Technical Report No. CAF-9901

Imprecise Probabilistic Prediction for Categorical Data

by

Peter WALLEY

and

Jean-Marc BERNARD

berj@univ-paris8.fr

Université de Paris 8

Laboratoire Cognition et Activités Finalisées
CNRS ESA 7021
2 rue de la Liberté
F-93526 Saint-Denis Cedex, France

January 1999

Imprecise Probabilistic Prediction for Categorical Data

Peter WALLEY

and

Jean-Marc BERNARD
berj@univ-paris8.fr

January 1999

Abstract

We consider the problem of predictive inference: after observing a sample of categorical data, how can we make probabilistic predictions about a future sample? This includes the problems of inference about a finite population from a sample taken without replacement, and predictive inference about future observations from a multinomial process. We describe an imprecise probability model, based on a set of Dirichlet-multinomial distributions, which gives predictive inferences in these problems. It can be used when there is no prior information about frequencies and even when there is no information about what types of observation are possible. Inferences from our model can be qualitatively different from objective Bayesian inferences even when the observed sample size is large. An application to the analysis of economic survey data is described.

Keywords: Dirichlet-multinomial distribution; Finite population; Foundations of statistics; Imprecise probability; Inductive logic; Multinomial process; Multiple hypergeometric data; Objective Bayesian inference; Predictive inference; Prior ignorance; Upper and lower probability

1 Introduction

Suppose that we observe a sample of n units, each of which is classified into one of finitely many categories or types, and we wish to make inferences about the types of units in a future sample of size n'. Suppose also that the past and future observations are linked through an assumption of exchangeability, that all orderings of the combined (past and future) observations are equally probable. Initially we know nothing about the relative frequencies of different types, and we may not even know what types of observation are possible, but we can learn something from the observed sample that enables us to make probabilistic predictions about the types of future observations. This paper describes a general model for making predictive inferences of this kind.

Another way of stating the problem is to consider that we wish to make inferences about the composition of a finite population of size $n+n'$, based on an observed sample of n units from the population. Thus the general problem includes two important kinds of inference: inference about a finite population from a sample taken without replacement (a multiple hypergeometric process); and predictive inference about a future sample from a multinomial process. Under the model we describe in this paper, these two kinds of inference are essentially the same. To illustrate our model, we shall apply it to some special kinds of predictive inference and use it to analyse data from a survey of French company managers.

The general problem of probabilistic prediction for categorical data has been of great historical importance in the development of statistical methodology. Karl Pearson (1920) called it "the fundamental problem of practical statistics". The famous paper of Bayes (1763) discussed the problem of probabilistic prediction in the case where observations are classified into two categories. Laplace (1778) proposed a generalisation to the case of multiple categories, and around 1830 Lubbock and Drinkwater-Bethune calculated the predictive probabilities that are implied by Laplace's model. Laplace's method and the "rule of succession" it generates (Laplace 1812/1825, pp. 44–46) were widely discussed and criticised during the nineteenth century; see Bru (1986), Dale (1991) and Stigler (1986) for summaries of this early work. For later discussions of the problem and other references, see Keynes (1921, Ch. 30), Johnson (1932) and its discussion in Zabell (1982), Carnap (1952), Fisher (1956, Ch. 5), Thatcher (1964), Good (1965), Dempster (1966), Geisser (1984), Walley (1996a), and Bernard (1998a).

In agreement with Geisser (1993), we think that predictive inference is a particularly natural approach to the general problem of inference since it can be defined wholly in terms of observable quantities, without reference to unobservable parameters. Predictive inference can also be considered to be more general than parametric inference, because inferences about underlying parameters can always be obtained as a limit of predictive inferences.

The finite, predictive view of inference that we adopt here is also more realistic in many practical applications. Most actual populations are finite and continuous models are generally used only as convenient approximations to the correct finite models. This point of view has been advocated especially by de Finetti (1974/1975), Basu (1975), Geisser (1984, 1993) and Lad (1996). The predictive approach is widely used in applications such as survey sampling, quality control, calibration, classification and

discrimination, and sequential clinical trials. See Aitchison & Dunsmore (1975) and Geisser (1993) for a general presentation of the predictive approach to inference and some of its applications.

1.1 The exchangeability assumption

In mathematical notation, the problem can be stated as follows. Each unit that is observed is classified into exactly one of k categories or types which are labelled as $1, 2, \ldots, k$. Of course the set of possible types and the value of k may not be known a priori; this problem is discussed below. We observe the types of n distinct units. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ denote the frequencies of each type that are observed in the sample of n, with $\sum_{i=1}^{k} x_i = n$. The problem is to make inferences about $\boldsymbol{x}' = (x'_1, x'_2, \ldots, x'_k)$, the frequencies of each type in a future sample of n' new units, where $\sum_{i=1}^{k} x'_i = n'$.

As a simple example of sampling without replacement from a finite population, suppose that an urn contains three balls, each coloured black or white. We draw two balls at random from the urn without replacement and observe their colours. Here $k = 2, n = 2$, and x_1 and x_2 denote the number of black balls and white balls respectively in the two drawn. We want to predict the colour of the remaining ball, so $n' = x'_1 + x'_2 = 1$.

The past and future frequencies, $\boldsymbol{x}$ and $\boldsymbol{x}'$, are related through the assumption that, conditional on the frequencies of each type in the $n + n'$ observations, each possible ordering of the $n + n'$ observations is equally probable a priori. In the example of the previous paragraph, this means that, conditional on there being, for instance, two black balls and one white ball in the urn initially, the three possible orders in which they can be drawn (BBW, BWB and WBB) are equally probable a priori. This is an assumption of exchangeability or order-invariance. The *hypergeometric* sampling process in the preceding example obviously satisfies the exchangeability assumption, and so do many other sampling models in which the observed sample size n may be random. For example, exchangeability also holds for *inverse multiple hypergeometric sampling*, where we continue sampling until we observe a fixed number x_i of type i units, and for *multinomial sampling* (fixed n) or *inverse multinomial sampling* (fixed x_i) from an infinite population.

In the case of *multiple hypergeometric sampling*, let $x^*_i = x_i + x'_i$ denote the combined frequency of type i in the n^* past and future observations. The probability of obtaining the observed frequencies $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ conditional on the combined frequencies $\boldsymbol{x}^* = (x^*_1, x^*_2, \ldots, x^*_k)$ is

$$P(\boldsymbol{x}|\boldsymbol{x}^*) \;=\; \prod_{i=1}^{k} \binom{x^*_i}{x_i} \Big/ \binom{n^*}{n}, \tag{1}$$

where x_i and x^*_i are nonnegative integers, $x_i \le x^*_i$, $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} x^*_i = n^*$. As a function of $\boldsymbol{x}^*$, the probability in (1) represents the likelihood of $\boldsymbol{x}^*$ given $\boldsymbol{x}$,

$$L(\boldsymbol{x}^*|\boldsymbol{x}) \;\propto\; \prod_{i=1}^{k} \binom{x^*_i}{x_i} = \prod_{i=1}^{k} \binom{x_i + x'_i}{x_i}. \tag{2}$$

More generally, under the exchangeability assumption, it is (2) which relates the observed frequencies $\boldsymbol{x}$ to the future frequencies $\boldsymbol{x}' = \boldsymbol{x}^* - \boldsymbol{x}$. All four sampling methods mentioned above, multiple hypergeometric or multinomial, direct (fixed n) or inverse (fixed x_i), lead to the same likelihood function (2).

1.2 Frequentist solutions to the problem

The problem of predictive inference has received much less attention than that of parametric inference, especially within the frequentist framework. Aitchison & Dunsmore (1975, pp. 79–87 and Sections 5 and 6) and Geisser (1993, Section 2) review various non-Bayesian solutions to the predictive problem. Bjørnstad (1990) reviews several methods based on the concept of *predictive likelihood* introduced by Fisher (1956, pp. 128–133). Hampel (1993, 1996) has recently proposed a new frequentist approach to predictive inference which produces imprecise probabilistic predictions in the form of upper and lower probabilities, and he has applied his approach to binary categorical data.

Within the standard frequentist framework, the most common approach is to look for *tolerance regions* which have good frequentist properties (e.g. specified coverage probability). The work of Thatcher (1964) is especially relevant to our present purpose as it deals with prediction from binary data and derives some important relationships between frequentist and objective Bayesian predictions. Guttman (1970) describes the frequentist theory of tolerance regions and compares them with Bayesian prediction sets, but focuses almost exclusively on the normal model.

A major difficulty encountered in the frequentist approach to prediction, which is pointed out by Bjørnstad (1990, p. 242), is that the usual parameter typically becomes a nuisance parameter for the predictive problem. For that reason there are many different, ad hoc frequentist solutions, depending on how the parameter is eliminated: by integration, maximization or conditioning. Also, because frequentist methods do not satisfy the likelihood principle, the solutions depend on the stopping rule for the experiment, and particularly on whether the data are obtained by direct or inverse sampling. Finally, the usual frequentist methods do not produce probability statements about future observations, and hence they do not allow us to express predictive inferences in a natural way. Because of these difficulties with the frequentist approach, in this paper we consider Bayesian predictive inference to be the only serious competitor to our method. Nevertheless, we are interested in finding methods of prediction that have good properties from a frequentist point of view, and these properties will be studied in later sections.

1.3 Objective Bayesian solutions to the problem

Bayesian inferences can be made in this problem by constructing a prior distribution $P(\boldsymbol{x}^*)$ for the unknown vector of combined frequencies and applying Bayes' theorem in the form $P(\boldsymbol{x}^*|\boldsymbol{x}) \propto L(\boldsymbol{x}^*|\boldsymbol{x})P(\boldsymbol{x}^*)$, where $L(\boldsymbol{x}^*|\boldsymbol{x})$ is given by (2) and the proportionality constant can be obtained by normalization. The difficulty is to construct an appropriate prior distribution, especially when there is little or no prior information about the possible types of observations.

In cases of prior ignorance, the most common Bayesian model for predictive inference appears to be the one first suggested by Laplace (1778, Article 33, pp. 482–485) and later studied in more detail by Lubbock & Drinkwater-Bethune (see references in Dale, 1991, pp. 280–286). This model, which we shall call the LLDB model, is also advocated by Geisser (1984). In the LLDB model, the prior distribution is defined by assuming that all the possible combined frequency vectors $\boldsymbol{x}^*$ are equally probable, each with prior probability $P(\boldsymbol{x}^*) = \binom{n^*+k-1}{n^*}^{-1}$. In the special case of only two possible types ($k = 2$), the LLDB prior distribution is uniform on the set $\{0, 1, 2, \ldots, n^*\}$ of possible frequencies x_1^* and agrees with the prior distribution proposed in Bayes (1763).

Geisser (1984) suggests that this prior distribution is an appropriate model when "prior information is lacking" about the types of units. But the LLDB model is based on the assumption that, before any observations have been made, we know exactly what types are possible. Inferences depend on what types are initially regarded as equally probable and on the number of distinct types, k. In the example of balls in an urn, if we classify the possible colours according to $\Omega_1 = \{\textit{black}, \textit{other colours}\}$ then there is prior probability $\frac{1}{2}$ that the first ball drawn from the urn will be black, but this probability changes to $\frac{1}{3}$ if we reclassify the colours according to $\Omega_2 = \{\textit{black}, \textit{white}, \textit{other colours}\}$.

This dependence of the LLDB inferences on the choice of Ω, and the paradoxical results it produces, was pointed out and criticised by, amongst others, Boole (1854), Peirce (1878), Bing and Hardy (see Dale, 1991, pp. 334–338, 353–357). It appears that this criticism was a very important factor in bringing about the widespread rejection of Bayesian methods in the late nineteenth and early twentieth centuries.

The LLDB prior distribution can be derived also by assuming that the type of each of the n^* observations is determined independently according to a multinomial process with chances $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ for each type, and assuming a uniform prior density for $\boldsymbol{\theta}$. Several other prior densities have been proposed to model ignorance about the chances $\boldsymbol{\theta}$. We will consider especially the models of Jeffreys (1946, 1961), Perks (1947), and Haldane (1948), each of which generates predictive inferences that differ from the LLDB model. Because parametric inferences about $\boldsymbol{\theta}$ (infinite population) can be regarded as limits of predictive inferences about $\boldsymbol{x}^*$ (finite population), there appears to be a serious inconsistency in adopting the LLDB model as the appropriate noninformative prior for predictive inference, and at the same time adopting a different model, such as Haldane's, Jeffreys' or Perks', as the appropriate one for inferences about $\boldsymbol{\theta}$. In this paper we consider the LLDB model as the main objective Bayesian alternative to our model, since the invocation of an underlying multinomial process seems unnecessary in most problems and may be inappropriate if the population that is sampled is actually finite.

1.4 Motivation for our model

To apply any of these objective Bayesian methods we must know, before any observations have been made, exactly what outcomes are possible. This is a restrictive and unrealistic assumption for many kinds of categorical data. For instance, suppose that, while on holiday, we find a small pond and start fishing: we would be unable to list

all the types of fish that we might catch, and the inferences we draw from the first three fish we catch should not depend on what types we identified beforehand. "Prior ignorance" often involves ignorance about what outcomes are possible. Typically a *possibility space* Ω (a set of possible types) can be identified only after some outcomes are observed, and it evolves as new observations are obtained.

It is therefore desirable to find a method of predictive inference for which predictions do not depend on the initial choice of possibility space. There is no proper objective Bayesian model which satisfies this property and is symmetric between the possible types. But there are models involving *upper and lower probabilities* which satisfy these properties.

The most basic property we require is the *embedding principle* (Walley, 1991): the *prior* probabilities, or upper and lower probabilities, assigned to an observable event should not depend on the possibility space in which the event is represented. A second property, called the *representation invariance principle* (Walley, 1996a), strengthens the embedding principle to apply also to *posterior* probabilities. The LLDB model violates these principles: in the urn example of subsection 1.3, the prior probability of drawing a black ball, and the posterior probability after observing that the first ball drawn is black, depend on whether the possibility space is taken to be Ω_1 or Ω_2. The models of Jeffreys and Perks violate the two principles in a similar way.

The model studied in this paper satisfies both principles. A related property is that the model produces highly imprecise prior probabilities concerning future events. For example, the event that the first observation will be of a specific type has prior upper probability 1 and lower probability 0. This is an important property in a model for prior ignorance.

Because the prior probabilities are imprecise, our model encompasses a range of objective Bayesian prior distributions. In the case where only two types are distinguished, our model encompasses the Bayesian predictive models suggested by Bayes and Laplace (the LLDB model), Haldane, and Jeffreys (which agrees with Perks' model in this case), and it covers what has been called an *ignorance zone* (Bernard, 1996).

Our model also satisfies the likelihood principle, as the LLDB model does. That is not true of some of the other objective Bayesian methods, as pointed out by Geisser (1984). In particular, the predictive prior obtained from Jeffreys' method depends on the sampling rule.

Another property of our model is that, as observations are sampled, the posterior upper and lower probabilities of a future event converge, and the difference between them reflects the observed sample size. This differs from the behaviour of Bayesian inferences. For example, suppose we observe $x_1 = \frac{n}{2}$ successes in a sample of n, and consider the predictive probability of success on a future trial. Under the LLDB model this probability is $\frac{1}{2}$ whatever the sample size n, even if $n = 0$. Under our model, the upper and lower probabilities are $\frac{1}{2} + \varepsilon(n)$ and $\frac{1}{2} - \varepsilon(n)$, where $\varepsilon(n)$ decreases from $\frac{1}{2}$ to 0 as n increases from 0 to ∞. Thus our model distinguishes between ignorance and an equal balance of data, through the precision of posterior probabilities.

Unlike some of the objective Bayesian approaches, our approach produces the same predictive inferences whether we model the observables directly or introduce a

multinomial model with unobservable parameters. The model presented here is closely related to the imprecise Dirichlet model for uncertainty about multinomial parameters (Walley, 1996a), but here we give an alternative formulation which refers only to observable quantities and we study in detail the predictive inferences from the model. The special case of two categories (a Bernoulli process) was also studied in Walley (1991) and Bernard (1996).

1.5 Outline of the paper

The model for predictive inference that is presented in this paper, which we call the IDMM, is defined in Sections 2 and 3 as a set of Dirichlet-multinomial prior or posterior distributions. Section 2 summarises the general properties of the IDMM as a model for prior ignorance about population frequencies, and Section 3 describes the general properties of the predictive inferences it produces.

Some special types of predictive inference are studied in more detail in Sections 4–6, and the predictions from the IDMM are compared with objective Bayesian and frequentist inferences. Section 4 is concerned with predictions about a single future observation. Section 5 concerns predictions about the number of successes in a future sample, which can be summarised by graphing posterior upper and lower cdfs or calculating prediction sets. Section 6 concerns the problem of confirming a universal hypothesis, i.e., a hypothesis that all members of a population have some property, when all members of a sample are found to have the property.

In Section 7, the IDMM is used to analyse some economic data collected in a survey of French companies, in order to make inferences about indices which measure the outlook for French industrial production. The concluding comments in Section 8 summarise the properties of the IDMM and identify some important problems for future research.

2 The IDMM as a model for prior ignorance

2.1 Notation

We use the symbols $n, \boldsymbol{x}$ and $\boldsymbol{f} = \boldsymbol{x}/n$ respectively to denote the sample size, frequencies and relative frequencies of each category in the observed sample. The corresponding quantities in the unobserved (future) sample are denoted by $n', \boldsymbol{x}'$ and $\boldsymbol{f}'$, and generally a prime superscript ($'$) always indicates a characteristic of the future sample. Similarly an asterisk superscript (*) is used to denote a characteristic of the combined (past plus future) observations, so that $n^* = n + n', \boldsymbol{x}^* = \boldsymbol{x} + \boldsymbol{x}'$, and $\boldsymbol{f}^* = (n\boldsymbol{f} + n'\boldsymbol{f}')/(n + n')$. Bold symbols ($\boldsymbol{x}$, etc.) are used to denote vectors, and capitals ($\boldsymbol{X}$, etc.) to denote random variables.

Formulae for probabilities and upper and lower probabilities will be expressed in terms of *generalized binomial coefficients* $\binom{y}{u}$, which are defined, for all real y and nonnegative integers u, by $\binom{y}{u} = y_{[u]}/u!$, where $y_{[u]}$ denotes the descending factorial function, $y_{[u]} = \prod_{i=0}^{u-1}(y - i)$ for $u > 0$ and $y_{[0]} = 1$ (Feller 1968, p. 50).

For all reals y and nonnegative integers u such that $y \geq u$, the generalized binomial coefficient can also be expressed in terms of the gamma function through $\binom{y}{u} = \Gamma(y+1)/\Gamma(u+1)\Gamma(y-u+1)$, or in terms of the beta function. When both y and u are nonnegative integers, $\binom{y}{u} = y!/[u!(y-u)!]$ is the usual binomial coefficient. Most of the later formulae can therefore be expressed in terms of gamma or beta functions, and many of them can be completely expressed in terms of factorials.

Most later formulae involve sums or products over an index i which represents a category or type and runs from 1 to k (the number of types). To simplify equations, $\sum_i$ and $\prod_i$ will be used as shorthands for $\sum_{i=1}^{k}$ and $\prod_{i=1}^{k}$ respectively.

2.2 Definition of the IDMM

We will define an imprecise (upper and lower) probability model in terms of a set of prior probability distributions for $\boldsymbol{X}^* = (X_1^*, X_2^*, \ldots, X_k^*)$, the combined frequencies from n^* observations. This set will be composed of distributions from the Dirichlet-multinomial family.

The *Dirichlet-multinomial* distribution with hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ and n^*, in short $DiMn(\boldsymbol{\alpha}, n^*)$, assigns probabilities

$$P_{\boldsymbol{\alpha}}(\boldsymbol{x}^*) = \prod_i \binom{x_i^* + \alpha_i - 1}{x_i^*} \Big/ \binom{n^* + s - 1}{n^*} \tag{3}$$

for nonnegative integers $x_1^*, x_2^*, \ldots, x_k^*$ such that $\sum_i x_i^* = n^*$, where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are positive reals and $s = \sum_i \alpha_i$. The hyperparameters $\alpha_1, \alpha_2, \ldots, \alpha_k$ can be thought of as "prior strengths" allocated to each of the k categories as they are homogeneous with the observed frequencies. As done in Walley (1996a) for the Dirichlet distribution, it is sometimes convenient to reparametrise the *DiMn* distribution in terms of the total prior strength s and the relative prior strengths $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_k)$, with $\varphi_i = \alpha_i/s$ so that $\sum_i \varphi_i = 1$. Hence $P_{\boldsymbol{\alpha}}(\boldsymbol{x}^*)$ can also be written as $P_{s\boldsymbol{\varphi}}(\boldsymbol{x}^*)$. The *DiMn* distribution has means $E(X_i^*) = n^*\varphi_i$, so that $E(F_i^*) = \varphi_i$. Thus the relative prior strengths $\boldsymbol{\varphi}$ are the prior means of the relative frequencies $\boldsymbol{F}^*$.

The uniform prior distribution proposed by LLDB is *DiMn* with $\alpha_i = 1$ for all i, i.e., $\varphi_i = \frac{1}{k}$ and $s = k$. The *DiMn* distribution is also referred to in some of the literature as a *Dirichlet-compound multinomial*, a *compound multinomial*, a *multivariate beta-binomial* or a *negative multivariate hypergeometric* distribution. Further properties of the *DiMn* may be found in Mosimann (1962), Hoadley (1969) and Johnson & Kotz (1997, pp. 80–83 and 202–211).

To model prior ignorance about the frequencies $\boldsymbol{X}^*$, we use the set of *DiMn* distributions:

$$\begin{aligned} &\{P_{\boldsymbol{\alpha}} : \alpha_i > 0 \text{ for } i = 1, \ldots, k, \sum_i \alpha_i = s\} \\ &= \{P_{s\boldsymbol{\varphi}} : \varphi_i > 0 \text{ for } i = 1, \ldots, k, \sum_i \varphi_i = 1\}, \end{aligned} \tag{4}$$

where s and n^* are fixed. In (4), $\boldsymbol{\varphi}$ ranges over the interior of the unit simplex. We call the imprecise probability model (4) the *imprecise Dirichlet-multinomial model* (IDMM)

with hyperparameter s, which we write as IDMM(s). This is the model studied in the rest of this paper. Strictly the model also depends on k, the number of categories, and on n^*, the number of observations that are considered, but this dependence is not crucial as the model produces essentially the same inferences when different values of k or n^* are used. We shall therefore omit the reference to k and n^*.

If D is any set of possible values for $\boldsymbol{x}^*$, the prior *upper and lower probabilities* of D under the IDMM(s), denoted by $\overline{P}(D)$ and $\underline{P}(D)$, are calculated by maximising and minimising the probability $P_{s\boldsymbol{\varphi}}(D)$ with respect to $\boldsymbol{\varphi}$ in the interior of the unit simplex. Clearly the upper and lower probabilities are related by $\overline{P}(D) = 1 - \underline{P}(D^c)$. The prior *degree of imprecision* concerning event D is defined as $\Delta(D) = \overline{P}(D) - \underline{P}(D)$, which varies between 0 (*precise* probabilities) and 1 (*vacuous*, or maximally imprecise, probabilities).

Similarly, for any real-valued function $V(\boldsymbol{x}^*)$, prior *upper and lower expectations* $\overline{E}(V)$ and $\underline{E}(V)$ are calculated by maximising and minimising the expectation $E_{s\boldsymbol{\varphi}}(V) = \sum_{\boldsymbol{x}^*} V(\boldsymbol{x}^*) P_{s\boldsymbol{\varphi}}(\boldsymbol{x}^*)$ with respect to $\boldsymbol{\varphi}$. They are related by $\overline{E}(V) = -\underline{E}(-V)$. In many problems it is necessary to report inferences in terms of upper and lower expectations because upper and lower probabilities may not be sufficiently informative. The reason is that, unlike the case of ordinary probability where there is a one-to-one correspondence between probability measures and expectation functions, many different upper and lower expectation functions may produce the same upper and lower probabilities. See Walley (1991, 1996b) for examples.

2.3 Near-ignorance properties of the prior IDMM

To model prior ignorance, the IDMM is defined to include the widest possible range of values for the prior means $E(X_i^*) = n^* \varphi_i$, which range from 0 to n^* as φ_i ranges from 0 to 1. The IDMM has several other properties which are important for modelling prior ignorance.

(a) It is *symmetric* in the k categories. For example, the prior upper and lower probabilities of any set of possible observations are invariant under any permutations of the k categories.

(b) It satisfies the *embedding principle*. Suppose that we pool categories i and h into a single category, denoted by $i + h$. From a single *DiMn* prior on the initial k-categorisation, the marginal prior on the new $(k-1)$-categorisation is still *DiMn*, with $\alpha_{i+h} = \alpha_i + \alpha_h$ and all other hyperparameters unchanged, so that the total prior strength $s = \sum_m \alpha_m$ is unchanged too. By applying this result recursively we see that, for any pooling of the k categories into $j < k$ pooled ones, both the *DiMn* form and the value of s are preserved. Thus the upper and lower probabilities of any event which can be described in terms of a pooled frequency vector $\boldsymbol{y}^*$, with $j < k$ pooled categories, can be obtained by either of two equivalent methods: (i) apply the IDMM to the k-categorization and then derive a marginal model for $\boldsymbol{y}^*$; or (ii) pool the k categories into a j-categorization and then apply the IDMM to the j-categorization. The two operations, pooling categories and formalizing uncertainty by the IDMM, commute.

(c) The prior upper and lower probabilities and expectations that are produced by the IDMM are *highly imprecise*, which indicates a high level of indeterminacy or

cautiousness in decisions based on the prior IDMM. As a first example, let C_i be the event that a particular unit to be observed will be of type i. Taking $x_i^* = n^* = 1$ in (3), we see that $P_{s\boldsymbol{\varphi}}(C_i) = \varphi_i$. By maximising and minimising with respect to φ_i in $(0,1)$, we obtain $\overline{P}(C_i) = 1$ (achieved as $\varphi_i \to 1$) and $\underline{P}(C_i) = 0$ (achieved as $\varphi_i \to 0$), and hence $\Delta(C_i) = 1$. Thus the upper and lower probabilities that any particular observation will be of any particular type are vacuous.

Next consider the prior *upper and lower cumulative distribution functions (cdfs)* for X_i^*, which are defined by $\overline{F}_i(u) = \overline{P}(X_i^* \leq u)$ and $\underline{F}_i(u) = \underline{P}(X_i^* \leq u)$. It can be verified that $P_{s\boldsymbol{\varphi}}(X_i^* = 0) \to 1$ as $\varphi_i \to 0$ and $P_{s\boldsymbol{\varphi}}(X_i^* = n^*) \to 1$ as $\varphi_i \to 1$, from which it follows that $\overline{F}_i(u) = 1$ if $u \geq 0$ and $\underline{F}_i(u) = 0$ if $u < n^*$. Thus the prior upper and lower cdfs for X_i^* are also vacuous. Similarly we find that $\overline{P}(X_i^* < X_j^*) = 1$ and $\underline{P}(X_i^* < X_j^*) = 0$ whenever $i \neq j$, so that prior beliefs about which of two categories will have greater frequency are also vacuous.

However, not every observable event D has vacuous prior upper and lower probabilities. As a simple example, suppose that there are two balls in an urn and each ball is either black or white. Let X_1^* denote the number of black balls, so the possible values of X_1^* are 0, 1 and 2. Under the IDMM(s), the prior upper and lower probabilities for $X_1^* = 0$ are 1 and 0, the vacuous probabilities, and similarly for $X_1^* = 2$. For $X_1^* = 1$ the lower probability is 0 but the upper probability is $\frac{s}{2(s+1)}$, which is less than $\frac{1}{2}$ and tends to $\frac{1}{2}$ as $s \to \infty$. Thus the IDMM appears to contain nonnegligible prior information that X_1^* is not equal to 1.

This example indicates that the IDMM does embody some assumptions about the observation process. In effect, as shown in Section 3.4, the IDMM involves an assumption that the types are chosen randomly and independently according to a multinomial process. In the last example, this means that the colours of the two balls are chosen independently, with some constant chance θ_1 that a particular ball is black. For any possible value of θ_1, $P(X_1^* = 1|\theta_1) = 2\theta_1(1-\theta_1) \leq \frac{1}{2}$. This explains why $\overline{P}(X_1^* = 1) \leq \frac{1}{2}$ under the IDMM.

2.4 Choice of the hyperparameter s

Inferences from the IDMM depend on the value of s, which may be any positive real number. If B is any event concerning future observations, the IDMM(s) produces intervals of posterior probabilities $[\underline{P}(B|\boldsymbol{x}), \overline{P}(B|\boldsymbol{x})]$ which are nested and become wider as s increases. This means that the inferences produced by two IDMMs with different values of s are always consistent with each other, and the effect of increasing s is simply to make inferences more cautious and less informative. Large values of s are unreasonable because the IDMM(s) produces usefully informative predictions only when the observed sample size is larger than s.

Values of s close to 0 are also unreasonable, because they imply very strong prior beliefs that all the observations in a future sample will be identical. For example, the prior lower probability that the first two observations will be either both successes or both failures is $\frac{2+s}{2+2s}$, which approaches 1 as $s \to 0$. As $s \to 0$, the posterior upper and lower probabilities from the IDMM(s) (based on at least one observation) converge to

a precise probability $P(B|\boldsymbol{x})$ which agrees with the posterior probability generated by Haldane's improper prior density.

Several other arguments which can guide us in the choice of s will be discussed when we consider inferences from the IDMM, in Sections 4–6 (especially 5.5) and 8. In this paper we will use $s = 1$ and $s = 2$ as standard values for the IDMM. We currently prefer the value $s = 1$, which is supported by comparisons with objective Bayesian inferences and with frequentist inferences in hypergeometric and binomial problems. In these problems, statistical procedures based on the IDMM with $s = 1$ are both valid from a frequentist point of view and reasonably powerful. For comparison we also include some more cautious inferences based on $s = 2$, which may be more acceptable to those who do not find these arguments convincing.

3 Inferences from the IDMM

3.1 The observed likelihood function

Suppose that we make a sequence of observations. We first observe an initial sample of size n, which is used as data to make predictions about a future sample of size n'. Here n and n' may be random, but we assume that n' is either fixed or a known function of the observations $\boldsymbol{x}$ and n. This includes the most important cases where either we intend to make predictions about a fixed number of future observations, n', or we want to make inferences about the composition of a finite population of known size n^* so that $n' = n^* - n$ is a known function of n, which may be random. We also assume that the stopping rule which determines n is *deterministic*, meaning that whether we stop sampling at any time is determined by the sequence of observations up to that time. This includes the most common types of stopping rules such as direct sampling (with fixed n) and inverse sampling (with fixed x_i). More generally, we could allow any "noninformative" stopping rule; see Raiffa & Schlaifer (1961, Ch. 2) or Bernardo & Smith (1994, pp. 250–252).

Let $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)$ denote the ordered vector of types that is observed in the initial sample, which constitutes all the observed data using the assumption of a deterministic stopping rule. After the initial sample $\boldsymbol{\omega}$ is observed, the observed sample size n and frequencies $\boldsymbol{x}$ are known, and hence n' and $n^* = n + n'$ are known, using the assumption that n' is a known function of $\boldsymbol{x}$ and n. We make the following assumption of *exchangeability*: conditional on the combined frequencies $\boldsymbol{x}^*$, all the $n^*!$ possible orderings of the n^* observations are equally probable.

Given $\boldsymbol{x}^*$ and $\boldsymbol{\omega}$, there are $x_i^*!/(x_i^* - x_i)! = x_i! \binom{x_i^*}{x_i}$ ways of choosing x_i ordered observations of type i for the initial sample $\boldsymbol{\omega}$, and there are $n'!$ ways of ordering the future observations. Hence there are $n'! \prod_i \left[x_i! \binom{x_i^*}{x_i} \right]$ orderings of the n^* observations which produce the ordered initial sample $\boldsymbol{\omega}$. It follows from the exchangeability assumption that, for any possible $\boldsymbol{\omega}$, the probability of obtaining data $\boldsymbol{\omega}$ conditional on $\boldsymbol{x}^*$ is

$$P(\boldsymbol{\omega}|\boldsymbol{x}^*) = (n^*!)^{-1}\ n'!\ \prod_i \left[x_i! \binom{x_i^*}{x_i} \right]. \tag{5}$$

As a function of the unknown $\boldsymbol{x}^*$, $P(\boldsymbol{\omega}|\boldsymbol{x}^*)$ is proportional to $\prod_i \binom{x_i^*}{x_i}$. Hence the observed likelihood function for $\boldsymbol{x}^*$ that is generated by the initial sample $\boldsymbol{\omega}$ is

$$L(\boldsymbol{x}^*|\boldsymbol{\omega}) = L(\boldsymbol{x}^*|\boldsymbol{x}) \propto \prod_i \binom{x_i^*}{x_i}, \tag{6}$$

for nonnegative integers $x_i^* \geq x_i$ such that $\sum_i x_i^* = \sum_i x_i + n'$.

Inferences about $\boldsymbol{x}^*$ will depend on the initial sample $\boldsymbol{\omega}$ only through the observed likelihood function for $\boldsymbol{x}^*$, which depends only on the vector of observed frequencies $\boldsymbol{x}$. Thus $\boldsymbol{x}$ is a *sufficient statistic* for inferences about $\boldsymbol{x}^*$.

One can gain a more intuitive understanding of how the exchangeability assumption determines the likelihood function (6) by thinking of the observation process as a random path, as in de Finetti (1974/1975, Vol. 2, pp. 20–21) for the case $k = 2$. First consider multiple hypergeometric sampling (n fixed). Let $\binom{n^*}{\boldsymbol{x}^*}$ denote the multinomial coefficient. There are $\binom{n^*}{\boldsymbol{x}^*}$ paths leading from the initial state $(0, 0, \ldots, 0)$ to the combined frequency vector $\boldsymbol{x}^*$. Amongst these, the number of paths that pass through the observed frequencies $\boldsymbol{x}$ is $\binom{n}{\boldsymbol{x}}$ (number of paths from the origin to $\boldsymbol{x}$) times $\binom{n'}{\boldsymbol{x}'}$ (number of paths from $\boldsymbol{x}$ to $\boldsymbol{x}^*$). Since all paths leading to a single point $\boldsymbol{x}^*$ are equally probable, by the exchangeability assumption, the conditional probability of $\boldsymbol{x}$ given $\boldsymbol{x}^*$ is $\binom{n}{\boldsymbol{x}}\binom{n'}{\boldsymbol{x}'} / \binom{n^*}{\boldsymbol{x}^*}$. Because n, $\boldsymbol{x}$ and n' are fixed, the likelihood for $\boldsymbol{x}^*$ is proportional to $(\prod_i x_i^*)/(\prod_i x_i')$, which is equivalent to (6).

For inverse sampling with x_1 fixed, the number of paths passing through $\boldsymbol{x}$ is $\binom{n-1}{\boldsymbol{y}}\binom{n'}{\boldsymbol{x}'}$, where $\boldsymbol{y} = (x_1 - 1, x_2, \ldots, x_k)$, since every path which produces observed frequencies $\boldsymbol{x}$ must pass through $\boldsymbol{y}$. In this case, the conditional probability of $\boldsymbol{x}$ given $\boldsymbol{x}^*$ is $\binom{n-1}{\boldsymbol{y}}\binom{n'}{\boldsymbol{x}'} / \binom{n^*}{\boldsymbol{x}^*}$, which again leads to a likelihood proportional to (6).

3.2 Model for posterior uncertainty: the posterior IDMM

To calculate inferences about the future frequencies $\boldsymbol{X}'$ after observing the frequencies $\boldsymbol{x}$, we need to combine the likelihood function (6) with the prior IDMM. The prior IDMM was defined in (4) as a set of *DiMn* prior distributions $P_{\boldsymbol{\alpha}}$. To update the prior IDMM to a set of posterior distributions, we use Bayes' theorem to update each of the prior distributions $P_{\boldsymbol{\alpha}}$. This method of updating guarantees coherence of upper and lower probabilities.

For a single *DiMn* prior $P_{\boldsymbol{\alpha}}$, the posterior distribution, regarded as a function of the variable $\boldsymbol{x}^*$ with $\boldsymbol{x}$ fixed, is

$$P_{\boldsymbol{\alpha}}(\boldsymbol{x}^*|\boldsymbol{x}) \propto L(\boldsymbol{x}^*|\boldsymbol{x}) P_{\boldsymbol{\alpha}}(\boldsymbol{x}^*) \propto \prod_i \binom{x_i^*}{x_i} \prod_i \binom{x_i^* + \alpha_i - 1}{x_i^*} \propto \prod_i \binom{x_i^* + \alpha_i - 1}{x_i^* - x_i}. \tag{7}$$

Substituting $x_i^* = x_i + x_i'$ and comparing with (3), we see that a $DiMn(\boldsymbol{\alpha}, n^*)$ prior distribution for $\boldsymbol{X}^*$ produces a $DiMn(\boldsymbol{x}+\boldsymbol{\alpha}, n')$ posterior distribution for $\boldsymbol{X}' = \boldsymbol{X}^* - \boldsymbol{x}$:

$$P_{\boldsymbol{\alpha}}(\boldsymbol{x}'|\boldsymbol{x}) = \prod_i \binom{x_i' + x_i + \alpha_i - 1}{x_i'} \Big/ \binom{n' + n + s - 1}{n'} \tag{8}$$

for nonnegative integers $x'_1, x'_2, \ldots, x'_k$ such that $\sum_i x'_i = n'$. The prior strengths $\boldsymbol{\alpha}$ are increased by the observed frequencies $\boldsymbol{x}$ to give posterior strengths $\boldsymbol{x} + \boldsymbol{\alpha}$, while the sample size $n^* = n' + n$ is reduced by n so that the posterior prediction bears on the n' unobserved units. For example, the LLDB posterior distribution is given by (8) with $s = k$ and each $\alpha_i = 1$.

The IDMM(s) is the set of all $DiMn(\boldsymbol{\alpha}, n^*)$ prior distributions which satisfy $\boldsymbol{\alpha} = s\boldsymbol{\varphi}, \varphi_i > 0$ and $\sum_i \varphi_i = 1$. Hence the IDMM prior can be easily updated: after observing the frequencies $\boldsymbol{x}$, the IDMM is updated to the set of all $DiMn(\boldsymbol{x} + s\boldsymbol{\varphi}, n')$ posterior distributions on $\boldsymbol{X}'$ with the same constraints on $\boldsymbol{\varphi}$, i.e.,

$$\{P_{\boldsymbol{x}+s\boldsymbol{\varphi}} : \varphi_i > 0 \text{ for } i = 1, 2, \ldots, k, \sum_i \varphi_i = 1\}. \tag{9}$$

This set of posterior distributions models the uncertainty about the future observations $\boldsymbol{X}'$ after observing the data $\boldsymbol{x}$.

We can make inferences from the IDMM by calculating posterior upper and lower probabilities of any event B or expectations of any function $V = V(\boldsymbol{x}')$, which are denoted by $\overline{P}(B|\boldsymbol{x}), \underline{P}(B|\boldsymbol{x}), \overline{E}(V|\boldsymbol{x})$ and $\underline{E}(V|\boldsymbol{x})$, by maximising and minimising $P_{\boldsymbol{\alpha}}(B|\boldsymbol{x}) = \sum_{\boldsymbol{x}' \in B} P_{\boldsymbol{\alpha}}(\boldsymbol{x}'|\boldsymbol{x})$ or $E_{\boldsymbol{\alpha}}(V|\boldsymbol{x}) = \sum_{\boldsymbol{x}'} V(\boldsymbol{x}') P_{\boldsymbol{\alpha}}(\boldsymbol{x}'|\boldsymbol{x})$ with respect to $\boldsymbol{\alpha}$, where $P_{\boldsymbol{\alpha}}(\boldsymbol{x}'|\boldsymbol{x})$ is given by (8). Some examples of these calculations will be given in the following sections.

3.3 General properties of the inferences

The inferences derived from the IDMM have the following general properties.
(a) They are *coherent*, in the strongest sense of Walley (1991, Section 7.1). The coherence property requires the decisions produced by several imprecise probability models to be mutually consistent. It generalises, and in some ways strengthens, the Bayesian concept of coherence (de Finetti, 1974/1975). In this problem, coherence of the inferences means that three models are mutually consistent: the prior upper and lower probabilities and expectations generated by the prior IDMM; the statistical sampling models which produce the likelihood functions (6); and the posterior upper and lower expectations generated by the posterior IDMM for all possible data $\boldsymbol{x}$.
(b) As can be seen from (7), inferences depend on the data and the sampling model only through the observed likelihood function $L(\cdot|\boldsymbol{x})$, since the class of prior distributions (4) does not depend on the particular sampling model. Therefore inferences from the IDMM satisfy the *likelihood principle* (Berger & Wolpert, 1984). For example, inferences are the same whether the data $\boldsymbol{x}$ are obtained from direct sampling with fixed n and random x_i, or from inverse sampling with fixed x_i and random n, because the likelihood function $L(\cdot|\boldsymbol{x})$ is proportional in the two cases.
(c) Inferences do not depend on what types are distinguished, nor even on the number of types, k. Formally they satisfy the *representation invariance principle* (Walley, 1996a), which holds for the same reasons as the embedding principle (subsection 2.3(b)): when categories are pooled, both the form of a *DiMn* distribution and the value of $s = \sum_i \alpha_i$ are preserved, so that the set of posterior *DiMn* distributions produced by the IDMM(s) is essentially unchanged. See Walley (1996a, p. 16) for a formal proof.

3.4 Relationships between the IDMM and the IDM

Next we explore the relationships between predictions from the IDMM and parametric inference. In the Bayesian framework, two relationships always exist. Firstly, when the future sample size n' tends to ∞, the posterior predictive distribution $P(\boldsymbol{x}'|\boldsymbol{x})$ tends to the corresponding posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{x})$ for population parameters $\boldsymbol{\theta}$. Secondly, applying Bayes' theorem to $P(\boldsymbol{\theta}|\boldsymbol{x})$, regarded as a prior distribution, and the hypothetical future observations $\boldsymbol{x}'$, sampled according to $P(\boldsymbol{x}'|\boldsymbol{\theta})$, yields the posterior predictive distribution $P(\boldsymbol{x}'|\boldsymbol{x})$. Thus the predictive distribution can also be viewed as a by-product of the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{x})$.

Since the updating process for the IDMM uses Bayes' theorem, the same two relationships exist here and both involve the *imprecise Dirichlet model* (IDM) proposed in Walley (1996a) for inferences from multinomial data. (For applications of the IDM, see also Walley, Gurrin & Burton (1996) and Bernard (1998b).) Under the IDM with parameter s, prior ignorance about the chances $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ of a multinomial process is modelled by the set of all Dirichlet($\boldsymbol{\alpha}$) prior distributions for $\boldsymbol{\theta}$ such that each $\alpha_i > 0$ and $\sum_i \alpha_i = s$. This set is updated to the set of all Dirichlet($\boldsymbol{x} + \boldsymbol{\alpha}$) posterior distributions for $\boldsymbol{\theta}$ satisfying the same constraints on $\boldsymbol{\alpha}$.

Let $f'_i = x'_i/n'$ be the proportion of the future sample that is of type i. It can be shown that, as $n' \to \infty$ with $\boldsymbol{f}'$ fixed, the posterior probability $P_{\boldsymbol{\alpha}}(\boldsymbol{f}'|\boldsymbol{x})$ obtained from (8) is asymptotically proportional to $\prod_i (f'_i)^{x_i+\alpha_i-1}$, which is a Dirichlet density with hyperparameters $\boldsymbol{x}+\boldsymbol{\alpha}$. Hence, when n' is large, the set of posterior distributions for $\boldsymbol{f}'$ under the IDMM(s) can be approximated by the set of Dirichlet($\boldsymbol{x} + \boldsymbol{\alpha}$) distributions. This shows that inference about a multinomial process using the IDM(s) can be regarded as a limit of inferences about frequencies in a finite population using the IDMM(s), as the population size becomes arbitrarily large.

Now consider the second type of relationship. The IDMM(s) can also be derived from the IDM(s). The simplest way to understand this is to imagine that the type of each observation in the past and future samples is chosen randomly and independently, according to a multinomial process with chances $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ for each type. If prior ignorance about the chances $\boldsymbol{\theta}$ is modelled by the IDM(s), each component of the model is a prior Dirichlet($\boldsymbol{\alpha}$) distribution for $\boldsymbol{\theta}$ which, together with the multinomial distribution of $\boldsymbol{X}^*$ given $\boldsymbol{\theta}$, generates (by integrating out $\boldsymbol{\theta}$) a $DiMn(\boldsymbol{\alpha}, n^*)$ distribution for $\boldsymbol{X}^*$. The set of all such distributions, as $\boldsymbol{\alpha}$ takes all its possible values, is just the IDMM(s). Thus the IDMM(s) represents the predictive distributions for observable frequencies that are implied by the IDM(s). This derivation suggests an interesting interpretation of the IDMM(s). It produces the same upper and lower probabilities as if the types of all units were chosen at random, according to a multinomial process, but we were initially ignorant about the chances of each type and used the IDM(s) to model prior ignorance.

3.5 Objective Bayesian models

Any objective Bayesian prior distribution for $\boldsymbol{\theta}$ generates a predictive distribution for $\boldsymbol{X}'$ given $\boldsymbol{x}$, and it is of interest to compare these predictive distributions with the predictions from the IDMM. Four symmetric Dirichlet densities have been proposed as

models for prior ignorance about the parameters $\boldsymbol{\theta}$. The LLDB predictive distribution corresponds to a uniform prior density for $\boldsymbol{\theta}$, which is a Dirichlet($\boldsymbol{\alpha}$) density with $\alpha_i = 1$ for all i. Jeffreys (1946, 1961) suggested the Dirichlet prior with $\alpha_i = \frac{1}{2}$. Perks (1947) proposed $\alpha_i = \frac{1}{k}$. Finally, the improper prior density of Haldane (1948) is characterized by $\alpha_i = 0$. These densities can be used to generate Bayesian alternatives to the LLDB model and the IDMM. The predictive probabilities produced by each model can be obtained by substituting the appropriate values of α_i and s in (8).

Haldane's model is a limit of the IDMM(s) as $s \to 0$ and it is effectively contained in the IDMM for any positive value of s. The IDMM(s) contains Perks' model if $s \geq 1$, but it contains the models of Jeffreys or LLDB only if $s \geq \frac{k}{2}$ or $s \geq k$ respectively. In the case where only two categories are considered ($k = 2$), the IDMM(2) contains all four Bayesian predictive models.

Another objective Bayesian approach, involving "ordered group reference priors", has been applied to multinomial data by Berger & Bernardo (1992) and Bernardo & Ramon (1998). In this approach the reference prior varies with the parameter of interest, and the method can produce inconsistent inferences when several parameters are considered. In the simple case of binomial data and inference about θ_1, the reference prior coincides with Jeffreys' prior. Kuboki (1998) uses this approach to find reference priors for predictive inference and also obtains Jeffreys' prior in the binomial case. An example of this approach will be discussed in Section 6. See Kass & Wasserman (1996) for a review of objective Bayesian priors and other references.

Most Bayesian authors, including Jeffreys, who have searched for noninformative priors for both the discrete case (e.g. multiple hypergeometric sampling) and the continuous case (e.g. multinomial sampling) have been led to propose prior distributions which produce incompatible predictive inferences: the LLDB prior in the discrete case, but Jeffreys', Haldane's or Perks' priors in the continuous case. Asymptotic arguments such as those in Section 3.4 suggest that we should require compatibility. The IDMM and IDM are compatible in this way.

4 Predictions about the next observation

To study more closely the predictive inferences produced by the IDMM, in the next three sections we examine three important types of inference. In this section we consider the simplest type of predictive inference: predictions about just one future observation. In this case $n' = 1$ and the possible frequency vectors $\boldsymbol{x}'$ are those vectors which contain 1 one and $k - 1$ zeroes. For example, a doctor might want to predict the outcome of using a particular medical treatment on his next patient, from data concerning the outcomes when the treatment was used on earlier patients.

This problem has been one of the most discussed in the history of statistics, especially in the case of $k = 2$ categories for which Laplace's rule of succession (Laplace, 1812/1825, pp. 44–46) has been vigorously debated during the last 200 years (see Dale, 1991).

4.1 The IDMM rule of succession

Let C be any nontrivial set of possible types. Write $x_C = \sum_{i \in C} x_i$, $\alpha_C = \sum_{i \in C} \alpha_i$ and $\varphi_C = \alpha_C / s$ for the observed frequency and prior strengths of the compound category C. Applying (8) with $n' = 1$, the probability of the event C under the $DiMn(\boldsymbol{x} + \boldsymbol{\alpha}, 1)$ posterior distribution is

$$P_{\boldsymbol{\alpha}}(C|\boldsymbol{x}) = \frac{x_C + \alpha_C}{n + s} = \frac{n f_C + s\varphi_C}{n + s}. \tag{10}$$

From the last expression, this probability is a weighted average of the observed relative frequency $f_C = x_C / n$ and the prior mean φ_C, with weights n and s.

By maximising and minimising this probability subject to $0 < \varphi_C < 1$, we see that the IDMM(s) produces the upper and lower probabilities

$$\overline{P}(C|\boldsymbol{x}) = \frac{x_C + s}{n + s}, \quad \underline{P}(C|\boldsymbol{x}) = \frac{x_C}{n + s}. \tag{11}$$

The range of probabilities in (11) covers the observed relative frequency f_C.

If s is fixed, $\overline{P}(C|\boldsymbol{x})$ and $\underline{P}(C|\boldsymbol{x})$ depend only on the values of x_C and n, and not on the way in which C and its complementary event are divided into simple categories. The IDMM therefore satisfies Johnson's *sufficientness postulate*, which is discussed in Johnson (1932), Good (1965) and Zabell (1982). This property is a consequence of the representation invariance principle, which implies that predictive inferences about C are unchanged if we consider that there are only two possible types of observations, C and its complementary event. For example, if we use the IDMM with $s = 1$ and we observe 1 dark ball in 3 drawings from an urn, then the upper and lower probabilities that the next ball drawn will be dark are 0.5 and 0.25, irrespective of how the possible colours are classified.

From (11), the posterior degree of imprecision concerning C is $\Delta(C|\boldsymbol{x}) = \overline{P}(C|\boldsymbol{x}) - \underline{P}(C|\boldsymbol{x}) = s/(n+s)$, which equals 1 when $n = 0$ (prior upper and lower probabilities are vacuous) and equals $\frac{1}{2}$ when $n = s$. Thus s can be interpreted as the sample size that is needed to reduce the imprecision $\Delta(C|\boldsymbol{x})$ from 1 to $\frac{1}{2}$. As $s \to \infty$, $\Delta(C|\boldsymbol{x}) \to 1$ and $[\underline{P}(C|\boldsymbol{x}), \overline{P}(C|\boldsymbol{x})] \to [0, 1]$, meaning that inferences become completely uninformative. On the other hand, if $n \geq 1$ and $s \to 0$, $\Delta(C|\boldsymbol{x}) \to 0$ and the posterior upper and lower probabilities converge to a precise probability, which is the posterior probability under Haldane's model.

4.2 Comparison with objective Bayesian inferences

Consider objective Bayesian inferences for the event that the type of the next observation belongs to some nontrivial set of types, C, and let $|C| = j$ with $1 \leq j \leq k - 1$. For a symmetric $DiMn(s\boldsymbol{\varphi}, n^*)$ prior distribution with each $\varphi_i = k^{-1}$, the posterior probability of C is found by setting $\varphi_C = \frac{j}{k}$ in (10):

$$P_{s\boldsymbol{\varphi}}(C|\boldsymbol{x}) = \frac{x_C + s\frac{j}{k}}{n + s}. \tag{12}$$

The four common objective Bayesian models are obtained from (12) by taking s to be 0 (Haldane), 1 (Perks), $\frac{k}{2}$ (Jeffreys) or k (LLDB).

For example, the LLDB prior produces $P(C|\boldsymbol{x}) = (x_C + j)/(n + k)$. This formula is due to Laplace (1778, p. 483). In the case of two categories ($k = 2$), the LLDB predictive probability reduces to Laplace's rule of succession, $P(C|\boldsymbol{x}) = (x_C + 1)/(n + 2)$. This probability is exactly halfway between the upper and lower probabilities generated by the IDMM with $s = 2$. Similarly the predictive probability from the models of Perks and Jeffreys, which is $P(C|\boldsymbol{x}) = (x_C + 0.5)/(n + 1)$ in the case $k = 2$, is exactly halfway between the upper and lower probabilities from the IDMM with $s = 1$.

It is always possible to argue that there is some arbitrariness in the initial k-categorization. The most extreme view of this arbitrariness is to consider all possible coarsenings and refinements of the possibility space. For the inferences in (12), this amounts to considering all possible values of j and k satisfying the constraint $1 \leq j \leq k - 1 < \infty$. Under the LLDB and Jeffreys' models, $P_{s\boldsymbol{\varphi}}(C|\boldsymbol{x})$ can take all rational values in $(0, 1)$ when j and k vary in this way, whatever the observed data. In the example where we observe 1 dark ball in 3 drawings, if "dark" comprises j simple categories then the LLDB predictive probability that the next ball will be dark is $\frac{1+j}{3+k}$, which can range from 0 to 1 depending on the sizes of j and k.

Thus the LLDB and Jeffreys' models produce vacuous probabilistic statements when categorization arbitrariness is taken into account. On the other hand, the value of $P_{s\boldsymbol{\varphi}}(C|\boldsymbol{x})$ for Haldane's model is always x_C/n, the observed relative frequency of the event. Finally, varying j and k in Perks' model, $P_{s\boldsymbol{\varphi}}(C|\boldsymbol{x})$ can take all rational values in $(x_C/(n + 1), (x_C + 1)/(n + 1))$, which is the interval derived from the IDMM with $s = 1$. (See Perks 1947, p. 308.) More generally, the predictive upper and lower probabilities that are produced by the IDMM(s) can be obtained from a single symmetric $DiMn(s\boldsymbol{\varphi}, n^*)$ prior with each $\varphi_i = \frac{1}{k}$, by considering all possible coarsenings and refinements of the possibility space.

These considerations show that taking categorization arbitrariness into account leads naturally to imprecise probabilities. In particular, Perks' model leads to the IDMM(1). Perks (1947, pp. 305–308) seems to have been close to proposing the idea of upper and lower probabilities in his discussion of categorization arbitrariness.

5 Predictions about the number of successes in future trials

In this section we consider predictive inferences about the number of *successes* in n' future observations, where an observation is counted as a success if and only if it belongs to a fixed set of categories C. Because the IDMM satisfies the representation invariance principle, predictions about the number of successes are unchanged if we redefine the possibility space to contain only two categories, the first representing success (the occurrence of C) and the second representing failure (the nonoccurrence of C). The number of successes in the future sample is therefore denoted by X'_1 and the number of failures by X'_2. The possible values of X'_1 are $0, 1, 2, \ldots, n'$.

5.1 Beta-binomial distributions

Under the *DiMn* prior distribution (3), the marginal distribution of $X_1^* = X_1 + X_1'$ is a *beta-binomial* distribution, written as $BeBi(\alpha_1, \alpha_2, n^*)$, which assigns probabilities

$$P_{\boldsymbol{\alpha}}(x_1^*) = \binom{x_1^* + \alpha_1 - 1}{x_1^*} \binom{x_2^* + \alpha_2 - 1}{x_2^*} \Big/ \binom{n^* + s - 1}{n^*} \tag{13}$$

for $x_1^* = 0, 1, \ldots, n^*$, where $x_1^* + x_2^* = n^*$ and $\alpha_1 + \alpha_2 = s$. This distribution has mean $E(X_1^*) = n^*\varphi_1$ and variance $Var(X_1^*) = n^*(n^* + s)(s + 1)^{-1}\varphi_1(1 - \varphi_1)$.

After observing frequencies $\boldsymbol{x} = (x_1, x_2)$ in a sample of size n, the $BeBi(\alpha_1, \alpha_2, n^*)$ prior distribution is updated to a $BeBi(x_1 + \alpha_1, x_2 + \alpha_2, n')$ posterior distribution for X_1',

$$P_{\boldsymbol{\alpha}}(x_1'|\boldsymbol{x}) = \binom{x_1' + x_1 + \alpha_1 - 1}{x_1'} \binom{x_2' + x_2 + \alpha_2 - 1}{x_2'} \Big/ \binom{n' + n + s - 1}{n'} \tag{14}$$

for $x_1' = 0, 1, \ldots, n'$, where $x_1' + x_2' = n', x_1 + x_2 = n$ and $\alpha_1 + \alpha_2 = s$.

The set of posterior distributions for X_1' that is generated by the IDMM(s) prior consists of all the *BeBi* distributions $P_{\boldsymbol{\alpha}}(\cdot|\boldsymbol{x})$ such that $0 < \alpha_1 < s$, or, equivalently, all $P_{s\boldsymbol{\varphi}}(\cdot|\boldsymbol{x})$ such that $0 < \varphi_1 < 1$. Posterior upper and lower probabilities or expectations concerning X_1' can be calculated from (14) by maximising or minimising with respect to the single variable α_1.

For example, the mean of the posterior *BeBi* distribution (14) is $n'(x_1+\alpha_1)/(n+s)$. By maximising or minimising this with respect to α_1, we obtain the posterior upper and lower expectations for the number of successes

$$\overline{E}(X_1'|\boldsymbol{x}) = \left(\frac{x_1 + s}{n + s}\right) n', \quad \underline{E}(X_1'|\boldsymbol{x}) = \left(\frac{x_1}{n + s}\right) n'. \tag{15}$$

5.2 Posterior upper and lower cdfs for the number of successes

Let $\overline{F}_1(\cdot|\boldsymbol{x})$ and $\underline{F}_1(\cdot|\boldsymbol{x})$ denote the posterior upper and lower cdfs of X_1' under the IDMM, $\overline{F}_1(u|\boldsymbol{x}) = \overline{P}(X_1' \leq u|\boldsymbol{x})$ and $\underline{F}_1(u|\boldsymbol{x}) = \underline{P}(X_1' \leq u|\boldsymbol{x})$. Because the *BeBi* distribution $P_{\boldsymbol{\alpha}}(\cdot|\boldsymbol{x})$ in (14) is stochastically increasing in α_1, the upper and lower cdfs are achieved as $\alpha_1 \to 0$ and as $\alpha_1 \to s$ respectively. Hence we obtain the formulae, for $u = 0, 1, \ldots, n'$,

$$\overline{F}_1(u|\boldsymbol{x}) = \sum_{x_1'=0}^{u} \binom{x_1' + x_1 - 1}{x_1'} \binom{n^* - x_1' - x_1 + s - 1}{n' - x_1'} \Big/ \binom{n^* + s - 1}{n'} \tag{16}$$

and

$$\underline{F}_1(u|\boldsymbol{x}) = \sum_{x_1'=0}^{u} \binom{x_1' + x_1 + s - 1}{x_1'} \binom{n^* - x_1' - x_1 - 1}{n' - x_1'} \Big/ \binom{n^* + s - 1}{n'}. \tag{17}$$

The prior upper and lower cdfs of X_1', which are obtained by setting $n = x_1 = 0$, are vacuous, as seen in subsection 2.3(c).

The posterior upper and lower cdfs can also be expressed in the following equivalent forms. (If s is not an integer, some of the terms in these sums may be negative.)

$$\overline{F}_1(u|\boldsymbol{x}) = \sum_{j=0}^{u} \binom{x_1+u}{j} \binom{n^*-x_1-u+s-1}{n'-j} \Big/ \binom{n^*+s-1}{n'} \tag{18}$$

$$= \sum_{j=0}^{u} \binom{n'}{j} \binom{n+s-1}{x_1+u-j} \Big/ \binom{n^*+s-1}{x_1+u} \quad \text{and} \tag{19}$$

$$\underline{F}_1(u|\boldsymbol{x}) = \sum_{j=0}^{u} \binom{x_1+u+s}{j} \binom{n^*-x_1-u-1}{n'-j} \Big/ \binom{n^*+s-1}{n'} \tag{20}$$

$$= \sum_{j=0}^{u} \binom{n'}{j} \binom{n+s-1}{n-x_1-u+j-1} \Big/ \binom{n^*+s-1}{n^*-x_1-u-1}. \tag{21}$$

Equality between (18) and (19), and between (20) and (21), can be verified by expressing the binomial coefficients in terms of gamma functions and rearranging the functions.

Assuming that s is an integer, the posterior upper and lower cdfs in equations (16)–(21) can be interpreted as three different kinds of cumulative probabilities in sampling without replacement from a population of size n^*+s-1. Firstly, the cdfs in (16) and (17) can be regarded as cumulative negative hypergeometric probabilities, in inverse sampling without replacement from a population of size n^*+s-1 which contains n' successes: $\overline{F}_1(u|\boldsymbol{x})$ is the chance that the $x_1 th$ failure occurs on or before the $(x_1+u)th$ trial, and $\underline{F}_1(u|\boldsymbol{x})$ is the chance that the $(x_1+s)th$ failure occurs on or before the $(x_1+u+s)th$ trial.

Secondly, the expressions in (18) and (20) can be interpreted as cumulative hypergeometric probabilities, in direct sampling without replacement from a population of size n^*+s-1 which contains n' successes: $\overline{F}_1(u|\boldsymbol{x})$ is the chance that a random sample of size x_1+u will contain no more than u successes, and $\underline{F}_1(u|\boldsymbol{x})$ is the chance that a random sample of size x_1+u+s will contain no more than u successes. The difference between the upper and lower cdfs is equivalent to a difference of s between the sample sizes. For integer s, equality between (16) and (18) follows from the fact that the $x_1 th$ failure occurs on or before the $(x_1+u)th$ trial if and only if the first x_1+u observations contain no more than u successes; see Thatcher (1964, p. 183).

A third interpretation comes from (19) and (21): $\overline{F}_1(u|\boldsymbol{x})$ is the chance that, in sampling without replacement from a population of size n^*+s-1 which contains x_1+u successes, a sample of size n' will contain no more than u successes; and $\underline{F}_1(u|\boldsymbol{x})$ is the chance of the same event when the population contains x_1+u+s successes. This means that we can regard the future sample as if it was taken without replacement from a population of size n^*+s-1 which contains either x_1+u successes (to give the upper probability) or x_1+u+s successes (to give the lower probability). Here the difference between the upper and lower probabilities is equivalent to a difference of s between the numbers of successes in the population.

All three interpretations are especially simple in the case $s=1$, since then they involve sampling from a population of size $n^*=n+n'$, which is the total number of observations in the past and future samples.

It can be shown that the posterior upper cdf for the proportion of successes in a future sample converges to a beta$(x_1, n - x_1 + s)$ cdf as the future sample size $n' \to \infty$. Similarly the posterior lower cdf for the proportion of successes converges to a beta$(x_1 + s, n - x_1)$ cdf. If successes and failures are regarded as the outcomes of an underlying binomial process, as outlined in Section 3.4, then the two limiting beta cdfs can be interpreted as the posterior upper and lower cdfs of θ_1, the unknown chance that a single observation will be a success.

5.3 Numerical examples

To illustrate the behaviour of the posterior upper and lower cdfs, we give some numerical examples. In each case we assume that the future sample size is $n' = 10$ and that 20% of the n observations in the observed sample are successes. Figures 1, 2 and 3 display the posterior upper and lower cdfs for the number of successes, based on the IDMM with $s = 1$ or $s = 2$, for the observed sample sizes $n = 5$, 25 and 100.

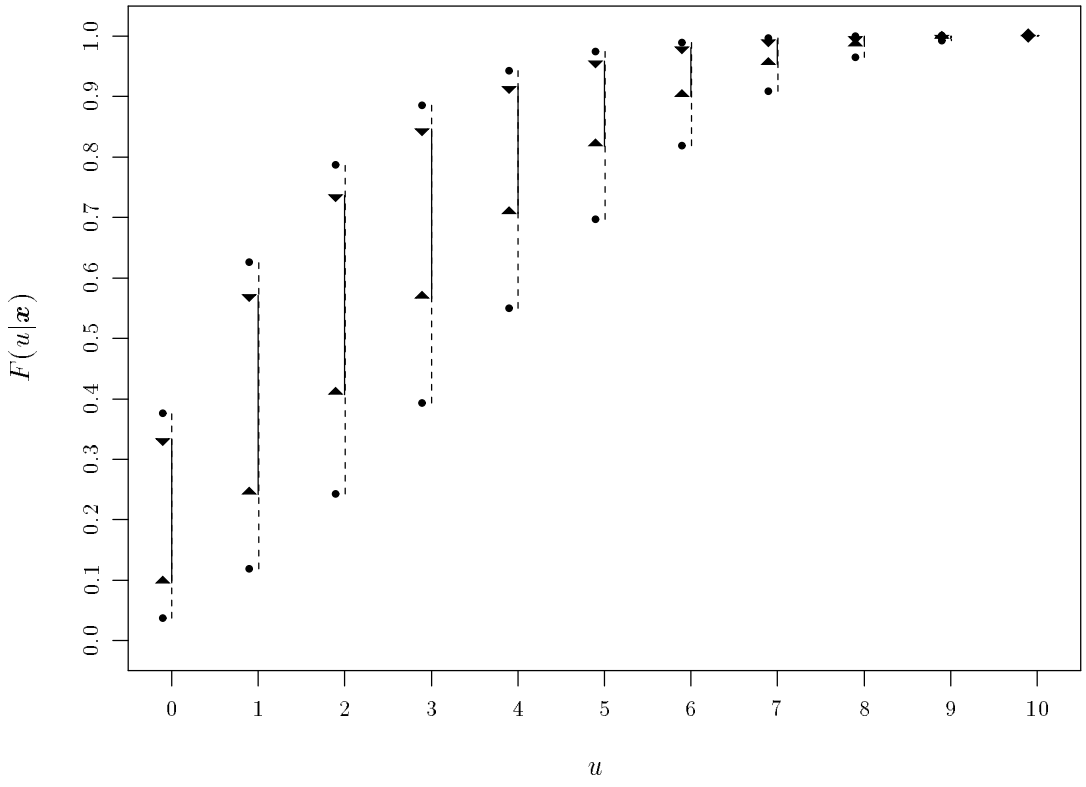

Figure 1: *Posterior upper and lower cdfs for the number of successes in $n' = 10$ future observations, based on the IDMM with $s = 1$ (inner solid lines) and $s = 2$ (outer dashed lines), after obtaining $x_1 = 1$ success in $n = 5$ previous observations.*

First consider the inferences obtained from $s = 1$, represented by the inner solid lines. By comparing the three figures, it is clear that the upper and lower cdfs tend to converge as the observed sample size n increases. For the smallest sample $n = 5$, shown

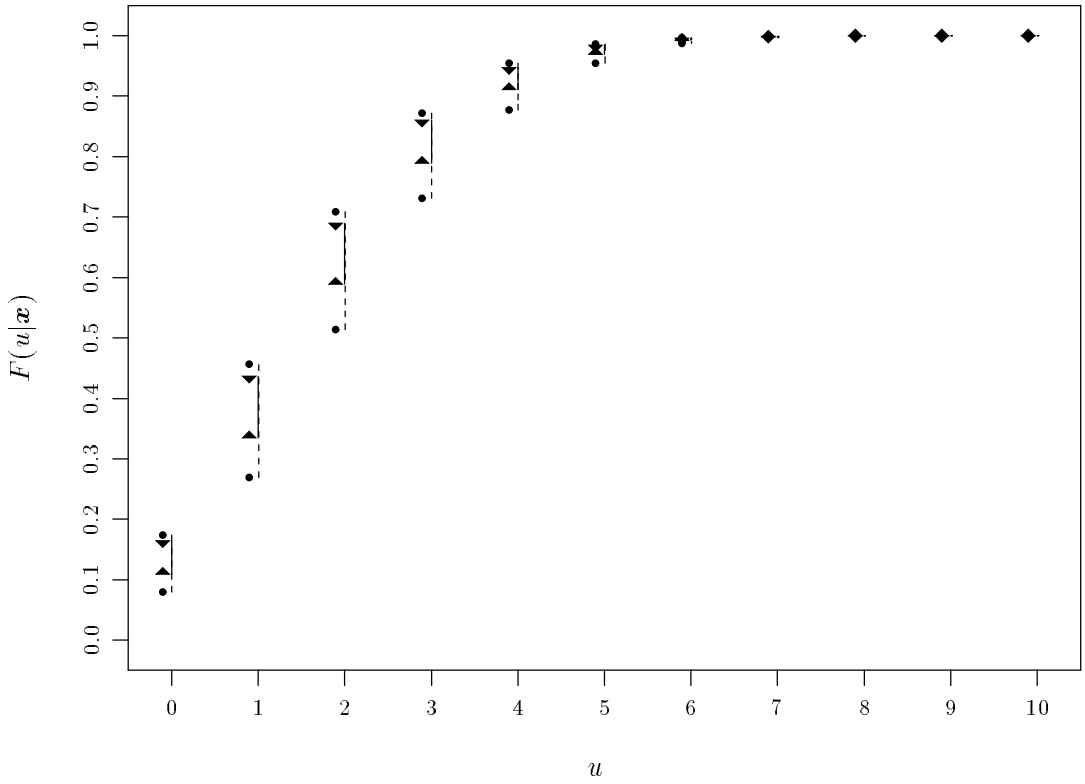

Figure 2: *Posterior upper and lower cdfs for the number of successes in $n' = 10$ future observations, based on the IDMM with $s = 1$ (inner solid lines) and $s = 2$ (outer dashed lines), after obtaining $x_1 = 5$ successes in $n = 25$ previous observations.*

in Figure 1, the upper and lower cdfs are far apart and it is therefore difficult to make useful predictions about the future number of successes. For the largest sample $n = 100$ in Figure 3, the upper and lower cdfs are close together and, for example, we can be very confident (lower probability 0.954) that there will be no more than 4 successes in the next 10 observations. For comparison, the lower probability of this event is 0.706 when $n = 5$ (Figure 1) and 0.911 when $n = 25$ (Figure 2).

The effect of increasing the value of s is to increase the difference between the upper and lower cdfs, especially for the smallest sample size $n = 5$. For example, the (lower, upper) probabilities of no more than 2 successes in 10 future observations are $(0.407, 0.736)$ for $s = 1$, and $(0.242, 0.786)$ for $s = 2$ (see Figure 1). For the largest sample size, $n = 100$, there is little difference between the inferences for $s = 1$ and $s = 2$; the (lower, upper) probabilities of the same event are $(0.652, 0.681)$ for $s = 1$, and $(0.629, 0.687)$ for $s = 2$ (see Figure 3).

5.4 Prediction sets

Predictions about X_1', the number of successes in n' future observations, can be summarised by giving one-sided or two-sided prediction sets which have a guaranteed posterior probability of containing the actual number of successes x_1'. The one-sided

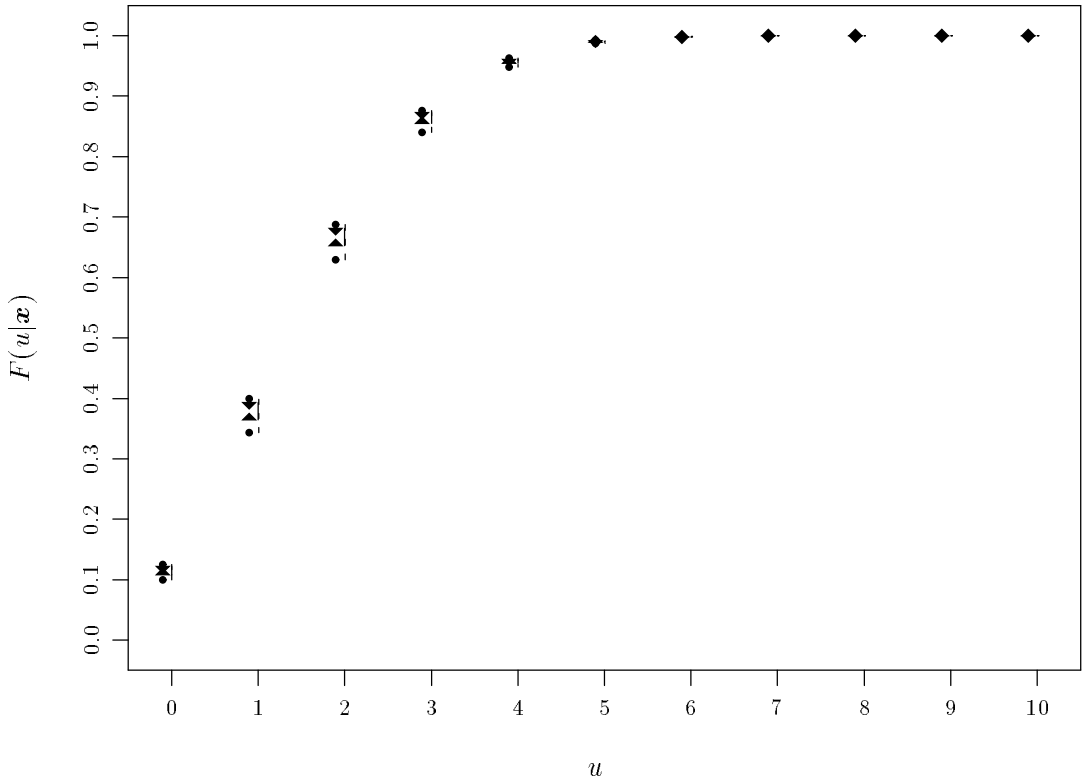

Figure 3: *Posterior upper and lower cdfs for the number of successes in $n' = 10$ future observations, based on the IDMM with $s = 1$ (inner solid lines) and $s = 2$ (outer dashed lines), after obtaining $x_1 = 20$ successes in $n = 100$ previous observations.*

prediction set for X_1' at level γ is the smallest set $\{0, 1, \ldots, y^\circ\}$ whose posterior lower probability of containing X_1' is at least γ. Thus y° is the smallest integer such that $\underline{F}_1(y^\circ|\boldsymbol{x}) \geq \gamma$, and it can be determined from the formulae for $\underline{F}_1(u|\boldsymbol{x})$ in equations (17), (20) and (21).

Similarly $y_\circ$, a lower bound for X_1' at level γ, can be found as the largest integer such that $\underline{P}(X_1' \geq y_\circ|\boldsymbol{x}) \geq \gamma$, i.e., such that $\overline{F}_1(y_\circ - 1|\boldsymbol{x}) \leq 1 - \gamma$. The set $\{y_\circ, \ldots, y^\circ\}$ can be regarded as a conservative two-sided prediction set for X_1' at level $2\gamma - 1$, since $\underline{P}(y_\circ \leq X_1' \leq y^\circ|\boldsymbol{x}) \geq 1 - \overline{P}(X_1' < y_\circ|\boldsymbol{x}) - \overline{P}(X_1' > y^\circ|\boldsymbol{x}) = \underline{P}(X_1' \geq y_\circ|\boldsymbol{x}) + \underline{P}(X_1' \leq y^\circ|\boldsymbol{x}) - 1 \geq 2\gamma - 1$.

Numerical examples of one-sided prediction sets can be read off Figures 1–3. From Figure 1 and $s = 1$, for example, we see that $\{0, 1, \ldots, 6\}$ is a 90% prediction set for the number of successes in 10 future observations based on 1 success in 5 previous observations, since $\underline{F}_1(6|\boldsymbol{x}) = .900$. Similarly $\{0, 1, \ldots, 7\}$ is a 95% prediction set in this case. These prediction sets are large; on the basis of so few observations, we cannot make useful predictions with much confidence. (The prediction sets are even larger if we use the IDMM with $s = 2$.) For the larger sample size $n = 25$ in Figure 2, the smaller sets $\{0, 1, \ldots, 4\}$ and $\{0, 1, \ldots, 5\}$ are respectively 91% and 97% prediction sets. For the largest sample size $n = 100$ in Figure 3, $\{0, 1, \ldots, 4\}$ is a 95% prediction set, and

we can make a more useful prediction about the future number of successes than we could from a small sample.

5.5 Relationship to frequentist inferences

First consider hypothesis testing. In the case of random sampling without replacement from a population of size $n^* = n + n'$, it is natural to regard $X_1^* = X_1 + X_1'$, the total number of successes in the population, as an unknown parameter. A frequentist could test a one-sided hypothesis of the form $H_0 : X_1^* \leq v$ by computing a *p-value* or observed significance level. Because larger values of X_1 indicate stronger evidence against H_0, the p-value is equal to $p = \max \{P(X_1 \geq x_1 | X_1^* = x_1^*) : x_1^* \leq v\} = P(X_1 \geq x_1 | X_1^* = v)$.

Assuming that the observed sample was obtained by hypergeometric sampling, the p-value is the chance of obtaining at least x_1 successes in a sample of size n taken without replacement from a population of size $n + n'$ which contains v successes. Since the observed sample contains at least x_1 successes if and only if the unobserved population contains no more than $v - x_1$ successes, the p-value agrees with the posterior upper cdf at $u = v - x_1$ under the IDMM(1), using the third interpretation in Section 5.2. It follows that $p = \overline{F}_1(u|\boldsymbol{x}) = \overline{P}(X_1' \leq u|\boldsymbol{x}) = \overline{P}(X_1^* \leq u + x_1 = v|\boldsymbol{x}) = \overline{P}(H_0|\boldsymbol{x})$. Thus the p-value for testing H_0 agrees with the posterior upper probability of H_0 that is produced by the IDMM with $s = 1$. Symmetrically, the p-value for testing $H_0 : X_1^* \geq v$ is $P(X_1 \leq x_1 | X_1^* = v) = \overline{P}(X_1' \geq v - x_1|\boldsymbol{x}) = \overline{P}(X_1^* \geq v|\boldsymbol{x}) = \overline{P}(H_0|\boldsymbol{x})$. Other links between the IDMM(1) and frequentist tests can be found in Bernard (1998a).

Next consider prediction sets. A frequentist prediction set or *tolerance region* for X_1' at level γ has the property that, when both X_1' and X_1 are hypergeometric random variables, the chance that the random tolerance region will include the random value of X_1' is at least γ. The standard frequentist tolerance region for X_1' is constructed from Fisher's exact test for a 2×2 contingency table, by classifying the $n + n'$ observations according to whether they are successes or failures and whether they belong to the past or future sample (Thatcher, 1964). Fisher's exact test is equivalent to the test of $H_0 : X_1^* \leq v$ discussed earlier in this subsection, in the sense that the one-sided p-value for Fisher's test, based on data (x_1, x_1'), agrees with the p-value for testing $H_0 : X_1^* \leq v$, where $v = x_1 + x_1'$, based on data x_1. From the previous result, the latter p-value agrees with $\overline{F}_1(x_1'|\boldsymbol{x})$ when this is calculated from the IDMM with $s = 1$. The one-sided tolerance region at level γ is the set of all x_1' for which the p-value $\overline{F}_1(x_1'|\boldsymbol{x})$ exceeds $1-\gamma$. Hence the tolerance region is $\{y_\circ, \ldots, n'\}$, where $y_\circ$ is the smallest integer such that $\overline{F}_1(y_\circ|\boldsymbol{x}) > 1 - \gamma$, i.e., the largest integer such that $\overline{F}_1(y_\circ - 1|\boldsymbol{x}) \leq 1 - \gamma$.

Thus the one-sided tolerance region at level γ agrees with the prediction set based on the IDMM with $s = 1$, which was defined in subsection 5.4. In other words, the IDMM with $s = 1$ produces one-sided prediction sets which have frequentist coverage probability at least γ and which are therefore valid from a frequentist point of view. The other one-sided tolerance region $\{0, \ldots, y^\circ\}$ is found in a symmetrical way from the lower cdf of the IDMM(1). It follows that the conservative two-sided prediction sets from the IDMM(1) also have frequentist coverage probability at least γ. This holds also for values $s \geq 1$ which produce wider prediction sets. All the prediction sets given in the numerical examples of subsection 5.4 are therefore valid (conservative)

frequentist prediction sets. Thatcher (1964) gives some other useful relationships between frequentist and Bayesian prediction sets.

5.6 Comparison with the LLDB model

Under the LLDB model with k categories, the number of successes X_1^* has a $BeBi(1, k-1, n^*)$ prior distribution. Inferences again depend on the number of categories, k. Consider the case $k = 2$, where observations are classified only as "success" or "failure". In this case X_1^* has a uniform prior distribution on $\{0, 1, \ldots, n^*\}$, as suggested in Bayes (1763).

After observing frequencies $\boldsymbol{x}$, the LLDB prior is updated to a $BeBi(x_1 + 1, n - x_1 + 1, n')$ posterior distribution for X_1', given by (14). The posterior expectation of the future number of successes is $E(X_1'|\boldsymbol{x}) = n'(x_1 + 1)/(n + 2)$, and this value lies between the posterior upper and lower expectations given in (15) provided that $s \geq 1$. If $s < 1$ then the IDMM posterior lower expectation is larger than the LLDB posterior expectation when x_1 and n are sufficiently large.

Still assuming that $k = 2$ and $s \geq 1$, it can be shown that the posterior cdf for the number of successes under the LLDB model (and also the models of Haldane, Perks and Jeffreys) always lies between the posterior upper and lower cdfs from the IDMM. It follows that the one-sided prediction sets produced by all these objective Bayesian models are contained in the prediction set from the IDMM(1).

6 Confirmation of a universal hypothesis

The case $x_1 = n$, where all the previous n observations have been of the same type, is especially interesting because it relates to the problem of confirming a universal hypothesis: if all n observed units are of the same type, how confident can we be about the hypothesis that all $n + n'$ units in the population are of the same type? An early solution was proposed by Laplace (1812/1825, pp. 45–46), who applied his rule of succession to calculate the probability that the sun will rise tomorrow, given that it has risen every day since the beginning of human history (which he took to be a period of 5000 years, i.e., 1,826,213 days). Other discussions are in Laplace (1820), Perks (1947) and Jeffreys (1961, p. 129).

6.1 Inferences from the IDMM

Suppose the observed sample of size n consists entirely of successes, so that $x_1 = n$. Let H_0 denote the hypothesis that $X_1^* = n^*$, i.e., that all $n^* = n + n'$ members of the population are successes. The posterior probability of H_0 that is produced by a $BeBi(\alpha_1, s - \alpha_1, n^*)$ prior distribution can be found by substituting $x_1 = n$ and $x' = n'$ in (14), giving

$$P_{\alpha_1}(H_0|x_1 = n) = \binom{n + n' + \alpha_1 - 1}{n'} \Big/ \binom{n + n' + s - 1}{n'} . \tag{22}$$

Since (22) is increasing in α_1, the posterior upper and lower probabilities produced by the IDMM(s) are achieved as $\alpha_1 \to s$ and $\alpha_1 \to 0$ in (22). Hence the IDMM(s) gives

$$\begin{aligned} \overline{P}(H_0|x_1 = n) &= 1, \qquad \text{and} \\ \underline{P}(H_0|x_1 = n) &= \binom{n+n'-1}{n'} \Big/ \binom{n+n'+s-1}{n'} = \prod_{i=0}^{n'-1} \left(\frac{n+i}{n+s+i}\right). \end{aligned} \tag{23}$$

When s is a positive integer, the lower probability can be expressed as a product of s terms,

$$\underline{P}(H_0|x_1 = n) = \prod_{i=0}^{s-1} \left(\frac{n+i}{n+n'+i}\right). \tag{24}$$

This formula is especially simple when $s = 1$: then $\underline{P}(H_0|x_1 = n) = n/(n+n') = n/n^*$, which is the ratio of the observed sample size to the total population size, i.e., the *sampling ratio*.

It can be seen from (23) that $\underline{P}(H_0|x_1 = n) \to 1$ as $sn'/n \to 0$. Hence, for any fixed nonzero values of n and n', $\underline{P}(H_0|x_1 = n) \to 1$ as $s \to 0$. For example, based on just one observation which falls into category i, the predictive lower probability that the next 100 observations will all be of the same type i converges to one as $s \to 0$. But if n is small and n' is large then it is unreasonable to have much confidence in H_0, and this again shows that very small values of s and Haldane's model ($s = 0$) can produce unacceptable inferences.

For fixed values of s and n', we become increasingly confident that H_0 is true as the observed sample size n increases. If $n' = 5$ and $n = 100$, for example, then the posterior lower probability of H_0 is 0.952 when $s = 1$ and 0.907 when $s = 2$.

On the other hand, if s and n are fixed then $\underline{P}(H_0|x_1 = n) \to 0$ as $n' \to \infty$. Since $\overline{P}(H_0|x_1 = n) = 1$ irrespective of n', this means that inferences concerning H_0 become increasingly uninformative as the unobserved population size increases, and tend to vacuous probabilities in the limit. Even if the observed sample size n is large, we cannot be confident that H_0 is true when n' is much larger than n.

To illustrate this, consider the numerical example of Bernardo (1984). It is estimated that there are about 100,000 iguanas on Fernandina Island. All of the 90 iguanas captured on the island have the same distinctive pattern on their skin. What inferences can we draw about the hypothesis, H_0, that all the iguanas on the island have this skin pattern? Applying the IDMM in this problem, the posterior upper probability of H_0 is 1, and the lower probability is given by (23) with $n = 90$ and $n + n' = 100,000$. The lower probability is 0.0009 for $s = 1$ and 8.1×10^{-8} for $s = 2$. According to the IDMM, because the unobserved population size is so much larger than the observed sample size, we should have very little confidence that H_0 is true, but nor can we have any confidence that it is false.

6.2 Comparison with objective Bayesian inferences

These conclusions are qualitatively different from the Bayesian conclusions of LLDB and Bernardo (1984), which also disagree strongly with each other. From (22), the

posterior probability of H_0 under the LLDB model is

$$P(H_0|x_1 = n) = \binom{n+n'}{n'} \Big/ \binom{n+n'+k-1}{n'} = \prod_{i=1}^{k-1} \left(\frac{n+i}{n+n'+i}\right). \tag{25}$$

Again this probability depends on the number of categories, k, but it cannot be larger than $(n+1)/(n+n'+1)$, which is the value for $k = 2$ that was derived by Laplace (1820, Article 32).

A different Bayesian model was suggested by Bernardo (1984). Bernardo's model is based on a so-called "reference prior" distribution which varies with the event whose probability is to be calculated. For that reason, several inferences based on the same data may be incoherent. For calculating the posterior probability of H_0, Bernardo's model assigns prior probability $\frac{1}{2}$ to H_0 and distributes the remaining probability uniformly over the other n^* possible values of X_1^*. This produces the posterior probability

$$P(H_0|x_1 = n) = \left[1 + \frac{1}{n+1}\left(\frac{n'}{n+n'}\right)\right]^{-1} \geq \left[1 + \frac{1}{n+1}\right]^{-1} = \frac{n+1}{n+2}. \tag{26}$$

(See Bernardo and Smith, 1994, pp. 322–323.)

Bernardo's model appears to produce absurd inferences for small values of n. For example, if an urn contains 1000 balls and the first ball drawn from it is red, then (26) tells us that there is posterior probability greater than $\frac{2}{3}$ that all 1000 balls are red! Here the LLDB model gives a very different posterior probability, smaller than 0.002.

The two objective Bayesian analyses of LLDB and Bernardo can produce highly discrepant answers even when the observed sample size n is large. If also the unobserved population size n' is much larger than n, the posterior probability of H_0 is close to zero under the LLDB model but close to one under Bernardo's model. In the iguanas example, the posterior probability of H_0 is 0.0009 for LLDB ($k = 2$) but 0.989 for Bernardo and 1 for Haldane. If we accept Bernardo's or Haldane's model then we can be very confident that H_0 is true, whereas if we accept the LLDB model then we can be extremely confident that H_0 is false! The large discrepancy between the two conclusions suggests that both conclusions are over-confident.

The conclusion from the IDMM, which seems somewhat more reasonable, is that we learn very little from the observations about the truth of H_0. After observing that all 90 iguanas in the sample have the same skin pattern, it is certainly plausible that all 100,000 iguanas on the island have the same pattern, because a sample in which no exceptions have been observed gives absolutely no evidence to suggest that H_0 is false. The IDMM therefore assigns upper probability one to H_0. But it is also plausible that H_0 is false and a few of the iguanas in the population are exceptions, because if this is the case then it is very likely that none of the exceptional iguanas will appear in a sample of 90. (If there are q exceptional iguanas in the population then the probability that any of them will appear in a sample of 90 is approximately $9q \times 10^{-4}$.) The IDMM therefore assigns a large upper probability to the complement of H_0 and (equivalently) a small lower probability to H_0.

If $s \geq 1$ then the upper and lower posterior probabilities for H_0 produced by the IDMM always encompass the posterior probabilities of both LLDB (assuming $k = 2$) and Bernardo. Except when n and n' are both small, the lower probability resulting from the IDMM(1) is close to the LLDB posterior probability for $k = 2$.

The differences in posterior probabilities between the three models reflect major differences in *prior* probabilities. The IDMM gives vacuous prior probabilities $\overline{P}(H_0) = 1$ and $\underline{P}(H_0) = 0$, which model prior ignorance about the truth of H_0. Bernardo's model is designed to give $P(H_0) = \frac{1}{2}$. The LLDB prior probability (when $k = 2$) is $P(H_0) = (n^* + 1)^{-1}$. As noted by Jeffreys (1961, p. 129), this prior probability is very small when the population size n^* is very large, despite the absence of prior information, and it is therefore not surprising that the LLDB posterior probability is also very small.

Authors generally agree that the differences between inferences from different objective Bayesian priors are minor whenever the observed sample size n is large. We see here a clear counter-example to this rule. A related feature of this example is that, under the IDMM, even a very large sample of observations may be quite uninformative about some properties of the population, such as whether or not the hypothesis H_0 is true.

Of course we do learn something useful from the data $x_1 = n$ using the IDMM, even if n^* is much larger than n. It can be shown that, for any positive value of ε, if n is sufficiently large then we can be very confident, after observing $x_1 = n$, that H_0 is "almost true" in the sense that the population relative frequency f_1^* exceeds $1 - \varepsilon$, no matter how large n^* is. That is true because, if n^* is very large, f_1^* can be regarded as an unknown binomial chance of success θ_1, and $\underline{P}(\theta_1 > 1 - \varepsilon | x_1 = n) \to 1$ as $n \to \infty$.

6.3 Comparison with frequentist inferences

In this problem, inferences from the IDMM with $s = 1$ again agree with frequentist inferences. Assume that the sample is obtained by direct sampling (n fixed). If we first consider a frequentist test of $H_0 : X_1^* = n^*$, the observed frequency of success, $x_1 = n$, is the only possible one under H_0, so the p-value for the test is $p_1 = 1$, whatever the values of n and n', indicating that there is absolutely no evidence against H_0.

Next consider a frequentist test of the hypothesis $H_1 : X_1^* \leq n^* - 1$, which states that the universal hypothesis H_0 is false. The p-value here is the chance of obtaining no failures in a sample of n from a population of size $n^* = n + n'$ which contains just one failure, which is clearly $p_2 = n'/(n + n')$. As noted in subsection 5.5, the p-value for this test, p_2, is equal to the posterior upper probability of H_1 under the IDMM with $s = 1$. This second test will reject H_1, and conclude that there is strong evidence in favour of H_0, when p_2 is sufficently small, i.e., when $1 - p_2 = n/(n + n')$ exceeds some fixed level γ. The same p-values p_1 and p_2 would be obtained by assuming inverse sampling with fixed x_1.

To summarise the conclusions from both tests, H_0 is compatible with the observed data at any fixed level, but there is strong evidence in favour of H_0 only if $n/(n + n')$ exceeds a fixed level γ. These frequentist conclusions agree perfectly with those obtained from the IDMM with $s = 1$, which gave $\overline{P}(H_0 | x_1 = n) = 1$ and $\underline{P}(H_0 | x_1 = n) = n/(n + n')$.

7 An application to economic survey data

The French "Institut National de la Statistique et des Etudes Economiques" (INSEE) specialises in collecting, analysing and communicating statistics concerning the French economy. One of the studies it runs is a monthly poll of business managers from private companies with more than 20 employees. The polling is done by means of stratified sampling where the strata correspond to various industrial sectors. The sampling ratio n/n^* varies between sectors. One of the questions, called the GPP question, asks the managers to categorise their current beliefs about the outlook for French industrial production, by choosing the most appropriate category from the set $\Omega = \{Rise, Stable, Decline\}$. The answer "rise" means that the manager forecasts that the level of production in French industry as a whole will rise in the immediate future.

The following data concern the chemical manufacturing sector, which comprises approximately 450 companies. We are interested in the data for two months, January and February 1998. A random sample of 148 companies from the 450 was sent the questionnaire, but only 66 of them replied to the GPP question in both months. We will regard the $n = 66$ companies as a random sample from the population of $n^* = 450$ companies in the sector, although there may have been some selection bias in the response. The observed joint frequencies $\boldsymbol{x}$ for the two months involve $k = 9$ basic categories and are shown in Table 1.

Table 1: *Observed joint frequencies of the answers to the GPP question (forecasts for change in the level of French industrial production), in January 1998 and February 1998, from* 66 *managers in the chemical manufacturing sector (INSEE, 1998).*

		February		
		Rise	Stable	Decline
January	Rise	17	4	0
	Stable	10	25	2
	Decline	2	2	4

7.1 Analysis of the January opinion balance

The first problem is to make inferences about the unknown frequencies in the chemical manufacturing sector for a single month, which we take to be January. Because the IDMM satisfies the representation invariance principle, inferences of this kind will be unchanged if we consider only the 3 pooled categories relating to January (the rows of Table 1), with observed frequencies $\boldsymbol{x} = (21, 37, 8)$. More specifically, what is required is to make inferences about the January *opinion balance* which is defined as

$$\delta_{Jan} = f_1^* - f_3^*, \tag{27}$$

where f_1^*, f_2^*, f_3^* are the population relative frequencies of the three pooled categories. The possible values of δ_{Jan} vary by steps of $1/n^*$ and satisfy $-1 \leq \delta_{Jan} \leq 1$.

Figure 4 shows the posterior upper and lower cdfs for δ_{Jan}, $\overline{F}(u) = \overline{P}(\delta_{Jan} \leq u|\boldsymbol{x})$ and $\underline{F}(u) = \underline{P}(\delta_{Jan} \leq u|\boldsymbol{x})$, that are obtained from the IDMM with $s = 1$. The upper and lower cdfs are actually step functions, but the step size $1/450$ is small enough for them to be closely approximated by the smooth functions shown in Figure 4.

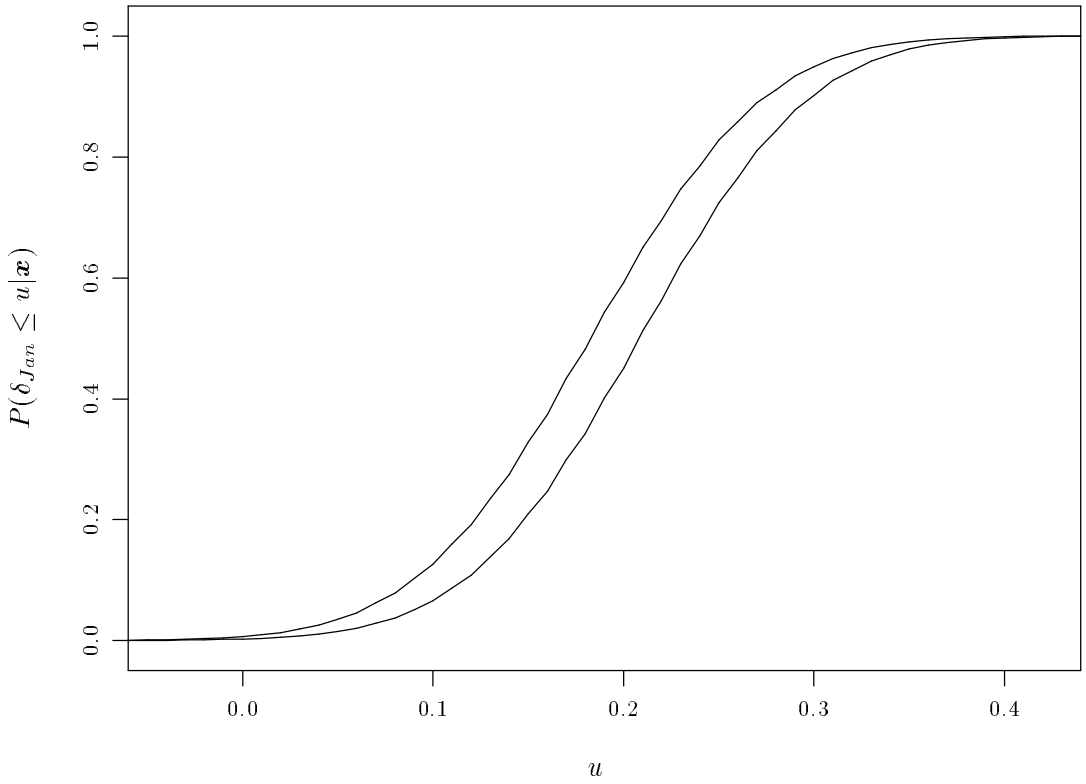

Figure 4: *Posterior upper and lower cdfs for the opinion balance* $\delta_{Jan} = f_1^* - f_3^*$*, using the IDMM with* $s = 1$ *and* $n^* = 450$*, based on observed frequencies* $\boldsymbol{x} = (21, 37, 8)$.

If δ is any linear combination of the population frequencies, it can be shown that the upper cdf $\overline{P}(\delta \leq u|\boldsymbol{x})$ is obtained from a limit of *DiMn* posterior distributions as the hyperparameter $\alpha_i \to s$, where i is the category whose frequency has the *smallest* coefficient in δ, and $\alpha_j \to 0$ if $j \neq i$. Similarly the lower cdf $\underline{P}(\delta \leq u|\boldsymbol{x})$ is obtained from a limit of *DiMn* posteriors as $\alpha_i \to s$, where i is the category whose frequency has the *largest* coefficient in δ. For the particular linear combination δ_{Jan} and $s = 1$, the upper and lower cdfs are therefore obtained from the *DiMn* prior distributions with $\boldsymbol{\alpha} \to (0, 0, 1)$ and $\boldsymbol{\alpha} \to (1, 0, 0)$ respectively.

From Figure 4 we can see how the posterior degree of imprecision concerning the event $\delta_{Jan} \leq u$ varies with u. When $u \leq 0$, the upper and lower probabilities for this event are both close to 0, so that we can be very confident that δ_{Jan} is greater than zero. Similarly we can be almost certain that δ_{Jan} is less than 0.4. The posterior degree of imprecision is greatest in the middle of this range; e.g., the upper and lower probabilities are 0.59 and 0.45 when $u = 0.20$.

The first question of interest is to make inferences about the sign of δ_{Jan}. A simple estimate of δ_{Jan} is $d_{Jan} = 21/66 - 8/66 = 0.20$, so that the opinion balance in the sample is positive. Can this be extended to the population opinion balance? That is, how confident can we be about the hypothesis $H_0 : \delta_{Jan} > 0$? We answer that question by finding the posterior upper and lower probabilities of H_0, which can be read off Figure 4. From the graph we see that $\underline{P}(\delta_{Jan} \leq 0|\boldsymbol{x}) = 0.002$ and $\overline{P}(\delta_{Jan} \leq 0|\boldsymbol{x}) = 0.007$, which give $\overline{P}(H_0|\boldsymbol{x}) = 0.998$ and $\underline{P}(H_0|\boldsymbol{x}) = 0.993$. Since the lower probability $\underline{P}(H_0|\boldsymbol{x})$ is very close to one, we can be almost certain that δ_{Jan} is greater than 0. In practical terms, the data give extremely strong evidence that, in January 1998, more managers in the chemical manufacturing sector expected French industrial production to rise in the immediate future than expected it to decline. (We are assuming here that the observed sample can be regarded as a random sample from the population.)

A second question of interest is to assess how much greater than zero δ_{Jan} is likely to be. One-sided prediction sets for δ_{Jan} can be read off Figure 4. For a given level γ, say $\gamma = 0.95$, we first look for the lower 95% prediction limit $\underline{\delta}$, which is the largest u satisfying $\underline{P}(\delta_{Jan} \geq u|\boldsymbol{x}) \geq 0.95$, or equivalently $\overline{P}(\delta_{Jan} < u|\boldsymbol{x}) \leq 0.05$, which from the upper cdf is found to be $\underline{\delta} = 0.06$. Similarly the upper 95% prediction limit $\overline{\delta}$ is the smallest u such that $\underline{P}(\delta_{Jan} \leq u|\boldsymbol{x}) \geq 0.95$, which from the lower cdf is found to be $\overline{\delta} = 0.32$. Since $2\gamma - 1 = 0.90$, the interval $[\underline{\delta}, \overline{\delta}] = [0.06, 0.32]$ is a conservative 90% two-sided prediction set. It is conservative in the sense that its lower probability $\underline{P}(\underline{\delta} \leq \delta_{Jan} \leq \overline{\delta}|\boldsymbol{x})$ is at least 0.90. So we can be confident, with lower probability at least 0.90, that the opinion balance for January in the 450 companies lies between 0.06 and 0.32.

7.2 Analysis of the difference of two opinion balances

Another problem is to measure the evolution of δ_m from one month to the next, in our case from January to February 1998. Now the population parameter of interest is the *opinion balance evolution index*, $\delta = \delta_{Feb} - \delta_{Jan}$. This can be expressed in terms of the $k = 9$ relative frequencies f^*_{ij} (indexed by row and column) as

$$\delta = 2f^*_{31} + f^*_{21} + f^*_{32} - f^*_{12} - f^*_{23} - 2f^*_{13}. \tag{28}$$

Here δ varies between -2 and 2 by steps of $1/n^*$. Because cells 21 and 32 have the same coefficient (1) in δ, cells 12 and 23 have the same coefficient (-1), and the IDMM satisfies the representation invariance principle, inferences about δ from the IDMM can be solely based on 5 pooled categories which correspond to the coefficients $(2, 1, 0, -1, -2)$ in δ and have observed frequencies $\boldsymbol{x} = (2, 12, 46, 6, 0)$.

Figure 5 shows the posterior upper and lower cdfs for δ under the IDMM with $s = 1$. Using the result stated in subsection 7.1, the upper and lower cdfs are obtained from *DiMn* prior distributions with $\boldsymbol{\alpha} \to (0, 0, 0, 0, 1)$ and $\boldsymbol{\alpha} \to (1, 0, 0, 0, 0)$ respectively. From Figure 5, we see that we can be almost certain that δ lies between -0.1 and 0.4. The upper and lower cdfs for δ in Figure 5 are further apart than those for δ_{Jan} in Figure 4, reflecting the weaker evidence concerning δ. In Figure 5, the maximum degree of imprecision is obtained around $u = 0.15$, where the lower and upper probabilities are equal to 0.37 and 0.64.

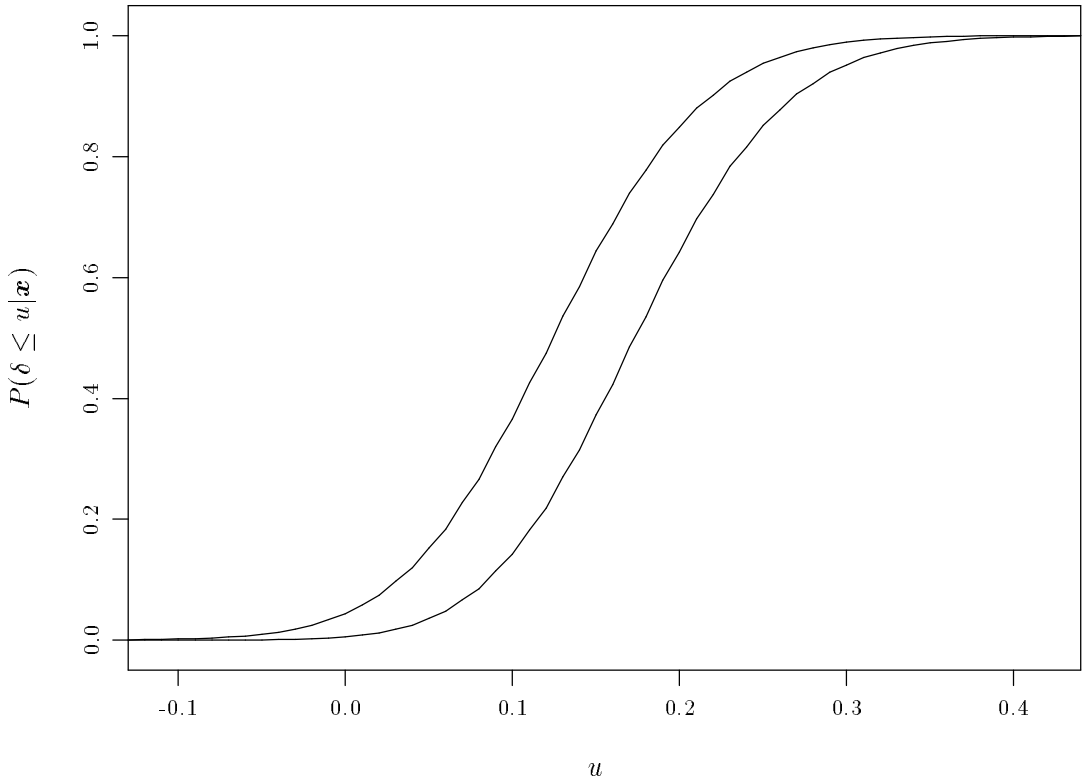

Figure 5: *Posterior upper and lower cdfs for the opinion balance evolution index* $\delta = 2f_1^* + f_2^* - f_4^* - 2f_5^*$*, using the IDMM with* $s = 1$ *and* $n^* = 450$*, based on observed frequencies* $\boldsymbol{x} = (2, 12, 46, 6, 0)$*.*

The simple estimate of δ is $d = 10/66 = 0.15$, so that the estimated difference in opinion balance is positive. Following the same steps as in subsection 7.1, the posterior lower and upper probabilities of the hypothesis that $\delta > 0$ are found to be 0.957 and 0.995. The posterior lower probability is sufficiently close to one for us to be confident that $\delta > 0$. There is strong evidence in the data that the opinion balance increased between January and February 1998, i.e., that the perceived outlook for industrial production improved amongst managers in the chemical manufacturing sector. The conservative 90% prediction set for δ, as defined in subsection 7.1, is found to be $[0.01, 0.30]$.

7.3 Comparison with other methods

The observed sample size in this example, $n = 66$, is large enough to dominate most of the difference between prior distributions, as seen from the fact that the upper and lower cdfs are quite close together in Figures 4 and 5. For that reason, there are no major differences between the conclusions from the preceding analysis and those obtained from the LLDB model or the other objective Bayesian models. For example, the posterior probability of the hypothesis that $\delta_{Jan} > 0$ under the LLDB model is 0.996 using $k = 3$ and 0.990 using $k = 9$, and the corresponding conservative 90%

prediction sets for δ_{Jan} are $[0.07, 0.31]$ and $[0.07, 0.29]$. These are close to the inferences from the IDMM(1).

Hoadley (1969) described a Bayesian approach for the type of problem considered here, of making inferences about a linear combination of population frequencies. He used a single *DiMn* prior and obtained inferences from a single *DiMn* posterior distribution.

Frequentist asymptotic confidence intervals for the population parameters δ_{Jan} and δ are studied in Caron, Ravalet & Sautory (1996), but their approach encounters difficulties when some cells have frequency zero, especially if n is small. An analysis based on the IDMM avoids such difficulties, as shown in the analysis of δ which included a cell with frequency zero. Furthermore, all the inferences we have presented for the INSEE data are exact. More generally, there is no need in using the IDMM to resort to asymptotic arguments. However, when the number of categories k is large, exact computation of the necessary *DiMn* distributions can become time consuming. In such cases it is possible to approximate each *DiMn* by a random sample from it, whose size can be chosen to be large enough to guarantee any desired level of accuracy.

8 Conclusions

We have proposed the IDMM as a method for making predictive inferences about categorical data. The method can be used to make inferences about a finite population from a sample without replacement (multiple hypergeometric data), or to make predictions about a future sample from an infinite population (multinomial data). We recommend the IDMM only for cases where there is little or no prior information about the population frequencies $\boldsymbol{x}^*$, and not for problems in which there are expected to be particular kinds of correlations or patterns amongst the frequencies.

The IDMM has several properties which are desirable, and arguably even necessary, when there is little prior information that can be used to predict frequencies. It satisfies the following general desiderata: (a) coherence; (b) the likelihood principle; (c) symmetry in the categories; (d) the embedding principle; (e) the representation invariance principle; (f) vacuous prior probabilities concerning a future observation; (g) consistency between predictive inference and parametric inference (i.e., between finite and infinite populations); (h) frequentist validity of inferences, provided that $s \geq 1$; and (i) if $s \geq 1$, it encompasses several objective Bayesian inferences for binary data ($k = 2$). None of the objective Bayesian methods satisfies all these desiderata; the LLDB method, for example, violates (d), (e), (f) and (h). In particular, because the IDMM satisfies the representation invariance principle, inferences are invariant under modifications to the possibility space, and the IDMM can be used even when there is no prior information about what types of observation are possible.

Predictive inferences from the IDMM, expressed in the form of posterior upper and lower probabilities and expectations or prediction sets, are easy to interpret and use in making decisions. As illustrated in Sections 5 and 7, predictive inferences can often be displayed graphically, in the form of posterior upper and lower cdfs for the quantities of interest. The degree of imprecision of the predictive probabilities reflects the amount

of information on which they are based; predictions tend to become more precise as the observed sample size increases. The IDMM has no difficulty in handling cases where some categories have observed frequency zero. As seen in Sections 4–7, calculations of inferences are relatively simple because they involve only *DiMn* distributions. Finally, the IDMM seems to give reasonable answers in the special cases studied in Sections 4–6, and to improve on objective Bayesian inferences in those cases where there is a substantial disagreement (e.g. Section 6).

In conclusion, we will outline some of the most important problems that need to be addressed in future research.

(i) *Frequentist validity and the choice of* s. As shown in subsections 5.5 and 6.3 and in Walley (1996a), inferences from the IDMM with $s = 1$ agree in some cases with standard frequentist inferences. In these cases, inferences that are based on the IDMM(s), such as prediction sets or hypothesis tests, are valid from a frequentist point of view if and only if $s \geq 1$. This is a strong argument for taking $s \geq 1$ in the IDMM. On the other hand, the posterior upper and lower probabilities produced by the IDMM become increasingly imprecise, and thus increasingly uninformative, as s increases. In frequentist terms, procedures based on the IDMM have decreasing *power* as s increases. For inferences to be useful, and for procedures to be reasonably powerful in the frequentist sense, s should not be much greater than one.

The cases of agreement between frequentist inferences and the IDMM(1) in subsections 5.5 and 6.3 involve a particular sampling method (fixed n), and essentially they involve only two categories. It needs to be investigated whether the same kind of agreement holds for other kinds of sampling rules, particularly inverse sampling (fixed x_i), and for inferences involving more than two categories. It appears that inferences for binary data from the IDMM(1) do have frequentist validity under a wide range of sampling rules, including inverse sampling, but it is not clear whether this remains true for more complicated inferences involving multiple categories. We believe that further investigation of this question will help to determine an optimal value of s. We suggest choosing the smallest value of s for which inferences from the IDMM have frequentist validity under these more general conditions.

(ii) *Axiomatic characterization.* Another important problem is to characterize the IDMM and a unique value of s, or possibly an alternative model, by adding further desiderata to those listed earlier in this section. The aim here is to justify a unique method of predictive inference or *inductive logic*, along the lines attempted by Carnap (1952). The extra desiderata could include frequentist properties such as those suggested in (i), as well as exchangeability, parametrisation invariance, and other invariance axioms suggested by Carnap.

(iii) *Enlarging the IDMM.* It is possible to embed the prior IDMM class (4) in a larger class of exchangeable prior distributions for $\boldsymbol{x}^*$, in order to produce prior upper and lower probabilities that are even closer to vacuous than those for the IDMM. In the example of subsection 2.3, where X_1^* denotes the number of black balls in an urn which contains two balls, the IDMM produces a prior upper probability $\overline{P}(X_1^* = 1) < \frac{1}{2}$. In some applications, to better model prior ignorance about the frequencies, we may require a value greater than $\frac{1}{2}$. That can be achieved by adding prior distributions which satisfy $P(X_1^* = 1) > \frac{1}{2}$ to the IDMM class.

The IDMM class (4) contains only *DiMn* distributions, which are *infinitely exchangeable* in the sense that they can be extended to exchangeable probability distributions on an infinite sequence of observations. Infinite exchangeability is a necessary requirement for multinomial sampling from an infinite population, but it is not clear that it is always necessary when we are sampling from a finite population. In the latter case we might want to enlarge the IDMM by adding prior distributions, such as multiple hypergeometric distributions, that are finitely but not infinitely exchangeable. It is not clear whether this can be done in such a way as to satisfy the embedding and representation invariance principles.

(iv) *Substantive prior information.* The IDMM can be generalized to model substantive prior information, by using a more general set of values for the hyperparameters $\boldsymbol{\alpha}$ of the *DiMn* distributions. For example, the hyperparameter s can be allowed to vary between upper and lower bounds $\overline{s}$ and $\underline{s}$, and $\boldsymbol{\varphi}$ can be restricted to a suitable subset of the unit simplex. Possible models for binary data ($k = 2$) are suggested in Walley (1991, Section 5.4) and Walley et al. (1996). There is great potential for developing models for various kinds of prior information, because imprecise probabilities can model many effects that cannot be modelled by a single Bayesian prior. For example, they can respond in a useful way to conflict between prior beliefs and observed data, as explained in Walley (1991) and Walley et al. (1996). A specific problem is to find ways of modelling prior information about correlations between frequencies. Such information is often present when the observations are discrete measurements on an ordered scale, especially when they are real-valued measurements made with finite precision; then it is often reasonable to expect a positive correlation between frequencies in adjacent cells.

Acknowledgements

The authors would like to thank M. Reynaud and F. Donzel of INSEE (Paris, France) for providing them with the data of Section 7. Peter Walley would like to thank the Laboratoire Cognition et Activités Finalisées, Université Paris 8, and the Departamento de Estadistica y Matematica Aplicada, Universidad de Almeria, for support in carrying out this research.

References

Aitchison, J. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis.* Cambridge: Cambridge University Press.

Basu, D. (1975) Statistical information and likelihood. *Sankhya*, Series A, **37**, 1–71.

Bayes, T. R. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, London*, Series A, **53**, 370–418. Reprinted in *Biometrika* (1958), **45**, 243–315.

Berger, J. O. and Bernardo, J. M. (1992) Ordered group reference priors with applications to a multinomial problem. *Biometrika*, **79**, 25–37.

Berger, J. O. and Wolpert, R. L. (1984) *The Likelihood Principle.* Volume 6 of IMS Lecture Notes, Monograph Series. Hayward, California: Institute of Mathematical Statistics.

Bernard, J.-M. (1996) Bayesian interpretation of frequentist procedures for a Bernoulli process. *American Statistician*, **50**, 7–13.

Bernard, J.-M. (1998a) Bayesian inference for categorized data. In *Statistical Inference in the Strategy of the Researcher* (eds H. Rouanet et al.). Berne: Peter Lang.

Bernard, J.-M. (1998b) Bayesian implicative analysis for multivariate binary data using an imprecise Dirichlet model. Submitted for publication.

Bernardo, J. M. (1984) Comment, in discussion of Geisser (1984). *American Statistician*, **38**, 247–248.

Bernardo, J. M. and Ramon, J. M. (1998) An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician*, **47**, 101–135.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. New York: Wiley.

Bjørnstad, J. F. (1990) Predictive likelihood: a review. *Statistical Science*, **5**, 242–265.

Boole, G. (1854) *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities*. London: Macmillan.

Bru, B. (1986) Postface. In Laplace (1812/1825), 245–300.

Carnap, R. (1952) *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Caron, N., Ravalet, P. and Sautory, O. (1996) Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises. *Research Report* **9602**, Méthodologie Statistique, INSEE, Paris.

Dale, A. I. (1991) *A History of Inverse Probability: From Thomas Bayes to Karl Pearson*. New York: Springer-Verlag.

Dempster, A. P. (1966) New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, **37**, 355–374.

Feller, W. (1968) *Introduction to Probability Theory and its Applications*, Vol. 1, 3rd edn. New York: Wiley.

de Finetti, B. (1974/1975) *Theory of Probability*, Vols. 1 and 2. New York: Wiley.

Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Geisser, S. (1984) On prior distributions for binary trials (with discussion). *American Statistician*, **38**, 244–251.

Geisser, S. (1993) *Predictive Inference: An Introduction*. Monographs on Statistics and Applied Probability 55. New York: Chapman and Hall.

Good, I. J. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, Massachusetts: Massachusetts Institute of Technology Press.

Guttman, I. (1970) *Statistical Tolerance Regions: Classical and Bayesian*. London: Griffin.

Haldane, J. B. S. (1948) The precision of observed values of small frequencies. *Biometrika*, **35**, 297–300.

Hampel, F. (1993) Some thoughts about the foundations of statistics. In *New Directions in Statistical Data Analysis and Robustness* (eds S. Morgenthaler, E. Ronchetti and W. A. Stahel), pp. 125-137. Basel: Birkhauser.

Hampel, F. (1996) On the philosophical foundations of statistics: Bridges to Huber's work, and recent results. In *Robust Statistics, Data Analysis, and Computer Intensive Methods; In Honor of Peter Huber's 60th Birthday* (ed. H. Reider), pp. 185-196. New York: Springer.

Hoadley, B. (1969) The compound multinomial distribution and Bayesian analysis of categorical data from a finite population. *Journal of the American Statistical Association*, **64**, 216–229.

Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, Series A, **186**, 453–461.

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

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*. New York: Wiley.

Johnson, W. E. (1932) Probability: the deductive and inductive problems. *Mind*, **49**, 409–423.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Keynes, J. M. (1921) *A Treatise on Probability.* Vol. 8 of Collected Writings (1973 edition). London: Macmillan.

Kuboki, H. (1998) Reference priors for prediction. *Journal of Statistical Planning and Inference*, **69**, 295–317.

Lad, F. (1996) *Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction.* New York: Wiley.

Laplace, P.-S. (1778) Mémoire sur les probabilités, Article 33. In *Oeuvres Complètes de Laplace* (edited in 1893), pp. 383–485. Paris: Gauthiers-Villars et fils.

Laplace, P.-S. (1812/1825) *Essai Philosophique sur les Probabilités*, 5th edition (1825). Reprinted in 1986 by Christian Bourgois éditeur, Paris.

Laplace, P.-S. (1820) *Théorie Générale des Probabilités*, 3rd edn. Reprinted in Laplace, *Théorie Analytique des Probabilités*, Volume 2 (1995). Paris: Editions Jacques Gabay.

Mosimann, J. E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**, 65–82.

Pearson, K. (1920) The fundamental problem of practical statistics. *Biometrika*, **13**, 1–16.

Peirce, C. S. (1878) The probability of induction. *Popular Science Monthly.* Reprinted in *Philosophical Writings of Peirce* (ed. J. Buchler, 1955), Ch. 13. New York: Dover.

Perks, F. J. A. (1947) Some observations on inverse probability including a new indifference rule (with discussion). *Journal of the Institute of Actuaries*, **73**, 285–334.

Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory.* Cambridge, Massachusetts: Harvard University Press.

Stigler, S. M. (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900.* Cambridge, Massachusetts: Belknap Press.

Thatcher, A. R. (1964) Relationships between Bayesian and confidence limits for predictions (with discussion). *Journal of the Royal Statistical Society*, Series B, **26**, 176–192.

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

Walley, P. (1996a) Inferences from multinomial data: Learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society*, Series B, **58**, 3–57.

Walley, P. (1996b) Measures of uncertainty in expert systems. *Artificial Intelligence*, **83**, 1–58.

Walley, P., Gurrin, L. and Burton, P. (1996) Analysis of clinical data using imprecise prior probabilities. *The Statistician*, **45**, 457–485.

Zabell, S. L. (1982) W. E. Johnson's 'sufficientness' postulate. *Annals of Statistics*, **10**, 1091–1099.