The Statistical Analysis of Mitochondrial DNA Polymorphisms: χ^2 and the Problem of Small Samples[1]

Derek A. Roff and Paul Bentzen[2]
Department of Biology, McGill University

Significance levels obtained from a χ^2 contingency test are suspect when sample sizes are small. Traditionally this has meant that data must be combined. However, such an approach may obscure heterogeneity and hence potentially reduce the power of the statistical test. In this paper, we present a Monte Carlo solution to this problem: by this method, no lumping of data is required, and the accuracy of the estimate of α (i.e., a type 1 error) depends only on the number of randomizations of the original data set. We illustrate this technique with data from mtDNA studies, where numerous genotypes are often observed and sample sizes are relatively small.

Introduction

Small sample size frequently makes the analysis of geographic or temporal variation in isozyme or mtDNA frequencies problematic. This is particularly true in mtDNA studies, where restriction-enzyme surveys often reveal many relatively rare variants. For large samples, a χ^2 contingency test is appropriate. However, if the expected values within cells are very small, the calculated χ^2 may be inflated upward, and hence one cannot make use of the usual tabulated values of χ^2 to assess the significance of the observed value. Cochran (1954) suggested that a reasonable rule of thumb is that no expected frequency should be <1.0 and that $\leq 20\%$ of the expected frequencies should be <5.0. When data do not conform to these criteria, the general solution is to group data, most often by considering only the relative frequency of the most common type. Such a procedure necessarily loses information and is to be avoided if possible.

An alternate approach is to use Fisher's exact permutation test. This test suffers from the problem that the number of permutations of the data matrix increases factorially with the number of rows, columns, and total sample size. As a consequence, the number of required permutations increases astronomically with even modest increases in either the number of cells or sample size. For this reason the method is generally used only for 2×2 tables. Pagano and Halvorsen (1981) proposed for this method an algorithm that does not increase as fast as the complete enumeration method, although it still increases in a nonlinear fashion with the number of rows, columns, and total sample size. mtDNA data sets may comprise >100 cells and sample sizes >100 observations. For such sets neither the complete enumeration method nor the method of Pagano and Halvorsen (1981) is computationally feasible, particularly on a microcomputer.

1. Key words: mitochondrial DNA, χ^2, statistical analysis, Monte Carlo method.

2. Present address: Department of Biology, Dalousie University, Halifax, Nova Scotia, Canada, B3H 4J1.

Address for correspondence and reprints: Department of Biology, McGill University, 1205 Dr. Penfield Avenue, Montreal, Quebec, Canada, H3A 1B1.

Mol. Biol. Evol. 6(5):539–545. 1989.

0737-4038/89/0605-0009$02.00

The solution to this general problem is to generate, using a Monte Carlo technique, the distribution of χ^2 expected if the null hypothesis (of homogeneity) were true for the particular data set under study. The exact probability of obtaining a value of χ^2 as large as or larger than that observed can be obtained by considering all possible arrangements of the data set, subject, of course, to the constraint that row and column totals remain constant. As discussed above, for all but the smallest data sets, this approach is impractical; however, this probability can be approximated as closely as desired by a Monte Carlo approach. In the present paper we present a computer method for generating such distributions and illustrate the utility of the method with mtDNA data drawn from the literature.

The Algorithm

To obtain the expected distribution of χ^2, a large number of randomizations of the data set must be generated, subject to the constraint that row and column totals remain equal to those of the original data. Since the expected distribution of this statistic may not in fact conform to that of the χ^2, we shall designate the statistic by X^2, the observed value by X_o^2, and that for each randomized set by X_r^2. The estimated probability, P, of obtaining a value of X^2 as large as or larger than that obtained from the actual observations (X_o^2) is given by $\hat{P} = n/N$, where n is the number of randomizations that generate $X_r^2 > X_o^2$ and where N is the total number of randomized sets. The standard error of $\hat{P}$ is $\sqrt{[\hat{P}(1 - \hat{P})/N]}$.

Let the number of rows and columns in the original data matrix be R and C, respectively, and let the total number of observations be M. Without loss of generality, we shall assume that the original data matrix is constructed such that the rows correspond to sites and the columns to genotypes. Marginal column (= genotype) totals are designated by $c(1)$, $c(2)$, $c(3)$, . . . , $c(C)$, and marginal row (= site) totals are designated by $r(1)$, $r(2)$, $r(3)$, . . . , $r(R)$. To produce a randomization of the data matrix, we construct an $M \times 2$ matrix as follows: in column 1 are placed M random numbers, and column 2 is filled with $c(1)$ ones, $c(2)$ twos, $c(3)$ threes, etc. Column 1 is now sorted into ascending (or descending) rank, moving the value in column 2 along with the number in column 1. The randomized number of each genotype at site 1 is obtained by computing the number of ones, twos, threes, etc. in the first $r(1)$ rows of the second column; similarly, the numbers of each genotype at site 2 are obtained by considering the number of ones, twos, threes, etc. in rows $r(1) + 1$ to $r(1) + r(2)$. The same procedure is repeated for all sites, resulting in a new data matrix that is randomized but in which the marginal totals equal those of the original matrix. A new randomization is produced by generating a new sequence of M random numbers. The above procedure is illustrated with the following hypothetical data set comprising two genotypes sampled at two sites (such a data set could be analyzed using the Fisher exact test and is used here only for simplicity of demonstration):

	Genotype		Total
	1	2	3
Site	2	0	2
Total	3	2	5

The $M \times 2$ matrices, before and after sorting, are

$$\begin{bmatrix} 2017 & 1 \\ 7449 & 2 \\ 9470 & 2 \\ 2215 & 1 \\ 9329 & 1 \end{bmatrix} \begin{bmatrix} 2017 & 1 \\ 2215 & 1 \\ 7449 & 2 \\ 9329 & 1 \\ 9470 & 2 \end{bmatrix}.$$

Note that the order in which the $c(1)$ ones, $c(2)$ twos, etc. are placed initially in the above matrix is immaterial. Collecting genotypes according to site, we obtain from the sorted matrix the final randomized matrix,

	Genotype		Total
Site	2	1	3
	1	1	2
Total	3	2	5

We illustrate the above procedure by using data drawn from five separate studies of mtDNA, using in each case 1,000 randomizations of the original data sets.

Examples

Data from Bentzen et al. (1988)

From 14 separate rivers Bentzen et al. (1987) identified 10 different mtDNA genotypes of American shad (*Alosa sapidissima*). The total number of fish sampled was 244, many genotypes being at very low frequency. Of the 140 cells, 92 (66%) have expected values <1.0 and only 13 (9.3%) have expected values >5.0. Thus, to make use of tabulated χ^2, it is necessary to combine cells. Only the most frequent genotype is represented at reasonably high frequency in all samples, and hence, following the usual procedure, we combined all the rare genotypes. Now no cells have an expected value <1.0, and <20% (17.8%) have expected values <5.0.

The calculated χ^2 for the combined data set is 22.96, which just exceeds the critical value of χ^2 (22.36) at the 5% level. The estimated value of X_o^2 for the uncombined data is 236.5, which is highly significant ($P < 0.001$) based on the tabulated values of χ^2. However, this result is suspect because of the very low frequencies within many cells. But none of the X_r^2 values obtained from the 1,000 randomizations of the data set exceed X_o^2, the largest value obtained being only 166.2. Thus, we conclude that, far from being marginally significant, the heterogeneity among samples is highly significant (probability $X_r^2 > X_o^2 \approx 0/1{,}000$, i.e., $P < 0.001$). The cumulative frequency distribution of X_r^2 matches very closely the cumulative frequency distribution of χ^2 (fig. 1a).

Data from Edwards and Skibinski (1987)

Although designated as separate species, the common mussel (*Mytilus edulis*) and the Mediterranean mussel (*M. galloprovincialis*) hybridize and introgress extensively in parts of Britain and Ireland (Skibinski et al. 1983). On the basis of mtDNA

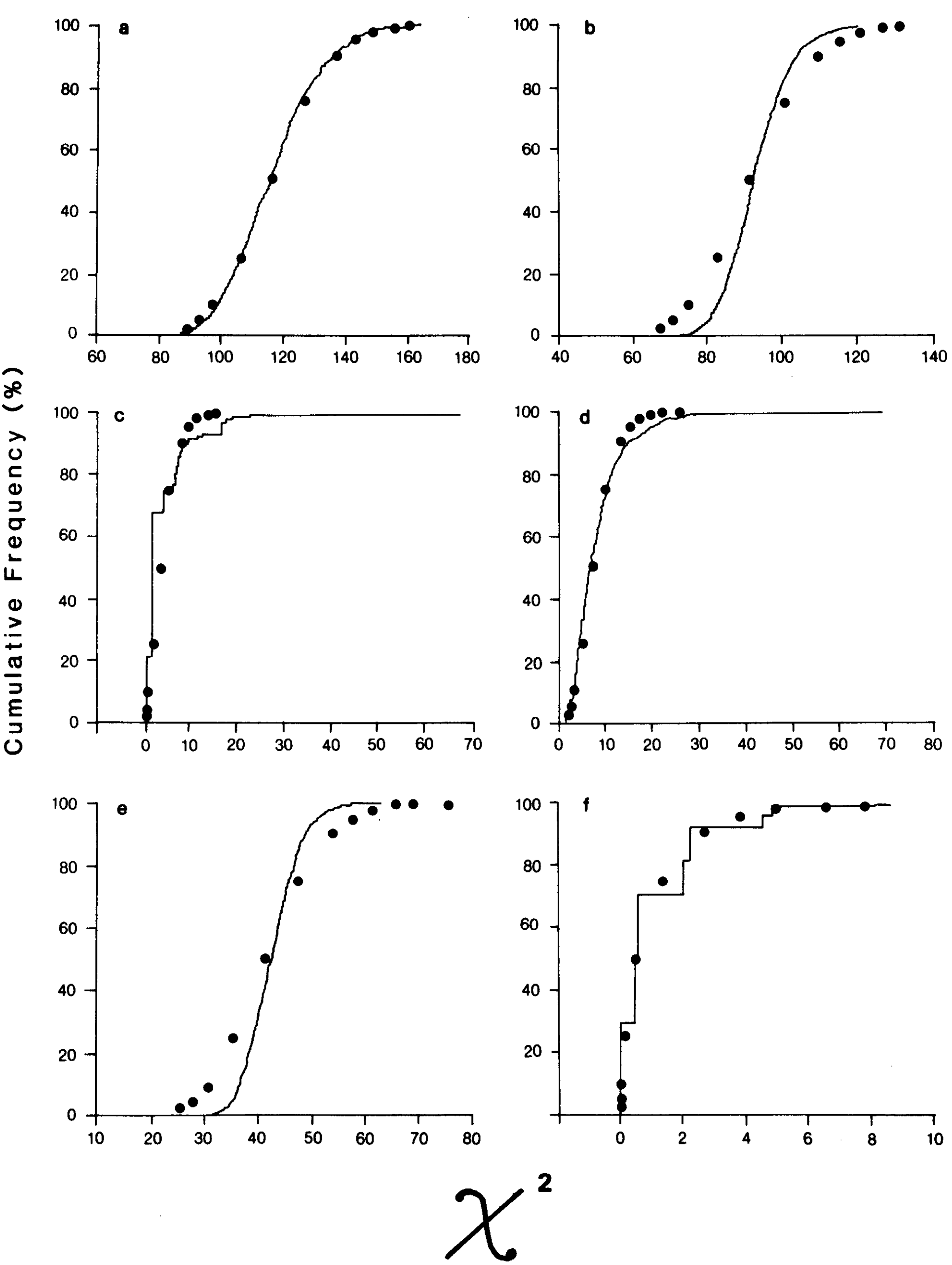

FIG. 1.—Cumulative distribution of X_r^2 (solid line). Solid dots indicate the tabulated value of χ^2 from Zar (1984). Original data sets are from (a) Bentzen et al. (1988), (b) Edwards and Skibinski (1987), (c) and (d) DeSalle et al. (1987), (e) Kornfield and Bogdanowicz (1987), and (f) Avise et al. (1987).

analysis, Edwards and Skibinski (1987) found 24 separate genotypes across five different localities.

In this data set, 75% (90 of 120 cells) of the expected values are <1.0, thus making a comparison of the estimated value of χ^2 (X_o^2) with the appropriate tabulated value theoretically unsound. However, as with the previous data, none of the 1,000 randomizations produced an X^2 value greater than that obtained from the original data (144.7). Thus we conclude that there is significant heterogeneity among samples at $P < 0.001$. The median values of X_r^2 and χ^2 are virtually identical, but the variance of X_r^2 is smaller than χ^2 (fig. 1b). The probability level obtained from the statistical tables is also $P < 0.001$.

Data from DeSalle et al. (1987)

DeSalle et al. (1987) isolated three mtDNA genotypes of *Drosophila mercatorum* distributed among five sites. Because of low sample size they first transformed their data by

$$a_i = [\arcsin \sqrt{(x_i/(n_i + 1)} + \arcsin \sqrt{(x_i + 1)/(n_i + 1)}]/2 ,$$

where n_i are the number of haploid genomes sampled at site i and where x_i are the number that show the variant type of interest.

A statistic, V, was then computed from

$$V = 4 \sum_{i=1}^{r} n_i(a_i - \bar{a})^2 ,$$

where r is the number of sites, $\bar{a} = \sum_{i=1}^{r} n_i a_i/N$, and $N = \sum_{i=1}^{r} n_i$. Under the null hypothesis of no heterogeneity, V is distributed as a χ^2 with $r - 1$ degrees of freedom. There are two problems with this approach. First, the predicted distribution of V depends on the asymptotic normality of a_i, which cannot be relied on when sample sizes are small. Second, only two types (type x and "the rest") can be analyzed at one time. This means that multiple tests are required, which increases the probability of obtaining a significant result by chance alone. The method proposed in the present paper circumvents these problems.

For the purpose of the present analysis, we selected data taken in 1983 (DeSalle et al. 1987, table 2). Three genotypes over five sites are represented. To compare the present method with that of DeSalle et al., we first combined the data for two genotypes and considered type *Bst*NI and the rest. The type *Bst*NI was represented by only three individuals in one site. Of the 10 cells, four contain expected values <1.0, and five (50%) contain expected values <5.0. Therefore, a χ^2 test using the tabulated values is not legitimate. The test proposed by DeSalle et al. (1987) gives a significant probability at $P < 0.025$. Application of the randomization procedure gives $P < 0.0048$ ($\hat{P}$ + 2 SE), only two replicates exceeding the observed value of 66.6. The test of DeSalle et al. (1987) is clearly very conservative and considerably underestimates the level of significance. Depsite the large fraction of small values in the data set, the cumulative distributions of X_r^2 and χ^2 are very similar (fig. 1c), the randomization procedure producing a slight excess of large values compared with that expected by χ^2. The probability obtained from the tabulated values of χ^2 is $P < 0.001$.

We next applied the randomization procedure to all three genotypes, a test not possible with the approach of DeSalle et al. (1987). No values of X_r^2 exceeded X_o^2 (95.8), and hence we conclude that for the complete data set $P < 0.001$. As with the earlier test, the cumulative distributions of X_r^2 and χ^2 are very similar, X_r^2 being slightly skewed and containing a few relatively large values (fig. 1d).

Data from Kornfield and Bogdanowicz (1987)

Among three different geographic sites, Kornfield and Bogdanowicz (1987) distinguished 26 different genotypes of Atlantic herring (*Clupea harengus*). One site was sampled in two consecutive years; for the present analysis we selected the year with the largest sample, which reduced the number of observed genotypes to 22.

The total number of fish sampled was only 61, and hence most genotypes were represented by only one or two individuals (though the marginal totals by genotype were >1 for seven of the genotypes). Of the 66 cells, 52 contain expected values <1.0 and only one contains an expected value >5.0. Obviously the usual χ^2 test is not recommended for these data. The randomization procedure generated 713 values of X^2 that exceeded the observed value of 39.6. The probability of observing a value of $X^2 = 39.6$ by chance alone is thus 0.713 ± 0.029 (± 2 SE), and hence the null hypothesis of no significant geographic variation cannot be rejected.

As with the data from Edwards and Skibinski (1987), the distribution of X_r^2 is less variable than that of χ^2, though the medians are very similar (fig. 1e; the lack of variation is probably due to the large number of marginal totals that sum to 1). From the tabulated values, the probability of obtaining a value of 39.6 by chance alone is $0.75 < P < 0.5$.

Data from Avise et al. (1987)

Avise et al. (1987) examined mtDNA variation in Atlantic and Gulf Coast populations of the hardheaded catfish (*Arius felis*). Because of extremely low sample sizes, they combined different genotypes into two clusters and combined different geographic sites into Gulf or Atlantic categories (i.e., two "genotypes" and two "sites"). Even in this radical lumping, one of the four cells contained only three fish and two of the cells contained expected values <5. For this reason Avise et al. (1987) computed the G statistic with correction for continuity.

The probability obtained using the randomization method is $P = 0.289 \pm 0.029$ (± 2 SE), and from the tabulated values $0.25 < P < 0.1$. Because of the low sample sizes and restricted number of cells, the cumulative distribution of X_r^2 is distinctly stepped (fig. 1f), but it closely corresponds to the cumulative distribution of χ^2.

Discussion

Although the cumulative distributions of X^2 and χ^2 did not always correspond, the difference was in no case sufficiently great to lead to an incorrect statistical decision if the tabulated values of χ^2 were used. Of particular importance was that, in one case (American shad; Bentzen et al. 1988), lumping the data to increase the expected cell frequencies greatly decreased the level of significance. Similarly, the approximate method adopted by DeSalle et al. (1987) can lead to an underestimate of the level of significance.

These results indicate that the significance level of an estimated χ^2 value for small sample sizes may frequently be gauged reasonably accurately from the tabulated values. However, since it is also clear that the distributions are not identical, the actual level should be estimated by the randomization procedure outlined in the present paper. This is particularly critical for values that are close to the 0.05 level.

mtDNA is believed to be a highly sensitive indicator of population divisions (Birky et al. 1983). As the examples above illustrate, the method described here can help realize this potential by making maximal use of small data sets divided among relatively numerous genotypes.

The general procedure can also be extended beyond its use for χ^2 and be used for other types of analysis, such as regression or one-way ANOVA. The advantage of the method is that no assumptions are made concerning the underlying distribution.

Acknowledgments

We are most grateful for the helpful suggestions of Drs. S. Stewart and W. M. Fitch and by two anonymous referees. This work was supported by a grant to D.A.R. from the National Science and Engineering Research Council of Canada.

LITERATURE CITED

AVISE, J. C., C. A. REEB, and N. C. SANDERS. 1987. Geographic population and species differences in mitochondrial DNA of mouthbrooding marine catfishes (*Ariidae*) and demersal spawning toadfishes (*Batrachoididae*). Evolution **41**:991–1002.

BENTZEN, P., W. C. LEGGETT, and C. G. BROWN. 1988. Length and restriction site heteroplasmy in the mitochondrial DNA of American shad (*Alosa Sapidissima*). Genetics **118**:509–518.

BIRKY, C. W. JR., T. MARUYAMA, and P. FUERST. 1983. An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts and some results. Genetics **103**:513–527.

COCHRAN, W. G. 1954. Some methods for strengthening the common χ^2 tests. Biometrics **10**: 417–451.

DESALLE, R., A. TEMPLETON, I. MORI, S. PLETSCHER, and J. S. JOHNSTON. 1987. Temporal and spatial heterogeneity of mtDNA polymorphism in natural populations of *Drosophila mercatorum.* Genetics **116**:215–223.

EDWARDS, C. A., and D. O. F. SKIBINSKI. 1987. Genetic variation of mitochondrial DNA in mussel (*Mytilus edulis* and *M. galloprovincialis*) populations from southwest England and south Wales. Mar. Biol. **94**:547–556.

KORNFIELD, I., and S. M. BOGDANOWICZ 1987. Differentiation of mitochondrial DNA in Atlantic herring, *Clupea harengus.* Fisheries Bull. **85**:561–568.

PAGANO, M., and K. T. HALVORSEN. 1981. An algorithm for finding the exact significance levels of r × c contingency tables. J. Am. Stat. Assoc. **76**:931–934.

SKIBINSKI, D. O. F., J. A. BEARDMORE, and T. F. CROSS. 1983. Aspects of the population genetics of *Mytilus* (Mytilidae: Mollusca) in the British Isles. Biol. J. Linnaean Soc. **19**:137–183.

ZAR, J. H. 1984. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, N.J.

WALTER M. FITCH, reviewing editor

Received November 1, 1988; final revision received April 21, 1989

Accepted April 21, 1989