The Number of Markers in the HapMap Project: Some Notes on Chi-Square and Exact Tests for Hardy-Weinberg Equilibrium

To the Editor: Pearson's chi-square test was, until recently, the most widely used procedure for assessing Hardy-Weinberg equilibrium (HWE) in random samples of unrelated individuals.[1–3] Over the last few years, however, Haldane's exact test for HWE has gained popularity. Procedures for testing HWE have been extensively investigated.[4–9] Bayesian and other alternatives for the classical tests have also been proposed.[10–13]

A recent study[14] compared type 1 error rates for the chi-square test with those of Haldane's exact test, and it reported above nominal type 1 error rates for the chi-square test and therefore recommended the exact test in all situations. However, in the comparison,[14] Yates' continuity correction[15,16] had apparently not been applied. In statistics, the continuity correction is widely accepted as a device for improving the accuracy of the results when working with discrete variables.[17]

The p value in an exact test is usually computed as the sum of the probabilities of all samples that are as extreme or more extreme than the current one.[14] An alternative approach is to define the p value as twice the p value of a one-sided test. Because of the nonsymmetrical nature of the Levene-Haldane distribution of the number of heterozygotes given the allele frequency, the two definitions give different results. Yates[16] advocated the use of a doubled one-tail probability as the p value for Fisher's exact test.

In the light of these remarks, a new comparison of the type 1 error rates for chi-square and exact procedures is needed, in which we consider the continuity correction and both definitions of the p value in an exact test. We briefly summarize both tests and compare their type 1 error rates below. The practical implications of using the various procedures are illustrated with HapMap data.[18]

The Pearson chi-square statistic for a test for HWE is given by:

$$X_c^2 = \sum_{i=1}^{3} \frac{(|n_i - e_i| - c)^2}{e_i},$$

in which the n_i represents one of the three genotypic counts (n_{AA}, n_{AB}, and n_{BB}) and e_i the respective expected value under HWE. This is a test for the goodness of fit of a multinomial distribution. Parameter c represents the continuity correction. Setting $c = 0$ gives the ordinary chi-square statistic, and setting c=1/2 gives the corrected chi-square statistic. The p value of the test is obtained by comparing the chi-square statistic with a chi-square distribution with one degree of freedom.

The exact test for HWE[19–21] uses the conditional distribution of the number of heterozygotes, N_{AB}, given the allele count N_A, and is given by

$$\begin{aligned} &P(N_{AB} = n_{AB} \,|\, N_A = n_A) \\ &\quad = \frac{n!n_A!n_B!2^{n_{AB}}}{(2n)!n_{AB}!\left(\frac{1}{2}(n_A - n_{AB})\right)!\frac{1}{2}((n_B - n_{AB}))!}, \end{aligned}$$

in which n_A and n_B refer to the sample counts of A and B alleles. We will refer to this distribution as the Levene-Haldane distribution. Geneticists usually wish to perform a two-sided test, because there is no a priori reason to suppose that a SNP deviating from HWE will show a lack or an excess of heterozygotes. If there are reasons to expect a lack (e.g., inbreeding) or an excess (e.g., overdominance) then a one-sided test is needed. The p value of the exact test is usually calculated as the sum of the probabilities of all possible samples as extreme or more extreme than the observed sample, given the allele count of the observed sample. We refer to this p value as the SELOME p value (sum equally likely or more extreme). An alternative is to define the p value as twice the one-sided tail area, and we will call this p value the DOST p value (double one-sided tail). If the observed number of heterozygotes is below that expected under HWE, the DOST p value is twice the sum of the probabilities of observing the number of heterozygotes in the sample or less. If it is above that expected, then the p value is twice the sum of the probabilities of observing the number of heterozygotes in the sample or more. We argue that DOST p values are the most sensible p values, and we motivate this with the following example.

If we have a sample of n = 100 individuals, and if there are 93 copies of the minor allele, then the corresponding Levene-Haldane distribution of $N_{AB}|N_A$, given in the left panel of Figure 1, is a virtually symmetric distribution with expectation 50.005. Very low and very high heterozygote frequencies both constitute evidence against HWE. Because of the near-symmetric nature of the distribution in this case, it should be evident that observing 61 heterozygotes in a sample constitutes virtually as much evidence against HWE as observing 39 heterozygotes. Observing 61 heterozygotes or more has a probability of 0.021670, and observing 39 or fewer heterozygotes has nearly the same probability, 0.021674 (Table 1). Suppose we observed 61 heterozygotes. If we wish to perform a two-sided test, the obvious DOST p value is 2 × 0.021670 = 0.04334. When we use the SELOME rule, the probability of observing a sample as extreme as 61 heterozygotes or more extreme is 0.04334, but the probability of observing a sample as extreme as 39 heterozygotes is *different*: 0.029243. For a symmetrical

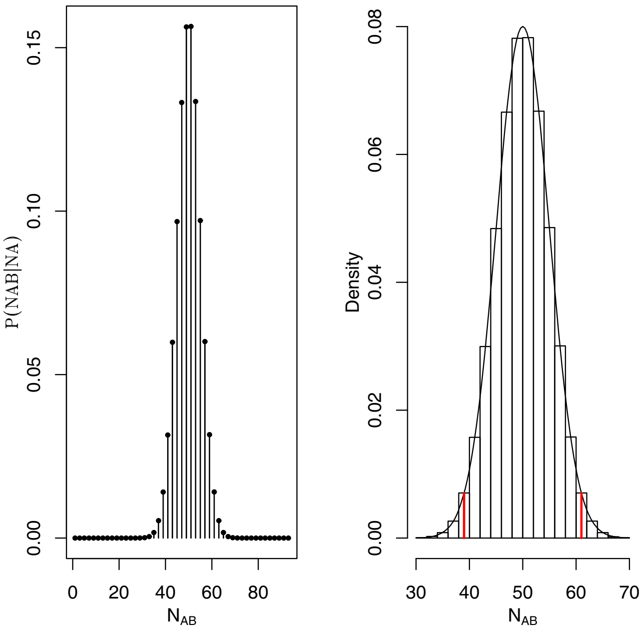

Figure 1. Levene-Haldane Distribution of the Number of Heterozygotes for a Given Allele Count without and with Normal Approximation

Left panel: Levene-Haldane distribution for n = 100, $n_A = 93$. Right panel: Levene-Haldane distribution with normal approximation.

Table 1. Sample Probabilities and p Values

n_{AA}	n_{AB}	n_{BB}	$P(N_{AB}\|N_A)$	$P(N_{AB} \geq n_{AB})$	$P(N_{AB} \leq n_{AB})$	p_{selome}	p_{dost}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
35	23	42	0.000000	1.000000	0.000000	0.000000	0.000000
34	25	41	0.000000	1.000000	0.000000	0.000001	0.000001
33	27	40	0.000003	1.000000	0.000003	0.000006	0.000007
32	29	39	0.000019	0.999997	0.000022	0.000044	0.000045
31	31	38	0.000102	0.999978	0.000124	0.000245	0.000249
30	33	37	0.000455	0.999876	0.000579	0.001148	0.001158
29	35	36	0.001697	0.999421	0.002277	0.004532	0.004553
28	37	35	0.005322	0.997723	0.007598	0.015168	0.015196
27	**39**	34	**0.014076**	0.992402	**0.021674**	**0.029243**	**0.043348**
26	41	33	0.031516	0.978326	0.053190	0.074861	0.106380
25	43	32	0.059891	0.946810	0.113081	0.166375	0.226163
24	45	31	0.096794	0.886919	0.209875	0.323289	0.419751
23	47	30	0.133237	0.790125	0.343113	0.553643	0.686225
22	49	29	0.156350	0.656887	0.499462	0.843528	0.998925
21	51	28	0.156472	0.500538	0.655935	1.000000	1.000000
20	53	27	0.133535	0.344065	0.789470	0.687178	0.688131
19	55	26	0.097117	0.210530	0.886586	0.420405	0.421060
18	57	25	0.060120	0.113414	0.946706	0.226495	0.226827
17	59	24	0.031623	0.053294	0.978330	0.106484	0.106588
16	**61**	23	**0.014101**	**0.021670**	0.992431	**0.043344**	**0.043341**
15	63	22	0.005314	0.007569	0.997745	0.009846	0.015139
14	65	21	0.001686	0.002255	0.999431	0.002835	0.004511
13	67	20	0.000448	0.000569	0.999879	0.000693	0.001138
12	69	19	0.000099	0.000121	0.999979	0.000143	0.000242
11	71	18	0.000018	0.000021	0.999997	0.000025	0.000043
10	73	17	0.000003	0.000003	1.000000	0.000004	0.000006
9	75	16	0.000000	0.000000	1.000000	0.000000	0.000001
8	77	15	0.000000	0.000000	1.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

All possible samples for n = 100, $n_A = 93$. The table shows the probabilities of observing the sample $P(N_{AB}|N_A)$, cumulative probabilities, and the SELOME and DOST p values of an exact test for each possible sample. Bold entries indicate probabilities that are referred to in the text.

distribution, this difference seems absurd. The reason for this difference is that the probability of $P(N_{AB} = 61|N_A = 93) = 0.014101$ is omitted in the calculation of the latter p value, because it is slightly larger than $P(N_{AB} = 39|N_A = 93) = 0.014076$ (Table 1). On common-sense grounds, the practice of summing probabilities "as extreme or more extreme as those observed" seems mistaken. This is further exemplified by approximating the Levene-Haldane distribution with a normal distribution, as is done in the right panel of Figure 1. Under the approximating normal curve, the evidence against a null value of 50.005 is evidently twice the probability of exceeding 61, and this *equals* twice the probability of observing 39 or less. In practice, the discrete Levene-Haldane distribution is more asymmetric than in the example above, but it can often be well approximated by a normal curve. It is markedly asymmetric for extreme allele frequencies, but it can then be approximated by a normal curve after proper transformation. In short, doubling the one-sided tail area seems a more adequate way to compute the p value in Haldane's exact test and is much more in line with statistical procedures for continuous variables, as well as with the classical chi-square test, the latter also being essentially a two-sided test when considered as the square of an $N(0, 1)$ variate. Table 1 also shows that SELOME p values are generally smaller than DOST p values in both tails and therefore more easily lead to rejection of HWE.

We compared the type 1 error rates of chi-square tests and exact tests with both types of p values. Type 1 error rates can be computed exactly by summing the probabilities of all genotypic compositions that pertain to the rejection region. We computed rejection rates for the same combinations of parameters used previously,[14] with 100 or 1000 individuals and three significance levels (0.05, 0.01, and 0.001). Figure 2 shows the error rates for the exact test with DOST p values, the chi-square test, and the chi-square test with continuity correction. DOST p values in fact form the natural choice, because the tests being compared are now both actually two-tailed.

Figure 2 shows the inflated type 1 error rates for the ordinary chi-square test (blue) in comparison with the exact

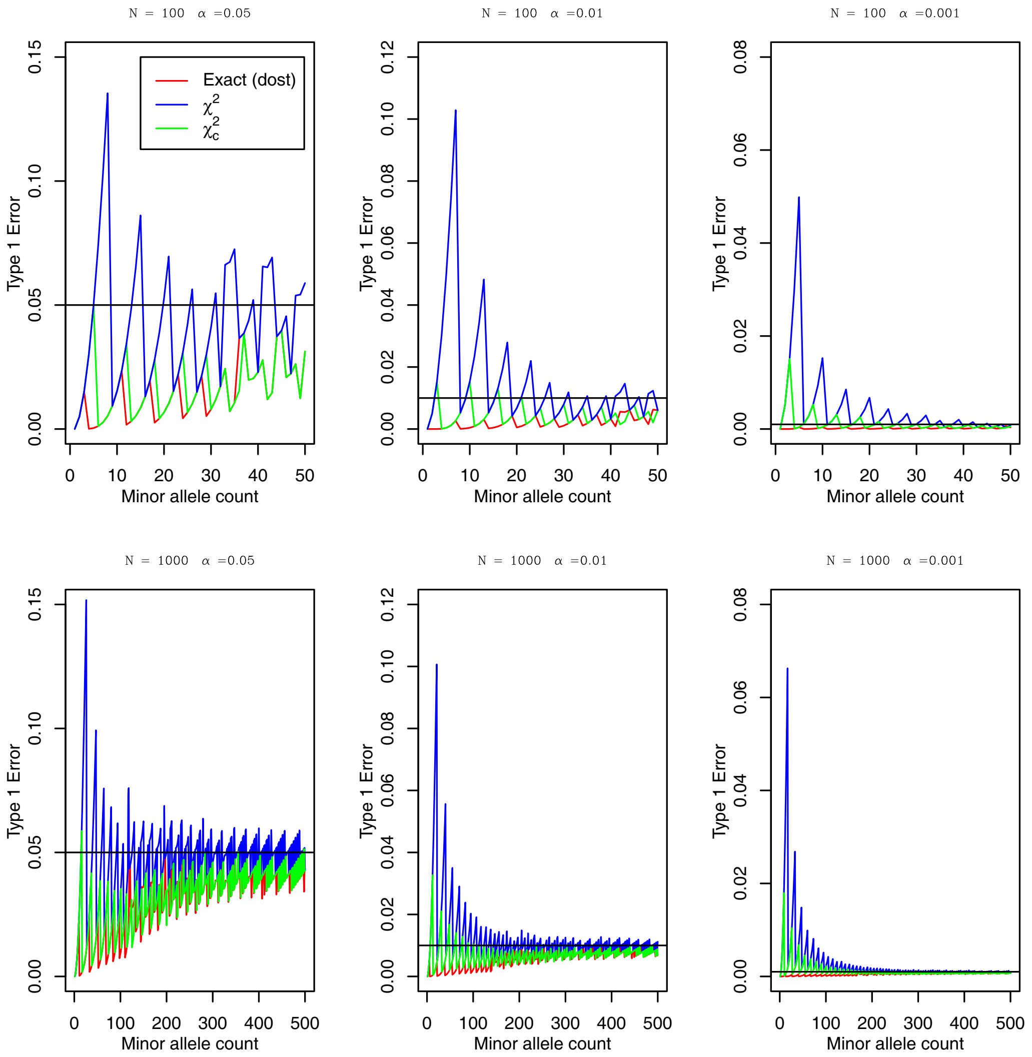

Figure 2. Type 1 Error Rates as a Function of Sample Size and α for Different Statistical Tests
Type 1 error rates for different sample sizes (100, 1000) and significance levels (0.05, 0.01, and 0.001) for the exact test (red), the chi-square test (blue), and the chi-square test with continuity correction (green).

test reported previously.[14] However, the graph also shows that the continuity correction (green) effectively reduces this inflation, bringing the chi-square test into very close agreement with the exact test. The better agreement of the corrected chi-square test with the exact test has been noted before with numerical examples.[22]

The chi-square test with correction has highly inflated rates (100%) for very small minor allele counts. This is due to an edge effect of the continuity correction.[23] This edge effect is easily avoided by using a cutoff for the continuity correction for low minor allele frequencies, as was done in Figure 2. The test with correction has a rejection rate that is mostly below the nominal level for $\alpha = 0.05$ or 0.01. Often the test with correction is the closest to the nominal level. Results of HWE tests are often poorly reported in association studies.[2] We add that it is typically not reported whether a continuity correction has been applied or not.

Figure 3 compares the error rates of the exact test for both definitions of the p value. Both tests have a rejection rate that is always below the nominal level. The SELOME rates are closer to the nominal level and are larger than or equal to the DOST rates. When the distribution of the number of heterozygotes is asymmetric, the exact test that uses the SELOME p values is essentially a *one-sided* test, because all probabilities that contribute to the p value are in one tail of the distribution only. Evidently a one-tailed test has, as Figure 3 shows, better power, but this gain in power is irrelevant if one really needs a *two-sided* test. We therefore recommend the use of DOST p values in the exact test for HWE.

We use a HapMap database[18] from chromosome 1 to illustrate the effects on marker admission of the choices made in chi-square and exact tests. The HapMap project currently uses the exact test for HWE with criterion $p > 0.001$ as a filter for the inclusion of a SNP in the database. This is based on the idea that strong deviations from HWE may be the result of genotyping error. Violation of HWE may, however, be due to many alternative

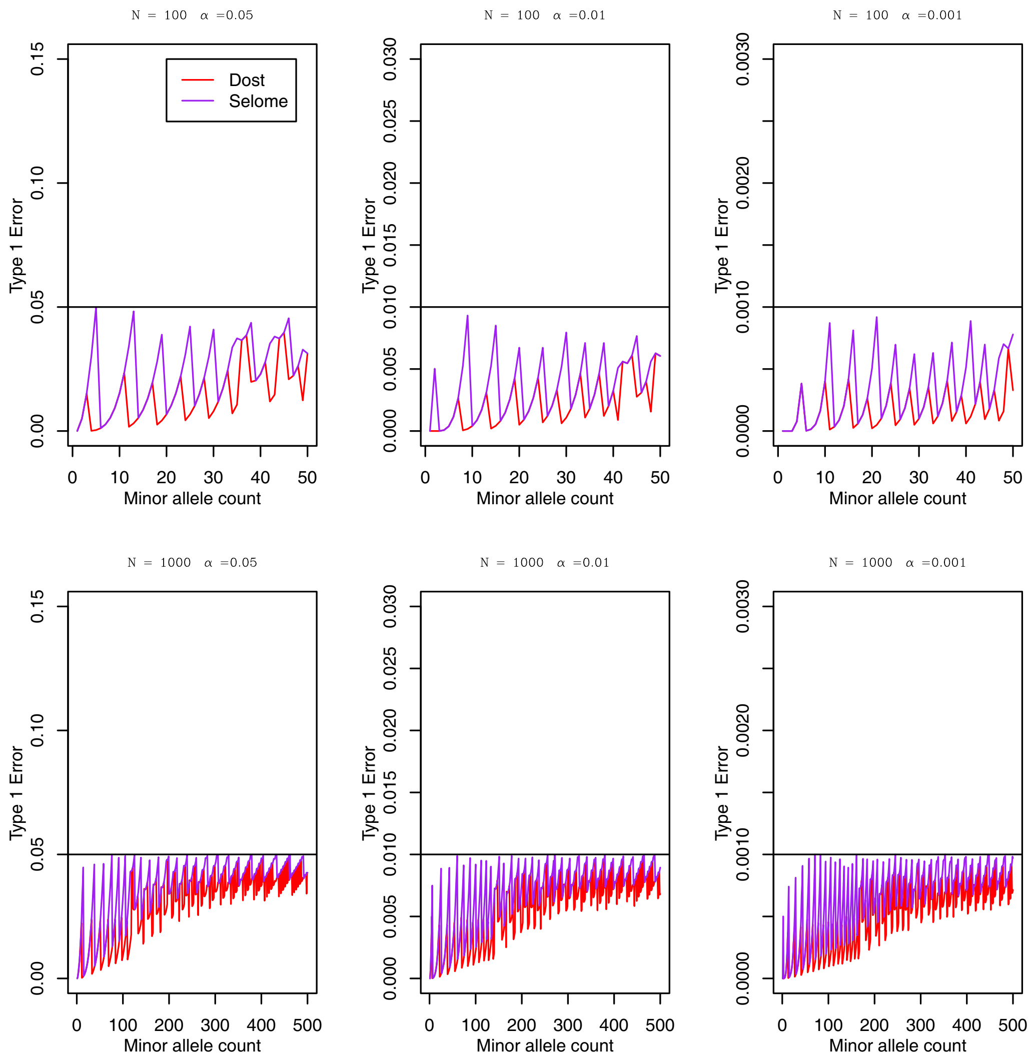

Figure 3. DOST and SELOME Type 1 Error Rates as a Function of Sample Size and α
Type 1 error rates for different sample sizes (100, 1000) and significance levels (0.05, 0.01, and 0.001) for exact tests with SELOME p values (purple) and DOST p values (red).

explanations, such as selection, nonrandom mating, population substructure, and, not in the least, disease association.[1,24] Several scholars[25–27] have therefore argued that HWE tests should be performed but not used as a criterion for excluding markers prior to association study. We used the Han Chinese sample from Beijing (CHB), consisting of 45 unrelated individuals (phase II, NCBI build 35). This database contains 529,081 redundant, unfiltered markers. The database has three additional duplicate individuals, and many submitted SNPs are repeated. Of each repeated SNP, we selected the one which had the fewest missing values. Next, SNPs were filtered according to HapMap criteria,[18,28,29] by eliminating SNPs that had more than one inconsistency over the three duplicates and by eliminating SNPs with more than 20% missing values. After filtering, the database consisted of 45 individuals typed for 337,746 SNPs. Of these, 42% were monomorphic, and 16.8% of the polymorphic SNPs had a minor allele frequency below 0.05. We analyzed this filtered database by using the four different tests for HWE described above. We used the R package[30] HardyWeinberg (version 1.4) for the computation of all test results. Rejection rates for the different tests are given in Table 2.

Table 2 shows that DOST p values have the lowest rejection rate and form the most conservative approach to testing HWE. The ordinary chi-square test has the highest rejection rates, followed by exact SELOME and corrected chi-square. When the criterion for inclusion of a SNP is changed from SELOME to DOST p values, an additional amount of 0.73% of the SNPs would be admitted at the 5% level, or 0.1% at the 0.1% level. These percentages look small, but genome-wide they correspond to a large amount of markers. With 3.1 million *admitted* SNPs genome-wide[18] this corresponds roughly to minimally 22,630 additional SNPs admitted at the 5% level or minimally 3100 additional SNPs at the 0.1% level. In practice, the number of additionally admitted SNPs will be larger, because the number of unfiltered SNPs in the project is

Table 2. Rejection Rates for HWE Tests

	Rejected (%)		
HWE Test	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
χ^2	4.73	2.77	1.87
χ^2_c (with cutoff)	3.19	2.01	1.40
Exact (DOST)	2.86	1.70	1.20
Exact (SELOME)	3.59	1.90	1.30

Rejection rates for different tests for HWE for a HapMap database of 337,746 SNPs from the CHB population of 45 individuals, for three different levels of α.

well over 3.1 million. We note that the HapMap database is an empirical database and that the rejection rates in Table 2 are therefore not expected to coincide with the theoretical levels of 5%, 1%, or 0.1%, the true number of markers out of HWE being unknown. We investigated the "newly admitted" markers in some detail. Figure 4 shows a ternary plot of the newly admitted markers without missing data (sample size 45). The plot shows the acceptance regions of the chi-square test with and without continuity correction and the acceptance regions of the exact test with the SELOME and the DOST criterion, with $\alpha = 0.001$. The zigzag lines for the exact tests connect samples for which the exact test is just significant. The newly admitted SNPs cover the whole range of allele frequencies and are typically around the boundary of the acceptance region of a corrected chi-square test for HWE. The exact test using the DOST criterion has the largest acceptance region. Note that for some intermediate allele frequencies, equilibrium is rejected according to a SELOME exact test but accepted by a corrected chi-square. The ordinary chi-square test has the smallest acceptance region.

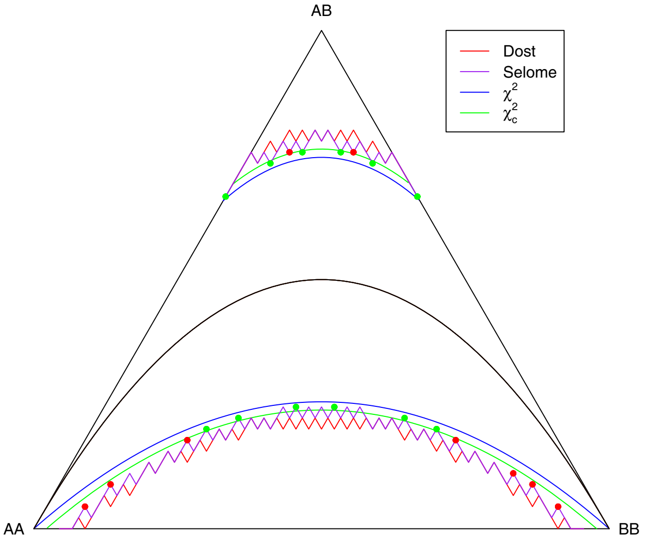

Figure 4. Ternary Plot of Extra Admitted Markers
Ternary plot of newly admitted markers without missing data. The black curve represents HWE. Acceptance regions of the chi-square test with and without correction (green and blue, respectively) and the exact tests with SELOME p values (purple) and DOST p values (red) are shown. Green and red dots indicate nonsignificant and significant SNPs, respectively, for the corrected chi-square test with $\alpha = 0.001$.

Constructing reliable SNP assays in the laboratory is expensive and time consuming. We have no sound statistical reasons to reject HWE for SNPs that have a significant SELOME p value but a nonsignificant DOST p value. The logical consequence is to admit these markers to the HapMap project. This will increase the genomic coverage of the project, and, after all, these markers may be associated with disease.

Jan Graffelman[1,*]

[1]Universitat Politècnica de Catalunya, Departament d'Estadística i Investigació Operativa, Avinguda Diagonal 647, 6th floor, 08028 Barcelona, Spain

*Correspondence: jan.graffelman@upc.edu

Acknowledgments

This study was supported by grants SEC2003-04476 and CODA-RSS MTM2009-13272 of the Spanish Ministry of Education and Science. The author thanks the referees for their comments on the paper.

Web Resources

The URLs for data presented herein are as follows:

R software, http://www.r-project.org
R package HardyWeinberg, version 1.4. http://www-eio.upc.es/~jan and http://www.r-project.org
HapMap databases, http://www.hapmap.org

References

1. Wittke-Thompson, J.K., Pluzhnikov, A., and Cox, N.J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. Am. J. Hum. Genet. *76*, 967–986.
2. Salanti, G., Amountza, G., Ntzani, E.E., and Ioannidis, J.P.A. (2005). Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. Eur. J. Hum. Genet. *13*, 840–848.
3. Yu, C., Zhang, S., Zhou, C., and Sile, S. (2009). A likelihood ratio test of population Hardy-Weinberg equilibrium for case-control studies. Genet. Epidemiol. *33*, 275–280.
4. Elston, R.C., and Forthofer, R. (1977). Testing for Hardy-Weinberg equilibrium in small samples. Biometrics *33*, 536–542.
5. Emigh, T.H. (1980). A comparison of tests for Hardy-Weinberg equilibrium. Biometrics *36*, 627–642.
6. Hernández, J.L., and Weir, B.S. (1989). A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics *45*, 53–70.
7. Guo, S.W., and Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics *48*, 361–372.
8. Gomes, I., Collins, A., Lonjou, C., Thomas, N.S., Wilkinson, J., Watson, M., and Morton, N. (1999). Hardy-Weinberg quality control. Ann. Hum. Genet. *63*, 535–538.
9. Cox, D.G., and Kraft, P. (2006). Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum. Hered. *61*, 10–14.

10. Lindley, D.V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In Bayesian Statistics, 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds. (Oxford, UK: Oxford University Press), pp. 307–326.
11. Montoya-Delgado, L.E., Irony, T.Z., de B Pereira, C.A., and Whittle, M.R. (2001). An unconditional exact test for the Hardy-Weinberg equilibrium law: sample-space ordering using the Bayes factor. Genetics *158*, 875–883.
12. Wellek, S. (2004). Tests for establishing compatibility of an observed genotype distribution with Hardy-Weinberg equilibrium in the case of a biallelic locus. Biometrics *60*, 694–703.
13. Pereira, C.A.B., Nakano, F., Stern, J.M., and Whittle, M.R. (2006). Genuine Bayesian multiallelic significance test for the Hardy-Weinberg equilibrium law. Genet. Mol. Res. *5*, 619–631.
14. Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. *76*, 887–893.
15. Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. Journal of the Royal Statistical Society (Supplement) *1*, 217–235.
16. Yates, F. (1984). Tests of significance for 2 × 2 contingency tables. J. R. Stat. Soc. [Ser A] *147*, 426–463.
17. Fleiss, J.L. (1981). Statistical methods for rates and proportions, Second Edition (New York: John Wiley & Sons).
18. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
19. Haldane, J.B.S. (1954). An exact test for randomness of mating. J. Genet. *52*, 631–635.
20. Levene, H. (1949). On a matching problem arising in genetics. Ann. Math. Stat. *20*, 91–94.
21. Weir, B.S. (1996). Genetic Data Analysis II (Massachusetts: Sinauer Associates).
22. Rohlfs, R.V., and Weir, B.S. (2008). Distributions of Hardy-Weinberg equilibrium test statistics. Genetics *180*, 1609–1616.
23. Graffelman, J., and Camarena, J.M. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum. Hered. *65*, 77–84.
24. Li, M., and Li, C. (2008). Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet. Epidemiol. *32*, 589–599.
25. Fardo, D.W., Becker, K.D., Bertram, L., Tanzi, R.E., and Lange, C. (2009). Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium. Eur. J. Hum. Genet. *17*, 1676–1682.
26. Minelli, C., Thompson, J.R., Abrams, K.R., Thakkinstian, A., and Attia, J. (2008). How should we use information about HWE in the meta-analyses of genetic association studies? Int. J. Epidemiol. *37*, 136–146.
27. Zou, G.Y., and Donner, A. (2006). The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann. Hum. Genet. *70*, 923–933.
28. International HapMap Consortium. (2003). The international hapmap project. Nature *426*, 789–796.
29. International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
30. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL http://www.R-project.org. ISBN 3-900051-00-3.

DOI 10.1016/j.ajhg.2009.11.019. ©2010 by The American Society of Human Genetics. All rights reserved.

Response to Graffelman: Tests of Hardy-Weinberg Equilibrium

To the Editor: Testing for Hardy-Weinberg equilibrium (HWE) is perhaps the most common quality-control procedure in all of human genetics. Although there are many potential explanations for departures from HWE, the prototypical causes of departure from HWE are genotyping error and differential missing-data rates among genotypes.[1] These two are critically important because they can give rise to false positives in genetic association studies.[2] Standard practice in association studies is to test for HWE in all samples (or control samples) and to reject any marker with a p value for HWE $< \alpha$. For the HapMap project,[3,4] $\alpha = 0.001$, but other studies might elect different values.

For large samples and common alleles, a convenient means of calculating these p values is to use a simple χ^2 test. However, this χ^2 test requires two simplifying assumptions that are never true: (1) that heterozygote counts are approximately normally distributed and (b) that these counts are continuous. In a Letter to the Editor, Graffelman suggests that a continuity correction mitigates problems associated with the second assumption. In our view, the best solution to the problems associated with using a χ^2 test is the use of an exact test. A major impediment to exact tests is the associated computational burden, but that burden is greatly diminished with the use of the algorithm of Wigginton et al.[5] for calculating exact probabilities and test statistics.

Wigginton et al. note that with exact probabilities in hand, there are four possible tests of HWE. Specifically, they outline two one-tailed tests (P_{low}, P_{high}) and two two-tailed tests (P_{HWE}, $P_{2\alpha}$). They define P_{HWE} as the probability of observing a genotype configuration at least as unlikely as that actually observed and $P_{2\alpha}$ as $\min(1.0, 2P_{high}, 2P_{low})$. Wigginton et al. recommend that P_{HWE} should be used in almost all circumstances and discard $P_{2\alpha}$ as too conservative (i.e., as producing incorrect probability values).

$P_{DOST} = \min(2P_{high}, 2P_{low})$, the statistic proposed by Graffelman, is just an imperfect approximation of $P_{2\alpha}$. P_{DOST} often takes values > 1.0 and still produces