RESEARCH ARTICLE

A new method of testing for Hardy–Weinberg equilibrium and ordering populations

NADER EBRAHIMI* and DEVRIM BILGILI

Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA

Abstract

The assumption of Hardy–Weinberg equilibrium (HWE) among alleles in a nonevolving population is of fundamental importance in genetic studies. Deviation from HWE in a population usually indicates inbreeding, stratification and sometimes problems in genotyping. In populations of affected individuals, these deviations can also provide evidence for association. In this paper, we introduce a measure based on the Kullback–Leibler discrimination information function that quantifies the deviation from HWE in a population. We use this measure to order populations. We also propose a test for HWE based on an estimate of this measure. The test is a statistically consistent test of the null hypothesis for all alternatives and is very easy to implement. Our proposed test statistic is compared with an earlier, widely used, test. Finally, the use of the proposed new test is shown in an illustrative example.

[Ebrahimi N. and Bilgili D. 2007 A new mehtod of testing for Hardy–Weinberg equilibrium and ordering populations. *J. Genet*. **86**, 1–7]

Introduction

In biology, evolution is often defined as being the sum total of the genetically inherited changes in the individuals who are the members of a population's gene pool. It is clear that, although the effects of evolution are felt by individuals, it is the population as a whole that actually evolves. In this context, microevolution can be treated simply as a change in frequencies of alleles in the gene pool of a population. For instance, suppose there is a trait that is determined by the inheritance of a gene with two alleles A_1 and A_2. If the parent generation has 90% A_1 and 10% A_2 and their offspring collectively have 92% A_1 and 8% A_2, evolution has occurred between the generations. The entire population's gene pool has evolved in the direction of lower frequency of the A_2 allele.

This definition of evolution was developed as a result of independent work by G. H. Hardy and W. Weinberg in 1908. Through mathematical modelling based on probability, they concluded that gene pool frequencies are inherently stable but that evolution should be expected in all populations virtually all of the time. Hardy, Weinberg, and the population geneticists who followed them came to understand that evolution will not occur in a population if the following conditions are met: (i) mutations are negligible; (ii) natural selection is not operating in the population; (iii) there is no migration in or out of the population; (iv) the population is large (i.e. there is no genetic drift); (v) all members of the population breed, individuals are mating randomly, and everyone produces the same number of offspring.

Hardy and Weinberg went on to develop a simple equation or formula in a nonevolving population that can be used to determine genotype frequencies at any locus in terms of allele frequencies. This has become known as the Hardy–Weinberg equilibrium (HWE) equation. For an m-allele autosomal locus with alleles $A_1, \ldots, A_m$, the genotypic array as given by the HWE equation is $\sum_i p_i^2 A_i A_i + 2 \sum_{i<j} p_i p_j A_i A_j$, where p_i is the allelic frequency of A_i, $i = 1, \ldots, m$. For example, for $m = 2$ in the equation $p_1^2 + 2p_1(1 - p_1) + (1 - p_1)^2 = 1$, p_1 is defined as the frequency of the dominant allele A_1 and $(1 - p_1)$ as the frequency of the recessive allele A_2 for a trait controlled by a pair of alleles A_1 and A_2. In other words, p_1 equals all of the alleles in the individuals who are homozygous dominant (A_1A_1) and half of the alleles in people who are heterozygous (A_1A_2) for this trait in a pop-

*For correspondence. E-mail: nader@math.niu.edu.

Keywords. population genetics; Kullback–Leibler discrimination information function; genetic data analysis; Hardy–Weinberg equilibrium.

ulation. In a mathematical term $p_1 = A_1A_1 + \frac{1}{2}A_1A_2$. In the HWE equation, p_1^2 is the predicted frequency of homozygous dominant A_1A_1 individuals in a population, $2p_1(1-p_1)$ is the predicted frequency of heterozygous (A_1A_2) individuals, and $(1-p_1)^2$ is the predicted frequency of homozygous recessive (A_2A_2) individuals.

The original descriptions of HWE are an important landmark in the history of population genetics, and now it is common practice to check whether observed genotypic frequencies in a population conform to Hardy–Weinberg expectations. As pointed out by Wigginton et al. (2005), these expectations appear to hold for most human populations, and deviations from HWE at a particular marker locus may suggest problems with genotyping or population structure or, in samples of affected individuals, an association between the marker and disease susceptibility.

In this paper, we introduce a measure based on Kullback–Leibler discrimination information function that quantifies the deviation from HWE in a population (see Kullback (1978) for more details about this function). We use this measure to order populations on the basis of deviations from HWE. We also propose a test based on an estimate of this measure. Several methods proposed so far for testing HWE may be seen in Yuan and Bonney (2003) and Wigginton *et al*. (2005), and references cited there. An advantage of our test is that it is consistent and very easy to implement. Also, it detects arbitrarily small deviations from HWE.

The paper is organized as follows: in the next section we describe the Kullback–Leibler discrimination information measure. In the sections after that, we derive our measure and use it to order populations, and then obtain the test statistics based on this measure for both $m = 2$ and $m > 2$ to test for HWE and discuss how to implement the proposed test. The practicality of our method is demonstrated by an example. We also compare the power of the proposed test with Yuan and Bonney (2003) test. Finally, we give some concluding remarks.

Kullback–Leibler discrimination information function

Consider a sample of genotypes for N unrelated diploid individuals measured at an autosomal locus. The sample has $2N$ alleles, including n_{A_1} copies of rarer allele and n_{A_2} copies of the common allele, with $n_{A_1} + n_{A_2} = 2N$. In this section, for simplicity we assume that $m = 2$. Let the number of heterozygous A_1A_2 genotypes be $N_{A_1A_2}$. It is clear that the numbers of A_1A_1 and A_2A_2 homozygous genotypes are $N_{A_1A_1} = \frac{n_{A_1} - N_{A_1A_2}}{2}$ and $N_{A_2A_2} = \frac{n_{A_2} - N_{A_1A_2}}{2}$ respectively.

Consider the problem of discriminating between two probability mass functions (models), say P_1 and P_2, for the random prospect $N_{A_1A_2}$ that ranges over the space S. In our setup, if n_{A_1} is an even number then $S = S_1 = \{0, 2, 4, 6, \ldots, n_{A_1}\}$, and if n_{A_1} is an odd number then $S = S_2 = \{1, 3, 5, \ldots, n_{A_1}\}$. Assuming that $N_{A_1} = n_{A_1}$ and given an observation $N_{A_1A_2} = k$, Bayes's theorem relates the likelihood ratio to the prior and posterior odds in favour of the model P_1 as follows:

$$\log \frac{P_1(N_{A_1A_2} = k)}{P_2(N_{A_1A_2} = k)} = \log \frac{P(P_1|N_{A_1A_2} = k)}{P(P_2|N_{A_1A_2} = k)} - \log \frac{P(P_1)}{P(P_2)},$$

where $P(\cdot)$ and $P(\cdot|N_{A_1A_2} = k)$ denote the prior and posterior probabilities of the model. As the difference between the posterior and prior log-odds, the logarithm of the likelihood ratio quantifies the information in $N_{A_1A_2} = k$ in favour of P_1 against P_2 (Kullback 1978).

When there is no specific information on the value of k, other than $k \in S_1$ for n_{A_1} is an even number ($k \in S_2$ for n_{A_1} is an odd number), the mean observation per k from P_1 for the discrimination information between P_1 and P_2 is

$$K(P_1 : P_2) = \sum_{k \in S} \left(\log \frac{P_1(N_{A_1A_2} = k|N_{A_1} = n_{A_1})}{P_2(N_{A_1A_2} = k|N_{A_1} = n_{A_1})} \right) \times P_1(N_{A_1A_2} = k|N_{A_1} = n_{A_1}). \quad (1)$$

Here S is S_1 if n_{A_1} is an even number and it is S_2 if n_{A_1} is an odd number. The discrimination information function (equation 1) introduced by Kullback and Leibler (1951) is the fundamental information measure for comparing two models P_1 and P_2. In this equation (1), $K(P_1 : P_2) \geq 0$ and equality holds if the two models are identical. Therefore, $K(P_1 : P_2)$ is a measure of discrepancy between two models, P_1 and P_2. Note that $K(P_1 : P_2)$ is not symmetric and, thus, it is not a distance function. It is a measure of directed divergence between two models P_1 and P_2, where P_2 is referred to as the reference model.

Ordering populations based on Hardy–Weinberg equilibrium

In this section, we introduce our measure to quantify deviation from HWE in a population and use it to order populations based on HWE. Following the notation of the previous section, it is clear that under the assumption of HWE, the conditional probability of observing exactly k heterozygotes in a sample of N individuals, given n_{A_1} rarer alleles, is

$$P_2(N_{A_1A_2} = k|N_{A_1} = n_{A_1}, N) = \frac{(2^k)N!}{k!\left(\frac{n_{A_1} - k}{2}\right)!\left(\frac{n_{A_2} - k}{2}\right)!} \times \frac{n_{A_1}!\, n_{A_2}!}{2N!}. \quad (2)$$

When n_{A_1} is odd, possible numbers of heterozygotes, k, are in $S_2 = \{1, 3, 5, \ldots, n_{A_1}\}$. When n_{A_1} is even, possible numbers of heterozygotes are in $S_1 = \{0, 2, 4, 6, \ldots, n_{A_1}\}$. Note that here HWE is used as our reference model.
Also, in general,

$$P_1(N_{A_1A_2} = k|N_{A_1} = n_{A_1}, N) = \frac{P(N_{A_1A_2} = k, N_{A_1} = n_{A_1}|N)}{P(N_{A_1} = n_{A_1}|N)}$$

$$= \frac{f(k, n_{A_1}; N)}{C_{n_{A_1}}}, \quad (3)$$

where $f(k, n_{A_1}; N) = \dfrac{(2^k)N!(P_{12})^k(P_{11})^{\frac{n_{A_1}-k}{2}}(P_{22})^{\frac{n_{A_2}-k}{2}}}{k!\left(\frac{n_{A_1}-k}{2}\right)!\left(\frac{n_{A_2}-k}{2}\right)!}$, $C_{n_{A_1}} = \sum_{k\in S_1} f(k, n_{A_1}; N)$ if n_{A_1} is an even number, $C_{n_{A_1}} = \sum_{k\in S_2} f(k, n_{A_1}; N)$ if n_{A_1} is an odd number, P_{12} is the probability of heterozygotes, P_{11} is the probability of A_1A_1 and P_{22} is the probability of A_2A_2. Note that in equation 3 $P_{11}, P_{12}, P_{22} > 0$ and $P_{11} + 2P_{12} + P_{22} = 1$.

We note that if $P_{12} = 0$, then $f(0, n_{A_1}; N) = \dfrac{(P_{11})^{\frac{n_{A_1}}{2}}(P_{22})^{\frac{n_{A_2}}{2}}}{\left(\frac{n_{A_1}}{2}\right)!\left(\frac{n_{A_1}}{2}\right)!}$ and $f(k, n_{A_1}; N) = 0$, $k \neq 0$. Similarly, if $P_{11} = 0$, then $f(k, k; N) = \dfrac{2^k(P_{12})^k(P_{22})^{\frac{n_{A_2}-k}{2}}}{k!\left(\frac{n_{A_2}-k}{2}\right)!}$ and $f(k, n_{A_1}; N) = 0$ if $k \neq n_{A_1}$. If $P_{22} = 0$, then $f(k, 2N-k; N) = \dfrac{2^kN!(P_{12})^k(P_{11})^{N-k}}{k!(N-k)!}$. If $P_{11} = P_{12} = 0$, then $P_{22} = 1$ and $f(0, 0; N) = 1$ and $f(k, n_{A_1}; N) = 0$ if $k \neq 0$ and $n_{A_1} = 0$. If $P_{11} = P_{22} = 0$, then $P_{12} = \dfrac{1}{2}$ and $f(k, k; N) = 1$. Finally, if $P_{12} = P_{22} = 0$, then $P_{11} = 1$, and $f(0, 2N; N) = 1$.

To quantify deviation from HWE for a population, using equations 1, 2 and 3, we define

$$\begin{aligned} K(P_1 : P_2) &= \sum_{k\in S} P_1(N_{A_1A_2} = k|N_{A_1} = n_{A_1}, N)\times \\ &\qquad \log\frac{P_1(N_{A_1A_2} = k|N_{A_1} = n_{A_1}, N)}{P_2(N_{A_1A_2} = k|N_{A_1} = n_{A_1}, N)} \\ &= \sum_{k\in S}(f(k, n_{A_1}; N)/C_{n_{A_1}})\times \\ &\qquad \log\frac{(P_{12})^k(P_{11})^{\frac{n_{A_1}-k}{2}}(P_{22})^{\frac{n_{A_2}-k}{2}}}{\left(\frac{n_{A_1}!\,n_{A_2}!}{(2N)!}\right)C_{n_{A_1}}}. \end{aligned} \quad (4)$$

In equation 4, $S = S_1$ if n_{A_1} is an even number and $S = S_2$ if n_{A_1} is an odd number. Now, using the fact that in general, when n_{A_1} is an even number,

$$\begin{aligned} \mathrm{E}_1(N_{A_1A_2}|N_{A_1} = n_{A_1}, N) &= \frac{2NP_{12}}{C_{n_{A_1}}}\times \\ &\sum_{\substack{k=2\\ k \text{ is even}}}^{n_{A_1}} \frac{(N-1)!2^{k-1}}{(k-1)!\left(\frac{n_{A_1}-k}{2}\right)!\left(\frac{n_{A_2}-k}{2}\right)!\left(\frac{n_{A_2}-k}{2}\right)!}\times \\ &P_{12}^{k-1}(P_{11})^{\frac{n_{A_1}-k}{2}}(P_{22})^{\frac{n_{A_2}-k}{2}} \\ &= \frac{2NP_{12}}{C_{n_{A_1}}}\sum_{\substack{y=1\\ y \text{ is odd}}}^{n_{A_1}-1} f(y, n_{A_1}-1; N-1) \\ &= \frac{2NP_{12}}{C_{n_{A_1}}}C^*_{n_{A_1}}, \end{aligned} \quad (5)$$

where $C^*_{n_{A_1}} = \sum_{\substack{y=1\\ y \text{ is odd}}}^{n_{A_1}-1} f(y, n_{A_1}-1; N-1)$. When n_{A_1} is an odd number,

$$\mathrm{E}_1(N_{A_1A_2}|N_{A_1} = n_{A_1}, N) = \frac{2NP_{12}}{C_{n_{A_1}}}C^*_{n_{A_1}}, \quad (6)$$

where $C^*_{n_{A_1}} = \sum_{\substack{y=0\\ y \text{ is even}}}^{n_{A_1}-1} f(y, n_{A_1}-1; N-1)$. Here E_1 stands for the expected value under the general model P_1.

Using equations 5 and 6, equation 4 reduces to

$$\begin{aligned} K(P_1 : P_2) = {} & \log 2N! - \log n_{A_1}! - \log n_{A_2}! - \log Cn_{A_1} \\ & + \frac{2NP_{12}C^*_{n_{A_1}}}{C_{n_{A_1}}}\log P_{12} + \left(\frac{n_{A_1}}{2} - \frac{NP_{12}C^*_{n_{A_1}}}{C_{n_{A_1}}}\right)\times \\ & \log P_{11} + \left(\frac{n_{A_2}}{2} - \frac{NP_{12}C^*_{n_{A_1}}}{C_{n_{A_1}}}\right)\log P_{22}. \end{aligned} \quad (7)$$

For a given population of size N with a known number of A_1 alleles, n_{A_1}, and P_{11}, P_{12} and P_{22}, $K(P_1 : P_2)$ in equation 7 quantifies the deviation from HWE. Thus, $K(P_1 : P_2)$ can be used as a measure of disequilibrium.

We can therefore use the following criteria to order populations. Population 1 with a known number of A_1 alleles, n_{A_1}, and P_{11}, P_{12} and P_{22} is said to be closer to HWE than Population 2 with a known number of A_1 alleles, $n^*_{A_1}$, and $P^*_{11}P^*_{12}$ and P^*_{22} if

$$K(P_1 : P_2) \leq K(P^*_1 : P^*_2). \quad (8)$$

Table 1 gives the value of $K(P_1 : P_2)$ for different values of $N, n_{A_1}, P_{11}, P_{12}$ and P_{22}. For example, if $N = 100$, then from table 1 it is clear that the population with $n_{A_1} = 40$ and $P_{11} = 0.2, P_{12} = 0.3$ and $P_{22} = 0.2$ is closer to HWE than the population with $n^*_{A_1} = 40$ and $P^*_{11} = 0.8$, $P^*_{12} = 0.05$ and $P^*_{22} = 0.1$.

Testing for Hardy–Weinberg equilibrium

In this section we derive a test statistic to test for HWE, describe the implementation of the test, and illustrate it with an example.

Test statistic

To construct our test statistic for testing HWE based on equation 7, since evaluation of $K(P_1 : P_2)$ requires a complete knowledge of P_{11}, P_{12} and P_{22}, we operationalize this equation by developing the discrimination information statistic as follows. It is clear that a reasonable estimate of P_{11}, P_{12} and P_{22} are $\hat{P}_{11} = \dfrac{N_{A_1A_2}}{N}, \hat{P}_{12} = \dfrac{N_{A_1A_2}}{2N}$ and $\hat{P}_{22} = \dfrac{N_{A_2A_2}}{N}$

Table 1. $K(P_1 : P_2)$ for different values of n_{A_1}, P_{11}, P_{12} and P_{22}.

N	n_{A_1}	$P_{11} = 0.8, P_{12} = 0.05, P_{22} = 0.1$	$P_{11} = 0.6, P_{12} = 0.1, P_{22} = 0.2$	$P_{11} = 0.2, P_{12} = 0.3, P_{22} = 0.2$
10	3	1.1966	0.7135	0.0294
10	4	2.0343	1.0485	0.0564
10	5	1.9710	1.2015	0.0888
10	7	2.4488	1.5125	0.1579
10	8	2.8699	1.6384	0.1868
20	3	1.3013	0.6293	0.0147
20	5	2.3345	1.2234	0.0474
20	8	3.5966	1.9834	0.1238
20	10	4.2075	2.3921	0.1868
20	15	5.1995	3.0803	0.3432
50	5	2.1673	0.9220	0.0194
50	8	3.7445	1.7845	0.0532
50	10	4.7264	2.3580	0.0843
50	20	8.7786	4.8918	0.3237
50	40	13.1167	7.7852	0.9090
100	10	4.3705	1.8967	0.0431446
100	20	9.4127	4.7237	0.1761470
100	40	17.4602	9.7417	0.6562550
100	80	26.0990	15.4921	1.8121200
100	90	26.8816	16.0249	1.9690700

respectively. This produces an estimate of $K(P_1 : P_2)$, the discrimination information statistic, as

$$K_N(P_1 : P_2) = \log 2N! - \log n_{A_1}! - \log n_{A_2}! - \log \hat{C}_{n_{A_1}} + \frac{2N\hat{P}_{12}\hat{C}^*_{n_{A_1}}}{\hat{C}_{N_{A_1}}} \log \hat{P}_{12} + \left(\frac{n_{A_1}}{2} - \frac{N\hat{P}_{12}\hat{C}^*_{n_{A_1}}}{\hat{C}_{n_{A_1}}}\right) \times \log \hat{P}_{11} + \left(\frac{n_{A_2}}{2} - \frac{N\hat{P}_{12}\hat{C}^*_{n_{A_1}}}{\hat{C}_{n_{A_1}}}\right) \log \hat{P}_{22}, \quad (9)$$

where $\hat{C}_{n_{A_1}}$ and $\hat{C}^*_{n_{A_1}}$ are obtained by replacing P_{11}, P_{22} and P_{12} with $\hat{P}_{11}$, $\hat{P}_{22}$, and $\hat{P}_{12}$ respectively in $C_{n_{A_1}}$ and $C^*_{n_{A_1}}$. This test is consistent, and large values of $K_N(P_1 : P_2)$ indicate that the population does not follow HWE.

Implementation of the test

The discrimination information statistic $K_N(P_1 : P_2)$ is very easy to compute. However, the sampling distribution is difficult to obtain in closed form. To implement our method we need to study the sampling distribution of the proposed test.

From Kullback (1978), under the HWE for large N, $2K_N(\hat{P}_1 : \hat{P}_2)$ has an approximately chi-square distribution with two degrees of freedom. Thus, for large N, we reject the hypothesis that the population does not follow HWE at the significance level α if

$$2K_N(P_1 : P_2) \geq \chi^2_{2,\alpha}. \quad (10)$$

For small N and different values for n_{A_1} we determine the $C_{N,n_{A_1}}(\alpha)$. Here $C_{N,n_{A_1}}(\alpha)$ is the upper α-quantile of the distribution $K_N(P_1 : P_2)$ under HWE. That is, $P(K_N(P_1 : P_2) \geq C_{N,n_{A_1}}(\alpha)) = \alpha$. We reject the null hypothesis that the population does not follow HWE at the significance level α if $K_N(P_1 : P_2) \geq C_{N,n_{A_1}}(\alpha)$. Table 2 gives $C_{N,n_{A_1}}(\alpha)$ for selected values of N and n_{A_1}. The program to obtain $C_{N,n_{A_1}}(\alpha)$ for other values of N and n_{A_1}, written in Mathematica, is available on request from the authors. As applications of table 2, consider a population of size $N = 50$ with $n_{A_1} = 25$. Suppose $K_N(P_1 : P_2) = 1.98$. From table 2, it is clear that the null hypothesis should not be rejected at $\alpha = 0.1$. That is, the population follows HWE at the significance level $\alpha = 0.1$. Now, suppose $K_N(P_1 : P_2) = 5.3$, then from table 2 it is clear that the null hypothesis should be rejected at $\alpha = 0.01$. That is, the population does not follow HWE at the significance level $\alpha = 0.01$.

Example

The following example illustrates the procedure proposed. Spitze (1993) reported the following number of genotypes at the *PGI* locus in his *Daphnia* population in Nothing Pond: $A_1A_1 = 11$, $A_1A_2 = 55$, $A_2A_2 = 61$. In his population $N = 127$, $n_{A_1} = 77$ and $n_{A_2} = 177$. It is clear that $\hat{P}_{11} = \frac{11}{127} = 0.087$, $\hat{P}_{22} = \frac{61}{127} = 0.48$, and $\hat{P}_{12} = \frac{55}{254} = 0.216$. Using equation 9, we obtain $K_N(P_1 : P_2) = 0.037$. Therefore, from equation 10, there is no evidence ($P = 0.96$) for deviation from HWE, and thus no evidence for lack of random mating.

Simulation study

We compare the proposed test statistic K_N with the exact test statistic proposed by Yuan and Bonney (2003) based on 1000 simulations for different N, P_{11}, P_{12} and P_{22}. The exact test proposed by these authors has been shown to perform reliably and efficiently. Simulations were performed based on two models. The first model (in the first column of table 3)

Table 2. $C_{N,n_{A_1}}(\alpha)$ values of the $K_N(P_1 : P_2)$ statistics under Hardy–Weinberg equilibrium.

$N = 20$	$n_{A_1} = 6$	$C(.003) = 8.12$, $C(.0303) = 2.86$, $C(.65) = 0.44$,
	$n_{A_1} = 13$	$C(.00009) = 8.28$, $C(.0051) = 4.5$, $C(.057) = 2.94$, $C(.12) = 1.86$.
	$n_{A_1} = 18$	$C(1.48 \times 10^{-6} = 13.4)$, $C(.0004) = 7.73$, $C(.0007 = 7.55)$, $C(.01007) = 4.2$, $C(.018) = 3.88$, $C(.098) = 1.83$, $C(.18) = 1.63$.
$N = 50$	$n_{A_1} = 10$	$C(1.22 \times 10^{-7}) = 15.9$, $C(.000051) = 9.2$, $C(.0031) = 4.73$, $C(.06) = 1.8$, $C(.66) = 0.49$.
	$n_{A_1} = 25$	$C(3.8 \times 10^{-11}) = 22.7$, $C(1.3 \times 10^{-8}) = 17.4$, $C(9 \times 10^{-7}) = 12.82$, $C(.000031) = 9.2$, $C(.00051) = 6.25$, $C(.0056) = 5.02$, $C(.022) = 4.04$, $C(.05) = 3.9$, $C(.15) = 2.18$.
	$n_{A_1} = 40$	$C(3.4 \times 10^{-15}) = 33.3$, $C(4.15 \times 10^{-12}) = 25.8$, $C(7.6 \times 10^{-10}) = 20.08$, $C(5.1 \times 10^{-8}) = 15.6$, $C(8.3 \times 10^{-7}) = 14.02$, $C(1.7 \times 10^{-6}) = 11.98$, $C(.00003) = 9.8$, $C(.00006) = 8.9$, $C(.0005) = 6.5$, $C(.004) = 4.6$, $C(.0065) = 4.4$, $C(.026) = 2.92$, $C(.036) = 2.82$, $C(.087) = 1.62$, $C(.13) = 1.58$.
$N = 100$	$n_{A_1} = 25$	$C(.408 \times 10^{-15}) = 31.67$, $C(2.9 \times 10^{-12}) = 25.68$, $C(5.4 \times 10^{-10}) = 20.2$, $C(4.4 \times 10^{-8}) = 15.7$, $C(1.84 \times 10^{-6}) = 11.83$, $C(.00005) = 8.6$, $C(.00065) = 5.87$, $C(.007) = 3.7$, $C(.04) = 1.94$, $C(.22) = 1.17$.
	$n_{A_1} = 50$	$C(5.34 \times 10^{-25}) = 55.89$, $C(2.0039 \times 10^{-21}) = 47.23$, $C(2.66 \times 10^{-16}) = 40.28$, $C(3.005 \times 10^{-14}) = 34.7$, $C(1.99 \times 10^{-12}) = 29.8$, $C(8.45 \times 10^{-11}) = 25.49$, $C(2.44 \times 10^{-9}) = 21.7$, $C(5 \times 10^{-8}) = 18.25$, $C(7.4 \times 10^{-7}) = 12.4$, $C(8.3 \times 10^{-6}) = 10$, $C(.00007) = 8.9$, $C(.0026) = 8.3$, $C(.0071) = 7.8$, $C(.01) = 5.97$, $C(.011) = 4.35$, $C(.051) = 2.99$, $C(.12) = 1.89$.
	$n_{A_1} = 80$	$C(8.34 \times 10^{-30}) = 66.95$, $C(4.0056 \times 10^{-26}) = 58.1$, $C(3.073 \times 10^{-23}) = 50.84$, $C(9.03 \times 10^{-21}) = 44.96$, $C(1.37 \times 10^{-18}) = 39.8$, $C(1.22 \times 10^{-16}) = 35.2$, $C(7.2 \times 10^{-15}) = 31.04$, $C(3.9 \times 10^{-13}) = 28$, $C(7.8 \times 10^{-13}) = 27$, $C(8.4 \times 10^{-12}) = 23.84$, $C(1.9 \times 10^{-10}) = 20.7$, $C(2.2 \times 10^{-10}) = 19.78$, $C(1.2 \times 10^{-9}) = 19.7$, $C(4.23 \times 10^{-9}) = 17.87$, $C(2.41 \times 10^{-8}) = 16.62$, $C(6.4 \times 10^{-8}) = 15.27$, $C(3.5 \times 10^{-7}) = 13.89$, $C(7.5 \times 10^{-7}) = 12.9$, $C(3.8 \times 10^{-6}) = 11.46$, $C(7.14 \times 10^{-6}) = 10.8$, $C(.00024) = 9.3$, $C(.00046) = 8.85$, $C(.00057) = 7.5$, $C(.00068) = 7.13$, $C(.0015) = 5.84$, $C(.0021) = 5.6$, $C(.0066) = 4.43$, $C(.0086) = 4.28$, $C(.017) = 3.23$, $C(.024) = 3.13$, $C(.045) = 2.26$, $C(.07) = 2.17$, $C(.12) = 1.46$.

Table 3. Simulation results: power comparison.

Test	$P_{11} = 0.8\ P_{12} = 0.05\ P_{22} = 0.1$		$P_{11} = 0.3\ P_{12} = 0.3\ P_{22} = 0.1$	
K_N	$N = 20,$	Power = 0.94	$N = 20,$	Power = 0.58
	$N = 50,$	Power = 0.97	$N = 50,$	Power = 0.72
	$N = 100,$	Power = 1	$N = 100,$	Power = 0.81
Yuan and	$N = 20,$	Power = 0.90	$N = 20,$	Power = 0.41
Bonney's	$N = 50,$	Power = 0.93	$N = 50,$	Power = 0.66
(2003) test	$N = 100,$	Power = 1	$N = 100,$	Power = 0.77

does not satisfy HWE, that is, the null hypothesis is false. The second model (in the second column of table 3) also does not satisfy HWE (the null hypothesis is false). However, the second model is very close to HWE.

The results are tabulated in table 3. From the results in the first column, we see that in terms of statistical power our proposed test outperformed the exact test for small N. Also, from the second column, we see that our test does very well even when we have small discrepancies from HWE.

Test for m (m > 2) alleles

To preserve consistency with previous discussion and notation, consider a sample of genotypes of N unrelated diploid individuals measured at an autosomal locus that has m alleles $A_1, A_2 \ldots, A_m$. Let n_{A_1} be the number of ith allele, $i = 1, \ldots, m$. Then it is clear that $\sum_{i=1}^{m} n_{A_i} = 2N$. Also, let the number of heterozygous A_iA_j genotypes be N_{ij}, $1 \leq i < j \leq m$. Then the number of homozygous A_iA_i genotypes is $N_{A_iA_i} = \dfrac{N_{A_i} - \sum_{j=i+1}^{m} N_{A_iA_j}}{2}$, $i = 1, \ldots, m$. Under the assumption of HWE,

$$P_2(N_{A_iA_j} = k_{ij},\ 1 \leq i \leq j \leq m | n_{A_1}, \ldots, n_{A_m}, N) = \frac{(N!)(2)^{\sum_{i=1}^{m}\sum_{j=i+1}^{m} k_{ij}} \prod_{i=1}^{m}(n_{A_i})!}{(2N)! \prod_{i=1}^{m}(k_{ii}!) \prod_{1\leq i<j\leq m}(k_{ij}!)}. \quad (11)$$

For the general case,

$$P_1(N_{A_iA_j} = k_{ij}, 1 \leq i \leq j \leq m | n_{A_1}, \ldots, n_{A_m}, N) = \frac{f(k_{ij}, n_{A_i}, 1 \leq i \leq j \leq m; N)}{C_{n_{A_1}, \ldots, n_{A_{m-1}}}}, \quad (12)$$

where both equations 11 and 12 are defined over the set

$$S = \left\{ k_{ij}, 1 \leq i \leq j \leq m \quad \text{such that} \sum_{j=i+1}^{m} k_{ij} + 2k_{ii} = n_{A_i}, \quad i = 1, \ldots, m \right\},$$

$$f(k_{ij}, n_{A_i}, 1 \leq i \leq j \leq m; N) = \frac{(N!)(2)^{\sum_{i=1}^{m}\sum_{j=i+1}^{m} k_{ij}} \prod_{1\leq i\leq j\leq m}(P_{ij})^{k_{ij}}}{\prod_{i=1}^{m}(k_{ii})! \prod_{1\leq i<j\leq m}(k_{ij}!)},$$

P_{ij} is the probability of getting genotype A_iA_j, and $C_{n_{A_1},\ldots,n_{A_{m-1}}} = \sum_S f(k_{ij}, n_{A_i}, 1 \leq i \leq j \leq m; N)$. It should be noted that $\sum_{i=1}^{m} P_{ii} + 2\sum_{i=1}^{m}\sum_{j=i+1}^{m} P_{ij} = 1$.

From equations 1, 11 and 12,

$$K(P_1 : P_2) = \log 2N! - \sum_{i=1}^{m} \log n_{A_i} - \log C_{n_{A_1},\ldots,n_{A_{m-1}}} + \sum_{i=1}^{m}\sum_{j=i}^{m} a_{ij} \log P_{ij}, \quad (13)$$

where $a_{ij} = \sum_S n_{ij} f(n_{ij}, n_{A_i}, 1 \leq i \leq j \leq m; N)$.

As an example, suppose $m = 3$. Then, equation 13 reduces to

$$\begin{aligned} K(P_1 : P_2) = {} & \log 2N! - \log n_{A_1}! - \log n_{A_2}! \\ & - \log(2N - n_{A_1} - n_{A_2})! \\ & - \log C_{n_{A_1},n_{A_2}} + \left[\frac{NP_{11}C^*_{n_{A_1},n_{A_2}}(1,1)}{C_{n_{A_1},n_{A_2}}}\right] \log P_{11} \\ & + \left[\frac{2NP_{12}C^*_{n_{A_1},n_{A_2}}(1,2)}{C_{n_{A_1},n_{A_2}}}\right] \log P_{12} \\ & + \left[\frac{2NP_{13}C^*_{n_{A_1},n_{A_2}}(1,3)}{C_{n_{A_1},n_{A_2}}}\right] \log P_{13} \\ & + \left[\frac{2NP_{23}C^*_{n_{A_1},n_{A_2}}(2,3)}{C_{n_{A_1},n_{A_2}}}\right] \log P_{23} \\ & + \left[\frac{NP_{22}C^*_{n_{A_1},n_{A_2}}(2,2)}{C_{n_{A_1},n_{A_2}}}\right] \log P_{22} \\ & + \left[\frac{NP_{33}C^*_{n_{A_1},n_{A_2}}(3,3)}{C_{n_{A_1},n_{A_2}}}\right] \log P_{33}, \end{aligned} \quad (14)$$

where $C^*_{n_{A_1},n_{A_2}}(i,j) = \sum_S f(k_{11}, \ldots, k_{ij} - 1, \ldots, k_{33}, n_{A_1}, n_{A_2}, n_{A_3}; N - 1)$, $1 \leq i \leq j \leq 3$. As an example, if $P_{11} = P_{22} = P_{33} = 0.1$, $P_{12} = P_{13} = 0.1$, $P_{23} = 0.3$, $N = 5$, and $n_{A_1} = n_{A_2} = 3$, then $K(P_1 : P_2) = 0.63$, which is close to zero.

Using arguments similar to the previous section, the discrimination information statistic can be obtained from equation 13 simply by replacing P_{ii} with $\hat{P}_{ii} = \dfrac{N_{ii}}{N}$, $i = 1, \ldots, m$, and P_{ij} by $\hat{P}_{ij} = \dfrac{N_{ij}}{2N}$, $1 \leq i \leq j \leq m$. Again, for large N, $2K_N(P_1 : P_2)$ has an approximately chi-square distribution. Thus, one can reject the assumption of HWE if

$2K_N(P_1 : P_2) > \chi^2_{\alpha,\nu}$ where ν is the number of parameters that must be estimated.

Concluding remarks

A criterion was proposed to order populations based on HWE. Also, a test of HWE based on an estimate of the discrimination information measure between the underlying model supported by given data and HWE has been presented. Our proposed test is consistent and it is very easy to implement. The proposed test performs very well in terms of power, especially for small deviations from HWE, compared with the widely used test of Yuan and Bonney (2003).

References

Kullback S. 1978 *Information theory and statistics*. Dover, New York.

Kullback S. and Leibler R. A. 1951 On information and sufficiency. *Ann. Math. Stat*. **22**, 79–86.

Spitze K. 1993 Population structure in *Daphnia obtusa*: quantitative genetic and allozmic variation. *Genetics* **135**, 365–374.

Wigginton J. E., Cutler D. S. and Abecasis G. R. 2005 A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet*. **76**, 877–883.

Yuan A. and Bonney G. E. 2003 Exact test of Hardy–Weinberg equilibrium by Markov chain Monte Carlo. *Math. Med. Biol*. **20**, 327–340.

Received 1 August 2006