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, 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.

DEFINITION OF TESTS

Definition of alternative hypotheses: Consider a population in which there are k alleles at frequencies p = (p1, ..., pk). Any deviation from Hardy-Weinberg proportions can be summarized by the set f of parameters f = (f1, ..., fk) such that the probability of some sample S of n individuals, composed of ng genotypes AiA′j (i ≠ j), is as follows (model 1) :

\[ L_1(S, f, p) = \frac{N!}{n_1! n_2! \cdots n_k!} \left( p_1^{n_1} + p_2 \sum_{j=1}^k f_j p_j \right)^{n_1} \times \left( 2p_1 p_2 (1 - f_{12}) \right)^{n_{12}} \cdots \left( p_k^2 + p_k \sum_{j=k}^k f_{jk} p_j \right)^{n_k}. \]

In some situations (e.g., regular inbreeding) all f are identical to some value f0 so that this probability is as follows (model 2) :

\[ L_2(S, f, p) = \frac{N!}{n_1! n_2! \cdots n_k!} \left( p_1^{n_1} + f_0 p_k (1 - p_k) \right)^{n_1} \times \left( 2p_1 p_k (1 - f) \right)^{n_{12}} \cdots \left( p_k^2 + f_0 (1 - p_k) \right)^{n_k}. \]

Exact tests: The null hypothesis is f = 0. This is a composite hypothesis, because it does not specify unique values for p (STUART and ORD 1991). For this reason it is necessary to consider a probability distribution deduced from the above one and independent of p when f = 0. Considering the set ′ of all possible samples S, LEVENE (1949) has shown that the conditional probability

\[ \Pr(S) = \frac{L_1(S, 0, p)}{L_1(S, f, p) + L_1(S, 0, p)} = \frac{L_2(S, 0, p)}{L_2(S, f, p) + L_2(S, 0, p)} \]

is independent of p, and its value is

\[ \Pr(S) = \frac{N! \prod_{j=1}^k n_j!}{(2N)! \prod_{j=1}^k n_j!} 2^{(\sum_j r_j - n)} \]

A deviation from Hardy-Weinberg proportions indicates either selection, population mixing or non-random 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 RENZO 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 (PAMILIO 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.
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., \( \sum_{Pr(S)<Pr(S')} Pr(S) \). This is the exact test of most authors (Louis and Dempster 1987; Hernaldez 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 \( p \), 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 choice of the probability of a particular configuration as a test statistic.

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

\[ X^2 \text{ (exact) test: Its statistic is the well-known chi square, i.e.,} \]

\[ X^2 = \left( \sum_i \left( \frac{n_i - Np_i^2}{Np_i^2} \right)^2 \right) + \left( \sum_{i<j} \left( \frac{n_{ij} - 2Np_ip_j}{2Np_ip_j} \right)^2 \right) \]

(Li and Horvitz 1953), where \( p = (p_1 \cdots p_k) \) are the observed allele frequencies in the sample.

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

\[ G(S) = \sum_i n_i \log \left( \frac{\hat{p}_i}{\hat{p}_i} \right) + \sum_{i<j} n_{ij} \log \left( \frac{\hat{p}_i \hat{p}_j}{\hat{p}_i \hat{p}_j} \right), \]

where the \( \hat{p}_i \) are the observed genotype frequencies (e.g., Hernaldez and Weir 1989).

\[ \hat{f}_c \text{ or } N_{Het}: \hat{f}_c \text{ is the estimator of } f \text{ defined by Cock-} \]

\[ \text{erham (e.g., Cockerham 1973; Weir and Cockerham 1984). It can be used as a test statistic and has a simple relation with the number of heterozygotes } N_{Het} \text{ (easily derived from Cockerham 1973):} \]

\[ \hat{f}_c = \frac{4N^2 - \sum n_i^2 - 2(2N-1)N_{Het}}{4N^2 - \sum n_i^2 - 2N_{Het}}. \]

Within the set \( C \) 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 \( \sum_{N_{Het}(s) = N_{Het}} Pr(S) \) if the alternative hypothesis is heterozygote deficiency and \( \sum_{N_{Het}(s) = N_{Het}} Pr(S) \) for heterozygote excess. Thus two one-tailed tests can be defined by these two definitions of \( P \) value.

\[ f \text{ and } \hat{f}_a: \text{The second model (eq. 2) suggests other statistics, such as the maximum likelihood estimate } f \text{ of } \]

\[ f \text{ that, jointly with maximum likelihood estimates } \hat{p} \text{ of allele frequencies, maximizes } L_a(S, f, p). \text{ When there are more than two alleles, numerical methods are necessary to find this maximum and the estimates } f \text{ and } \hat{p}, \text{ so that another estimate } \hat{f}_a \text{ of } f \text{ that maximizes the likelihood } L_a \text{ if the vector of observed frequencies } \hat{p} \text{ is taken as estimate of } p \text{ has been suggested as a suitable alternative (Li and Horvitz 1953; Curie-Cohen 1982). } \hat{f}_a \text{ is the unique root of} \]

\[ -\frac{\partial \log L_a}{\partial f} \left|_{p=\hat{p}} \right. = \sum_i n_i \frac{1-\hat{p}_i}{\hat{p}_i + (1-\hat{p}_i)f} \]

\[ - \frac{N_{Het}}{1-f} = 0 \]

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

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

\[ \left. \frac{\partial \log L_a}{\partial f} \right|_{f=0} = \sum_{i=1}^k \frac{n_i}{\hat{p}_i} - N. \]

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 \( C, \hat{f}_T \) is a (monotone) function of \( U \) only:

\[ \hat{f}_T = \frac{(2N-1) \left( 1 + U/N \right) - (2N-k)}{2(N-1)(k-1)}. \]

**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 \( \lambda \) of likelihoods under the null hypothesis and under an alternative hypothesis, such as \( \lambda(S) = L_a(S, 0, p) / L_a(S, f, p) \). \( \lambda \) must not be confused with a likelihood ratio statistic as it depends on unknown
2 kinds of deviations from Hardy-Weinberg proportions

A completely specified alternative hypothesis; these power comparisons where samples are generated from more than one allelic type. The upper bound is computed as described in text parameters, \( f \) and \( p \) in this example. In the two-allele case, it can be checked that the ranking of \( \lambda \) is independent of \( p \) within \( \ell \) and depends only on the sign of \( f \) and the value of \( n_2 \), 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 \( \lambda \) 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 \( \lambda \) 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 \( \lambda \), 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 \( \S 10.4 \) of PRESS et al. (1988) was used to estimate \( f \) and \( f_\ell \) 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 1

Heterozygote Excess and Deficiency

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( p )</th>
<th>( N )</th>
<th>Power of tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ell = 2 ) alleles:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-1/6)</td>
<td>100</td>
<td>((0.25 0.75))</td>
<td>0.436</td>
</tr>
<tr>
<td>(1/8)</td>
<td>50</td>
<td>((0.45 0.55))</td>
<td>0.1562</td>
</tr>
<tr>
<td>(1/4)</td>
<td>50</td>
<td>((0.45 0.55))</td>
<td>0.4505</td>
</tr>
<tr>
<td>(1/2)</td>
<td>20</td>
<td>((0.25 0.75))</td>
<td>0.4700</td>
</tr>
<tr>
<td>( \ell = 3 ) alleles:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-1/8)</td>
<td>50</td>
<td>(\pi_{3,1} = (0.5 0.3 0.2))</td>
<td>0.3541</td>
</tr>
<tr>
<td>(-1/8)</td>
<td>50</td>
<td>(\pi_{3,2} = (0.36 0.33 0.33))</td>
<td>0.3410</td>
</tr>
<tr>
<td>(1/10)</td>
<td>100</td>
<td>(\pi_{3,1})</td>
<td>0.392</td>
</tr>
<tr>
<td>(1/10)</td>
<td>100</td>
<td>(\pi_{3,2})</td>
<td>0.422</td>
</tr>
<tr>
<td>(1/10)</td>
<td>100</td>
<td>(\pi_{3,3} = (0.7 0.2 0.1))</td>
<td>0.3676</td>
</tr>
<tr>
<td>(1/8)</td>
<td>50</td>
<td>(\pi_{3,1})</td>
<td>0.3390</td>
</tr>
<tr>
<td>(1/8)</td>
<td>50</td>
<td>(\pi_{3,2})</td>
<td>0.3437</td>
</tr>
<tr>
<td>(1/8)</td>
<td>50</td>
<td>(\pi_{3,3})</td>
<td>0.2997</td>
</tr>
<tr>
<td>(1/4)</td>
<td>20</td>
<td>(\pi_{3,1})</td>
<td>0.4218</td>
</tr>
<tr>
<td>(1/4)</td>
<td>20</td>
<td>(\pi_{3,2})</td>
<td>0.4226</td>
</tr>
<tr>
<td>(1/4)</td>
<td>20</td>
<td>(\pi_{3,3})</td>
<td>0.3405</td>
</tr>
<tr>
<td>(1/4)</td>
<td>50</td>
<td>(\pi_{3,1})</td>
<td>0.7536</td>
</tr>
<tr>
<td>(1/4)</td>
<td>50</td>
<td>(\pi_{3,2})</td>
<td>0.7736</td>
</tr>
<tr>
<td>(1/4)</td>
<td>50</td>
<td>(\pi_{3,3})</td>
<td>0.6844</td>
</tr>
<tr>
<td>(1/2)</td>
<td>20</td>
<td>(\pi_{3,1})</td>
<td>0.8871</td>
</tr>
<tr>
<td>(1/2)</td>
<td>20</td>
<td>(\pi_{3,2})</td>
<td>0.7636</td>
</tr>
</tbody>
</table>

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 \( f \), computed as in GUO and THOMPSON (1992), was below 0.005. Lower \( P \) values have lower SEs. \( \pi_{3,1} = (0.22 0.21 0.20 0.19 0.18) \), \( \pi_{3,2} = (0.30 0.25 0.20 0.15 0.10) \), and \( \pi_{3} = (0.16 0.15 0.14 0.13 0.12 0.11 0.10 0.09) \).
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{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 $X^2$ 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 $\hat{f}_T$, $U$, and $N_{\text{HET}}$ 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 $N_{\text{HET}}$, 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, $\hat{f}_T$ being slightly better for heterozygote deficiency. $N_{\text{HET}}$ 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 $\hat{f}_T$ 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 $f_0$ (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 Wåhund 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 $f_0$ 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; Robertsson and Hill 1984; and see APPENDIX). The variance of $\hat{f}_T$ is also $1/(N(k-1))$ when allele frequencies are identical but is generally larger (Chakraborty and Danker-Hoppe 1991, eq. 4.2b). The score ($U$) test being identical to the $\hat{f}_T$ test and $N_{\text{HET}}$ to $\hat{f}_T$, it is expected that the score and $\hat{f}_T$ tests will be generally more powerful than $N_{\text{HET}}$ when $f$ is small. This is indeed observed even for large values of $f$, $N_{\text{HET}}$ being less efficient. An estimate of power can be deduced from the large sample normality and variance of $\hat{f}_T$ and $\hat{f}_A$ 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 $\geq 0$ 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 Wahlund effect. In all of these cases, the $\hat{f}_A$, $U$, and $N_{\text{HET}}$ 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)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Power of tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>( N )</td>
</tr>
<tr>
<td>(-63/208)</td>
<td>0</td>
</tr>
<tr>
<td>(-9/16)</td>
<td>0</td>
</tr>
<tr>
<td>(-1/6)</td>
<td>3/22</td>
</tr>
<tr>
<td>(-1/11)</td>
<td>9/319</td>
</tr>
<tr>
<td>(-1/33)</td>
<td>0</td>
</tr>
<tr>
<td>(-1/33)</td>
<td>1/9</td>
</tr>
<tr>
<td>(-9/49)</td>
<td>9/91</td>
</tr>
<tr>
<td>(-9/16)</td>
<td>0</td>
</tr>
</tbody>
</table>

Each comparison is based on 10,000 independent samples. See Table 1 for SEs. For convenience we give \( f \) as a semimatrix in which \( jz = X_{ij} / (1 - p_i) \) (since \( f_{ij} = f_j \)).

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

\[
\sum_{i=1}^{N_{\text{HWE}}} \frac{n_i}{1 - f} - \sum_{i=1}^{N_{\text{HWE}}} \frac{n_i}{n_i - \hat{f}_i} = 0
\]

and

\[
U = \sum_i \sum_j \frac{n_{ji}}{1 - \hat{f}_j} - \sum_i N_i,
\]

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 (\textit{e.g.}, \textit{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 (\textit{e.g.}, \textit{JARNE} and \textit{CHARLESWORTH} 1993; \textit{THORNHILL} 1993), but the power remains low for some applications (\textit{ROBERTSON} and \textit{HILL} 1984). In such cases, the multisample \( U \) test will be useful.

M.R. thanks P. \textit{PAMULO} for the opportunity to spend a year at the Department of Genetics of the University of Uppsala. We thank P. \textit{DAVID}, J. \textit{GOLODET}, P. \textit{JARNE}, P. \textit{PAMULO} and particularly J.-D. \textit{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 Bioystematique de Montpellier. This is paper 95-038 of the Institut des Sciences de l'\'evoluti0n.

LITERATURE CITED


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 \( \{ f = 0 \} \),

\[
\min = \inf_{n \geq 0} \left\{ \frac{\hat{p}_i}{1 - \hat{p}_i} \right\}.
\]

If there are at least two nonempty homozygote classes, then \( \hat{f}_a > \min \approx -1/2 \). If there is only one nonempty homozygote class (say \( n_0 > 0 \)), then \( \hat{f}_a = (N \hat{p}_0 - n_0) / (N(1 - \hat{p}_0)) \approx -1 \). 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

\[
g(S, c) = \frac{-\partial \log L(S, f, \text{p})}{N df} \bigg|_{p=p^*, f=f^*}
\]

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 (\( \hat{p}_i \)) and genotypic (\( \hat{p}_y \)) frequencies converge in probability to the expected values. Then

\[
g(S, 0) = -\sum_{j<i} \hat{p}_j + \sum_i \hat{p}_i \frac{1 - \hat{p}_i}{\hat{p}_i} (A.1)
\]

converges in probability to 0, and for all \( c (1 > c > \min \approx -1) \),

\[
g'(S, c) = -\frac{\sum_{j<i} \hat{p}_j}{(1 - c)^2} - \sum_i \hat{p}_i \left( \frac{1 - \hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i) c} \right)^2 (A.2)
\]

\[
< -\frac{\sum_{j<i} \hat{p}_j}{4} - \sum_i \hat{p}_i (1 - \hat{p}_i)^2 (A.3)
\]

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 \( \hat{f}_a \) (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.

\[
\frac{\partial \log L}{df} \bigg|_{0, \text{p}} = \frac{\partial \log L}{df} \bigg|_{\text{p}^*, \text{p}} = \frac{\partial^2 \log L}{df^2} \bigg|_{\text{p}^*, \text{p}}
\]

The first term on the right member is null (by definition of \( \hat{f} \)), so
By standard arguments (e.g., Stuart and Ord 1991),

\[
\frac{1}{N} \frac{\partial^2 \log L}{\partial f^2} \bigg|_{\hat{f} \neq \hat{p}}
\]

and

\[
\frac{1}{N} \frac{\partial^2 \log L}{\partial f \partial p_j} \bigg|_{\hat{f} \neq \hat{p}}
\]

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

\[
\frac{1}{N} \frac{\partial \log L}{\partial f} \bigg|_{0,p}
\]
is asymptotically normal with variance \(k - 1\). Therefore \(\hat{f}\) has variance \(1/(N(k - 1))\).

\[\hat{f}_A: \text{For some} f^* \text{ between 0 and} \hat{f}_A \text{ and} p^* \text{ between} p \text{ and} \hat{p},\]

\[
\frac{\partial \log L}{\partial f} \bigg|_{0,p} = \frac{\partial \log L}{\partial f} \bigg|_{\hat{f}_A \neq \hat{p}} - \hat{f}_A \frac{\partial^2 \log L}{\partial f^2} \bigg|_{\hat{f}_A \neq \hat{p}}
\]

\[
+ \sum_{j=1}^{k-1} (p_j - \hat{p}_j) \frac{\partial^2 \log L}{\partial f \partial p_j} \bigg|_{\hat{f}_A \neq \hat{p}}. \quad (A.6)
\]

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 \(f\), it is shown that \(\hat{f}_A\) is asymptotically normal with variance \(1/(N(k - 1))\).

\[\hat{f}_T: \text{In the same way}\]

\[
\frac{\partial \log L}{\partial f} \bigg|_{0,p} = \frac{\partial \log L}{\partial f} \bigg|_{\hat{f}_T \neq \hat{p}}
\]

\[
+ \sum_{j=1}^{k-1} (p_j - \hat{p}_j) \frac{\partial^2 \log L}{\partial f \partial p_j} \bigg|_{\hat{f}_T \neq \hat{p}}. \quad (A.7)
\]

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