On the Bayesian Analysis of
Categorical Data: The Problem of Nonresponse

by

D. Basu¹ and Carlos A. B. Pereira²

September, 1980
FSU Statistics Report No. M558

The Florida State University
Department of Statistics
Tallahassee, Florida 32306

¹Research partially supported by NSF Grant No. 79-04693
²Research supported by CAPES and USP - Brazil
ON THE BAYESIAN ANALYSIS OF CATEGORICAL DATA:

THE PROBLEM OF NONRESPONSE

ABSTRACT: It is demonstrated how a suitably chosen prior for the frequency parameters can streamline the Bayesian analysis of Categorical data with missing entries due to nonresponse or other causes. The two cases where the data follow the Multinomial or the Hypergeometric model are treated separately. In the first case it is adequate to restrict the prior (for the cell probabilities) to the class of Dirichlet distributions. In the case of the Hypergeometric model it is convenient to select a prior from the class of Dirichlet-Multinomial (DM) distributions. The DM distributions are studied in some details.
ON THE BAYESIAN ANALYSIS OF CATEGORICAL DATA:
THE PROBLEM OF NONRESPONSE

I - INTRODUCTION

The simplest case of the problem of nonresponse is as follows. Let $\pi_1$ be the unknown proportion of individuals in a certain population, $P$, that belong to a particular category $A_1$. With $\pi_1$ as the only parameter of interest, a survey is conducted using a simple random sample of size $n$. Of the $n$ individuals surveyed, $n_1$ respond to the question "Do you belong to category $A_1$?" with a yes/no answer, but $n_2 = n - n_1$ individuals do not respond. Denoting the category of respondents by $R$, and the complementary category by $R'$, the survey data may be summarized as:

\[
\begin{array}{c|cc|}
 & R & R' \\
\hline
A_1 & x_1 & n_2 \\
A_2 & x_2 & n_2 \\
\hline
n_1 & n_2 & n
\end{array}
\]

with $A_2$ being the complement of $A_1$. 
In many practical problems, it is understood that the non-response of an individual is highly dependent on the value of the measurement under study. For example, suppose that one is surveying a population of students in order to estimate the proportion of cannabis smokers. In this case, it should be expected that a student who smokes has a higher chance of being a nonrespondent than one who does not. In this instance, at least, a nonresponse is a strong source of information.

The above understanding of the problem suggests that the population must also be partitioned into the categories R and R'; that is, the class of elements which would respond to the question, if selected, and its complement. The population proportions may be displayed in a $2 \times 2$ - tabular form as:

\[
\begin{array}{c|cc|c}
 & R & R' & \\
\hline
A_1 & p_{11} & p_{12} & \Pi_1 \\
A_2 & p_{21} & p_{22} & \Pi_2 \\
\hline
q & 1 - q & 1
\end{array}
\]

\[\Pi_2 = 1 - \Pi_1 \]

\[q = p_{11} + p_{21}\]

How can the data (1.1) be analysed vis-à-vis the parameter of interest $\Pi_1 = p_{11} + p_{12}$?

If the population of size $N$ is regarded to be infinitely large compared to the sample size $n$; that is, if a multinomial model for the data is adopted, then the likelihood function is:
introduces the Dirichlet-Multinomial distribution and some of its properties. This distribution, besides being the marginal distribution of the data, plays an important role in the rest of the paper.

Sections 4 and 5 deal with the case of sampling from a finite population; that is, the case where the statistical model is Hypergeometric or, more generally, Multivariate Hypergeometric.

For the case where \( k = 2 \), instead of \( p_{11}, p_{21}, p_{12}, \) and \( p_{22} \), the unknown frequency counts \( \theta_{11}, \theta_{21}, \theta_{12}, \) and \( \theta_{22} \) must be considered. As in (1.2), the population parameters may be displayed as:

\[
(1.4) \\
\begin{array}{c|c|c|c}
 & R & R' \\
\hline
A_1 & \theta_{11} & \theta_{12} & \theta_1 \\
A_2 & \theta_{21} & \theta_{22} & \theta_2 \\
\hline
\psi & N - \psi & N
\end{array}
\]

with \( \theta_2 = N - \theta_1 \), and the parameter of interest being \( \theta_1 = \theta_{11} + \theta_{12} \).

A Dirichlet-Multinomial prior for \( \Theta = (\theta_{11}, \theta_{21}, \theta_{12}, \theta_{22}) \) greatly simplifies the analysis of the data (1.1) vis-à-vis the parameter of interest \( \theta_1 \).

**NOTATIONS:** Let \( x, y, \) and \( z \) be either random variables or random vectors. When \( x, y, \) and \( z \) are mutually independent we write \( x \perp \perp y \perp \perp z \). By \( x \perp \perp y | z \) it is meant that \( x \) and \( y \) are conditionally independent given \( z \), and if \( x \) and \( y \) have the same distribution we write \( x \sim y \).
\[(1.3) \quad L = p_{11} x_1 p_{21} x_2 (1 - q)^{n_2}.\]

We represent the data by \(X = (x_1, x_2, n_2)\) with \(n_2 = n - (x_1 + x_2)\).

Since \(p_{12}\) cannot be defined in terms of the sampling distribution of \(X\), an orthodox non-Bayesian would characterize \(\Pi_1 = p_{11} + p_{12}\) as nonidentifiable, and would have little else to say on the matter. None of the many non-Bayesian methods of nuisance parameter elimination listed in Basu [1977] apply to the present case. On the other hand, a Bayesian regards a parameter as an unknown entity that exists in its own right. It enters into the sampling distribution of a properly planned experiment but is not defined by the experiment. Nonidentifiability is, therefore, a non-problem from the Bayesian viewpoint.

With a suitable representation \(\xi\) of his/her opinion about \(p = (p_{11}, p_{21}, p_{12}, p_{22})\), the Bayesian will proceed to derive the posterior distribution by matching \(\xi\) with the likelihood function \((1.3)\). The posterior marginal distribution of the parameter of interest \(\Pi_1\) will be obtained by integration.

In Section 2 we demonstrate how the choice of a Dirichlet prior for \(p\) simplifies the Bayesian operation. The more general case where the respondents are classified into \(k\) (instead of 2) categories, \(A_1, \ldots, A_k\), is analyzed in a similar fashion. Since the inference is based on the data, it is of interest to study the distribution of the data under the considered prior. Section 3
Let $p = (p_1, \ldots, p_k)$ be a $k$-dimensional positive random vector such that $\sum_{i=1}^{k} p_i = 1$. We write $p \sim D(\alpha_1, \ldots, \alpha_k)$ to indicate that the distribution of $p$ is a Dirichlet with nonnegative real parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$. For $k = 2$, instead of $(p_1, p_2) \sim D(\alpha_1, \alpha_2)$, we use the conventional Beta distribution notation, $p_1 \sim B(\alpha_1, \alpha_2)$.

Let $x = (x_1, \ldots, x_k)$ be a $k$-dimensional nonnegative integer random vector with fixed $n = \sum_{i=1}^{k} x_i$. We write $x | p \sim M(n; p)$, where $p$ is defined as above, to indicate that the conditional distribution of $x$ given $p$ is Multinomial with parameters $n$ and $p$. For $k = 2$, instead of $(x_1, x_2) | (p_1, p_2) \sim M(n; (p_1, p_2))$. We use the conventional Binomial distribution notation, $x_1 | p_1 \sim Bi(n; p_1)$.

When $\theta = (\theta_1, \ldots, \theta_k)$ is a nonnegative integer random vector with $\sum_{i=1}^{k} \theta_i = N$ fixed, we write $x | \theta \sim H(N, n, \theta)$ to indicate that the conditional distribution of $x$ given $\theta$ is Multivariate Hypergeometric with parameter $(N, n, \theta)$. For $k = 2$, instead of $(x_1, x_2) | (\theta_1, \theta_2) \sim H(N, n, (\theta_1, \theta_2))$, we use the conventional notation for Hypergeometric distributions, $x_1 | \theta_1 \sim h(N, n, \theta_1)$. The probability function correspondent to $H(N, n, \theta)$ may be expressed in the two following ways:

$$f(x | \theta) = \frac{\binom{\theta_1}{x_1} \binom{\theta_2}{x_2} \cdots \binom{\theta_k}{x_k}}{\binom{N}{n}}$$

$$= \frac{\binom{n}{x_1} \cdots \binom{N-n}{x_k} \binom{\theta_1-x_1}{\theta_1-x_1} \cdots \binom{\theta_k-x_k}{\theta_k-x_k}}{\binom{N}{\theta_1} \cdots \binom{\theta_k}{\theta_k}}$$
2 - NONRESPONSE: THE MULTINOMIAL MODEL

First we consider the case of $k = 2$, where the data, the population parameters, and the likelihood are described by (1.1), (1.2), and (1.3) respectively.

In the full response model, it is well known that the family of Dirichlet distributions of the correct dimension, is the natural conjugate family for the Bayesian analysis. That is, if $y_1$ and $y_2$ were the observations in $\mathbb{R}^1$, and

\[(2.1) \quad p = (p_{11}, p_{21}, p_{12}, p_{22}) \sim \text{D}(\alpha_{11}, \alpha_{21}, \alpha_{12}, \alpha_{22})\]

a priori, then the posterior distribution would be

\[\text{D}(\alpha_{11} + x_1, \alpha_{21} + x_2, \alpha_{12} + y_1, \alpha_{22} + y_2).\]

To introduce a Bayesian solution to the nonresponse case, it is useful to consider the following reparametrization:

\[(2.2) \quad q = p_{11} + p_{21}, \quad q_{11} = \frac{p_{11}}{q}, \quad \text{and} \quad q_{12} = \frac{p_{12}}{1 - q}\]

with the reverse transformation being

\[(2.3) \quad p_{11} = q q_{11}, \quad p_{12} = (1 - q) q_{12}\]
\[p_{21} = q (1 - q_{11}), \quad \text{and} \quad p_{22} = (1 - q) (1 - q_{12})\]

The following general results for Dirichlet distributions is a key to the solution. Let $m \in \{2, \ldots, k - 1\}$ be fixed.
LEMMA 1

The following set of conditions is necessary and sufficient to have \((p_1, \ldots, p_k) \sim D(\alpha_1, \ldots, \alpha_k)\):

(i) \[ y = \sum_{i=1}^{m} p_i \sim B\left(\sum_{i=1}^{m} p_i, \frac{y^k}{\sum_{i=1}^{m} \alpha_i}\right) \]

(ii) \[ \frac{1}{y} (p_1, \ldots, p_m) \sim D(\alpha_1, \ldots, \alpha_m) \]

\[ \frac{1}{1 - y} (p_{m+1}, \ldots, p_k) \sim D(\alpha_{m+1}, \ldots, \alpha_k), \]

and (iii) \[ y \perp \perp \frac{1}{y} (p_1, \ldots, p_m) \perp \perp \frac{1}{1 - y} (p_{m+1}, \ldots, p_k). \]

The proof of this result is straightforward and therefore is omitted.

Suppose that, a priori, (2.1) is considered. By Lemma 1, this is equivalent to

\[ q \sim B(\alpha_1, \alpha_2), q_{11} \sim B(\alpha_{11}, \alpha_{21}), \]

(2.4)

\[ q_{12} \sim B(\alpha_{12}, \alpha_{22}), \text{ and } q \perp \perp q_{11} \perp \perp q_{12} \]

where \[ \alpha_j = \alpha_{1j} + \alpha_{2j}, (j = 1, 2). \]

The reparametrization (2.2) changes the likelihood (1.3) to

(2.5) \[ L = q^n (1 - q)^n_2 q_{11} (1 - q_{11}) x_1 x_2. \]

By matching the prior (2.4) with (2.5), we derive the posterior distribution of \((q, q_{11}, q_{12})\):
(i) \( q \perp \perp q_{11} \perp \perp q_{12} \mid x \),

(ii) \( q_{11} \mid x \sim B(\alpha_{11}, x_1, \alpha_{21} + x_2) \),

\( q_{12} \mid x \sim q_{12} \sim B(\alpha_{12}, \alpha_{22}) \),

and (iii) \( q \mid x \sim q \mid n_1 \sim B(\alpha_{11} + n_1, \alpha_{21} + n_2) \).

As expected, \( n_1 \) is sufficient to predict \( q \), and \( q_{12} \) is independent of the data. Since \( \alpha_{12} = \alpha_{12} + \alpha_{22} \leq \alpha_{12} + n_2 \), the posterior distribution of the original parameter \( p \) is again Dirichlet if and only if \( n_2 = 0 \). It is, however, a mixture of Dirichlet distributions, and \( (p_{11}, p_{21}, (1 - q)) \mid x \sim D(\alpha_{11} + x_1, \alpha_{21} + x_2, \alpha_{12} + n_2) \). Note that these properties of the posterior allow one to define a "nice" conjugate family of distributions for the nonresponse case. That is, the prior given by (2.4) would be conjugate if, instead of \( q \sim B(\alpha_{11}, \alpha_{12}) \), we had \( q \sim B(\alpha_{11}, \beta) \), where \( \beta \geq \alpha_{12} \).

To proceed with the estimation of \( \Pi_1 \), the parameter of interest, we recall (2.5) to write \( \Pi_1 = q q_{11} + (1 - q) q_{12} \), and consider

\( \alpha = \alpha_{11} + \alpha_{21} + \alpha_{12} + \alpha_{22} \), and \( \alpha_i = \alpha_{11} + \alpha_{12} (i = 1, 2) \). Under the squared error loss function, the Bayes estimator of \( \Pi_1 \) is given by:

\[
\hat{\Pi}_1 = E(\Pi_1 \mid x) = E(q q_{11} + (1 - q) q_{12} \mid x).
\]

In view of the posterior distribution (2.6), we finally have.
\[
\hat{\pi}_1 = \frac{1}{\alpha + n} (\alpha_{12} + x_1 + \frac{\alpha_{12} n_2}{\alpha_{22}})
\]

We notice that (see Example in Section 3) \(\frac{\alpha_{12}}{\alpha_{22}} n_2\) is the conditional expectation of \(y_1\) - the sample frequency of nonrespondents that belong to \(A_1\) - given the data. Therefore, \(\hat{\pi}_1\) is an intuitive estimator since in the case of full response we would have \(y_1\) in place of \(\frac{\alpha_{12}}{\alpha_{22}} n_2\).

The generalization of the above analysis to the case of \(k\) categories, \(A_1, \ldots, A_k\) (\(k \geq 2\)), is straightforward. Tables (1.1) and (1.2) are replaced respectively by:

\[
\begin{array}{ccc}
\hat{\pi}_1 & n_1 & n_2 \\
A_1 & \vdots & \vdots \\
& \vdots & \vdots \\
A_k & x_k & \\
& r_1 & n_2 & n \\
\end{array}
\]

\[
\begin{array}{ccc}
\hat{\pi}_1 & p_{11} & p_{12} & \Pi_1 \\
A_1 & \vdots & \vdots & \vdots \\
& \vdots & \vdots & \vdots \\
A_k & p_{k1} & p_{k2} & \Pi_k \\
& q & 1 - q & 1 \\
\end{array}
\]
The parameter of interest is now $\Pi = (\Pi_1, \ldots, \Pi_k)$, and the data is $X = (x_1, \ldots, x_k, n_2)$. In place of (2.1), a priori, we consider that

$\mathbf{p} = (p_{11}, \ldots, p_{k1}, p_{12}, \ldots, p_{k2}) \sim D(\alpha_{11}, \ldots, \alpha_{k1}, \alpha_{12}, \ldots, \alpha_{k2})$.

(2.10)

Analogous to (2.2) and (2.3) the following reparametrization is considered:

$q = \frac{p_{i1}}{p_{i1} + p_{i2}}, \quad q_{i1} = \frac{p_{i1}}{q}, \quad q_{i2} = \frac{p_{i2}}{1 - q} \quad (i = 1, \ldots, k)$

(2.11)

$Q_1 = (q_{i1}, \ldots, q_{k1})$, and $Q_2 = (q_{i2}, \ldots, q_{k2})$.

Conversely,

$p_{i1} = qq_{i1}, \quad p_{i2} = (1 - q)q_{i2} \quad (i = 1, \ldots, k)$

(2.12)

and $\Pi = qQ_1 + (1 - q)Q_2$.

With the reparametrization (2.11) the likelihood is given by:

$L = q^{n_1}(1 - q)^{n_2} \frac{k}{\Pi} \frac{x_i}{q_{i1}}$

(2.13)

Again, by Lemma 1, to consider (2.10) a priori is equivalent to considering the following set of conditions:

$q \perp \perp Q_1 \perp \perp Q_2, \quad q \sim B(\alpha_{11}, \alpha_{22})$,

(2.14)

$Q_1 \sim D(\alpha_{11}, \ldots, \alpha_{k1})$, and $Q_2 \sim D(\alpha_{12}, \ldots, \alpha_{k2})$. 

where $\gamma_j = \sum_i \alpha_{ij} (j = 1, 2)$.

By matching (2.14) with (2.13), we obtain the posterior distribution which is defined by the conditions below.

$$q \perp \perp Q_1 \perp \perp Q_2 | X, q | X \sim q | n_1 \sim B(\alpha_{11} + n_1, \alpha_{12} + n_2),$$

(2.15) $Q_1 | X \sim D(\alpha_{11} + x_1, \ldots, \alpha_{k1} + x_1)$, and

$$Q_2 | X \sim Q_2 \sim D(\alpha_{12}, \ldots, \alpha_{k2}).$$

Again, $p | X$ is distributed as Dirichlet if and only if $n_2 = 0$.

It is, however, a mixture of Dirichlet distributions and

$$(p_{11}, \ldots, p_{k1}, (1 - q)) | X \sim D(\alpha_{11} + x_1, \ldots, \alpha_{k1} + x_1, \alpha_{12} + n_2).$$

As before, we might consider a conjugate family of distributions by taking $\beta \geq \alpha_{12}$ for $\alpha_{12}$ in (2.14).

The Bayes estimator for the parameter of interest

$\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$, analogous to (2.7), has the following form:

(2.16) $\hat{\theta} = E[\theta | X] = \frac{1}{\alpha + n} [(\alpha_{11}, \ldots, \alpha_{k1}) + XM]$

where $M$ is a $(k + 1) \times k$-matrix with the $(k + 1)$th row being $[(\alpha_{12}, \ldots, \alpha_{k2}) \frac{1}{\alpha_{12}}]$, the diagonal elements being the unity, and the remaining elements being zero.

The next section deals with the study of the distribution of the data $X$. The covariance matrix of $\hat{\theta}$ is presented at the end of the section.
3 - THE DIRICHLET - MULTINOMIAL DISTRIBUTION: PROPERTIES

When the discrete data follow the Multinomial model, the family of Dirichlet distributions has been widely used by Bayesians since it is a conjugate family large enough to accommodate various shades of prior opinion. The study of the mixture of Multinomial distributions by a Dirichlet distribution therefore becomes relevant because the (marginal) distribution of the data is then a mixture of this kind. Generalizing the definition of the Beta-Binomial (Ferguson [1967]) this mixture is called here the Dirichlet-Multinomial distribution. More specifically, for \( k \geq 2 \), let \( \mathbf{x} = (x_1, \ldots, x_k) \) be a nonnegative integer random vector such that \( \sum_{i=1}^{k} x_i = n \) is fixed, and let \( \mathbf{p} = (p_1, \ldots, p_k) \) be a nonnegative real random vector with \( \sum_{i=1}^{k} p_i = 1 \).

**DEFINITION**

If \( \mathbf{p} \sim D(\alpha_1, \ldots, \alpha_k) \) and \( \mathbf{x} | \mathbf{p} \sim \text{M}(n; \mathbf{p}) \), then the distribution of \( \mathbf{x} \) is called Dirichlet-Multinomial (DM) with parameter \( (n; \alpha_1, \ldots, \alpha_k) \), and we write \( \mathbf{x} \sim \text{DM}(n; \alpha_1, \ldots, \alpha_k) \). When \( k = 2 \), in place of \( (x_1, x_2) \sim \text{DM}(n; \alpha_1, \alpha_2) \), we write \( x_1 \sim \text{BB}(n; \alpha_1, \alpha_2) \) to indicate that \( x_1 \) is distributed as Beta-Binomial.

It is easy to check that the probability function (p.f.) associated with the DM distribution is given by:

\[
(3.1) \quad f(\mathbf{x}) = \frac{n! \prod \Gamma(\alpha_i)}{\Gamma(\alpha + n)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + x_i)}{x_i! \Gamma(\alpha_i)} ,
\]

where \( \alpha = \sum_{1}^{k} \alpha_i \).
Some of the important properties of the DM distributions are given below. Let \( x = (x_1, \ldots, x_k) \sim \text{DM}(n; \alpha_1, \ldots, \alpha_k) \).

**Proposition 1**

If \((i_1, \ldots, i_k)\) is a permutation of \((1, \ldots, k)\), then
\[
(x_{i_1}, \ldots, x_{i_k}) \sim \text{DM}(n; \alpha_{i_1}, \ldots, \alpha_{i_k}).
\]

**Proposition 2**

If \(m \in \{1, 2, \ldots, k\}\) is fixed, then for \(\beta = \frac{\nu_m}{\nu_1} \alpha_1\)
\[
(x_1, \ldots, x_m, n - \frac{\nu_m}{\nu_1} x_1) \sim \text{DM}(n; \alpha_1, \ldots, \alpha_m, \alpha - \beta),
\]
and
\[
n_1 = \sum_{i=1}^{m} x_i \sim \text{BB}(n; \beta, \alpha - \beta).
\]

These two results are immediate consequences of analogous properties of the Multinomial and the Dirichlet distributions.

**Proposition 3**

For \(m\) and \(n_1\) defined as above, we have that
\[
(x_1, \ldots, x_m) | n_1 \sim \text{DM}(n_1; \alpha_1, \ldots, \alpha_k).
\]

**Proof**

Note that the conditional probability function of
\[
(x_1, \ldots, x_m) | n_1 \text{ is obtained by dividing the p.f. of }
(x_1, \ldots, x_m, n - n_1) \text{ by the p.f. of } n_1, \text{ which is the p.f. of a DM}(n; \alpha_1, \ldots, \alpha_k).
\]

The result we present next is an important characterization of the DM distribution which will be used in the sequel.

Let \((x_1, \ldots, x_k)\) be a nonnegative integer random vector with \(\sum_{1}^{k} x_i = n\) fixed. Choose an integer \(m \in \{2, \ldots, k - 1\}\), and
-14-

denote \( n_1 = \sum_{i=1}^{m} x_i \) with \( n_2 = n - n_1 \). Consider now the following set of conditions:

(i) \((x_1, \ldots, x_m) \mid (x_{m+1}, \ldots, x_k) \mid n_1\)

(ii) \((x_1, \ldots, x_m) \mid n_1 \sim \text{DM}(n_1; a_1, \ldots, a_m), \quad (x_{m+1}, \ldots, x_k) \mid n_1 \sim \text{DM}(n_2; a_{m+1}, \ldots, a_k),\)

and (iii) \( n_1 \sim \text{BB}(n; \sum_{i=1}^{m} x_i, \alpha - \sum_{i=1}^{m} a_i). \)

**THEOREM 1**

The above set of conditions (3.2) are necessary and sufficient to have:

(iv) \((x_1, \ldots, x_k) \sim \text{DM}(n; a_1, \ldots, a_k).\)

**PROOF**

By Propositions 1, 2, and 3 (iv) \(\Rightarrow\) (ii), and (iii). To conclude the remaining implications we only note that (3.1) may be factorized as:

\[
f(x) = \frac{n! \Gamma(n) \Gamma(n_1)}{\Gamma(n + m) n_1! n_2!} \prod_{i=1}^{m} \frac{\Gamma(n_1 + a_i + x_i) \Gamma(\alpha - x_i)}{\Gamma(\alpha - n_1) \Gamma(\alpha + n_2)} \prod_{i=m+1}^{k} \frac{\Gamma(n_2 + a_i + x_i) \Gamma(\alpha - x_i)}{\Gamma(\alpha - n_2) \Gamma(\alpha + n_2)}
\]

where, as before, \( \alpha = \sum_{i=1}^{k} a_i \), and \( \beta = \sum_{i=1}^{m} a_i \). The first factor is the p.f. of a BB\((n; \beta, \alpha - \beta), \) the second is the p.f. of a DM\((n_1; a_{m+1}, \ldots, a_k), \) and the third is the p.f. of a DM\((n_2; a_{m+1}, \ldots, a_k). \)

\( \Box \)
EXAMPLE

Recalling the Bayes estimator \( \hat{n}_1 \) presented in (2.7), we notice that \((x_1, x_2) \mid (y_1, y_2) \mid n_1\), and then
\[ y_1 \mid x \sim y_1 \mid n_2 \sim \text{BB}(n_2; \alpha_{12}, \alpha_{22}) \]
which implies (see (3.3) below) that \( E\{y_1 \mid X\} = E\{y_1 \mid n_2\} = \frac{n_2 \alpha_{12}}{\alpha_{22}}. \)

An interesting property of the DM distribution is given below

where we consider the finite sequence \((z_1, \ldots, z_k)\) with
\[ z_j = \sum_{i=1}^{j} x_i \quad (j = 1, \ldots, k). \]
Clearly, \( z_1 = x_1 \), \( z_m = n_1 \), and \( z_k = n \).

COROLLARY

If \((x_1, \ldots, x_k) \sim \text{DM}(n; \alpha_1, \ldots, \alpha_k)\), then \((z_1, \ldots, z_k)\) forms a Markov Chain.

It is intuitive that we might give a characterization of the DM distribution in terms of \((z_1, \ldots, z_k)\). This, however, would go beyond our needs.

To present the mean vector and the covariance matrix of the DM distribution we introduce the vector \( a = (\alpha_1, \ldots, \alpha_k) \), and the matrix
\[
A = \begin{bmatrix}
\alpha_1 & 0 & \cdots & 0 \\
0 & \alpha_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \alpha_k
\end{bmatrix}
\]

From Propositions 1 and 2, we notice that \( x_i \sim \text{BB}(n; \alpha_i, \alpha - \alpha_i) \), and \( x_i + x_j \sim \text{BB}(n; \alpha_i + \alpha_j, \alpha - \alpha_i - \alpha_j) \) for \( i, j = 1, \ldots, k \) with
i \neq j$. From easy computations when using the definition of $BB$ we have that

$$E(x_i) = n \frac{\alpha_i}{\alpha}$$

$$\text{Var}(x_i) = \alpha_i - \frac{\alpha_i^2}{\alpha} \frac{\alpha + n}{\alpha(\alpha + 1)}$$

and

$$\text{Var}(x_i + x_j) = [\alpha_i + \alpha_j - \frac{(\alpha_i + \alpha_j)^2}{\alpha}] \frac{\alpha + n}{\alpha(\alpha + 1)} = \text{Var}(x_i) + \text{Var}(x_j) + 2 \text{cov}(x_i, x_j).$$

From this last equation, it follows that

$$\text{cov}(x_i, x_j) = \frac{\alpha_i \alpha_j}{\alpha} \frac{\alpha + n}{\alpha(\alpha + 1)}.$$

Finally, the mean vector and the covariance matrix are given by:

$$E(x) = \frac{n}{\alpha} a$$

$$\text{Cov}(x) = [A - \frac{1}{\alpha} a' a] \frac{\alpha + n}{\alpha(\alpha + 1)}$$

where $a'$ is the transpose of $a$.

The data vector $X = (x_1, \ldots, x_k, n_2)$, for the nonresponse data presented in Section 2, follows the DM model; that is, $X \sim DM(n; \alpha_{11}, \ldots, \alpha_{k1}, \alpha_{2})$. In this case
(3.4) \( a = (a_{11}, \ldots, a_{k1}, a_{.2}) \), and
\[
A = \begin{bmatrix}
  a_{11} & 0 & \cdots & 0 & 0 \\
  0 & a_{21} & \cdots & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  0 & 0 & \cdots & a_{k1} & 0 \\
  0 & 0 & \cdots & 0 & a_{.2}
\end{bmatrix}
\]

The mean vector and the covariance matrix for \( \hat{\mu} \), the Bayes estimator given by (2.16), are:
\[
E(\hat{\mu}) = \frac{1}{\alpha + n} \begin{bmatrix} a_{11}, \ldots, a_{k1}\end{bmatrix} + E(X)M, \quad \text{and}
\]

(3.5)
\[
Cov(\hat{\mu}) = \left(\frac{1}{\alpha + n}\right)^2 M' \text{Cov}(X)M
\]

Using (3.4), we have that
\[
\begin{align*}
E(\hat{\mu}) &= \frac{1}{\alpha} \begin{bmatrix} a_{11}, \ldots, a_{k1}\end{bmatrix}, \\
M' \text{M} &= \begin{bmatrix} a_{11} & \cdots & 0 \\
0 & \ddots & \vdots \\
0 & \cdots & a_{k1} \end{bmatrix} + \frac{1}{a_{.2}} (a_{12}, \ldots, a_{k2})' (a_{12}, \ldots, a_{k2}),
\end{align*}
\]

and
\[
\frac{1}{\alpha} M' \text{M} a' a M = \frac{1}{\alpha} (a_{11}, \ldots, a_{k1})' (a_{11}, \ldots, a_{k1}),
\]

which imply
\[
\text{Var}(\hat{\mu}_1) = \frac{n}{\alpha(\alpha + 1)(\alpha + n)} \left[\alpha_{11} + \frac{a_{12}^2}{a_{.2}} - 1\right]
\]

(3.6)
\[
\text{Cov}(\hat{\mu}_i, \hat{\mu}_j) = \frac{n}{\alpha(\alpha + 1)(\alpha + n)} \left[\frac{a_{12} a_{j2}}{a_{.2}} - 1\right]
\]

for \( i, j = 1, \ldots, k \) and \( i \neq j \).
To conclude this section we notice that some of the important discrete distributions are particular cases of the DM distribution. For example, for $a_1 = \ldots = a_k = 1$, the probability function is $f(x) = \binom{n + k - 1}{n}$ which goes by the name Bose-Einstein statistic in Statistic Mechanics. See Feller (1968) for additional discussion.

4 - THE DM DISTRIBUTION: A NATURAL FAMILY OF PRIORS FOR FINITE POPULATION STUDIES

A sample of fixed size $n$ is taken from a population of finite size $N$ which is partitioned in $k \leq N$ categories. The category frequency counts are represented by $\theta_1, \ldots, \theta_k$ with $\sum_1^k \theta_i = N$. From the sample, an inference about $(\theta_1, \ldots, \theta_k)$ is required. Corresponding to each $\theta_i (i = 1, \ldots, k)$, $x_i$ is the sample frequency count of the $i$-th category where $\sum_1^k x_i = n$.

The above problem may be viewed in a simple way by considering a bag with $N$ balls of $k \leq N$ different colors that are identified by $c_1, c_2, \ldots, c_k$. The number of balls with the $i$-th color ($i = 1, \ldots, k$) is represented by $\theta_i$ where $\sum_1^k \theta_i = N$. Suppose that this $N$ balls are separated into two bags in such a way that one bag (bag number 1) contains $n$ balls, and the other (bag number 2) $N-n$ balls. The statistician is allowed to look at the composition of bag 1 and record the numbers $x_1, \ldots, x_k$, the frequency counts of the $k$ colors. Now, the unknown quantities for the statistician are $\theta_1 - x_1, \ldots, \theta_k - x_k$; that is, the composition of bag 2.
We restrict the choice of the prior distribution for $\theta$ to the family of DM distributions. Given the sample $x = (x_1, \ldots, x_k)$, the composition of the first bag, we want to derive the posterior distribution of $\theta - x = (\theta_1 - x_1, \ldots, \theta_k - x_k)$, the composition of the second bag. In order to reach this goal, we only use intuitive arguments since, an algebraic analysis, besides being tedious (albeit easy), would bury the beauty of the argument.

Let $e_1 = (1, 0, \ldots, 0), \ldots, e_k = (0, \ldots, 0, 1)$ be the standard orthonormal basis for $\mathbb{R}^k$. To each unit $j$ ($j = 1, \ldots, N$) of the population $P$, we associate an incidence vector $y_j$ which is equal to $e_i$ if the color of $j$ is $c_i$. More specifically, let $P = \{1, \ldots, N\}$ be an enumeration of the population of balls. Associated with $P$ are the incidence vectors $y_1, \ldots, y_N$ described above. The unknown vector is $\theta = (\theta_1, \ldots, \theta_k) = \sum_{j=1}^{N} y_j$. We are considering the case where the sampling selection is noninformative. That is, the selection of the $n$ balls (sample) from $P$, is based only on the labels $1, \ldots, N$, which are themselves uninformative about the incidence vectors $y_1, \ldots, y_N$.

A natural way to introduce the prior model $\theta \sim \text{DM}(N; \alpha_1, \ldots, \alpha_k)$ is to consider a random vector $p = (p_1, \ldots, p_k) \sim D(\alpha_1, \ldots, \alpha_k)$, and to stipulate that for $j = 1, \ldots, N$

$$ y_j | p \sim \text{M}(1; p), \text{ and } y_1 \perp \ldots \perp y_N | p. $$
In other words, given \( p \) the \( y_j \)'s are i.i.d. with common distribution \( M(1; p) \). Since \((y_1, \ldots, y_N)\) is an exchangeable finite sequence, without loss of generality, we can consider our sampled items as being the first \( n \) population items, say \( \{1, 2, \ldots, n\} \). That is, the two bags are \( \{1, \ldots, n\} \) and \( \{(n+1), \ldots, N\} \). Now, the sample is represented by the vector \( \mathbf{x} = \sum_{j=1}^{N} y_j \), and the unknown quantity of interest is the vector \( \mathbf{\theta} - \mathbf{x} = \sum_{j=n+1}^{N} y_j \).

In terms of the pseudo parameter \( p \), we then have, a priori, the following:

(i) \( p \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k) \),

(ii) \( \mathbf{x} | p \sim \mathcal{M}(n; p) \),

(iii) \( \mathbf{x} \sim \mathcal{D}(\mathcal{M}(n; \alpha_1, \ldots, \alpha_k), \mathcal{D}(\mathcal{M}(N-n; \alpha_1, \ldots, \alpha_k)), \mathcal{D}(\mathcal{M}(N-n; \alpha_1, \ldots, \alpha_k)), \mathcal{D}(\mathcal{M}(N-n; \alpha_1, \ldots, \alpha_k))) \),

(iv) \( (\mathbf{\theta} - \mathbf{x}) | p \sim \mathcal{M}(N - n; p) \),

(v) \( (\mathbf{\theta} - \mathbf{x}) \sim \mathcal{D}(\mathcal{M}(N - n; \alpha_1, \ldots, \alpha_k)) \),

(vi) \( (\mathbf{\theta} - \mathbf{x}) \prod \mathbf{x} | p \).

and (vii) \( \mathbf{\theta} | p \sim \mathcal{M}(N; p) \).

The result below is useful to our discussion.

**Lemma 2**

If two independent random vectors \( X \) and \( Y \) are such that \( X \sim \mathcal{M}(n_1; p) \) and \( Y \sim \mathcal{M}(n_2; p) \), then the conditional distribution of \( X | X + Y \) is the Multivariate Hypergeometric with parameter \( (n_1 + n_2, n_1, X + Y) \); that is, \( X | X + Y \sim \mathcal{H}(n_1 + n_2, n_1, X + Y) \). (Note that this distribution does not depend on the value of \( p \).)
The following conclusion based on (4.1), and Lemma 2 is the first important result since defines the likelihood function:

\[(4.2) \quad x| (\theta, p) \sim x| \theta \sim H(N, n, \emptyset)\]

From the Bayesian analysis of the multinomial case, we recall that \(p|x \sim D(\alpha_1 + x_1, \ldots, \alpha_k + x_k)\). On the other hand we notice that the conditional distribution of \((\theta - x)|x\) may be viewed as a composition of the distribution of \((\theta - x)|p \sim (\theta - x)|(p, x)\) (see (4.1)) by the distribution of \(p|x\). Now, by the definition of the DM distribution, we have that \((\theta - x)|x \sim DM(N - n; \alpha_1 + x_1, \ldots, \alpha_k + x_k)\). This is the main result of this section and may be summarized as:

**THEOREM 2**

For the finite population sampling described, if 
\(\theta \sim DM(N; \alpha_1, \ldots, \alpha_k)\) a priori, then \((\theta - x)|x \sim \) 
\(DM(N - n; \alpha_1 + x_1, \ldots, \alpha_k + x_k)\) a posteriori.

The next section is devoted to the nonresponse problem in finite populations.

5 - **NONRESPONSE: THE MULTIVARIATE HYPERGEOMETRIC MODEL**

The data for the nonresponse problem is presented in the \(k \times 2\)-tabular form as in (2.7). Instead of (1.4), the population parameters have the following representation:
(5.1) \[
\begin{array}{c|ccc}
 & R & R' \\
\hline
A_1 & \theta_{11} & \theta_{12} & \theta_1 \\
A_2 & \theta_{21} & \theta_{22} & \theta_2 \\
\vdots & \vdots & \vdots & \vdots \\
A_k & \theta_{k1} & \theta_{k2} & \theta_k \\
\hline
\psi & N - \psi & N \\
\end{array}
\]

Now, the parameter of interest is \( \theta = (\theta_1, \ldots, \theta_k) \), and the likelihood may be written as:

\[
L = \frac{k}{n_2} \prod_{i=1}^{k} \frac{\psi}{n_1} \left( \frac{\theta_i}{(n_i)} \right)^{n_i} \left( N - \psi \right)^{N - n_2 - \sum_{i=1}^{k} n_i}
\]

(5.2)

Suppose that, a priori, a DM distribution for
\( \theta = (\theta_{11}, \ldots, \theta_{k1}, \theta_{12}, \ldots, \theta_{k2}) \) is considered; that is, a priori

(5.3) \( \theta \sim \text{DM}(N; \alpha_{11}, \ldots, \alpha_{k1}, \alpha_{12}, \ldots, \alpha_{k2}) \).

An additional notation is introduced below:

\( \theta_1 = (\theta_{11}, \ldots, \theta_{k1}) \), and \( \theta_2 = (\theta_{12}, \ldots, \theta_{k2}) \).

Recall that, \( \theta = \theta_1 + \theta_2 \) is the parameter of interest,

\( x = (x_1, \ldots, x_k, n_2) \) is the data vector which may also be represented in the slightly abbreviated form \( x = (x_1, \ldots, x_k) \).

\( \alpha_{ij} = \sum_{i=1}^{k} \alpha_{ij} \) (j = 1, 2), \( \alpha_1 = \alpha_{11} + \alpha_{12} \) (i = 1, \ldots, k), and \( \alpha = \sum \alpha_{ij} \).
Writing the parameters in the extended form \((\psi, \theta_1, \theta_2)\), it is convenient to describe the prior (5.3) in the following equivalent form (see Theorem 1):

\[
(i) \quad \psi \sim BB(N; \alpha_1, \alpha_2),
\]

(5.4) \(\quad (ii) \quad \theta_1 | \psi \sim DM(\psi; \alpha_{11}, \ldots, \alpha_{k1}), \quad \theta_2 | \psi \sim DM(N - \psi; \alpha_{12}, \ldots, \alpha_{k2}), \)

and (iii) \(\theta_1 \perp \perp \theta_2 | \psi\)

The Theorem presented below is the main result of this section. It allows a simple derivation of the Bayes estimator.

**THEOREM 3**

The posterior distribution derived from the Bayes operation, when (5.2) is the likelihood and (5.4) defines the prior, is given by the following set of conditions:

\[
(i)' \quad (\psi - n_1\mid x \sim (\psi - n_1) | n_1 \sim BB(N - n; \alpha_{11} + n_1, \alpha_{2} + n_2)
\]

(5.5) \(\quad (ii)' \quad (\theta_1 - x\mid (\psi, x) \sim DM(\psi - n_1; \alpha_{11} - x_1, \ldots, \alpha_{k1} - x_k), \quad \theta_2 | (\psi, x) \sim \theta_2 | \psi, \)

and (iii)' \(\theta_1 \perp \perp \theta_2 | (\psi, x)\)

**PROOF**

The second condition of (ii)' follows from the fact that the likelihood does not depend on \(\theta_2\) when \(\psi\) is known. This fact together with the prior condition (iii), implies (iii)'.

To prove (i)' and the first condition of (ii)', we consider (as in Section 2) the invisible nonresponse sample frequency counts,
say \( y = (y_1, \ldots, y_k) \). If we had full response, the data would have been represented by \((x, y)\). From Theorems 1 and 2 we have that

\[
a) \quad \psi - n_1 \mid (x, y) \sim BB(N - n_1; \alpha_{11} + n_1, \alpha_{2} + n_2)
\]

\[
b) \quad \theta - x \mid (\psi, x, y) \sim DM(\psi - n_1; \alpha_{11} + x_1, \ldots, \alpha_{kl} + x_k)
\]

From a) and b) it follows that \( (\psi - n_1) \mid (x, y) \sim (\psi - n_1) \mid n_1 \), and that \( \theta - x \mid (\psi, x, y) \sim \theta - x \mid (\psi, x) \) which imply (i)' and first condition of (ii)' respectively. □

Note that we showed above that \( \psi \perp\!\!\!\perp x \mid n_1 \); that is, \( n_1 \) is partially Bayes sufficient to predict \( \psi \). See Basu (1977) for a more complete discussion of this concept.

As in the multinomial case, the posterior (5.5) does not define a distribution in the same class in which the prior was chosen from; that is, (5.5) does not define a DM distribution. It is easy to check, however, that \((\theta_{11} - x_1, \ldots, \theta_{kl} - x_k, N - \psi - n_2) \sim DM(N - n; \alpha_{11} + x_1, \ldots, \alpha_{kl} + x_k, \alpha_{2} + n_2)\). A more complete class might be considered by taking in (5.4) a \( \beta \geq \alpha_{2} \) for \( \alpha_{2} \) in (i).

From the posterior (5.5) we obtain the following results:

\[
E\{\psi - n_1 \mid X\} = (N - n) \frac{\alpha_{11} + n_1}{\alpha + n}
\]

\[
E\{N - \psi \mid X\} = n_2 + (N - n) \frac{\alpha_{2} + n_2}{\alpha + n}
\]

\[
E\{\theta_{11} - x_1 \mid (\psi, X)\} = (\psi - n_1) \frac{\alpha_{11} + x_1}{\alpha_{11} + n_1}
\]
Using now the properties of conditional expectation we have the Bayes estimators,

\[
\hat{\theta}_i = E(\theta_i | X) = E(\theta_{i1} + \theta_{i2} | X) \\
= E(\theta_{i1} | X) + E(\theta_{i2} | X) \\
= x_i + \frac{\alpha_{i1}}{a_{i1} + n_1} E(\psi - n_1 | X) + \frac{\alpha_{i2}}{a_{i2}} E(N - \psi | X) \\
= \frac{a + N}{a + n} (x_i + n_2 \alpha_{i2}) + (N - n) \frac{\alpha_i}{a + n}
\]

Similarly to (2.15), the Bayes estimator of the parameter of interest \( \theta = \theta_1 + \theta_2 \) is given by:

\[
\hat{\theta} = E(\theta | X) = \frac{a + N}{a + n} x_M + \frac{N - n}{a + n} (\alpha_{1,}, \ldots, \alpha_k)
\]

Using the results (3.4), and (3.5) we finally have:

\[
E(\hat{\theta}) = \frac{N}{a_1} (\alpha_{1,}, \ldots, \alpha_k)
\]

\[
Cov(\hat{\theta}) = \left( \frac{a + N}{a + n} \right)^2 M' Cov(X)M
\]

which implies that

\[
(5.6) \quad Cov(\hat{\theta}_i, \hat{\theta}_j) = \frac{n(a + N)^2}{(a + n)(a + 1)a} \left( \delta_{ij} \frac{\alpha_{i1}}{a_{i1}} + \frac{\alpha_{i2}}{a_{i2}} - \frac{\alpha_i}{a} \right)
\]
where $\delta_{ij}$ is the Kronecker's delta, and $\text{Var}\{\theta_i\} = \text{Cov}\{\theta_i, \theta_i\}$.

6 - FINAL REMARKS

(i) There are many follow-up techniques used to obtain response among some of the $n_2$ units that have not responded initially. For example, from the $n_2$ nonrespondents in our sample, we select a subsample of size $n'_2 \leq n_2$ and offer an incentive to those who now would respond. In that way, information about $Q_2$ in (2.14) or about $\theta_2|\psi$ in (5.4) might be improved. See Kaufman and King (1973), and Singh and Sedrausk (1978) for a more specific discussion on this two stage sampling.

(ii) Although we have restricted ourselves to the nonresponse problem, it should be understood that our method applies equally well to the general problem of categorical data with missing entries. Consider, for instance, the categorical data where all but the first $n$ cell entry data are missing. By using Lemma 1 for the multinomial case or Theorem 1 for the hypergeometric case, we would, analogously to (2.6) or (5.4), obtain the posterior distribution for the cell parameters.

(iii) One word about the relevance of the variance of Bayes estimators as presented in (3.6), and (5.6). Note that we are not talking about conditional variances (with the parameter fixed) but the variance of the marginal distribution of the estimator. Consider the $k = 2$ multinomial case for instance. It is clear that $\text{Var}\{\hat{\Pi}_1\} = \text{Var}\{\Pi_1\} - E\{\text{Var}\{\Pi_1|X\}\}$; that is, the variance of $\hat{\Pi}_1$ may be
regarded as the expected amount of uncertainty removed, when uncertainty (De Groot (1962)) about the parameter is measured by its variance. Thus, the variance of the Bayes estimator is a kind of a measure of the amount of information in the experiment. The larger the variance of \( \hat{\theta}_1 \) is, the better off we are!
REFERENCES


