Ming-Hui Chen, Dipak K. Dey, Peter Müller,
Dongchu Sun, and Keying Ye

# Frontiers of Statistical Decision Making and Bayesian Analysis

— In Honor of James O. Berger —

May 7, 2010

Springer

# Contents

# Chapter 2
# Objective Bayesian Inference with Applications

It is natural to start a review of research frontiers in Bayesian analysis with a discussion of research challenges related to prior choices. In particular, in this chapter we discuss the definition of reference priors in some non-standard settings as well as the use of reference priors to define objective Bayesian testing.

## 2.1 Bayesian Reference Analysis of the Hardy-Weinberg Equilibrium

*José M. Bernardo and Vera Tomazella*

An important problem in genetics, testing whether or not a trinomial population is in Hardy-Weinberg equilibrium, is analyzed from an objective Bayesian perspective. 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 quantity of interest in this problem is the divergence of the null model from the true model; as a consequence, the analysis is made using the reference prior for the trinomial model which corresponds to that divergence being the parameter of interest. The results are illustrated using examples both with simulated data and with data previously analyzed in the relevant literature.

## *2.1.1 Problem Statement*

### 2.1.1.1 The Hardy-Weinberg (HW) Equilibrium in Genetics

At a single autosomal locus with two alleles, a diploid individual has three possible genotypes, typically denoted $\{AA, aa, Aa\}$, with (unknown) population frequencies $\{\alpha_1, \alpha_2, \alpha_3\}$, where $0 < \alpha_i < 1$ and $\sum_{i=1}^{3} \alpha_i = 1$.

The population is said to be in HW equilibrium if there exists a probability $p = P(A)$, $0 < p < 1$, such that $\{\alpha_1, \alpha_2, \alpha_3\} = \{p^2, (1-p)^2, 2p(1-p)\}$. To determine whether or not a population is in HW equilibrium, which is often the case when random mating takes place, is an important problem in biology.

Given a random sample of size $n$ from the population, and observed $\{n_1, n_2, n_3\}$ individuals (with $n = n_1 + n_2 + n_3$) from each of the three possible genotypes $\{AA, aa, Aa\}$, the question is whether or not these data support the hypothesis of HW equilibrium.

This is an important example of *precise* hypothesis in the sciences, for HW equilibrium corresponds to a zero measure set within the original parameter space.

### 2.1.1.2 Statistical Formulation

Since $\sum_{i=1}^{3} \alpha_i = 1$, there are only two independent parameters. In terms of the population frequencies $\alpha_1$ and $\alpha_2$ of the two pure genotypes $AA$ and $aa$, the relevant statistical model is the trinomial

$$\mathrm{Tri}(n_1, n_2 | n, \alpha_1, \alpha_2) = \frac{n!}{n_1!\, n_2!\, (n - n_1 - n_2)!}\, \alpha_1^{n_1} \alpha_2^{n_2}\, (1 - \alpha_1 - \alpha_2)^{n - n_1 - n_2}$$

with $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, and $0 < \alpha_1 + \alpha_2 < 1$ and, in conventional language, it is required to test the null hypothesis

$$H_0 = \{(\alpha_1, \alpha_2);\ \alpha_1 = p^2, \alpha_2 = (1-p)^2,\ 0 < p < 1\}.$$

This is the parametric form of the equation of the line $\sqrt{\alpha_1} + \sqrt{\alpha_2} = 1$, represented with a solid line in Figure 2.1, and it is a set of zero measure within the parameter space, the simplex $\mathscr{A} = \{(\alpha_1, \alpha_2);\ 0 < \alpha_1 < 1,\ 0 < \alpha_2 < 1,\ 0 < \alpha_1 + \alpha_2 < 1\}$.

Testing a trinomial population for HW equilibrium is a problem that has received a fair amount of attention in the statistical literature. Main pointers include the frequentist analysis of Haldane (1954), an "exact" test based on the distribution $p(n_1, n_2 | H_0, n_1 - n_2, n)$, and the Bayesian analysis of Lindley (1988) who reparametrizes to

$$\psi(\alpha_1, \alpha_2) = \frac{1}{2} \log \frac{4\, \alpha_1\, \alpha_2}{(1 - \alpha_1 - \alpha_2)^2},$$

so that $\psi = 0$ when $H_0$ is true, and then obtains approximations to the posterior density of $\psi$, $\pi(\psi | n_1, n_2, n_3)$ for a range of different prior choices.

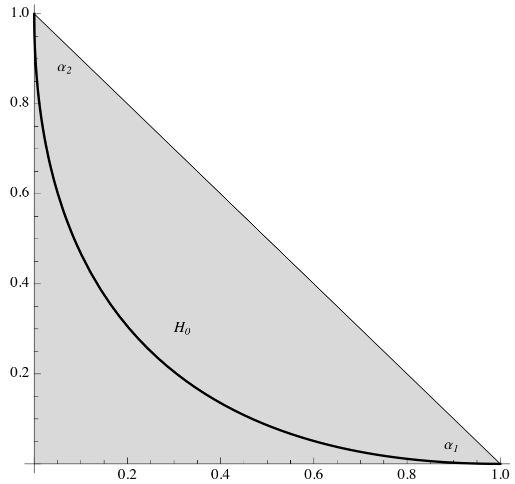

FIGURE 2.1. Precise null (solid line) within the parameter space (shaded region).

## 2.1.2 *Objective Precise Bayesian Testing*

### 2.1.2.1 The Decision Problem and the Intrinsic Loss Function

If data $\mathbf{z}$ are assumed to have been generated from the probability model $\mathscr{M} \equiv \{p_{\mathbf{z}}(\cdot|\phi,\omega),\ \mathbf{z} \in \mathscr{Z},\ \phi \in \Phi,\ \omega \in \Omega\}$, then testing whether of not the observed data $\mathbf{z}$ are compatible with the precise hypothesis $H_0 = \{\phi = \phi_0\}$ may be seen as a simple decision problem with only two alternatives:

1. $a_0$: To accept $H_0$, and work *as if* data were generated from the *reduced* model $\mathscr{M}_0 \equiv \{p_{\mathbf{z}}(\cdot|\phi_0,\omega),\ \mathbf{z} \in \mathscr{Z},\ \omega \in \Omega\}$; and
2. $a_1$: To reject $H_0$, and keep working with the *assumed* model $\mathscr{M}$.

Foundations then dictate (see, e.g., Bernardo and Smith, 1994, Chapter 2 and references therein) that one must

1. Specify a loss function $\ell\{a_i,(\phi,\omega)\}$, $i = 0,1$.
2. Specify a prior function $p(\phi,\omega)$, on $\Phi \times \Omega$, and use Bayes to obtain
$$p(\phi,\omega|\mathbf{z}) \propto p(\mathbf{z}|\phi,\omega)\, p(\phi,\omega).$$
3. Reject $H_0$ if, and only if, $l(a_0|\mathbf{z}) > l(a_1|\mathbf{z})$, where
$$l(a_i|\mathbf{z}) = \int_\Phi \int_\Omega \ell\{a_i,(\phi,\omega)\} p(\phi,\omega|\mathbf{z}) d\phi d\omega.$$

One should then reject $H_0$ if, and only if, $l(a_0|\mathbf{z}) > l(a_1|\mathbf{z})$, hence if, and only if,

$$\int_\Phi \int_\Omega \left[\,\ell\{a_0,(\phi,\omega)\} - \ell\{a_1,(\phi,\omega)\}\right] p(\phi,\omega|\mathbf{z}) d\phi d\omega > 0,$$

which only depends on the loss increase from rejecting $H_0$, given by

$$\Delta(\phi,\omega) = \ell\{a_0,(\phi,\omega)\} - \ell\{a_1,(\phi,\omega)\}.$$

Without loss of generality, the loss increase $\Delta(\phi,\omega)$ may be written in the form $\delta\{\phi_0,(\phi,\omega)\} - d_0$, where

1. $\delta\{\phi_0,(\phi,\omega)\}$ is the non-negative terminal loss to be suffered by accepting $\phi = \phi_0$ as a function of $(\phi,\omega)$; and
2. $d_0$ is the strictly positive utility of accepting $H_0$ when it is true.

With this notation, one should reject the null if, and only if

$$\int_{\Phi}\int_{\Omega} \delta\{\phi_0,(\phi,\omega)\}\, p(\phi,\omega|\mathbf{z})\, d\phi d\omega > d_0,$$

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 $\psi = \psi(\phi)$ the conditions to reject $\phi = \phi_0$ should certainly be *precisely the same* as the conditions to reject $\psi = \psi(\phi_0)$ (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 $\mathscr{F} \equiv \{p_{\mathbf{z}}(\cdot|\psi), \psi \in \Psi\}$, the loss suffered from using an estimate $\tilde{\psi}$ is of the form

$$\ell(\tilde{\psi},\psi) = \delta\{p_{\mathbf{z}}(\cdot|\tilde{\psi}), p_{\mathbf{z}}(\cdot|\psi)\},$$

defined in terms of the discrepancy of $p_{\mathbf{z}}(\cdot|\tilde{\psi})$ from $p_{\mathbf{z}}(\cdot|\psi)$, rather than on the discrepancy of $\tilde{\psi}$ from $\psi$. 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:

$$\delta\{\phi_0,(\phi,\omega)\} = \inf_{\omega_0 \in \Omega} \int_{\mathscr{Z}} p(\mathbf{z}|\phi,\omega) \log \frac{p(\mathbf{z}|\phi,\omega)}{p(\mathbf{z}|\phi_0,\omega_0)}\, d\mathbf{z}.$$

This may be also be described as the minimum (Kullback-Leibler) logarithmic divergence of $\mathscr{M}_0$ from the assumed model.

#### 2.1.2.2 Reference Analysis and Precise Hypothesis Testing

Given a model $\mathscr{M} \equiv \{p_{\mathbf{z}}(\cdot|\phi,\omega),\ \mathbf{z} \in \mathscr{Z},\ \phi \in \Phi,\ \omega \in \Omega\}$, the $\theta$-reference prior function $\pi_\theta(\phi,\omega)$ (see Bernardo, 2005, and references therein) is that which maximizes the missing information about $\theta = \theta(\phi,\omega)$. The corresponding marginal reference posterior $\pi(\theta|\mathbf{z})$ 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) to test $H_0 \equiv \{\phi = \phi_0\}$ is the solution to the hypothesis testing decision problem corresponding to the *intrinsic loss* and the relevant *reference prior*. It only requires computing the *intrinsic test statistic*, defined as the reference posterior expectation,

$$d(H_0|\mathbf{z}) = \int_0^\infty \delta\ \pi(\delta|\mathbf{z}) d\delta,$$

of the intrinsic discrepancy loss $\delta(\phi,\omega) = \delta\{\phi_0, (\phi,\omega)\}$, which is in this case of the quantity of interest.

The intrinsic test statistic is a direct *measure of evidence against* $H_0$, in a *log-likelihood ratio scale*, which is independent of the sample size, the dimensionality of the problem, and the parametrization used. For further details and many examples, see Bernardo (2005) and references therein.

### *2.1.3 Testing for Hardy-Weinberg Equilibrium*

#### 2.1.3.1 The Quantity of Interest

Within the trinomial model,

$$\text{Tri}\{n_1, n_2|n, \alpha_1, \alpha_2\} = \frac{n!}{n_1!\, n_2!\, (n-n_1-n_2)!}\, \alpha_1^{n_1} \alpha_2^{n_2}\, (1-\alpha_1-\alpha_2)^{n-n_1-n_2},$$

the logarithmic divergence of a member $\text{Tri}\{n_1, n_2|n, p_0^2, (1-p_0)^2\}$ of the null

$$H_0 = \{(\alpha_1, \alpha_2);\ \alpha_1 = p^2, \alpha_2 = (1-p)^2, \quad 0 < p < 1\}$$

from the assumed model $\text{Tri}\{n_1, n_2|n, \alpha_1, \alpha_2\}$ is

$$k\{p_0|\alpha_1, \alpha_2\} = E_{(n_1,n_2|\alpha_1,\alpha_2)}\left[\log \frac{\text{Tri}\{n_1, n_2|n, \alpha_1, \alpha_2\}}{\text{Tri}\{n_1, n_2|n, p_0^2, (1-p_0)^2\}}\right]$$

which, after some algebra, reduces to

$$n[(\alpha_2-\alpha_1-1)\log(p_0) + (\alpha_1-\alpha_2-1)\log(1-p_0) - (1-\alpha_1-\alpha_2)\log(2) - \text{H}\{\alpha\}],$$

where $\text{H}\{\alpha\} = -\alpha_1 \log\alpha_1 - \alpha_2 \log\alpha_2 - (1-\alpha_1-\alpha_2)\log(1-\alpha_1-\alpha_2)$ is the entropy of $\alpha = \{\alpha_1, \alpha_2, 1-\alpha_1-\alpha_2\}$. The last expression is minimized, for $0 < p_0 < 1$, when $p_0 = (1+\alpha_1-\alpha_2)/2$, and substitution yields the intrinsic loss function,

$$\delta\{H_0, (\alpha_1, \alpha_2)\} = \inf_{0<p_0<1} k\{p_0|\alpha_1, \alpha_2\} = n\,\theta(\alpha_1, \alpha_2),$$

where

$$\theta(\alpha_1,\alpha_2) = 2\,\mathrm{H}\{\omega, 1-\omega\} - \mathrm{H}\{\alpha_1,\alpha_2,1-\alpha_1-\alpha_2\} - (1-\alpha_1-\alpha_2)\log(2),$$

and $\omega = \omega(\alpha_1,\alpha_2) = (1+\alpha_1-\alpha_2)/2$ is the value of $p$ for a trinomial population $\mathrm{Tri}\{n_1,n_2|n,p^2,(1-p)^2\}$ in HW equilibrium which is closest, in the logarithmic divergence sense, to the trinomial population $\mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\}$. The function $\delta\{H_0,(\alpha_1,\alpha_2)\}$ measures the discrepancy of the null from the trinomial model $\mathrm{Tri}\{\cdot|n,\alpha_1,\alpha_2\}$.

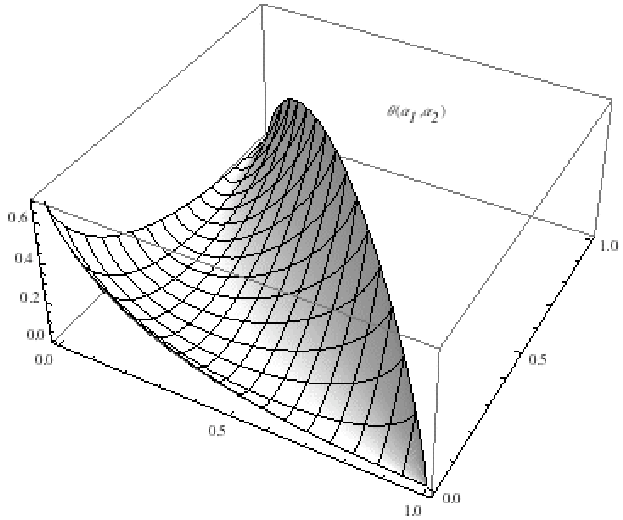

FIGURE 2.2. The quantity of interest, $\theta = \theta(\alpha_1,\alpha_2)$.

The quantity of interest in this problem is clearly the function $\theta = \theta(\alpha_1,\alpha_2)$ since $\delta\{H_0,(\alpha_1,\alpha_2)\} = n\,\theta(\alpha_1,\alpha_2)$ precisely measures how far the null $H_0$ is from the assumed model. In particular, the population is in HW equilibrium if, and only if, $\theta = 0$, in which case, $\sqrt{\alpha_1}+\sqrt{\alpha_2} = 1$ or $\alpha_2 = (1-\sqrt{\alpha_1})^2$. Figure 2.2 provides a 3D plot of the surface $\theta(\alpha_1,\alpha_2)\}$. It is zero for all HW equilibrium values and achieves its maximum value, $\log(2)$, at both $(0,0)$ and $(1/2,1/2)$. Hence, in this problem, the intrinsic loss is a bounded function.

### 2.1.3.2 The Reference Prior

To obtain the joint reference prior $\pi_\theta(\alpha_1,\alpha_2)$ when $\theta = \theta(\alpha_1,\alpha_2)$ is the quantity of interest, a complementary parameter $\omega = \omega(\alpha_1,\alpha_2)$ must be chosen, so that $(\theta,\omega)$ is a one-to-one transformation of $(\alpha_1,\alpha_2)$. A convenient choice is the function $\omega(\alpha_1,\alpha_2) = (1+\alpha_1-\alpha_2)/2$, which occurs in the expression of $\delta\{H_0,(\alpha_1,\alpha_2)\}$ obtained above. The reference prior in this parametrization when $\theta$ is the parameter of interest is then obtained as $\pi_\theta(\theta,\omega) = \pi(\omega|\theta)\,\pi(\theta)$. Finally, the required reference prior in the original parametrization is obtained as

$$\pi_\theta(\alpha_1,\alpha_2) = |\mathrm{J}(\alpha_1,\alpha_2)|\,\pi_\theta(\theta(\alpha_1,\alpha_2),\omega(\alpha_1,\alpha_2)),$$

where $\mathrm{J}(\alpha_1,\alpha_2) = \left(\frac{\partial\theta\;\partial\omega}{\partial\alpha_1\,\partial\alpha_2}\right)$ is the corresponding Jacobian matrix.

The required transformation, represented in Figure 2.6, is delicate. Indeed, the Jacobian determinant $|\mathrm{J}(\alpha_1,\alpha_2)| = \log(1-\alpha_1-\alpha_2) - \frac{1}{2}\log(4\alpha_1\,\alpha_2)$ is null at the HW line, positive below, negative above, and diverges at the simplex borders. A one-to-one transformation is only obtained in each of the two separate regions defined by the equilibrium line. Thus a one-to-one transformation is $\{\alpha_1,\alpha_2\} \iff \{\theta,\omega,\lambda\}$ where $\lambda \in \{1,2\}$ indicates region, with $\lambda = 1$ when $\sqrt{\alpha_1}+\sqrt{\alpha_2} < 1$, and $\lambda = 2$ when $\sqrt{\alpha_1}+\sqrt{\alpha_2} > 1$. Formally,

$$\pi_\theta(\alpha_1,\alpha_2) = \pi_\theta(\alpha_1,\alpha_2|\lambda=1) + \pi_\theta(\alpha_1,\alpha_2|\lambda=2).$$

The joint reference priors in each of the two regions must be be separately computed.

This model is regular. Hence, the reference prior $\pi(\omega|\theta)\,\pi(\theta)$ may be found in terms of the relevant Fisher information matrix. In the original parametrization, the inverse of Fisher matrix $\mathrm{F}_1$ is

$$\mathrm{F}_1^{-1}(\alpha_1,\alpha_2) = \begin{pmatrix} \alpha_1(1-\alpha_1) & -\alpha_1\,\alpha_2 \\ -\alpha_1\,\alpha_2 & \alpha_1(1-\alpha_1) \end{pmatrix},$$

so that, in the new parametrization, Fisher matrix is $F_2$ such that

$$\mathrm{F}_2^{-1}(\theta,\omega) = \mathrm{J}(\alpha_1,\alpha_2)\cdot\mathrm{F}_1^{-1}(\alpha_1,\alpha_2)\cdot\mathrm{J}^t(\alpha_1,\alpha_2),$$

evaluated with the inverse functions $\alpha_1(\theta,\omega)$ and $\alpha_2(\theta,\omega)$. Fisher matrix $\mathrm{F}_2$ has a complex, but analytical expression, in terms of $\alpha_1$ and $\alpha_2$, but the inverse functions $\alpha_i(\theta,\omega)$ must be numerically computed.

The reference prior $\pi(\omega|\theta)\,\pi(\theta)$ may be found in terms of $\mathrm{H} = \mathrm{F}_2$ and $\mathrm{V} = \mathrm{F}_2^{-1}$ (Berger and Bernardo, 1992a), from

$$\pi(\omega|\theta) \propto h_{22}^{1/2}(\theta,\omega)$$

and

$$\pi(\theta) \propto \exp\left[\int_{\Omega(\theta)} \pi(\omega|\theta)\log\{v_{11}^{-1/2}(\theta,\omega)\}\,d\omega\right].$$

**Lower region:** $R_1 = \{(\alpha_1,\alpha_1);\ \sqrt{\alpha_1}+\sqrt{\alpha_2} \le 1\}$**.** The reference conditional priors are numerically found to be approximate the Beta densities (see Figure 2.3)

$$\pi_1(\omega|\theta) \approx \frac{1}{\omega_1(\theta)-\omega_0(\theta)}\,\mathrm{Be}\left(\frac{\omega-\omega_0(\theta)}{\omega_1(\theta)-\omega_0(\theta)}\,\middle|\,\frac{1}{2},\frac{1}{2}\right),\quad \omega_0(\theta) < \omega < \omega_1(\theta),$$

where $\omega_0(\theta)$ and $\omega_1(\theta)$ are respectively the inverse functions of

$$\theta_1(\omega) = (2\omega - 1)\log(2\omega - 1) - 2\omega\log(\omega), \quad 1/2 < \omega < 1,$$

$$\theta_0(\omega) = (1 - 2\omega)\log(1 - 2\omega) - 2(1 - \omega)\log(1 - \omega), \quad 0 < \omega < 1/2.$$

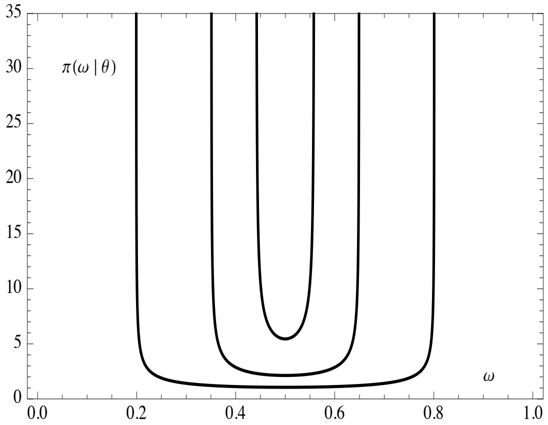

FIGURE 2.3. Conditional reference priors of $\omega \in (\omega_0(\theta), \omega_1(\theta))$, in the lower region of the parameter space, for $\theta = 0.05, 0.20$ and $0.40$.

Using the analytical approximation for the conditional reference priors, the marginal reference prior for the quantity of interest results

$$\pi_1(\theta) \approx \frac{1}{\log(2)}\,\text{Be}\left(\frac{\theta}{\log(2)}\,\middle|\,\frac{1}{2}, \frac{1}{2}\right), \quad 0 < \theta < \log(2).$$

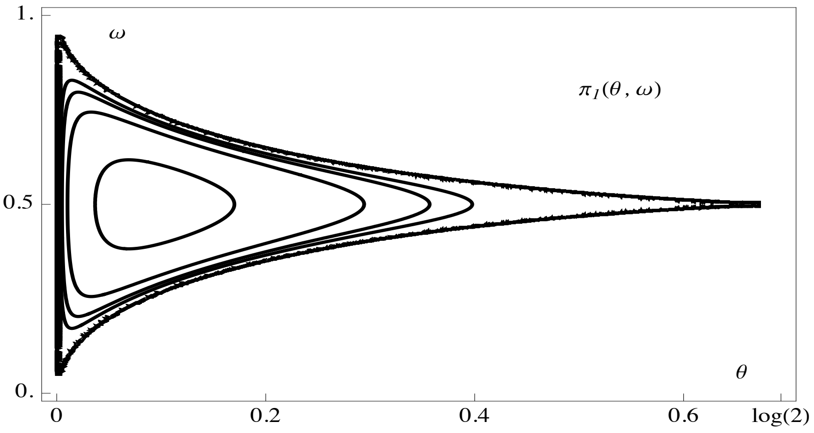

FIGURE 2.4. Contour plot of the joint reference prior $\pi_1(\theta, \omega)$ in the lower region.

The joint reference prior for this region is then $\pi_1(\theta,\omega) = \pi_1(\omega|\theta)\,\pi_1(\theta)$. The contour plot of this joint refernce prior is shown in Figure 2.4. Notice that these reference priors are all proper.

**Upper region:** $R_2 = \{(\alpha_1,\alpha_1);\ \sqrt{\alpha_1}+\sqrt{\alpha_2} \le 1\}$**.** Similarly, in the region over the HW equilibrium line, the reference conditional priors are numerically found to be

$$\pi_2(\omega|\theta) \approx \frac{1}{\omega_1(\theta)-\omega_0(\theta)}\,\mathrm{Be}\left(\frac{\omega-\omega_0(\theta)}{\omega_1(\theta)-\omega_0(\theta)}\,\middle|\,\frac{1}{2},\frac{1}{2}\right), \quad \omega_0(\theta) < \omega < \omega_1(\theta),$$

where $\omega_1(\theta)$ and $\omega_0(\theta)$ are respectively the inverse functions of

$$\theta_1(\omega) = -\omega\log(\omega) - (1-\omega)\log(1-\omega), \quad 1/2 < \omega < 1$$

$$\theta_0(\omega) = -\omega\log(\omega) - (1-\omega)\log(1-\omega), \quad 0 < \omega < 1/2.$$

The marginal reference prior for $\theta$ in the upper region is

$$\pi_2(\theta) \approx \frac{1}{\log(2)}\,\mathrm{Be}\left(\frac{\theta}{\log(2)}\,\middle|\,\frac{1}{2},\frac{1}{2}\right), \quad 0 < \theta < \log(2).$$

The joint reference prior for the upper region is then $\pi_2(\theta,\omega) = \pi_2(\omega|\theta)\,\pi_2(\theta)$. Again, all these reference priors are all proper.

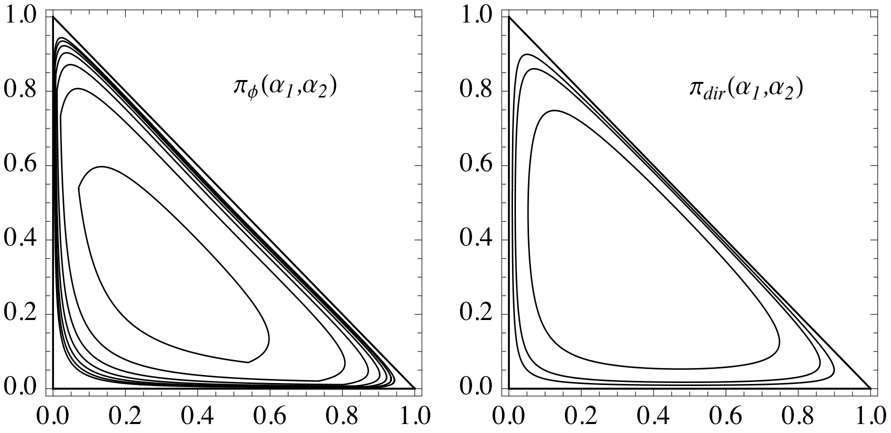

FIGURE 2.5. Contour plots of the joint reference prior in the original parametrization and a Dirichlet density with parameter $(1/3,1/3,1/3)$.

**Joint reference prior in the original parametrization.** Returning to the original parametrization and combining the results from the two regions produces $\pi_\theta(\alpha_1,\alpha_2)$, whose contour plot is represented in the left panel of Figure 2.5. For comparison, the right panel represents the contour plot of a Dirichlet density with parameter vector $(1/3,1/3,1/3)$. This could be used as an approximation if exact computation is not needed.

#### 2.1.3.3 Posterior Inference: Estimation and Testing

**Joint reference posterior.** For any data set, $\{n_1,n_2,n_3\}$, where $n_1$ and $n_2$ are respectively the number of observed pure genotypes *AA* and *aa*, and $n_3$ is the number of observed mixed genotypes *Aa*, the joint reference posterior is

$$\pi_\theta(\alpha_1,\alpha_1|n_1,n_2,n_3) = c(n_1,n_2|n)\ \mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\}\ \pi_\theta(\alpha_1,\alpha_2),$$

where $n = n_1+n_2+n_3$ and

$$c(n_1,n_2|n) = \int_0^1 \left\{ \int_0^{1-\alpha_1} \mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\}\ \pi_\theta(\alpha_1,\alpha_2)\, d\alpha_2 \right\} d\alpha_1,$$

a delicate numerical integral given the prior shape.

The posterior probabilities of the two non-equilibrium regions are

$$P[R_1|n_1,n_2,n_3] = \int_0^1 \left\{ \int_0^{(1-\sqrt{\alpha_1})^2} \pi_\theta(\alpha_1,\alpha_1|n_1,n_2,n_3)\, d\alpha_2 \right\} d\alpha_1,$$

and $P[R_2|n_1,n_2,n_3] = 1 - P[R_1|n_1,n_2,n_3]$.

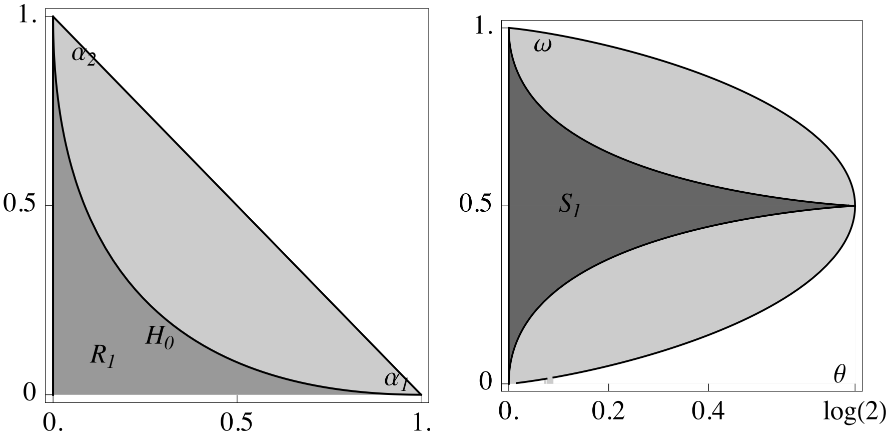

FIGURE 2.6. Original and transformed parameter spaces.

Since the transformation between $(\alpha_1,\alpha_2)$ and $(\theta,\omega)$ is not one-to-one, computing the joint posterior density in terms of the $(\theta,\omega)$ requires identification of the two possible inverse values $\alpha_1(\theta,\omega)$ and $\alpha_2(\theta,\omega)$. This is done in terms of

$S_1 = \mathrm{Image}(R_1)$, where $R_1$ is the region below $H_0$, and $S_2 = \mathrm{Image}(R_2)$, where $R_2$ is the region above $H_0$. Thus, if $(\theta, \omega) \in S_1$, which is contained in $S_2$), then there are two diffferent pairs of $(\alpha_1, \alpha_2)$ values which map into $(\theta, \omega)$ (see Figure 2.6).

It follows that, for any data $\mathbf{z} = \{n_1, n_2, n_3\}$,

$$\pi(\theta, \omega|\mathbf{z}) = \pi(\theta, \omega|\mathbf{z}, S_1)P(R_1|\mathbf{z}) + \pi(\theta, \omega|\mathbf{z}, S_2)P(R_2|\mathbf{z})$$

$$\pi(\theta, \omega|\mathbf{z}, S_i) = \frac{\pi(\alpha_1, \alpha_2|\mathbf{z}, R_i)}{|\mathrm{J}(\alpha_1, \alpha_2)|}, \quad \alpha_j \to \alpha_{ji}(\theta, \omega), \quad i = 1, 2,$$

where $\{\alpha_{1i}, \alpha_{2i}\}$ is the inverse function mapping $S_i$ into $R_i$.

The required marginal reference posterior for the quantity of interest $\theta$ is then

$$\pi(\theta|\mathbf{z}) = \int_{\Omega(\theta)} \pi(\theta, \omega|\mathbf{z})\, d\omega.$$

This will concentrate on its extreme value $\theta = 0$ if, and only if, the population is in approximate HW equilibrium.

**Intrinsic test statistic.** As described in Section 2.1.2, the intrinsic test statistic $d(H_0|\mathbf{z})$ is the reference posterior expectation of $\delta\{H_0, (\alpha_1, \alpha_2)\}$, defined as the minimum logarithmic divergence of the null model from the true model. Since $\delta\{H_0, (\alpha_1, \alpha_2)\} = n\, \theta(\alpha_1, \alpha_2)$, the intrinsic statistic is simply

$$d(H_0|\mathbf{z}) = n \int_0^{\log(2)} \theta\ \pi(\theta|\mathbf{z})\, d\theta = n\, E[\theta|\mathbf{z}],$$

the reference posterior expectation of the quantity of interest times the sample size. This is precisely the reference posterior expectation of the log-likelihood ratio against the null and, therefore, $d(H_0|\mathbf{z})$ has an immediate meaning as an objective measure of the evidence against the null provided by the data.

## 2.1.4 Examples

### 2.1.4.1 Simulations

**Data simulated under HW equilibrium.** A trinomial sample of size $n = 30$ from a population in HW equilibrium was simulated with $P[A] = p = 0.3$, so that $\{\alpha_1, \alpha_2\} = \{p^2, (1-p)^2\} = \{0.09, 0.49\}$, $\omega = p = 0.3$, and $\theta = 0$. The simulation yielded $\{n_1, n_2, n_3\} = \{2, 15, 13\}$.

Figure 2.7 represents the marginal reference posterior of $\delta = n\,\theta$ which, as expected, concentrates around the null value $\delta = 0$, with $d(H_0|\mathbf{z}) = n$, $E[\theta|\mathbf{z}] = 0.321 = \log(1.38)$, so that the likelihood ratio against the null is expected to be only about 1.38, and the null is accepted: one may safely proceed as if the population where in HW equilibrium, suggesting random mating.

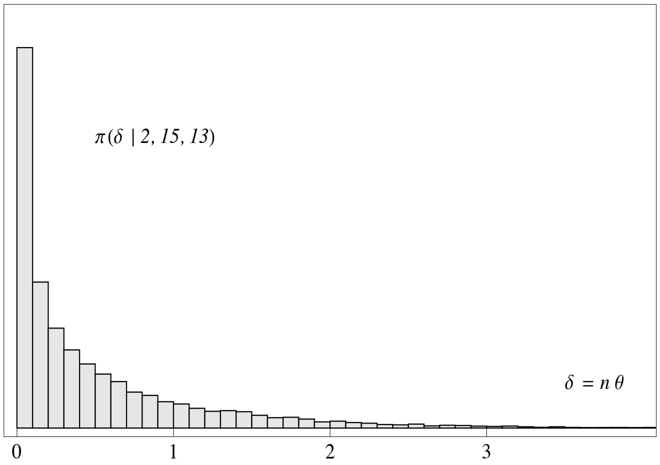

FIGURE 2.7. Reference posterior distribution of $\delta = n\,\theta$ with data simulated from a population in HW equilibrium.

**Data simulated under non-HW equilibrium.** A trinomial sample of size $n = 30$ was simulated with $\{\alpha_1, \alpha_2\} = \{0.45, 0.40\}$, so that $\omega = 0.525$, $\theta = 0.269$, and the population is *not* in HW equilibrium. The simulation then yielded $\{n_1, n_2, n_3\} = \{12, 12, 6\}$.

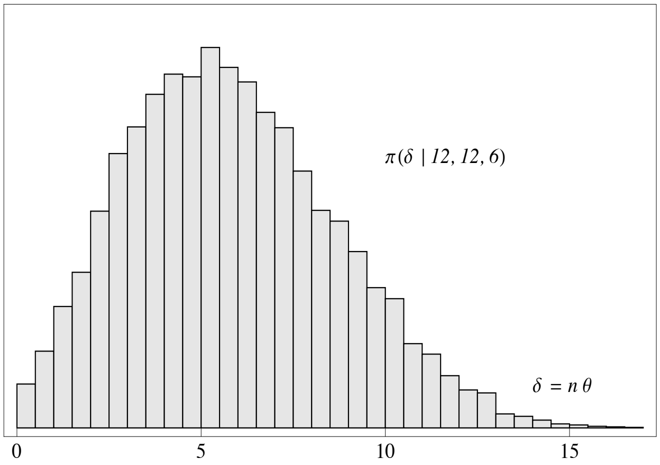

FIGURE 2.8. Reference posterior distribution of $\delta = n\,\theta$ with data simulated from a population not in HW equilibrium.

As Figure 2.8 illustrates, the marginal reference posterior of $\delta = n\,\theta$ has an interior mode, $d(H_0|\mathbf{z}) = n$, and $E[\theta|\mathbf{z}] = 5.84 \approx \log(344)$, so that the likelihood ratio against the null is expected to be about 344. Thus, the null should certainly be *rejected*, and one should work under the assumption that the population is *not* in HW equilibrium, thus suggesting non random mating.

### 2.1.4.2 An Example from the Literature

**Lindley data.** Lindley (1988) analyzed the data $\mathbf{z} = \{0, 90, 10\}$ from a Bayesian viewpoint, noting that asymptotic results are scarcely satisfactory in this case, and performing an analysis of the clear dependence of the results on the prior chosen. This could be expected, for these data are somewhat extreme due to the fact that there are no observations from the pure *AA* genotype. Conclusions from extreme data are often very sensitive to the prior, and they cannot be usually be well approximated with asymptotic arguments.

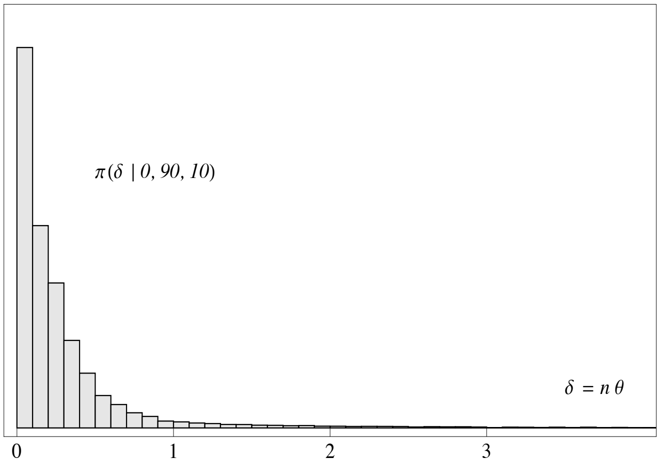

FIGURE 2.9. Marginal reference posterior distribution of $\delta = n\,\theta$ for Lindley (1988) data.

Reference analysis has been known to perform fine in many other problems with extreme data. This provides yet another example. As Figure 2.9 illustrates, it is found that $\pi(\delta|\mathbf{z})$, the reference posterior density of the expected discrepancy from the null is again very concentrated around the null value $\delta = 0$. Indeed, $d(H_0|\mathbf{z}) = n$, $E[\theta|\mathbf{z}] = 0.51 = \log(1.66)$ and hence the likelihood ratio against the null may the expected to be just about 1.66.

We must therefore conclude that the HW equilibrium hypothesis is compatible with these data.