97-F-29

Some Superpopulation Models for Estimating the Number of Population Uniques

Akimichi Takemura
University of Tokyo

September 1997

Some superpopulation models for estimating the number of population uniques

Akimichi Takemura
Faculty of Economics, University of Tokyo

September, 1997

Abstract

The number of the unique individuals in the population is of great importance in evaluating the disclosure risk of a microdata set. We approach this problem by considering some basic superpopulation models including the gamma-Poisson model of Bethlehem et al. (1990). We introduce Dirichlet-multinomial model which is closely related but more basic than the gamma-Poisson model, in the sense that binomial distribution is more basic than Poisson distribution. We also discuss the Ewens model and show that it can be obtained from the Dirichlet-multinomial model by a limiting argument similar to the law of small numbers. The multivariate Ewens distribution is a basic mathematical model used in genetics. Estimation of the number of the population uniques is particularly simple under the Ewens model.

Although these models might not necessarily well fit actual populations, they can be considered as basic mathematical models for our problem, as binomial and Poisson distributions are considered as basic models for count data.

Key words: Dirichlet distribution, Ewens sampling formula, microdata, multinomial distribution, statistical disclosure control

1 Introduction

The number of the population uniques is one of the key quantities in evaluating the disclosure risk of a microdata set. For example Willenborg and de Waal (1996) discuss the notion of the population uniques repeatedly throughout their monograph. In this paper we adopt a parametric approach to this problem in the form of superpopulation models. Although there may be some criticisms against superpopulation models as being hypothetical, the models discussed in this paper have nice mathematical theory and give many insights into the problem.

The plan of this paper is as follows. In this section we give a mathematical formulation of our problem and set up appropriate notations. In Section 2 we propose Dirichlet-multinomial model and investigate its properties. In Section 3 we discuss the gamma-Poisson model by Bethlehem et al. (1990) and compare it with the Dirichlet-multinomial model. We show that if we fix the population sample size N and consider the conditional

model given N in the gamma-Poisson model, then we obtain our Dirichlet-multinomial model. In section 4 we derive the multivariate Ewens distribution from the Dirichlet-multinomial distribution by a limiting argument similar to the law of small numbers. We also show that estimation of the number of population uniques is particularly simple in the Ewens model.

1.1 Notation and formulation of our problem

Consider a finite population of N individuals

$$\{y_1, y_2, \ldots, y_N\}$$

where y_i is the value of the characteristic of interest of the i's individual. In this paper we only consider categorical variable y. Let K be the total number of categories or cells. For simplicity let the cells be numbered as $1, \ldots, K$. $y_j = i$ means that the j-th individual falls in the i-th cell. Let F_i denote the population frequency of the cell i

$$F_i = \sharp\{j \mid y_j = i\}$$

and let $\boldsymbol{F}$ denote the population frequency vector

$$\boldsymbol{F} = (F_1, \ldots, F_K).$$

If $F_i = 1$ then there exists a unique individual falling in the i-th cell. This individual is called population unique. Our problem is how to estimate the number of the population uniques

$$\theta = S_1 = \sharp\{i \mid F_i = 1\}.$$

When a sample of size n is drawn from the population, $\hat{\theta}$ denotes the estimator of θ based on the sample.

In addition to the population uniques, it might be necessary to consider some more rare individuals. For example we might consider "population doubles", who fall in a cell i with $F_i = 2$. In general we consider the number of the cells of size l. Let

$$S_l = \sharp\{i \mid F_i = l\}$$

and

$$\boldsymbol{S} = (S_0, S_1, \ldots, S_N).$$

$\boldsymbol{S}$ is called size index vector (Sibuya (1993b), Sibuya and Yamato (1995)) or frequencies of frequencies (Good (1965)). In the classical occupancy problem (see Section IV.2 of Feller (1968) or Korwar (1988)) the parameter of interest is $n - S_0$, which is the number of occupied cells. On the other hand S_1 is of the primary interest in microdata disclosure problem.

When a particular sample of size n has been drawn, it may be more relevant to estimate the number of the population uniques included in the sample, rather than the total number of the uniques in the population. However under simple random sampling, an obvious estimator of the number of the population uniques included in the sample

is given by $\hat{\theta}n/N$. With this simple estimator in mind, we only discuss the problem of estimating θ.

In microdata disclosure problem the total number of the cells K is usually very large and the estimation of the whole frequency vector $\boldsymbol{F} = (F_1, \ldots, F_K)$ seems to be difficult. One approach to cope with this difficulty is the superpopulation approach, where $\boldsymbol{F}$ is considered as a realization of a random vector whose distribution (prior distribution) is determined by a small number of hyperparameters. This prior distribution can be interpreted as a hypothetical sampling from a larger superpopulation. Let τ denote the hyperparameter and let $p(\boldsymbol{F} \mid \tau)$ denote the probability mass function of $\boldsymbol{F}$. Under a particular superpopulation model with given τ, the number of population uniques $\theta = S_1$ can be estimated by its expected value with respect to the prior distribution $E_\tau(S_1) = \sum_{i=1}^{K} P(F_i = 1 \mid \tau)$. Therefore if we can estimate the hyperparameter τ from the sample by $\hat{\tau}$, then a natural estimator of θ is given as

$$\hat{\theta} = E_{\hat{\tau}}(S_1) = \sum_{i=1}^{K} P(F_i = 1 \mid \hat{\tau}).$$

Note that this whole approach is the empirical Bayes approach discussed extensively in statistical literature (see e.g. Carlin and Louis (1996), Maritz and Lwin (1989)).

The gamma-Poisson model[1] by Bethlehem et al. (1990) is a primary example of superpopulation models proposed for the study of disclosure risk problem. We will clarify why this model is of basic importance for our problem by introducing Dirichlet-multinomial model, which is more basic than gamma-Poisson model in the sense that binomial distribution is more basic than Poisson distribution. A slightly annoying and confusing nature of the gamma-Poisson model is that the population size N is a random variable in the prior distribution. In our Dirichlet-multinomial model N is fixed and the model is less confusing.

By a limiting argument similar to the law of small numbers we obtain the multivariate Ewens distribution from the Dirichlet-multinomial distribution in Section 4. The multivariate Ewens distribution (called the Ewens sampling formula in genetics) is a basic mathematical model of random clustering. See e.g. Ewens (1990), Sibuya (1993b), Sibuya and Yamato (1995), or Chapter 41 of Johnson et al. (1997). We show that estimation of the number of the population uniques is particularly simple under the multivariate Ewens distribution.

1.2 Sampling

Consider sampling of n ($n < N$) individuals from a finite population. The sampling design we consider is the simple random sampling without replacement. In addition we consider Bernoulli sampling (Section 2.2 of Särndal et al. (1992)) in Appendix A in order to clarify some properties of the gamma-Poisson model.

Let

$$\boldsymbol{f} = (f_1, \ldots, f_K)$$

[1]We use the term "gamma-Poisson model" instead of the term "Poisson-gamma model" used in microdata disclosure literature, because beta-binomial or Dirichlet-multinomial distributions are standard terms used in Bayesian literature in general.

be the sample frequency vector of the cells and

$$\boldsymbol{s} = (s_0, s_1, \ldots, s_n)$$

be its size index vector. Under simple random sampling without replacement the probability distribution of $\boldsymbol{f} = (f_1, \ldots, f_K)$ given $\boldsymbol{F} = (F_1, \ldots, F_K)$ is the multivariate hypergeometric distribution with the probability mass function

$$p(f_1, \ldots, f_K) = p(f_1, \ldots, f_K \mid \boldsymbol{F}) = \frac{\binom{F_1}{f_1} \cdots \binom{F_K}{f_K}}{\binom{N}{n}}.$$

Consider a superpopulation model given in terms of the probability mass function $p(F_1, \ldots, F_K \mid \tau)$ where τ is the hyperparameter. Then the joint probability mass function of $\boldsymbol{f}$ and $\boldsymbol{F}$ is given by

$$p(f_1, \ldots, f_K \mid F_1, \ldots, F_K) \times p_F(F_1, \ldots, F_K \mid \tau). \tag{1}$$

In the empirical Bayes approach the hyperparameter τ is usually estimated based on the marginal distribution of the sample. Summing up (1) with respect to $F_1, \ldots, F_K$, the marginal probability mass function of $\boldsymbol{f} = (f_1, \ldots, f_K)$ is given as

$$p_f(f_1, \ldots, f_K \mid \tau) = \sum_{\boldsymbol{F}} p(f_1, \ldots, f_K \mid \boldsymbol{F}) \cdot p_F(\boldsymbol{F} \mid \tau). \tag{2}$$

This marginal probability mass function serves as the likelihood function of τ. We are interested in a convenient superpopulation model, where the summation on the right hand side of (2) can be explicitly evaluated.

A conceptually very simple model is as follows. We draw n individuals from the realized finite population of size N. Suppose that we can think of these n individuals as directly generated from the superpopulation. Then as far as the marginal distribution is concerned we can forget the intermediate finite population. In other words we look for a sufficient condition such that the following equality holds:

$$\sum_{\boldsymbol{F}} p(f_1, \ldots, f_K \mid \boldsymbol{F}) \cdot p_F(\boldsymbol{F} \mid \tau) = p_F(f_1, \ldots, f_K \mid \tau).$$

A simple sufficient condition can be stated in terms of the exchangeability of the prior distribution with respect to the individuals.

Suppose that the prior distribution of the values of the population $\boldsymbol{y} = (y_1, \ldots, y_N)$ is exchangeable with respect to the individuals, i.e.,

$$(y_1, \ldots, y_N) \stackrel{\mathrm{d}}{=} (y_{i_1}, \ldots, y_{i_N}),$$

where $(i_1, \ldots, i_N)$ is an arbitrary permutation of $(1, \ldots, N)$ and $\stackrel{\mathrm{d}}{=}$ denotes the equality of the distributions. We draw n individuals from the realized finite population by simple random sampling without replacement. Suppose that $j_1, \ldots, j_n$ are the labels of the individuals drawn from the population. By the assumed exchangeability, the distribution of $y_{j_1}, \ldots, y_{j_n}$ is the same as the distribution of $y_1, \ldots, y_n$. Furthermore this equality does not depend on the values of $j_1, \ldots, j_n$. Therefore we have the following basic lemma.

Lemma 1 *Suppose that the prior distribution of the values of N individuals is exchangeable with respect to the individuals. Let n individuals be drawn from the realized population with simple random sampling without replacement. Then the marginal distribution of the values of n individuals coincides with the prior distribution of values of n individuals directly drawn from the superpopulation.*

For example suppose that $y_1, \ldots, y_N$ are independent and identical (multivariate) Bernoulli trials such that $\boldsymbol{F} = (F_1, \ldots, F_K)$ has the multinomial distribution $\mathrm{Mult}(N, \pi_1, \ldots, \pi_K)$. Then the marginal distribution of the sample frequency vector $\boldsymbol{f} = (f_1, \ldots, f_K)$ is again the multinomial distribution $\mathrm{Mult}(n, \pi_1, \ldots, \pi_K)$.

We can also consider mixture of multinomial distributions. Let μ denote a probability distribution on the simplex $\mathcal{S} = \{(\pi_1, \ldots, \pi_K) \mid \pi_i \geq 0, \sum \pi_i = 1\}$ and consider the probability mass function of $\boldsymbol{F}$ of the form

$$p_F(F_1, \ldots, F_K) = \int_{\mathcal{S}} \binom{N}{F_1, \ldots, F_K} \pi_1^{F_1} \cdots \pi_K^{F_K} d\mu(\pi_1, \ldots, \pi_K). \tag{3}$$

Then by Lemma 1 the marginal distribution of the sample frequency vector $\boldsymbol{f}$ is given by

$$p_f(f_1, \ldots, f_K) = \int_{\mathcal{S}} \binom{n}{f_1, \ldots, f_K} \pi_1^{f_1} \cdots \pi_K^{f_K} d\mu(\pi_1, \ldots, \pi_K).$$

A convenient mixture distribution μ for the multinomial distribution is the Dirichlet distribution, which is the natural conjugate prior distribution for the multinomial distribution. This is the Dirichlet-multinomial model discussed in the next section.

Here we note the implication of the well known de Finetti theorem to our problem. The assumption of the exchangeability of $y_1, \ldots, y_N$ in Lemma 1 seems to be a natural one. If we further assume that $(y_1, \ldots, y_N)$ is a part of an infinite sequence of exchangeable random variables for any N, then the distribution of $(y_1, \ldots, y_N)$ has to be a mixture of the multivariate Bernoulli trials, i.e., the prior distribution of $F_1, \ldots, F_K$ is necessarily of the form (3). Diaconis and Freedman (1980) extends the de Finetti theorem and shows that it holds approximately even for finite exchangeable sequences.

Finally we mention the distribution of the sample size index vector. We have discussed distribution of the sample frequency vector $\boldsymbol{f}$. The distribution of the sample size index vector $\boldsymbol{s} = (s_0, \ldots, s_n)$ involves a combinatorial complication and can not be easily treated in general. If the marginal distribution of $\boldsymbol{f}$ is exchangeable with respect to the cells, then it is possible to write down the marginal distribution of $\boldsymbol{s}$ as discussed in Appendix B. The result in Appendix B is used for discussing the multivariate Ewens distribution in Section 4.

2 Dirichlet-multinomial model

We have already given the basic idea for the Dirichlet-multinomial model in the last subsection. We here give more detailed definition of the model. Let

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$$

be a probability vector and let the prior distribution of $\boldsymbol{F} = (F_1, \ldots, F_K)$ (given $\boldsymbol{\pi}$) be the multinomial distribution $\text{Mult}(N, \pi_1, \ldots, \pi_K)$. Furthermore assume that $\boldsymbol{\pi}$ follows the Dirichlet distribution with parameter $\alpha_1, \ldots, \alpha_K$ with the probability density function

$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1}$$

on the simplex $\mathcal{S} = \{(\pi_1, \ldots, \pi_K) \mid \pi_i \geq 0, \sum \pi_i = 1\}$. It is well known that the unconditional prior distribution of $\boldsymbol{F} = (F_1, \ldots, F_K)$ is given by the Dirichlet-multinomial distribution with the probability mass function

$$\begin{aligned} & p(F_1, \ldots, F_K) \\ &= \binom{N}{F_1, \ldots, F_K} \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \int_{\mathcal{S}} \pi_1^{\alpha_1 + F_1 - 1} \cdots \pi_K^{\alpha_K + F_K - 1} d\pi_1 \cdots d\pi_{K-1} \\ &= \binom{N}{F_1, \ldots, F_K} \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \frac{\Gamma(\alpha_1 + F_1) \cdots \Gamma(\alpha_K + F_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K + N)} \\ &= \frac{N! \Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K + N)} \frac{\Gamma(\alpha_1 + F_1)}{\Gamma(\alpha_1) F_1!} \cdots \frac{\Gamma(\alpha_K + F_K)}{\Gamma(\alpha_K) F_K!}. \end{aligned} \tag{4}$$

Note that $\alpha_1, \ldots, \alpha_K$ are the hyperparameters of the prior distribution of $\boldsymbol{F}$.

The Dirichlet-multinomial distribution is studied by many researchers under different names. Mosimann (1962) gave a systematic study of this distribution and called it compound multinomial distribution. Janardan and Patil (1972) and Janardan (1973) gave detailed study of family of distributions including the Dirichlet-multinomial distribution. See Chapter 35 of Johnson et al. (1997) for further references.

Write $A = \sum_{i=1}^K \alpha_i$. The first and the second order moments of the Dirichlet-multinomial distribution are easily evaluated as

$$E(F_i) = N \frac{\alpha_i}{A} \tag{5}$$

and

$$\begin{aligned} \text{Var}(F_i) &= N \frac{\alpha_i}{A} (1 - \frac{\alpha_i}{A}) \cdot \frac{A + N}{A + 1}, \\ \text{Cov}(F_i, F_j) &= -N \frac{\alpha_i}{A} \frac{\alpha_j}{A} \cdot \frac{A + N}{A + 1}. \end{aligned} \tag{6}$$

Furthermore

$$P(F_i = 1) = N \alpha_i \frac{\Gamma(A) \Gamma(A - \alpha_i + N - 1)}{\Gamma(A + N) \Gamma(A - \alpha_i)}$$

and the expected value of the number of the population uniques is given as

$$E(S_1) = N \frac{\Gamma(A)}{\Gamma(A + N)} \sum_{i=1}^K \alpha_i \frac{\Gamma(A - \alpha_i + N - 1)}{\Gamma(A - \alpha_i)}.$$

Now consider simple random sampling without replacement from the realized population. By Lemma 1 the marginal distribution of the sample frequency vector $\boldsymbol{f} =$

$(f_1, \ldots, f_K)$ is again the Dirichlet-multinomial with the same $\alpha_1, \ldots, \alpha_K$ with the probability mass function

$$p(f_1, \ldots, f_K \mid \alpha_1, \ldots, \alpha_K) = \frac{n!\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K + n)} \frac{\Gamma(\alpha_1 + f_1)}{\Gamma(\alpha_1) f_1!} \cdots \frac{\Gamma(\alpha_K + f_K)}{\Gamma(\alpha_K) f_K!}. \tag{7}$$

It is also straightforward to verify (7) by explicitly evaluating the sum in (2). The first and second order moments of $\boldsymbol{f} = (f_1, \ldots, f_K)$ are given by (5) and (6) with N replaced by n.

In the above full formulation of the Dirichlet-multinomial model $\alpha_1, \ldots, \alpha_K$ are free hyperparameters. Since we are considering the situation where K is large, it is desirable to introduce some simplifying assumption on the values of $\alpha_1, \ldots, \alpha_K$. If we assume that all α_i's are equal and put

$$\alpha_1 = \alpha_2 = \cdots = \alpha, \tag{8}$$

then we obtain a model which is exchangeable with respect to the cells. This exchangeable case of the Dirichlet-multinomial model corresponds to the gamma-Poisson model of Bethlehem et al. (1990) and to the multivariate Ewens distribution of Section 4. In this paper we mainly consider this exchangeable case. However the exchangeability of the cells might be an unrealistic assumption for microdata sets as discussed in Section 5.

We now consider estimation of the hyperparameters of the Dirichlet-multinomial model based. Given the sample frequency vector $\boldsymbol{f} = (f_1, \ldots, f_K)$ we can employ maximum likelihood estimation using (7) as the likelihood function. Alternatively we can consider moment estimators based on (5) and (6) with the population size N replaced by the sample size n. Janardan (1976) and Levin and Reeds (1977) investigated the estimation of the Dirichlet-multinomial distribution. In particular, Levin and Reeds (1977) gave a detailed analysis of the likelihood function (7). Here we discuss only salient features of the estimation of the Dirichlet-multinomial distribution based on Janardan (1976) and Levin and Reeds (1977).

We make the following reparameterization. Let

$$\gamma_i = \frac{\alpha_i}{A}, \quad i = 1, \ldots, K, \qquad A = \sum_{i=1}^{K} \alpha_i.$$

Consider letting $A \to \infty$ with γ_i, $i = 1, \ldots, K$, fixed. Then the Dirichlet-multinomial distribution converges to the multinomial distribution $\mathrm{Mult}(N, \gamma_1, \ldots, \gamma_K)$. In this sense $A = \infty$ is a valid parameter value. As we will see below, this causes difficulty in estimation of A, whereas the estimation of γ_i, $i = 1, \ldots, K$, seems to be more straightforward.

We briefly discuss estimation of γ_i's, when they are free parameters subject only to $\sum \gamma_i = 1$. Noting

$$E(f_i) = n\frac{\alpha_i}{A} = n\gamma_i$$

an unbiased estimator of γ_i is given as

$$\hat{\gamma}_i = \frac{f_i}{n}. \tag{9}$$

It can be easily seen that (9) is also approximately equal to the maximum likelihood estimator of γ_i. Therefore (9) seems to be a natural estimator of γ_i, when γ_i's are free

parameters. In the cell exchangeable model (8) $\gamma_i = 1/K, i = 1, \ldots, K$, and there is no need to estimate γ_i's.

We now discuss estimation of A given the values (or estimated values) of γ_i, $i = 1, \ldots, K$. As already mentioned above, the difficulty lies in the possibility that $\hat{A} = \infty$. We consider the behavior of the log likelihood function as $A \to \infty$. When A is large it can be easily shown that the logarithm of (7) is written as

$$\log p(f_1, \ldots, f_K \mid A, \gamma_1, \ldots, \gamma_K) = \text{const} - \frac{1}{2A}[n(n-1) - \sum_{i=1}^{K} \frac{f_i(f_i-1)}{\gamma_i}] + o(\frac{1}{A}).$$

Therefore the behavior of the log likelihood function as $A \to \infty$ depends on the quantity

$$B = n(n-1) - \sum_{i=1}^{K} \frac{f_i(f_i-1)}{\gamma_i}.$$

Levin and Reeds (1977) solved a conjecture of Good (1965) and proved that the likelihood function $p(f_1, \ldots, f_K \mid A, \gamma_1, \ldots, \gamma_K)$ has at most one local maximum in A for given $\gamma_1, \ldots, \gamma_K$. Therefore the maximum likelihood estimator $\hat{A}_{ML}$ is indeed $\hat{A}_{ML} = \infty$ when $B > 0$.

In the cell exchangeable case, $\gamma_i = 1/K$ and $B = n(n-1) - K\sum_{i=1}^{K} f_i(f_i-1)$. In this case B can be simplified as

$$B = K(K-1)(\bar{f} - s_f^2) = n(K - 1 - \chi^2) \tag{10}$$

where

$$s_f^2 = \frac{\sum_{i=1}^{K}(f_i - \bar{f})^2}{K-1} \tag{11}$$

is the sample variance of $\boldsymbol{f} = (f_1, \ldots, f_K)$ and

$$\chi^2 = \sum_{i=1}^{K} \frac{(f_i - \bar{f})^2}{\bar{f}}$$

is the chi-square statistic for testing the equality of the probability of the cells for the multinomial distribution. Therefore if $\bar{f} > s_f^2$ or equivalently if $\chi^2 < K-1$ then $\hat{A}_{ML} = \infty$.

In the cell exchangeable case we can also use moment estimator for $A = K\alpha$. Let $\hat{\alpha}_E$ denote a moment estimator of α. In the cell exchangeable case

$$E(\sum_{i=1}^{K} f_i(f_i-1)) = n(n-1)\frac{\alpha+1}{K\alpha+1}.$$

Therefore

$$E(T) = \frac{\alpha+1}{K\alpha+1},$$

where

$$T = \frac{1}{n(n-1)} \sum_{i=1}^{K} f_i(f_i-1).$$

Solving

$$T = \frac{\hat{\alpha}_E + 1}{K\hat{\alpha}_E + 1}$$

we obtain a moment estimator

$$\frac{1}{\hat{\alpha}_E} = \frac{KT - 1}{1 - T}.$$

It can be easily shown that $0 < T < 1$ and the sign of $\hat{\alpha}_E$ depends on the sign of $KT - 1$. Note that $KT - 1$ and hence $1/\hat{\alpha}_E$ can be negative with positive probability, which is annoying. However in this case we should interpret $1/\hat{\alpha}_E = 0$ or $\hat{\alpha}_E = \infty$. Indeed B in (10) can be written as

$$B = n(n-1)(1 - KT)$$

and $1/\hat{\alpha}_E < 0$ corresponds to $\hat{A}_{ML} = K\hat{\alpha}_{ML} = \infty$.

In terms of s_f^2 of (11) the moment estimator $\hat{\alpha}_E$ can also be written as

$$\frac{n-1}{K\hat{\alpha}_E + 1} = \frac{K}{n} s_f^2 - 1.$$

If we approximate the left hand side by $n/(K\hat{\alpha}_E)$ then a modified moment estimator $\tilde{\alpha}_E$ is given by

$$\frac{1}{\tilde{\alpha}_E} = \frac{K}{n}\left(\frac{K}{n} s_f^2 - 1\right). \tag{12}$$

This is the estimator proposed by Bethlehem et al. (1990) for the gamma-Poisson model.

3 Gamma-Poisson model

In this section we investigate the relation between the gamma-Poisson model proposed by Bethlehem et al. (1990) and the Dirichlet-multinomial model of the previous section. We shall show that the Dirichlet-multinomial model is the conditional model when the random population size N is fixed in the gamma-Poisson model. Conversely the gamma-Poisson model is obtained by randomizing N of the Dirichlet-multinomial model with negative binomial distribution.

Let Π_i, $i = 1, \ldots, K$, be i.i.d. random variables from the gamma distribution Gamma(α, β) with the density

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}.$$

Given the realized value of $\pi_i = \Pi_i$, let F_i be distributed according to the Poisson distribution with mean parameter $N_0\pi_i$. F_i, $i = 1, \ldots, K$, are mutually independent.

In the gamma-Poisson model α and β are assumed to satisfy the following restriction

$$\alpha\beta = 1/K.$$

The population size $N = F_1 + \cdots + F_K$ is a random variable in this model and $N_0 = E(N)$ in the model specification is the expected population size[2].

[2]We make the notational distinction between N and its expectation N_0 for clarity.

By integrating out with respect to the gamma density, we see that F_i, $i = 1, \ldots, K$, are i.i.d. random variables having the following negative binomial distribution.

$$\begin{aligned} P(F_i = x) &= \int_0^\infty \frac{(N_0\pi)^x}{x!} e^{-N_0\pi} \frac{1}{\Gamma(\alpha)\beta^\alpha} \pi^{\alpha-1} e^{-\pi/\beta} d\pi \\ &= \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!} \frac{(N_0\beta)^x}{(1+N_0\beta)^{x+\alpha}}, \qquad x = 0, 1, \ldots . \end{aligned} \tag{13}$$

In Bethlehem et al. (1990) the parameters of the gamma distribution are taken as $(N_0\alpha, \beta/N_0)$ instead of (α, β) and the parameters of the negative binomial distribution in (13) have to be changed accordingly. We think that the parameterization of Bethlehem et al. (1990) is somewhat confusing and we prefer to use the parameterization in (13).

In summary the gamma-Poisson model is written as

$$p(F_1, \ldots, F_K \mid \beta) = \prod_{i=1}^K \frac{\Gamma(\alpha+F_i)}{\Gamma(\alpha)F_i!} \frac{(N_0\beta)^{F_i}}{(1+N_0\beta)^{F_i+\alpha}},$$

where K, N_0 are constants and $\alpha = 1/(\beta K)$.

Note that the negative binomial distribution is closed under convolution. Therefore N is again a negative binomial random variable with probability mass function

$$P(N = x) = \frac{\Gamma(K\alpha+x)}{\Gamma(K\alpha)x!} \frac{(N_0\beta)^x}{(1+N_0\beta)^{x+K\alpha}}. \tag{14}$$

Now we show that this model can be obtained by randomizing N of the Dirichlet-multinomial model with the negative binomial distribution in (14). Consider the cell exchangeable model (8). In (4) N is a constant and hence (4) can be regarded as the conditional probability mass function given N. Multiplying (4) by (14) we obtain the unconditional probability mass function of $\boldsymbol{F} = (F_1, \ldots, F_K)$ as

$$\begin{aligned} p(F_1, \ldots, F_K) &= \frac{N!\Gamma(K\alpha)}{\Gamma(K\alpha+N)} \frac{\Gamma(\alpha+F_1)}{\Gamma(\alpha)F_1!} \cdots \frac{\Gamma(\alpha+F_K)}{\Gamma(\alpha)F_K!} \\ &\qquad \times \frac{\Gamma(K\alpha+N)}{\Gamma(K\alpha)N!} \frac{(N_0\beta)^N}{(1+N_0\beta)^{N+K\alpha}} \\ &= \frac{\Gamma(\alpha+F_1)}{\Gamma(\alpha)F_1!} \cdots \frac{\Gamma(\alpha+F_K)}{\Gamma(\alpha)F_K!} \frac{(N_0\beta)^N}{(1+N_0\beta)^{N+K\alpha}} \\ &= \frac{\Gamma(\alpha+F_1)}{\Gamma(\alpha)F_1!} \cdots \frac{\Gamma(\alpha+F_K)}{\Gamma(\alpha)F_K!} \frac{(N_0\beta)^{F_1+\cdots+F_K}}{(1+N_0\beta)^{F_1+\cdots+F_K+K\alpha}} \\ &= \prod_{i=1}^K \frac{\Gamma(\alpha+F_i)}{\Gamma(\alpha)F_i!} \frac{(N_0\beta)^{F_i}}{(1+N_0\beta)^{F_i+\alpha}}. \end{aligned} \tag{15}$$

Therefore F_i, $i = 1, \ldots, K$, are i.i.d. random variables with the negative binomial distribution given in (13).

Conversely, if we divide (15) by (14) we immediately obtain the Dirichlet-multinomial model as a conditional model from the gamma-Poisson model with N fixed. We summarize our result in the following theorem.

Theorem 1 *Let $N = F_1 + \cdots + F_K$ be fixed in the gamma-Poisson model, then the conditional model is the Dirichlet-multinomial model with $\alpha = \alpha_i$, $i = 1, \ldots, K$. Conversely the gamma-Poisson model is obtained from the Dirichlet-multinomial model with $\alpha = \alpha_i$, $i = 1, \ldots, K$, by randomizing N with the negative binomial distribution in (14).*

We can give an alternative explanation of this theorem as follows. Let G_i, $i = 1, \ldots, K$, be independently distributed according to the the Poisson distribution Poisson($N_0\pi_i$). It is well known that given $N = G_1 + \cdots + G_K$ the conditional distribution of $(G_1, \ldots, G_K)$ is the multinomial distribution Mult($N, p_1, \ldots, p_K$), where

$$p_i = \frac{N_0\pi_i}{\sum_j N_0\pi_j} = \frac{\pi_i}{\sum_j \pi_j}, \qquad i = 1, \ldots, K.$$

Now let π_i, $i = 1, \ldots, K$, be distributed according to gamma distribution Gamma(α_i, β), $i = 1, \ldots, K$. Then $(p_1, \ldots, p_K) = (\pi_1/\sum_j \pi_j, \ldots, \pi_K/\sum_j \pi_j)$ has the Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_K)$. From this consideration it becomes almost obvious that Theorem 1 holds.

Note that in the above explanation the scale parameter β of the gamma distribution is irrelevant because the distribution of G_i, $i = 1, \ldots, K$, depends only on the ratios $\pi_i/\sum_j \pi_j$, $i = 1, \ldots, K$. In the gamma-Poisson model $1/\beta = K\alpha$ actually represents $\sum_j \alpha_j$ for the cell exchangeable case.

We now consider simple random sampling of size n (n fixed) without replacement from the finite population generated by the gamma-Poisson model. Concerning the marginal distribution of the sample frequency vector $\boldsymbol{f} = (f_1, \ldots, f_K)$ the following corollary follows immediately from Theorem 1.

Corollary 1 *Let $\boldsymbol{f} = (f_1, \ldots, f_K)$ be the sample frequency vector obtained by simple random sampling of size n without replacement taken from the finite population generated by the gamma-Poisson model. Then $\boldsymbol{f}$ has the Dirichlet-multinomial distribution with parameter $\alpha = \alpha_i$, $i = 1, \ldots, K$, i.e., its probability mass function is*

$$p(f_1, \ldots, f_K) = \frac{n!\Gamma(K\alpha)}{\Gamma(K\alpha + n)} \frac{\Gamma(\alpha + f_i)}{\Gamma(\alpha) f_i!} \cdots \frac{\Gamma(\alpha + f_K)}{\Gamma(\alpha) f_K!}. \tag{16}$$

Proof. By Theorem 1 for each realized value of N the conditional model of the gamma-Poisson model is the Dirichlet-multinomial model. Now the sample frequency vector from the Dirichlet-multinomial population has the Dirichlet-multinomial distribution (16). Note that (16) does not depend on the value of N. Therefore the unconditional distribution of the sample frequency vector is given by (16). Q.E.D.

By this corollary we see that the estimation of the gamma-Poisson model is exactly the same as the cell exchangeable Dirichlet-multinomial model. In particular the moment estimator (12) of Bethlehem et al. (1990) can be used to estimate the hyperparameter α.

Remark 1 *In Section 6 of Bethlehem et al. (1990) it is claimed that the sample variance s_f^2 of (12) has to be calculated using only nonempty cells (i.e. non-zero f_i's). This does not seem to be correct.*

4 Ewens model

In this section we investigate the relation between the Dirichlet-multinomial distribution and the multivariate Ewens distribution frequently used in genetics. The multivariate Ewens distribution has been investigated in detail by M. Sibuya. See Sibuya (1992), Sibuya (1993a), Sibuya (1993b), Sibuya and Yamato (1995). A related model was already given by Taga and Isii (1959).

The Ewens model is a stochastic clustering model which describes how N (distinguishable) individuals form random clusters of various sizes. The description and the notation of the model here follow that of Sibuya (1993b) and the reader is referred to Sibuya (1993b) for more detailed description.

Let $\mathcal{U}_n = \{1, \dots N\}$ and let A denote a partition of $\mathcal{U}$:

$$A = \{\{i, j, k, \dots\}, \{l, m, \dots\}, \dots\}, \qquad 1 \le i, j, k, \dots \le N$$

Let $S(A) = (S_1, \dots, S_N)$ denote the size index vector of the partition A. Unlike the multinomial model, empty cells do not make sense in stochastic clustering models and S_0 is not defined. Let u be the total number of clusters

$$u = S_1 + \cdots + S_N.$$

The multivariate Ewens distribution has one-dimensional parameter $\alpha > 0$ and the probability of the partition A is given by

$$p(A) = \frac{\alpha^u}{\alpha^{[N]}} \prod_{j=1}^{N} ((j-1)!)^{S_j}, \tag{17}$$

where $\alpha^{[N]} = \alpha(\alpha+1)\cdots(\alpha+N-1)$. Note that $p(A)$ depends only on the size index vector $S(A)$. $\alpha^{[N]}$ is the normalizing constant of the distribution and does not depend on A. Furthermore $\prod_{j=1}^{N}((j-1)!)^{S_j}$ does not involve the parameter α. Therefore u is the sufficient statistic of the distribution. Writing

$$\alpha^u = \exp(u \log \alpha)$$

we see that the multivariate Ewens distributions form a one-parameter exponential family. In particular maximum likelihood estimation is straightforward in the multivariate Ewens distribution as discussed in Sibuya (1992).

(17) gives the probability of partition A of N distinguishable individuals. If N individuals are indistinguishable, then we can only observe the size index of A and the probability of $(S_1, \dots, S_N)$ is shown to be

$$p(S_1, \dots, S_N) = \frac{\alpha^u}{\alpha^{[N]}} \frac{N!}{\prod_{j=1}^{N} j^{S_j} S_j!}. \tag{18}$$

Concerning the sampling, the idea of Lemma 1 holds for the Ewens model as well. Note that the individuals are exchangeable in the Ewens model. Therefore if we sample from the multivariate Ewens distribution using simple random sampling without replacement,

the marginal distribution of the sample is again the multivariate Ewens distribution with N replaced by n and the same parameter α.

Now consider moment estimation of α. The expected value of the number of the sample uniques s_1 is obtained as

$$E(s_1) = \alpha \frac{n}{\alpha + n - 1}.$$

If n is large compared to α then

$$E(s_1) \doteq \alpha.$$

Therefore a simple estimator of α is given by the number of the sample uniques

$$\hat{\alpha} = s_1.$$

With this $\hat{\alpha}$ the number of the population uniques is estimated as

$$\hat{S}_1 = \hat{\alpha} \frac{N}{\hat{\alpha} + N - 1} \doteq \hat{\alpha} = s_1.$$

Namely the estimate of the number of the population uniques is basically given by the number of the sample uniques. This is a remarkable feature of the Ewens model. Note that this simplicity might suggest the limitation of this model being one-parameter model.

A simple diagnostic check of the Ewens model can be given as follows. Sibuya (1993b) and Arratia et al. (1992) derived the asymptotic distribution of $S_1, S_2, \ldots, S_m$ as $N \to \infty$ in the following form.

Lemma 2 *Fix m and let $N \to \infty$. Then S_i, $i = 1, \ldots, m$, converge in distribution to independent Poisson variables with mean α/i, $i = 1, \ldots, m$.*

Lemma 2 is stated in terms of the population size indices. However clearly it holds also for the sample size indices $s_1, \ldots, s_m$ as $n \to \infty$. Therefore if the Ewens model holds, then the sample size indices $(s_1, s_2, s_3, \ldots)$ should decrease roughly in the proportion $(1, 1/2, 1/3, \ldots)$.

Now we will show that the Ewens model can be obtained from the Dirichlet-multinomial model by a limiting argument similar to the law of small numbers. Consider the cell exchangeable case and let $\alpha_i = \alpha$, $i = 1, \ldots, K$, in (4). Then as shown in Appendix B the probability mass function of the size index vector $(S_0, \ldots, S_N)$ is given by

$$\binom{K}{S_0, \ldots, S_N} \frac{N!\Gamma(A)}{\Gamma(A+N)} \prod_{j=0}^{N} \left(\frac{\Gamma(\alpha + j)}{\Gamma(\alpha) j!} \right)^{S_j} = \frac{N! K! \Gamma(A)}{\Gamma(A+N) S_0!} \prod_{j=1}^{N} \left(\frac{\Gamma(\alpha + j)}{\Gamma(\alpha) j!} \right)^{S_j} \frac{1}{S_j!}.$$

We now perform the following limiting operation. Let $A = K\alpha$ be fixed and let $K \to \infty$, $\alpha \to 0$. Ignoring S_0 we consider the limit of the marginal probability mass function of $(S_1, \ldots, S_N)$. $S_0 = K - (S_1 + \cdots + S_N) = K - u$ diverges to infinity as $K \to \infty$. Now as $K \to \infty$

$$\begin{aligned} K \frac{\Gamma(\alpha + j)}{\Gamma(\alpha) j!} &= K \frac{A/K(A/K + 1) \cdots (A/K + j - 1)}{j!} \\ &= A \frac{(A/K + 1) \cdots (A/K + j - 1)}{j!} \\ &\to A \frac{(j-1)!}{j!} = \frac{A}{j}. \end{aligned}$$

Furthermore

$$\frac{K!}{(K-u)!K^u} \to 1.$$

Therefore

$$\frac{N!K!\Gamma(A)}{\Gamma(A+N)(K-u)!K^u} \prod_{j=1}^{N} \left(K\frac{\Gamma(\alpha+j)}{\Gamma(\alpha)j!} \right)^{S_j} \frac{1}{S_j!} \to \frac{N!\Gamma(A)A^u}{\Gamma(A+N)} \prod_{j=1}^{N} \frac{1}{j^{S_j}S_j!}, \tag{19}$$

Comparing (19) and (18) we see that the right hand side of (19) is the probability mass function of the multivariate Ewens distribution with parameter A. Hence we have proved the following theorem.

Theorem 2 *Consider the Dirichlet-multinomial model with interchangeable cells $\alpha_i = \alpha$, $i = 1, \ldots, K$. Let $K\alpha = A$ be fixed and let $K \to \infty$, $\alpha \to 0$. Then the marginal distribution of $(S_1, \ldots, S_N)$ converges in distribution to the multivariate Ewens distribution with parameter A.*

In genetics this theorem is known in a different context. See Watterson (1976) and Section 5 of Ewens (1990).

The motivation behind Theorem 2 can be given using a version of Polya's urn model with continuous parameter. Let $\alpha_1, \ldots, \alpha_K$ be nonnegative real numbers and let $A = \alpha_1 + \cdots + \alpha_K$. Suppose that a ball of "color i" is observed (or a ball falls in the i-th cell) with probability α_i/A, $i = 1, \ldots, K$. Having observed color i, we replace α_i and A by

$$\alpha_i \to \alpha_i + 1, \qquad A \to A + 1.$$

With these replaced values of α_i's we observe the color of the second ball. We repeat the procedure until we observe N balls with colors $y_1, \ldots, y_N$. Here $y_j = i$ means that j'th ball has color i. Consider the probability of the particular vector $(y_1, \ldots, y_N)$ which has the frequency vector of the colors $(F_1, \ldots, F_K)$. It can be easily seen that this probability depends only on $(F_1, \ldots, F_K)$ and is given by

$$p(y_1, \ldots, y_N) = \frac{\prod_{i=1}^{K} \alpha_i(\alpha_i+1)\cdots(\alpha_i+F_i-1)}{A(A+1)\cdots(A+N-1)}.$$

Therefore the distribution of $(y_1, \ldots, y_N)$ is exchangeable with respect to the order of the balls and the probability of the frequency vector is given by

$$\begin{aligned} p(F_1, \ldots, F_K) &= \binom{N}{F_1, \ldots, F_K} \times \frac{\prod_{i=1}^{K} \alpha_i(\alpha_i+1)\cdots(\alpha_i+F_i-1)}{A(A+1)\cdots(A+N-1)} \\ &= \frac{N!}{A(A+1)\cdots(A+N-1)} \prod_{i=1}^{K} \frac{\alpha_i(\alpha_i+1)\cdots(\alpha_i+F_i-1)}{F_i!} \\ &= \frac{N!\Gamma(A)}{\Gamma(A+N)} \prod_{i=1}^{K} \frac{\Gamma(\alpha_i+F_i)}{\Gamma(\alpha_i)F_i!}. \end{aligned} \tag{20}$$

We see that this form of Polya's urn model is equivalent to the Dirichlet-multinomial distribution. Now consider the limiting case $K \to \infty, \alpha \to 0$ with $A = K\alpha$ fixed. We see that the urn model above corresponds exactly to the Ewens model described in Sibuya (1993b).

5 Some discussions

We have discussed mathematical properties of the Dirichlet-multinomial and the related models. The theory of these models are beautiful and these models are used in many fields. However these models might be too simple to fit the real populations. In fact Skinner and Holmes (1993) report that the gamma-Poisson model does not fit actual populations very well.

One problem is the exchangeability of the cells. The actual classifications used in microdata sets are determined based on their relevances in the population and exchangeability assumption seems to be a mathematical convenience. In the Dirichlet-multinomial model α_i's need not be all equal and we may be able to construct a better fitting model by allowing some variability in α_i's.

Other problem we have to consider is that the cells in the microdata sets are actually determined by cross classifications of many categories, whereas the theory developed in this paper treats the cells as essentially one-dimensional. If we consider cross-classified cells, we face the problem of modeling correlations among categories. This seems to be a very challenging topic for further investigation.

Appendix

A Bernoulli sampling for the gamma-Poisson model

We have so far considered the simple random sampling without replacement. Here we consider the Bernoulli sampling for the gamma-Poisson model because of its simplicity.

The Bernoulli sampling is defined in Section 2.2 and Section 3.2 of Särndal et al. (1992) as follows. Let a frame of N individuals of the population be given. Let $0 < p < 1$ be fixed and consider a coin with the probability of heads p. For each individual of the population we toss the coin and draw the individual if and only if the coin results in heads. In the Bernoulli sampling the sample size n is a random variable having the binomial distribution $\mathrm{Bin}(N, p)$. Given a value of n, Bernoulli sampling reduces to the simple random sampling without replacement.

Consider the Bernoulli sampling from a discrete population with the population frequency vector $\boldsymbol{F} = (F_1, \ldots, F_K)$. Then the sample frequencies f_i, $i = 1, \ldots, K$, are independent and have the Binomial distribution $\mathrm{Bin}(F_i, p)$, $i = 1, \ldots, K$. This independence of f_i's is the simplicity gained by using the Bernoulli sampling.

In the gamma-Poisson model F_i's are i.i.d. random variables having the negative binomial distribution in (13). Combined with the Bernoulli sampling we see that (F_i, f_i), $i = 1, \ldots, K$, are i.i.d random vectors. Hence f_i, $i = 1, \ldots, K$, are marginally i.i.d. random variables.

Let the sampling probability p in the Bernoulli sampling be determined by $p = n_0/N_0$, where n_0 is a predetermined expected sample size. Then it can be easily shown that the marginal distribution of f_i is the negative binomial distribution (13) with N_0 replaced by

n_0, i.e.,

$$P(f_i = x) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)x!} \frac{(n_0\beta)^x}{(1 + n_0\beta)^{x+\alpha}}, \qquad x = 0, 1, \ldots,$$

where $\alpha\beta = 1/K$. The first and the second order moments of f_i are given by

$$E(f_i) = \frac{n_0}{K}, \qquad \mathrm{Var}(f_i) = E(f_i) \cdot (1 + n_0\beta) = \frac{n_0}{K}(1 + n_0\beta) \tag{21}$$

In this i.i.d. case it is easy to estimate β. In particular from (21) a moment estimator of $n_0\beta = n_0/(K\alpha)$ based on the variance is given by

$$n_0\hat{\beta} = \frac{Ks_f^2}{n_0} - 1,$$

where s_f^2 is the sample variance of $\boldsymbol{f} = (f_1, \ldots, f_K)$ given in (11). Replacing N_0 and n_0 by the actual values N and n, we again obtain the estimator (12) by Bethlehem et al. (1990)

B Probability distribution of size index vector for cell exchangeable case

Suppose that we have an explicit expression $p_f(f_1, \ldots, f_K)$ of the probability mass function of the sample frequency vector $\boldsymbol{f} = (f_1, \ldots, f_K)$. Then the probability mass function of the sample size index vector $\boldsymbol{s}$ can be obtained by summing up $p_f(f_1, \ldots, f_K)$ for $\boldsymbol{f}$'s which have the same size index vector $\boldsymbol{s}$. However in general this summation can not be evaluated explicitly.

One simple case is when the probability distribution of $\boldsymbol{f}$ is exchangeable with respect to the cells, i.e.,

$$p_f(f_1, \ldots, f_K) = p_f(f_{i_1}, \ldots, f_{i_K})$$

for any permutation $(i_1, \ldots, i_K)$ of $(1, \ldots, K)$. In this case $p_f(f_1, \ldots, f_K)$ only depends on the size index vector $\boldsymbol{s} = (s_0, s_1, \ldots, s_n)$ and it remains to count the number of sample frequency vectors $\boldsymbol{f}$ which has the same size index vector. Consider a particular case

$$f_1 = \cdots = f_{s_0} = 0, \ f_{s_0+1} = \cdots = f_{s_0+s_1} = 1, \ f_{s_0+s_1+1} = \cdots = f_{s_0+s_1+s_2} = 2, \ldots.$$

The number of ways to place s_i i's, $i = 1, \ldots, n$, into K positions is

$$\binom{K}{s_0, \ldots, s_n}.$$

Therefore the probability mass function of the size index vector for the cell exchangeable case is given by

$$p_s(s_0, \ldots, s_n) = \binom{K}{s_0, \ldots, s_n} p_f(\underbrace{0, \ldots, 0}_{s_0}, \underbrace{1, \ldots, 1}_{s_1}, \underbrace{2, \ldots, 2}_{s_2}, \ldots,)$$

For example in the case of the multinomial distribution with equal cell probability $(1/K)$ we have

$$p_f(\underbrace{0,\ldots,0}_{s_0},\underbrace{1,\ldots,1}_{s_1},\underbrace{2,\ldots,2}_{s_2},\ldots,) = \left(\frac{1}{K}\right)^n \frac{n!}{\prod_{i=1}^n (i!)^{s_i}}.$$

Therefore

$$\begin{aligned} p_s(s_0,\ldots,s_n) &= \binom{K}{s_0,\ldots,s_n}\left(\frac{1}{K}\right)^n \frac{n!}{\prod_{i=1}^n (i!)^{s_i}} \\ &= \frac{K!n!}{K^n \prod_{i=1}^n (i!)^{s_i} s_i!}. \end{aligned}$$

References

[1] Arratia, R, Barbour, A.D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.*, **2**, 519–535.

[2] Bethlehem,J.G., Keller,W.J. and Pannekoek,J., (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

[3] Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London.

[4] Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *Ann. Probab.*, **8**, 745–764.

[5] Ewens, W.J. (1990). Population genetics theory – the past and the future. in *Mathematical and Statistical Development of Evolutionary Theory*, S. Lessard ed., 177–227, Kluwer, Dordrecht.

[6] Feller, W. (1968). *An Introduction to Probability Theory and its Applications.* 3rd ed., Volume 1. Wiley. New York.

[7] Good, I.J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, Massachusetts.

[8] Janardan, K. G. (1973). Chance mechanisms for multivariate hypergeometric models. *Sankhya*, Series A, **35**, 465–478

[9] Janardan, K.G. (1976). Certain estimation problems for multivariate hypergeometric models. *Ann. Inst. Statist. Math*, **28**, 429–444.

[10] Janardan, K. G. and Patil, G. P. (1972). A unified approach for a class of multivariate hypergeometric models. *Sankhya*, Series A, **34**, 363–376

[11] Johnson, N.L., Kots, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* John Wiley. New York.

[12] Korwar, R.M. (1988). On the observed number of classes from multivariate power series and hypergeometric distributions. *Sankhya*, Series B, **50**, 39–59

[13] Levin, B. and Reeds, J. (1977). Compound multinomial likelihood functions are unimodal: Proof of a conjecture of I. J. Good. *Annals of Statistics*, **5**, 79–87.

[14] Maritz, J.S. and Lwin, T. (1989). *Empirical Bayes Methods.* 2nd ed., Chapman and Hall, London.

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

[16] Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer, New York.

[17] Sibuya, M. (1992). A cluster-number distribution and its application to the analysis of homonyms. *Japan. J. Appl. Statist.*, **20**, 139–153 (in Japanese).

[18] Sibuya, M. (1993a). Random partitions of a finite set by cycles of permutations, *Japan. J. Indust. Appl. Math.*, **10**, 69–84.

[19] Sibuya, M. (1993b). A random clustering process. *Ann. Inst. Stat. Math.*, **45**, 459–465.

[20] Sibuya, M. and H. Yamato (1995). Characterization of some random partitions. *Japan J. Indust. Appl. Math.*, **12**, 237–263.

[21] Skinner, C.J. and Holmes, D.J. (1993). Modelling population uniques. in *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin.

[22] Taga Y. and Isii, K. (1959). On a stochastic model concerning the pattern of communication - Diffusion of news in a social group - . *Ann. Inst. Statist. Math.*, **11**, 25–43.

[23] Watterson, G.A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *J. Appl. Probab.*, **13**, 639–651.

[24] Willenborg, L. and de Wall, T. (1996). *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics 111, Springer, New York.