Mathematical Population Studies

An International Journal of Mathematical Demography

Two Variance Results in Population Genetics Theory

To cite this Article: , 'Two Variance Results in Population Genetics Theory',
Mathematical Population Studies, 14:2, 93 - 110
To link to this article: DOI: 10.1080/08898480701298376
URL: http://dx.doi.org/10.1080/08898480701298376

Mathematical Population Studies, 14:93–110, 2007
Copyright © Taylor & Francis Group, LLC
ISSN: 0889-8480 print/1547-724X online
DOI: 10.1080/08898480701298376

Two Variance Results in Population Genetics Theory

Warren J. Ewens
Department of Biology, University of Pennsylvania, USA

Arindam Roy Choudhury
Department of Organismic and Evolutionary Biology,
Harvard University, USA

Richard C. Lewontin
Museum of Comparative Zoology, Harvard University, USA

Carsten Wiuf
Bioinformatics Research Center, University of Aarhus,
Denmark

The assessment of the degree of genetic variation in a natural population, and the nature of that variation, is of central importance in both theoretical and applied population studies. Two "variance" results in population genetics theory are presented. For the first, expressions are found for the expected difference in the estimates of genetic variation in a population obtained by two investigators sampling from the same population in the same generation. The second result concerns the question of whether the degree of genetic variation in a population is best estimated by using the number of alleles observed in a sample of genes or by the number of polymorphic sites observed in the sample. For some combinations of the actual degree of variation and the sample size the former is preferred while for other combinations the latter is preferred. The reason for this is discussed.

Keywords: alleles; estimation; genetics; optimality; sites; variance

INTRODUCTION

Knowledge of the degree of genetic variability in a population is important for several reasons. This variability can only be estimated, in

Address correspondence to Professor Warren J. Ewens, Department of Biology, University of Pennsylvania, Philadelphia, PA 19104-6018, USA, Tel. (215) 898-7109. E-mail: wewens@sas.upenn.edu.

practice, by a sample taken from the population, with the sample size usually being far less than the population size. We assume that this estimation is based on a sample consists of n aligned DNA segments, each of course of the same length. These segments possibly correspond to some gene, and for convenience we shall refer throughout to these segments as genes, although the theory is unchanged for segments of any form. The genetic variability exhibited by this sample can be estimated using either "alleles" data or "sites" data. An allele is a particular DNA sequence for this segment, and the sample of n sequences will reveal some number k of different sequences, or alleles, where $1 \le k \le n$, together with their sample frequencies. The same sample will exhibit some number s of polymorphic sites, at each of which, in the model that we consider, two different bases are observed in the sample. Large values of k and s suggest substantial genetic variability in the population.

We focus on that component of genetic variation caused by base substitutions. The total mutation rate for the segment under consideration is denoted by u, this being the mean number of base changes in any meiotic event. In the simple infinitely many alleles neutral Wright-Fisher model (Ewens, 2004: 111–117), which we assume governs the population evolution, genetic variation is normally measured by the composite parameter θ, defined as $4Nu$, where N is the (diploid) population size. More general definitions of θ apply for more complex models, in which case the effective population size N_e is used, and $\theta = 4\ N_e\ u$. In these models the stationary probability that two segments in the sample have different DNA sequences (equivalently, determine two different alleles) is $\theta/(1+\theta)$. This and other formulae given in this section are diffusion, or equivalently coalescent theory, approximations, and are calculated under the assumption that the sample is small compared to the population size. The case where the sample size is not small relative to the population size is discussed in the "finite populations" section.

In the infinitely many sites model, the mean number of sites at which the two segments differ is θ. Thus in both the infinitely many alleles and the infinitely many sites models, estimation of population genetic variation is in effect estimation either of θ or of $\theta/(1+\theta)$, and we adopt this viewpoint.

Our results are limited to the case where the population remains constant over time, and assume that theory derived for the Kingman coalescent process (Kingman, 1982a,b,c) is sufficiently accurate for the evolutionary model assumed. Some aspects of the accuracy of the approximations involved are discussed. Generalizations of the results to cases beyond those considered would be interesting, for example to the case where there are several investigators whose sample sizes are not necessary equal, but these generalizations appear to be difficult to obtain.

The two questions we address are

1. What can be said about the difference in the estimates of θ made by two different investigators, each taking a sample of n (aligned) segments from the same population in the same generation?
2. Is θ best estimated by using the number k of different alleles observed in the data or by using the number s of segregating sites observed in the data?

The first question arises from the possibility, given the volume of genetic data now available, that several investigators will estimate the value of θ for the same population at the same time, and the difference between their estimates is related to the consistency with which we expect θ to be estimated. It was claimed by Ewens (2004: 313) that the answer to the second question is that θ is better estimated from s rather than k, but it was noted by one of us (AR) that if the standard approximate formula (15) for the mean square error of the alleles-based estimator is used, this is not always the case. We discuss the circumstances in which k or s provides the better estimate of θ.

THE NARROW VARIANCE OF ESTIMATES OF θ, USING ALLELES DATA

We assume a sample of n genes taken at stationarity from a diploid population of size N evolving according to the neutral Wright-Fisher model. In the infinitely many alleles model, appropriate for describing alleles variation, the joint distribution of the number of allelic types seen in this sample, together with their frequencies, is well known (Ewens, 2004: 114). The form of this distribution implies that the number K of alleles seen in the sample is a sufficient statistic for θ, so that from standard statistical theory, optimal estimation of θ using alleles data is carried out by using the observed value k of K only. It is therefore necessary, in discussing the extent to which the estimate of θ found by one investigator can be expected to differ from that of another investigator, to discuss first the extent to which the number of alleles observed by the two investigators can be expected to differ.

Suppose that two investigators take a sample, each of n sequences, from the same population in the same generation. The first will see some (random) number K_1 alleles and the second some (random) number K_2. The "narrow" or "sampling" variance of K is defined as the mean value of $(1/2)(K_1 - K_2)^2$. This is identical to the broad variance of K minus the covariance between K_1 and K_2, where the broad variance is the mean of $(K - E(K))^2$. A somewhat different definition of the narrow variance is given in Hein et al. (2005: 65). Our first aim is to find an expression for this narrow variance.

To illustrate one approach to the required calculation we consider the case $n = 3$. We think of the three genes drawn by each investigator as a single sample of six genes. There are 11 allelic configurations of these six genes, which we write as $\{6\}, \{5,1\}$, $\{4,2\}$, $\{4,1,1\}, \{3,3\}, \{3,2,1\}, \{3,1,1,1\}, \{2,2,2\}, \{2,2,1,1\}, \{2,1,1,1,1\}$ and $\{1,1,1,1,1,1\}$. As an example, the configuration $\{3,1,1,1\}$ implies that four different alleles were observed among the six genes, with one allele seen three times and the remaining three alleles once each.

We denote the observed numbers of alleles seen by the two investigators as k_1 and k_2, respectively. For the configuration $\{6\}$, for which only one allele is seen in the six genes, necessarily $k_1 = k_2 = 1$, so that $|k_1 - k_2| = 0$. Further, $|k_1 - k_2|$ must be 0 for other configurations, for example $\{2,2,2\}$ and $\{1,1,1,1,1,1\}$. On the other hand, $|k_1 - k_2|$ must be 1 for the configuration $\{5, 1\}$, for which one of the investigators must see one allele in the three genes sampled and the other must see one allele arising twice and another once.

For some configurations $|k_1 - k_2|$ can be 0 or 1, for others 1 or 2 and for others again 0 or 2. Thus for the configuration $\{4,1,1\}$ (three alleles, one (A) seen four times, two (B and C) seen once each) we find $|k_1 - k_2| = 0$ if the first investigator's sample is AAB and second's is AAC, while $|k_1 - k_2| = 2$ if the first investigator's sample is AAA and second's is ABC. Similar calculations arise for all other configurations.

The probabilities of all the above possibilities are known (Ewens, 2004: 114), and from these and the arguments in the previous paragraph it is found that

$$\frac{E(K_1 - K_2)^2}{2} = \frac{\theta(90 + 114\theta + 35\theta^2 + 3\theta^3)}{S(\theta)}, \tag{1}$$

where

$$S(\theta) = (\theta + 1)(\theta + 2)(\theta + 3)(\theta + 4)(\theta + 5). \tag{2}$$

It is interesting to compare this with the “broad” variance of K, which for the case $n = 3$ is $\theta(6 + 8\theta + 3\theta^2)/[(\theta + 1)^2(\theta + 2)^2]$. The ratio of the narrow to the broad variance is

$$\frac{(\theta + 1)(\theta + 2)[30 + 38\theta + (35/3)\theta^2 + \theta^3]}{(\theta + 3)(\theta + 4)(\theta + 5)[2 + (8/3)\theta + \theta^2]}. \tag{3}$$

This always lies between 1/2 and 1, approaching 1/2 as θ approaches 0 and approaching 1 as θ increases without limit. Thus the narrow variance is never less than half the broad variance for samples of size 3.

Although it is in principle possible to find the narrow variance for any arbitrary value of n by the above method, in practice this approach leads to extremely complicated and extensive calculations even for n as small as 7 or 8, and another approach is needed. This approach is based on the (random) population frequencies p_1, $p_2, \ldots,$ of the alleles in the population in any generation. These frequencies have a complicated stationary distribution (Watterson, 1976). However, despite this complexity, much is known about the mean of sums of functions of these frequencies. For example, if $\phi(p)$ is a function of order p or less near $p = 0$, then

$$E \sum_m \phi(p_m) = \theta \int_0^1 \phi(p) p^{-1} (1-p)^{\theta-1}\, dp. \tag{4}$$

The function $\theta p^{-1}(1-p)^{\theta-1}$ involved in the right-hand side is called the (univariate) frequency spectrum of the population allelic frequencies, and has the interpretation that if terms of order $(\delta p)^2$ and smaller are ignored, the probability that there exists an allele in the population with frequency between p and $p + \delta p$ is $\theta p^{-1}(1-p)^{\theta-1}\delta p$.

We write the number K_{j} of alleles seen by investigator $j (j = 1, 2)$ as $K_{\mathrm{j}} = I_{\mathrm{j}1} + I_{\mathrm{j}2} + \ldots\ldots,$ where the indicator function I_{jm} takes the value 1 if investigator j sees allele m in his sample and 0 otherwise. From this, the narrow variance $(1/2)E(K_1 - K_2)^2$ of K is

$$\frac{1}{2} E \sum_m (I_{1m} - I_{2m})^2 + \frac{1}{2} E \sum_{m \neq r} \sum_r (I_{1m} - I_{2m})(I_{1r} - I_{2r}). \tag{5}$$

Now $(I_{1\mathrm{m}} - I_{2\mathrm{m}})^2 = 1$ if and only if allele m is seen in the sample of one observer and not in that of the other. Given sample sizes n for each observer, and given a population frequency p_m of allele m, the conditional probability of this event, which is the conditional expected value of $(I_{1m} - I_{2m})^2$, is $2(1-p_m)^n\{1-(1-p_m)^n\}$. Using (4), the unconditional expected value of the first term on the right-hand side of (5) is

$$\int_0^1 (1-p)^n \{1 - (1-p)^n\} \theta p^{-1} (1-p)^{\theta-1} dp, \tag{6}$$

which reduces to

$$S(n, \theta) = \frac{\theta}{\theta + n} + \frac{\theta}{\theta + n - 1} + \cdots + \frac{\theta}{\theta + 2n - 1}. \tag{7}$$

when n is large, to a close approximation $S(n, \theta) \approx \theta\ \ln\ 2$.

Eq. (7) can be checked by using another argument. The first term in the expression (5) is the mean number of alleles seen in a combined sample of $2n$ genes minus the mean number of alleles seen in a single sample of n genes. The well-known formula for the mean number of alleles seen in a sample of j genes, namely $\sum_{i=1}^{j-1}\theta/(\theta+j-1)$, leads directly to Eq. (7).

The second term on the right-hand side of the expression (5) can be evaluated by using the bivariate frequency spectrum $\theta^2(xy)^{-1}$ $(1-x-y)$, having the interpretation that if small-order terms are ignored, the probability that there exists one allele in the population with population frequency in $(x, x+\delta x)$ and another with population frequency in $(y, y+\delta y)$ is $\theta^2(xy)^{-1}(1-x-y)\delta x\,\delta y$. This function has the property that for any function $\phi(p_r, p_m)$ that is of order $p_r p_m$ or less when p_r and p_m are close to zero,

$$E\left[\sum\sum_{r\neq m}\phi(p_r, p_m)\right] = \theta^2\int_0^1\int_0^{1-x}(xy)^{-1}(1-x-y)^{\theta-1}\phi(x,y)\,dy\,dx. \tag{8}$$

This equation allows the calculation of the second term on the right-hand side of the expression (5) in a manner analogous to that leading to Eq. (6). The details are not given here, and the final conclusion is that the expression (5) becomes

$$S(n,\theta) - \theta^2\sum_{j=1}^{n}\frac{n![(j-1)!]^2(n-j+\theta-1)!}{j!(n-j)!(n+j+\theta-1)!}, \tag{9}$$

where $x!$ is defined in the standard way for non-integer x through the gamma function.

The expression (9) agrees with that given in Eq. (1) for the case $n=3$, confirming both modes of calculation. For large n the first term $S(n,\theta)$ in (9) dominates the second term, so that for large n the narrow variance of the number of alleles seen is approximately θ ln 2. It is interesting that this value does not depend on the sample size n.

Numerical computations show that the narrow variance (9) behaves in a complex way as a function of n and θ when n is small. This variance appears to increase monotonically to the asymptotic limit θ ln 2 as n increases when $\theta > 0.1663\ldots$, whereas when $\theta < 0.1663\ldots$ it appears to increase as a function of n, reach a maximum, and then decrease to θ ln 2 as n increases. We see no particular significance in the numerical value $0.1663\ldots.$

A second approach to finding the narrow variance of K depends on the following theoretical result. Let X_1, X_2, Z_1 and Z_2 be random variables such that the vector (X_1, Z_1) has the same distribution as the vector $(X_2,\ Z_2)$. Write the expected value of X_{i} given Z_{i} as E_{i} $(i = 1, 2)$ and assume that given (Z_1, Z_2), the distribution of X_{i} depends only on Z_{i} $(i = 1, 2)$. It follows that

$$1/2\ E\ [(X_1 - X_2)^2] = E\ [\mathrm{Var}\ (X_1|Z_1)] + 1/2\ E\ [(E_1 - E_2)^2], \tag{10}$$

where both outer expected values on the right-hand side are with respect to the distribution of Z_1 and Z_2. Eq. (10) is related to writing the variance of X_1 as the sum of the expectation of the conditional variance $E\ [\mathrm{Var}\ (X_1|Z_1)]$ and the variance of the conditional expectation $\mathrm{Var}\ (E_1)$.

Suppose now that X_{i} is K_{i} $(i = 1, 2)$ and that Z_{i} is the (random) allele frequency vector $(p_1, p_2, \ldots.)$ in the population. In this case $E_1 = E_2$, so that Eq. (10) reduces to

$$1/2\ E\,[(K_1 - K_2)^2] = E\,[\mathrm{Var}\ (K_1|Z)]. \tag{11}$$

This equation leads to an expression for the narrow variance of K as

$$S(n, \theta) - 1 - \sum_{i=1}^{n-1} \frac{i\theta^2}{(i+\theta)^2} + \theta^2 \left(\sum_{i=0}^{n-1} \frac{1}{i+\theta} \right)^2$$
$$- \theta^2 \sum_{i=2}^{2n} (-1)^i \frac{(n!)^2}{\prod_{j=0}^{i-1} (j+\theta)} \sum_{k=\max(1,i-n)}^{\min(n,i-1)} \frac{1}{k(i-k)(n-k)!(n-i+k)!}. \tag{12}$$

Here $S(n, \theta)$ is defined in Eq. (7), so that the first term in the above expression agrees with that in the expression (9). While it is not immediately obvious that the entire expression is identical to that in (9), the two agree in the small number of cases we have checked by hand and in the extensive number of cases that we have checked numerically.

Finite Populations

The calculations leading to the expressions (9) and (12) for the narrow variance of K assume a large population size and a comparatively small sample size. In effect they assume that conclusions from the Kingman coalescent process, which approximates the ancestry of genes whose evolution obeys the Wright-Fisher model, are sufficiently accurate, or equivalently that diffusion approximations for the Wright-Fisher model are sufficiently accurate. When the population size is not large, and in particular when the sample size is a non-negligible fraction of the total population size, this assumption does not hold and these expressions are no longer necessarily accurate.

The extent to which coalescent theory approximations are accurate in a finite population has been examined by Fu (2006). In broad terms, Fu found that coalescent approximations tend to break down when the sample size is of the order of the square root of the population size. On the other hand, he concluded that because the approximation has both positive and negative effects that largely cancel out, “in many situations beyond [those justified] by Kingman’s analysis . . . the Kingman coalescent remains close to the exact coalescent for the [Wright-Fisher] model.”

This broad statement does not give a specific numerical indication of the accuracy of the expression (9) for small populations. To investigate this matter we conducted simulations for various values of θ, n and the total population size N. It is necessary to sample from a simulated stationary population, and this was carried out by starting the population with one allele only, and then not sampling until this allele had left the population. Mutational events in the simulation followed a Poisson process.

The results of these simulations are given in Table 1. The empirical narrow variances are those found from simulations, and are subject to statistical errors. The “theoretical” narrow variances listed in the table are those given by the expression (9). When the sample size is small compared to the population size the empirical and theoretical values agree, as expected. As the sample size increases relative to the population size, however, the empirical variances increasing differ from the theoretical variances. However, the departure is not as large as might be expected, in line with the observations made by Fu (2006). Thus even if $n = 50$ and $N = 2{,}500$, so that the sample size is exactly equal to the square root of the population size, they differ only by 2.5% ($\theta = 0.2$), by 2.1% ($\theta = 1.0$), and by 7.6% ($\theta = 2.0$).

Estimation of θ

We now turn from the narrow variance of the number of alleles seen by two observers to the narrow variance of their respective estimators of θ. The maximum likelihood estimator $\hat{\theta}_k$ of θ, given the number K of alleles seen by an investigator, is defined implicitly by the equation

$$K = \sum_{j=1}^{n} \frac{\hat{\theta}_k}{(\hat{\theta}_k + j - 1)}, \tag{13}$$

This estimator is biased, and there is no unbiased estimator of θ available from alleles data.

TABLE 1 Empirical and Theoretical Narrow Variances of K, the Number of Alleles Seen in a Sample of n Genes, for Various Values of θ and n. The Theoretical Variance is Given in Eq. (9)

θ	n	N	Empirical variance	Theoretical variance
	100	1,250	.068	.069414
.1	150	2,500	.072	.069381
				Decreasing to .06932
.2	50	500	.141	
		1,000	.144	.138412
		2,500	.142	
.2	100	500	.162	
		1,000	.157	.138525
		2,500	.154	
.2	150	1,500	.129	
		2,500	.113	.138561
				Increasing to .1386
1	50	1,000	.688	
		2,500	.682	.668
		5,000	.645	
1	100	1,000	.666	
		2,500	.696	.681
		5,000	.690	
1	150	1,000	.712	
		2,500	.707	.684
		5,000	.672	
				Increasing to .693
2	50	500	1.25	
		2,500	1.19	1.28
		5,000	1.28	
2	100	500	1.25	
		2,500	1.26	1.33
		5,000	1.30	
2	250	500	1.20	1.39
				Increasing to 1.39

Equation (13) shows that, to a close approximation, $\hat{\theta}_k = K/\ln n$. Thus if the estimator of θ found by investigator i ($i = 1, 2$) is denoted $\hat{\theta}_{ki}$, we have to a close approximation

$$\frac{(\hat{\theta}_{k1} - \hat{\theta}_{k2})^2}{2} = \frac{(K_1 - K_2)^2}{2(\ln n)^2}. \tag{14}$$

Taking expectations throughout in Eq. (14), and using the approximate formula θ ln 2 for the narrow variance of K, we find that the narrow variance of the estimator $\hat{\theta}_k$ is approximately

$$\frac{[\theta \ln 2]}{[(\ln n)^2]}. \tag{15}$$

Three Comparisons

It is interesting to compare the approximate narrow variance of $\hat{\theta}_k$ given in the expression (15) with three other variances. The first of these is with the broad variance of $\hat{\theta}_k$. As already mentioned, there is no alleles-based unbiased estimator of θ, but in large samples the bias of the maximum likelihood estimator $\hat{\theta}_k$ is small, and the broad mean square error of $\hat{\theta}_k$ is, to a close approximation,

$$MSE(\hat{\theta}_k) = \frac{\theta}{\sum_{j=1}^{n-1} j/(j+\theta)^2}, \tag{16}$$

(Ewens, 2004: 304). An accurate calculation shows that for the range of parameter values that we consider, this approximation is typically within about 2% of the true mean square error. We therefore use Eq. (16) here and in the discussion in the following section. The right-hand side in Eq. (16) is approximately $\theta/\ln\ n$, so that the narrow variance (15) is smaller than this by a multiplicative factor of approximately ln $2/\ln\ n$.

Second, it is found from the expression (15) that the narrow variance of the estimate of the heterozygosity probability $\theta/(1+\theta)$, when based on the number of alleles k observed in the sample of n genes, is approximately

$$\frac{[\theta \ln 2]}{[(1+\theta)^4 (\ln n)^2]}. \tag{17}$$

The corresponding approximate broad variance of $\theta/(1+\theta)$ is

$$\frac{\theta}{[(1+\theta)^4 (\ln n)]}, \tag{18}$$

and these two expressions also differ by a multiplicative factor of approximately ln $2/\ln\ n$.

Third, the "natural" estimator of the heterozygosity probability $\theta/(1+\theta)$ is the sample heterozygosity $1 - \sum_j n_j(n_j - 1)/n(n-1)$, where allele j is observed n_j times in the sample. The sufficiency of K for θ shows that this has a larger variance, as an estimator of $\theta/(1+\theta)$, than that of the estimator of $\theta/(1+\theta)$ deriving from K. However, it does not follow that a corresponding statement holds about the narrow variance of the sample heterozygosity, measuring the extent to which the two investigators will have different values of their respective sample heterozygosities. This narrow variance can be shown to be

$$\frac{4\theta}{[n(1+\theta)(2+\theta)(3+\theta)]}, \tag{19}$$

This is of order n^{-1} and is thus of a smaller order of magnitude than the narrow variance given by the expression (17). The implication of this is that the two investigators should have quite close values for their sample heterozygosities, even though their respective value will not necessary be close to the true mean heterozygosity $\theta/(1+\theta)$.

The following result is parallel to that in the expression (17). An alternative way of estimating the sample heterozygosity is to use

$$\hat{\pi} = 2\sum_i \sum_j I_{ij}/n(n-1),$$

where $I_{ij} = 1$ is sequences i and j are different, $I_{ij} = 0$ otherwise. This estimator also has mean $\theta/(1+\theta)$. The narrow variance of $\hat{\pi}$, that is $1/2E(\hat{\pi}_1 - \hat{\pi}_2)^2$, is

$$\frac{4\theta[(n-1)\theta+n)]}{n(n-1)(1+\theta)(2+\theta)(3+\theta)}. \tag{20}$$

The comments following the expression (19) apply for this estimator also. For large n, the expressions (19) and (20) differ by a multiplicative factor $1+\theta$.

THE NARROW VARIANCE OF ESTIMATES OF θ, USING "SITES" DATA

Our emphasis is on those DNA segments corresponding to a gene, so that the assumption of no recombination between the sites in the segment considered is made throughout.

We start with the theoretical result in Eq. (10). We consider the coalescent tree spanning the joint sample of $2n$ genes, and, in the various expressions in Eq. (10), we take Z_i to be the length L_i of the tree spanning sample i ($i = 1, 2$). Let S_i be the number of mutations in the subtree corresponding to sample i. S_i has a Poisson distribution with parameter $\theta L_i/2$, and our aim is to find an expression for $E(S_1 - S_2)^2/2$ and to find an asymptotic ($n \to \infty$) expression for this variance.

The total length L of all branches in the coalescent tree can be written as $L_1 + L_2 - L_{12} + L_0$, where L_{12} is the sum of the lengths of the branches shared between the two subtrees corresponding to the two samples and L_0 is the length of the branch in the subtree for both samples not included in the subtrees of either sample. L_0 is positive only if the subtrees for both samples form monophyletic groups, and the probability of this rapidly approaches 0 as n increases.

Define X_i to be $X_i = S_i - S_{12}$, the number of mutations in sample i not shared by the other sample. It follows that

$$E(S_1 - S_2)^2 = E(X_1 - X_2)^2,$$

so that for this case, Eq. (10) becomes

$$\frac{E(S_1 - S_2)^2/2}{\theta E(L_1 - L_{12})/2 + E(L_1 - L_2)^2/2},$$

because X_i is Poisson with intensity $\theta(L_1 - L_{12})/2$. In the Appendix we prove that for large sample sizes

$$E(S_1 - S_2)^2/2 \;\approx\; \theta E(L_1 - L_{12})/2 \;\approx\; \theta \ln 2, \tag{21}$$

a result similar to that for alleles data.

The standard estimator of θ obtained from segregating sites data, given by Watterson (1975), is

$$\hat{\theta}_s = S \sum_{i=1}^{n-1} 1/i \approx S/(\ln n) \tag{22}$$

where S is defined as the number of segregating sites seen in a sample of n genes. Using the approximate result in (21), we find the narrow variance of this estimator to be close to the narrow variance of the estimator based on “alleles” data given by the expression (15).

An alternative to the estimator of θ given in Eq. (22) is the pairwise estimator

$$\hat{\theta}_s^* = 2 \sum_{i \neq j} m_{ij}/n(n-1), \tag{23}$$

where m_{ij} is the number of aligned nucleotide differences observed when comparing sequence i with sequence j. This is an unbiased estimator of θ, and its narrow variance is $2/\theta\ [3(\mathrm{n}-1)] + 2(2n+3)\theta^2/[9n(n-1)]$. This is of order n^{-1}, so that we obtain a conclusion similar to that obtained for the pairwise alleles estimator, namely that two investigators are likely to have quite close estimates of θ if each uses the pairwise estimator given in Eq. (23), despite the well-known fact that the estimator in Eq. (23) is not consistent (i:e., its variance does not approach 0 as $n \to \infty$). It follows that for large sample sizes the two estimates, although likely to be close, might differ substantially from the true value.

ESTIMATING θ USING ALLELES AND SITES DATA: A COMPARISON

We now take up the question of whether broad estimation of θ is best carried out by using alleles or sites data. It is assumed in the sites case that all sites are completely linked, since the case of interest for the DNA segments considered is that where they correspond to a gene.

We consider first the estimation of θ using sites data, specifically by using the number S of segregating sites in a sample of n genes. The variance of the standard unbiased estimator of θ, using S, given by Eq. (22), is

$$\mathrm{Var}(\hat{\theta}_s) = \theta/g_1 + g_2\theta^2/g_1^2\cdot \tag{24}$$

where $g_1 = \sum_{j=1}^{n-1} j^{-1}$, $g_2 = \sum_{j=1}^{n-1} j^{-2}\cdot$ To the extent that Eq. (16) gives a sufficiently accurate value for the mean square error of the alleles-based estimator $\hat{\theta}_k$, the comparison of the efficiencies of $\hat{\theta}_k$ and $\hat{\theta}_s$ reduces to a comparison of the numerical values derived from the expressions on the respective right-hand sides of Eq. (16) and Eq. (24).

When θ is very small we expect these numerical values to be close, since in this case we expect only a small number of segregating sites and a matching small number of alleles. This expectation is confirmed by observing that when θ is small, the expressions on the right-hand sides of both Eq. (16) and Eq. (24) are approximately $\theta/\ln n$ and that their ratio approaches 1 as θ approaches 0. It is possible to show that when $\theta < 1$, the expression on the right-hand side in Eq. (24) is always less than that on the right-hand side in Eq. (16) whatever the value of n. Numerical evidence suggests that this inequality is also true whenever $n < 50$ for all values of θ. This bound appears to be sharp: when $n = 51$ we can find values of θ for which the expression in Eq. (16) is less than that in Eq. (24). Examples are given by some of the values in Table 2, which gives the ratio of the expression in Eq. (24) to that in Eq. (16).

TABLE 2 Selected Values of Var $(\hat{\theta}_s)/\mathrm{MSE}(\hat{\theta}_k)$ for Various Values of n and θ

	$\theta = .5$	$\theta = 1$	$\theta = 3$	$\theta = 5$
$n = 50$	.902	.874	.891	.928
$n = 100$	.918	.903	.960	1.038
$n = 500$	.943	.942	1.047	1.178

For the values of n and θ listed in the table, the approximate MSE of $\hat{\theta}_k$ is less than the variance of $\hat{\theta}_s$ when n and θ are both large. However, this observation is partly misleading: for any given value of n, the MSE of $\hat{\theta}_k$ appears to be less than the variance of $\hat{\theta}_s$ only for a bounded range of values of θ. Thus when $n = 100$, the MSE of $\hat{\theta}_k$ is less than the variance of $\hat{\theta}_s$ when $4.01 < \theta < 158.9$. When $n = 200$ the corresponding range of values of θ is even wider, extending from about 4 to about 650. On the other hand, the values of n for which the approximate MSE of $\hat{\theta}_k$ is less than the variance of $\hat{\theta}_s$ appears to be of the form $n \geq n(\theta)$, for some $n(\theta)$ depending on θ. For $\theta = 3$, $n(\theta) = 177$ and for $\theta = 4$, $n(\theta) = 101$.

Neither $\hat{\theta}_k$ nor $\hat{\theta}_s$ make full use of the data in the n sample sequences. The calculation of $\hat{\theta}_k$ is found by observing whether two sequences are the same or different, but if they are different the calculation does not make use of the way in which they are different. The calculation of $\hat{\theta}_s$ is found by observing whether a site is polymorphic, but if it is, the calculation does not use the frequencies of the segregating nucleotides. More complete information about θ would be available if the times of the coalescent of the sample were known and the branches on which mutations took place, as well as the numbers of these mutations, were available (Felsenstein, 1992). In practice this information cannot be expected to be available, and a less extreme assumption is that the number of mutation events on each branch of the coalescent is available. The Fisher information bound for the asymptotic mean square error of the estimator of θ, given this information, was found by Fu and Li (1993). As expected, both the expressions in Eq. (16) and Eq. (24) exceed the value that they find, which can be written as

$$\frac{\theta}{\sum_{j=1}^{n-1} \frac{1}{j+\theta}}. \tag{25}$$

If the sites had been unlinked, the variance of $\hat{\theta}_s$ would be θ/g_1, and this also exceeds the value in Eq. (25). Thus linkage between sites, as well as the times at which various mutational events occurred, is a factor in variance calculations.

The comparison of the MSE of $\hat{\theta}_k$ and the variance of $\hat{\theta}_s$ is best discussed through the coalescent of the sample. It is a standard property of the coalescent that, when the population size is constant over time, its longest arms tend to occur just before the final coalescence to the most recent ancestor of all genes in the sample. When the sample size is small, most mutations tend to occur on these arms. Sites data record all these mutations but alleles data do not. Thus for small sample sizes, estimation of θ using sites information should have smaller mean square error than estimation of θ using alleles data, and this is what is observed. For large sample sizes there are many shorter, recent arms of the coalescent, and thus an increased chance of mutations on short arms, with perhaps only a small number of mutations on each. Single mutations on such arms are recorded in alleles data, leading to significant information about θ from these data. Thus for large sample sizes and moderate values of θ it is plausible that alleles data provides more information about θ than does sites data. Finally, as the mutation rate, and hence θ, increases, even short arms can accumulate several mutations, all of which are recorded in sites data but not in alleles data. Thus for large θ we once again expect sites data to give more information about θ than alleles data, and this agrees with the calculations derived from Eq. (16) and Eq. (24).

ACKNOWLEDGEMENT

We acknowledge the valuable comments by two anonymous referees.

REFERENCES

Ewens, W.J. (2004). *Mathematical Population Genetics I. Theoretical Introduction*. New York: Springer.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research 56*: 139–147.

Fu, Y.-X. (2006). Exact coalescent for the Wright-Fisher model. *Theoretical Population Biology 69*: 385–394.

Fu, Y.-X. and Li, W.-H. (1993). Maximum likelihood estimation of population parameters. *Genetics 134*: 1261–1270.

Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution*. Oxford: Oxford University Press.

Kingman, J.F.C. (1982a). The coalescent. *Stochastic Processes and their Applications 13*: 235–248.

Kingman, J.F.C. (1982b). Exchangeability and the evolution of large populations. In G. Koch, and F. Spizzichino (Eds.), *Exchangeability in Probability and Statistics*. Amsterdam: North-Holland, pp. 97–112.

Kingman, J.F.C. (1982c). On the genealogy of large populations. *Journal of Applied Probability 19A*: 27–43.

Saunders, I.W., Tavaré, S., and Watterson, G.A. (1984). On the genealogy of nested sub-samples from a haploid population. *Advances in Applied Probability 16*: 471–491.

Watterson, G.A. (1975). On the number of segregating sites in a genetical model without recombination. *Theoretical Population Biology 7*: 256–276.

Watterson, G.A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *Journal of Applied Probability 13*: 639–651.

APPENDIX

In this appendix we prove Eq. (21).

The variable L_i relates to the total length of the tree of the whole sample of size $2n$ through $L = L_1 + L_2 - L_{12} + L_0$, where L_{12} is the sum of branches shared between the two trees, and L_0 is all branches in the whole tree that is not counted in either L_1 or L_2. The variable L_0 is only non-zero if either sample forms a monophyletic group. If it is necessary to emphasize that these quantities depend on the sample size n for each investigator, we write them as $L(n)$, $L_1(n)$, $L_2(n)$, $L_{12}(n)$, and $L_0(n)$.

Rewriting the equality in the previous paragraph yields $L_1 - L_{12} = L + L_2 - L_0$. The proof of Eq. (21) is carried out in several steps. First we will find asymptotic expressions for the mean and variance of $L_1 - L_{12}$. To find these we find various asymptotic expressions as shown below.

A) Mean and variance of L_0:

First, $L_0 \to 0$ almost surely as $n \to \infty$, because eventually sample 1 and 2 share a MRCA (Saunders et al., 1984). Also $L_0 \leq 2H_\infty$, where H_∞ is the height of the entire (infinite) population coalescent. Since H_∞ and $(H_\infty)^2$ are integrable, it follows that $E(L_0) \to 0$ and $\mathrm{Var}(L_0) \to 0$ 0 as $n \to \infty$.

B) The mean of $L(n) - L_2(n)$:

Taking expectation yields

$$E[L(n) - L_2(n)] = 2\sum_{i=1}^{2n-1} 1/i - 2\sum_{i=1}^{n-1} 1/i. \qquad (26)$$

For large n, this gives

$$E[L(n) - L_2(n)] \approx 2(\ln(2n-1) + \gamma) + 2(\ln(n-1) + \gamma) \approx 2\ln(2), \qquad (27)$$

where γ is Euler's constant.

C) The variance of $L(n) - L_2(n)$:

Let $L(n) = \sum_{i=2}^{2n} T_i(n)$ and $L_2(n) = \sum_{i=2}^{n} T_{2i}(n)$, where $T_i(n)$, $i = 2, \ldots, 2n$ are independent exponential variableswith intensities $i(i-1)/2$, and> $T_{2i}(n)$, $i = 2, \ldots, n$ also are independent exponential variables with intensities $i(i-1)/2$. We note that $L(n)$ and $L_2(n)$ are not independent, so that $T_i(n)$ and $T_{2j}(n)$ are not independent. For any fixed K, we write

$$L(n) = \sum_{i=2}^{K} T_i(n) + \sum_{i=K+1}^{2n} T_i(n), \qquad (28)$$

and

$$L_2(n) = \sum_{i=2}^{K} T_{2i}(n) + \sum_{i=K+1}^{2n} T_{2i}(n). \tag{29}$$

For any given $\varepsilon > 0$, we can choose K large enough so that

$$\sum_{i=K+1}^{2n} \mathrm{Var}(T_{\mathrm{i}}(n)) < \varepsilon$$

and

$$\sum_{i=K+1}^{2n} \mathrm{Var}(T_{2i}(n)) < \varepsilon$$

for all $n > K$. This can be done because

$$\sup_n \mathrm{Var}(L(n)) = \sup_n \mathrm{Var}(L_1(n)) = \sum_{i=2}^{\infty} 4/(i-1)^2 = 2\pi^2/3.$$

Next, consider the sum $\sum_{i=2}^{K} T_i(n) - T_{2i}(n)$. It follows from Saunders et al. (1984) that the tree of the total sample and the tree of sample 2 eventually (for large n) share the K oldest ancestors. Hence $D_n = \sum_{i=2}^{K} T_i(n) - T_{2i}(n) \to 0$ almost surely as $n \to \infty$. The difference D_n is bounded, so that $|D_n| < KH_\infty$, and it follows that $E(D_n) \to 0$ and $\mathrm{Var}(D_n) \to 0$ as $n \to \infty$. This implies that for any $\varepsilon > 0$ there exists M such that $\mathrm{Var}(D_n) < \varepsilon$ for $n > M$. Finally, combining these results, we find that $\mathrm{Var}(L(n) - D_2(n)) < 3\varepsilon$ for any given $\varepsilon > 0$ and $n > \max(M, K)$, where M and K are as given above. It follows that

$$\mathrm{Var}(L(n) - L_2(n) \to 0 \text{ as } n \to \infty.$$

Next we turn to the mean and variance of $L_1 - L_{12}$.

D) The mean of $L_1 - L_{12}$:

Taking expectations yield $E[L_1 - L_{12}] = E[L - L_2] - E[L_0] \approx 2\ \ln(2)$ for large n (from A and B).

E) The variance of $L_1 - L_{12}$ (and $L_2 - L_{12}$):

The variance is bounded by

$$\mathrm{Var}(L_1 - L_{12}) \leq 3\ \mathrm{Var}(L - L_2) + 3\ \mathrm{Var}(L_0),$$

so that $\mathrm{Var}(L_1 - L_{12}) \to 0$ as $n \to \infty$ (from A and C).

Finally, we return to the evaluation of $E[(S_1 - S_2)^2/2$. We apply Eq. (10) with $X_i = S_i - S_{12}$ and $Z_i = L_i$. Then

$$(1/2)E[(S_1 - S_2)^2] = (1/2)E[(X_1 - X_2)^2]$$

and it follows that

$$(1/2)E[(S_1 - S_2)^2] = (\theta/2)E[L_1 - L_{12}] + (1/2)E[(L_1 - L_2)^2]$$

because $X_i = S_i - S_{12}$ is Poisson with parameter $(\theta/2)(L_1 - L_{12})$. From the fact that $E[(L_1 - L_2)^2] = E[\{(L_1 - L_{12}) - (L_2 - L_{12})\}^2] \rightarrow 0$ as $n \rightarrow \infty$, it follows that $(1/2)E[(S_1 - S_2)^2] \rightarrow \theta \ln(2)$, as we set out to prove.