Bayesian Methods for Multinomial Sampling with Missing Data Using Multiple Hypergeometric Functions[1]

by

James M. Dickey, Jyh-Ming Jiang,

and

Joseph B. Kadane

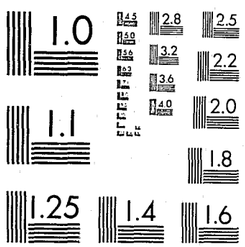
Research Report 6/83 - #15

State University of New York at Albany

Department of Mathematics and Statistics

5/30/83

[1]Work by Professor Dickey and Mr. Jiang sponsored by National Science Foundation, Grant MCS-8301335. Work by Professor Kadane sponsored by National Institutes of Justice, Grant 81-IJ-CX-0087, and the Office of Naval Research Contract, N00014-82-K-0622.



95230

0.1

Bayesian Methods for Multinomial Sampling with Missing Data Using Multiple Hypergeometric Functions[1]

by

James M. Dickey, Jyh-Ming Jiang,

and

Joseph B. Kadane

95230

Research Report 6/83 - #15

State University of New York at Albany

Department of Mathematics and Statistics

5/30/83

[1]Work by Professor Dickey and Mr. Jiang sponsored by National Science Foundation, Grant MCS-8301335. Work by Professor Kadane sponsored by National Institutes of Justice, Grant 81-IJ-CX-0087, and the Office of Naval Research Contract, N00014-82-K-0622.

ABSTRACT

Distribution theory is given for Bayesian inference from multinomial (or multiple Bernoulli) sampling with missing category distinctions, such as a contingency table with supplemental purely marginal counts. A new conjugate family generalizes the usual Dirichlet prior distributions. The posterior moments and predictive probabilities are found to be proportional to ratios of Carlson's hypergeometric functions of matrix argument. Dimension-reducing integral identities and expansions are given for statistical use. Closed-form expressions are developed for cases of nested missing distinctions. Examples are given, together with a simple method for assessment of a Dirichlet subjective prior distribution.

1. INTRODUCTION

In his two books (1950, 1965) and various articles, I.J. Good popularized Bayesian conjugate-prior inference for multinomial sampling, in which the Dirichlet distributions form the conjugate family of prior distributions. This mathematics is tractable and elegant. Densities, moments, and predictive probabilities take closed forms. Furthermore, the family of Dirichlet distributions and mixtures of Dirichlets with a small number of terms is large enough in many situations to offer realistic models for predata subjective uncertainty. (An exception occurs when there is a prior prejudice for local smoothness between neighboring category probabilities, such as when the data frequencies result from grouping a sampled continuous variable (Dickey 1968b).)

However, in the case of inference from data with some missing distinctions between categories this tractability seems to fade. The likelihood then contains factors which are powers of the sums of probabilities of the confused categories. The more data values that miss a particular distinction, the greater the power of the corresponding sum; and the greater the variety of such missed distinctions, the more such sums there are in the likelihood. In the least tractable cases, the sample contains moderate amounts of data of assorted kinds: complete data and data with various missing distinctions, without nesting or other logical constraints between the missed distinctions.

Bayesian treatments of multinomial sampling with missing data were given a decade ago by Karson and Wrobleski (1970), Antelman (1972), and Kaufman and King (1973). These concerned two-way contingency tables containing data with missing information regarding row or column variables. The studies were restricted to the 2x2 case. Basu and Pereira (1982) extended consideration to Kx2 tables and summarized properties of the Dirichlet distribution under relevant changes of variable. For an account of frequentist results, see Chen and Fienberg (1974, 1976), Dempster et al (1977), and references cited therein (for example, Hartley 1958).

After formulating here the general problem of Bayesian multinomial inference for missing category distinctions, in Section 2, we shall introduce in Sections 3 and 4 a representation of posterior integrals from Dirichlet prior distributions as multiple hypergeometric functions, specifically, the elegant functions R and R^* of Carlson (1977). The posterior distributions are found to be identical to distributions developed in Dickey (1983). This is then extended in Section 4 to a new conjugate prior family for such sampling. A parallel development based on changes of variable in the Dirichlet distribution yields closed forms for the case of nested missing distinctions, Section 5. This leads to new simplications of R^*. Straightforward expansions of posterior distributions as probability mixtures of Dirichlet distributions are generalized in Section 6, and related expansions of R^* are given. Examples are briefly given in Section 7. An appendix outlines a simple method for assessment of a Dirichlet subjective prior distribution.

2. SAMPLING WITH MISSING DISTINCTIONS

To begin, we shall need to establish notation for the non-missing (uncensored) case. Denote by $\underline{x} = (x_1,\ldots,x_K)'$ the vector of frequency counts for sampling from a distribution on a finite sample space having the unknown probability vector $\underline{u} = (u_1,\ldots,u_K)'$. That is, x_i denotes the number of sample values falling in the i^{th} category of probability u_i, $i = 1,\ldots,K$. For Bayesian inference concerning $\underline{u}$ under any noninformative-stopping process, the vector $\underline{x}$ can be treated as if it were the frequency count from a multiple-Bernoulli sequence $\underline{\delta}_x = (\delta_1,\ldots,\delta_M)$ of prespecified length M. Then $x_\cdot = M$, where $x_\cdot$ denotes $x_1+\ldots+x_K$. The corresponding likelihood function is proportional to the probability mass of such $\underline{\delta}_x$,

$$\mathrm{pr}(\underline{\delta}_x|\underline{u}) = \Pi_{i=1}^{K} u_i^{x_i} \quad . \tag{2.1}$$

Note that this function of $\underline{u}$ is parameterized by the frequencies $\underline{x}$, a sufficient statistic. (The term "multiple-Bernoulli" refers to a fixed-length sampling distribution for $\underline{\delta}_x$, for which $\underline{x}$ has the "multinomial" distribution with mass equal to the product of $\binom{M}{\underline{x}}$ multiplied by (2.1).)

The Bayesian predictive distribution is the subjective probability of an outcome sequence $\underline{\delta}_x$ without conditioning on the unknown sampling parameters $\underline{u}$. That is, $\underline{u}$ is averaged out of the sampling probability (2.1) according to a distribution of $\underline{u}$. Whatever the prior distribution, the prior predictive probability takes the form of a prior moment,

$$pr(\underline{\delta}_x) = E_u pr(\underline{\delta}_x|\underline{u}) \qquad (2.2)$$

$$= E_u \prod_1^K u_i^{x_i} .$$

This gives the subjective probability for the outcome sequence $\underline{\delta}_x$ conditional only on the total count $x_.$. A similar statement holds for posterior predictive probabilities of additional future outcomes in terms of posterior moments. Hence, to provide Bayesian predictive probabilities, we need merely develop the prior and posterior moments. (A similar statement will hold in the case of censored sampling.)

The conjugate prior family for uncensored sampling is the Dirichlet. The random vector $\underline{u}$ is said to have the Dirichlet distribution $D(\underline{b})$ with parameter vector $\underline{b} = (b_1,\ldots,b_K)'$, each $b_i>0$, if $\underline{u}$ has the density in any K-1 of its coordinates,

$$f(\underline{u};\underline{b}) \equiv B(\underline{b})^{-1} \prod_1^K u_i^{b_i-1}, \qquad (2.3)$$

$$B(\underline{b}) = [\prod_1^K \Gamma(b_i)]/\Gamma(b_.),$$

for all $\underline{u}$ in the probability simplex $\{\underline{u}: \text{each } u_i\geq 0,\ u_.=1\}$.

Consider the prior distribution,

$$\underline{u} \sim D(\underline{b}) . \qquad (2.4)$$

The corresponding prior moment is

$$g(\underline{c};\underline{b}) \equiv E_{u|b} \prod_1^K u_i^{c_i}$$

$$= B(\underline{b}+\underline{c})/B(\underline{b}). \qquad (2.5)$$

The predictive distribution is then the Dirichlet-Bernoulli with mass $pr(\underline{\delta}_x) = g(\underline{x};\underline{b})$. Note that $g(\underline{x})$ is the probability mass of the outcome sequence $\underline{\delta}_x$, rather than the frequency count $\underline{x}$.

The posterior distribution corresponding to the conjugate prior (2.4) and likelihood function (2.1) is again Dirichlet, with updated parameters,

$$\underline{u}|\underline{x} \sim D(\underline{b}+\underline{x}). \qquad (2.6)$$

The posterior Dirichlet density is then $f(\underline{u};\underline{b}+\underline{x})$ and the posterior moment is $g(\underline{c};\underline{b}+\underline{x})$.

Consider now the generalization to sampling from the distribution having probabilities $\underline{u}$ when some observations are censored, that is, do not report unique categories, but rather mere sets of categories. For example, an observation may be reported as falling either in category i or in category j, for a particular pair i<j. Denote the frequency count of such observations by y_{ij}. Let y_{ijk} be the frequency count of observations confusing a triple of categories, i<j<k. An analogous notation applies for any proper subset σ of at least two categories: y_σ for $\sigma\subset\{1,\ldots,K\}$. Denote a sequence of N such censored outcomes by $\underline{\varepsilon}_y = (\varepsilon_1,\ldots,\varepsilon_N)$ and

denote by $\underline{y}$ the corresponding vector counting the confusions of each possible type,

$$\underline{y} = (y_{12},\ldots,y_{123},\ldots,y_{23\ldots K})' \qquad (2.7)$$

The dimensionality of $\underline{y}$ is 2^K-K-2, although in practice, $\underline{y}$ will be sparse with most coordinates zero.

Consider sampling under noninformative stopping and noninformative censoring (Dawid and Dickey 1977). The likelihood function is proportional to the likelihood function from an outcome sequence $(\underline{\delta}_x,\underline{\varepsilon}_y)$, $\underline{\varepsilon}_y = (\varepsilon_1,\ldots,\varepsilon_N)$, where $\underline{\delta}_x$ and $\underline{\varepsilon}_y$ are independent and the coordinates ε_i independently arise from prespecified partitions of the sample space. Then

$$pr(\underline{\delta}_x,\underline{\varepsilon}_y|\underline{u}) = pr(\underline{\delta}_x|\underline{u})pr(\underline{\varepsilon}_y|\underline{u}), \qquad (2.8)$$

where

$$pr(\underline{\varepsilon}_y|\underline{u}) = \Pi_{k=2}^{K-1}\,\Pi_{i_1<\ldots<i_k}(u_{i_1}+\ldots+u_{i_k})^{y_{i_1\ldots i_k}}$$

$$= \Pi_\sigma(\Sigma_{i\varepsilon\sigma}u_i)^{y_\sigma}. \qquad (2.9)$$

Now, consider the effect of this likelihood function for inference from the Dirichlet prior density (2.3). The independence of the two factors in the likelihood (2.8) implies the following result.

Theorem 2.1. A data sequence $(\underline{\delta}_x,\underline{\varepsilon}_y)$ of prespecified length and censoring pattern has the predictive probability

$$pr(\underline{\delta}_x,\underline{\varepsilon}_y) = E_{u|b}[pr(\underline{\delta}_x|\underline{u})pr(\underline{\varepsilon}_y|\underline{u})]$$

$$= g(\underline{x};\underline{b})h(\underline{y};\underline{b}+\underline{x}), \qquad (2.10)$$

where the first factor g is given by (2.5) and the second factor is the expectation of (2.9) under the Dirichlet partial-posterior distribution, $\underline{u}|\underline{x} \sim D(\underline{b}+\underline{x})$,

$$h(\underline{y};\underline{b}+\underline{x}) = E_{u|b+x}pr(\underline{\varepsilon}_y|\underline{u}). \qquad (2.11)$$

The posterior density of the unknown probability vector $\underline{u}$ has the expression in terms of this same quantity h;

$$f(\underline{u};\underline{b}+\underline{x};-\underline{y}) = f(\underline{u};\underline{b}+\underline{x})pr(\underline{\varepsilon}_y|\underline{u})/h(\underline{y};\underline{b}+\underline{x}). \qquad (2.12)$$

(The function f of three arguments generalizes the previous notation f of two arguments (2.3); the motivation for defining a negative third argument will follow later.) The corresponding posterior moment is proportional to a ratio of such quantities,

$$E_{u|b+x,y}[(\Pi_1^K u_i^{c_i})\Pi_\sigma(\Sigma_{i\varepsilon\sigma}u_i)^{e_\sigma}]$$

$$= g(\underline{c};\underline{b}+\underline{x})h(\underline{y}+\underline{e};\underline{b}+\underline{x}+\underline{c})/h(\underline{y};\underline{b}+\underline{x}). \qquad (2.13)$$

where the proportionality factor $g(\underline{c};\underline{b}+\underline{x})$ is the usual Dirichlet posterior moment (2.5), based on the uncensored part $\underline{x}$ of the data. □

Attention focuses naturally, then, on the properties and calculation of the quantities h (2.11).

3. MULTIPLE HYPERGEOMETRIC FUNCTIONS

B. C. Carlson (1963,1971,1974,1977) developed the following class of multiple hypergeometric functions for an organizing and unifying role in the field of special functions. See also Dickey (1983) for probabilistic interpretations and statistical uses. We define the function R as a moment of a homogeneous linear form in the random vector $\underline{u} \sim D(\underline{b})$,

$$R_a(\underline{b},\underline{z}) = E^{(k-1)}_{\underline{u}|\underline{b}}(\underline{u}'\underline{z})^a \ , \qquad (3.1)$$

where $\underline{z} = (z_1,\ldots,z_K)'$ and $\underline{u}'\underline{z} = u_1z_1+\ldots+u_Kz_K$. (In this section we shall indicate the dimensionality of integrals in the notation we use for expectation.). This definition requires $b_i \geq 0$ for all i, but more general definitions, including a function that generates R, appear in Carlson (1977). Dickey (1983) exhibits R as being, itself, the probability generating function of the Dirichlet-multinomial distribution, the conjugate Bayesian predictive distribution for multinomial sampling.

The following classical identity, attributed to Picard by Appell and Kampé de Fériet (1926), reduces the dimensionality of the integral representation, thus permitting computation of R by simple quadrature. Although this refers to a restricted range of parameters, a contour integral applies more generally (Carlson 1977, Theorem 6.8-2, p.155).

Theorem 3.1 (Picard's identity). For $-b_. < a < 0$, $\underline{w} = (w_1,w_2)'$, and $\underline{d} = (-a, b_.+a)'$,

$$R_a(\underline{b},\underline{z}) = E^{(1)}_{w|d} \prod_{i=1}^{K}(w_1z_i+w_2)^{-b_i}$$
$$= B(\underline{d})^{-1}\int_0^1 v^{-a-1}(1-v)^{b_.+a-1}[\prod_1^K (vz_i+1-v)^{-b_i}]dv. \ \square$$

Under the parameter restriction $a = -b_.$, R takes the simple limiting form,

$$R_{-b_.}(\underline{b},\underline{z}) = \Pi z_i^{-b_i} \ . \qquad (3.3)$$

A two-way multiple hypergeometric function generalizing R, a function of matrix argument, can be defined by considering a bilinear form $\underline{u}'Z\underline{v}$ in independent random vectors, $\underline{u} \sim D(\underline{b})$ (K coordinates) and $\underline{v} \sim D(\underline{c})$ (L coordinates) for nonrandom matrix Z(K×L). Define the function R as a moment of the bilinear form,

$$R_a(\underline{b},Z,\underline{c}) = E^{(K-1)}_{\underline{u}|\underline{b}}E^{(L-1)}_{\underline{v}|\underline{c}}(\underline{u}'Z\underline{v})^a \qquad (3.4)$$
$$= E^{(K-1)}_{\underline{u}|\underline{b}}R_a(\underline{c};\underline{u}'\underline{z}_{*1},\ldots,\underline{u}'\underline{z}_{*L}),$$

where

$$Z = \begin{pmatrix} \underline{z}_{1*} \\ \underline{z}_{2*} \\ . \\ . \\ . \\ \underline{z}_{K*} \end{pmatrix} = (\underline{z}_{*1},\ldots,\underline{z}_{*L}). \qquad (3.5)$$

A simpler integral representation and a dimension-reducing integral identity are available for R_a, as follows, under the

parameter restriction $a = -c_\cdot$.

Theorem 3.2.

$$R_{-c_\cdot}(\underline{b},Z,\underline{c}) = E^{(K-1)}_{\underline{u}|b} \Pi_1^L (\underline{u}'\underline{z}_{*j})^{-c_j}$$

$$= E^{(L)}_{\underline{w}|d} \Pi_1^K (\underline{z}_{i*}'\underline{v} + w_{L+1})^{-b_i} , \qquad (3.6)$$

where

$$\underline{w} = \begin{pmatrix} \underline{v} \\ w_{L+1} \end{pmatrix} , \quad \underline{d} = \begin{pmatrix} \underline{c} \\ b_\cdot - c_\cdot \end{pmatrix} . \qquad (3.7)$$

The second equality in (3.6), first given by Dickey (1968a), generalizes Picard's identity (L = 1). □

Note that if $b_\cdot = c_\cdot$ in (3.6), $d_{L+1} = 0$ and the L-fold integral becomes an (L-1)-fold integral with $w_{L+1} \equiv 0$. More generally, as the following Corollary shows, the dimensionality of the integral can be reduced when the parameter vector $\underline{c}$ has zero coordinate values. This property is useful in missing data problems in which the high-dimensional vector $\underline{y}$ is sparse.

Corollary 3.3. If, without loss of generality, the vector $\underline{c}$ is taken in the form $\underline{c} = (\underline{c}^{(1)'},\underline{0}')'$ where the subvector $\underline{c}^{(1)}$ has $L^{(1)} \leq L$ coordinates, then

$$R_{-c_\cdot}(\underline{b},Z,\underline{c}) = R_{-c_\cdot^{(1)}}(\underline{b},Z^{(1)},\underline{c}^{(1)}) \qquad (3.8)$$

where $Z^{(1)}$ consists of the corresponding $L^{(1)}$ columns of Z, $Z = (Z^{(1)},Z^{(2)})$. A similar dimension-reduction occurs whenever there are some columns of Z which are proportional to each other, or proportional to the vector of unit entries $\underline{1} = (1,\ldots,1)'$. □

In principle, the integral identity (3.6) can be useful for computation of R by quadrature when L<<K (or $L^{(1)}$<<K). However, we have found this awkward in practice, because for interesting values of the parameters, the integrands tend to have poles in the range of integration. At present, series expansions of R appear more practical, as given later in the paper.

4. INFERENCE AND A NEW CONJUGATE FAMILY

Under the parameter restriction $a = -c_.$, Carlson's two-way multiple hypergeometric function R has the same form (3.6) as the mement quantities h (2.11) needed in Bayesian Dirichlet-prior inference for censored multinomial sampling. To express this identification, specialize the matrix Z to the full-subsets indicator $Z = \tilde{Z}$, consisting of zeros and ones whose L columns $\underset{\sim}{Z}_{*j}$ indicate all the proper subsets of two or more elements, $\sigma_j \subset \{1,\ldots,K\}$,

$$\tilde{Z}_{ij} = \begin{cases} 1 \text{ if } i\varepsilon\sigma_j \\ 0 \text{ otherwise,} \end{cases} \tag{4.1}$$

$j=1,\ldots L$, where

$$L = 2^K - K - 2. \tag{4.2}$$

Theorem 4.1.

$$h(\underline{d};\underline{b}) = E_{u|b} \Pi_{j=1}^{L} (\Sigma_{i\varepsilon\sigma_j} u_i)^{d_j} \tag{4.3}$$

$$= R_{d_.}(\underline{b},\tilde{Z},-\underline{d}). \quad \square$$

This identification ties the problems of Bayesian statistical inference for missing data to a mainstream segment of the theory of special functions. The Dirichlet prior distribution $\underline{u} \sim D(\underline{b})$ implies the predictive probability (2.10) with (4.3),

$$pr(\underline{\delta}_x,\underline{\varepsilon}_y) = [B(\underline{b}+\underline{x})/B(\underline{b})] \cdot R_{y_.}(\underline{b}+\underline{x},\tilde{Z},-\underline{y}), \tag{4.4}$$

and the corresponding posterior moment (2.13) becomes

$$[B(\underline{b}+\underline{x}+\underline{c})/B(\underline{b}+\underline{x})] \cdot R_{y_.+e_.}(\underline{b}+\underline{x}+\underline{c},\tilde{Z},-\underline{y}-\underline{e})/R_{y_.}(\underline{b}+\underline{x},\tilde{Z},-\underline{y}). \tag{4.5}$$

Note that our posterior distribution (2.12) is not in the same family as the prior, and thus the Dirichlet family is not conjugate for the missing-data likelihood. This motivates the following development.

4.1 A General Family.

Dickey (1983) generalized the Dirichlet distributions to the family, for arbitrary matrix parameter Z(K×L for arbitrary K,L),

$$\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d}), \tag{4.6}$$

having density on the simplex of K probabilities,

$$B(\underline{b})^{-1}(\Pi_1^K u_i^{b_i-1})[\Pi_1^L (\underline{u}'\underline{z}_{*j})^{d_j}] / R_{d_.}(\underline{b},Z,-\underline{d}). \tag{4.7}$$

(For the powers $\underline{d}$ in (4.7), the sign of the third parameter $-\underline{d}$ in (4.6) is defined to match the usage of Carlson's R; this avoids confusion in the long run.)

The moments of $\mathcal{D}$ are proportional, as in (4.5), to ratios of R functions,

$$E^{(K-1)}_{u|b,Z,-d}[(\Pi_1^K u_i^{c_i})\Pi_1^L(\underline{u}'\underline{z}_{*j})^{e_j}] \tag{4.8}$$

$$= [B(\underline{b}+\underline{c})/B(\underline{b})]$$

$$\cdot R_{d_\cdot+e_\cdot}(\underline{b}+\underline{c},Z,-\underline{d}-\underline{e})/R_{d_\cdot}(\underline{b},Z,-\underline{d}).$$

Zero parameter coordinates yield important reductions in complexity, as follows.

Lemma 4.2. If $\underline{d} = (\underline{d}^{(1)'},\underline{0}')'$ (without loss of generality) $Z = (Z^{(1)},Z^{(2)})$, and $\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d})$, then

$$\underline{u} \sim \mathcal{D}(\underline{b},Z^{(1)},-\underline{d}^{(1)}). \tag{4.9}$$

In particular, $\mathcal{D}(\underline{b},Z,\underline{0}) \sim D(\underline{b})$, regardless of the matrix Z. □

Lemma 4.3. If $\underline{b} = (\underline{b}_1',\underline{0}')'$, $Z = (Z_1',Z_2')'$ and $\underline{u} = (\underline{u}_1',\underline{u}_2')'$, then $\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d})$ if and only if $\underline{u}_2 = \underline{0}$ with probability one and

$$\underline{u}_1 \sim \mathcal{D}(\underline{b}_1,Z_1,-\underline{d}). \quad \square \tag{4.10}$$

Lemma 4.3 holds, in particular, for the Dirichlet distribution ($\underline{d} = \underline{0}$). Compare to Corollary 3.3 for R.

4.2 Inference.

To return to Dirichlet-prior inference for censored multinomial sampling, note that our posterior distribution (2.12) has the representation in the notation of (4.6),

$$\underline{u}|\underline{x},\underline{y} \sim \mathcal{D}(\underline{b}+\underline{x},\tilde{Z},-\underline{y}). \tag{4.11}$$

Hence, the posterior density $f(\underline{u};\underline{b}+\underline{x};-\underline{y})$ (2.12) is given by (4.7) with parameter values $\underline{b}+\underline{x},\tilde{Z},-\underline{y}$.

As stated earlier, the Dirichlet subfamily is not closed under censored multinomial sampling. However, the family of possible posterior distributions (4.11) does have this property, and thus we consider prior distributions of the following form.

Theorem 4.4. Consider the distributions

$$\underline{u} \sim \mathcal{D}(\underline{b},\tilde{Z},-\underline{d}). \tag{4.12}$$

As a prior distribution, (4.12) yields the predictive probability, by (4.8),

$$pr(\underline{\delta}_x, \underline{\varepsilon}_y) = [B(\underline{b}+\underline{x})/B(\underline{b})] \tag{4.13}$$

$$\cdot R_{d_\cdot+y_\cdot}(\underline{b}+\underline{x},\tilde{Z},-\underline{d}-y)/R_{d_\cdot}(\underline{b},\tilde{Z},-\underline{d}),$$

and the posterior distribution as an updating of parameters,

$$\underline{u}|\underline{x},\underline{y} \sim \mathcal{D}(\underline{b}+\underline{x},\tilde{Z},-(\underline{d}+\underline{y})). \tag{4.14}$$

The corresponding posterior density $f(\underline{u};\underline{b}+\underline{x};-(\underline{d}+\underline{y}))$ is given by (4.7) with parameter values $\underline{b}+\underline{x},\tilde{Z},-(\underline{d}+\underline{y})$. The Dirichlet-prior theory is the special case $\underline{d} = \underline{0}$. □

The posterior moments for (4.14) are obtained from (4.8) under the posterior parameter values. For example, the posterior mean is

$$E(u_i|\underline{x},\underline{y}) = [(b_i+x_i) / (b_\cdot+x_\cdot)] \tag{4.15}$$

$$\cdot R_{d_\cdot+y_\cdot}(\underline{b}+\underline{x}+\underline{\gamma}_i,\tilde{Z},-\underline{d}-\underline{y})/R_{d_\cdot+y_\cdot}(\underline{b}+\underline{x},\tilde{Z},-\underline{d}-\underline{y})$$

where the coordinates γ_{ij} of the vector $\underline{\gamma}_i$ are defined to be zero except for unity in the i^{th} coordinate. The posterior

second moment is

$$E(u_i u_j | \underline{x},\underline{y}) = \frac{(b_i+x_i+\gamma_{ij})(b_j+x_j)}{(b_.+x_.+1)(b_.+x_.)}$$

$$\cdot R_{d_.+y_.}(\underline{b}+\underline{x}+\underline{\gamma}_i+\underline{\gamma}_j,\tilde{z},-\underline{d}-\underline{y})/R_{d_.+y_.}(\underline{b}+\underline{x},\tilde{z},-\underline{d}-\underline{y}). \quad (4.16)$$

Dickey (1983) proposed the use of $\mathcal{D}$ as prior distributions for ordinary (noncensored) multinomial sampling. We suggest, however, that in practice, even for censored data, the usual Dirichlet distributions (or mixtures thereof with a small number of terms) may be preferred to $\mathcal{D}$ as prior distributions, thus yielding the new distributions $\mathcal{D}$ as posterior distributions. The Dirichlet has the advantage as a prior distribution of being convenient for subjective assessment (see Appendix), though more rigid in the subjective opinions it permits.

5. NESTED CENSORING

The posterior moments and distribution theory are considerably simplified when censoring is nested, that is, when for every two sets of confused categories, either one set is contained in the other or they are disjoint. For example, consider a multiple-choice questionnaire. If every person in the sample either answers all questions or is a complete nonrespondent then the censoring is nested. The Dirichlet-prior inference for such sampling was given by Basu and Pereira (1982).

The example can be extended to allow a particular question which the respondents can choose to ignore, or even a set of such questions that can be ignored as a whole. A series of sets of questions, linearly ordered by set-inclusion and the choice to answer all of each such set or only its subset is also permitted under nested censoring. For example, suppose the questions are presented in the same order to each respondent, who chooses some arbitrary place to stop answering questions, such as how far he/she gets before time runs out. As long as the respondents answer all questions until the places they stop, the data is nested. However, unconstrained decisions on whether to answer different questions would lead to non-nested censoring patterns in the data. Nested censoring is best treated in terms of nested partitions, which are defined in the following.

Define a matrix $Z(K\times L)$ to be a <u>partition indicator</u> if its columns indicate disjoint and exhaustive subsets, that is, a partition of $\{1,\ldots,K\}$. Formally,

1. Each entry of Z is a zero or a one.
2. Orthogonal columns: $\underline{z}'_{*j}\underline{z}_{*K} = 0$ for each $j \neq K$.
3. No row has all zero entries: $\underline{z}_{i*} \neq \underline{0}'$, for each $i = 1,\ldots,K$.

Given a partition indicator Z and K-tuple $\underline{u}$, define the corresponding <u>partition-sum vector</u> $\underline{s}(\underline{u}) = (s_1(\underline{u}),\ldots,s_L(\underline{u}))'$,

$$\underline{s}(\underline{u}) = Z'\underline{u} \quad . \tag{5.1}$$

Then each $s_j(\underline{u}) = \underline{z}'_{*j}\underline{u}$. If $\underline{u}$ is a probability vector, then the coordinates of $\underline{s}(\underline{u})$ are the probabilities of the partition-element events.

Define the <u>coordinatewise product</u> of two vectors as the vector of products of corresponding coordinates, $\underline{z}\times\underline{u} = (z_1u_1,\ldots,z_Ku_K)'$. Then given a partition indicator Z and K-tuple $\underline{u}$, define the <u>partition-inner-proportion vectors</u>,

$$\underline{r}_j(\underline{u}) = \underline{z}_{*j}\times\underline{u}/s_j(\underline{u}), \tag{5.2}$$

$j=1,\ldots,L$. Note that if $\underline{u}$ is a probability vector, each $\underline{r}_j$ consists of the conditional probabilities given the j^{th} partition-element event. As such, it has zero coordinates for all i not in the j^{th} subset. We have the invertible mapping corresponding to (5.1) and (5.2): $\underline{u} \leftrightarrow \underline{s}(\underline{u}),\underline{r}_j(\underline{u})(j=1,\ldots,L)$. The following theorem extends a result of Wilks (1962).

<u>Theorem 5.1</u>. Given a partition indicator Z, the vector $\underline{u}$ has the Dirichlet distribution $\underline{u} \sim D(\underline{b})$ if and only if the L+1 image vectors in (5.1), (5.2) are independently distributed and

$$\underline{s}(\underline{u}) \sim D(\underline{s}(\underline{b})) \tag{5.3}$$

$$\underline{r}_j(\underline{u}) \sim D(\underline{z}_{*j}\times\underline{b}), \tag{5.4}$$

$j = 1,\ldots,L$. Note that $\underline{z}_{*j}\times\underline{b} = s_j(\underline{b})\underline{r}_j(\underline{b})$, and that by Lemma 4.3, unless $z_{ij} = 1$, the i^{th} coordinate of $\underline{r}_j(\underline{u})$ equals zero with probability one. □

The following simple form of R appears to be new in the literature.

<u>Corollary 5.2</u>. If Z is a partition indicator,

$$R_{d.}(\underline{b},Z,-\underline{d}) = B(\underline{s}(\underline{b})+\underline{d})/B(\underline{s}(\underline{b})) \quad . \ \square \tag{5.5}$$

<u>Proof</u>. By (5.3),

$$E^{(K-1)}_{\underline{u}|\underline{b}}\Pi_1^L(\underline{z}'_{*j}\underline{u})^{d_j} = E^{(L-1)}_{\underline{s}(u)|\underline{s}(b)}\Pi_1^L s_j(\underline{u})^{d_j} \quad . \tag{5.6}$$

Theorem 5.1 generalizes immediately to distributions $\mathcal{D}$, as follows.

<u>Theorem 5.3.</u> Given a partition indicator Z, $\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d})$ if and only if the independent distributions (5.3), (5.4) hold for this same Z, with $\underline{s}(\underline{b})$ replaced by $\underline{s}(\underline{b})+\underline{d}$. □

<u>Corollary 5.4</u>. Given a partition indicator Z, the distribution $\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d})$ has closed-form moments (in terms of the same Z),

$$\begin{aligned} &E[(\Pi u_i^{c_i})\Pi s_j(\underline{u})^{e_j}] \\ &= [B(\underline{b}+\underline{c})R_{d.+e.}(\underline{b}+\underline{c},Z,-\underline{d}-\underline{e})] \\ &\quad /[B(\underline{b})R_{d.}(\underline{b},Z,-\underline{d})] \end{aligned} \tag{5.7}$$

$$= [B(\underline{b}+\underline{c})B(\underline{s}(\underline{b}+\underline{c})+\underline{d}+\underline{e})/B(\underline{s}(\underline{b}+\underline{c}))]$$
$$/[B(\underline{b})B(\underline{s}(\underline{b})+\underline{d})/B(\underline{s}(\underline{b}))]$$
$$= \left[\prod \frac{B(\underline{z}_{*j}\times(\underline{b}+\underline{c}))}{B(\underline{z}_{*j}\times\underline{b})}\right]\cdot\left[\frac{B(\underline{s}(\underline{b}+\underline{c})+\underline{d}+\underline{e})}{B(\underline{s}(\underline{b})+\underline{d})}\right] \quad . \quad \square$$

By Lemma 4.2, this yields the moments of the prior distribution (4.12) and the posterior distributions (4.11), (4.14), under censoring by a partition, provided the prior distribution refers to the same partition (if to any).

The elements of a partition of a set are subsets, any one of which could then be partitioned further. Two partitions, of which one is a partition of an element of the other, are said to be directly nested. Then one is a partition of a set and the other is a partition of a subset. Two partitions are said to be nested if they belong to a sequence of successively directly nested partitions. Finally, define a collection of partitions to be nested if they form a tree; that is, each partition is nested with respect to a particular partition, the root of the collection. Then the root is the only partition in the collection that is a partition of the original set.

Define matrix $Z(K\times L)$ to be a two-level nested-partitions indicator if it indicates a collection of partitions in which each nonroot partition is directly nested in the root partition. Namely, $Z = (Z^{(1)},Z^{(2)})$ where $Z^{(1)}$ indicates the root partition and

$$Z^{(2)} = (Z_j^{(2)})_{j=1}^{L^{(1)}} , \tag{5.8}$$

in which each submatrix $Z_j^{(2)}$ indicates a partition of the subset indicated by the j^{th} column $\underline{z}_{*j}^{(1)}$ of $Z^{(1)}$, $j=1,\ldots,L^{(1)}$. The submatrices $Z_j^{(2)}$ are filled out with zero rows for zero coordinates of $\underline{z}_{*j}^{(1)}$.

Theorem 5.5. Given the two-level nested-partitions indicator Z, $\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d})$ if and only if independently

$$\underline{s}^{(1)}(\underline{u}) \sim D(\underline{s}^{(1)}(\underline{b})+\underline{d}^{(1)}+\overline{\underline{d}}^{(2)}) \tag{5.9}$$

$$\underline{r}_j^{(1)}(\underline{u}) \sim \mathcal{D}(\underline{z}_{*j}^{(1)}\times\underline{b},\ Z_j^{(2)},-\underline{d}_j^{(2)}), \tag{5.10}$$

$j=1,\ldots,L^{(1)}$, where $\underline{s}^{(1)}$ and $\underline{r}_j^{(1)}$ $(j=1,\ldots,L^{(1)})$ refer to $Z^{(1)}$. In conformity to $(Z^{(1)},Z^{(2)}),\underline{d}' = (\underline{d}^{(1)}{}',\underline{d}^{(2)}{}')$ and

$$\underline{d}^{(2)}{}' = (\underline{d}_j^{(2)}{}')_{j=1}^{L^{(1)}} ; \tag{5.11}$$

and $\overline{\underline{d}}^{(2)}$ has the j^{th} coordinate $\underline{d}_j^{(2)}{}'\underline{1}$; $j=1,\ldots,L^{(1)}$. $\square$

Such a two-level nested-partitions distribution $\mathcal{D}$ can be expressed in terms of independent Dirichlet-distributed vectors, by application of Theorem 5.3 to (5.10). By iteration of Theorem 5.5, a nested-partitions distribution of any number of levels can be so expressed. Thus, the moments of any nested-partitions distribution can be obtained in closed form, as in Corollary 5.4. This yields further new closed forms for the function R for a nested-partitions indicator.

6. EXPANSION BY POSSIBLE DATA

A well known tautology in Bayesian statistics equates a posterior distribution based on censored data to a probability mixture of posterior distributions from the possible uncensored versions of the observed data. The following form of this tautology exhibits a $\mathcal{D}$ distribution as being, itself, a finite mixture of Dirichlet distributions. A useful expansion of R is obtained as a consequence.

Consider, for this purpose, matrices W of possible frequency counts refining the observed censored frequency vector $\underline{y}$. Write

$$W = (w_{i\sigma}: i=1,\ldots,K,\ \sigma\subset\{1,\ldots,K\}) \tag{6.1}$$

This K×L matrix has columns indexed by subsets $\sigma(L=2^K-K-2)$. For each subset $\sigma\subset\{1,\ldots,K\}$, the column $\underline{w}_{*\sigma}$ is supported on σ,

$$w_{i\sigma} = 0 \text{ for } i\notin\sigma, \tag{6.2}$$

and these possible frequencies for the categories in σ sum to the observed count for σ,

$$w_{\cdot\sigma} = y_\sigma \ . \tag{6.3a}$$

This says, for the row-vector of column sums,

$$\underline{w}_{\cdot *} = \underline{1}'W = \underline{y}' \ . \tag{6.3b}$$

Note that the column-vector of row sums, $\underline{w}_{*\cdot} = W\underline{1}$, is a possible uncensored frequency vector for $\{1,\ldots,K\}$ harmonizing with the observed censored frequencies $\underline{y}$.

Lemma 6.1. The sampling probability of a censored sequence is the weighted sum of probabilities over all possible refinements of

the censored frequencies,

$$pr(\underline{\varepsilon}_y|\underline{u}) = \sum_{W|y}[\Pi_\sigma \binom{y_\sigma}{\underline{w}_{*\sigma}}]\Pi_i u_i^{w_{i\cdot}} \ . \tag{6.4}$$

The notation $W|y$ indicates that the summation is over nonnegative integer $w_{i\sigma}$ constrained by (6.2), (6.3). □

Proof.

$$pr(\underline{\varepsilon}_y|\underline{u}) = \Pi_\sigma(\textstyle\sum_{i\varepsilon\sigma}u_i)^{y_\sigma} \tag{6.5}$$

$$= \Pi_\sigma \sum_{\underline{w}_{*\sigma}|y_\sigma} \binom{y_\sigma}{\underline{w}_{*\sigma}} \Pi_i\ u_i^{w_{i\sigma}}$$

$$= \sum_{W|y} \Pi_\sigma[\binom{y_\sigma}{\underline{w}_{*\sigma}} \Pi_i\ u_i^{w_{i\sigma}}] \ .$$

In a more detailed notation, one can write,

$$\Pi_\sigma \binom{y_\sigma}{\underline{w}_{*\sigma}} = \prod_{k=2}^{K-1} \prod_{i_1<\ldots<i_k} \binom{y_{i_1\ldots i_k}}{\underline{w}_{*i_1\ldots i_k}} \ . \tag{6.6}$$

Theorem 6.2. For a Dirichlet prior distribution $\underline{u} \sim D(\underline{b})$, the posterior distribution $\mathcal{D}$ (2.12), (4.11) based on $\underline{x}$, $\underline{y}$ is a posterior-probability mixture of posterior Dirichlet distributions based on $\underline{x}$, $\underline{w}_{*\cdot}$,

$$f(\underline{u};\underline{b}+\underline{x};-\underline{y}) \tag{6.7}$$

$$= \sum_{W|y} f(\underline{u};\underline{b}+\underline{x}+\underline{w}_{*\cdot})pr(W|\underline{b}+\underline{x},\underline{y}),$$

where f with two and f with three arguments are defined by (2.3) and (4.11), (4.7), respectively,

$$\text{pr}(W|\underline{b}+\underline{x},\underline{y}) \tag{6.8}$$

$$= \left[\Pi_\sigma \binom{y_\sigma}{\underline{w}_{*\sigma}}\right] g(\underline{w}_{*\cdot};\underline{b}+\underline{x})/h(\underline{y};\ \underline{b}+\underline{x})$$

and

$$h(\underline{y};\ \underline{b}+\underline{x}) \tag{6.9}$$

$$= \sum_{W|y} \left[\Pi_\sigma \binom{y_\sigma}{\underline{w}_{*\sigma}}\right] g(\underline{w}_{*\cdot};\ \underline{b}+\underline{x}),$$

where g and h are defined by (2.5), (2.11), (4.3). □

Related results appear in Shefrin (1981). Theorem 6.2 can be generalized immediately for use of a prior distribution $\mathcal{D}(\underline{b},Z,-\underline{d})$, in the spirit of Theorem 4.4, merely by substituting $\underline{d}+\underline{y}$ for $\underline{y}$.

The representation (6.7) can be reexpressed in a general notation for probability mixtures,

$$\underline{u}|\underline{x},\underline{y} \sim D(\underline{b}+\underline{x}+\underline{w}_{*\cdot})*W \tag{6.10}$$

where W has the mass function (6.8), or its $\mathcal{D}$-prior form. In this notation, our theorem generalizes simply to give a mixture-representation for any $\mathcal{D}$ distribution, as follows.

Theorem 6.3. Consider arbitrary Z(K×L, for arbitrary K,L). Conformably partition $Z = (Z^{(1)},Z^{(2)})$, $\underline{d} = (\underline{d}^{(1)'},\underline{d}^{(2)'})'$ and $L = L^{(1)}+L^{(2)}$. Then $\underline{u} \sim \mathcal{D}(\underline{b},Z,-\underline{d})$ if and only if

$$\underline{u} \sim \mathcal{D}(\underline{b}+\underline{w}_{*\cdot},Z^{(1)},-\underline{d}^{(1)})*W, \tag{6.11}$$

where $W(K\times L^{(2)})$ has probability mass

$$\text{pr}(W) = \left[\Pi_{j=1}^{L^{(2)}} \binom{d_j^{(2)}}{\underline{w}_{*j}} \Pi_{i=1}^{K} z_{ij}^{(2)\,w_{ij}}\right]$$

$$\frac{B(\underline{b}+\underline{w}_{*\cdot})}{B(\underline{b})} \cdot \frac{R_{d_\cdot^{(1)}}(\underline{b}+\underline{w}_{*\cdot},Z^{(1)},-\underline{d}^{(1)})}{R_{d_\cdot}(\underline{b},Z,-\underline{d})}, \tag{6.12}$$

for $\underline{w}_{*\cdot} = \underline{d}^{(2)}$. □

Theorem 6.2 is the special case, $L^{(1)} = 0$ with parameters $\underline{b}+\underline{x},\tilde{Z},-\underline{y}$. Note that the generalization of restriction (6.2), $w_{ij} = 0$ for $z_{ij}^{(2)} = 0$, holds automatically for (6.12). Since (6.12) must sum to unity, we have an apparently new representation for R under the parameter restriction of (3.6).

Corollary 6.4.

$$R_{d_\cdot}(\underline{b},Z,-\underline{d}) \tag{6.13}$$

$$= \sum_{w|d^{(2)}} \left[\Pi^{L^{(2)}} \binom{d_j^{(2)}}{\underline{w}_{*j}} \Pi^{K} z_{ij}^{(2)\,w_{ij}}\right]$$

$$[B(\underline{b}+\underline{w}_{*\cdot})/B(\underline{b})]R_{d_\cdot^{(1)}}(\underline{b}+\underline{w}_{*\cdot},Z^{(1)},-\underline{d}^{(1)}) .$$

In case $L^{(1)} = 0$ in (6.13), $R_{d_\cdot^{(1)}} = 1$ as in (6.9). □

In our experience so far, the expansions of this section have been more convenient and economical for computations than the naive quadrature of analogs of the dimension-reduced integral (3.6).

Such expansion methods can be easily combined with the partition-nesting methods of the preceding section to minimize the number of terms involved, as in Antelman (1972) for special cases. Recursion relations are easily set up for the multinomial coefficients.

7. EXAMPLES

Example 1. Contingency Tables with Supplemental Margins.

An interesting special case of censored multinomial sampling is provided by contingency tables with supplemental purely marginal data. Consider a two-way table with three independent multinomial data sets, respectively for: the cross-classified table, itself; the row variable alone; and the column variable alone. By the likelihood principle, for suitable models, these data can be directly combined into the inferential equivalent of a single censored multinomial sample.

Chen and Fienberg (1974) analyze such data, given here in Table 1, on M+N = 456 premature live births, classified by serum bilirium level (mg. per 100 ml, "Low" 0-1.0) and/or a composite health index (0-10, "Low" ("healthier") 0-6).

Table 1. Data on premature live births (Chen and Fienberg 1974).

Serum bilirium	Health Index Low	Health Index High	Supplemental data on serum bilirium
Low	35	75	11
High	57	112	13
Supplemental data on health index	117	36	

For the cross-classified cell probabilities u_{ij} (i serum bilirium, j health index), a uniform prior distribution, $(u_{11},u_{12},u_{21},u_{22}) \sim D(\underline{b})$, $\underline{b} = (1,1,1,1)'$, yields a posterior distribution $\mathcal{D}(\underline{b}+\underline{x},Z,-\underline{d})$, where $\underline{x}$ is given by the cross-classified body of Table 1,

$$Z = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \qquad (7.1)$$

and $\underline{d} = (117,36,11,13)'$. This posterior distribution can be expressed as a mixture of nested-partition distributions by Theorem 6.3, by expanding the two lower powers $(u_{11}+u_{12})^{11}$ and $(u_{21}+u_{22})^{13}$. The resulting posterior means are given in Table 2 with the posterior standard deviations and correlations, and the maximum likelihood solution from Chen and Fienberg (1974). These Bayesian estimates are not very different from the maximum likelihood values, as should be expected from a uniform prior distribution and reasonable sample sizes. Hence, as is true of the likelihood inference, appreciable information is gained in the Bayesian analysis from the supplemental marginal data.

Table 2. Estimates of cross-classified cell probabilities for data on premature live births.

Cell probabilities		u_{11}	u_{12}	u_{21}	u_{22}
Max.-likelihood est.		0.1880	0.2090	0.2960	0.3090
Posterior mean		0.1888	0.2096	0.2950	0.3065
Posterior SD		0.0255	0.0206	0.0277	0.0230
Posterior correlations	u_{11}	1	-0.21	-0.60	-0.20
	u_{12}		1	-0.21	-0.40
	u_{21}			1	-0.35

Example 2. Combining Surveys with Different Questionnaires

Kadane (1982) analyzed data from two sample surveys of attitudes on the death penalty. The primary categories are, for i = 1,2,3,4:

1. Would not decide guilt versus innocence in a fair and impartial manner.
2. Fair and impartial on guilt versus innocence; and, on sentencing, would always vote for the death penalty, regardless of circumstances.
3. Fair and impartial; and would never vote for the death penalty.
4. Fair and impartial; and would sometimes and sometimes not vote for the death penalty.

A survey by the Field Research Corporation produced data, $x_1=68$, $x_3=97$, $y_{2,4}=674$ (M+N=839), and a Harris survey produced data, $x_2=15$, $y_{1,3,4}=1484$ (M+N=1499). By the likelihood principle, these two multinomial samples can be directly combined in the form of a single censored multinomial sample, for which the censoring is not nested (M+N=2338).

A genuine Bayesian analysis reports the coherent effect of data on prior distributions expressing actual expert opinion prior to knowledge of the data. To simulate aspects of such a process, we assessed a Dirichlet distribution by interactive elicitation of the opinion of a social psychologist with interests in legal

matters, a person who was not yet familiar with the two surveys. (Our assessment method is outlined in the Appendix.) The resulting prior and posterior distributions are, respectively, $D(\underline{b})$ with $\underline{b}$ as given in Table 3, and $\mathcal{D}(\underline{b}+\underline{x}, Z, -\underline{d})$ where $\underline{b}+\underline{x} = (70.8, 26.2, 118.0, 105.0)'$,

$$Z = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \qquad (7.2)$$

and $\underline{d} = (674, 1484)'$.

Using the method of Theorem 6.3 to expand $(u_2+u_4)^{674}$ in $\mathcal{D}$, we obtain a probability mixture of nested-partition distributions, each having closed-form moments. The resulting posterior means, standard deviations, and correlations for the four cell probabilities are given in Table 3, together with the posterior means from a uniform prior distribution, and the estimates obtained by simply solving the following simultaneous linear equations. Setting each u_i and u_σ equal to the corresponding observed frequency ratio, we have the system,

$$u_1 = 68/839,\ u_3 = 97/839,\ u_2+u_4 = 674/839; \qquad (7.3)$$

$$u_2 = 15/1499,\ u_1+u_3+u_4 = 1484/1499. \qquad (7.4)$$

Table 3. Estimation of death-penalty attitudes, combining two surveys analyzed by Kadane (1982).

Cell probabilities	u_1	u_2	u_3	u_4
Combined survey data				
$\underline{x}$	68	15	97	0
$\underline{y}$	$y_{1,3,4}=1484$	$y_{2,4}=674$		
(M+N = 2337)				
Estimate by observed frequency ratios	0.081	0.010	0.116	0.793
Posterior mean from uniform prior	0.082	0.011	0.116	0.791
Expert prior mean $\underline{b}/b_.$ ($b_. = 140$)	0.020	0.080	0.150	0.750
Expert prior SD	0.012	0.023	0.030	0.036
Posterior mean	0.073	0.016	0.122	0.789
Posterior SD	0.008	0.003	0.010	0.013
Posterior correlations				
u_1	1	-0.01	-0.10	-0.55
u_2		1	-0.01	-0.23
u_3			1	-0.74

The three vector estimates are not very different. Note that the expert posterior mean vector is farther away from the

observed-frequency estimate than is the posterior mean yielded by the uniform prior distribution, and this is true in each coordinate separately. This seems reasonable in that the uniform prior expresses greater uncertainty than the expert prior. The effect of either prior is to move each coordinate of the observed-frequency estimate toward the corresponding prior mean. The uniform prior, however, moves the first coordinate estimate in the opposite direction than does the expert prior. Hence, the uniform prior distribution seems less reasonable than the expert prior in this case.

APPENDIX. A Simple Method for Assessment of a Dirichlet Subjective Prior Distribution in Multinomial Sampling.

A subjective prior distribution in the Dirichlet family can be assessed, as follows, in the multiple-Bernoulli or multinomial sampling context and notation of Section 2. Our method is a "device of imaginary results", in the language of I.J. Good. That is, the expert whose opinion is being assessed imagines hypothetical data and the method elicits his pretended reaction to it. Of course, real data can be used in a similar way. The elicitations refer only to the Dirichlet-Bernoulli predictive distribution. Indeed the predictive distribution is primary and can be assessed by the method, regardless of the significance or not of assuming an imbedded multiple-Bernoulli process and Dirichlet mixing distribution. The assessment proceeds in three steps.

1. Elicit predictive probabilities for a single future observation,

$$\text{pr}(\delta=i) = b_i/b_\cdot, \quad i=1,\ldots,K. \tag{A.1}$$

2. Condition opinion on an imaginary (or real) future sample with frequencies $\underline{x} = (x_1,\ldots,x_K)$. Elicit conditional predictive probabilities for a further future observation,

$$\text{pr}(\delta=i|\underline{x}) = \frac{b_\cdot}{b_\cdot+x_\cdot}\,\text{pr}(\delta=i)+\frac{x_\cdot}{b_\cdot+x_\cdot}\hat{u}_i\,, \tag{A.2}$$

where $pr(\delta=i)$ was obtained in Step 1 and $\hat{u}_i$ denotes the usual relative-frequency estimate,

$$\hat{u}_i = x_i/x_. \quad . \tag{A.3}$$

Solve for $b_.$,

$$b_. = x_. \left[\frac{\hat{u}_i - pr(\delta=i|\underline{x})}{pr(\delta=i|\underline{x}) - pr(\delta=i)}\right] \quad . \tag{A.4}$$

3. Calculate b_i ,

$$b_i = pr(\delta=i)b_. \quad , \quad i=1,\ldots,K \quad . \tag{A.5}$$

In practice, the numerator and denominator in (A.4) should have the same sign. Under the conjugate-prior model, $pr(\delta=i|\underline{x})$ lies between $pr(\delta=i)$ and $\hat{u}_i$. The ratio of its distance to $\hat{u}_i$ to its distance to $pr(\delta=i)$ is $b_./x_.$, a (positive) constant in i. If $pr(\delta=i|\underline{x})$ is elicited and (A.4) solved for $b_.$ for various i values, then these $b_.$ values can be averaged. Similarly, an average $b_.$ can be obtained from consideration of several samples $\underline{x}$. Pooling of samples would permit nested conditioning and avoid the need for forgetting or unpretending.

The values given in Table A.1 were elicited from a social psychologist for Example 2 of Section 7. The simple average of the assessed $b_.$ values is $\bar{b}_. = 140$.

Table A.1. Assessment of an expert prior distribution for death-penalty attitude surveys.

i	1	2	3	4
Prior-predictive probabilities $pr(\delta=i)$ [Elicited]	0.02	0.08	0.15	0.75
Imaginary sample data x_i $(M = 200)$	16	20	32	132
Relative-frequency estimate $\hat{u}_i$	0.08	0.10	0.16	0.66
Posterior-predictive probabilities $pr(\delta=i\|\underline{x})$ [Elicited]	0.05	0.09	0.16	0.70
Solution $b_.$ $(\bar{b}_. = 140)$	200	200	0	160
$b_i = \bar{b}_. pr(\delta=i)$	2.8	11.2	21.0	105.0

REFERENCES AND BIBLIOGRAPHY

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

Appell, P., and J. Kampé de Fériet (1926). Fonctions Hypergéométriques et Hypersphériques: Polynomes d'Hermite. Gauthier-Villars, Paris.

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. (1963). Lauricella's hypergeometric function F_D. J. of Mathl. Anal. and Appls., Vol. 7, 452-470.

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.

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

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.

END