Testing for Hardy-Weinberg Equilibrium in a Biological Population: An Objective Bayesian Analysis

Vera Lucia TOMAZELLA ¹
José Miguel BERNARDO ²

• ABSTRACT: Many of the problems which traditionally have been formulated in terms of hypothesis testing are really complex decision problems on model choice, whose appropriate solution naturally depends on the structure of the problem. In this paper a probability model for the formation of genotypes from two alleles is given and expressed in terms of two parameters, $\alpha$ and $\beta$; $\alpha = 0$ corresponding to Hardy-Weinberg equilibrium (Lindley, 1988). A particular scientific hypothesis of genetical equilibrium is discussed, special attention is paid to considering that in some genetical applications the proportion of $A$ alleles is known fairly precisely before sampling, the posterior distribution of $\alpha$ considering $\beta$ known is found providing estimation of $\alpha$. The corresponding precise hypothesis testing problem is considered from a decision-theoretical viewpoint, where the null hypothesis is rejected if the null model is expected to be too far from the true model in the logarithmic divergence (Kullback-Leibler) sense. The results are illustrated using examples with data previously analyzed in the literature.

• Key Words: Reference Analysis, Hardy-Weinberg equilibrium, Precise hypothesis testing, information-theory, Intrinsic discrepancy.

1 Introduction

Ever since its discovery in the early 1900s the Hardy-Weinberg equilibrium (HWE) has been a subject of interest and powerful research tool in population genetics (Crow, 1988). Several statistical tests to detect and measure deviation from HWE have been devised, each having advantages and limitations. The limitations become more obvious when testing for deviation within multiallelic DNA loci is attempted. To test the HWE, usually the $\chi^2$ test, the conditional exact test, and the likelihood-ratio test are used. For a complete discussion on these procedures, including some comparisons, see, for instance, Hernandez and Weir (1989), Guo and Thompson (1992), Maiste (1993), Weir (1996), and Lazzeroni and Lange (1997). The conditional exact test is analogous to Fisher’s exact test for contingency tables, there is a sufficient statistic, under the null hypothesis, that is considered to be known, hence the $p$ value is based on the conditional probabilities of the sample points given the value of the statistic.

¹Departamento de Estatística, Universidade Federal de São Carlos - UFSCar, CP 676 - 13565-905 - São Carlos - SP, E-mail: vera@power.ufscar.br
²Departamento de Estadística Universitat de València - València-Espanha, e-mail: bernardo@uv.es
Maiste (1993) and Maiste and Weir (1995) contrasted these tests and claimed to show that the exact conditional test has a better performance. A problem with this test is to define an order in the sample space to calculate the $p$ value. In fact, this is the main difficulty when dealing with high-dimension sample spaces (Kempthorne and Folks, 1971). In addition, Emigh (1980) makes a useful comparison of various equilibrium tests. The problem of estimating the allelic frequencies, under the Bayesian perspective, was considered by Gunel and Wearden (1995). Chow and Fong (1992) studied this same problem but as a particular case of the simultaneous estimation of the related proportions. A complete bibliographic discussion on the many HWE tests has been presented by Montoya-Delgado et al. (2001). They developed an exact test on the basis of the comparison between weighted likelihoods (Dickey and Lientz 1970) under the null and alternative hypotheses. The ratio of these two functions is the Bayes factor (BF). A distribution of the BF under the null hypothesis defines a natural order in the sample space. The aim of this study is to develop a precise hypothesis test on the assumption of HWE, where we treat the problem as a Bayesian decision problem with only two alternatives: either to accept the hypothesis that the population is HW equilibrium, or to conclude that the observed data are, under the assumed model, incompatible with that hypothesis. The combined use of an information-theory based loss function, the intrinsic discrepancy (Bernardo and Rueda, 2002), and an objective prior function, the reference prior (Bernardo, 1979; Berger and Bernardo, 1992), produces a new solution to this problem which, has the invariance properties one should presumably require.

In the paper attention focuses on one of the simplest problems of hypothesis testing within the context of testing Hardy-Weinberg Equilibrium. Special attention is to $\beta$ known. Aside from providing a solution to genetics problems, the paper provides an example of general, rational, Bayesian procedure for testing a scientific hypothesis.

Section 2 contains the Hardy-Weinberg equilibrium model formulation. In Section 3, asymptotic results are shown considering two examples. In Section 4, objective Bayesian counterparts to these traditional inference problem of estimation and testing which are based on the joint of intrinsic loss functions and reference analysis, are briefly considered. In Section 5, inference about Hardy-Weinberg equilibrium is present. In Section 6 we consider the problem of the null hypothesis $\alpha = 0$. In Section 7 concluding remarks.

2 The Hardy-Weinberg equilibrium in genetics

In a large, random mating population in which the evolutionary forces such as selection, migration, and mutation are not acting, allele and genotype frequencies do not change. Furthermore, they are related in a simple way.

At a single autosomal locus with two alleles, $\{A, a\}$, a diploid individual can be one of three possible genotypes $\{AA, Aa, aa\}$. The genotype frequencies in a population will be written by $p_1$, $p_2$, $p_3$, with $p_i \geq 0$, $i = 1, 2, 3$ and $p_1 + p_2 + p_3 = 1$.

In a panmictic population obeying Mendelian rules, equilibrium is attained in one reproductive generation and this assures the existence of a real number $p \in (0, 1)$, such that the
genotype proportions satisfy the relations
\[ p_1 = p^2 \quad p_2 = 2p(1 - p) \quad p_3 = (1 - p)^2 \]
for some \( p \) satisfying \( 0 \leq p \leq 1 \)

and the locus is said to exhibit HWE. One generation random mating is sufficient to produce HWE.

Hence, to decide as to the existence or not of equilibrium it is necessary to test the null hypothesis \( H_0 \), \( w = p_1, p_2, p_3 \in \Omega \), where,
\[
\Omega = \left\{ (p_1, p_2, p_3) \mid 0 \leq p_1 \leq 1; \quad p_2 = 2\sqrt{p_1}(1 - \sqrt{p_1}); \quad p_3 = (1 - \sqrt{p_1})^2 \right\}
\]

Consequently, the statistical problem of interest is the construction of a procedure to test the following two alternative hypotheses,
\[ H_0 : w \in \Omega \]
and
\[ H_1 : w \notin \Omega \]
where \( \Omega \) is defined as above (see figure 1).

![Figure 1: Hardy-Weinberger equilibrium locus](image)

Alternatively expressed, we wish to test the hypothesis that \( 4p_1p_3 = p_2^2 \), a relation clearly true in HW equilibrium, or equivalently that \( 4p_1p_3/p_2^2 = 1 \).

Let us suppose that the system is codominant; that is, distinct genotypic classes define distinct phenotypic classes. In this way, in a sample of size \( n \), the frequencies of members in each class \( n_1, n_2, n_3 \) satisfying the condition \( n = n_1 + n_2 + n_3 \), can be observed. Assuming
that the sample elements are obtained independently, by using a Bernoulli multivariate
process prefixing the sample size $n$, the trinomial model is

$$\text{Tri} (n_1, n_3 \mid n, p_1, p_3) = \frac{n!}{n_1! (n - n_1 - n_3)! n_3!} p_1^{n_1} (1 - p_1 - p_3)^{n - n_1 - n_3} p_3^{n_3}$$

with $0 \leq p_1 \leq 1, 0 \leq p_3 \leq 1, 0 \leq p_1 + p_3 \leq 1$ and $n_1 = 0, \ldots, n$, $n_3 = 0, \ldots, n - n_1$.

Lindley (1988) framed their analysis in Bayesian terms on two new parameters

$$\alpha = \frac{1}{2} \log \frac{4 p_1 p_3}{p_2^2}, \quad \alpha \in \mathbb{R}$$
$$\beta = \frac{1}{2} \log \frac{p_1}{p_3}, \quad \beta \in \mathbb{R}$$

His formulation has the advantage that $\alpha$ and $\beta$ have bounds that do not depend on the other, making it straightforward to determine the marginal posterior distribution for each variable separately. Equilibrium is clearly $\alpha = 0$ and it is wished to test $\alpha = 0$ against $\alpha \neq 0$, $\beta$ being nuisance. Notice that having a parameter $\alpha$ enables the geneticist not merely to measure equilibrium, $\alpha = 0$, but also to describe the amount of disequilibrium through non-zero values.

So the values of $p_1$, $p_2$ and $p_3$ in terms of $\alpha$ and $\beta$ are given by:

$$p_1 = \frac{e^{\alpha + \beta}}{2(1 + e^\alpha \cosh(\beta))}$$
$$p_2 = \frac{1}{1 + e^\alpha \cosh(\beta)}$$
$$p_3 = \frac{e^{\alpha - \beta}}{2(1 + e^\alpha \cosh(\beta))}$$

3 The likelihood and asymptotic results

The likelihood function for data $\mathcal{D} = (n_1, n_2, n_3)$ in terms of $\alpha$ and $\beta$ with simple calculations show that this is

$$L (\alpha, \beta \mid \mathcal{D}) = \frac{e^{\alpha (n_1 + n_3)} e^{\beta (n_1 - n_3)}}{(1 + e^\alpha \cosh(\beta))^n}$$

According to the likelihood principle the likelihood provides all the information about $\alpha$ and $\beta$ contained in the data, and in particular, Haldane’s conditionally and the distribution of any statistic are irrelevant. We therefore concentrate on (4) and begin by considering its asymptotic behavior when all of $n_1$, $n_2$ and $n_3$ are large.

**Example 1:** A numeric example was developed in Lindley (1998), where he considered 4 situations of different sample, for each situation was considered $n = 100$, with sample 1 $(31, 38, 31)$, sample 2 $(6, 22, 72)$, sample 3 $(2, 6, 92)$ and sample 4 $(1, 8, 91)$. 
Lindley classified the cases according to the observed proportion of A alleles $\hat{p} = (2n_1 + 2n_2)/2n$. Figure (2) shows the likelihood function (4) for all samples with values of $\beta = 0$ and $\beta = 2$, and Figure (3) show the likelihood function (4) in three dimension and contour graphic for sample 1.

Figure 2: Likelihood function (4), for all samples with $\beta = 0$ (left) and $\beta = 2$ (right)

The logarithm of the likelihood function (4) is given by

$$l(\alpha, \beta|D) = \alpha(n_1 + n_3) + \beta(n_1 - n_3) - n \log [1 + e^\alpha \cosh(\beta)]$$  \hspace{1cm} (5)

The first derivatives of $l(\alpha, \beta|D)$ in (5) are given by

$$\frac{\partial l}{\partial \alpha} = (n_1 + n_3) - \frac{n \cosh(\beta)}{\cosh(\beta) + e^{-\alpha}}$$

$$\frac{\partial l}{\partial \beta} = (n_1 - n_3) - \frac{n \sinh(\beta)}{\cosh(\beta) + e^{-\alpha}}$$  \hspace{1cm} (6)
Equating these to zero and solving the resulting equations for \( \alpha \) and \( \beta \), the maximum likelihoods are

\[
\hat{\alpha} = \frac{1}{2} \log \left[ \frac{4n_1n_3}{(n-n_1-n_3)^2} \right]
\]

\[
\hat{\beta} = \frac{1}{2} \log \left[ \frac{n_1}{n_3} \right]
\]

The second derivatives of \( l(\alpha, \beta|D) \) most conveniently depend on the data through \( n \), the sample size. They are

\[
\frac{\partial^2 l}{\partial \alpha^2} = -\frac{n \cosh(\beta)e^{-\alpha}}{(\cosh(\beta) + e^{-\alpha})^2}
\]

\[
\frac{\partial^2 l}{\partial \alpha \partial \beta} = -\frac{n \sinh(\beta)e^{-\alpha}}{(\cosh(\beta) + e^{-\alpha})^2}
\]

\[
\frac{\partial^2 l}{\partial \beta^2} = -\frac{n (1 + \cosh(\beta)e^{-\alpha})}{(\cosh(\beta) + e^{-\alpha})^2}
\]

By standard theory it follows that the likelihood is, for large values of \( n \), approximately a normal form, centered at \((\hat{\alpha}, \hat{\beta})\). The Fisher information matrix is therefore

\[
I(\alpha, \beta) = \frac{ne^{-\alpha}}{(\cosh(\beta) + e^{-\alpha})^2} \begin{bmatrix}
\cosh(\beta) & \sinh(\beta) \\
\sinh(\beta) & \cosh(\beta) + e^\alpha
\end{bmatrix}
\]

The dispersion matrix given by inverting the Fisher information matrix is

\[
I^{-1}(\alpha, \beta) = \frac{(\cosh(\beta) + e^{-\alpha})}{n} \begin{bmatrix}
\cosh(\beta) + e^{-\alpha} & -\sinh(\beta) \\
-\sinh(\beta) & \cosh(\beta)
\end{bmatrix}
\]

Table 1 shows estimates results for maximum likelihood \( \hat{\alpha} \) and \( \hat{\beta} \) given in (7), \( p_1, p_2 \) and \( p_3 \) given in (3) and \( \hat{p} = (2n_1+n_2)/2n \) the population of Alleles A in the samples for example 1.

<table>
<thead>
<tr>
<th>parameters</th>
<th>(31, 38, 31)</th>
<th>(6, 22, 72)</th>
<th>(2, 6, 92)</th>
<th>(1, 8, 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} )</td>
<td>0.50</td>
<td>0.17</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>( \hat{p}_1 )</td>
<td>0.31</td>
<td>0.06</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>( \hat{p}_2 )</td>
<td>0.38</td>
<td>0.22</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>( \hat{p}_3 )</td>
<td>0.31</td>
<td>0.72</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>( \hat{\alpha} )</td>
<td>0.489</td>
<td>0.636</td>
<td>1.509</td>
<td>0.869</td>
</tr>
<tr>
<td>( \hat{\beta} )</td>
<td>0.00</td>
<td>-1.224</td>
<td>-1.914</td>
<td>-2.255</td>
</tr>
</tbody>
</table>
Example 2: Lidicker and McCollum (1997) examined genetic variation in two populations of sea otters (*Enhydra lutris*) in the eastern Pacific. Sea otters were distributed throughout this region before fur hunting nearly led to their local extinction. The table below contains counts of the number of individuals with a given genotype for six variable (polymorphic) two-allele loci.

Table 2: data from Lidicker and McCollum 1997

<table>
<thead>
<tr>
<th>Locus</th>
<th>Genotype</th>
<th>California</th>
<th>Alaska</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST</td>
<td>SS</td>
<td>37</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>ICD</td>
<td>SS</td>
<td>48</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>LA</td>
<td>SS</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>PAP</td>
<td>SS</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>ME</td>
<td>SS</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>NP</td>
<td>SS</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SF</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Similarly to example 1, Table 3 show the results for estimates of maximum likelihood \( \hat{\alpha} \) and \( \hat{\beta} \) given in (7), \( p_1, p_2 \) and \( p_3 \) given in (3) and \( \hat{p} = (2n_1 + n_2)/2n \) the population of Alleles A considering the data of the example 2.

Table 3: estimates results for the parameters of the model

<table>
<thead>
<tr>
<th>parameters</th>
<th>(37, 20, 7)</th>
<th>(48, 4, 3)</th>
<th>(20, 11, 2)</th>
<th>(16, 7, 10)</th>
<th>(16, 11, 5)</th>
<th>(17, 4, 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha} )</td>
<td>0.73</td>
<td>0.91</td>
<td>0.77</td>
<td>0.63</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>( \hat{\beta} )</td>
<td>0.58</td>
<td>0.87</td>
<td>0.61</td>
<td>0.48</td>
<td>0.50</td>
<td>0.65</td>
</tr>
<tr>
<td>( \hat{\beta} )</td>
<td>0.31</td>
<td>0.07</td>
<td>0.33</td>
<td>0.21</td>
<td>0.34</td>
<td>0.15</td>
</tr>
<tr>
<td>( \hat{p}_3 )</td>
<td>0.11</td>
<td>0.06</td>
<td>0.06</td>
<td>0.31</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>( \hat{\alpha} )</td>
<td>0.475</td>
<td>1.792</td>
<td>0.139</td>
<td>1.285</td>
<td>0.486</td>
<td>1.528</td>
</tr>
<tr>
<td>( \hat{\beta} )</td>
<td>0.832</td>
<td>0.386</td>
<td>1.151</td>
<td>0.235</td>
<td>0.581</td>
<td>0.612</td>
</tr>
</tbody>
</table>
4 Objective precise bayesian testing

4.1 The decision problem and the intrinsic loss function

If data $z$ are assumed to have been generated from a probability model

$$M \equiv \{p_z(\cdot | \phi, \omega), z \in \Phi, \omega \in \Omega\},$$

then testing whether or not the observed data $z$ are compatible with the precise hypothesis $H_0 \equiv \phi = \phi_0$ is a simple decision problem with only two alternatives:

- $a_0$: To accept $H_0$, and work as if data were generated from the reduced model
  $$M_0 \equiv \{p_z(\cdot | \phi_0, \omega), z \in \Phi, w \in \Omega\}$$

- $a_1$: To reject $H_0$, and keep working with the assumed model $M$.

Foundations dictate (see e.g., Bernardo and Smith, 1994, and references therein) that, to solve this decision problem, one must specify utility functions $u\{a_i, (\phi, \omega)\}$ for the two alternatives $a_0$ and $a_1$, and a joint prior distribution $\pi(\phi, \omega)$ for the unknown parameters $(\phi, \omega)$; then, $H_0$ should be rejected if, and only if,

$$\int_{\Phi} \int_{\Omega} [u\{a_1, (\phi, \omega)\} - u\{a_0, (\phi, \omega)\}] \pi(\phi, \omega | z) d\omega > 0 \quad (11)$$

where, using Bayes theorem, $\pi(\phi, \omega | z) \propto p(z | \phi, \omega) \pi(\phi, \omega)$ is the joint posterior which corresponds to the prior $\pi(\phi, \omega)$. Thus, only the utilities difference must be specified, and this may usefully be written as

$$u\{a_1, (\phi, \omega)\} - u\{a_0, (\phi, \omega)\} = \ell\{\phi_0, (\phi, \omega)\} - d_0 \quad (12)$$

where $\ell\{\phi_0, (\phi, \omega)\}$ is the non-negative terminal loss suffered by accepting $\phi = \phi_0$ given $(\phi, \omega)$, and $d_0 > 0$ is the utility of accepting $H_0$ when it is true. Hence, $H_0$ should be rejected if, and only if,

$$t(\phi_0|z) = \int_{\Theta} \int_{\Omega} \ell\{\phi_0, (\phi, \omega)\} \pi(\phi, \omega | z) d\phi d\omega > d_0 \quad (13)$$

that is, if (and only if) the null model is expected to be too divergent from the true model.

For any one-to-one function $\varphi = \varphi(\phi)$ the conditions to reject $\phi = \phi_0$ should certainly be precisely the same as the conditions to reject $\varphi = \varphi(\phi_0)$ (a property unfortunately not satisfied by many published hypothesis testing procedures). This requires the use of an invariant loss function.

Model-based loss functions are loss functions defined in terms of the discrepancy measures between probability models. Within a family $F \equiv \{p_z(\cdot | \varphi), \phi \in \varphi\}$, the loss suffered from using an estimate $\widehat{\varphi}$ is of the form

$$l(\widehat{\varphi}, \varphi) = \delta\{p_z(\cdot | \widehat{\varphi}), p_z(\cdot | \varphi)\}$$

defined in terms of the discrepancy of $p_z(\cdot | \widehat{\varphi})$ from $p_z(\cdot | \varphi)$, rather than on the discrepancy of $\widehat{\varphi}$ from $\varphi$. Model-based loss functions are obviously invariant under one-to-one reparametrization.
A model-based loss function with unique additive properties and built in calibration, is the intrinsic loss function, defined as the minimum expected log-likelihood ratio against the null

$$\delta \{ \phi_0, (\phi, \omega) \} = \inf_{\omega \in \Omega} \delta \{ p(x|\theta, \omega), p(x|\theta_0, \omega_0) \}$$  (14)

As it is apparent from its definition, the intrinsic loss (14) is the minimum conditional expected log-likelihood ratio (under repeated sampling) against $H_0$.

### 4.2 Reference analysis and precise hypothesis testing

Given a model $M \equiv \{ p_z(.|\phi, \omega), z \in Z, \phi \in \Phi, \omega \in \Omega \}$, the $\theta$-reference prior function $\pi_\theta(\phi, \omega)$ is that which maximizes the missing information about $\theta = \theta(\phi, \omega)$. The corresponding marginal reference posterior $\pi(\theta|z)$ summarizes inferential statements about a quantity of interest $\theta$ which only depend on the model assumed and the data obtained.

**The Bayesian Reference Criterion (BRC)**

An objective Bayesian procedure (objective in the sense that it depends exclusively on the assumed model and the observed data), requires an objective noninformative prior which mathematically describes lack on relevant information about the quantity of interest, and which only depends on the assumed statistical model and on the quantity of interest. Recent literature contains a number of requirements which may be regarded as necessary properties of any algorithm proposed to derive these baseline priors; those requirements include general applicability, invariance under reparametrization, consistent marginalization, and appropriate coverage properties. The reference analysis algorithm, introduced by Bernardo (1979) and further developed by Berger and Bernardo (1989, 1992) is, to the best of our knowledge, the only available method to derive objective priors which satisfy all these desiderata. For an introduction to reference analysis, see Bernardo and Ramón (1998); for a textbook level description see Bernardo and Smith (1994, Ch. 5); for a critical overview of the topic, see Bernardo (2005), references therein and ensuing discussion.

The Bayesian reference criterion (Bernardo and Rueda, 2002) is the normative Bayes solution to the decision problem of hypothesis testing described above which corresponds to use of the intrinsic loss and the reference prior. Given a parametric model $\{ p_z(.|\phi, \omega), \phi \in \Phi, \omega \in \Omega \}$, this prescribes to reject the hypothesis $\{ H_0 \equiv \phi = \phi_0 \}$ if, and only if,

$$d(H_0|z) = \int_{-\infty}^{\infty} \delta \pi(\delta|z) d\delta > d_0$$  (15)

where $d(H_0|z)$, termed the intrinsic (test) statistic, is the reference posterior expectation of the intrinsic loss $\delta_z \{ H_0, (\phi, \omega) \}$ defined by (14), and where $d_0$ is a context dependent positive utility constant, the largest acceptable average log-likelihood ratio against $H_0$ under repeated sampling. This is a continuous measure of the evidence provided by the data against the
(null) hypothesis that a model of the form \( p(x|\phi_0, \omega_0) \) for some \( \omega_0 \in \Omega \) may safely be used as a proxy for the assumed model, \( p(x|\phi, \omega) \). In particular, values of \( d_0 \approx \ln(10) = 2.3 \) should be regarded as mild evidence against \( H_0 \), while values \( d_0 \approx \ln(100) = 4.6 \) suggest strong evidence against \( H_0 \), and values larger than \( \ln(1000) = 6.9 \) may be safely used to reject \( H_0 \).

In traditional language, \( d(H_0|z) \) is a (monotone) test statistic for \( H_0 \), and the null hypothesis should be rejected if the value of \( d(H_0|z) \) exceeds some critical value \( d_0 \).

**Proposition 1.** (One-Dimensional Asymptotic Behaviour). If \( x = \{x_1, \ldots, x_n\} \) is a random sample from a regular model \( \{p(x|\theta), \theta \in \Theta \subset \mathcal{R}, x \in X \subset \mathcal{R}\} \), with one continuous parameter, and \( \phi(\theta) = \int \theta i(\theta)^{1/2} d\theta \), where \( i(\theta) = -E_{x|\theta}[\partial^2 \log p(x|\theta)/\partial \theta^2] \), then the intrinsic statistic \( d(\theta_0, x) \) to test \( \{\theta = \theta_0\} \) is

\[
d(\theta_0, x) = \frac{1}{2} \left[ 1 + z^2 \left( \theta_0, \hat{\theta} \right) \right] + o(n^{-1}), \quad z^2 \left( \theta_0, \hat{\theta} \right) = \sqrt{n} \left[ \phi(\hat{\theta}) - \phi(\theta_0) \right] \tag{16}
\]

where \( \hat{\theta} = \hat{\theta}(x) = \arg \max p(x|\theta) \). Moreover, the expected value \( E_{\theta}(d(\theta_0, x)) \) under repeated sampling is

\[
E_{x|\theta}[d(\theta_0, x)] = 1 + n [\phi(\theta) - \phi(\theta_0)]^2 + o(n^{-1}) \tag{17}
\]

So that \( d(\theta_0, x) \) will concentrate around the value one if \( \theta = \theta_0 \) and will linearly increase with \( n \) otherwise. The arguments leading to Proposition 1 may be extended to multivariate situations, with or without nuisance parameters.

## 5 Objective Inference about Hardy-Weinberg equilibrium with \( \beta \) known

In some genetical applications the proportion of \( A \) alleles is known fairly precisely before sampling. This is \( p_1 + \frac{1}{2} p_2 \), or equivalently \( (p_1 - p_3 + 1)/2 \), so that \( p_1 - p_3 \) is well determined. In this parametrization the equivalent situation would be that \( \log p_1 - \log p_3 \) is known, implying knowledge of \( \beta \). So we develop the reference analysis for the case of one parameter. The likelihood function of \( \alpha \) conditional the \( \beta \) known and the data is

\[
L(\alpha|\beta, D) = \frac{e^{\alpha(n_1 + n_3)}}{[1 + e^{\alpha \cosh(\beta)}]^n} \tag{18}
\]

In practice, if \( \beta \) is not known, a first approximation may be obtained by using its MLE estimator \( \hat{\beta} = (1/2) \log[n_1/n_3] \).

Considering the data of example 1, and example 2, Figure (4) shows the likelihood function (18) respectively taking the MLEs of \( \beta \).

The logarithm of the likelihood (18) is

\[
l(\alpha, \beta|D) = \alpha(n_1 + n_3) - n \log [1 + e^{\alpha \cosh(\beta)}] \tag{19}
\]
Figure 4: Likelihood function (18) for example 1 (right) and example 2 (left)

Hence

$$\frac{\partial l}{\partial \alpha} = (n_1 + n_3) - \frac{n \cosh(\beta)}{\cosh(\beta) + e^{-\alpha}}$$

Equating these to zero and solving the resulting equation for $\alpha$, the maximum likelihood estimate is

$$\hat{\alpha} = \log \left[ \frac{n_1 + n_3}{(n - n_1 - n_3) \cosh(\beta)} \right]$$

(20)

The second derivative of $l$ is given by

$$\frac{\partial^2 l}{\partial \alpha^2} = - \frac{n \cosh(\beta) e^{-\alpha}}{(\cosh(\beta) + e^{-\alpha})^2}$$

The Fisher information for $\alpha$ considering $\beta$ known is

$$I(\alpha|\beta) \propto \frac{e^{-\alpha}}{(\cosh(\beta) + e^{-\alpha})^2}$$

Thus the reference prior of $\alpha$ conditioned the knowledge of $\beta$ is given by

$$\pi(\alpha|\beta) = [I(\alpha|\beta)]^{1/2} \propto \frac{e^{\alpha/2}}{(1 + e^\alpha \cosh(\beta))}$$

(21)

Since that

$$\pi(\alpha|\beta) = \frac{e^{\alpha/2} (1 + e^\alpha \cosh(\beta))^{-1}}{\int_{-\infty}^{\infty} e^{\alpha/2} (1 + e^\alpha \cosh(\beta))^{-1} d\alpha}$$

(22)

Solving the integral in (22), we have that the reference prior of $\alpha$ conditional the knowledge of $\beta$ is
\[
\pi(\alpha|\beta) = \frac{1}{\pi} \left( e^{\alpha \cosh(\beta)} \right)^{1/2} \left( 1 + e^{\alpha \cosh(\beta)} \right)^{-1/2}, \quad -\infty < \alpha < \infty \tag{23}
\]

It is proper

Figure (5) show the reference prior distribution for values of \( \beta = 0.5, \beta = 1 \) and \( \beta = 2 \).

\[
\pi(\alpha|\beta, \mathcal{D}) = \frac{\Gamma(n+1)}{\Gamma(n_1 + n_3 + \frac{1}{2}) \Gamma(n - n_1 - n_3 + \frac{1}{2})} \left( e^{\alpha \cosh(\beta)} \right)^{n_1 + n_3 + \frac{1}{2}} \left( 1 + e^{\alpha \cosh(\beta)} \right)^{n+1} \tag{24}
\]

where \(-\infty < \alpha < \infty\).

Since (24) has been derived from a proper prior, it is obviously proper for any data. Moreover, as expected, it depend on \( n_1 \) and \( n_3 \). In the particular case where \((n_1 = n_2 = n_3 = 0)\) so that all \( n \) observations belong to the third category, we do not have any information about the parameters of interest and hence, the reference posterior distribution reduces to the marginal reference prior (23), shown in Figure (6).

Figure 5: Reference priori (23)

Figure 6: Reference prior and reference posterior for any data with \((n_1 = n_2 = n_3 = 0)\)
Considering the data of the example 1, and example 2, Figure (7) and Figure (8) show the reference posterior distribution (24) taking the MLEs of \( \beta \). Also Figure (9) shows the reference prior and reference posterior for sample 1 with \( \hat{\beta} = 0 \).

Figure 7: Reference posterior distribution for example 1 samples

Figure 8: Reference posterior distribution for example 2 samples

Figure 9: Reference prior distribution and posterior for samples 1 and \( \hat{\beta} = 0 \)
5.0.1 Point Estimation

The idea of estimation of parameters by point leads in the Bayesian scene to take like estimate typical points of a posterior distribution for which the determination of its measures of location is revealed useful. The election of the Bayesian estimate depends naturally on the distribution a posterior. The used estimates are the mode, mean, median and variance. So considering the reference posterior distribution (24) and after simple calculations we find that the estimate of the mean, mode and variance respectively are given by,

\[ \mu_{\alpha|D} = \left[ \psi(n_1 + n_3 + 1/2) - \psi(n - n_1 - n_3 + 1/2) - \log(\cosh(\beta)) \right], \]

\[ mo_{\alpha|D} = \log\left[ \frac{1 + 2(n_1 + n_3)}{(1 + 2(n - n_1 - n_3) \cosh(\beta))} \right], \]

\[ \sigma_{\alpha|D}^2 = \left[ \psi(n_1 + n_3 + 1/2) - \psi(n - n_1 - n_3 + 1/2) \right]^2 + 
\psi(1, n_1 + n_3 + 1/2) + \psi(1, (n - n_1 - n_3) + 1/2) - 
2 \log(\cosh(\beta))[\psi(n_1 + n_3 + 1/2) - \psi(n - n_1 - n_3 + 1/2)] + (\log(\cosh(\beta)))^2 - 
(\psi(n_1 + n_3 + 1/2) - \psi(n - n_1 - n_3 + 1/2) - \log(\cosh(\beta)))^2. \]

Table 4 and Table 5 show the posterior results for example 1 and example 2 samples respectively: the mode \( mo_{\alpha|D} \), mean \( \mu_{\alpha|D} \), median \( me_{\alpha|D} \), and standard deviation \( \sigma_{\alpha|D} \).

Table 4: Point estimation for Lindley data taking the MLEs of \( \beta \)

<table>
<thead>
<tr>
<th>measures</th>
<th>(31, 38, 31)</th>
<th>(6, 22, 72)</th>
<th>(2, 6, 92)</th>
<th>(1, 8, 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{\alpha</td>
<td>D} )</td>
<td>0.4895</td>
<td>0.6518</td>
<td>1.5080</td>
</tr>
<tr>
<td>( mo_{\alpha</td>
<td>D} )</td>
<td>0.4845</td>
<td>0.6358</td>
<td>1.4344</td>
</tr>
<tr>
<td>( me_{\alpha</td>
<td>D} )</td>
<td>0.4870</td>
<td>0.6438</td>
<td>1.4712</td>
</tr>
<tr>
<td>( \sigma_{\alpha</td>
<td>D} )</td>
<td>0.0424</td>
<td>0.0583</td>
<td>0.1769</td>
</tr>
</tbody>
</table>

Table 5: Point estimation for Lidicker and McCollum data taking the MLEs of \( \beta \)

<table>
<thead>
<tr>
<th>measures</th>
<th>(37, 20, 7)</th>
<th>(48, 4, 3)</th>
<th>(20, 11, 2)</th>
<th>(16, 7, 10)</th>
<th>(16, 11, 5)</th>
<th>(17, 4, 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{\alpha</td>
<td>D} )</td>
<td>0.4761</td>
<td>2.4703</td>
<td>0.1396</td>
<td>1.2840</td>
<td>0.4863</td>
</tr>
<tr>
<td>( mo_{\alpha</td>
<td>D} )</td>
<td>0.4628</td>
<td>2.3648</td>
<td>0.1179</td>
<td>1.2349</td>
<td>0.4656</td>
</tr>
<tr>
<td>( me_{\alpha</td>
<td>D} )</td>
<td>0.4694</td>
<td>2.4175</td>
<td>0.1288</td>
<td>1.2595</td>
<td>0.4758</td>
</tr>
<tr>
<td>( \sigma_{\alpha</td>
<td>D} )</td>
<td>0.0727</td>
<td>0.2683</td>
<td>0.1363</td>
<td>0.1812</td>
<td>0.1384</td>
</tr>
</tbody>
</table>

6 Testing the Hardy-Weinberg Equilibrium for \( \beta \) known

The statistical description of the genetics is that there is a random sample \( D = (n_1, n_2, n_3) \) of \( n \) individuals from a Trinomial distribution with parameters \( p_1, p_1 \) and \( p_3 \), as is described in Section 1. It is desired to test the hypothesis that it really only depends on
one parameter \( p \), yielding Trinomial parameters \( p^2, 2p(1-p), (1-p)^2 \), that is, if the population is equilibrium. Considering the parametrization (2), equilibrium is \( \alpha = 0 \) and we wish to test \( H_0 : \alpha = 0 \) against \( H_1 : \alpha \neq 0 \).

In this direction considering the model in (1), and the methodology described in section (4), the logarithmic divergence of \( p(n_1, n - n_1 - n_3, n_3) | \alpha_0, \beta \) of \( p(n_1, n - n_1 - n_3, n_3 | \alpha, \beta) \), where \( \alpha_0 = 0 \) is given by

\[
k(\alpha_0 | \alpha, \beta) = \sum p(n_1, (n - n_1 - n_3, n_3) | \alpha_0, \beta) \log \left( \frac{p(n_1, (n - n_1 - n_3, n_3) | \alpha_0, \beta)}{p(n_1, (n - n_1 - n_3, n_3) | \alpha, \beta)} \right)
= n \left[ \log \left( \frac{1 + \exp(\alpha \cosh(\beta))}{1 + \exp(\alpha \cosh(\beta))} \right) - \frac{\alpha \cosh(\beta)}{1 + \cosh(\beta)} \right]
\]

This is a nonnegative amount that is zero if, and only if \( \alpha = 0 \). The logarithmic divergence \( p(n_1, n - n_1 - n_3, n_3 | \alpha, \beta) \) of \( p(n_1, n - n_1 - n_3, n_3 | \alpha_0, \beta) \), where \( \alpha_0 = 0 \) is

\[
k(\alpha | \alpha_0, \beta) = \sum p(n_1, (n - n_1 - n_3, n_3) | \alpha, \beta) \log \left( \frac{p(n_1, (n - n_1 - n_3, n_3) | \alpha, \beta)}{p(n_1, (n - n_1 - n_3, n_3) | \alpha_0, \beta)} \right)
= n \left[ \log \left( \frac{1 + \exp(\alpha \cosh(\beta))}{1 + \exp(\alpha \cosh(\beta))} \right) + \frac{\alpha \exp(\alpha \cosh(\beta))}{1 + \exp(\alpha \cosh(\beta))} \right]
\]

Considering the intrinsic discrepancy \( \delta(\alpha_0, \alpha) = \min_{\beta} \{ k(\alpha_0 | \alpha, \beta), k(\alpha | \alpha_0, \beta) \} \). Thus it is possible to be shown that the logarithmic divergence in (29) is smaller and

\[
\delta(\alpha_0, \alpha) = n \left[ \log \left( \frac{1 + \exp(\alpha \cosh(\beta))}{1 + \exp(\alpha \cosh(\beta))} \right) - \frac{\alpha \cosh(\beta)}{1 + \cosh(\beta)} \right]
\]

Figure (10) represents the intrinsic discrepancy for different values of \( n \) and \( \beta = 0 \)

![Figure 10: Intrinsic discrepancy](image)

Considering the reference posterior (24) the intrinsic test statistics is a function given by

\[
d(\alpha_0, D) = \int_{0}^{\infty} \delta(\alpha_0, \alpha) \pi(\alpha | \beta, D) d\alpha,
\]

this has a simple analytical expression given by

\[
d(\alpha_0, \alpha) = \psi(n + 1) - \frac{1}{1 + \cosh(\beta)} \psi((n - n_1 - n_3) + 1/2) - \log(1 + (\cosh(\beta)))
- \left( 1 - \frac{1}{1 + \cosh(\beta)} \right) (\psi(n_1 + n_3 + 1/2) - \log(\cosh(\beta)))
\]
Considering the problem of testing $H_0 \equiv \{ \alpha = \alpha_0 \}$, where $\alpha_0 = 0$, the intrinsic statistic approximated is given by,

$$d(\alpha_0, D) = \int_0^\infty \delta (\alpha_0, \alpha) \pi (\alpha|\beta, D) d\alpha = \frac{1}{\pi} \left[ 1 + z^2 (\alpha_0, \hat{\alpha}) \right] + o(n^{-1})$$  \hspace{1cm} (33)

where

$$z (\alpha_0, \hat{\alpha}) = \sqrt{\pi} [\phi (\hat{\alpha}) - \phi (\alpha_0)],$$

and

$$\phi (\hat{\alpha}) = \frac{2}{\pi} \arctan \left( e^{\hat{\alpha} \sqrt{\cosh(\beta)}} \right)$$

and

$$\hat{\alpha} = \log \left( \frac{n_1 + n_3}{(n - n_1 - n_3) \cosh(\beta)} \right).$$

So the expression of the intrinsic discrepancy for the model is then

$$d(\alpha_0, D) = \frac{1}{2} \left[ 1 + \frac{4n}{\pi^2} \left[ \arctan \left( \frac{n_1 + n_3}{n - n_1 - n_3} \right) - \arctan \left( \sqrt{\cosh(\beta)} \right) \right]^2 \right] + o(n^{-1}) \hspace{1cm} (34)$$

The expected value of $d(\alpha_0, D)$ under repeated sampling is

$$E_{x|\alpha,\beta} [d(\alpha_0, D)] = 1 + \frac{4n}{\pi^2} \left[ \arctan(\exp(\alpha/2) \sqrt{\cosh(\beta)}) - \arctan(\sqrt{\cosh(\beta)}) \right]^2 + o(n^{-1}) \hspace{1cm} (35)$$

Table 11 shows the measure of the intrinsic statistic approximated (34) and the exact value of the intrinsic statistic (32) for the data of example 1

<table>
<thead>
<tr>
<th>measure intrinsic</th>
<th>(31, 38, 31)</th>
<th>(6, 22, 72)</th>
<th>(2, 6, 92)</th>
<th>(1, 8, 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d(\alpha_0, D)_{Apro}$</td>
<td>3.437</td>
<td>4.769</td>
<td>12.562</td>
<td>4.456</td>
</tr>
<tr>
<td>$d(\alpha_0, D)_{Exato}$</td>
<td>3.464</td>
<td>4.973</td>
<td>15.106</td>
<td>4.710</td>
</tr>
</tbody>
</table>

We observe that samples 1, 2 and 4 have “strong evidence against the null hypothesis” while Sample 3 has “very strong evidence against the null hypothesis”.

For the data of example 2 we have the results of the intrinsic statistic approximated (34) and the exact value of the intrinsic statistic (32) shown in Table 12, where we observed that the samples 1,3,4,5 and 6 do not have any evidence against the hypothesis $H_0$ and sample 3 has mild evidence against the hypothesis $H_0$.

Table 12: Intrinsic measures $d(\alpha_0, D)$ exact and approximated for example 2

<table>
<thead>
<tr>
<th>measure intrinsic</th>
<th>(37, 20, 7)</th>
<th>(48, 4, 3)</th>
<th>(20, 11, 2)</th>
<th>(16, 7, 10)</th>
<th>(16, 11, 5)</th>
<th>(17, 4, 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d(\alpha_0, D)_{Apro}$</td>
<td>0.669</td>
<td>3.222</td>
<td>0.507</td>
<td>1.102</td>
<td>0.5913</td>
<td>1.103</td>
</tr>
<tr>
<td>$d(\alpha_0, D)_{Exato}$</td>
<td>0.527</td>
<td>3.954</td>
<td>0.565</td>
<td>1.836</td>
<td>0.4107</td>
<td>1.097</td>
</tr>
</tbody>
</table>
7 Concluding Remarks

This article focuses on the derivation of objective methods where the results only depend on the data obtained and the model assumed. The BRC criterion provides a general reference Bayesian solution to hypothesis testing. The formulation of genotypes from two alleles was considered in terms of two parameters $\alpha$ and $\beta$. An advantage of this approach is the provision, in $\alpha$, of the measure of departure from equilibrium, that is, $\alpha = 0$ corresponding to equilibrium. We considered the case special when the proportion of A alleles is known, implying knowledge of $\beta$, where the posterior distribution conditional of $\alpha$ given $\beta$ known were considered. Another possibility for study is to consider $\beta$ as nuisance parameter and $\alpha$ the parameter of interest and using the methodology of reference analysis to find a reference prior joint to $\alpha$ and $\beta$. Lindley’s paper of (1988) dealing with this important problem considering inference about equilibrium with a uniform prior.

8 Reference


