Statistical inference for Hardy-Weinberg equilibrium using log-ratio coordinates

J. GRAFFELMAN[1]

[1]Department of Statistics and Operations Research - Universitat Politècnica de Catalunya, Spain
jan.graffelman@upc.edu

Abstract

Testing markers for Hardy-Weinberg equilibrium (HWE) is an important step in the analysis of large databases used in genetic association studies. Gross deviation from HWE can be indicative of genotyping error. There are many approaches to testing markers for HWE. The classical chi-square test was, till recently, the most widely used approach to HWE-testing. Over the last decade, the computationally more demanding exact test has become more popular. Bayesian approaches, where the full posterior distribution of a disequilibrium parameter is obtained, have also been developed. As far as CODA is concerned, Aitchison described how the HWE law can be "discovered" when a set of samples, all genotyped for the same marker, is analyzed by log-ratio principal component analysis. A well-known tool in CODA, the ternary plot, is known in genetics as a de Finetti diagram. The Hardy-Weinberg law defines a parabola in a ternary plot of the three genotypes frequencies of a bi-allelic marker. Ternary plots of bi-allelic genetic markers typically show points that "follow" the parabola, though with certain scatter that depends on the sample size. When represented in additive, centered or isometric log-ratio coordinates, the HW parabola becomes a straight line. Much of CODA is concerned with data sets where *each individual row* in the data set (an individual, a sample, an object) constitutes a composition. In data sets comprising genetic markers, individual rows (persons) are not really compositions, but it is the *total sample of all individuals* that constitutes a composition. The CODA approach to genetic data has shown useful in supplying interesting graphics, but to date CODA seems not to have provided formal statistical inference for HWE, probably because the distribution of the log-ratio coordinates is not known. Nevertheless, the log-ratio approach directly suggests some statistics that can be used for measuring disequilibrium: the second clr and the second ilr coordinate of the sample. Similar statistics have been used in the genetics literature. In this contribution, we will use the multivariate delta method to derive the asymptotic distribution of the isometric log-ratio coordinates. This allows hypothesis testing for HWE and the construction of confidence intervals for large samples that contain no zeros. The type 1 error rate of the test is compared with the classical chi-square test.

1 Introduction

The Hardy-Weinberg law states that, in the absence of disturbing forces (selection, migration, mutation, non-random mating and others), the genotypic composition of a population with respect to a genetic marker attains a stable equilibrium in one generation. For a bi-allelic marker (e.g. a single nucleotide polymorphism (SNP)), with alleles A and B with respective allele frequencies p and $q = 1-p$, this implies that the genotypes AA, AB and BB will occur in the proportions p^2, $2pq$ and q^2 respectively. We will use $\mathbf{f} = (f_{AA}, f_{AB}, f_{BB})$ to indicate the relative population frequencies of the three genotypes, and $\hat{\mathbf{f}} = (\hat{f}_{AA}, \hat{f}_{AB}, \hat{f}_{BB})$ to denote their sample estimators. A common alternative formulation of the Hardy-Weinberg law is obtained by squaring the heterozygote frequency: $f_{AB}^2 = 4f_{AA}f_{BB}$. Several statistical tests are available for testing genetic markers for HWE. Standard chi-square and exact tests for HWE can be found in Weir (1996). Here, we consider a test based on the isometric logratio transformation (Egozcue et al., 2003). For composition $\mathbf{f}$, the ilr transformation of the genotype counts is given by:

$$\mathbf{y} = \text{ilr}(\mathbf{f}) = \begin{cases} \left(\frac{1}{\sqrt{2}} \ln \frac{f_{AA}}{f_{BB}}, \frac{1}{\sqrt{6}} \ln \frac{f_{AA}f_{BB}}{f_{AB}^2}\right) \\ \left(\frac{1}{\sqrt{2}} \ln \frac{f_{AB}}{f_{BB}}, \frac{1}{\sqrt{6}} \ln \frac{f_{AB}f_{BB}}{f_{AA}^2}\right) \\ \left(\frac{1}{\sqrt{2}} \ln \frac{f_{AA}}{f_{AB}}, \frac{1}{\sqrt{6}} \ln \frac{f_{AA}f_{AB}}{f_{BB}^2}\right) \end{cases} . \quad (1)$$

We will use the first of these three transformations in the remainder of this paper. Under HWE, we have $\mathbf{y} = (\sqrt{2} \ln \frac{p}{1-p}, -\sqrt{2/3} \ln(2))$. The first ilr coordinate is the logit of the allele frequency,

multiplied by $\sqrt{2}$, and the second coordinate turns out to be a constant, $-\sqrt{2/3}\ln(2) = -0.5660$. In a previous contribution (Graffelman and Egozcue, 2011) we used a bootstrap approach to test for HWE using this second ilr coordinate. In this paper, we use a parametric approach, and obtain the asymptotic distribution of the ilr coordinates.

2 A parametric test for HWE based on isometric logratio coordinates

The sample counts n_{AA}, n_{AB}, n_{BB} constitute a sample of size n from a multinomial distribution with parameter vector $\mathbf{f} = (f_{AA}, f_{AB}, f_{BB}) = (p^2, 2pq, q^2)$. The maximum likelihood estimator for $\mathbf{f}$ is given by the vector of relative sample frequencies $\hat{\mathbf{f}} = (1/n)(n_{AA}, n_{AB}, n_{BB})$. The asymptotic distribution of $\hat{\mathbf{f}}$ is the multivariate normal distribution $N(\mathbf{f}, \mathbf{D}_f - \mathbf{f}\mathbf{f}')$, where $\mathbf{D}_f = \text{diag}(\mathbf{f})$. We denote the theoretical population logratio coordinates by $\mathbf{y}$, and use $\hat{\mathbf{y}}$ to refer to the corresponding sample estimates. By applying the delta method (Wasserman, 2010), we obtain the asymptotic distribution of $\sqrt{n}(\hat{\mathbf{y}} - \mathbf{y})$ as multivariate normal $N(\mathbf{0}, \mathbf{\Sigma})$, with the covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \frac{1}{2}\left(\frac{1}{f_{AA}} + \frac{1}{f_{BB}}\right) & \frac{1}{2\sqrt{3}}\left(\frac{1}{f_{AA}} - \frac{1}{f_{BB}}\right) \\ \frac{1}{2\sqrt{3}}\left(\frac{1}{f_{AA}} - \frac{1}{f_{BB}}\right) & \frac{1}{6}\left(\frac{1}{f_{AA}} + \frac{4}{f_{AB}} + \frac{1}{f_{BB}}\right) \end{bmatrix}. \tag{2}$$

Thus, under HWE we have $\hat{y}_2 \approx N\left(-\sqrt{2/3}\ln(2), \frac{1}{6n}\left(\frac{1}{f_{AA}} + \frac{4}{f_{AB}} + \frac{1}{f_{BB}}\right)\right)$. The following hypothesis test for HWE

$$H_0 : y_2 = -\sqrt{2/3}\ln(2),$$
$$H_1 : y_2 \neq -\sqrt{2/3}\ln(2),$$

can be carried out by computing

$$Z = \frac{\hat{y}_2 + \sqrt{2/3}\ln(2)}{\sqrt{\frac{1}{6n}\left(\frac{1}{\hat{f}_{AA}} + \frac{4}{\hat{f}_{AB}} + \frac{1}{\hat{f}_{BB}}\right)}},$$

where $\hat{y}_2$ indicates the sample estimate of the second ilr coordinate. A $100(1-\alpha)\%$ confidence interval for the second theoretical ilr coordinate can be constructed as

$$CI_{1-\alpha}(y_2) = \hat{y}_2 \pm z_{\alpha/2}\sqrt{\frac{1}{6n}\left(\frac{1}{\hat{f}_{AA}} + \frac{4}{\hat{f}_{AB}} + \frac{1}{\hat{f}_{BB}}\right)}, \tag{3}$$

where HWE would be rejected at level α if -0.5660 is outside this interval. Instead of using the ilr coordinates directly, it may be more natural to reparametrize the test as

$$\theta = -y_2 - \sqrt{2/3}\ln(2) = \frac{1}{2}\sqrt{\frac{2}{3}} \cdot \ln\left(\frac{f_{AB}^2}{4 f_{AA} f_{BB}}\right) \tag{4}$$

and test the null hypothesis $H_0 : \theta = 0$. The interpretation is then more simple since $\theta > 0$ indicates heterozygote excess, and $\theta < 0$ heterozygote dearth. The ratio $(4f_{AA}f_{BB})/(f_{AB}^2)$ or its inverse is used in genetics as a measure of disequilibrium (Olson and Foley, 1996). We estimate θ by $\hat{\theta} = -\hat{y}_2 - \sqrt{\frac{2}{3}}\ln(2)$, and our test statistic becomes

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{6n}\left(\frac{1}{\hat{f}_{AA}} + \frac{4}{\hat{f}_{AB}} + \frac{1}{\hat{f}_{BB}}\right)}}, \tag{5}$$

and a $100(1-\alpha)\%$ confidence interval for θ is given by

$$CI_{1-\alpha}(\theta) = \hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{6n}\left(\frac{1}{\hat{f}_{AA}} + \frac{4}{\hat{f}_{AB}} + \frac{1}{\hat{f}_{BB}}\right)}. \quad (6)$$

3 Simulations

We perform a small simulation study concerning the rejection rate of the asymptotic test developed in the previous section. We generate genotypic compositions under the HWE assumption by drawing samples from a trinomial distribution with parameters p^2, $2pq$ and q^2, and three different sample sizes, (50, 100 or 500 trials) for a given allele frequency. We vary the allele frequency from 0.1 to 0.9 in steps of 0.1. We use 10,000 Monte Carlo simulations, and estimate the rejection rate of our test using $\alpha = 0.05$. The results are shown in Table 1.

	n=50		n=100		n=500	
p	Z_{ilr}	χ^2	Z_{ilr}	χ^2	Z_{ilr}	χ^2
0.1	0.0491	0.0780	0.0456	0.0634	0.0403	0.0509
0.2	0.0299	0.0440	0.0298	0.0433	0.0449	0.0468
0.3	0.0267	0.0451	0.0419	0.0480	0.0459	0.0465
0.4	0.0434	0.0483	0.0469	0.0482	0.0481	0.0486
0.5	0.0461	0.0465	0.0475	0.0475	0.0510	0.0516
0.6	0.0441	0.0503	0.0476	0.0489	0.0538	0.0542
0.7	0.0298	0.0481	0.0436	0.0487	0.0512	0.0527
0.8	0.0294	0.0433	0.0308	0.0443	0.0495	0.0517
0.9	0.0498	0.0782	0.0452	0.0619	0.0358	0.0448

Table 1: Monte Carlo rejection rates (10,000 simulations) for two tests for HWE, the standard chi-square test (χ^2) and the Z-test based on the second ilr coordinate (Z_{ilr}). p = allele frequency.

Table 1 shows that the standard chi-square test has above nominal rejection rates for extreme allele frequencies in smaller samples. This has been described in the literature (Emigh, 1980; Graffelman, 2010). The Z-test based on the second ilr coordinate has somewhat lower rejection rates and for the smaller samples seems closer to the nominal level at extreme allele frequencies.

4 Examples

We use four different data sets and apply 6 tests for HWE to these data sets. Two datasets concern large samples of bloodgroup data for the MN locus. The first large sample is a sample of 1,000 donors for which the MN bloodgroup was determined (Hedrick, 2005; Cleghorn, 1960). The genotype counts were (298,489,213) for MM, MN and NN individuals respectively. This sample has an intermediate allele frequency ($p_M = 0.54$). The second large sample is taken from a human population in Oceania (Mourant et al., 1976) and has genotype counts (12,110,1026) and a more extreme allele frequency ($p_M = 0.06$). Two smaller samples consist of 28 individuals of a French population for which we consider the genotype frequencies of two AG polymorphisms (see Table 2), one with an intermediate allele frequency ($p_G = 0.52$), and one with an extreme allele frequency ($p_G = 0.91$). We apply six statistical tests to each sample: the classical χ^2 test with and without continuity correction, the exact test with its usual p-value and the more conservative dost p-value (Graffelman, 2010), the bootstrap procedure for the second ilr coordinate previously described by Graffelman and Egozcue (2011), and the asymptotic test developed in the previous section. Results of all tests are shown in Table 2.
Table 2 shows that for the MN data all tests give very similar results, in particular for the intermediate allele frequency. In this case the test based on the second ilr coordinate gives virtually the same result as a classical chi-square test without continuity correction ($(-0.4706)^2 = 0.2215$). For the large sample with the more extreme allele frequency the asymptotic ilr test (Z_{ilr}) is more conservative than the standard χ^2 and exact tests, though all tests qualitatively agree in that they reject HWE. For the

	MN data				GA data			
	(298,489,213)		(12,110,1026)		(8,13,7)		(24,3,1)	
	$n = 1000$	$p_M = 0.54$	$n = 1148$	$p_M = 0.06$	$n = 28$	$p_G = 0.52$	$n = 28$	$p_G = 0.91$
	Stat	p-value	Stat	p-value	Stat	p-value	Stat	p-value
$\chi^2_{cc=0.5}$	0.1790	0.6723	16.7364	0.0000	0.0170	0.8973	0.5903	0.4423
χ^2	0.2215	0.6379	18.8752	0.0000	0.1380	0.7101	3.2592	0.0710
Z_{ilr}	-0.4706	0.6379	-4.0407	0.0001	-0.3714	0.7104	-1.5360	0.1245
Exact (standard)	-	0.6557	-	0.0003	-	0.7149	-	0.1767
Exact (dost)	-	0.6723	-	0.0005	-	0.9282	-	0.3533
Bootstrap	-0.5407	0.6300	0.0203	0.0010	-0.4288	0.6390	0.4485	0.0730

Table 2: Test statistics and p-values of six tests for HWE using 4 different samples. χ^2: standard chi-square statistic for HWE. $\chi^2_{cc=0.5}$: chi-square test with continuity correction of 0.5. Z_{ilr}: Z-statistic for the test based on the second ilr coordinate.

intermediate allele frequency of the GA polymorphism there is again a close agreement between the standard χ^2 test and our asymptotic ilr test. The largest differences occur for a small sample size in combination with a small minor allele frequency. The χ^2 and bootstrap tests are close to rejecting HWE, whereas the other tests do not. The asymptotic χ^2 and Z_{ilr} tests may not be applicable under these conditions, and probably the exact tests are to be preferred for this data set. We note that the asymptotic ilr based test (Z_{ilr}) seems again to be more conservative with extreme allele frequencies. The confidence intervals for y_2 and θ for the same data sets are shown in Table 3.

Sample	n	MAF	$\hat{y}_2$	CI(y_2)	$\hat{\theta}$	CI(θ)
MN	1000	0.458	-0.541	(-0.643, -0.440)	-0.024	(-0.126, 0.077)
MN	1148	0.058	0.007	(-0.271, 0.285)	-0.573	(-0.851, -0.295)
AG	28	0.482	-0.451	(-1.058, 0.156)	-0.115	(-0.722, 0.492)
AG	28	0.089	0.400	(-0.833, 1.634)	-0.966	(-2.199, 0.267)

Table 3: Point estimates and confidence intervals for y_2 and θ using four data sets.

The negative estimates $\hat{\theta}$ indicate that there is some degree of heterozygote dearth in all 4 samples, though the deviation is only significant in the second MN sample, where the value 0 is outside the confidence interval for θ.

5 Conclusions

We have obtained the asymptotic distribution of the ilr coordinates of a 3-way composition by applying the delta method. This has allowed us to construct a parametric test for HWE based on the second ilr coordinate. For intermediate allele frequencies the test is similar to the ordinary χ^2 test, whereas for extreme allele frequencies it seems more conservative. A drawback of the test is that an adjustment for zero genotype counts is needed. In the simulations, we used Jeffreys' 1946 estimator for the genotype frequencies which sums 0.5 to all genotype counts (Graffelman and Egozcue, 2011) in order to avoid problems with zeros. A more sophisticated approach for treating the zeros (Fry et al., 2000; Martín-Fernández and Thió-Henestrosa, 2006) could be used. An advantage of the statistics $\hat{\theta}$ and Z_{ilr} over the standard chi-square statistic is that their sign is informative, and directly indicates if the sample is characterized by a lack or an excess of heterozygotes.

6 Software

The various statistical tests for Hardy-Weinberg equilibrium discussed in this paper are available in the R-package `HardyWeinberg` (Graffelman and Morales-Camarena, 2008).

Acknowledgments

This study was supported by grants SEC2003-04476 and CODA-RSS MTM2009-13272 of the Spanish Ministry of Education and Science.

References

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Emigh, T. H. (1980). A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics 36*, 627–642.

Fry, J. M., T. R. L. Fry, and K. R. Mclaren (2000). Compositional data analysis and zeros in micro data. *Applied Economics 32*, 953–959.

Graffelman, J. (2010). The number of markers in the hapmap project: some notes on chi-square and exact tests for Hardy-Weinberg equilibrium. *The American Journal of Human Genetics 86*, 813–818.

Graffelman, J. and J. J. Egozcue (2011). Hardy-Weinberg equilibrium: a non-parametric compositional approach. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 207–215. John Wiley & Sons, Ltd.

Graffelman, J. and J. Morales-Camarena (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity 65*(2), 77–84. DOI: 10.1159/000108939.

Jeffreys, H. (1946, September). An Invariant Form for the Prior Probability in Estimation Problems. *Royal Society of London Proceedings Series A 186*, 453–461.

Martín-Fernández, J. A. and S. Thió-Henestrosa (2006). Rounded zeros: some practical aspects for compositional data. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn (Eds.), *Compositional Data Analysis in the Geosciences: from Theory to Practice.*, Volume 160, pp. 191–201. Geological Society, London, Special Publication, 264: Elsevier Science Publishers. North-Holland.

Mourant, A. E., A. C. Kopeć, and K. Domaniewska-Sobczak (1976). *The Distribution of the Human Blood Groups and other Polymorphisms* (Second ed.). London: Oxford University Press.

Olson, J. M. and M. Foley (1996). Testing for homogeneity of Hardy-Weinberg disequilibrium using data sampled from several populations. *Biometrics 52*(3), 971–979.

Wasserman, L. A. (2010). *All of Statistics*. Pittsburgh: Springer.

Weir, B. S. (1996). *Genetic Data Analysis II*. Massachusetts: Sinauer Associates.