Biostatistics (2014), **15**, 1, *pp.* 74–86
doi:10.1093/biostatistics/kxt028
Advance Access publication on August 22, 2013

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

RACHEL WARD*

Department of Mathematics, RLM 10.144, University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA

rward@math.utexas.edu

RAYMOND J. CARROLL

Department of Statistics, 3143 TAMU, Texas A&M University, College Station, TX 77843, USA

Summary

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 χ^2 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.

Keywords: Absolute discrepancies; Hardy–Weinberg equilibrium; Relative discrepancies; Root mean square.

1. Introduction

Hardy (1908) and 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). Statistical tests of deviation from HWE are fundamental for validating such assumptions. Traditionally, Pearson's χ^2 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 χ^2 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 and Thompson (1992) campaigned for an exact test of HWE based on the likelihood function. While their work renewed interest in conditional exact tests for HWE (Raymond and Rousset, 1995; Maiste and Weir, 1995; Diaconis and Sturmfels, 1998; Wigginton *and others*, 2005), likelihood-based

*To whom correspondence should be addressed.

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 and 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 HWE. 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 *and others*, 2013), which highlight the power of the root-mean-square statistic over classic statistics in detecting meaningful discrepancies in non-uniform 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 HWE. 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 with the root-mean-square statistic in distinguishing important classes of deviations from HWE, and end with concluding remarks in Section 6. Supplementary material available at *Biostatistics* online includes pseudocode for algorithms and proofs of technical results.

2. HWE: 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 the lower-triangular array of indices (j, k) satisfying $j \geqslant k$.

A population is said to be in HWE if the following holds. 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

$$p_{j,k} = p_{j,k}(\theta_j, \theta_k) = \begin{cases} 2\theta_j\theta_k, & j > k, \\ \theta_k^2, & j = k. \end{cases} \tag{2.1}$$

3. TESTING HWE

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 $\mathrm{pr}(X_i = \{A_j, A_k\}) = p_{j,k}$, $1 \leqslant k \leqslant j \leqslant r$. 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 $n_j = \sum_{k=j}^{r} n_{k,j} + \sum_{k=1}^{j} n_{j,k}$, $j = 1, \ldots, r$. In order to gauge the consistency of the sample counts $(n_{j,k})$ with HWE, 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 (2.1). 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 (2.1); these parameter specifications give rise to the *model counts* of genotypes under HWE, $m_{jk} = (2 - \delta_{jk})(n_j\, n_k)/(4n)$, 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 the 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 non-parametric 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 (Lindley, 1964; Chen and Thomson, 1999; Shoemaker *and others*, 1998; Ayres and Balding, 1998; Lauretto *and others*, 2009; Li and Graubard, 2009; Wakefield, 2010; Consonni *and others*, 2011). In this paper, we will focus only on non-parametric (or nearly non-parametric) 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 the 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

$$X^2 = \sum_{1 \leqslant k \leqslant j \leqslant r} (n_{j,k} - m_{j,k})^2 / m_{j,k}, \tag{3.1}$$

the log-likelihood ratio or G^2 divergence, $G^2 = 2\sum_{1 \leqslant k \leqslant j \leqslant r} n_{j,k} \log(n_{j,k}/m_{j,k})$, and the Hellinger distance $H^2 = 4\sum_{1 \leqslant k \leqslant j \leqslant r} (\sqrt{n_{j,k}} - \sqrt{m_{j,k}})^2$. Another classic measure of discrepancy is the negative log-likelihood function, which is based directly on the likelihood function for the multinomial distribution, $L = -\log(\mathcal{L})$, where $\mathcal{L}(n_{j,k}; n, m_{j,k}) = (n!/n_{1,1}!n_{1,2}! \cdots n_{r,r}!n^n) m_{1,1}^{n_{1,1}} m_{1,2}^{n_{1,2}} \cdots m_{r,r}^{n_{r,r}}$. The negative log-likelihood statistic 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 and Pratt, 1975; Radlow and 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 χ^2 random variable with $r(r-1)/2$ degrees of freedom as the number of draws n goes to infinity and the 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* (FC) 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 $N_j^{(i)} = (\sum_{k=j}^{r} N_{k,j}^{(i)} + \sum_{k=1}^{j} N_{j,k}^{(i)})$, 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 and Tibshirani (1993), Henze (1996), and Bickel *and others* (2006).

The FC p-value corresponds to imposing additional restrictions on the probability space associated to the null hypothesis. To compute the FC 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 and Thompson (1992) provided an efficient means for performing such a simulation: apply a random permutation to the sequence

$$\mathcal{A} = \left\{ \overbrace{A_1, A_1, \ldots, A_1}^{n_1}, \overbrace{A_2, \ldots, A_2}^{n_2}, \ldots, \overbrace{A_r, \ldots, A_r}^{n_r} \right\}_{2n}, \tag{3.2}$$

and identify the pairs $\{A_{2j}, A_{2j+1}\}$. The FC 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 FC p-values is provided in Algorithms 1 and 2 of Appendix S.1 in supplementary material available at *Biostatistics* online.

3.2 *The root-mean-square statistic*

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

$$F = \left\{ \frac{2}{n^2 r(r+1)} \sum_{1 \leqslant k \leqslant j \leqslant r} (n_{j,k} - m_{j,k})^2 \right\}^{1/2}. \tag{3.3}$$

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 *and others* (2011, 2012). Using the pseudocode provided in Algorithms 1 and 2 of Appendix S.1 in supplementary material available at *Biostatistics* online, 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 HWE. We first evaluate the performance of the various statistics on three benchmark

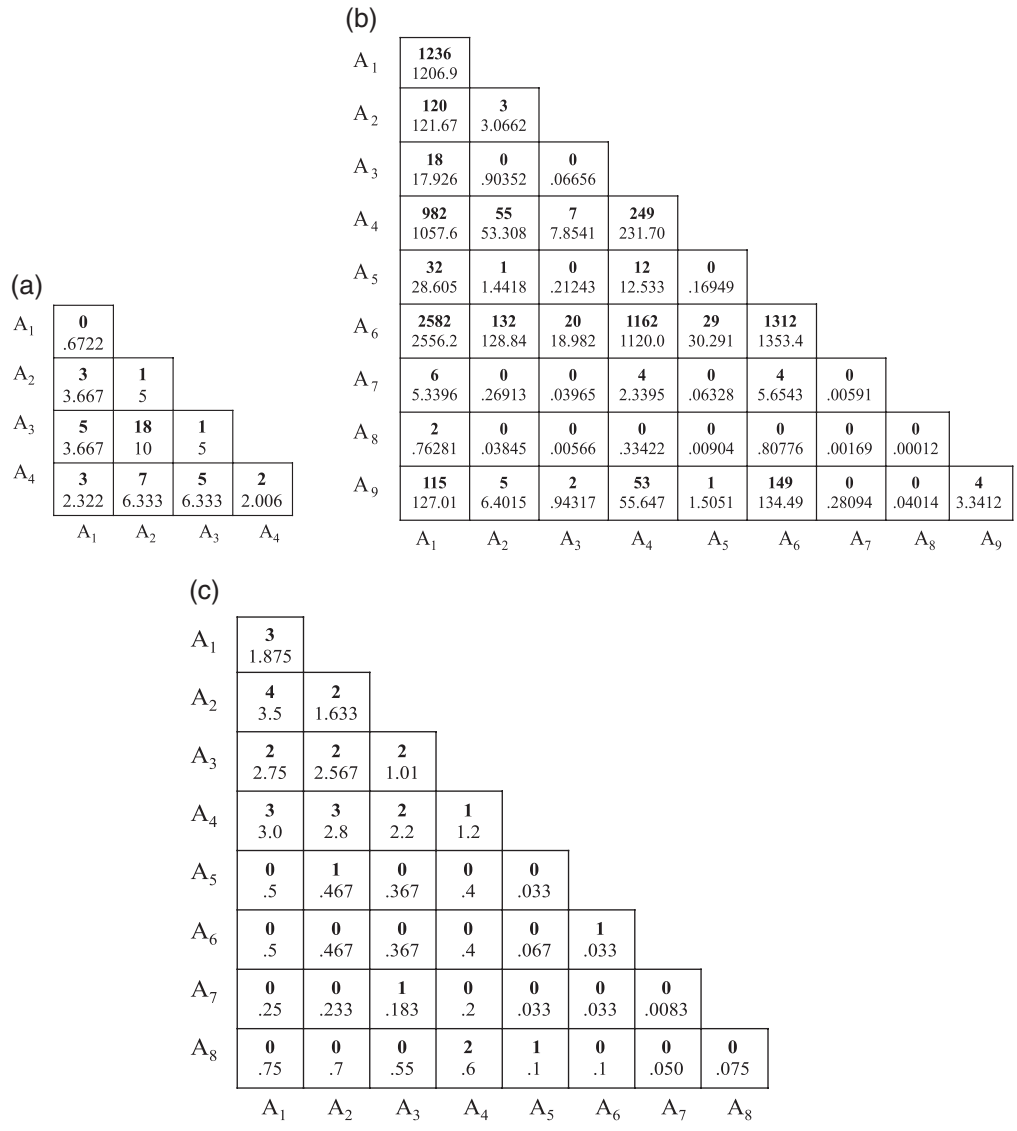

Fig. 1. The three data sets from Guo and Thompson (1992). Observed counts are in bold and expected counts under HWE are below. (a) Example 1: $n = 45$, (b) Example 2: $n = 8297$, and (c) Example 3: $n = 30$.

data sets from Guo and Thompson (1992). The three data sets, which we refer to as Examples 1–3, are represented in Figure 1 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 FC 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.

Table 1. *Plain and 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 HWE model (2.1)*

	Example 1		Example 2		Example 3	
Statistic	Plain p-value	FC p-value	Plain p-value	FC p-value	Plain p-value	FC p-value
X^2	0.693	0.709	0.020	0.020	0.015	0.026
G^2	0.600	0.630	0.013	0.013	0.181	0.276
H^2	0.562	0.602	0.027	0.025	0.307	0.449
L	0.648	0.714	0.016	0.018	0.155	0.207
F	0.039	0.039	0.002	0.002	0.885	0.917

With 99% confidence, p-values are correct to ± 0.001.

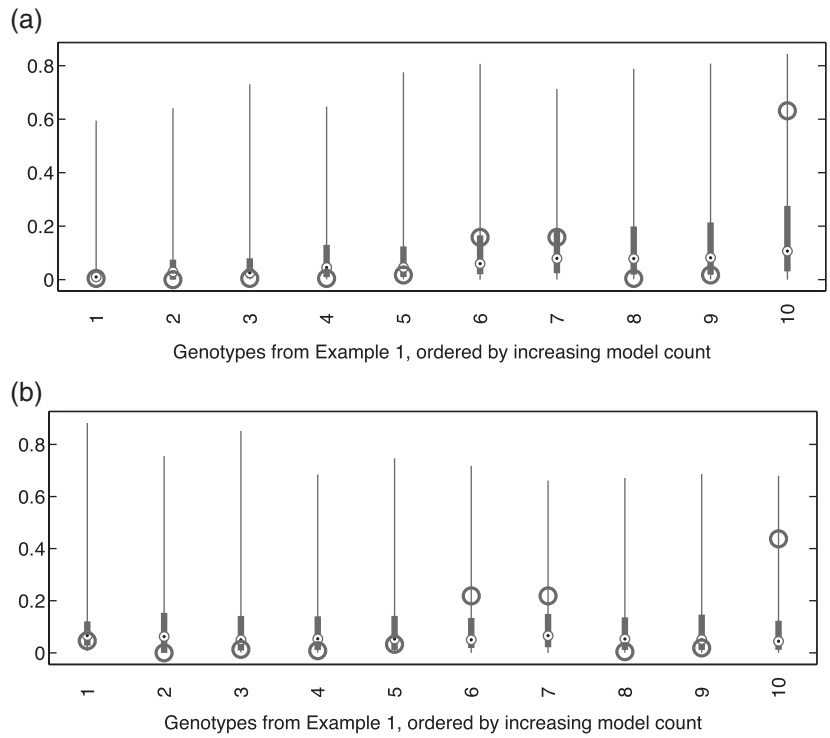

Fig. 2. (a) Expected vs. observed root-mean-square discrepancies and (b) expected vs. observed χ^2 discrepancies. By root-mean-square and χ^2 discrepancies, we mean the terms within the summations in formulas (3.3) and (3.1), normalized to sum to 1.

4.1.1 *Graphical views of the data.* Figures 2–4 contain boxplots displaying the median, upper and lower quartiles, and whiskers reaching from the 1st to 99th percentiles for root-mean-square discrepancies and χ^2 discrepancies simulated under the plain HWE null hypothesis for the data sets from Examples 1–3. The boxplots are for simulated data, whereas the large open circles indicate the observed data. In the χ^2 boxplots, we see the division by expected proportion in the summands of the χ^2 discrepancy (3.1) 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 (3.3) reflected in the larger contribution of

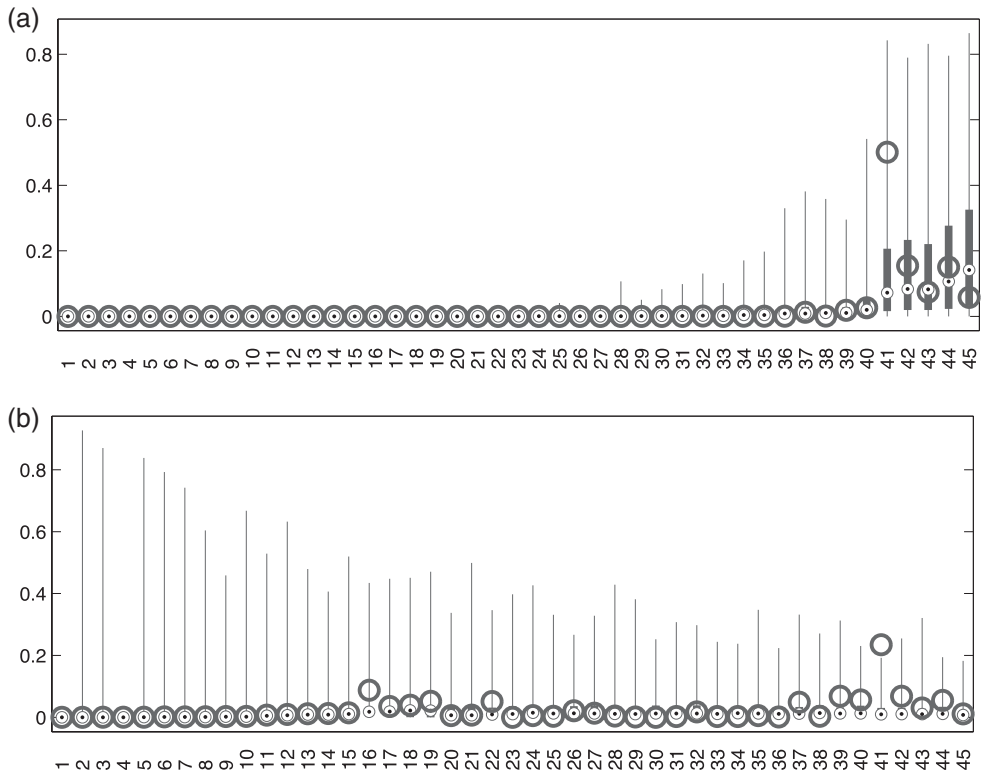

Fig. 3. (a) Expected vs. observed root-mean-square discrepancies and (b) expected vs. observed χ^2 discrepancies. By root-mean-square and χ^2 discrepancies, we mean the terms within the summations in formulas (3.3) and (3.1), normalized to sum to 1.

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 χ^2 statistic, are sensitive to relative rather than absolute discrepancies.

4.1.2 *Interpretation of the results for Example* 1. Comparing the boxplots in Figure 2, we see that both χ^2 and root-mean-square tests report a significant deviation in the largest index, among others. The largest index corresponds to the 18 observed counts vs. 10 expected counts of genotype $\{A_3, A_2\}$ in Example 1. However, as reported in Table 1, the p-value of 0.039 given by the root-mean-square test is an order of magnitude smaller than the p-value of 0.693 reported by the χ^2 test, as this discrepancy is larger compared with expected root-mean-square fluctuations than it is compared with expected χ^2 fluctuations. As indicated by the boxplots in Figure 2, the statistical significance of the deviation in index 10 (as well as the deviations in indices 6 and 7) is masked by large expected relative deviations in the rare genotypes in the χ^2 summation.

4.1.3 *Interpretation of the results for Example* 2. The distribution of discrepancies in Figure 3 can be interpreted similarly to the boxplots from Figure 2. In contrast to the $n = 45$ draws from Example 1, however, this data set contains $n = 8297$ draws; we infer that the qualitative differences between the root-mean-square and χ^2 statistic are not unique to small sample-size data.

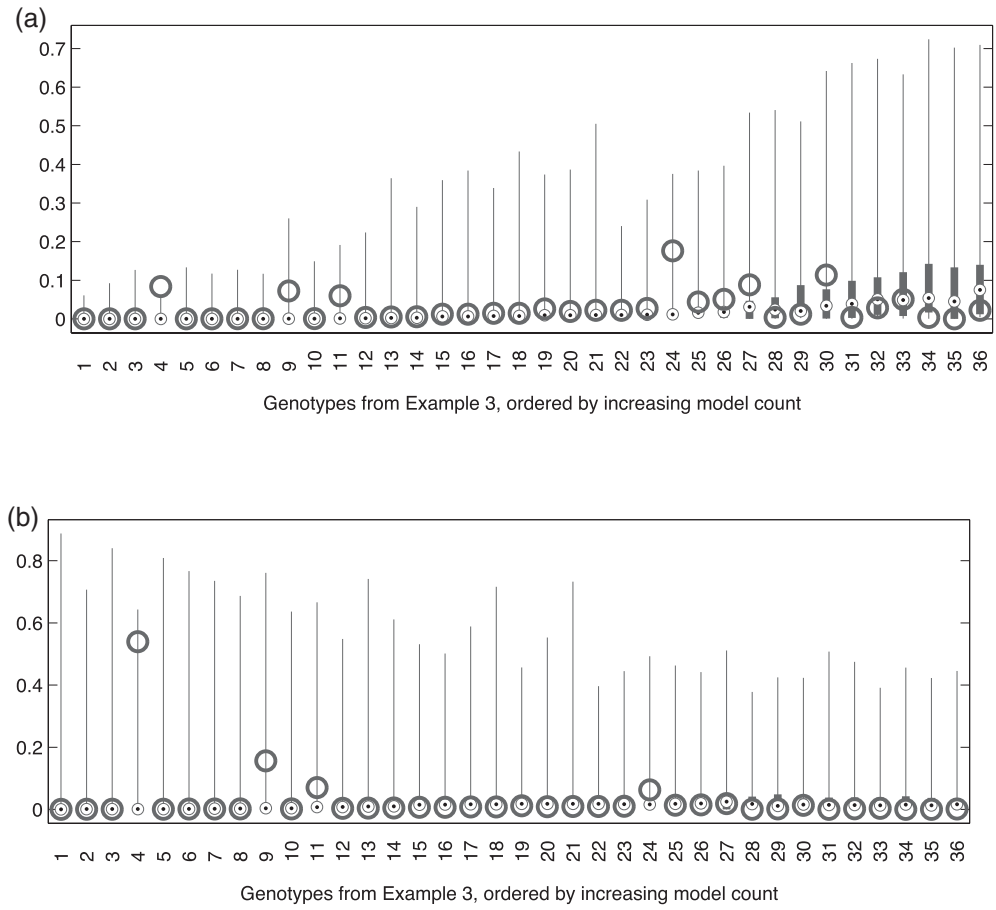

Fig. 4. (a) Expected vs. observed root-mean-square discrepancies and (b) expected vs. observed χ^2 discrepancies. By root-mean-square and χ^2 discrepancies, we mean the terms within the summations in formulas (3.3) and (3.1), normalized to sum to 1.

4.1.4 *Interpretation of the results for Example* 3. Comparing the expected and observed χ^2 discrepancies in Figure 4(b), we might posit that the small p-value of 0.015 that the χ^2 test gives to the data in Example 3 depends strongly on the discrepancy at the fourth 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 χ^2 goodness-of-fit test on the remaining $n = 29$ draws, the χ^2 statistic X^2 returns a p-value of 0.207, well over an order of magnitude larger than the previous p-value, confirming that the small p-value given by the χ^2 statistic for the data set in Figure 4 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 the statistics X^2, G^2, H^2, L, and 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 and Thomson, 1999; Ayres and Balding, 1998; Lauretto *and others*,

Table 2. *Statistical power and Type* I *error of the various tests of HWE against deviations due to* selection*, i.e. deviations of the form (4.1) with parameters as specified in Alternatives* 1–4 *and fitness parameters (4.2) and deviations due to* inbreeding*, i.e. deviations of the form (4.3) with parameters as specified in Alternatives* 1–4 *and inbreeding parameter* $f = \frac{1}{10}$

	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
Statistic	Power	Type I	Power	Type I	Power	Type I	Power	Type I
Deviations due to selection for the common allele—plain p-value test								
X^2	0.06	0.06	0.04	0.05	0.04	0.04	<0.01	0.06
G^2	0.07	0.08	0.07	0.06	0.07	0.06	0.01	0.08
H^2	0.07	0.07	0.08	0.06	0.08	0.05	0.01	0.07
L	0.03	0.04	0.03	0.04	0.04	0.04	<0.01	0.03
F	0.09	0.05	0.13	0.05	0.19	0.05	0.23	0.05
Deviations due to selection for the common allele—FC p-value test								
X^2	0.04	0.05	0.03	0.04	0.05	0.06	0.03	0.05
G^2	0.05	0.05	0.04	0.04	0.06	0.06	0.04	0.05
H^2	0.04	0.05	0.05	0.05	0.07	0.06	0.04	0.05
L	0.04	0.05	0.03	0.04	0.03	0.06	0.02	0.05
F	0.09	0.05	0.11	0.05	0.13	0.06	0.15	0.05
Deviations due to inbreeding—plain p-value test								
X^2	0.20	0.06	0.34	0.05	0.60	0.04	0.64	0.06
G^2	0.25	0.08	0.29	0.06	0.48	0.06	0.64	0.08
H^2	0.19	0.07	0.18	0.06	0.28	0.05	0.42	0.07
L	0.23	0.04	0.39	0.04	0.63	0.04	0.70	0.03
F	0.11	0.05	0.16	0.05	0.26	0.05	0.29	0.05
Deviations due to inbreeding—FC p-value test								
X^2	0.21	0.05	0.35	0.04	0.61	0.06	0.68	0.05
G^2	0.18	0.05	0.26	0.04	0.48	0.06	0.56	0.05
H^2	0.14	0.05	0.16	0.05	0.30	0.06	0.36	0.05
L	0.25	0.05	0.37	0.04	0.63	0.06	0.74	0.05
F	0.12	0.05	0.15	0.05	0.27	0.06	0.32	0.05

Power and Type I errors are at the 5% significance level, and computed using 5000 simulations from the alternative distribution and expected distribution, respectively, and 5000 Monte-Carlo trials per each simulation.

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*: $r = 10$, $n = 50$, and $\theta_1 = \theta_2 = \frac{1}{3}$, and $\theta_j = \frac{1}{24}$ for $3 \leqslant j \leqslant 10$;
(2) *Alternative*: $r = 10$, $n = 100$, and $\theta_1 = \theta_2 = \frac{1}{3}$, and $\theta_j = \frac{1}{24}$ for $3 \leqslant j \leqslant 10$;
(3) *Alternative*: $r = 10$, $n = 200$, and $\theta_1 = \theta_2 = \frac{1}{3}$, and $\theta_j = \frac{1}{24}$ for $3 \leqslant j \leqslant 10$;
(4) *Alternative*: $r = 20$, $n = 200$, and $\theta_j \sim 1/j$ for $1 \leqslant j \leqslant 20$.

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 with what would be expected under HWE. To account for selection, one introduces

fitness parameters $w_{j,k} > 0$ into the HWE equations,

$$p_{j,k} = \begin{cases} 2(w_{j,k}/\bar{w})\theta_j\theta_k, & 1 \leqslant k < j \leqslant r, \\ (w_{k,k}/\bar{w})\theta_k^2, & j = k, \end{cases} \tag{4.1}$$

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:

$$w_{j,k} = \begin{cases} 1.5, & k = 1, \\ 1, & \text{otherwise}. \end{cases} \tag{4.2}$$

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. In all examples, 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:

$$p_{j,k} = \begin{cases} 2\theta_j\theta_k(1-f), & j > k, \\ \theta_k^2 + f\theta_k(1-\theta_k), & j = k, \end{cases} \quad 1 \leqslant k \leqslant j \leqslant r. \tag{4.3}$$

HWE 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 (4.3) with positive inbreeding coefficient. The root-mean-square statistic appears to be less powerful than the classic statistics in detecting deviations due to inbreeding. Moreover, it is often desired to estimate the inbreeding coefficient f itself, for example, because of its role in quantifying the behavior of marker-trait association tests in non-HWE populations, and the χ^2 test statistic is equal to $n\widehat{f}$, where $\widehat{f}$ is the maximum likelihood estimator for f (see Rori and Weir (2008)).

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 HWE. 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 χ^2 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 1 while the χ^2 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.

$$\textit{Common Allele data set}: \quad \begin{cases} n_{1,1} = r \text{ of type } \{A_1, A_1\}, \\ n_{1,k} = 2 \text{ of type } \{A_1, A_k\}, & 2 \leqslant k \leqslant r+1, \\ n_{j,k} = 0 \text{ of type } \{A_j, A_k\}, & 2 \leqslant j \leqslant k \leqslant r+1. \end{cases} \tag{5.1}$$

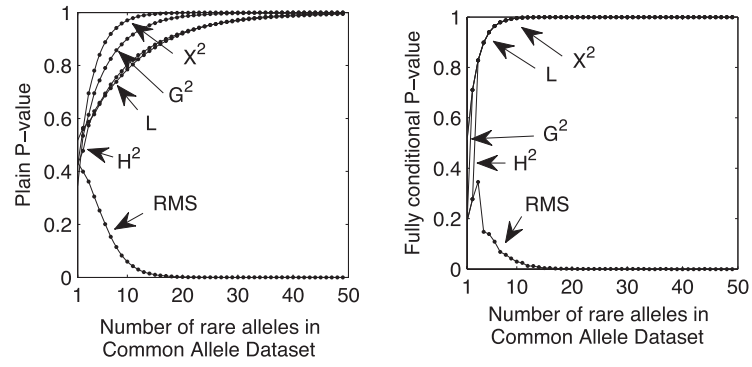

Fig. 5. The p-values (accurate to three digits with 99% confidence) for Pearson's statistic X^2, the log-likelihood-ratio statistic G^2, the Hellinger statistic H^2, and the root-mean-square statistic F in the Common Allele data set to be consistent with the HWE model (2.1), as a function of the number of alleles r. The left plot is for the plain p-values, while the right plot is for the FC p-values.

Note that the Common Allele data set consists of $n_1 = 4r$ alleles of type A_1 and $n_k = 2$ alleles of type A_k, $2 \leqslant k \leqslant r + 1$. The maximum-likelihood model counts are

$$\begin{cases} m_{1,1} = 4r/3, & \\ m_{1,k} = 4/3, & 2 \leqslant k \leqslant r+1, \\ m_{k,k} = 1/(3r), & 2 \leqslant k \leqslant r+1, \\ m_{j,k} = 2/(3r), & 2 \leqslant j < k \leqslant r+1, \quad j < k. \end{cases} \tag{5.2}$$

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 5.1 In the limit as $r \to \infty$, the plain p-values (as computed via Algorithm 1 of Appendix S.1 in supplementary material available at *Biostatistics* online) 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 HWE model all converge to 1, while the plain p-value for the root-mean-square statistic converges to 0.

The proof of Theorem 5.1 is given in Appendix S.2 of supplementary material available at *Biostatistics* online. Figure 5 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 FC p-values, even though Theorem 5.1 applies directly only to the plain p-values. Finally, the particular distribution of the draws in the Common Allele data set was somewhat arbitrary; a similar asymptotic analysis holds for many other data sets. We could have considered instead a data set involving two or three common alleles, one common 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 HWE. The classic tests, tuned to detect *relative* discrepancies, can be blind to 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, easily detects large discrepancies in common genotypes. We demonstrated this in the analysis of three benchmark data sets of Guo and Thompson (1992). We also found that the root-mean-square test can be significantly more powerful at detecting deviations from HWE 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 HWE (see Figure 3), and their combined p-values provide a more fortified test than either p-value used on its own.

7. Software

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

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Mark Tygert, Andrew Gelman, and Abhinav Nellore for their helpful contributions. *Conflict of Interest*: None declared.

Funding

References

Ayres, K. and 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. and Stoker, T. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Annals of Statistics* **34**, 721–741.

Brownlee, K. (1965). *Statistical Series and Methodology in Science and Engineering.* New York: Wiley.

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

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

Council, National Research. (1996). *The Evaluation of Forensic DNA Evidence.* Washington, DC: National Academy Press.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* **26**(1), 363–397.

Efron, B. and 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**(4), 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. AND PRATT, J. (1975). P-values: interpretation and methodology. *The American Statistician* **29**(1), 20–25.

GUO, S. AND 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**(1), 81–93.

LAURETTO, M., NAKANO, F., FARIA, S., PEREIRA, C. AND STERN, J. (2009). A straightforward multiallelic significance test for the Hardy–Weinberg equilibrium law. *Genetics and Molecular Biology* **32**(3), 619–625.

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

LINDLEY, D. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics* **35**(4), 1622–1643.

MAISTE, P. AND WEIR, B. (1995). A comparison of tests for independence in the FBI RFLP data bases. *Genetica* **96**(1–2), 125–138.

PERKINS, W., TYGERT, M. AND WARD, R. (2011). 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. AND WARD, R. (2012). Computing the confidence levels for a root-mean-square test of goodness of fit, II, in preparation.

PERKINS, W., TYGERT, M. AND WARD, R. (2013). Some deficiencies of chi-square and classical exact tests of significance. *Applied and Computational Harmonic Analysis*, to appear.

RADLOW, R. AND 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**(352), 811–813.

RAYMOND, M. AND ROUSSET, F. (1995). An exact test for population differentiation. *Evolution* **49**(6), 1280–1283.

RORI, R. AND WEIR, B. (2008). Distributions of Hardy–Weinberg equilibrium test statistics. *Genetics* **180**(3), 1609–1616.

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

SHOEMAKER, J., PAINTER, I. AND 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, in preparation.

WAKEFIELD, J. (2010). Bayesian methods for examining Hardy–Weinberg equilibrium. *Biometrics* **66**(1), 257–265.

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. AND ABECASIS, G. (2005). A note on exact tests of Hardy–Weinberg equilibrium. *American Journal of Human Genetics* **76**(5), 887–893.

[*Received April 1, 2013; revised June 12, 2013; accepted for publication July 25, 2013*]