Testing Hardy-Weinberg equilibrium with a simple root-mean-square statistic

Rachel Ward[1] and Raymond J. Carroll[2]

Abstract

We provide evidence that, in certain circumstances, a root-mean-square test of goodness-of-fit can be significantly more powerful than state-of-the-art tests in detecting deviations from Hardy-Weinberg equilibrium. Unlike Pearson's chi-square test, the log–likelihood-ratio test, and Fisher's exact test, which are sensitive to *relative* discrepancies between genotypic frequencies, the root-mean-square test is sensitive to *absolute* discrepancies. This can increase statistical power, as we demonstrate using benchmark data sets and simulations, and through asymptotic analysis.

Some Key Words: Chi-square test; Common genotypes; Exact tests; Goodness-of-fit; Hardy-Weinberg equilibrium; p-value; Power; Root-mean-square.

Short title: Hardy-Weinberg equilibrium

[1]Department of Mathematics, RLM 10.144, University of Texas at Austin, 2515 Speedway, Austin, TX 78712, U.S.A., rward@math.utexas.edu. Supported in part by a Donald D. Harrington Faculty Fellowship, Alfred P. Sloan Research Fellowship, and DOD-Navy grant N00014-12-1-0743.

[2]Department of Statistics, 3143 TAMU, Texas A&M University, College Station, TX 77843, U.S.A., carroll@stat.tamu.edu. Research supported by a grant from the National Cancer Institute (R37-CA057030). This publication is based in part on work supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

1   Introduction

In 1908, G. H. Hardy (Hardy, 1908) and W. Weinberg (Weinberg, 1908) independently derived mathematical equations to corroborate the theory of Mendelian inheritance, proving that in a large population of individuals subject to random mating, the proportions of alleles and genotypes at a locus stay unchanged unless specific disturbing influences are introduced. Today, Hardy-Weinberg equilibrium (HWE) is a common hypothesis used in scientific domains ranging from botany (Weising, 2005) to forensic science (Council, 1996) and genetic epidemiology (Sham, 2001; Khoury et al., 2004). Statistical tests of deviation from Hardy-Weinberg equilibrium are fundamental for validating such assumptions. Traditionally, Pearson's chi-square goodness-of-fit test, or an asymptotically-equivalent variant such as the log–likelihood-ratio test, was used for this assessment. Before computers became readily available, the asymptotic chi-square approximation for the statistics used in these tests, however poor, was the only practical means for drawing inference. With the now widespread availability of computers, exact tests can be computed effortlessly, opening the door to more powerful goodness-of-fit tests. In their seminal paper, Guo & Thompson (1992) campaigned for an exact test of HWE based on the likelihood function. While their work renewed interest in conditional exact tests for Hardy-Weinberg equilibrium (Raymond & Rousset, 1995; Diaconis & Sturmfels, 1998; Wigginton et al., 2005), likelihood-based tests have also been subject to criticism, and there is little evidence that such tests are more powerful than other exact tests, such as those based on likelihood-ratios (Engels, 2009) or the root-mean-square.

In this article, we demonstrate, using the classical data sets from Guo & Thompson (1992) and several numerical experiments, that goodness-of-fit tests based on the root-mean-square distance can be up to an order-of-magnitude more powerful than all of the classic tests at detecting meaningful deviations from Hardy-Weinberg equilibrium. The classic tests, tuned to detect *relative* discrepancies, can be blind to overwhelmingly large discrepancies among common genotypes that are drowned out by expected finite-sample size fluctuations in rare genotypes. The root-mean-square statistic, on the other hand, is tuned to detect deviations in *absolute* discrepancies, and easily detects large discrepancies in common genotypes.

None of the statistics we consider produces a test that is uniformly more powerful than any other. At the very least, the root-mean-square statistic and the classic statistics focus on complementary classes of alternatives, and their combined p-values provide a more informative test than either p-value used on its own.

The results of our analysis are consistent with the numerous experiments conducted in recent work (Perkins et al., 2013), which highlight the power of the root-mean-square statistic over classic statistics in detecting meaningful discrepancies in nonuniform distributions. Tygert (2012) provides several representative examples for which the root-mean-square test is more powerful than Fisher's exact test for homogeneity in contingency-tables.

This article is structured as follows: in Section 2 we recall the set-up and motivation for testing Hardy-Weinberg equilibrium. We describe the relevant test statistics in Section 3, and in Section 4 we compare the performance of these statistics on the classic data sets from Guo and Thompson, and also compare the power and Type I error of the statistics in detecting deviations due to inbreeding and selection. We provide an asymptotic analysis of the various statistics in Section 5 to highlight the limited power of the classic statistics compared to the root-mean-square statistic in distinguishing important classes of deviations from Hardy-Weinberg equilibrium, and end with concluding remarks in Section 6. **Supplementary Material** includes pseudocode for algorithms and proofs of technical results.

2 Hardy-Weinberg equilibrium: set-up and motivation

Recall that a *gene* refers to a segment of DNA at a particular location (locus) on a chromosome. The gene may assume one of several discrete variations, and these variants are referred to as *alleles*. An individual carries two alleles for each autosomal gene — one allele selected at random from the pair of alleles carried by the mother, and one allele selected at random from the pair of alleles carried by the father. These two alleles, considered as an unordered pair, constitute the individual's *genotype*. A gene having r alleles $A_1, A_2, \ldots, A_r$ has $r(r+1)/2$ possible genotypes. These genotypes are naturally indexed over a lower-triangular array as in Figure 1.

A population is said to be in *Hardy-Weinberg Equilibrium* (HWE) if the following holds.

Figure 1: Enumeration of genotypes for a gene having r alleles $A_1, A_2, \ldots, A_r$.

If $p_{j,k}$ is the relative proportion of genotype $\{A_j, A_k\}$ in the population, and if θ_k is the proportion of allele A_k in the population, then the system is in HWE if

$$p_{j,k} = p_{j,k}(\theta_j, \theta_k) = \begin{cases} 2\theta_j\theta_k, & j > k \\ \theta_k^2, & j = k. \end{cases} \tag{1}$$

3 Testing Hardy-Weinberg equilibrium

A random sample of n genotypes $X_1, X_2, \ldots X_n$ from this population can be regarded as a sequence of independent and identical draws from the multinomial distribution specified by probabilities

$$\mathrm{pr}(X_i = \{A_j, A_k\}) = p_{j,k}, \quad 1 \leq k \leq j \leq r. \tag{2}$$

If $n_{j,k}$ realizations of genotype $\{A_j, A_k\}$ are observed in the sample of n genotypes, then the number of instances of allele A_j in the observed sample of $2n$ alleles is

$$n_j = \sum_{k=j}^{r} n_{k,j} + \sum_{k=1}^{j} n_{j,k}, \qquad j = 1, \ldots, r. \tag{3}$$

In order to gauge the consistency of the sample counts $(n_{j,k})$ with Hardy-Weinberg equilibrium, we must first specify the $r - 1$ free parameters $\theta_1, \theta_2, \ldots, \theta_{r-1}$ corresponding to the underlying allele proportions in the HWE model (1). The *observed* proportions of alleles, $n_1/(2n), n_2/(2n), \ldots, n_{r-1}/(2n)$, are the maximum likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_{r-1}$ in the family of HWE equilibrium equations (1); these parameter specifications give rise to the *model counts* of genotypes under Hardy-Weinberg equilibrium,

$$m_{jk} = (2 - \delta_{jk})(n_j \ n_k)/(4n), \tag{4}$$

3

(a) Example 1: $n = 45$.

	A_1	A_2	A_3	A_4
A_1	**0** / .6722			
A_2	**3** / 3.667	**1** / 5		
A_3	**5** / 3.667	**18** / 10	**1** / 5	
A_4	**3** / 2.322	**7** / 6.333	**5** / 6.333	**2** / 2.006

(b) Example 2: $n = 8297$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_1	**1236** / 1206.9								
A_2	**120** / 121.67	**3** / 3.0662							
A_3	**18** / 17.926	**0** / .90352	**0** / .06656						
A_4	**982** / 1057.6	**55** / 53.308	**7** / 7.8541	**249** / 231.70					
A_5	**32** / 28.605	**1** / 1.4418	**0** / .21243	**12** / 12.533	**0** / .16949				
A_6	**2582** / 2556.2	**132** / 128.84	**20** / 18.982	**1162** / 1120.0	**29** / 30.291	**1312** / 1353.4			
A_7	**6** / 5.3396	**0** / .26913	**0** / .03965	**4** / 2.3395	**0** / .06328	**4** / 5.6543	**0** / .00591		
A_8	**2** / .76281	**0** / .03845	**0** / .00566	**0** / .33422	**0** / .00904	**0** / .80776	**0** / .00169	**0** / .00012	
A_9	**115** / 127.01	**5** / 6.4015	**2** / .94317	**53** / 55.647	**1** / 1.5051	**149** / 134.49	**0** / .28094	**0** / .04014	**4** / 3.3412

(c) Example 3: $n = 30$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	**3** / 1.875							
A_2	**4** / 3.5	**2** / 1.633						
A_3	**2** / 2.75	**2** / 2.567	**2** / 1.01					
A_4	**3** / 3.0	**3** / 2.8	**2** / 2.2	**1** / 1.2				
A_5	**0** / .5	**1** / .467	**0** / .367	**0** / .4	**0** / .033			
A_6	**0** / .5	**0** / .467	**0** / .367	**0** / .4	**0** / .067	**1** / .033		
A_7	**0** / .25	**0** / .233	**1** / .183	**0** / .2	**0** / .033	**0** / .033	**0** / .0083	
A_8	**0** / .75	**0** / .7	**0** / .55	**2** / .6	**1** / .1	**0** / .1	**0** / .050	**0** / .075

Figure 2: The three data sets from Guo & Thompson (1992). Observed counts are in bold and expected counts under HWE are below.

4

where δ_{jk} is the Kronecker delta function with $\delta_{jk} = 1$ if $j = k$ and $= 0$ otherwise. A *goodness-of-fit* test serves as an omnibus litmus test to gauge consistency of the data with HWE. Ideally, the goodness-of-fit test should be sensitive to a wide range of possible local alternatives; more realistically, several different goodness-of-fit tests can be used jointly, each sensitive to its own class of alternatives. If a nonparametric test as such indicates deviation from equilibrium, different parametric tests can then be used to elucidate particular effects of the deviation such as directions of disequilibrium or level of inbreeding. Several parametric Bayesian methods have been proposed as well (Chen & Thomson, 1999; Shoemaker et al., 1998; Ayres & Balding, 1998; Lauretto et al., 2009; Li & Graubard, 2009; Consonni et al., 2011). In this paper we will focus only on nonparametric (or nearly nonparametric) tests of fit, but we emphasize that goodness-of-fit tests should be combined with Bayesian approaches and other types of evidence for and against the HWE hypothesis before drawing final inference.

3.1 Goodness-of-fit testing

A goodness-of-fit test compares the model and empirical distributions using one of many possible measures. Three classic measures of discrepancy, all special cases of Cressie-Read power divergences, are Pearson's χ^2-divergence

$$X^2 = \sum_{1 \le k \le j \le r} (n_{j,k} - m_{j,k})^2 / m_{j,k}, \tag{5}$$

the log–likelihood-ratio or g^2 divergence,

$$G^2 = 2 \sum_{1 \le k \le j \le r} n_{j,k} \log(n_{j,k}/m_{j,k}), \tag{6}$$

and the Hellinger distance

$$H^2 = 4 \sum_{1 \le k \le j \le r} \left(\sqrt{n_{j,k}} - \sqrt{m_{j,k}} \right)^2. \tag{7}$$

Another classic measure of discrepancy is the negative log–likelihood function, which is based directly on the likelihood function for the multinomial distribution,

$$L = -\log(\mathcal{L}), \text{ where} \tag{8}$$

$$\mathcal{L}(n_{j,k}; n, m_{j,k}) = \text{Prob}\Big(N_{1,1} = n_{1,1},\ N_{2,1} = n_{2,1},\ \ldots, N_{r,r} = n_{r,r}\Big)$$

$$= \frac{n!}{n_{1,1}! n_{1,2}! \ldots n_{r,r}! n^n} m_{1,1}^{n_{1,1}} m_{1,2}^{n_{1,2}} \ldots m_{r,r}^{n_{r,r}}. \tag{9}$$

The negative log-likelihood statistic (8) looks similar to the log–likelihood-ratio statistic G^2, but there is an important distinction to be made: the log–likelihood-ratio, which sums the logarithms of *ratios* between observed and expected counts, is a proper divergence. The negative log–likelihood function is not a divergence, and this results in several undesirable properties that have led many to criticize its use (Gibbons & Pratt, 1975; Radlow & Alf, 1975; Engels, 2009).

The negative log–likelihood function does have something in common with the power-divergence discrepancies: under the null-hypothesis, the negative log–likelihood statistic L and the power divergence statistics X^2, G^2, and H^2 all become a chi-square random variable with $r(r-1)/2 - 1$ degrees of freedom as the number of draws n goes to infinity and number of alleles remains fixed (Brownlee, 1965). Before computers became widely available, using a statistic with known asymptotic approximation was necessary for obtaining any sort of approximate p-value. The exact (non-asymptotic) p-values for these statistics or any other measure of discrepancy can now be computed effortlessly using Monte-Carlo simulation.

In this paper, we distinguish two types of commonly-used p-values, which we refer to as the *plain* p-value and *fully conditional* p-value. One could also consider Bayesian p-values (Gelman, 2003), among other formulations.

To compute the plain p-value, one repeatedly simulates n independent and identically distributed draws from the model multinomial distribution $(m_{j,k}/n)$. For each simulation i, the genotype counts $N_{j,k}^{(i)}$, allelic counts $N_j^{(i)} = \big(\sum_{k=j}^{r} N_{k,j}^{(i)} + \sum_{k=1}^{j} N_{j,k}^{(i)} \big)$, allelic proportions $\Theta_j^{(i)} = N_j^{(i)}/(2n)$, and equilibrium model counts associated to this sample, $M_{j,k}^{(i)} = (2 - \delta_{j,k}) N_j^{(i)} N_k^{(i)}/(4n)$, are computed. The plain p-value is the fraction of times the discrepancy between the simulated counts $(N_{j,k}^{(i)})$ and their model counts $(M_{j,k}^{(i)})$ is at least as large as the measured discrepancy between the observed counts $n_{j,k}$ and their model counts $m_{j,k}$. Henze

(1996) shows that this procedure has an asymptotically correct Type I error for fixed r as $n \to \infty$. This procedure for producing p-values can be viewed as a parametric bootstrap approximation, as discussed for example, by Efron & Tibshirani (1993), Henze (1996), and Bickel et al. (2006).

The fully conditional p-value corresponds to imposing additional restrictions on the probability space associated to the null hypothesis. To compute the fully conditional p-value, the observed counts of alleles, $n_1, \ldots, n_r$, are treated as known quantities in the model, to remain fixed upon hypothetical repetition of the experiment. This would hold, for example, if the sample population used in the experiment were the entire population of individuals. More specifically, one repeatedly simulates n i.i.d. draws from the *hypergeometric* distribution that results from conditioning the multinomial model distribution $(m_{j,k}/n)$ on the observed allele counts, $N_1 = n_1, N_2 = n_2, \ldots, N_r = n_r$. Guo & Thompson (1992) provided an efficient means for performing such a simulation: apply a random permutation to the sequence

$$\mathcal{A} = \Big\{ \overbrace{A_1, A_1, \ldots, A_1}^{n_1}, \overbrace{A_2, \ldots, A_2}^{n_2}, \ldots, \overbrace{A_r \ldots A_r}^{n_r} \Big\}, \tag{10}$$

and identify the pairs $\{A_{2j}, A_{2j+1}\}$. The fully conditional p-value is the fraction of times the discrepancy between the simulated counts $(N_{j,k}^{(i)})$ and the model counts $(m_{j,k})$ is at least as large as the measured discrepancy.

Pseudocode for calculating plain and fully conditional p-values is provided in Algorithms 1 and 2 of Appendix S.1 in the **Supplementary Material**.

3.2   The root-mean-square statistic

A natural measure of discrepancy for goodness-of-fit testing which has not received as much attention in the literature is the root-mean-square distance,

$$f = \left\{ \frac{2}{n^2 r(r+1)} \sum_{1 \le k \le j \le r} (n_{j,k} - m_{j,k})^2 \right\}^{1/2}. \tag{11}$$

In contrast to the classic statistics, the asymptotic distribution for the root-mean-square statistic F in the limit of infinitely many draws and fixed alleles, while completely well-defined and efficient to compute, depends on the model distribution, as described by Perkins

7

et al. (2011a,b). Using the pseudocode provided in Algorithms 1 and 2 of Appendix S.1 in the **Supplementary Material**, we can compute p-values for the root-mean-square statistic.

4 Numerical results

4.1 Benchmark data sets

We next compare the performances of the root-mean-square statistic and the classic statistics in detecting deviations from Hardy-Weinberg equilibrium. We first evaluate the performance of the various statistics on three benchmark data sets from Guo & Thompson (1992). The three data sets, which we refer to as Examples 1, 2, and 3, are represented in Figure 2 as lower-triangular arrays of counts. The bold entry in each cell corresponds to the number $n_{j,k}$ of observed counts of genotype $\{A_j, A_k\}$ in the sample, and the second entry in each cell corresponds to the expected number $m_{j,k}$ of counts under HWE.

For each example, and for each of the five test statistics X^2, G^2, H^2, L, and F, we calculate both the plain and fully conditional p-values using $16,000,000$ Monte-Carlo simulations for each calculation. The results of the analyses of Examples 1-3 are displayed in Table 1. We next discuss the results for each example.

4.1.1 Graphical views of the data

Figures 3, 4, and 5 contain boxplots displaying the median, upper and lower quartiles, and whiskers reaching from the 1st to 99th percentiles for relative root-mean-square discrepancies and relative chi-square discrepancies simulated under the plain Hardy-Weinberg equilibrium null hypothesis for the data sets from Examples 1, 2, and 3. The boxplots are for simulated data, whereas the large open circles indicate the observed data. For a detailed description of these plots, we refer the reader to Appendix A.1. In the chi-square boxplots, we see the division by expected proportion in the summands of the chi-square discrepancy (5) reflected in the larger contribution of relative discrepancies to the reported p-values; in contrast, we see the equal-weighting of the summands of the root-mean-square distance (11) reflected in the larger contribution of absolute discrepancies to the reported root-mean-square p-values.

In Section 5, we will see that all of the classic statistics, not just the chi-square statistic, are sensitive to relative rather than absolute discrepancies.

4.1.2 Interpretation of the results for Example 1

Comparing the boxplots in Figure 3, we see that both chi-square and root-mean-square tests report a significant deviation in the largest index, among others. The largest index corresponds to the 18 observed counts versus 10 expected counts of genotype $\{A_3, A_2\}$ in Example 1. However, the p-value reported by the root-mean-square test is an order of magnitude smaller than the p-value reported by chi-square test, as this discrepancy is larger compared to expected root-mean-square fluctuations than it is compared to expected chi-square fluctuations. In the chi-square summation, the statistical significance of this deviation (as well as the deviations in indices 6 and 7) is washed out by large expected relative deviations in the rare genotypes.

Table 1: Plain and fully-conditional (FC) p-values for Pearson's statistic X^2, the log–likelihood-ratio statistic G^2, the Hellinger distance H^2, the negative log–likelihood statistic L, and the root-mean-square statistic F, for the observed genotypic counts in Examples 1-3 to be consistent with the Hardy-Weinberg equilibrium model (1). With 99% confidence, p-values are correct to $\pm.001$.

	Example 1		Example 2		Example 3	
Statistic	Plain p-val	FC p-val	Plain p-val	FC p-val	Plain p-val	FC p-val
X^2	.693	.709	.020	.020	.015	.026
G^2	.600	.630	.013	.013	.181	.276
H^2	.562	.602	.027	.025	.307	.449
L	.648	.714	.016	.018	.155	.207
F	**.039**	**.039**	**.002**	**.002**	.885	.917

4.1.3 Interpretation of the results for Example 2

The distribution of discrepancies in Figure 4 can be interpreted similarly to the boxplots from Figure 3: both the chi-square and root-mean-square tests report a **statistically** significant deviation in the 5th-largest index, corresponding to the 982 observed counts versus 1057.6 expected counts of genotype $\{A_4, A_1\}$ in Example 2. However, the p-value reported by the

root-mean-square test is an order of magnitude smaller than the p-value reported by chi-square test, as this discrepancy is larger compared to expected root-mean-square fluctuations than it is compared to expected chi-square fluctuations. In the chi-square summation, the statistical significance of this deviation is washed out by large expected relative deviations in the rare genotypes. In contrast to the $n = 45$ draws from Example 1, this data set contains $n = 8297$ draws; we infer that the qualitative differences between the root-mean-square and chi-square statistic are not unique to small sample-size data.

4.1.4 Interpretation of the results for Example 3

Comparing the expected and observed chi-square discrepancies in Figure 5(b), we might posit that the small p-value of .015 that the chi-square test gives to the data in Example 3 depends strongly on the discrepancy at the 4th index on the plot, corresponding to a single draw of genotype $\{A_6, A_6\}$. By removing this draw from the data set and re-running the chi-square goodness-of-fit test on the remaining $n = 29$ draws, the chi-square statistic X^2 returns a p-value of .207, well over an order of magnitude larger than the previous p-value, confirming that the small p-value given by the chi-square statistic for the data set in Figure 2(c) is the result of observing a single rare genotype. The root-mean-square statistic is not as sensitive to this discrepancy.

4.2 Power analyses

We now compare the power and Type I error for Pearson's statistic X^2, the log–likelihood-ratio statistic G^2, the Hellinger distance H^2, the negative log–likelihood statistic L, and the root-mean-square statistic F, in detecting practical deviations of genotype frequencies from those expected under HWE, namely populations with increased homozygosity (as due to inbreeding), populations with increased heterozygosity, and populations of genotypes undergoing selection (Chen & Thomson, 1999; Ayres & Balding, 1998; Lauretto et al., 2009). The results in Table 2 support the assertion that the root-mean-square statistic and the classic statistics focus their power on complementary classes of alternatives. In this section we will consider four parameter specifications:

1. **Alternative:** $r = 10$, $n = 100$, and $\theta_1 = \theta_2 = 1/3$, and $\theta_j = 1/24$ for $3 \leq j \leq 10$

2. **Alternative:** $r = 10$, $n = 100$, and $\theta_j \sim 1/j$ for $1 \leq j \leq 10$

3. **Alternative:** $r = 10$, $n = 200$, and $\theta_1 = \theta_2 = 1/3$, and $\theta_j = 1/24$ for $3 \leq j \leq 10$

4. **Alternative:** $r = 20$, $n = 200$, and $\theta_j \sim 1/j$ for $1 \leq j \leq 20$

4.2.1 Deviations due to selection

When there is selection for or against a particular allele or genotype in the population, the result is an excess or deficiency of genotypes carrying a particular allele or pair of alleles compared to what would be expected under HWE. To account for selection, one introduces fitness parameters $w_{j,k} > 0$ into the HWE equations,

$$p_{j,k} = \begin{cases} 2(w_{j,k}/\bar{w})\theta_j\theta_k, & 1 \leq k < j \leq r \\ (w_{k,k}/\bar{w})\theta_k^2, & j = k. \end{cases} \tag{12}$$

where $\bar{w}$ is a normalization constant.

We consider the scenario where the common allele A_1 is undergoing selection, so that genotypes carrying allele A_1 have higher fitness in the population:

$$w_{j,k} = \begin{cases} 1.5, & k = 1, \\ 1, & \text{else.} \end{cases} \tag{13}$$

The power and Type I errors of the various statistical tests in detecting deviations from HWE due to selection for common alleles are listed in Table 2. The root-mean-square statistic appears to be uniformly more powerful than the classic statistics while maintaining the correct asymptotic Type I error rate. We will provide theoretical justification for these observations through an asymptotic analysis in Section 5.

4.2.2 Deviations due to inbreeding

We now consider genotypic distributions parameterized by an inbreeding coefficient, f, which describes the extent to which members of the population with similar genetic make-up are more or less likely to mate with each other:

$$p_{j,k} = \begin{cases} 2\theta_j\theta_k(1-f), & j > k \\ \theta_k^2 + f\theta_k(1-\theta_k), & j = k. \end{cases} \quad 1 \leq k \leq j \leq r \tag{14}$$

Table 2: Statistical power and Type I error of the various tests of HWE against deviations due to <u>selection</u>, i.e., deviations of the form (12) with parameters as specified in Alternatives 1-4 and fitness parameters (13) and deviations due to <u>inbreeding</u>, i.e. deviations of the form (14) with parameters as specified in Alternatives 1-4 and inbreeding parameter $f = 1/10$. Power and Type I errors are at the 5% significance level, and computed using 5000 simulations from the alternative distribution and 5000 Monte Carlo trials per each simulation.

Deviations due to selection for common allele

Statistic	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
	Power	Type I	Power	Type I	Power	Type I	Power	Type I
X^2	.04	**.05**	.01	**.05**	.04	**.05**	< .01	**.05**
G^2	.07	.06	.02	.07	.07	.06	.01	.08
H^2	.08	.06	.05	.07	.08	**.05**	.01	.07
L	.03	.04	.01	.04	.04	.04	< .01	.04
F	**.13**	**.05**	**.12**	**.05**	**.19**	**.05**	**.23**	**.05**

Deviations due to inbreeding

Statistic	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
	Power	Type I	Power	Type I	Power	Type I	Power	Type I
X^2	.34	**.05**	.34	**.05**	.60	.04	.64	.06
G^2	.29	.06	.33	.06	.48	.06	.64	.08
H^2	.18	.07	.22	.06	.28	**.05**	.42	.07
L	**.39**	**.05**	**.36**	.04	**.63**	.04	**.70**	.03
F	.16	**.05**	.16	**.05**	.26	**.05**	.29	**.05**

Hardy Weinberg Equilibrium corresponds to $f = 0$. A negative value $f < 0$ corresponds to a deficiency of homozygotes, while a positive value of f corresponds to an excess of homozygotes. Table 2 displays the power of the various tests against alternatives of the form (14) with positive inbreeding parameter. The root-mean-square statistic appears to be less powerful than the classic statistics in detecting deviations due to inbreeding, but does appear to reach the correct asymptotic Type I error rate more quickly.

5   An asymptotic power analysis

In this section we give theoretical justification to our assertion that the root-mean-square statistic can be more powerful than the classic statistics in detecting deviations from Hardy-

Weinberg equilibrium. To model the setting where the number of draws and number of genotypes are of the same magnitude, we consider the limit in which the number of alleles and number of draws go to infinity *together*, so that the asymptotic chi-square approximation to the classic statistics is not valid in this limit. Our method is to create data sets such that the root-mean-square statistic has asymptotic power one while the chi-square statistics have asymptotic power zero.

We consider a gene having $r + 1$ alleles, one common allele and r rare alleles. The *Common Allele* data set we consider involves $n = 3r$ observed genotypes, distributed as indicated below.

Table 3: Common Allele data set

$n = 3r$ observed genotypes
$n_{1,1} = r$ of type $\{A_1, A_1\}$, $\quad n_{1,1}/n = 1/3$
$n_{1,k} = 2$ of type $\{A_1, A_k\}$, $\quad n_{1,k}/n = 2/(3r)$, $\ 2 \leq k \leq r + 1$
$n_{j,k} = 0$ of type $\{A_j, A_k\}$, $\quad n_{j,k}/n = 0$, $\qquad 2 \leq j \leq k \leq r + 1$
$n_1 = 4r$ alleles of type A_1, $\quad n_1/(2n) = 2/3$
$n_k = 2$ $\ $ alleles of type A_k, $\quad n_k/(2n) = 1/(3r)$, $\quad 2 \leq k \leq r + 1$.

The maximum-likelihood model counts for the Common Allele data set are

$$
\begin{cases}
m_{1,1} = 4r/3, \\
m_{1,k} = 4/3, & 2 \leq k \leq r + 1, \\
m_{k,k} = 1/(3r), & 2 \leq k \leq r + 1, \\
m_{j,k} = 2/(3r), & 2 \leq j < k \leq r + 1, \quad j < k.
\end{cases}
\tag{15}
$$

To see that the Common Allele data set becomes increasingly inconsistent with the Hardy-Weinberg model as r increases, observe that, under the null hypothesis, we would expect in a sample of $n = 3r$ genotypes to see $r/3 = \sum_{j=2}^{r+1} \sum_{k=2}^{r+1} m_{j,k}$ genotypes containing only rare alleles. The Common Allele data set however contains *no* genotypes containing only rare alleles. In spite of this inconsistency, we will prove that the plain p-values for each of the four classic statistics X^2, G^2 and H^2, *converge to 1* as $r \to \infty$, indicating zero asymptotic power. In contrast, the p-value for the root-mean-square statistic converges to zero.

Theorem 1 *In the limit as $r \to \infty$, the plain p-values (as computed via Algorithm 1 of Appendix S.1 in the* **Supplementary Material***) given by X^2, the log–likelihood-ratio statistic G^2 and the Hellinger distance H^2 for the Common Allele data set to be consistent with the Hardy-Weinberg equilibrium model all converge to 1, while the plain p-value for the root-mean-square statistic converges to 0.*

The proof is given in Appendix S.2. of the **Supplementary Material**.

Figure 6 shows that the convergence of the classic p-values to 1, and of the root-mean-square p-value to 0, occurs very quickly. This convergence is demonstrated for both the plain and fully conditional p-values, even though Theorem 1 applies directly only to the plain p-values.

Finally, the particular distribution of the draws in the Common Allele data set was somewhat arbitrary. However, a similar asymptotic analysis holds for many other data sets. For example, we could have considered instead a data set involving two, three, or four common alleles, or one common allele and three fairly-common alleles, and so on.

6   Concluding Remarks

We have proposed the use of a simple root-mean-square statistic for testing deviations from Hardy-Weinberg equilibrium. The classic tests, tuned to detect *relative* discrepancies, can be blind to overwhelmingly large discrepancies among common genotypes that are drowned out by expected finite-sample size fluctuations in rare genotypes. The root-mean-square statistic, on the other hand, is tuned to detect deviations in *absolute* discrepancies, and easily detects large discrepancies in common genotypes. We demonstrated this in the analysis of three benchmark data sets of Guo & Thompson (1992), in one of which only the root-mean-square statistic gave overwhelming evidence of a departure from Hardy-Weinberg equilibrium, and another of which it gave an order of magnitude smaller p-value. We also found that the root-mean-square test can be significantly more powerful at detecting deviations from Hardy-Weinberg equilibrium arising from selection. These numerical results were complemented by the asymptotic power analysis of Section 5. At the very least, the root-mean-square statistic

and the classic statistics focus on *complementary* classes of deviations from Hardy-Weinberg equilibrium (see Figure 4), and their combined P-values provide a more fortified test than either P-value used on its own.

Acknowledgment

We thank Mark Tygert, Andrew Gelman, Abhinav Nellore, and Will Perkins for their helpful contributions.

Supplementary Material

Supplementary Material contains Algorithms 1 and 2 of Appendix S.1 and the proof of Theorem 1 in Appendix S.2.

Software availability

Code for calculating plain and fully conditional p-values using the root-mean-square test statistic is available in R from http://math.utexas.edu/~rward. With appropriate citation, the code is freely available for use and can be incorporated into other programs.

References

AYRES, K. & BALDING, D. (1998). Measuring departures from Hardy-Weinberg: a Markov chain Monce Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769–777.

BICKEL, P., RITOV, Y. & STOKER, T. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Annals of Statististics* **34**, 721–741.

BROWNLEE, K. (1965). *Statistical series and methodology in science and engineering.* New York: Wiley, Inc.

CHEN, J. & THOMSON, G. (1999). The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test. *Biometrics* **55**, 1269–1272.

CONSONNI, G., MORENO, E. & VENTURINI, S. (2011). Testing Hardy-Weinberg equilibrium: an objective Bayesian analysis. *Statistics in Medicine* **30**, 62–74.

COUNCIL, N. R. (1996). *The evaluation of forensic DNA evidence.* National Academy Press.

DIACONIS, P. & STURMFELS, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* **26**, 363–397.

EFRON, B. & TIBSHIRANI, R. (1993). *An Introduction To The Bootstrap.* Boca Raton, FL: Chapman and Hall.

ENGELS, W. (2009). Exact tests for Hardy-Weinberg proportions. *Genetics* **183**, 1431–1441.

GELMAN, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* **71**, 369–382.

GIBBONS, J. & PRATT, J. (1975). P-values: Interpretation and methodology. *The American Statistician* **29**, 20–25.

GUO, S. & THOMPSON, E. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.

HARDY, G. (1908). Mendelian proportions in a mixed population. *Science* **28**, 49–50.

HENZE, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models. *The Canadian Journal of Statistics* **24**, 81–93.

KHOURY, M., LITTLE, J. & BURKE, W. (2004). *Human Genome Epidemiology: A Scientific Foundation For Using Genetic Information To Improve Health And Prevent Disease.* New York: Oxford University Press.

LAURETTO, M., NAKANO, F., FARIA, S., PEREIRA, C. & STERN, J. (2009). A straight-forward multiallelic significance test for the Hardy-Weinberg equilibrium law. *Genetics and Molecular Biology* **32**, 619–625.

LI, Y. & GRAUBARD, B. (2009). Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data. *Biometrics* **65**, 1096–1104.

MOTWANI, R. & RAGHAVAN, P. (1995). *Randomized Algorithms.* New York, NY: Cambridge University Press.

PERKINS, W., TYGERT, M. & WARD, R. (2011a). Computing the confidence levels for a root-mean-square test of goodness of fit. *Applied Mathematics and Computation* **217**, 9072–9084.

PERKINS, W., TYGERT, M. & WARD, R. (2011b). Computing the confidence levels for a root-mean-square test of goodness of fit, II ArXiv:1009.2260.

PERKINS, W., TYGERT, M. & WARD, R. (2013). χ^2 and classical exact tests often wildly misreport significance; the remedy lies in computers. *Applied and Computational Harmonic Analysis* To appear.

RADLOW, R. & ALF, E. (1975). An alternate multinomial assessment of the accuracy of the χ^2 test of goodness of fit. *Journal of the American Statistical Association* **70**, 811–813.

RAYMOND, M. & ROUSSET, F. (1995). An exact test for population differentiation. *Evolution* **49**, 1280–1283.

SHAM, P. (2001). *Statistics In Human Genetics.* London: Arnold Publishers.

SHOEMAKER, J., PAINTER, I. & WEIR, B. (1998). A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics* **149**, 2079–2088.

TYGERT, M. (2012). Testing the significance of assuming homogeneity in contingency-tables/cross-tabulations ArXiv:1201.1421.

WEINBERG, W. (1908). Über den nachweis der vererbung beim menschen. *Jh. Ver. vaterl. Naturk. Wurttemb.* **64**, 369–382. (English translations in BOYER 1963 and JAMESON 1977).

WEISING, K. (2005). *DNA Fingerprinting In Plants: Principles, Methods, And Applications.* Boca Raton, Florida: Taylor and Francis.

WIGGINTON, J., CUTLER, D. & ABECASIS, G. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893.

Appendix

A.1   Description of Figures 3, 4, and 5

Consider for a sample of genotype counts the linear ordering given by the nondecreasing rearrangement of the Hardy-Weinberg equilibrium model counts: if $m_{[j]}$ denotes the jth smallest expected frequency among all the model genotype frequencies, $1 \leq j \leq r(r+1)/2$, then we denote the corresponding number of draws by $n_{[j]}$, and the corresponding number of observed and expected simulated draws under the (plain) HWE null hypothesis by $N_{[j]}$ and $M_{[j]}$.

The observed root-mean-square discrepancies are

$$d_j^{rms} = \left(m_{[j]} - n_{[j]} \right)^2,$$

while the observed chi-square discrepancies are

$$d_j^{chi} = \frac{\left(m_{[j]} - n_{[j]} \right)^2}{m_{[j]}}.$$

The random vectors of expected root-mean-square discrepancies in n i.i.d. draws from the model distribution are

$$D_j^{rms} = \left(M_{[j]} - N_{[j]} \right)^2,$$

17

and

$$D_j^{chi} = \frac{\left(M_{[j]} - N_{[j]}\right)^2}{M_{[j]}}.$$

To generate the boxplots for the relative root-mean-square discrepancies, we simulated $K = 1000$ realizations of n i.i.d. draws from the HWE model in the respective examples. For each simulation, we computed the vector of root-mean-square discrepancies (A.1) and normalized the vector to sum to 1. We displayed the distribution of discrepancies using a boxplot: for each term j, the median of the distribution $D_j^{(\cdot)} = (D_j^{(i)})_{i=1}^{1000}$ is indicated by the bulls-eye mark $\odot$. The rectangular box around the median extends to the 25th and 75th percentiles of the data, and the whiskers extending from each side of the box reach out to the 1 and 99th percentiles of the data. On top of the boxplot, the observed discrepancies, d_j^{rms}, normalized to sum to 1, are indicated by large open circles.

The chi-square plot for each figure was created by repeating the same set-up as above using the relative chi-square discrepancies.
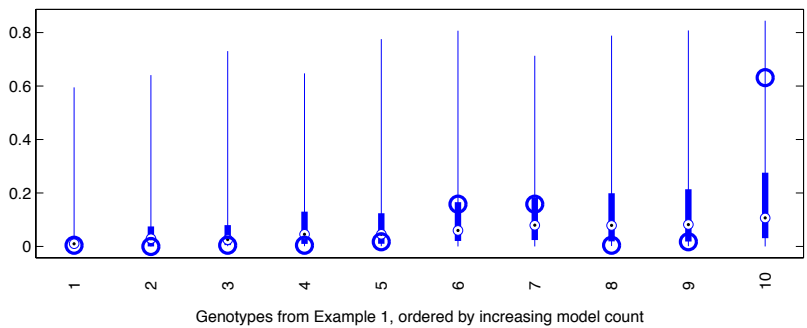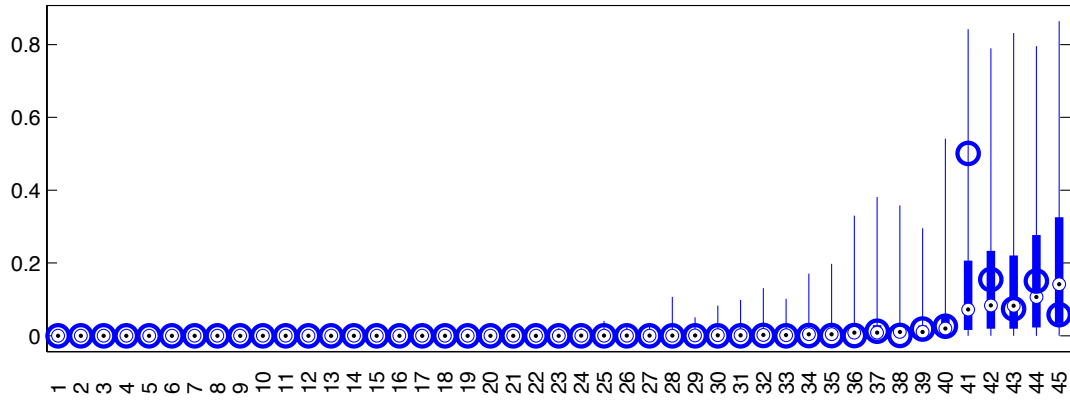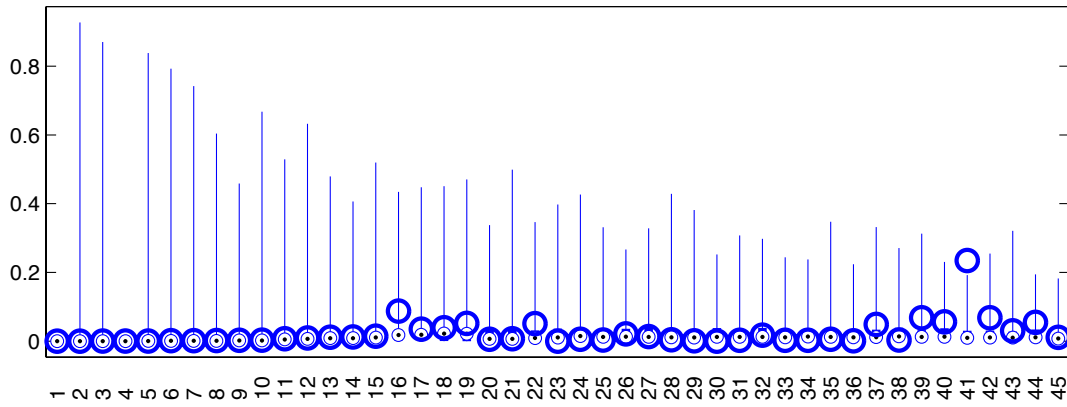
(a)
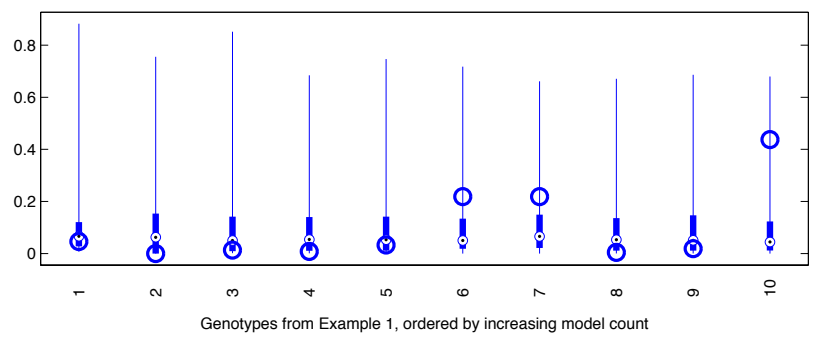
(b)

Figure 3: Example 1. Expected vs. observed relative root-mean-square discrepancies (top plot) and Expected vs. observed relative χ^2 discrepancies (bottom plot).
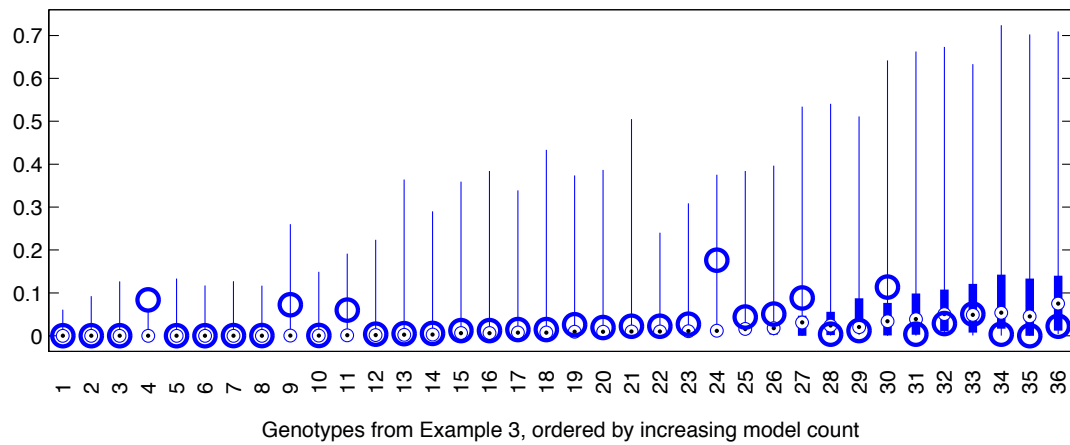
Figure 4: Example 2. Expected vs. observed relative root-mean-square discrepancies (top plot) and Expected vs. observed relative χ^2 discrepancies (bottom plot).
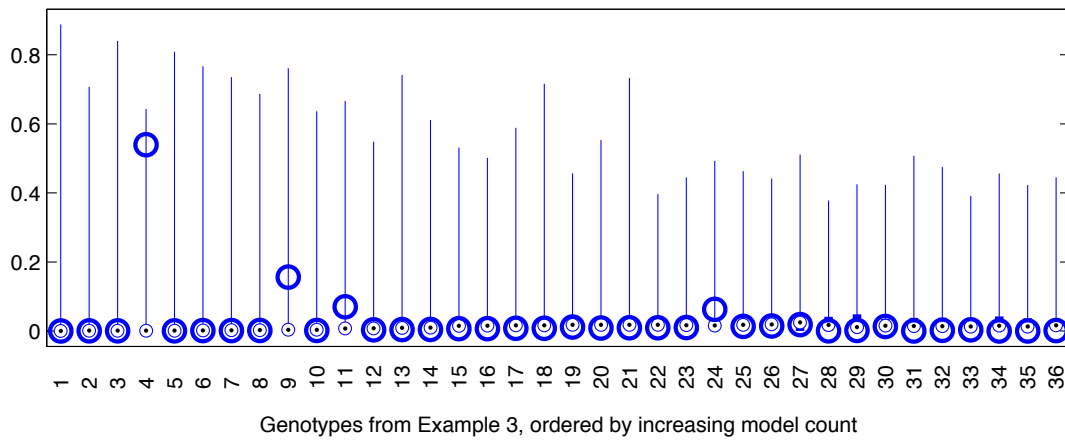
(a)

(b)

Figure 5: Example 3. Expected vs. observed relative root-mean-square discrepancies (top plot) and Expected vs. observed relative χ^2 discrepancies (bottom plot).
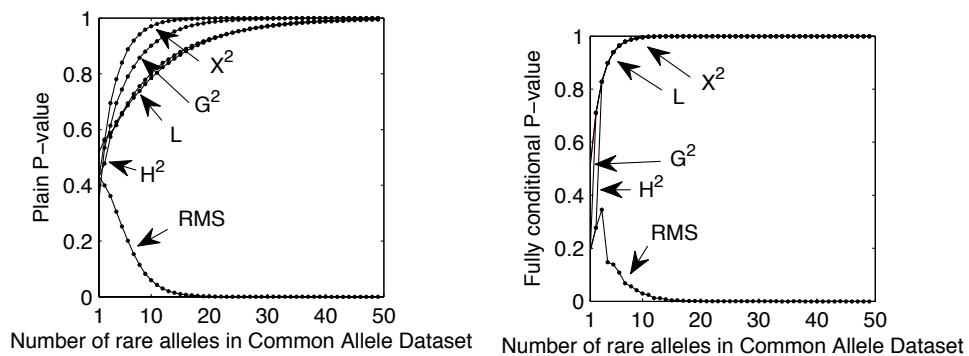
Figure 6: The p-values (accurate to three digits with 99% confidence) for Pearson's statistic X^2, the log–likelihood-ratio statistic G^2, the Hellinger statistic H^2 and the root-mean-square statistic F in the Common Allele data set to be consistent with the Hardy-Weinberg equilibrium model (1), as a function of the number of alleles r. The top plot if for the plain p-values, while the bottom plot is for the conditional p-values.

Supplementary Material to

Testing Hardy-Weinberg equilibrium with a simple root-mean-square statistic

Rachel Ward and Raymond J. Carroll

S.1 Pseudocode for calculating exact p-values

Algorithm 1: Computing the plain p-value

Input: Observed genotype counts $n_{j,k}$, number of Monte Carlo simulations ℓ, and test statistic S (e.g. $S = X^2, G^2, H^2, \dots$)

Output: plain p-value associated to test statistic $S(n_{j,k}, m_{j,k})$

Compute maximum-likelihood model counts $m_{j,k} = (2 - \delta_{jk})(n_j\, n_k)/(4n)$

Measure the discrepancy $s = S(n_{j,k}, m_{j,k})$.

$i \leftarrow 0$

repeat

- $i \leftarrow i + 1$
- Draw n genotypes $X_1^{(i)}, \dots, X_q^{(i)}, \dots, X_n^{(i)}$ i.i.d. from the multinomial model distribution $(m_{j,k}/n)$
- Aggregate simulated genotype counts $N_{j,k}^{(i)} = \#\{q : X_q^{(i)} = \{A_j, A_k\}\}$
- Aggregate simulated allele counts $N_j^{(i)} = \left(\sum_{k=j}^{r} N_{k,j}^{(i)} + \sum_{k=1}^{j} N_{j,k}^{(i)}\right)$ and proportions $\Theta_j^{(i)} = N_j^{(i)}/(2n)$.
- Compute maximum-likelihood counts $M_{j,k}^{(i)} = (2 - \delta_{jk})N_j^{(i)} N_k^{(i)}/(4n)$
- Evaluate simulated discrepancy $S_i = S(N_{j,k}^{(i)}, M_{j,k}^{(i)})$

until $i = \ell$

return plain p-value, $P = \#\{i : S_i \geq s\}/\ell$

S.2 Proof of Theorem 1

The crux of the proof is that, as r increases, relative fluctuations in the rare genotypes simulated under HWE become sufficiently large that the sum of relative discrepancies expected under the null hypothesis exceeds the sum of the observed relative discrepancies. However, the sum of absolute fluctuations expected under the HWE model remains bounded below the sum of the observed absolute discrepancies.

Algorithm 2: Computing the fully conditional p-value

Input: Observed genotype counts $n_{j,k}$ and allele counts n_j, number of Monte Carlo simulations ℓ, and test statistic S (e.g. $S = X^2, G^2, H^2, \dots$)

Output: fully conditional p-value associated to test statistic $S(n_{j,k}, m_{j,k})$

Compute maximum-likelihood model counts $m_{j,k} = (2 - \delta_{jk}) n_j n_k / (4n)$.

Measure the discrepancy $s = S(n_{j,k}, m_{j,k})$.

$i \leftarrow 0$

repeat

- $i \leftarrow i + 1$

- Apply a random permutation to the sequence of alleles as in (10) to obtain n simulated genotypes $X_1^{(i)}, \dots, X_q^{(i)}, \dots, X_n^{(i)}$ with fixed allele counts n_j.

- Aggregate simulated genotype counts $N_{j,k}^{(i)} = \#\{q : X_q^{(i)} = \{A_j, A_k\}\}$

- Evaluate simulated discrepancy $S_i = S(N_{j,k}^{(i)}, m_{j,k})$

until $i = \ell$

return fully conditional p-value, $P = \#\{i : S_i \geq s\}/\ell$

In the proof of Theorem 1, we will use the notation $u_n \gtrsim v_n$ to indicate that there exists some absolute constant $C > 0$ such that $u_n \geq C v_n$ for all $n = \{1, 2, \dots\}$. We use the notation $u \lesssim v$ accordingly. We will use $C > 0$ to denote a positive universal constant that might be different in each occurrence. We write $X(r) \to y$ to mean that the distribution $X(r)$ converges to the value y as $r \to \infty$.

<u>Proof of Theorem 1</u> Recall the relevant notation for computing plain p-values in Algorithm 1, along with the Common Allele data set in Table 4 and its maximum-likelihood HWE model counts (15). Here and throughout, we will refer to A_1 as the *common* allele and to $\{A_1, A_1\}$ as the common genotype; we will refer to the remaining r alleles as *rare*, to genotypes of the form $\{A_1, A_j\}$, $2 \leq j \leq r+1$, as *rare observed* genotypes, and to genotypes of the form $\{A_j, A_k\}$, $2 \leq j \leq k \leq r+1$ as *unobserved* genotypes.

1. Because the model proportion $\theta_1 = 2/3$ remains constant as r increases but the number of draws $n = 3r$ tends to infinity, the law of large numbers implies that $\Theta_1 \to \theta_1 = 2/3$. Accordingly, $M_{1,1}/n \to m_{1,1}/n = 4/9$ and $\sum_{j=2}^{r+1} \Theta_j = 1 - \Theta_1 \to 1/3$. In words,

eventually 2/3 of the simulated alleles and 4/9 of the simulated genotypes from the model will be common.

2. Similarly,

$$\sum_{k=2}^{r+1} M_{k,1}/n \;\rightarrow\; \sum_{k=2}^{r+1} m_{1,k}/n = 4/9;$$
$$\sum_{k=2}^{r+1}\sum_{j=2}^{r+1} M_{k,j}/n \;\rightarrow\; \sum_{k=2}^{r+1}\sum_{j=2}^{r+1} m_{k,j}/n = 1/9.$$

In words, roughly 4/9 of the draws simulated from the model will be *rare observed* genotypes, while 1/9 of the simulated draws will *unobserved* genotypes.

3. With probability approaching 1 as $r \to \infty$, each of the roughly $n/9 = r/3$ simulated draws from the pool of $(r^2 - r)/2$ unobserved genotypes will have a different genotype from the others. At this point, roughly $r/3$ of the unobserved simulated proportions $N_{j,k}/n$, $2 \le k \le j \le r+1$, will equal $1/(3r)$, while the others will equal 0.

4. The coupon collector's problem (see, for example, Motwani & Raghavan (1995)) implies that with probability approaching 1 as $r \to \infty$, among the roughly $2r$ simulated draws from the pool of r rare alleles, no rare allele will be drawn more than $\log(r)$ times (fixing the base of the logarithm at any real number greater than 1 that does not depend on r), and at least $3r/4$ among the r rare alleles will be drawn at least twice.

In particular, the last point above implies that, with probability approaching 1 as $r \to \infty$, all of the simulated rare proportions $\Theta_j = \Theta_j(r)$, $2 \le j \le r+1$, will satisfy

$$\Theta_j(r) \le \log(r)/r \tag{S.1}$$

and, for at least $3r/4$ among the r simulated rare proportions,

$$1/(3r) \le \Theta_j(r) \le \log(r)/r. \tag{S.2}$$

1. **The p-value for the root-mean-square goes to 0 when $r \to \infty$.** The measured sum-square discrepancy $\widetilde{f}^2 = r(r+1)f^2/2$ between the observed proportions $n_{j,k}/n$ and the model proportions $m_{j,k}/n$ is

$$
\begin{aligned}
\widetilde{f}^2 &= \left(\frac{n_{1,1}}{n} - \frac{m_{1,1}}{n}\right)^2 + \sum_{k=2}^{r+1}\left(\frac{n_{k,1}}{n} - \frac{m_{k,1}}{n}\right)^2 + \sum_{2 \le k \le j \le r+1}\left(\frac{m_{j,k}}{n}\right)^2 \\
&= \left(\frac{1}{9}\right)^2 + \frac{4}{81r} + \frac{1}{81r^3} + \frac{2(r-1)}{81r^3}.
\end{aligned}
$$

As $r \to \infty$,

$$\widetilde{f} \to 1/9. \tag{S.3}$$

If we instead consider the sum-square statistic $\widetilde{F}^2 = \frac{(r+1)(r+2)}{2} F^2$ resulting from drawing $n = 3r$ genotypes i.i.d. from the model distribution (15), points 1, 3, and 4 above give

$$\widetilde{F}^2 \lesssim \frac{(N_{1,1} - 4r/3)^2}{9r^2} + \sum_{k=2}^{r+1} \left(\frac{(\log r)^2}{r}\right)^2$$

$$+ \sum_{2 \leq k \leq j \leq r+1 : N_{j,k}=1} \left(\frac{1}{3r}\right)^2 + \sum_{2 \leq k \leq j \leq r+1 : N_{j,k}=0} \left(\frac{\log r}{r}\right)^4$$

$$\sim \frac{Z^2}{27r/4} + \frac{(\log r)^4}{r} + \frac{r}{3}\frac{1}{9r^2} + \left(\frac{r(r+1)}{2} - \frac{r}{3}\right)\left(\frac{\log r}{r}\right)^4, \qquad \text{(S.4)}$$

where $Z = (N_{1,1} - 4r/3)/\sqrt{4r/3}$ converges in distribution to a standard normal distribution as $r \to \infty$. Therefore, as $r \to \infty$, $\widetilde{F} \to 0$. Combining (S.3) and (S.4) shows that the p-value for the root-mean-square statistic, $P = \mathrm{pr}\{F \geq f\} = \mathrm{pr}\{\widetilde{F} \geq \widetilde{f}\}$, goes to 0 as $r \to \infty$.

2. **The p-value for X^2 goes to 1 as $r \to \infty$.** Similar to the measured sum-square discrepancy $\widetilde{f}$, the measured χ^2 discrepancy $\tilde{\chi}^2 = \chi^2/n$ converges to some finite positive real number as $r \to \infty$. Alternatively, if we simulate $n = 3r$ genotypes from the model distribution and (following point 3 above) consider only those roughly $r/3$ summands in the normalized χ^2 statistic $\tilde{X}^2 = X^2/n$ corresponding to the unobserved genotypes with one simulated draw,

$$\tilde{X}^2 \gtrsim \frac{r}{3} \min_{2 \leq k \leq j \leq r+1 : N_{j,k}=1} \left(\frac{N_{j,k}}{n} - \frac{M_{j,k}}{n}\right)^2 / \left(\frac{M_{j,k}}{n}\right)$$

$$\gtrsim \frac{r}{3}\left(\frac{1}{3r} - \left(\frac{\log r}{r}\right)^2\right)^2 / \left(\frac{\log r}{r}\right)^2. \qquad \text{(S.5)}$$

It follows from (S.5) that $\tilde{X}^2 \gtrsim r/\{\log(r)\}^2 \to \infty$, and so the p-value for the χ^2 statistic, $P = \mathrm{pr}(X^2 \geq \chi^2) = \mathrm{pr}(\tilde{X}^2 \geq \tilde{\chi}^2)$, goes to 1 as $r \to \infty$.

3. **The p-value for the Hellinger statistic H^2 goes to 1 when $r \to \infty$.** We have to be a bit more careful with the analysis of the Hellinger discrepancy $\tilde{h}^2 = h^2/(4n)$. The observed discrepancy is

$$\tilde{h}^2 = \frac{(\sqrt{3} - 2)^2}{9} + \sum_{j=2}^{r+1}\left(\sqrt{\frac{2}{3r}} - \sqrt{\frac{4}{9r}}\right)^2 + \sum_{2 \leq k < j \leq r+1} \frac{2}{9r^2} + \sum_{j=2}^{r+1} \frac{1}{9r^2}$$

$$= \frac{(\sqrt{3} - 2)^2}{9} + \frac{10 - 4\sqrt{6}}{9} + \frac{1}{9}$$

$$= .14.... \qquad \text{(S.6)}$$

Alternatively, suppose we simulate $n = 3r$ genotypes from the model distribution and consider r sufficiently large. Each estimated rare allele proportion will be bounded:

$\Theta_j \leq \log(r)/r$, as stated in (S.1) . Furthermore, by (S.2), at least $3/4$ of these propor-tions will satisfy $\Theta_j \geq 1/(3r)$, ensuring that at least $\frac{(3/4)^2 r^2}{2} - r$ among the $r(r+1)/2$ simulated proportions for the unobserved genotypes satisfy $M_{j,k}/n \geq 2/(9r^2)$. Then, for sufficiently large r,

$$
\begin{aligned}
\tilde{H}^2 \quad &\geq \quad \sum_{2 \leq j \leq k \leq r+1} \left(\sqrt{N_{j,k}/n} - \sqrt{M_{j,k}/n} \right)^2 \\
&\geq \quad \#\{j, k : N_{j,k} = 1\} \left(\frac{1}{\sqrt{3r}} - \frac{\log(r)}{r} \right)^2 \\
&+ \quad \left(\left(\frac{3}{4}\right)^2 \frac{r^2}{2} - r - \#\{j, k : N_{j,k} = 1\} \right) \left(\frac{2}{9r^2} \right) \\
&\sim \quad \frac{r}{3} \left(\frac{1}{\sqrt{3r}} - \frac{\log r}{r} \right)^2 + \left(\left(\frac{3}{4}\right)^2 \frac{r^2}{2} - r - \frac{r}{3} \right) \left(\frac{2}{9r^2} \right) \\
&\rightarrow \quad .17....
\end{aligned}
\tag{S.7}
$$

Combining (S.6) and (S.7), we conclude that the p-value for the Hellinger distance, $P = \mathrm{pr}(H^2 \geq h^2) = \mathrm{pr}(\tilde{H}^2 \geq \tilde{h}^2)$, goes to 1 as $r \rightarrow \infty$.