Revue Internationale de

ISSN 0980-1472

systēmique

Vol. 11, N° 1, 1997

afcet

DUNOD

AFSCET

Revue Internationale de systēmique

Revue Internationale de Sytémique

volume 11, numéro 1, pages 11 - 29, 1997

Bayesian Analysis of tree-strutured categorized data

Jean-Marc Bernard

Numérisation Afscet, mars 2016.

REVUE INTERNATIONALE DE SYSTÉMIQUE
Vol. 11, N° 1, 1997, pp. 11 à 29

BAYESIAN ANALYSIS OF TREE-STRUCTURED CATEGORIZED DATA*

Jean-Marc BERNARD [1]

Résumé

Le modèle probabiliste bayésien, actuellement largement utilisé en psychologie cognitive, est connu des statisticiens en tant qu'une méthode d'inférence statistique : *l'inférence bayésienne*. nous présentons ce second point de vue en considérant le problème de l'analyse de données catégorisées présentant une structure d'arbre. L'approche bayésienne fait intervenir des distributions de Dirichlet dont nous donnons les principales propriétés et, en particulier, une propriété d'indépendance qui fait intervenir la structure d'arbre, et qui justifie l'approche de l'*inférence spécifique*. Nous discutons le choix de la distribution initiale dans le cadre non-informatif, et la structure d'arbre suggère une solution particulière : l'*initiale standard*. Enfin, l'analyse d'un exemple de données séquentielles provenant de l'éthologie permet d'illustrer certains avantages décisifs de l'inférence bayésienne pour analyser des données.

Abstract

The Bayesian probabilisitic model, widely used by cognitive psychologists, is known to statisticians as a method of statistical inference: *Bayesian inference*. We present this latter viewpoint by considering tree-structured categorized data. The Bayesian approach involves Dirichlet distributions, of which we give the main properties and, in particular, an independence property, related to the tree-structure, which provides a justification to the *specific inference* approach. We discuss the choice of the prior adopting a non-informative viewpoint, and the tree-structure guides us to a particular solution: the *standard prior*. Finally, the analysis of an example of sequential data from the

* This paper is an outgrowth of a paper presented by the author and Henry Rouanet at the "Mathematical Psychology Meetings" held in Cambridge (MASS), USA, August 1986.

1. Laboratoire de Psychologie Cognitive, CNRS ER 125, Université Paris-8, 2, rue de la Liberté, 93526 Saint-Denis Cedex.

Rev. intern. systémique. 0980-1472 Vol. 11/97/01/ $ 7.00/© Afcet Gauthier-Villars

field of ethology enables us to illustrate some decisive advantages of Bayesian inference for analyzing data.

I. INTRODUCTION

Probabilistic models meet an increasing interest amongst cognitive psychologists and AI researchers. Within this context, the *Bayesian model*, named after the clergyman Thomas Bayes because of his famous "Bayes' theorem", is a particularly privileged model. Fundamentaly, the Bayesian model is a cumulative probabilistic model: the initial state of knowledge of the "system", which is expressed by means of probabilities, is updated into a final state of knowledge, by the taking into account of one or several "observations" issued from the "world"; in its turn, this final state becomes the new initial state, that could be updated by some future observations. The strength of the Bayesian approach, as clearly advocated by De Finetti (1974, 1975), is that it provides a coherent means of updating the "system" 's probabilities through observable events.

Within the context of Statistics, this model is the foundation of *Bayesian inference*, in which the "observations" are statistical data, the "system" is the statistician, *i.e.* the one who analyzes the data, and the initial and final states are described respectively by a *prior* and a *posterior* probability distribution.

Because of what preceeds, the Bayesian approach is an important object of study for both psychologists and statisticians, and what is presented in this paper may be considered, either as a learning probabilistic model, or as a statistical method of inference. In the following, we shall only adopt this latter viewpoint by considering Bayesian inference for categorized data that are underlied by some hierarchical tree-structure. But, from the former viewpoint, that we shall leave aside here, most of what follows has profound connections with the problem of the probabilistic modelling of human categorization (see, e.g., Anderson, 1991).

The current existing methods of statistical inference may be grouped into two major classes: the *frequentist methods,* from which significance tests and confidence intervals are derived, and the *Bayesian methods*. Because of the known limitations of the frequentist methods – they do not provide answers to some questions that are quite natural –, more and more statisticians are led to envisage the Bayesian methodology as a necessary complement. From the statistical viewpoint of *analyzing data*, one crucial issue, within the Bayesian approach, is that of specifying an "initial state of knowledge of the

statistician" which can be considered as "neutral" or "vague" enough, in such a way as providing a final state which conveys solely the information brought by the data, and, thus, in such a way as confering an "objective" status on this final state of knowledge. We have deliberately set the words "neutral", "vague" and "objective" within quotes in the previous sentence, because this goal is not as simple as it might sound and it has led to much work, debate, and proposals, including the *non-informative* approach of Jeffreys (1961) and the more recent *reference* approach of Bernardo (1979) and Bernardo and Smith (1994).

The high desirability of the use of Bayesian non-informative methods within the context of ANOVA, was particularly stressed by Lépine and Rouanet (1975), Rouanet and Lecoutre (1983) and Rouanet (1996). Following this line, we investigated similar methods for categorized data (Bernard, 1986; Bernard, 1991), with a particular focus on *structured data*, *i.e.* the analysis of complex designs. It appears that this framework can be easily extended to *tree-structured data.* For the analysis of categorical data, within the Bayesian approach, the "multinomial-Dirichlet" model plays a central role. This will be the main concern of this article. Further, we shall also be particularly concerned here with the properties of this model when considering that the data are underlied by some particular tree-structure.

This paper is organized as follows. Section II, defines tree-structured categorized data and the concept of measures on a tree. Assuming that the data are a random sample from a multinomial population, the usual Bayesian approach, that involves Dirichlet prior and posterior distributions, is presented in Section III. Section IV gives some basic properties of Dirichlet distributions, and Section V describes an important independence property related to the tree-structure of the data; some of the methodological implications of this property are emphasized, and related literature is briefly reviewed. In Section VI, we discuss the choice of the prior distribution for the purpose of analyzing data, which leads us to a "standard prior" for tree-structured data. Finally, Section VII presents an example of ethological sequential data that is analyzed by the proposed methodology.

II. TREE-STRUCTURED CATEGORIZED DATA

II.1. Tree representations: an example

Let us take, as a first example, a simple experiment in which n subjects are tested on three successive occasions or trials, each trial having only two possible outcomes: a success (S) or a failure (F). When a failure occurs,

the subject cannot be tested any further. This design can be represented by the tree-structure of Figure 1.

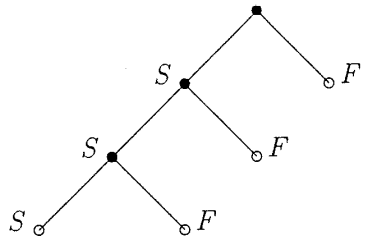

Figure 1. A design involving four basic categories, from three successive "success-failure" (S-F) trials, represented with its tree-structure; each level of the tree corresponds to one trial.

Another way to describe this design would be to only consider the leaves of the preceding tree, *i.e.* the four basic categories into which any subject must fall: SSS, SSF, SF, or F. Both descriptions may be envisaged, but adopting the tree-structure representation stresses that some of the basic categories may be lumped together preferentially, or alternatively, that the data analysis will focus on some specific conditional frequencies. In brief, the tree-structure might either be dictated by the design structure itself, or be suggested by a specific question asked to the data. Whichever the case is, we shall consider, in the following, that the set of categories and their tree-structure are both given and fixed, and proceed from that point onwards.

II.2. Trees

A partial-order relation ">" ("covers") on the set T defines a tree if the associated graph is connected, without cycles, and such that any pair of elements have a supremum. One element covers all the other elements: the *root*. The relation ">" partitions the set T into two subsets: N, the *nodes, i.e.* the subset of elements that cover some other elements (they appear as filled circles in figures); U, the *leaves*, *i.e.* the subset of elements that do not cover any element (they appear as empty circles in figures).

In terms of categorized data, the set U represents the K basic observable categories, and N the privileged lumpings. From now on, the example of a categorized variable U_5, with $K = 5$ categories, with two privileged lumpings, will be taken as a typical example that we shall use illustratively throughout this article (*see* Figure 2). It is convenient to label u_1, u_2, etc., the leaves, and to label u_{234} a node that covers leaves u_2, u_3 and u_4.

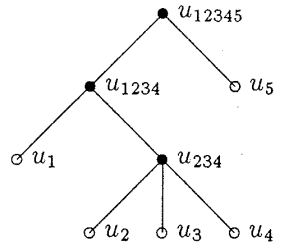

Figure 2. Example of a tree underlying the set of $K = 5$ categories U_5; leaves are labeled u_1, u_2, u_3, u_4, u_5, and nodes u_{1234}, u_{234}, including the root u_{12345}.

II.3. Measures on a tree

For the purposes of this paper, we need to consider several measures on the set U, for instance the observed counts $a = (a_k)_{k\in[1,\ldots,K]}$ of a sample of size n, or the corresponding observed frequencies $f = (f_k)_{k\in[1,\ldots,K]}$, with $f_k = a_k/n$. Any measure can be extended to a tree T underlying U in an obvious additive way: the measure-value for a node $c \in N$ is the sum of measure-values of elements of U covered by c. These measures, when extended to tree T, will be noted a_T or f_T, and the value of a_T for node u_{234} will be noted a_{234}.

Two types of measures are involved when analyzing categorized data: *frequency-measures,* e.g. f_T, that are always normalized ($f_{\text{root}} = f_{12345} = 1$), and *force-measures*, e.g. a_T, that are not necessarily so [1]. When considering a sub-tree T' of a larger tree T, as we shall do in what follows, force-measures are simply restricted to the sub-tree, whereas frequency-measures are normalized. When a normalization occurs for a sub-tree, the frequencies are *conditional frequencies*, or *transition frequencies*, noted e.g. $f_2^{234} = f_2/f_{234}$.

III. BAYESIAN INFERENCE

The basic observable categories are the K leaves, so that the inferential model, including the sampling model and the prior distribution, may be stated on the set U only.

Let us consider that the data consist of a random sample of n observations from an infinite population and that each observation falls into one of the $K = 5$ categories of $U_5 = \{u_1, u_2, u_3, u_4, u_5\}$. This

defines a multinomial sampling model for the categories counts $a = (a_1, a_2, a_3, a_4, a_5)$ conditionally on the population parameters: the parent or true frequencies $\varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5)$, with $\sum \varphi_k = 1$:

$$a|\varphi \sim MN(n, \varphi). \tag{1}$$

Bayesian inference requires the introduction of a *prior distribution* on φ, which is the probabilistic expression of the knowledge about φ, prior to the data. For multinomial data, we choose, as is usual, the prior distribution in the family of Dirichlet distributions – the *conjugate family* for multinomial sampling. A Dirichlet distribution $D(\alpha)$ depends on K hyperparameters, one for each category: $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$. If the prior on φ is chosen $D(\alpha)$, Bayes' theorem leads to a Dirichlet posterior distribution $D(a + \alpha)$ on $\varphi|a$:

$$\begin{array}{c} a|\varphi \sim MN(n, \varphi) \\ \text{and} \\ \varphi \sim D(\alpha) \end{array} \quad \Rightarrow \quad \varphi|a \sim D(a + \alpha) \tag{2}$$

The hyperparameters α are called *prior* (or *initial*) forces, and $\nu = \sum \alpha_k$, the *total prior force*. Similarly, the observed counts a are called the *observed forces* and n the *total observed force*. These two kind of forces are combined in an additive way, through equation (2), to give the *posterior* (or *final*) forces $a + \alpha$.

Seen as a learning model, Bayes' theorem can be expressed in the following way: the initial state of knowledge about the unknown parameters φ, specified by the prior distribution, is updated by the data, thus giving the posterior distribution. This updating is quite simple here, since it just amounts to adding, for each category u_k, the corresponding observed force (a_k) to the prior one (α_k), to get the posterior force $(a_k + a_k)$.

The posterior distribution, $\varphi|a \sim D(a + \alpha)$, is the basic result of the Bayesian approach, from which all relevant inferential statements are drawn. Of course, for this distribution to be fully specified, it is necessary to choose some particular prior forces α; this issue will be discussed in section 6.

In this section, we have introduced some new measures on set U. Let us summarize the several measures in presence and their respective roles. The observed frequencies, f, represent a known frequency-measure, determined by the force-measure a. The true frequencies, φ, represent an unknown frequency-measure, the knowledge on which is expressed probabilistically by means of a Dirichlet distribution characterized by some force-measure, α prior to the data, and $a + \alpha$ posterior to the data.

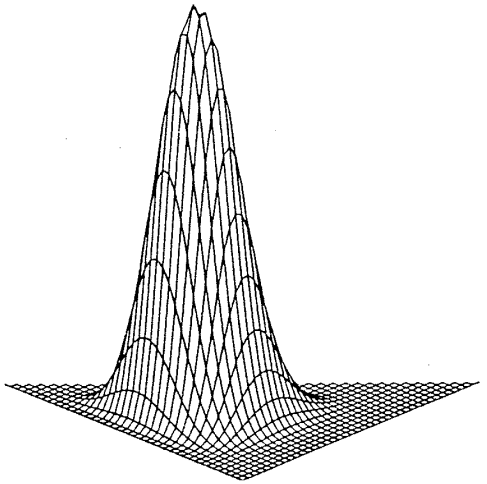

Figure 3. Example of a Dirichlet distribution $D(\alpha)$, with $K = 3$ and $\alpha = (10, 8, 6)$; the leftmost point of the simplex corresponds to $\varphi = (1, 0, 0)$, the frontmost point to $\varphi = (0, 1, 0)$, and the rightmost point to $\varphi = (0, 0, 1)$.

IV. DIRICHLET DISTRIBUTIONS

As may be seen from equation (2), the state of information on φ is at both stages, prior or posterior, conveyed by some particular Dirichlet distribution, so that these distributions require our particular attention. Let us consider a distribution $D(\alpha)$, where α represents either the prior forces or the posterior ones. The Dirichlet distribution is the multivariate generalization (K categories) of the Beta distribution (2 categories). Its major properties may be found in Wilks (1962). The density of the Dirichlet distribution $D(\alpha)$, $\alpha_k > 0$, with total force $\nu = \sum \alpha_k$, is defined over the $(K-1)$-dimensional simplex (because of the constraint $\sum \varphi_k = 1$) by,

$$h(\varphi) = \frac{\Gamma(\nu)}{\prod \Gamma(\alpha_k)} \prod \varphi_k^{\alpha_k - 1}. \tag{3}$$

The mean of the Dirichlet distribution is solely determined by the relative forces of the categories, α_k/ν, but its dispersion mainly depends on the total force ν: the larger ν is, the more concentrated the distribution is. Figure 3 gives an example of a Dirichlet distribution for $K = 3$, $\alpha = (10, 8, 6)$ and $\nu = 24$.

Two important properties of the Dirichlet distribution will be relevant for the remaining of this article. They will be stated on the example of a set U of $K = 5$ categories, but are nethertheless general.

Lumping property: When lumping several categories together, e.g. u_2, u_3 and u_4, into a single one, the corresponding forces add: $(\varphi_1, \varphi_{234}, \varphi_5) \sim D(\alpha_1, \alpha_{234}, \alpha_5)$. Thus, the Bayesian results are compatible with any tree structure underlying U, so that equation (2) can be rewritten in terms of the measures extended to the three T : $\varphi_T \sim D(\alpha_T)$ and $\varphi_T | a_T \sim D(a_T + \alpha_T)$.

Restriction property: If the inference is restricted to the relative frequencies of categories u_2, u_3 and u_4 only, *i.e.* the conditional frequencies φ_2^{234}, φ_3^{234}, and φ_4^{234}, then only the corresponding forces need to be considered: $(\varphi_2^{234}, \varphi_3^{234}, \varphi_4^{234}) \sim D(\alpha_2, \alpha_3, \alpha_4)$.

V. AN INDEPENDENCE PROPERTY OF THE DIRICHLET DISTRIBUTION

As we said earlier, considering that data are tree-structured, according to tree T, indicates that the analysis focusses on some conditional frequencies associated with tree T. The definition of the conditional frequencies of interest can be described by means of the operation of "cutting" a tree at some of its nodes.

V.1. "Cutting" a tree

The *node-cutting* operation of T at a particular node c splits the tree into two sub-trees. The upper sub-tree, $\overline{T}$, contains the initial root; c becomes a leaf of $\overline{T}$; the measures associated with $\overline{T}$ remain unchanged, apart from their restriction to the sub-tree. The lower sub-tree, $\underline{T}$, does not contain the initial root and c becomes its new root; The force-type measures, $a_{\underline{T}}$ and $\alpha_{\underline{T}}$, remain unchanged, while the frequency-type measures, $f_{\underline{T}}$ and $\varphi_{\underline{T}}$, are normalized.

Figure 4 gives the example of the tree T of figure 2 cut at node u_{234} and the two resulting sub-trees, $\overline{T}$ and $\underline{T}$, for both of which we have displayed the associated measures α and φ.

V.2. An independence property

THEOREM 1. – *For a node-cutting operation on tree T, spliting T into $\overline{T}$ and $\underline{T}$, and if, $\varphi_T \sim D(\alpha_T)$, then,*

$$\begin{aligned} &(a)\ \varphi_{\overline{T}} \sim D(\alpha_{\overline{T}}) \\ &(b)\ \varphi_{\underline{T}} \sim D(\alpha_{\underline{T}}) \\ &(c)\ \varphi_{\overline{T}} \perp\!\!\!\perp \varphi_{\underline{T}} \end{aligned} \tag{4}$$

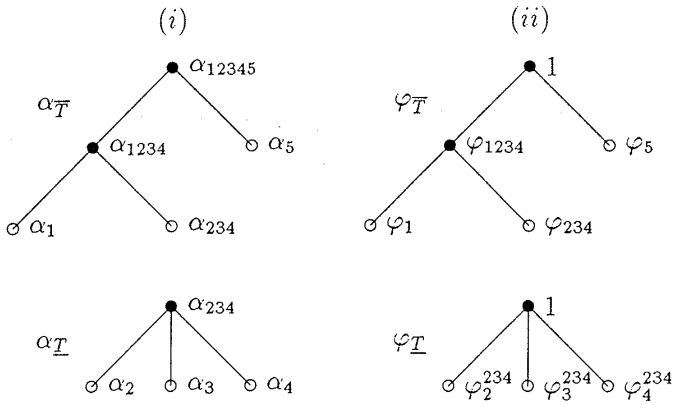

Figure 4. Cutting the tree of figure 2 at node u_{234}: (i) the two sub-trees, $\overline{T}$ and $\underline{T}$, after cutting, with the associated force-measures $\alpha_{\overline{T}}$ and $\alpha_{\underline{T}}$; and (ii) the same sub-trees, with the associated frequency-measures $\varphi_{\overline{T}}$ and $\varphi_{\underline{T}}$.

As can be seen, (a) is simply a restatement of the lumping property, as well as (b) is a restatement of the restriction property. The (c) part is the new interesting result: it expresses that the two distributions of $\varphi_{\overline{T}}$ and $\varphi_{\underline{T}}$ are independent [2]. The next theorem follows from the recursive use of theorem 1.

THEOREM 2. – *Let N' be a subset of N, with $|N'| = p$. The p node-cutting operations performed at every node of N' transform the Dirichlet distribution over the parent frequency-measure φ_T, into $(p+1)$ independent Dirichlet distributions over the frequency-measures $\varphi_{T'}$ on the $(p+1)$ sub-trees T' obtained.*

Applied to the example of figure 2, for $N' = \{u_{1234}, u_{234}\}$, theorem 2 leads to the three following independent distributions:

$$\begin{aligned} (\varphi_{1234}, \varphi_5) &\sim D(\alpha_{1234}, \alpha_5) \\ (\varphi_1^{1234}, \varphi_{234}^{1234}) &\sim D(\alpha_1, \alpha_{234}) \\ (\varphi_2^{234}, \varphi_3^{234}, \varphi_4^{234}) &\sim D(\alpha_2, \alpha_3, \alpha_4) \end{aligned} \tag{5}$$

How can we express in words what theorem 2 states ? One aspect is that it basically describes an *inheritance* property of the multinomial-Dirichlet model. We first started by setting a general multinomial-Dirichlet model on the overall tree, and we find out here that, when considering solely any local or specific level, *i.e.* some sub-tree, we have, in fact, a local model which has a similar mathematical expression: it still involves a Dirichlet distribution, dealing with forces and frequencies that have been respectively restricted or

normalized. Further, theorem 2 also expresses that such a local model can be, because of independence, disconnected from the rest of the tree, so that the expression "cutting a tree" may really be given its full meaning. We shall turn back to some implications of that in section V.4.

Let us also notice that theorem 2 does not rely on any assumption about the tree T; it is thus true for *any tree* underlying U. For that reason, one may say that Dirichlet distributions are, in some way, "in harmony" with tree-structures.

V.3. Related literature

The Dirichlet distribution can be characterized in terms of independence properties (Darroch, Ratcliff, 1971; Fang, Kotz, Ng, 1990, ch. 1; Mosimann, 1962). In Lindley (1964), a particular case of theorem 2 is given for the case of contingency tables.

Connor and Mosimann (1969) put forward the concepts of *neutrality* and of *complete neutrality*. Though these authors do not refer to trees nor to tree-structures, each of these two notions may actually be defined by considering a *particular* set of node-cutting operations on a *particular* tree, for which theorem 2 holds. The Dirichlet distribution is thus "highly neutral" since, as we previously noticed, theorem 2 holds for any tree and any set of node-cutting operations.

In fact, all independence properties of the Dirichlet distribution found in the literature are several viewpoints on the fundamental property expressed by theorem 2. The explicit use of tree-structures, that was not envisaged in previous work, provides a unifying framework for all these previous results.

V.4. Methodological implications

Specific inference: Let us consider that some question of interest bears on a particular conditional frequency, θ, within a larger tree-structured design, and that the observed value of that frequency is: $t = a/(a+b)$, were a and b are two observed counts. From the lumping and restriction properties only, the posterior distribution of θ is $\mathrm{Beta}(a+\alpha, b+\beta)$, where α and β are two prior forces determined by the overall prior distribution. From the overall dataset, only the relevant data, a and b, are involved. The independence property tells us more: the state of knowledge about θ does not depend on the value of *any* other conditional frequency appearing at another level of the tree. Thus theorem 2 is a justification for the "specific inference approach"

(Rouanet, Lecoutre, 1983): for a specific question, involving a restricted number of parameters, θ instead of φ, the general model can be replaced by a specific model including less parameters and less data. The overall tree will be replaced by the minimal relevant sub-tree. But, the general model and the specific one will fully agree, *i.e.* give precisely the same posterior distribution, only if the overall prior and the specific one are "compatible". Thus point will be discussed in section VI.

Design equivalence: Theorem 2 shows a technical equivalence between the analysis of sequential data and the analysis of data collected according to a design with independent groups. Let us come back to the example of successive "success-failure" trials of section 2.1. In order to compare the "frequency of success in trial 1" with the "frequency of success in trial 2 after a successful first trial", the two following designs could be envisaged by an experimenter:

(a) One single group of subjects is tested on the first trial; the remaining successful subjects are then tested on the second trial.

(b) A first group of subjects is tested on the first trial. A second group of subjects goes through a selection phase: only subjects that are successful on the first trial are kept; these subjects are then tested on the second trial, being the experimental stage.

According to theorem 2, for the question of interest considered, these two designs are to be treated in exactly the same way: in either case, the two frequencies of interest follow independent Beta distributions. (Let us recall that a Dirichlet distribution bearing on two categories is a Beta distribution.) Again, the only difference might come from the selection of incompatible priors for the two situations.

VI. SPECIFYING THE PRIOR DISTRIBUTION

All results given so far are general and valid whatever the prior distribution $D(\alpha)$. How should this one be chosen? We shall only answer this question by adopting the *data analysis approach*, in which the goal of the analysis is to bring out the information on the unknown true frequencies φ that is provided by the data only, without taking into account any possible external or previous information. This amounts to choosing a prior formalizing an initial state of "ignorance".

A considerable amount of work has been devoted to the search of a suitable ignorance prior for categorical data (*see* e.g. references in Bernardo, 1979). All

proposed priors lie between $D(0)$, *i.e.* $\alpha_k = 0$ for all k, (Haldane, 1948) and $D(1)$ (the Bayes-Laplace solution), including $D(1/2)$ the Jeffreys' (1961) indifference prior, and $D(1/K)$ proposed by Perks (1947). These various solutions differ very little, in terms of the posterior distribution, when the number of categories K is small relative to the sample size n.

For the general model and a specific model to coïncide, for any specific inference, the only possible prior is Haldane's $D(0)$. Despite this strong argument in its favor, this solution is not satisfactory for the case of small counts or unobserved categories (some a_k are 0), as it then leads to inferential statements that are too "data-glued". On the other hand, in Jeffreys' and Bayes-Laplace solutions, the total prior force depends on the number K of categories, so that undesirable properties appear when lumping or splitting categories. Perks' solution overcomes this problem by fixing the total prior force to $\nu = 1$ and by sharing it evenly among the K categories.

The previous solutions are suited for the case of symmetrical categories, as they all give the same prior force to each category. The case of asymmetrical categories has, comparatively, less been considered. In the *reference prior* as defined by Bernardo (1979) and Berger and Bernardo (1992), asymmetry is introduced by distinguishing, among the φ_k, the parameters of interest from the nuisance parameters; the proposed solution amounts to adopting what we called earlier the "specific approach" with a Jeffreys' prior, $D(1/2)$, on the leaves of the relevant sub-tree.

This last solution, though, does not take into account the tree-structure of the relevant sub-tree itself. This is the reason why we suggest the use of the following *standard prior* defined as an adaptation of Perks' (1947) solution: the total prior force $\nu = 1$ is put onto the root of the tree, as in Perks' prior, and is evenly split among the elements just covered by the root; for a node-element, this force is then split again onto the next level of the tree; this process is applied recursively until all leaves have been reached (*see* Figure 5). Finally, if the inference focuses on a restricted number of parameters, the specific inference approach is adopted: the minimal sub-tree containing all parameters of interest is considered, and the prior, as defined previously, is specified on this relevant sub-tree only.

Instead of focusing on the selection of one single prior, an alternative solution is to consider a *set* of admissible ignorance priors. We recently suggested this idea of an *ignorance zone* for the case of binomial data (Bernard, 1996), while parallelly Walley (1996) was proposing an *imprecise Dirichlet model* (IDM) for multinomial data. In the IDM with fixed total prior force ν, prior ignorance is formalized by the set of all Dirichlet priors $D(\alpha)$

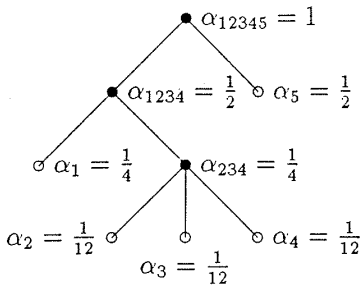

Figure 5. The standard prior for U underlied by the tree T; the total force $\nu = 1$ is evenly split at each level of the tree.

satisfying: $\forall k$, $\alpha_k \geq 0$ and $\sum_k \alpha_k = \nu$. Such a set of priors, that we shall denote $ID(\nu)$, leads, for each inference, to an interval for the associated probability, instead of a single probability value. When ν is small (the value $\nu = 1$ is a typical choice) and if the data size n is not too small, the lower and upper probabilities of an inferential statement of interest will generally be not too far apart, so that valuable inferences may be drawn.

In any case, no particular prior should be thought as *the good solution*, but rather as *a reasonable solution* that "have a minimal impact on the Bayesian analysis when compared with the impact provided by the data" (Berger, Bernardo, 1992, p. 26). We think that, if a single prior is to be chosen, the standard prior defined above is a quite reasonable solution for the case of tree-structured data that is our concern in this article. As a complement, we suggest that several of the previously described priors should be used, and corresponding results compared, in order to ensure that the conclusion is little affected by the prior's choice. In this respect, one possible recommandation is to use the IDM $ID(\nu = 1)$ which defines an ignorance zone including both the standard and Perks' priors.

VII. APPLICATION TO THE "ANTS" DATA

The approach presented in the preceding sections has already been applied to sequential data for the study of the predatory behaviour of an earwig (*see* e.g. Bernard, Blancheteau, Rouanet, 1985).

The following data have been communicated to us by Fresneau (1994) and deal with the survival of a colony of ants. The insects have been observed for several days, and the number of exits from the nest have been recorded for

each insect. It is assumed that the insects died during their last recorded exit. "Death" and "survival" are respectively designated by, 1 and I for the first exit, 2 and II for the second one, etc. The data for 162 ants and the first seven exits are shown in figure 6. One question of interest is to study the evolution of death-rate according to the number of the preceding successful exits.

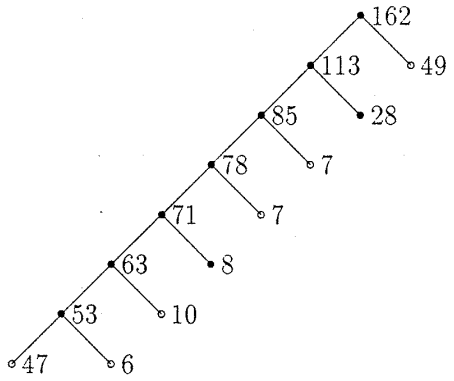

Figure 6. The ants survival data for the first seven exits. Out of 162 ants who exited at least once, 49 exited only once and 113 more than once, etc. After seven exits, only 47 ants survived.

The successive death-rates correspond to transition frequencies associated with the tree of figure 6; the observed death-rates are given in table 1. Descriptively, the death-rate appears to be much larger on the first two exits, than on the following five ones. The average death-rate on the first two exits is $t_1 = 27.5\%$, to be compared with $t_2 = 11.1\%$ for the next five, so that the observed difference between t_1 and t_2, $d_{obs} = 16.4\%$, is important. In the following, we shall focus on the *property of interest*, according to which the difference is greater than 10%. The property is true for the observed difference: $d_{obs} > 10\%$. The goal of the inferential analysis is to assess whether this property also holds for the corresponding true difference $\delta : \delta > 10\%$? The Bayesian way of answering such a question is straightforward as it simply amounts to calculating the posterior probability of the property of interest, *i.e.* $Prob\,(\delta > 10\%)$.

Table I. Observed death-rates in % for the first 7 exits: $f_1 = 49/162 = 30.2\%$, $f_2^I = 28/113 = 24.8\%$, etc.

f_1	f_2^I	f_3^{II}	f_4^{III}	f_5^{IV}	f_6^V	f_7^{VI}
30.2	24.8	8.2	9.0	11.3	15.9	11.3

According to theorem 2, the joint distribution over the true death-rates, *i.e.* the transition frequencies $(\varphi_1, \varphi_2^I, \varphi_3^{II}, ..., \varphi_7^{VI})$, can be easily determined: each of these frequencies is distributed Beta and all are mutually independent. The parameter of interest δ is defined as a contrast between these transition frequencies:

$$\delta = (\varphi_1 + \varphi_2^I)/2 - (\varphi_3^{II} + \varphi_4^{III} + ... + \varphi_7^{VI})/5 \qquad (6)$$

Using a standard prior, as defined in section 6, the mean and variance of the posterior distribution for δ are easily found to be: $mean = 0.164$ and $var = 0.032^2$; the standard posterior distribution is shown in figure 7. Due to the finite range of δ, we can approximate its distribution by a Beta distribution with same mean and variance (an exact algorithm could be used instead, of course). From this posterior distribution, we get the statement, $Prob(\delta > 10\%) = 0.978$, thus assessing the largeness of the true difference with a good guarantee. If, instead, we choose one of the usual symmetrical ignorance priors, $Prob\,(\delta > 10\%)$ becomes: 0.976 with a $D\,(0)$, 0.973 with a $D\,(1/K)$, 0.961 with a $D\,(1/2)$ and 0.938 with a $D\,(1)$. Last, with an imprecise Dirichlet prior $ID\,(\nu = 1)$, the required probability belongs to the interval [0.969; 0.980]. The sensitivity to the choice of the prior is seen to be moderate though not negligible. However, in any case the required probability can be assessed to be greater than 0.938: from the sample of 162 observed data, the property of interest ($d_{obs} > 10\%$) can thus be extended to the population ($\delta > 10\%$) with a guarantee of, at least, 0.938.

For a weaker statement such as $\delta > 7.5\%$, each of the preceding priors lead of course to a larger probability than for the previous statement, and all of these probabilities range from 0.990 to 0.998. So, in this case, the impact of the choice of the prior is of no practical importance.

If it is preferred to compare the average death-rates t_1 and t_2 through their ratio, rather than through their difference, the Bayesian approach is again straightforward: the observed ratio $r_{obs} = t_1/t_2 = 2.48$ presents the property $r_{obs} > 1.5$, which, using a standard prior, may be extended inferentialy with a good guarantee, as we have $Prob\,(\rho > 1.5) = .998$.

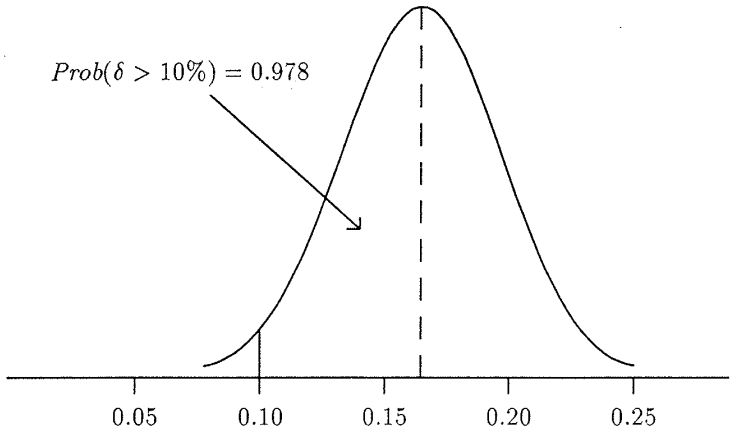

Figure 7. Posterior distribution of δ obtained with a standard prior; the region for which $\delta > 10\%$ has a probability of 0.978.

VIII. CONCLUSION

We would like to conclude by some general remarks refering to two different, though related, aspects of the present paper: The first ones, are about the statistical use of the multinomial-Dirichlet model, and the second ones, about some extensions that enable its use within the context of cognitive psychology.

As a statistical method, the multinomial-Dirichlet model is a simple powerful tool for the Bayesian analysis of categorized data. It is highly general since it makes no assumption about the population from which data are sampled, apart from stating that the population is characterized by a fixed, though unknown, set of frequencies φ.

Theorem 2 is an important property of this model. First of all, it expresses the fundamental "inheritance" property of the model. Whichever level of analysis is adopted, general or specific, the model takes the same mathematical form: At both of these levels, any state of knowledge – prior or posterior – is described by a Dirichlet probability distribution.

Secondly, we have pointed out that theorem 2 does not make any assumption about how categories are structured, so that it applies to any tree-structure (and possibly to several ones), thus allowing us to say that Dirichlet distributions are in "harmony" with trees. Finally, a major consequence of theorem 2, at the statistical level, is that it justifies the *specific inference*

approach which consists of solely modeling a sub-tree of interest rather than the overall tree. One can easily perceive the deep (and quite reassuring) meaning of that last point: it ensures that, as far as multinomial data are concerned, one is allowed to say something about a specific aspect of reality without having to look at it all at once [34].

The example analyzed in section 7, with its associated rather complex property of interest, has given us the opportunity to stress the extremely powerful nature of the Bayesian approach, which is not shared by the frequentist approach: it enables to extend *any* descriptive conclusion into the corresponding inductive one, by providing straightforward answers to questions of the following type: "In the light of some observed data having some property of interest, may the property be generalized to the population from which data were sampled?"

Though it was not discussed in the paper, it is worth mentionning that the Bayesian approach also provides a *predictive* answer to such questions: instead of focusing on statements about the population, the inference then deals with a possible future set of data.

"Going from observed events to a prediction about future events" is the very essence of the Bayesian paradigm, which basically involves an updating process from a prior prediction to a posterior prediction through already observed events; these predictions are expressed by probabilities and Bayes' theorem guarantees their overall coherence. This viewpoint has been particularly emphasized by De Finetti (1974, 1975) and Bernardo and Smith (1994).

As we have indicated in the introduction, the goal of "analyzing data" requires to consider an initial state of knowledge that may be considered as formalizing prior ignorance.

This view corresponds to a learning process starting "from point 0". But a more general view of the Bayesian approach is to envisage any possible initial state of knowledge. For categorized observations, this might be achieved by considering a Dirichlet prior with large (greater than one) initial forces α_k, or a prior obtained by a mixture of such Dirichlet distributions.

With this last feature, we already have a quite general probabilistic learning model. Several other extensions might be envisaged, depending on what is being modelled, but we shall just give here a few hints for the modelling of human categorization.

We have considered, throughout this paper, that the categories and their tree-structure were given in advance. This restriction needs to be relaxed if

we want to describe a sequential process where categories, and their possible structure, is progressively constructed.

Such a process would start from a limited set of pre-existing categories and should allow, either to create new categories, or to merge or split old ones. We suggest that, again, theorem 2 will be the key instrument for so doing. The three basic operations, just mentionned, amount to no more than changing a previous tree-structure into a new one, and the multinomial-Dirichlet model will propagates from one tree to the other by simply adding or spliting the concerned forces. As a matter of fact, the model proposed by Anderson (1991) involves Dirichlet distributions and already implements some of the above mentioned ideas, but, without any doubt, more work remains to be done in this direction.

Notes and references

1. In this article, the word "frequency" should always be understood as meaning "relative frequency". In the present context, the term "force" ("strength" could be used instead) will always refer to counts or quantities that are homogeneous to counts without necessarily being integers.

2. This theorem may be seen as a particular case of theorem 1.4 of Fang, Kotz and Ng (1990, p. 19); its proof is straightforward and proceeds by reexpressing the variable φ in terms of leaves' frequencies of $\varphi_{\overline{T}}$ on one hand and of $\varphi_{\underline{T}}$ on the other hand.

3. The fact that a specific aspect of reality can be analyzed independently from the context (*i.e.* more global aspects of the reality) does not mean that the context is absent; on the contrary, it is integrated into the "conditional part" of the relevant conditional frequencies involved.

4. As pointed out by an anonymous referee, theorem 2 might be thought to legitimate current statistical practice in which the selection of the data and the analysis of the selected data are implicitly considered as two independent stages. It should be kept in mind, though, that theorem 2 is a property of the multinomial-Dirichlet model which only constraints the frequencies φ_k to add up to one. More constrained models could be envisaged which would not share this property (*see* e.g. Connor and Mosimann, 1969).

J. R. ANDERSON, The Adaptive Nature of Human Categorization, *Psychological Review,* 1991, 98, No. 3, p. 409-429.

J. O. BERGER and J. M. BERNARDO, Ordered Group Reference Priors with Application to the Multinomial Problem, *Biometrika*, 1992, 79, No. 1, p. 25-37.

J.-M. BERNARD, Méthodes d'Inférence Bayésienne sur des Fréquences, *Informatique et Sciences Humaines*, 1986, 68, p. 89-133.

J.-M. BERNARD, Inférence Bayésienne et Prédictive sur les Fréquences in *L'Inférence Statistique dans la Démarche du Chercheur,* by H. ROUANET *et al.,* European University Studies, Series 6, Psychology, Berne, Peter Lang, 1991, p. 121-153.

J.-M. BERNARD, Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process, *The American Statistician,* 1996, 50, No. 1, p. 7-13.

J.-M. BERNARD, M. BLANCHETEAU and H. ROUANET, Le Comportement Prédateur chez un Forficule, Euborellia Moesta (Géné). II Analyse Séquentielle au Moyen de Méthodes d'Inférence Bayésienne, *Biology of Behaviour,* 1985, 10, p. 1-22.

J. M. BERNARDO, Reference Posterior Distributions for Bayesian Inference (with discussion), *Journal of the Royal Statistical Society,* Series B, 1979, 41, No. 2, p. 113-147.

J. M. BERNARDO and A. F. M. SMITH, *Bayesian Theory,* New-York: John Wiley & Sons, 1994.

R. J. CONNOR and J. E. MOSIMANN, Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution, *Journal of the American Statistical Association,* 1969, 64, p. 194-206.

J. N. DARROCH and D. RATCLIFF, A Characterization of the Dirichlet Distribution, *Journal of the American Statistical Association*, 1971, 66, No. 335, p. 641-643.

B. De FINETTI, *Theory of Probability,* Volumes 1 and 2, New-York, John-Wiley, 1974, 1975.

K. T. FANG, S. KOTZ and K. W. NG, *Symmetric Multivariate and Related Distributions,* New-York, Chapman and Hall, 1990.

D. FRESNEAU, *Comportement social et structures des sociétés chez une fourmi ponérine: Pachycondyla Apicalis,* Thèse de doctorat d'état, Université Paris Nord, 1994.

J. B. S. HALDANE, The Precision of Observed Values of Small Frequencies, *Biometrika,* 1948, 35, p. 297-300.

H. JEFFREYS, *Theory of Probability,* 3rd ed., Oxford, Clarendon Press, 1961.

D. LÉPINE and H. ROUANET, Introduction aux Méthodes Fiduciaires : Inférence sur un Contraste entre Moyennes, *Cahiers de Psychologie,* 1975, 18, p. 193-218.

D. V. LINDLEY, The Bayesian Analysis of Contingency Tables, *The Annals of Mathematical Statistics,* 1965, 35, No. 4, p. 1622-1643.

J. E. MOSIMANN, On the Compound Multinomial Distribution, the Multivariate β-distribution, and Correlations Among Proportions, *Biometrika*, 1962, 49, No. 1-2, p. 65-82.

F. J. A. PERKS, Some Observations on Inverse Probability Including a New Indifference Rule (with discussion), *Journal of the Institute of Actuaries,* 1947, 73, p. 285-334.

H. ROUANET, Bayesian Methods for Assessing Importance of Effects, *Psychological Bulletin,* 1996, 119, No. 1, p. 149-158.

H. ROUANET and B. LECOUTRE, Specific Inference in ANOVA: from Significance Tests to Bayesian Procedures, *British Journal of Mathematical and Statistical Psychology,* 1983, 36, p. 252-268.

P. WALLEY, Inferences from Multinomial Data: Learning about a Bag of Marbles (with discussion), *Journal of the Royal Statistical Society,* 1996, Serie B, 58, No. 1, p. 3-57.

S. S. WILKS, *Mathematical Statistics,* New York, John Wiley.