BIOMETRICS 35, 757–763
December, 1979

A Comparison of Two Methods for Making Statistical Inferences on Nei's Measure of Genetic Distance

LAURENCE D. MUELLER

Department of Genetics, University of California,
Davis, California 95616, U.S.A.

Summary

The delta and jackknife methods can be used to estimate Nei's measure of genetic distance and calculate confidence intervals for this estimate. Computer simulations were used to study the bias and variance of each estimator and the accuracy of the corresponding approximate 95% confidence intervals. The simulations were conducted using 3 sets of data and several sample sizes. The results showed: (1) the jackknife reduced bias; (2) in 8 out of 9 cases the variance and mean square error of the jackknife estimator were less; (3) a second order jackknife reduced the bias the most but suffered a corresponding increase in variance; (4) both the first order jackknife and delta methods yielded intervals whose confidence levels were approximately equal but less than 95%.

1. Introduction

Introduction of the gel electrophoresis technique has made it possible for population geneticists to assay large samples of structural gene products in natural populations. The loci detectable by this technique are segments of DNA which code for proteins that perform enzymatic functions. A subclass of all genetic variants for a given locus will code for enzymes which can be differentiated by electrophoresis. These genetic variants for a given locus are thus called electrophoretic alleles. Allele and electrophoretic allele will be used interchangeably. Using estimates of electrophoretic allele frequencies, conservative estimates of genetic variation within and between populations can be obtained. Many problems of interest to population geneticists and evolutionary biologists require that genetic differences between populations be expressed in a single statistic. These problems include the process of speciation (Ayala 1975) and the construction of phylogenetic relationships between species (Ayala 1975, Sneath and Sokal 1973). Suggestions of appropriate statistics have not been lacking (Nei 1973). One widely used measure has been Nei's standard measure of genetic distance (Nei 1971, 1972).

Procedures for estimating Nei's distance and the sampling variance of these estimates has been described previously (Nei 1978, Nei and Roychoudhury 1974). In this paper numerical results will be presented that show Nei's standard estimator of genetic distance is biased upwards when only a small number of loci are sampled. The bias introduced when a small number of loci are sampled has been discussed previously by Nei (1973). When only a small number of individuals have been sampled at a large number of loci Nei (1978) has derived an unbiased estimator of his distance statistic. The problem of bias reduction when only a

Key Words: Jackknife; Delta method.

small number of loci have been sampled has not been studied. An alternate method of estimating Nei's distance, the jackknife method, is examined. For each estimator the bias, variance and mean square error are determined from 3 Monte Carlo studies. These properties are used as criteria for deciding which of these methods for estimating Nei's distance is best. In addition to studying the properties of these estimators of genetic distance the accuracy of confidence intervals constructed about these estimators will be examined.

2. *The Statistics*

2.1. *Nei's Distance Measure*

To define Nei's estimate of genetic distance for a given sample, let n = number of loci in the sample, m_i = number of alleles at the *ith* locus, $x_k^{(i)}$ = the frequency of the *kth* allele at locus i in population X, and $y_k^{(i)}$ = the frequency of the *kth* allele at locus i in population Y. It will be assumed that the allele frequencies are known exactly, so Nei's (1978) adjustment for bias is not needed. For the *ith* locus,

$$j_X^{(i)} = \Sigma_k[x_k^{(i)}]^2, \quad j_Y^{(i)} = \Sigma_k[y_k^{(i)}]^2, \quad j_{XY}^{(i)} = \Sigma_k[x_k^{(i)}y_k^{(i)}],$$

where the summation goes from 1 to m_i and i goes from 1 to n. The genetic distance, D, between populations X and Y is estimated by

$$\hat{D}_n = -\ln[n^{-1}\Sigma_i j_{XY}^{(i)}/(n^{-1}\Sigma_i j_X^{(i)} n^{-1}\Sigma_i j_Y^{(i)})^{1/2}] = -\ln[\bar{j}_{XY}/(\bar{j}_X\bar{j}_Y)^{1/2}]. \tag{1}$$

The genetic distance being estimated, D, is defined by (1) with the summations taken over all loci in the genome of the species being studied instead of the sample of size n. In general $E(\hat{D}_n) \neq D$, that is $\hat{D}_n$ is a biased estimator as indicated by Nei (1973).

2.2. *The Delta Method*

The delta method provides a recipe for determining approximately the expected value and variance of a random variable (or vector). This is accomplished by expanding the function in a Taylor series about the expected value of the random variable and taking the expected value of the first two terms (see Kendall and Stuart 1969, pp. 231–232, for a typical derivation). Nei and Roychoudhury (1974) have used this Taylor series approximation to obtain an estimator of $\mathrm{Var}(\hat{D}_n)$ applicable to populations having polymorphic loci which is,

$$\begin{aligned}\hat{\mathrm{Var}}(\bar{j}_X)/4\bar{j}_X^2 + \hat{\mathrm{Var}}(\bar{j}_Y)/4\bar{j}_Y^2 + \hat{\mathrm{Var}}(\bar{j}_{XY})/\bar{j}_{XY}^2 + \hat{\mathrm{Cov}}(\bar{j}_X\bar{j}_Y)/2\bar{j}_X\bar{j}_Y \\ - \hat{\mathrm{Cov}}(\bar{j}_X\bar{j}_{XY})/\bar{j}_X\bar{j}_{XY} - \hat{\mathrm{Cov}}(\bar{j}_Y\bar{j}_{XY})/\bar{j}_Y\bar{j}_{XY}.\end{aligned}$$

2.3. *The Jackknife*

For a recent review of jackknife methodology see Miller (1974). If $\hat{D}_{n,i}$ is defined as equation (1) except that the data for the *ith* locus has been deleted, n pseudovalues may be defined as $s_{n,i} = n\hat{D}_n - (n-1)\hat{D}_{n,i}$ $(i = 1, 2, \ldots, n)$. The jackknife estimator $\tilde{D}_n$ is simply the mean of the n pseudovalues

$$\tilde{D}_n = n^{-1}\Sigma_i s_{n,i} = n\hat{D}_n - n^{-1}(n-1)\Sigma_i \hat{D}_{n,i}. \tag{2}$$

Tukey (1958) suggested that the n pseudovalues be treated as approximately independent

and identically distributed random variables. The pseudovalues can then be used to define the variance of $\tilde{D}_n$ using the standard estimator

$$\hat{\text{Var}}(\tilde{D}_n) = (n^{-1})[(n-1)^{-1}\textstyle\sum_i (s_{n,i} - \tilde{D}_n)^2] .$$

If $\hat{D}_n$ is biased and its expected value is $D + a/n + b/n^2$, then the jackknife will eliminate the $1/n$ term from the bias. There is also a second order jackknife, $\tilde{D}_n^{(2)}$, such that $E(\tilde{D}_n^{(2)}) = D + O(1/n^3)$. It is defined as

$$\tilde{D}_n^{(2)} = [n^2\tilde{D}_n - (n-1)^2\textstyle\sum_i \tilde{D}_{n,i}/n]/[n^2 - (n-1)^2]$$

where $\tilde{D}_{n,i}$ is the same as (2) except that the data for the *ith* locus has been deleted.

No simple relationship between $\tilde{D}_n$ and the initial random variables $j_X^{(i)}$, $j_Y^{(i)}$, and $j_{XY}^{(i)}$ is apparent in (2). It is then quite difficult to partition the variance of $\tilde{D}_n$ into inter- and intralocus effects as Nei and Roychoudhury (1974) have done for $\text{Var}(\hat{D}_n)$. The use of (2) in theoretical studies could be cumbersome. Here attention will be restricted to data analysis only.

3. *The Monte Carlo Study*

3.1. *Simulation Procedures*

Initially, two genetic populations with n loci are defined. Ten thousand random samples of n loci are then drawn with replacement and $\hat{D}_n$ and $\tilde{D}_n$ are calculated from (1) and (2). 95% confidence intervals are constructed about each of these estimates in each sample assuming that

$$\phi(D_n) = (D_n - D)/[\text{Var}(D_n)]^{1/2}, \tag{3}$$

where D_n is either $\tilde{D}_n$ of $\hat{D}_n$ and has a t distribution with $n-1$ degrees of freedom (d.f.). The frequency with which the interval included D was tabulated and the bias, variance and mean square error of each estimator were calculated from the 10,000 observations in addition to the confidence level associated with each method.

The simulations were carried out using three sets of data to define the allelic frequencies for each of the genetic populations. Information concerning these are summarized in Table 1. From these data, the $\mathbf{j}^{(i)}$ vectors, $\mathbf{j}^{(i)} = (j_X^{(i)}, j_Y^{(i)}, j_{XY}^{(i)})'$, for the n loci randomly sampled with replacement were used to compute the Monte Carlo statistics. Given i, $\mathbf{j}^{(i)}$ was assumed known, which is akin to assuming the allelic frequencies were estimated without error. This procedure seems justified since Nei and Roychoudhury (1974) have shown that the intralocus contribution to the variance of $\hat{D}_n$ is generally quite small compared to the interlocus contribution. Although $\hat{D}_n \geq 0$ it is possible that $\tilde{D}_n < 0$, especially when a small number of loci are sampled. Since negative distance values have no logical interpretation during the course of the simulation whenever $\tilde{D}_n$ was negative it was set equal to 0. This procedure decreased both the bias and variance of $\tilde{D}_n$.

Samples from the data of Avise and Ayala frequently produced $\bar{j}_{XY} = 0$, especially when n was small. Because $\tilde{D}_n$ and $\hat{D}_n$ are undefined when all $j_{XY}^{(i)} = 0$ the following modification of these data were made. For each locus where $j_{XY}^{(i)} = 0$, $x_k^{(i)}$ was assigned the value 10^{-6}, where k is the most frequent allele in population Y. Since only 5 significant digits of $j_X^{(i)}$, $j_Y^{(i)}$ and $j_{XY}^{(i)}$ were recorded, this modification of the data had no effect on the values of $j_X^{(i)}$ or $j_Y^{(i)}$, but did produce non-zero $j_{XY}^{(i)}$ so that $\tilde{D}_n$ and $\hat{D}_n$ could be defined. D for this altered set of data was decreased by only 2×10^{-4} distance units.

TABLE 1
Summary of Data Used in the Monte Carlo Studies

Populations	Frequency of Polymorphic Loci† I	II	Number of Loci Sampled	D	Reference
Haupt Creek (I) and Upper Skaggs Springs (II) populations of the red bellied newt *Taricha rivularis*	0.17	0.20	35	0.0157	Hedgecock (1978)
Venezuelan populations of *Drosophila tropicalis*(I) and *D. equinoxialis*(II)	0.38	0.45	29	0.499	Ayala *et al* (1974)
California minnows *Pogonicthys macrolepidotus*(I) and *Notemigonus crysoleucus*(II)	0.087	0.26	23	1.08	Avise and Ayala (1976)

† A locus is considered polymorphic if the frequency of the most common allele is ≤ 0.95.

3.2. *Results*

A summary of the results are given in Tables 2, 3 and 4. One run calculating the second order jackknife was carried out with the data of Ayala, Tracey, Barr, McDonald and Perez-Salas (1974) using $n = 5$. The main advantage of the second order jackknife is its potential bias-reducing properties. This potential was realized but was coupled to a correspondingly high increase in variance. Because of this undesirable effect and the large number of computations necessary, the second order jackknife was not run for larger sample sizes or different data.

The simulation provided a large but finite set of data from which to estimate the bias, confidence level, etc. Therefore confidence intervals have been placed on most of these estimates in Tables 2, 3 and 4 to support conclusions drawn in the discussion.

The estimate of the bias of each estimator was obtained as the deviation of the mean value of each from the parametric value of D (Table 1). The variance of this estimated bias

TABLE 2
Monte Carlo Results: D for Hedgecock's (1978) Data = 0.0157
$\tilde{D}_n$ = the Jackknife Estimator, $\hat{D}_n$ = the Delta Estimator

Estimator	Expected % bias†	% Bias ± 95% C.I.	Variance of Estimator††,*	95% C.I. Estimate of Variance††	M.S.E. of Estimator††	Confidence Level ± 95% C.I., *t*-dist.
$\tilde{D}_5$		0.0072 ± 0.04	42.6	(41.4, 43.8)	42.6	72.0 ± 1.0
$\hat{D}_5$	6.6	7.95 ± 0.044	49.8	(48.4, 51.2)	50.0	70.9 ± 1.0
$\tilde{D}_{15}$		0.122 ± 0.023	13.5	(13.1, 13.9)	13.5	78.9 ± 0.80
$\hat{D}_{15}$	2.2	2.2 ± 0.023	14.2	(13.8, 14.6)	14.3	78.9 ± 0.80
$\tilde{D}_{30}$		0.135 ± 0.016	6.98	(6.79, 7.18)	6.98	85.7 ± 0.69
$\hat{D}_{30}$	1.1	1.26 ± 0.017	7.15	(6.96, 7.35)	7.16	86.0 ± 0.68

* Estimated from the 10,000 values of $\tilde{D}_n$ and $\hat{D}_n$.
† Based on numerical evaluation of (4).
†† In units of 10^{-5}.

TABLE 3

Monte Carlo Results: D for Ayala *et al.*'s (1974) Data = 0.499, $\tilde{D}_n$ = the Jackknife Estimator, $\hat{D}_n$ = the Estimator, $\tilde{D}_n^{(2)}$ = the Second Order Jackknife Estimator

Estimator	Expected % bias†	% Bias ± 95% C.I.	Variance of Estimator††	95% C.I. Estimate of Variance	M.S.E. of Estimator	Confidence level ± 95% C.I., t-dist.
$\tilde{D}_5$		17.4 ± 0.94	0.229	(0.223 , 0.236)	0.237	88.8 ± 0.62
$\hat{D}_5$	15	23.6 ± 1.1	0.334	(0.325 , 0.343)	0.347	89.2 ± 0.61
$\tilde{D}_n^{(2)}$		5.8 ± 5.8	2.16	(2.10 , 2.22)	2.16	—
$\tilde{D}_{15}$		0.066 ± 0.80	0.0416	(0.0405, 0.0428)	0.0416	93.7 ± 0.48
$\hat{D}_{15}$	4.8	5.7 ± 0.85	0.0490	(0.0477, 0.0504)	0.0498	94.3 ± 0.45
$\tilde{D}_{30}$		0.17 ± 0.56	0.0204	(0.0198, 0.0210)	0.0204	94.7 ± 0.44
$\hat{D}_{30}$	2.4	2.7 ± 0.58	0.0218	(0.0212, 0.0224)	0.0220	95.1 ± 0.42

† Based on numerical evaluation of (4).
†† Estimated from the 10,000 values of $\tilde{D}_n$ and $\hat{D}_n$.

was also estimated and used in calculating a confidence interval for the % bias as shown in Tables 2, 3 and 4 and equal tail confidence intervals were calculated for the variance.

During the simulation the computer kept track of the number of times, x, that each calculated confidence interval using $\tilde{D}_n$ or $\hat{D}_n$ included D. Obviously x has a binomial variance $[(x/10{,}000)(1 - (x/10{,}000))/10{,}000]$. To further examine the distribution properties of $\phi(D_n)$ the frequency of $|\phi(D_n)| > 2.5$ in several runs is presented along with that expected assuming $\phi(D_n) \sim t_{n-1}$ in Table 5.

4. *Discussion*

The major difficulty in obtaining a best estimator of D is that standard solutions such as maximum likelihood estimators are not available since nothing of sufficient accuracy can be

TABLE 4

Monte Carlo Results: D for Avise and Ayala's (1976) Data = 1.08, $\tilde{D}_n$ = the Jackknife Estimator, $\hat{D}_n$ = the Delta Estimator

Estimator	Expected % Bias†	% Bias ± 95% C.I.	Variance of Estimator††	95% C.I. Estimate of Variance	M.S.E. of Estimator	Confidence level ± 95% C.I., t-dist.
$\tilde{D}_5$		86.4 ± 8.59	1.92 × 10	(18.7 , 19.7)	20.1	84.3 ± 0.72
$\hat{D}_5$	18.9	143 ± 8.18	1.74 × 10	(16.9 , 17.9)	19.8	84.7 ± 0.71
$\tilde{D}_{15}$		3.35 ± 1.27	0.417	(0.406 , 0.429)	0.419	95.4 ± 0.41
$\hat{D}_{15}$	6.3	9.32 ± 1.36	0.480	(0.467 , 0.494)	0.490	95.7 ± 0.40
$\tilde{D}_{30}$		0.33 ± 0.50	0.0657	(0.0639, 0.0676)	0.0657	95.6 ± 0.40
$\hat{D}_{30}$	3.2	3.32 ± 0.54	0.0762	(0.0741, 0.0784)	0.0774	95.8 ± 0.39

† Based on numerical evaluation of (4).
†† Estimated from the 10,000 values of $\tilde{D}_n$ and $\hat{D}_n$.

TABLE 5
Observed and Expected Frequency of $|\phi(D_n)| > 2.5$ for Simulation Studies of Hedgecock's (1978) and Avise and Ayala's (1976) Data

	Estimator					
	$\hat{D}_5$	$\tilde{D}_5$	$\hat{D}_{15}$	$\tilde{D}_{15}$	$\hat{D}_{30}$	$\tilde{D}_{30}$
Observed (Hedgecock)	0.294	0.369	0.208	0.206	0.111	0.115
Observed (Avise and Ayala)	0.157	0.158	0.0295	0.0338	0.0196	0.0232
Expected	0.074		0.0257		0.0198	

said about the distribution of the random vector $(j_X^{(i)}, j_Y^{(i)}, j_{XY}^{(i)})$. The two alternative methods considered here both suffer from relatively large variances. This problem has been emphasized previously by Nei (Li and Nei 1975, Nei 1975, Nei and Roychoudhury 1974). Nei stressed the importance of studying a large number of loci. There are, however, many studies of natural populations where less than 30 loci have been sampled and very few with more than 40 studied.

$\hat{D}_n$ also has the drawback that it will be biased when a small number of loci are sampled. The approximate magnitude of this bias is given by

$$(1/2n)[\mathrm{Var}(j_{XY})/J_{XY}^2 - \tfrac{1}{2}(\mathrm{Var}(j_X)/J_X^2 + \mathrm{Var}(j_Y)/J_Y^2)] . \quad (4)$$

Equation (4) was derived from the expected value of the first 3 terms of a Taylor series expansion of the function $\hat{D}_{XY}$ about the point

$$(E(\bar{j}_X), E(\bar{j}_Y), E(\bar{j}_{XY})) = (J_X, J_Y, J_{XY}) .$$

In Tables 2, 3 and 4 the expected bias from equation (4) is given in addition to the observed bias in the simulations. In all cases $\hat{D}_n$ was biased upwards. It can be seen that equation (4) gives a good estimate of the magnitude of the bias for sample sizes of 15 and 30 but is rather poor for a sample size of 5. For small sample sizes third and higher order terms in the Taylor series are of sufficient magnitude that equation (4) is no longer a good approximation of the bias. The magnitude of the expected and the observed bias increase with increasing D. This result is in accord with equation (4). Large values of D imply small values of J_{XY} and the first term of equation (4) is inversely proportional to J_{XY}^2.

Examination of Tables 2, 3 and 4 show the jackknife to be quite effective at reducing the bias. In all cases with $n \geq 15$ $\tilde{D}_n$ has either no detectable bias or it is <0.1%. One exception is $\tilde{D}_{15}$ in Table 4. In this case the bias of the jackknife estimator is still $\frac{1}{3}$ that of the delta estimator.

In all but one case the variance and the mean square error of $\tilde{D}_n$ is less than $\hat{D}_n$. The jackknife's reduction in variance is not as dramatic as the bias reduction. For instance, the variance of $\tilde{D}_{30}$ is only 2% less than the variance of $\hat{D}_{30}$ for Hedgecock's data, 6% less for Ayala's data and 14% less for Avise's data. While these are modest reductions in variance this fact, coupled with the substantial bias reduction of $\tilde{D}_n$, make it the superior estimator. These desirable properties seem to hold over a wide range of sample sizes and values of D.

The last problem considered is the estimation of confidence intervals about $\tilde{D}_n$ and $\hat{D}_n$. There are no detectable differences between the confidence levels of intervals generated around $\tilde{D}_n$ and $\hat{D}_n$. Most intervals in Tables 2, 3 and 4 have confidence levels <95%. This indicates that the assumptions that allow one to infer that equation (3) has a t distribution

with $n - 1$ d.f. are not entirely correct. One assumption that does not hold is that $\tilde{D}_n$ and $\hat{D}_n$ are unbiased estimators of D. As n gets larger the confidence levels associated with each method approach 95%. This is due, at least partly, to the fact that for large n $E(\tilde{D}_n)$, $E(\hat{D}_n) \rightarrow D$.

An indication of how well $\tilde{D}_n$ and $\hat{D}_n$ are described by a t distribution is given in Table 5. One striking result is the large discrepancy between the expected and the observed values for Hedgecock's data even when $n = 30$. This problem occurs because the distribution of D_n is not symmetric but is truncated at 0. When D is small this truncation causes the distribution of $\phi(D_n)$ to deviate substantially from a t distribution. Potential solutions to this problem are currently under investigation.

Acknowledgments

I would like to thank Norman Matloff for introducing the jackknife technique to me and Francisco Ayala, Todd Bierbaum, Michael Turelli, and an anonymous referee for numerous comments that greatly improved the paper. This work was supported in part by Contract PA 200-14 Mod #4 with ERDA and by NIH Grant 1 PO1 GM22221, as well as by NIH training grant PHS 1 T32 GM07467-1. Computer time was generously provided by the U.C.D. Computer Center.

References

Avise, J. C. and Ayala, F. J. (1976). Genetic differentiation in speciose versus depauperate phylads: evidence from the California minnows. *Evolution 30,* 46–58.

Ayala, F. J. (1975). Genetic differentiation during the speciation process. In *Evolutionary Biology, Vol. 8.* T. Dobzhansky, M. K. Hecht and W. C. Steere (eds.). Plenum Press, New York, 1–78.

Ayala, F. J., Tracey, M. L., Barr, L. G., McDonald, J. F. and Perez-Salas, S. (1974). Genetic variation in natural populations of five *Drosophila* species and the hypothesis of the selective neutrality of protein polymorphisms. *Genetics 77,* 343–384.

Hedgecock, D. (1978). Population subdivisions and genetic divergence in the red-bellied newt, *Taricha rivularis. Evolution 32,* 271–286.

Kendall, M. G. and Stuart, A. (1969). *The Advanced Theory of Statistics,* Vol. 1. Hafner, New York.

Li, W. H. and Nei, M. (1975). Drift variances of heterozygosity and genetic distance. *Genetical Research 25,* 229–248.

Miller, R. G. (1974). The jackknife—a review. *Biometrika 61,* 1–15.

Nei, M. (1971). Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. *American Naturalist 105,* 385–398.

Nei, M. (1972) Genetic distance between populations. *American Naturalist 106,* 283–292.

Nei, M. (1973). The theory and estimation of genetic distance. In *Genetic Structure of Populations.* N. E. Morten (ed.), University Press of Hawaii, Honolulu, 45–54.

Nei, M. (1975). Mathematical models of speciation and genetic distance. In *Population Genetics and Ecology.* S. Karlin and E. Nevo (eds.), Academic Press, New York, 723–765.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics 89,* 583–590.

Nei, M. and Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distances. *Genetics 76,* 379–390.

Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy.* W. H. Freeman and Co., San Francisco, California.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (Abstract). *Annals of Mathematical Statistics 29,* 614.

Received June 1978; *Revised February and May* 1979