Testing Heterozygote Excess and Deficiency

François Rousset* and Michel Raymond*,†

** Équipe Génétique et Environnement, Institut des Sciences de l'Évolution, Université de Montpellier II, 34095 Montpellier, France and † Department of Genetics, Uppsala University, S-75007 Sweden*

Manuscript received January 3, 1995
Accepted for publication May 5, 1995

ABSTRACT

Currently used tests of Hardy-Weinberg proportions do not take into account the nature of the alternative hypothesis, which is generally a heterozygote deficiency. Different exact tests, appropriate for small sample size and large number of alleles, are proposed in this perspective, and their properties are evaluated by power comparisons. Some tests are found to be close to optimal for the detection of inbreeding or heterozygote excess, one of which is a score test closely related to Robertson and Hill's estimator of the inbreeding coefficient. This test is also easily applied to multiple samples. Such tests are not always the most appropriate if alternative hypotheses differ from those considered here.

A deviation from Hardy-Weinberg proportions indicates either selection, population mixing or nonrandom mating, and its detection is one of the first steps in the study of population structure. For example, population structure as well as selfing can result in heterozygote deficiency. Such deviations are usually tested for one locus by computing a chi-square statistic that follows asymptotically a chi-square distribution with $k(k-1)/2$ df under the null hypothesis, where k is the observed number of alleles (Li and Horvitz 1953). Also used is an exact test inspired from Fisher's exact test for contingency tables, which was first applied by Haldane (1954), and a large number of less successful proposals (see Lessios 1992 for further references). Exact tests are appropriate even when many rare alleles are present (Guo and Thompson 1992; Chakraborty and Zhong 1994) and should therefore be used for population genetic analyses of hypervariable markers such as microsatellite loci (*e.g.*, Di Rienzo *et al.* 1994; Estoup *et al.* 1994; Morin *et al.* 1994).

It has sometimes been pointed out that the chi-square statistic is not very appropriate when the alternative hypothesis of the test is heterozygote deficiency (Pamilo and Varvio-Aho 1984; Lessios 1992), and the same may be true of the exact test as generally used. Although heterozygote deficiency is the alternative most often considered in practice, such tests do not take this fact into account. For this reason, we discuss exact tests based on different statistics and compare them to previous ones.

DEFINITION OF TESTS

Definition of alternative hypotheses: Consider a population in which there are k alleles at frequencies $\mathbf{p} = (p_1 \cdots p_k)$. Any deviation from Hardy-Weinberg proportions can be summarized by the set $\mathbf{f}$ of parameters f_{ij} ($f_{ij} = f_{ji}$) such that the probability of some sample S of N individuals, composed of n_{ij} genotypes $A_iA_j (i \leq j)$, is as follows (model 1):

$$L_1(S, \mathbf{f}, \mathbf{p}) = \frac{N!}{n_{11}!n_{12}!\cdots n_{kk}!}\left(p_1^2 + p_1 \sum_{j\neq 1} f_{1j}p_j\right)^{n_{11}} \times (2p_1p_2(1-f_{12}))^{n_{12}} \cdots \left(p_k^2 + p_k \sum_{j\neq k} f_{kj}p_j\right)^{n_{kk}}. \quad (1)$$

In some situations (*e.g.*, regular inbreeding) all f_{ij} are identical to some value f so that this probability is as follows (model 2):

$$L_2(S, f, \mathbf{p}) = \frac{N!}{n_{11}!n_{12}!\cdots n_{kk}!}(p_1^2 + fp_1(1-p_1))^{n_{11}} \times (2p_1p_2(1-f))^{n_{12}} \cdots (p_k^2 + fp_k(1-p_k))^{n_{kk}}. \quad (2)$$

Exact tests: The null hypothesis is $f = 0$. This is a composite hypothesis, because it does not specify unique values for $\mathbf{p}$ (Stuart and Ord 1991). For this reason it is necessary to consider a probability distribution deduced from the above one and independent of $\mathbf{p}$ when $f = 0$. Considering the set $\mathcal{C}$ of all possible samples s_j that, for all allelic types i, have number of genes n_i identical to those of a particular sample S, Levene (1949) has shown that the conditional probability

$$\Pr(S) = \frac{L_1(S, 0, \mathbf{p})}{\Sigma_{\mathcal{C}} L_1(s_j, 0, \mathbf{p})} = \frac{L_2(S, 0, \mathbf{p})}{\Sigma_{\mathcal{C}} L_2(s_j, 0, \mathbf{p})} \quad (3)$$

is independent of $\mathbf{p}$, and its value is

$$\Pr(S) = \frac{N!\prod_1^k n_i!}{(2N)!\prod_{j\geq i} n_{ij}!}\, 2^{\Sigma_{j>i} n_{ij}}. \quad (4)$$

Corresponding author: François Rousset, Équipe Génétique et Environnement, Institut des Sciences de l'Évolution, CC065, USTL, Place E. Bataillon, 34095 Montpellier Cedex 05, France.
E-mail: rousset@isem.univ-montp2.fr

The probability test defines a ranking of possible samples from lower to higher $\Pr(S)$, and the P value of the sample is the sum of probabilities of samples with lower rank, *i.e.*, $\Sigma_{\Pr(s_i)\leq\Pr(S)} \Pr(s_i)$. This is the exact test of most authors (LOUIS and DEMPSTER 1987; HERNÁNDEZ and WEIR 1989; WEIR 1990; GUO and THOMPSON 1992; CHAKRABORTY and ZHONG 1994).

Exact tests generally share several features that need to be distinguished: (1) no use of asymptotic distributions (exactness), (2) a probability distribution independent of unknown parameters under the null hypothesis, an important requirement that together with (1) leads to the use of particular conditional distributions like LEVENE's one that is independent of unknown frequencies $\mathbf{p}$, and (3) use of the probability of a particular configuration as a test statistic.

The conditional distribution of any statistic under the null hypothesis can be computed using LEVENE's distribution. Thus, different test statistics define different rankings of possible samples, but the P value is similarly defined as a sum of exact probabilities of samples with more extreme ranks, so all tests are exact tests. The choice of the probability as a test statistic traces back to FISHER's exact test for contingency tables. FISHER rejected any theory for the choice of test statistics (LEHMANN 1993). Under the NEYMAN-PEARSON theory these are chosen on the basis of the probability of rejection of the null hypothesis (power) when some specified alternative hypothesis is true. In the following, the exact power of some classical tests (*e.g.*, EMIGH 1980; HERNÁNDEZ and WEIR 1989) will be compared to that of tests more specifically adapted to the detection of inbreeding.

Definition of test statistics: In addition to the probability (Pr) test, we considered the following tests:

X^2 (exact) test: Its statistic is the well-known chi square, *i.e.*,

$$X^2 = \left(\sum_i \frac{(n_{ii} - N\tilde{p}_i^2)^2}{N\tilde{p}_i^2}\right) + \left(\sum_{i<j} \frac{(n_{ij} - 2N\tilde{p}_i\tilde{p}_j)^2}{2N\tilde{p}_i\tilde{p}_j}\right) \quad (5)$$

(LI and HORVITZ 1953), where $\tilde{\mathbf{p}} = (\tilde{p}_1 \cdots \tilde{p}_k)$ are the observed allele frequencies in the sample.

G (exact) test: Its statistic is a log likelihood ratio statistic directly derived from the general model (eq. 1) of deviation from Hardy-Weinberg proportions, so that

$$G(S) = \sum_i n_{ii} \log\left(\frac{\tilde{p}_i^2}{\tilde{p}_{ii}}\right) + \sum_{i<j} n_{ij} \log\left(\frac{2\tilde{p}_i\tilde{p}_j}{\tilde{p}_{ij}}\right), \quad (6)$$

where the $\tilde{p}_{ij}$ are the observed genotype frequencies (*e.g.*, HERNÁNDEZ and WEIR 1989).

$\hat{f}_C$ or N_{Het} test: $\hat{f}_C$ is the estimator of f defined by COCKERHAM (*e.g.*, COCKERHAM 1973; WEIR and COCKERHAM 1984). It can be used as a test statistic and has a simple relation with the number of heterozygotes N_{Het} (easily derived from COCKERHAM 1973):

$$\hat{f}_C = \frac{4N^2 - \Sigma_i n_i^2 - 2(2N-1)N_{Het}}{4N^2 - \Sigma_i n_i^2 - 2N_{Het}}. \quad (7)$$

Within the set $\mathcal{C}$ of samples with identical n_i, $\hat{f}_C$ is function of N_{Het} only, decreasing for increasing N_{Het}. Therefore identical tests are defined from $\hat{f}_C$ and N_{Het} and it is simpler to use the N_{Het} test. The P value is $\Sigma_{N_{Het}(s_i)\leq N_{Het}(S)} \Pr(s_i)$ if the alternative hypothesis is heterozygote deficiency and $\Sigma_{N_{Het}(s_i)\geq N_{Het}(S)} \Pr(s_i)$ for heterozygote excess. Thus two one-tailed tests can be defined by these two definitions of P value.

$\hat{f}$ and $\hat{f}_A$ tests: The second model (eq. 2) suggests other statistics, such as the maximum likelihood estimate $\hat{f}$ of f that, jointly with maximum likelihood estimates $\hat{\mathbf{p}}$ of allele frequencies, maximizes $L_2(S, f, \mathbf{p})$. When there are more than two alleles, numerical methods are necessary to find this maximum and the estimates $\hat{f}$ and $\hat{\mathbf{p}}$, so that another estimate $\hat{f}_A$ of f that maximizes the likelihood L_2 if the vector of observed frequencies $\tilde{\mathbf{p}}$ is taken as estimate of $\mathbf{p}$ has been suggested as a suitable alternative (LI and HORVITZ 1953; CURIE-COHEN 1982). $\hat{f}_A$ is the unique root of

$$-\left.\frac{\partial \log L_2}{\partial f}\right|_{\mathbf{p}=\tilde{\mathbf{p}}} = \sum_i n_{ii} \frac{1-\tilde{p}_i}{\tilde{p}_i + (1-\tilde{p}_i)f} - \frac{N_{Het}}{1-f} = 0 \quad (8)$$

within $[-1, 1]$ ($\hat{f}_A = 1$ if $N_{Het} = 0$, $\hat{f}_A = -1$ if $N_{Het} = N$; see APPENDIX). Two one-tailed tests can be defined as above.

Score (U) test: It is constructed from the score

$$\left.\frac{\partial \log L_2}{\partial f}\right|_{f=0}$$

(*e.g.*, COX and HINKLEY 1974; STUART and ORD 1991). Here again the allelic frequencies $\mathbf{p}$ are unknown and we use their maximum likelihood estimates under the null hypothesis $f = 0$, which are the observed frequencies $\tilde{\mathbf{p}}$:

$$U = \left.\frac{\partial \log L_2}{\partial f}\right|_{f=0,\mathbf{p}=\tilde{\mathbf{p}}} = \sum_{i=1}^{k} \frac{n_{ii}}{\tilde{p}_i} - N, \quad (9)$$

which is the left side of eq. 8 evaluated at $f = 0$.

U is equivalent for testing purposes to the estimator $\hat{f}_T$ of ROBERTSON and HILL (1984, eq. 13) because within $\mathcal{C}$, $\hat{f}_T$ is a (monotone) function of U only:

$$\hat{f}_T = \frac{(2N-1)(1 + U/N) - (2N-k)}{2(N-1)(k-1)}. \quad (10)$$

Applications of the NEYMAN-PEARSON lemma: The NEYMAN-PEARSON lemma (*e.g.*, STUART and ORD 1991; LEHMANN 1994) states that the most powerful possible test of a null hypothesis *vs.* an alternative hypothesis would be defined by ordering samples according to the ranks of the ratio λ of likelihoods under the null hypothesis and under an alternative hypothesis, such as $\lambda(S) = L_2(S, 0, \mathbf{p}) / L_2(S, f, \mathbf{p})$. λ must not be confused with a likelihood ratio statistic as it depends on unknown

TABLE 1

Power comparisons (model 2)

Parameters			Upper bound	Power of tests					
f	N	**p**		$\hat{f}_A$	U	N_{Het}	Pr	X^2	G
2 alleles:									
−1/6	100	(0.25 0.75)	0.4363	0.4363	0.4363	0.4363	0.3046	0.3403	0.3582
1/8	50	(0.45 0.55)	0.1562	0.1562	0.1562	0.1562	0.1116	0.1074	0.1056
1/4	50	(0.45 0.55)	0.4505	0.4505	0.4505	0.4505	0.3730	0.3637	0.3573
1/2	20	(0.25 0.75)	0.4700	0.4700	0.4700	0.4700	0.4541	0.4475	0.4082
3 alleles:									
−1/8	50	$\pi_{3,1}$ = (0.5 0.3 0.2)	0.3541	0.3356	0.3498	0.2722	0.1106	0.1247	0.1748
−1/8	50	$\pi_{3,2}$ = (0.3$\overline{6}$ 0.3$\overline{3}$ 0.3)	0.3410	0.3322	0.3350	0.2821	0.1405	0.1523	0.1785
1/10	100	$\pi_{3,1}$	0.392	0.392	0.396	0.334	0.204	0.199	0.154
1/10	100	$\pi_{3,2}$	0.422	0.427	0.427	0.389	0.193	0.181	0.165
1/10	100	$\pi_{3,3}$ = (0.7 0.2 0.1)	0.3676	0.3666	0.3624	0.3033	0.2388	0.2323	0.1589
1/8	50	$\pi_{3,1}$	0.3390	0.3352	0.3322	0.2875	0.1790	0.1650	0.1235
1/8	50	$\pi_{3,2}$	0.3437	0.3395	0.3388	0.3011	0.1578	0.1451	0.1167
1/8	50	$\pi_{3,3}$	0.2997	0.2948	0.2908	0.2425	0.2176	0.1995	0.1388
1/4	20	$\pi_{3,1}$	0.4218	0.4165	0.4053	0.3623	0.2773	0.2373	0.1867
1/4	20	$\pi_{3,2}$	0.4226	0.4160	0.4100	0.3702	0.2493	0.2135	0.1593
1/4	20	$\pi_{3,3}$	0.3405	0.3380	0.3260	0.2831	0.2761	0.2480	0.2165
1/4	50	$\pi_{3,1}$	0.7536	0.7515	0.7445	0.7075	0.5411	0.5172	0.4273
1/4	50	$\pi_{3,2}$	0.7736	0.7688	0.7646	0.7388	0.5184	0.4970	0.4386
1/4	50	$\pi_{3,3}$	0.6844	0.6806	0.6561	0.6117	0.5538	0.5183	0.4175
1/2	20	$\pi_{3,1}$	0.8871	0.8832	0.8641	0.8569	0.7647	0.7194	0.6389
1/2	20	$\pi_{3,3}$	0.7636	0.7589	0.7352	0.7203	0.6938	0.6660	0.6206
>3 alleles:									
−1/10	50	$\pi_{5,1}$	0.431	0.381	0.428	0.349	0.052	0.067	0.111
1/8	50	$\pi_{5,2}$	0.499	0.504	0.480	0.430	0.215	0.192	0.132
1/4	50	$\pi_{5,2}$	0.922	0.918	0.908	0.886	0.669	0.629	0.491
1/4	20	π_8	0.801	0.770	0.723	0.705	0.524	0.358	0.316

Each comparison is based on 1000 or 10,000 independents samples (three or four decimal points, respectively) containing more than one allelic type. The upper bound is computed as described in text (NEYMAN-PEARSON lemma). The estimates of power are binomially distributed, therefore their SEs are at most 0.005 for 10,000 samples and 0.016 for 1000 samples. For each of the >three-alleles cases, the average SE of the estimate of P value for $\hat{f}$, computed as in GUO and THOMPSON (1992), was below 0.005. Lower P values have lower SEs. $\pi_{5,1}$ = (0.22 0.21 0.20 0.19 0.18), $\pi_{5,2}$ = (0.30 0.25 0.20 0.15 0.10), and π_8 = (0.16 0.15 0.14 0.13 0.12 0.11 0.10 0.09).

parameters, f and **p** in this example. In the two-allele case, it can be checked that the ranking of λ is independent of **p** within $\mathcal{C}$ and depends only on the sign of f and the value of n_{12}, so that any monotone function of this value will define a uniformly most powerful test of $f = 0$ *vs.* one-sided alternatives $f > 0$ or $f < 0$. But generally the ranking of λ depends on allele frequencies so that no real test can be defined from the lemma. In power comparisons where samples are generated from a completely specified alternative hypothesis, these parameters are known so λ can be computed and used to construct a test that is more powerful than any possible test (this is not strictly true when different possible samples have the same λ, see TOCHER 1950). The power of real tests can then be compared to this upper bound.

POWER COMPARISONS

The power of different statistics to detect several kinds of deviations from Hardy-Weinberg proportions have been computed for different numbers of alleles. Heterozygote deficiency can be generated by regular systems of inbreeding (eq. 2) or by population structure, in which case the general model (eq. 1) is appropriate. Heterozygote excess can also be the alternative hypothesis of interest, for example if symmetric overdominance is expected, or if allelic frequencies are expected to differ in fathers and mothers.

For each set of parameters (**p**, **f**), samples were generated according to multinomial models (eq. 1 or 2). The algorithm §10.4 of PRESS *et al.* (1988) was used to estimate $\hat{f}$, and $\hat{f}_A$ was estimated by Newton-Raphson method. Performing exact tests requires algorithms that generate Levene's probability distribution for multiple alleles and large sample sizes. Exact P values were computed in the two- and three-allele cases using the complete enumeration algorithm of LOUIS and DEMPSTER (1987) or estimated by a Markov chain algorithm (GUO and THOMPSON 1992) after 100 000 iterations for more than three alleles. The computer programs have been checked by comparison with published results (LOUIS and DEMPSTER 1987; GUO and THOMPSON 1992) for

TABLE 2

Power comparisons of $\hat{f}$ vs. $\hat{f}_A$

Parameters			Power of	
f	N	**p**	$\hat{f}$	$\hat{f}_A$
−1/8	50	(0.5 0.3 0.2)	0.332	0.324
−1/10	50	(0.22 0.21 0.20 0.19 0.18)	0.412	0.404
1/4	20	$(0.3\overline{6}\ 0.3\overline{3}\ 0.3)$	0.422	0.419
1/4	50	(0.5 0.3 0.2)	0.723	0.722

Power is evaluated on 1000 samples containing more than one allelic type. See Table 1 for SEs. In the five-alleles case, P values are estimated after 10,000 iterations of the Markov chain algorithm. In this case, the average SE of the $\hat{f}$ estimate was 0.013.

these algorithms and by independent calculations for maximum likelihood estimation. A computer program that performs the exact U test on one or several samples is available in the present version of the Genepop software (RAYMOND and ROUSSET 1995).

Power was estimated as the frequency of tests for which the P value is below 0.05, an arbitrary choice that does not bias the comparison of tests (results not shown). Because of the relatively heavy computational requirements of the joint estimation of $\hat{f}$ and $\hat{\mathbf{p}}$, the test based on $\hat{f}$ was considered only in a few comparisons.

Results are presented in Tables 1–3 for the two models and two to eight alleles (three in most cases). The present power comparisons agree with earlier results by EMIGH (1980) and HERNÁNDEZ and WEIR (1989) who showed for two- and three-allele cases that the exact Pr and χ^2 tests are slightly better than the exact G test to detect heterozygote deficiency, whereas the reverse is true for heterozygote excess. With two alleles, tests based on $\hat{f}_A$, U and N_{Het} are identical and most powerful, as found by application of the Newman-Pearson lemma since these statistics are monotone functions of the number of heterozygotes. With several alleles, the power of all tests increases with the number of alleles, but this increase is more pronounced for the tests based on $\hat{f}_A$, U and N_{Het}, as seen for $(f, N) = (1/8, 50)$ or $(1/4, 50)$ in Table 1. Among them, the best tests are always those based on $\hat{f}_A$ and U for model 2, $\hat{f}_A$ being slightly better for homozygote deficiency. N_{Het} performs less well than these two in all cases. The test based on $\hat{f}$ was found in a few comparisons (Table 2) to be slightly more powerful than the $\hat{f}_A$ test, but its computational requirements are much heavier. Comparison to the upper bound (NEYMAN-PEARSON lemma) shows that no substantial improvement is to be expected from yet unknown tests.

We investigated some cases where different genotypes have different f_{ij} (Table 3), subject to the restriction that there is no excess of any homozygote class, as found in offspring of crosses between males and females with different allelic frequencies or no deficiency of any homozygote class, as for the Wahlund effect. In both cases there may be an excess of some heterozygote classes and deficiency of some others. There is no simple rule as to which test will be the best one when f_{ij} values are very heterogeneous.

DISCUSSION

Power of the tests: The main result is the behavior of $\hat{f}_A$ and U under model 2 that are best for heterozygote deficiency and excess, respectively, and close to each other in both cases. When f is close to 0, they have almost identical power. When model 2 is true and f is small, it is expected that estimators of f that have low variance under the null hypothesis will provide asymptotically more powerful tests. $\hat{f}$, $\hat{f}_T$ and $\hat{f}_A$ are asymptotically normal with variance $1/(N(k-1))$ when $f = 0$ (YASUDA 1968; ROBERTSON and HILL 1984; and see APPENDIX). The variance of $\hat{f}_C$ is also $1/(N(k-1))$ when allele frequencies are identical but is generally larger (CHAKRABORTY and DANKER-HOPFE 1991, eq. 4.2b). The score (U) test being identical to the $\hat{f}_T$ test and N_{Het} to $\hat{f}_C$, it is expected that the score and $\hat{f}_A$ tests will be generally more powerful than N_{Het} when f is small. This is indeed observed even for large values of f, N_{Het} being less efficient. An estimate of power can be deduced from the large sample normality and variance of $\hat{f}_A$ and $\hat{f}_T$ under the null hypothesis as follows: $y = |f|\sqrt{N(k-1)} - x$, where x is such that the upper-tail probability at x of the normal distribution is the type I error and the power is the lower tail probability at y of the normal distribution. This appears to be an overestimate for low values of $N(k-1)$.

Consistency: Apart from power, the important difference between the tests is one of consistency. A test such as the exact X^2 test is consistent, *i.e.*, its power tends to 1 when sample size increases (*e.g.*, STUART and ORD 1991) for any deviation from Hardy-Weinberg proportions in model 1. The different estimators of f converge in probability to nonnull values, if for all alleles the expected homozygote deficiency is ≥ 0 and at least one is >0, or if all homozygote deficiencies are ≤ 0 and at least one is <0. Such situations include inbreeding or the Walhund effect. In all of these cases, the $\hat{f}_A$, U, and N_{Het} tests will be consistent. However, some simple situations in which a better test exists should not be overlooked. For example, when for an arbitrary number of alleles, the alternative hypothesis is selection for or against one particular genotype (homozygote or heterozygote), the best test statistic is the number of individuals with such a genotype in the sample. HERNÁNDEZ and WEIR (1989) discuss procedures to study deviations of each heterozygote class from Hardy-Weinberg proportions.

Multiple samples: The definitions of $\hat{f}_A$ and U can

TABLE 3

Power comparisons (model 1)

Parameters			Upper	Power of tests					
f	N	**p**	bound	$\hat{f}_A$	U	N_{Het}	Pr	X^2	G
$\begin{pmatrix} -63/208 & & \\ 0 & 0 & \\ -9/16 & 0 & -63/208 \end{pmatrix}$	50	(0.35 0.3 0.35)	0.9429	0.3984	0.7687	0.8603	0.8042	0.8186	0.8145
$\begin{pmatrix} -7/57 & & \\ -1/11 & -9/319 & \\ -1/6 & 3/22 & -1/16 \end{pmatrix}$	50	(0.525 0.275 0.2)	0.2141	0.1332	0.1551	0.1441	0.0794	0.0879	0.1084
$\begin{pmatrix} 1/99 & & \\ 1/33 & 1/21 & \\ -1/33 & 1/9 & 1/51 \end{pmatrix}$	100	(0.55 0.3 0.15)	0.1368	0.1061	0.1036	0.0817	0.0744	0.0703	0.0655
$\begin{pmatrix} 9/91 & & \\ 0 & 0 & \\ 9/49 & 0 & 9/91 \end{pmatrix}$	50	(0.35 0.3 0.35)	0.1931	0.1571	0.1548	0.1400	0.1095	0.1014	0.0907
$\begin{pmatrix} -9/91 & & \\ 0 & 0 & \\ -9/49 & 0 & -9/91 \end{pmatrix}$	50	(0.35 0.3 0.35)	0.1909	0.1441	0.1551	0.1273	0.0938	0.1016	0.1080
$\begin{pmatrix} 7/57 & & \\ 1/11 & 9/319 & \\ 1/6 & -3/22 & 1/16 \end{pmatrix}$	50	(0.525 0.275 0.2)	0.2276	0.1614	0.1537	0.1453	0.1229	0.1134	0.0871
$\begin{pmatrix} 1/6 & & \\ 1/10 & 1/75 & \\ 3/14 & -3/35 & 9/91 \end{pmatrix}$	100	(0.4 0.25 0.35)	0.5215	0.3425	0.3195	0.3465	0.2735	0.2620	0.2510
$\begin{pmatrix} 63/208 & & \\ 0 & 0 & \\ 9/16 & 0 & 63/208 \end{pmatrix}$	50	(0.35 0.3 0.35)	0.9279	0.7885	0.7107	0.8189	0.8244	0.7963	0.7754

Each comparison is based on 10,000 independent samples. See Table 1 for SEs. For convenience we give **f** as a semimatrix in which $f_{ii} = \Sigma_{j \neq i} f_{ij} p_j / (1 - p_i)$ (since $f_{ij} = f_{ji}$).

readily be extended to l samples, *i.e.*, different populations or different loci, if it can be assumed that there is no gametic disequilibrium between them:
$\hat{f}_A$ is the root of

$$\frac{\Sigma_l N^l_{Het}}{1-f} - \sum_l \sum_i^{k_l} n^l_{ii} \frac{1 - \tilde{p}^l_i}{\tilde{p}^l_i + (1 - \tilde{p}^l_i) f} = 0 \qquad (11)$$

and

$$U = \sum_l U_l = \sum_l \sum_{i=1}^{k_l} \frac{n^l_{ii}}{\tilde{p}^l_i} - \sum_l N_l, \qquad (12)$$

the latter test being easier to compute in practice, because the distribution of the score U under the null hypothesis can be obtained by summation of single sample scores which distributions are generated independently. Such multisample tests are preferable to a procedure like Fisher's combination of probabilities test (*e.g.*, YATES, 1955).

Conclusion: The score (U) or $\hat{f}_A$ tests are clearly to be used when the alternative hypothesis of importance for the problem at hand is described by a single parameter f similar for all genotypes. Inbreeding large enough to be detected by single-sample tests occurs in various biological contexts (*e.g.*, JARNE and CHARLESWORTH 1993; THORNHILL 1993), but the power remains low for some applications (ROBERTSON and HILL 1984). In such cases, the multisample U test will be useful.

M.R. thanks P. PAMILO for the opportunity to spend a year at the Department of Genetics of the University of Uppsala. We thank P. DAVID, J. GOUDET, P. JARNE, P. PAMILO and particularly J.-D. LEBRETON for discussion. This work was supported by the Programme Environnement du Centre National de la Recherche Scientifique (GDR 11.05) and the Centre de Biosystématique de Montpellier. This is paper 95-038 of the Institut des Sciences de l'Évolution.

LITERATURE CITED

CHAKRABORTY, R., and H. DANKER-HOPFE, 1991 Analysis of population structure: a comparative study of different estimators

of Wright's fixation indices, pp. 203–254 in *Handbook of Statistics*, vol. 8, edited by C. R. RAO and R. CHAKRABORTY, Elsevier, Amsterdam.

CHAKRABORTY, R., and Y. ZHONG, 1994 Statistical power of an exact test of Hardy-Weinberg proportions of genotypic data at a multiallelic locus. Hum. Hered. **44:** 1–9.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics*. Chapman & Hall, London.

CURIE-COHEN, M., 1982 Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics **100:** 339–358.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

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

ESTOUP, A., M. SOLIGNAC and J.-M. CORNUET, 1994 Precise assessment of the number of patrilines and of genetic relatedness in honeybee colonies. Proc. R. Soc. Lond. Ser. B **258:** 1–7.

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

HALDANE, J. B. S., 1954 An exact test for randomness of mating. Genetics **52:** 631–635.

HERNÁNDEZ, J. L., and B. S. WEIR, 1989 A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics **45:** 53–70.

JARNE, P., and D. CHARLESWORTH, 1993 The evolution of the selfing rate in functionally hermaphrodite plants and animals. Annu. Rev. Ecol. Syst. **24:** 441–466.

LEHMANN, E. L., 1993 The Fisher, Newman-Pearson theories of testing hypotheses: one theory or two? J. Am. Statist. Assoc. **88:** 1242–1249.

LEHMANN, E. L., 1994 *Testing Statistical Hypotheses*, Ed. 2, Chapman & Hall, New York.

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

LEVENE, H., 1949 On a matching problem arising in genetics. Ann. Math. Stat. **20:** 91–94.

LI, C. C., and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

LOUIS, E. J., and E. R. DEMPSTER, 1987 An exact test for Hardy-Weinberg and multiple alleles. Biometrics **43:** 805–811.

MORIN, P. A., J. J. MOORE, R. CHAKRABORTY, L. JIN, J. GOODALL *et al.*, 1994 Kin selection, social structure, gene flow, and the evolution of chimpanzees. Science **265:** 1193–1201.

PAMILO, P., and S. VARVIO-AHO, 1984 Testing genotype frequencies and heterozygosities. Mar. Biol. **79:** 99–100.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY and W. T. VETTERLING, 1988 *Numerical Recipes in C*. Cambridge University Press, Cambridge.

RAYMOND, M., and F. ROUSSET, 1995 GENEPOP Version 1.2, a population genetics software for exact tests and ecumenicism. J. Hered. (in press).

ROBERTSON, A., and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: Sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

STUART, A., and J. K. ORD, 1991 *Kendall's Advanced Theory of Statistics*, Vol. 2., Ed. 5, Edward Arnold, London.

THORNHILL, N. W., (Editor), 1993 *The Natural History of Inbreeding and Outbreeding*. University of Chicago Press, Chicago.

TOCHER, K. D., 1950 Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika **37:** 130–144.

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

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

YASUDA, N., 1968 Estimation of the inbreeding coefficient from phenotypic frequencies by a method of maximum likelihood scoring. Biometrics **24:** 915–935.

YATES, F., 1955 The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments. Biometrika **42:** 382–403.

Communicating editor: M. SLATKIN

APPENDIX

We first show that $\hat{f}_A$ converges in probability to $f = 0$. Generally eq. 8 has a unique root within $]f_{min}, 1[$, where

$$f_{min} = -\inf_{n_{ii}>0}\left\{\frac{\tilde{p}_i}{1-\tilde{p}_i}\right\}.$$

If there are at least two nonempty homozygote classes, then $\hat{f}_A > f_{\min} \geq -1/2$. If there is only one nonempty homozygote class (say $n_{kk} > 0$), then $\hat{f}_A = (N\tilde{p}_k - n_{kk})/(N(1-\tilde{p}_k)) \geq -1$. If there is no homozygote, $\hat{f}_A = -1$ is a suitable setting.

Asymptotically the probability that $\hat{f}_A = -1$ or 1 is null hypothesis. Otherwise, writing

$$g(S, x) = -\left.\frac{\partial \log L_2(S, f, \mathbf{p})}{N\partial f}\right|_{\mathbf{p}=\tilde{\mathbf{p}}, f=x},$$

by Taylor's theorem, for some c between $\hat{f}_A$ and 0, $g(S, \hat{f}_A) = g(S, 0) + g'(S, c)\hat{f}_A$. By definition of $\hat{f}_A$ (eq. 8), $g(S, \hat{f}_A) = 0$ so $\hat{f}_A = -g(S, 0)/g'(S, c)$.

All observed genic ($\tilde{p}_i$) and genotypic ($\tilde{p}_{ij}$) frequencies converge in probability to the expected values. Then

$$g(S, 0) = -\sum_{j<i}\tilde{p}_{ij} + \sum_i \tilde{p}_{ii}\frac{1-\tilde{p}_i}{\tilde{p}_i} \qquad (A.1)$$

converges in probability to 0, and for all $c(1 > c > f_{min} \geq -1)$,

$$g'(S, c) = -\frac{\Sigma_{j<i}\tilde{p}_{ij}}{(1-c)^2} - \sum_i \tilde{p}_{ii}\left(\frac{1-\tilde{p}_i}{\tilde{p}_i + (1-\tilde{p}_i)c}\right)^2 \qquad (A.2)$$

$$< -\frac{\Sigma_{j<i}\tilde{p}_{ij}}{4} - \sum_i \tilde{p}_{ii}(1-\tilde{p}_i)^2 \qquad (A.3)$$

that converges to a strictly negative value. Hence $\hat{f}_A$ converges in probability to $f = 0$.

Large sample variances: They have been computed for $\hat{f}$ (YASUDA 1968) and $\hat{f}_T$ (CURIE-COHEN 1982; ROBERTSON and HILL 1984). In the present case the standard line of reasoning for maximum likelihood estimates can be simplified, as first shown below, and it appears that the variances of the three estimates can be obtained by the same method.

$\hat{f}$: Considering $\partial \log L_2/\partial f$, by Taylor's theorem, for some f^* between 0 and $\hat{f}$ and $\mathbf{p}^*$ between $\mathbf{p}$ and $\hat{\mathbf{p}}$,

$$\left.\frac{\partial \log L}{\partial f}\right|_{0,\mathbf{p}} = \left.\frac{\partial \log L}{\partial f}\right|_{\hat{f},\hat{\mathbf{p}}} - \hat{f}\left.\frac{\partial^2 \log L}{\partial f^2}\right|_{f^*,\mathbf{p}^*} + \sum_j^{k-1}(p_j - \hat{p}_j)\left.\frac{\partial^2 \log L}{\partial f \partial p_j}\right|_{f^*,\mathbf{p}^*}. \qquad (A.4)$$

The first term on the right member is null (by definition of $\hat{f}$), so

$$\frac{1}{\sqrt{N}}\frac{\partial \log L}{\partial f}\bigg|_{0,\mathbf{p}} = -\sqrt{N}\,\hat{f}\,\frac{1}{N}\frac{\partial^2 \log L}{\partial f^2}\bigg|_{f^*,\mathbf{p}^*} + \sum_{j}^{k-1}\sqrt{N}\,(p_j - \hat{p}_j)\,\frac{1}{N}\frac{\partial^2 \log L}{\partial f\,\partial p_j}\bigg|_{f^*,\mathbf{p}^*}. \quad \text{(A.5)}$$

By standard arguments (*e.g.*, STUART and ORD 1991),

$$\frac{1}{N}\frac{\partial^2 \log L}{\partial f^2}\bigg|_{f^*,\mathbf{p}^*}$$

and

$$\frac{1}{N}\frac{\partial^2 \log L}{\partial f\,\partial p_j}\bigg|_{f^*,\mathbf{p}^*}$$

converge to their expectations at $(0, \mathbf{p})$, which are $1 - k$ and 0, respectively, and

$$\frac{1}{\sqrt{N}}\frac{\partial \log L}{\partial f}\bigg|_{0,\mathbf{p}}$$

is asymptotically normal with variance $k - 1$. Therefore $\hat{f}$ has variance $1/(N(k-1))$.

$\hat{f}_A$: For some f^* between 0 and $\hat{f}_A$ and $\mathbf{p}^*$ between $\mathbf{p}$ and $\tilde{\mathbf{p}}$,

$$\frac{\partial \log L}{\partial f}\bigg|_{0,\mathbf{p}} = \frac{\partial \log L}{\partial f}\bigg|_{\hat{f}_A,\tilde{\mathbf{p}}} - \hat{f}_A\,\frac{\partial^2 \log L}{\partial f^2}\bigg|_{f^*,\mathbf{p}^*} + \sum_{j}^{k-1}(p_j - \tilde{p}_j)\,\frac{\partial^2 \log L}{\partial f\,\partial p_j}\bigg|_{f^*,\mathbf{p}^*}. \quad \text{(A.6)}$$

The first term on the right member is null by definition of $\hat{f}_A$ and by exactly the same line of reasoning as for $\hat{f}$, it is shown that $\hat{f}_A$ is asymptotically normal with variance $1/(N(k-1))$.

$\hat{f}_T$: In the same way

$$\frac{\partial \log L}{\partial f}\bigg|_{0,\mathbf{p}} = \frac{\partial \log L}{\partial f}\bigg|_{0,\tilde{\mathbf{p}}} + \sum_{j}^{k-1}(p_j - \tilde{p}_j)\,\frac{\partial^2 \log L}{\partial f\,\partial p_j}\bigg|_{0,\mathbf{p}^*}, \quad \text{(A.7)}$$

where the first term on the right member is U. So $U/\sqrt{N}$ is also asymptotically normal with variance $k - 1$, and by eq. 10, $\hat{f}_T$ also has variance $1/(N(k-1))$.