Exact Tests for Hardy-Weinberg Proportions

William R. Engels

University of Wisconsin
Genetics Department
Madison, WI 53706
ABSTRACT

Exact conditional tests are often required to evaluate statistically whether a sample of diploids comes from a population with Hardy-Weinberg proportions. This requirement is especially common when the sample includes multiple alleles and sparse data, thus rendering asymptotic methods, such as the common $\chi^2$ test, unreliable. Such an exact test can be performed using the likelihood ratio as its test statistic rather than the more commonly-used probability test. Conceptual advantages in using the likelihood ratio are discussed. A substantially improved algorithm is described to permit the performance of a full-enumeration exact test on sample sizes which are too large for previous methods. An improved Monte Carlo algorithm is also proposed for samples which preclude full enumeration. These algorithms are about two orders of magnitude faster than those currently in use. Finally, methods are derived to compute the number of possible samples with a given set of allele counts, a useful quantity for evaluating the feasibility of the full enumeration procedure. Software implementing these methods, *ExactoHW*, is provided.

INTRODUCTION

When studying the genetics of a population, one of the first questions to be asked is whether the genotype frequencies fit Hardy-Weinberg (HW) expectations. They will fit HW if the population is behaving like a single randomly-mating unit without intense viability selection acting on the sampled loci. In addition, testing for HW proportions is often used for quality control in genotyping, as the test is sensitive to misclassifications or undetected null alleles. Traditionally, geneticists have relied on test statistics with asymptotic $\chi^2$ distributions to test for
goodness-of-fit with respect to HW proportions. However, as pointed out by several authors (Aoki 2003; Chakraborty and Zhong 1994; Elston and Forthofer 1977; Emigh 1980; Guo and Thompson 1992; Hernandez and Weir 1989; Kang 2008; Louis and Dempster 1987; Maiste and Weir 2004; Rohlfs and Weir 2008; Rousset and Raymond 1995; Wigginton et al. 2005), these asymptotic tests quickly become unreliable when samples are small or when rare alleles are involved. The latter situation is increasingly common as techniques for detecting large numbers of alleles become widely used. Moreover, loci with large numbers of alleles are intentionally selected for use in DNA identification techniques (e.g., Weir 1992). The result is often sparse-matrix data for which the asymptotic methods cannot be trusted.

A solution to this problem is to use an exact test (Haldane 1954; Levene 1949) analogous to Fisher’s Exact Test for independence in a 2 x 2 contingency table and its generalization to rectangular tables (Freeman and Halton 1951). In this approach, one considers only potential outcomes which have the same allele frequencies as observed, thus greatly reducing the number of outcomes which must be analyzed. One then identifies all such outcomes which deviate from the HW null hypothesis by at least as much the observed sample. The total probability of this subset of outcomes, conditioned on HW and the observed allele frequencies, is then the P-value of the test. When it is not possible to enumerate all outcomes, it is still feasible to approximate the P-value by generating a large random sample of tables.

The exact HW test has been used extensively and eliminates the uncertainty inherent in the asymptotic methods (Emigh 1980; Guo and Thompson 1992; Hernandez and Weir 1989; Rousset and Raymond 1995). However, there are two difficulties with the application of this method and its interpretation, both of which are addressed in this report.
The first issue is the question of how one decides which of the potential outcomes are assigned to the subset which deviate from HW proportions by as much or more than the observed sample. If the alternative hypothesis is specifically an excess or dearth of homozygotes, then the tables can be ordered by ROUSSET and RAYMOND’s (1995) $U$-score or, equivalently, ROBERTSON AND HILL’s (1984) minimum-variance estimator of the inbreeding coefficient. However, when no specific direction of deviation from HW is suspected, then there are several possible test statistics which can be used (EMIGH 1980). These include the $\chi^2$ statistic, the likelihood ratio (LR) and the conditional probability itself. The last option is by far the most widely used (CHAKRABORTY and ZHONG 1994; ELSTON and FORTHOFER 1977; LOUIS and DEMPSTER 1987; WEIR 1996; WIGGINTON et al. 2005) and implemented in the GENEPOP software package (ROUSSET 2008). The idea of using the null-hypothesis probability as the test statistic was originally suggested in the context of rectangular contingency tables (FREEMAN and HALTON 1951), but this idea has been criticized for its lack of discrimination between the null hypothesis and alternatives (CRESSIE and READ 1989; GIBBONS and PRATT 1975; RADLOW and ALF 1975). For example, suppose a particular sample was found to have a very low probability under the null hypothesis of HW. Such a result would usually tend to argue against the population being in HW equilibrium. However, if this particular outcome also has a very low probability under even the best-fitting alternative hypothesis, then it merely implies that a rare event has occurred regardless of whether the population is in random-mating proportions. The first part of this report compares the use of probability versus the likelihood ratio as the test statistic in HW exact tests. Reasons for preferring the likelihood ratio are presented.

The second difficulty in performing HW exact tests is the extensive computation needed for large samples when multiple alleles are involved. In this report I present a new algorithm for
carrying out these calculations. This method adapts some of the techniques originally developed for rectangular contingency tables in which each possible outcome is represented as a path through a lattice-like network (MEHTA and PATEL 1983). Unlike the loop-based method currently in use (LOUIS and DEMPSTER 1987), the new algorithm uses recursion and can be applied to any number of alleles without modification. In addition, it improves the efficiency by about two orders of magnitude, thus allowing the full enumeration procedure to be applied to larger samples and with greater numbers of alleles.

The recursion algorithm has been tested successfully on samples with as many as 17 alleles when most of those alleles are rare. However, there are still some samples for which a complete enumeration is not practical. For example, the data from the human Rh locus in Figure 1D would require examining $2 \times 10^{56}$ tables (see below). For such cases, GUO AND THOMPSON (1992) suggest two methods of approximation. In the first, many independent random tables are selected with the same allele frequencies as the observed sample to obtain an empirical distribution of the table probabilities and that of any test statistic used. The second approach uses a Markov chain to generate successive approximations of the distribution. The authors point out that the former approach is preferable to the Markov chain method owing to better control of the precision of the final $P$-value estimate. However, the Markov chain method is more often used (ROUSSET 2008) because it was thought to be applicable to larger tables. Several improvements to the method of independent random tables are suggested here to make that approach practical for even the largest of realistic samples, thus eliminating the need for a Markov chain approach.

Finally, I address the problem of determining the number of tables of genotype counts corresponding to a given set of allele counts. This number is needed for determining whether the
exact test can be performed by full enumeration. Previously, this number could not be obtained except by actually carrying out the complete enumeration.

The methods described are implemented in a software package, *ExactoHW*, for MacOS X10.5 or later. It is available in compiled form (supplementary File S1) or as source code on request.

**MATERIALS AND METHODS**

All calculations were performed on a MacPro3.1 computer from Apple Inc. with two Quad-Core Intel Xeon processors running at 2.8 GHz. The operating system was MacOS X10.5. Programming for power calculations as well as the *ExactoHW* software was done with Apple’s Xcode development system and Cocoa application programming interface. A version of GENEPOP 4.0.10 was compiled on the same equipment for use in comparisons with *ExactoHW*. Both GENEPOP and *ExactoHW* are written in dialects of C. Random permutations for the Monte Carlo procedure were obtained by the Fisher-Yates shuffle (Fisher and Yates 1943) truncated after the first $n$ swaps in a table of $2n$ elements. Random numbers were generated by the multiply-with-carry method (Marsaglia 2003).

**COMPARISON OF TEST STATISTICS**

**Formulation:** A sample of diploid genotypes at a $k$-allele locus can be represented by a triangular matrix such as:

$$
\mathbf{a} = \begin{bmatrix}
  a_{11} \\
  a_{21} & a_{22} \\
  \vdots & \vdots & \ddots \\
  a_{k1} & a_{k2} & \cdots & a_{kk}
\end{bmatrix}
$$
where $a_{ij}$ is the observed number of genotypes with alleles $i$ and $j$. The number of alleles of type $i$ is $m_i = 2a_{ii} + \sum_{i \geq j} a_{ij}$, and let $n$ be the total sample size ($n = \sum a_{ii}$), and $\sum m_i = 2n$. If we assume this sample was obtained by multinomial sampling from a population in HW proportions with the observed allele frequencies $\left(\frac{m_i}{2n}\right)$, then the conditional probability of the sample given the observed allele counts is

$$P(a | m) = \frac{2^{n-d} n! \prod m_i!}{(2n)! \prod_{i \geq j} a_{ij}!}$$

(HALDANE 1954; LEVENE 1949) where $d$ is the number of homozygotes ($d = \sum a_{ii}$). Equation 1 can be derived as the ratio of two multinomial probabilities. The numerator is the probability of the observed sample if the genotype frequencies fit HW expectations, and the denominator is the probability of obtaining the observed allele frequencies.

The likelihood ratio is given by

$$LR(a) = \frac{\prod_i m_i^{m_i}}{2^{n+d} n^n \prod_{i \geq j} a_{ij}!}$$

(e.g., WEIR 1996, p. 106) and can also be derived as the ratio of two multinomial probabilities. The numerator is the same as for Equation 1, and the denominator is the probability of obtaining the observed outcome under the best-fitting alternative hypothesis. This best-fitting hypothesis is that of sampling from a population whose genotype frequencies are identical to those of the observed sample: $a_{ij} / n$. These equations can also be derived from the assumption of Poisson
sampling. Interestingly, as pointed out by a reviewer, Equations 1 and 2 become interchangeable following the application of Stirling’s approximation: \( \ln x! \approx x \ln x - x \).

**Comparison of probability versus likelihood ratio test statistics:** To visualize the relationship between these two types of test, consider a sample of 10 diploids containing five alleles. The allele counts are 9, 6, 3, 1 and 1. That is: \( \mathbf{m} = \begin{bmatrix} 9 & 6 & 3 & 1 & 1 \end{bmatrix} \). There are 139 possible samples of this kind, and their probabilities and likelihood ratios are plotted in Figure 2. It is clear that the two quantities are strongly correlated, with a nearly linear relationship when plotted on a log-log scale.

One of the 139 tables, \( \mathbf{a} = \begin{bmatrix} 2 & 4 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \), is indicated by the intersection of the two dashed lines. This plot provides a graphical demonstration of the difference between the two kinds of exact test: The probability test for HW consists of summing the probabilities of all the samples which lie on or to the left of the vertical dashed line, whereas the likelihood ratio test selects those on or below the horizontal line. The positive correlation ensures that the subsets selected by these procedures contain many of the same points. However, these subsets are not identical. They differ by the points lying in the upper left and lower right quadrants. In this case, the points in these quadrants are enough to cause a threefold difference in the computed \( P \)-values.

Visualizing the tests in this way helps to clarify why the likelihood ratio may be seen to provide a better fit to our intuitive notion of what is being tested. The points in the upper left
quadrant are included in the probability test because they have a slightly lower probability than the observed sample. However, it can be argued that they are not more deviant from the HW hypothesis, since their probability is relatively low even under the best-fitting alternative hypothesis. They are simply rare outcomes regardless of the true state of the population. On the other hand, those in the lower right quadrant do seem to deviate from HW more than the observed case when compared with the best-fitting alternative. By this reasoning, the likelihood ratio $P$-value of 0.034 is to be preferred over the probability-based value of 0.010 as it better reflects the strength of evidence against the HW hypothesis relative to the alternatives.

It is interesting to note that samples showing a homozygote excess relative to HW tend to lie above the diagonal in Figure 2 while many of those with a dearth of homozygotes lie below it. This tendency appears to be a general characteristic, as it was equally clear in each of several other examples plotted in this way (see supplementary Figure S1). It implies that when there is a homozygote excess, as might be caused by inbreeding or hidden subdivisions of the population, the probability-based test will tend to give lower $P$-value as compared to the likelihood ratio test. The reverse is true when there is a heterozygote excess. This trend is reflected in the power comparisons conducted in several studies (EMIGH 1980; ROUSSET and RAYMOND 1995) as well as those described below.

**Different alternative hypotheses.** Another useful way to compare the probability test with the likelihood ratio test is to think of them as similar test statistics – *i.e.*, likelihood ratio-based – but directed against different alternative hypotheses. Note that the probability test could be thought of as a likelihood ratio test if the alternative hypothesis is that all possible conditional samples have an equal probability. That way the denominator of the likelihood ratio will be the same for all samples, and the resulting ordering of the possible samples will be identical to that
produced by the probability test. However, it is not clear that any sampling procedure or realistic population characteristics would lead to all possible tables being equally likely. By contrast, multinomial sampling from a population with a fixed set of genotype frequencies is probably a good approximation to what is typically done. Therefore, this way of comparing the two tests also argues against the use of probability itself as a test statistic, as it is equivalent to performing a likelihood ratio test against an unrealistic alternative hypothesis. It suggests that the use of LR as a test statistic may be a better choice in terms of matching a realistic set of alternative hypotheses.

**Power comparisons.** Finally, we can compare these two kinds of test in terms of their power. That is, we can compute the probability of the $P$-value falling below a given threshold, $\alpha$, when the population deviates from HW to various extents. The contour plots in Figure 3 compare the powers of the likelihood ratio test (numerator) with the probability test (denominator) for sample sizes of 50 and 500 and two alleles. Supplementary file S2 shows other sample sizes between 10 and 600. Power comparisons of this kind have been reported previously (CHAKRABORTY and ZHONG 1994; EMIGH 1980; HERNANDEZ and WEIR 1989; KANG 2008; MAISTE and WEIR 2004; ROUSSET and RAYMOND 1995) but not with full coverage of the parameter space. Each plot was constructed by computing the power of each test under multinomial sampling at 2687 points distributed evenly within the parameter space. The frequency, $q$, of the less-frequent allele can range from 0 to $\frac{1}{2}$, and the inbreeding coefficient, $F$, lies between $\frac{-q}{1-q}$ and 1.

There are many areas within the parameter space where the two tests have approximately the same power, as indicated by the white spaces in Figure 3. However, there are also substantial
areas where the probability test (shades of red) or the LR test (shades of blue) has significantly
greater power. Even when the sample size is 500, the red and blue regions are still prominent,
indicating that the two tests converge only slowly as sample size increases.

The minimum value for the ratio is around 0.6 (red region), but the maximum exceeds 4.0
(purple). In other words, the decrease in relative power associated with using the LR test in the
red areas is not great, but a fourfold decrease in power can result when the probability test is
used for populations in the purple areas. This comparison suggest an advantage to using the LR
test when there is no expectation concerning the sign of \( F \).

The blue and purple regions lie within the area where \( F \) is negative, and the red sectors occur
mainly in areas where \( F \) is positive, echoing the previous observation (EMIGH 1980) that the
probability test can have greater power when there is an excess of homozygotes whereas the LR
test’s power is greater when there is a heterozygote excess. The basis for this tendency can be
seen in Figure 2, where tables with a homozygote excess lie more often above the diagonal.

The red areas in Figure 3 need not be interpreted as advantageous for the probability test. On
the contrary, if one accepts the arguments above, these regions of the parameter space represent
situations where using the probability tests entails an increased risk of overestimating the
evidence for homozygote excess. The reason is that the probability test is actually aimed at a
subtly different alternative hypothesis which does not reflect realistic sampling procedures. On
the other hand, the blue areas can be considered situations where the LR test has a power
advantage in detecting heterozygote excess. Homozygote excess tends to be more common as it
can arise from inbreeding, population subdivision or undetected null alleles. Of course, if one
type of excess is suspected initially, then the maximum power can be obtained from using a one-
sided criterion such as the $U$-score (ROUSSET and RAYMOND 1995). A Bayesian approach can also be used to take account of prior expectations (MONTOYA-DELGADO et al. 2001).

Contour plots similar to those in Figure 3 were also constructed to compare the LR test with $\chi^2$ as the test statistic for ordering the tables (see supplementary figure S2). The results were very similar to Figure 3, suggesting that the $\chi^2$ statistic results in an ordering which is closer to that of the probability than to the LR. This similarity might be expected, as $\chi^2$ does not take explicit account of the probability of each table under the alternative hypothesis of multinominal sampling.

**ALGORITHMS**

**Full enumeration algorithm:** A significant advance in the exact analysis of rectangular contingency tables was obtained by MEHTA and PATEL (1983) who found that the set of tables with fixed marginal totals could be represented by a network of nodes connected by arcs. Each pathway from the initial node to the final one corresponds to one of the tables. The total lengths of the arcs in each pathway can also be used to calculate the probability and test statistic associated with each table. This representation suggested an efficient recursion-based algorithm for enumerating the tables and computing the associated $P$-value.

The approach taken here is analogous, but adapted to the triangular tables of genotype data with fixed allele counts. For example, consider a sample of four diploids, each homozygous for a different allele. Thus,

$$a = \begin{bmatrix} 1 \\ 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$
and \( \mathbf{m} = \begin{bmatrix} 2 & 2 & 2 & 2 \end{bmatrix} \). There are 17 possible tables with this set of allele counts. Figure 4A shows the network representation of this case. Each path from the initial node (2222) to the final one (0000) represents one of the 17 tables, and the observed table is indicated by the dashed line. The four digits identifying each node are the residual allele counts, and each column of nodes represents the genotype assignments for one of the rows of the table, starting with the bottom. These columns are referred to as levels in the contingency table literature (AGRESTI 1992; MEHTA and PATEL 1983). When tracing paths, arcs are followed only in the rightward direction. The 5-allele example with 139 tables used in Figure 2 is shown in Figure 4B. Each table corresponds to one of the paths from (96311) to (00000).

To traverse the network of tables while computing the desired probabilities and test statistics, I propose an algorithm in which a pair of functions, \( \text{Homozygote} \) and \( \text{Heterozygote} \), operate in a recursive fashion by calling themselves and each other. Each call to \( \text{Homozygote} \) corresponds to one of the nodes, whereas each arc corresponds to one or more calls to the \( \text{Heterozygote} \) function.

Calculation of the probability and statistics associated with each completed table is distributed through the lattice so that each new table requires minimal calculations. These calculations are greatly facilitated by noting from Equations 1 and 2 that the logs of the probability and LR can be written as:

\[
\ln P(a) = K_p - \sum_{i \geq j} f(a_{ij}) - d \ln 2
\]
\[
\ln LR(a) = K_g - \sum_{i \geq j} g(a_{ij}) - d \ln 2
\]

(3)
where $K_p$ and $K_g$ are constants which need be computed only once for the entire set of tables, and
the functions, $f(i)$ and $g(i)$, defined as $\ln(i!)$ and $i \ln(i)$ respectively, are also computed only
once for the integers up to the largest allele count, $m_i$ and retrieved when needed. Note that the
log of the probability is calculated initially in order to avoid underflow errors.

The functions $Homozygote$ and $Heterozygote$ take the following parameters, which must be
passed by value rather than by reference, as required by the recursion process: $r$ and $c$ represent
the current row and column, with $c$ being unnecessary in $Homozygote$; $f_p$ and $g_p$ represent partial
sums of $\sum f(a_{ij}) + d \ln 2$ and $\sum g(a_{ij}) + d \ln 2$; $R$ is an array $(R_1, R_2, \ldots, R_k)$ containing the
residual allele counts. Note that the quantity $f_p$ or $g_p$ can be thought of as sum of the arc lengths in
each path of the network diagram (Figure 4).

After the constants, $K_p$ and $K_g$, and lookup tables, $f(i)$ and $g(i)$, have been computed, the
main calculation is initiated with a call to $Homozygote$ with $r$ set to $k$, $R$ set to the allele counts,
$m_1, m_2, \ldots, m_k$, sorted in descending order, and the remaining parameters set to zero. Pre-sorting
of the allele counts greatly increases the efficiency but does not affect the numerical outcome.
The procedure below applies when there are three or more alleles. The two recursive functions
are defined as follows:

$Homozygote(r, f_p, g_p, R)$ : The first step is to compute the lower and upper bounds for $a_r$
given the current residual allele counts. These are:

$$lower = \left( R_r - \sum_{i=1}^{r-1} R_i \right) / 2$$
$$upper = R_r / 2$$
with lower set to zero if the above expression is negative (LOUIS and DEMPSTER 1987). Integer arithmetic is assumed where appropriate so that fractions are rounded down, thus making it unnecessary to specify whether quantities are even or odd. Now, for each value of \( a_{rr} \) between lower and upper, call *Heterozygote* with parameters:

\[
[r, r - 1, f_p + f(a_{rr}) + a_{rr} \ln 2, g_p + g(a_{rr}) + a_{rr} \ln 2, R']
\]

in which \( R' \) is modified from \( R \) by subtracting 2\( a_{rr} \) from \( R_r \).

*Heterozygote*(\( r, c, f_p, g_p, R \)) : As before, we start by finding the upper and lower bounds for genotype \( a_{rc} \):

\[
\text{lower} = R_r - \sum_{i=1}^{c-1} R_i \\
\text{upper} = \min(R_r, R_c)
\]

with any negative value for lower replaced by zero. The next step depends on the values of \( r \) and \( c \). If \( c > 2 \), then for each value of \( a_{rc} \) from lower to upper, call *Heterozygote* with parameters:

\[
[r, c - 1, f_p + f(a_{rc}) + a_{rc} \ln 2, g_p + g(a_{rc}) + a_{rc} \ln 2, R']
\]

in which \( R' \) is constructed by subtracting \( a_{rc} \) from each of \( R_r \) and \( R_c \).

If \( c = 2 \) and \( r > 3 \), then for each value of \( a_{r2} \) from lower to upper, let

\[
a_{r1} = \min(R_r - a_{r2}, R_1)
\]

and call *Homozygote* with parameters

\[
[r - 1, f_p + f(a_{r2}) + f(a_{r1}), g_p + g(a_{r2}) + g(a_{r1}) + g(a_{r1}), R']
\]

where \( R' \) is constructed by subtracting \( a_{r2} \) from \( R_r \), and \( R_2 \), and \( a_{r1} \) from \( R_r \) and \( R_1 \).

Finally, if \( c = 2 \) and \( r = 3 \), then for each value of \( a_{32} \) from lower to upper, let
\[ a_{31} = \min(R_3 - a_{32}, R_1) \]
\[ f' = f_p + f(a_{31}) + f(a_{32}) . \]
\[ g' = g_p + g(a_{31}) + g(a_{32}) . \]

At this point, we are left with the equivalent of a two-allele case in which the allele counts are

\[ m'_1 = R_1 - a_{31} \text{ and } m'_2 = R_2 - a_{32} . \]

If \( m'_1 \leq m'_2 \) then for each value of \( a_{ij} \) from zero to \( m'_1 / 2 \) we set

\[ a_{21} = m'_1 - 2a_{11} \]
\[ a_{22} = \left( m'_2 - a_{21} \right) / 2 . \]

If \( m'_1 > m'_2 \), then for each value of \( a_{22} \) from zero to \( m'_2 / 2 \) we set

\[ a_{21} = m'_2 - 2a_{22} \]
\[ a_{11} = \left( m'_1 - a_{21} \right) / 2 . \]

Either way, for each value we can process a completed table whose log-probability and \( \ln LR \) test statistic are

\[ \ln P = K_p - f' - f(a_{11}) - f(a_{21}) - f(a_{22}) - (a_{11} + a_{22}) \ln 2 \]
\[ \ln LR = K_g - g' - g(a_{11}) - g(a_{21}) - g(a_{22}) - (a_{11} + a_{22}) \ln 2 . \]

If the table is deemed to deviate from HW expectations at least as much as the observed table based on the \( LR \) or another criterion, then the actual probability is found by taking the antilog, and the \( P \)-value is incremented by this amount.

When the initial call to \textit{Homozygote} finally returns, the entire tree of tables has been traversed, all probabilities and test statistics have been computed and processed, and the exact \( P \)-values have been computed.
An enhancement to the above algorithm is the addition of the $U$-score test for homozygote- or heterozygote excess (ROUSSET and RAYMOND 1995), which can be thought as a “one sided” procedure for narrowing the alternative hypotheses. For the purpose of ordering the tables, the only quantity needed for each table is $\sum a_{ii} / m_i$. By adding one more parameter to each function, this sum can be computed distributively throughout the recursion in a way similar to the other two quantities (see Equation 3). With precomputed lookup tables for $a_{ii} / m_i$ ($i = 1, 2$), inclusion of this test statistic does not significantly increase the computation time. *ExactoHW* reports either $P(U \geq \text{observed})$ or $P(U \leq \text{observed})$ depending on whether the observed $U$ score is positive or negative.

To confirm that this procedure yields the same $P$-values as the algorithm of LOUIS and DEMPSTER (1987) implemented in GENEPOP (ROUSSET 2008), the $P$-values were computed by both methods for the samples in Figure 1A-C and listed in supplementary Table S1. To compare the relative speeds of the algorithms, both programs were compiled from their C dialects and run on the same computer. The comparison used 4-allele samples with the same allele frequencies and sample sizes ranging from $n = 500$ to $n = 2000$. The present algorithm was found to be about two orders of magnitude faster (supplementary Table S2). The speed advantage is especially apparent in the largest sample size, where the analysis by GENEPOP required more than 8 hours of computation compared to less than three minutes for *ExactoHW*, even though the latter operation performed all three tests (probability, LR and $U$-score) compared to probability alone.

**Monte Carlo method:** GUO and THOMPSON (1992) suggested generating random tables of genotypes with the observed allele counts by first obtaining a random permutation of an array containing the $2n$ haplotypes in the observed sample. Then each pair of adjacent haplotypes in
the permuted array is taken as one of the \( n \) genotypes. The probability and test statistic are then computed for each such random table, resulting in an estimate of the \( P \)-value after sufficiently many random tables have been generated. The authors concluded that this method might be useful in some cases but is not efficient enough to handle large tables owing to the necessity to compute the probability and test statistic for each table. Instead, they proposed a Markov chain alternative despite the inherent disadvantage of that method in terms of controlling the precision of the resulting \( P \)-value.

On reexamining GUO and THOMPSON’S (1992) random sampling method, it is found that a dramatic improvement in efficiency can be obtained with a few minor modifications. The most important of these is the use of Equations 3 and the pre-computed values for \( K_p, K_g, f, \) and \( g \) for finding the probability and test statistic. With this technique, the time needed for computing \( P \) and \( LR \) is small compared to that of generating the random table. An additional factor of two improvement can be achieved by noting that the random permutation process can be stopped after the first \( n \) elements of the randomly-permuted haplotype array, then pairing haplotype \( i \) with haplotype \( n + i \) to produce each diploid genotype (See Materials and Methods). Finally, one can take advantage of present-day multicore computers to generate multiple random tables simultaneously.

All of these techniques are incorporated in \textit{ExactoHW}. The result is that samples such as those in Figures 1A – 1C can be analyzed by the Monte Carlo method at the rate of approximately 400,000 tables per second while computing all three test statistics for each. Even the much larger sample in Figure 1D is amenable to this approach, with a rate of 38,000 tables per second (supplementary Table S3). The \( P \)-values in supplementary Table S1 confirm the accuracy of this algorithm.
COUNTING TABLES

For a given data set, the choice between the full enumeration test versus a Monte Carlo alternative depends on the number of tables needed for the full enumeration. If this number is small enough, the full enumeration is always preferable. For rectangular contingency tables, the number of possible tables with a fixed set of marginal totals has been examined by GAIL and MANTEL (1977) and subsequent authors (reviewed in GRESELIN 2004). Several exact and approximate approaches have been described with the latter being less computationally intensive. However, no similar analysis has been reported for the triangular tables associated with genotype data with fixed allele counts. The following analysis provides three alternatives to address the table-counting problem for genotypic data.

**Generating function method**: The first approach is to make use of a generating function, $G(x_1, x_2, \ldots, x_k)$, on a set of dummy variables corresponding to the alleles. The contribution to this function from genotype $ij$ is

$$\sum_{z=0}^{\infty} \left( \frac{x_i x_j}{z!} \right)^z \left( \frac{1}{1-x_i x_j} \right).$$

Therefore, the generating function is:

$$G(x) = \prod_{i \geq j} \left( \frac{1}{1-x_i x_j} \right)$$

and the number of tables is the coefficient of $x_1^{m_1} x_2^{m_2} x_3^{m_3} \ldots x_k^{m_k}$ in the expansion of this function. Finding this coefficient still requires computation, but with existing algorithms it can be more efficient than enumerating the entire set. In particular, the `SeriesCoefficient` function, which is
part of the Mathematica software (Wolfram), works well. A Mathematica function for this task can be defined as follows:

```mathematica
count[m_] := SeriesCoefficient @@ Flatten[
    {Product[1/(1 - Subscript[x, i]*Subscript[x, j]),
        {i, 1, Length[m]}, {j, 1, i}],
     Table[{Subscript[x, i], 0, Sort[m][[i]]}, {i, 1, Length[m]}]}, 1];
```

The set of allele counts is pre-sorted in this definition in order to facilitate the process, but such sorting is not needed to obtain the correct answer. With this definition, the total number of tables in the example of Figure 2 is obtained with the command: count[{9,6,3,1,1}] to yield 139 tables. For the example in Figure 1A, the command count[{11, 30, 30, 19}] yields 162,365 tables; for Figure 1B count[{15, 14, 11, 12, 2, 2, 1, 3}] yields 250,552,020 tables; for Figure 1C count[{68, 115, 192, 83}] yields 1,289,931,294 tables. These numbers are identical to those found by enumerating the entire sets.

**Algorithmic approach:** The recursive algorithm described above can be modified to provide a relatively efficient count of the number of tables. Note from Figure 4 that the number of tables downstream from any node is independent of how that node was reached. Therefore, if we are only interested in the number of tables rather than their probability and test statistics, it should only be necessary to traverse each node once. When the number of tables downstream from the node has been determined, this number is placed into a hash table keyed to the identifier of the node. This identifier consists of the residuals and the node’s level (see Figure 4). When this node is reached again, its downstream table count is retrieved and added to the total, eliminating the need to traverse any of the downstream nodes. This method is typically 50 times faster than complete enumeration. ExactoHW uses this algorithm to compute the needed number of tables in
a separate thread in order to provide a quick estimate of the time needed while the full enumeration calculation is in progress.

**Normal approximation:** The large sample in Figure 1D overwhelms the two exact methods described above and calls for an approximate approach. Following the strategy of *GAIL* and *MANTEL* (1977) for rectangular contingency tables, we start by considering a larger set of tables with fewer restrictions. Let \( S \) be the set of all possible samples of \( n \) diploids without regard to the allele counts but allowing genotypes involving any of the \( k \) alleles. The cardinality of \( S \) is known, as it represents the \( n \)-multisets of the set of \( k(k+1)/2 \) genotypes. Thus

\[
|S| = \binom{k(k+1)/2}{n} = \binom{n+k(k+1)/2-1}{k(k+1)/2-1}.
\] (5)

We wish to count the members of the subset \( S_m \), which includes only those tables with allele counts \( m \), by multiplying \( |S| \) by the probability that a randomly selected member of \( S \) has allele counts \( m \).

When considering a random sample from \( S \), it is not appropriate to use the multinomial distribution which does not assign equal probability to each distinguishable table. Instead, we will make use of the one-to-one correspondence between the elements of \( S \) and the linear arrangements of \( n \) “stars” and \( b = k(k+1)/2 - 1 \) “bars.” Each genotype count corresponds to the number of stars between adjacent bars (*FELLER* 1968, page 38). If a random permutation of these \( n + b \) elements is performed, then each genotype count will have expectation \( n/(b+1) \) and probability distribution

\[
P(a) = \left( \frac{b}{n+b-a} \right)^{a-1} \prod_{i=0}^{a-1} \left( \frac{n-i}{n+b-i} \right).
\] (6)
The variance of each genotype count is found using Equation 6 to be

\[ V_a = \frac{nb(n+b+1)}{(b+1)^2(b+2)}. \]  

(7)

Also, the covariance between any two genotype counts is

\[ CV(a,a') = \frac{-V_a}{b}. \]  

(8)

Using Equations 7 and 8 plus the definition of the allele count, \( m_i = 2a_i + \sum_{i>j} a_{ij} \), we can compute the variance of each allele count

\[ V_m = (k+3)V_a + \left(k^2 + k - 2\right)CV(a,a') \]
\[ = (k+1)V_a \]  

(9)

and the covariance between any pair of allele counts

\[ CV(m,m') = V_a + \left(k^2 + 2k\right)CV(a,a') \]
\[ = \frac{-V_m}{k-1}. \]  

(10)

With the variance-covariance matrix for \( m \) determined by Equations 9 and 10 and its mean given by \( \bar{m}_i = n / k \), it is possible to approximate the probability of \( m \) by the multivariate normal density. In order to avoid singularity in the variance-covariance matrix, we can reduce the dimension to \( k-1 \) by excluding one of the \( m \)’s. This change does not affect the probability, as the sum of the allele counts is fixed. At this point, the situation becomes equivalent to equation 3.5 of GAIL and MANTEL (1977), and analogous simplifications to the multivariate normal density function can be used. These simplifications arise because the \( m_i \) are equicorrelated and have a common variance. Thus
\[ P(m) = \sqrt[k]{\frac{k-1}{2\pi k V_m}} \frac{1}{(k-1)/2} e^{-Q/2} \]  
\hspace{1cm} (11)

where

\[ Q = \left( \frac{k-1}{V_m k} \right) \left( \sum m_i^2 - \frac{(2n)^2}{k} \right). \]

We can now estimate the desired number of tables from Equations 5 and 11 as \( |S_m| = |S|P(m) \).

Figure 5 compares the normal approximation to the exact numbers for all possible sets of allele counts when \( n = 100 \) diploids and \( k = 3 \) alleles. The approximation is most reliable in the central region. It tends to underestimate the number of tables in the corners of the simplex and overestimate the number near the midpoint of each edge. Applying this approximation to the example in Figure 1A yields 166,195 tables, which is reasonably close to the true value of 162,365. For Figure 1B the approximation is 210,540,416 compared to the exact count of 250,552,020. For the large sample in Figure 1D, this method estimates the number of tables as \( 2 \times 10^5 \), thus confirming that this sample cannot be analyzed by full enumeration.

**DISCUSSION**

This report aims to facilitate the use of exact tests for Hardy-Weinberg proportions. Exact tests, as opposed to large-sample asymptotic approximations, are increasingly needed as data from multi-allelic loci accumulate. Performing the exact tests consists of examining all – or a sampling of – the potential results having the same sample size and allele frequencies as the observed data, then finding the probability that such a sample would deviate from HW expectations by at least as much as the observed data. Though straightforward in concept, the
execution can involve extensive computations. Furthermore, complications arise when one realizes that there are different ways to define the degree of deviation from HW proportions, leading to very different results.

The general question of whether probability itself should be used as a test statistic for ordering the potential outcomes of a discrete-valued experiment as opposed to using the likelihood ratio, $\chi^2$, or other measures including Bayesian approaches, has been examined by several authors (CRESSIE and READ 1989; DAVIS 1986; GIBBONS and PRATT 1975; HORN 1977; MAISTE and WEIR 2004; MONTOYA-DELGADO et al. 2001 ; RADLOW and ALF 1975; WAKEFIELD 2009), and it is unlikely that the discussion will end here. However, it is hoped that the visualization provided in Figure 2 and the accompanying discussions will at least help to clarify some of the differences and raise the possibility that the likelihood ratio may be a closer fit to what most population geneticists aim to do when testing for goodness of fit to HW proportions.

All can agree, however, that the full exact test is preferable to a Monte Carlo simulation when the former is computationally feasible. To that end, there have been two previous attempts (AOKI 2003; MAURER et al. 2007) to improve on LOUIS and DEMPSTER’s (1987) original algorithm for full enumeration of all tables with a given set of genotype counts. In both of those efforts, the strategy consisted of “trimming” the tree of potential tables by skipping branches which cannot contribute to the P-value or by identifying branches where the contribution can be found without traversing the entire branch. AOKI (2003) was particularly successful in finding expressions for boundaries on the minimum and maximum probabilities of tables lying downstream of a given node in the network diagram. One drawback of trimming is that only a single test can be conducted at a time, as some tables which can be skipped for one test statistic must still be evaluated for another. Both of these trimming algorithms enhanced the
computational efficiency compared to the original algorithm of Louis and Dempster (1987), but they are still considerably slower than the algorithm proposed here. For example, Aoki’s method was applied to the 8-allele data set in Figure 4B to perform the probability test in 625 seconds, whereas ExactoHW performed all three tests (probability, likelihood ratio and $U$-score) in 44 seconds. This comparison is indirect, as different machines were used for the tests. However, the difference is large enough that even after considering the threefold difference in processor clock speeds used for the tests (930 MHz versus 2.8 GHz) there is still a significant speed advantage to the present algorithm.

It might seem surprising that the algorithm proposed here and used in ExactoHW is so much more efficient than other methods despite examining many more tables compared to the “trimming” methods and while performing three tests rather than one. The explanation lies in the efficiency gained by distributing the calculations for the probability and test statistics throughout the recursive process. That is, each time a recursive call is made to Homozygote or Heterozygote, partial calculations are passed along so that only minimal computation is needed at each step. When this technique is combined with the pre-computed tables implied by Equation 3, the computational time needed for the probability, LR and $U$-score is small compared to the time needed just for generating the tables.

Despite this efficiency, it is still easy to find data sets which would require generation of too many tables to allow full enumeration by any method. The data set in Figure 1D, for example, would require approximately $2 \times 10^{56}$ tables. For such cases, it is necessary to resort to a Monte Carlo simulation, for which two kinds of strategy have been proposed (Guo and Thompson 1992). The first approach is to generate a large number of independent random members of the set of tables with the same allele counts as the observed sample, and use as the $P$-value the
proportion of these tables which deviate from HW expectations as much or more than the observed sample. GUO and THOMPSON (1992) proposed one method for generating such tables. Their method, with some key enhancements described above, was used in ExactoHW. An alternative method proposed by HUBER et al. (2006) is optimal for very large sample sizes \( (n > 10^5) \). The other Monte Carlo strategy makes use of a Markov chain to approximate the distribution of the test statistic (GUO and THOMPSON 1992; LAZZERONI and LANGE 1997). This method has the disadvantage of requiring trial-and-error to determine the parameters needed to give the estimated \( P \)-value its desired precision (GUO and THOMPSON 1992) as opposed to the method of independent trials which yields an estimate of the \( P \)-value whose standard error is inversely proportional to the square root of the number of trials. Fortunately, sufficiently many independent trials can be generated for any realistic sample size. For example, ExactoHW generates independent trials for the large sample in Figure 1D at the rate of two million tables per minute while computing the probability, LR and \( U \)-score for each. Since this example is larger than most actual data sets, and since the number of random tables needed for an adequate estimate of the \( P \)-value is well below one million (GUO and THOMPSON 1992), it seems clear that the method is adequate for any realistic sample.

These speeds improve on existing methods of independent sampling by at least two orders of magnitude. With the Markov chain method speed is not usually an issue. However, it is worth noting that the independent trial method given here actually outpaces that of the Markov chain method when tested for a given degree of precision (see supplementary Table S3). The efficiency of independent-sampling Monte Carlo method as implemented in ExactoHW would seem to eliminate any necessity to resort to the Markov chain approach.
One concern with any statistical procedure based on discrete data, and with the exact HW tests in particular, is that the resulting $P$-value takes on only discrete values (HERNANDEZ and WEIR 1989; WEIR 1996). As a result, if an experimenter sets a threshold level for the $P$-value, $\alpha$ say, it may be that the actual probability of rejecting the null hypothesis when it is true is not close to $\alpha$. ROHLFS and WEIR (2008) derived the distribution of the $P$-value for the exact probability test for HW in the case of two alleles and used this information to correct the bias. This consideration can be important when it is necessary to make specific decisions based on the evidence against HW proportions (GOMES et al. 1999; SALANTI et al. 2005; ZOU and DONNER 2006). On the other hand, for most situations where no immediate decision is required, one can follow the advice of YATES (1984) who recommended for discrete data that researchers simply report the calculated $P$-value itself without worrying about whether it lies above or below an arbitrary cutoff point. That way, readers can interpret the exact $P$-value as a measure of the strength or weakness of the case against the population being in HW proportions and the genotyping being accurate and complete.

ACKNOWLEDGEMENTS

Carter Denniston and Jeff Rohl contributed many useful ideas concerning discrete statistical methods and the use of recursion and distributed computation. This work was supported by grant GM30948 from the National Institutes of Health.
REFERENCES


FREEMAN, G. H., and J. H. HALTON, 1951 Note on an Exact Treatment of Contingency, Goodness of Fit and Other Problems of Significance. *Biometrika* **38**: 141-149.


ROUSSET, F., 2008 genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources 8: 103-106.


FIGURES AND LEGENDS

**Figure 1. Sample data sets.** Examples which have been used in previous discussions of exact tests for HW proportions. For each data set, a triangular matrix of genotype counts is shown next to the vector of allele counts. (A) From Table 2, bottom row, of LOUIS and DEMPSTER (1987). (B) From Figure 2 of GUO and THOMPSON (1992). (C) From the documentation included with the GENEPOP software package (ROUSSET 2008). (D) From Figure 5 of GUO and THOMPSON (1992).
Figure 2. Distribution of test statistics. The likelihood ratio and probability was computed for each of the 139 possible tables with allele counts \( m = [9 \ 6 \ 3 \ 1 \ 1] \). Overlapping symbols are indicated by darker shading. Dashed lines intersect at the specific table (see text) whose \( P \)-value is being evaluated by the two exact tests. Both axes are logarithmically scaled.
**Figure 3. Power comparisons.** Contour plots of the ratio of the power of the exact likelihood ratio test (numerator) and the probability test (denominator) for the case of two alleles. Sample sizes and $\alpha$ levels are as shown. Each plot was constructed from a grid of 2687 points distributed uniformly throughout the parameter space of allele frequency and inbreeding coefficient. For each such point, the power was determined by generating all possible multinomial samples and summing the probabilities of those whose $P$-value is less than or equal to $\alpha$. *Mathematica* (Wolfram Research Inc.) was used to draw contour curves from the computed power ratios. Supplemental file S2 contains similar contour plots covering more sample sizes.
Figure 4. Network diagrams. The tables with a given set of allele counts can be represented by the paths through a network of nodes connected by arcs. Each path begins with the leftmost node and proceeds rightward. Each node is labeled with a string of digits indicating the residual allele counts at that point. (A) Network for the case of two copies of each of four alleles. There are 17 paths from (2222) to (0000). The dashed line represents the sample in which each homozygote is observed once. (B) Network showing the 139 paths for the case of \( m = [9 \ 6 \ 3 \ 1 \ 1] \). The dashed lines specify the table which is indicated by the intersection of dashed lines in Figure 2.
Figure 5. Numbers of tables. Contour plots show the numbers of tables for each possible set of allele counts when \( n = 100 \) diploids and \( k = 3 \) alleles. (A) The exact number of tables was computed by the generating function method of Equation 4. (B) The approximate numbers of tables were found by multiplying the multivariate normal density (Equation 11) by the total cardinality (Equation 5). Note that the counts of the first two alleles are indicated while the third allele is implicit, as \( m_3 = 2n - m_1 - m_2 \).
**Figure S1.** Distribution of test statistics. The likelihood ratio and probability was computed for each of a set of randomly drawn tables from the allele counts shown in Figure 1. Darker shading in symbols indicates overlapping points. Both axes are logarithmically scaled. (A-D) as in Figure 1.
**Figure S2.** Power Comparisons. Plots are similar to those in Figure 2 except that the denominator is the power computed using the Chi Square value to order tables instead of the probability.