Supplementary Material to

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

Rachel Ward and Raymond J. Carroll
Algorithm 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. \( S = X^2, G^2, H^2, \ldots \))

**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 \( s = S(n_{j,k}, m_{j,k}) \)

**Algorithm:**

\[ i \leftarrow 0 \]

repeat

- \( i \leftarrow i + 1 \)
- Draw \( n \) genotypes \( X_1^{(i)}, \ldots, X_q^{(i)}, \ldots, X_n^{(i)} \) 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 \( N_j^{(i)} = (\sum_{k=j}^r N_{k,j}^{(i)} + \sum_{k=1}^{j-1} N_{j,k}^{(i)}) \) 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.

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 \( u_n \geq C v_n \) for all \( n = \{1, 2, \ldots\} \). We use the notation \( u \gtrsim v \) 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 1, along with the Common Allele data set in Table 4 and its maximum-likelihood HWE model.
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 $\ell$, and test statistic $S$ (e.g., $S = X^2, G^2, H^2, \ldots$)

**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 (3.9) to obtain $n$ simulated genotypes $X_1^{(i)}, \ldots, X_q^{(i)}, \ldots, X_n^{(i)}$ 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, $P = \# \{i : S_i \geq s\}/\ell$
```

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,

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

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 $N_{j,k}/n$, $2 \leq k \leq j \leq r + 1$, will equal $1/(3r)$, while the others will equal 0.

4. The coupon collector’s problem (see, for example, Motwani and 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 $\Theta_j = \Theta_j(r)$, $2 \leq j \leq r + 1$, will satisfy

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

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

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

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

$$\tilde{f}^2 = \left(\frac{n_{1,1}}{n} - \frac{m_{1,1}}{n}\right)^2 + \sum_{k=2}^{r+1} \left(\frac{n_{k,1}}{n} - \frac{m_{k,1}}{n}\right)^2 + \sum_{2 \leq k \leq j \leq r+1} \left(\frac{m_{j,k}}{n}\right)^2$$

$$= \left(\frac{1}{9}\right)^2 + \frac{4}{81r} + \frac{1}{81r^2} + \frac{2(r - 1)}{81r^3}. $$

As $r \to \infty$,

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

If we instead consider the sum-square statistic $\bar{F}^2 = \frac{(r+1)(r+2)}{2}F^2$ resulting from drawing $n = 3r$ genotypes i.i.d. from the model distribution (5.2), points 1, 3, and 4 above give

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

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

$$\sim \frac{Z^2}{27r/4} + \frac{(\log(r))^4}{r} + \frac{r}{3} \frac{1}{9r^2} + \left(\frac{r(r + 1)}{2} - \frac{r}{3}\right) \left(\frac{\log(r)}{r}\right)^4. \tag{S.4}$$
where \( Z = (N_{1,1} - 4r/3)/\sqrt{4r/3} \) converges in distribution to a standard normal distribution as \( r \to \infty \). Therefore, as \( r \to \infty \), \( \tilde{F} \to 0 \). Combining (S.3) and (S.4) shows that the p-value for the root-mean-square statistic, \( P = \Pr\{F \geq f\} = \Pr\{\tilde{F} \geq \tilde{f}\} \), 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 \( \tilde{f} \), the measured \( \chi^2 \) discrepancy \( \tilde{X}^2 = \chi^2/n \) 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 \( \chi^2 \) statistic \( \tilde{X}^2 = X^2/n \) corresponding to the unobserved genotypes with one simulated draw,

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

\[
= r \left( \frac{1}{3r} \right)^2 \left( \frac{\log r}{r} \right)^2 / \left( \frac{\log r}{r} \right)^2.
\]  

(S.5)

It follows from (S.5) that \( \tilde{X}^2 \geq r/\{\log(r)\}^2 \to \infty \), and so the p-value for the \( \chi^2 \) statistic, \( P = \Pr(X^2 \geq \chi^2) = \Pr(\tilde{X}^2 \geq \tilde{\chi}^2) \), 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

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

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

\[=.14.\]

(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 proportions will satisfy \( \Theta_j \geq 1/(3r) \), ensuring that at least \( (3/4)^2 r^2 - r \) 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 $r$, we have

$$
\hat{H}^2 \geq \sum_{2 \leq j, k \leq r+1} \left( \sqrt{N_{j,k}/n} - \sqrt{M_{j,k}/n} \right)^2 \\
\geq \# \{ j, k : N_{j,k} = 1 \} \left( \frac{1}{\sqrt{3r}} - \frac{\log (r)}{r} \right)^2 \\
+ \left( \frac{3}{4} \right)^2 \frac{r^2}{2} - r - \# \{ j, k : N_{j,k} = 1 \} \left( \frac{2}{9r^2} \right) \\
\sim r \left( \frac{1}{\sqrt{3r}} - \frac{\log r}{r} \right)^2 + \left( \frac{3}{4} \right)^2 \frac{r^2}{2} - r - \frac{r}{3} \left( \frac{2}{9r^2} \right) \\
\rightarrow .17... 
$$

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