Bayesian Methods for Censored Categorical Data

by

James M. Dickey, Jyh-Ming Jiang,

and

Joseph B. Kadane*

Technical Report No. 484

January 1987

---

*James M. Dickey is Professor, Department of Theoretical Statistics, University of Minnesota, Minneapolis, MN 55455.
Jyh-Ming Jiang is Assistant Professor, Mathematical Systems Program, Sangamon State University, Springfield, IL 62708.
Joseph B. Kadane is Leonard J. Savage, Professor of Statistics and Social Sciences at Carnegie-Mellon University, Pittsburgh, PA 15213.
Work by Professor Dickey sponsored by National Science Foundation, Grant DMS-8614793. Work by Professor Dickey and Professor Jiang sponsored by National Science Foundation, Grant MCS-8301335. Work by Professor Kadane sponsored by National Institutes of Justice, Grant 81-IJ-CX-0087, the office of Naval Research Contracts, N00014-82-K-0622 and N00014-85-K-0539, and the National Science Foundation, Grant DMS-8503019.

The authors would like to thank Morris L. Eaton and an Associate Editor for helpful suggestions and members of the NSF-NBER sponsered seminar for Bayesian Inference in Econometrics for useful discussion.

Abstract. Bayesian methods are given for finite-category sampling when some of the observations suffer missing category distinctions. Dickey's (1983) generalization of the Dirichlet family of prior distributions is found to be closed under such censored sampling. The posterior moments and predictive probabilities are proportional to ratios of B.C. Carlson's multiple hypergeometric functions. Closed-form expressions are developed for the case of nested reported sets, when Bayesian estimates can be computed easily from relative frequencies. Effective computational methods are also given in the general case. An example involving surveys of death-penalty attitudes is used, throughout, to illustrate the theory.

A simple special case of categorical missing data is a two-way contingency table with cross-classified count data x_{ij} ($i = 1, \ldots, r$, $j = 1, \ldots, c$), together with supplementary trials counted only in the margin distinguishing the rows, y_i ($i = 1, \ldots, r$). There could also be further supplementary trials reported only by counts distinguishing the columns, z_j ($j = 1, \ldots, c$). Under assumptions that the censoring process, itself, is "noninformative" regarding the category probabilities θ_{ij} (for example the report for each possible outcome might be nonrandom and prespecified), the Bayesian inference regarding the θ_{ij}'s would be based on the likelihood function

$$(\Pi_{ij}\theta_{ij}^{x_{ij}})\,[\Pi_i(\Sigma_j\theta_{ij})^{y_i}]\,[\Pi_j(\Sigma_i\theta_{ij})^{z_j}].$$

Such a likelihood is ordinarily considered intractable and unsuited for Bayesian conjugate prior inference. We develop a Bayesian conjugate theory, however, by recognizing the complete integrals of such functions as Carlson functions and

the posterior distributions resulting from Dirichlet prior distributions as known generalized Dirichlet distributions. The corresponding posterior density functions are similar in form to the likelihood, and these constitute a family of distributions closed under sampling and tractable in various senses, including the convenient computability of moments and modes.

KEY WORDS: Bayesian inference; Generalized Dirichlet distributions; Missing data; Multinomial model; Multiple hypergeometric functions.

1. Introduction.

Bayesian conjugate-prior inference is tractable and elegant for categorical sampling, that is, independent identical sampling from a distribution with finite support (Laplace 1774, Good 1950, 1965). If the Dirichlet distributions are used as the conjugate family of prior and posterior distributions, the densities, moments, and predictive probabilities take closed forms. Also, the class of mixtures of Dirichlet distributions with a small number of mixands is similarly tractable and large enough in many situations to offer realistic descriptions of predata subjective uncertainty. (For completeness results, see Diaconis and Ylvisaker 1985, and Dalal and Hall 1983.) However, in cases of inference from censored data, that is, when some of the observations suffer missing distinctions between categories, so mere sets of categories are sometimes reported, tractability appears to fade. The likelihood contains further factors that are powers of sums of probabilities over sets of categories not distinguished.

Bayesian treatments of categorical sampling with such missing data have been given by Karson and Wrobleski (1970), Antelman (1972), Kaufman and King (1973), Albert and Gupta (1983), Gunel (1984), Smith and Gunel (1984), Smith et al (1985) Albert (1985), and Kadane (1985). All these deal with 2x2 contingency tables with information missing regarding row or column variables. Basu and Pereira (1982) extend consideration to kx2 tables and summarize properties of the Dirichlet distribution under relevant changes of variable. In general cases, Shefrin (1981, 1983) develop expansions with respect to the possible values of the unreported data. (For treatments from the frequentist viewpoint, see for example, Hartley 1958, Chen and Fienberg 1974, 1976, and Dempster et al 1977.)

After first setting up requisite tools and notation (Section 2), we formulate the problem of Bayesian inference from a sequence of quite general reports, where the reports are based, respectively, on the individual trials of a categorical sample (Section 3). We give a representation of the relevant posterior complete integrals from a Dirichlet prior distibution in terms of B.C. Carlson's (1977) two-way multiple hypergeometric functions R. The posterior distributions are cases of the generalized Dirichlet distributions of Dickey (1983). This theory extends to a new family of prior distributions that is closed under sampling. The theory is then specialized to censored data, that is, reported sets of categories, and conditions are given under which the inferences are made independently regarding the censoring process and the sampling process. For the special case of nested reported sets, a parallel development based on changes of variable in the Dirichlet distribution yields an

elementary theory in which the Bayesian posterior estimates are computed easily from relative frequencies (Section 4). For the case of nonnested censoring, successful straightforward expansions and asymptotic methods of computation are available and are discussed briefly (Section 5). The theory will be illustrated throughout by an example of surveys of death-penalty attitudes.

2. Preliminary Tools and Notation. Sampling Process Observed.

To begin, we review and summarize known results and establish notation for the case of sample data reported in full and accurate detail. Consider a finite-category independent and identically distributed sampling sequence $k_1, k_2, \ldots$ having possible-outcome categories $k = 1,2,\ldots,K$ (say), with sampling probabilities, $\theta_k = \text{pr}(k_n = k|\underline{\theta})$ with probability vector $\underline{\theta} = (\theta_1,\ldots,\theta_K)$, for $n = 1,2,\ldots$. For a sample sequence segment $k_1,\ldots,k_N$, define the category frequency counts $\underline{x} = (x_1,\ldots,x_K)$, where each $x_k = \Sigma_{n=1}^{N} \delta_k(k_n)$, $k=1,\ldots,K$.

(We define the indicator $\delta_k(k) = 1$ and $\delta_k(j) = 0$ for $j \neq k$.) Thus $x_+ = N$, where we denote summation over an index by a plus sign.

Bayesian inference depends on the data only through the likelihood function. We assume that the sample size N is noninformative regarding $\underline{\theta}$ (Raiffa and Schlaifer 1961, sec. 2.3), so the likelihood function of $\underline{\theta}$ is proportional to a likelihood as if N were fixed before the sampling. The latter likelihood depends on the data $k_1,\ldots,k_N$ only through the sufficient statistic $\underline{x}$,

$$pr(k_1,\ldots,k_N|\underline{\theta}) = \Pi_{k=1}^{K} \theta_k^{x_k} \quad . \tag{2.1}$$

Of course if N is prespecified, $\underline{x}$ is multinomially distributed, and $pr(\underline{x}|\underline{\theta}) = \binom{N}{\underline{x}} pr(k_1,\ldots,k_N|\underline{\theta})$, where $\binom{N}{\underline{x}} = N!/\Pi(x_k!)$.

A conjugate family for the likelihood (2.1) is the Dirichlet. The random vector $\underline{\theta}$ is said to have the Dirichlet distribution, $\underline{\theta} \sim D(\underline{b})$, with parameter vector $\underline{b} = (b_1,\ldots,b_K)$, each $b_k > 0$, if $\underline{\theta}$ has the density $f(\underline{\theta};\underline{b})$, as follows. Let

$$B(\underline{b}) = [\Pi_1^K \Gamma(b_k)]/\Gamma(b_+).$$

Then define the probability density with respect to any K-1 of the coordinates of $\underline{\theta}$,

$$f(\underline{\theta};\underline{b}) = B(\underline{b})^{-1} \Pi_1^K \theta_k^{b_k-1} \tag{2.2}$$

for all $\underline{\theta}$ in the probability simplex $\{\underline{\theta}$: each $\theta_k \geq 0$, $\theta_+ = 1\}$. The posterior distribution, consequent to the prior distribution $\underline{\theta} \sim D(\underline{b})$ (2.2) and the likelihood function (2.1), is the Dirichlet with updated parameters, $\underline{\theta}|\underline{x} \sim D(\underline{b}+\underline{x})$, with density $f(\underline{\theta};\underline{b}+\underline{x})$.

The Bayesian predictive distributions express uncertainty concerning observables, taking account of the uncertainty concerning parameters $\underline{\theta}$ by averaging out $\underline{\theta}$ in the sampling probability, (2.1). For any prior distribution, the prior predictive probability mass is the same as a prior moment,

$$pr(k_1,\ldots,k_N) = E_\theta \prod \theta_k^{x_k},$$

where $\underline{x}$ is a value of the count variable, not a condition in the expectation. Denote the Dirichlet $\underline{c}$th moment by

$$\begin{aligned} h(\underline{c};\underline{b}) &= E_\theta \prod \theta_k^{c_k} \\ &= B(\underline{b}+\underline{c})/B(\underline{b}). \end{aligned} \tag{2.3}$$

Then the predictive probability is the $\underline{x}$th moment, $pr(k_1,\ldots,k_N) = h(\underline{x};\underline{b})$. In particular, the predictive probability for a single observation $pr(k_1 = k)$ is the same as the prior mean, that is, the $\underline{\delta}(k)$th moment: $pr(k_1=k) = E(\theta_k) = h(\underline{\delta}(k);\underline{b}) = b_k/b_+$, where each jth coordinate of $\underline{\delta}(k)$ is the indicator $\delta_j(k)$.

The posterior Dirichlet $\underline{c}$th moment is

$$\begin{aligned} E_{\theta|x}(\prod \theta_k^{c_k}) &= h(\underline{c};\underline{b}+\underline{x}) \\ &= B(\underline{b}+\underline{x}+\underline{c}) \,/B(\underline{b}+\underline{x}) \quad . \end{aligned} \tag{2.4}$$

So the posterior predictive probability of further data $k_1^*,\ldots,k_{N^*}^*$ given $k_1,\ldots,k_N$, is just the posterior $\underline{x}^*$th moment,

$$pr(k_1^*,\ldots,k_{N^*}^*|\underline{x}) = E_{\theta|x}(\Pi\, \theta_k^{x_k^*}) = h(\underline{x}^*;\underline{b}+\underline{x}).$$

In particular, for a single further observation,

$$pr(k_{N+1} = k|\underline{x}) = E(\theta_k|\underline{x}) = h(\underline{\delta}(k);\ \underline{b}+\underline{x}) = (b_k+x_k)/(b_+ +x_+),$$

the posterior Dirichlet mean. Note also that, by definition or by (2.4), $pr(k_1^*,\ldots,k_{N^*}^*|\underline{x}) = pr(k_1,\ldots,k_N,k_1^*,\ldots,k_{N^*}^*)/pr(k_1,\ldots,k_N)$, that is, $h(\underline{x}^*;\underline{b}+\underline{x}) = h(\underline{x}^*+\underline{x};\underline{b})/h(\underline{x};\underline{b})$.

3. Reporting-Processes.

3.1 General Reporting. Consider a single trial k_n in a categorical sampling process, followed by a report r_n of arbitrary form. This could be any statement made by a reporter or reporting process - not necessarily truthful, or even relevant. We shall model randomness in the value r_n conditional on k_n. (Dawid and Dickey (1977) took the alternative approach of assuming the truthful reporting of the value of a statistic, with the form of the statistic chosen at random, and the report to include the information of which statistic was chosen.)

Denote the conditional probability of the report $r_n = r$ given the outcome $k_n = k$ by $\lambda_{r|k}$ and the array of these parameters by Λ, with i,jth entry $\lambda_{r|k}$ for row $i = k$ and column $j = r$, $k = 1,\ldots,K$, and $r\in S$, where S is the set of possible values for r, assumed finite. Then our model-probability of receiving

the report r_n (k_n itself being unknown) is

$$pr(r_n=r|\underline{\theta},\Lambda) = \sum_{k=1}^{K} \lambda_{r|k}\,\theta_k \quad . \tag{3.1}$$

Both the parameters $\underline{\theta}$ and Λ could be considered unknown.

Denote a sequence of sampling-process outcomes by k_1, k_2,..., and their respective reports by r_1, r_2,... . The reporting process applies separately to each trial k_n, that is, we assume the pairs (k_n,r_n) are independent of each other and are identically distributed, n=1,2,... . Assume a sample of (known) size N, with N noninformative regarding $\underline{\theta}$ and Λ. Again, as far as inference concerning $\underline{\theta}$ or Λ is concerned, noninformative optional stopping is permitted and the analysis is equivalent to the case of fixed N. Define the frequency counts of the reports as $\underline{y} = (y_r: r\varepsilon S)$, where each

$$y_r = \sum_{n=1}^{N} \delta_r(r_n); \quad \text{so } y_+ = N.$$

Then $\underline{y}$ is a sufficient summary of the r_n's, and our likelihood function of $\underline{\theta}$ and Λ is

$$\begin{aligned} & pr(r_1,\ldots,r_N|\underline{\theta},\Lambda) \qquad (3.2) \\ & = pr(\underline{y}|\underline{\theta},\Lambda) \;/\; \binom{N}{\underline{y}} \\ & = \Pi_{r\varepsilon S}\left(\sum_{k=1}^{K} \lambda_{r|k}\,\theta_k\right)^{y_r} \quad . \end{aligned}$$

In practice, although we do not treat such data here, an arbitrary function of the vector $(r_1,\ldots,r_N)$ could be received.

Our interest is in inference concerning $\underline{\theta}$. We begin by describing inference from (3.2) conditional on Λ. We need a lemma, first, concerning the moments of linear forms in a Dirichlet vector.

Lemma 3.1 Let $\underline{d} = (d_1,\ldots,d_J)$ and $Z = (z_{ij})$, with each $z_{ij}\geq 0$, $i=1,,\ldots,K$, $j=1,\ldots,J$. Then a generalized moment of $\underline{\theta} \sim D(\underline{b})$ is given by

$$E_{\theta|b} \prod_{j=1}^{J}\left(\sum_{k=1}^{K} z_{kj}\,\theta_k\right)^{d_j} \tag{3.3}$$

$$= R(\underline{b},Z,-\underline{d}) \quad . \qquad \square$$

The right-hand side of (3.3) indicates B.C. Carlson's (1971, 1974) multiple-hypergeometric function $R_a(\underline{b},Z,-\underline{d})$ in the special case $a=d_+$. The functions R_a are a matrix-argument generalization of the classical vector-argument hypergeometric functions of Appell and Lauricella. If $a = d_+$ and each coordinate $b_k>0$, the function $R_a=R$ has the multiple-integral representation (3.3). Interesting properties of R include an identity changing the dimensionality of the integral from K-1 to J (Dickey 1968a) and special dimension-reductions for special argument matrices Z. The identity (3.3) ties our problems of Bayesian statistical inference for missing data to a mainstream theory of special functions. (In his 1977 monograph B.C. Carlson based a unified introduction to special functions on a vector-argument form of R.) Dickey (1983) gives an introduction to Carlson's functions for statisticians.

Theorem 3.2 Assume the data take the form of reports, generated with probabilities Λ, conditional on a categorical sampling process, having probabilities $\underline{\theta}$. The likelihood function (3.2) for report counts $\underline{y}$, and a Dirichlet conditional prior distribution $\underline{\theta}|\Lambda \sim D(\underline{b})$ give the following Bayesian inferences. The predictive probability mass $E_{\underline{\theta}|\Lambda} pr(r_1,\ldots,r_N|\underline{\theta},\Lambda)$ is

$$pr(r_1,\ldots,r_N|\Lambda) = R(\underline{b},\Lambda,-\underline{y}) \quad , \tag{3.4}$$

and the posterior distribution is

$$(\underline{\theta}|\underline{y};\Lambda) \sim D(\underline{b},\Lambda,\underline{y}) \quad , \tag{3.5}$$

where the notation $\underline{\theta} \sim D(\underline{b},Z,\underline{d})$, for a matrix Z of nonnegative entries, means that $\underline{\theta}$ has the density on the probability simplex, in terms of the Dirichlet $D(\underline{b})$ density $f(\underline{\theta};\underline{b})$ (2.2),

$$g(\underline{\theta};\underline{b},Z,\underline{d}) = \tag{3.6}$$
$$f(\underline{\theta};\underline{b}) \prod_{j=1}^{J} \left(\sum_{k=1}^{K} z_{kj}\theta_k \right)^{d_j}$$
$$/R(\underline{b},Z,-\underline{d}) \quad . \quad \square$$

Dickey (1983) introduced the class of distributions $D(\underline{b},Z,\underline{d})$ (3.6). The Dirichlet $D(\underline{b})$ is the special case $\underline{d} = \underline{0}$ Note the difference in sign here between the last parameter vector of the general class D and the corresponding Carlson's R (a departure from the notation in Dickey 1983).

By the form of its density, the distribution (3.6) has for its general $(\underline{b}^*,\underline{d}^*)$th moment,

$$E_{\theta|b,Z,d}\Big(\prod_{k=1}^{K}\theta_k^{b_k^*}\Big)\prod_{j=1}^{J^*}\Big(\sum_{k=1}^{K}z_{kj}^*\,\theta_k\Big)^{d_j^*} \tag{3.7}$$

$$= h(\underline{b}^*;\underline{b})\; R(\underline{b}+\underline{b}^*,\,(Z,Z^*),\,-(\underline{d},\underline{d}^*))$$

$$/\; R(\underline{b},Z,-\underline{d})\quad .$$

This will serve as a source of moment formulae. It has the form of the Dirichlet (monomial) moment multiplied by a ratio of Carlson functions. Note that in the special case of a common coefficients matrix, $Z^*=Z$, the numerator of the ratio becomes $R(\underline{b}+\underline{b}^*,Z,-(\underline{d}+\underline{d}^*))$.

Corollary 3.3 In the inference setting of Theorem 3.2, the posterior (monomial) $\underline{c}$th moment is a simple multiple of the corresponding prior moment, (2.3),

$$E\Big(\prod_{k=1}^{K}\theta_k^{c_k}|\underline{y};\Lambda\Big) = \tag{3.8}$$

$$E\Big(\prod_{k=1}^{K}\theta_k^{c_k}|\Lambda\Big)\cdot R(\underline{b}+\underline{c},\,\Lambda,-\underline{y})/R(\underline{b},\Lambda,-y)\quad .\ \square$$

3.2 Censored Data. Consider now the case of sampling from a finitely supported distribution where reports of the outcomes are censored, so that

distinctions between the categories are sometimes missing and mere sets of categories are reported. For a trial k_n, let the corresponding report be a nonempty subset $r_n = s$, $s \subset \{1,\ldots,K\}$. Let the class S contain all subsets s having positive predictive probability of appearing in a report: if $s \notin S$ then $\lambda_{s|k}=0$, for all k, with prior probability one, and $y_s=0$. (It is assumed that the count $y_s=0$ when, with prior certainty, $\lambda_{s|k} = 0$ for all k.) We make the following assumptions concerning the censoring process $\lambda_{s|k}$.

(i) Reporting is truthful:

$$\lambda_{s|k}=0 \text{ whenever } k \notin s. \tag{3.9}$$

(ii) Every report $r=s$ is differentially noninformative among the categories within s:

$$\lambda_{s|k}=\lambda_{s|k'}=\lambda_s \text{ (say) whenever both } k \in s \text{ and } k' \in s. \tag{3.10}$$

This means we have a "report-based" reporting process, in the terminology of Dawid and Dickey (1977). This is related to Rubin's (1976) concept of "missing at random", in which the sampled distribution has a multivariate structure and variables are missed in some of the draws.

Assumptions (i) and (ii) for the model have the following simple consequence on each trial. The conditional probability of the report $r_n=s$ is $pr(r_n=s|k,\Lambda) = \lambda_s$ for all $k \in s$ (and 0 otherwise), and the probability of the

occurrence of s is the sum $pr(k_n \varepsilon s|\underline{\theta}) = \Sigma_{k\varepsilon s}\theta_k$. Thus, by (3.1), the unconditional probability of the report is just the product, $pr(r_n=s|\underline{\theta},\Lambda) = \Sigma_{k\varepsilon s}\lambda_s\theta_k = \lambda_s(\Sigma_{k\varepsilon s}\theta_k) = pr(r_n=s|(k_n\varepsilon s),\Lambda)\cdot pr(k_n\varepsilon s|\theta)$. This decomposes the likelihood (3.2) into separate factors pertaining to the censoring process and the sampling process,

$$pr(r_1,\ldots,r_N|\underline{\theta},\Lambda) \tag{3.11}$$

$$= pr(r_1,\ldots,r_N|s_1,\ldots,s_N,\Lambda) \cdot pr(s_1,\ldots,s_N|\underline{\theta})$$

$$= (\Pi_{s\varepsilon S}\,\lambda_s^{y_s})\;\Pi_{s\varepsilon S}\,(\Sigma_{k\varepsilon s}\,\theta_k)^{y_s},$$

where each $r_n=s_n$, $n=1,\ldots,N$. (As usual, $0^0=1$.)

Our final assumption concerns the uncertainty about the modeled processes.

(iii) Assume <u>prior independence</u> between the two parameter arrays $\underline{\theta}$ and Λ. (Known Λ is a special case.)

This, together with the factorization (3.11), implies posterior independence between $\underline{\theta}$ and Λ. So, the assumptions (i)-(iii) imply that the censoring process, per se, is noninformative regarding $\underline{\theta}$, in the sense that inference about $\underline{\theta}$ can be carried out marginally without conditioning on Λ. Under these assumptions, one can just use the likelihood of $\underline{\theta}$ based directly on the reported

outcome events,

$$\text{pr}(s_1,\ldots,s_N|\underline{\theta}) = \Pi_{s\in S}(\textstyle\sum_{k\in s}\theta_k)^{y_s} \quad . \tag{3.12}$$

It is as if each report $r_n=s_n$ were an observation on an ordinary two-category Bernoulli random variable, as if no other set could have been reported than s_n or its complement. (These would be independent, but not identically distributed, Bernoullis.) The following theorem gives the inferential distribution theory.

Theorem 3.4 Given categorical sampling with noninformative censoring (assumptions (i)-(iii) and hence likelihood (3.12)), the Dirichlet prior distribution $\underline{\theta} \sim D(\underline{b})$ gives the extended Dirichlet posterior distribution

$$\underline{\theta}|\underline{y} \sim D(\underline{b},Z_S,\underline{y}) \quad , \tag{3.13}$$

where the columns of z_S indicate the sets of the (ordered) class S, the k,sth entry of Z_S being

$$z_{ks} = \begin{cases} 1 \text{ if } k\in s \\ 0 \text{ otherwise,} \end{cases} \tag{3.14}$$

for $k=1,\ldots,K$, $s\in S$. □

This is related to our earlier general-report Theorem 3.2, but in the case of noninformative censoring the conditional reporting probabilities Λ need not

be known. Again, as in (3.8), the posterior monomial moment is a simple multiple of the corresponding prior moment, the posterior mean being given by

$$E(\theta_k|\underline{y}) = (E\theta_k)\, R(\underline{b}+\underline{\delta}(k),Z_S,-\underline{y}) \Big/ R(\underline{b},Z_S,-\underline{y}) \quad , \tag{3.15}$$

where $E\theta_k = b_k/b_+$.

As is evident by (3.13), the original Dirichlet family is not closed under censored sampling. However, the extended family (3.6) does have this property when the columns of Z are defined to include those of Z_S.

Theorem 3.5 Consider the extended Dirichlet family as prior distributions for noninformative censoring, $\underline{\theta} \sim D(\underline{b},Z,\underline{d})$, where $Z=(Z_S,\tilde{Z})$. Such a distribution yields the posterior distribution in the same family,

$$\underline{\theta}|\underline{y} \sim D(\underline{b},Z,\underline{d}+\underline{e}) \quad , \tag{3.16}$$

where $\underline{e} = (\underline{y},\underline{0})$. □

In cases when the sampling probability of a set of categories s is better known prior to the current data analysis than are its component category probabilities, the extended family of distributions can more accurately express one's prior uncertainty than can the Dirichlet. One can include in the prior

density a factor

$$(\sum_{k\epsilon s} \theta_k)^{d_1}(1-\sum_{k\epsilon s} \theta_k)^{d_2}$$

with large d_1+d_2 when $\sum_{k\epsilon s}\theta_k$ is believed near in value to $d_1/(d_1+d_2)$. Like the Dirichlet, the new family is not expressive for situations of local smoothness, when close categories are believed to have close probabilities (as in Dickey 1968, 1969).

3.3 Restrictions. No assumption has been made that the class of reported sets S is a partition, nor has S been restricted in any other way. It is clear, however, that (i) and (ii) themselves place restrictions on S, because of the following property. Define for each category $k=1,\ldots,K$, $S_k = \{s: k\epsilon s, s\epsilon S\}$. Then for each k,

$$\sum_{s\epsilon S_k} \lambda_s = 1 \quad . \tag{3.17}$$

So for any sample outcome k, the conditional reporting probabilities of all the sets containing k must sum to unity, and these probabilities depend only on the sets. This property rules out some configurations of partially overlapping sets. For example, it rules out the following simple case. Suppose there are three categories, $k = 1,2,3$, and two sets, $S=\{s,t\}$ with $s=\{1,2\}$ and $t=\{2,3\}$. Then (3.17) would require $\lambda_s=1$, $\lambda_s+\lambda_t=1$, and $\lambda_t=1$.

Clearly, if S is a partition, such difficulties cannot arise. More generally, a collection S combining successively nested partitions, in which any

two sets are either disjoint or one a subset of the other, will also be consistent with (3.17). These form the subject of our next section.

Note, however, that in practice, data containing reports with partially overlapping sets are not uncommon. For example, in two-way contingency tables, supplemental marginal row counts and further supplemental marginal column counts involve partially overlapping sets. Many cases of data reporting partially overlapping sets can be treated naturally from the theory by considering distinct portions of the data to be reports by different censoring processes of different samples from the same distribution. Then the overall combined likelihood is a product of likelihoods corresponding to the different portions of the data, for each of which the assumptions (i), (ii) may be satisfied. The censoring process is then independent but not identical from trial to trial. The following example is of this type.

3.4 Death Penalty Attitudes. Combining Data from Different Questionnaires. Kadane (1983) analyzed data from two sample surveys of potential juror's attitudes, in which respondents were instructed to assume the availability of a death penalty. The primary categories k are, with K=4:

1. Would not decide guilt versus innocence in a fair and impartial manner.
2. Fair and impartial on guilt versus innocence; and when sentencing, would ALWAYS vote for the death penalty, regardless of circumstance.
3. Fair and impartial on guilt; and when sentencing would NEVER vote for the death penalty.
4. Fair and impartial on guilt; and when sentencing would SOMETIMES AND

SOMETIMES NOT vote for the death penalty.

A survey by the Field Research Corporation produced the frequency data $y_{\{1\}}=68$, $y_{\{3\}}=97$, $y_{\{2,4\}}=674$, ($y_+=839$), and a survey by Harris yielded $y_{\{2\}}=15$, $y_{\{1,3,4\}}=1484$ ($y_+=1499$). (Kadane reduced the effective sample size to account for a dependence or "clustering" in the way the data were gathered, a complication we do not consider here.) Since the censoring is nonstochastic and imposed by the forms of the questionnaires, the censoring process itself is noninformative.

Assuming these are two censored multinomial samples with common underlying categories, one multiplies the two likelihoods to obtain a likelihood whose data is the counts obtained by pooling the two sets of counts. The combined likelihood is

$$(\Pi_1^4 \theta_k^{x_k})(\theta_2+\theta_4)^{y_{\{2,4\}}}(\theta_1+\theta_3+\theta_4)^{y_{\{1,3,4\}}}, \qquad (3.18)$$

in which $\underline{x}=(68, 15, 97, 0)$, $y_{\{2,4\}}=674$, and $y_{\{1,3,4\}}=1484$ ($x_+ + y_+ = 2338$). We have denoted the combined reported counts of singleton sets by $x_k=y_{\{k\}}$.

A thorough Bayesian analysis would report the coherent effect of the data on a range of prior distributions expressing a range of expert opinions prior to knowledge of the data. To simulate aspects of such a process, we assessed a Dirichlet distribution by eliciting the opinion of a single social psychologist with interests in legal matters and a familiarity with previous studies, but not yet familiar with the two surveys under discussion. Elicitation of his prior

means b_k/b_+ and elicitation of b_+ by a version of I.J. Good's (1950, p.35) "device of imaginary results" yielded the following Dirichlet parameter vector, in which $b_+=140$.

Table I. Expert Dirichlet Prior Parameters.

b_k/b_+	0.02	0.08	0.15	0.75
b_k	2.8	11.2	21.0	105.0

The corresponding expert posterior distribution ($D(\underline{b}+\underline{x},Z,\underline{y})$ has the density $g(\underline{\theta};\underline{b}+\underline{x},Z,\underline{y})$ (3.6),

$$[\Pi_1^4 \theta_k^{b_k+x_k-1}/B(\underline{b}+\underline{x})](\theta_2+\theta_4)^{y\{2,4\}}(\theta_1+\theta_3+\theta_4)^{y\{1,3,4\}}/R(\underline{b}+\underline{x},Z,-\underline{y}), \qquad (3.19)$$

where $\underline{b}+\underline{x}$ = (70.8, 26.2, 118.0, 105.0), y = (674, 1484), and

$$Z = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} . \qquad (3.20)$$

The uniform distribution, $\theta \sim D(\underline{1})$, where $\underline{1}=(1,\ldots,1)$, is frequently proposed as a "noninformative" prior distribution. This leads to a posterior density proportional to the likelihood function (3.18), itself, namely $g(\underline{\theta};\underline{1}+\underline{x},Z,\underline{y})$.

4. Partitions and Nested Censoring.

Inference is considerably simplified in cases where the censoring is nested, that is, where S is a union of nested partitions, so that for every pair of sets of confused categories, either one set is a subset of the other or they are disjoint. We begin with the case where a partition of the sampling domain is sometimes reported. Many of the results extend to more general nested censoring.

4.1 Grouped Date: Observing a Partition. It is well known that the Dirichlet form of inference is preserved by data-grouping. A Dirichlet prior distribution for the detailed category probabilities induces a Dirichlet marginal prior distribution for the probabilities for a partition. Data grouped according to the partition will then give a Dirichlet posterior distribution for these partition probabilities; and this posterior distribution would be the same as the marginal distribution from a posterior distribution based on any ungrouped (but actually unavailable) version of the data. Below, we extend this property to the generalized Dirichlet family.

Anderson (1957) finds in the normal case that if multivariate data has values of the variables missing in a nested pattern, the sampling model can be reparameterized according to successive conditional distributions of the levels of missing variables, so that the likelihood function factorizes into functions of the respective new parameters. Our problem could be treated through a general Anderson scheme by reformulating our sampling model of outcome k_n,

$n=1,\ldots,N$, to have an equivalent K-variate indicator-type outcome $\underline{\delta}(k_n)$, $n=1,\ldots,N$, where vector function $\underline{\delta}(k_n)$ has coordinates $\delta_k(k_n)=1$ if $k_n=k$, and 0 if $k_n \neq k$. Then the censored categorical data could be viewed as multivariate data with missing values of coordinate variables, and in the nested censoring case the likelihood would be seen to factorize usefully. We shall achieve essentially this same factorization.

Consider the case of data $\underline{y}$ reporting the counts of a partition U of $\{1,\ldots,K\}$. Without loss of tractability, we can also consider further fully detailed data $\underline{x}$ counting additional occurences of the categories, per se. Then $S = U \cup \{1,\ldots,K\}$ and a (noninformative) random censoring mechanism could allocate a trial to be counted either in $\underline{y}$ or in $\underline{x}$. For each set s_j of $U = (s_j: j=1,\ldots,J)$ let $u_j(\underline{\theta})$ be the sampling probability of s_j; so

$$u_j(\underline{\theta}) = \sum_{k \in s_j} \theta_k, \qquad j=1,\ldots,J. \tag{4.1a}$$

Let $t_{k|j}(\underline{\theta})$, for $k=1,\ldots,K$, be the conditional sampling probability of category k given s_j; so

$$t_{k|j}(\underline{\theta}) = \begin{cases} \theta_k/u_j(\underline{\theta}), & \text{if } k \in s_j \\ 0, & \text{otherwise,} \end{cases} \tag{4.1b}$$

$k=1,\ldots,K$; $j=1,\ldots,J$. In terms of the partition-indicator matrix Z_U (KxJ) whose columns indicate (as in (3.14)) the sets s_j listed in U, we can write the vector form of (4.1a),

$$\underline{u}(\underline{\theta}) = \underline{\theta}\, Z_U\,, \tag{4.2a}$$

where $\underline{u}(\underline{\theta})$ and $\underline{\theta}$ denote the obvious row vectors. We also have the vector form of (4.1b),

$$\underline{t}_{*|j}(\underline{\theta}) = \underline{\theta}_{(s_j)}/u_j(\underline{\theta}), \tag{4.2b}$$

where the vector $\underline{\theta}_{(s_j)} = (\theta_k$ if $k\varepsilon s_j$, 0 otherwise: $k=1,\ldots,K)$, for each $j=1,\ldots,J$. (Categories k not in the set s_j correspond to zero coordinates of $\underline{t}_{*|j}(\underline{\theta})$.) In the special case of multinomial data (fixed y_+, x_+ and nonrandom censoring), $\underline{u}(\underline{\theta})$ and $\underline{t}_{*|*}(\underline{\theta})$ would be the respective parameters of marginal and conditional multinomial distributions,

$$\underline{y}+\underline{u}(\underline{x}) \sim \text{Multinomial}\ (\underline{u}(\underline{\theta}),\ y_+ + x_+) \tag{4.3}$$

$$\underline{x}_{(s_j)}|u_j(\underline{x}) \sim \text{Multinomial}\ (\underline{t}_{*|j}(\underline{\theta}),\ u_j(\underline{x})), \tag{4.4}$$

independently for $j=1,\ldots,J$, where each $\underline{x}_{(s_j)}=(x_k$ if $k\varepsilon s_j$, 0 otherwise: $k=1,\ldots,K)$, and where $\underline{u}(\underline{x}) = \underline{x}\, Z_U$, the same function as (4.1a), (4.2a), but with $\underline{\theta}$ replaced by $\underline{x}$.

Lemma 4.1 (Factorized Likelihood). Under noninformative stopping, noninformative censoring, and censoring by a partition, the likelihood

factorizes with respect to the two types of data and parameters,

$$(\Pi_k \theta_k^{x_k}) \ \Pi_j \ u_j(\underline{\theta})^{y_j} = \tag{4.5}$$

$$\Pi_j u_j(\underline{\theta})^{y_j+u_j(\underline{x})} \Pi_{k\epsilon s_j} t_{k|j}(\underline{\theta})^{x_k} \quad . \quad \square$$

This factorization holds without assuming multinomially distributed count statistics. It is central to the nested-censoring theory. For now, we merely note that it shows that the maximum likelihood estimate of $\underline{\theta}$ is the same as the value from the multinomial-model method of moments, and it can be obtained by just setting the transformed parameters $\underline{u}(\underline{\theta})$ and $\underline{t}_{*|*}(\underline{\theta})$ equal to their corresponding sample fractions,

$$\underline{u}(\hat{\underline{\theta}}) = (\underline{y}+\underline{u}(\underline{x}))/(y_+ + u_+) \ , \tag{4.6}$$

$$\underline{t}_{*|j}(\hat{\underline{\theta}}) = \underline{t}_{*|j}(\underline{x}), \qquad j=1,\ldots,J, \tag{4.7}$$

where $\underline{u}(\underline{x})$ and $\underline{t}_{*|*}(\underline{x})$ are given by (4.1ab), (4.2ab) with $\underline{\theta}$ replaced by $\underline{x}$.

We turn now to prior distributions and their use in the inference. First, what distribution of the new parameters is implied by a Dirichlet distribution for $\underline{\theta}$?

Lemma 4.2. (Version of Wilks 1962, pp. 180-181.) If the random probability vector $\underline{\theta}$ has the Dirichlet distribution $\underline{\theta} \sim D(\underline{b})$, then the partition-grouped probabilities and associated conditional probabilities have the independent Dirichlet distributions, $\underline{u}(\underline{\theta}) \sim D(\underline{u}(\underline{b}))$ and $\underline{t}_{*|j}(\underline{\theta}) \sim D(\underline{b}_{(s_j)})$, $j=1,\ldots,J$, where $\underline{b}_{(s_j)}=(b_k$ if $k\varepsilon s_j$, 0 otherwise: $k=1,\ldots,K)$. □

Theorem 4.3. For a partition indicator matrix Z_U, the generalized-Dirichlet vector $\underline{\theta} \sim D(\underline{b},Z_U,\underline{d})$ is expressible in terms of independent Dirichlet vectors under the transformation (4.1ab), (4.2ab),

$$\underline{u}(\underline{\theta}) \sim D(\underline{u}(\underline{b})+\underline{d}) \tag{4.8}$$

$$\underline{t}_{*|j}(\underline{\theta}) \sim D(\underline{b}_{(s_j)}) \ , \tag{4.9}$$

$j=1,\ldots,J$. □

A consequent simple form of R for partitions appears to be new in the special-functions literature.

Corollary 4.4. For a partition indicator Z_U,

$$R(\underline{b},Z_U,-\underline{d}) = h(\underline{d};\underline{u}(\underline{b})) \tag{4.10}$$

$$= B[\underline{u}(\underline{b})+\underline{d}]/B[\underline{u}(\underline{b})] \ . \quad \square$$

Corollary 4.5. (Moments.) Given partition indicator Z_U, the generalized-Dirichlet distribution $\theta \sim D(\underline{b},Z_U,\underline{d})$ has a closed-form general $(\underline{c},\underline{e})$th moment in product form,

$$E[(\Pi\theta_k^{c_k})\ \Pi u_j(\underline{\theta})^{e_j}] \tag{4.11}$$

$$= E[\Pi u_j(\underline{\theta})^{u_j(\underline{c})+e_j}]\ \Pi_j[E\Pi_{k\varepsilon s_j} t_{k|j}(\underline{\theta})^{c_k}]$$

$$= h(\underline{u}(\underline{c})+\underline{e};\ \underline{u}(\underline{b})+\underline{d})\cdot\Pi_j h(\underline{c}_{(s_j)};\ \underline{b}_{(s_j)})\ .\ \square$$

Theorem 4.6 (Inference.) The frequency data $\underline{x},\underline{y}$, with $\underline{y}$ censored by the partition U, and the generalized-Dirichlet prior distribution $\underline{\theta} \sim D(\underline{b},Z_U,\underline{d})$, in which Z_U indicates the same partition U, yield the posterior distribution in the same family,

$$\underline{\theta}|\underline{x},\underline{y} \sim D(\underline{b}+\underline{x},Z_U,\underline{d}+\underline{y})\ . \tag{4.12}$$

This is equivalent to the independent posterior Dirichlet distributions,

$$\underline{u}(\underline{\theta})|\underline{x},\underline{y} \sim D(\underline{u}(\underline{b}+\underline{x}) + (\underline{d}+\underline{y}))\ , \tag{4.13}$$

$$\underline{t}_{*|j}(\underline{\theta})|\underline{x} \sim D(\underline{b}_{(s_j)}+\underline{x}_{(s_j)}) \quad , \tag{4.14}$$

$j=1,\ldots,J.$ □

Theorem 4.7 (Three Estimates.) In cases of censoring by a partition, (I) the posterior mean of $\underline{\theta}$ and (II) the posterior mode of $\underline{\theta}$ are given by

$$\hat{\theta}_k = u_j(\hat{\underline{\theta}})\cdot t_{k|j}(\hat{\underline{\theta}}), \qquad k\varepsilon s_j, \tag{4.15}$$

$k=1,\ldots,K$, with factors satisfying equations (4.6), (4.7), under the respective replacements in both equations,

$$\text{(I)} \qquad (\underline{y},\underline{x}) \leftarrow (\underline{y}+\underline{d},\underline{x}+\underline{b}) \tag{4.16}$$

or

$$\text{(II)} \qquad (\underline{y},\underline{x}) \leftarrow (\underline{y}+\underline{d},\ \underline{x}+\underline{b}-\underline{1}_K) \quad , \tag{4.17}$$

$\underline{1}_K = (1,\ldots,1)$. (III) A third estimate $\underline{\theta}$ is constructed by (4.15) from the mode of the joint posterior distribution of $\underline{u}(\underline{\theta})$, $\underline{t}_{*|*}(\underline{\theta})$. This mode is given by equation (4.6) with the replacement.

$$\text{(III)} \qquad (4.6)\ (\underline{y},\underline{x}) \leftarrow (\underline{y}+\underline{d}-\underline{1}_J,\ \underline{x}+\underline{b}) \quad , \tag{4.18}$$

and equation (4.7) with the replacement,

$$\text{(III)} \qquad (4.7)\ \underline{x} \leftarrow \underline{x}+\underline{b}-\underline{1}_K \ . \ \square \qquad (4.19)$$

The estimates (I), (II), (III) will be approximately the same for large sample sizes y_+, x_+.

4.2 General Nested Censoring. The censoring in data from a multiple-choice questionnaire is nested if there is a subset of the questions that each respondent can choose to answer or to ignore as a whole set. Or, suppose the questions are presented in the same order to each respondent, who chooses some arbitrary place to stop answering, such as how far he/she gets before time runs out. As long as each respondent answers all questions in order until stopping, the censoring will be nested. However, unconstrained decisions on whether to answer different questions would lead to data having non-nested censoring.

In practice, the condition of differential noninformativeness of the censoring seems likely to be seriously violated for examination questionnaires, even if the censoring is nested. Condition (3.10) would claim that how well a student ansers questions is unrelated to how many questions the student answers. This seems overly restrictive in the context of grading student exams.

Nested censoring includes contingency tables having supplementary purely marginal counts for one of the variables. It would rule out data having supplementary marginal counts for more than one variable. It would also rule

out situations where a trial is simultaneously counted for reporting in supplemental marginal counts of more than one variable.

It is easily seen that data showing nested censoring can be organized according to a tree of successively nested partitions and successively conditional data. That is, the likelihood function can be rewritten as in (4.5), as proportional to a likelihood for a multinomial process and conditional multinomial processes, whose data first give the total counts for a crudely grouped partition of outcome categories, then the breakdown of some of these counts according to refinements of groups, then further a breakdown of some of these refined counts, and so on. Theorem 4.3 and its consequences, as stated for a single partition, are generalized easily to a tree of successively nested partitions.

4.3 Example (Cont.) Although the censoring is not nested in our example involving the combined data from two questionnaires, each survey can be analyzed separately as a case of observing a partition (Section 4.1). When our Dirichlet prior distribution is combined with either survey's likelihood function, a posterior distribution is obtained in the form of Theorems 4.3 and 4.6, $\underline{\theta}|\underline{x},\underline{y} \sim D(\underline{b}+\underline{x},Z,\underline{y})$ where $\underline{b}$ is the Dirichlet prior parameter, $\underline{x}$, $\underline{y}$ are the data reported in the particular single survey, and Z is the partition indicator corresponding to $\underline{y}$. For the Field Research Corp. survey alone, $\underline{b}+\underline{x} = (b_1+68, b_2, b_3+97, b_4)$,

$$Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad (4.24)$$

and $\underline{y} = (0, 674)$. The posterior distribution then has the representation, $(\theta_1+\theta_3, \theta_2+\theta_4) \sim \text{Beta}(b_1+b_3+165,\ b_2+b_4+674)$, $\theta_1/(\theta_1+\theta_3) \sim \text{Beta}(b_1+68,\ b_3+97)$, and $\theta_2/(\theta_2+\theta_4) \sim \text{Beta}(b_2, b_4)$, all independent. This yields the posterior estimates in Table II. Our posterior mode of $\underline{\theta}$ follows from Theorem 4.7 (II), using the total exponents in the joint posterior density, for example, $\hat{\theta}_2=[(b_2+b_4+674-2)/(b_+ +839-4)]. \ [(b_2-1)/(b_2+b_4-2)]$. The posterior mean calculation by Theorem 4.7 (I) uses the posterior exponents similarly, while merely omitting the negative integer terms.

Table II. Posterior estimates from survey by Field Research Corp. (by Theorem 4.7)

Category k	1	2	3	4
Expert Prior				
Mode (II)	0.072	0.072	0.120	0.736
Mean (I)	0.072	0.078	0.121	0.729
(Standard Deviation)	(0.008)	(0.022)	(0.010)	(0.025)
Uniform Prior				
Mean (I)	0.082	0.401	0.116	0.401

Note that the uniform prior distribution gives a posterior distribution exhibiting through its mean an unreasonable judgement of exchangeability between θ_2 and θ_4. Also, this prior distribution does not produce a mode in its

posterior distribution, since for merely one survey the likelihood is underidentified and there is no unique maximum likelihood estimate. Even the expert prior distribution leads to the appreciable posterior correlation for the Field survey alone, $\mathrm{Corr}(\theta_2,\theta_4) = -0.86$.

A parallel analysis of the Harris survey yields the values given in Table III. A combined analysis appears later below.

Table III. Posterior estimates from survey by Harris.

Category k	1	2	3	4
Expert Prior				
Mode (II)	0.014	0.015	0.157	0.814
Mean (I)	0.021	0.016	0.160	0.802
(Standard Deviation)	(0.013)	(0.003)	(0.032)	(0.034)
Uniform Prior				
Mean (I)	0.330	0.011	0.330	0.330

5. Calculation Methods.

In general, the density (3.6) of our posterior distribution for censored data, $\underline{\theta}|x,y \sim D(\underline{b}+\underline{x},Z_S,\underline{y})$, takes the form of a likelihood function for the "combined sample" $\underline{x}+\underline{b}-\underline{1}$, $\underline{y}$. Hence any maximum likelihood algorithm for categorical missing data can also be used to produce a posterior mode. As is well known and easily shown (for example, see the treatment of a similar model in Vardi et al 1986 Sec. 2.1), multinomial missing data tend to have a well

behaved likelihood, log-concave with a unique point of maximum at a single stationary point. Thus, many algorithms work well. We mention, in particular, the convenient E-M algorithm of Hartley (1958) and Dempster et al (1977).

In nonnested situations, an approximate maximum point can be obtained by factorizing the posterior density into a product of functions resembling likelihoods, as if there were independent subsamples in each of which the censoring pattern is nested. Each such subsample yields a linear system in $\underline{\theta}$ of the form (4.6), (4.7), and together these can be treated as a combined linear system that one solves by least squares, or by weighted least squares with weights proportional to the subsample sizes. Since a sample can be broken up into subsamples with apparent nested censoring in a variety of ways, this linear approach does not give a unique estimate. However, it does provide crudely informative summary statistics and a starting value for iterative maximizing routines.

The posterior mean and other moments require computation of ratios of Carlson multiple integrals. Numerical quadrature and Monte Carlo methods are inconvenient when the dimension is even moderate. But two other methods are highly effective: multinomial expansions of the integrand and Laplace's integral method. The former is exact for any nonnegative integer values of the exponents, but the number of terms grows as a product of powers of those exponents. One version of the expansion method can be found in Jiang (1984), and another in Kadane (1985). Laplace's method (see Tierney and Kadane 1986) is approximate, uses the point of maximum, and becomes more accurate as the sizes of the exponents increase. Our experience with these methods suggests the

existence of a broad middle range of exponent values in which both the expansion method is not too burdensome computationally and the Laplace method is quite accurate.

Example (Cont.) In our survey example, combining the likelihoods from the two surveys, the posterior density (3.19) has modes as given in Table IV. The posterior mode from the uniform prior is the same as the maximum likelihood estimate. (This is also the same as the linear-system "approximate" mode from the uniform prior, because the combined linear system happens to have an exact solution.)

Table IV also gives the posterior moments from the combined survey. To the accuracy reported here, the multinomial expansion method and the Laplace method delivered the same numerical values for posterior moments. The two prior distributions, however, differed noticeably in the posterior estimates they produced. Notice that in the case of the first category probability θ_1, the posterior means from the two prior distributions are located in opposite directions from the maximum likelihood estimate, the simple frequency ratio $\hat{\theta}_1 = 0.081$. This seems to be an unreasonable aspect of the uniform prior, in this example.

Table IV. Posterior estimates from combined data, surveys by Field Research and Harris Survey.

Category k	1	2	3	4
Expert Prior				
Approx. Mode (II)	0.077	0.014	0.117	0.792
Mode	0.072	0.015	0.121	0.791
Mean	0.073	0.016	0.122	0.789
(Standard Deviation)	(0.008)	(0.003)	(0.010)	(0.013)
Uniform Prior				
Mode	0.081	0.010	0.116	0.793
Mean	0.082	0.011	0.116	0.791
(Standard Deviation)	(0.009)	(0.003)	(0.011)	(0.014)

6. Summary

This paper shows that a Bayesian analysis of noninformatively censored categorical data reduces to elementary computations from Dirichlet posterior distributions, in nested-censoring cases, and to generalized Dirichlet distributions and the computation of specially restricted Carlson functions and their ratios, in more general cases. Since these computations are quite tractable, such analyses are now feasible.

REFERENCES

Albert, J.H., and Gupta, A.K. (1983). Bayesian estimation methods for 2x2 contingency tables using mixtures of Dirichlet distributions. J. Amer. Statist. Assoc., Vol 78, 708-717.

Albert, James H. (1985). Bayesian estimation methods for incomplete two-way contingency tables using prior belief of association. Bayesian Statistics 2. Ed. by J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith. North-Holland, Amsterdam, 589-602.

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Amer. Statist. Assoc., Vol. 52, 200-203.

Antelman, Gordon R. (1972). Interrelated Bernoulli processes. J. Amer. Statist. Assoc. Vol. 67, 831-841.

Basu, D., and Pereira, Carlos A. de B. (1982). On the Bayesian analysis of categorical data: The problem of nonresponse. J. of Statistical Planning and Inference, Vol. 6, 345-362.

Carlson, B.C. (1971). Appell functions and multiple averages. S.I.A.M. J. of Math. Anal. Vol. 2, No. 3, (Aug '71), 420-430.

Carlson, B.C. (1974). Inequalities for Jacobi polynomials and Dirichlet averages. S.I.A.M. J. of Math. Anal. Vol. 5, No. 4, (Aug '74), 586-596.

Carlson, B.C. (1977). Special Functions of Applied Mathematics. Academic Press, New York.

Cnen, T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. Biometrics, Vol. 30, 629-642.

Chen, T. and Fienberg, S.E. (1976). The analysis of contingency tables with incompletely classified data. Biometrics, Vol. 32, 133-144.

Dalal, S.R. and Hall, W.J. (1983). Approximating priors by mixtures of natural conjugate priors. J. Roy. Statist. Soc., B, Vol. 45, pp. 278-286.

Dawid, A.P. and Dickey, James M. (1977). Likelihood and Bayesian inference from selectively reported data. Journal of the American Statistical Association, Vol. 72, 845-850.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). J. Roy. Statist. Soc., Ser. B, Vol. 39, 1-38.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. Bayesian Statistics 2. North-Holland, Amsterdam, 133-156.

Dickey, James, M. (1968a). Three multidimensional-integral identities with Baysian applications. The Annals of Mathematical Statistics, Vol. 39, 1615-27.

Dickey, James M. (1968b). Smoothed estimates for multinomial cell probabilities. Annals of Mathematical Statistics, Vol. 39, 561-566.

Dickey, James M. (1969). Smoothing by cheating. Annals of Mathematical Statistics, Vol. 40, 1477-1482.

Dickey, James M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. J. Amer. Statist. Assoc., Vol. 78, 628-637.

Good, I.J. (1950). Probability and the Weighing of Evidence. Hafner, New York.

Good, I.J. (1965). The Estimation of Probabilities. M.I.T. Press, Cambridge, Mass.

Gunel, Erdogan (1984). A Bayesian analysis of the multinomial model for a dichotomous response with nonrespondents. Commun. Statist.--Theor. Meth. Vol. 13(6), 737-751.

Hartley, (1958). Maximum likelihood estimation from incomplete data. Biometrics. Vol. 14, 174-194.

Jiang, Jyh-Ming (1984). Distributional properties of linear forms in a Dirichlet vector and applications. Ph.D. dissertation (Tech. Rept. 7/84-#14). Dept. of Mathematics and Statistics, State Univ. of N.Y. at Albany.

Kadane, Joseph B. (1983). Juries hearing death penalty cases: Statistical analysis of a legal procedure. J. Amer. Statist. Assoc., Vol. 78, 544-552.

Kadane, Joseph B. (1985). Is victimization chronic? A Bayesian analysis of multinomial missing data. J. of Econometrics Vol. 29, 47-67.

Karson, M.J., and Wrobleski, W.J. (1970). A Bayesian analysis of binomial data with a partially informative category. Procs. of Bus. & Economics Sec., Amer. Statist. Assoc., 532-534.

Kaufman, G.M., and King, B. (1973). A Bayesian analysis of nonresponse in dichotomous processes. J. Amer. Statist. Assoc. Vol. 68, 670-678.

Laplace, Pierre Simon (1774). Memoir on the probability of the causes of events. Translation by Stephen M. Stigler (1986). Laplace's 1774 memoir on inverse probability. Statistical Science, Vol. 1, No. 3, 359-378.

Raiffa, Howard, and Schlaifer, Robert (1961). Applied Statistical Decision Theory. Division of Research, Graduate School of Business Administration, Harvard University, Boston.

Rubin, Donald B. (1976). Inference and missing data. Biometrika, Vol. 63, 581-592.

Shefrin, H.M. (1981). On the combinatorial structure of Bayesian learning with imperfect information. Discrete Mathematics 37, 245-254.

Shefrin, H.M. (1983). Markov chains, imperfect state information, and Bayesian learning. Mathematical Modelling, Vol. 4, pp. 1-7.

Smith, Philip J., Choi, Sung C., and Gunel, Erdogen (1985). Bayesian analysis of a 2x2 contingency table with both completely and partially cross-classified data. Journal of Educational Statistics Vol. 10(1), 31-43.

Smith, Philip J., and Gunel, Erdogen (1984). Practical Bayesian approaches to the analysis of a 2x2 contingency table with incompletely categorized data. Commun. Statist.--Theor. Meth. Vol. 13(16), 1941-1963.

Tierney, Luke, and Kadane, Joseph B. (1986). Accurate approximations for posterior moments and marginal densities. J. Amer. Statist. Assoc. Vol. 81 (393), 82-86.

Vardi, Y., Shepp, L.A., and Kaufman, L. (1985). A statistical model for positron emission tomography (with Discussion). J. Amer. Statist. Assoc. Vol. 80 (389), 8-37.

Wilks, S.S. (1962). Mathematical Statistics. John Wiley and Sons, New York.