}. \tag{1.7}$$

The accuracy of $\hat{P}$ as an estimate of P_{neut} depends on (i) the accuracy of the continuous integration formulas (1.1) and (1.2) for what, strictly speaking, should be discrete summations (ii) the accuracy of the approximations (1.3) and (1.4) and (iii) the accuracy of the sample averages $\overline{H}$ and $\overline{n}_a(x < q)$ as estimates of the corresponding expectations μ_H and μ_q.

In the next section we shall discuss these problems in the very special case when *all* mutations are neutral so that $P_{\text{neut}} = 1$. In the following section, we shall concentrate on the case when there are two classes of alleles, those which are selectively neutral and those which have a selective disadvantage, s. In the final section, we shall study bottleneck effects. An appendix includes a description of a new method to simulate samples having neutral and deleterious genes.

The conclusion of these studies is that the estimate (1.7) is not fully satisfactory. It is subject to large sampling variations, to inaccuracies in the various approximations leading to (1.6) and to non-stationary effects. The problem is obviously not an easy one to tackle, and it may well be impossible to produce a satisfactory method of estimation which is robust against the various complications which arise in population genetics.

2. The stationary, neutral case

If (i) all mutations produce new alleles, (ii) all mutations are neutral, so that $\theta = \theta_T$ and $P_{\text{neut}} = 1$, and (iii) the population is at statistical stationarity, then (Kimura & Crow (1964); Ewens (1964))

$$\Phi(x) = \theta(1-x)^{\theta-1}/x \quad (0 < x \leqslant 1) \tag{2.1}$$

is the (continuous) frequency spectrum. It is not exact; the discrete frequency spectrum for the expected number of alleles which are of relative frequency $x = j/2n$, in a sample of size $2n$ genes, is

$$E(\beta(j)) = \theta(2n)_{[j]}/[j(\theta+2n-1)_{[j]}] \quad (j = 1, 2, \ldots, 2n), \tag{2.2}$$

where $\beta(j)$ is the number of alleles having j representative genes in the sample, and where $(2n)_{[j]} = 2n(2n-1)\ldots(2n-j+1)$; see Watterson (1974*a*). This expression is exact for sampling without replacement from a Moran model population, and it is a close approximation for sampling from a Wright–Fisher model. However, for $2n$ large, there is an almost perfect agreement between (2.1) and (2.2):

$$\Phi(j/2n)/2n \approx E(\beta(j)).$$

The true expected heterozygosity in a sample is

$$1 - \sum_{j=1}^{2n} (j/2n)^2 E(\beta(j)) = (1-1/2n)\,\theta/(\theta+1),$$

whereas (1.1) yields the very similar

$$\mu_H \approx \theta/(\theta+1), \tag{2.3}$$

and for samples of sizes investigated by Kimura (1983), these two results are essentially identical. Nevertheless, the true expected number of alleles of frequency less than q in a sample of $2n$ genes is

$$\mu_q = \sum_{j=1}^{(2nq)^*} E(\beta(j)), \tag{2.4}$$

where $(2nq)^*$ is the largest integer *less* than $2nq$. But (1.2) yields the approximation

$$\mu_q \approx \int_{1/2n}^{q} \Phi(x)\,dx \approx \theta \int_{1/2n}^{q} x^{-1}\,dx = \theta \log(2nq), \tag{2.5}$$

which can be substantially in error, due mostly to the lack of continuity correction at the terminals of the integral. Almost perfect accuracy can be achieved by the use of the summation

$$\mu_q \approx \sum_{j=1}^{(2nq)^*} \Phi(j/2n)/2n.$$

The following approximations yield quite satisfactory accuracies, in terms of the digamma function ψ and Euler's constant γ:

$$\mu_q \approx \theta \sum_{j=1}^{(2nq)^*} 1/j = \theta[\psi((2nq)^*+1)+\gamma]$$
$$\approx \theta\{\log[(2nq)^*+1] + 0{\cdot}57722 - 1/[2(2nq)^*+2]\}. \tag{2.6}$$

As a numerical example, consider the first set of data discussed by Kimura (1983), in which the average sample size was $2n = 3912$ and q was taken as $q = 0{\cdot}01$. Then $(2nq)^* = 39$, so that (2.6) yields $\mu_q \approx 4.27\theta$ whereas (2.5) yields $\mu_q \approx 3.67\theta$. Thus (2.5) can be in error by about 14%, and similarly, the estimate $\hat{P}$ in (1.7) can be 14% too low for this reason. For smaller samples, the error would be worse. For instance, the final example discussed by Kimura had an average size of $2n = 568{\cdot}06$, with $q = 0{\cdot}01$ so that $(2nq)^* = 5$. The proportionate error in $\hat{P}$, because of the use of (2.5), is that it is 24% too low. In effect, $\hat{P}$ would be estimating 0·76 rather than the correct value $P_{\text{neut}} = 1$.

Kimura's reason for defining $\hat{P}$ as in (1.7), is that it is mainly the neutral alleles which influence the heterozygosity, whereas deleterious alleles, along with neutral alleles, could be among those with frequencies below q. See also Nei (1977). Thus in cases when there is a distinction between the neutral mutation parameter θ, and the total mutation parameter, θ_T, it is the former which should appear in (2.3), see (1.3), and the latter which should appear in (2.5), see (1.4). If that were so, then the definition (1.7) would be an appropriate way to estimate P_{neut}, because then, $\overline{H}/(1-\overline{H})$ would be an estimate of θ, and $\overline{n}_a(x < q)/\log(2nq)$ would be an estimate of θ_T. In view of the above, a better estimate would be

$$\hat{P} = \frac{\overline{H}}{1-\overline{H}} \frac{\log[(2nq)^*+1] + 0{\cdot}57722 - 1/[2(2nq)^*+2]}{\overline{n}_a(x < q)}, \tag{2.7}$$

so that Kimura's first example would yield an estimate $\hat{P} = 0{\cdot}18$ rather than Kimura's own estimate of 0·15.

https://doi.org/10.1017/S0016672300023569 Published online by Cambridge University Press

And his last example would also yield $\hat{P} = 0{\cdot}18$ rather than Kimura's 0·13.

We now turn to consider the influence of the number of loci on the accuracy of the estimates. An indication that this is important can be gained from Kimura's first example. Kimura used the mean heterozygosity, $\overline{H} = 0{\cdot}106$, from 46 loci, whereas for the calculation of $\overline{n}_a(x < q)$, he used data from 11 loci. When the mean heterozygosity was calculated by Kimura from those same 11 loci, it became 0·147. Had the latter value been used, Kimura's estimate $\hat{P} = 0{\cdot}15$ would have become $\hat{P} = 0{\cdot}23$. It might be expected that sampling variation would be even more important in the calculation of $\overline{n}_a(x < q)$.

There are two aspects to sampling variation. There is the variation due to the choice of genes from the population, which is not a serious matter if large random samples are chosen. Perhaps more important is the variation from population to population (or from locus to locus in the one population). Their combined effect can be checked theoretically by the use of the second-order frequency spectrum to help calculate variances of $\overline{H}$ and $\overline{n}_a(x < q)$. This spectrum was obtained by Watterson (1974*a*) in the discrete neutral case:

$$E(\beta(j)\beta(k)) = \theta^2(2n)_{[j+k]}/[jk(\theta+2n-1)_{[j+k]}] \quad \text{for} \quad (j \neq k) \quad (2.8a)$$

and

$$E(\beta(j)(\beta(j)-1)) = \theta^2(2n)_{[2j]}/[j^2(\theta+2n-1)_{[2j]}]. \quad (2.8b)$$

An approximation is

$$E(\beta(j)\beta(k)) \approx \theta^2(1-(j+k)/(2n))^{\theta-1}/[jk] \quad \text{for} \quad (j \neq k, j+k \leqslant 2n), \quad (2.9a)$$

$$E(\beta(j)(\beta(j)-1)) \approx \theta^2(1-(2j)/(2n))^{\theta-1}/j^2 \quad \text{for} \quad (2j \leqslant 2n), \quad (2.9b)$$

see Watterson (1974*a*). For large values of $2n$, there is a very close agreement between (2.8) and (2.9); indeed, there is exact agreement when $\theta = 1$. For *low* values of j, the $\beta(j)$ are approximately independent Poisson variables with $E(\beta(j)) \approx \theta/j$.

Using (2.9) we can obtain the approximations that, for the average over l samples each of size $2n$ genes,

$$\mathrm{Var}(\overline{H}) = \sum_{j=1}^{2n}\sum_{k=1}^{2n}(j/2n)^2(k/2n)^2\,\mathrm{Cov}(\beta(j),\beta(k))/l \approx 2\theta/[l(\theta+1)^2(\theta+2)(\theta+3)], \quad (2.10)$$

$$\mathrm{Var}(\overline{n}_a(x<q)) = \sum_{j=1}^{(2nq)^*}\sum_{k=1}^{(2nq)^*}\mathrm{Cov}(\beta(j),\beta(k))/l \approx \mu_q/l, \quad (2.11)$$

and

$$\mathrm{Cov}(\overline{H},\overline{n}_a(x<q)) = \sum_{j=1}^{2n}\sum_{k=1}^{(2nq)^*}(j/2n)^2\,\mathrm{Cov}(\beta(j),\beta(k))/l \approx (2nq)^*\theta\{-2/[2n(\theta+1)]+[(2nq)^*+1]/[2(2n)^2]\}/l \approx -2q\theta/[l(\theta+1)]. \quad (2.12)$$

The result (2.10) was found by Stewart (1976) and Watterson (1974*b*): (2.11) and (2.12) seem to be new. In (2.11) μ_q is given by (2.6) and $\overline{n}_a(x < q)$ is essentially the mean of l independent Poisson variables, each having a mean μ_q.

As a numerical illustration, we consider Kimura's first example, but assuming a fully neutral model with $\theta = \theta_T = 0{\cdot}114$ (one of Kimura's estimates), $2n = 3912$, $q = 0{\cdot}01$ and $l = 11$. Then $\overline{H}$ has a mean 0·102 and a standard error of 0·050, while $\overline{n}_a(x < q)$ has a mean of 0·485 and a standard error of 0·210. The covariance is $-0{\cdot}000\,19$. As a check, we have simulated samples with these parameters, and averaged them over 11 independent loci. In all, 20 replicates were calculated. The simulation method of Watterson (1984) and Hoppe (1984) was used. The resulting $\overline{H}$ values varied from 0·028 to 0·249, with a mean of 0·117 and standard deviation 0·059. The $\overline{n}_a(x < q)$ values varied from 0·182 to 0·636, with mean 0·391 and standard deviation 0·151. The covariance of $\overline{H}$ and $\overline{n}_a(x < q)$ in the replicates was $+0{\cdot}001$. The $\hat{P}$ values ranged from 0·316 to 3·347, with a mean of 1·417 and standard deviation of 0·888. Strictly speaking, $\hat{P}$ has no finite mean or variance,since it is possible that either $\overline{H} = 1$, and/or $\overline{n}_a(x < q) = 0$, in which case $\hat{P} = \infty$ in (1.7); this did not happen in our 20 replicates. (But see section 4, when it did!)

As another illustration, consider Kimura's final example, with $2n = 568{\cdot}06$ (on average), $q = 0{\cdot}01$, $l = 31$; we assume that only neutral mutations can occur and take one of Kimura's estimates for $\theta = \theta_T = 0{\cdot}215$. Then from (2.10)–(2.12), we find that $\overline{H}$ has a mean of 0·177 with standard error 0·036, $\overline{n}_a(x < q)$ has a mean of 0·491 and standard error 0·126, while the covariance between them is $-0.000\,11$. If $\overline{H}$ and $\overline{n}_a(x < q)$ take their expected values, then $\hat{P}$ becomes 0·76 (not $P_{\text{neut}} = 1$). If $\overline{H}$ were one standard deviation above its mean and $\overline{n}_a(x < q)$ were one standard deviation below its mean, then $\hat{P} = 1{\cdot}29$. If the variations went in the opposite directions, then $\hat{P} = 0{\cdot}46$. As a further check, 10 replicate simulations, each of $l = 31$ loci, produced $\overline{H}$ values ranging between 0·138 and 0·230, with mean 0·180, standard deviation 0·033, $\overline{n}_a(x < q)$ values ranging between 0·290 and 0·645, with mean 0·452 and standard deviation 0·092, while the $\hat{P}$ values themselves ranged between 0·624 and 1·148, with mean 0·864 and standard deviation 0·170. The covariance between $\overline{H}$ and $\overline{n}_a(x < q)$ was $+0{\cdot}0016$.

The conclusion here is that $\overline{H}$, $\overline{n}_a(x < q)$ and $\hat{P}$ are all very variable, if based on $l = 11$ or even as many as $l = 31$ loci.

3. Deleterious mutations

In the previous section, all mutations lead to neutral alleles. The technique suggested by Kimura was designed for cases when both neutral and deleterious alleles could occur. We mainly discuss genic selection,

https://doi.org/10.1017/S0016672300023569 Published online by Cambridge University Press

although recessive deleterious alleles will be mentioned later.

Suppose that the mutation parameter for deleterious mutations is θ_d, so that $\theta_T = \theta + \theta_d$. Further, suppose that the deleterious mutants each have the same additive effect; each deleterious gene causes a reduction of amount s in fitness. A diploid having two neutral genes has fitness 1, while one having one neutral and one deleterious gene has fitness $1-s$ and one having two deleterious genes has fitness $1-2s$. Then, at stationarity, the frequency spectrum is given by

$$\Phi(x) = x^{-1}(1-x)^{\theta_T - 1}[\theta e^{Sx} + \theta_d] \\ M(\theta, \theta_T, S(1-x))/M(\theta, \theta_T, S), \quad (3.1)$$

where M is the confluent hypergeometric function

$$M(a,b,x) = \frac{\Gamma(b)}{\Gamma(a)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\, x^n}{\Gamma(b+n)\, n!},$$

and where $S = 4N_e s$. See Ewens & Li (1980, (36)).

For large S and small x, Ewens & Li (1980), (39), (40)) show that

$$\Phi(x) \sim x^{-1}(1-x)^{\theta-1}(\theta + \theta_d e^{-Sx}),$$

while for small S,

$$\Phi(x) \sim x^{-1}(1-x)^{\theta_T - 1}\{\theta_T - S^2\theta\theta_d x[2 - x(2+\theta_T)]/ \\ [2\theta_T(\theta_T+1)]\},$$

see Watterson (1978, p. 410).

Ewens and Li (1980) also quote the heterozygosity formula

$$\mu_H = 1 - \frac{\theta M(\theta+2, \theta_T+2, S) + \theta_d M(\theta, \theta_T+2, S)}{\theta_T(\theta_T+1)\, M(\theta, \theta_T, S)} \quad (3.2)$$

which for large S becomes

$$\mu_H \approx \theta/(1+\theta) + 2\theta_d/[S(1+\theta)] + O(S^{-2}) \quad (3.3)$$

while for small S,

$$\mu_H \approx \theta_T/(1+\theta_T) \\ -S^2\theta\theta_d/[\theta_T(\theta_T+1)^2(\theta_T+2)(\theta_T+3)] + O(S^3). \quad (3.4)$$

Similarly, for large S we find, analogously to (2.6),

$$\mu_q \approx \theta\{\log[(2nq)^*+1] + 0{\cdot}57722 - 1/[2(2nq)^*+2]\} \\ + \theta_d \sum_{j=1}^{(2nq)^*} e^{-Sj/2n}/j, \quad (3.5)$$

while for small S we find

$$\mu_q \approx \theta_T\{\log[(2nq)^*+1] + 0{\cdot}57722 - 1/[2(2nq)^*+2]\} \\ - S^2\theta\theta_d(2nq)^*/[2n\theta_T(\theta_T+1)]. \quad (3.6)$$

Whether S is extreme or not, a possible computational formula for μ_q would be to use

$$\mu_q \approx \sum_{j=1}^{(2nq)^*} \Phi(j/2n)/2n \quad (3.7)$$

with Φ given now by (3.1).

Comparing (3.5) with (1.4), or even with (2.6), we see that the latter two are not satisfactory when S is large. Hence the estimator $\hat{P}$ in (1.7) is not likely to yield accurate estimates of P_{neut} except perhaps for values of S in some intermediate range. When S is very large even compared to $2n$ the θ_d terms in (3.3) and (3.5) can be neglected, and μ_H and μ_q are given approximately by (1.3) and (2.6) respectively. $\hat{P}$ would be estimating the right side of (1.6), which then becomes

$$\log(2nq)/\{\log[(2nq)^*+1] \\ + 0{\cdot}57722 - 1/[2(2nq)^*+2]\} \quad (3.8)$$

and which bears no relation to $P_{\text{neut}} = \theta/\theta_T$.

At the other extreme, when S is very small (1.6), (3.4) and (3.6) show that $\hat{P}$ would again be estimating the quantity in (3.8) rather than P_{neut}. When S is small, it is difficult for any estimator based on allele frequencies to detect the difference between neutral and deleterious alleles. On the other hand, when S is large, usually a sample would contain only neutral alleles, and again the estimator cannot be expected to perform well. The fact that $\hat{P}$ is not estimating 1, but rather, (3.8), in these circumstances is because the approximation (1.2) is inadequate when used to derive $\hat{P}$.

To illustrate the way in which S can influence $\hat{P}$, we have calculated its value as in (1.7), but assuming that infinitely many loci have been sampled so that $\overline{H}$ is equal to its mean μ_H in (3.2) and $\overline{n}_a(x<q)$ is equal to μ_q as in (3.7), subject to (3.1). The calculations were done using two sets of parameter values as estimated by Kimura. For his first example, he estimated $\theta = 0{\cdot}114$, $\theta_T = 0{\cdot}744$ (so that $\theta_d = 0{\cdot}63$), with $q = 0{\cdot}01$ and $2n = 3912$. If these were the true parameter values, $P_{\text{neut}} = 0{\cdot}153$ (as estimated by Kimura), except when $S = 0$ for which $P_{\text{neut}} = 1$ since all mutants are then neutral. The results are given in Table 1 under the heading '$\hat{P}$, ∞ loci', and in Fig. 1. We see that when S is small, say $S \leqslant 1$, or large, say $S \geqslant 50000$, $\hat{P}$ is estimating 0·86, the quantity in (3.8), rather than P_{neut}. For an intermediate range, say $25 \leqslant S \leqslant 200$, $\hat{P}$ is approximately 0·18 rather than the correct $P_{\text{neut}} = 0{\cdot}15$. So $\hat{P}$ is positively biased (except at $S = 0$), but it is reasonably close to the true value over the range $25 \leqslant S \leqslant 200$.

Similarly, in Table 2 and Fig. 2, we show the corresponding results for Kimura's last example, supposing that $\theta = 0{\cdot}215$, $\theta_d = 1{\cdot}385$, $\theta_T = 1{\cdot}6$, $q = 0{\cdot}01$ and $2n = 568$. When $\hat{P}$ is based on infinitely many loci, then it is too low when $S = 0$, it is too high otherwise, but it is fairly reasonable for cases when $25 \leqslant S \leqslant 200$, say.

In Figs. 1 and 2, the scale on the S axis is linear for $\log(1+S)$, in order to cope with the wide range of S values required. Not only do the figures show the $\hat{P}$ (∞ loci) values, but also the $\overline{H}$ and $\overline{n}_a(x<q)$ values. It is clear that $\overline{H}$ is not being much influenced by deleterious alleles having $S \geqslant 100$, say, whereas

Table 1. *Variation of $\hat{P}$ with selection*

S	True P_{neut}	$\hat{P}$		
		∞ loci	11 loci	
			Mean[a]	S.D.[a]
0	1	0·86	1·01	0·39
0·1	0·153	0·86	0·95	0·43
1	0·153	0·85	0·89	0·32
5	0·153	0·51	0·62	0·17
10	0·153	0·31	0·30	0·14
25	0·153	0·21	0·18	0·05
50	0·153	0·18	0·16	0·10
75	0·153	0·18	0·15	0·06
100	0·153	0·18	0·24	0·13
200	0·153	0·19	0·19	0·11
500	0·153	0·24	0·25	0·08
1000	0·153	0·30	0·23	0·08
5000	0·153	0·61	0·65	0·52
10000	0·153	0·78	1·36	0·70
50000	0·153	0·86	1·56	0·72

Parameters $2n = 3912$, $q = 0{\cdot}01$, $\theta = 0{\cdot}114$, $\theta_T = 0{\cdot}744$
[a] Based on 10 replicate simulations.

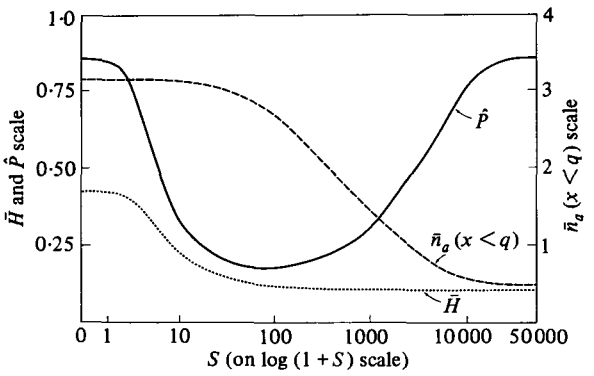

Fig. 1. Variation of $\overline{H}$, $\bar{n}_a(x < q)$ and $\hat{P}$ with Selection. Parameters: $2n = 3912$, $q = 0{\cdot}01$, $\theta = 0{\cdot}114$, $\theta_T = 0{\cdot}744$, $l = \infty$.

$\bar{n}_a(x < q)$ is being influenced by such alleles, even up to $S = 10000$ say. It is for reasons like this that Kimura proposed his estimator. However, the estimator $\hat{P}$ is reasonably effective only over the region where the $\hat{P}$ graph is near its minimum.

In section 5 of Ewens & Li (1980), formulae are given for the alleles' frequency spectrum, $\Phi(x)$, when deleterious alleles are recessive. Using those formulae, it may be shown that biases in $\hat{P}$ under recessive deleterious alleles are similar to those found above for additive deleterious alleles.

As in section 2, we can here expect sampling fluctuations to be very important when $\hat{P}$ is calculated from averages, $\overline{H}$ and $\bar{n}_a(x < q)$, based on only $l = 11$ or $l = 31$ loci. Theoretical studies would be possible using Griffiths' (1983, (7)) result for the higher order frequency spectra, analogous to (2.9) above for the neutral case. Instead, however, we proceed by simulation studies. In the appendix we give a new method for simulating samples from a stationary population which may include both neutral and selectively deleterious genes. We have used this method to obtain 10 replicates for each of the two examples discussed already, with $l = 11$ and $l = 31$ loci each, and for the same values of S as studied above. When $\hat{P}$ is calculated from finite samples from a finite number of loci, it does not have a finite mean or standard deviation ($\hat{P} = \infty$ is possible). Nevertheless, in Tables 1 and 2 we show the means and standard deviations observed over the 10 replicates, as a way of indicating the sort of $\hat{P}$ values obtained, and their variability.

The results in Tables 1 and 2 indicate that $\hat{P}$ (l loci) takes values around its theoretical value (∞ loci), but with considerable sampling variation. For instance, when $S = 50$ so that $\hat{P}$ (∞ loci) is reasonably close to P_{neut}, still the 10 replicates varied between $\hat{P} = 0{\cdot}06$ and 0.33 in the Table 1 example, and between $\hat{P} = 0{\cdot}08$ and 0·24 in the Table 2 example. There is evidence that $\hat{P}$ in Table 2 is less variable than in Table 1, a not-unexpected result considering the higher number of loci (more than offsetting the smaller sample size) in the Table 2 example compared with the Table 1 example. It is interesting that, in Table 1, with $l = 11$ loci the distribution of $\hat{P}$ seems skewed towards higher values (compared with the ∞ loci results) in the extreme cases $S = 0$ and $S = 10000$, or 50000, when virtually all genes in the sample would be neutral. The same tendency is much less marked in Table 2, with $l = 31$ loci.

Variation increases as $\hat{P}$ itself increases. Thus for the 10 replicates for $S = 50000$ in Table 1, the smallest was $\hat{P} = 0.48$ and the largest was $\hat{P} = 2.88$. Of course, $\hat{P}$ values greater than 1 would, presumably, be truncated to 1 in practice.

https://doi.org/10.1017/S0016672300023569 Published online by Cambridge University Press

Table 2. *Variation of $\hat{P}$ with selection*

		$\hat{P}$		
			31 loci	
S	True P_{neut}	∞ loci	Mean[a]	S.D.[a]
0	1·000	0·76	0·80	0·10
0·1	0·134	0·76	0·72	0·06
1	0·134	0·76	0·78	0·15
5	0·134	0·53	0·60	0·08
10	0·134	0·30	0·30	0·07
25	0·134	0·19	0·17	0·02
50	0·134	0·16	0·14	0·05
75	0·134	0·16	0·14	0·03
100	0·134	0·16	0·15	0·02
200	0·134	0·19	0·16	0·05
500	0·134	0·31	0·26	0·05
1000	0·134	0·50	0·42	0·11
5000	0·134	0·76	0·51	0·11
10000	0·134	0·76	0·59	0·15
50000	0·134	0·76	0·93	0·36

Parameters $2n = 568$, $q = 0{\cdot}01$, $\theta = 0{\cdot}215$, $\theta_T = 1{\cdot}6$.
[a] Based on 10 replicate simulations.

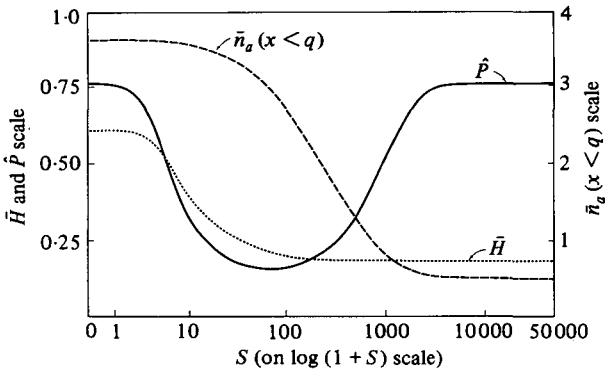

Fig. 2. Variation of $\overline{H}$, $\bar{n}_a(x<q)$ and $\hat{P}$ with Selection. Parameters: $2n = 568$, $q = 0{\cdot}01$, $\theta = 0{\cdot}215$, $\theta_T = 1{\cdot}6$, $l = \infty$.

4. Bottleneck effects, neutral alleles only

In this section, we investigate the effects that bottlenecks may have on the estimation of P_{neut}. In the previous sections, calculations were done assuming the population was at a statistically stationary equilibrium. Now, we discuss cases when that assumption does not hold.

Suppose that time is measured backwards into the past, in units of $2N_e$ generations (where N_e is the diploid effective population size at the time in question). The sample will be taken to be at time $t = 0$, and for the period $0 \leqslant t \leqslant t_1$ in the recent past the population was of constant size N_1 and the mutation parameter was $\theta_1 = 4N_1\nu_1$. For the more remote past with $t_1 < t \leqslant t_1 + t_2$, the population was of constant size N_2 and the mutation parameter was $\theta_2 = 4N_2\nu_2$. For still more remote past times, we assume that the population size varies periodically between N_1 and N_2, over respective time intervals of lengths t_1 and t_2, and having respective mutation parameters θ_1 and θ_2.

Here, we emphasize that we are now assuming all mutants are new and selectively neutral, so that the mutation parameters θ_1 and θ_2 both correspond to $\theta = \theta_T$, but they apply at different time periods in the past.

Suppose that a random sample of $2n$ genes is chosen from each of l independent loci. Unknown to the experimenter is the fact that the true proportion of neutral mutants is $P_{neut} = 1$. The sample estimate $\hat{P}$, if it differs from 1, does so partly because of the intrinsic bias in (1.7), partly because of small-sample effects, as discussed above, and now, partly because of the non-stationarity in the population caused by the periodic bottlenecks in the past.

In Watterson (1988), a method of studying such samples by computer simulation is presented. We use that method here to investigate the behaviour of $\hat{P}$. As a particular example, we again use parameters suggested by Kimura's (1983) first example. We take a sample of $2n = 3912$ genes, from each of $l = 11$ loci, and use the cut-off $q = 0{\cdot}01$. For θ values, we take $\theta_1 = 0{\cdot}114$ and $\theta_2 = 0{\cdot}744$ (which are numerically identical with θ and θ_T used earlier, but which now have different interpretations). With such θ values, our sample is being taken at the end of a period of small population size which was preceded by alternating large and small population phases. We also consider the opposite case $\theta_1 = 0{\cdot}744$, $\theta_2 = 0{\cdot}114$ in which our sample is drawn from a large population, following on from earlier small and large phases. The lengths of the time periods, t_1 and t_2, are varied, to see their effects on $\hat{P}$. In particular, by taking $t_1 = \infty$ we can study the

https://doi.org/10.1017/S0016672300023569 Published online by Cambridge University Press

Table 3. *Variation of $\hat{P}$ after periodic bottlenecks (true $P_{neut} = 1$)*

t_2 \ t_1	0·01	0·1	1	10	∞
0·01	[a]	1·07	1·96	∞[b]	
		(0·39)	(1·29)	(∞[b])	
0·1	3·47	3·79	1·39	0·88	
	(1·40)	(1·00)	(0·53)	(0·46)	
1	2·52	6·13	2·75	∞[c]	1·42[d]
	(0·81)	(2·57)	(2·67)	(∞[c])	(0·89)
10	2·21	5·85[d]	2·81	1·91	
	(0·68)	(2·07)	(1·95)	(2·83)	
∞	2·78	4·97	3·64	1·49	
	(1·12)	(2·40)	(2·29)	(1·18)	

Parameters $2n = 3912$, $l = 11$, $q = 0{\cdot}01$, $\theta_1 = 0.114$, $\theta_2 = 0{\cdot}744$; *tabulated*: mean $\hat{P}$ (*s.d.* $\hat{P}$) for 10 replicates.
[a] Not computed; thousands of periods required.
[b] One replicate had $\bar{n}_a(x < q) = 0$, $\hat{P} = \infty$. The other 9 replicates had average $\hat{P} = 1{\cdot}48$, *s.d.* = 0.57.
[c] One replicate had $\bar{n}_a(x < q) = 0$, $\hat{P} = \infty$. The other 9 replicates had average $\hat{P} = 1{\cdot}26$, *s.d.* = 1.19.
[d] Based on 20 replicates.

Table 4. *Variation of $\hat{P}$ after periodic bottlenecks (true $P_{neut} = 1$)*

t_2 \ t_1	0·01	0·1	1	10	∞
0·01	[a]	0·85	0·96	0·90	
		(0·19)	(0·23)	(0·29)	
0·1	0·33	0·49	0·92	0·90	
	(0·10)	(0·12)	(0·17)	(0·27)	
1	0·22	0·26	0·83	0·88	0·89
	(0·13)	(0·07)	(0·27)	(0·21)	(0·18)
10	0·18	0·17	0·79	0·94	
	(0·11)	(0·06)	(0·19)	(0·21)	
∞	0·19	0·18	0·59	0·90	
	(0·09)	(0·11)	(0·24)	(0·21)	

Parameters $2n = 3912$, $l = 11$, $q = 0{\cdot}01$, $\theta_1 = 0.744$, $\theta_2 = 0{\cdot}114$; *tabulated*: mean $\hat{P}$ (*s.d.* $\hat{P}$) for 10 replicates.
[a] Not computed; thousands of periods required.

stationary case, while if $t_2 = \infty$ with $t_1 < \infty$, we have a case when only *one* change of population size occurs.

Each set of parameter combinations was replicated 10 times, so that the means and standard deviations of $\overline{H}$ and $\bar{n}_a(x < q)$ could be estimated and sample means and standard deviations for $\hat{P}$ obtained. Recall that each replicate itself involves averaging over $l = 11$ loci, as per Kimura's example. The results for $\hat{P}$ are presented in Table 3, where $\theta_1 < \theta_2$, and in Table 4 when $\theta_1 > \theta_2$.

In Table 3, we see that $\hat{P}$ usually takes values which are impossible for P_{neut}, namely values greater than 1. The variability is substantial; for instance when $t_1 = t_2 = 1$, the ten replicate values varied from $\hat{P} = 0{\cdot}77$ to 10·18. The two worst instances were, however, when $t_1 = 10$ and $t_2 = 0{\cdot}01$ or $t_2 = 10$. In both cases, one of the ten replicates had 11 loci all with $\bar{n}_a(x < q) = 0$, so that $\bar{n}_a(x < q) = 0$ and $\hat{P} = \infty$. It is possibly fortuitous that in both cases, $t_1 = 10$. There was a replicate (for $t_1 = t_2 = 1$) in which ten of the eleven loci had $n_a(x < q) = 0$, while the eleventh had $n_a(x < q) = 1$; but the same also occurred with $t_1 = t_2 = 10$.

In Table 4, with $\theta_1 > \theta_2$, the $\hat{P}$ values are typically below the true value $P_{neut} = 1$, due to bottleneck effects. These are most marked, as might be expected, when t_1 is small (0·01 or 0·1) so that there has recently been a rapid population expansion, but the heterozygosity has not yet risen correspondingly. Indeed, the $\hat{P}$ values then are of similar magnitude to those found by Kimura, and indicate that a recent population expansion could produce effects which might be erroneously attributed to the presence of deleterious alleles.

https://doi.org/10.1017/S0016672300023569 Published online by Cambridge University Press

Of course it may be possible to correct $\hat{P}$ for the effects of any *known* bottlenecks. Simulation studies, similar to those used here, may help.

Appendix

It is a straightforward matter to simulate samples from a population with two classes of alleles. We first discuss some preliminary results. Let X denote the proportion of genes in the population which are of neutral allelic type. Then (Griffiths 1983 (4)) X has a probability density

$$f(x) = \frac{\Gamma(\theta_T)}{\Gamma(\theta)\,\Gamma(\theta_d)\,M(\theta,\theta_T,S)}\, e^{Sx}\, x^{\theta-1}(1-x)^{\theta_d-1}, \quad \text{for} \quad (0<x<1),$$

which has mean

$$E(X) = \frac{\theta M(\theta+1,\theta_T+1,S)}{\theta_T\, M(\theta,\theta_T,S)}.$$

Suppose that a random sample of $2n$ genes is to be chosen. The number, Y say, of these genes which are of the neutral types will have the mixed-binomial distribution

$$P(Y=y) = \int_0^1 \binom{2n}{y} x^y(1-x)^{2n-y} f(x)\,dx$$

$$= \frac{\Gamma(\theta_T)}{\Gamma(\theta)\,\Gamma(\theta_d)\,M(\theta,\theta_T,S)} \binom{2n}{y} \int_0^1 e^{Sx}\, x^{y+\theta-1} (1-x)^{2n-y+\theta_d-1}\,dx \quad \text{(A 1)}$$

$$= \binom{2n}{y} \frac{\Gamma(y+\theta)\,\Gamma(2n-y+\theta_d)\,\Gamma(\theta_T)}{\Gamma(\theta)\,\Gamma(\theta_d)\,\Gamma(2n+\theta_T)} \frac{M(y+\theta,2n+\theta_T,S)}{M(\theta,\theta_T,S)}, \quad y=0,1,2,\ldots,2n.$$

That is,

$$P(Y=y) = \binom{2n}{y} \frac{\theta_{(y)}(\theta_d)_{(2n-y)}}{(\theta_T)_{(2n)}} \frac{M(y+\theta,2n+\theta_T,S)}{M(\theta,\theta_T,S)}, \quad \text{for} \quad (y=0,1,2,\ldots,2n), \quad \text{(A 2)}$$

where $\theta_{(y)} = \theta(\theta+1)\ldots(\theta+y-1)$.

While simple recursion relations exist from which such probabilities can be computed fairly easily, it is perhaps preferable to proceed differently. From (A 1), we have

$$P(Y=y) = \frac{\Gamma(\theta_T)}{\Gamma(\theta)\,\Gamma(\theta_d)\,M(\theta,\theta_T,S)} \binom{2n}{y} \sum_{m=0}^{\infty} \frac{S^m}{m!} \frac{\Gamma(m+y+\theta)\,\Gamma(2n-y+\theta_d)}{\Gamma(m+2n+\theta_T)}$$

$$= \sum_{m=0}^{\infty} f_S(m)\, g_m(y), \quad \text{(A 3)}$$

where

$$f_S(m) = \frac{\Gamma(\theta_T)}{\Gamma(\theta)\,M(\theta,\theta_T,S)} \frac{S^m\,\Gamma(m+\theta)}{m!\,\Gamma(m+\theta_T)}$$

$$= \frac{S^m\,\theta_{(m)}}{m!(\theta_T)_{(m)}} [M(\theta,\theta_T,S)]^{-1} \quad (m=0,1,2,\ldots), \quad \text{(A 4)}$$

and

$$g_m(y) = \binom{2n}{y} \frac{(m+\theta)_{(y)}\,(\theta_d)_{(2n-y)}}{(m+\theta_T)_{(2n)}} \quad (y=0,1,\ldots,2n). \quad \text{(A 5)}$$

Both $\{f_S(m)\}$ and $\{g_m(y)\}$ are probability distributions. The former is denoted E_1CB by Gurland and Tripathi (1975), and it is a three parameter hyper-Poisson distribution. If $\theta=\theta_T$, it reduces to a Poisson distribution, mean S. In general, it has a probability generating function $G(\phi) = M(\theta,\theta_T,S\phi)/M(\theta,\theta_T,S)$, from which it can be shown that, for large S, m is approximately normal with mean and variance approximately equal to $S-\theta_d$ and S respectively. Apart from the first term

$$f_S(0) = 1/M(\theta,\theta_T,S), \quad \text{(A 6)}$$

the remaining terms are given by the recurrence

$$f_S(m+1) = \frac{S(\theta+m)}{(m+1)\,(\theta_T+m)} f_S(m). \quad \text{(A 7)}$$

The second factor in (A 3), $g_m(y)$, is the Polya–Eggenberger distribution, corresponding to the probability of drawing y neutral genes in $2n$ draws from an urn, in which initially there were $m+\theta$ neutral genes, θ_d deleterious genes, and in which the drawing is with replacement and, moreover, in which one extra gene is added to the urn after each drawing, of the same class (neutral or deleterious) as the one just drawn.

Suppose that the proportion, X, of neutral genes in the population is given. Kingman (1980 p. 58) and Griffiths (1983) have shown that those neutral genes are distributed among the neutral alleles according to the Poisson–Dirichlet distribution with mutation parameter θ. Similarly, the deleterious allele relative frequencies, when divided by $1-X$, have the Poisson–Dirichlet distribution with mutation parameter θ_d. The corresponding results within a sample of $2n$ genes, given that $Y=y$ of them are neutral and $2n-y$ are deleterious, are that the neutral genes are distributed among the neutral alleles according to Ewens' (1972) sampling distribution with parameter θ and sample size y, whereas the deleterious alleles also have Ewens' distribution with parameter θ_d and sample size $2n-y$.

Now it is known, Hoppe (1984) and Watterson (1984), that the allelic types of genes in a sample can be easily simulated by a Polya urn scheme, consistent with Ewens' distribution. So it is very neat that, for a given m, the number y of neutral alleles, their allelic

https://doi.org/10.1017/S0016672300023569 Published online by Cambridge University Press

types, and the allelic types of the remaining $2n-y$ deleterious alleles, can all be simulated by the one pass through a Polya-urn scheme computer subroutine. The total simulation scheme is then as follows:

(*a*) For given S, θ, θ_T, calculate the probabilities (A 4), using (A 6) and (A 7), or, for large S (say $S \geqslant 50$) use a normal approximation, mean $S-\theta_d$, variance S.

(*b*) Use a standard method to simulate a variate, m, having (A 4) or the approximating normal as its distribution, e.g. the alias method (Kennedy & Gentle, (1980) pp. 197–200).

(*c*) For the jth gene in the sample ($j = 1, 2, \ldots, 2n$), suppose that in the $j-1$ previous draws, x neutral and $j-1-x$ deleterious genes have been drawn. Then, consistent with (A 5), decide whether the jth gene is neutral (with probability $(m+\theta+x)/(m+\theta_T+j-1)$) or deleterious (with probability $(\theta_d+j-1-x)/(m+\theta_T+j-1)$). If the gene is to be neutral, it is either a new mutant allele (with probability $\theta/(\theta+x)$) or it copies the allelic type of one of the x already-allocated neutral genes (each having probability $1/(\theta+x)$ of being copied). On the other hand if the jth gene is to be deleterious, its type is either a new mutant allele (with probability $\theta_d/(\theta_d+j-1-x)$) or a copy of the type of an allocated deleterious gene (each having probability $1/(\theta_d+j-1-x)$ of being copied).

The above scheme for two classes of genes can be generalized to any number of classes, using Griffiths' (1983, pp. 9, 10) results.

I thank Bob Griffiths for very helpful suggestions concerning the simulation scheme presented in the appendix.

References

Ewens, W. J. (1964). The maintenance of alleles by mutation. *Genetics* **50**, 891–898.

Ewens, W. J. & Li, W.-H. (1980). Frequency spectra of neutral and deleterious alleles in a finite population. *Journal of Mathematical Biology* **10**, 155–166.

Griffiths, R. C. (1983). Allele frequencies with genetic selection. *Journal of Mathematical Biology* **17**, 1–10.

Gurland, J. & Tripathi, R. (1975). Estimation of parameters on some extensions of the Katz family of discrete distributions involving hypergeometric functions. In *Statistical Distributions in Scientific Work*, vol. 1 (ed. G. P. Patil, S. Kotz and J. K. Ord), pp. 59–79. Boston: Reidel.

Hoppe, F. M. (1984). Polya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology* **20**, 91–94.

Kennedy, W. J. & Gentle, J. E. (1980). *Statistical Computing*. New York: Marcel Dekker.

Kimura, M. (1983). Rare variant alleles in the light of the neutral theory. *Molecular Biology and Evolution* **1**, 84–93.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Kingman, J. F. C. (1980). Mathematics of genetic diversity. *SIAM CBMS-NSF regional conference series in applied mathematics*, **34**.

Nei, M. (1977). Estimation of mutation rate from rare protein variants. *American Journal of Human Genetics* **29**, 225-232.

Stewart, F. M. (1976). Variability in the amount of heterozygosity maintained in neutral populations. *Theoretical Population Biology* **9**, 188–201.

Watterson, G. A. (1974*a*). The sampling theory of selectively neutral alleles. *Advances in Applied Probability* **6**, 463–488.

Watterson, G. A. (1974*b*). Models for the logarithmic species abundance distributions. *Theoretical Population Biology* **6**, 217–250.

Watterson, G. A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Watterson, G. A. (1984). Estimating the divergence time of two species. Statistics Research Report 94, Monash University.

Watterson, G. A. (1988). The neutral alleles model with bottlenecks. In *Mathematical Evolutionary Theory* (ed. M. W. Feldman) Princeton: Princeton University Press (in press).