Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 22

Disequilibrium Coefficient: A Bayesian Perspective

Helena Brentani* Eduardo Y. Nakano† Camila B. Martins‡
Rafael Izbicki** Carlos Alberto Pereira††

*University of São Paulo, helena.brentani@gmail.com
†University of Brasilia, eynakano@gmail.com
‡University of São Paulo, cabertinim@gmail.com
**University of São Paulo, rafaelizbicki@gmail.com
††University of São Paulo, cadebp@gmail.com

Disequilibrium Coefficient: A Bayesian Perspective*

Helena Brentani, Eduardo Y. Nakano, Camila B. Martins, Rafael Izbicki, and Carlos Alberto Pereira

Abstract

Hardy-Weinberg Equilibrium (HWE) is an important genetic property that populations should have whenever they are not observing adverse situations as complete lack of panmixia, excess of mutations, excess of selection pressure, etc. HWE for decades has been evaluated; both frequentist and Bayesian methods are in use today. While historically the HWE formula was developed to examine the transmission of alleles in a population from one generation to the next, use of HWE concepts has expanded in human diseases studies to detect genotyping error and disease susceptibility (association); Ryckman and Williams (2008). Most analyses focus on trying to answer the question of whether a population is in HWE. They do not try to quantify how far from the equilibrium the population is. In this paper, we propose the use of a simple disequilibrium coefficient to a locus with two alleles. Based on the posterior density of this disequilibrium coefficient, we show how one can conduct a Bayesian analysis to verify how far from HWE a population is. There are other coefficients introduced in the literature and the advantage of the one introduced in this paper is the fact that, just like the standard correlation coefficients, its range is bounded and it is symmetric around zero (equilibrium) when comparing the positive and the negative values. To test the hypothesis of equilibrium, we use a simple Bayesian significance test, the Full Bayesian Significance Test (FBST); see Pereira, Stern and Wechsler (2008) for a complete review. The disequilibrium coefficient proposed provides an easy and efficient way to make the analyses, especially if one uses Bayesian statistics. A routine in R programs (R Development Core Team, 2009) that implements the calculations is provided for the readers.

KEYWORDS: Bayesian methods, FBST, Hardy-Weinberg equilibrium

---

*Helena Brentani, INPD, LIM23, Institute of Psychiatry, University of São Paulo. Eduardo Y. Nakano, Statistics Department, University of Brasilia. Camila B. Martins, Institute of Mathematics and Statistics, University of São Paulo. Rafael Izbicki, Institute of Mathematics and Statistics, University of São Paulo. Carlos Alberto Pereira, Institute of Mathematics and Statistics, University of São Paulo. Carlos Pereira would like to dedicate this paper to the late Professor Oswaldo Frota Pessoa, his genetics master, who was a pioneer in mastering Genetics and Psychiatry Practice to important Brazilian scientists. All the authors have been supported by three of the Brazilian Research Institutions: FAPESP, CAPES and CNPq.

1. INTRODUCTION

HWE, also known as principle or law of Hardy-Weinberg, plays a key role in population genetics. It was proposed, independently, by Hardy (1908) and Weinberg, (1908). By intuition, one could wrongly guess that rare alleles became rarer as time passes. However, the Hardy-Weinberg principle shows that in a Mendelian population, under certain restrictions, allele frequencies will be constant throughout generations. HWE is based on the independent assortment of alleles in a population of sufficient size.

Suppose there are two alleles in a locus, *A* and *a*. The Hardy-Weinberg principle states that, for this locus, equilibrium holds if there exists a parameter $0<p<1$ such that genotype *AA* frequency is p^2, *Aa* frequency is $2p(1-p)$ and *aa* frequency is $(1-p)^2$. p can be interpreted as the frequency of allele *A*. Clearly, the sum of these three genotypic proportions is equal to 1, as should the two allelic frequencies. Having reached equilibrium, which takes only one generation as proved by the two mathematicians, the population stays in this way. Hence one could say that if a population is not in genetic equilibrium, some adverse situations should be occasionally present.

The null hypothesis for HWE testing is that no significant lack of equilibrium exists in the population, for the genotype in study. Both classical and Bayesian methods have been proposed in order to test this hypothesis. These methods include likelihood ratio tests, chi-square test, Bayes factors, FBST and other procedures. See, for instance, Emingh (1980); Pereira and Rogatko (1984); Lindley (1988); Hernández and Weir (1989); Singer et al. (1991); Chow and Fong (1992); Weir (1996); Ayres and Balding (1998); Shoemaker and Weir (1998); Pereira and Stern (1999); Rogatko et al. (2000); Montoya-Delgado et al. (2001), Pereira et al. (2006), Lauretto et al. (2009) and Wakefield (2010). It is important to consider that both classical and Bayesian tests have their weakness and strengthens, favoring one against the other. We are not going to enter in the choice between the two perspectives because we see no simple way to use classical statistics to handle the coefficient introduced here. Hence, the perspective of the present paper is Bayesian.

Assumptions for the Hardy-Weinberg law to be applicable are: infinitely large population size or a population size large enough that random fluctuations in allele and genotype frequencies are small, mating is random, no mutation, no considerable migration or emigration, there is no natural selection and the numbers of males and females should not differ considerably. Departure from HWE can occur in a population if it is either under current selective pressures or has recently undergone a drastic increase or decrease in size or its mating process is far from random or population admixture or stratification exists.

The goal of population association studies is to identify patterns of polymorphisms that vary systematically between individuals with different disease states and could represent effects of risk-enhancing or protective alleles. Recently, much emphasis has been given to tests that aim to evaluate if a population is in HWE. While historically the HWE formula was developed to examine the transmission of alleles in a population from one generation to the next, usage of HWE concepts has expanded in human diseases studies to detect genotyping error and disease susceptibility (association): see Ryckman and Williams (2008) for interesting discussion. Understanding how HWE testing can be used in the process of disease gene discovery or genotyping error is becoming increasingly important as the number of SNPs in studies increases to the hundreds.

Usually the answer from an HWE test is based only on the accept/reject rule. Tests for HWE have been used to detect genotyping error, but those tests have low power to detect this kind of error at common allele frequencies (Hosking et al, 2004) unless the sample size is very large. The procedure of mechanistically discarding markers that are out of HWE may cause investigators to miss important biological signals in their analyses. Usually, we are not interested in just evaluating if a population is in HWE, but also in trying to quantify how far from the equilibrium it is. A recently published paper by Attia et al. (2010) proposes to look at measures that quantify the degree of deviation from HWE in order to detect genotyping errors. This procedure should be used predominantly as screening rather than testing for genotyping errors. They proposed the use of three measures – the "inbreeding coefficient" (f), Weir (1996), the "disequilibrium parameter" (D), Hernandez and Weir (1989), and the alpha parameter (α), Lindley (1998).

In this paper we propose a simple unique coefficient that measures how far from HWE a population is. As it is going to be seen, this coefficient has the advantage of being easy to interpret: one can look at the original parametric space, as in Figure 2, so that it can be said what a "small" or a "large" coefficient means. Moreover, it varies from -1 to 1, and is 0 if, and only if, HWE holds. We also show how to conduct a complete Bayesian analysis based on it. For this purpose, Bayesian analysis is much simpler to implement than classical methods (mainly for small samples), and conclusions are easier to be drawn. The use of this new coefficient is shown to be very useful in the analysis of polymorphisms that are out of HWE in cases, but not in controls. Such SNPs contribute to the disease: the larger the absolute value of the coefficient is, the more probable is SNP association with the disease. Our coefficient is a modification of that introduced by Pereira and Rogatko (1984). This modification is based in the same argument Yule (1912) proposed for his Homogeneity Coefficient of Association. He was looking for a function of the odds ratio and we are looking for a similar function

DOI: 10.2202/1544-6115.1636

of that of Pereira and Rogatko (1984). By calculating the posterior density coefficient we can simply define credible intervals and perform the FBST – Full Bayesian Significance Test – of Pereira and Stern (1999). For a FBST review see Pereira, Stern and Wechsler (2008). In the appendix we discuss a comparison of the coefficient discussed here with two important alternatives already discussed in the literature. Also in the Appendix we present a sensitivity study for the prior used in this paper.

2. MATERIAL AND METHODS

2a. Model and Hypothesis

Hardy-Weinberg law shows that in a Mendelian population, under certain restrictions, allele frequencies will be constant through generations; alternatively, been in HWE means that the relation of the proportions of the genotypes in a specific locus exhibit a special association. In order to see if a population is in HWE one takes a sample of size $n=n_1+n_2+n_3$ and observe n_1, n_2 and n_3, the absolute frequencies of the genotypes *AA*, *Aa* and *aa*, respectively. Let π_1, π_2 and π_3 be the population frequencies of this genotypes, with $\pi_1+\pi_2+\pi_3=1$ and $\pi_i \geq 0$, $i=1, 2, 3$. If we consider genotypes from different individuals to be statistically independent, likelihood function can be expressed as

$$L(\pi_1,\pi_2,\pi_3 \mid n_1,n_2,n_3) \propto \pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3} \tag{1}$$

with $\propto$ denoting proportionality. Note that the parametric space is

$$\Theta = \{(\pi_1,\pi_2,\pi_3) : \pi_1 \geq 0, \pi_2 \geq 0, \pi_3 \geq 0, \pi_1+\pi_2+\pi_3 = 1\}. \tag{2}$$

HWE holds if, and only if, there exists $0 < p < 1$ such that

$$\pi_1 = p^2,\ \pi_2 = 2p(1\text{-}p),\ \&\ \pi_3 = (1\text{-}p)^2.$$

Hence, the HWE null hypothesis is $H : \theta \in \Theta_H$ – for $\theta = (\pi_1,\pi_2,\pi_3)$ and

$$\Theta_H = \left\{(\pi_1,\pi_2,\pi_3) : \exists\ p \in [0,1] \therefore [\pi_1 = p^2], [\pi_2 = 2p(1-p)], [\pi_3 = (1-p)^2]\right\} \subset \Theta. \tag{3}$$

2b. Bayesian Modeling

Following Pereira and Rogatko (1984), a Dirichlet distribution is taken as a prior for θ. If the chosen prior is a Dirichlet distribution of order 3, with positive real parameters a_1, a_2, and a_3, denoted by $D(a_1,a_2,a_3)$, then the posterior distribution happens to be $D(A_1,A_2,A_3)$ with $A_i=a_i+n_i$, $i=1,2,3$ – the Dirichlet family of distributions of order 3 is a conjugate family for the trinomial sampling distribution. Note that the likelihood function is obtained from the observed data $d=(n_1,n_2,n_3)$ and the fact that the sampling distribution is proportional to a trinomial with parameter vector θ.

The posterior density function of our posterior distribution is given by

$$f(\theta \mid d) = \Gamma(n+a)\prod_{i=1}^{3}\frac{\pi_i^{(n_i+a_i-1)}}{\Gamma(n_i+a_i)} \quad \text{for } \theta \in \Theta$$

Denoting $n=n_1+n_2+n_3$, $a=a_1+a_2+a_3$ and $A=A_1+A_2+A_3=n+a$, we obtain the posterior mean vector and the co-variance matrix as follows:

$$\hat{\theta} = E(\theta \mid d) = \left(\tfrac{A_1}{A}, \tfrac{A_2}{A}, \tfrac{A_3}{A}\right) = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right) \quad \&$$

$$\Sigma = \frac{1}{A+1}\left(\begin{pmatrix} \hat{\theta}_1 & 0 & 0 \\ 0 & \hat{\theta}_2 & 0 \\ 0 & 0 & \hat{\theta}_3 \end{pmatrix} - \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \end{pmatrix}\left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right)\right) \tag{4}$$

We have presented a complete description of the posterior distribution, the complete tool for a Bayesian analysis. A detailed review of the multinomial data with Dirichlet conjugate prior analysis can be found in Pereira and Stern (2008): important properties of these distributions are discussed in detail.

The disequilibrium coefficient, denoted by λ, is now defined and will be explained and motivated in the next sections. In fact, this coefficient is a modification of the one presented by Pereira and Rogatko (1984). It is based on the correlation coefficient proposed by Yule (1912) to measure association in a 2 by 2 contingency table. Since we have the complete specification of the parameter θ and the fact that λ is simply a function of θ, we may obtain also the complete specification of the distribution of λ.

DOI: 10.2202/1544-6115.1636

Definition: The Hardy-Weinberg disequilibrium coefficient is defined as follows:

$$\lambda = \frac{\sqrt{\pi_1 \pi_3} - \frac{1}{2}\pi_2}{\sqrt{\pi_1 \pi_3} + \frac{1}{2}\pi_2} \tag{5}$$

This coefficient is a number between -1 and 1, just as the Pearson correlation coefficient, and it is zero if, and only if, the population is in HWE. The next section is the motivation for the use of this coefficient.

3. MOTIVATION

The most celebrated parameter for measuring association in a 2 by 2 contingency table is the cross product ratio (CPR). Such a contingency table is a twofold table in which the columns indicate a binary classification – success and failure of an event – and the lines an additional binary classification. Denoting success and failure by *g* and *G* for columns and *t* and *T* for lines, Table 1 illustrates the notation used for observed frequencies and for the parameters.

Table 1: Frequency data, *n*, (parameters, π) in a multinomial classification.

	g	***G***
t	n_1 (p_1)	n_{12} (p_{12})
T	n_{21} (p_{21})	n_3 (p_3)

The cross product ratio associated to this table is the following parameter:

$$\psi = \frac{\pi_1 \pi_3}{\pi_{12} \pi_{21}} \tag{6}$$

Under the Dirichlet Bayesian analysis as described in the former section, the posterior of ψ is the product of two independent beta distributions of second type (beta prime) random variables – if π has a beta distribution with parameter (a;b), then $\theta = \pi/(1-\pi)$, the odds, is said to have a second type beta distribution with parameter (a;b), Rao (1965). See Basu and Pereira (1982), Irony et al. (2000) and Pereira and Stern (2008) for complete details. Considering, as before, the prior parameter vector (a_1,a_{12},a_{21},a_3), data d=(n_1,n_{12},n_{21},n_3) and A_i=a_i+n_i, the posterior mean and variance of ψ can be expressed as follows:

$$E\{\psi \mid d\} = \frac{A_1 A_3}{(A_{12}-1)(A_{21}-1)}$$
$$\& \qquad (7)$$
$$V\{\psi \mid d\} = \frac{A_1(A_1 + A_{12} - 1)}{(A_{12}-1)(A_{12}-2)} \times \frac{A_3(A_3 + A_{21} - 1)}{(A_{21}-1)(A_{21}-2)}$$

Clearly, this holds only if $A_{12} > 2$ and $A_{21} > 2$.

Yule (1912) understood that although ψ could be considered an association coefficient, it is unbounded and unbalanced: negative association occurs if $\psi \in (0;1)$, independence if $\psi = 0$, and positive association if $\psi \in (1;\infty)$. In a clinical application of the CPR it is important to understand that if results indicate 'negative association' they will be represented by a value within the small interval: from 0 to 1. On the other hand, if the indication is 'positive association' the representation is a number ranging from 1 to infinite, a very large interval. Taking this into consideration, it is clear that comparing positive associations with negative ones is not a simple visual task. In other words, when one says that a negative association is represented by CPR=0.2 the positive counterpart will be CPR=5. The ideal would be using something like the Pearson Correlation Coefficient which goes from (-1;1). It is not adequate to use Pearson correlation in a contingency table environment since it was defined for continuous random quantities and we are dealing with dichotomous ones. After some important consideration Yule finally defined what he called a stable correlation coefficient for two-fold contingency tables:

$$\dot{\lambda} = \frac{\sqrt{\pi_1 \pi_3} - \sqrt{\pi_{12} \pi_{21}}}{\sqrt{\pi_1 \pi_3} + \sqrt{\pi_{12} \pi_{21}}} = \frac{\sqrt{\psi} - 1}{\sqrt{\psi} + 1} \qquad (8)$$

This is a bounded and balanced coefficient: negative association occurs if $\dot{\lambda} \in (-1;0)$, independence if $\dot{\lambda} = 0$, and positive association if $\dot{\lambda} \in (0,1)$. Positive and negative association points vary in bounded intervals of the same length.

From now on the statistical problem is to make inferences about this interesting parameter. Having completely specified the posterior distribution for the original parameter vector, θ, and consequently for the two independent factors of the cross product ratio, Bayesians consider that the task of estimating $\dot{\lambda}$ is no longer a problem. With today's computing power, using the R program (R Development Core Team, 2009) and a simulation procedure, one can fully obtain numerically the posterior density of $\dot{\lambda}$. A simple program in R language is presented in http://www.ime.usp.br/~cpereira/programs/association.r.

DOI: 10.2202/1544-6115.1636

A classical statistician could have difficulty to perform exact inferences about the Yule coefficient. The problem is to find a statistic whose distribution depends directly on λ: This could be a difficult chore.

In order to understand the behavior of Yule's measure of association, we consider here examples showing the role of sample sizes and the direction of association: negative, positive and independence. To start looking for solutions one should specify the prior distribution. The specification is on the original parameter θ: for non informative prior we choose to use the Jeffreys' prior for multinomial inferences (Jeffreys, 1931), $a_1=a_{12}=a_{21}=a_3=½$. For a multinomial of order four, this is interesting since the precision represented by the prior is the same as in the uniform distribution in the interval (0,1): the sum of the prior parameters is two, just like in the uniform or Beta(1,1) distribution, taking values in the interval (0,1).

The data described by Table 2 are from Hospital das Clínicas, São Paulo. They refer to 93 children with hepatic obstruction which can take two forms: intra-hepatic or extra-hepatic. To help to discriminate between the states, intra and extra, two clinical tests were available; both tests having dichotomous alternative responses, Positive or Negative. The two possible outcomes for the first test are ε+ and ε- and for the second test, E+ and E-. It is important to note that the conduct will be different whether the patient has intra or extra hepatic obstruction. The questions to be answered are: i) Having both results on hand, which clinical conduct should be adopted? ii) Should the two tests be taken? As demonstrated below the two tests need to be undertaken because the symptoms of the two forms are the same. On the other hand the first (second) test favors diagnosis of the intra-hepatic (extra-hepatic) form.

Considering the Jeffreys' prior for the two contingency tables, one obtains the posterior Dirichlet densities, of order four, with the following parameter vectors:

(5.5;9.5;28.5;6.5) for extra-hepatic and
(28.5;12.5;1.5;4.5) for Intra-hepatic.

Figure 1 presents the prior density and the posterior densities of λ for both disease states: intra- and extra-hepatic. These posterior densities are all Bayesian and should produce inferences about this contingency coefficient parameter. The Bayesian estimates, the posterior means, the Bayesian interval estimates, credible intervals, and the Bayesian significance indices, e-values, are listed below. Introduced by Pereira and Stern (1999) and reviewed in Pereira, Stern and Wechsler (2008), the e-values are the Bayesian alternative for p-values. To compute the e-values we integrate the posterior density over the set of all parameter points that have density smaller than the density value obtained for the

sharp hypothesis, $\dot{\lambda} = 0$. The e-value is the tail of the posterior obtained from the hypothesis value. Note that none of these statistical Bayesian end-points depend on asymptotic results to be evaluated.

1. Posterior means: $E(\dot{\lambda} \mid d) = \begin{cases} -.4750 & \text{for Extra – hepatic} \\ .4723 & \text{for Intra – hepatic} \end{cases}$
2. 95% credible set: $\begin{cases} (-.72;-.21) & \text{for Extra – hepatic} \\ (.08;.85) & \text{for Intra – hepatic} \end{cases}$
3. Bayesian significance index (e-value): $\begin{cases} .13\% & \text{for Extra – hepatic} \\ 1.96\% & \text{for Intra – hepatic} \end{cases}$
4. Chi-squared significance index (p-value): $\begin{cases} .12\% & \text{for Extra – hepatic} \\ 1.81\% & \text{for Intra – hepatic} \end{cases}$

Since the significance test consists of the judgment of the significance indices, we believe both numbers are small and therefore independence between the two tests should be rejected in both groups of studies: Bayesian and classical significance indices produce equivalent conclusions. Here, the p-values are obtained by the standard chi-squared statistics for independence. The e-values are the tail area of the posterior densities that is defined from the value of this density at point $\dot{\lambda} = 0$. That is, the tail area is the set $\{\dot{\lambda} : f(\dot{\lambda} \mid d) < f(0 \mid d)\}$: here f is the posterior density of $\dot{\lambda}$.

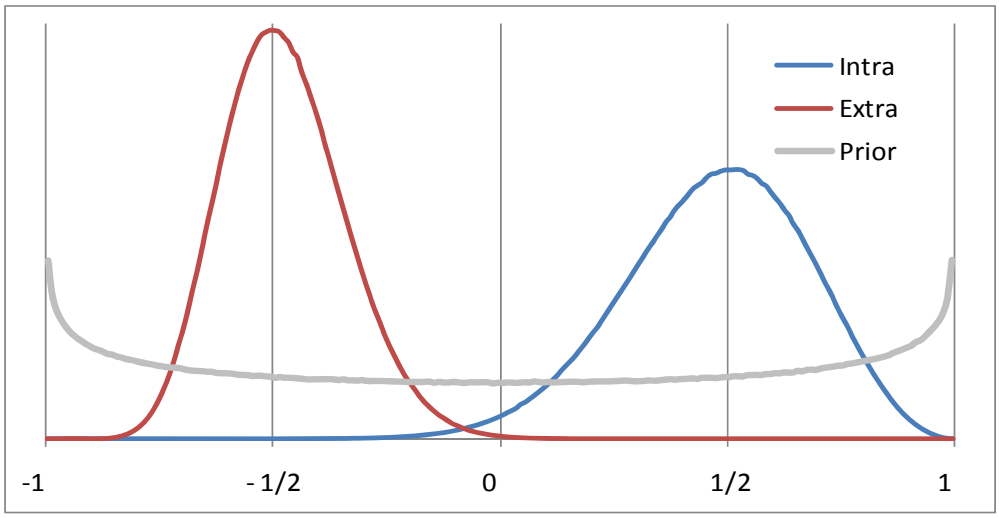
Table 2: Outcomes of the diagnostic tests for 93 children.

	Extra-hepatic			**Intra-hepatic**			**Total**
Results	E^+	E^-	**Sum**	E^+	E^-	**Sum**	
***e*+**	**5**	**9**	**14**	**28**	**12**	**40**	**54**
***e*-**	**28**	**6**	**34**	**1**	**4**	**5**	**39**
Sum	**33**	**15**	**48**	**29**	**16**	**45**	**93**

Using this example it is clear that the interpretation of the coefficient here proposed is simple with an intuitive visual task.

DOI: 10.2202/1544-6115.1636

Figure 1: Jeffreys' prior and posterior densities for extra- and intra-hepatic groups.



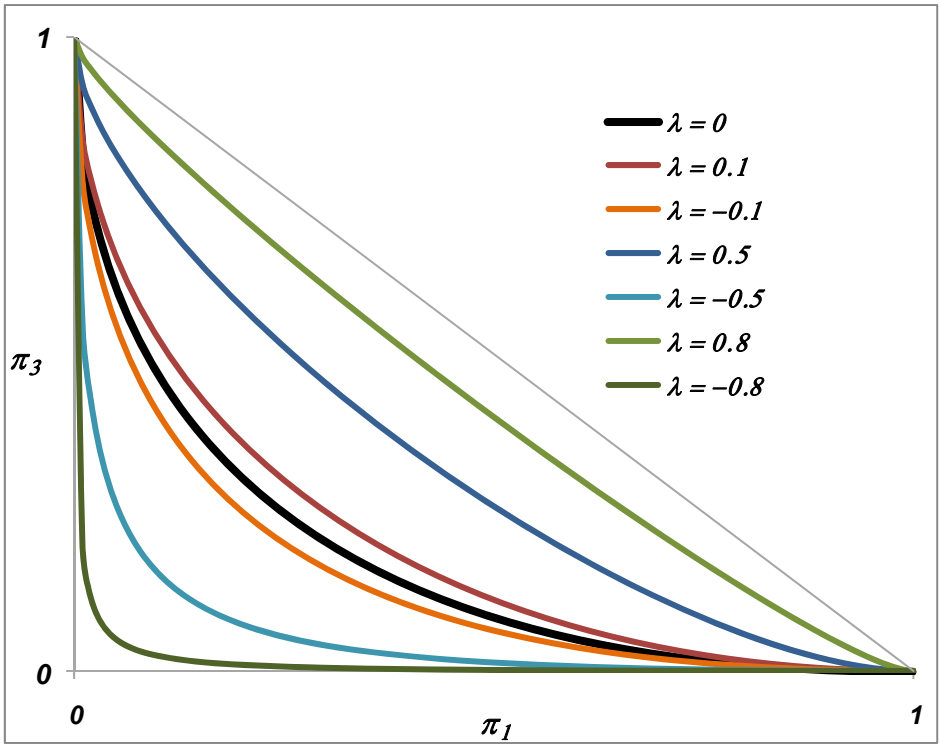
To end this section we must discuss the choice of the prior distribution in situations as contingency tables for independence studies. A prior sensitivity analysis for this kind of problem is based on looking at different priors in order to see their influence in the final results. We follow the same view of Campos and Benavoli (2011) for considering different kinds of weak prior information.

Our original parameter space is the four categories multinomial parameter that is a simplex of three dimensions. The natural conjugate class of priors is the Dirichlet class with positive real parameters. The choice of these parameters, (a_1,a_{12},a_{21},a_3), could be viewed, in a metaphoric way, as our prior sample with size $n_0=a_1+a_{12}+a_{21}+a_3$. The sample size $n=x_1+x_{12}+x_{21}+x_3$ usually should be larger than the prior sample size in order to have more information from data than from prior. The larger the ratio n over n_0, the less informative is the prior. Also, by choosing the prior parameters one may pay attention to the choice of each of the a's in relation to the observed x's. For instance, in the Intra-hepatic sample, $x_{21}=1$ and in the Jeffreys' prior a_{21} = ½. Table A1 in the Appendix presents the comparison of the posterior end-points for different weak priors, highlighting the fact that radical posterior differences could occur only for highly informative priors – large n_0's –. Another important fact is that there is a closed relation between the CPR and the chi-squared statistics and this is the reason to obtain closed values of the e-value and the p-value, as seen in the examples. We cannot forget that there is a one to one correspondence between CPR, ψ, and the association coefficient $\dot{\lambda}$. This implies that this known coefficient ψ produces the same inferences as does $\dot{\lambda}$. Section 4 discusses the disequilibrium coefficient λ defined by (5).

4. THE DISEQUILIBRIUM COEFFICIENT

In Table 1 replacing both the first column (g,G) and the first line (t,T) by (A,D) and considering them as the alternative alleles of the mother and the father, respectively, the cells can be viewed as genotype frequencies that are to be observed. Testing independence in such a table would correspond to test the Hardy-Weinberg equilibrium. However, in genetics one observes the sum of the two possible heterozygote frequencies, namely $n_2=n_{12}+n_{21}$. Hence, one can only count the frequencies of the three genotypes: homozygous, for both alleles, and heterozygous. See Ryckman and Williams (2008) for a similar introduction to the association view discussed here. It is impossible to discriminate the heterozygous alleles: one cannot identify the parent that transmits a specific allele. The corresponding heterozygous parameter would be $\pi_2=\pi_{12}+\pi_{21}$. This is a special kind of censured data. In order to reproduce Yule's association coefficient , we estimate each heterozygote parameter by considering ½ of π_2. With this arrangement we only repeat the formula (8) to obtain expression (5). Hence, λ is $\dot{\lambda}$ with $\frac{\pi_2}{2}$ replacing both π_{12} and π_{21}.

Figure 2: Example of Disequilibrium Curves.



DOI: 10.2202/1544-6115.1636

Figure 2 illustrates alternative curves defined by alternative values of λ. In this way, one may understand that by using this coefficient it can be said not only whether HWE holds or not, but also how far from the equilibrium the population is. It is relevant to recognize that the curve for λ=0 is the equilibrium curve.

Alternative coefficients to study equilibrium/disequilibrium of Hardy-Weinberg are the ones presented by Pereira and Rogatko (1984), which is a one-to-one function of λ, the inbreeding coefficient f and the disequilibrium D, both discussed by Weir (1996).

The inbreeding coefficient f is the constant that in general relates the genotype frequencies to the allelic frequencies in the following way:

$$\begin{cases} \pi_1 = p^2 + p(1-p)f \\ \pi_2 = 2p(1-p)(1-f) \\ \pi_3 = (1-p)^2 + p(1-p)f \end{cases}$$

Analogous to f, D is a constant that relates allele and genotyping frequencies as in the following equations:

$$\begin{cases} \pi_1 = p^2 + D \\ \pi_2 = 2p(1-p) - 2D \\ \pi_3 = (1-p)^2 + D \end{cases}$$

An important fact is that both f and D have restricted support since their range of possibilities depends strongly on the value of the allelic frequency. $D=p(1-p)f$ ranges from $Max\left\{-p^2;-(1-p)^2\right\}$ to $p(1-p)$ and f clearly goes from $Max\left\{-\frac{p}{1-p};-\frac{1-p}{p}\right\}$ to one. These facts imply that both parameterizations from $(\pi_1;\pi_3)$ to $(p;D)$ or $(p;f)$ are not variation independent. This violates the principle of elimination of nuisance parameters as described by Basu (1977), whenever analyses are based either in D or f. On the other hand, our coefficient λ is variation independent from any of the original parameters and goes from -1 to 1 as the correlation coefficient of two random quantities. This is illustrated by Figure 2. We also illustrate in the Appendix the strong dependence of the range of D with the values of the original parameters, see Figure A1. The relationship between λ and D is illustrated by Figures A2 and A3.

The functions relating the three coefficients are as follows:

$$\lambda = \frac{\sqrt{(\frac{p}{1-p}+f)(\frac{1-p}{p}+f)}-(1-f)}{\sqrt{(\frac{p}{1-p}+f)(\frac{1-p}{p}+f)}+(1-f)} = \frac{\sqrt{(p^2+D)((1-p)^2+D)}-p(1-p)+D}{\sqrt{(p^2+D)((1-p)^2+D)}+p(1-p)-D}$$

To compare the performance of these three coefficients, the Appendix illustrates Bayesian statistical analysis for each one of them, for all the examples presented in Section 5. We have two kinds of sensitivity analyses, the first is related to the change of prior parameters and the second is about the 3 coefficients λ, f and D. Our objective is to show that the coefficient λ is more sensitive to sample and prior changes.

In order to described the usefulness of the disequilibrium coefficient λ,the next section illustrate cases of polymorphisms showing how flexible our analysis can be.

5. RESULTS

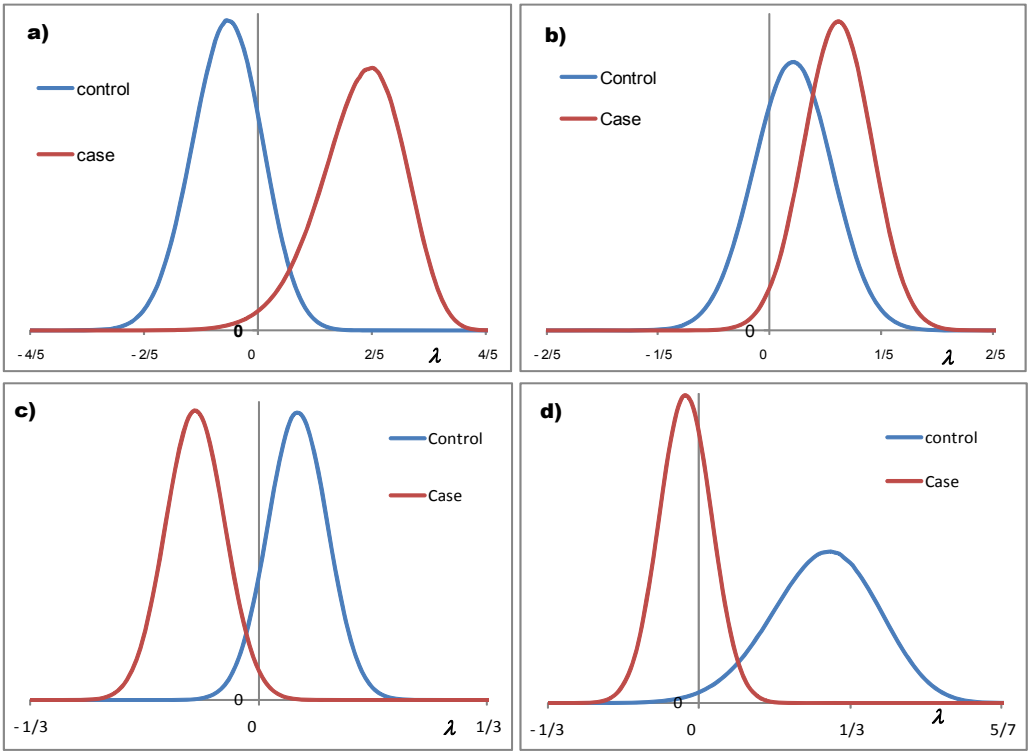
Using data from published studies we calculated the disequilibrium coefficient and plotted the posterior density curves. To date, APOE-ε4 is the only recognized risk factor for Alzheimer Disease (AD). It is well established that whenever one carries an ε4 allele, his AD risk increases in an allele dose dependent manner. This situation is associated with earlier age of onset of the disease. However, possessing an ε4 allele does not guarantee disease development. Previous reports have suggested that additional factors within the APOE locus, in other genes and from the environment might also modulate risk. So there are several available data from papers studying other polymorphisms in the APOE locus as well as other genes concerning function of the APOE genotype. Lambert et al. (2002), with the objective to determine whether the effects of APOE promoter polymorphism on AD are independent of the APOE-ε4 allele, studied -491 $A \rightarrow T$ and -219 $G \rightarrow T$ SNPs. These APOE polymorphisms increase risk for AD in addition to risk associated with the APOE-ε4 allele. We decided to use such important cases as our examples. The first example is from a 491 $A \rightarrow T$ in APOE promoter genotyped in cases and controls from the spanish population, (AD-121-1). Table 3 presents all statistical results and looking at Figure 3a one should see that controls are in HWE. As expected, the density curve for the control group includes the zero with high density (tails having large areas). On the other hand, the density curve for the disease state seems to be out of HWE since the zero has low frequency (tails having small areas).

DOI: 10.2202/1544-6115.1636

The second and third examples are a 219 $G \rightarrow T$ in APOE promoter: b) from a United Kingdom population (AD1-121-2) and c) from a French population (AD-121-3).

The last example is from cases and controls of APOE-ε4 carriers described in Bhojak et al. (2000): The -174 $G \rightarrow C$ SNP in IL6 gene is the object of Figure 3d. Table 3 presents the estimates – Bayesian and Maximum Likelihood Estimates, MLE –, Credibility 95% intervals and e- and p-values for all examples of Figure 3.

Figure 3: Posterior Density distributions of cases and controls genotyped for SNPs in APOE gene (a, b, c) and IL6 gene (d).



It is important to notice that the e-values are similar to the standard p-values in all cases. However, e-values are more illustrative since they were obtained here from the univariate curves. The p-values, on the other hand, are obtained from the trinomial distribution – dimension 2 –, using the chi-squared with the standard observed/expected computation. Recall that e-values are the areas of the tails defined from the point λ=0. If $f(\lambda)$ is the posterior density , we look for a point x of λ such that $f(x)=f(0)$, the same density of the hypothesis. Compute the probability 1–e of the set $(0;x)$ or $(x;0)$ according to x being positive

or negative, respectively: the evidence e is the complement of this probability, the tail area. The tails have large (small) areas whenever zero has high (low) densities. As can be seen from Figures 3a, 3b and 3c, the two (case and control) posterior modes could be in the same side of zero or in its opposite sides. Note that in the first three examples the modes of controls are closer to zero than the cases are.

The advantage of this new disequilibrium coefficient is that one can understand the disequilibrium status by looking at the figure. Evidence is higher against HWE favoring disequilibrium in cases of the AD121(1) SNP than in cases of AD121(3) SNP. As has been reviewed in some studies, it is perfectly normal to accept that we are interested in searching for SNPs out of HWE in cases but not in controls. Our Bayesian analysis can not only say that the SNP is out of HWE but can also inform how far from the equilibrium it is. Surprisingly, even with a smaller sample AD121(1) SNP is farther from HWE than AD121(3) SNP .The last example, Figure 3d, describes an unexpected result since we do not reject HWE in cases with a large sample and reject it in controls with smaller samples. Situations like this suggest that we should abandon such data or perform re-sequence. This is a clear indication of genotyping error.

In the original study of SNPs (Lambert et al., 2002) looking to the promoter region of APOE in AD cases, the auhtors found association between AD and -491 $A \rightarrow T$ SNP. Using the λ coefficient, we can indicate that the Spanish sample is responsible for the Odds Ratio heterogeneity discussed in that paper. For instance, considering the observed frequencies *AA*, *AT* and *TT* as being 94, 18 and 4, for the cases, we would obtain λ=0.3 and an e-value of 2.2%. Consequently, since no other frequency genotyping data of the set of distribution rejects HWE, the Spanish sample is the source of the problem.

Table 3: Statistical end-points for evaluating disequilibrium.

		Fenotype			**λ estimate**		**95% Credible Interval for λ**		**e-value**	**p-value**
Study	**Sample**	***AA***	***AD***	***DD***	**Bayes**	**MLE**	**L.Inf**	**L.Sup**	**FBST**	χ^2
AD121(1)	**case**	4	18	94	0.357	0.366	0.066	0.649	0.016	0.018
	control	6	53	74	-0.112	-0.114	-0.352	0.128	0.361	0.362
AD121(2)	**case**	57	118	100	0.122	0.123	0.003	0.242	0.045	0.046
	control	58	97	48	0.042	0.042	-0.095	0.179	0.550	0.550
AD121(3)	**case**	120	361	194	-0.084	-0.084	-0.160	-0.007	0.032	0.032
	control	206	309	142	0.051	0.051	-0.026	0.128	0.198	0.197
AD108(4)	**case**	110	148	44	-0.031	-0.031	-0.149	0.088	0.611	0.611
	control	34	22	12	0.290	0.295	0.051	0.529	0.018	0.022

DOI: 10.2202/1544-6115.1636

The routine R program for the Disequilibrium Coefficient is presented in http://www.ime.usp.br/~cpereira/programs/dishw.r.

These 8 samples are used for a sensitivity study presented in the Appendix. We change priors and change coefficients for all samples. We consider the two alternative coefficients discussed by Weir (1996). In the authors' opinion, coefficient λ is the one with more sensitivity; it changes with prior and small sample modifications. Table A2 presents the posterior end-points results for λ, our disequilibrium coefficient, for f, the inbreeding coefficient, and for D, the Weir's disequilibrium coefficient, using different priors. Looking at their variation as function of the prior variation, we note that range attained by D is very small and λ has the largest range of values comparing the 3 coefficients.

6. DISCUSSION

In this article we discuss the importance of creating a disequilibrium coefficient so that one can measure discrepancy from HWE. There are other measures of disequilibrium like the ones discussed and used by Attia et al. (2010). The coefficient introduced here has the advantage to be simpler, bounded and balanced. It varies just like the correlation coefficient and may have a simple interpretation as illustrated by Figure 2.

Bayesian approach allows one to incorporate prior knowledge to the inference. Also there is no restriction related to sample size that can be sequentially obtained up to the time of the study. Furthermore, no use of asymptotic results is needed.

Considering that e- and p-values in our example hardly differ, why using e instead of p? The main argument is that the e-value is directly calculated from the posterior distribution of λ. Also, the stopping rule does not influence the posterior after the sample is obtained. The p-value on the other hand is calculated from a transformation of the sample space in a chi-squared variable. In this sense, the inference does depend on the sample space that changes with the stopping rule.

Looking at the literature of choosing genes associated with a disease, we understand that here one has a strong alternative candidate to perform such work. Recall that in Genome Wide Scans we have a very large number of SNPs to choose the elected ones. Our suggestion is to order the absolute value of the disequilibrium coefficient estimates, $|\lambda|$, of controls. Comparing to a threshold, defined by the FBST for the fixed observed sample, T_α - α being the significance level chosen; all the ones that are greater than this number are left out. From those remaining SNPs we order the $|\lambda|$ estimates for cases. Now, the ones smaller than the threshold are also left out. The remaining SNPs from this cut ranking process should be the ones of great relevance for the disease in study.

APPENDIX – A Sensitivity Study

Table A1 shows, for the association coefficient, how distinct priors influence the end points of the statistical analysis. As expected, the more informative the prior is, the more the posterior is influenced by it. We consider here only weak information priors in order to allow one to compare our analysis with any other frequentist statistical alternative. Note that our prior sample size here goes from 2 to 4 showing that for the cases of weak sample information (Intra-Hepatic sample), the changes can be drastic.

Table A1: End-points posterior parameters for alternative weak priors for $\dot{\lambda}$: Independence contingency table hepatic studies using the coefficient of association.

Group	Sample Frequencies				Dirichlet Prior	Posterior		
	e+/E+	e+/E-	e-/E+	e-/E-	Parameters	mean	sd	e-valor
extra-hepatic	5	9	28	6	(1/2;1/2;1/2;1/2)	-0.4750	0.1346	0.0013
					(1;1;1;1)	-0.4561	0.1329	0.0015
					(1/2 ,1;1;1/2)	-0.4886	0.1315	0.0007
					(1;1/2;1/2;1)	-0.4419	0.1359	0.0023
Intra-hepatic	28	12	1	4	(1/2;1/2;1/2;1/2)	0.4723	0.2009	0.0196
					(1;1;1;1)	0.4231	0.1895	0.0267
					(1/2;1;1;1/2)	0.3958	0.1974	0.0445
					(1;1/2;1/2;1)	0.4973	0.1923	0.0117

Table A2 illustrates the sensitivity of all three coefficients, λ, f and D, for small changes in the prior parameters. Here our prior sample size, n_0, goes from 1.5 to 4. We call attention for the fact that small changes in the prior sample do not affect the posterior analyses whenever the sample is large. However, if the sample is small, the prior helps to improve the information used in the final analysis. Study AD121(1) is the best example of this case. If we sum 1 to 4, for instance, 20% of the posterior sample frequency is coming from the prior. For Jeffreys' prior the prior frequency is only about 10% of the posterior frequency.

Figures A1, A2 and A3 provide some more insight about the relationship between λ, f and D, as discussed in the text.

DOI: 10.2202/1544-6115.1636

Table A2: End-points posterior parameters for alternative weak priors for λ: Polymorphism examples introduced in Section 5.

Study	Group	Sample			Prior	Posterior Mean			e-value		
		AA	*AD*	*DD*	Parameters	λ	f	D	λ	f	D
AD 121 (1)	case	4	18	94	(.5;.5;.5)	.3575	.2303	.0241	.0163	.0558	.0906
					(1;1;1)	.3724	.2462	.0266	.0087	.0386	.0686
					(1;2;1)	.3503	.2307	.0255	.0141	.0482	.0785
					(.5;1;.5)	.3459	.2224	.0235	.0205	.0618	.0965
	control	6	53	74	(.5;.5;.5)	-.1120	-.0719	-.0134	.3606	.3678	.3744
					(1;1;1)	-.0955	-.0621	-.0117	.4251	.4400	.4452
					(1;2;1)	-.1046	-.0686	-.0130	.3803	.3894	.3953
					(.5;1;.5)	-.1165	-.0752	-.0141	.3403	.3446	.3515
AD 121 (2)	case	57	118	100	(.5;.5;.5)	.1222	.1198	.0292	.0444	.0452	.0455
					(1;1;1;1)	.1235	.1211	.0295	.0416	.0423	.0427
					(1;2;1)	.1194	.1171	.0285	.0486	.0494	.0497
					(.5;1;.5)	.1202	.1178	.0287	.0480	.0488	.0492
	control	58	97	48	(.5;.5;.5)	.0419	.0418	.0104	.5488	.5486	.5488
					(1;1;1)	.0441	.0439	.0109	.5269	.5267	.5269
					(1;2;1)	.0390	.0389	.0097	.5746	.5744	.5746
					(.5;1;.5)	.0394	.0392	.0098	.5729	.5726	.5728
AD 121 (3)	case	120	361	194	(.5;.5;.5)	-.0837	-.0824	-.0203	.0312	.0309	.0311
					(1;1;1)	-.0827	-.0814	-.0201	.0331	.0328	.0330
					(1;2;1)	-.0841	-.0828	-.0204	.0301	.0298	.0301
					(.5;1;.5)	-.0844	-.0831	-.0205	.0298	.0295	.0297
	control	206	309	142	(.5;.5;.5)	.0507	.0503	.0124	.1962	.1964	.1969
					(1;1;1)	.0514	.0509	.0126	.1898	.0190	.1905
					(1;2;1)	.0498	.0494	.0122	.2037	.2038	.2043
					(.5;1;.5)	.0499	.0495	.0122	.2034	.2035	.2040
AD 108 (4)	case	110	148	44	(.5;.5;.5)	-.0307	-.0288	-.0068	.6108	.6139	.6142
					(1;1;1)	-.0285	-.0266	-.0063	.6364	.6396	.6399
					(1;2;1)	-.0318	-.0298	-.0071	.5966	.5995	.5999
					(.5;1;.5)	-.0324	-.0304	-.0072	.5912	.5941	.5945
	control	34	22	12	(.5;.5;.5)	.2901	.2733	.0610	.0176	.0213	.0245
					(1;1;1)	.2926	.2764	.0620	.0151	.0183	.0211
					(1;2;1)	.2730	.2579	.0580	.0232	.0271	.0303
					(.5;1;.5)	.2800	.2638	.0590	.0218	.0259	.0293

Figure A1: Coefficient D: Curves.

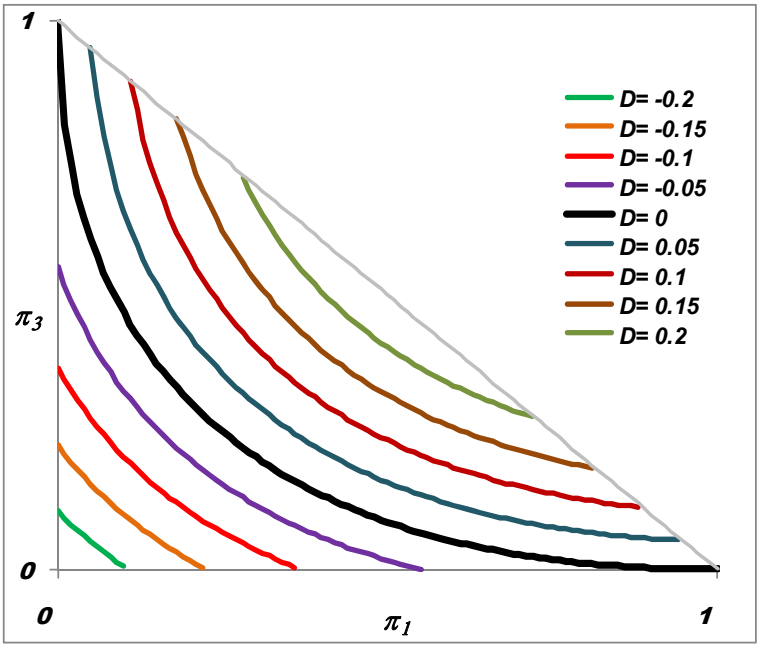
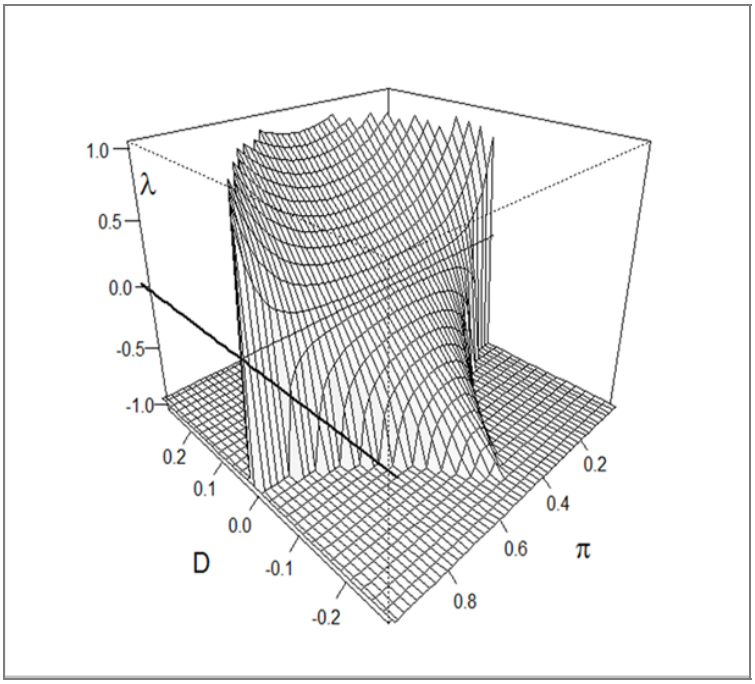
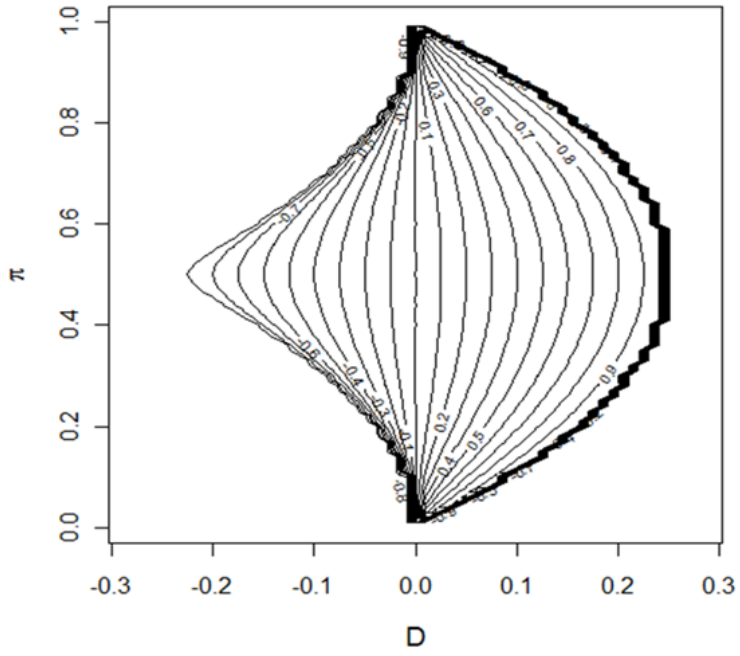


Figure A2: λ as a function of (π,D).



DOI: 10.2202/1544-6115.1636

Figure A3: Contour curves of Figure A2.



REFERENCES

Attia J, Thakkinstian A, McElduff P, Milne E, Dawson S, Scott RJ, Klerk N, Armstrong B and Thompson J (2010), Detecting genotyping error using measures of degree of Hardy-Weinberg disequilibrium, Statistical ***Applications in Genetics & Molecular Biology*** 9(1):5.

Ayres KL and Balding DJ (1998), Measuring departures from Hardy-Weinberg: A Markov chain Monte Carlo method for estimating the inbreeding coefficient, ***Heredity*** 80:769-77.

Basu D (1977), On the elimination of nuisance parameters, ***J American Statistical Society*** 72(358):355-66.

Basu D and Pereira CAB (1982), On the Bayesian analysis of categorical data: the problem of non response***. J Statistical Planning & Inference***, 6(4):345-62.

Bhojak TJ, DeKosky ST, Ganguli M, Kamboh MI (2000), Genetic polymorphisms in the cathespin D and interleukin-6 genes and the risk of Alzheimer's disease, ***Neuroscience Letters*** 288(1):21-4.

Campos CP de and Benavoli A (2011), Inference with multinomial data: why to weaken the prior strength. In: Proceedings of the 22nd International Joint Conference in Artificial Intelligence, to appear.

Chow M and Fong DKH (1992). Simultaneous estimation of the Hardy-Weinberg proportions, ***Canadian J Statistics*** 20:291-6.

Emigh TH (1980). A comparison of tests for Hardy-Weinberg equilibrium, ***Biometrics*** 36: 627-42.

Hardy GH (1908). Mendelian proportions in a mixed population, ***Science*** 28:49-50.

Hernández JL and Weir BS (1989), A disequilibrium coefficient approach to Hardy-Weinberg testing, ***Biometrics*** 45:53-70.

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF (2004), Detection of genotyping errors by Hardy-Weinberg equilibrium testing, ***European J Human Genetics*** 12(5):395-9

Irony TZ, Pereira CAB, and RC Tiwari RC (2000), Analysis of Opinion Swing: Comparison of Two Correlated Proportions, ***The American Statistician***, 54(1): 57-62.

Jeffreys H (1931). ***Scientific Inference***, Cambridge University Press.

Lambert JC, Araria-Goumidi L, Myllykangas L, Ellis C, Wang JC, Bullido MJ, Harris JM, Artiga MJ, Hernandez D, Kwon JM, Frigard B, Petersen RC, Cumming AM, Pasquier F, Sastre I, Tienari PJ, Frank A, Sulkava R, Morris JC, St Clair D, Mann DM, Wavrant-DeVrièze F, Ezquerra-Trabalon M, Amouyel P, Hardy J, Haltia M, Valdivieso F, Goate AM, Pérez-Tur J, Lendon CL, Chartier-Harlin MC (2002), Contribution of APOE promoter polymorphisms to Alzheimer's disease risk, ***Neurology*** 59(1):59-66.

Lauretto MS, Nakano F, Faria SRJ, Pereira CAB and Stern JM (2009), A straightforward multiallelic significance test for the Hardy-Weinberg equilibrium law, ***Genetics and Molecular Biology*** 32(3):619-25.

DOI: 10.2202/1544-6115.1636

Lindley DV (1988), Statistical inference concerning Hardy-Weinberg equilibrium, in: Bernardo JM, DeGroot MH, Lindley DV and Smith AFM Editors, ***Bayesian Statistics***. Oxford U Press, Oxford: 307-26.

Montoya-Delgado L, Irony T, Pereira CAB and Whittle M (2001). An unconditional exact test for the Hardy-Weinberg Equilibrium Law: sample space ordering using the Bayes factor, ***Genetics*** 158:875-83.

Pereira CAB, Nakano F, Stern JM and Whittle MR (2006), Genuine Bayesian multiallelic significance test for the Hardy-Weinberg equilibrium law, ***Genetics and Molecular Research*** 5(4):619-31.

Pereira CAB and Rogatko A (1984), The Hardy-Weinberg equilibrium under a Bayesian perspective, ***Brazilian Journal of Genetics*** 7(4):689-707.

Pereira CAB and Stern JM (1999), Evidence and credibility: full Bayesian significance test of precise hypothesis, ***Entropy*** 1:99-110.

Pereira CAB and Stern JM (2008), Especial characterizations of standard discrete models, ***Statistical Journal*** 6(3):199-230

Pereira CAB, Stern JM and Wechsler S (2008), Can a significance test be genuinely Bayesian? ***Bayesian Analysis*** 3(1):79-100

R Development Core Team (2009), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Ryckman K and Williams SM (2008), Calculation and Use of the Hardy-Weinberg Model in Association Studies, ***Current Protocols in Human Genetics*** 57:1.18.1-1.18.11.

Rao CR (1965), Linear Statistical Inference and Its Applications, Wiley

Rogatko A, Slifker MJ and Babb JS (2000), Hardy-Weinberg equilibrium diagnostics, ***Theoretical Population Biology*** 62:251-7.

Shoemaker J, Painter I and Weir BS (1998), A Bayesian characterization of Hardy-Weinberg disequilibrium, ***Genetics Society of America*** 149: 2079-88.

Singer JM, Peres CA and Harle CE (1991), A note on the Hardy-Weinberg equilibrium in generalized ABO systems, ***Statistics and Probability Letters*** 11:173-5.

Wakefield J (2010), Bayesian Methods for Examining Hardy–Weinberg Equilibrium, ***Biometrics*** 66(1):257-65

Weinberg W (1908), Über den Nachweis der Vererbung beim Menschen, *Jahreshefte des Vereins für vaterländische Naturkunde in* Württemberg 64:369-382.

Weir BS (1996), ***Genetic data analysis II - methods for discrete population genetic data,*** Sinauer Associates, Sunderland.

Yule, GU (1912), On the methods of measuring association between two attributes. ***J Royal Statistical Society*** 75:579-642.

DOI: 10.2202/1544-6115.1636