Marine Biology 112, 517–523 (1992)

Marine Biology
International Journal on Life in Oceans and Coastal Waters
© Springer-Verlag 1992

# Testing electrophoretic data for agreement with Hardy-Weinberg expectations

**H. A. Lessios**

Smithsonian Tropical Research Institute, P.O. Box 2072, Balboa, Panamá

Date of final manuscript acceptance: November 8, 1991. Communicated by J. M. Lawrence, Tampa

**Abstract.** Electrophoretic data from marine organisms are routinely tested for conformity to expectations of the Hardy-Weinberg rule, but the statistical procedures used and the interpretation of their results are often flawed. This paper summarizes literature on statistical testing for Hardy-Weinberg equilibrium and suggests an analytical strategy based on carrying out computationally simple goodness-of-fit $\chi^2$ tests (with pooling and correction factors for continuity if necessary) when appropriate, and resorting to computationally tedious, exact tests when necessary. It recommends adjustments of significance levels to avoid the large Type-I error that may result from multiple tests for Hardy-Weinberg equilibrium, one for each locus and each population. It points out the obvious but common error of interpreting non-significant tests as evidence of conformity to Hardy-Weinberg expectations, and makes suggestions as to how tests that produce significance can be used to reach conclusions of biological relevance.

## Introduction

Electrophoretic surveys commonly test genotype frequencies statistically to see if they conform to Hardy-Weinberg (HW) expectations. Although such tests can be useful, they are frequently done erroneously, they routinely have the wrong significance values applied to them, and they are often misinterpreted. One has the impression from the literature that most investigators proceed with what is believed to be standard methodology, but that they are sufficiently aware of its problems to dismiss results from comparisons of their data to HW expectations with little or no comment. This practice seems wasteful; either the tests should be performed correctly and relied upon to provide information relevant to the study, or they should be bypassed entirely. This paper identifies some of the pitfalls associated with comparisons of observed genotype frequencies to those expected from HW equilibrium as they are commonly done, and provides a set of guidelines as to how such comparisons can be performed and interpreted. Specifically, this paper deals with the questions: (1) What tests should be employed for the comparisons? (2) How should the existence of rare genotypes be treated? (3) How should significance levels be adjusted for multiple tests? (4) How should the results of statistical testing be interpreted?

## Hardy-Weinberg equilibrium

The HW rule states that in a population in which gametes associate entirely at random, genotype frequencies within one generation will follow the multinomial distribution, with allele frequencies as the distribution parameters. HW equilibrium is expected when (a) the entities for which it is calculated segregate in a Mendelian fashion (i.e., are alleles at one locus), (b) alleles are co-dominant, so that all heterozygotes can be recognized, (c) the locus is autosomal, (d) gene frequencies are the same in the two sexes, (e) all genotypes are selectively equivalent, (f) all reproduction is sexual, (g) mating is random throughout the population, (h) no mutations occur, (i) no migration takes place, and (j) populations are infinitely large. As no natural population conforms exactly to all of these assumptions, no natural population is exactly at HW equilibrium (Smith 1970). The question is how well genotype frequencies at every locus in each population approach this ideal, and the object of comparing observed genotype frequencies with those expected from the rule is to find out whether one or more of the preconditions necessary for HW equilibrium are violated.

It would appear simple to carry out a goodness-of-fit test to see how well observed genotype frequencies conform to expected ones. However, the existence of a large body of literature that presents new tests or modifications to traditional ones in HW comparisons (Dobzhansky and Levene 1948, Levene 1949, Haldane 1954, Li 1955, p. 13, Cannings and Edwards 1969, Brown 1970, Smith 1970, Mantel and Li 1974, Chapco 1976, Elston and Forthofer 1977, Emigh 1980, Haber 1981, Louis and

Dempster 1987, Lindley 1988, Hernandez and Weir 1989, Weir 1990, p. 74), compares tests against each other (Emigh and Kempthorne 1975, Elston and Folthofer 1977, Emigh 1980, Haber 1981, Hernandez and Weir 1989), or points out mistakes in the manner in which previous tests have been applied (Dobzhansky and Levene 1948, Crisp et al. 1978, Pamilo and Varvio-Aho 1984) belies this apparent simplicity. The complications arise in part from the lack of independence of observed and expected genotype frequencies, in part from the discrete nature of Mendelian inheritance, but most often from the properties of natural polymorphisms, which often include alleles in low frequency. An additional problem, identified by Cooper (1968) and by Rice (1989), but ignored in practically all studies, comes from the fact that electrophoretic surveys usually include many loci, and carry out many simultaneous tests, thus greatly increasing the probability of Type-I error.

## Procedures for testing for fit to Hardy-Weinberg expectations

Tests used to compare observed genotype frequencies to those expected from HW equilibrium fall into three general categories: (a) exact tests, (b) goodness-of-fit $\chi^2$ tests, and (c) likelihood-ratio tests and log-linear models.

Exact tests calculate the probability that the observed sample (and others less probable than the observed one) could be drawn from the population by chance if the null hypothesis held true. An exact test for HW ratios in a locus with two alleles was proposed by Haldane (1954) and extended for multiple alleles by Emigh (1980) and by Louis and Dempster (1987). Chapco (1976) offered a slightly different procedure based on a model of Cannings and Edwards (1969). This model has been criticized by Emigh and Kempthorne (1975) as logically unsound. The advantage of exact tests is that they are not adversely affected by small expected values, and thus may be the only valid tests when sample sizes are small and some alleles are rare. Their disadvantage is that they require a prodigious amount of computation in all but the simplest cases.

Goodness-of-fit $\chi^2$ tests have intuitive appeal for comparisons between observed and expected distributions, they are directly related to Wright's (1969, p. 174) fixation index, they are computationally simple, and they have been exhaustively studied by statisticians. Although various modifications specific to HW testing have been proposed to the standard $\chi^2$ test (e.g. Mantel and Li 1974, Elston and Forthofer 1977), it has been found by Chapco (1976), Emigh (1980), Haber (1981), and Hernandez and Weir (1989) to perform quite well in comparison to exact tests. The major disadvantage of $\chi^2$ testing is that it is severely affected by small expected frequencies, thus making it necessary to pool genotype classes.

Likelihood-ratio tests, although conceptually different from goodness-of-fit tests, can be used interchangeably with the latter to test the hypothesis that deviations of observed from expected genotype frequencies are equal to zero. Despite the opinion of some authors to the contrary (e.g. Zar 1974, p. 50, Ferguson 1980, p. 164), likelihood-ratio tests share the disadvantages of $\chi^2$ in dealing with small expected frequencies, except under some very restricted conditions (Larntz 1978, Sokal and Rohlf 1981, p. 709).

Thus, the advantages and disadvantages of goodness-of-fit tests and likelihood-ratio tests in simple HW testing are about equal, and either one would serve in a given situation. The distribution, behavior in relation to exact tests, and appropriate correction factors of $\chi^2$ are better known. Likelihood-ratio tests will, therefore, not be treated further here. Clear explanations of their rationale and instructions on how to carry them out can be found in Sokal and Rohlf (1981, p. 692) and Weir (1990, p. 82). Given the computational labor necessary to calculate exact probabilities, especially when there are more than two alleles (Hernandez and Weir 1989), the analytical strategy proposed here consists of using a $\chi^2$ test whenever appropriate, and resorting to an exact test when necessary.

How should one test observed genotype frequencies to see how well they approximate HW expectations? The first decision regards the null hypothesis. Expected genotype frequencies can be calculated in several ways. One way to avoid ad hoc pooling of genotypes is to test heterozygosities rather than genotype frequencies; i.e., to pool all homozygotes and pool all heterozygotes at each locus regardless of allele, and test them against predictions of the HW rule about the expected pooled respective numbers (Dobzhansky and Epling 1944, Pamilo and Varvio-Aho 1984). The test statistic will be:

$$\chi^2 = \frac{\left(\left|2N\sum_{i<j}^{n} p_i p_j - \sum_{i<j} O_{ij}\right| - c\right)^2}{2N\sum_{i<j}^{n} p_i p_j} + \frac{\left(\left|N\sum_{i=1}^{n} p_i^2 - \sum_{i=1} O_{ii}\right| - c\right)^2}{N\sum_{i=1}^{n} p_i^2}, \quad (1)$$

where $N$ is the number of individuals sampled, $p_i$ and $p_j$ are estimated frequencies of alleles $i$ and $j$ ($i<j$), $n$ is the number of alleles at the locus, $O_{ij}$ the number of heterozygotes of alleles $i$ and $j$, and $O_{ii}$ is the number of homozygotes of allele $i$. There is always one degree of freedom in this test, and $c$ is the correction factor for continuity. The value of $c$ is usually 0.5 (Yates 1934), but Emigh (1980) suggested that a value of 0.25 is preferable for HW testing.

This approach is appropriate when deviations from HW equilibrium are suspected to be due to aspects of the mating structure of the population that affect overall heterozygosity regardless of allele. It may be the only possible $\chi^2$ test when sample sizes are small, and there is one common allele and many rare ones. Curie-Cohen (1982) stated that it is the most relevant test when the inbreeding coefficient is calculated as the overall heterozygote deficiency. Its major disadvantage is that it has less statistical power than tests that incorporate information about individual genotypes (Dobzhansky and Levene 1948,

Pamilo and Varvio-Aho 1984). It is almost powerless to detect deviations due to forces that act on the frequency of individual alleles or genotypes, such as selection and mutation. Although significant results from this test indicate significant deviations from HW expectations, nonsignificant results may be due to pooling rather than a good fit between expected and observed genotype distributions. In the frequent case of disequilibrium suspected to come from mixing of populations (Wahlund 1928), it will not be an adequate test when there are more than two alleles. Although the Wahlund effect decreases overall heterozygosity in a locus, it may increase the frequencies of particular heterozygotes, depending on the covariances between gene frequencies of the source populations (Nei 1965). Such increases will not only go undetected, but will actually detract from the overall power of this test. It is, therefore, often desirable to carry out an additional test of the frequency of each observed genotype against what the HW rule expects. This is a procedure that requires several decisions along the way, unless sample sizes are in the hundreds and there are no rare alleles.

For testing individual genotype frequencies one could calculate the expected numbers from the HW formula as $Np_i^2$ for homozygotes, and as $2Np_i p_j$ for heterozygotes. However, for small sample sizes (<100 individuals, according to Spiess 1989, p. 42) a special problem arises. Levene (1949), Haldane (1954), and Smith (1970) pointed out that a finite sample from a population at HW equilibrium overrepresents the number of homozygotes; they used a correction (usually referred to as Levene's correction) for the asymptotic nature of gene frequencies. The corrected expected values are:

for heterozygotes:

$$E_{ij}=\frac{4N^2 p_i p_j}{2N-1}, \tag{2}$$

for homozygotes:

$$E_{ij}=\frac{N p_i(2N p_i-1)}{2N-1}. \tag{3}$$

Emigh (1980) found that the use of these expectations in goodness-of-fit $\chi^2$ tests produces results that are, by-and-large, compatible with exact tests. No matter how individual expected genotype frequencies are estimated, the goodness-of-fit $\chi^2$ statistic is:

$$\chi^2=\sum_{i<j}^{n}\frac{(|O_{ij}-E_{ij}|-c)^2}{E_{ij}}+\sum_{i=1}^{n}\frac{(|O_{ii}-E_{ii}|-c)^2}{E_{ii}}. \tag{4}$$

$c$ is the correction factor, which could be 0.5 or 0.25 in the case of two alleles, and should be 0 for more than two alleles. Because one degree of freedom is lost for each estimated allele frequency, the degrees of freedom for this test are not equal to the number of genotype classes minus one, but rather (Dobzhansky and Levene 1948, Ward and Sing 1970, Crisp et al. 1978):

$$df=\frac{n(n-1)}{2}. \tag{5}$$

A problem that invariably arises in goodness-of-fit testing with the sample sizes usually employed in electrophoretic surveys, is that rare alleles lead to small expected genotype frequencies. Because expected values appear in the denominator, the $\chi^2$ test can produce artificially inflated significance values. When Levene's correction of expected values is used, a $\chi^2$ test will be impossible whenever $p_i=1/(2N)$, i.e., whenever a single allele of a particular kind has been observed, because the expected number of homozygotes for this allele will be zero. Above this limiting value, there is uncertainty as to when it is necessary to seek special remedies (Horn 1977). Roscoe and Byars (1971) suggested that it is sufficient to have a sample size that exceeds twice the number of genotype classes when the significance level $\alpha=0.05$, and four times the number of genotype classes when $\alpha=0.01$. Kendall (1952) and Tate and Hayer (1973), on the other hand, recommended that the minimum expected frequency should be 20. Many authors have made recommendations that fall within the range of these two extremes (see Tate and Hayer 1973). The rule most often given in statistical textbooks is that of Cochran (1954), who advised that for an unbiased $\chi^2$ all expected frequencies should be larger than 1, and that more than 80% of the frequencies should be more than 5. In $\chi^2$ tests, the problem of low expected values is circumvented by pooling adjacent classes with low frequencies. However, one cannot do so indiscriminately when testing genotype frequencies. As Pamilo and Varvio-Aho (1984) pointed out, the correct procedure is to pool rare alleles, and recalculate expected and observed genotype frequencies. However, in the case of a locus with two alleles, one of them rare, such pooling results in no test at all, since it converts all individuals to homozygotes. In this particular case, the only possible way to find out whether the genotype frequencies deviate from HW expectations is an exact test. Computational labor can be avoided by the use of the tables published by Vithayasai (1973), in which the number of heterozygotes compatible with HW equilibrium is given as a function of sample size and frequency of the rarest homozygote.

## Significance levels

Electrophoretic studies usually include many loci, a large fraction of which are polymorphic, and often sample many populations. The number of simultaneous tests carried out for HW equilibrium is often large, occasionally exceeding 100 [see Winans (1980) for an example of 132 such tests]. The most serious and most common error in HW testing of electrophoretic data is failure to adjust significance levels for the numbers of tests carried out, a failure that leads to conclusions of significance for deviations that may be due to chance. Such adjustments are not very complicated, and are ignored at the risk of vitiating the conclusions of the entire study.

To adjust significance levels for multiple tests, one can use the standard Bonferroni technique (Miller 1980, p. 8) of dividing the predetermined significance level, $\alpha$, by the number of tests, $k$, to obtain a corrected significance level, $\alpha'=\alpha/k$. Each $\chi^2$ value is than considered significant at $\alpha$ if it exceeds the $\chi^2$ value corresponding to $\alpha/k$. A more

powerful technique is the sequential Bonferroni, devised by Holm (1979) and brought to the attention of evolutionary biologists by Rice (1989). An even sharper (and usually simpler) method of sequential comparisons is that of Hochberg (1988). This procedure orders tests according to their probability value. The highest probability value, $p_m$, is compared to the significance level $\alpha$. If $p_m < \alpha$, it and all smaller probabilities are judged significant, and testing stops. If $p_m \geq a$, the test to which it pertains is deemed not significant, and the comparisons continue with subsequent probabilities, each compared to a modified significance level, $\alpha'_{i+1} = \alpha/(1+i)$, where $i$ is the number of tests already performed. When a test is significant, testing stops, and it and all subsequent tests are deemed significant.

The practical difficulty with applying any Bonferroni technique in HW testing comes from the large number of tests generally performed. Significance levels are usually set at a maximum probability level of 0.05, but when $k$ exceeds 50, $\alpha/k$ is $<0.001$. Statistical tables that give $\chi^2$ values for such small probabilities are hard to find. I am only aware of one (Pearson and Hartley 1976, p. 128). Three alternatives to the use of these tables are: (a) If the $\chi^2$ values are not too close to the critical region, a graphic table can be used, such as that in Crow and Kimura (1970, p. 516), which graphs them from $P=0.0001$ to $P=1.00$. (b) If Emigh's (1980) correction for continuity is not used, a statistical package, such as SPSS (Norusis 1986) that gives probability values to four significant digits can be employed to perform the $\chi^2$ tests. (c) One can calculate the probability value corresponding to each value of $\chi^2$ from the incomplete gamma function. Press et al. (1986, p. 165) provided FORTRAN and PASCAL routines for doing so. An implementation of their algorithms in DBASE-III+ can be obtained on disk from me. The last alternative is desirable if the techniques of Holm (1979) or Hochberg (1988) are used, even if the number of tests does not exceed 50, because in these procedures the ordering of tests and the determination of significance depends on the specific values of probability corresponding to each test statistic. When degrees of freedom in all tests are the same, such ordering can be done using the value of $\chi^2$ alone, but it will rarely be the case that all polymorphic loci have the same number of alleles. Loci with two alleles and low expected frequencies, for which it is necessary to carry out exact tests rather than goodness-of-fit tests, present a special practical problem when the number of tests is $>5$, because the lowest significance level in Vithayasai's (1973) tables is 0.01. In this case, the only alternative is to compute the exact probabilities from the binomial, using Haldane's (1954) method. This can be done in terms of the observed number of one of the two alleles, $n_i$ (where $n_i = 2O_{ii} + O_{ij} = 2Np_i$). Conditional probabilities, $Pr(E_{ij}|n_i)$ are calculated for all possible values of $E_{ij}$. The relation is:

$$\ln[Pr(E_{ij}|n_i)] = \ln(N!) + \ln(n_i!) + \ln[(2N-n_i)!] + E_{ij}\ln(2) - \\ -\ln\left[\left(\frac{n_i - E_{ij}}{2}\right)!\right] - \ln(E_{ij}!) - \ln\left[\left(N - \frac{(n_i + E_{ij})}{2}\right)!\right] - \ln[(2N)!]\,. \tag{6}$$

The logarithmic transformation is necessary, because the untransformed function would cause computers and calculators to overflow, even with small sample sizes. Even so, it may be necessary to calculate the logarithms of some factorials as sums of the logarithms of the factors. At the end of the calculations the logarithms can be converted back to probability values. Note that these probabilities are conditional on $n_i$ being fixed, so the possible number of expected heterozygotes ($E_{ij}$) corresponding to every hypothetical value of homozygotes ($E_{ii}$) will be $E_{ij} = n_i - 2E_{ii}$ (i.e., every increase by one homozygote requires a corresponding decrease by two heterozygotes). The cumulative probability of the observed genotype frequencies in the sample plus the sum of the probabilities of all less likely combinations is then compared to the significance level to determine whether the deviations are significantly different from HW expectations. More details about Haldane's exact test can be found in Weir (1990, p. 78).

Whether one can mix probabilities obtained from $\chi^2$ and from exact tests in a sequential Bonferroni is debatable. When deviations in all comparisons are in the same direction, the error that may arise from the slight differences of probabilities estimated from an approximate test and calculated from an exact one will certainly be smaller than the Type-I error that would result if the number of tests is not taken into account. This is particularly true if Levene's correction is used in the calculation of expected genotype frequencies. However, one should be careful in mixing probabilities from $\chi^2$ and exact tests when there is an excess of homozygotes in one and of heterozygotes in the other. This is because the $\chi^2$ treats positive and negative deviations from expected values as equal, while in exact tests heterozygote deficiencies come out as less probable than excesses (Louis and Dempster 1987, Hernandez and Weir 1989). If the incompatible tests are close to the critical region for assigning significance to all subsequent tests, the only remedy is to calculate exact tests in every case. One hopes that this problem will not appear often, because the amount of computation required is prohibitive even on a computer. For 4 alleles and a sample size of 20, there are 6671 probabilities to be calculated (Hernandez and Weir 1989). In such cases, a pseudo-probability test, based on a sample of the outcomes (Hernandez and Weir 1989) may be necessary.

## Interpretation of results from tests for Hardy-Weinberg expectations

Let us assume that tests have been performed, significance levels have been adjusted, and the results indicate that genotype frequencies in a few loci are significantly different from HW expectations, while those in most loci are not. How is one to interpret such results?

When genotype frequencies do not significantly deviate from expected values, it is often assumed that they are "in conformity with HW equilibrium", and thus that the populations and loci meet the requirements for reaching this equilibrium. Such lack of significance is even used to

support the assertion that the employed isozymes are Mendelian characters (see Fairbairn and Roff 1980). This is wrong. In tests for HW expectations, as in any statistical test, inability to reject the null hypothesis does not necessarily indicate that the hypothesis holds, only that the sample size was insufficient for its rejection. Although the distinction between failing to reject and accepting a null hypothesis is obvious to most biologists, a large number of papers in the marine biological literature assume – explicitly or tacitly – that loci are at HW equilibrium for no other reason than that significant differences could not be demonstrated. This is done without any consideration of the statistical power of the employed test (see Cohen 1988, Fairbairn and Roff 1980, Peterman 1989), even though it has long been clear that HW testing with realistic sample sizes is, even under the best of conditions, a procedure prone to Type-II error in detecting statistically significant effects of inbreeding (Ward and Sing 1970, Haber 1980), or deviations from Mendelian inheritance (Fairbairn and Roff 1980). In addition to the general problem of asymmetry in hypothesis testing, the HW rule has some unique problems. It is, for example, possible for a population in which individuals do not mate at random to have a genotype frequency distribution that mimics the multinomial (Li 1988), or for a locus under strong selection to have no appreciable deviations from HW proportions (Wallace 1958, Lewontin and Cockerham 1959, Li 1959, Workman 1969, Schaap 1980). It is also possible that factors that would cause deviations from HW expectations may pull genotype frequencies in opposite directions, so that the end result is non-significant differences between numbers of observed and expected genotypes (Workman 1969, Cavalli-Sforza and Bodmer 1971, p. 58).

Thus, little can be made of non-significant results, except in comparison with significant ones. Significant deviations, on the other hand, definitely mean that one or more of the preconditions for observing HW equilibrium do not hold. The problem is to determine which of the possible causes of deviation can account for the observed departures. Obviously, one needs to consider sources of error, such as biased sampling or incorrect scoring of gels, as well as the possibility that isozymes at the loci that control the particular enzyme may not behave as simple Mendelian traits, due to post-translational modifications or dominance arising from null alleles (Fairbairn and Roff 1980). Attention to these possibilities is necessary for all loci, whether or not they show significance in tests for HW proportions. The question is whether significant deviations (and thus tests for HW equilibrium) can provide useful information about the genetics of the sampled populations. I believe that such tests, carried out in an extensive fashion with no intention of testing any particular biological hypothesis, are of limited value and not worth reporting in detail. Tests for HW proportions can be useful – and significant deviations can be meaningful – in testing a specific hypothesis based on the biology of the population or on the genetics of the locus. This is the case when genetic data are gathered to test the possibility (or determine the consequences) of asexual reproduction (e.g. Ayre 1984, Hoffmann 1986, Mladenov and Emson 1990), self fertilization (e.g. Brown 1979), or parthenogenesis (Lynch 1983). Such tests can also be useful when examining the possibility that the sample includes two or more assortatively mating populations, such as sibling species (e.g. Makela and Richardson 1977) or immigrants from locally mating populations (e.g. Johnson and Black 1984). In all cases, the multiplicity of possible causes of deviation from HW expectation constrain comparisons between observed and expected genotype frequencies to the role of circumstantial evidence in favor of the existence of the alleged phenomenon. This weakness can be partially overcome in cases where a hypothesis alternate to HW equilibrium can be stated in terms of expected genotype frequencies and tested. This usually takes the form of calculating the fixation index expected under each condition. For an example of the utility of this approach see Brown (1979), and for calculations of expected fixation indices in cases of sex linkage, mating between relatives, various degrees of self-fertilization, biased sex ratios, polysomy, and changing population size see Wright (1969, p. 174–210). More elaborate models, involving a varying number of assumptions, have been offered for specific cases, such as variable degrees of self-fertilization (Hedrick and Cockerham 1986), finite parental mating populations (Robertson 1965, Kirby 1975, Robertson and Hill 1984), isoloci resulting from polyploidization (Stoneking et al. 1981, Waples 1988), and population subdivision (Nei 1965, Sinnock 1975, Makela and Richardson 1977). When no alternative hypothesis can be formulated and tested, inferences as to the true cause of discrepancies from equilibrium can be strengthened through comparisons between populations or between loci. If a population is suspected to be influenced by a factor that draws it away from equilibrium, comparisons of deviations with other populations in which this factor is absent (if such a population can be found) should reveal whether the suspected factor is likely to be the true cause of the discrepancies. Similarly, certain destabilizing forces, such as selection and mutation, act on a single locus, while others, such as population subdivision or inbreeding, act on all loci. Thus, comparisons between loci and testing for linkage disequilibrium (Weir and Cockerham 1989) may be instructive. Neither comparisons to expectations of alternate hypotheses, nor comparisons between populations or between loci will necessarily reveal the processes taking place in natural populations, because the needed parameters are sometimes impossible to measure. However, the evidence that comes from testing against HW expectations, when combined with other sources of information, can serve as a basis for further investigation. For this reason alone, it is important that the statistical tests be performed correctly.

*Acknowledgements.* I thank E. B. Bermingham, M.-A. Coffroth, W. G. Eberhard, H. R. Lasker, N. Knowlton, and L. Weigt for comments on the manuscript. Supported by the Molecular Evolution Program of the Smithsonian Institution.

## Literature cited


Ayre, D. (1984). The effects of sexual and asexual reproduction on geographic variation in the sea anemone *Actinia tenebrosa*. Oecologia 62: 222–229

Brown, A. H. D. (1970). The estimation of Wright's fixation index from genotypic frequencies. Genetica 41: 399–406

Brown, A. H. D. (1979). Enzyme polymorphism in plant populations. Theor. Popul. Biol. 15: 1–42

Cannings, C., Edwards, A. W. F. (1969). Expected genotypic frequencies in a small sample: deviation from Hardy-Weinberg equilibrium. Am. J. hum. Genet. 21: 245–247

Cavalli-Sforza, L. L., Bodmer, W. F. (1971). The genetics of human populations. W. H. Freeman & Co., San Francisco

Chapco, W. (1976). An exact test of the Hardy-Weinberg law. Biometrics 32: 183–189

Cochran, W. G. (1954). Some methods for strengthening the common $\chi^2$ tests. Biometrics 10: 417–451

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd ed. L. Erlbaum Associates, Hillsdale, New Jersey

Cooper, D. W. (1968). The significance level in multiple tests made simultaneously. Heredity, Lond. 23: 614–617

Crisp, D. J., Beaumont, A. R., Flowerdew, M. W., Vardy, A. (1978). The Hardy-Weinberg test – a correction. Mar. Biol. 46: 181–183

Crow, J. F., Kimura, M. (1970). An Introduction to population genetics theory. Harper & Row, New York

Curie-Cohen, M. (1982). Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics, Austin, Tex. 100: 339–358

Dobzhansky, T., Epling, C. (1944). Contributions to the genetics, taxonomy and ecology of *Drosophila pseudoobscura* and its relatives. Publs Carnegie Instn 554: 1–183

Dobzhansky, T., Levene, H. (1948). Genetics of natural populations. XVII. Proof of operation of natural selection in wild populations of *Drosophila pseudoobscura*. Genetics, Princeton 33: 537–547

Elston, R. C., Forthofer, R. (1977). Testing for Hardy-Weinberg equilibrium in small samples. Biometrics 33: 536–542

Emigh, T. H. (1980). A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36: 627–642

Emigh, T. H., Kempthorne, O. (1975). A note on goodness of fit of a population to Hardy-Weinberg structure. Am. J. hum. Genet. 27: 778–783

Fairbairn, D. J., Roff, D. A. (1980) Testing genetic models of isozyme variability without breeding data: can we depend on the $\chi^2$? Can. J. Fish. aquat. Sciences 37: 1149–1159

Ferguson, A. (1980). Biochemical systematics and evolution. John Wiley & Sons, New York

Haber, M. (1980). Detection of inbreeding effects by the $\chi^2$ test on genotypic and phenotypic frequencies. Am. J. hum. Genet. 32: 754–760

Haber, M. (1981). Exact significance levels of goodness-of-fit tests for the Hardy-Weinberg equilibrium. Hum. Hered. 1: 161–166

Haldane, M. B. S. (1954). An exact test for randomness of mating, J. Genet. 52: 631–635

Hedrick, P. W., Cockerham, C. C. (1986). Partial inbreeding: equilibrium heterozygosity and the heterozygosity paradox. Evolution 40: 856–861

Hernandez, J. L., Weir, B. S. (1989). A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45: 53–70

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75: 800–802

Hoffmann, R. J. (1986). Variation in contributions of asexual reproduction to the genetic structure of populations of the sea anemone *Metridium senile*. Evolution 40: 357–365

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6: 65–70

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: a review and an application to a health impairment scale. Biometrics 33: 237–248

Johnson, M. S., Black, R. (1984). The Wahlund effect and the geographical scale of variation in the intertidal limpet *Siphonaria* sp. Mar. Biol. 79: 295–302

Kendall, M. G. (1952). The advanced theory of statistics. 5th ed. Vol. 1. Griffin, London

Kirby, G. C. (1975). Heterozygote frequencies in small subpopulations. Theor. Popul. Biol. 8: 31–48

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. J. Am. statist. Ass. 73: 253–263

Levene, H. (1949). On a matching problem arising in genetics. Ann. math. Statist. 20: 91–94

Lewontin, R. C., Cockerham, C. C. (1959). The goodness-of-fit test for detecting natural selection in random mating populations. Evolution 13: 561–564

Li, C. C. (1955). Population genetics. University of Chicago Press, Chicago

Li, C. C. (1959). Notes on relative fitness of genotypes that forms a geometric progression. Evolution 13: 564–567

Li, C. C. (1988). Pseudo-random mating populations. In celebration of the 80th anniversary of the Hardy-Weinberg law. Genetics, Austin, Tex. 119: 731–737

Lindley, D. V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In: Bernardo, J., DeGroot, M. H., Lindley, D. V., Smith, A. F. M. (eds.) Bayesian statistics. Oxford University Press, Oxford, p. 307–326

Louis, E. J., Dempster, E. R. (1987). An exact test for Hardy-Weinberg and multiple alleles. Biometrics 43: 805–811

Lynch, M. (1983). Ecological genetics of *Daphnia pulex*. Evolution 37: 358–374

Makela, M., Richardson, R. H. (1977). The detection of sympatric sibling species using genetic correlation analysis. I. Two loci, two gamodemes. Genetics, Austin, Tex. 86: 664–678

Mantel, N., Li, C. C. (1974). Estimation and testing of a measure of non-random mating. Ann. hum. Genet. 37: 445–454

Miller, R. G. (1980). Simultaneous statistical inference. Springer Verlag, New York

Mladenov, P. V., Emson, R. H. (1990). Genetic structure of populations of two closely related brittle stars with contrasting sexual and asexual life histories, with observations on the genetic structure of a second asexual species. Mar. Biol. 104: 265–274

Nei, M. (1965). Variation and covariation of gene frequencies in subdivided populations. Evolution 19: 256–258

Norusis, M. J. (1986). SPSS/PC+ for the IBM PC/XT/AT. SPSS Inc., Chicago

Pamilo, P., Varvio-Aho, S. (1984). Testing genotype frequencies and heterozygosities. Mar. Biol. 79: 99–100

Pearson, E. S., Hartley, H. O. (1976). Biometrika tables for statisticians. Vol. I. Biometrika Trust, London

Peterman, R. M. (1990). Statistical power analysis can improve fisheries research management. Can. J. Fish. aquat. Sciences 47: 1–15

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1986). Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge

Rice, W. R. (1989). Analyzing tables of statistical tests. Evolution 43: 223–225

Robertson, A. (1965). The interpretation of genotypic ratios in domestic animal populations. Anim. Prod. 7: 319–324

Robertson, A., Hill, W. G. (1984). Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics, Austin, Tex. 107: 703–718

Roscoe, J. T., Byars, J. A. (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. J. Am. statist. Ass. 70: 811–813

Schaap, T. (1980). The application of the Hardy-Weinberg principle to the study of populations. Ann. hum. Genet. 44: 211–215

Sinnock, P. (1975). The Wahlund effect for the two locus model. Am. Nat. 109: 565–570

Smith, C. A. B. (1970). A note on testing the Hardy-Weinberg law. Ann. hum. Genet. 33: 377–383

Sokal, R. R., Rohlf, F. J. (1981). Biometry. The principles and practice of statistics in biological research. 2nd ed. W. H. Freeman & Co., New York

Spiess, E. B. (1989). Genes in populations. 2nd ed. John Wiley & Sons, New York

Stoneking, M., May, B., Wright, J. E. (1981). Loss of duplicate gene expression in salmonids. Evidence for a null allele polymorphism at the duplicate aspartate amino transferase loci in brook trout *Salvelinus fontinalis*. Biochem. Genet. 19: 1063–1078

Tate, M. W., Hayer, L. A. (1973). Inaccuracy of the $\chi^2$ test of goodness-of-fit when expected frequencies are small. J. Am. statist. Ass. 68: 836–841

Vithayasai, C. (1973). Exact critical values of the Hardy-Weinberg test statistic for two alleles. Communs Statist. 1: 229–242

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11: 65–106

Wallace, B. (1958). The comparison of observed and calculated zygotic distributions. Evolution 12: 113–115

Waples, R. S. (1988). Estimation of allele frequencies at isoloci. Genetics, Austin, Tex. 118: 371–384

Ward, R. H., Sing, C. F. (1970). A consideration of the power of the $\chi^2$ test to detect inbreeding effects in natural populations. Am. Nat. 104: 355–365

Weir, B. S. (1990). Genetic data analysis. Sinauer, Sunderland, Mass.

Weir, B. S., Cockerham, C. C. (1989). Analysis of disequilibrium coefficients. In: Hill, W. G., Mackay, T. F. C. (eds.) Evolution and animal breeding. C. A. B. International, Oxon, p. 45–51

Winans, G. A. (1980). Geographic variation in the milkfish *Chanos chanos*. I. Biochemical evidence. Evolution 34: 558–574

Workman, P. L. (1969). The analysis of simple genetic polymorphisms. Hum. Biol. 41: 97–114

Wright, S. (1969). Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies. University of Chicago Press, Chicago

Yates, F. (1934). Contingency tables involving small numbers and the $\chi^2$ test. Jl R. statist. soc. (Suppl.) 1: 217–235

Zar, J. H. (1974). Biostatistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey