Heredity **74** (1995) 129–135

Received 14 March 1994

One- and two-sample tests for single-locus inbreeding coefficients using the bootstrap

S. VAN DONGEN* & T. BACKELJAU†

Department of Biology, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk and †Royal Belgian Institute of Natural Sciences, Vautierstraat 29, B-1040 Brussels, Belgium

Two bootstrap procedures are proposed to perform one- and two-sample tests on inbreeding coefficients for single loci by resampling over the genotypes. These tests allow testing against a broad range of new alternative hypotheses in addition to panmixis. Monte Carlo simulations show that the coverage probability of these tests behaves satisfactorily if the number of bootstrap resamples is larger than or equal to 2500 and the sample size is larger than or equal to 20, for the case of two alleles. As the fixation index of the underlying distribution becomes more extreme, higher sample sizes are required to obtain a reliable test. Two explicit formulae for the power of the two tests are estimated from Monte Carlo simulations, and a comparison with the classical chi-square test is made. A Turbo-Pascal computer program is available to perform the two presented bootstrap tests.

Keywords: bootstrap, coverage probability, Hardy–Weinberg law, inbreeding coefficient, Monte Carlo simulation, power.

Introduction

The Hardy–Weinberg (HDW) law is of great importance in population genetics and states that within one generation, genotype frequencies will follow the multinomial distribution with the allele frequencies as the distribution parameters, provided that the gametes in that population associate completely at random and no selective forces are active (Lessios, 1992).

Statistical tests for conformity of observed genotype frequencies to HDW-expectations often lack generality and may only be applied when specific assumptions are met. Large sample goodness-of-fit tests can give false and misleading P-values when sample sizes are small and/or some expected cell-counts are sparse or equal to zero, except under some very restricted conditions (Sokal & Rohlf, 1981; Lessios, 1992; but see Agresti, 1990). To overcome this problem many corrections have been proposed, but the improvement in correctness of the statistical test is often limited (Guo & Thompson, 1992). Exact tests are restricted to cases with few alleles (say, 4–5) because of the very high computation time (Hernández & Weir, 1989; Guo & Thompson, 1992; Lessios, 1992). Pooling of genotypes or alleles is often applied to reduce the number of cells in the contingency table so that goodness-of-fit tests can be performed correctly or, if necessary, the complete enumeration of the exact significance level is possible. However, pooling may obscure deviations from HDW-expectations (Swofford & Selander, 1989; Lessios, 1992; Zaykin & Pudovkin, 1993). The use of a Monte Carlo procedure has been suggested by Hernández & Weir (1989) and a general Monte Carlo algorithm to approximate the exact significance level of a test of HDW-conformity has been proposed by Guo & Thompson (1992).

Lessios (1992) pointed out that extensive statistical tests for HDW-conformity are of limited value unless they are performed with the intention of testing a particular biological or genetical hypothesis. This is, for example, the case when comparing an observed fixation index with the fixation index expected under each theoretical condition (for calculation of expected fixation indices under various conditions see Wright, 1969, p. 174–210 and Brown, 1979), or when comparing the observed fixation indices of two populations that differ with respect to a factor which is expected to cause the deviations from HDW-expectations (Lessios, 1992). The above cited test procedures obviously lack the generality to test against all these

*Correspondence.

alternatives. This problem can be rectified by the jackknife or bootstrap procedure. The use of these methods has been proposed to estimate the distribution of F-statistics for multiple loci by resampling over all the scored loci (Weir & Cockerham, 1984; Weir, 1990a,b), but no general resampling algorithm is available to estimate the distribution of observed fixation indices for single loci by resampling over genotypes. In this paper we present two general bootstrap algorithms: the first can be applied to compare statistically an observed fixation index with a fixed expected value and the second one to compare two observed fixation indices. The behaviour of these two test procedures will be investigated by Monte Carlo simulations. A Turbo-Pascal (version 3.0) program which is available free of charge upon request, was written on an IBM 386 PC to perform these tests. (To obtain the program send a formatted 3.5 inch disc to the first author.)

The bootstrap

Efron (1979) introduced a very general resampling procedure, the 'bootstrap', for estimating the distributions of statistics based on independent observations. Let $\boldsymbol{X} = X_1, X_2, ..., X_n$ be a random sample of size n of independent, identically distributed (i.i.d.) random variables with common but unknown distribution function F, and $\boldsymbol{x} = x_1, x_2, ..., x_n$ its observed realization, and let $R(\boldsymbol{X},F)$ be a statistic of interest for which one wants to estimate the distribution. The bootstrap method proceeds by constructing the empirical distribution function F_n, drawing a random sample of size n, with replacement, from F_n and approximating the sampling distribution of $R(\boldsymbol{X},F)$ by the bootstrap distribution of $R^* = R(\boldsymbol{X}^*,F_n)$. Efron (1979) suggests a Monte Carlo approximation to estimate the bootstrap distribution of R^*. Based on this bootstrap distribution, several methods have been proposed to estimate confidence intervals (C.I.) and to perform hypothesis tests (for reviews see Efron & Tibshirani, 1986; Efron, 1987; Hall, 1988; Hall & Wilson, 1991). Hall and Wilson (1991) provide two guidelines for bootstrap hypothesis testing. They recommend resampling the sample data B times, calculating $t^*_i = |R^*(\boldsymbol{X}^*_i,F_n) - R(\boldsymbol{X},F)|/\sigma^*$ for each bootstrap sample, taking $t = t^*_{(B(1-\alpha))}$ where $t^*_{(i)}$ is the ordered statistic of t^*_i and rejecting H_0 if $|R(\boldsymbol{X},F) - R_0|/\sigma > t$. This method performs well if a 'good' estimate of σ is available. With fixation indices the variance must be estimated by the bootstrap, resulting in a nested procedure with high computation times. Hall and Wilson (1991) suggest that in such cases the bootstrap pivoting (i.e. dividing by σ) may be disregarded.

Monte Carlo simulations

Two errors can be made while performing a statistical test: (a) falsely rejecting the null hypothesis which is known as a type I error, or (b) falsely accepting the null hypothesis, the so-called type II error. Two error probabilities are associated with these error types, namely α and β respectively. Ideally, one would like to keep both α and β close to zero. In practice this is not possible, and usually one keeps α fixed (<0.05) while β and consequently the power of the test depends on the sample size, the minimal difference one wants to detect and the test procedure (Siegel & Castellan, 1988).

The two error probabilities can be investigated by Monte Carlo simulations (Bickel & Krieger, 1989). In each simulation step, a dataset is generated from an underlying distribution, and the test is performed. This is repeated many (M) times and one counts for example the number (N) of tests where $P<0.05$. If the data were generated from the null distribution, one would ideally expect that $1-N/M$ (i.e. the coverage probability) is close to 0.95 (i.e. the nominal value $= 1-\alpha$). Analogously, the power of a test can be investigated. The only difference to the previous approach is that datasets are generated from an alternative distribution.

Estimation of the inbreeding coefficient

The F-statistics F_{ST}, F_{IT} and F_{IS} introduced by Wright (1951) offer a convenient way to summarize the population structure (Weir and Cockerham, 1984). F_{IS} and F_{IT} measure the deviations from HDW-proportions in the subpopulation and the total population respectively, while F_{ST} measures the genetic differentiation between subpopulations (Nei, 1977). Nei (1977) showed that the gene diversity of the total population can be partitioned into its intra- and inter-subpopulational components when gene diversity is defined as the frequency of heterozygotes expected under HDW-equilibrium. Therefore he reformulated the three F-statistics and obtained for F_{IS}, which is the statistic of interest here, the following expression:

$$F_{IS} = \frac{H_S - H_I}{H_S},$$

where H_I = observed heterozygosity in subpopulation S and H_S = expected heterozygosity under HDW-conditions. Nei (1978) proposed an unbiased estimator for H_S for a single locus which is given by:

$$H_S = \frac{2n(1-\Sigma f_i^2)}{2n-1},$$

where n = sample size and f_i = frequency of the ith allele.

One-sample test

To test if an observed F_{IS} differs from an expected value F_e we propose the following bootstrap algorithm:

(1) construct F_n, where the x_is contain the genotypes,
(2) draw a bootstrap sample $\boldsymbol{X}^*$ of size n from F_n,
(3) calculate $F^*_{IS\,i}$ based on $\boldsymbol{X}^*$, and $t^*_i = |F^*_{IS\,i} - F_{IS}|$,
(4) repeat steps 2 to 3 B times, and
(5) P-value = proportion of times that $t^*_i > |F_{IS} - F_e|$.

By taking, for example, $t^*_i = F^*_{ISi} - F_{IS}$ and comparing this value to $F_{IS} - F_e$ without taking the absolute values, a one-sided hypothesis test at the 0.05 significance level with H_a:$F_{IS} > F_e$ can be conducted.

We performed Monte Carlo simulations of size 1000 for four different sample sizes (10, 20, 50 and 100) and six different sizes of bootstrap resamples (100, 500, 1000, 2500, 5000 and 10 000) with $\alpha = 0.05$, to investigate the behaviour of the coverage probability, for the case of two alleles with both the allele frequencies equal to 0.5 and H_0:$F_{IS} = 0$. Data were generated from a multinomial distribution in Turbo-Pascal (version 3.0). The pseudorandom number generator was seeded before each series of random numbers was generated in order to make the pseudorandom numbers as 'random' as possible (Duntemann, 1985). The distribution parameters were calculated according to Wright's (1969, p.175) formulae. These simulations (Table 1) show that the coverage probability differs maximally only 2 per cent from the nominal level if the number of bootstrap resamples (B) is larger than or equal to 2500 and if the sample size is larger than or equal to 20. For very small samples (< 20), the coverage probability seems to be lower than expected, even for very large numbers of bootstrap resamples. For large sample sizes (100) a B equal to 500 seems already to result in a coverage probability close to the nominal level. The coverage probabilities for the one-sample test with H_0:$F_{IS} = F_e$ where F_e ranges from 0 to 0.9, estimated with Monte Carlo simulations of size 1000 and 2500 bootstrap resamples, are summarized in Table 2. The coverage probability becomes much lower than the expected nominal level as the fixation index of the underlying distribution becomes more extreme for the sample sizes 10 and 20. This effect seems to be strongest for the smallest sample size and is absent for the larger sample sizes (≥ 50).

We estimated the power of the one-sample test, with Monte Carlo simulations of size 1000 for three sample sizes (20, 50 and 100), $B = 2500$, $\alpha = 0.05$, F_{IS} ranging from 0.1 to 1, and H_0:$F_{IS} = 0$ (Table 3). The sample size equal to 10 was omitted from this analysis since the bootstrap test seems to fail here in the sense that the type I error rate is higher than expected. In order to be able to predict the power of the test for conditions other than in the simulation, we used a logistic regression approach to model the power of the test in relation to the underlying F_{IS} and the sample size (S). Under the null hypothesis, the reduction in deviance ($DEV_{reduction}$) by adding a variable to the model follows approximately a chi-square distribution with 1 degree of freedom for continuous variables (Dobson, 1990).

Table 1 Coverage probabilities for the one-sample bootstrap test, as estimated by Monte Carlo simulations for four different sample sizes and six different numbers of bootstrap resamples with the underlying F_{IS} equal to zero and H_0:$F_{IS} = 0$

Sample size	Number of bootstrap resamples					
	100	500	1000	2500	5000	10000
10	0.60	0.73	0.83	0.79	0.84	0.83
20	0.80	0.92	0.92	0.94	0.93	0.96
50	0.84	0.93	0.92	0.93	0.95	0.94
100	0.93	0.96	0.94	0.96	0.96	0.95

Table 2 Coverage probabilities estimated by Monte Carlo simulations ($M = 1000$) with the number of bootstrap resamples equal to 2500 for four different sample sizes and the underlying F_{IS} ranging from 0 to 0.9.

Sample size	F_{IS} of the underlying distribution									
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.09
10	0.87	0.90	0.91	0.86	0.80	0.77	0.65	0.47	0.32	0.02
20	0.94	0.90	0.90	0.92	0.91	0.89	0.86	0.89	0.83	0.76
50	0.96	0.95	0.95	0.95	0.94	0.93	0.94	0.96	0.92	0.94
100	0.95	0.94	0.94	0.93	0.95	0.95	0.96	0.95	0.94	0.95

Table 3 Power of the one-sample bootstrap test as estimated by Monte Carlo simulations for three different sample sizes and F_{IS} ranging from 0.1 to 1

Sample size	F_{IS} of the underlying distribution									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
20	0.09	0.17	0.33	0.55	0.74	0.86	0.92	0.98	0.99	1
50	0.11	0.28	0.51	0.82	0.96	0.99	1	1	1	1
100	0.13	0.52	0.83	0.99	1	1	1	1	1	1

The residual deviance follows approximately a chi-square distribution with the residual number of degrees of freedom under the hypothesis that the model gives a good fit (Dobson, 1990). A model fit in GLIM resulted in the following model:

$$1-\beta = 1-\exp(-\log_e(1+\exp(-3.308+6.585F_{IS}+0.1059F_{IS}S))).$$

($DEV_{reduction}$ equals 1003, 441 and 244 respectively for the intercept, F_{IS} and $F_{IS}S$ ($P<0.0001$ in all three cases); S on its own was not included in the model: $DEV_{reduction}=1.89$, 1 d.f., $P=0.17$.) The residual deviance of the model equals 16.57 with 30 d.f. ($P=0.98$) indicating a good model fit.

Two-sample test

To test if two fixation indices differ significantly from each other one could take the 95 per cent confidence intervals and reject the null hypothesis of no difference if the intervals do not overlap (Weir, 1990b). This conservative approach (Crowley, 1992) can be improved by using approximately the same bootstrap algorithm as for the one-sample case. The test statistic of interest here is the difference between the two observed fixation indices ($d=F_{IS1}-F_{IS2}$) instead of F_{IS}. If d differs significantly from zero, the two fixation indices can be considered significantly different from each other. By keeping the resampling separate for the two fixation indices, this bootstrap procedure should be less dependent on the similarity of the underlying distributions than most other statistical procedures (Crowley, 1992).

As before, one-sided hypotheses can be tested by omitting the absolute values. To investigate the coverage probability of the two-sample bootstrap test we generated for each Monte Carlo simulation two datasets of equal sample size independently from a multinomial distribution with $F_{IS}=0$. We performed for three samples sizes (20, 50 and 100), and six different numbers of bootstrap resamples (100, 500, 1000, 2500, 5000 and 10000), Monte Carlo simulations of size 1000 for the case of two alleles with both allele frequencies equal to 0.5 and H_0:$F_{IS1}=F_{IS2}$. The results are summarized in Table 4 and show that the coverage probability deviates by maximally 3 per cent from the nominal level. A higher accuracy can be achieved by increasing the number of bootstrap resamples.

Table 4 Coverage probabilities for the two-sample bootstrap test, as estimated by Monte Carlo simulations for three different sample sizes and six different numbers of bootstrap resamples

Sample size	Number of bootstrap resamples					
	100	500	1000	2500	5000	10000
20	0.96	0.98	0.96	0.94	0.95	0.94
50	0.92	0.95	0.96	0.95	0.96	0.95
100	0.98	0.98	0.93	0.95	0.95	0.95

The results of the power estimations are summarized in Table 5. We generated the two datasets seperately. For the first, F_{IS} was kept equal to 0 while for the second F_{IS} ranged from 0.1 to 1. The logistic regression model fit in GLIM resulted in the following formula:

$$1-\beta = 1-\exp(-\log_e(1+\exp(-3.200+4.667F_{IS}+0.0719F_{IS}S))).$$

($DEV_{reduction}$ equals 1176, 449 and 306 respectively for the intercept, F_{IS} and $F_{IS}S$ ($P<0.0001$ in all three cases); S on its own was not included in the model: $DEV_{reduction}=0.24$, 1 d.f., $P=0.62$.) The residual deviance of the model equals 36.56 with 30 d.f. ($P=0.19$) indicating a good model fit.

Comparison with the power of the chi-square test

When the null hypothesis is false, the chi-square and the likelihood ratio test statistics have approximate noncentral chi-square distributions (Agresti, 1990). This property can be used to estimate the power of these tests by estimating the noncentrality parameter.

Table 5 Power estimations of the two-sample bootstrap test as estimated by Monte Carlo simulations for three different sample sizes and the difference between the two underlying fixation indices ranging from 0.1 to 1

Sample size	$F_{IS1} - F_{IS2}$									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
20	0.04	0.10	0.24	0.32	0.46	0.60	0.69	0.77	0.93	1
50	0.07	0.25	0.31	0.47	0.79	0.87	0.94	0.97	0.99	1
100	0.12	0.30	0.58	0.86	0.93	0.95	0.98	1	1	1

For Wright's model this parameter is given by:

$\lambda = N(m-1)F^2$ (Haber, 1980),

where N = sample size, m = number of alleles and F = fixation index. The power of a chi-square test to detect the given F with sample size N and m alleles can be found in noncentrality tables (Agresti, 1990). We compared the power of the chi-square test to that of the one-sample bootstrap test for sample sizes equal to 20, 50 and 100, F_{IS} ranging from 0.1 to 0.9, $m = 2$ (1 d.f.), and H_0:$F_{IS} = 0$. We calculated the proportional difference between the power of the bootstrap test and the chi-square test relative to the power of the bootstrap. The chi-square test has up to ± 20 per cent less power than the bootstrap test for small sample sizes and $F_{IS} < 0.6$ (Table 6). For larger samples, the powers of the bootstrap and chi-square test seem to become comparable.

Table 6 Proportional (per cent) difference between the power of the one-sample bootstrap and chi-square test for three sample sizes, F_{IS} between 0.1 and 1, and H_0:$F_{IS} = 0$

Sample size	F_{IS} of the underlying distribution									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
20	19	14	18	21	18	12	4	3	1	0
50	0	−4	−10	2	2	0	0	0	0	0
100	0	−6	0	−2	0	0	0	0	0	0

Positive values indicate that the bootstrap test has higher power.

Discussion

We presented two bootstrap tests which allow us to compare statistically: (a) an observed fixation index with any expected value between −1 and 1, and (b) two observed fixation indices, both for single loci and by resampling over the genotypes. Whereas all other statistical procedures test for HDW-conformity, these two bootstrap tests provide a way to test against many other alternative hypotheses.

The bootstrap can be applied in almost every situation where the data are independent and are a random sample of the population (Crowley, 1992). This feature does, however, also incorporate the danger that it will be used without considering the behaviour of the test in a new situation. There has been insufficient basic research to determine when the bootstrap can be expected to be reliable (Noreen, 1989; Manly, 1991), which may not always be the case (Crowley, 1992). The nonparametric bootstrap distribution is asymptotically highly accurate (Bickel & Freedman, 1981; Singh, 1981; Efron & Tibshirani, 1986), but that is no guarantee for a good small sample behaviour (Efron & Tibshirani, 1986; Noreen, 1989), and the meaning of 'large' in this context is usually not clear (Manly, 1991). Empirical investigation of the behaviour of the test is very important (Manly, 1991; Crowley, 1992). Monte Carlo simulations offer a convenient way to investigate the behaviour of a test (Bickel & Krieger, 1989). For both the one- and two-sample bootstrap tests proposed here, these simulations show that the type I error probability is close to the expected level (i.e. 0.05) if the number of bootstrap resamples is larger than or equal to 2500 and for sample sizes ≥ 20, confirming the reliability of the tests. For smaller sample sizes however, the bootstrap test seems to fail. Also, as the fixation index of the underlying distribution becomes more extreme the bootstrap test seems to become less reliable for the smaller sample sizes. The reason for that is that the bootstrap implicitly assumes that the sample contains all the 'important' information from the total population (Crowley, 1992). For example, as the underlying fixation index increases, the probability of having only homozygotes in the sample increases. In that case the estimated fixation index equals 1 and all the fixation indices based on a resample of the data will also be equal to 1 resulting in a P-value equal to zero. The presence of one heterozygote in that sample

contains very important information since it provokes variability in the bootstrap resamples. Thus, if one expects an extreme underlying fixation index, the sample size must be large enough to avoid obtaining pure homo- or heterozygous samples. The same problem may arise if there are for example two alleles with one rare allele. The probability of having a pure homozygous sample increases and sample sizes must be taken that are sufficiently large.

As Lessios (1992) and Guo & Thompson (1992) pointed out, the fact that the null hypothesis cannot be rejected does not necessarily indicate that the null hypothesis holds. This may simply be the result of a lack of power due to a small sample size. The power of the statistical test for the given sample size needs to be estimated to get an idea of the type II error probability (β). The formulae found for the two bootstrap tests may prove useful in such cases. These formulae should, however, be used with caution since it has not been tested whether they hold under different conditions, with more alleles and/or other allele frequencies.

Since for the two presented bootstrap tests the distribution of fixation indices for single loci is considered, the tests can be expected to have maximal power to detect factors that affect all alleles or genotypes at that locus, whereas for factors that may act on only one or a few alleles or genotypes, such as selection, the presented methods may have reduced power. The methods can be easily extended, however, to the estimation of the distribution of fixation indices for single alleles which will result in a higher power to detect forces like selection. This extension will be prone, however, to type I errors because of the multiple testing within each locus. Significance levels may be adjusted by a classic Bonferroni technique or by a sharper method of sequential comparisons (Hochberg, 1988; Lessios, 1992).

The comparison of two observed fixation indices with the two-sample bootstrap test is not restricted to the comparison of the same loci with comparable alleles, but can be done between different loci of different species. By repeatedly resampling the data independently for the two samples and calculating the bootstrap estimate of the fixation indices, the distributions of the two fixation indices are compared. In fact one compares the degree of deviation from HDW-expectations and thus of the magnitude of the processes leading to the fixation indices for the compared loci. Since this method should be less dependent on the similarity of the underlying distributions (Crowley, 1992), it can even be applied to compare different loci from different species in contrast to the randomization method. It is, of course, up to the investigator to decide which comparisons are biologically relevant.

Although more simulations must be performed, especially to investigate the behaviour of the tests for multiple allele situations, these bootstrap procedures seem to be a promising technique in statistically comparing inbreeding coefficients. We recommend the application of these bootstrap tests if hypotheses other than testing against panmixis are required. The sample sizes should be kept > 20, and should be increased if the underlying fixation index is expected to be extreme and/or if certain alleles are rare such that the probability of obtaining samples from which resamples show no variability is high. The number of bootstrap resamples should be > 2500 and preferably equal to 5000. Since it takes only a few minutes to perform the test on an IBM 386 PC, it may be better to set the number of bootstrap resamples as high as possible.

Acknowledgements

We are indebted to Yannis Michalaikis, Bart Kempenaers, Erik Matthijsen and Andre A. Dhondt for their critical comments on the manuscript. S. Van Dongen is Research Assistant at the National Fund for Scientific Research (Belgium). Financial support was received from FJBR grants 2.0004.91 and 2.0128.94.

References

AGRESTI, A. 1990. *Categorical Data Analysis.* John Wiley, New York.

BICKEL, P. J. AND FREEDMAN, D. A. 1981. Some asymptotic theory for the bootstrap. *Ann. Stat.*, **9**, 1196–1217.

BICKEL, P. J. AND KRIEGER, A. M. 1989. Confidence bands for a distribution function using the bootstrap. *J. Am. Stat. Ass.*, **84**, 95–100.

BROWN, A. H. D. 1979. Enzyme polymorphism in plant populations. *Theor. Pop. Biol.*, **15**, 1–42.

CROWLEY, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Ann. Rev. Ecol. Syst.*, **23**, 405–447.

DOBSON, A. J. 1990. *An Introduction to Generalized Linear Models.* Chapman and Hall, London.

DUNTEMANN, J. 1985. *Complete Turbo Pascal.* Glenview, IL.

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

EFRON, B. 1987. Better bootstrap confidence intervals. *J. Am. Stat. Ass.*, **82**, 171–185.

EFRON, B. AND TIBSHIRANI, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–77.

GUO, S. W. AND THOMPSON, E. A. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

HABER, M. 1980. Detection of inbreeding effects by the chi-square test on genotypic and phenotypic frequencies. *Am. J. Hum. Genet.*, **32**, 754–760.

HALL, P. 1988. Theoretical comparison of bootstrap confidence intervals. *Ann. Stat.*, **16**, 927–953.

HALL, P. AND WILSON, S. R. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics*, **47**, 757–762.

HERNANDEZ, J. L. AND 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.

LESSIOS, H. A. 1992. Testing electrophoretic data for agreement with Hardy–Weinberg expectations. *Mar. Biol.*, **112**, 517–523.

MANLY, B. F. J. 1991. *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London.

NEI, M. 1977. *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.*, **41**, 225–233.

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

NOREEN, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley, New York.

SIEGEL, S. AND CASTELLAN, N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw–Hill, New York.

SINGH, K. 1981. On the asymptotic accuracy of Efron's bootstrap. *Ann. Stat.*, **9**, 1187–1195.

SOKAL, R. R. AND ROHLF, F. J. 1981. *Biometry*, 2nd edn. Freeman, San Fransisco.

SWOFFORD, D. L. AND SELANDER, R. B. 1989. BIOSYS-1. A computer program for the analysis of allelic variation in population genetics and biochemical systematics. Release 1.7. University of Illinois, Urbana, IL.

WEIR, B. S. 1990a. Intraspecific Differentiation. In: Hillis, D. M. and Moritz, C. (eds) *Molecular Systematics*, pp. 373–410. Sinauer, Sunderland, MA.

WEIR, B. S. 1990b. *Genetic Data Analysis*. Sinauer, Sunderland, MA.

WEIR, B. S. AND COCKERHAM, C. C. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

WRIGHT, S. 1951. The genetical structure of populations. *Ann. Eugen.*, **15**, 323–354.

WRIGHT, S. 1969. *Evolution and the Genetics of Populations*, vol 2, *The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

ZAYKIN, D. V. AND PUDOVKIN, A. I. 1993. Two programs to estimate significance of chi-square values using pseudo-probability tests. *J. Hered.*, **84**, 152.