ICES REPORT 12-24

June 2012

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

by

Rachel Ward

The Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Reference: Rachel Ward, Testing Hardy-Weinberg equilibrium with a simple root-mean-square statistic, ICES REPORT 12-24, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, June 2012.

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

Rachel Ward*

October 15, 2012

Abstract

We provide evidence that a root-mean-square test of goodness-of-fit can be significantly more powerful than state-of-the-art exact 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 datasets and through asymptotic analysis. With the aid of computers, exact P-values for the root-mean-square statistic can be calculated effortlessly, and can be easily implemented using the author's freely available code.

1 Introduction

In 1908, G. H. Hardy [Har08] and W. Weinberg [Wei08] 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 [Wei05] to forensic science [Cou96] and genetic epidemiology [Sha01, KLB04]. 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 [GT92], Guo and Thompson 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 [RR95, DS98, WCA05], 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 [Eng09] or the root-mean-square.

In this article, we demonstrate using the classical datasets from Guo and Thompson [GT92] that goodness-of-fit tests based on the root-mean-square distance can be more powerful than all of the classic

*The author is with the Department of Mathematics and the Institute for Computational Engineering and Sciences at the University of Texas at Austin, 2515 Speedway, Austin, TX, 78712, rward@math.utexas.edu. She has been supported in part by a Donald D. Harrington Faculty Fellowship, Alfred P. Sloan Research Fellowship, and DOD-Navy grant N00014-12-1-0743.

tests at detecting deviations from Hardy-Weinberg equilibrium, by up to an *order of magnitude*. Our results should not be confused with good luck or anomaly. Upon further analysis of the datasets, it is revealed that 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.

To make these observations precise, we show that in the asymptotic limit where the number of draws and number of alleles tend to infinity together, the root-mean-square statistic has asymptotic power one while the classic statistics have asymptotic power zero against a family of alternatives involving one common allele, several rare alleles, and an excess of observed common genotypes. We observe numerically that this asymptotic limit is reached very quickly, on datasets involving no more than ten alleles and 30 draws. In the classical asymptotic limit, where the number of draws tends to infinity but the number of alleles remains fixed, the root-mean-square statistic is a linear combination of Gaussian random variables; as shown in [PTW11b, PTW11c, PTW11a], the root-mean-square statistic is often more powerful in this asymptotic limit as well as others.

We keep in mind that none of the statistics we consider produces a test that is uniformly more powerful than any other, but while each statistic focuses power on its own class of alternatives, we maintain that the root-mean-square statistic is more relevant for many deviations of interest in practice. 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 the recent work [PTW11a], which highlight the power of the root-mean-square statistic over classic statistics in detecting meaningful discrepancies in nonuniform distributions. The recent paper [Tyg12] provides several representative examples for which the root-mean-square test is more powerful than Fisher's exact test for homogeneity in contingency-tables. We should remark that the root-mean-square statistic is not completely unrelated to other statistics used in practice; as shown in [CTW12], the root-mean-square statistic can be interpreted as an analog to the discrete Kolmogorov-Smirnov statistic for nominal data.

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 we compare the performance of these statistics on the classic datasets from Guo and Thompson in Section 4. 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.

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 of her autosomal genes — one allele selected at random from the pair of alleles carried by her mother, and one allele selected at random from the pair of alleles carried by her father. These two alleles, considered as an unordered pair, constitute the individual's *genotype*. A gene having r alleles $\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_r$ 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) with respect to the system of alleles

$\{\mathrm{A}_1, \mathrm{A}_1\}$			
$\{\mathrm{A}_2, \mathrm{A}_1\}$	$\{\mathrm{A}_2, \mathrm{A}_2\}$		
$\vdots$	$\vdots$	$\ddots$	
$\{\mathrm{A}_\mathrm{r}, \mathrm{A}_1\}$	$\{\mathrm{A}_\mathrm{r}, \mathrm{A}_2\}$	$\dots$	$\{\mathrm{A}_\mathrm{r}, \mathrm{A}_\mathrm{r}\}$

Figure 1: Enumeration of genotypes for a gene having r alleles $\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_r$.

if the proportion of individuals in the population with two distinct alleles is twice the product of the allele proportions and the proportion of individuals in the population with two copies of the same allele is the square of that allele's frequency in the population. That is, if $p_{j,k}$ is the relative proportion of genotype $\{\mathrm{A}_j, \mathrm{A}_k\}$ in the population, and if θ_k is the proportion of allele A_k in the population, then the system is in equilibrium if

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

A large population of genotypes satisfying Hardy-Weinberg equilibrium will remain in Hardy-Weinberg equilibrium assuming random mating and no disturbing forces (e.g., no selection, no mutation, no migration, and so on). Moreover, Hardy-Weinberg equilibrium is neutral: if any assumptions are violated in a particular generation and a population is perturbed, then in the next generation the population will be in a new equilibrium (assuming assumptions are restored) specified by the new allele proportions. Hardy-Weinberg equilibrium is then a robust and reliable certificate that a population is not evolving with respect to the gene of interest, and the detection of deviations from Hardy-Weinberg equilibrium is paramount in genetic analyses.

3 Testing for deviations from Hardy-Weinberg equilibrium

In practice, one rarely has access to the genetic profile of every individual in a population, and statistical inference must be based on a random sampling of the population. If a population of genotypes $\{\mathrm{A}_j, \mathrm{A}_k\}$ with underlying genotypic proportions $p_{j,k}$ is sufficiently large, a random sample of n genotypes $X_1, X_2, \dots X_n$ from this population can be regarded as a sequence of independent and identical draws from the multinomial distribution specified by probabilities

$$\mathrm{Prob}\Big(X_i = \{\mathrm{A}_j, \mathrm{A}_k\}\Big) = p_{j,k}, \quad 1 \leqslant k \leqslant j \leqslant r. \tag{2}$$

If $n_{j,k}$ realizations of genotype $\{\mathrm{A}_j, \mathrm{A}_k\}$ 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

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

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 $\theta_1, \theta_2, \dots, \theta_{r-1}$ corresponding to the underlying allele proportions in the HWE model (1). Intuitively, the *observed* proportions of alleles, $n_1/(2n), n_2/(2n), \ldots, n_{r-1}/(2n)$, serve as our best estimates for $\theta_1, \theta_2, \ldots, \theta_{r-1}$; these parameter specifications give rise to the *model counts* of genotypes under Hardy-Weinberg equilibrium,

$$m_{j,k} = m_{j,k}(n_1, n_2, \ldots, n_r) = \begin{cases} (n_j n_k)/(2n), & j > k \\ \\ n_j^2/(4n), & j = k. \end{cases}$$

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

where δ_{jk} is the Kronecker delta function,

$$\delta_{jk} = \begin{cases} 0, & \text{if } j \neq k \\ 1, & \text{if } j = k. \end{cases}$$

It is not difficult to verify that the observed proportions of alleles are indeed the maximum-likelihood estimates for the underlying parameters $\theta_1, \theta_2, \ldots, \theta_{r-1}$ in the family of HWE equilibrium equations (1).

The observed counts of genotypes $n_{j,k}$ in a random sample from a population in Hardy-Weinberg equilibrium should not deviate too much from their model counts $m_{j,k}$. However, a systematic approach is needed to distinguish inevitable finite-population and finite-sample size fluctuations from model mismatch. Without additional prior information, a *goodness-of-fit* test serves as an omnibus litmus test to gauge consistency of the data with the Hardy-Weinberg equilibrium model. 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, and we refer the reader to the discussions in [CT99, SPW98, AB98, LNF$^+$09, LG09, CMV11]. 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

$$\chi^2 = \sum_{1 \leqslant k \leqslant j \leqslant r} \frac{\left(n_{j,k} - m_{j,k}\right)^2}{m_{j,k}}, \tag{5}$$

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

$$g^2 = 2 \sum_{1 \leqslant k \leqslant j \leqslant r} n_{j,k} \log\left(\frac{n_{j,k}}{m_{j,k}}\right), \tag{6}$$

and the Hellinger distance

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

The end result of a goodness-of-fit test is the *P-value*, the probability of observing a discrepancy between model and sample proportions of genotypes at least as extreme as the measured discrepancy, under the null hypothesis of i.i.d. draws from the model. If a goodness-of-fit test returns a sufficiently small P-value — e.g. .01 or .001 — then one can be highly confident that the model assumptions do not hold. A more *powerful* measure of discrepancy for a given data set will produce a smaller P-value if the null hypothesis is not satisfied. We remark that there are subtleties involved with the definition and interpretation of Pvalues, as discussed, for example, in Section 3 of [PTW11a].

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. We remark that one could also consider Bayesian P-values [Gel], among other formulations.

To compute the plain P-value, one repeatedly simulates n i.i.d. 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 $m_{j,k}$.

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, $N_1 = n_1, N_2 = n_2, \ldots, N_r = n_r$. In [GT92], Guo and Thompson provided an efficient means for performing such a simulation: apply a random permutation to the sequence

$$\mathcal{A} = \Big\{ \underbrace{\overbrace{\mathrm{A}_1, \mathrm{A}_1, \ldots, \mathrm{A}_1}^{n_1}, \overbrace{\mathrm{A}_2, \ldots, \mathrm{A}_2}^{n_2}, \ldots, \overbrace{\mathrm{A}_r \ldots \mathrm{A}_r}^{n_r}}_{2n} \Big\}, \tag{8}$$

and identify the pairs $\{\mathrm{A}_{2j}, \mathrm{A}_{2j+1}\}$. 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 5.1 and 5.2 of the Appendix.

Remark 3.1. It is important to note that P-values computed by repeated Monte-Carlo simulation are really *exact*: Given any specified precision ε, Hoeffding's inequality guarantees with 99.9% certainty that the Pvalue obtained using ℓ simulations will equal the P-value P obtained using infinitely many simulations to within precision ε, as long as the number of Monte Carlo simulations exceeds $\ell = 4P(1-P)/\varepsilon^2$. In all of our experiments, we used $\ell = 16,000,000$ Monte Carlo simulations so that the reported three digits of precision in our P-values are correct with 99.9% certainty.

Remark 3.2. The $r-1$ parameters in the family of HWE distributions (1) are often referred to as *nuisance* parameters. The fully conditional test we describe is a variant of the conditional test proposed by R.A. Fisher for dealing with nuisance parameters in the context of contingency table analysis [Fis25, MP83], and amounts to conditioning on a minimally sufficient statistic for estimating the nuisance parameters. Conditioning in the context of HWE testing dates back to the works of [Lev49, Hal54], but was not considered feasible for large data sets until Guo and Thompson derived the aforementioned method for efficiently simulating draws from the conditional distribution. Note that while conditioning on the counts of alleles in the observed population does effectively remove parameters from the null hypothesis, it also imposes additional assumptions on the experiment that are not necessarily reflective of reality. In small samples drawn from a large genotype population, the allele counts are not known a priori and estimates of the allele counts are subject to change upon repetition of the experiment. Unlike the fully conditional P-value, the plain P-value takes this into account; for more details, see Section 3 of [PTW11a].

3.2 The negative log-likelihood statistic

A popular alternative to testing Hardy-Weinberg equilibrium with the power divergence discrepancies χ^2, g^2, and h^2 is to use a discrepancy based directly on the likelihood function for the multinomial distribution,

$$\begin{aligned}\mathcal{L}(n_{j,k};\, n,\, m_{j,k}) &= \text{Prob}\Big(N_{1,1} = n_{1,1},\ N_{2,1} = n_{2,1},\ \dots, N_{r,r} = n_{r,r}\Big) \\ &= \frac{n!}{n_{1,1}!n_{1,2}!\dots n_{r,r}!n^n} m_{1,1}^{n_{1,1}} m_{1,2}^{n_{1,2}} \dots m_{r,r}^{n_{r,r}}. \end{aligned} \tag{9}$$

Because the likelihood function has an excessively large dynamic range, the negative of the logarithm of the likelihood is instead used as a test statistic:

$$l = -\log(\mathcal{L}). \tag{10}$$

Because the logarithm is nondecreasing over its domain, the negative likelihood function and negative log–likelihood function produce the same P-value, and this P-value can be interpreted as the probability of observing data with equal or lesser likelihood than that observed under the null hypothesis. The negative log-likelihood function (10) looks similar to the log–likelihood-ratio function 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 [GP75, RA75, Eng09].

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 [Bro65]. 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.

3.3 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,

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

The square of the root-mean-square distance is proportional to Pearson's χ^2 discrepancy when the model distribution is uniform, but takes on a very different character when the model distribution diverges from uniformity. Note that in practice, multiallelic distributions of genotypes are often very nonuniform, due to the presence of a few common alleles and several rare alleles.

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 [PTW11b, PTW11c]. This has likely contributed to its underrepresentation in the literature, as much of the classical statistical methodology was canonized before computers became readily accessible. Using the pseudocode provided in Algorithms 5.1 and 5.2, we can now obtain exact P-values for the root-mean-square statistic just as easily as we can compute exact P-values for any of the classic statistics.

4 Numerical results

We are now ready to compare the performances of the root-mean-square statistic and the classic statistics in detecting deviations from Hardy-Weinberg equilibrium. We evaluate the performance of the various statistics on three benchmark datasets from Guo and Thompson [GT92]. The three datasets, 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 $\{\mathrm{A}_j, \mathrm{A}_k\}$ 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 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. Recall that a small P-value P lets us infer, with $(100(1-P))\%$ confidence, that the draws are not i.i.d. or the draws are inconsistent with the HWE model.

The results of the analyses of Examples 1,2, and 3 — displayed in Tables 1, 2, and 3 — suggest that for both plain and fully conditional exact tests of goodness-of-fit, the root-mean-square statistic can be significantly more powerful than the classic statistics in detecting deviations.

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 datasets 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 the Appendix. 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 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 statistically significant deviation in the largest index, among others. The largest index corresponds to the 18 observed counts versus 10 expected counts of genotype $\{\mathrm{A}_3, \mathrm{A}_2\}$ 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

Table 1: P-values with 99.9% confidence intervals 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 Example 1 to be consistent with the Hardy-Weinberg equilibrium model (4).

Statistic	plain Pvalue	fully conditional Pvalue
X^2	.693 ±.001	.709 ±.001
G^2	.600 ±.001	.630 ±.001
H^2	.562 ±.001	.602 ±.001
L	.648 ±.001	.714 ±.001
F	**.039 ± .001**	**.039 ± .001**

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.

4.2 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 $\{\mathrm{A}_4, \mathrm{A}_1\}$ 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 dataset 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.

4.3 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 $.015 \pm .001$ 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 $\{\mathrm{A}_6, \mathrm{A}_6\}$. By removing this draw from the dataset 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 $.207 \pm .001$, 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 dataset in Figure 2(a) is the result of observing a single rare genotype. The root-mean-square statistic is not as sensitive to this discrepancy.

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.

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
	A_1	A_2	A_3	A_4

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

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
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9

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

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
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8

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

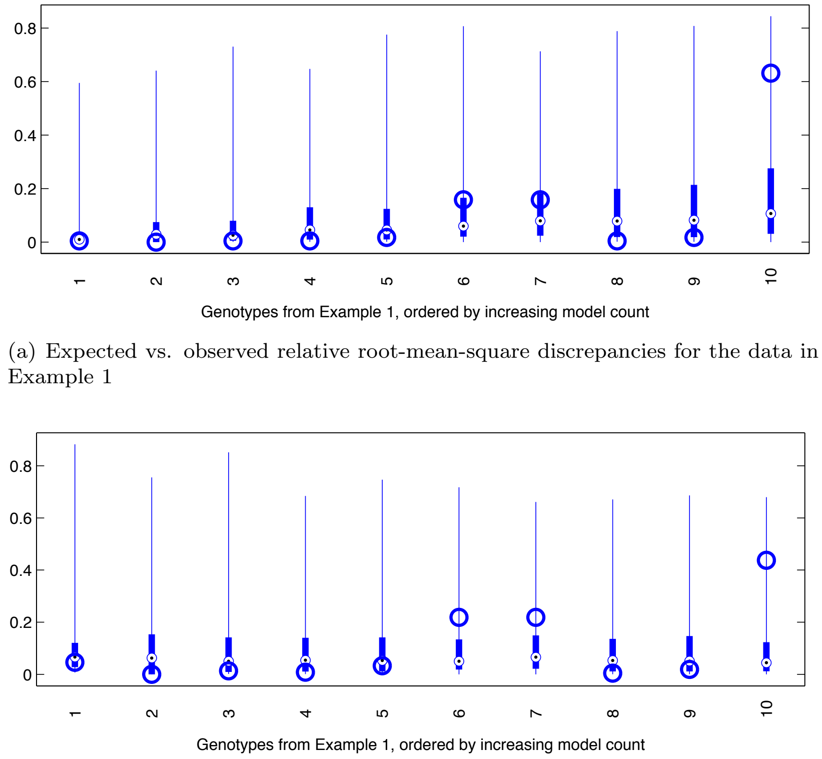
Figure 2: The three datasets from Guo and Thompson [GT92]. Observed counts are in bold and model counts are below.

Table 2: P-values with 99.9% confidence intervals 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 Example 2 to be consistent with the Hardy-Weinberg equilibrium model (4).

Statistic	plain Pvalue	fully conditional Pvalue
X^2	.020 ±.001	.020 ±.001
G^2	.013 ±.001	.013 ±.001
H^2	.027 ±.001	.025 ±.001
L	.016 ±.001	.018 ±.001
F	**.002 ± .001**	**.002 ± .001**

Table 3: P-values with 99.9% confidence intervals 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 Example 3 to be consistent with the Hardy-Weinberg equilibrium model (4).

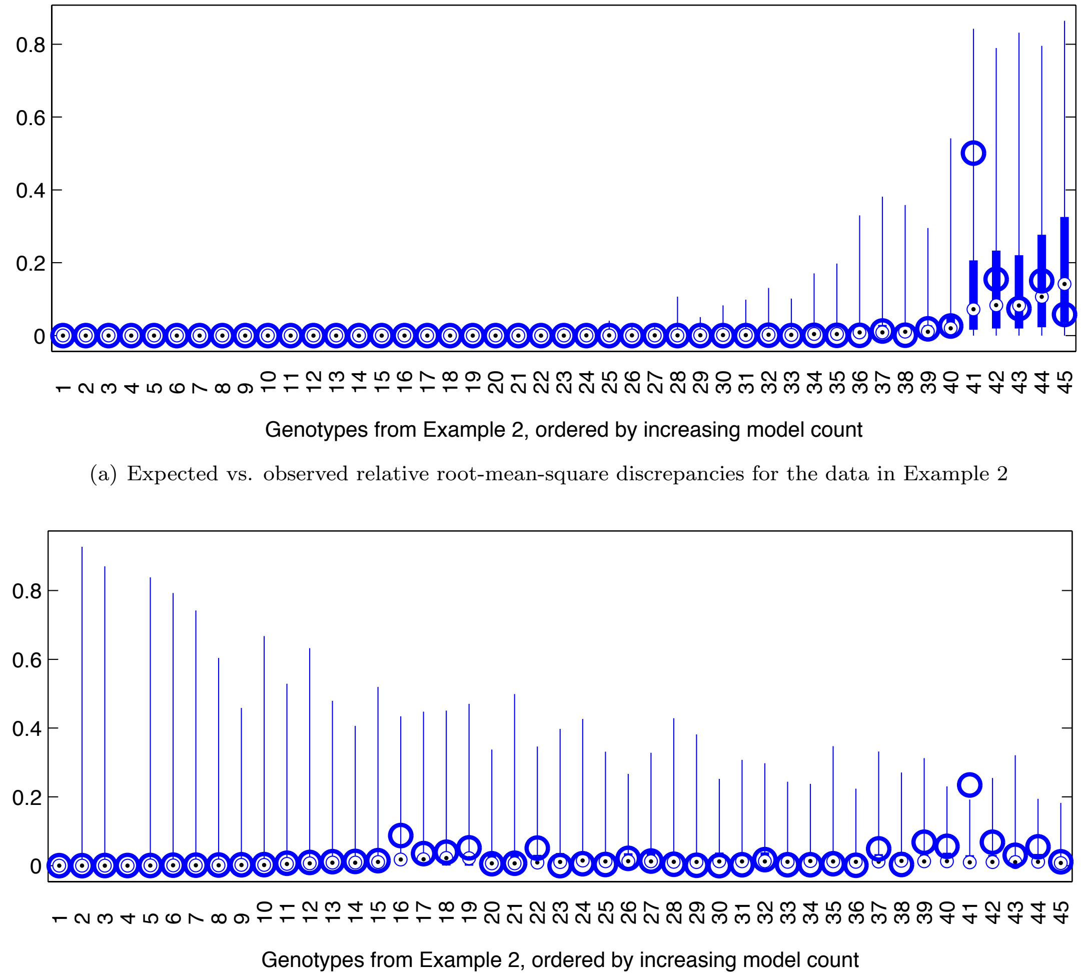
Statistic	plain Pvalue	fully conditional Pvalue
X^2	.015 ± .001	.026 ± .001
G^2	.181 ±.001	.276 ±.001
H^2	.307 ±.001	.449 ±.001
L	.155 ±.001	.207 ±.001
F	.885 ± .001	.917 ± .001



(a) Expected vs. observed relative root-mean-square discrepancies for the data in Example 1

(b) Expected vs. observed relative χ^2 discrepancies for the data in Example 1

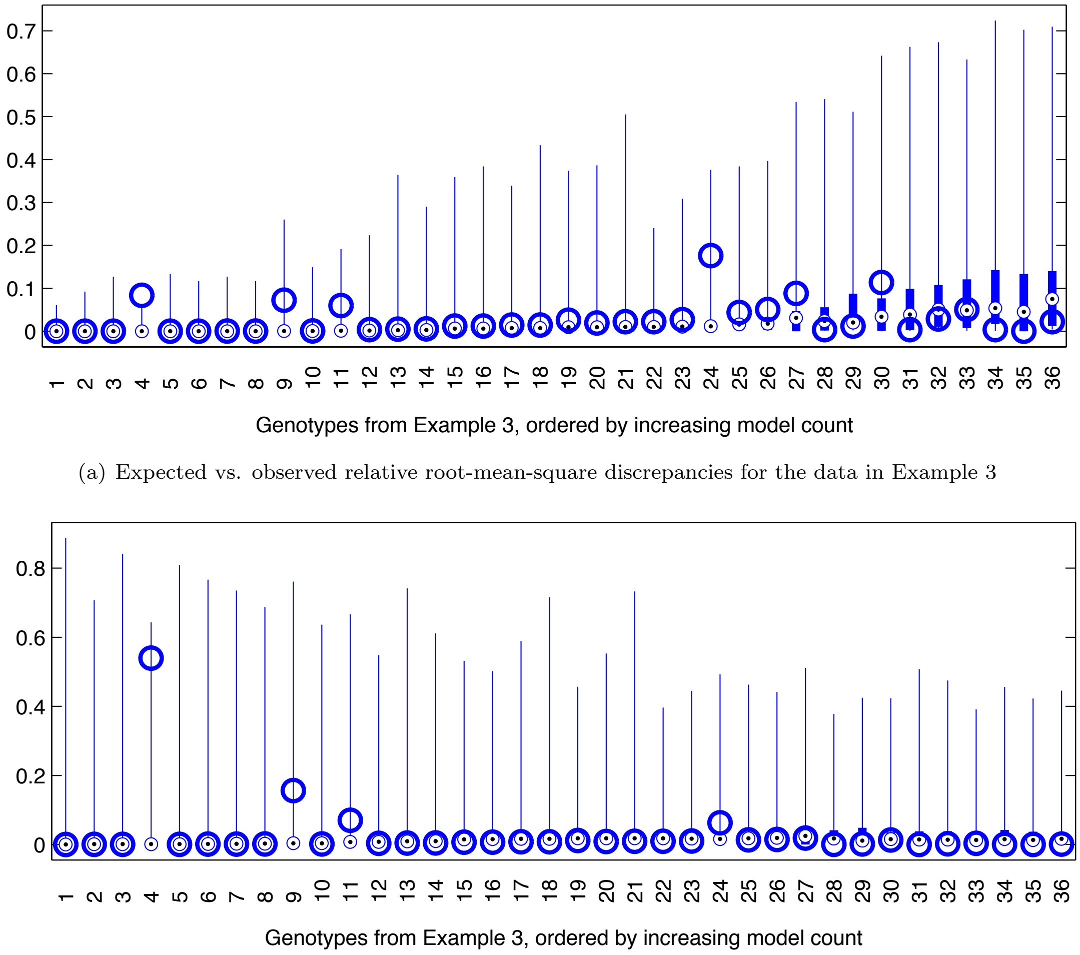
Figure 3



(a) Expected vs. observed relative root-mean-square discrepancies for the data in Example 2

(b) Expected vs. observed relative χ^2 discrepancies for the data in Example 2

Figure 4



(a) Expected vs. observed relative root-mean-square discrepancies for the data in Example 3

(b) Expected vs. observed relative χ^2 discrepancies for the data in Example 3

Figure 5

We will show for a representative family of datasets that the root-mean-square statistic has asymptotic power one while the chi-square statistics have asymptotic power zero. 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*. Note that the asymptotic chi-square approximation to the classic statistics is not valid in this limit.

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

Table 4: Common Allele dataset

$n = 3r$ observed genotypes		
$n_{1,1} = r$ of type $\{\mathrm{A}_1, \mathrm{A}_1\}$,	$n_{1,1}/n = 1/3$	
$n_{1,k} = 2$ of type $\{\mathrm{A}_1, \mathrm{A}_k\}$,	$n_{1,k}/n = 2/(3r)$,	$2 \leqslant k \leqslant r + 1$
$n_{j,k} = 0$ of type $\{\mathrm{A}_j, \mathrm{A}_k\}$,	$n_{j,k}/n = 0$,	$2 \leqslant j \leqslant k \leqslant r + 1$
$n_1 = 4r$ alleles of type A_1,	$n_1/(2n) = 2/3$	
$n_k = 2$ alleles of type A_k,	$n_k/(2n) = 1/(3r)$,	$2 \leqslant k \leqslant r + 1$.

The maximum-likelihood model counts for the Common Allele dataset are

$$\begin{cases} m_{1,1} = 4r/3, & \\ m_{1,k} = 4/3, & 2 \leqslant k \leqslant r + 1, \\ m_{k,k} = 1/(3r), & 2 \leqslant k \leqslant r + 1, \\ m_{j,k} = 2/(3r), & 2 \leqslant j < k \leqslant r + 1, \quad j < k. \end{cases} \tag{12}$$

To see that the Common Allele dataset 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 dataset 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, H^2, and L *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 5.1. *In the limit as $r \to \infty$, the plain P-values (as computed via Algorithm 5.1) given by X^2, the log–likelihood-ratio statistic G^2, the Hellinger distance H^2, and the negative log–likelihood statistic L for the Common Allele dataset 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 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.

In the proof of Theorem 5.1, we will use the notation $u_n \gtrsim v_n$ to indicate that there exists some absolute constant $C > 0$ such that $u_n \geqslant C v_n$ 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 $r \to \infty$.

Proof of Theorem 5.1. Recall the relevant notation for computing plain P-values in Algorithm 5.1, along

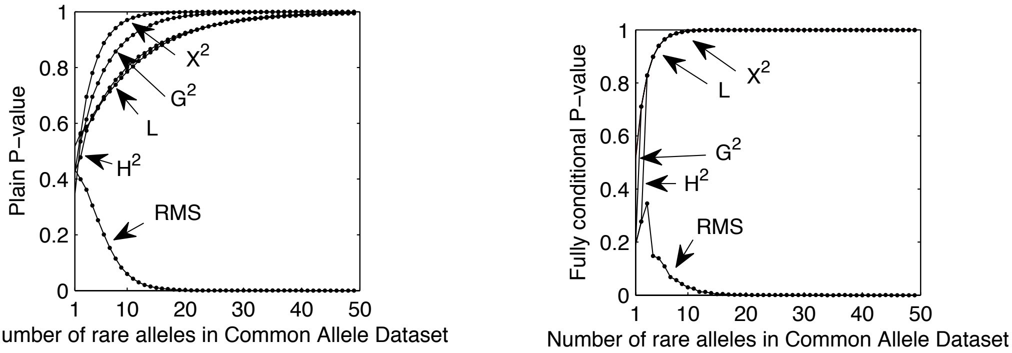


Figure 6: P-values (accurate to three digits with 99% confidence) for Pearson's statistic X^2, the log–likelihood-ratio statistic G^2, the Hellinger statistic H^2, the negative log–likelihood statistic L, and the root-mean-square statistic F, for the observed genotypic counts in the Common Allele dataset to be consistent with the Hardy-Weinberg equilibrium model (4), as a function of the number of alleles r.

with the Common Allele dataset in Table 4 and its maximum-likelihood HWE model counts (12). 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 $\{A_1, A_j\}$, $2 \leqslant j \leqslant r+1$, as *rare observed* genotypes, and to genotypes of the form $\{A_j, A_k\}$, $2 \leqslant j \leqslant k \leqslant r+1$ 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, $\sum_{k=2}^{r+1} M_{k,1}/n \to \sum_{k=2}^{r+1} m_{1,k}/n = 4/9$ and $\sum_{k=2}^{r+1} \sum_{j=2}^{r+1} M_{k,j}/n \to \sum_{k=2}^{r+1} \sum_{j=2}^{r+1} m_{k,j}/n = 1/9$. 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 $N_{j,k}/n$, $2 \leqslant k \leqslant j \leqslant r+1$, will equal $1/(3r)$, while the others will equal 0.
4. The coupon collector's problem (see, for example, [MR95]) 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 $\Theta_j = \Theta_j(r)$, $2 \leqslant j \leqslant r+1$, will satisfy

$$\Theta_j(r) \leqslant \log(r)/r \tag{13}$$

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

$$1/(3r) \leqslant \Theta_j(r) \leqslant \log(r)/r. \tag{14}$$

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

$$\begin{aligned} \widetilde{f}^2 &= \Big(\frac{n_{1,1}}{n} - \frac{m_{1,1}}{n}\Big)^2 + \sum_{k=2}^{r+1} \Big(\frac{n_{k,1}}{n} - \frac{m_{k,1}}{n}\Big)^2 + \sum_{2 \leqslant k \leqslant j \leqslant r+1} \Big(\frac{m_{j,k}}{n}\Big)^2 \\ &= \Big(\frac{1}{9}\Big)^2 + \frac{4}{81r} + \frac{1}{81r^3} + \frac{2(r-1)}{81r^3}. \end{aligned} \tag{15}$$

As $r \to \infty$,

$$\widetilde{f} \to \frac{1}{9}. \tag{16}$$

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 (12), points 1, 3, and 4 above give

$$\begin{aligned} \widetilde{F}^2 &\lesssim \frac{(N_{1,1} - 4r/3)^2}{9r^2} + \sum_{k=2}^{r+1} \Big(\frac{(\log r)^2}{r}\Big)^2 \\ &+ \sum_{2 \leqslant k \leqslant j \leqslant r+1 : N_{j,k} = 1} \Big(\frac{1}{3r}\Big)^2 + \sum_{2 \leqslant k \leqslant j \leqslant r+1 : N_{j,k} = 0} \Big(\frac{\log r}{r}\Big)^4 \\ &\sim \frac{Z^2}{27r/4} + \frac{(\log r)^4}{r} + \frac{r}{3}\frac{1}{9r^2} + \Big(\frac{r(r+1)}{2} - \frac{r}{3}\Big)\Big(\frac{\log r}{r}\Big)^4, \end{aligned} \tag{17}$$

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 (16) and (17) shows that the P-value for the root-mean-square statistic, $P = \text{Prob}\{F \geqslant f\} = \text{Prob}\{\widetilde{F} \geqslant \widetilde{f}\}$, 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,

$$\begin{aligned} \tilde{X}^2 &\gtrsim \frac{r}{3} \min_{2 \leqslant k \leqslant j \leqslant r+1 : N_{j,k} = 1} \Big(\frac{N_{j,k}}{n} - \frac{M_{j,k}}{n}\Big)^2 / \Big(\frac{M_{j,k}}{n}\Big) \\ &\gtrsim \frac{r}{3}\Big(\frac{1}{3r} - \Big(\frac{\log r}{r}\Big)^2\Big)^2 / \Big(\frac{\log r}{r}\Big)^2. \end{aligned} \tag{18}$$

It follows that $\tilde{X}^2 \gtrsim \frac{r}{(\log r)^2} \to \infty$, and so the P-value for the χ^2 statistic, $P = \text{Prob}(X^2 \geqslant \chi^2) = \text{Prob}(\tilde{X}^2 \geqslant \tilde{\chi^2}^2)$, goes to 1 as $r \to \infty$.

3. **The P-values for the log–likelihood-ratio G^2 and negative log–likelihood L go to 1 when $r \to \infty$** by an argument analogous to that used for the χ^2 P-value.

4. **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

$$\begin{aligned}
\tilde{h}^2 &= \frac{(\sqrt{3}-2)^2}{9} + \sum_{j=2}^{r+1} \Big(\sqrt{\frac{2}{3r}} - \sqrt{\frac{4}{9r}}\Big)^2 + \sum_{2 \leqslant k < j \leqslant r+1} \frac{2}{9r^2} + \sum_{j=2}^{r+1} \frac{1}{9r^2} \\
&= \frac{(\sqrt{3}-2)^2}{9} + \frac{10-4\sqrt{6}}{9} + \frac{1}{9} \\
&= .14.... \qquad (19)
\end{aligned}$$

Alternatively, suppose we simulate $n = 3r$ genotypes from the model distribution and consider r sufficiently large. Each estimated rare allele proportion will be bounded: $\Theta_j \leqslant \log(r)/r$, as stated in (13). Furthermore, by (14), at least 3/4 of these proportions will satisfy $\Theta_j \geqslant 1/(3r)$, 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 $M_{j,k}/n \geqslant 2/(9r^2)$. Then, for sufficiently large r,

$$\begin{aligned}
\tilde{H}^2 &\geqslant \sum_{2 \leqslant j \leqslant k \leqslant r+1} \Big(\sqrt{N_{j,k}/n} - \sqrt{M_{j,k}/n}\Big)^2 \\
&\geqslant \#\{j,k : N_{j,k} = 1\}\Big(\frac{1}{\sqrt{3r}} - \frac{\log{(r)}}{r}\Big)^2 \\
&+ \Big(\Big(\frac{3}{4}\Big)^2 \frac{r^2}{2} - r - \#\{j,k : N_{j,k} = 1\}\Big)\Big(\frac{2}{9r^2}\Big) \\
&\sim \frac{r}{3}\Big(\frac{1}{\sqrt{3r}} - \frac{\log r}{r}\Big)^2 + \Big(\Big(\frac{3}{4}\Big)^2 \frac{r^2}{2} - r - \frac{r}{3}\Big)\Big(\frac{2}{9r^2}\Big) \\
&\to .17.... \qquad (20)
\end{aligned}$$

Combining (19) and (20), we conclude that the P-value for the Hellinger distance, $P = \text{Prob}(H^2 \geqslant h^2) = \text{Prob}(\tilde{H}^2 \geqslant \tilde{h}^2)$, goes to 1 as $r \to \infty$.

□

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 5.1 applies directly only to the plain P-values.

To conclude this section, we remark that the particular distribution of the draws in the Common Allele dataset was rather arbitrary, and that a similar asymptotic analysis holds for many other datasets. For example, we could have considered instead a dataset involving two, three, or four common alleles, or one common allele and three fairly-common alleles, and so on.

Acknowledgment

The author would like to thank Mark Tygert, Andrew Gelman, Abhinav Nellore, and Will Perkins for their helpful contributions.

Software availability

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

References

[AB98] K. Ayres and D. Balding, *Measuring departures from Hardy-Weinberg: a Markov chain Monce Carlo method for estimating the inbreeding coefficient*, Heredity **80** (1998), 769–777.

[Bro65] K. Brownlee, *Statistical series and methodology in science and engineering*, Wiley, Inc., New York, 1965.

[CMV11] G. Consonni, E. Moreno, and S. Venturini, *Testing Hardy-Weinberg equilibrium: an objective Bayesian analysis*, Statistics in Medicine **30** (2011), 62–74.

[Cou96] National Research Council, *The evaluation of forensic dna evidence*, National Academy Press, 1996.

[CT99] J. Chen and G. Thomson, *The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test*, Biometrics **55** (1999), 1269–1272.

[CTW12] J. Carruth, M. Tygert, and R. Ward, *The discrete Kolmogorov-Smirnov versus the Euclidean distance in testing goodness-of-fit*, In preparation (2012).

[DS98] P. Diaconis and B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Annals of Statistics **26** (1998), no. 1, 363–397.

[Eng09] W. Engels, *Exact tests for Hardy-Weinberg proportions*, Genetics **183** (2009), no. 4, 1431–1441.

[Fis25] R. Fisher, *Statistical methods for research workers*, Oliver and Boyd, Edinburgh, 1925.

[Gel] A. Gelman, *A Bayesian formulation of exploratory data analysis and goodness-of-fit testing*, International Statistical Review **71**, 369–382.

[GP75] J. Gibbons and J. Pratt, *P-values: Interpretation and methodology*, The American Statistician **29** (1975), no. 1, 20–25.

[GT92] S. Guo and E. Thompson, *Performing the exact test of Hardy-Weinberg proportion for multiple alleles*, Biometrics **48** (1992), 361–372.

[Hal54] J. Haldane, *An exact test for randomness of mating*, Journal of Genetics **52** (1954), 631–635.

[Har08] G. Hardy, *Mendelian proportions in a mixed population*, Science **28** (1908), 49–50.

[KLB04] M. Khoury, J. Little, and W. Burke, *Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease*, Oxford University Press, New York, 2004.

[Lev49] H. Levene, *On a matching problem arising in genetics*, Annals of Mathematical Statistics **20** (1949), 91–94.

[LG09] Y. Li and B. Graubard, *Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data*, Biometrics **65** (2009), 1096–1104.

[LNF$^+$09] M. Lauretto, F. Nakano, S. Faria, C. Pereira, and J. Stern, *A straightforward multiallelic significance test for the Hardy-Weinberg equilibrium law*, Genetics and Molecular Biology **32** (2009), no. 3, 619–625.

[MP83] C. Mehta and N. Patel, *A network algorithm for performing Fusher's exact test in $r \times c$ contingency tables*, Journal of the American Statistical Association (1983), 427–434.

[MR95] R. Motwani and P. Raghavan, *Randomized algorithms*, Cambridge University Press, New York, NY, 1995.

[PTW11a] W. Perkins, M. Tygert, and R. Ward, χ^2 *and classical exact tests often wildly misreport significance; the remedy lies in computers*, arXiv:1201.1431 (2011).

[PTW11b] ———, *Computing the confidence levels for a root-mean-square test of goodness of fit*, Appl. Math. Comput. **217** (2011), 90729084.

[PTW11c] ———, *Computing the confidence levels for a root-mean-square test of goodness of fit, II*, arXiv:1009.2260 (2011).

[RA75] R. Radlow and E. Alf, *An alternate multinomial assessment of the accuracy of the* χ^2 *test of goodness of fit*, Journal of the American Statistical Association **70** (1975), no. 352, 811–813.

[RR95] M. Raymond and F. Rousset, *An exact test for population differentiation*, Evolution **49** (1995), no. 6, 1280–1283.

[Sha01] P. Sham, *Statistics in human genetics*, Arnold Publishers, London, 2001.

[SPW98] J. Shoemaker, I. Painter, and B. Weir, *A Bayesian characterization of Hardy-Weinberg disequilibrium*, Genetics **149** (1998), 2079–2088.

[Tyg12] M. Tygert, *Testing the significance of assuming homogeneity in contingency-tables/cross-tabulations*, arXiv:1201.1421 (2012).

[WCA05] J. Wigginton, D. Cutler, and G. Abecasis, *A note on exact tests of Hardy-Weinberg equilibrium*, American Journal of Human Genetics (2005), 887–893.

[Wei08] W. Weinberg, *Über den nachweis der vererbung beim menschen*, Jh. Ver. vaterl. Naturk. Wurttemb. **64** (1908), 369–382, (English translations in BOYER 1963 and JAMESON 1977).

[Wei05] K. Weising, *DNA fingerprinting in plants: Principles, methods, and applications*, Taylor and Francis, Boca Raton, Florida, 2005.

Appendix I: Pseudocode for calculating exact P-values

Algorithm 5.1: Computing the plain P-value

```
Input: Observed genotype counts $n_{j,k}$, number of Monte Carlo simulations $\ell$, 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 $N_j^{(i)} = \big(\sum_{k=j}^{r} N_{k,j}^{(i)} + \sum_{k=1}^{j} N_{j,k}^{(i)}\big)$ 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, $P = \#\{i : S_i \geqslant s\}/\ell$
```

Appendix II: 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, $1 \leqslant j \leqslant r(r+1)/2$, 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

$$d_j^{rms} = \big(m_{[j]} - n_{[j]}\big)^2, \tag{21}$$

while the observed chi-square discrepancies are

$$d_j^{chi} = \frac{\big(m_{[j]} - n_{[j]}\big)^2}{m_{[j]}}. \tag{22}$$

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

$$D_j^{rms} = \big(M_{[j]} - N_{[j]}\big)^2, \tag{23}$$

Algorithm 5.2: Computing the fully conditional P-value

```
Input: Observed genotype counts $n_{j,k}$ and allele counts $n_j$, number of Monte Carlo simulations $\ell$, 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 (8) 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, $P = \#\{i : S_i \geqslant s\}/\ell$
```

and

$$D_j^{chi} = \frac{\left(M_{[j]} - N_{[j]}\right)^2}{M_{[j]}}. \tag{24}$$

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 (23) 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.