Non-parametric Inference about an Unknown Mean using the Imprecise Dirichlet Model

Jean-Marc Bernard
Université Paris 8 & CNRS ESA 7021
berj@univ-paris8.fr

Abstract

We observe a random sample of n observations from an unknown distribution having mean μ. In this paper we consider the problem of making inferences about the unknown parameter μ and other characteristics of the unknown distribution by adopting a non-parametric and objective-minded framework, in which the observations are considered bounded and discrete. We first review the Bayesian approach based on Dirichlet priors and discuss the problems encountered by the usual vague priors. An alternative approach is then proposed which models prior ignorance by an imprecise Dirichlet model (IDM) with parameter ν (Walley, 1996). The comparison of inferences produced by the IDM with the ones from alternative objective Bayesian models gives support for chosing $\nu \leq 2$ in the IDM. The new method of inference is illustrated on Darwin's maize data.

Keywords. Bayesian inference, Dirichlet distribution, L-Dirichlet distribution, Prior ignorance, IDM, Upper and lower probabilities.

1 Introduction

Consider that we observe a random sample of n observations $(x_i)_{i=1,\dots,n}$ from an unknown parent distribution having mean μ. Here we consider the problem of making inferences about the unknown parameter μ adopting a point of view different from the usual ones.

The most common approaches to this problem — the parametric ones —, assume the normality of the parent distribution, that is consider the x_i's as independently distributed $N(\mu, \sigma^2)$, where μ and σ are unknown. Under that assumption, the statistic $t = (m - \mu)\sqrt{n}/s$, where m and s^2 are the sample mean and corrected variance respectively, follows a Student's t distribution with $(n-1)$ degrees of freedom. In the frequentist approach to inference, hypothesis tests about μ and confidence intervals for μ are derived from that sampling distribution. The objective Bayesian approach assumes a vague prior on (μ, σ^2), with pdf $p(\mu, \sigma) \propto \sigma^{-1}$ (Jeffreys, 1961; Box & Tiao, 1973, Section 2.4; Bernardo & Smith, 1994, Example 5.17). With that prior, the posterior distribution for μ is a scaled $t(m, s^2/n)$ with $(n-1)$ degrees of freedom, which provides straightforward probabilistic reinterpretations of the frequentist p-values.

Although the preceding solutions to the problem are deeply rooted in statistical practice, we consider that they are not realistic in most real problems. The three underlying assumptions that the observations x_i are (i) unbounded, (ii) continuous and (iii) normal, all seem debatable in most practical contexts. Instead we believe that almost all numerical data are actually bounded, discrete and thus non-normal.

Regarding boundedness, it is typical to consider some of the most extreme observations as outliers and, often then, to discard them from the analysis. Typically, a priori bounds on possible values exist in most contexts. Ericson (1969, p. 206) insists on "... the inherent discreteness of almost all observation due to the limitation of measuring instruments, etc.". Geisser (1984, p. 250) claims "... that in all experimentation the observable are essentially discrete and our continuous models are merely convenient, though useful, approximations to discrete situations." In our view too, the continuity assumption appears to be motivated by mathematical convenience rather than by realistic considerations. More specifically, we believe that the normality assumption has played an important role in the early developments of statistical inference primarily because of its mathematical tractability and the initial lack of computing devices. But, nowadays, integrating such a constraint has become unnecessary, and alternative inferential methods can now be envisaged in which the criteria of relevance and of realism should be the only ones. This aim is one of the motivations for this paper.

An early attempt to consider our problem from a

Bayesian non-parametric viewpoint, free from any assumptions, is the article of Bloch & Watson (1967) who consider discrete bounded data as a multinomial sample from an infinite population and use a Dirichlet prior to describe prior uncertainty. Another closely related approach is the one that considers the case of a finite population (Ericson, 1969), which merely amounts to considering the predictive problem of making inferences about the mean of n' future real-valued observations from the n observed ones, using a Dirichlet-multinomial prior on the $N = n + n'$ combined observations. The two models, Dirichlet prior with an infinite population and Dirichlet-multinomial prior with a finite population of size N, are related in the sense that each one can be derived from the other. What is remarkable is that, under these two models, some inferences produced by some of the usual vague priors are extremely close to the ones obtained from the parametric normal setting, as if the assumption of normality were almost irrelevant. One problem, however, is that, since the number of categories K can be very large, the several recommended vague priors for multinomial data (Haldane, 1948; Perks, 1947; Jeffreys, 1946; or Bayes-Laplace's uniform prior) may lead to considerably different inferences.

Another approach to Bayesian non-parametric inference is the one based on the Dirichlet process prior. However, since it assumes that the observations are continuous and unbounded, it is subject to the same criticisms previously addressed to the normal model and thus of little relevance to the discrete view adopted here.

Walley (1996) recently proposed a new method of inference for categorical data based on the imprecise Dirichlet model (IDM). This imprecise probability model can be viewed as a generalization of Bayesian conjugate inference for categorical data, in which prior uncertainty is described by a set of Dirichlet priors, each of which being updated into a Dirichlet posterior using Bayes' theorem. The IDM is devised to model a prior state of near ignorance and is thus objective-minded. Posterior uncertainty is described by the set of Dirichlet posterior distributions. As in every imprecise probability model, inferences can be summarized by giving upper and lower posterior probabilities of events, or upper and lower posterior cdfs, expectations, variances, *etc.* of parameters of interest. Any of these upper and lower quantities is obtained by maximization or minimization over the set of Dirichlet posterior distributions.

The IDM has proved particularly valuable when the number of categories is large, especially as compared to precise Bayesian models (see, *e.g.*, Bernard, 2000). A thorough analysis of predictive inferences from the IDM is done in Walley & Bernard (1999). For our present non-parametric problem, Walley (1996, Section 6) suggests that the IDM could be used to make non-parametric inferences from real-valued observations made with finite precision measurement. Some simple results for the upper and lower cdfs of a single future observation are given. Geyer (1996) in his discussion of Walley's (1996) paper notes that the IDM can be used to make inferences about an unknown mean, though implicitly introduces the unnecessary and unrealistic constraint that all future observations have values among the ones already observed.

The aim of the present paper is simply to extend the work of Bloch & Watson (1967) and of Ericson (1969) in following Walley's (1996) suggestion of using the IDM in this context. The IDM has several desirable properties as a model for making inferences from a prior state of near ignorance. The two major properties are that *(i)* it satisfies the *likelihood principle (LP)*, and *(ii)* it satisfies the *representation invariance principle (RIP)*. The RIP states that inferences do not depend on the way observations are divided into categories as long as the set of categories retained is fine enough for the definition of some parameter of interest. With the exception of Haldane's model, which can be viewed as a limiting case of the IDM, the usual objective Bayesian models do not satisfy these principles (Walley, 1996; Walley & Bernard, 1999). Despite some of its good properties, Haldane's model is also unsatisfactory because it is too "data-glued" and usually produces too strong inferences, especially if data have a null variance (*e.g.* Bernard, 1996). These general properties of the IDM, and the fact that the usual Bayesian objective models vary too much in the inferences they produce when the number of categories K is large, make the IDM a good candidate as an alternative general method of inference. However the IDM depends on a parameter ν, which basically expresses the extent of imprecision, and, in our view, the problem of chosing a specific value for ν is not settled yet. Walley (1996, 2000)[1] and Walley & Bernard (1999) give several arguments for chosing ν between 1 and 2, but most of them rely on the case $K = 2$. One major issue is then to see if the case of a possibly large K also provides arguments in favour of the same value(s) or not. The present study of making inferences from real-valued discrete and bounded data appears to be a privileged problem to consider for that goal.

The paper is organized as follows. Section 2 presents our basic discrete model. Section 3 mostly reviews earlier work on the Bayesian non-parametric approach to the problem (Bloch & Watson, 1967; Ericson, 1969), but provides simpler, more compact, formu-

[1] The parameter ν is denoted s in Walley (1996)

las. It also studies in more detail the core distribution involved, the *L-Dirichlet* distribution. The IDM is introduced in Section 4 and compared with parametric approaches. Finally, Section 5 provides the analysis of Darwin's maize data by our method.

2 Our basic discrete model

Our basic discrete model considers that each observation $x_i, i = 1, \ldots, n$, can only take one value in the finite set of reals $\mathcal{Y} = \{y_1, \ldots, y_K\}$. Another way of seeing this is to consider that the observations are categorized into K categories, and that each category k is associated with a precise numerical value y_k. We denote a_k the number of observations falling in category k, with $\sum_k a_k = n$.[2] Some values in $\mathcal{Y}$ may not be observed, so some a_k's may be zero. Without loss of generality, we assume that the y_k's are written in increasing order, *i.e.* $y_1 < y_2 < \cdots < y_K$. We denote $D_y = y_K - y_1$ the length of $\mathcal{Y}$. One important particular case, likely to be quite common in practice, is the case of equidistant values, $y_{k+1} - y_k = d_y$ for all $k \in [1, K-1]$ where d_y can be thought of as the fixed precision of the observations.

We are interested in an arbitrarily large population of such observations. The population is characterized by some unknown *parent (or true) proportions* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ of observations falling in categories $[1, \ldots, K]$, *i.e.* having value $y_1, \ldots, y_K$. The vector $\boldsymbol{\theta}$ is an unknown parameter of the model, and belongs to the $(K-1)$-dimensional unit simplex, $\mathcal{S}(1, K)$, *i.e.* each $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. The two vectors $\boldsymbol{y}$ and $\boldsymbol{\theta}$ define the *parent distribution*, that is the distribution of $\boldsymbol{y}$ with normalized weights $\boldsymbol{\theta}$.

We observe a sample of n observations from that population, among which a_k observations fall in category k, *i.e.* have value y_k. The vector of data counts is denoted by $\boldsymbol{a} = (a_1, \ldots, a_K)$, and $\boldsymbol{f} = \boldsymbol{a}/n$ is the vector of observed relative frequencies. The vector of data counts, $\boldsymbol{a}$ is assumed to be a multinomial sample of size n from the population:

$$\boldsymbol{a}|\boldsymbol{\theta} \quad \sim \quad Mn(n, \boldsymbol{\theta}). \tag{1}$$

The general problem is to make inferences about some characteristic of the parent distribution, *e.g.* its mean, $\mu = \sum_k \theta_k y_k$, its variance $\sigma^2 = \sum_k \theta_k (y_k - \mu)^2$, its cdf or its quantiles. Each of these unknown quantities is a derived parameter, *i.e.* a function of $\boldsymbol{\theta}$, since $\boldsymbol{y}$ is considered fixed and known. Though the model presented below enables one to study any of these derived parameters, we shall focus here on the study of the parent mean μ.

[2] In this paper, all sums and products with index k run from 1 to K.

3 The Bayesian approach

As indicated in the introduction, the IDM involves a set of Dirichlet priors, which is updated into a set of Dirichlet posteriors. Before introducing imprecision, it is thus necessary to examine in detail the Bayesian approach using a single Dirichlet prior.

3.1 Dirichlet model for θ

In the most common Bayesian approach, namely the Bayesian conjugate analysis, prior uncertainty about the unknown multidimensional parameter $\boldsymbol{\theta}$ is described by a Dirichlet prior distribution,

$$\boldsymbol{\theta} \quad \sim \quad Diri(\boldsymbol{\alpha}), \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and each hyperparameter $\alpha_k > 0$. We call the α_k's the *prior strengths* and $\nu = \sum_k \alpha_k$ the *total prior strength*. We shall often use an alternative parameterization in terms of ν and of the *prior (relative) frequencies* $\boldsymbol{\varphi} = \boldsymbol{\alpha}/\nu$, where $\boldsymbol{\varphi} \in \mathcal{S}^\star(1, K)$ and $\mathcal{S}^\star(1, K)$ denotes the interior of the unit simplex $\mathcal{S}(1, K)$.[3] In order to simplify the notations, we shall sometimes use $\alpha_k = 0$, $\varphi_k = 0$ and $\nu = 0$ in place of $\alpha_k \to 0$, $\varphi_k \to 0$ and $\nu \to 0$ respectively, when no difficulty arises from such a notation. The Dirichlet prior density at any point $\boldsymbol{\theta}$ is given by

$$h(\boldsymbol{\theta}) \quad = \quad \frac{\Gamma(\nu)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}. \tag{3}$$

The expectations, variances and covariances of components of $\boldsymbol{\theta}$ are given by $E(\theta_k) = \varphi_k$, $V(\theta_k) = \varphi_k(1-\varphi_k)/(\nu+1)$, and $Cov(\theta_j, \theta_k) = -\varphi_j \varphi_k/(\nu+1)$.

The posterior distribution on $\boldsymbol{\theta}$ is then an updated Dirichlet distribution, namely

$$\boldsymbol{\theta}|\boldsymbol{a} \quad \sim \quad Diri(\boldsymbol{\alpha} + \boldsymbol{a}). \tag{4}$$

3.2 Inference about μ using a Dirichlet prior

We first need to introduce some general notation. For any K-dimensional real-valued vector $\boldsymbol{y} = (y_1, \ldots, y_K)$ and positive weights $\boldsymbol{w} = (w_1, \ldots, w_K)$ whose sum, $\sum_k w_k$, is strictly positive, we call $(\boldsymbol{y}, \boldsymbol{w})$ a *(weighted) protocol*, and define as usually its *mean* and *variance* as

$$Mean(\boldsymbol{y}, \boldsymbol{w}) \quad = \quad M = \sum_k w_k y_k \;/\; \sum_k w_k \tag{5}$$

$$Var(\boldsymbol{y}, \boldsymbol{w}) \quad = \quad \sum_k w_k (y_k - M)^2 \;/\; \sum_k w_k. \tag{6}$$

[3] Note that observed quantities are denoted by latin letters (a_k, f_k, n) and the analog prior quantities by the corresponding greek letters $(\alpha_k, \varphi_k, \nu)$.

In the following, we shall often use several standard properties of $Mean(\cdot)$ and $Var(\cdot)$. In particular, we shall often use the fact that $(\boldsymbol{y}, \boldsymbol{w})$ and $(\boldsymbol{y}, \boldsymbol{p})$, where $\boldsymbol{p} = \boldsymbol{w} / \sum_k w_k$, have same mean and variance.

Now let us consider the derived unknown parameter μ defined as the weighted mean of the y_k's, with respective weights θ_k, that is

$$\mu \quad = \quad Mean(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_k \theta_k y_k. \tag{7}$$

Generally, we define a *L-Dirichlet* (*L-Diri*) distribution as follows.

Definition 1 (L-Dirichlet distribution) *Let* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ *be distributed* Diri($\boldsymbol{\alpha}$), *where* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ *is a vector of strictly positive reals, and let* $\mu = \sum_k \theta_k y_k$, *where* $\boldsymbol{y} = (y_1, \ldots, y_k)$ *is a vector of reals. Then* μ *follows a* L-Diri($\boldsymbol{\alpha}, \boldsymbol{y}$).

From this general definition, the prior and posterior distributions for μ are

$$\mu \quad \sim \quad L\text{-}Diri(\boldsymbol{y}, \boldsymbol{\alpha}) \tag{8}$$

$$\mu | \boldsymbol{a} \quad \sim \quad L\text{-}Diri(\boldsymbol{y}, \boldsymbol{\alpha} + \boldsymbol{a}) \tag{9}$$

3.3 Properties of the L-Dirichlet distribution

We need to examine some general properties of the L-Dirichlet distribution. Let us consider a random variable $\mu = Mean(\boldsymbol{y}, \boldsymbol{\theta})$ distributed $L\text{-}Diri(\boldsymbol{y}, \boldsymbol{\alpha})$. The distribution of μ is studied in Bloch & Watson (1967). Several of the properties given below are either given in that paper or can be derived from results in Ericson (1969). What is new here is the introduction of more compact notation and the derivation of the uniform distribution as a special limiting case.

3.4 L-Dirichlet in standard format

First, let us note that, for reasons of symmetry, if we consider any permutation $\pi(\cdot)$ of $\{1, 2, \ldots, K\}$, permuting simultaneously the components of $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{y}$ leaves the problem unchanged. Formally we have, $\mu = Mean(\boldsymbol{y}, \boldsymbol{\theta}) = Mean(\pi(\boldsymbol{y}), \pi(\boldsymbol{\theta}))$ and $L\text{-}Diri(\boldsymbol{y}, \boldsymbol{\alpha}) = L\text{-}Diri(\pi(\boldsymbol{y}), \pi(\boldsymbol{\alpha}))$.

Consider that the vector $\boldsymbol{y}$ contains two identical values, $y_j = y_k$, and define the $(K-1)$-dimensional vectors $\boldsymbol{\theta}'$, $\boldsymbol{\alpha}'$ and $\boldsymbol{y}'$ as follows: delete the k'th coordinate, replace θ_j by $\theta_j + \theta_k$ and α_j by $\alpha_j + \alpha_k$. Then again, the problem remains unchanged, that is formally $\mu = Mean(\boldsymbol{y}', \boldsymbol{\theta}')$ and $L\text{-}Diri(\boldsymbol{y}, \boldsymbol{\alpha}) = L\text{-}Diri(\boldsymbol{y}', \boldsymbol{\alpha}')$. The latter statement follows easily from the "pooling property" of the Dirichlet (*e.g.*, Bernard, 1997; Fang *et al.*, 1990, Thm 1.4).

From these two properties, it is clear that any *L-Diri* distribution can be reparameterized so as to satisfy $y_1 < y_2 < \cdots < y_K$. Such a *standard format* is assumed in the sequel of this paper.

3.4.1 Generalities, cdf

The L-Dirichlet distribution is continuous. The distribution of μ is bounded by $min\ (\boldsymbol{y}) = y_1$ and $max\ (\boldsymbol{y}) = y_K$. These bounds are obtained for the two extreme compositions of the population, $(\theta_1 = 1, \theta_2 = \cdots = \theta_K = 0)$ and $(\theta_1 = \cdots = \theta_{K-1} = 0, \theta_K = 1)$ respectively.

The cdf $F(m) = P(\mu \leq m)$ can be expressed in terms of independent Gamma variables:

$$F(m) \quad = \quad P(\sum_k c_k \gamma_k \leq 0), \tag{10}$$

where $c_k = y_k - m$, $\gamma_k \sim Ga(\alpha_k)$ and $\perp\!\!\!\perp (\gamma_k)_{k=1,\ldots,K}$.[4] From that expression, Bloch & Watson (1967) derive analytic formulas for both the pdf and cdf of μ when the α_k's are integers. Even with that restriction, these authors qualify the formulas "... too cumbersome for practical use" (p. 1426).

Since no general nor simple analytic formulas for either the pdf or the cdf of the L-Dirichlet distribution are available, a general and simple solution is to resort to Monte-Carlo sampling from the required Dirichlet distribution, as suggested by Gelman *et al.* (1995, pp. 76–77) and Bernard (1998). That method is easy to implement using either the characterization of the Dirichlet in terms of independent Gamma distributions (*e.g.*, Wilks, 1962, p. 179), or one of its decomposition as a product of independent Beta variables (see *e.g.* Bernard, 1997).

It is difficult to give general statements about the shape of the L-Dirichlet distribution without having an analytic expression for its cdf. We shall only give a few indications that arise from a few computations we have carried out. For large ν, the distribution seems to be always unimodal. We particularly explored the case of symmetric L-Dirichlet distributions, *i.e.* $\varphi_k = 1/K$ for all k's and an equidistant vector of values $\boldsymbol{y}$. It is easy to see that, in that case, the distribution of μ is always symmetric. When varying ν and K with $K > 2$, we first found that $\nu \geq K/2$ seems to always produce a unimodal distribution, more concentrated as ν grows. Note that this covers Jeffreys' prior ($\nu = K/2$) and Bayes-Laplace's prior ($\nu = K$). For $\nu = 1$, which corresponds to Perks' prior, the distribution has K modes (at each y_k). As ν increases, its overall shape is roughtly unimodal when looked from

[4] The symbol $\perp\!\!\!\perp$ indicates that two (or several) random variables are *independent*.

afar, the modes only consisting in small local peaks. On the other side, the limiting case $\nu \to 0$ leads to Haldane's prior for $\boldsymbol{\theta}$ and the discrete uniform on the set $\{y_1, \ldots, y_K\}$ for μ. These remarks should be considered as indicative and not as formally proved facts.

3.4.2 Standardized L-Dirichlet distribution

Because of the finite range of the distribution of μ, it will sometimes be convenient to introduce a standardized version of the L-Dirichlet distribution. Let $\zeta(\cdot)$ be the function from $[y_1, y_K]$ to $[0,1]$, defined by $\zeta(y) = (y - y_1)/(y_K - y_1)$, and let $\boldsymbol{z} = \zeta(\boldsymbol{y}) = (z_1, \ldots, z_K)$, which clearly satisfies $z_1 = 0 < z_2 < \cdots < z_K = 1$.

Then we say the variable $\lambda = \zeta(\mu)$ has a *standardized L-Dirichlet* distribution. Hence we have the following

Definition 2 (Standardized L-Dirichlet)
A L-Diri$(\boldsymbol{z}, \boldsymbol{\alpha})$ *such that* $z_1 = 0 < z_2 < \cdots < z_K = 1$ *is called a standardized L-Dirichlet distribution.*

3.4.3 Approximation by a Beta distribution

Due to the difficulty for calculating the exact pdf and cdf of the L-Dirichlet distribution, Bloch & Watson (1967) suggest to approximate it by a scaled Beta distribution. The idea is simply to approximate $\lambda = \zeta(\mu) = (\mu - y_1)/(y_K - y_1)$, defined on $[0,1]$, by a Beta distribution with same first two moments. Since both distributions have the same support $[0,1]$, that approximation has a much better tail behaviour than the normal one.

3.4.4 Expectation of μ

Since μ is a linear combination of the θ_k's, its expectation is given by

$$\begin{aligned} E(\mu) &= \sum_k y_k E(\theta_k) = \sum_k y_k \varphi_k \\ &= Mean(\boldsymbol{y}, \boldsymbol{\varphi}) = Mean(\boldsymbol{y}, \boldsymbol{\alpha}) \end{aligned} \quad (11)$$

3.4.5 Variance of μ

Using a general property of linear combinations of random variables given in Wilks (1962, Eq. (3.6.3), p. 82), the variance of μ is the sum over all $(j,k) \in [1, \ldots, K]^2$ of $C_{jk} y_j y_k$, where $C_{kk} = V(\theta_k)$ and $C_{jk} = Cov(\theta_j, \theta_k)$ for $j \neq k$. Using the expressions of the variance and covariance of components of $\boldsymbol{\theta}$, given in Section 3.1, and after some simplifications, we obtain

$$\begin{aligned} V(\mu) &= \sum_k \left[\frac{\varphi_k(1-\varphi_k)}{(\nu+1)} y_k^2 + \sum_{j \neq k} \frac{-\varphi_j \varphi_k}{(\nu+1)} y_j y_k \right] \\ &= \frac{\sum_k \varphi_k y_k^2 - (\sum_k \varphi_k y_k)^2}{(\nu+1)} \\ &= \frac{Var(\boldsymbol{y}, \boldsymbol{\varphi})}{(\nu+1)} = \frac{Var(\boldsymbol{y}, \boldsymbol{\alpha})}{(\nu+1)} \end{aligned} \quad (12)$$

3.4.6 The uniform as a special case

Let $\boldsymbol{\theta} \sim Diri(\boldsymbol{\alpha})$ and consider the following reparameterization: $\phi = \theta_1 + \theta_K$, $\theta_K^\star = \theta_K/(\theta_1 + \theta_K)$ and $\theta_k^\circ = \theta_k/(1 - \theta_1 - \theta_K)$, $k = 2, \ldots, K-1$. Denote $\boldsymbol{\theta}^\circ = (\theta_2^\circ, \ldots, \theta_{K-1}^\circ)$, $\nu^\star = \alpha_1 + \alpha_K$ and $\nu^\circ = \alpha_2 + \cdots + \alpha_{K-1}$. Then, it is easily shown that $\phi \sim Beta(\nu^\star, \nu^\circ)$, $\theta_K^\star \sim Beta(\alpha_K, \alpha_1)$, and $\boldsymbol{\theta}^\circ \sim Diri(\alpha_2, \ldots, \alpha_{K-1})$, and that all three distributions are independent (Fang *et al.*, 1990, Thm. 1.4, p. 19; Bernard, 1997). Under that reparameterization, $\lambda = \zeta(\mu)$ becomes

$$\lambda = \phi \theta_K^\star + (1 - \phi)(z_2 \theta_2^\circ + \cdots + z_{K-1} \theta_{K-1}^\circ). \quad (13)$$

Now, let $\nu^\circ \to 0$; then ϕ converges to a punctual distribution concentrated at value 1, and hence the distribution of λ converges to that of $\theta_K^\star$, that is to a $Beta(\alpha_K, \alpha_1)$. Consequently, the distribution of μ converges to a scaled Beta distribution, $y_1 + (y_K - y_1) Beta(\alpha_K, \alpha_1)$. This shows that the L-Dirichlet family is a generalization of the Beta family.

Since the $Beta(1,1)$ distribution is the uniform distribution on $[0,1]$, $Uni(0,1)$, a corollary of the preceding result is that, if $\alpha_1 = \alpha_K = 1$ and $\nu^\circ \to 0$, $\lambda \simeq Uni(0,1)$ and $\mu \simeq Uni(y_1, y_K)$. Using the simplified notation introduced in Section 3.1, we have thus proved the following

Theorem 1 *For any real-valued vector* $\boldsymbol{y}$ *such that* $y_1 < y_2 < \cdots < y_K$,

$$\text{L-Diri}(\boldsymbol{y}, (1, 0, \ldots, 0, 1)) = \text{Uni}(y_1, y_K). \quad (14)$$

3.5 Prior and posterior characteristics of the distribution on μ

Under a Dirichlet $Diri(\boldsymbol{\alpha})$ on $\boldsymbol{\theta}$, we saw that the induced prior and the posterior distributions on μ are respectively $\mu \sim$ *L-Diri*$(\boldsymbol{y}, \boldsymbol{\alpha})$ and $\mu | \boldsymbol{a} \sim$ *L-Diri*$(\boldsymbol{y}, \boldsymbol{\alpha} + \boldsymbol{a})$.

From results of Section 3.3, the prior mean and variance of μ are thus simply

$$E(\mu) = Mean(\boldsymbol{y}, \boldsymbol{\alpha}) = Mean(\boldsymbol{y}, \boldsymbol{\varphi}) \quad (15)$$

$$V(\mu) = \frac{Var(\boldsymbol{y}, \boldsymbol{\alpha})}{(\nu+1)} = \frac{Var(\boldsymbol{y}, \boldsymbol{\varphi})}{(\nu+1)}. \quad (16)$$

Using elementary properties of $Mean(\cdot)$ and $Var(\cdot)$, the posterior mean and variance of μ are given by

$$E(\mu | \boldsymbol{a}) = Mean(\boldsymbol{y}, \boldsymbol{\alpha} + \boldsymbol{a})$$

$$= \frac{\nu Mean(\boldsymbol{y}, \boldsymbol{\varphi}) + n Mean(\boldsymbol{y}, \boldsymbol{f})}{\nu + n} \quad (17)$$

$$V(\mu|\boldsymbol{a}) = \frac{Var(\boldsymbol{y}, \boldsymbol{\alpha} + \boldsymbol{a})}{(\nu + n + 1)}$$

$$= \frac{\nu\, Var(\boldsymbol{y}, \boldsymbol{\varphi}) + n\, Var(\boldsymbol{y}, \boldsymbol{f})}{(\nu + n)(\nu + n + 1)} + \frac{\nu n \left[Mean(\boldsymbol{y}, \boldsymbol{f}) - Mean(\boldsymbol{y}, \boldsymbol{\varphi})\right]^2}{(\nu + n)^2(\nu + n + 1)} \quad (18)$$

3.6 Bayesian objective inference

For multinomial data, four Dirichlet priors have been proposed to model prior ignorance. All four are symmetric Dirichlet, that is $\varphi_k = 1/K$ for all k's. They only differ in their respective total prior strength: $\nu \to 0$ (Haldane, 1948), $\nu = 1$ (Perks, 1947), $\nu = K/2$ (Jeffreys, 1946) and $\nu = K$ (Bayes-Laplace's uniform prior on $\mathcal{S}(1, K)$). It is often claimed that the differences, regarding posterior inferences, between these solutions are small when n is large, but this is only true when n is much larger than K. Regarding the problem of making inferences about μ or, equivalently, $\lambda = \zeta(\mu) = Mean(\boldsymbol{z}, \boldsymbol{\theta})$, the typical case is that of a large K and the inferences produced by the four Bayesian priors can vary in an extreme way.

To see this, consider the parameter λ for the case of equidistant values in $\boldsymbol{z}$ and $K \geq 2$. The four Bayesian priors agree for the prior expectation of λ: all four priors lead to $E(\lambda) = 1/2$. The situation is quite different if we look at the prior variance of λ. For a constant vector $\boldsymbol{\varphi}$, the quantity $Var(\boldsymbol{z}, \boldsymbol{\varphi})$ is equal to $(K+1)/[12(K-1)]$ and hence, from equation (12), the prior variance of λ for any symmetric Dirichlet prior is thus $(K+1)/[12(K-1)(\nu+1)]$. For $\nu \to 0$ (which corresponds to Haldane's prior) and large K, this is close to $1/12$ and the prior has then approximately the same first two moments as the uniform distribution on λ. On the contrary, for a large value of ν, the prior induced on λ from a symmetric Dirichlet prior on $\boldsymbol{\theta}$ can have a much smaller variance than the uniform, thus indicating a strong prior belief that λ is close to its expectation $1/2$. This is particularly true when either Jeffreys' prior ($\nu = K/2$) or Bayes-Laplace's prior ($\nu = K$) are used.

Another argument in favour of Haldane's prior is the remarkable result that the posterior inferences it produces almost agree with the ones from traditional models assuming normality. In effect, from (17) and (18), the posterior expectation and variance of μ, under Haldane's prior, are

$$E(\mu|\boldsymbol{a}) = Mean(\boldsymbol{y}, \boldsymbol{a}), \quad V(\mu|\boldsymbol{a}) = \frac{Var(\boldsymbol{y}, \boldsymbol{a})}{n+1}. \quad (19)$$

Note that under the normal model, $E(\mu|\boldsymbol{a}) = Mean(\boldsymbol{y}, \boldsymbol{a})$, and $V(\mu|\boldsymbol{a}) = Var(\boldsymbol{y}, \boldsymbol{a})/(n-3)$, so that Haldane's and the normal models will almost agree if n is not too small.

The preceding discussion might induce one to think that the only reasonable Bayesian solution is to use Haldane's prior. Unfortunately, it can be shown that the resulting posterior pdf has a non-zero density only in the range $[z'_1, z'_K]$ of the *observed data*, and not in the range $[z_1 = 0, z_K = 1]$ known *a priori* to bound the observations. In other words, Haldane's prior fails to capture the prior knowledge that observable values have a priori bounds. The most extreme situation is the one when a single category k, with value y_k, is observed, *i.e.* $a_k = n$; in this case Haldane's prior leads to the unreasonable conclusion that $\mu = y_k$ with certainty. It should also be noted that, in the general case, Haldane's model leads to a more concentrated posterior distribution than the normal model, and can thus be considered as over-confident that μ is close to the observed mean. These are examples of Haldane's model being too "data-glued".

This is certainly the existence of such difficulties with the general objective-minded priors that has led some authors to propose *reference Bayesian priors* (Bernardo, Smith, 1994, Section 5.4), whose declared purpose is to provide a non-informative prior *relative to some specific parameter of interest*. Following that view, the problem would then be to specify a non-informative prior directly on μ and not on $\boldsymbol{\theta}$ as we did previously. Since the parameter μ is bounded, the most obvious choice is the uniform distribution, $Uni(y_1, y_K)$, which is a limiting case of an L-Dirichlet distribution, as showed in Theorem 1.[5] The implied Dirichlet prior on $\boldsymbol{\theta}$ is, however, far from being symmetric (unlike the previous priors), since the prior relative strengths are $\boldsymbol{\varphi} = (1/2, 0, \ldots, 0, 1/2)$. One difficulty with the reference prior approach is that inferences for various parameters of interest are not necessarily coherent. In addition, since that approach constitutes a generalization of Jeffreys' prior, again it does not satisfy the likelihood principle.

One fact emerging from the preceding discussion is that, in the Bayesian framework, the idea of prior ignorance is operationalized in quite different ways whether one considers inference about $\boldsymbol{\theta}$ or inference about μ. Considerations about how to formalize prior ignorance for $\boldsymbol{\theta}$ seem to be inconsistent with the ones for μ. In the next section, we shall see that an imprecise model, such as the imprecise Dirichlet model, does not suffer from that criticism.

[5] We hypothetize that the uniform prior is the reference prior on μ, in the sense of Bernardo, though we haven't checked that.

4 Imprecise Dirichlet model

4.1 Presentation of the model

Walley (1996) proposed the *imprecise Dirichlet model (IDM)* as a model for prior ignorance in the case of categorical data. The model consists in describing prior uncertainty by a set of Dirichlet priors instead of a single one. Bernard (1998, 2000) explored several applications of the IDM, and often refers to the set of priors as defining an *ignorance zone*. The special case of two categories was also studied in Walley (1991, Section 5.3) and in Bernard (1996).

The prior IDM(ν) is defined as the set of all Dirichlet priors on $\boldsymbol{\theta}$ with a fixed prior strength $\nu > 0$, that is $\{Diri(\boldsymbol{\alpha}) : \alpha_k > 0 \text{ for all } k, \sum_k \alpha_k = \nu\}$ or equivalently

$$\{Diri(\nu\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \mathcal{S}^{\star}(1, K)\}, \tag{20}$$

where $\mathcal{S}^{\star}(1, K)$ is the interior of the unit simplex $\mathcal{S}(1, K)$.

Let $P_{\nu\boldsymbol{\varphi}}(\cdot)$ and $E_{\nu\boldsymbol{\varphi}}(\cdot)$ be respectively a prior probability and a prior expectation provided by a particular $Diri(\nu\boldsymbol{\varphi})$ in the set (20). The uncertainty about any event A concerning $\boldsymbol{\theta}$ is described by *prior lower and upper probabilities*, denoted by $\underline{P}(A)$ and $\overline{P}(A)$, and calculated by minimizing and maximizing $P_{\nu\boldsymbol{\varphi}}(A)$ with respect to $\boldsymbol{\varphi} \in \mathcal{S}^{\star}(1, K)$. Similarly, for any real-valued function $G(\boldsymbol{\theta})$, *prior lower and upper expectations* $\underline{E}(G)$ and $\overline{E}(G)$ are calculated by minimizing or maximizing the expectation $E_{\nu\boldsymbol{\varphi}}(G)$ with respect to $\boldsymbol{\varphi}$. The upper and lower probabilities and expectations are related by $\overline{P}(A) = 1 - \underline{P}(A^c)$ and $\overline{E}(G) = -\underline{E}(-G)$. For a thorough presentation of the theory of imprecise probabilities and, in particular, of its consequences in manipulating such expressions, see Walley (1991).

Each Dirichlet prior in the prior IDM is updated into a Dirichlet posterior using Bayes' theorem. This updating procedure guarantees coherence of the inferences (Walley, 1991). Hence the posterior uncertainty about $\boldsymbol{\theta}$ from the IDM(ν) is expressed by the set

$$\{Diri(\boldsymbol{a} + \nu\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \mathcal{S}^{\star}(1, K)\} \tag{21}$$

As for the prior IDM, posterior lower and upper probabilities or expectations are obtained respectively through minimization or maximization with respect to $\boldsymbol{\varphi} \in \mathcal{S}^{\star}(1, K)$.[6]

[6] In the general theory, minimization or maximization must actually be performed over the set of all convex combinations of priors in the set (20) or of posteriors in the set (21). In most problems, however, the sets (20) or (21) are sufficient to provide the upper and lower component. A problem where the convex set is necessary will appear in Sections 4.2 and 4.3.3 for finding upper variances (see also Walley, 1991, Appendix G).

4.2 The prior IDM for μ

The prior lower and upper expectations of μ are obtained by respectively minimizing and maximizing $E(\mu) = Mean(\boldsymbol{y}, \boldsymbol{\varphi})$ with respect to $\boldsymbol{\varphi} \in \mathcal{S}^{\star}(1, K)$; they are simply

$$\underline{E}(\mu) = y_1 \quad \text{and} \quad \overline{E}(\mu) = y_K, \tag{22}$$

and are respectively obtained for $\varphi_1 \to 1$ and $\varphi_K \to 1$ (*e.g.*, in the limit, $\boldsymbol{\varphi} = (1, 0, \ldots, 0)$ and $\boldsymbol{\varphi} = (0, \ldots, 0, 1)$ respectively). In terms of the standardized parameter $\lambda = \zeta(\mu)$ defined on $[0, 1]$, we have $\underline{E}(\lambda) = 0$ and $\overline{E}(\lambda) = 1$. The prior expectations of μ (or equivalently of λ) are *vacuous* (*i.e.* maximally imprecise).

Since λ is defined on $[0, 1]$, we also see that these lower and upper expectations correspond respectively to $\lambda = 0$ with probability 1 and to $\lambda = 1$ with probability 1. Hence the prior lower and upper prior cdf's of μ are also vacuous:

$$\underline{F}(m) = \underline{P}(\mu \leq m) = 0, \quad \text{for } m \in [y_1, y_K) \tag{23}$$

$$\overline{F}(m) = \overline{P}(\mu \leq m) = 1, \quad \text{for } m \in [y_1, y_K]. \tag{24}$$

The prior lower and upper variances of μ are given by

$$\underline{V}(\mu) = 0 \quad \text{and} \quad \overline{V}(\mu) = (y_K - y_1)^2/4, \tag{25}$$

or, equivalently, in terms of the standardized parameter $\lambda = \zeta(\mu)$,

$$\underline{V}(\lambda) = 0 \quad \text{and} \quad \overline{V}(\lambda) = 1/4. \tag{26}$$

The prior lower variance is obtained by minimizing $V(\mu) = Var(\boldsymbol{y}, \boldsymbol{\varphi})/(\nu + 1)$ with respect to $\boldsymbol{\varphi} \in \mathcal{S}^{\star}(1, K)$. The minimum is attained by a single Dirichlet prior in the set (20) of the type $\varphi_k \to 1$ for some k.

As explained in Walley (1991, Appendix G), the prior upper variance is more difficult to compute since it requires the maximization to be performed over all the convex combinations of couples Dirichlet priors in the set (20). Using elementary properties of $Var(\cdot)$ and of convex combinations, the solution to that maximizing problem is reached by the mixture, with equal weights, of the two priors corresponding respectively to $\varphi_1 \to 1$ and $\varphi_K \to 1$ (*i.e.* in the limit, $\boldsymbol{\varphi} = (1, 0, \ldots, 0)$ and $\boldsymbol{\varphi} = (0, \ldots, 0, 1)$ respectively).

All these properties of the prior on μ induced by the IDM indicate that, whatever the value of ν, the prior uncertainty on μ is maximally imprecise.

4.3 The posterior IDM for μ

4.3.1 Posterior lower and upper expectation

Equation (17) expresses the posterior mean of μ, $E(\mu|\boldsymbol{a})$, for a single Dirichlet prior, as a weigthed average of the prior mean $Mean(\boldsymbol{y}, \boldsymbol{\varphi})$ and the observed mean $Mean(\boldsymbol{y}, \boldsymbol{f})$. Using (22), we easily find the lower and upper posterior expectations of μ as

$$\underline{E}(\mu|\boldsymbol{a}) = \frac{\nu y_1 + n Mean(\boldsymbol{y}, \boldsymbol{f})}{\nu + n} \tag{27}$$

$$\overline{E}(\mu|\boldsymbol{a}) = \frac{\nu y_K + n Mean(\boldsymbol{y}, \boldsymbol{f})}{\nu + n}, \tag{28}$$

respectively obtained for $\boldsymbol{\varphi} = (1, 0, \ldots, 0)$ and $\boldsymbol{\varphi} = (0, \ldots, 0, 1)$. Since $y_1 \leq Mean(\boldsymbol{y}, \boldsymbol{f}) \leq y_K$, the interval $[\underline{E}(\mu|\boldsymbol{a}, \overline{E}(\mu|\boldsymbol{a}]$ always covers $Mean(\boldsymbol{y}, \boldsymbol{f})$. As n increases, both expectations tend to the observed mean, $Mean(\boldsymbol{y}, \boldsymbol{f})$.

4.3.2 Posterior lower and upper cdf's

Since μ is a linear combination of the population frequencies, the θ_k's, it can be easily shown, using (10), that the posterior lower cdf $\underline{F}(m|\boldsymbol{a}) = \underline{P}(\mu \leq m|\boldsymbol{a})$ is obtained from the Dirichlet posterior distribution with $\varphi_k \to 1$ where k is the category whose frequency has the *largest* coefficient in μ. The largest coefficient in μ is y_K, so the lower cdf is obtained for $\boldsymbol{\varphi} = (0, \ldots, 0, 1)$. Similarly, the posterior upper cdf $\overline{F}(m|\boldsymbol{a}) = \overline{P}(\mu \leq m|\boldsymbol{a})$ is obtained from the Dirichlet posterior with $\varphi_k \to 1$ where k is the category whose frequency has the *smallest* coefficient in μ, that is $k = 1$, and thus $\boldsymbol{\varphi} = (1, 0, \ldots, 0)$. The same argument was used in Walley & Bernard (1999, Section 7) for a similar problem. The lower and upper posterior cdfs for μ are respectively those of

$$\textit{L-Diri}(\boldsymbol{y}, \boldsymbol{a} + \boldsymbol{\alpha} = (a_1, \ldots, a_{K-1}, a_K + \nu))$$
$$\textit{L-Diri}(\boldsymbol{y}, \boldsymbol{a} + \boldsymbol{\alpha} = (a_1 + \nu, a_2, \ldots, a_K))$$

Notice that the prior $\textit{L-Diri}(\boldsymbol{y}, (1, 0, \ldots, 0))$ is the one that simultaneously provides the prior and posterior *lower expectation and upper cdf*. Symmetrically, the prior $\textit{L-Diri}(\boldsymbol{y}, (0, \ldots, 0, 1))$ provides the prior and posterior *upper expectation and lower cdf*. The former allocates the entire prior strength to the value y_1, while the latter allocates it entirely to y_K.

4.3.3 Posterior lower and upper variances

In order to simplify the notations in this section, let us define $D_y = y_K - y_1$, $M_0 = Mean(\boldsymbol{z}, \boldsymbol{\varphi})$, $V_0 = Var(\boldsymbol{z}, \boldsymbol{\varphi})$, $M = Mean(\boldsymbol{z}, \boldsymbol{f})$, and $V = Var(\boldsymbol{z}, \boldsymbol{f})$, where $\boldsymbol{z} = \zeta(\boldsymbol{y})$.

From (18), finding the lower posterior variance of μ requires the minimization, with respect to $\boldsymbol{\varphi} \in \mathcal{S}^*(1, K)$, of the quantity

$$Q = [(n + \nu)V_0 + n(M - M_0)^2]$$

The analysis of the minimizing problem shows that the solution consists in allocating the entire prior strength to the category m whose value z_m is the nearest to M. Formally, let $m \in \{1, \ldots, K\}$ be the category for which $|M - z_m|$ is minimal (if two adjacent categories satisfy this, any of them will do). The solution is to chose $\varphi_m \to 1$ and hence $\varphi_k \to 0$ for $k \neq m$. The lower variance of μ is then given by

$$\underline{V}(\mu) = \frac{nD_y^2}{(n + \nu)(n + \nu + 1)} \left[V + \frac{\nu(M - z_m)^2}{n + \nu} \right]. \tag{29}$$

Similarly to what we saw for the prior upper variance, the posterior upper variance of μ is more difficult to compute, since it requires the maximization to be performed for all convex combinations of couples of posterior distributions from the set (21). We give the result below without proof. First let

$$r = \frac{1}{2} + \frac{n(M - 1/2)}{\nu(n + \nu + 1)}.$$

The upper variance is given by

$$\overline{V}(\mu) = \frac{nD_y^2}{(n + \nu)(n + \nu + 1)}(V + Q_r), \tag{30}$$

where

$$Q_r = \begin{cases} (\nu + 1)\left[\frac{(M-1/2)^2}{n+\nu+1} + \frac{\nu}{4n}\right] & \text{if } 0 < r < 1 \\ \nu M^2/(n + \nu) & \text{if } r \leq 0 \\ \nu(1 - M)^2/(n + \nu) & \text{if } r \geq 1 \end{cases}.$$

In the first case, $0 < r < 1$, the upper variance is attained by the mixture of two Dirichlet priors $Diri(\nu\boldsymbol{\varphi})$, with $\boldsymbol{\varphi} = (1, 0, \ldots, 0)$ and $\boldsymbol{\varphi} = (0, \ldots, 0, 1)$ respectively, and with respective weights $(r, 1 - r)$. Note that this is the only case to consider when $\nu \geq 1$. In the second case, $r \leq 0$, the maximum is attained by the single Dirichlet prior corresponding to $\varphi_K \to 1$, *i.e.* to $\boldsymbol{\varphi} = (0, \ldots, 0, 1)$. Symetrically, in the last case, $r \geq 1$, the maximum is attained by the single Dirichlet prior corresponding to $\varphi_1 \to 1$, *i.e.* to $\boldsymbol{\varphi} = (1, 0, \ldots, 0)$.

It can be seen that, as n increases both lower and upper variances will converge to $D_y^2 V/n = Var(\boldsymbol{y}, \boldsymbol{a})/n$. This quantity is also the asymptotic approximation of $V(\mu|\boldsymbol{a})$ in the normal model.

4.4 Choice of ν

The IDM as defined in (20) and (21) depends on the choice of ν. The constant ν determines how fast the lower and upper probabilities converge one towards the other as n increases, and can thus be interpreted as a measure of the *caution* of the inferences. The larger ν is, the more cautious the inferences are. For instance, let us consider the expectation of $\lambda = \zeta(\mu)$. Its prior imprecision, $\Delta(\lambda) = \overline{E}(\lambda) - \underline{E}(\lambda) = 1$, is maximal, as we already noted. Its posterior imprecision, $\Delta(\lambda|\boldsymbol{a}) = \overline{E}(\lambda|\boldsymbol{a}) - \underline{E}(\lambda|\boldsymbol{a})$, is equal to $\nu/(n+\nu)$, and thus to $1/2$ if $\nu = n$. The parameter ν can thus be considered as the number of observations that are necessary to divide the imprecision about the expectation of λ (and hence of μ) by 2.

If we consider two IDMs with ν_1 and ν_2 respectively, such that $\nu_2 > \nu_1$, then the probability interval associated with any event under the IDM(ν_2) will cover the corresponding interval under the IDM(ν_1). Various IDMs are thus nested according to their respective ν parameters.

These preliminary comments about the interpretation of ν being made, on which criteria should ν be chosen? Since, despite Walley's (1996) wish, we are still lacking a new principle of inference that would entail a specific value for ν, the most important criterion that is available is the requirement that the IDM should be cautious enough to encompass frequentist or objective Bayesian alternatives. On the other side, being too cautious means that the inferences are too weak. Consequently, the ideal value for ν appears to be the smallest one among those encompassing alternative statistical models.

Bernard (1996), Walley (1996, 2000) and Walley & Bernard (1999) give several convincing arguments for chosing $n \geq 1$. The other arguments they give for chosing $\nu \leq 2$ are less so, because most of them involve binary data, *i.e.* the case $K = 2$. But, as we saw in Section 3, some usual Bayesian alternatives use Dirichlet priors where ν depends on K: $\nu = K/2$ for Jeffreys' prior, or $\nu = K$ for Bayes-Laplace's prior. Since ν must not depend on K for the IDM to satisfy the representation invariance principle, the study of situations with possibly large K, like the present one, is thus of particular importance if we want to find an upper bound for ν.

It first appears from Section 3.6 that, in our present problem, neither Jeffrey's prior nor Bayes-Laplace's can be considered, by any Bayesian, as appropriate for representing a state of prior ignorance about μ: both priors strongly favor values of μ around $(y_1 + y_K)/2$, and so much the larger K is. Thus, the only objective Bayesian competitors to our model seem to be Haldane's prior ($\varphi = 1/K$ and $\nu \to 0$), Perks' prior ($\varphi = 1/K$ and $\nu = 1$) and the uniform prior for μ. In addition, Theorem 1 tells us that the prior IDM on $\boldsymbol{\theta}$ with $\nu = 2$ always contains the uniform prior for μ. Hence prior or posterior inferences from the IDM($\nu = 2$) encompass inferences from any of the latter three objective Bayesian models. Globally, it seems that we now have strong arguments for $1 \leq \nu \leq 2$, with an example with possibly large K. We hope that future research will lead to narrow that interval. In the next example, we use the IDM with both $\nu = 1$ and $\nu = 2$.

5 An example: Darwin's maize data

Our purpose here is merely to illustrate the inferences from the IDM on a well-known example. The example is discussed in Fisher (1935, Table 3) and, in a context of non-parametric inference, in Geyer (1996) and Walley (1996). The data are $n = 15$ discrete measurements on maize made with a finite precision $d_y = 1$. The observed data are given in Table 1.

Table 1: *Darwin data. Observed values x_i, $n = 15$.*

−67	−48	6	8	14	16	23	24	28	29	41	49	56	60	75

The basic descriptive statistics for this data set are $Mean(\boldsymbol{y}, \boldsymbol{a}) = 20.93$ and $Var(\boldsymbol{y}, \boldsymbol{a}) = 36.46^2$.

Let us first consider the usual inferences from the normal model. Under the normal sampling model, if we want to test the extended hypothesis $H_0 : \mu \leq 0$, we find $t_{obs} = 2.15$ for the t-test statistic with $n - 1 = 14$ degrees of freedom, which leads to an observed one-sided significance level $p_{obs} = 0.025$. We can also compute a confidence interval $[m_1, m_2]$ for μ at, say, the 0.90 level, and we find: $\mu \in [3.8, 38.1]$. If, in addition, we assume the non-informative prior on (μ, σ) (see Introduction), the two previous results can be reinterpreted in Bayesian terms: $P(\mu < 0|\boldsymbol{a}) = p_{obs} = 0.025$ and $P(\mu \in [3.8, 38.1]|\boldsymbol{a}) = 0.90$. The posterior expectation and variance of μ are then $E(\mu|\boldsymbol{a}) = M = 20.93$ and $V(\mu|\boldsymbol{a}) = 10.5^2$. These results are summarized in the first row of Table 2.

We now consider several models that assume that the observations are discrete and bounded. For illustrative purposes, we shall study two possible intervals $[y_1, y_K]$ of bounds, namely $[-67, 75]$ (the bounds of the observed values) and $[-100, 100]$. The next three rows of Table 2 give the results to the same questions for three of the objective Bayesian models considered in Section 3, each one corresponding to a single L-Dirichlet prior: Haldane's prior, Perks' prior and the uniform distribution on $[y_1, y_K]$. Note that Haldane's model does not depend on the a priori bounds $[y_1, y_K]$.

Finally, the bottom part of Table 2 gives the corresponding results for four IDM models obtained by crossing the two intervals of bounds and $\nu = 1, 2$.[7]

Table 2: *Darwin data. Comparison of posterior inferences from several prior models: normal, Haldane, Perks, uniform on $[y_1, y_K]$ for μ, and four IDM models. The expectations, standard deviations and probabilities are conditional on the observed data, $\boldsymbol{a}$; m_1 and m_2 are credibility limits at levels 0.95 and 0.05 respectively. For each IDM, the two rows give the lower and upper component respectively.*

Model	y_1	y_K	ν	$E(\mu)$	$\sigma(\mu)$	$P(\mu < 0)$	m_1	m_2
Normal	—	—	—	20.9	10.5	0.025	3.8	38.1
Haldane	—	—	0	20.9	9.4	0.021	4.8	34.6
Perks	-67	75	1	19.9	9.0	0.023	4.1	33.5
	-100	100	1	19.6	9.3	0.028	3.1	33.8
Uniform	-67	75	2	18.9	10.0	0.039	1.5	34.2
	-100	100	2	18.5	11.5	0.064	-1.8	36.1
IDM	-67	75	1	15.4	8.6	0.010	-2.3	30.6
				24.3	10.6	0.072	8.3	38.1
	-67	75	2	10.6	8.1	0.005	-7.7	26.8
				27.3	12.9	0.159	11.4	41.1
	-100	100	1	13.4	8.6	0.009	-6.5	29.9
				25.9	12.2	0.119	9.1	41.2
	-100	100	2	6.7	8.1	0.004	-14.9	25.1
				30.2	16.3	0.273	13.0	46.2

6 Conclusions

6.1 Remarks about our model

For reasons of simplicity, we have considered that the data resulted from multinomial sampling. In fact, since the IDM satisfies the likelihood principle, the only relevant component of the sampling model is the likelihood function. Any other sampling model that leads to the same likehood function as the multinomial is thus covered by the present work, for instance negative multinomial sampling. More generally, the sampling model could be replaced by an the assumption that the observed data are exchangeable with any set of future observations of an arbitrarily large size.

Our model makes no assumption whatsoever about the shape of the parent distribution. It could be modified so as to incorporating some prior knowledge that the distribution is smooth. Some of our results permit to foresee that this assumption may have little practical consequences for inferences about μ. Anyway, our personal inclination is to avoid making assumptions that could be falsified once the data are observed. But clearly, some will disagree with that view.

One central result of the present paper is that the IDM(ν) encompasses three objective Bayesian models which appear as the most reasonable non-parametric models for our discrete problem, *whenever the parameter ν is ≥ 2*. We think that this is an argument in favor of chosing $\nu \leq 2$, or even $\nu = 2$. Previous work about the IDM has already given strong arguments for chosing $\nu \geq 1$. So it now appears that the values $1 \leq \nu \leq 2$ should be privileged in using the IDM.

6.2 Directions for future research

We would like to conclude by briefly mentioning some directions in which the present work could be completed and extended. First, we need to study the predictive inferences from our model; a reasonable amount of material is already available in Ericson (1969) and Hoadley (1969), but for the Bayesian approach only. Secondly, several other characteristics of the parent distribution — *e.g.*, its variance or its quantiles —, could be subject to inferences from the IDM; Walley (1996, Section 6) already gives the IDM answer for the quantiles, which is extremely simple, but a comparison with other models would perhaps bring more insight in the IDM. Thirdly, the inferences from the IDM should also be compared with the ones from the Dirichlet process prior, which is also a non-parametric model but assumes that the observations are continuous and unbounded. A more systematic comparison with frequentist or Bayesian inferences under the normal model should also be done.

Finally, the problem of the inference on an unknown mean that we studied here naturally leads to consider several extensions for more sophisticated ANOVA problems, such as inferences about contrasts on means from several groups of observations, either in crossed designs or in nested ones; again the work of Ericson (1969) and Hoadley (1969) already contains some material about that matter, but only using the Bayesian framework, not the IDM.

Acknowledgements

I am extremely grateful to Peter Walley for his valuable remarks and suggestions and for pointing out mistakes in a previous draft.

[7] For any model based on a L-Dirichlet distribution, including the IDM, the computations of $P(\mu < 0|a)$ and of the credibitlity limits were done by means of Monte-Carlo simulations (see Section 3.4.1). Samples of size 200,000 from each L-Dirichlet distribution were used, so that any (upper or lower) probability has a standard error which is never larger than 0.0011.

References

Bernard, J.-M. (1996), "Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process", *The American Statistician*, 50, 7–13.

Bernard, J.-M. (1997), "Bayesian Analysis of Tree-Structured Categorized Data", *Revue Internationale de Systémique*, 11, 11–29.

Bernard, J.-M. (1998), "Bayesian Inference for Categorized Data", In H. Rouanet *et al.*, *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference*, European University Studies, Series 6: Psychology, Bern: Peter Lang, pp. 159–226.

Bernard, J.-M. (2000), "Implicative Analysis for Multivariate Binary Data using an Imprecise Dirichlet Model", to appear in *J. Statist. Planning and Inference* .

Bernardo, J. M., Smith, A. F. M. (1994), *Bayesian Theory*, New-York: John Wiley & sons.

Bloch, D. A., Watson, G. S. (1967), "A Bayesian Study of the Multinomial Distribution", *The Annals of Mathematical Statistics*, 38, 1423–1435.

Box, G. E. P., Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, Mass.: Addison–Wesley.

Ericson, W. A. (1969), "Subjective Bayesian Models in Sampling Finite Populations", *J. Roy. Statist. Soc. B* , 31, 195–233.

Fang, K. T., Kotz, S., Ng, K. W. (1990), *Symetric Multivariate and Related Distributions*, New-York: Chapman and Hall.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Geisser, S. (1984), "On prior distributions for Binary Trials" (with discussion), *The Amercian Statistician*, 38, 244–251.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Geyer, C. J. (1996), in Discussion of Walley, P., "Inferences from Multinomial Data: Learning about a Bag of Marbles", *Journal of the Royal Statistical Society*, Ser. B., 58, 3–57, pp. 40–41.

Haldane, J. B. S. (1948), "The Precision of Observed Values of Small Frequencies," *Biometrika*, 35, 297–300.

Hoadley, B. (1969), "The compound multinomial distribution and Bayesian analysis of categorical data from finite population", *Journal of the American Statistical Association*, Vol. 64, pp. 2–16–229.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems", *Proceedings of the Royal Society of London*, Ser. A, 186, pp. 4–53–461.

Jeffreys, H. (1961), *Theory of Probability*, 3rd ed., Oxford: Clarendon Press.

Perks, F. J. A. (1947), "Some Observations on Inverse Probability Including a New Indifference Rule," *Journal of the Institute of Actuaries*, 73, 285–334.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, Monographs on Statistics and Applied Probability 42, London: Chapman & Hall.

Walley, P. (1996), "Inferences from Multinomial Data: Learning about a Bag of Marbles", *Journal of the Royal Statistical Society*, Ser. B., 58, 3–57.

Walley, P. (2000), "Reconciling Frequentist Properties with the Likelihood Principle", to appear in *J. Statist. Planning and Inference* .

Walley, P., Bernard, J.-M. (1999), *Imprecise Probabilistic Prediction for Categorical Data*, Technical report CAF-9901, Laboratoire Cognition et Activités Finalisées, Université Paris 8, Saint-Denis (France).

Wilks, S. S. (1962), *Mathematical Statistics*, New York: John Wiley.